4
Protein Structure Hierarchy

Chapter 4 Notation

SYMBOL	DEFINITION
α_R	classic right-handed α-helix
α/β	protein class
$\alpha+\beta$	protein class
β-sheet	aggregating amino-acid strands
β_2	hairpin motif
β_4	Greek key motif
C^α	α-Carbon
π	helix form looser than α_R
ϕ	$\{\text{N–C}^\alpha\}$ rotation about peptide bond
ψ	$\{\text{C}^\alpha\text{–C}\}$=O rotation about peptide bond
3_{10}	helix form tighter than α_R

Try to learn something about everything and everything about something.

Thomas Henry Huxley (1825–1895).

T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*, Interdisciplinary Applied Mathematics 21, DOI 10.1007/978-1-4419-6351-2_4,

4.1 Structure Hierarchy

The complexity of protein structures requires a description of their structural components. This chapter describes the elements of protein secondary structure — regular local structural patterns — such as helices, sheets, turns, and loops. Helices and sheets tend to fall into specific regions in the $\{\phi, \psi\}$ space of the Ramachandran plot (see Figures 3.18 and 3.19). The corresponding width and shape of each region reflects the spread of that motif as found in proteins.

Following this description of each secondary structural element, we discuss the basic four *classes* of protein supersecondary or tertiary structure (the 3D spatial architecture of a protein), namely α-proteins, β-proteins, α/β-proteins, and $\alpha + \beta$-proteins. This is followed by a presentation of the *fold* motifs for each such class. Classes and folds are at the top of protein structure classification, as introduced in the last section. Describing these folds and structural motifs is far from an exact science, so variations in some of these aspects are common.

4.2 Helices: A Common Secondary Structural Element

4.2.1 Classic α-Helix

In the classic, right-handed α-helix (α_R), a *hydrogen bonding* network connects each backbone carbonyl (C=O) oxygen of residue i to the backbone hydrogen of the NH group of residue $i + 4$ (see Figure 4.1). This hydrogen bonding provides substantial stabilization energy.

The regular spiral network of the α-helix is ubiquitous in proteins. It is associated with a $\{\phi, \psi\}$ pair of about $\{-60^o, -50^o\}$. The resulting helix has 3.6 residues per turn, and each residue occupies approximately 1.5 Å in length. The helix may be curved or kinked depending on the amino acid sequence, as well as on solvation and overall packing effects. Such distortions are reflected by the $\{\phi, \psi\}$ distribution around the α_R region in typical Ramachandran plots. **Hemoglobin**, **myoglobin**, **bacteriorhodopsin**, **human lysozyme**, **T4 lysozyme**, **Trp repressor**, and **repressor-of-primer (Rop)** are all examples of proteins that are virtually entirely α-helical. See Figures 4.2 and 4.3 for illustrations of such α-proteins (see below) and Figure 3.10 for Rop.

An α-helix is associated with a dipole moment: the amino terminus of the helix has a positive charge and the carboxyl end has a negative charge clustered about it. Thus, residues that are negatively charged on the amino end and positively-charged on the carboxyl end stabilize the helix; residues with the opposite charge allocation destabilize the helix.

Experimental and theoretical work has shown that both intrinsic and extrinsic (inter-residue interactions) factors are important for helix formation in proteins. Residues with restricted sidechain conformations, due to long or bulky groups, are poorer α-helix participants than other residues. Glutamine, methionine,

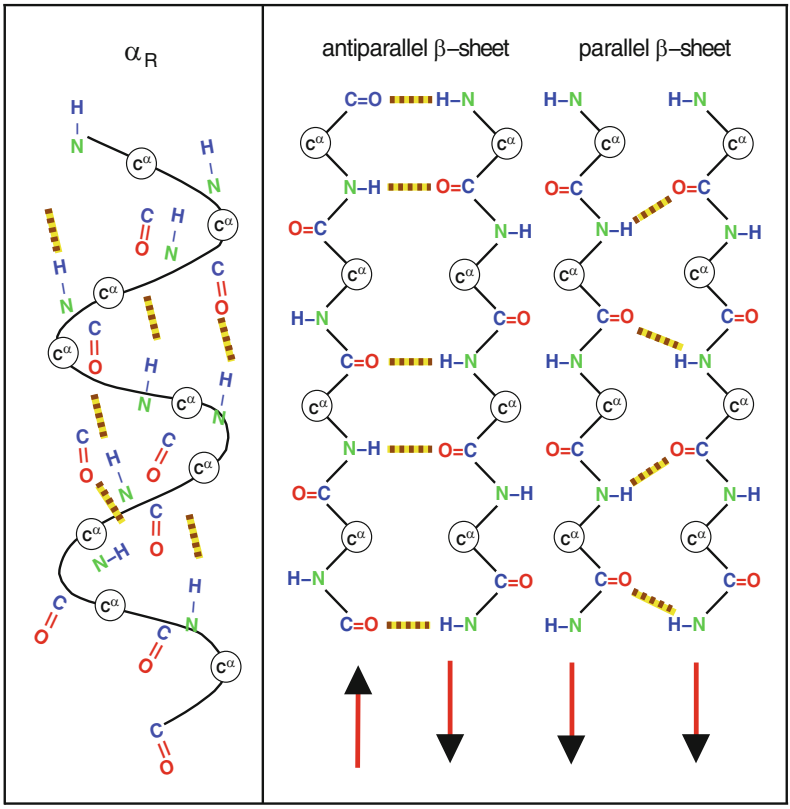


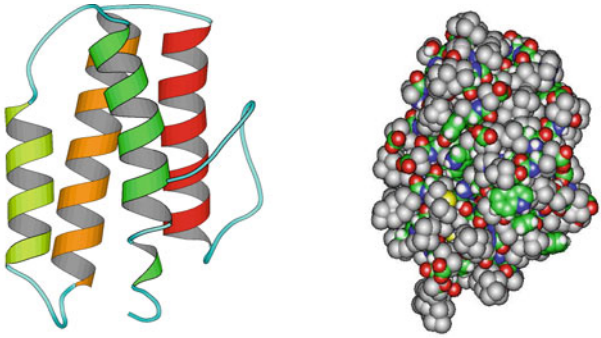
Figure 4.1. Hydrogen bonding patterns in the classic α-helix (α_R), with the ribbon tracing the α-carbons (left), anti-parallel β-sheet (middle), and parallel β-sheet (right).

and leucine favor α-helix formation, while valine, serine, aspartic acid, and asparagine tend to destabilize α-helices (e.g., due to steric and electrostatic considerations).

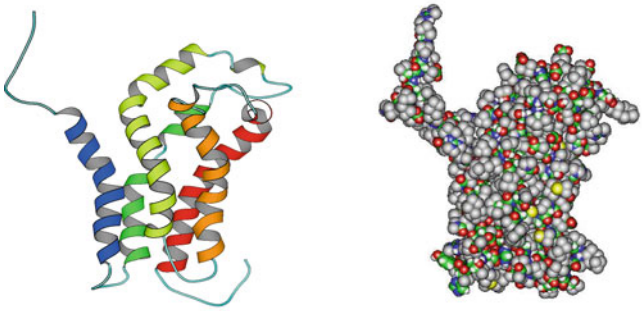
4.2.2 3_{10} and π Helices

There are more common variants of the α-helix motif that are typically not stable in solution but can play a part in overall protein structure. These include the tighter 3_{10} and looser π helices, with $\{\phi, \psi\}$ angles around $\{-50^o, -25^o\}$ and $\{-60^o, -70^o\}$, respectively.

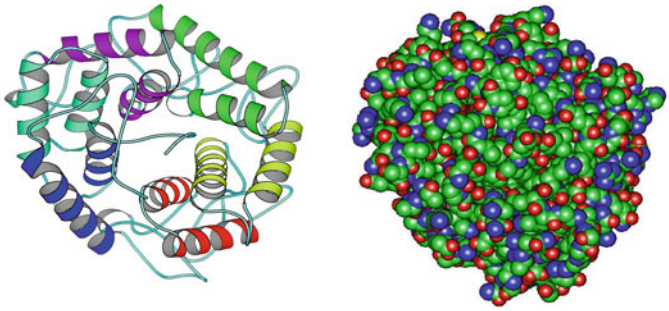
The tighter 3_{10} helix of three residues per turn (instead of 3.6 in the classic α-helix) involves hydrogen bonds between residues i and $i+3$ instead of i and $i+4$ as in α_R. There are 10 atoms within the hydrogen bond; hence the nomenclature 3_{10}. The more loosely coiled π helix has hydrogen bonds between residues i and $i + 5$ of the polypeptide.



Myohemerythrin (2MHR, 118 residues, four–helix bundle)



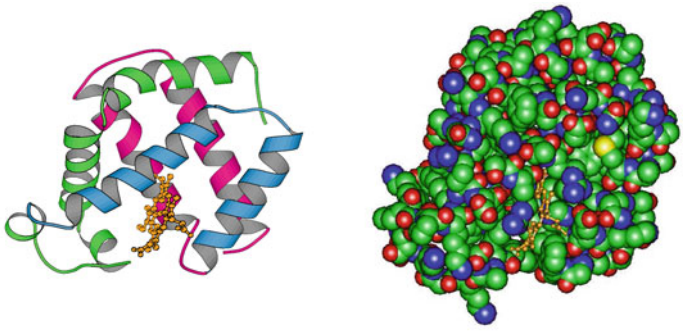
Pix (1BY1, 209 residues, five–helix bundle)



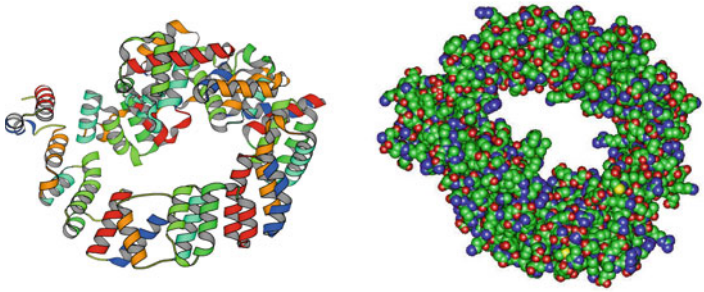
Cellulase Cela (1CEM, 363 residues, six–alpha hairpins)

Figure 4.2. Examples of α-proteins: **myohemerythrin**, **pix**, and **cellulase cela**.

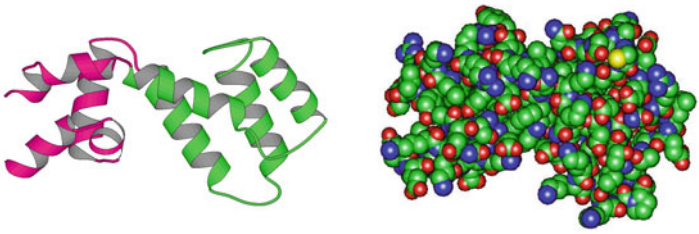
Because of their close packing, 3_{10} helices generally form for a few residues only, often at the C-terminus end of classic α-helices where the helix tends to tighten. Similarly, the π helix occurs rarely since the backbone atoms are so loosely packed that they leave a hole.



Myoglobin (5MBA, 147 residues, folded leaf)



Soluble Lytic Transglycosylase Slt70
(1QSA, 618 residues, folded leaf superhelix)



Guanine Nucleotide–Binding Protein G(I)
(1AGR, 205 residues, two all–alpha domains)

Figure 4.3. Examples of α-proteins: **myoglobin**; **soluble lytic transglycosylate** protein of bacterial muramidase, in the N-terminal region of the enzyme muramidase in bacterial cell walls; and **guanine nucleotide-binding protein**, an irregular α-helical protein with a fold containing a 4-helix bundle with left-handed twist.

4.2.3 Left-Handed α-Helix

A left-handed α-helix is theoretically possible, with $\{\phi, \psi\} = \{+60^{\circ}, +60^{\circ}\}$. However, this motif is generally unstable. The chirality preference for α-helices follows the chirality of L-amino acids.

4.2.4 *Collagen Helix*

The triple-stranded **collagen** helix is often considered a specific secondary element. It is associated with $\{\phi, \psi\} = \{-60^{\circ}, +125^{\circ}\}$. A large body of structural data has suggested that extensive hydration networks in the collagen triple helix (among the protein residues and with water molecules) are responsible for collagen stability and assembly (see [115, 680] and references cited therein). A recent hypothesis — that inductive effects by electron-withdrawing residue moieties might play a key factor in collagen's stability [562] — remains to be proven.

4.3 β-Sheets: A Common Secondary Structural Element

Another common motif is a β-sheet. These sheet regions form by aggregating amino-acid strands, termed β-strands, via hydrogen bonds. Typical lengths of β-strands are 5–10 residues. The aggregation can occur in a parallel or anti-parallel orientation of the strands, as shown in Figure 4.1, each with a distinct hydrogen bonding pattern. Each such β-strand has two residues per turn and can be considered a special type of helix. The hydrogen bond crosslinking between strands — alternating C=O $\cdots$ H–N and N–H $\cdots$ O=C — is such that the sheet has a pleated appearance. Thus, in comparison to α-helices, β-sheets require connectivity interactions that are much longer in range.

For parallel β-sheets, $\phi \approx -120^{\circ}$ and $\psi \approx +115^{\circ}$. For anti-parallel β-sheets, $\phi \approx -140^{\circ}$ and $\psi \approx +135^{\circ}$. As for α-helices, the ring of proline does not adapt well into β-sheets since it cannot participate in the hydrogen bond network between strands. Valine, isoleucine, and phenylalanine have been found to enhance β-sheet formation.

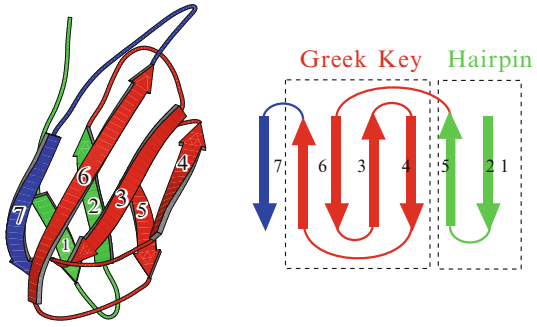
Often, at the edges of β-sheets, an additional residue that cannot be included in the normal hydrogen bonding pattern produces a *β-bulge* of the extra residue. Figures 4.4 and 4.5 show the structures of proteins that are mostly β-sheets.

4.4 Turns and Loops

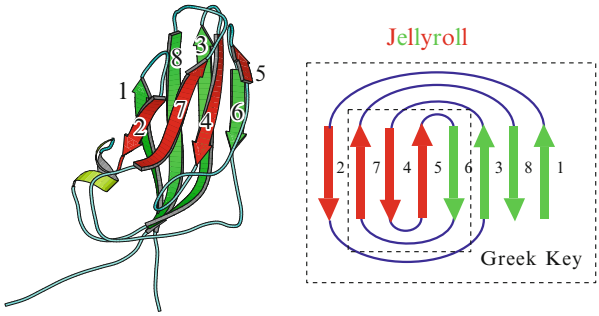
Other common structural motifs in proteins are turns and loops.

Turns (also called β-turns or reverse turns) occur in regions of sharp reversal of orientation, such as the junction of two anti-parallel β-strands. Such motifs are classified as turns based on distance criteria (e.g., the C^{α} atoms of residues i and $i + 3$ are less than 7 Å distant).

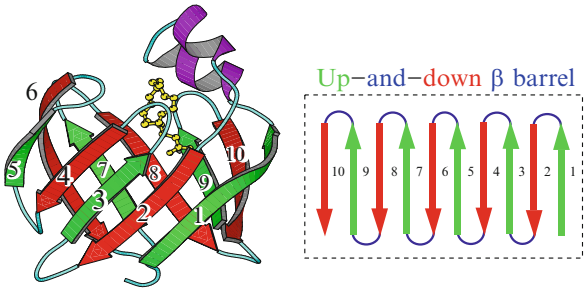
Loops occur often in short (five residues or less) regions connecting various motifs. Loop regions that connect two adjacent anti-parallel β-strands are known as hairpin loops. Short hairpin loops are found at protein surfaces.



Fibronectin (1TTG, 94 residues, β sandwich)



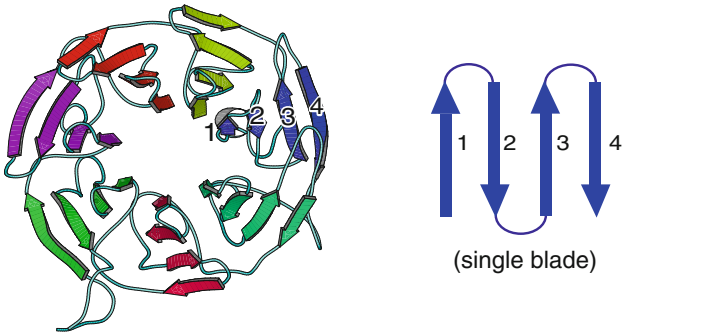
Satellite Panicum Mosaic Virus (1STM, 157 residues)



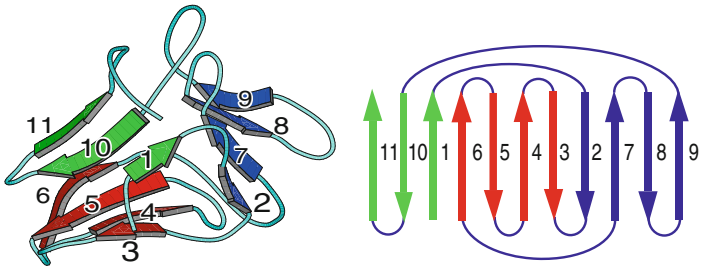
Fatty Acid Binding Protein (1HMS, 132 residues, β barrel)

Figure 4.4. Examples of β-proteins and common motifs: **fibronectin**, β-sandwich illustrating hairpin and Greek key motifs; coat protein of **satellite panicum mosaic virus**; and **fatty acid binding protein**, up-and-down β-barrel.

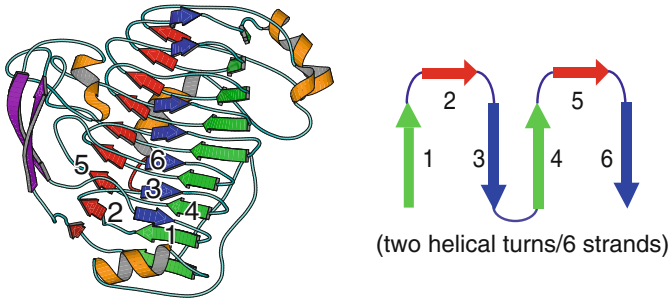
The majority of turns and loops lies on the protein surface because of solvation considerations. They are important elements that allow, and possibly drive, protein compaction. Most loops interact with solvent and are highly hydrophilic (water soluble). Since protein core regions are more stable than short



Galactose Oxidase (1GOF, 639 residues, 7–bladed β–propeller)



Agglutinin (1BWU, 106 residues, β–prism)



Pectin Lyase A (1IDK, 359 residues, right–handedβ–helix)

Figure 4.5. Examples of β-proteins and common motifs: **galactose oxidase**, **agglutinin**, and **pectin lyase A**.

connective elements of helices and strands, evolutionary differences among homologous sequences are often localized to loop and turn regions. Non-coding regions (introns) are similarly found in genes that correspond to loops and turns in protein structures.

4.5 Formation of Supersecondary and Tertiary Structure

4.5.1 Complex 3D Networks

The secondary structural elements described above often combine into simple motifs that occur frequently in protein structures. Such motifs (or folds) are also called *supersecondary structure*. Examples are β hairpin (β-loop-β units), Greek key, and β-α-β units (see below).

Supersecondary and tertiary structures of proteins can be described by the specific topological arrangement of the secondary or supersecondary structural motifs. Although the 3D architecture of a protein can be a complex composite of various secondary and supersecondary structural motifs, the majority of the residues — roughly 90% — are found to be involved in secondary structural elements. In fact, on average 30% of the residues are found as helices, 30% as sheets, and 30% as loops and turns. Proteins can be monomeric or multimeric, with subunits that fold in a dependent or independent manner with respect to other domains.

The different polypeptide domains can be connected by disulfide bonds, hydrogen bonds, or the weaker van der Waals interactions. Tertiary structure is also affected by the environment. Hydrogen bonding with solvent water molecules can stabilize the native conformation, and the salt concentration can affect the compact arrangement of the folded chain.

Molecular graphics packages often display the secondary structural motifs clearly by using ribbon diagrams in which helices are depicted as coils and sheets are shown as twisting planes with arrows (see Figures 4.2, 4.3, 4.4, and 4.5, for example).

4.5.2 Classes in Protein Architecture

Based on the known protein structures at atomic resolution, four major *classes* can be used to describe the arrangement in space of the various secondary structural elements (or domains) of polypeptides:

- α-proteins — proteins which form compact aggregates by packing mainly α-helices, often in a symmetric arrangement around a central hydrophobic core;
- β-proteins — proteins which pack together mainly β-sheets, with adjacent strands linked by turns and loops and various hydrogen bonding networks formed among the individual strands, often resulting in layered or barrel structures;
- α/β-proteins — proteins that are folded with alternating α-helices and β-strands, often forming layered or barrel-like structures; and
- $\alpha + \beta$-proteins — proteins that combine largely-separated (i.e., non-alternating) helical and strand regions, often by hairpins.

Figures 4.2–4.7 illustrate members of each such class.

Recent statistics for PDB protein structures reveal that approximately 24% belong to the all-α class, 15% to all-β, 12% to α/β, and 32% to $\alpha+\beta$. The remaining 17% includes multidomain proteins, membrane and cell-surface proteins, and peptides, and small proteins (see Figures 4.8–4.10). For updated statistical information, check scop.mrc-lmb.cam.ac.uk/scop/, click on 'Statistics here'. (See last section of this chapter for SCOP description).

Other classes are defined for proteins found on membrane and cell surfaces, small and/or irregular proteins with multiple disulfide bridges, proteins with multiple domains or with bound ligands, and more. Included, for example, are small proteins like **rubredoxin** (PDB entry 1rb9), various zinc-finger and metal-binding proteins like the cysteine-rich domain of **protein kinase** (PDB entry 1ptq), disulphide-rich proteins like **sea anemone toxin k** (PDB entry 1roo), and **proteinase inhibitor PMP-C** (PDB entry 1pmc).

4.5.3 Classes are Further Divided into Folds

The protein *classes* are further divided into observed *folds* for protein structures. Folds describe the arrangement of secondary structural elements and/or chain topology. Each protein class has common folds, as described in turn in the next three sections.

4.6 α-Class Folds

In the α-class of proteins (Figures 4.2 and 4.3), bundles, folded leafs, and hairpin arrays are major fold groups.

4.6.1 Bundles

Bundles occur when α-helices pack together to produce a hydrophobic core. Typically, an array of α-helices is roughly aligned around a central axis. The bundle can be right or left-handed depending on the twist that each helix makes with respect to this axis. A *coiled coil* (two intertwined helices) can be a building block of these bundles. A simple example of a coiled coil is seen in the **DNA-binding leucine zipper protein** shown in Figure 6.5 of Chapter 6.

Among the α-protein bundles, the four-helix bundle motif (often written as α_4) is common, as in **myohemerythrin**, Figure 4.2, and **Rop** (a small RNA-binding protein involved in replication), Figure 3.10. Other α_4 proteins are **ferritin** (a storage molecule for iron in eukaryotes), **cytochrome c′** (heme-containing electron carrier), the **coat protein of tobacco mosaic virus**, and **human growth hormone**.

Multi-helical bundles are also observed in α-proteins; 3–6 and 8-helix aggregates are more frequent than others. Figure 4.2 shows a 5-helix bundle for the transport protein **pix**.

4.6.2 Folded Leafs

Complex packing patterns involving layered arrangements are often features of long α-proteins. For example, in folded leaf folds, a layer of α-helices wraps around a central hydrophobic core. Like bundles, such multihelical assemblies (usually 3 or more) pack together as well as form layers. The longest helices are usually in the center, and often the arrangement contains internal pseudosymmetry.

The globin fold of **myoglobin** (Figure 4.3) shows such a compact arrangement of a folded leaf arrangement formed by 8 helices, leaving a pocket for heme binding. **Cytochrome C6** in Figure 3.12 also displays a folded leaf.

A more complex layered topology is the two-layered ring structure of one α-helical domain in the N-terminal region of the enzyme **muramidase** in bacterial cell walls, **soluble lytic transglycosylate** (Figure 4.3). It is built from 27 α-helices, arranged in a two-layered superhelix, leaving a large central hole, thought to be important in its catalytic activity.

4.6.3 Hairpin Arrays

Other α-helix assemblies that cannot be described by bundle or folded leaf motifs are often described as hairpin arrays (arrays of α-helix /loop /α-helix motifs). The calcium binding protein **calmodulin**, for example, has a helix/loop/helix motif where the loop region between two helices binds calcium (see Figure 3.11). Figure 4.2 also shows **cellulase cela**, a toroid-like circular array composed from 6 hairpins.

An irregular α-protein from an all-α subdomain of the regulator of G-protein signaling 4, namely **guanine nucleotide-binding protein**, is also shown in Figure 4.3. This protein's motif contains a 4-helical bundle with left-handed twist and up-and-down topology.

4.7 β-Class Folds

Proteins in the β-class display a flexible and rich array of folds, as seen in Figures 4.4 and 4.5. Various connectivity topologies can exist within networks of *parallel*, *anti-parallel*, or *mixed* β-sheets that twist, coil and bend in various ways. Indeed, note the much wider regions of the Ramachandran plot associated with β-sheets than with α-helices (Figs. 3.18 and 3.19).

4.7.1 Anti-Parallel β Domains

To describe these intriguing folds, it is simpler to begin with folds associated with the large subclass of β-proteins made exclusively of *anti-parallel* β domains. Such proteins tend to form distorted barrel structures. They can be described in terms of building blocks of two-strand, four-strand, eight-strand units, etc., as follows.

Two-Strand Units

The basic two-strand unit, the hairpin (denoted β_2), involves a β/loop/β motif. It has adjacent anti-parallel β-strands linked head-to-tail by a turn or loop; see the β-strands connected as $1 \rightarrow 2$ or $4 \rightarrow 5$ for the head-to-tail direction in **fibronectin** in Figure 4.4.

Four-Strand Units

Proceeding to connections of four β-strands, there are 24 ways to combine two β-hairpin units to form a 4-stranded anti-parallel β-sheet unit. The most common topology is the Greek key (or β_4). The four strands of a Greek key produce a β-sandwich through the head-to-tail connectivity of $3 \rightarrow 4 \rightarrow 5 \rightarrow 6$, as shown in the diagrams for **fibronectin** and **satellite panicum mosaic virus** in Figure 4.4. The β-strands in these illustrations are labeled according to their connectivity in the protein.

Eight-Strand Units

Correspondingly, there are many more ways to combine a larger number of β-strands from motifs of smaller systems. The two most common folds for 8 anti-parallel β-strands are jellyrolls (β_8) and up-and-down β-sheet.

- The appetizing jellyroll is illustrated in Figure 4.4 for the β-sandwich coat protein of **satellite panicum mosaic virus**. It is a network of 8 anti-parallel β-sheets with the connectivity $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8$, where strands are *shuffled* when viewed in the diagram left to right. Note the Greek key submotif in the $4 \rightarrow 5 \rightarrow 6 \rightarrow 7$ subunit of the jellyroll.
- In the up-and-down β-sheet, each β-strand is connected to the next by a short loop. It has the simpler connectivity $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8$, where strands 1 through 8 are written left to right (no shuffling required). Figure 4.4 shows this fold for **fatty acid binding protein** ($1 \rightarrow 2 \rightarrow \cdots \rightarrow 9 \rightarrow 10$).

4.7.2 Parallel and Antiparallel Combinations

More generally, β-protein topologies made of composites of parallel and anti-parallel strands usually form layered or barrel structures. The sandwich, barrel, and β-propeller are three general reference fold groups.

Sandwiches and Barrels

In sandwiches, β-sheets twist and pack with aligned strands, whereas in barrels the sheets twist and coil so that often the first strand is hydrogen bonded to the last strand to produce closed structures. See the sandwich protein **fibronectin** and barrel in **fatty acid binding protein** in Figure 4.4. The immunoglobulins in Figure 3.12(d) are also β-sandwiches where seven strands form two sheets.

Propellers

In β-propeller folds, 6 to 8 β-sheets, each with 4 anti-parallel and twisted strands, arrange radially to resemble a propeller. The 7-bladed propeller of **galactose oxidase** is shown in Figure 4.5.

Other β-Folds

Other β-folds include β-prisms (3 sheets that pack around an approximate 3-fold axis), barrel/sandwich hybrids (2 β-sheets, each shaped as a half barrel and packing like a sandwich), and β-clips (3 two-stranded β-sheets, forming a long hairpin folded upon itself in two locations). **Agglutinin** in Figure 4.5 shows a β-prism fold.

Recently, β-helix structures have been identified [238]. The polypeptides contain up to 16 helical turns, each of which contains 2 or 3 β-sheet strands. Unlike the β-sandwiches, the β-sheet strands of a β-helix have little or no twist. Most such β-helix folds known to date are right-handed, as seen in **pectin lyase A** in Figure 4.5. The β-helix motif has been suggested to occur in the infectious scrapie prion protein [1371].

4.8 α/β and $\alpha+\beta$-Class Folds

Even more diverse fold patterns are known for the α/β and $\alpha+\beta$-classes of proteins depending on the sheet types (parallel, anti-parallel, or mixed network) and the location of the helices (exterior, interior, or on both faces) with respect to the sheet assembly.

We can broadly classify three fold motifs in this class (see Figure 4.6): barrels — closely packed β-strands (usually 8) with α-helices on the exterior, open structures made of twisted β-sheets (parallel or mixed) surrounded by α-helices on both the exterior and interior, and leucine-rich motifs of curved β-sheets with exterior α-helices in leucine-rich regions.

4.8.1 α/β Barrels

A classic example of a barrel core is the barrel structure of **triosephosphate isomerase** (TIM), an $(\alpha/\beta)_8$ topology (see Figure 4.6). The TIM barrel is one of the most common polypeptide-chain folds known today. TIM's 8 parallel

β-strands coil to form a central core, and its 8 α-helices pack along the exterior. The central barrel 'mouth' is the active site of the protein.

4.8.2 *Open Twisted α/β Folds*

An example from the highly-variable class of open twisted α/β structures is **flavodoxin** (Figure 4.6). Note that its helices lie on opposite sides of the β-sheet. Typically, the active sites of proteins in this fold class are near the loop regions that connect β-strands to α-helices. Another member of this class is **maltate dehydrogenase**, characterized by the *Rossmann* fold (named after its discoverer Michael Rossmann). This $(\beta\alpha\beta\alpha\beta)_2$ topology has a central, parallel twisted β-sheet surrounded by α-helices and/or loops. It is an important motif in proteins that bind to nucleic acids.

4.8.3 *Leucine-Rich α/β Folds*

Ribonuclease inhibitor is an example in the leucine-rich class of α/β folds. Its *horseshoe* structure is formed by homologous repeats of right-handed β-loop-α units (see Figure 4.6). The 17 parallel β-strands lie on the inside of this horseshoe, with the 16 α-helices clustering on the outside. The leucine residues present in all three segments of the repeating unit — the β-strand, the loop, and the α-helix — pack snuggly together to form a hydrophobic core between the β-strand and α-helix regions.

4.8.4 *$\alpha+\beta$ Folds*

Yet more complex fold patterns have been observed for the $\alpha+\beta$-class of proteins (see Figure 4.7). This diversity reflects the various topologies of the subdomains (or layers) as well as the richness of connectivity patterns among them.

4.8.5 *Other Folds*

Examples of multi-domain proteins, membrane and cell surface proteins, and small proteins are shown in Figures 4.8, 4.9, and 4.10.

4.9 Number of Folds

It has been postulated that the number of folding motifs is finite and that the entire catalog of folds will eventually be known with the rapidly-increasing number of solved globular proteins [157, 237, 560]. Such postulates come from stereochemical considerations — for example, there is a small number of ways to link compactly α-helices and β-strands — database analyses, and statistical sampling approaches.

4.9.1 Finite Number?

The exact number of folds has not been determined. Some studies estimate this number to be several thousand [266, 780], while others yield only several hundred [1340, 1434] (around 10,000 or 3000 total folds in the former group and 850 total folds in the latter works), so a minimal estimate of around 1000 [1259] and the range of 1000–10,000 seem reasonable [168]. Only time will tell how many folds Nature has produced.

Since many computational folding-prediction schemes use known folds for closely-related sequences or closely-related functions of proteins, a finite number of folds suggests that *eventually* we will be able to describe 3D structures from sequence quite successfully!

Zhang and DeLisi estimated in 1998 [1434], however, that with the technology available at that time, 95% of the folds will only be determined only in 90 years. They argued that, aside from technological improvements, we should carefully select new sequences for structure determination so as to maximize new fold detection and thereby reduce that time substantially. This is important since the annual number of new folds discovered during 1995–2000 has only averaged around 10%, with even less during 2000–2002. Certainly, careful selection of targets is even more critical if the number of folds is actually larger (e.g., of order 10,000) and associated with single sequence families [266]. The structural genomics initiatives (see beginning of Chapter 2) are certainly accelerating the discovery of new folds (see, for example, [48,214]), but the effect of these projects will take time to assess (see, for example, differing opinions in [87, 993]). For updated fold information, search PDB holdings.

4.10 Quaternary Structure

Quaternary structures describe complex interactions for multiple polypeptide chains, each independently folded, with possibly other molecules (nucleic acids, lipids, ions, etc.). The interactions are stabilized by hydrogen bonds, salt bridges, and various other complex intermolecular and intramolecular associations in space. The classic example for a quaternary structure is that of the protein **hemoglobin**, which consists of four polypeptide chains. The four subunits, each of which contains an oxygen-binding heme group, are arranged symmetrically. Other examples of quaternary structure are DNA polymerases (with catalytic and regulatory components) and ion channels, and protein/nucleic acid complexes with complex structures involving many subunits like viruses, nucleosomes, and microtubules.

4.10.1 Viruses

Virus coats are often comprised of many protein molecules and have intriguing quaternary structures. These protein coats envelope the inner domain which

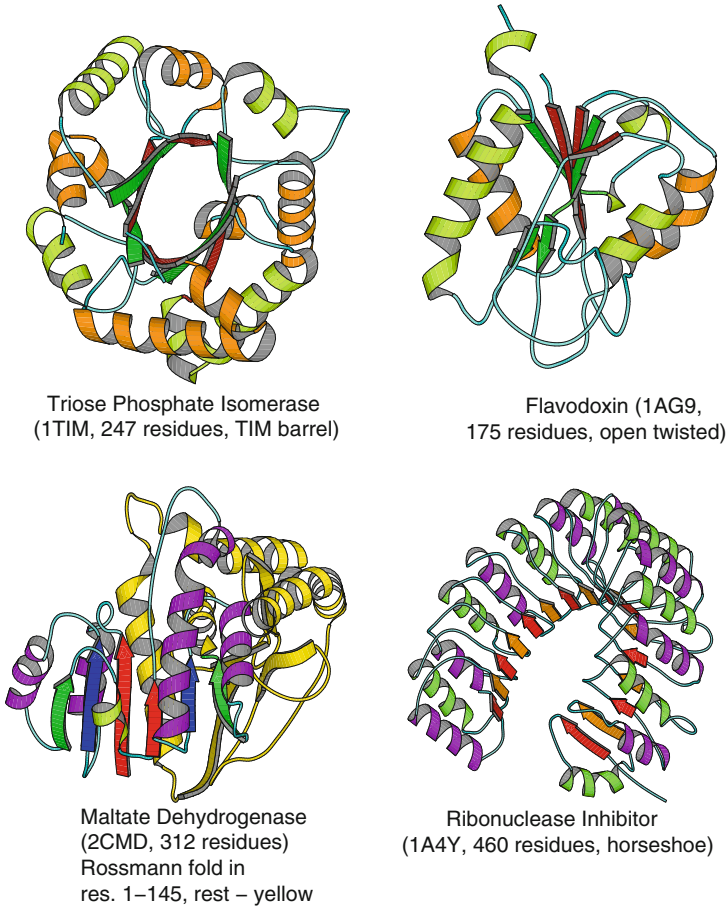


Figure 4.6. Examples of α/β-proteins. **TIM** (triosephosphate isomerase) displays an architecture of 8 twisted parallel β-strands which form a barrel surrounded by α-helices. **Flavodoxin**, an electron transport protein that binds to a flavin mononucleotide prosthetic group, displays an open twisted α/β fold made of three layers (2 helices at left, 5 β-strands in the middle, and 2 helices at right). **Maltate dehydrogenase** contains the $(\beta\alpha\beta\alpha\beta)_2$ *Rossmann* fold in the subunit shown. **Ribonuclease inhibitor**, in the leucine-rich class of α/β folds, displays a *horseshoe* structure.

consists of infectious nucleic acids. For example, the **poliovirus** — a spherical complex of 310 Å in diameter — has a shell of 60 copies of each of four proteins. The coat of **tobacco mosaic virus** combines 2130 identical protein units, each of 158 residues, arranged in a helix around a coiled RNA structure of 6400 nucleotides. This results in a rod-shaped complex 3000 Å long and 18 Å in diameter.

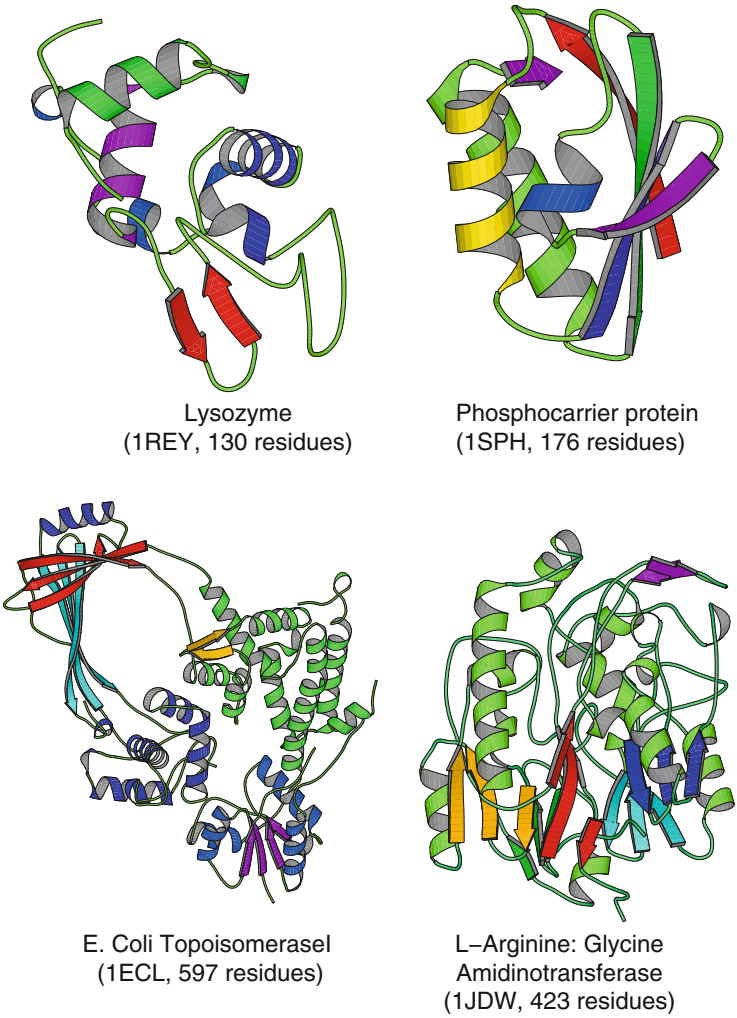


Figure 4.7. Examples of $\alpha + \beta$-proteins: **lysozyme**, **phosphocarrier protein**, **DNA topoisomerase I**, and **glycine amidinotransferase**.

Figure 4.11 illustrates the structure of the 180-chain **tomato bushy stunt virus** that infects many plants, including tomatoes and cherry trees. Interestingly, virus coats are assemblies of similar proteins rather than one huge protein or combinations of different proteins, because the relatively small amount of viral nucleic acids must encode this protein coat; at the same time, the nucleic acids must be covered completely. Hence a large protein shell consisting of repetitive motifs satisfies both of these criteria.

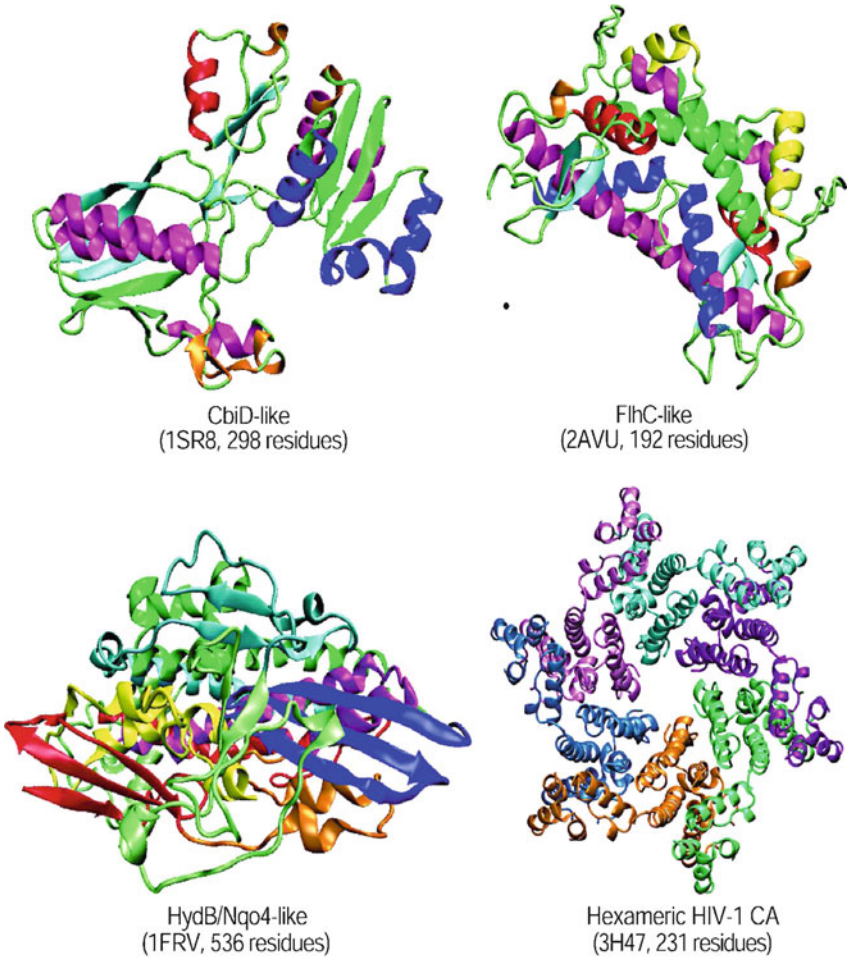


Figure 4.8. Examples of multidomain proteins: **CbiD-like** protein with two domains; **FlhC-like** protein with three domains; **HydB/Nqo4-like** with four domains; and **Hexameric HIV-1 CA** with two domains.

Among the larger molecular structures determined by X-ray crystallography at moderate resolution (i.e., approaching 3.5 Å) is the core particle of **bluetongue virus**, an agent of disease in both plants and mammals. Its transcriptionally active compartment measures 700 Å in diameter and is composed of two principal structural proteins that assemble in two layers, a core and a subcore, together encapsulating the RNA genome (10 segments of doubled-stranded RNA, ∼19,000 base pairs total). The crystal structure revealed how these approximately 1000 protein components self-assemble through a complex mixture of packing mechanisms involved in each of the two layers, using triangulation and geometrical quasi-equivalence packing motifs [484].

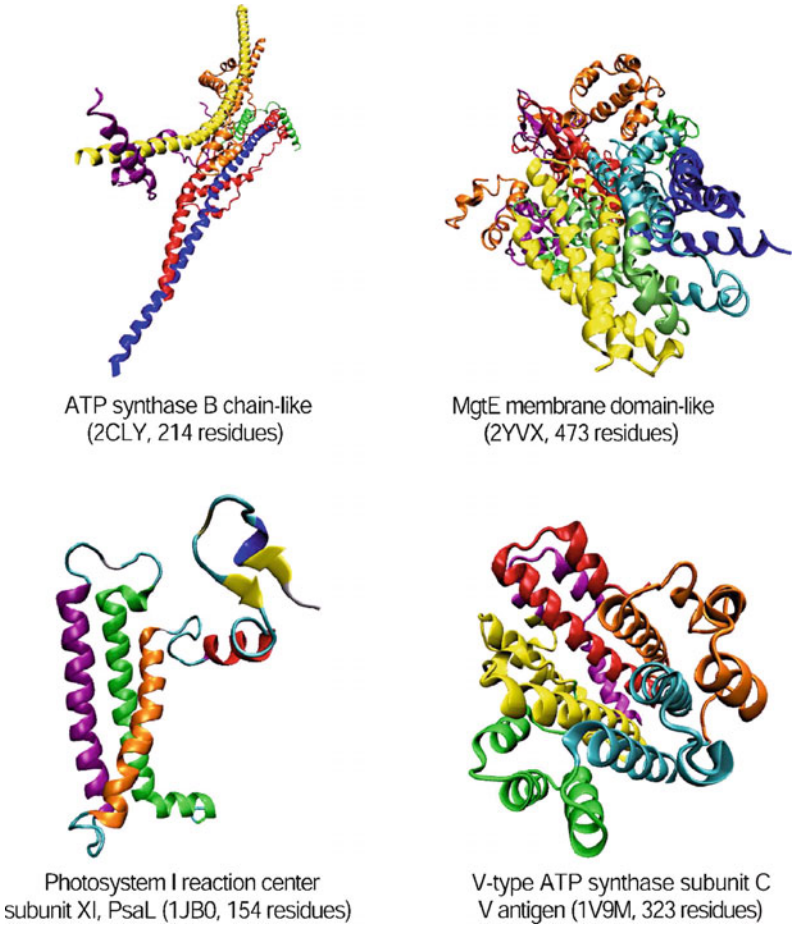


Figure 4.9. Examples of membrane and cell surface proteins and peptides: **ATP synthase B chain-like** protein, with a long helix; **MgtE membrane domain-like** protein, with five transmembrane helices; **Photosystem I reaction center subunit** protein, with three transmembrane helices; and **V-type ATP synthase subunit C** protein, with nine transmembrane helices.

4.10.2 From Ribosomes to Dynamic Networks

Other examples of quaternary structure are noted for the ribosome, muscle-fiber complexes, bacterial flagellar filaments, and photosynthetic assemblies of membrane proteins.

The *E. Coli* **ribosome** is a ribonucleoprotein complex with a diameter of about 200 Å constructed from 3 RNA molecules and 55 protein chains [419]. The Nobel Prize in Chemistry was awarded in 2009 to three scientists who independently obtained atomic-level crystallographic views of this magnificent

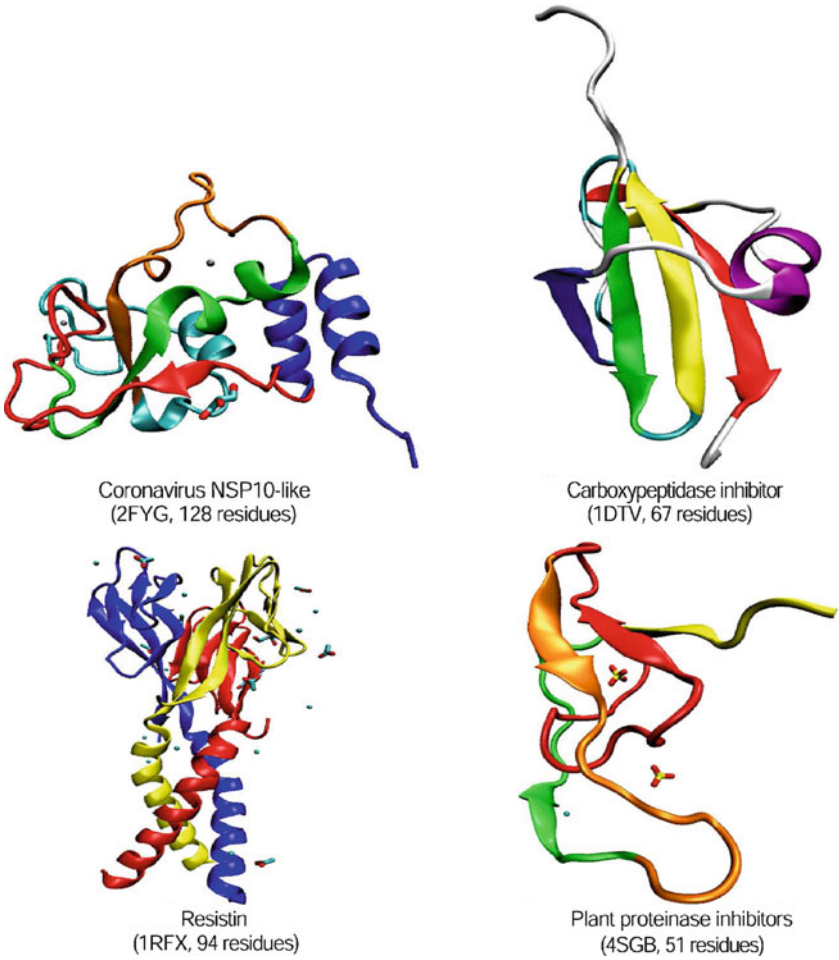


Figure 4.10. Examples of small proteins: **Coronavirus NSP10-like**, binds two zinc ion per subunit; **Carboxypeptidase inhibitor**, disulfide-rich, α+β; **Resistin**, disulfide-rich six-stranded β-sandwich; and **Plant proteinase inhibitor** complexed with calcium and SO_4.

RNA/protein machine: Ada Yonath, Venkatraman Ramakrishnan, and Thomas Steitz. For example, the Yonath lab solved the large ribosomal subunit from *Deinococcus radiodurans* [516] and the small ribosomal subunit from *Thermus thermophilus* [1135] (see Fig. 1.1). The Steitz lab reported the structure of the large ribosomal subunit from *Haloarcula marismortui* (2833 of the subunit's 3045 nucleotides and 27 of its 31 proteins) [85], and Ramakrishnan's group reported the structure of the small subunit of *T. thermophilus* [1379]. These eagerly awaited structures of the bacterial ribosome were aided by cryo-electron microscopy reconstructions — first reported in 1995 for the ribosome from *E. Coli* (see recent

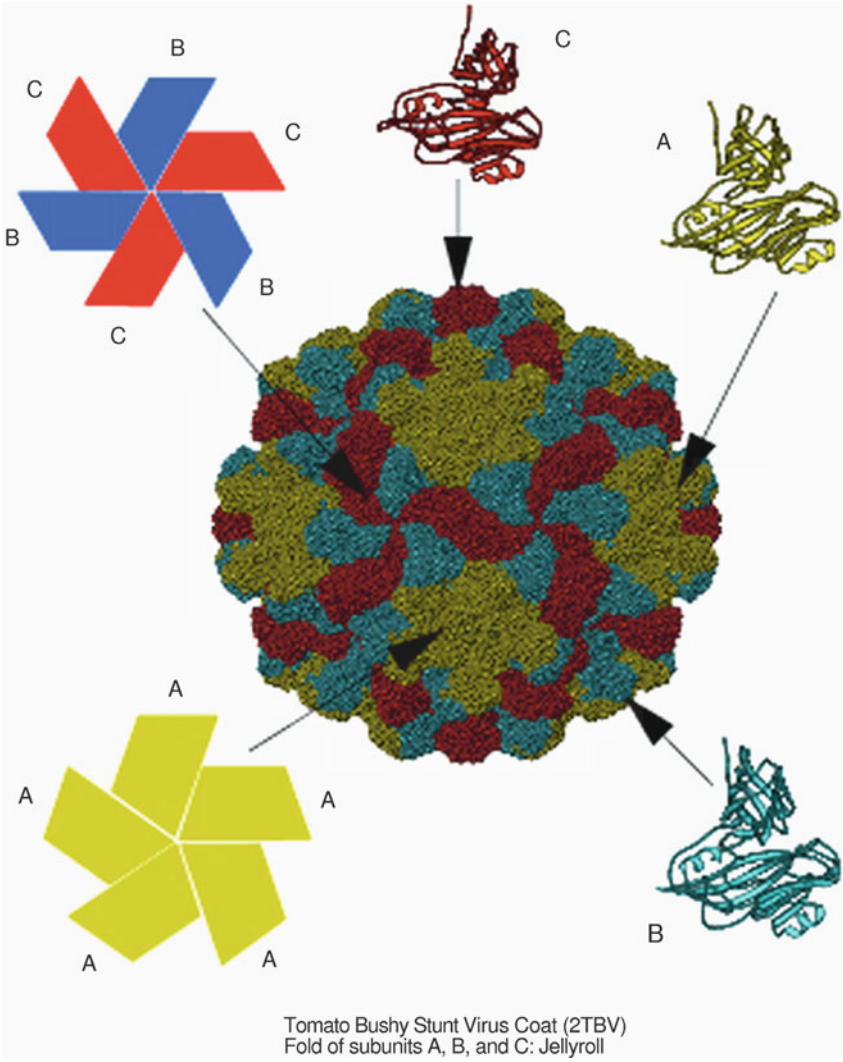


Figure 4.11. The structure of **tomato bushy stunt virus**, a spherical arrangement of 180 polypeptide chains, each of 387 amino acids, with every 3 chains making up an asymmetric unit (the subunits are colored blue, green, and red).

views in Figure 1.2 [710]) — which helped crystallographers estimate the initial phasing of their X-ray data (see [171] for a perspective). The combined structural characterizations of the ribosome provided clear evidence that the ribosome is a ribozyme — that is, that the ribosome RNA's component likely catalyzes peptide bond formation (see Chapter 7, RNA sections).

Muscle cells contain parallel myofibrils composed of two kinds of filaments, each with the following proteins: **myosin** (thick filament), and **actin**,

tropomyosin, and **troponin** (thin filament); around these filaments, **titin** — itself two extremely long proteins — plus nebulin form a flexible mesh. Muscle contraction is produced by the interaction of actin and myosin.

The bacterial flagellar motor of the protein **flagellin** [1085] represents another challenging motor complex solved recently. Filaments of flagellin are formed by an arrangement of stacked flagellin proteins ('protofilaments') lined up side by side; an arrangement like loosely rolled sheets of paper results. The remarkable cooperativity among the different filaments leads to conversions between a macroscopic left-handed form — used for swimming — and a right-handed form — used for reorientation of motion. The high-resolution flagellin crystal suggests how this possible structural switch (between left and right-handed supercoiled forms) might occur to direct function.

Insights into the solar energy converters in the membranes of bacteria and plants were provided by the crystal structure of photosystem I, a large photosynthetic assembly of membrane proteins and other cofactors from the thermophilic cyanobacterium *S. elogatus* [616]. The detailed atomic picture (at 2.5 Å resolution) of the network of 12 proteins subunits and 127 cofactors (chlorophylls, lipids, ions, waters, others) shows the beautiful coordination of all components for efficient absorption and conversion of solar energy into chemical energy.

4.11 Protein Structure Classification

Many groups worldwide are working on classifying known protein structures; see [47, 48, 952, 1259] for a perspective of protein structure and function evolution. Several classification schemes and associated software products exist.A popular program is SCOP: "Structural Classification of Proteins" [887]. (See scop.mrc-lmb.cam.ac.uk/scop/ or connect to SCOP through links available in many mirror sites such as PDB) [262]. These classifications are currently assigned manually, by visual inspection, but some automated tools are being used for assistance.

Also noteworthy is the PROSITE (www.expasy.ch/prosite/) database of protein families and domains intended to help researchers associate new sequences with known protein families. Other databases of patterns and sequences of protein families are PFAM and PRODOM; see [881] for a comprehensive list.

The SCOP levels (top-to-bottom) are: class, fold, superfamily, family, and domain. The sequence, or reference PDB structure, can be considered at the very bottom of this tree.

The top level of the SCOP hierarchy is the *class* (all-α, all-β, α/β, $\alpha+\beta$, multi-domain, membrane and cell-surface, and small proteins). Each *class* denotes common, global topologies of secondary structure.

Next comes the *fold*, which clusters proteins that have the same global structure, that is, similar packing and connectivity schemes for the secondary structural elements. Folds are often also called *supersecondary structure*. From 50 to several

hundred folds are currently known for each class, with the repertoire increasing steadily. An example mentioned above, the α/β barrel fold, groups **TIM** with other proteins like **RuBisCo(C)**, **Trp biosynthesis**, and **glycosyltransferase** into a *superfamily*, the next level of the classification hierarchy.

The *superfamily* groups proteins with low sequence identity but likely evolutionary similarity, as judged by similar overall folds and/or related functions. Members of the same superfamily are thus thought to evolve from a common ancestor. Another superfamily, for example, contains **actin**, the ATPase domain of the **heat shock protein**, and **hexokinase**. Superfamilies often pose the greatest challenge in the task of protein classification.

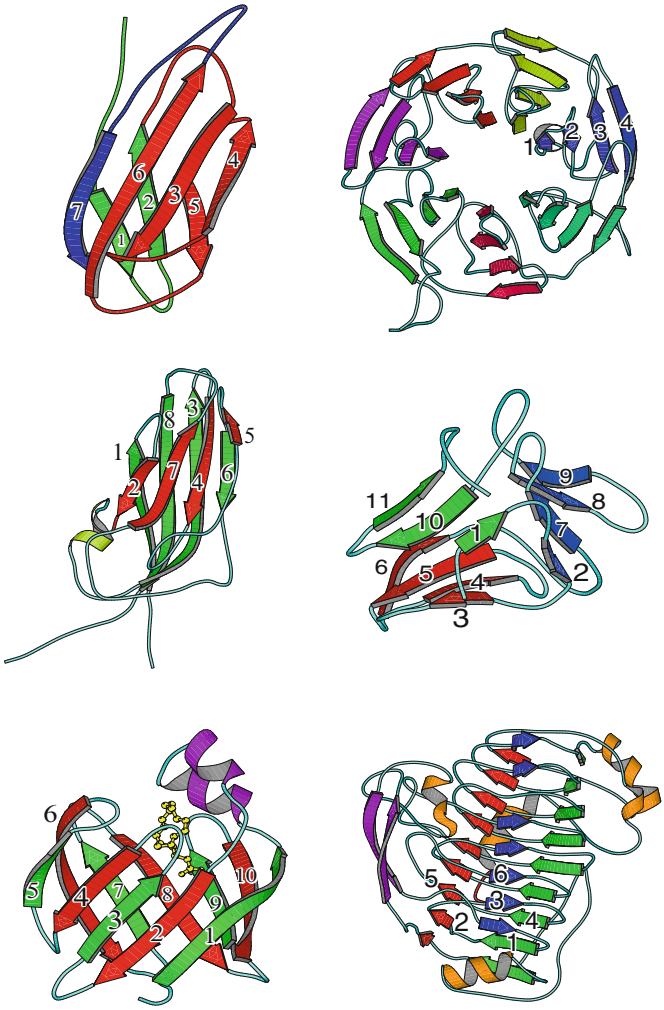
Superfamilies are further divided into *families*, which cluster proteins with substantial sequence, structure, and function similarity. Generally, this requirement implies a sequence identity of at least 30%, but there are instances of low sequence identity (e.g., 15%) but definitive structural and functional similarities, as in the case of globin proteins. For example, families of glycosyltransferase include β-galactosidases, β-glucanase, α-amylase, and β-amylase.

Finally, at the bottom of the tree of the SCOP classification lies the *domain* category, to distinguish further structurally-independent regions that may be found in larger proteins.

For updated information on the number of identified folds, superfamilies, and domains, check scop.mrc-lmb.cam.ac.uk/scop/count.html.

As our knowledge of protein structure increases, our classification schemes and software tools will evolve quickly. Automation of the classification is important for rapid structural analysis and ultimately for relating the sequence and structure to biological function.

The reader is encouraged to re-read at this point the sections in Chapter 2 on protein folding/misfolding (Sections 2.2 and 2.3).


6
4
7
2
3
5
1
1
2
3
4
8
3
1
5
2
7
4
6
11
9
8
10
1
7
6
5
4
2
3
6
5
4
7
8
9
10
3
2
1
5
6
3
2
4
1