1

From Sequence to Structure

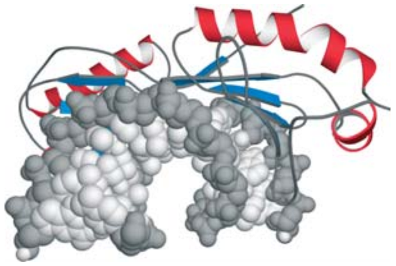
The genomics revolution is providing gene sequences in exponentially increasing numbers. Converting this sequence information into functional information for the gene products coded by these sequences is the challenge for post-genomic biology. The first step in this process will often be the interpretation of a protein sequence in terms of the three-dimensional structure into which it folds. This chapter summarizes the basic concepts that underlie the relationship between sequence and structure and provides an overview of the architecture of proteins.

1-0 Overview: Protein Function and Architecture

Binding

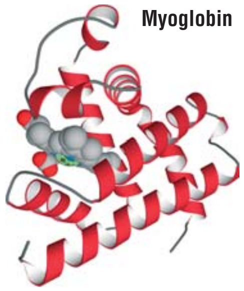
Specific recognition of other molecules is central to protein function. The molecule that is bound (the ligand) can be as small as the oxygen molecule that coordinates to the heme group of myoglobin, or as large as the specific DNA sequence (called the TATA box) that is bound—and distorted—by the TATA binding protein. Specific binding is governed by shape complementarity and polar interactions such as hydrogen bonding.

TATA binding protein



The TATA binding protein binds a specific DNA sequence and serves as the platform for a complex that initiates transcription of genetic information. (PDB 1tgh)

Myoglobin

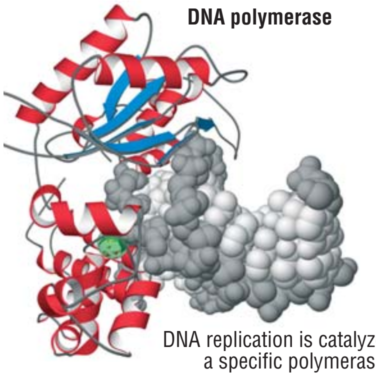


Myoglobin binds a molecule of oxygen reversibly to the iron atom in its heme group (shown in grey with the iron in green). It stores oxygen for use in muscle tissues. (PDB 1a6k)

Catalysis

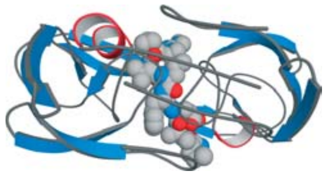
Essentially every chemical reaction in the living cell is catalyzed, and most of the catalysts are protein enzymes. The catalytic efficiency of enzymes is remarkable: reactions can be accelerated by as much as 17 orders of magnitude over simple buffer catalysis. Many structural features contribute to the catalytic power of enzymes: holding reacting groups together in an orientation favorable for reaction (proximity); binding the transition state of the reaction more tightly than ground state complexes (transition state stabilization); acid-base catalysis, and so on.

DNA polymerase



DNA replication is catalyzed by a specific polymerase that copies the genetic material and edits the product for errors in the copy. (PDB 1pbx)

HIV protease

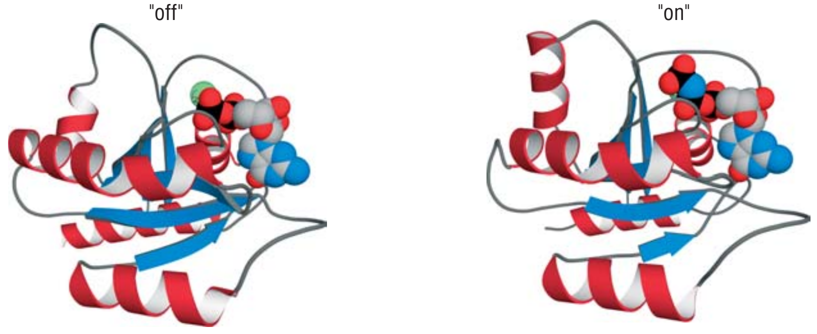


Replication of the AIDS virus HIV depends on the action of a protein-cleaving enzyme called HIV protease. This enzyme is the target for protease-inhibitor drugs (shown in grey). (PDB 1a8k)

Switching

Proteins are flexible molecules and their conformation can change in response to changes in pH or ligand binding. Such changes can be used as molecular switches to control cellular processes. One example, which is critically important for the molecular basis of many cancers, is the conformational change that occurs in the small GTPase Ras when GTP is hydrolyzed to GDP. The GTP-bound conformation is an "on" state that signals cell growth; the GDP-bound structure is the "off" signal.

Ras

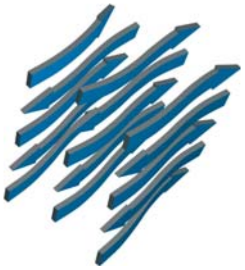


The GDP-bound ("off"; PDB 1pll) state of Ras differs significantly from the GTP-bound ("on"; PDB 121p) state. This difference causes the two states to be recognized by different proteins in signal transduction pathways.

Structural Proteins

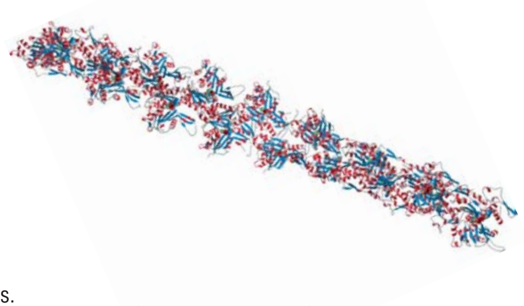
Protein molecules serve as some of the major structural elements of living systems. This function depends on specific association of protein subunits with themselves as well as with other proteins, carbohydrates, and so on, enabling even complex systems like actin fibrils to assemble spontaneously. Structural proteins are also important sources of biomaterials, such as silk, collagen, and keratin.

Silk



Silk derives its strength and flexibility from its structure: it is a giant stack of antiparallel beta sheets. Its strength comes from the covalent and hydrogen bonds within each sheet; the flexibility from the van der Waals interactions that hold the sheets together. (PDB 1slk)

F-actin



Actin fibers are important for muscle contraction and for the cytoskeleton. They are helical assemblies of actin and actin-associated proteins. (Courtesy of Ken Holmes)

Figure 1-1 Four examples of biochemical functions performed by proteins

Proteins are the most versatile macromolecules of the cell

This book is concerned with the functions that proteins perform and how these are determined by their structures. “Protein function” may mean the biochemical function of the molecule in isolation, or the cellular function it performs as part of an assemblage or complex with other molecules, or the phenotype it produces in the cell or organism.

Major examples of the biochemical functions of proteins include binding; catalysis; operating as molecular switches; and serving as structural components of cells and organisms (Figure 1-1). Proteins may bind to other macromolecules, such as DNA in the case of DNA polymerases or gene regulatory proteins, or to proteins in the case of a transporter or a receptor that binds a signaling molecule. This function exploits the ability of proteins to present structurally and chemically diverse surfaces that can interact with other molecules with high specificity. Catalysis requires not only specific binding, to substrates and in some cases to regulatory molecules, but also specific chemical reactivity. Regulated enzymes and switches, such as the signaling G proteins (which are regulated enzymes that catalyze the hydrolysis of GTP), require large-scale conformational changes that depend on a delicate balance between structural stability and flexibility. Structural proteins may be as strong as silk or as tough and durable as keratin, the protein component of hair, horn and feathers; or they may have complex dynamic properties that depend on nucleotide hydrolysis, as in the case of actin and tubulin. This extraordinary functional diversity and versatility of proteins derives from the chemical diversity of the side chains of their constituent amino acids, the flexibility of the polypeptide chain, and the very large number of ways in which polypeptide chains with different amino acid sequences can fold.

There are four levels of protein structure

Proteins are polymers of 20 different amino acids joined by peptide bonds. At physiological temperatures in aqueous solution, the polypeptide chains of proteins fold into a form that in most cases is globular (see Figure 1-2c). The sequence of the different amino acids in a protein, which is directly determined by the sequence of nucleotides in the gene encoding it, is its *primary structure* (Figure 1-2a). This in turn determines how the protein folds into higher-level structures. The *secondary structure* of the polypeptide chain can take the form either of alpha helices or of beta strands, formed through regular hydrogen-bonding interactions between N–H and C=O groups in the invariant parts of the amino acids in the polypeptide **backbone** or main chain (Figure 1-2b). In the globular form of the protein, elements of either alpha helix, or beta sheet, or both, as well as loops and links that have no secondary structure, are folded into a *tertiary structure* (Figure 1-2c). Many proteins are formed by association of the folded chains of more than one polypeptide; this constitutes the *quaternary structure* of a protein (Figure 1-2d).

For a polypeptide to function as a protein, it must usually be able to form a stable tertiary structure (or *fold*) under physiological conditions. On the other hand, the demands of protein function require that the folded protein should not be too rigid. Presumably because of these constraints, the number of folds adopted by proteins, though large, is limited. Whether the limited number of folds reflects physical constraints on the number of stable folds, or simply the expedience of divergent evolution from an existing stable fold, is not known, but it is a matter of some practical importance: if there are many possible stable folds not represented in nature, it should be possible to produce completely novel proteins for industrial and medical applications.

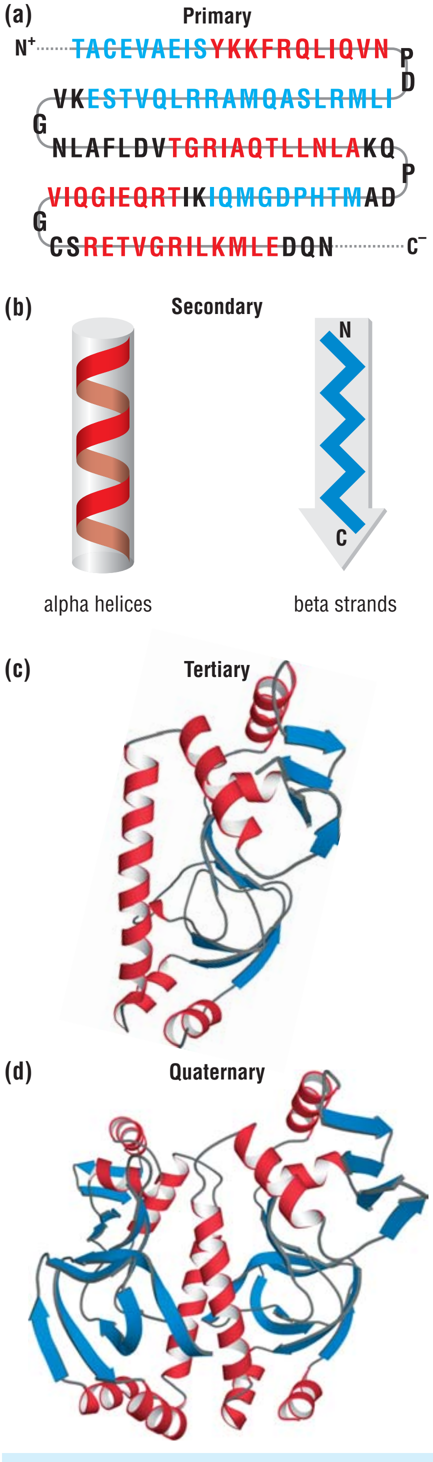


Figure 1-2 Levels of protein structure illustrated by the catabolite activator protein **(a)** The amino-acid sequence of a protein (primary structure) contains all the information needed to specify **(b)** the regular repeating patterns of hydrogen-bonded backbone conformations (secondary structure) such as alpha helices (red) and beta sheets (blue), as well as **(c)** the way these elements pack together to form the overall fold of the protein (tertiary structure) (PDB 2cgp). **(d)** The relative arrangement of two or more individual polypeptide chains is called quaternary structure (PDB 1cgp).

Definitions

backbone: the regularly repeating part of a polymer. In proteins it consists of the amide –N–H, alpha carbon –C–H and the carbonyl –C=O groups of each amino acid.

References

Alberts, B. *et al.*: *Molecular Biology of the Cell* 4th ed. Chapter 3 (Garland, New York, 2002).

Jansen, R. and Gerstein, M.: **Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins.** *Nucleic Acids Res.* 2000, **28**:1481–1488.

Michal, G., ed.: *Boehringer Mannheim Biochemical Pathways Wallcharts*, Roche Diagnostics Corporation, Roche Molecular Biochemicals, P.O. Box 50414, Indianapolis, IN 46250-0414, USA.

Voet, D. and Voet, J.G.: *Biochemistry* 2nd ed. Chapters 4 to 7 (Wiley, New York, 1995).

http://www.expasy.ch/cgi-bin/search-biochem-index (searchable links to molecular pathways and maps).

The chemical characters of the amino-acid side chains have important consequences for the way they participate in the folding and functions of proteins

The amino-acid **side chains** (Figure 1-3) have different tendencies to participate in interactions with each other and with water. These differences profoundly influence their contributions to protein stability and to protein function.

Hydrophobic amino-acid **residues** engage in *van der Waals* interactions only. Their tendency to avoid contact with water and pack against each other is the basis for the *hydrophobic effect*. Alanine and leucine are strong helix-favoring residues, while proline is rarely found in helices because its backbone nitrogen is not available for the hydrogen bonding required for helix formation. The aromatic side chain of phenylalanine can sometimes participate in weakly polar interactions.

Hydrophilic amino-acid residues are able to make *hydrogen bonds* to one another, to the peptide backbone, to polar organic molecules, and to water. This tendency dominates the interactions in which they participate. Some of them can change their charge state depending on their pH or the microenvironment. Aspartic acid and glutamic acid have pK_a values near 5 in aqueous solution, so they are usually unprotonated and negatively charged at pH 7. But in the hydrophobic interior of a protein molecule their pK_a may shift to 7 or even higher (the same effect occurs if a negative charge is placed nearby), allowing them to function as proton donors at physiological pH. The same considerations apply to the behavior of lysine, which has a pK_a greater than 10 in water and so is usually depicted as positively charged. But in a nonpolar environment, or in the presence of a neighboring positive charge, its pK_a can shift to less than 6, and the resulting neutral species can be a proton acceptor. Histidine is perhaps the most versatile of all the amino acids in this regard, which explains why it is also the residue most often found in enzyme active sites. It has two titratable –N–H groups, each with pK_a values around 6. When one of these –N–H groups loses a proton, however, the pK_a of the other one becomes much greater than 10. When both are protonated, the residue as a whole is positively charged. When only one is protonated (usually it is the one farthest from the main chain of the protein) the side chain is neutral and has the ability both to donate and to accept a proton. The fully deprotonated form is negatively charged, and occurs rarely. Arginine is always completely protonated at neutral pH; its positive charge is localized primarily at the carbon atom of the guanidium head. Serine, threonine, glutamine and asparagine do not ionize but are able both to donate and to accept hydrogen bonds simultaneously. Cysteine, like histidine, is commonly found in enzyme active sites, because the thiolate anion is the most powerful *nucleophile* available from the naturally occurring amino acids.

Amphipathic residues have both polar and nonpolar character, making them ideal for forming interfaces. It may seem surprising to consider the charged side chain of lysine as amphipathic, but its long hydrophobic region is often involved in van der Waals interactions with hydrophobic side chains. Tyrosine does not usually ionize at physiological pH (its pK_a is about 9) but in some enzyme active sites it can participate in acid-base reactions because the environment can lower this pK_a. The –O–H group is able both to donate and to accept hydrogen bonds, and the aromatic ring can form weakly polar interactions. Tryptophan behaves similarly, but the indole –N–H group does not ionize. Methionine is the least polar of the amphipathic amino acids, but the thioether sulfur is an excellent ligand for many metal ions.

Definitions

amphipathic: having both polar and nonpolar character and therefore a tendency to form interfaces between **hydrophobic** and **hydrophilic** molecules.

hydrophilic: tending to interact with water. Hydrophilic molecules are polar or charged and, as a consequence, are very soluble in water. In polymers, hydrophilic **side chains** tend to associate with other hydrophilic side chains, or with water molecules, usually by means of hydrogen bonds.

hydrophobic: tending to avoid water. Hydrophobic molecules are nonpolar and uncharged and, as a consequence, are relatively insoluble in water. In polymers, hydrophobic **side chains** tend to associate with each other to minimize their contact with water or polar side chains.

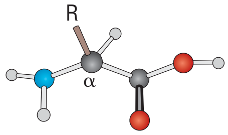
residue: the basic building block of a polymer; the fragment that is released when the bonds that hold the polymer segments together are broken. In proteins, the residues are the amino acids.

side chain: a chemical group in a polymer that protrudes from the repeating backbone. In proteins, the side chain, which is bonded to the alpha carbon of the backbone, gives each of the 20 amino acids its particular chemical identity.

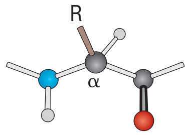
References

Creighton, T.E.: *Proteins: Structure and Molecular Properties* 2nd ed. Chapter 1 (Freeman, New York, 1993).

A website summarizing the physical-chemical properties of the standard amino acids may be found at: http://prowl.rockefeller.edu/aainfo/contents.htm



The chemical structure of an amino acid. The backbone is the same for all amino acids and consists of the amino group (NH_2), the alpha carbon and the carboxylic acid group (COOH). Different amino acids are distinguished by their different side chains, R. The neutral form of an amino acid is shown: in solution at pH 7 the amino and carboxylic acid groups ionize, to NH_3^+ and COO^-. Except for glycine, where R=H, amino acids are chiral (that is, they have a left–right asymmetry). The form shown is the L-configuration, which is most common.



An amino-acid residue as it is incorporated into a polypeptide chain. The R group is the side chain. The 20 different side chains that occur in proteins are depicted below. For proline, the side chain is fused back to the nitrogen of the backbone. The configuration about the alpha carbon is L for most amino acids in proteins.

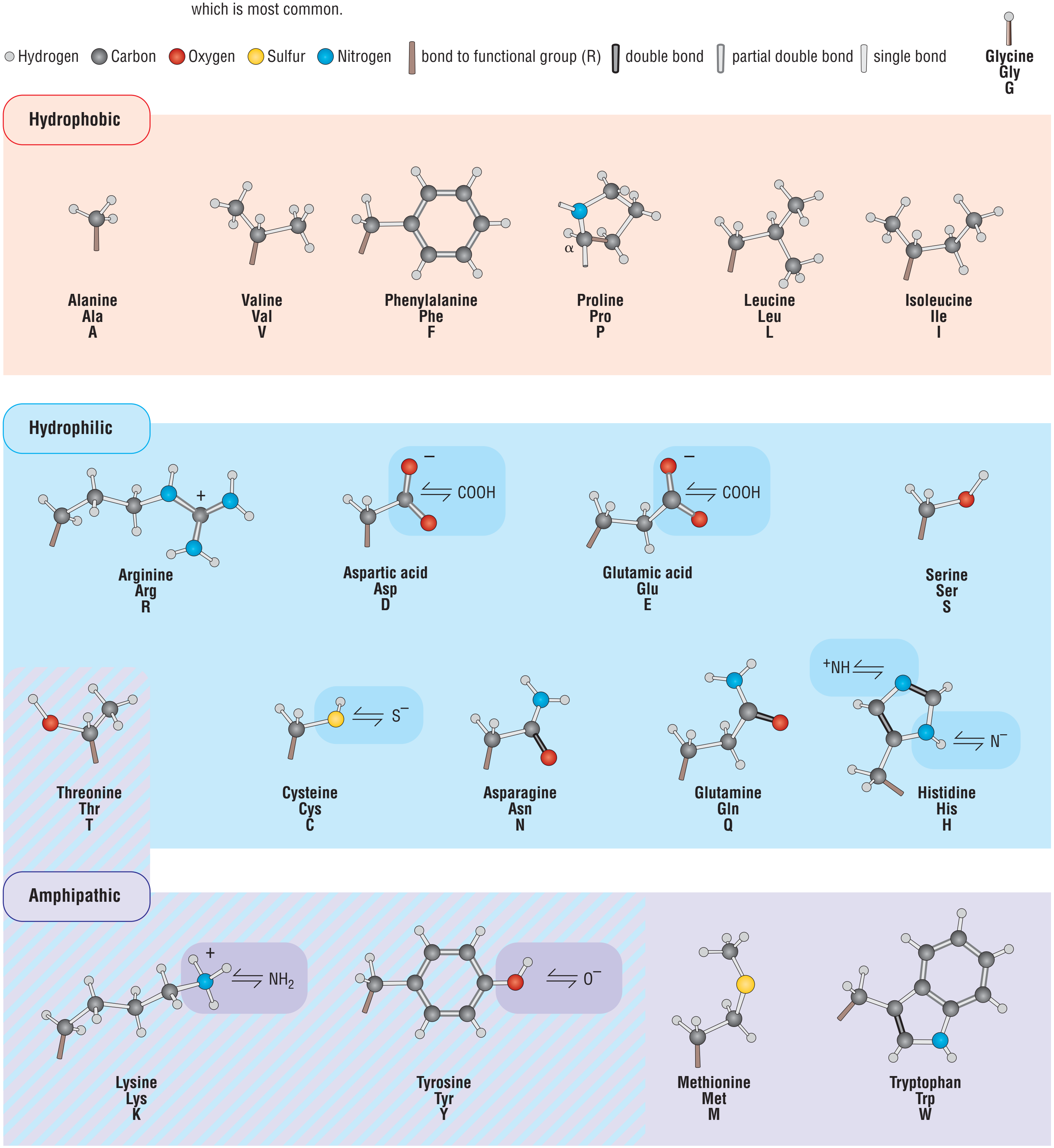


Figure 1-3 Amino-acid structure and the chemical characters of the amino-acid side chains Charged side chains are shown in the form that predominates at pH 7. For proline, the nitrogen and alpha carbon are shown because the side chain is joined to the nitrogen atom to form a ring that includes these atoms.

1-2 Genes and Proteins

	2nd position				
Ist position (5' end)	U	C	A	G	3rd position (3' end)
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Amino acids	Abbreviations		Codons
Alanine	Ala	A	GCA GCC GCG GCU
Cysteine	Cys	C	UGC UGU
Aspartic acid	Asp	D	GAC GAU
Glutamic acid	Glu	E	GAA GAG
Phenylalanine	Phe	F	UUC UUU
Glycine	Gly	G	GGA GGC GGG GGU
Histidine	His	H	CAC CAU
Isoleucine	Ile	I	AUA AUC AUU
Lysine	Lys	K	AAA AAG
Leucine	Leu	L	UUA UUG CUA CUC CUG CUU
Methionine	Met	M	AUG
Asparagine	Asn	N	AAC AAU
Proline	Pro	P	CCA CCC CCG CCU
Glutamine	Gln	Q	CAA CAG
Arginine	Arg	R	AGA AGG CGA CGC CGG CGU
Serine	Ser	S	AGC AGU UCA UCC UCG UCU
Threonine	Thr	T	ACA ACC ACG ACU
Valine	Val	V	GUA GUC GUG GUU
Tryptophan	Trp	W	UGG
Tyrosine	Tyr	Y	UAC UAU

Figure 1-4 The genetic code Each of the 64 possible three-base codons codes for either an amino acid or a signal for the end of the coding portion of a gene (a stop codon). Amino acids shaded pink have nonpolar (hydrophobic) side chains; those shaded blue have polar or charged side chains. Those shaded mauve are amphipathic. Glycine has no side chain. Almost all of the amino acids can be specified by two or more different codons that differ only in the third position in the codon. Single-base changes elsewhere in the codon usually produce a different amino acid but with similar physical-chemical properties.

There is a linear relationship between the DNA base sequence of a gene and the amino-acid sequence of the protein it encodes

The **genetic code** is the formula that converts hereditary information from genes into proteins. Every amino acid in a protein is represented by a **codon** consisting of three consecutive **nucleotides** in the gene. DNA contains four different nucleotides, with the **bases** adenine (A), guanine (G), thymidine (T) and cytosine (C), whose sequence in a gene spells out the sequence of the amino acids in the protein that it specifies: this is the **primary structure** of the protein. The nucleotide sequence of the DNA is **transcribed** into **messenger RNA** (**mRNA**), with uridine (U) replacing thymine (T). Figure 1-4 shows the correspondence between the 64 possible three-base codons in mRNA and the 20 naturally occurring amino acids. Some amino acids are specified by only one codon, whereas others can be specified by as many as six different codons: the genetic code is **degenerate**. There are three codons that do not code for amino acids, but signal the termination of the polypeptide chain (**stop codons**). The process by which the nucleotide sequence of the DNA is first transcribed into RNA and then **translated** into protein is outlined in Figure 1-5.

In bacteria and other lower organisms, the relationship between the base sequence of the gene and the amino acid sequence of the corresponding protein is strictly linear: the protein sequence can be read directly from the gene sequence (Figure 1-5 left-hand side). In higher organisms, however, genes are typically segmented into coding regions (**exons**) that are interrupted by non-coding stretches (**introns**). These non-coding introns are transcribed into RNA, but are enzymatically excised from the resulting transcript (the **primary transcript**), and the exons are then spliced together to make the mature mRNA (Figure 1-5 right-hand side).

The process of intron removal and exon ligation has been exploited in the course of evolution through **alternative splicing**, in which exon segments as well as intron segments may be differentially excised from the primary transcript to give more than one mRNA and thus more than one protein. Depending on the arrangement of the introns, alternative splicing can lead to truncated proteins, proteins with different stretches of amino acids in the middle, or frameshifts in which the sequence of a large part of the protein is completely different from that specified by an in-frame reading of the gene sequence. Coding sequences can also be modified by **RNA editing**. In this process, some nucleotides are changed to others, and stretches of additional nucleotides can be inserted into the mRNA sequence before translation occurs. Modification of the coding sequences by RNA processing in these ways complicates the interpretation of genomic sequences in terms of protein structure, though this complication does not apply to cDNA sequences, which are artificially copied by reverse transcription from mRNA.

The organization of the genetic code reflects the chemical grouping of the amino acids

The amino acids fall into groups according to their physical-chemical properties (see Figure 1-3). The organization of the genetic code reflects this grouping, as illustrated in Figure 1-4. Note that single-base changes (**single-nucleotide polymorphism**) in the third position in a codon will often produce the same amino acid. Single-base changes elsewhere in the codon will usually produce a different amino acid, but with the same physical-chemical properties: for example, the second base specifies if the amino acid is polar or hydrophobic. Changes of this sort are known as **conservative substitutions** and when they are found in comparisons of protein sequences they are taken to indicate conservation of structure between two proteins. Examination of protein sequences for the same gene product over a large evolutionary distance

Definitions

alternative splicing: the selection of different coding sequences from a gene by the removal during RNA processing of portions of the RNA containing or affecting coding sequences.

base: the aromatic group attached to the sugar of a **nucleotide**.

codon: three consecutive **nucleotides** in a strand of DNA or RNA that represent either a particular amino acid or a signal to stop translating the transcript of the gene.

conservative substitution: replacement of one amino acid by another that has similar chemical and/or physical properties.

degenerate: having more than one **codon** for an amino acid.

exon: coding segment of a gene (compare **intron**).

genetic code: the relationship between each of the 64 possible three-letter combinations of A, U (or T), G and C and the 20 naturally occurring amino acids that make up proteins.

intron: noncoding DNA within a gene.

messenger RNA (mRNA): the RNA molecule transcribed from a gene after removal of **introns** and editing.

nucleotide: the basic repeating unit of a nucleic acid polymer. It consists of a **base** (A, U [in RNA, T in DNA], G or C), a sugar (ribose in RNA, deoxyribose in DNA) and a phosphate group.

primary structure: the amino-acid sequence of a polypeptide chain.

primary transcript: the RNA molecule directly

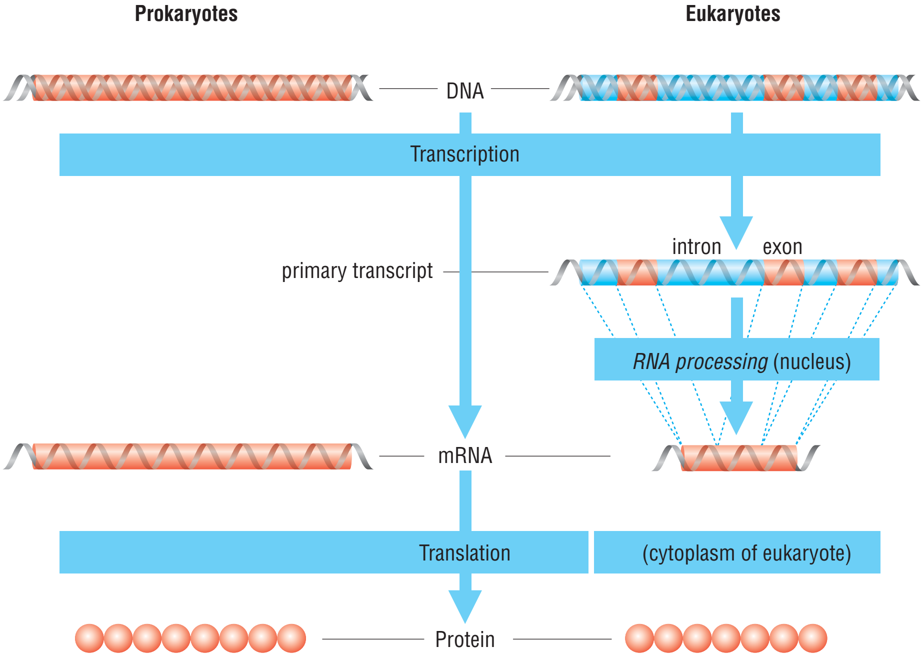


Figure 1-5 The flow of genetic information in prokaryotes (left) and eukaryotes (right) The amino-acid sequences of proteins are coded in the base sequence of DNA. This information is transcribed into a complementary base sequence in messenger RNA (mRNA). In prokaryotes, the mRNA is generated directly from the DNA sequence (left-hand side of diagram). Eukaryotic genes (right-hand side) are often interrupted by one or more noncoding intervening segments called introns. These are transcribed along with the exons to produce a primary transcript, from which the introns iare excised in the nucleus and the coding segments, the exons, joined together to generate the mRNA. Finally, the mRNA base sequence is translated into the corresponding amino-acid sequence on the ribosome, a process that occurs in the cytoplasm of eukaryotic cells. (Diagram not to scale.)

illustrates this principle (Figure 1-6). An amino acid that is altered from one organism to another in a given position in the protein sequence is most often changed to a residue of similar physical-chemical properties, exactly as predicted by the organization of the code.

	Gly	Ala	Val	Leu	Ile	Met	Cys	Ser	Thr	Asn	Gln	Asp	Glu	Lys	Arg	His	Phe	Tyr	Trp	Pro
Gly																				
Ala	58																			
Val	10	37																		
Leu	2	10	30																	
Ile		7	66	25																
Met	1	3	8	21	6															
Cys	1	3	3		2															
Ser	45	77	4	3	2	2	12													
Thr	5	59	19	5	13	3	1	70												
Asn	16	11	1	4	4			43	17											
Gln	3	9	3	8	1	2		5	4	5										
Asp	16	15	2		1			10	6	53	8									
Glu	11	27	4	2	4	1		9	3	9	42	83								
Lys	6	6	2	4	4	9		17	20	32	15		10							
Arg	1	3	2	2	3	2	1	14	2	2	12	9		48						
His	1	2	3	4			1	3	1	23	24	4	2	2	10					
Phe	2	2	1	17	9	2		4	1	1					1	2				
Tyr		2	2	2	1		3	2	2	4			1	1		4	26			
Trp				1				2							3		1	1		
Pro	5	35	5	4	1		1	27	7	3	9	1	4	4	7	5	1			

Figure 1-6 Table of the frequency with which one amino acid is replaced by others in amino-acid sequences of the same protein from different organisms The larger the number, the more common a particular substitution. For example, glycine is commonly replaced by alanine and vice versa; this makes chemical sense because these are the amino acids with the smallest side chains. Similarly, aspartic acid and glutamic acid, the two negatively charged residues, frequently substitute for one another. There are some surprises: for example, serine and proline often substitute for each other, as do glutamic acid and alanine. Serine may substitute for proline because the side-chain OH can receive a hydrogen bond from its own main-chain NH, mimicking the fused ring of proline.

transcribed from a gene, before processing.

RNA editing: enzymatic modification of the RNA **base** sequence.

single-nucleotide polymorphism (SNP): a mutation of a single **base** in a **codon**.

stop codon: a **codon** that signals the end of the coding sequence and usually terminates **translation**.

transcription: the synthesis of RNA from the coding strand of DNA by DNA-dependent RNA polymerase.

translation: the transfer of genetic information from the sequence of **codons** in **mRNA** into a sequence of amino acids in a polypeptide chain.

References

Alberts, B. *et al.*: *Molecular Biology of the Cell* 4th ed. Chapters 3 and 6 (Garland, New York, 2002).

Argyle, E.: **A similarity ring for amino acids based on their evolutionary substitution rates.** *Orig. Life* 1980, **10**:357–360.

Dayhoff, M.O. *et al.*: **Establishing homologies in protein sequences.** *Methods Enzymol.* 1983, **91**:524–545.

Jones, D.T. *et al.*: **The rapid generation of mutation data matrices from protein sequences.** *Comput. Appl. Biosci.* 1992, **8**:275–282.

Topham, C.M. *et al.*: **Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables.** *J. Mol. Biol.* 1993, **229**:194–220.

1-3 The Peptide Bond

Proteins are linear polymers of amino acids connected by amide bonds

Amino acids are crucial components of living cells because they are easy to polymerize. α-Amino acids are preferable to β-amino acids because the latter are too flexible to form spontaneously folding polymers. The amino acids of a protein chain are covalently joined by amide bonds, often called **peptide bonds**: for this reason, proteins are also known as **polypeptides**. Proteins thus have a repeating **backbone** from which 20 different possible kinds of side chains protrude (see Figure 1-8). On rare occasions, nonstandard side chains are found. In plants, a significant number of unusual amino acids have been found in proteins. In mammals, however, they are largely confined to small hormones. Sometimes, post-translational modification of a conventional amino acid may convert it into a nonstandard one. Examples are the non-enzymatic carbamylation of lysine, which can produce a metal-ion ligand, thereby activating an enzyme; and the deamidation of asparagine, which alters protein stability and turnover rate.

Chemically, the peptide bond is a covalent bond that is formed between a carboxylic acid and an amino group by the loss of a water molecule (Figure 1-7). In the cell, the synthesis of peptide bonds is an enzymatically controlled process that occurs on the ribosome and is directed by the mRNA template. Although peptide bond formation can be reversed by the addition of water (**hydrolysis**), **amide bonds** are very stable in water at neutral pH, and the hydrolysis of peptide bonds in cells is also enzymatically controlled.

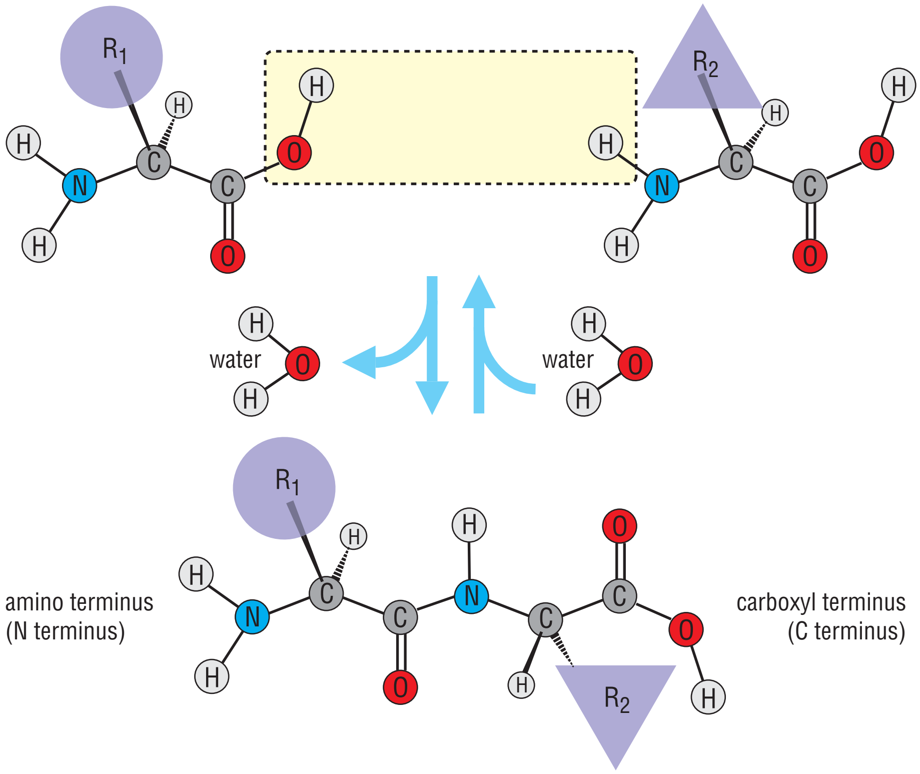


Figure 1-7 Peptide bond formation and hydrolysis Formation (top to bottom) and hydrolysis (bottom to top) of a peptide bond requires, conceptually, loss and addition, respectively, of a molecule of water. The actual chemical synthesis and hydrolysis of peptide bonds in the cell are enzymatically controlled processes that in the case of synthesis nearly always occurs on the ribosome and is directed by an mRNA template. The end of a polypeptide with the free amino group is known as the amino terminus (N terminus), that with the free carboxyl group as the carboxyl terminus (C terminus).

Definitions

amide bond: a chemical bond formed when a carboxylic acid condenses with an amino group with the expulsion of a water molecule.

backbone: the repeating portion of a **polypeptide** chain, consisting of the N–H group, the alpha-carbon C–H group, and the C=O of each amino-acid residue. Residues are linked to each other by means of **peptide bonds**.

dipole moment: an imaginary vector between two separated charges that may be full or partial. Molecules or functional groups having a dipole moment are said to be polar.

hydrolysis: breaking a covalent bond by addition of a molecule of water.

peptide bond: another name for **amide bond**, a chemical bond formed when a carboxylic acid condenses with an amino group with the expulsion of a water molecule. The term peptide bond is used only when both groups come from amino acids.

phi torsion angle: see **torsion angle**.

polypeptide: a polymer of amino acids joined together by **peptide bonds**.

psi torsion angle: see **torsion angle**.

resonance: delocalization of bonding electrons over more than one chemical bond in a molecule. Resonance greatly increases the stability of a molecule. It can be represented, conceptually, as if the properties of the molecule were an average of several structures in which the chemical bonds differ.

The properties of the peptide bond have important effects on the stability and flexibility of polypeptide chains in water

The properties of the amide bond account for several important properties of polypeptide chains in water. The stability of the peptide bond, as well as other properties important for the behavior of polypeptides, is due to **resonance**, the delocalization of electrons over several atoms. Resonance has two other important consequences. First, it increases the polarity of the peptide bond: the **dipole moment** of each peptide bond is shown in Figure 1-8. The polarity of the peptide bond can make an important contribution to the behavior of folded proteins, as discussed later in section 1-6.

Second, because of resonance, the peptide bond has partial double-bond character, which means that the three non-hydrogen atoms that make up the bond (the carbonyl oxygen O, the carbonyl carbon C and the amide nitrogen N) are coplanar, and that free rotation about the bond is limited (Figure 1-9). The other two bonds in the basic repeating unit of the polypeptide backbone, the N–C_α and C_α–C bonds (where C_α is the carbon atom to which the side chain is attached), are single bonds and free rotation is permitted about them provided there is no steric interference from, for example, the side chains. The angle of the N–C_α bond to the adjacent peptide bond is known as the **phi torsion angle**, and the angle of the C–C_α bond to the adjacent peptide bond is known as the **psi torsion angle** (see Figure 1-9). Thus a protein is an unusual kind of polymer, with rotatable covalent bonds alternating with rigid planar ones. This combination greatly restricts the number of possible conformations that a polypeptide chain can adopt and makes it possible to determine from simple steric considerations the most likely backbone conformation angles for polypeptide residues other than glycine.

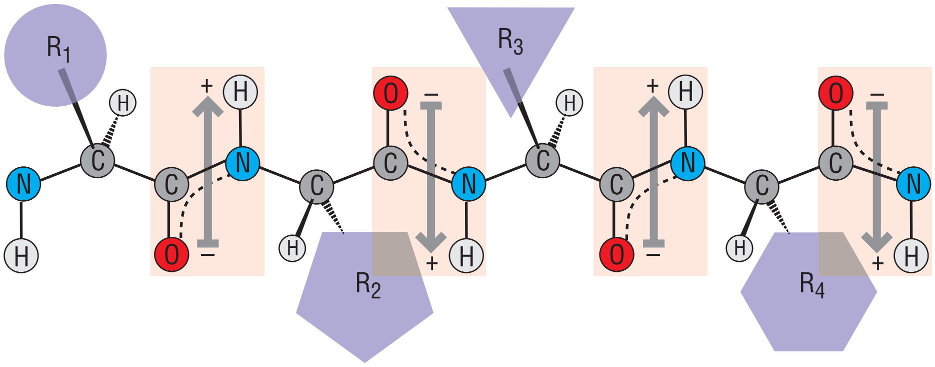


Figure 1-8 Schematic diagram of an extended polypeptide chain The repeating backbone is shown, with schematized representations of the different side chains (R_1, R_2 and so on). Each peptide bond is shown in a shaded box. Also shown are the individual dipole moments (arrows) associated with each bond. The dashed lines indicate the resonance of the peptide bond.

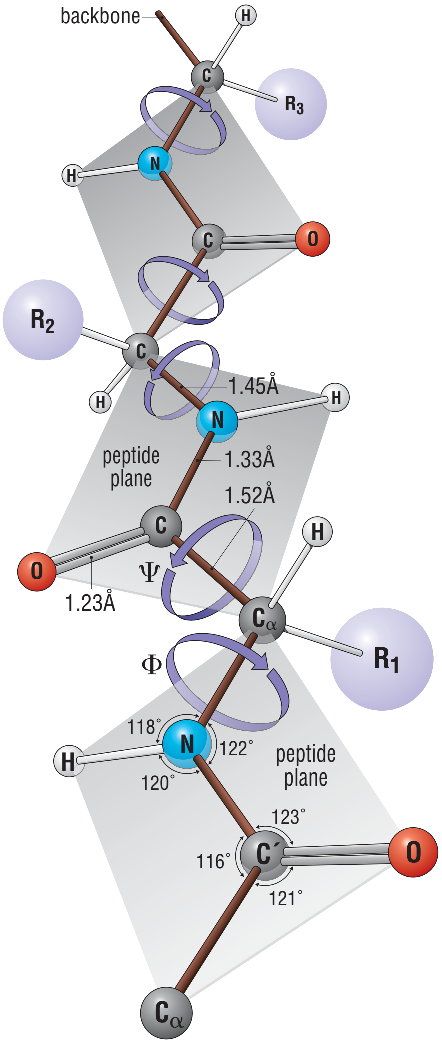


Figure 1-9 Extended polypeptide chain showing the typical backbone bond lengths and angles The planar peptide groups are indicated as shaded regions and the backbone torsion angles are indicated with circular arrows, with the phi and psi torsion angles marked. The omega torsion angle about the C–N peptide bond is usually restricted to values very close to 180° (*trans*), but can be close to 0° (*cis*) in rare cases. X–H bond lengths are all about 1 Å.

torsion angle: the angle between two groups on either side of a rotatable chemical bond. If the bond is the C_α–N bond of a peptide backbone the torsion angle is called **phi**. If the bond is the C_α–C backbone bond, the angle is called **psi**.

References

Martin, R.B.: **Peptide bond characteristics**. *Met. Ions Biol. Syst.* 2001, **38**:1–23.

Pauling, L.C.: *The Nature of the Chemical Bond and the Structure of Molecules and Crystals* 3rd ed. Chapter 8 (Cornell Univ. Press, Ithaca, New York, 1960).

Voet, D. and Voet, J.G.: *Biochemistry* 2nd ed. (Wiley, New York, 1995), 67–68.

1-4 Bonds that Stabilize Folded Proteins

Folded proteins are stabilized mainly by weak noncovalent interactions

The amide bonds in the backbone are the only covalent bonds that hold the residues together in most proteins. In proteins that are secreted, or in the extracellular portions of cell-surface proteins, which are not exposed to the **reducing environment** in the interior of the cell, there may be additional covalent linkages present in the form of **disulfide bridges** between the side chains of cysteine residues. Except for cross-links like these, however, the remainder of the stabilization energy of a folded protein comes not from covalent bonds but from noncovalent weakly polar interactions. The properties of all the interactions that hold folded proteins together are listed in Figure 1-10. Weakly polar interactions depend on the electrostatic attraction between opposite charges. The charges may be permanent and full, or fluctuating and partial. In general, the term **electrostatic interaction** is reserved for those interactions due to full charges, and this convention is observed in Figure 1-10. But in principle, all polar interactions are electrostatic and the effect is the same: positively polarized species will associate with negatively polarized ones. Such interactions rarely contribute even one-tenth of the enthalpy contributed by a single covalent bond (see Figure 1-10), but in any folded protein structure there may be hundreds to thousands of them, adding up to a very large contribution. The two most important are the **van der Waals interaction** and the **hydrogen bond**.

Van der Waals interactions occur whenever the fluctuating electron clouds on an atom or group of bonded atoms induce an opposite fluctuating dipole on a non-bonded neighbor, resulting in a very weak electrostatic interaction. The effect is greatest with those groups that are the most polarizable; in proteins these are usually the methyl groups and methylene groups of hydrophobic side chains such as leucine and valine. Van der Waals interactions diminish rapidly as the interacting species get farther apart, so only atoms that are already close together (about 5 Å apart or less) have a chance to participate in such interactions. A given van der Waals interaction is extremely weak (see Figure 1-10), but in proteins they sum up to a substantial energetic contribution.

Hydrogen bonds are formed when a hydrogen atom has a significant partial positive charge by virtue of being covalently bound to a more electronegative atom, such as oxygen, and is attracted to a neighboring atom that has a significant partial negative charge (see Figure 1-10). This electrostatic interaction draws the two non-hydrogen atoms closer together than the sum of their atomic radii would normally allow. So, if two polar non-hydrogen atoms in a protein, one of which has a hydrogen attached, are found to be less than 3.5 Å apart, a hydrogen bond is assumed to exist between them. It is thought that the hydrogen-bonding effect is energetically most favorable if the three-atom system is roughly linear. The atom to which the hydrogen is covalently attached is called the **donor atom**; the non-bonded one is termed the **acceptor atom**. If the donor, the acceptor or both are fully charged, the hydrogen bond is stronger than when both are uncharged. When both the donor and acceptor are fully charged, the bonding energy is significantly higher and the hydrogen-bonded ion pair is called a **salt bridge** (see Figure 1-10).

The strengths of all polar weak interactions depend to some extent on their environment. In the case of hydrogen bonding, the strength of the interaction depends critically on whether the groups involved are exposed to water.

Figure 1-10 Table of the typical chemical interactions that stabilize polypeptides Values for the interatomic distances and free energies are approximate average values; both can vary considerably. Any specific number is highly dependent on the context in which the interaction is found. Therefore values such as these should only be taken as indicative of the approximate value.

Definitions

disulfide bridge: a covalent bond formed when the reduced –S–H groups of two cysteine residues react with one another to make an oxidized –S–S– linkage.

electrostatic interaction: noncovalent interaction between atoms or groups of atoms due to attraction of opposite charges.

hydrogen bond: a noncovalent interaction between the **donor atom**, which is bound to a positively polarized hydrogen atom, and the **acceptor atom**, which is negatively polarized. Though not covalent, the hydrogen bond holds the donor and acceptor atom close together.

reducing environment: a chemical environment in which the reduced states of chemical groups are favored. In a reducing environment, free –S–H groups are favored over –S–S– bridges. The interior of most cells is a highly reducing environment.

salt bridge: a **hydrogen bond** in which both donor and acceptor atoms are fully charged. The bonding energy of a salt bridge is significantly higher than that of a hydrogen bond in which only one participating atom is fully charged or in which both are partially charged.

van der Waals interaction: a weak attractive force between two atoms or groups of atoms, arising from the fluctuations in electron distribution around the nuclei. Van der Waals forces are stronger between less electronegative atoms such as those found in hydrophobic groups.

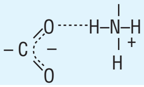
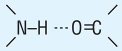
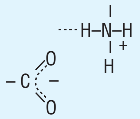
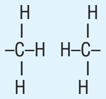
The hydrogen-bonding properties of water have important effects on protein stability

Water, which is present at 55 M concentration in all aqueous solutions, is potentially both a donor and an acceptor of hydrogen bonds. Water molecules hydrogen bond to one another, which is what makes water liquid at ordinary temperatures (a property of profound biological significance) and has important energetic consequences for the folding and stability of proteins. The ability of water molecules to hydrogen-bond to the polar groups of proteins has important effects on the energy, or strength, of the hydrogen bonds formed between such groups. This is most clearly seen by comparing hydrogen bonds made by polar groups on the surface and in the interior of proteins.

The strengths of polar weak interactions depend to some extent on their environment. A polar group on the surface of a protein can make interactions with water molecules that are nearly equivalent in energy to those it can make with other surface groups of a protein. Thus, the difference in energy between an isolated polar group and that of the same species when involved in a hydrogen bond with another polar group from that protein, is small. If, however, the interaction occurs in the interior of the protein, away from bulk solvent, the net interaction energy reflects the difference between the group when hydrogen-bonded and when not.

It is energetically very unfavorable not to make a hydrogen bond, because that would leave one or more uncompensated partial or full charges. Thus, in protein structure nearly all potential hydrogen-bond donors and acceptors are participating in such interactions, either between polar groups of the protein itself or with water molecules. In a polypeptide chain of indeterminate sequence the most common hydrogen-bond groups are the peptide C=O and N–H; in the interior of a protein these groups cannot make hydrogen bonds with water, so they tend to hydrogen bond with one another, leading to the secondary structure which stabilizes the folded state.

Chemical Interactions that Stabilize Polypeptides

Interaction	Example	Distance dependence	Typical distance	Free energy (bond dissociation enthalpies for the covalent bonds)
Covalent bond	–Cα–C–	-	1.5 Å	356 kJ/mole (610 kJ/mole for a C=C bond)
Disulfide bond	–Cys–S–S–Cys–	-	2.2 Å	167 kJ/mole
Salt bridge		Donor (here N), and acceptor (here O) atoms <3.5 Å	2.8 Å	12.5–17 kJ/mole; may be as high as 30 kJ/mole for fully or partially buried salt bridges (see text), less if the salt bridge is external
Hydrogen bond		Donor (here N), and acceptor (here O) atoms <3.5 Å	3.0 Å	2–6 kJ/mole in water; 12.5–21 kJ/mole if either donor or acceptor is charged
Long-range electrostatic interaction		Depends on dielectric constant of medium. Screened by water. 1/r dependence	Variable	Depends on distance and environment. Can be very strong in nonpolar region but very weak in water
Van der Waals interaction		Short range. Falls off rapidly beyond 4 Å separation. $1/r^6$ dependence	3.5 Å	4 kJ/mole (4–17 in protein interior) depending on the size of the group (for comparison, the average thermal energy of molecules at room temperature is 2.5 kJ/mole)

References

Burley, S.K. and Petsko, G.A.: **Weakly polar interactions in proteins.** *Adv. Prot. Chem.* 1988, **39**:125–189.

Dunitz, J.D.: **Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions.** *Chem. Biol.* 1995, **2**:709–712.

Fersht, A.R.: **The hydrogen bond in molecular recognition.** *Trends Biochem. Sci.* 1987, **12**:301–304.

Jaenicke, R.: **Stability and stabilization of globular proteins in solution.** *J. Biotechnol.* 2000, **79**:193–203.

Pauling, L.C.: *The Nature of the Chemical Bond and the Structure of Molecules and Crystals* 3rd ed. Chapter 8 (Cornell Univ. Press, Ithaca, New York, 1960).

Sharp, K.A. and Englander, S.W.: **How much is a stabilizing bond worth?** *Trends Biochem. Sci.* 1994, **19**:526–529.

Spearman, J. C.: *The Hydrogen Bond and Other Intermolecular Forces* (The Chemical Society, London, 1975).

1-5 Importance and Determinants of Secondary Structure

Folded proteins have segments of regular conformation

Although proteins are linear polymers, the structures of most proteins are not the random coils found for synthetic non-natural polymers. Most soluble proteins are globular and have a tightly packed core consisting primarily of hydrophobic amino acids. This observation can be explained by the tendency of hydrophobic groups to avoid contact with water and interact with one another. Another striking characteristic of folded polypeptide chains is that segments of the chain in nearly all proteins adopt conformations in which the phi and psi torsion angles of the backbone repeat in a regular pattern. These regular segments form the elements of the **secondary structure** of the protein. Three general types of secondary structure elements have been defined (see section 1-0): helices, of which the most common by far is the **alpha helix**; **beta sheets** (sometimes called **pleated sheets**), of which there are two forms, parallel and antiparallel; and beta turns, in which the chain is forced to reverse direction and which make the compact folding of the polypeptide chain possible.

Secondary structure contributes significantly to the stabilization of the overall protein fold. Helices and pleated sheets consist of extensive networks of hydrogen bonds in which many consecutive residues are involved as we shall see in the next two sections. The hydrogen bonding in these elements of structure provides much of the enthalpy of stabilization that allows the polar backbone groups to exist in the hydrophobic core of a folded protein.

The arrangement of secondary structure elements provides a convenient way of classifying types of folds

Prediction of the location of secondary structure elements from the amino-acid sequence alone is accurate to only about 70% (see section 1-8). Such prediction is sometimes useful because the pattern of secondary structure elements along the chain can be characteristic of certain overall protein folds. For example, a beta-sheet strand followed by an alpha helix, repeated eight times, usually signifies a type of fold called a TIM barrel. All TIM barrels known to date are enzymes, so recognition of a TIM-barrel fold in a sequence suggests that the protein has a catalytic function. However, it is a general rule that while classification of a protein may suggest function it cannot define it, and TIM barrels are known that catalyze many different reactions, so prediction of a more specific function cannot be made from recognition of the fold alone. Moreover, relatively few folds can be recognized in this way. Individual secondary structure elements are rarely associated with specific functions, although there are some interesting exceptions such as the binding of alpha helices in the major groove of DNA in two families of DNA-binding proteins.

Steric constraints dictate the possible types of secondary structure

The physical size of atoms and groups of atoms limits the possible phi and psi torsion angles (see Figure 1-9) that the backbone of a polypeptide chain can adopt without causing protruding groups like the carbonyl and side chains to bump into each other. These allowed values can be plotted on a phi, psi diagram called a **Ramachandran plot** (Figure 1-11). Two broad regions of phi, psi space are permitted by steric constraints: the regions that include the torsion angles of the right-handed alpha helix and of the extended beta or pleated sheet. Residues may have phi, psi values that lie outside the allowed regions in cases where the protein fold stabilizes a locally strained conformation.

Definitions

alpha helix: a coiled conformation, resembling a right-handed spiral staircase, for a stretch of consecutive amino acids in which the backbone –N–H group of every residue n donates a hydrogen bond to the C=O group of every residue n+4.

beta sheet: a **secondary structure** element formed by backbone hydrogen bonding between segments of extended polypeptide chain.

beta turn: a tight turn that reverses the direction of the polypeptide chain, stabilized by one or more backbone hydrogen bonds. Changes in chain direction can also occur by loops, which are peptide chain segments with no regular conformations.

hairpin turn: another name for **beta turn**.

pleated sheet: another name for **beta sheet**.

Ramachandran plot: a two-dimensional plot of the values of the backbone torsion angles phi and psi, with allowed regions indicated for conformations where there is no steric interference. Ramachandran plots are used as a diagnosis for accurate structures: when the phi and psi torsion angles of an experimentally determined protein structure are plotted on such a diagram, the observed values should fall predominantly in the allowed regions.

reverse turn: another name for **beta turn**.

secondary structure: folded segments of a polypeptide chain with repeating, characteristic phi, psi backbone torsion angles, that are stabilized by a regular pattern of hydrogen bonds between the peptide –N–H and C=O groups of different residues.

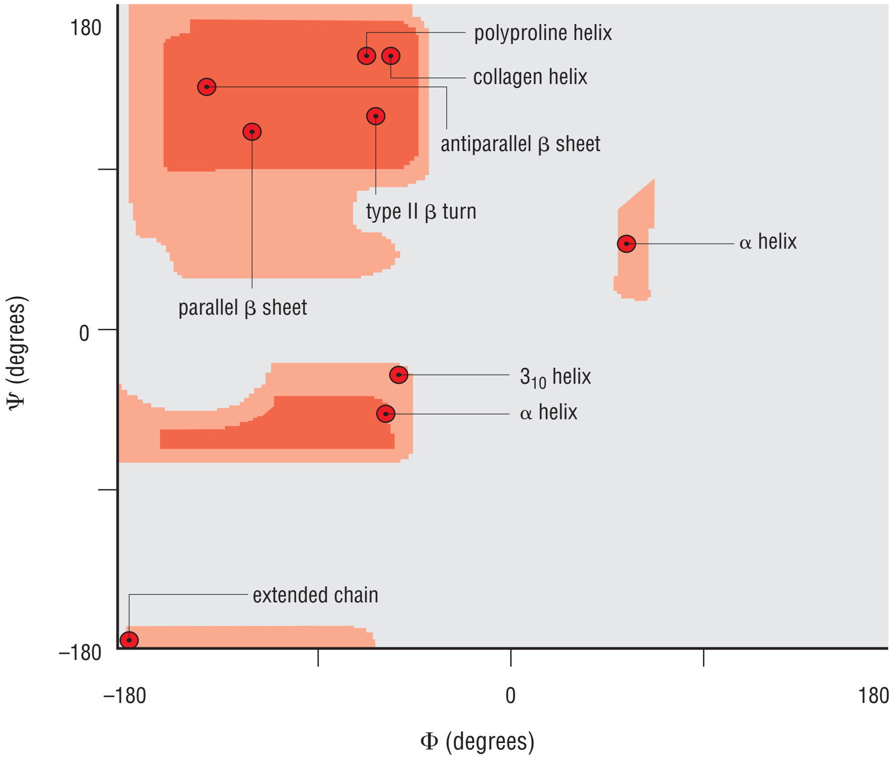


Figure 1-11 Ramachandran plot Shown in red are those combinations of the backbone torsion angles phi and psi (see Figure 1-9) that are "allowed" because they do not result in steric interference. The pink regions are allowed if some relaxation of steric hindrance is permitted. Common protein secondary structure elements are marked at the positions of their average phi, psi values. The isolated pink alpha-helical region on the right is actually for a left-handed helix, which is only rarely observed in short segments in proteins. The zero values of phi and psi are defined as the *trans* configuration.

The simplest secondary structure element is the beta turn

The simplest secondary structure element usually involves four residues but sometimes requires only three. It consists of a hydrogen bond between the carbonyl oxygen of one residue (n) and the amide N–H of residue n+3, reversing the direction of the chain (Figure 1-12). This pattern of hydrogen bonding cannot ordinarily continue because the turn is too tight. This tiny element of secondary structure is called a **beta turn** or **reverse turn** or, sometimes, a **hairpin turn** based on its shape. In a few cases, this interaction can be made between residue n and n+2, but such a turn is strained. Although the reverse turn represents a simple way to satisfy the hydrogen-bonding capability of a peptide group, inspection of this structure reveals that most of the C=O and N–H groups in the four residues that make up the turn are not making hydrogen bonds with other backbone atoms (Figure 1-12). Water molecules can donate and accept hydrogen bonds to these groups if the turn is not buried. Therefore, beta turns are found on the surfaces of folded proteins, where they are in contact with the aqueous environment, and by reversing the direction of the chain they can limit the size of the molecule and maintain a compact state.

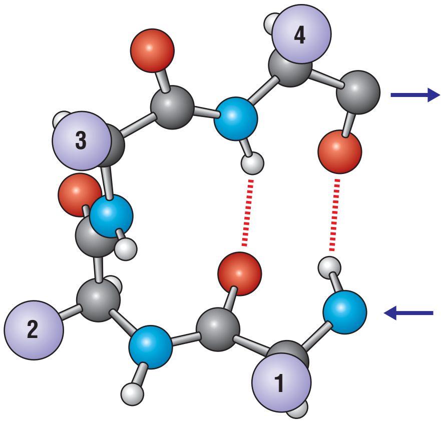


Figure 1-12 Typical beta turn Schematic diagram showing the interresidue backbone hydrogen bonds that stabilize the reversal of the chain direction. Side chains are depicted as large light-purple spheres. The tight geometry of the turn means that some residues, such as glycine, are found more commonly in turns than others.

References

Deane, C.M. *et al.*: **Carbonyl-carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid.** *Protein Eng.* 1999, **12**:1025–1028.

Mattos, C. *et al.*: **Analysis of two-residue turns in proteins.** *J. Mol. Biol.* 1994, **238**:733–747.

Ramachandran, G.N. *et al.*: **Stereochemistry of polypeptide chain configurations.** *J. Mol. Biol.* 1963, **7**:95–99.

Richardson, J.S. and Richardson, D.C.: **Principles and patterns of protein conformation** in *Prediction of Protein Structure and the Principles of Protein Conformation* 2nd ed. Fasman, G.D. ed. (Plenum Press, New York, 1990), 1–98.

1-6 Properties of the Alpha Helix

Alpha helices are versatile cylindrical structures stabilized by a network of backbone hydrogen bonds

Alpha helices are the commonest secondary structural elements in a folded polypeptide chain, possibly because they are generated by local hydrogen bonding between C=O and N–H groups close together in the sequence. In an alpha helix, the carbonyl oxygen atom of each residue (n) accepts a hydrogen bond from the amide nitrogen four residues further along (n+4) in the sequence (Figure 1-13c), so that all of the polar amide groups in the helix are hydrogen bonded to one another except for the N–H group of the first residue in the helical segment (the amino-terminal end) and the C=O group of the last one (the carboxy-terminal end). The result is a cylindrical structure where the wall of the cylinder is formed by the hydrogen-bonded backbone, and the outside is studded with side chains. The protruding side chains determine the interactions of the alpha helix both with other parts of a folded protein chain and with other protein molecules.

The alpha helix is a compact structure, with approximate phi, psi values of –60° and –50° respectively: the distance between successive residues along the helical axis (translational rise) is only 1.5 Å (Figure 1-13a). It would take a helix 20 residues long to span a distance of 30 Å, the thickness of the hydrophobic portion of a **lipid bilayer** (alpha helices are common in the trans-membrane portions of proteins that span the lipid bilayer in cell membranes; see section 1-11). Alpha helices can be right-handed (clockwise spiral staircase) or left-handed (counterclockwise), but because all amino acids except glycine in proteins have the L-configuration, steric constraints favor the right-handed helix, as the Ramachandran plot indicates (see Figure 1-11), and only a turn or so of left-handed alpha helix has ever been observed in the structure of a real protein. There appears to be no practical limit to the length of an alpha helix; helices hundreds of Ångstroms long have been observed, such as in the keratin fibers that make up human hair. There are variants of the alpha helix with slightly different **helical parameters** (Figure 1-14), but they are much less common and are not very long because they are slightly less stable.

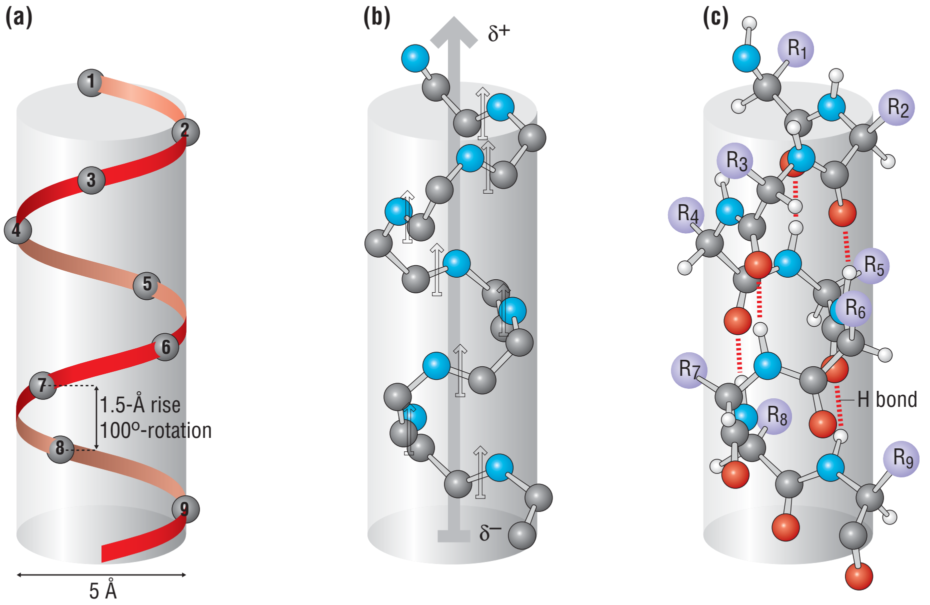


Figure 1-13 The alpha helix The chain path with average helical parameters is indicated showing **(a)** the alpha carbons only, **(b)** the backbone fold with peptide dipoles and **(c)** the full structure with backbone hydrogen bonds in red. All three chains run from top to bottom (that is, the amino-terminal end is at the top). Note that the individual peptide dipoles align to produce a macrodipole with its positive end at the amino-terminal end of the helix. Note also that the amino-terminal end has unsatisfied hydrogen-bond donors (N–H groups) whereas the carboxy-terminal end has unsatisfied hydrogen-bond acceptors (C=O groups). Usually a polar side chain is found at the end of the helix, making hydrogen bonds to these donors and acceptors; such a residue is called a helix cap.

Definitions

amphipathic alpha helix: an alpha helix with a hydrophilic side and a hydrophobic side.

helical parameters: set of numerical values that define the geometry of a helix. These include the number of residues per turn, the translational rise per residue, and the main-chain torsional angles.

helix dipole: the macrodipole that is thought to be formed by the cumulative effect of the individual peptide dipoles in an alpha helix. The positive end of the dipole is at the beginning (amino terminus) of the helix; the negative end is at the carboxyl terminus of the helical rod.

lipid bilayer: the structure of cellular membranes, formed when two sheets of lipid molecules pack against each other with their hydrophobic tails forming the interior of the sandwich and their polar head-groups covering the outside.

References

Hol, W.G.: **The role of the alpha helix dipole in protein function and structure.** *Prog. Biophys. Mol. Biol.* 1985, **45**:149–195.

Pauling, L. *et al.*: **The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain.** *Proc. Natl Acad. Sci. USA* 1951, **37**:205–211.

Scott, J.E.: **Molecules for strength and shape.** *Trends Biochem. Sci.* 1987, **12**:318–321.

Average Conformational Parameters of Helical Elements

Conformation	Phi	Psi	Omega	Residues per turn	Translation per residue
Alpha helix	–57	–47	180	3.6	1.5
3-10 helix	–49	–26	180	3.0	2.0
Pi-helix	57	–70	180	4.4	1.15
Polyproline I	–83	+158	0	3.33	1.9
Polyproline II	–78	+149	180	3.0	3.12
Polyproline III	–80	+150	180	3.0	3.1

Figure 1-14 Table of helical parameters Average conformational parameters of the most commonly found helical secondary structure elements.

In a randomly coiled polypeptide chain the dipole moments of the individual backbone amide groups point in random directions, but in an alpha helix the hydrogen-bonding pattern causes all of the amides—and their dipole moments—to point in the same direction, roughly parallel to the helical axis (Figure 1-13b). It is thought that, as a result, the individual peptide dipoles in a helix add to make a macrodipole with the amino-terminal end of the helix polarized positively and the carboxy-terminal end polarized negatively. The magnitude of this **helix dipole** should increase with increasing length of the helix, provided the cylinder remains straight. Because favorable electrostatic interactions could be made between oppositely charged species and the ends of the helix dipole, one might expect to find, at frequencies greater than predicted by chance, negatively charged side chains and bound anions at the amino-terminal ends of helices, and positively charged side chains and cations interacting with the carboxy-terminal ends. Experimentally determined protein structures and studies of model peptides are in accord with these predictions. Indeed, the helix dipole in some cases contributes significantly to the binding of small charged molecules by proteins.

Alpha helices can be amphipathic, with one polar and one nonpolar face

The alpha helix has 3.6 residues per turn, corresponding to a rotation of 100° per residue, so that side chains project out from the helical axis at 100° intervals, as illustrated in Figure 1-15, which shows the view down the helix axis. This periodicity means that, broadly speaking, residues 3-4 amino acids apart in the linear sequence will project from the same face of an alpha helix. In many alpha helices, polar and hydrophobic residues are distributed 3-4 residues apart in the sequence, to produce an alpha helix with one hydrophilic face and one hydrophobic face; such a helix is known as an **amphipathic alpha helix**, which can stabilize helix–helix packing. Helices with this character frequently occur on the surfaces of proteins, where their polar faces are in contact with water, or at interfaces where polar residues interact with one another: the distribution of polar and hydrophobic residues in a sequence is therefore useful in positioning alpha helices in structure prediction, and in predicting their positions at interfaces.

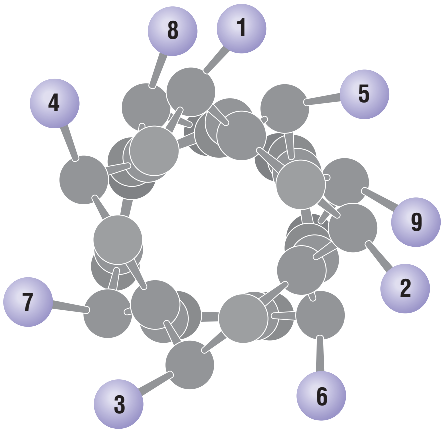


Figure 1-15 View along the axis of an idealized alpha-helical polypeptide The view is from the amino-terminal end. Side chains project outward from the helical axis at 100° intervals. Note that side chains four residues apart in the sequence tend to cluster on the same face of the helix, for shorter helices. For long helices any such pattern would slowly coil about the helix axis, so if two long helices had a pattern of hydrophobic groups four residues apart they would interact by forming a coiled coil (see Figure 1-67).

Collagen and polyproline helices have special properties

Although the amino acid proline, which lacks an N–H group, is not frequently found in an alpha helix, two interesting helical structures can be formed from sequences rich in proline residues. The first is the collagen triple helix (Figure 1-16). Collagen is the main constituent of the bones, tendons, ligaments and blood vessels of higher organisms and consists of a repeating tripeptide in which every third residue is a glycine (GlyXY)n. X and Y are usually proline residues, although lysine occurs sometimes. Many of the proline residues are hydroxylated post-translationally. Each collagen strand forms a (left-handed) helical conformation and three such strands coil around each other like those of a rope. The effect is to create a fibrous protein of great tensile strength. Collagen molecules more than 2 μm in length have been observed. Denaturing the collagen triple helix by heating converts it to a disordered, dissociated, random mass that we call gelatin.

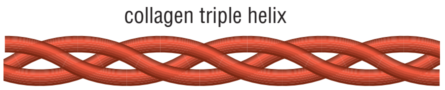


Figure 1-16 The structure of collagen Collagen is a three-chain fibrous protein in which each chain winds round the others. The rise per residue is much larger than in an alpha helix.

The second proline-rich conformation is that formed by polyproline sequences. When the peptide bonds in a polyproline sequence are all *trans* it forms a left-handed helix with three residues per turn. Such a conformation is easily recognized by other proteins, and helical polyproline sequences often serve as docking sites for protein recognition modules, such as SH3 domains in signal transduction pathways.

1-7 Properties of the Beta Sheet

Beta sheets are extended structures that sometimes form barrels

In contrast to the alpha helix, the beta pleated sheet, whose name derives from the corrugated appearance of the extended polypeptide chain (Figure 1-17), involves hydrogen bonds between backbone groups from residues distant from each other in the linear sequence. In beta sheets, two or more strands that may be widely separated in the protein sequence are arranged side by side, with hydrogen bonds between the strands (Figure 1-17). The strands can run in the same direction (**parallel beta sheet**) or **antiparallel** to one another; **mixed sheets** with both parallel and antiparallel strands are also possible (Figure 1-17).

Nearly all polar amide groups are hydrogen bonded to one another in a beta-sheet structure, except for the N–H and C=O groups on the outer sides of the two edge strands. Edge strands may make hydrogen bonds in any of several ways. They may simply make hydrogen bonds to water, if they are exposed to solvent; or they may pack against polar side chains in, for example, a neighboring alpha helix; or they may make hydrogen bonds to an edge strand in another protein chain, forming an extended beta structure that spans more than one subunit and thereby stabilizes quaternary structure (Figure 1-18). Or the sheet may curve round on itself to form a barrel structure, with the two edge strands hydrogen bonding to one another to complete the closed cylinder (Figure 1-19). Such **beta barrels** are a common feature of protein architecture.

Parallel sheets are always buried and small parallel sheets almost never occur. Antiparallel sheets by contrast are frequently exposed to the aqueous environment on one face. These observations suggest that antiparallel sheets are more stable, which is consistent with their hydrogen bonds being more linear (see Figure 1-17). Silk, which is notoriously strong, is made up of stacks of antiparallel beta sheets. Antiparallel sheets most commonly have beta turns connecting the strands, although sometimes the strands may come from discontiguous regions of the linear sequence, in which case the connections are more complex and may include segments of alpha

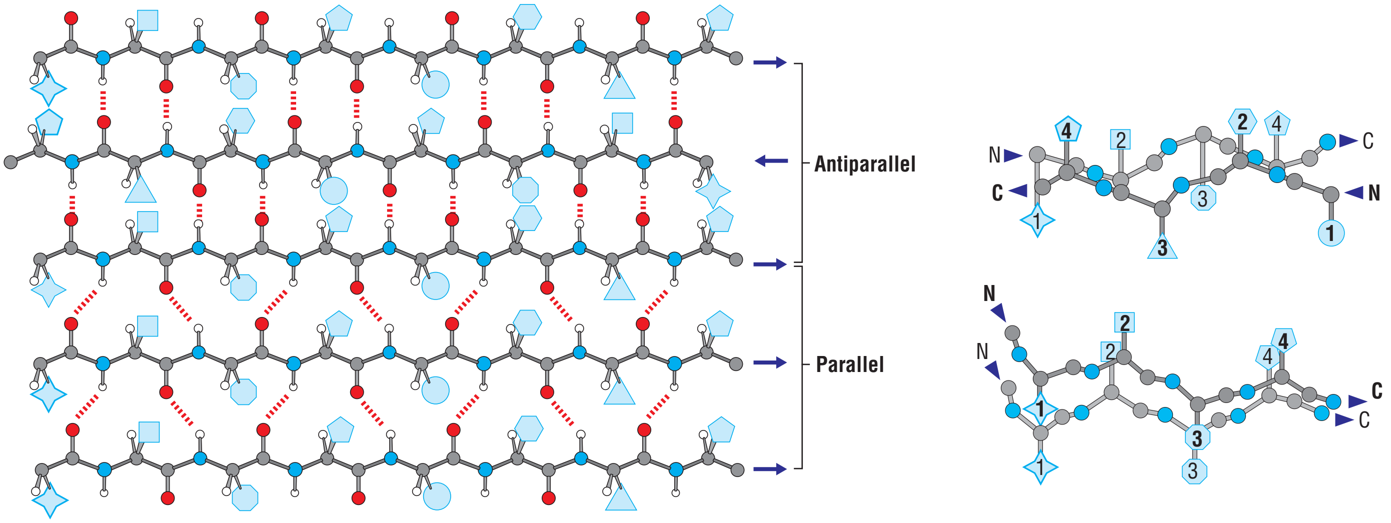


Figure 1-17 The structure of the beta sheet The left figure shows a mixed beta sheet, that is one containing both parallel and antiparallel segments. Note that the hydrogen bonds are more linear in the antiparallel sheet. On the right are edge-on views of antiparallel (top) and parallel sheets (bottom). The corrugated appearance gives rise to the name "pleated sheet" for these elements of secondary structure. Consecutive side chains, indicated here as numbered geometric symbols, point from alternate faces of both types of sheet.

Definitions

antiparallel beta sheet: a beta sheet, often formed from contiguous regions of the polypeptide chain, in which each strand runs in the opposite direction from its immediate neighbors.

beta barrel: a beta sheet in which the last strand is hydrogen bonded to the first strand, forming a closed cylinder.

mixed beta sheet: beta sheet containing both parallel and antiparallel strands.

parallel beta sheet: a beta sheet, formed from non-contiguous regions of the polypeptide chain, in which every strand runs in the same direction.

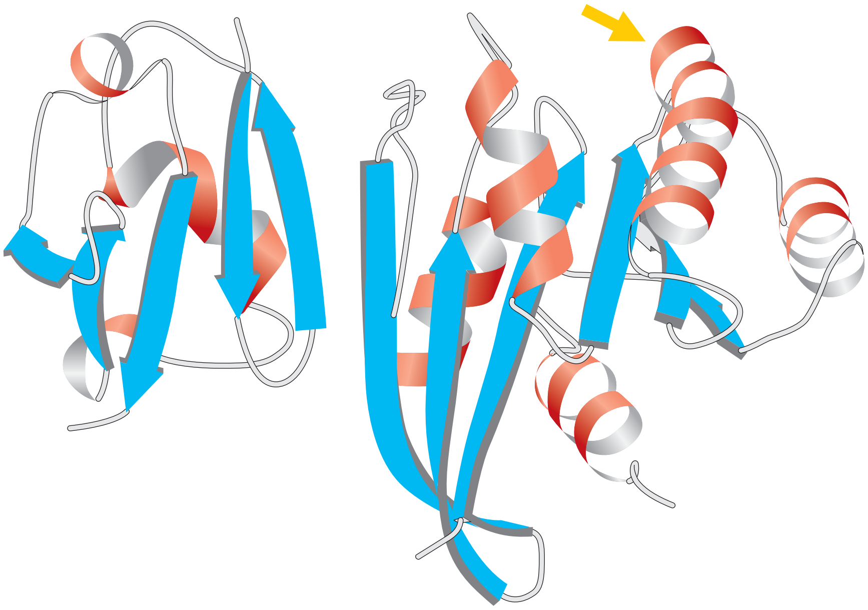


Figure 1-18 **Two proteins that form a complex through hydrogen bonding between beta strands (the Rap–Raf complex, PDB 1gua)** Two antiparallel edge strands of individual beta sheets hydrogen bond to each other at the protein–protein interface, forming a continuous mixed sheet that stabilizes the complex. The protein on the right contains a parallel beta sheet where each strand is connected to the next by an alpha helix, such as the one indicated with the yellow arrow. These helices pack against the faces of the sheet.

helix. Parallel sheet strands are of necessity always discontiguous, and the most common connection between them is an alpha helix that packs against a face of the beta sheet (see for example the helix indicated in Figure 1-18).

The polypeptide chain in a beta sheet is almost fully extended. The distance between consecutive residues is 3.3 Å and the phi and psi angles for peptides in beta sheets are approximately –130° and +125° respectively. Beta strands usually have a pronounced right-handed twist (see for example the sheets in Figure 1-19), due to steric effects arising from the L-amino acid configuration. Parallel strands are less twisted than antiparallel ones. The effect of the strand twist is that sheets consisting of several long strands are themselves twisted.

Because the polypeptide chain in a beta sheet is extended, amino-acid side chains such as those of valine and isoleucine, which branch at the beta carbon, can be accommodated more easily in a beta structure than in a tightly coiled alpha helix where side chains are crowded more closely together. Although unbranched side chains can fit in beta structures as well, branched side chains appear to provide closer packing so they are found more frequently in sheets than other residues. However, it is generally easier to identify helical stretches in sequences than to identify sections of beta structure, and locating the ends of beta sections from sequence alone is particularly difficult.

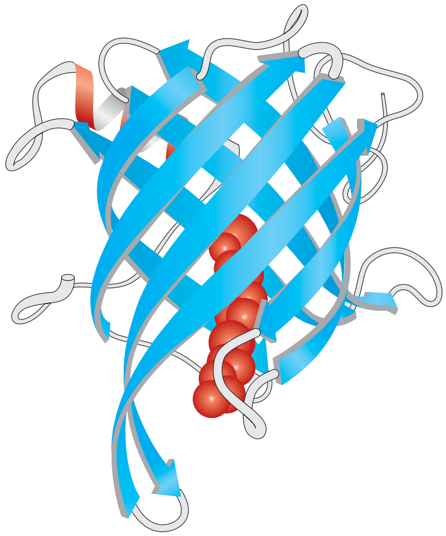


Figure 1-19 **Beta barrel** In this retinol-binding protein (PDB 1rlb), a large antiparallel beta sheet curves all the way around so that the last strand is hydrogen bonded to the first, forming a closed cylinder. The interior of this beta barrel is lined with hydrophobic side chains; nonpolar molecules such as retinol (shown in red) can bind inside.

Amphipathic beta sheets are found on the surfaces of proteins

Like alpha helices (see section 1-6), beta strands can be amphipathic. Because nearly all peptide bonds are *trans* (that is, the C=O and N–H groups point in opposite directions to avoid collision between them), as one proceeds along a beta strand the side chains point in opposite directions (see Figure 1-17). Thus, a stretch of sequence with alternating hydrophobic and hydrophilic residues could have one hydrophobic and one hydrophilic face, forming an amphipathic beta strand (or, depending on its length, several strands of amphipathic antiparallel sheet). Such strands and sheets are found on the surface of proteins.

References

Gellman, S.H.: **Minimal model systems for beta sheet secondary structure in proteins.** *Curr. Opin. Chem. Biol.* 1998, **2**:717–725.

Pauling, L. and Corey, R.B.: **Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets.** *Proc. Natl Acad. Sci. USA* 1951, **37**:729–740.

Richardson, J.S.: **The anatomy and taxonomy of protein structure.** *Adv. Prot. Chem.* 1981, **34**:167–339.

Tutorial for using Protein Explorer to view structures:

http://www.umass.edu/microbio/chime/explorer/pe_tut.htm

1-8 Prediction of Secondary Structure

Certain amino acids are more usually found in alpha helices, others in beta sheets

Analysis of the frequency with which different amino acids are found in different types of structure shows some general preferences. For example, long side chains such as those of leucine, methionine, glutamine and glutamic acid are often found in helices, presumably because these extended side chains can project out away from the crowded central region of the helical cylinder. In contrast, residues whose side chains are branched at the beta carbon, such as valine, isoleucine and phenylalanine, are more often found in beta sheets, because every other side chain in a sheet is pointing in the opposite direction, leaving room for beta-branched side chains to pack. Such tendencies underlie various empirical rules for the prediction of secondary structure from sequence, such as those of Chou and Fasman.

In the Chou-Fasman and other statistical methods of predicting secondary structure, the assumption is made that local effects predominate in determining whether a stretch of sequence will be helical, form a turn, compose a beta strand, or adopt an irregular conformation. This assumption is probably only partially valid, which may account for the failure of such methods to achieve close to 100% success in secondary structure prediction. The methods take proteins of known three-dimensional structure and tabulate the preferences of individual amino acids for various structural elements. By comparing these values with what might be expected randomly, conformational preferences can be assigned to each amino acid (Figure 1-20). To apply these preferences to a sequence of unknown structure, a moving window of about five residues is scanned along a sequence, and the average preferences are tallied. Empirical rules are then applied to assign secondary structural features based on the average preferences.

Conformational Preferences of the Amino Acids

Amino acid	Preference α-helix	 β-strand	 Reverse turn
Glu	**1.59**	0.52	1.01
Ala	**1.41**	0.72	0.82
Leu	**1.34**	1.22	0.57
Met	**1.30**	1.14	0.52
Gln	**1.27**	0.98	0.84
Lys	**1.23**	0.69	1.07
Arg	**1.21**	0.84	0.90
His	**1.05**	0.80	0.81
Val	0.90	**1.87**	0.41
Ile	1.09	**1.67**	0.47
Tyr	0.74	**1.45**	0.76
Cys	0.66	**1.40**	0.54
Trp	1.02	**1.35**	0.65
Phe	1.16	**1.33**	0.59
Thr	0.76	**1.17**	0.90
Gly	0.43	0.58	**1.77**
Asn	0.76	0.48	**1.34**
Pro	0.34	0.31	**1.32**
Ser	0.57	0.96	**1.22**
Asp	0.99	0.39	**1.24**

Figure 1-20 Table of conformational preferences of the amino acids The normalized frequencies for each conformation were calculated from the fraction of residues of each amino acid that occurred in that conformation, divided by this fraction for all residues. Random occurrence of a particular amino acid in a conformation would give a value of unity. A value greater than unity indicates a preference for a particular type of secondary structure. Adapted, with permission, from Table II of Williams, R.W. *et al.*: *Biochim. Biophys. Acta* 1987, **916**:200–204.

Unfortunately, these tendencies are only very rough, and there are many exceptions. It is probably more useful to consider which side chains are disfavored in particular types of secondary structures. With specialized exceptions (see section 1-6), proline is disfavored in both helices and sheets because it has no backbone N–H group to participate in hydrogen bonding. Glycine is also less commonly found in helices and sheets, in part because it lacks a side chain and therefore can adopt a much wider range of phi, psi torsion angles in peptides. These two residues are, however, strongly associated with beta turns, and sequences such as Pro–Gly and Gly–Pro are sometimes considered diagnostic for turns. Although predictive schemes based on residue preferences have some value, none is completely accurate, and the one rule that seems to be most reliable is that any amino acid can be found in any type of secondary structure, if only infrequently. Proline, for instance, is sometimes found in alpha helices; when it is, it simply interrupts the helical hydrogen-bonding network and produces a kink in the helix. One set of conformational preferences of the different amino acids, based on empirical data, is given in Figure 1-20, and an example of secondary structure prediction is shown in Figure 1-21.

Because secondary structure formation is driven by the burial of peptide groups when hydrophobic side chains associate with each other and exclude water, one might expect that isolated segments of secondary structure (for example, sequences corresponding to a single alpha helix or an antiparallel sheet) would not be very stable on their own in aqueous solution. For most sequences, that is exactly what is observed. A few sequences form semistable helices in water, especially at reduced temperatures, and it has been suggested that these might serve as nucleation sites for protein folding.

One place where a single, isolated alpha helix might be expected to be stable would be in the hydrophobic interior of a membrane. With no water to compete for hydrogen bonding, the amide groups would have an extremely strong tendency to form hydrogen bonds with each other, even in the absence of a nonpolar protein core. The membrane would, in such a case, substitute for such a core. Many transmembrane proteins span the membrane by means of a single alpha helix consisting of about 20 hydrophobic residues; the presence of a stretch of around 20 such residues is often considered as diagnostic for a transmembrane helix. Of course, the same considerations would apply to a beta sheet, with one problem: the edge strands of a beta sheet in a membrane would have unsatisfied hydrogen-bonding groups, with no water or polar side chains to interact with them. Thus, all of those transmembrane beta sheets whose structures have been observed thus far form closed barrels and therefore have no edge structures, and all such structures found to date are parts of pores or channels.

```
                    10        20        30        40        50        60        70
                     |         |         |         |         |         |         |
UNK_7585500 MTKNESYSGIDYFRFIAALLIVAIHTSPLFSFSETGNFIFTRIVAPVAVPFFFMTSGFFLISRYTCNAEK
DPM         cccttttctccchhhhhhhheehecccccccccttcceeeeeeehceheceeehccceeeeeccccthhc
DSC         ccccccccccchhhhhhheeeeeecccccccccccccceeeeeccccccceeeccccceeeeecccchhh
GOR4        cccccccccchhhhhhhhhhhhhccccccccccccccceeeeecccccccccccccceeeeecccccccc
HNNC        cccccccchhhhhhhhhhhhhhhhhcccceeeccccceeeeeeeccccccheehccchhhhhccccchhh
PHD         cccccccccchhhhhhhhhhhhhhhcccceeeecccceeeeeeeeeeeeeeeeeccceeeeecccchhhh
Predator    ccccccccccchhhhhhhhhhhhhcccccccccccccceeeeeeeccceeeeecccceeeeecccchhhh
SIMPA96     ccccccccchhhhhhhhhhheeecccccccccccccceeeeeeccccccceeecccceeeeecccccchhh
SOPM        hccttccttchhhhhhhhhheeeecccccceeecttcceeeeeeccccccceeeecttceeeehcccchhh
Sec.Cons.   cccccccccchhhhhhhhhh??h?ccccccccccccceeeeee?cccccceeecccceeeeecccccchhh

                    80        90       100       110       120       130       140
                     |         |         |         |         |         |         |
UNK_7585500 LGAFIKKTTLIYGVAILLYIPINVYNGYFKMDNLLPNIIKDIVFDGTLYHLWYLPASIIGAAIAWYLVKK
DPM         hchhhhhcceeeeeeeeeeececectcccchcccccceecceeeccccccccccccceechhhhhheehh
DSC         hhhhhcccceeeeeeeeeeeccccccccccccccccccccceeecccccccccccccccccccchhhhhhhh
GOR4        cccccccchhhhceeeeeecccceecccccccccccccceeeeeeccceeceeecccchhhhhhhhhhhhh
HNNC        hhhhhhhhhhhhhhhheeeeccceccccceehchhhhhhhhhheeccchheeheecchhhhhhhhhhhhhhh
PHD         hhhhhceeeeeecceeeeeeeecccccccccccccchhhhhhhhhcchhhhhhhhhhhhhhhhhhhhhhhh
Predator    hhhhhcceeecccceeeeeeccccccccccccccccchhhhhcccceeeeeeeccchhhhhhhhhhhhhh
SIMPA96     hhhhhhhhhhhhceeeeeecccccccccccccccchhhhhhhcccccceeeeecchhhhhhhhhhhhhhh
SOPM        hhhhhhhheeeeeeeeeeecccceттcchhhhtcchhhhhhheettceeeeeecccchhhhhhhhhhhhh
Sec.Cons.   hhhhh?h??eeeceeeeeeeccc?ccccccccccc??hhhhheeccc?eeeeeccc?hhhhhhhhhhhhh

                   150       160       170       180       190       200       210
                     |         |         |         |         |         |         |
UNK_7585500 VHYRKAFLIASILYIIGLFGDSYYGIVKSVSCLNVFYNLIFQLTDYTRNGIFFAPIFFVLGGYISDSPNR
DPM         ehhhhhhhhheeeeeeccctttccceeeeeeeeceeeceeehhcccccccchhceeeeecccctctttc
DSC         hccchhhhhhhhheeeeecccccccceeecccccccccceeeeccccccccccccceeeeecccccccccc
GOR4        hhhhhhhhhhhhheeeecccccceeeeeeeccccccccccеecccccccccееceeeecccccccccccc
HNNC        hhhhhhhhhhhhhhhhhhcccccheheeeehhhhhhhhhhhhhhcccccchеehhhhhhhcccccccccc
PHD         hhhhhhhhhhhhhhhhhhccccceeeeeeecceхhhhhhhhhhhhhhhcccceceeeeeceсeeeccccc
Predator    chhhhhhhhhhhhhhhhcccccceeeeecccccchhhhhhhhhcccccccceeeeeeeeeecccccccccc
SIMPA96     hhhhhhhhhhhheeeecccccchhhhhhhhhhhhhhhhhhhhcccccccceeeeeeeeeecccccccccc
SOPM        cchtthhhhhhhhheeeecccchhhhhhhhhhhhhhhhhheeehcccttteeecceeeeetcccccсttt
Sec.Cons.   hhhhhhhhhhhhh?ee?cccccc?eeeeeecc?hhhhhhh?hccccccccceecceeeeecccccccccc

                   220       230       240       250       260       270       280
                     |         |         |         |         |         |         |
UNK_7585500 YRKKNYIRIYSLFCLMFGKTLTLQHFDIQKHDSMYVLLLPSVWCLFNLLLHFRGKRRTGLRTISLDQLYH
DPM         cctccceeeeeeeehhhccchchhhhhhhcccchеееееcceeeehhhhhhhhtcctcccccechccccc
DSC         ccccccceeeeeeeccccccccccccccccccceeeeecccchhhhhhheeccccccccceeeccccccc
GOR4        ccccceeeeeeeeeecccccceecceeeecccceeeeeecccccccchhhhccccccccceeeccccееc
HNNC        cchchheeehhhhhhhccccceeeeeeecccccеееееhhhhhhhhhhhehcccccccceeeeeeeeecc
PHD         ccccceeeeeecceecccccееееccccccccceeeeeeccchhhhhhhhhhccceccccceeccccccce
Predator    ccccceeeeeehhhhhhcceeeecccccccccccеееееccchhhhhhhhhcccccccceeeeeeccccc
SIMPA96     cccccceeeeeeeeeeccceeeeecccccccccеееееeccchhhhheeeeecccccccceeeehhhhcc
SOPM        ccccchеееееееееettttеееееееcctccceeeeecchhhhhhhhhheettcccccceeeeehhhhh
Sec.Cons.   ccccc?eeeeeeeee?cccceeeecccccccccеееееeccchhhhhhhhhccccccccceeee?ccccc

                   290       300       310       320       330       340       350
                     |         |         |         |         |         |         |
UNK_7585500 SSVYDCCNTIVCAELLHLQSLLVENSLVHYIAVCFASVVLAVVITALLSSLKPKKAKHTADTDRAYLEIN
DPM         ccecccccceehhhhhhhhhhhhhcceeeeeheeeheeeeeeeeehhhccctccchchchccchhhhhhc
DSC         ccccccccceeehhhhhhhhhhhcccccceeeeecccceeeeeeeecccccccccccccccchhhhheee
GOR4        cceeccccеееecchhhhhhhhhhcceeeeeeeccccccchhhhhhhhhhhcchhhhhcccccchhhhhhh
HNNC        ccchhhhhhheehhhhhhhhhehcchhhhhhhhhhhhhhhhhhhhhhhhhccccccccccchhhhhhheh
PHD         cccchhhhhhhhhhhhhhhhhhhhhhеееееееееееееhhhhhhhhhhhccccccccccchhhhhhhhh
Predator    cccccccccchhhhhhhhhhhhcccccceeehhhhhhhhhhhhhhhhccccccccccccccccceeeeec
SIMPA96     cccccccceeehhhhhhhhhhhhccchhhhhhhhhhhhhhhhhhhhhhhhcccccccccccchhhhhhhc
SOPM        hhhhhtthhhhhhhhhhhhhhhhhtthhheeehhhhhhhhhhhhhhhhhtccttccccccccchhheehh
Sec.Cons.   cccccccc?eehhhhhhhhhhhh?cc??eee??hhhhhhhhhhhhhhhhhcccccccccccc?hhhhhhh

                   360       370       380       390       400       410       420
                     |         |         |         |         |         |         |
UNK_7585500 LNNLEHNVNTLQKAMSPKCELMAVVKAEAYGHGMYEVTTYFEPIGVFYLAVATIDEGIRLRKYGIFSEIL
DPM         cccccctecchhhhhttchhhhhhhhhhhtccccchеeccccccceeehhehhechchhhhhccеееhее
DSC         hhhhhhhhhhhhhcccccceeeeeeeeccccccccchhhhhhhhcccchheehhhhhhhhhhhhhcccceе
GOR4        hhhhhhchhhhhhccccchhhhhhhhhhhcccceeeeeeecccccеееееhhhhhhhhhhhhcccceeee
HNNC        ccccccchhhhhhcccchhhhhhhhhhhhhcccceeeeeecccchhhhhhhhhhhhhhhhhchccechеее
PHD         hhhhhhhhhhhhhhccccceeeeeeeeccccccchhhhhhhhhhhhhhhhecccchhhhhhhhhhhcccceе
Predator    cccccccccccccccccchhhhhhhhhhhcccceeeeeeccccccchhhhhhhcccccccccccccceeee
SIMPA96     hhhhhhhhhhhhhhccchhhhhhhhhhcccccccеееееccccccеееееhhhhccchhhhhccccceee
SOPM        hhhhhhhhhhhhhhhcсthhhhhhhhhhhtttthheeeeeccctteeeeeehhhhttceechtteehhee
Sec.Cons.   hhhhhhhhhhhhh?cccchhhhhhhhhhhccccceeeee?cccccee?hehhhhhhhhhhhhcccc?eee

                   430       440       450       460       470       480       490
                     |         |         |         |         |         |         |
UNK_7585500 ILGYTSPSRAKELCKFELTQTLIDYRYLLLLNKQGYDIKAHIKIDTGMHRLGFSTEDKDKILAAFFLKHI
DPM         eecccctcchhhhhhhhhhhceeccheehhhhttccccchhhehccccccccccttctcchhhhhhhhhhe
DSC         eecccccchhhhhcceeecccccccchhhhhhhhhhccceeeeeeccccceeeecchhhhhhhhhhhcchh
GOR4        eecccccchhhhhhhhcccchhhhhhhhhhhcccccceeeeeccccceeeecccchhhhhhhhhhhhhhh
HNNC        eeecccccchhhhhhhhhhhhhhhhhhhhhhhcccccheeeeeecccccceccccccchhhhhhhhhhhhh
PHD         eeecccchhhhhhhhhhhheeccchhhhhhhhhhhhhheeeeeeecccccccccchhhhhhhhhhhhhce
Predator    eeccccchhhhhhhhhhhhhhchhhhhhhhhccccceeeeeccccccccccccccccccchhhhhhhhhhh
SIMPA96     eecccccchhhhhhhhhhhhhhhhhhhheeeecccccccеееееcccceeeccccccchhhhhhhhhhhhhc
SOPM        eeeccccchhhhhhhhhhhhhhhhhhheeeettttcceeeeeeecttceeetccccchhhhhhhhhhhhc
Sec.Cons.   eecccccchhhhhhhhhhhhhhhhhhhhhhhhcccccceeeeeeccccc?ecccccchhhhhhhhhhhhhh
```

Figure 1-21 An example of secondary structure prediction An example of the prediction of secondary structure from sequence for a protein of unknown function from the *Enterococcus faecalis* genome. Only the first 490 residues are shown. Eight different statistical prediction schemes have been applied to this sequence. What is striking is that all of the schemes agree on the approximate locations of the alpha helices (h) and beta strands (e), but they disagree considerably on the lengths and end positions of these segments. Note also that the probable positions of loops (indicated by a c) and turns (indicated by a t) are very inconsistently predicted. Such results are typical, but the application of many methods is clearly more informative than the use of a single one. The bottom line shows the consensus prediction.

References

Chou, P.Y. and Fasman, G.D.: **Prediction of protein conformation.** *Biochemistry* 1974, **13**:222–245.

Deleage, G. *et al.*: **Protein structure prediction. Implications for the biologist.** *Biochimie* 1997, **79**:681–686.

McKessar, S.J. *et al.*: **Genetic characterisation of vanG, a novel vancomycin resistance locus of *Enterococcus faecalis*.** *Antimicrob. Agents Chemother.* 2000, **44**:3224–3228.

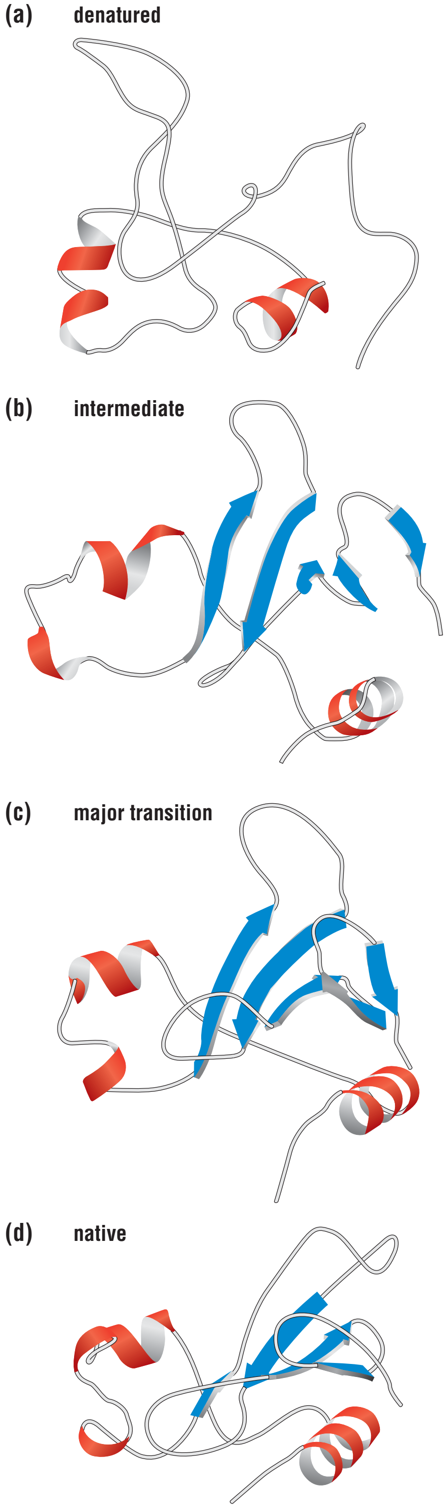
Swindells, M.B. *et al.*: **Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures.** *Nat. Struct. Biol.* 1995, **2**:596–603.

Williams, R.W. *et al.*: **Secondary structure predictions and medium range interactions.** *Biochim. Biophys. Acta* 1987, **916**:200–204.

Zhu, Z.Y. and Blundell, T.L.: **The use of amino acid patterns of classified helices and strands in secondary structure prediction.** *J. Mol. Biol.* 1996, **260**:261–276.

Secondary structure prediction resources on the Internet:

http://antheprot-pbil.ibcp.fr/ns_sommaire.html



The folded structure of a protein is directly determined by its primary structure

The three-dimensional or tertiary structure of a protein is determined by the sequence of amino acids encoded by the gene that specifies the protein. Translation of the mRNA produces a linear polymer of amino acids that usually folds spontaneously into a more compact, stable structure. Sometimes folding is assisted by other proteins called **chaperones**, but most proteins can be unfolded and refolded in dilute solution, demonstrating that the primary structure contains all the information necessary to specify the folded state. Protein folding can occur quite rapidly, but there is evidence that one or more partially folded intermediate states often exist, transiently, along the path to the final structure (Figure 1-22). The structures of these intermediates are not as well characterized as the native structures, but have many of the secondary structure elements of the fully folded protein without the closely packed interior and full complement of weak interactions that characterize what is termed the **native state**.

Competition between self-interactions and interactions with water drives protein folding

Consider a protein of arbitrary sequence emerging from the ribosome. If the chain is made up of only polar and charged amino acids, nearly every chemical group in it can hydrogen bond to water whether the chain is folded up or not, so there will be no driving force to form a compact or regular structure. Many such sequences are known in nature, and, as expected, they have no stable folded structure on their own in solution. The amino-acid sequences of soluble proteins tend to be mixtures of polar and nonpolar residues, sometimes in patches, but most often distributed along the chain with no discernible pattern. When such a sequence is synthesized in water, it cannot remain as a fully extended polymer. True, the polar and charged side chains, and the polar peptide groups, will be able to form hydrogen bonds with water; but the nonpolar side chains cannot. Their physical presence will disrupt the hydrogen-bonded structure of water without making any compensating hydrogen bonds with the solvent. To minimize this effect on the water structure, these side chains will tend to clump together the way oil droplets do when dispersed in water. This **hydrophobic effect**—the clustering of hydrophobic side chains from diverse parts of the polypeptide sequence—causes the polypeptide to become compact (Figure 1-23). From an energetic point of view the compactness produces two favorable results: it minimizes the total hydrophobic surface area in contact with water, and it brings the polarizable hydrophobic groups close to each other, allowing van der Waals interactions between them. Polar side chains do not need to be shielded from the solvent because they can hydrogen bond to water, so they will tend to be distributed on the outside of this “oil drop” of hydrophobic residues.

Figure 1-22 Folding intermediates Structures of **(a)** denatured, **(b)** intermediate, **(c)** major transition and **(d)** native states of barnase. Structures were determined from molecular dynamics calculations and NMR experiments, illustrating a possible folding pathway. Note that during folding, segments of secondary structure form that do not completely coincide with their final positions in the sequence, and that the non-native states are considerably expanded and more flexible relative to the final folded form. These characteristics appear to be common to most, if not all, protein-folding pathways. (Bond, C.J. *et al.*: *Proc. Natl Acad. Sci. USA* 1997, **94**:13409–13413.)

Definitions

chaperone: a protein that aids in the folding of another protein by preventing the unwanted association of the unfolded or partially folded forms of that protein with itself or with others.

hydrophobic effect: the tendency of nonpolar groups in water to self-associate and thereby minimize their contact surface area with the polar solvent.

native state: the stably folded and functional form of a biological macromolecule.

But this hydrophobic collapse has a negative consequence. When the nonpolar side chains come together to form a hydrophobic core, they simultaneously drag their polar backbone amide groups into the greasy interior of the protein (see Figure 1-23). These polar groups made hydrogen bonds to water when the chain was extended, but now they are unable to do so. Leaving these groups with unsatisfied hydrogen bonds would lead to a significant energy penalty. Yet they cannot make hydrogen bonds to polar side chains because most such side chains are on the surface interacting with water. The result is that most peptide N–H and carbonyl groups of folded proteins hydrogen bond to each other. It is this tendency of the amide groups of polypeptide chains to satisfy their hydrogen-bonding potential through self-interactions that gives rise to secondary structure, as described in section 1-4.

Although most hydrophobic side chains in a protein are buried, some are found on the surface of the folded polypeptide chain in contact with water. Presumably, this unfavorable situation is offset by the many favorable interactions that provide a net stability for the folded protein. As a rule, such residues occur in isolation; when hydrophobic side chains cluster on the surface they are usually part of a specific binding site for other molecules, or form a patch of mutually interacting nonpolar groups.

Computational prediction of folding is not yet reliable

Recently there have been a number of efforts to fold amino-acid sequences into the correct three-dimensional structures *ab initio* purely computationally. Such methods vary in detail (for example, some start with secondary structure prediction, others do not) but in the end all depend on several assumptions. First, it is assumed that the equilibrium conformation is the global free-energy minimum on a folding pathway. This assumption is likely to be correct, but no one knows for sure. Second, it is assumed that the current empirical potential energy parameters used to compute the contributions of hydrogen bonds, van der Waals interactions and so forth to the overall stabilization energy are sufficiently accurate. This is uncertain. Third, very many globular proteins are oligomeric, and many will not fold as monomers; but only monomeric proteins are treated by these methods. The problem of recognizing that a given sequence will produce a dimeric or tetrameric protein, and how to treat oligomerization in computational approaches to folding, has not even begun to be addressed.

Helical membrane proteins may fold by condensation of preformed secondary structure elements in the bilayer

The hydrophobic environment of a membrane interior allows formation of the same secondary structure elements as does aqueous solution, but the range of protein architectures appears to be much more limited. Thus far, only all-helical and all-beta-barrel integral membrane proteins have been observed (see section 1-11).

Much less is known about the mechanism of folding of proteins whose structure is largely embedded in the hydrophobic interior of a lipid bilayer. Because water molecules do not occupy stable positions in this region of a membrane, the polar N–H and C=O groups of a peptide backbone have no option but to hydrogen bond to one another. Thus, it is thought that transmembrane segments of integral membrane proteins form secondary structure (usually alpha helices) very early in the folding process, and that these elements then assemble to give the final structure by diffusional motion in the bilayer until the most favorable set of side-chain interactions is found. The folding pathway of all-beta-sheet membrane proteins is unclear.

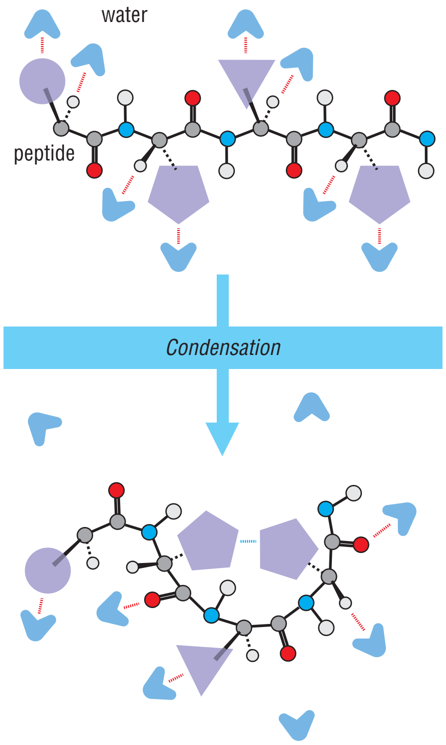


Figure 1-23 Highly simplified schematic representation of the folding of a polypeptide chain in water In the unfolded chain, side-chain and main-chain groups interact primarily with water, even if they are hydrophobic and the interaction is unfavorable. Burying the hydrophobic groups in the interior of a compact structure enables them to interact with each other (blue line), which is favorable, and leaves polar side chains on the surface where they can interact with water (red lines). The polar backbone groups that are buried along with the hydrophobic side chains must make hydrogen bonds to each other (not shown), as bulk water is no longer available.

References

Bond, C.J. *et al.*: **Characterization of residual structure in the thermally denatured state of barnase by simulation and experiment: description of the folding pathway.** *Proc. Natl Acad. Sci. USA* 1997, **94**:13409–13413.

Cramer, W.A. *et al.*: **Forces involved in the assembly and stabilization of membrane proteins.** *FASEB J.* 1992, **6**:3397–3402.

Dinner, A.R. *et al.*: **Understanding protein folding via free-energy surfaces from theory and experiment.** *Trends Biochem. Sci.* 2000, **25**:331–339.

Eaton, W.A. *et al.*: **Fast kinetics and mechanisms in protein folding.** *Annu. Rev. Biophys. Biomol. Struct.* 2000, **29**:327–359.

Fersht, A.: *Structure and Mechanism in Protein Science. A Guide to Enzyme Catalysis and Protein Folding* (Freeman, New York, 1999).

The condensing of multiple secondary structural elements leads to tertiary structure

In a folded protein, the secondary structure elements fold into a compact and nearly solid object stabilized by weak interactions involving both polar and nonpolar groups. The resulting compact folded form is called the **tertiary structure** of the protein. Because tertiary structure is not regular, it is hard to describe it simply. One way to characterize tertiary structure is by the topological arrangement of the various secondary structure elements as they pack together. In fact, the same types of secondary structure elements can come together in many different ways depending on the sequence (Figure 1-24). Tertiary structure is sometimes classified according to the arrangement of secondary structure elements in the linear sequence and in space. One effect of tertiary structure is to create a complex surface topography that enables a protein to interact specifically either with small molecules that may bind in clefts, or with other macromolecules, with which it may have regions of complementary topology and charge. These recognition sites are often formed from the stretches of amino acids joining secondary structure elements.

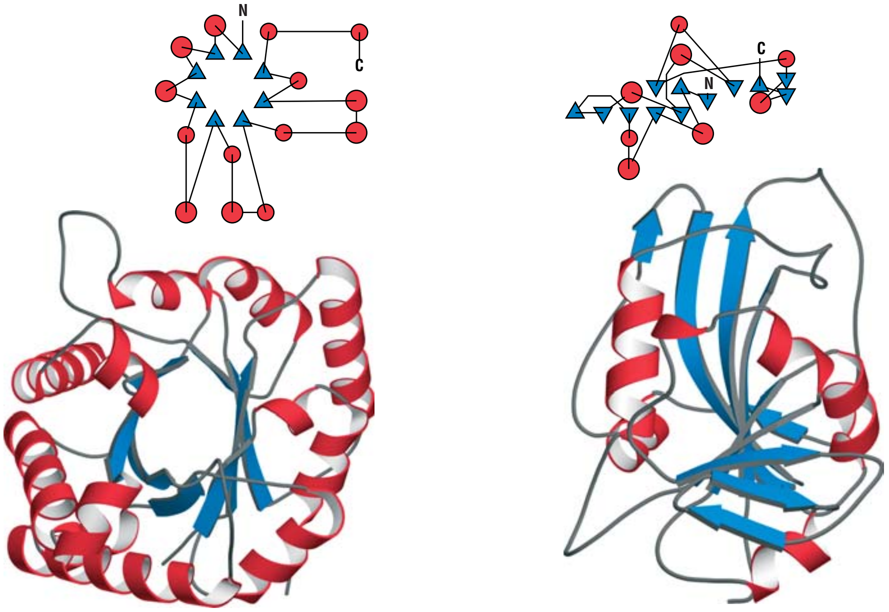


Figure 1-24 Comparison of the structures of triosephosphate isomerase and dihydrofolate reductase Two proteins with similar secondary structure elements but different tertiary structures. Approximately the same secondary structure elements can be arranged in more than one way. Both TIM (left) and DHFR (right) consist of eight beta strands with connecting alpha helices, yet the former is a singly wound parallel alpha/beta barrel whereas the latter is a doubly wound alpha/beta domain with a mixed sheet. (PDB 1tim and 1ai9)

Although helical segments and beta strands are often connected by tight turns, more often there are long stretches of amino acids in between secondary structural elements that do not adopt regular backbone conformations. Such loops are found at the surface of proteins and typically protrude into the solvent. Consequently they provide convenient sites for protein recognition, ligand binding and membrane interaction. For example, the antigen-binding site in immunoglobulins is made up of a series of loops that project up from the core beta structure like the fingers of a cupped hand (Figure 1-25). Because these protruding loops often contribute little to the stabilization of the overall fold, they can tolerate mutations more readily than can the core of the protein. Since

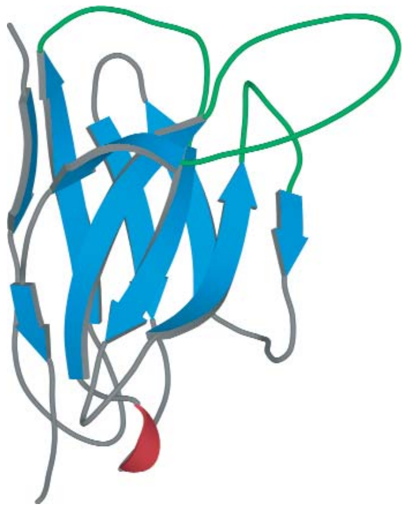


Figure 1-25 Variable loops Three-dimensional structure of the V domain of an immunoglobulin light chain showing the hypervariable loops (green) protruding from the ends of a sandwich formed by two antiparallel beta sheets. The structure resembles a cupped hand, with the hypervariable loops forming the fingers. These loops form the antigen-binding site. (PDB 1ogp)

Definitions

packing motif: an arrangement of secondary structure elements defined by the number and types of such elements and the angles between them. The term motif is used in structural biology in a number of contexts and thus can be confusing.

tertiary structure: the folded conformation of a protein, formed by the condensation of the various secondary elements, stabilized by a large number of weak interactions.

References

Barlow, D.J. and Thornton, J.M.: **Helix geometry in proteins.** *J. Mol. Biol.* 1988, **201**:601–619.

Eilers, M. *et al.*: **Internal packing of helical membrane proteins.** *Proc. Natl Acad. Sci. USA* 2000, **97**:5796–5801.

Lesk, A.M. and Chothia, C.: **Solvent accessibility, protein surfaces and protein folding.** *Biophys. J.* 1980, **32**:35–47.

Richards, F.M. and Richmond, T.: **Solvents, interfaces and protein structure**. *Ciba. Found. Symp.* 1997, **60**:23–45.

Rose, G.D. and Roy, S.: **Hydrophobic basis of packing in globular proteins.** *Proc. Natl Acad. Sci. USA* 1980, **77**:4643–4647.

Walther, D. *et al.*: **Principles of helix-helix packing in proteins: the helical lattice superposition model.** *J. Mol. Biol.* 1996, **255**:536–553.

they are also often involved in function, their mutability provides a mechanism for the evolution of new functions. Although surface loops are often drawn as being open, like a lariat, in reality their side chains frequently pack together so that the loop is nearly solid. This means that when loops undergo conformational changes they often move as rigid bodies.

Bound water molecules on the surface of a folded protein are an important part of the structure

When the polar backbone groups of a polypeptide chain become involved in secondary and tertiary structure interactions, the water molecules that were interacting with them in the unfolded protein are freed to rejoin the structure of liquid water. But there are many polar groups, both backbone and side-chain, on the surface of a folded protein that must remain in contact with water. Atomic-resolution structures of proteins show a layer of bound water molecules on the surfaces of all folded soluble proteins (Figure 1-26); these waters are making hydrogen bonds with polar backbone and side-chain groups and also with one another. There are several such water molecules per residue. Some are in fixed positions and are observed every time the structure is determined. However, others are in non-unique positions and reflect an ensemble of water–protein interactions that hydrate the entire surface. A few additional water molecules are trapped inside the protein in internal cavities. Because bound water molecules make important interactions with groups that would otherwise make none, the waters in fixed positions should be considered as part of the tertiary structure, and any detailed structure description that does not include them is incomplete.

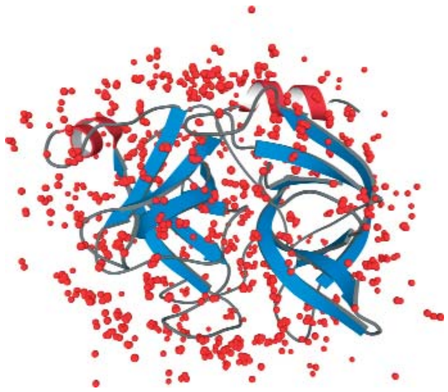


Figure 1-26 Porcine pancreatic elastase showing the first hydration shell surrounding the protein In any one structure determination, only a subset of these water molecules is seen. This picture is a composite of the results of parallel structure determinations of the same protein.

Tertiary structure is stabilized by efficient packing of atoms in the protein interior

The individual secondary structure elements in a protein pack together in part to bury the hydrophobic side chains, forming a compact molecule with very little empty space in the interior (Figure 1-27). The interactions that hold these elements together are the weak interactions described earlier: polar interactions between hydrophilic groups and van der Waals interaction between nonpolar groups. Close packing of atoms maximizes both the probability that these interactions will occur and their strength.

Maintaining a close-packed interior can be accomplished by many different modes of packing of helices with each other and of sheets with each other, and between helices and sheets. These various types of packing arrangements can be described in terms of a set of **packing motifs** that have been used to classify protein tertiary structures in general terms. For example, in helix–helix interactions, the protruding side chains of one helix fit into grooves along the cylindrical surface of the other helix in what has been described as a "ridges and grooves" arrangement. This principle is best illustrated in alpha-helical dimers and we return to it later (see Figure 1-67). These steric considerations permit several different interhelical crossing angles, each set of which constitutes a distinct packing motif (Figure 1-28).

Although the density of atoms in the hydrophobic core of a folded protein is high, the packing is not perfect. There are many cavities that range in size from subatomic packing defects to ones large enough to accommodate several water molecules. If the cavity walls are lined with hydrophobic side chains, the cavity is usually found to contain no ordered waters, but more commonly there are some polar groups lining the cavity and these interact with buried water molecules that fill the space.

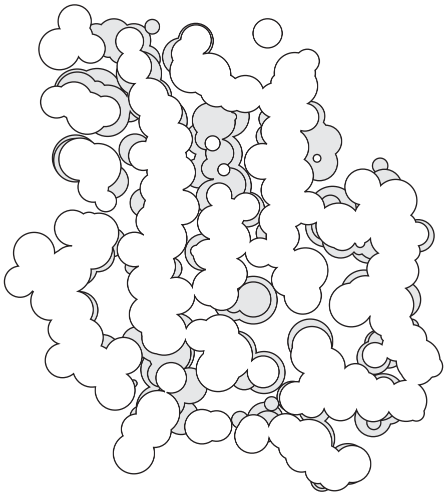


Figure 1-27 Cut-away view of the interior of a folded protein The atoms in the interior of a protein are packed almost as closely as in a solid. Note that there are a few cavities and small channels in some parts of the structure. These packing defects provide room for neighboring atoms to move, allowing the structure to have some flexibility.

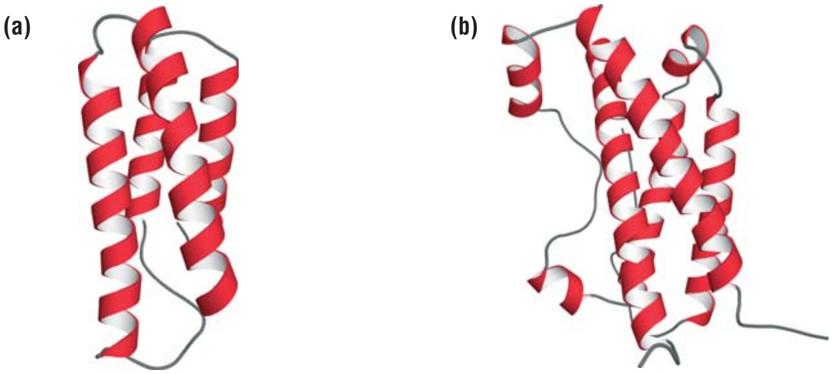


Figure 1-28 Packing motifs of a helical structure When two helices pack together, their side chains interdigitate. Because several interhelical crossing angles allow good interdigitation, a number of distinct arrangements of helical bundles are possible. The two examples illustrated here are **(a)** cytochrome b562 (PDB 256b) and **(b)** human growth hormone (PDB 3hhr).

1-11 Membrane Protein Structure

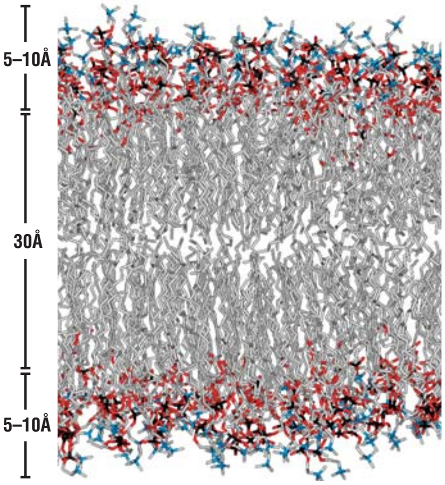


Figure 1-29 A segment of a simulated membrane bilayer http://www.lrz-muenchen.de/~heller/membrane/membrane.html

The principles governing the structures of integral membrane proteins are the same as those for water-soluble proteins and lead to formation of the same secondary structure elements

Not all proteins in the cell exist in an aqueous environment. Some are embedded in the hydrophobic interior of the membranes that form the surfaces of cells, organelles and vesicles. Most biological membranes are bilayers of lipid molecules (derived from fatty acids) with polar or charged head-groups (Figure 1-29). The bilayer resembles a sandwich with the head-groups as the bread and the lipid tails as an almost completely hydrophobic filling. The nonpolar interior of the membrane is approximately 30 Å across; the head-group layers contribute an additional 5–10 Å on each side to the total thickness of the membrane.

A protein that is inserted into a membrane is exposed to an almost completely nonpolar environment. The side chains of amino acids forming transmembrane segments of proteins are usually hydrophobic, and can be accommodated with no energetic cost; but the polar backbone carbonyl and amide groups will all have unfavorable interactions with the nonpolar lipid tails. There will thus be the same strong driving force for these groups to hydrogen bond with one another as there is in the hydrophobic interior of a soluble protein when it folds up in water, and with the same results. Formation of alpha-helical and beta-sheet secondary structure elements is thus strongly favored in the membrane interior. Because hydrogen bonds in a completely nonpolar environment are considered stronger than if the same groups were exposed to solvent, an isolated alpha helix can exist stably in a membrane, whereas non-interacting helices are rare in water-soluble proteins. Any polar side chains will be found either on the protein surface that protrudes out of the membrane, interacting with the polar head-groups of the lipids, or in the core of the membrane-embedded part of the protein, where they can interact with each other or form a polar surface that often constitutes a pore or ion channel through the bilayer.

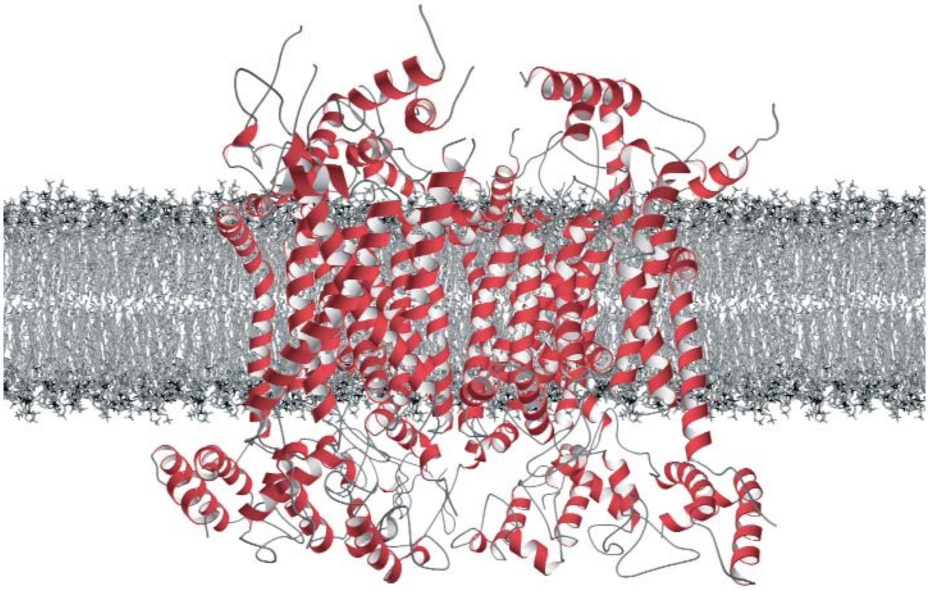


Figure 1-30 The three-dimensional structure of part of the cytochrome bc1 complex The protein (PDB 1bgy) is shown with a simulated lipid bilayer showing the transmembrane parts and some of the cytosolic segments both above and below the membrane.

References

Doyle, D.A. *et al.*: **The structure of the potassium channel: molecular basis of K$^+$ conduction and selectivity.** *Science* 1998, **280**:69–77.

Ferguson, A.D. *et al.*: **Siderophore-mediated iron transport: crystal structure of FhuA with bound lipopolysaccharide.** *Science* 1998, **282**:2215–2220.

Heller, H. *et al.*: **Molecular dynamics simulation of a bilayer of 200 lipids in the gel and in the liquid crystal phases.** *J. Phys. Chem.* 1993, **97**:8343–8360.

Koebnik, R. *et al.*: **Structure and function of bacterial outer membrane proteins: barrels in a nutshell.** *Mol. Microbiol.* 2000, **37**:239–253.

Kyte, J. and Doolittle, R.F.: **A simple method for displaying the hydropathic character of a protein.** *J. Mol. Biol.* 1982, **157**:105–132.

Popot, J.L. and Engelman, D.M.: **Helical membrane protein folding, stability and evolution.** *Annu. Rev. Biochem.* 2000, **69**:881–922.

Popot, J.L. and Engelman, D.M.: **Membrane protein folding and oligomerization: the two-stage model.** *Biochemistry* 1990, **29**:4031–4037.

von Heijne, G.: **Recent advances in the understanding of membrane protein assembly and structure.** *Q. Rev. Biophys.* 1999, **32**:285–307.

Xia, D. *et al.*: **Crystal structure of the cytochrome bc1 complex from bovine heart mitochondria.** *Science* 1997, **277**:60–66.

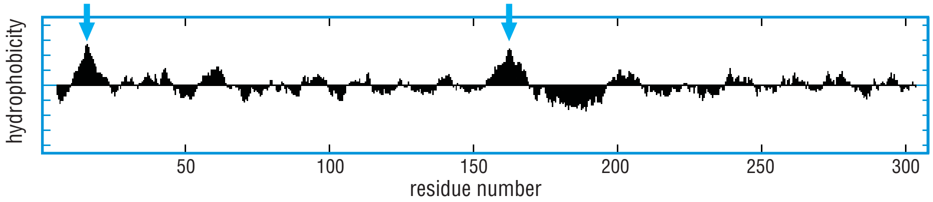


Figure 1-31 Hydropathy plot of the *Rhizobium meliloti* protein DctB The plot shows that two membrane-spanning alpha-helical regions are predicted. The plot represents the average hydrophobicity of an eight-residue moving window that slides along the sequence.

Because the backbone hydrogen bonds of an alpha helix are local, alpha helices are by far the most common secondary structure element in membrane proteins (Figure 1-30). As the translation per residue in a helix is 1.5 Å, a stretch of about 20 consecutive hydrophobic residues can form a helix that spans the bilayer if the helix axis is not tilted with respect to the membrane plane. Such stretches are easily recognized in protein sequences and are considered diagnostic for internal membrane proteins in analysis of genome sequences, because they do not occur frequently in soluble proteins. Figure 1-31 illustrates a hydropathy plot (plot of mean residue hydrophobicity) as a function of sequence for a carboxylic acid transport sensor (DctB) in the nitrogen-fixing bacterium *Rhizobium meliloti*. Two membrane-spanning alpha-helical regions are predicted. Many membrane-associated proteins are embedded in the lipid bilayer via only one or two membrane-spanning segments. These are always helical.

Beta sheets also occur in membrane proteins, but they are harder to recognize in the sequence. A beta strand 8–9 residues long would span the membrane (the translation per residue is about 3.5 Å) if the chain were perpendicular to the plane of the bilayer, but such stretches occur in soluble proteins and the variable twist of beta sheets makes it likely that the strand will be tilted. In those membrane proteins in which beta sheets have been found so far, they are antiparallel sheets with short polar turns. Because the edge strands in a beta sheet that is embedded in a membrane would have many unsatisfied hydrogen-bond donor and acceptor groups in their backbones, all such sheets examined to date form closed barrels with the first and last strands hydrogen bonded to each other (Figure 1-32). These beta sheets will have hydrophobic side chains covering their exterior surface, but can have polar or charged side chains lining the interior of the barrel. Such barrels seem to be used primarily as channels to permit water or ions to diffuse across the membrane. Channels can also be made from primarily helical proteins, as in the case of the potassium channel (Figure 1-33).

No integral membrane proteins with both helical and beta-sheet secondary structure have yet been found. There is good reason to expect that these are less common than all-helical or all-beta types: the need to hydrogen bond the polar groups on the edge strands of a beta sheet would be difficult to satisfy in a mixed structure. At present there are too few membrane protein structures determined to permit us to generalize with confidence on this point, or to allow creation of a detailed taxonomy of membrane protein fold families.

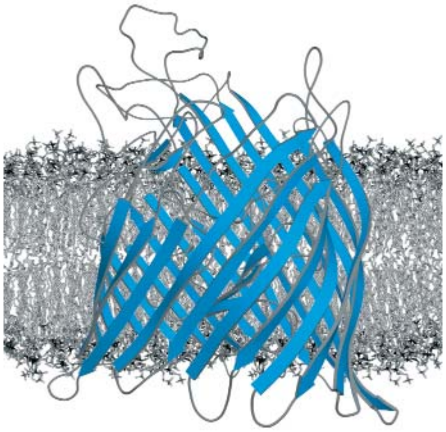


Figure 1-32 The three-dimensional structure of the all-beta transport protein FhuA The protein (PDB 1by3) is shown with a simulated lipid bilayer in the position it would occupy in the bilayer. The beta strands form a barrel that serves as a pore in the membrane, with hydrophobic side chains on the outside of the barrel and polar side chains lining the pore.

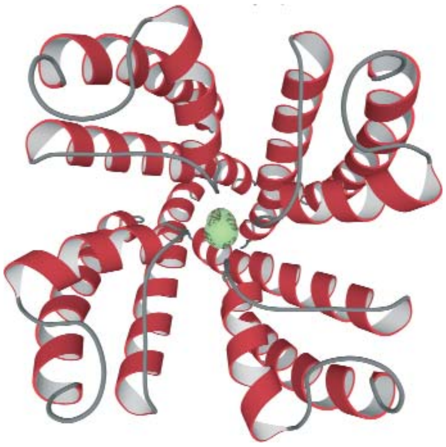


Figure 1-33 Three-dimensional structure of the bacterial potassium channel The protein is a homotetramer with a single channel formed at the interface of the four subunits. Although the protein is primarily helical, the pore of the channel is formed, in part, by two extended strands (grey) from each subunit. Shown here is the view looking down the channel from the extracellular side of the membrane: the outer ends of the helices in the figure are at the top of the chanel looking down, and the pore-forming strands can be seen looping inward. Two potassium ions are depicted as green spheres within the pore. (PDB 1bl8)

Tool for producing hydropathy plots on the Internet:

http://arbl.cvmbs.colostate.edu/molkit/hydropathy/index.html

Membrane models on the Internet:

http://www.lrz-muenchen.de/~heller/membrane/membrane.html

The folded protein is a thermodynamic compromise

Protein tertiary structure is maintained by the sum of many weak forces, some of which are stabilizing and some of which are destabilizing, some of which are internal to the protein and some of which are between the protein and its environment. The net effect is a folded structure that is only marginally stable in water at room temperature.

The contributions of the forces to protein stability are usually quantified in terms of the energy associated with any one of them. The heat released when such an interaction is formed in an isolated system is the **enthalpy** of the bond. However, bond enthalpies do not give a complete picture of the energetics of interactions in biological systems, in part because they neglect the contributions of water. Water plays two major roles in modulating the strengths of weak interactions. First, interactions between polar groups contribute only the difference in enthalpy between the groups when they are bonded to each other and the same groups when they are bonded to water. Because the interactions of water molecules with, for example, hydrogen-bond donors and acceptors are often similar, and of nearly equivalent enthalpy, to those that these groups can make to one another, the net enthalpy term is small.

Second, the contribution of water to the **entropy** of a weak interaction is also considerable. Entropy is a measure of randomness or disorder. The second law of thermodynamics states that spontaneous processes such as protein folding tend to increase the total entropy of a system plus its surroundings. An example of the importance of entropic contributions from water is found in the hydrophobic effect. Nonpolar groups in water tend to be surrounded by a cluster of water molecules that are more ordered than in the normal structure of liquid water (Figure 1-34). When such hydrophobic groups clump together, expelling water, the water molecules that are released undergo an increase in entropy. Although there will be a shell of ordered water around the clump, the total number of these ordered water molecules will be smaller than if all the hydrophobic groups were exposed to solvent individually. The gain in solvent entropy that results from the association of hydrophobic groups together is the driving force behind the hydrophobic effect. Thus, in evaluating the energetic consequences of a weak interaction, the changes in entropy of the interacting groups and the water around them all need to be considered simultaneously.

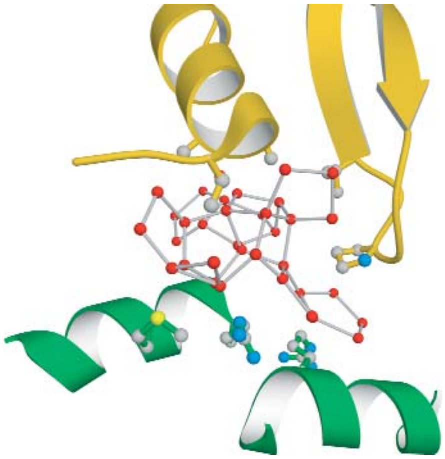


Figure 1-34 Illustration of the ordered arrays of water molecules surrounding exposed hydrophobic residues in bovine pancreatic ribonuclease A Such waters (red) often form pentagonal arrays. It is thought that this ordering of water around exposed polar groups is the driving force for the hydrophobic effect. (PDB 1dyg)

Stability is defined as a net loss of **free energy**, a function of the combined effects of entropy and enthalpy. Such a loss may result predominantly from a loss of enthalpy when a bond forms, or predominantly from a gain in entropy when the disorder of a system (protein) plus its surroundings (water) increases, or from a balance between enthalpy and entropy changes. Most weak interactions release about 4–13 kJ/mole of free energy when they occur in water and therefore contribute only a small amount to the total stability of a protein. However, there are a large number of them, adding up to a very large free-energy decrease when secondary and tertiary structures form.

Finally, even though many hundreds of hydrogen bonds and van der Waals interactions occur in a folded protein, the net free energy of stabilization of most folded proteins—the difference in free energy between the folded and unfolded states—is actually rather small, about 21–42 kJ/mole, or only about 10 times the average thermal energy available at physiological temperature. Most folded proteins are marginally stable because the free energy released when hundreds of weak interactions form is almost exactly counterbalanced by the enormous loss of conformational flexibility (loss of entropy) that occurs when the unfolded chain folds into a compact, ordered structure. A folded protein is a thermodynamic compromise.

Definitions

denaturant: a chemical capable of unfolding a protein in solution at ordinary temperatures.

denatured state: the partially or completely unfolded form of a biological macromolecule in which it is incapable of carrying out its biochemical and biological functions.

enthalpy: a form of energy, equivalent to work, that can be released or absorbed as heat at constant pressure.

entropy: a measure of the disorder or randomness in a molecule or system.

free energy: a function, designed to produce a criterion for spontaneous change, that combines the **entropy** and **enthalpy** of a molecule or system. Free energy decreases for a spontaneous process, and is unchanged at equilibrium.

mesophilic: favoring moderate temperatures. Mesophilic organisms normally cannot tolerate extremes of heat or cold. Mesophilic enzymes typically denature at moderate temperatures (over 40 °C or so).

temperature-sensitive: losing structure and/or function at temperatures above physiological or room temperature. A temperature-sensitive mutation is a change in the amino-acid sequence of a protein that causes the protein to inactivate or fail to fold properly at such temperatures.

thermophilic: favoring high temperatures. A thermophilic organism is one that requires high temperatures (above approximately 50 °C) for survival. A thermophilic enzyme is one that functions optimally and is stable at temperatures at which **mesophilic** proteins denature.

Protein structure can be disrupted by a variety of agents

High temperatures break the weak interactions that stabilize the folded or native form of a protein and eventually convert the structure to a largely unfolded or denatured one, in which these interactions are replaced by hydrogen bonds with water. The **denatured state** is usually defined empirically, either by loss of biological or biochemical activity, or by spectroscopic signals characteristic of an unfolded polypeptide (Figure 1-35). Because the free-energy difference between the native and denatured states is small, loss of a single interaction in the native state can sometimes bring the free-energy difference close to the thermal energy available at ordinary temperatures. A mutation that causes a normally stable protein to unfold at relatively low temperatures is called a **temperature-sensitive** (ts) mutation. These mutations are widely used in experimental biology to test the function of a protein in cells by raising the temperature and thereby disabling the mutant protein. Similarly, just a few additional interactions can greatly increase the stability of a protein at elevated temperatures, producing a protein that is more able to withstand heating, prolonged storage or shipping for industrial applications. One example is the thermostable Taq DNA polymerase used in PCR.

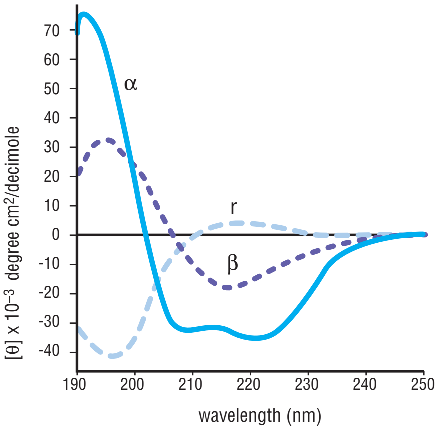


Figure 1-35 Computed circular dichroism spectra for the evaluation of protein conformation Circular dichroism spectrum of poly(Lys) in the alpha-helical (α), anti-parallel beta sheet (β) and random coil (r) conformations. (From Greenfield, N.J. and Fasman, G.D.: *Biochemistry* 1969, **8**:4108–4116.)

Another way to unfold a protein is by the use of chemical **denaturants** such as urea or guanidinium hydrochloride, or detergents like SDS. In contrast to thermal denaturation, these compounds are thought to unfold proteins in large part by competing for hydrogen bonds with the polar groups of the backbone and side chains.

Some proteins are naturally very stable to thermal or chemical denaturation. One important class of very stable proteins consists of those from microorganisms that normally live at high temperatures. These so-called **thermophilic** proteins sometimes retain their structure—and activity—at temperatures approaching the boiling point of water. No single type of interaction or effect accounts for such hyperthermostability. Comparisons of structures of proteins with similar sequences and functions isolated from thermostable microbes and their **mesophilic** counterparts show a variety of differences: some thermophilic proteins have more salt bridges, while others appear to have more hydrophobic interactions and shorter protruding loops, and so forth. There seem to be many ways to achieve the same effect, and when this is the case it is usual in biology to find all of them.

The marginal stability of protein tertiary structure allows proteins to be flexible

Above absolute zero, all chemical bonds have some flexibility: atoms vibrate and chemical groups can rotate relative to each other. In proteins, because most of the forces that stabilize the native state are noncovalent, there is enough thermal energy at physiological temperatures for weak interactions to break and reform frequently. Thus a protein molecule is more flexible than a molecule in which only covalent forces dictate the structure. Protein structures continuously fluctuate about the equilibrium conformation observed by techniques such as X-ray diffraction and nuclear magnetic resonance (NMR) (Figure 1-36). Thermally driven atomic fluctuations range in magnitude from a few hundredths of an Ångstrom for a simple atomic vibration to many Ångstroms for the movement of a whole segment of a protein structure relative to the rest. These fluctuations are large enough to allow small molecules such as water to penetrate into the interior of the protein. They are essential for protein functions such as ligand binding and catalysis, for they allow the structure to adjust to the binding of another molecule or to changes in the structure of a substrate as a reaction proceeds.

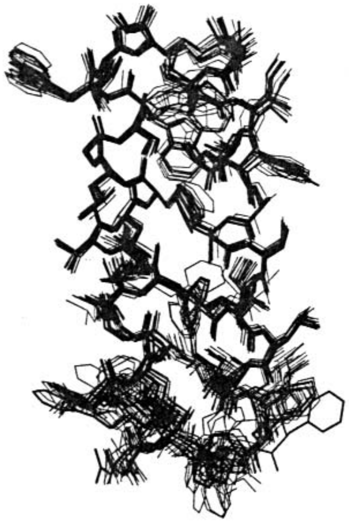


Figure 1-36 Results of a molecular dynamics simulation of two interacting alpha helices The diagram shows fluctuations of portions of the structure. Some parts of the protein seem to be more mobile than others.

References

Dill, K.A and Bromberg, S.: *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology* (Garland, New York and London, 2003).

Ferreira, S.T. and De Felice, F.G.: **Protein dynamics, folding and misfolding: from basic physical chemistry to human conformational diseases.** *FEBS Lett.* 2001, **498**:129–134.

Greenfield, N.J. and Fasman, G.D.: **Computed circular dichroism spectra for the evaluation of protein conformation.** *Biochemistry* 1969, **8**:4108–4116.

Jaenicke, R.: **Stability and stabilization of globular proteins in solution.** *J. Biotechnol.* 2000, **79**:193–203.

Kauzmann, W.: **Some factors in the interpretation of protein denaturation.** *Adv. Protein Chem.* 1959, **14**:1–63.

Sharp, K.A. and Englander, S.W.: **How much is a stabilizing bond worth?** *Trends Biochem. Sci.* 1994, **19**:526–529.

1-13 Protein Stability: Post-Translational Modifications

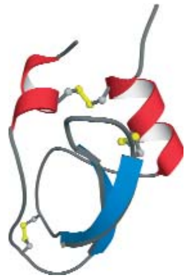


Figure 1-37 The structure of the small protein bovine pancreatic trypsin inhibitor, BPTI The three disulfide bonds are yellow, beta strands are blue, and alpha helices are red. If these disulfide bonds are reduced this small protein unfolds, presumably because there is not enough secondary structure to stabilize the fold without them. (PDB 1bpi)

Covalent bonds can add stability to tertiary structure

Noncovalent forces are the principal interactions that stabilize protein tertiary structure but they are not the only ones. Many proteins also are stabilized by additional, covalent interactions that provide a form of cross-linking between segments of secondary structure in the native state. The most common of these covalent bonds is the disulfide bridge that can form between two cysteine side chains that are brought close together by the tertiary structure (Figure 1-37). Formation of a disulfide bridge (also called an S–S bridge or a disulfide bond) involves the oxidation of the two sulfhydryl groups as a coupled redox reaction in the endoplasmic reticulum. Conversely, the bridge can be broken by reduction. Thus, S–S bridges are uniquely sensitive to their environment. They are not found in most intracellular proteins, because the environment inside the cell is highly reducing, but they are common in proteins that are secreted from that environment into the oxidizing conditions found outside the cell.

The second most common cross-linking interaction in proteins is the coordination of a metal ion to several protein side chains; the **coordinate covalent bonds** between the protein and the metal ion form a type of internal metal chelate (Figure 1-38). The strength of the binding of the metal ion to the protein varies from very loose (**K_d** of mM) to very tight (K_d of nM) depending on the nature of the metal ion and the protein ligands. Not all of the ligands are contributed by the protein; one or more water molecules can also occur in the coordination sphere. A given protein can have more than one stabilizing metal ion binding site. The metal ions that most commonly form such chelates are calcium (Ca^{2+}) and zinc (Zn^{2+}), although monovalent cations such as potassium and sodium can also function in this way. These stabilizing metal ions carry out no chemistry and are distinct from metal ions in active sites of metalloproteins which carry out the biochemical function of a protein (see below). Sometimes when these metal ions are removed by chelating agents such as EDTA, the resulting protein remains folded, although less stable, under physiological conditions. In other cases, removal of the metal ions from the protein leads to denaturation.

Finally, some proteins are stabilized by the covalent binding of a dissociable organic or organometallic **cofactor** at the active site, or by the formation of a covalent cross-link between amino-acid side chains that is different from a disulfide bridge (Figure 1-39a). So far, these cross-links have always been found at the active site, where they contribute critically to the chemical function of the protein. The covalent bond between cofactor and protein may be formed with the organic part of some cofactors as in the case of D-amino acid aminotransferase (DaAT) (Figure 1-39a), or with a metal ion that is an integral part of some cofactors as in the case of vitamin B12, chlorophyll, and the heme group in some heme-containing proteins (Figure 1-39b), or with both, as in the case of the heme group in cytochrome c (Figure 1-39c). However, in some cases the cofactor is not a separable molecule, but is created by the chemical cross-linking of two amino-acid side chains, as in the case of the redox active cofactor PQQ (Figure 1-39d) and the bioluminescent chromophore in green fluorescent protein. Although many proteins that are stabilized in this way remain folded when the cofactor is dissociated, some do not.

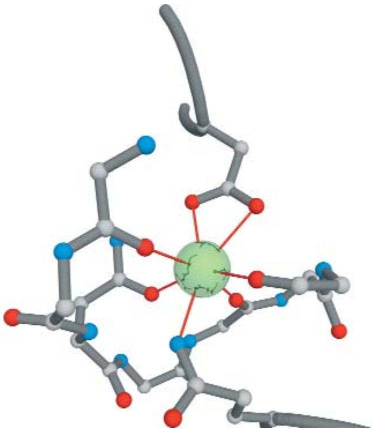


Figure 1-38 Stabilization by coordinate covalent bonds Close-up of one of the three calcium ion binding sites in the bacterial protein subtilisin, showing the coordination of the metal ion by the protein. This site is used only for protein stability, not for catalysis. Removal of this metal ion significantly destabilizes the protein. (PDB 1sca)

Post-translational modification can alter both the tertiary structure and the stability of a protein

Proteins in eukaryotic cells that are destined to be placed on the cell surface or secreted into the environment are often modified by the covalent attachment of one or more chains of carbohydrate molecules at specific serine, threonine or asparagine residues. This is known as **glycosylation**, and along with the covalent attachment of lipids, is among the most important

Definitions

cofactor: an organic or organometallic molecule that binds to a protein and provides an essential chemical function for that protein.

coordinate covalent bond: a bond formed when a lone pair of electrons from an atom in a ligand is donated to a vacant orbital on a metal ion.

glycosylation: the post-translational covalent addition of sugar molecules to asparagine, serine or threonine residues on a protein molecule. Glycosylation can add a single sugar or a chain of sugars at any given site and is usually enzymatically catalyzed.

K_d: the dissociation constant for the binding of a ligand to a macromolecule. Typical values range from 10^{-3} M to 10^{-10} M. The lower the K_d, the tighter the ligand binds.

limited proteolysis: specific cleavage by a protease of a limited number of the peptide bonds in a protein substrate. The fragments thus produced may remain associated or may dissociate.

***N*-acetylation:** covalent addition of an acetyl group from acetyl-CoA to a nitrogen atom at either the amino terminus of a polypeptide or in a lysine side chain. The reaction is catalyzed by *N*-acetyltransferase.

phosphorylation: covalent addition of a phosphate group, usually to one or more amino-acid side chains on a protein, catalyzed by protein kinases.

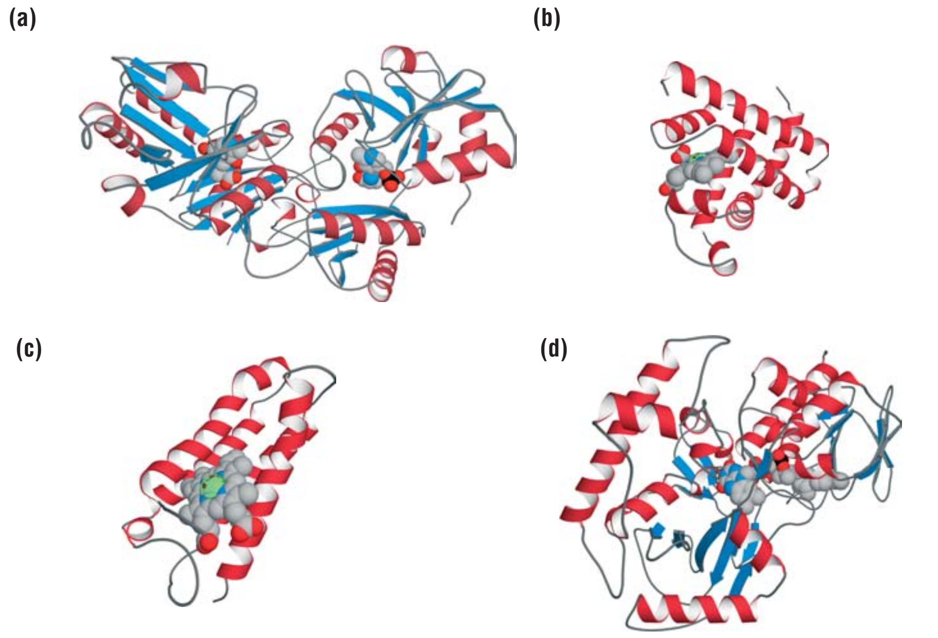


Figure 1-39 Examples of stabilization by cofactor binding **(a)** DaAT is a dimeric protein with covalently bound pyridoxal phosphate bound in each subunit. (PDB 3daa) **(b)** The heme group of myoglobin is attached to the protein by a coordinate covalent bond between a histidine side chain and the heme iron. (PDB 1a6k) **(c)** In cytochrome c the heme group is attached to the protein in two places: by a coordinate covalent bond between the sulfur of a methionine residue and the heme iron, and by covalent attachment of the porphyrin cofactor to the protein. (PDB 1a7v) **(d)** The cofactor PQQ in polyamine oxidases is formed by the reaction of two side chains with each other, leaving it attached to the protein. (PDB 1b37) In all cases, the apoprotein (the protein without the cofactor) is thermally less stable and/or more susceptible to proteolytic digestion than the protein with the covalently attached cofactor.

of the post-translational modifications. Both are enzymatically catalyzed by specific enzymes. Probably the most important for protein stability is glycosylation. The roles of the attached sugars are not known precisely for most proteins, but on proteins expressed on the surface of blood cells, for example, they are believed to be important in preventing the cells from sticking to one another or to vessel walls and obstructing blood flow. Some protein glycosylation sites are involved in protein–protein recognition. In many instances the removal of carbohydrates from glycosylated proteins leads either to unfolding or to aggregation. This characteristic limits the use of prokaryotic systems in the application of recombinant DNA technology to such proteins, since prokaryotes do not carry out this post-translational modification. It is generally believed that glycosylation does not alter the tertiary structure of a protein, but can significantly influence thermal stability, stability to degradation and quaternary structure.

There are other forms of post-translational modification that can also alter the stability—and in some cases the tertiary structure—of folded proteins. In contrast to glycosylation, these modifications usually alter the function of the protein. Some of these modifications, such as **phosphorylation** and ***N*-acetylation**, are reversible and thus can act as conformational switches. The functional consequences of post-translational modification will be discussed in Chapter 3. Others, such as **limited proteolysis**, the cleavage of the polypeptide chain at one or more sites, are irreversible, and thereby change the structure and function of a protein permanently. Limited proteolysis, for example, sometimes generates an active protein from an inactive precursor. Common kinds of post-translational modification that influence the stability of proteins are summarized in Figure 1-40. Although many types of post-translational modification can increase or decrease the stability of a protein, such modifications often also serve additional functions, such as signaling or activation of catalysis.

Most Common Post-translational Modifications

Reversible	Irreversible
disulfide bridge	cofactor binding
cofactor binding	proteolysis
glycosylation	ubiquitination
phosphorylation	peptide tagging
acylation	lysine hydroxylation
ADP-ribosylation	methylation
carbamylation	
N-acetylation	

Figure 1-40 Table of post-translational modifications affecting protein stability

References

Imperiali, B. and O'Connor, S.E.: **Effect of *N*-linked glycosylation on glycopeptide and glycoprotein structure.** *Curr. Opin. Chem. Biol.* 1999, **3**:643–649.

Morris, A.J. and Malbon, C.C.: **Physiological regulation of G protein-linked signaling.** *Physiol. Rev.* 1999, **79**:1373–1430.

Palade, G.E.: **Protein kinesis: the dynamics of protein trafficking and stability.** *Cold Spring Harbor Symp. Quant. Biol.* 1995, **60**:821–831.

Rattan, S.I. *et al.*: **Protein synthesis, posttranslational modification, and aging.** *Annls N.Y. Acad. Sci.* 1992, **663**:48–62.

Tainer, J.A. *et al.*: **Protein metal-binding sites.** *Curr. Opin. Biotechnol.* 1992, **3**:378–387.

Zhang, T. *et al.*: **Entropic effects of disulfide bonds on protein stability.** *Nat. Struct. Biol.* 1994, **1**:434–438.

For a database of all known post-translational modifications, organized by type of amino acid modified, see:

http://pir.georgetown.edu/pirwww/dbinfo/resid.html

1-14 The Protein Domain

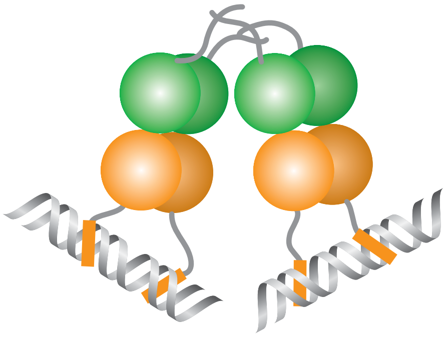


Figure 1-41 **Schematic diagram of the Lac repressor tetramer binding to DNA** Each monomer of the Lac repressor is made up of a tetramerization domain (green) and a DNA-binding domain (orange).

Globular proteins are composed of structural domains

Some proteins, such as the keratin of hair, are fibrous: their polypeptide chains are stretched out in one direction. Most proteins, however, are globular: their polypeptide chains are coiled up into compact shapes. Since proteins range in molecular weight from a thousand to over a million, one might have thought that the size of these globular folds would increase with molecular weight, but this is not the case. Proteins whose molecular weights are less than about 20,000 often have a simple globular shape, with an average molecular diameter of 20 to 30 Å, but larger proteins usually fold into two or more independent globules, or structural domains. A **domain** is a compact region of protein structure that is often, but not always, made up of a continuous segment of the amino-acid sequence, and is often capable of folding stably enough to exist on its own in aqueous solution. The notion that the domains of large proteins are independently stable has been verified by cloning the corresponding DNA sequences and expressing them independently. Not only do many of them form stable, folded structures in solution, they often retain part of the biochemical function of the larger protein from which they are derived. The bacterial Lac repressor, which is a tetrameric protein that binds tightly to a specific DNA sequence, is a good example (Figure 1-41). One of the two domains in the monomer can dimerize by itself, and binds to DNA with an affinity that nearly matches that of the intact protein. The function of the other domain is to form the tetramer by making protein–protein interactions; by itself it tetramerizes but does not bind to DNA.

Not all domains consist of continuous stretches of polypeptide. In some proteins, a domain is interrupted by a block of sequence that folds into a separate domain, after which the original domain continues. The enzyme alanine racemase has an interrupted domain of this type (Figure 1-42).

Domains vary in size but are usually no larger than the largest single-domain protein, about 250 amino acids, and most are around 200 amino acids or less. Forty-nine per cent of all domains are in the range 51 to 150 residues. The largest single-chain domain so far has 907 residues, and the largest number of domains found in a protein to date is 13. As the domains in a protein associate with one another by means of the same interactions that stabilize their internal structures, what is true for domains is true for whole proteins, and vice versa: the same structural principles apply to both.

Domains have hydrophobic cores

Hydrophobic cores appear to be essential for the stability of domains. Concentrating hydrophobic groups in the core is energetically favorable because it minimizes the number of unfavorable interactions of hydrophobic groups with water, and maximizes the number of van der Waals interactions the hydrophobic groups make with each other. All the different polypeptide folding patterns presented in this book can be thought of as alternative solutions to a single problem: how to fold a polypeptide chain so as to maximize the exposure of its hydrophilic groups to water while minimizing the exposure of its hydrophobic groups. It is the presence of a hydrophobic core that usually allows protein domains to fold stably when they are expressed on their own.

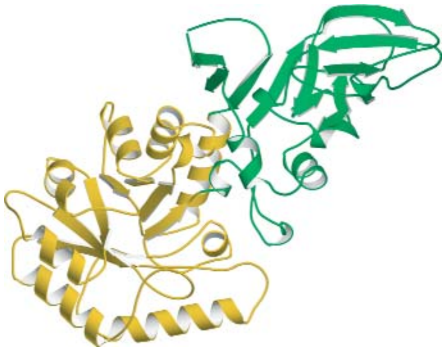


Figure 1-42 **Structure of alanine racemase** The diagram shows that one of its structural domains (yellow) is interrupted by insertion of another domain (green).

Definitions

domain: a compact unit of protein structure that is usually capable of folding stably as an independent entity in solution. Domains do not need to comprise a contiguous segment of peptide chain, although this is often the case.

Multidomain proteins probably evolved by the fusion of genes that once coded for separate proteins

There are many examples of proteins with two or more domains of nearly identical structure. The *Escherichia coli* thioesterase, for example, is organized into two equal-sized domains of almost identical structure (Figure 1-43a). The domains can be overlaid on top of one another with almost perfect overlap in the paths of their polypeptide chains, except for a few of the external loops and the polypeptide that links the domains.

It is likely that this protein, and others that have internal similarity of structure, evolved by gene duplication. A single gene coding for a protein resembling one domain is assumed to have been duplicated in tandem, and the two genes to have fused so that their sequences are expressed as a single polypeptide. In some proteins the duplicated domains retain some sequence identity, but in other proteins they do not. Whether duplicated domains display sequence identity depends on how long ago the duplication occurred, and the nature of the functional constraints that guided their divergence. The more ancient the gene duplication, the more time for mutation to obscure the sequence relationship. In thioesterase, for example, completely different sequences give rise to the same overall fold. The original sequence identity between the two thioesterase domains has been largely obliterated by random mutations over millennia. Sometimes the original gene that is duplicated can be identified. In the case of thioesterase, the protein thioester dehydrase, which carries out a similar function to thioesterase, is a homodimeric protein. Each monomer has the same fold as one of the domains of thioesterase (Figure 1-43b).

Sometimes, gene duplication can occur within a single structural domain. An example is shown in Figure 1-44, which depicts the fold of the eye-lens protein gamma-crystallin. The protein as a whole is made up of two similar domains, which are 40% identical in sequence. Closer inspection of the two domains reveals that they, too, are made up of two essentially identical halves. Each eight-stranded beta-sheet domain is composed of two four-stranded antiparallel sheets of the same topology. Internal symmetry such as this does not prove tandem duplication of a smaller gene, as this arrangement of beta strands could simply be a stable configuration for two antiparallel beta sheets to pack against one another. For the gamma-crystallins, however, there is enough residual sequence identity to justify the conclusion that the individual domain did indeed evolve by gene duplication and fusion. Additional evidence for this evolutionary history can be found in the sequences of the genes for other crystallins. The gene for mouse beta-crystallin, a protein closely related in amino-acid sequence to gamma-crystallin, is divided into four exons, and each exon codes for one four-stranded beta-sheet segment. In other words, the positions of the three introns correspond to the junctions between each of the subdomains that were presumably encoded by the primordial gene from which the modern genes arose by duplication. This is persuasive evidence for gene duplication in crystallin evolution.

If the fusion of tandem genes can account for proteins with internal symmetry, then it is likely that this mechanism also explains the origin of multidomain proteins where the domains are structurally unrelated, as in the Lac repressor (see Figure 1-41).

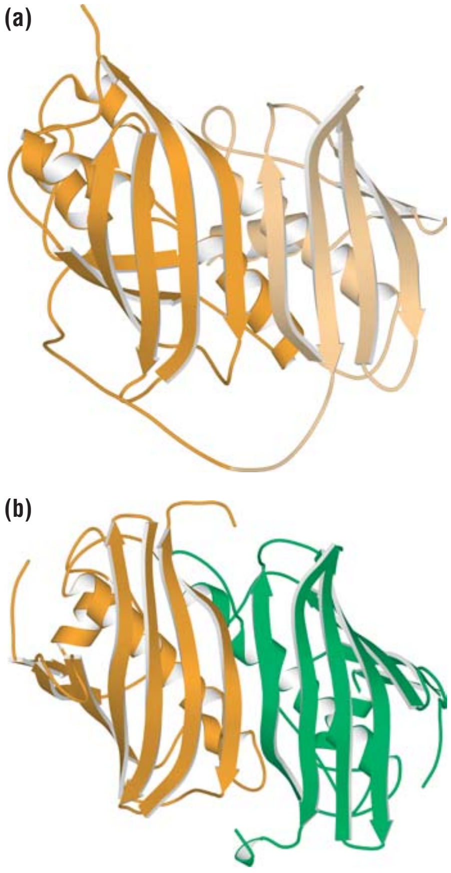


Figure 1-43 Structures of thioesterase and thioester dehydrase **(a)** Structure of *E. coli* thioesterase, a protein composed of two nearly identical domains (dark and light orange) fused together. (PDB 1c8u) Each domain resembles the subunit of thioester dehydrase (PDB 1mkb) **(b)**, a protein composed of two identical subunits.

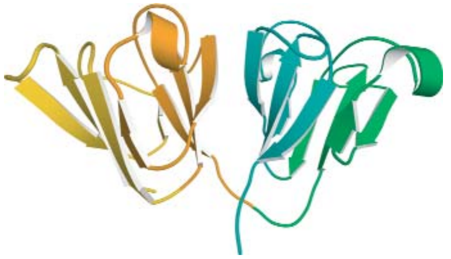


Figure 1-44 Structure of gamma-crystallin Gamma-crystallin is composed of two nearly identical domains. Each domain is also made up of two nearly identical halves. (PDB 1gcs)

References

Burchett, S.A.: **Regulators of G protein signaling: a bestiary of modular protein binding domains.** *J. Neurochem.* 2000, **75**:1335–1351.

Campbell, I.D. and Downing, A.K.: **Building protein structure and function from modular units.** *Trends Biotechnol.* 1994, **12**:168–172.

Dengler, U. *et al.*: **Protein structural domains: analysis of the 3Dee domains database.** *Proteins* 2001, **42**:332–344.

Hawkins, A.R. and Lamb, H.K.: **The molecular biology of multidomain proteins. Selected examples.** *Eur. J. Biochem.* 1995, **232**: 7–18.

Hegyi, H. and Bork, P.: **On the classification and evolution of protein molecules.** *J. Prot. Chem.* 1997, **16**: 545–551.

Richardson, J.S.: **The anatomy and taxonomy of protein structure.** *Adv. Protein Chem.* 1981, **34**:167–339.

Thornton, J.W. and DeSalle, R.: **Gene family evolution and homology: genomics meets phylogenetics.** *Annu. Rev. Genomics Hum. Genet.* 2000, **1**:41–73.

1-15 The Universe of Protein Structures

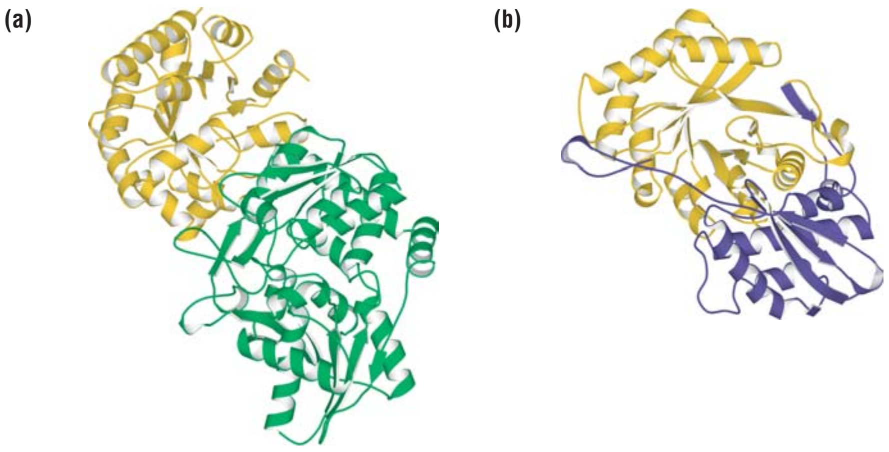


Figure 1-45 Structures of tryptophan synthase and galactonate dehydratase **(a)** Tryptophan synthase (PDB 1ttp). **(b)** Galactonate dehydratase. Both proteins have one domain that is an alpha/beta barrel (yellow), even though the other domains in both proteins are very different from one another. There is no sequence similarity or functional relationship between the similar domains.

The number of protein folds is large but limited

As more protein structures are determined experimentally, it is increasingly found that new structures look like old structures. Sometimes an entire "new" structure will resemble that of another protein whose structure is already known. In most cases, however, the overall polypeptide fold of the protein will be "new", but the structure will be divisible into a number of domains, at least one of which resembles the tertiary structure previously observed in another protein (Figure 1-45).

It appears that the number of different protein folds in nature is limited. They are used repeatedly in different combinations to create the diversity of proteins found in living organisms. Building new proteins, it would seem, is like assembling a four-course dinner from a set of *á la carte* choices—the possible **domain folds**. Although the size of the menu is not yet known, it is much smaller than the total number of gene products—perhaps as small as a few thousands—and almost all of the tertiary structure folds that have been discovered so far are known to appear in many different proteins. Thus, a complete protein can be described by specifying which folds each domain has and how they interact with each other. This approach to describing protein structure is appealing both for its logical form and because it reflects our prejudice that proteins fold up domain by domain.

Protein structures are modular and proteins can be grouped into families on the basis of the domains they contain

Although many proteins are composed of a single structural domain, most proteins are built up in a modular fashion from two or more domains fused together. In some cases, each domain has a characteristic biochemical function and the function of the entire protein is determined by the sum of the individual properties of the domains. Proteins involved in signal transduction and cell-cycle control are often constructed in this fashion (Figure 1-46). One example is the cancer-associated kinase Src-Lck, which has a catalytic kinase domain that phosphorylates proteins on tyrosine residues, an SH2 domain that binds phosphotyrosine residues, an SH3 domain that recognizes proline-rich sequences, and a phosphotyrosine region that can interact with its own or other SH2 domains. When the modules that form proteins of this type fold and function independently, the order in which they occur in the polypeptide is not necessarily always important. Thus module swapping and the recruitment of new functions by adding modules is often simple, either through the course of evolution or artificially.

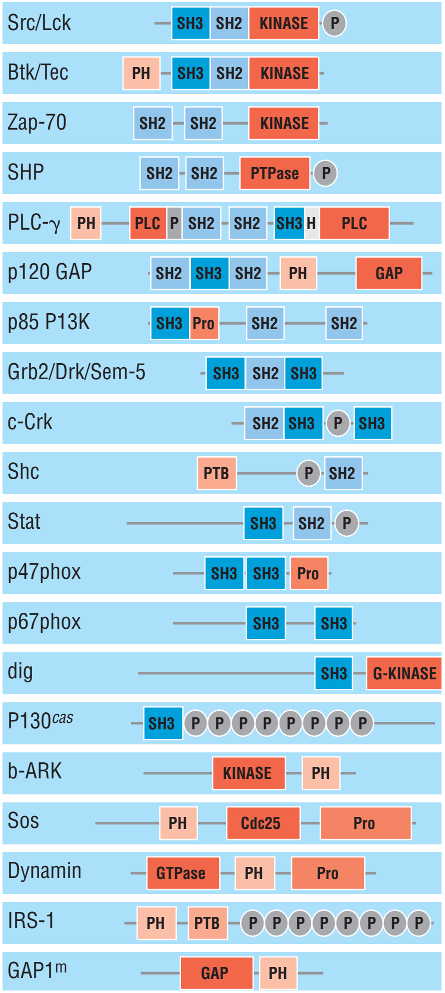


Figure 1-46 Schematic diagram of the domain arrangement of a number of signal transduction proteins The different modules have different functions; Pro = proline-rich regions that bind SH3 domains; P = phosphotyrosine-containing regions that bind SH2 domains; PH = pleckstrin homology domains that bind to membranes; PTPase = phosphatase domain; kinase = protein kinase domain; G-kinase = guanylate kinase domain; GAP = G-protein activation domain; PLC = phospholipase C catalytic domain. The function of the individual modules is sometimes, but not always, independent of the order in which they appear in the protein.

Definitions

domain fold: the particular topographical arrangement of secondary structural elements that characterizes a single domain. Examples are an antiparallel arrangement of four helices in a four-helix bundle, or an open twisted beta sandwich with a particular sequence that binds nucleotides.

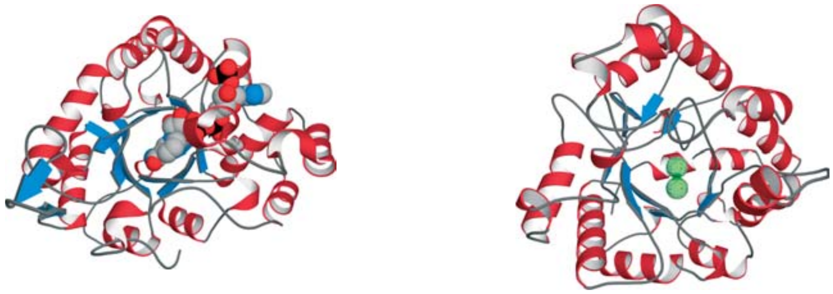


Figure 1-47 Structures of aldose reductase (left) and phosphotriesterase (right) Although the overall folds are very similar, aldose reductase uses NADPH to reduce sugars, while phosphotriesterase hydrolyzes phosphate groups using a bimetallic cofactor. (PDB 1ads and 1dpm)

Because sequence determines structure, which in turn determines function, it is tempting to classify proteins whose function cannot be recognized from sequence similarity alone into families based on the structures of the domains they contain. Often this approach is successful: proteins with a kinase domain are nearly always kinases; proteins with an alpha/beta hydrolase domain nearly always hydrolyze small-molecule substrates, and so forth. But often it is not the case that structural families share a common function. There are hundreds of proteins that contain a particular eight-stranded parallel beta barrel with surrounding alpha helices called a TIM barrel, but even two very similar single-domain TIM-barrel proteins can have completely different biochemical functions (Figure 1-47). Nor is it always the case that all proteins that perform the same biochemical function will have the same domains: amino-acid transamination, for example, can be catalyzed by two completely different folds (Figure 1-48). The coupling between overall structure and function can be quite loose. Nevertheless, grouping proteins into families on the basis of their domain architecture is, at a minimum, very useful for studying the way new protein functions may have evolved.

The modular nature of protein structure allows for sequence insertions and deletions

Deletions and insertions of amino acids can obscure evolutionary relationships, but how is it that long stretches of amino acids, sometimes an entire domain, can be inserted in or deleted from a protein sequence (see Figure 1-42) without disrupting the basic structure of a domain? The answer lies in the nature of domain folds. Domains are made up of secondary structure elements that are packed together to form tertiary structure. The loops that join the helices and sheets in most proteins are usually located on the surface, and often make few contacts with the rest of the domain. Within a given protein family, insertions and deletions nearly always occur in these surface loops, where variation in length has little effect on the packing of helices and sheets. Indeed, a rough rule of domains, and ultimately of the structural evolution of proteins, is that the framework tends to remain fairly constant in both sequence and structure while the loops change a great deal over evolutionary time. In the case of immunoglobulin (see section 1-10), the loops form the antigen-binding site and variation due to somatic recombination and mutation of immunoglobulin genes accounts for the diversity of antibody molecules.

Many models for protein evolution propose the shuffling of exon-coded segments to produce new protein molecules. Insertion of a new exon into an existing domain could change its properties dramatically, but of course the new molecule would still have to fold stably. Stable folding would be more likely if the new exon were inserted into a surface loop. Examination of intron/exon junctions in proteins whose three-dimensional structures are known shows that many exon boundaries do indeed occur in sequence positions corresponding to loops in the structure. Important exceptions include the immunoglobulins.

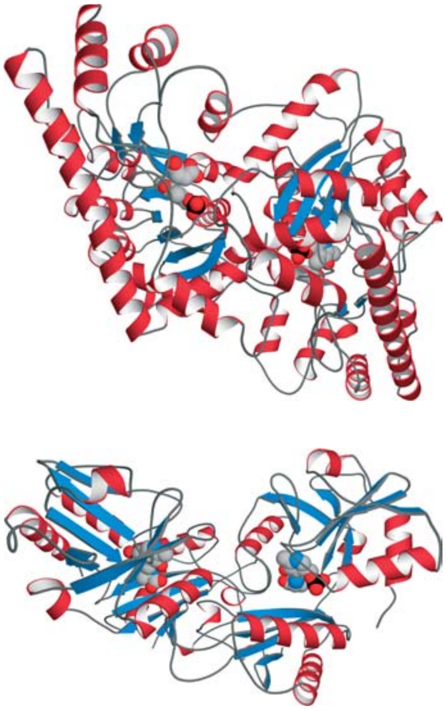


Figure 1-48 Structures of aspartate aminotransferase (top) and D-amino acid aminotransferase (bottom) Both enzymes catalyze the same reaction; but they have no structural similarity to each other at either the sequence or the tertiary level. Only the active sites, shown by the presence of the cofactor in space-filling representation, are very similar. (PDB 1yaa and 3daa)

References

Branden, C. and Tooze, J.: *Introduction to Protein Structure* 2nd ed. (Garland, New York, 1999).

Patthy, L.: **Genome evolution and the evolution of exon-shuffling.** *Gene* 1999, **238**:103–114.

Richardson, J.S.: **Introduction: protein motifs.** *FASEB J.* 1994, **8**:1237–1239.

Richardson, J.S. *et al.*: **Looking at proteins: representation, folding, packing and design.** *Biophys. J.* 1992, **63**:1185–1209.

Richardson, J.S. and Richardson, D.C.: **Principles and patterns of protein conformation** in *Prediction of Protein Structure and the Principles of Protein Conformation* Fasman, G.D. ed. (Plenum Press, New York, 1990), 1–98.

Salem, G.M. *et al.*: **Correlation of observed fold frequency with the occurrence of local structural motifs.** *J. Mol. Biol.* 1999, **287**:969–981.

Thornton, J.M. *et al.*: **Protein folds, functions and evolution.** *J. Mol. Biol.* 1999, **293**:333–342.

Internet resources on protein structure comparison and classification:

http://www.ebi.ac.uk/dali/

http://scop.mrc-lmb.cam.ac.uk/scop/

1-16 Protein Motifs

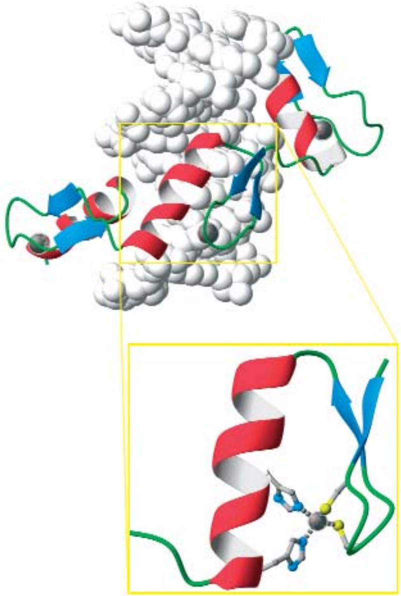


Figure 1-49 **Zinc finger motif** A fragment derived from a mouse gene regulatory protein is shown, with three zinc fingers bound spirally in the major groove of a DNA molecule. The inset shows the coordination of a zinc atom by characteristically spaced cysteine and histidine residues in a single zinc finger motif. The image is of Zif268. (PDB 1aay)

Protein motifs may be defined by their primary sequence or by the arrangement of secondary structure elements

The term **motif** is used in two different ways in structural biology. The first refers to a particular amino-acid sequence that is characteristic of a specific biochemical function. An example is the so-called zinc finger motif, CXX(XX)CXXXXXXXXXXXXHXXXH, which is found in a widely varying family of DNA-binding proteins (Figure 1-49). The conserved cysteine and histidine residues in this **sequence motif** form ligands to a zinc ion whose coordination is essential to stabilize the tertiary structure. Conservation is sometimes of a class of residues rather than a specific residue: for example, in the 12-residue loop between the zinc ligands, one position is preferentially hydrophobic, specifically leucine or phenylalanine. Sequence motifs can often be recognized by simple inspection of the amino-acid sequence of a protein, and when detected provide strong evidence for biochemical function. The protease from the human immunodeficiency virus was first identified as an aspartyl protease because a characteristic sequence motif for such proteases was recognized in its primary structure.

The second, equally common, use of the term motif refers to a set of contiguous secondary structure elements that either have a particular functional significance or define a portion of an independently folded domain. Along with the functional sequence motifs, the former are known generally as **functional motifs**. An example is the helix-turn-helix motif found in many DNA-binding proteins (Figure 1-50). This simple **structural motif** will not exist as a stably folded domain if expressed separately from the rest of its protein context, but when it can be detected in a protein that is already thought to bind nucleic acids, it is a likely candidate for the recognition element. Examples of structural motifs that represent a large part of a stably folded domain include the four-helix bundle (Figure 1-51), a set of four mutually antiparallel alpha helices that is found in many hormones as well as other types of proteins; the Rossmann fold, an alpha/beta twist arrangement that usually binds NAD cofactors; and the *Greek-key motif*, an all-beta-sheet arrangement found in many different proteins and which topologically resembles the design found on ancient vases. As these examples indicate, these structural motifs sometimes are suggestive of function, but more often are not: the only case here with clear functional implications is the Rossmann fold.

Identifying motifs from sequence is not straightforward

Because motifs of the first kind—sequence motifs—always have functional implications, much of the effort in bioinformatics is directed at identifying these motifs in the sequences of newly discovered genes. In practice, this is more difficult than it might seem. The zinc finger motif is always uninterrupted, and so is easy to recognize. But many other sequence motifs are discontinuous, and the spacing between their elements can vary considerably. In such cases, the term sequence motif is almost a misnomer, since not only the spacing between the residues but also the order in which they occur may be completely different. These are really functional motifs whose presence is detected from the structure rather than the sequence. For example, the "catalytic triad" of the serine proteases (Figure 1-52), which consists of an aspartic acid, a histidine and a serine, all interacting with one another, comprises residues aspartic acid 102, histidine 57

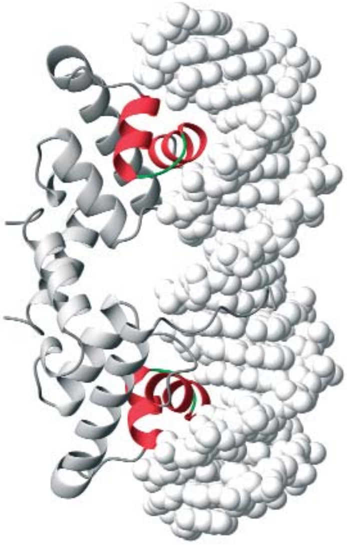


Figure 1-50 **Helix-turn-helix** The DNA-binding domain of the bacterial gene regulatory protein lambda repressor, with the two helix-turn-helix motifs shown in color. The two helices closest to the DNA are the reading or recognition helices, which bind in the major groove and recognize specific gene regulatory sequences in the DNA. (PDB 1lmb)

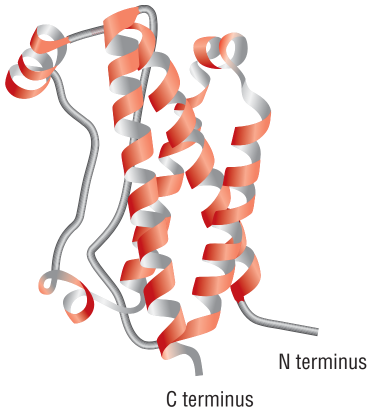
Definitions

convergent evolution: evolution of structures not related by ancestry to a common function that is reflected in a common **functional motif**.

functional motif: sequence or structural **motif** that is always associated with a particular biochemical function.

motif: characteristic sequence or structure that in the case of a **structural motif** may comprise a whole domain or protein but usually consists of a small local arrangement of secondary structure elements which then coalesce to form domains. **Sequence motifs**, which are recognizable amino-acid sequences found in different proteins, usually indicate biochemical function. Structural motifs are less commonly associated with specific biochemical functions.

Figure 1-51 Four-helix bundle motif The four-helix bundle motif can comprise an entire protein domain, and occurs in proteins with many different biochemical functions. Shown here is human growth hormone, a signaling molecule; shown in Figure 1-28a is cytochrome b562, an electron-transport protein. In Figure 1-54 the protein myohemerythrin is shown; its function is oxygen transport.



and serine 195 in one family of serine proteases. However, in another, unrelated family of serine proteases, the same triad is made up by aspartic acid 32, histidine 64, and serine 221 (see Figure 4-35). This is a case in which both the spacing between the residues that define the motif and the order in which they occur in the primary sequence are different. Nevertheless, these residues form a catalytic unit that has exactly the same geometry in the two proteases, and that carries out an identical chemical function. This is an example of **convergent evolution** to a common biochemical solution to the problem of peptide-bond hydrolysis. One of the major tasks for functional genomics is to catalog such sequence-based motifs, and develop methods for identifying them in proteins whose overall folds may be quite unrelated.

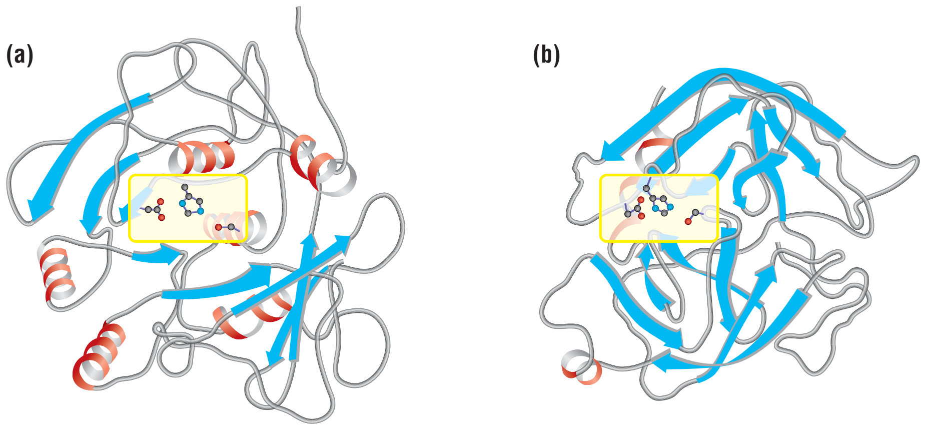


Figure 1-52 Catalytic triad The catalytic triad of aspartic acid, histidine and serine in **(a)** subtilisin, a bacterial serine protease, and **(b)** chymotrypsin, a mammalian serine protease. The two protein structures are quite different, and the elements of the catalytic triad are in different positions in the primary sequence, but the active-site arrangement of the aspartic acid, histidine and serine is similar.

Identifying structural motifs from sequence information alone presents very different challenges. First, as we have seen, many different amino-acid sequences are compatible with the same secondary structure; so there may be literally hundreds of different unrelated sequences that code for four-helix bundles. Sequence similarity alone, therefore, cannot be used for absolute identification of structural motifs. Hence, such motifs must be identified by first locating the secondary structure elements of the sequence. However, secondary structure prediction methods are not completely accurate, as pointed out earlier. Second, a number of structural motifs are so robust that large segments of additional polypeptide chain, even specifying entire different domains, can sometimes be inserted into the motif without disrupting it structurally. A common example is the so-called TIM-barrel domain, which consists of a strand of beta sheet followed by an alpha helix, repeated eight times. Protein domains are known that consist of nothing but this set of secondary structure elements; others are known in which an additional structural motif is inserted; and yet others are found in which one or more additional entire domains interrupt the pattern, but without disrupting the barrel structure (Figure 1-53).

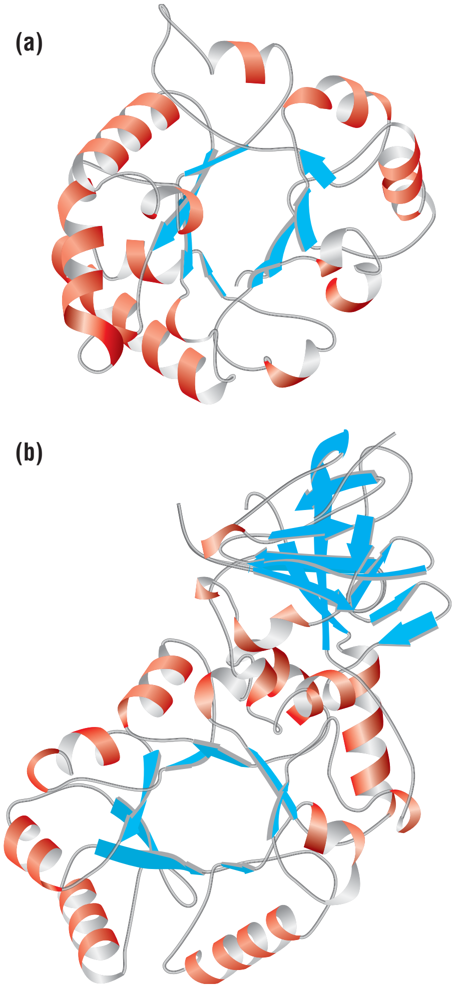


Figure 1-53 TIM-barrel proteins Triose phosphate isomerase **(a)** is shown together with alanine racemase **(b)**. In alanine racemase, the TIM-barrel domain is interrupted by an inserted domain.

References

Aitken, A.: **Protein consensus sequence motifs.** *Mol. Biotechnol.* 1999, **12**:241–253.

de la Cruz, X. and Thornton, J.M.: **Factors limiting the performance of prediction-based fold recognition methods.** *Protein Sci.* 1999, **8**:750–759.

Ponting, C.P. *et al.*: **Evolution of domain families.** *Adv. Protein Chem.* 2000, **54**:185–244.

1-17 Alpha Domains and Beta Domains

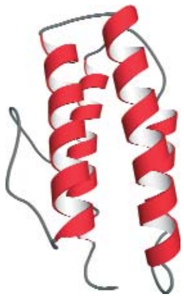


Figure 1-54 **Myohemerythrin** A protein composed of a single four-helical bundle domain. (PDB 2mhr)

Protein domains can be classified according to their secondary structural elements

It is useful to group domain folds into five broad classes, based on the predominant secondary structure elements contained within them. **Alpha domains** are comprised entirely of alpha helices. **Beta domains** contain only beta sheet. **Alpha/beta domains** contain beta strands with connecting helical segments. **Alpha+beta domains** contain separate beta sheet and helical regions. And **cross-linked domains** have little, if any, secondary structure but are stabilized by several disulfide bridges or metal ions. Within each class, many different arrangements of these elements are possible; each distinct arrangement is a structural motif.

Two common motifs for alpha domains are the four-helix bundle and the globin fold

The preference for certain helix-crossing angles (see section 1-10) leads to two common motifs for interacting helices. One of them is a bundle of four antiparallel alpha helices, each crossing the next at an angle of about –20°, so that the entire motif has a left-handed twist. This **four-helix bundle** has been found in a wide variety of alpha domains, where it serves such diverse functions as oxygen transport, nucleic acid binding, and electron transport. Examples of four-helix bundle proteins include myohemerythrin, an oxygen-storage protein in marine worms (Figure 1-54), and human growth hormone, which helps promote normal body growth.

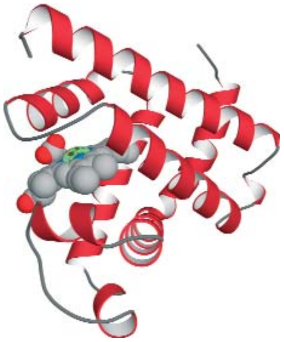


Figure 1-55 **Myoglobin** A protein composed of a single globin fold domain. (PDB 1a6k)

Another common alpha-domain motif, the **globin fold**, consists of a bag of about eight alpha helices arranged at +90° and +50° angles with respect to each other. This motif leads to the formation of a hydrophobic pocket in the domain interior in which large, hydrophobic organic and organometallic groups can bind (Figure 1-55). This fold gets its name from the protein myoglobin, a single-domain oxygen-storage molecule in which eight helices wrap around a heme group. It reappears in somewhat different form in the electron transport proteins called cytochromes, which also have bound heme groups. Interestingly, at least one heme-binding protein, cytochrome b562, is a four-helix bundle instead of a globin fold.

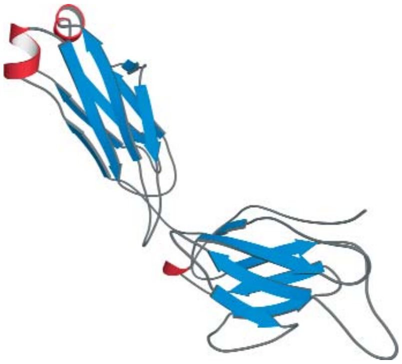


Figure 1-56 **Immunoglobulin** A protein composed of several beta domains (light chain only shown). (PDB 1a3l)

Beta domains contain strands connected in two distinct ways

Domains that contain only beta sheet, tight turns and irregular loop structures are called beta domains. Proteins made up of beta domains include immunoglobulins (Figure 1-56), several enzymes such as superoxide dismutase, and proteins that bind to sugars on the surfaces of cells. Because there are no helices to make long connections between adjacent strands of the beta sheet, all-beta domains contain essentially nothing but antiparallel beta structure, the strands of which are connected with beta turns and larger loops.

The patterns of connections between strands give rise to beta sheets with two distinct topologies. The directionality of the polypeptide chain dictates that a strand in an antiparallel beta sheet can only be linked to a strand an odd number of strands away. The most common connections are to an immediately adjacent strand or to one three strands away. If all the connections link adjacent strands, the beta sheet has an **up-and-down structural motif** (Figure 1-57). A particularly striking example is found in the enzyme neuraminidase from the influenza virus, which consists of a repeating structural motif of four antiparallel strands. Each up-and-down motif forms the blade of a so-called beta-propeller domain.

Definitions

alpha domain: a protein domain composed entirely of alpha helices.

alpha/beta domain: a protein domain composed of beta strands connected by alpha helices.

alpha+beta domain: a protein domain containing separate alpha-helical and beta-sheet regions.

beta domain: a protein domain containing only beta sheet.

beta sandwich: a structure formed of two antiparallel beta sheets packed face to face.

cross-linked domain: a small protein domain with little or no secondary structure and stabilized by disulfide bridges or metal ions.

four-helix bundle: a structure of four antiparallel alpha helices. Parallel bundles are possible but rare.

globin fold: a predominantly alpha-helical arrangement observed in certain heme-containing proteins.

Greek-key motif: an arrangement of antiparallel beta strands in which the first three strands are adjacent but the fourth strand is adjacent to the first, with a long connecting loop.

jelly roll fold: a beta sandwich built from two sheets with topologies resembling a Greek key design. The sheets pack almost at right-angles to each other.

up-and-down structural motif: a simple fold in which beta strands in an antiparallel sheet are all adjacent in sequence and connectivity.

Connection to the third strand leads to a motif called a **Greek key**, so named because it resembles the Greek-key design on ancient vases (Figure 1-58). An example of this motif is provided by pre-albumin, which contains two Greek-key motifs. The characteristic fold of the immunoglobulins, which is also found in a number of proteins that interact with other proteins on the cell surface, is a central Greek-key motif flanked on both sides by additional antiparallel strands.

Antiparallel beta sheets can form barrels and sandwiches

Antiparallel sheets in beta domains tend to be oriented with one face on the surface of the protein, exposed to the aqueous surroundings, and the other face oriented toward the hydrophobic core. This internal face is packed against another section of beta sheet with the inward-facing side chains of both packing together to form a hydrophobic core. Thus, in beta domains, the sheet tends to be amphipathic, with one face predominantly hydrophilic while the other is almost entirely composed of hydrophobic amino acids. This characteristic may make it possible to recognize such domains from the distribution of polar and nonpolar residues in the amino-acid sequence if secondary structure prediction methods become more accurate.

There are two ways to form structures in which antiparallel beta sheets can pack against each other. These give rise to beta barrels and **beta sandwiches**. In a beta-barrel motif, a single beta sheet forms a closed cylindrical structure in which all strands are hydrogen bonded to one another; the last strand in the sheet is hydrogen bonded to the first. Both types of beta-sheet connectivity are compatible with a beta barrel: pre-albumin is an example of a beta barrel constructed using the Greek-key motif (Figure 1-58), and human plasma retinol-binding protein, which carries vitamin A (retinol) in the serum, is an example of a beta barrel that is formed from an up-and-down motif (see Figure 1-19).

In a beta sandwich two separate beta sheets pack together face-to-face like two slices of bread. This arrangement differs from a barrel because the end strands of each sheet segment are not hydrogen bonded to one another. Their hydrogen-bonding potentials are satisfied chiefly by interactions with side chains or with water molecules. The two sheets in a beta sandwich are often at right angles to one another. Once again, both types of antiparallel sheet connectivity can be accommodated in this arrangement. The immunoglobulin fold (see Figure 1-56) is an example of a beta sandwich built with two Greek-key motifs. A variation of this theme is the **jelly roll fold** that comprises the major domain of the coat proteins of many spherical viruses. Bacteriochlorophyll A protein contains an antiparallel beta sandwich with up-and-down topology (Figure 1-59). The sandwich/barrel distinction is useful but not absolute: in a number of immunoglobulin domains the first or seventh strand switches sheets, forming a partial barrel.

The fibrous protein silk provides a particularly striking example of a beta sandwich. Silk is a beta-sheet protein composed largely of glycine, alanine, and serine, and every other residue in its sequence is a glycine. Because of the up-down alternation of residues in beta sheets, all glycines are on one side of the sheet, and the alanines and serines are on the other. The alanine and serine side chains of one sheet pack nicely between the alanine and serine side chains of another, producing a two-sheet structure. The tensile strength of silk derives from this interaction plus the hydrogen-bonded stability of the individual beta sheets themselves (see Figure 1-1).

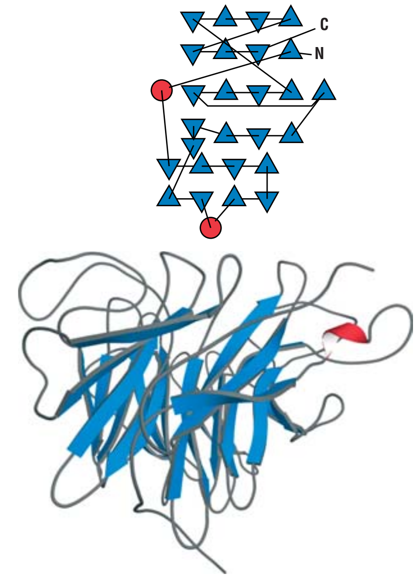


Figure 1-57 Neuraminidase beta-propeller domain A subunit of the four-subunit neuraminidase protein composed of repeating up-and-down beta motifs. (PDB 1a4q)

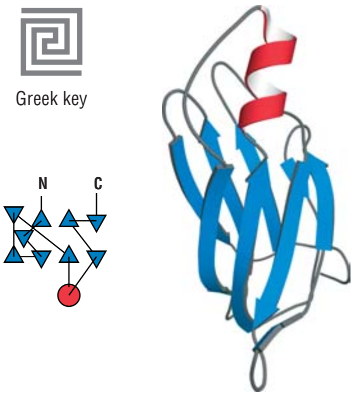


Figure 1-58 Pre-albumin An example of a beta domain made up of Greek-key motifs. (PDB 1tta) Only one subunit of the two-subunit structure is shown.

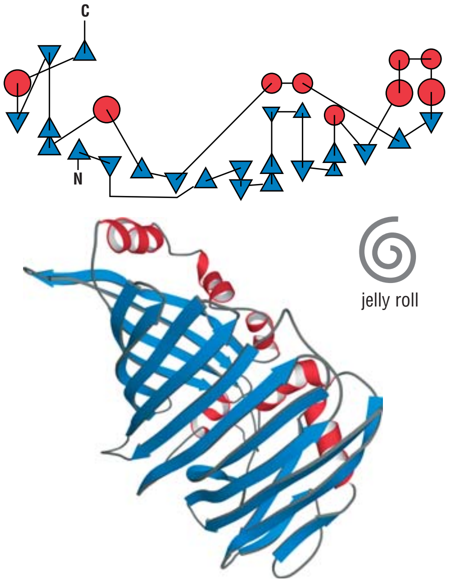


Figure 1-59 Bacteriochlorophyll A protein This protein contains a domain with an up-and-down beta sandwich, a motif known as a jelly roll. (PDB 1ksa)

References

Bork, P. *et al.*: **The immunoglobulin fold. Structural classification, sequence patterns and common core.** *J. Mol. Biol.* 1994, **242**:309–320.

Branden, C. and Tooze, J.: *Introduction to Protein Structure* 2nd ed. (Garland, New York, 1999).

Richardson, J.S. and Richardson, D.C.: **Principles and patterns of protein conformation** in *Prediction of Protein Structure and the Principles of Protein Conformation* 2nd ed. Fasman, G.D. ed. (Plenum Press, New York, 1990), 1–98.

Weber, P.C. and Salemme, F.R.: **Structural and functional diversity in 4 alpha-helical proteins.** *Nature* 1980, **287**:82–84.

For a comprehensive analysis of domain folds, an all-against-all structure comparison can be found at:

http://www.ebi.ac.uk/dali/

1-18 Alpha/Beta, Alpha+Beta and Cross-Linked Domains

In alpha/beta domains each strand of parallel beta sheet is usually connected to the next by an alpha helix

In alpha/beta domains the beta sheet is composed of parallel or mixed strands; the parallel strands must be joined by long connections because the linking segment has to traverse the length of the sheet, and these connections are usually made by alpha helices connecting parallel adjacent strands, giving rise to beta-alpha-beta-alpha units. As illustrated in Figure 1-60, the crossover connection between the two parallel beta strands can be either right-handed or left-handed. The right-handed twist of the beta strand (see section 1-7), however, produces an enormous bias toward the right-handed crossover topology: it is observed in more than 95% of alpha/beta structures. This crossover rule is obeyed even when the connected strands are not adjacent or when the connecting segment is a loop, not a helix.

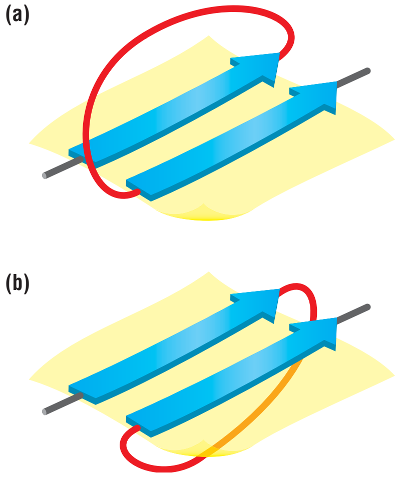


Figure 1-60 Crossover connection between parallel beta strands (a) A right-handed crossover connection. **(b)** A left-handed crossover connection.

There are two major families of alpha/beta domains: barrels and twists

Just as two motifs predominate in antiparallel barrels and sandwiches, two motifs also account for nearly all alpha/beta domains. One of these is a closed structure called an **alpha/beta barrel** (Figure 1-61a). The other is an open twisted beta structure that looks somewhat like a saddle; we will call it an **alpha/beta twist** (Figure 1-61b).

The most regular form of alpha/beta structure is the alpha/beta barrel, in which the beta-alpha-beta-alpha motif is repeated four or more times. In this motif, the strand order is consecutive, and the combination of the twist of the beta sheet itself and the adjacent laying down of strands produces a closed barrel. This fold is particularly stable when there are eight strands in the barrel (Figure 1-61a). It is often called a **TIM barrel** because it was first discovered in the three-dimensional structure of the enzyme triosephosphate isomerase, which is abbreviated TIM.

The core of the alpha/beta barrel motif is its parallel beta sheet, which is surrounded by alpha helices that shield it from solvent. The helices are amphipathic and their nonpolar sides pack against the hydrophobic face of one side of the sheet. The center of a beta barrel is usually filled with hydrophobic side chains from the other face of the beta sheet; thus in alpha/beta barrels, the sheet is almost entirely hydrophobic. The TIM-barrel structure is one of the few domain folds that is relatively easy to recognize from the amino-acid sequence. Because it is the most common domain fold yet observed, occuring in 10% of all enzyme structures, it is a good bet that any sequence predicted to have a relatively nonpolar beta strand followed by an amphipathic alpha helix, repeated eight times, will form a TIM barrel.

The parallel beta strands in alpha/beta twists form an open sheet that is twisted into a saddle-shaped structure. The strand order in the sheet is not consecutive because the sheet is built in two halves. The first beta strand in the primary sequence forms a strand in the middle of the sheet. Additional strands are laid down consecutively outward to one edge, whereupon the chain returns to the middle of the sheet (the so-called “switch point”) and forms the strand that hydrogen bonds to the outside of the first strand (Figure 1-61b). From there the chain continues out to the other edge. This mode of winding places the helices on one side for half of the sheet, and on the opposite side for the other half of the sheet. Again, the helices tend to

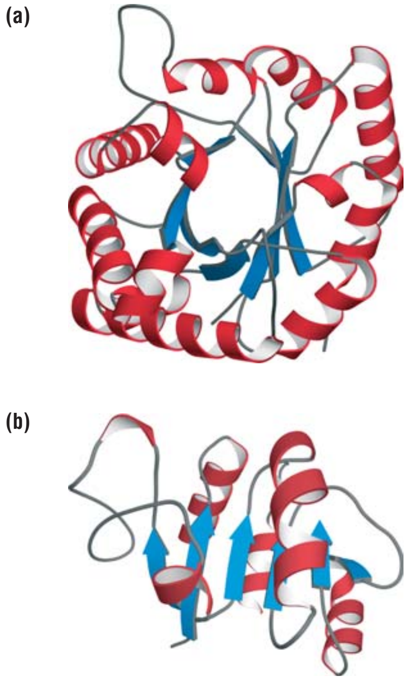


Figure 1-61 Alpha/beta domains (a) Alpha/beta barrel: the TIM barrel. (PDB 1tim) **(b)** Alpha/beta twist: aspartate semi-aldehyde dehydrogenase. (PDB 1brm) The connecting segments are usually alpha helices.

Definitions

alpha/beta barrel: a parallel beta barrel formed usually of eight strands, each connected to the next by an alpha-helical segment. Also known as a **TIM barrel**.

alpha/beta twist: a twisted parallel beta sheet with a saddle shape. Helices are found on one side of the sheet for the first half and the other side for the second half.

nucleotide-binding fold: an open parallel beta sheet with connecting alpha helices that is usually used to bind NADH or NADPH. It contains a characteristic sequence motif that is involved in binding the cofactor. Also known as the Rossmann fold.

TIM barrel: another name for the alpha/beta barrel fold.

zinc finger: a small, irregular domain stabilized by binding of a zinc ion. Zinc fingers usually are found in eukaryotic DNA-binding proteins. They contain signature metal-ion binding sequence motifs.

be amphipathic whereas the sheet is predominantly hydrophobic. In its classic form, the alpha/beta twist motif has six parallel beta strands and five connecting helices, as shown in Figure 1-61b. Whenever this fold occurs in an enzyme, the switch region is always part of the catalytic site of the protein. Another name for this structure is the **nucleotide-binding fold**, which is indicative of the function it performs in many proteins.

In contrast to the antiparallel beta sheet, which always has one face in contact with water, most parallel beta structures are shielded from direct interaction with water by their coating of alpha helices. In the alpha/beta barrel motif, the interhelical packing angle is always +50°, and the same value is common for the helices that coat the surfaces of the alpha/beta twist motifs as well. The preference for this angle over the –20° and +90° alternatives reflects the need to nest the helices in the grooves on the surface of the twisted beta structure.

Alpha+beta domains have independent helical motifs packed against a beta sheet

Alpha+beta domains contain both beta sheets and alpha helices, but they are segregated. No special organizing principles can be stated for this class, but their individual secondary structure regions follow all of the principles we have described for alpha helices and beta sheets separately. The helical motifs in alpha+beta domains are usually just clusters of interacting helices, while the beta sheets tend to be antiparallel or mixed. One example is a saddle-shaped, antiparallel sheet with a layer of alpha helices covering one face (Figure 1-62). This arrangement leaves the other face of the sheet exposed to the solvent, which is a preference of antiparallel beta structures that we have already noted. Sometimes the layer of helices is used to form a recognition site, such as the peptide-binding groove in the major histocompatibility proteins.

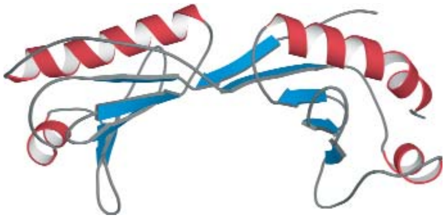


Figure 1-62 Alpha+beta saddle The structure of the TATA-binding protein that binds to DNA at the so-called TATA box that specifies the site at which gene transcription is initiated in eukaryotes. The beta sheet that forms the seat of the saddle binds in the minor groove of the DNA, bending it significantly. (PDB 1tgh)

Metal ions and disulfide bridges form cross-links in irregular domains

The final class of domain structure, the cross-linked irregular domain, is found in small single-domain intra- and extracellular proteins. There are two subclasses, which represent distinct solutions to the problem of structural stability in a domain that is too small to have an extensive hydrophobic core or a large number of secondary structural interactions. Both solutions involve cross-linking different parts of the domain via covalent interactions. In small irregular extracellular domains this cross-linking derives from disulfide bond formation, usually involving a number of cysteine pairs. In small irregular intracellular domains, metal ions (usually zinc but sometimes iron) form the cross-links, connecting different parts of the domain through ligation by nucleophilic side chains.

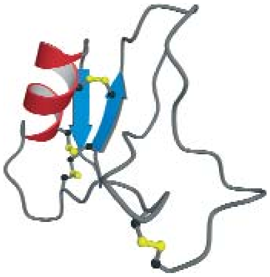


Figure 1-63 Disulfide-linked protein Scorpion toxin: a small irregular extracellular protein with no large hydrophobic core and minimal secondary structure. It is stabilized by four disulfide bridges. (PDB 1b7d)

Disulfide-linked extracellular small proteins are often toxins that inhibit essential cellular proteins and prevent them from functioning. Most of these proteins are unusually stable to proteolytic digestion and heat denaturation. This class includes cobra venom neurotoxin, scorpion toxin (Figure 1-63), the ragweed pollen allergy factor Ra5, several secreted protease inhibitors, and toxic proteins from marine snails.

Metal ion cross-linked domains are found, for example, in **zinc finger** transcription factors (Figure 1-64) and iron–sulfur proteins called ferredoxins. A number of other metal-stabilized domains have been found. Although their structures are not as well characterized as that of the zinc finger, they too can be recognized at the sequence level because of characteristic sequence patterns in the vicinity of the residues that contribute metal ligands.

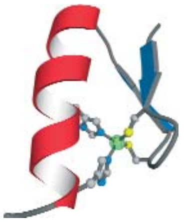


Figure 1-64 Zinc finger A domain from a larger transcription factor, that is stabilized by the coordination of two histidines and two cysteines to a zinc ion. In the absence of the metal ion, this domain is unfolded, presumably because it is too small to have a hydrophobic core. This domain is the most abundant one in the human genome. (PDB 1aay)

References

Bellamacina, C.R.: **The nicotinamide dinucleotide binding motif: a comparison of nucleotide binding proteins.** *FASEB J.* 1996, **10**:1257–1269.

Branden, C. and Tooze, J.: *Introduction to Protein Structure*, 2nd ed. (Garland, New York, 1999).

Chothia, C.: **Asymmetry in protein structure.** *Ciba Foundation Symp.* 1991, **162**:36–49.

Leon, O. and Roth, M.: **Zinc fingers: DNA binding and protein-protein interactions.** *Biol. Res.* 2000, **33**:21–30.

Reardon, D. and Farber, G.K.: **The structure and evolution of alpha/beta barrel proteins.** *FASEB J.* 1995, **9**:497–503.

Richardson, J.S. and Richardson, D.C.: **Principles and patterns of protein conformation** in *Prediction of Protein Structure and the Principles of Protein Conformation* 2nd ed. Fasman, G.D. ed. (Plenum Press, New York, 1990), 1–98.

1-19 Quaternary Structure: General Principles

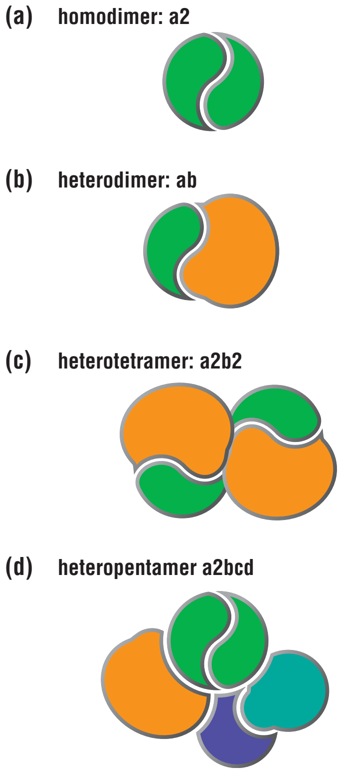
Many proteins are composed of more than one polypeptide chain

Many proteins self-associate into assemblies composed of anything from two to six or more polypeptide chains. They may also associate with other, unrelated proteins to give mixed species of the form (ab), (a2b2), and so on (Figure 1-65a-c). The acetylcholine receptor, a membrane protein of vital importance for neuromuscular communication, is a five-chain molecule of the form (a2bcd) (Figure 1-65d). Proteins also assemble with other kinds of macromolecules.

Protein assemblies composed of more than one polypeptide chain are called **oligomers** and the individual chains of which they are made are termed **monomers** or subunits. Oligomers containing two, three, four, five, six or even more subunits are known as **dimers**, **trimers**, **tetramers**, **pentamers**, **hexamers**, and so on. Much the commonest of these are dimers. Some oligomers, as we have mentioned, contain only one kind of monomer, while others are made up of two or more different chains. Oligomers composed of only one type of monomer are sometimes prefixed homo-: for example, keratin, which is made up of three alpha-helical polypeptides coiled around one another, is composed of three identical chains and is thus a **homotrimer**. Oligomers composed of monomers encoded by different genes are prefixed hetero-: for example, hemoglobin, which contains two alpha and two beta chains, is a **heterotetramer** built from two homodimers. The number and kinds of subunits in an assembly, together with their relative positions in the structure, constitute the quaternary structure of an assembly.

It is very common for the subunits of hetero-oligomers to resemble one another structurally, despite being encoded by different genes and in some cases having little or no sequence similarity. This is true, for example, for hemoglobin, where the alpha and beta chains have nearly identical folds, and for the acetylcholine receptor, where the four different gene products that make up the pentamer are closely related structurally. One can speculate that this pattern reflects the origin of many hetero-oligomers in the duplication of a gene that coded for the single subunit of an ancestral homo-oligomeric protein.

Macromolecular assemblies form spontaneously when the right amounts of the appropriate components are present. The interactions between subunits are tight and specific, and they exclude "wrong" molecules from interfering with self assembly.


(a) homodimer: a2

(b) heterodimer: ab

(c) heterotetramer: a2b2

(d) heteropentamer a2bcd

Figure 1-65 Schematic representations of different kinds of oligomers (a) a2 **(b)** ab **(c)** a2b2 **(d)** a2bcd. Many other arrangments are possible and are observed (see Figure 1-74).

All specific intermolecular interactions depend on complementarity

Protein surfaces are irregular. This is what enables proteins to bind specific ligands and to associate specifically with other proteins, and it underlies the formation of quaternary structure. The "fit" between one protein surface and another depends on much more than shape. It extends to the weak bonds that hold complexes together; hydrogen-bond donors are opposite acceptors, nonpolar groups are opposite other nonpolar groups, and positive charges are opposite negative charges (Figure 1-66). This property of complementarity is observed in all binding interactions, whether between a protein and a small molecule or between a protein and another kind of macromolecule.

Complementarity is necessary because an intermolecular interface is composed of many weak interactions. Any single hydrogen bond or van der Waals interaction will break quite often at body temperature (see section 1-12). For a complex to be stable long enough to function, the strength of binding must be greater than about 15–20 kJ/mole. As free energies are additive, tight binding can be achieved if there is a large number of weak interactions, and the number

Definitions

coiled coil: a protein or a region of a protein formed by a dimerization interaction between two alpha helices in which hydrophobic side chains on one face of each helix interdigitate with those on the other.

dimer: an assembly of two identical (homo-) or different (hetero-) subunits. In a protein, the subunits are individual folded polypeptide chains.

heptad repeat: a sequence in which hydrophobic residues occur every seven amino acids, a pattern that is reliably indicative of a **coiled-coil** interaction between two alpha helices in which the hydrophobic side chains of each helix interdigitate with those of the other.

heterotetramer: an assembly of four subunits of more than one kind of polypeptide chain.

hexamer: an assembly of six identical or different subunits. In a protein the subunits are individual folded polypeptide chains.

homotrimer: an assembly of three identical subunits: in a protein, these are individual folded polypeptide chains.

monomer: a single subunit: in a protein, this is a folded polypeptide chain.

oligomer: an assembly of more than one subunit: in a protein, the subunits are individual folded polypeptide chains.

pentamer: an assembly of five identical or different subunits: in a protein, these are individual folded polypeptide chains.

quaternary structure: the subunit structure of a protein.

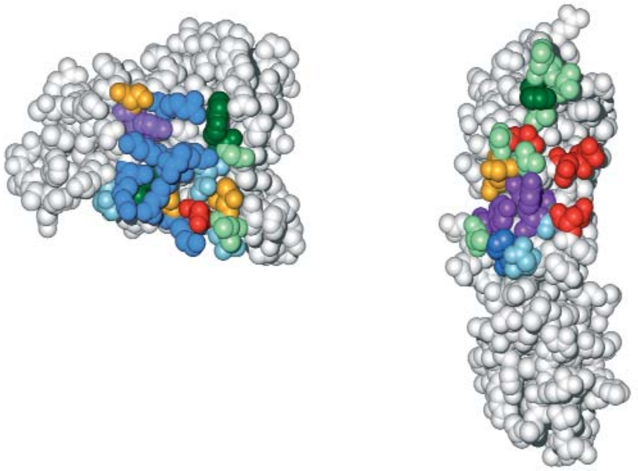


Figure 1-66 “Open-book” view of the complementary structural surfaces that form the interface between interleukin-4 (left) and its receptor (right) The contact residues are colored as follows: red, negatively charged; dark blue, positively charged; light blue, histidine; cyan, glutamine and asparagine; purple, tyrosine; yellow, serine/threonine; green, hydrophobic. Note that this interface contains a mixture of interaction types. Graphic kindly provided by Walter Sebald and Peter Reineme.

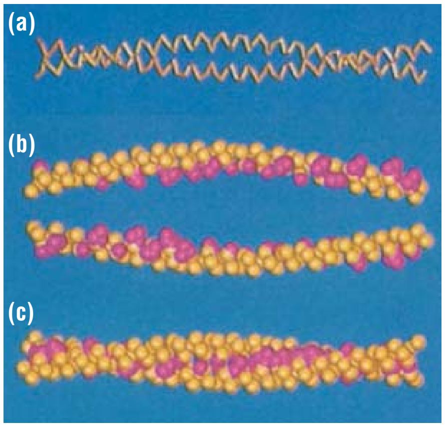


Figure 1-67 Coiled-coil alpha-helical interactions (a) Two interacting alpha helices of tropomyosin shown in a chain representation; **(b)** a space-filling representation of the separate alpha helices of tropomyosin with the hydrophobic side chains shown as dark protrusions; **(c)** the tropomyosin dimer showing how the hydrophobic side chains interdigitate in the coiled coil in a knobs in holes arrangement. (Taken from Cohen, C. and Parry, D.A.: **Alpha-helical coiled coils and bundles: how to design an alpha-helical protein.** *Proteins* 1990, **7**:1–15.)

and strength of weak interactions is maximized if contact surfaces fit closely together. Complementarity ensures that all possible van der Waals contacts are made, and that hydrogen-bond donors and acceptors at the interface between the two molecules pair with each other instead of making hydrogen bonds to water.

A particularly well characterized example of complementarity between interacting surfaces occurs in the case of coiled-coil structures (Figure 1-67). **Coiled coils** are dimers of alpha helices formed through the ridges and grooves arrangement we have already mentioned as the basis for tertiary structural interactions between alpha helices (see section 1-10). In such interacting helices, hydrophobic side chains, often those of leucines, are repeated at intervals of seven amino acids in the chain, forming the “ridge” of hydrophobic side chains that fit into spaces on the interacting helix. This pattern is known as the **heptad repeat**, and is characteristic of all dimeric structures formed through interacting alpha helices. It is one of the few cases in which structure can reliably be predicted from sequence.

Although all intermolecular interactions depend on surface complementarity, not all of them occur between preexisting complementary surfaces: one of the surfaces involved, or both, may be an unfolded region of the peptide in the absence of its partner. In coiled-coil proteins, for example, the two subunits are frequently unfolded as monomers and assume their folded structure only on dimerization. This is the case for the so-called leucine zipper family of transcriptional regulators which bind DNA on dimerization through a leucine-rich heptad repeat (Figure 1-68).

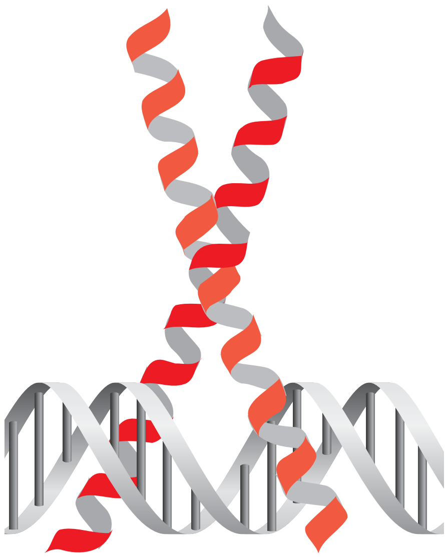


Figure 1-68 Peptide–peptide interactions in the coiled coil of the leucine zipper family of DNA-binding proteins The monomers of the leucine zipper are disordered in solution but fold on dimerization through hydrophobic coiled-coil interactions in their carboxy-terminal regions and on contact with DNA through their basic amino-terminal regions.

tetramer: an assembly of four identical or different subunits.

trimer: an assembly of three identical or different subunits.

References

Anston, A.A. *et al.*: **Circular assemblies.** *Curr. Opin. Struct. Biol.* 1996, **6**:142–150.

Bosshard, H.R. *et al.*: **Energetics of coiled coil folding: the nature of the transition state.** *Biochemistry* 2001, **40**:3544–3552.

Creighton, T.E.: *Proteins: Structure and Molecular Properties* 2nd ed. (Freeman, New York, 1993), 233–236.

Gonzalez, L. Jr *et al.*: **Buried polar residues and structural specificity in the GCN4 leucine zipper.** *Nat. Struct. Biol.* 1996, **3**:1011–1018.

Jones, S. and Thornton, J.M.: **Principles of protein-protein interactions.** *Proc. Natl Acad. Sci. USA* 1996, **93**:13–20.

Myers, J.K. and Oas, T.G.: **Reinterpretation of GCN4-p1 folding kinetics: partial helix formation precedes dimerization in coiled coil folding.** *J. Mol. Biol.* 1999, **289**:205–209.

Perham, R.N.: **Self-assembly of biological macromolecules.** *Philos. Trans. R. Soc. Lond. B.* 1975, **272**:123–136.

Zielenkiewicz, P. and Rabczenko, A.: **Methods of molecular modelling of protein-protein interactions.** *Biophys. Chem.* 1988, **29**:219–224.

1-20 Quaternary Structure: Intermolecular Interfaces

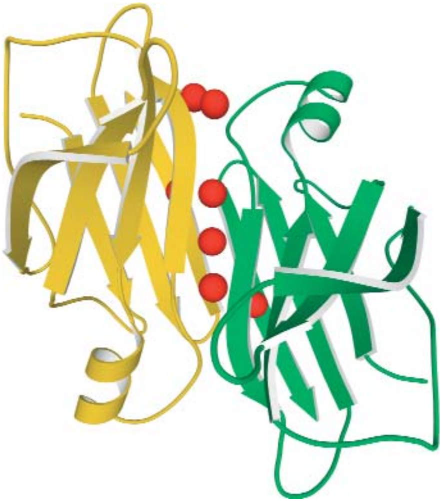


Figure 1-69 **Water molecules at a subunit interface** Pre-albumin is a dimeric plasma protein that binds iodinated hormones. The structure of the dimer clearly shows a network of water molecules (red spheres) trapped between the two subunits. (PDB 1bm7)

All types of protein-stabilizing interactions contribute to the formation of intermolecular interfaces

The weak bonds that hold subunits together are the same as those that stabilize the folded structures of monomeric proteins (see section 1-4). Thus, hydrophobic interactions, hydrogen bonds and salt bridges are all observed at the interfaces of protein–protein and protein–peptide complexes. Cross-linking interactions, such as disulfide interactions and metal-ion ligation, also occur at some interfaces. Of these interactions, the hydrophobic effect deserves special mention. The portion of the surface area of a subunit that becomes buried when an oligomer forms is, in general, less polar than is typically the case for surface regions of soluble, monomeric proteins. The amount of surface area that is actually buried at an interface varies greatly. Salt bridges, which would usually be found at the exposed surface as they involve charged residues, are surprisingly common at interfaces. Examples are found in virus structures, in coiled-coil interactions and in the recognition of phosphate groups by signaling proteins, among others. Hemoglobin has several intersubunit salt bridges, which can break in response to a change in pH, altering the relative orientations of the subunits and the affinity of the protein for oxygen. A rough correlation exists between the stability of the oligomer and the type of interaction that predominates at the interface. Very stable oligomers tend to bury a large hydrophobic surface area between subunits, whereas subunits that assemble and disassemble more easily as part of their function seem to employ more polar interactions.

The atomic-packing density at the interface between subunits usually approaches that of the interior of a monomeric protein, but water molecules are present more often at subunit interfaces than they are in protein interiors. Some water molecules may be trapped when individual folded subunits associate, but others are likely to have more important roles. The subunits of many oligomeric proteins are only partially folded before oligomerization (as we have seen in the case of the leucine zipper proteins: see Figure 1-68), and only in the context of their neighbors do they assume their final, correct, tertiary structure. Thus, it may be that water molecules found at interfaces are essential for preserving the structures of the partially folded monomer units before aggregation (Figure 1-69).

Hydrogen-bonding potentials at the interface between subunits must be satisfied just as they are throughout any folded protein; many of the hydrogen bonds are between subunits, and the rest are with interfacial waters. Because hydrogen bonds are highly directional, they orient interactions between subunits, and provide much of the specificity for complex oligomerizations. Sometimes, intersubunit hydrogen bonding can be part of the secondary structure of a protein; there are many cases where a beta sheet continues smoothly from one subunit to the next (Figure 1-70).

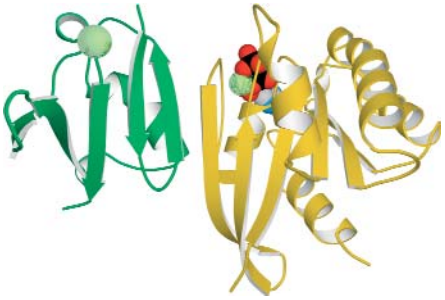


Figure 1-70 **Oligomerization by beta sheet formation** The signal transduction proteins Rap (left) and Raf (right) both contain beta sheets with exposed edge strands. These proteins form a heterodimer by using the edge strands to complete a continuous extended beta sheet that traverses both molecules. (PDB 1gua)

Inappropriate quaternary interactions can have dramatic functional consequences

A number of genetic diseases, or striking phenotypes, originate from hydrophobic surfaces that are inappropriately created as a result of mutation. Sickle-cell anemia is one example. The mutation of a glutamate to valine on the surface of the beta subunit of hemoglobin creates a hydrophobic patch that causes hemoglobin tetramers to polymerize into long fibrils (Figure 1-71): the polymer is formed by burial of the new patch at the tetramer–tetramer interface.

In the case of the sickle-cell mutation, abnormal polymerization occurs through the generation of an inappropriate hydrophobic interaction. But the normal polymerization interactions of

Definitions

dominant-negative: dominant loss of function due to a single mutant copy of a gene. This can occur when the mutant subunit is able to oligomerize with normal subunits to form a non-functional protein, thereby producing a loss-of-function phenotype even in the presence of a normal copy of the gene.

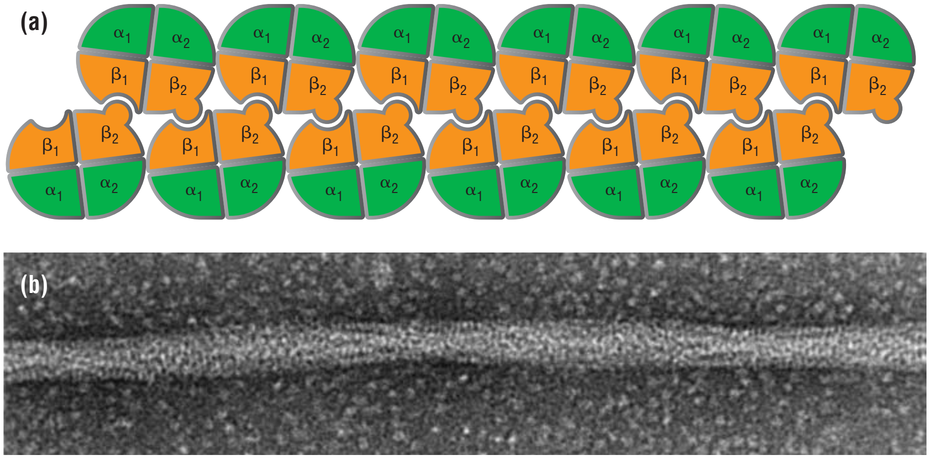


Figure 1-71 **Sickle-cell hemoglobin** Hemoglobin molecules form long polymers when they carry the sickle-cell mutation, in which a hydrophobic patch is created on the surface of the tetramer by the substitution of a hydrophobic valine for a hydrophilic glutamine in the beta subunit. **(a)** The hydrophobic patch created by the mutant valine is represented by a bump in the beta2 subunit, which binds in a hydrophobic pocket in the beta1 subunit of another hemoglobin molecule. Because the hydrophobic pocket into which the mutant valine binds is present only in the deoxy form of hemoglobin, the formation of the fibers, which constrains the molecule in the deoxy state, also functionally disables it. **(b)** Polymers of sickle-cell hemoglobin aggregate to form thicker fibers. These rigid fibers distort the hemoglobin-carrying red blood cells, causing them to rupture or to block blood vessels, with painful and sometimes fatal consequences. Severe anemia is thus not the only pathological consequence of the sickle-cell mutation. (Courtesy of Stuart J. Edelstein.)

oligomeric proteins can also make them more susceptible to disruption by mutation than are monomeric proteins. In the case of monomeric proteins, loss of function by mutation usually has to occur in both copies of the gene in question before the individual is affected: that is, the individual must be homozygous, and the effect is said to be recessive. In oligomeric proteins, however, mutant subunits produced by one copy of the gene may disrupt the function of normal subunits produced by the other, unmutated, copy so that the effect of the mutation is seen even in the presence of one normal gene: that is, when the individual is heterozygous. This is known as the **dominant-negative** effect, and can occur when the interaction surface is intact in the mutant subunit but the active site is abnormal or missing. Thus, for example, the introduction of a premature STOP codon may result in the expression of a protein fragment that can bind to a normal subunit but lacks the functional domain that contributes to the active site of the protein (Figure 1-72). The fragment will thus act as an inhibitor for the function of the intact subunit. This is the basis, for example, of an immune deficiency in the response to mycobacteria, in which anti-mycobacterial mechanisms normally induced in cells by interferon are abrogated by a mutant subunit of the heterotetrameric interferon receptor.

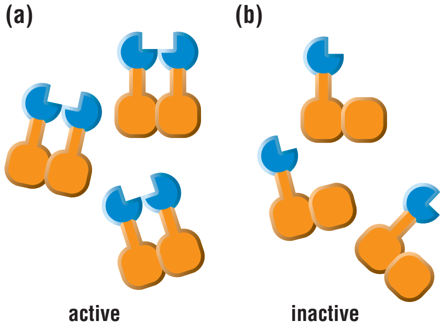


Figure 1-72 **Dominant-negative phenotype resulting from hydrophobic interactions between mutant and normal subunits of a dimeric protein** **(a)** Schematic representation of a protein, active only as a dimer, in which one domain (yellow) is necessary for dimerization while the other (blue) forms an active site on dimerization. **(b)** A truncated protein comprising the dimerization domain only and produced from a mutant copy of the gene can associate with normal protein to form an inactive dimer lacking an active site. Some active dimers will also form between normal monomers, so this will produce a dominant lack-of-function phenotype only if either normal function requires normal levels of the protein, or the truncated mutant is produced in excess, so that almost all of the dimers are inactive.

References

Bunn, H.F. and Forget, B.G.: *Hemoglobin: Molecular, Genetic and Clinical Aspects* Chapter 11 (Saunders, New York, 1986).

Harrington, D.J. *et al.*: **The high resolution structure of deoxyhemoglobin S.** *J. Mol. Biol.* 1997, **272**:398–407.

Herskowitz, I.: **Functional inactivation of genes by dominant negative mutations.** *Nature* 1987, **329**:219–222.

Jouanguy, E. *et al.*: **A human IFNGR1 small deletion hotspot associated with dominant susceptibility to mycobacterial infection.** *Nat. Genet.* 1999, **21**:370–378.

1-21 Quaternary Structure: Geometry

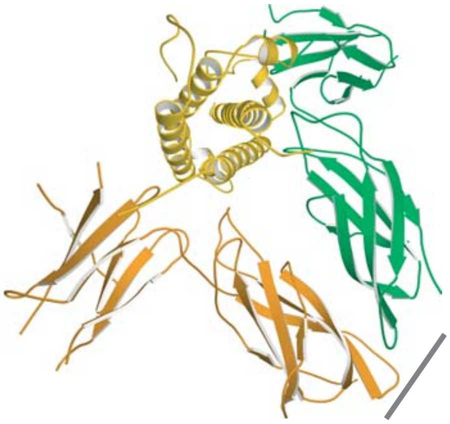


Figure 1-73 **The human growth hormone–receptor complex** Structure of the human growth hormone (yellow) complexed with two identical molecules of its receptor (orange and green). The receptor is a membrane protein, but only the extracellular hormone-binding portion is shown. The plane of the membrane is indicated by the slanted line. A molecule of the monomeric hormone binds to two identical receptor molecules. Similar regions of the two receptor molecules are used to bind two distinct regions of the hormone; the conformational flexibility of these regions allows for this versatility. (PDB 3hhr)

Protein assemblies built of identical subunits are usually symmetric

Protein complexes are built up through interactions across complementary binding surfaces. If one subunit has binding region A, the subunit it binds to must have the complementary region A′. If the interacting subunits are not identical, then nothing definite can be said about the spatial relationship of the monomers in the complex and the complex is said to be asymmetric. The human growth hormone–receptor complex is an example of an asymmetric complex (Figure 1-73).

If the subunits are identical, however, interactions across complementary surfaces nearly always produce symmetric complexes, in which the subunits are related to one another with one of a few kinds of geometry (Figure 1-74). Identical subunits form symmetric complexes because, in order to interact, each subunit must possess binding region A and its complement A′. (This is in contrast to non-identical subunits, each of which has only one or the other.) Depending on the location of A and A′ on the surface, subunits can associate to form closed structures (Figure 1-75a and b), with dimers and trimers being most common, or, much more rarely, open-ended chains, with helical arrangements being most common.

The repeating unit from which a symmetric complex is built can be either a monomer or an association of unlike polypeptide chains. For example, hemoglobin, which is constructed from four polypeptide chains, (a2b2), is a symmetric dimer of two (ab) units. The asymmetric unit from which a symmetric complex is built is referred to as the **protomer**.

If the subunit has a second set of complementary binding regions, B and B′, in addition to A and A′, it can associate to form more elaborate complexes (Figure 1-75c). A second binding region can allow symmetric rings to pair, with pairs of dimers that form tetramers and pairs of trimers that form hexamers being the most common. Insulin is an example of a hexameric protein that is built in this way (Figure 1-74f). In a similar way, open-ended chains can associate side-by-side to form multistranded helices. This is what happens in sickle-cell hemoglobin, when an additional binding site is created by mutation (see Figure 1-71). Subunits with two sets of complementary binding regions can also associate into more complex structures, usually described by reference to geometric figures—tetrahedra, octahedra, icosahedra—with the same symmetry. Type II 3-dehydroquinate dehydratase, for example, crystallizes as a dodecamer in which a tetramer of trimers forms a tetrahedron (Figure 1-74j); and the rhinovirus that causes the common cold is a large multisubunit icosahedron (Figure 1-74k).

So powerful is the tendency of subunits to form symmetric arrangements that this even influences the structure of oligomers made up of non-identical polypeptide chains. Many of these proteins are **pseudosymmetric**, as we have already seen for hemoglobin. In this protein the alpha and beta subunits are similar in sequence and hence nearly identical in structure, so it is a nearly symmetrical tetramer of four monomers. The giant multisubunit proteolytic complex called the proteasome is another example of a pseudosymmetric structure (Figure 1-74l).

Figure 1-74 **Examples of quaternary arrangements observed for oligomeric proteins** The structures shown in **a-k** are homo-oligomers. The proteasome **(l)** is a pseudo-symmetric structure, in which the subunits are not identical. **(a)** D-amino acid aminotransferase (PDB 3daa); **(b)** KDGP aldolase (PDB 1fq0); **(c)** neuraminidase (PDB 1a4q); **(d)** lactate dehydrogenase (PDB 1ldn); **(e)** cholera toxin (PDB 1chp); **(f)** insulin (PDB 4ins); **(g)** molybdenum cofactor biosynthesis protein C (PDB 1ekr); **(h)** GroES co-chaperonin (PDB 1g31); **(i)** galactonate dehydratase; **(j)** 3-dehydroquinate dehydratase (PDB 2dhq); **(k)** rhinovirus (PDB 1aym): this multisubunit protein has the same geometry as a soccer ball; **(l)** proteasome (PDB 1g65).

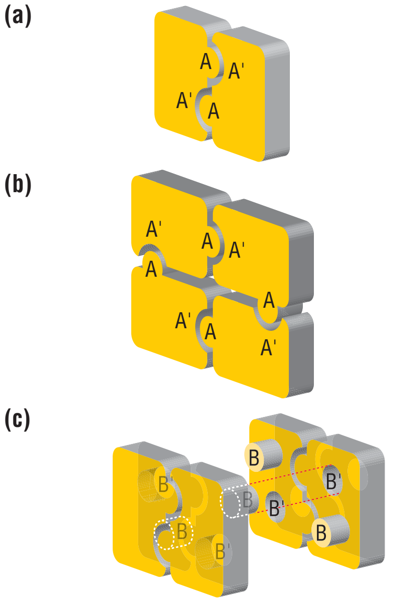


Figure 1-75 **Interactions underlying different geometric arrangements of subunits** Subunits with a pair of complementary binding sites A and A′ may form symmetric dimers **(a)** or tetramers **(b)** depending on the positions of the two binding sites. More complex assemblies may be formed by subunits with a second pair of complementary binding sites B and B′ that could for example allow the formation **(c)** of a tetrameric complex of two dimers.

Definitions

protomer: the asymmetric repeating unit (or units) from which an oligomeric protein is built up.

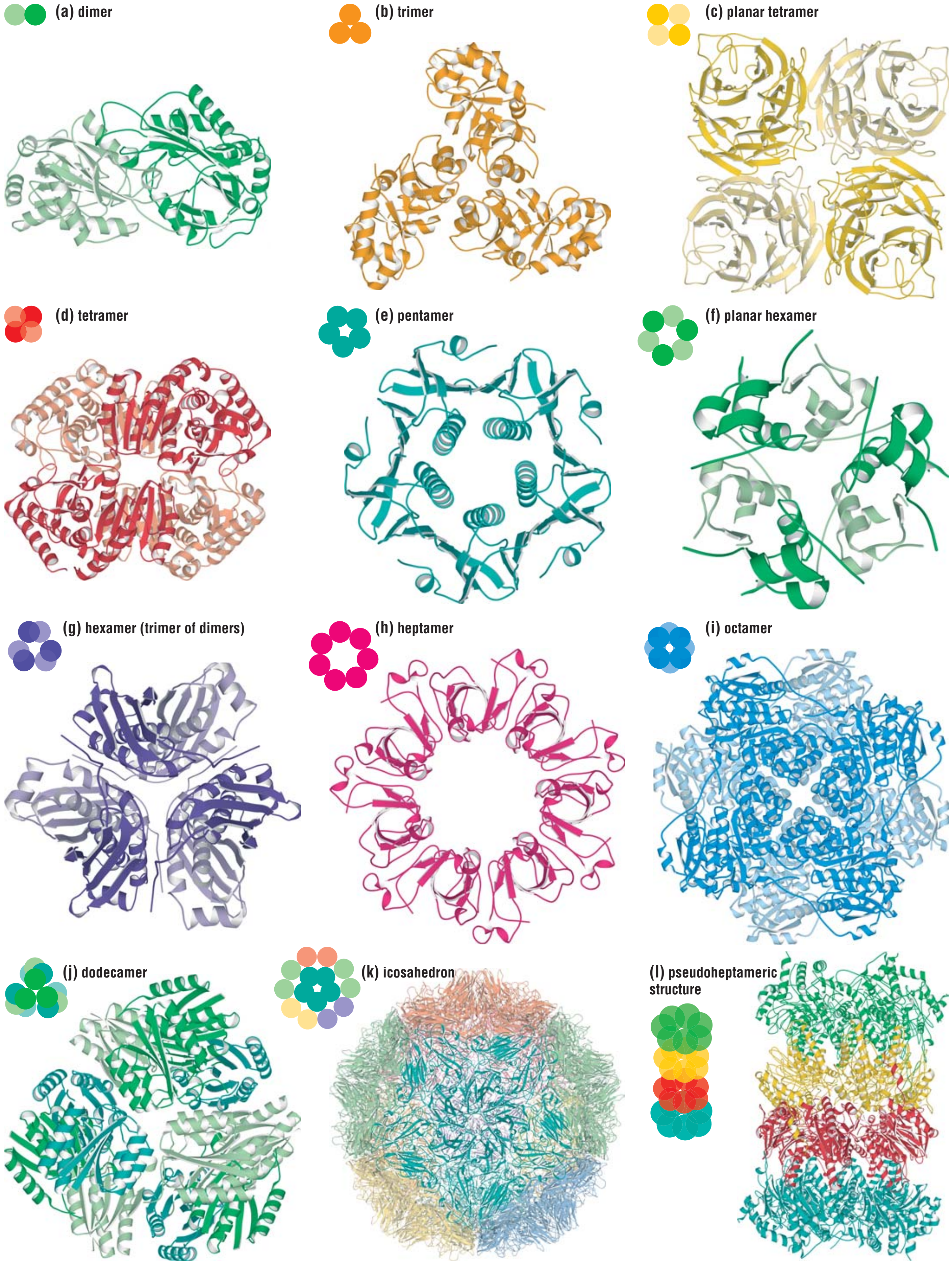
pseudosymmetric: having approximate but not exact symmetry. A protein with two non-identical subunits of very similar three-dimensional structure is a pseudosymmetric dimer.

References

Goodsell, D.S. and Olson, A.J.: **Structural symmetry and protein function.** *Annu. Rev. Biophys. Biomol. Struct.* 2000, **29**:105–153.

Matthews, B.W. and Bernhard, S.A.: **Structure and symmetry of oligomeric enzymes.** *Annu. Rev. Biophys. Bioeng.* 1973, **2**:257–317.

Milner-White, E.J.: **Description of the quaternary structure of tetrameric proteins. Forms that show either right-handed and left-handed symmetry at the subunit.** *Biochem. J.* 1980, **187**:297–302.


(a) dimer
(b) trimer
(c) planar tetramer
(d) tetramer
(e) pentamer
(f) planar hexamer
(g) hexamer (trimer of dimers)
(h) heptamer
(i) octamer
(j) dodecamer
(k) icosahedron
(l) pseudoheptameric structure

1-22 Protein Flexibility

Types of Motion Found in Proteins (all values approximate)

Motion	Spatial displacement (Å)	Characteristic time (s)	Energy source
Fluctuations (e.g., atomic vibrations)	0.01 to1	10^{-15} to 10^{-11}	k_bT
Collective motions (A) fast, infrequent (e.g., Tyr, Phe ring flips) (B) slow (e.g., domain movement; hinge-bending)	0.01 to > 5	10^{-12} to 10^{-3}	k_bT
Triggered conformational changes	0.5 to > 10	10^{-9} to 10^{3}	Binding interactions

Figure 1-76 Table of protein motions

Proteins are flexible molecules

The pictures of protein structures that emerge from X-ray crystallography and NMR seem rigid and static; in reality, proteins are highly flexible. Because the forces that maintain the secondary and tertiary folds are weak, there is enough energy available at body temperature to break any particular interaction. When existing weak interactions are broken, the groups that are released can make new interactions of comparable energy. These rearrangements can occur on a time scale that is faster than the time required to determine the structure by tools such as X-ray crystallography. Thus, the three-dimensional structures of proteins determined by physical techniques are average structures.

Protein motions can be classified in terms of their relationship to the average structure (Figure 1-76). The fastest motions are atomic fluctuations such as interatomic vibrations and the rotations of methyl groups. Next come collective motions of bonded and non-bonded neighboring groups of atoms, such as the wig-wag motions of long side chains or the flip-flopping of short peptide loops. The slowest motions are large-scale, ligand-induced conformational changes of whole domains.

Conformational fluctuations in domain structure tend to be local

It is almost as important to understand what types of conformational change are not observed in proteins as it is to realize that they are flexible in the first place. Whole folded domains never undergo large, thermally driven distortions at ordinary temperatures. Transitions from one type of folding motif to another are rarely seen except in pathological cases; an all alpha-helical protein will not normally refold to an all beta-sheet protein, except, for example, in the cases of amyloid and prion diseases. Smaller-scale refolding does occur in some proteins, however. Ligand binding may induce disordered polypeptide segments to become ordered. Ligands can also induce the disordering of a previously ordered strand, although this is less common. Association and dissociation of subunits can also be triggered by ligand binding, and the ligand can be as small as a proton if it changes the charge of a crucial residue.

Perhaps the most common ligand-induced conformational change is the lid-like movement of a polypeptide segment to cover a ligand-binding site (Figure 1-77). When the lid is open, there is free access to the ligand-binding site. Once the site is occupied, the loop interacts with the ligand to stabilize the closed conformation, and closure isolates the bound ligand from the surrounding solvent. Most loop closures involve rigid-body movement of the loop on two hinges. The internal conformation of the loop does not change appreciably because its side chains are packed closely together, making it function like a solid lid for the ligand-binding site. Mobile loops both act as gates for ligand binding and can make interactions that stabilize the complex. They play an important part in many enzymes.

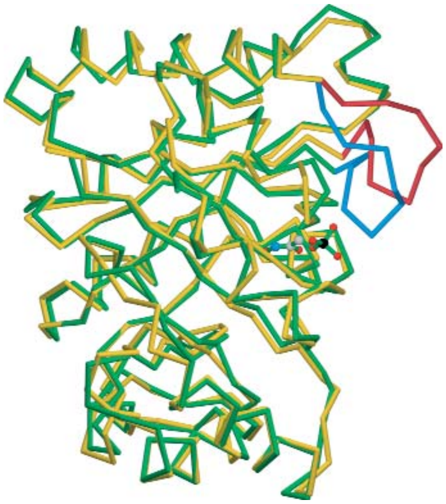


Figure 1-77 Triosephosphate isomerase Binding of substrate or inhibitor to the active site of the enzyme triosephosphate isomerase induces a 10 Å rigid-body movement in an eight-residue loop (red; open) which closes down over the active site (blue; closed) and shields the substrate from solvent. The inhibitor can be seen just below the loop.

Protein motions involve groups of non-bonded as well as covalently bonded atoms

At body temperature, the atoms in most protein molecules fluctuate around their average positions by up to an Ångstrom or occasionally even more, depending on their position in the protein (Figure 1-78). In the tightly packed interior, atomic motions are restricted to less than an Ångstrom. The closer to the surface of the molecule, the greater the increase in mobility until, for surface groups that are not surrounded by other atoms, the mean fluctuation may be several Ångstroms. Proteins have been called “semi-liquid” because the movements of their atoms are larger than those found in solids such as NaCl, but smaller than those observed in a liquid like water.

In a protein, the covalent structure of the polymer sets limits on the motions of atoms and groups of atoms. Chemical groups such as methyl groups or aromatic side chains display collective motions. Methyl groups rotate on a picosecond time-scale; aromatic rings, even those in the interior of the protein, flip at average rates of several thousand per second. The actual ring flip takes only about a picosecond, but it happens only about once every 10^9 picoseconds.

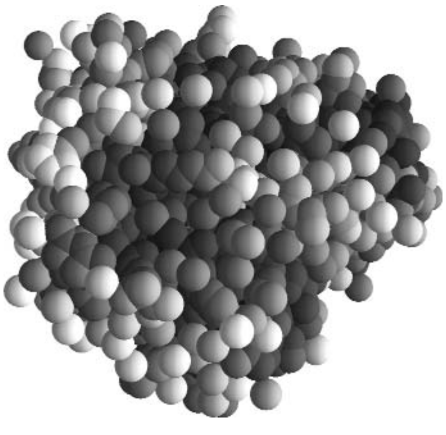


Figure 1-78 Protein shaded according to flexibility Space-filling model of sperm whale myoglobin in which each atom is shaded according to its average motion as determined by X-ray crystallography. The darker the atom, the more rigid it is. Note that the surface is not uniform in its flexibility. (PDB 1a6k)

Flipping an aromatic ring inside a protein, where the packing density is high, requires that surrounding atoms move out of the way. The probability that they will all move in the right direction at the same time is very low: hence the relatively long interval between flips. In the interior of proteins, close atomic packing couples the motions of non-bonded neighboring atoms. If a methyl group in the center of a protein is next to another methyl group, the motions of both will be correlated by virtue of their tendency to collide. Thus both the extent of motion of every group and its preferred directions depend on non-bonded as well as bonded contacts. Only for surface side chains and protruding loops are non-bonded interactions of little importance, and residues in such unrestrained positions are always the most flexible parts of a protein structure.

At biological temperatures, some proteins alternate between well-defined, distinct conformations (Figure 1-79). In order for two conformational states to be distinct, there must be a free-energy barrier separating them. The motions involved to get from one state to the other are usually much more complex than the oscillation of atoms and groups about their average positions. It is often the case that only one of the alternative conformations of a protein is biologically active.

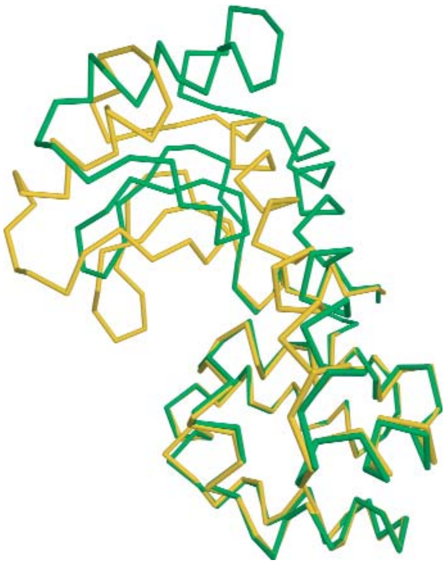


Figure 1-79 T4 lysozyme The enzyme T4 lysozyme contains two domains connected by a hinge. In different crystal forms of the protein, an open and closed state have been observed, related to each other by a hinge-bending motion. It is presumed that the protein in solution can exist in an equilibrium between both states at physiological temperature. (PDB 1l96 and 1l97)

Triggered conformational changes can cause large movements of side chains, loops, or domains

Of most importance for protein function are those motions that occur in response to the binding of another molecule. Ligand-induced conformational changes can be as modest as the rearrangement of a single side chain, or as complex as the movement of an entire domain. In all cases, the driving force is provided by ligand–protein interactions.

Often, the motion enables some part of the structure to make contact with a ligand. For example, the binding of aspartate to a large domain of the enzyme aspartate aminotransferase causes a smaller domain to rotate by 10°. This rotation moves the small domain by more than 5 Å, bringing it into closer contact with the rest of the protein (Figure 1-80). When the binding of a specific ligand causes a protein to change from an inactive to an active conformation, the process is described as **induced fit**. The driving force for induced fit in aspartate aminotransferase appears to be the formation of a salt bridge between an arginine residue in the mobile domain and the alpha-carboxylate of the bound aspartate. Mutant enzymes that are unable to carry out this triggered conformational change are inactive. We discuss the use of conformational changes to regulate enzymes in more detail in Chapter 2.

Ligand-induced conformational changes can also change the quaternary structure of proteins. This usually involves repacking of the interfaces between subunits so that the relative positions of the monomers are altered; this happens when oxygen binds to the tetramer hemoglobin. Sometimes, however, the stoichiometry of the oligomer changes on ligand binding. One example is the polymerization of actin monomers, driven by the binding of ATP, into linear helical polymers called thin filaments or microfilaments. Regulated polymerization of actin is essential for the formation and disassembly of cytoskeletal components needed for cell movement. Changing the oligomeric state of actin is a mechanism for controlling what it does.

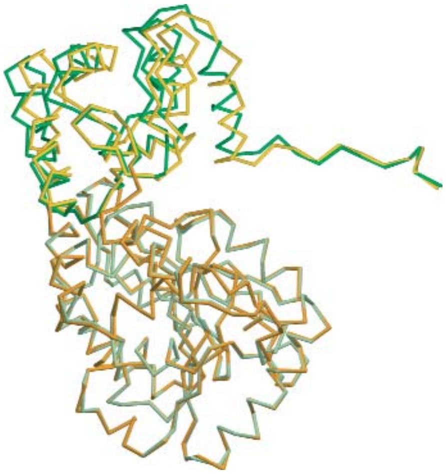


Figure 1-80 Aspartate aminotransferase, open and closed forms The enzyme L-aspartate aminotransferase contains two domains with the active site lying between them. Substrate binding induces a movement of the small domain (green) to a new position (yellow) in which the active site is more enclosed. This movement is essential to position some of the residues important for catalysis, and only the specific substrates of the enzyme induce it. (PDB 1ars and 1art)

Definitions

induced fit: a change in the conformation of a protein induced by the binding of a ligand. In the case of an enzyme, this may result in catalytic activation.

References

Arrondo, J.L. and Goni, F.M.: **Structure and dynamics of membrane proteins as studied by infrared spectroscopy.** *Prog. Biophys. Mol. Biol.* 1999, **72**:367–405.

Daggett, V.: **Long timescale simulations.** *Curr. Opin. Struct. Biol.* 2000, **10**:160–164.

Ishima, R. and Torchia, D.A.: **Protein dynamics from NMR.** *Nat. Struct. Biol.* 2000, **7**:740–743.

Karplus, M. and Petsko, G.A.: **Molecular dynamics simulations in biology.** *Nature* 1990, **347**:631–639.

Petsko, G.A. and Ringe, D.: **Fluctuations in protein structure from X-ray diffraction.** *Annu. Rev. Biophys. Bioeng.* 1984, **13**:331–371.

Ringe, D. and Petsko, G.A.: **Mapping protein dynamics by X-ray diffraction.** *Prog. Biophys. Mol. Biol.* 1985, **45**:197–235.

Wall, M.E. *et al.*: **Large-scale shape changes in proteins and macromolecular complexes.** *Annu. Rev. Phys. Chem.* 2000, **51**:355–380.