PROTEINS

1. Introduction

Proteins, ubiquitous to all living systems, are biopolymers (qv) built up of various combinations of 20 different naturally occurring amino acids (qv). The number of proteins in an organism may be as small as half a dozen, as in the case of the simple bacterial virus M13, or as large as 100,000, estimated to be in the human system. Proteins are encoded by the deoxyribonucleic acid (DNA) that is present in all living cells.

Protein uses are myriad. The large number of biochemical reactions within the living cell are catalyzed by enzyme proteins Traditional food processes such as baking, brewing, and cheesemaking involve the action of enzymes found in different microorganisms. Other proteins are involved in the transport of electrons, ions, and small molecules. Proteins are also key components of the immune system and control the genetic expression of other proteins. The smallest proteins have a molecular weight of only $\sim$400; the largest protein molecule discovered to date is the muscle protein titin with a molecular weight of 3,000,000.

The study of proteins has a long history. First isolated from both animal and plant sources in the late 1700s, their chemical composition was studied in the 1820s and 1830s. Glycine was the first constituent amino acid to be isolated from gelatin in 1820. The name protein, meaning primary, was coined in 1839. The concept of the linear linkage of amino acids each having the same backbone structure to form the protein was established by Emil Fischer in 1901. Great advances have taken place in protein research since the 1950s through biochemistry and molecular biology.

2. Properties

The different combinations of amino acid side chains in a protein molecule give the protein its unique structure and function. Although the size and chemical properties of the amino acid side chains vary greatly, these building blocks can be broadly classified into three categories: hydrophobic, charged, and uncharged polar. A set of classical experiments in the 1960s showed that the structure (conformation) of ribonuclease (RNase) is determined by the amino acid sequence under normal physiological conditions (1). This has proved to be true for most proteins.

2.1. Purification.

The properties of a protein such as size, net electric charge, and solubility can be exploited in protein purification. Centrifugation is used to separate proteins based on density and shape. Electrophoresis exploits the net charge on a protein as well as its mass and shape. Electrophoretic separation is usually carried out by passing the mixture through a gel medium in an electric field. In chromatography (qv), separation of proteins is carried out by passing them through a column packed with some porous material. The mixture is separated according to size in size-exclusion chromatography, ionic charge in ion-exchange chromatography, or selective adsorption in affinity-labeled chromatography (2). An accurate method for detecting a single protein in a mixture of proteins is through Western Blotting where an antibody is bound to the protein of interest.

Sequencing. The method for sequencing proteins was developed in the 1950s by Sanger (3). A protein was first cleaved into smaller fragments and then hydrolyzed into the constituent amino acids. The individual amino acids were separated on the basis of their charge or hydrophobicity and quantified. The complete sequence was then generated from the fragments. Sanger, among others, was also responsible for developing methods for nucleotide sequencing: protein sequences are now generally inferred from their corresponding DNA sequences. As of May 2004, the genomes of 20 eukaryotes have been sequenced, in addition to about 100 microbes and 1000 viruses (4). Large databases (qv) exist for the sequences of proteins from different organisms (5).

3. Biosynthesis

The whole of molecular biology is based on the central dogma: DNA→ ribonucleic acid (RNA) → protein. Each amino acid residue is encoded by a triplet of nucleotides which form the genetic material of chromosomes. The DNA and, in some cases the RNA, is transcribed onto a strand of messenger-RNA (mRNA) in the nucleus of the cell.

The translation of the mRNA to the protein occurs on ribosomes, specialized organelles that are present in the cytoplasm of the cell. This process is complex, the main events occur in the following series of steps: (*1*) The ribosome binds to one end of the mRNA strand; (*2*) the first triplet of the gene coding for a polypeptide, which always corresponds to the amino acid methionine (Met), binds to its specific transfer-RNA (tRNA) molecule, activating methionine; (*3*) the chain is elongated when the previous amino acid comes off its tRNA and combines with the next amino acid to yield a peptide bond; and (*4*) elongation continues until a specific termination triplet is encountered, when the polypeptide chain dissociates from the last tRNA and the ribosome.

Initially it was believed that every protein is uniquely coded by a continuous chromosomal segment. However, it was found that in many organisms including humans, the primary RNA transcript often consists of coding segments called exons sandwiched between stretches of noncoding DNA called introns. The introns are later cleaved off and the exons joined by a process called RNA splicing. Different arrangements of some or all of the exons from a set can give rise to a large number of different mRNAs and consequently various gene products. An analysis of the human genome sequence data has yielded an estimate of ∼20,000−25,000 for the total number of genes, far lower than previous estimates of 80,000; the difference can now be attributed to the existence of these splice variants. The number of proteins, however, can be much larger. The protein variants generated by alternative splicing are generally expressed in different cell types or at different stages of development. The term proteome commonly refers to the collection of proteins expressed by a particular cell or tissue type and the field of proteomics deals with the study of the expression, activity, and interaction of the proteins present in a cell type or metabolic pathway.

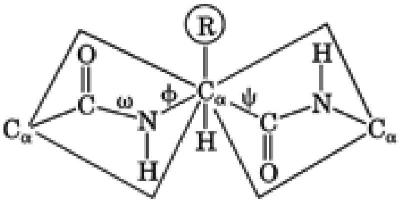
3.1. Post-Translational Modifications.

Proteins are often synthesized having 15–26 residue long signal peptides, usually at the amino terminus. The signal peptide serves to translocate the protein across a membrane when it is required in a location other than the cytoplasm. Once the protein has arrived, the signal peptide is cleaved. Certain enzymes and hormones (qv) undergo further cleavage to attain a specific functional form. Many viruses synthesize polyproteins that are then cleaved into individual polypeptide chains.

Proteins also undergo other covalent modifications after synthesis. These include (*1*) acetylation of the amino-terminal and amidation of the carboxy-terminal residue; (*2*) addition of fatty acid groups (myristoyl) or lipids, eg, farnesyl, for anchoring proteins to membranes; (*3*) glycosylation, one of the most common modifications, which serves many purposes such as cell–cell recognition through interaction with other molecules on the surface of cells; (*4*) phosphorylation, resulting in the addition of a negative charge, which serves as a switch to control the affinity for activators and inhibitors particularly in signal transduction; and (*5*) disulfide bond formation, in which cysteine (Cys) residues separated in sequence are brought spatially close together through covalent linkage.

4. Principles of Protein Structure

To understand the function of a protein at the molecular level, it is important to know its three-dimensional (3D) structure. The diversity in protein structure, as in many other macromolecules, results from the flexibility of rotation about single bonds between atoms. Each peptide unit is planar, ie, $\omega = 180^\circ$, and has two rotational degrees of freedom, specified by the torsion angles ϕ and ψ, along the polypeptide backbone. The number of torsion angles associated with the side chains, R, varies from residue to residue. The allowed conformations of a protein are those that avoid atomic collisions between nonbonded atoms. The allowed combinations of the torsion angles ϕ and Ψ were worked out in the 1960s and are referred to as the Ramachandran Plot after its inventor (5).



For any given protein, the number of possible conformations that it could adopt is astronomical. Yet each protein folds into a unique structure totally determined by its sequence. The basic assumption is that the protein is at a free energy minimum; however, calometric studies have shown that a native protein is more stable than its unfolded state by only 20–80 kJ/mol (5–20 kcal/mol) (6). This small difference can be accounted for by the favorable forces of the unfolded state, ie, conformational entropy owing to the large ensemble of conformers, and hydration owing to the favorable interaction of the polar atoms with the aqueous environment.

One of the principal driving forces determining the folded structure of a protein is the maintenance of peptide backbone hydrogen bonds on the removal of these bonds from the solvent to the protein interior. This is a force complementary to the hydrophobic effect that forces the nonpolar amino acid side chains away from the solvent into the interior. It is therefore natural to ask what types of structures allow for the maintenance of these hydrogen bonds. In the structure shown, it is clear that on opposite edges of the peptide plane there is a hydrogen bond donor (N−H) and a hydrogen bond acceptor (O). Therefore, if any set of peptide chains were stretched out side by side, a set of parallel hydrogen bonds between the chains could form. Another type of regular structure that can form is when the peptide chain twists into a long helix such that as one goes up the helix, one edge of the peptide plane always points upward (carboxy terminus) and the other points in the downward direction (amino terminus). The two basic structures of this kind, called the secondary structure, are the β-sheet and α-helix, respectively, shown in Figure 1.

Building on the two secondary structures and using the fact that amino acids can, to a first approximation, be classified as either hydrophilic (polar) or hydrophobic (nonpolar), allows for a simple understanding of the basic architecture of most proteins. Protein folding seeks to maximize the number of backbone hydrogen bonds while minimizing the number of exposed nonpolar amino acid side chains. Making the assumption that by starting with only helices and sheets the first part of the score has been optimized, only the number of solvent-exposed nonpolar residues has to be minimized.

There are basically three kinds of helices: hydrophobic, hydrophilic, and amphipathic. In the last, the two faces have opposite charge distributions. Sheets can be either hydrophobic or amphipathic. These structural units are strung together by a set of small connecting loops. The number of ways that these units can pack against each other, subject to the constraints of charge complementarity and exposure to solvent, gives rise to the various topologies of protein structures, referred to as its tertiary structure.

Depending on the predominance of the type of secondary structure, proteins can fall into any of three classes. Details are available in Ref. (7).

4.1. All-α-Proteins. The globins are well-known examples of all- α-proteins, the first protein structure to be solved. Globins consist of eight helices packed together around a heme group to form a box-like structure. The helices are so arranged that the ridges formed by the side chains on one helix fit into the grooves between the side chains of another helix. This packing arrangement, first predicted by Crick, has been proved to be true in the general packing of helices in proteins. The simplest motif of this class is the four-helix bundle found in the oxygen-carrying proteins hemerythrin and cytochrome *b*562. In these structures, hydrophobic side chains from all four helices point to the center of the bundle, forming a hydrophobic core. α-keratin, the main component of hair, consists of a number of long helices coiled around one another.

Most of the membrane-spanning regions of integral membrane proteins appear to be all-helical. Prototypical of such proteins is bacteriorhodopsin, which consists of seven transmembrane helices arranged in a circle, approximately perpendicular to the membrane. Owing to the hydrophobic nature of the membrane, the helices consist largely of hydrophobic residues except where they bind to the retinal molecule. The loops connecting the helices are more hydrophilic as these are in contact either with the polar head groups of the lipids or the aqueous environment. The complex photosynthetic reaction center from bacteria is made up of four subunits, three of which consist of membrane-spanning helices.

4.2. All-β-Proteins. The general topology of all-β-proteins as a structural class of proteins consists of two or more β-sheets packed against each other. The mode of packing depends on the nature of the faces of the sheets. In the immunoglobulin fold, found in recognition molecules of the immune system, two sheets pack against each other to form a β-sandwich having a hydrophobic core and polar surface. Structures of proteins of unrelated functions have revealed an interesting motif described as a four-blade β-propeller (8).

4.3. α/β-Proteins. The most frequent domain structures observed in proteins are those having a mixture of helices and sheets. One of the largest of all domain structures is the eight-stranded α/β-barrel structure which consists of a core of eight parallel β-strands surrounded by eight helices, forming a very symmetrical structure. A large number of proteins have the α/β-sandwich structures in which a β-sheet is packed against either a single row of helices or is sandwiched between two such rows.

Larger proteins usually have two or more structural units termed domains, each domain having structures similar to single-domain proteins. The interaction between individual domains is much less extensive than that within a single domain. In many cases each domain is responsible for carrying out a specific function.

In some proteins, such as hemoglobin, separate polypeptide chains must associate for the chains to be functional. This forms a quaternary structure.

4.4. Cofactors. Frequently, proteins exist in their native state in association with other nonprotein molecules or cofactors, which are crucial to their function. These may be simple metal ions, such as Fe^{n+} in hemerythrin or Ca^{2+} in calmodulin; a heme group, as for the globins; nucleotides, as for dehydrogenases, etc.

Protein Folding and Misfolding. Either during or soon after their biosynthesis, proteins assume their unique native conformation, which almost always corresponds to the thermodynamically most stable state under physiological conditions. It is possible that *in vivo* this process, known as protein folding, follows the same sequence as its synthesis, ie, from the amino to the carboxy terminus, which can usually be reproduced *in vitro* under suitable conditions, such as pH and temperature. It is difficult to imagine the protein searching through all possible conformational states during this process, since it would take an astronomically large amount of time. Instead, there is evidence to suggest that certain segments fold first and serve as seeds around which the rest of the protein folds. A number of proteins are known to pass through a transient intermediate state, the so-called molten globule state (9). The precise structural features of this state are not known, but appear to be compact, and contain most of the regular structure of the folded protein, yet have a large side-chain disorder (10). Recent studies indicate that a protein does not necessarily have to pass through a fixed set of intermediate states; but rather it samples a number of possible conformations accessible to a polypeptide chain. The intermediate states that a protein adopts possess energies that lie along the energy 'landscape' that serves to funnel it from its denatured conformation to its native state (11).

Many proteins frequently require the assistance of other protein molecules called molecular chaperones, for assuming the final tertiary structure *in vivo*. The best characterized is the bacterial complex involving GroEL and GroES, which provides a sequestered cavity that allows incompletely folded proteins to undergo the final stage of folding (12). Recent experiments suggest that a large fraction of freshly synthesized polypeptide chains still ends up misfolded. Cellular mechanisms are usually present to target these molecules for degradation. When they fail, the misfolded polypeptides aggregate and give rise to disorders such as cystic fibrosis, some cancers, and Alzheimer's and Parkinson's disease (13). The characteristic structure of these aggregates is the amyloid fiber that consists of cross-β sheets whose strands run perpendicular to the fibril axis based on X-ray fiber diffraction and birefringence data.

5. Protein Function

Proteins can be broadly classified into fibrous and globular. Many fibrous proteins serve a structural role (14). α-Keratin has been described. Fibroin, the primary protein in silk, has β-sheets packed one on top of another. Collagen, found in connective tissue, has a triple-helical structure. Other fibrous proteins have a motile function. Skeletal muscle fibers are made up of thick filaments consisting of the protein myosin, and thin filaments consisting of actin, troponin, and tropomyosin. Muscle contraction is achieved when these filaments slide past each other. Microtubules and flagellin are proteins responsible for the motion of cilia and bacterial flagella.

Globular proteins have biological function that they carry out by direct interaction with ligands that may be small atoms or large macromolecules. The protein without the bound ligand is known as the apo form, whereas the bound state is called the holo form. Many proteins are capable of binding more than one ligand, but usually do so on separate domains. The binding sites may be located in the protein interior as in the case of cofactors such as the hemes. Larger ligands most often bind at the surface including in the interdomain region. There is steric and physicochemical complementarity between the interacting surfaces resulting in close packing and favorable polar interactions. The structure of a protein domain generally does not alter significantly on binding. An exception is calmodulin, in which a large hinge movement has been observed between its two lobes on binding to a peptide analogue of myosin kinase (15). Sometimes the binding of a ligand at a particular site of a protein can have an effect on the affinity of binding of another ligand at a second site. The related site may be on a different domain of a multidomain protein or a different subunit of an oligomeric protein. This is termed allostery. The allosteric effect can be explained in terms of a conformational change at the primary site inducing a change at the secondary site. Allosteric control is frequently observed in gene regulation by a feedback mechanism. The binding of a repressor or activator to the DNA sites controlling the synthesis of a protein is modulated by its binding to the synthesized protein itself.

Binding sites comprise only a small fraction of the structure of most proteins. Most of the rest of the protein is responsible for providing the framework of the protein and in the process offering it stability. Consequently, functionally related proteins have the same overall topology, yet differ in the residues that form their binding sites, providing them their substrate specificity. During the folding process, functional residues distant in sequence are brought spatially close together. The chemical methods for determining these residues usually involve the covalent modification of the protein. The reactivity of the various groups to different reagents in the presence and absence of the ligand gives an indication of the binding of the group to its ligand. This assumes that the structure of the protein is unaffected by the modification. A more reliable method is affinity labeling, where a reactive group is incorporated into the ligand and its association with the group is monitored.

More detailed information about ligand binding has come from the 3D structures of protein–ligand complexes. In a number of cases, where the reaction is too fast to be followed using these methods, the substrate is replaced by an analogue or inhibitor where the mode of binding is presumably similar. The structures of proteins have shown that in a number of instances the same ligand, typified by heme, can be bound to different proteins having no structural similarity.

5.1. Oxygen Binding. The earliest investigations on proteins were carried out on the globins that bind oxygen molecules through the heme group. Its porphyrin ring packs against hydrophobic residues in the interior of the protein. The iron atom of the heme binds to O_2 on one side, and on the other to a histidine (His) residue that is one of only two residues conserved among all globins. This complex is stabilized by the interaction of O_2 with another His. Carbon monoxide, CO, which binds to isolated heme much more strongly than does O_2, has a much lower affinity for the globins owing to steric hindrance of the His residue. Theoretical studies have suggested that fluctuations in the protein structure produce channels for O_2 to diffuse in and out of its interior.

The allosteric effect is seen in hemoglobin, which can exist in two quaternary structural states: oxygenated (R) or deoxygenated (T). The binding of one O_2 or some other effector to one of the subunits stabilizes the R form as compared to the T form. Binding of a second and third O_2 stabilizes it even further.

5.2. Enzymes. Enzyme catalysis has been studied since the late 1800s and a great deal is known about this subject (16). Enzymes increase the rate of equilibrium of a chemical reaction by decreasing the barrier between the free energies of the products and reactants. This is believed to be achieved by the tighter binding of the enzyme to the transition state than to the substrate or product. An enzyme can function at very low concentrations by catalyzing the same reaction numerous times. The course of an enzymatic reaction can be studied using steady-state kinetics, where the dependence of the velocity of a reaction on substrate concentration is followed. An alternative method is relaxation spectrometry, in which the reaction is allowed to come to equilibrium, after which a rapid shift is made to one of the thermodynamic variables. The time course of the relaxation to the new equilibrium state can be followed using a variety of techniques.

A mechanism for enzyme action, the lock-and-key theory, was proposed by Emil Fischer in the 1890s, in which it was suggested that the complementary shapes of the enzyme and substrate help each to recognize the other. An alternative theory is that of induced fit, in which the substrate, on approaching the binding site of the enzyme, induces a conformational change making the reaction favorable. This was first observed in hexokinase where phosphorylation of glucose by adenosine triphosphate (ATP) was accomplished by a drastic rotation of the two domains of its structure. The crystal structures of a number of enzymes complexed with ligands have greatly aided the understanding of their mechanisms of action.

5.3. Molecules of the Immune System. One of the amazing capabilities of vertebrates is the capacity to recognize and bind foreign molecules called antigens for removal from the system. This is carried out by antibodies or immunoglobulin molecules, remarkable proteins in that having a limited gene pool, ie, $\sim$1000 segments in humans, the body is able to generate millions of such molecules. Each immunoglobulin is a Y-shaped molecule composed of two heavy and two light chains. Each of the chains is further composed of a constant and variable domain. The variable segments have hypervariable regions or complementarity determining regions (CDRs), which provide specificity to the antibody. The large repertoire of the antibody population is produced by the different rearrangements of the various available segments.

Many antibody structures in the presence and absence of antigens are known. All the domains have basically the same structure, termed the immunoglobulin fold, which consists of a β-sandwich held together by a disulfide bridge, having the CDR loops at the top of the sandwich. The variable loops from the different segments come together to form antigen-binding sites of vastly different shapes and sizes.

The major histocompatibility complex (MHC) molecules are responsible for binding foreign antigens and presenting these to T-cell receptor (TCR) molecules, which trigger a series of events leading to eventual destruction of the foreign invader. The MHC molecules accomplish this by possessing domains of immunoglobulin folds that form the antigen-binding site and other domains that are responsible for binding to TCR. The dual function allows a distinction between self-proteins and foreign invaders.

5.4. DNA-Binding Proteins. The process of differentiation, by which a single cell multiplies to form different types of cells at different generations, is accomplished by the turning on and off of genes at various stages. The expression and regulation of the genetic information are carried out by proteins that bind to segments of DNA that surround the gene coding for the protein to be synthesized. These can either act as repressors or activators, depending on whether they prevent or help the RNA polymerase from binding to the DNA and initiate its transcription. Both these proteins have the same helix–turn–helix motif in symmetrically related dimers that fit neatly into successive turns of the so-called major groove of the DNA double helix. The DNA undergoes a distortion on binding that is induced by nonspecific protein–DNA interactions involving the backbone atoms. The actual affinity, however, is dependent on the particular sequence of DNA.

Another class of DNA-binding proteins are the polymerases. These have a nonspecific interaction with DNA because the same protein acts on all DNA sequences. DNA polymerase performs the dual function of DNA replication, in which nucleotides are added to a growing strand of DNA, and acts as a nuclease to remove mismatched nucleotides. The domain that performs the nuclease activity has an α/β-structure, a deep cleft that can accommodate double-stranded DNA, and a positively charged surface complementary to the phosphate groups of DNA. The smaller domain contains the exonuclease active site at a smaller cleft on the surface that can accommodate a single nucleotide.

Another common DNA-binding motif is the zinc finger that is found in many gene-regulating proteins. It consists of a helix β-hairpin motif, stabilized by a chelated Zn^{2+} ion. Multiple zinc fingers can occur in tandem, making contact with base pairs along the major groove of DNA.

A number of repressor proteins, in particular bacterial MetJ and Arc repressors, have a ribbon–helix–helix motif, the β-strands being involved both in dimer formation and in interactions with DNA bases in the operator regions. Many transcription factors bind to DNA as dimers held together by the so-called leucine (Leu) zipper region, consisting of a pair of coiled-coil α-helices, where Leu occurs every two turns of the helix. The neighboring basic region appears to make a transition from random coil to helix on binding to DNA. In the case of another regulatory segment of DNA called the TATA-box, significant unwinding of the DNA occurs on binding to TATA-binding protein (TBP) (Fig. 2) (17).

5.5. Signal Transduction Proteins. Biological systems have the ability to respond to changes in external concentration of hormones (qv), growth factors, or other molecules or stimuli. These ligands can bind to specific transmembrane receptors called G-protein coupled receptors (GPCR), initiating a cascade of biochemical processes that produce an intracellular signal. For example, when light falls on the receptor protein rhodopsin, it transmits the signal to multiple copies of a transducing protein termed G-protein, which in turn, interacts with effector proteins that further amplify the signal. It is estimated that GPCRs are the targets of about a third of the current 100 top-selling drugs.

Protein kinases are another class of proteins that play an important role in signal transduction by phosphorylating other proteins resulting in their activation. Receptor tyrosine kinases have an additional extracellular domain that binds growth factors and leads to the activation of their kinase domain. This initiates a cascade of interactions that ultimately lead to gene activation, cell division, and increase in blood supply (angiogenesis). A number of diseases including inflammation, cancer, and diabetes are a result of malfunction of these kinases. In recent years, a great deal of research is being done to develop kinase inhibitors that target cancer cells by arresting their growth either directly or indirectly by restricting the blood supply to the cells.

6. Systems Biology

Until recently, the focus of biological research had been on the structure and function of individual proteins. Biologists are now beginning to look into "systems biology" (18), which seeks to understand, both in time and space, the complex interactions of molecular components within the living cell, such as enzymatic, signaling, and gene-regulatory pathways. Much effort is being devoted to applying theoretical methods such as network theory (19) to understand the interaction of the components. The development of the novel technique of microarrays has made it possible to study the expression of thousands of genes under different conditions.

7. Homology

Proteins that carry out the same or similar functions in different organisms generally have very similar structures. Such proteins also are often encoded by similar sequences of amino acids. These similarities are a result of the shared evolutionary history of the organisms and of their genes. Such similar proteins are termed homologous, and are believed to have a common ancestor. There is no direct way of determining whether two proteins are truly homologous. However, when the sequences of any group of proteins are recognizably similar it seems much more likely that they are homologues, having a common ancestor, than that they independently evolved to look the same. There are also cases where the last common ancestor appears to have been so long ago that the sequences are no longer recognizably similar, yet they are believed to be homologous.

In cases where the common function is well understood, such as the binding and transport of oxygen by the globins, identifying homology between proteins is relatively straightforward. In other cases, it must be inferred from the degree of structural and sequence similarity of two proteins. In the absence of any other information, only the degree of similarity between proteins at the sequence level is available to infer probable homology. Whereas two proteins having 30% or more of their amino acids identical over the whole sequence certainly implies homology, the converse is not necessarily true. For example, in the most distantly related globins, human alpha hemoglobin and bacterial leghemoglobin, there are only two amino acids that have not been modified out of >100 amino acids, since the time of the last common ancestor. Yet the 3D structures of these proteins are very similar and the two conserved amino acids are the very two which play the most important functional role.

Comparison of sequences is done by finding an alignment, ie, a position by position correspondence, which brings the maximum number of identical or similar amino acids into correspondence. This can be done mathematically (20). Favorable scores are assigned for the pairing of identical or similar residues in the alignment; penalties are given for positions where insertions or deletions are required to bring other positions into favorable alignment. The scores are based on the physicochemical similarity of the amino acids. In cases of very distantly related proteins, such alignments are often ambiguous. In those cases, attempts are made only to identify those regions along the sequence, or even only those positions that form the pattern of conserved amino acids, which are essential to the encoded function and/or structure. Sequence similarity alignments and functional diagnostic patterns form two of the most important tools of modern molecular biology and protein chemistry.

An initial comparison of the human genome with those of the rodents, mouse, and rat, has led to an estimate that ~90% of genes are conserved among the three species (21). It was also not entirely surprising to find that almost all the human genes associated with disease have homologues in the rat genome, which appears to explain the success in using the latter as a laboratory tool and model in drug development. There is no doubt that as the genomes of more organisms are sequenced, comparative genomics will increasingly prove useful in understanding the normal function of genes and in their malfunction in the cause of human diseases.

A comparison of proteins of similar function has shown that the 3D structure of proteins is more conserved than sequence. The function of a protein is assured by providing a particular shaped surface that fits the other molecular components with which the protein interacts. However, the core or basic internal structure of two proteins may be quite similar in terms of α-helices and β-sheets, yet the surfaces quite different. This is because the core structure acts as a stable scaffolding onto which loops of varying size and amino acid chemistry can be attached. Thus structural similarity of the cores of proteins does not necessarily imply homology. By one estimate, although there may be as many as 23,000 protein families in Nature (as determined by similarity in sequence or function), the number of distinct folds may only be one- tenth of that number. Most of these folds belong to one of the nine major superfolds shown in Figure 3 (23). Thus again, knowledge of existing structures, like sequences, allows inferences about new structures. Structures are cataloged in various protein data banks (22,24).

8. Structure Determination

The most common methods used to study the structure of proteins are as follows:

Technique	Information
electron microscopy	low resolution 3D structure
X-ray and neutron diffraction	high resolution 3D structure
electron diffraction	medium resolution 3D structure
nmr[a]	high resolution 3D structure
cd/ord[a]	secondary structure
infrared spectroscopy	secondary structure

[a] Nuclear Magnetic Resonance = nmr; cd = circular dichroism; ord = optical rotary despersion.

A detailed account is given in Ref. (25). The techniques giving the most detailed 3D structural information are X-ray and neutron diffraction, electron diffraction and microscopy (qv), and nuclear nmr spactroscopy.

8.1. Diffraction. When a beam of X-rays, neutrons, or electrons strikes a specimen, the beam is diffracted by the component atoms. If the atoms are arranged in a regular array as in a crystal, the diffracted rays interfere constructively in certain directions giving rise to a pattern. The rays can be recorded using photodetectors (qv) or on photographic film. The structure of the molecule that gave rise to the pattern can then be determined by calculating the Fourier transform of the diffraction spots. The intensities of the diffraction spots can be measured directly and different methods have been developed to determine the relative phases. X-ray crystallography is the most widely used diffraction method, yielding structural information to atomic detail. The most difficult part of this technique is in the crystallization of the protein, a process that is poorly understood. When the molecules are arranged in a fibrous form as in collagen, muscle, or some viruses, X-ray fiber diffraction is used. Neutron diffraction may be employed to locate the positions of the hydrogen atoms. Structures determined by electron diffraction are limited in resolution, but have the advantage that the phases can be derived directly from the images concurrently obtained by electron microscopy.

8.2. Nuclear Magnetic Resonance. This spectroscopy has developed into a powerful technique for protein structure determination. Spectra are obtained by placing a sample in a magnetic field and applying a radio frequency (rf) pulse. This pulse perturbs the equilibrium nuclear magnetization of atoms having nuclei of nonzero spin, such as ^{1}H and ^{13}C. The signals emitted as the system returns to equilibrium can be converted to a set of resonances at different frequencies by Fourier transform. The frequencies, called chemical shifts, are indicative of the type of nucleus as well as its chemical environment. Thus nmr can yield structural information about the protein. Both two-dimensional (2D) and 3D nmr have been developed to yield distances between atoms and thus obtain the atomic positions of a set of possible protein structures. One of the principal difficulties is in resolving the spectral peaks. The superiority of nmr over X-ray crystallography lies in its ability to determine the structure of a protein in the solution state, the natural state of the living cell.

9. Structure Prediction

Although the techniques described have resulted in the determination of many protein structures, the number is only a small fraction of the available protein sequences. Theoretical methods aimed at predicting the 3D structure of a protein from its sequence therefore form a very active area of research. This is important both to understanding proteins and to the practical applications in biotechnology and the pharmaceutical industries.

The most obvious approach for predicting the folded structure of a protein would be to search for its lowest energy conformation. In principle, knowledge of quantum chemistry should allow the necessary calculations to be carried out. However, the sheer size of the problem involving hundreds of thousands of interatomic interactions makes this extremely difficult. Simpler approaches, which fall into two basic categories, are used. The first is based on simplifying the energy calculations and conformational search, which usually involves reducing all the atomic level forces to simple classical mechanical forces. The second approach is not to attempt directly to predict the structure, but instead to use existing knowledge of protein structures to propose models most compatible with a given sequence.

The energy, E, of a protein can be expressed as the sum of different components:

$$E_{\mathrm{tot}} = E_{\mathrm{bl}} + E_{\mathrm{ba}} + E_{\mathrm{ta}} + E_{\mathrm{vdw}} + E_{\mathrm{es}} + E_{\mathrm{hb}}$$

where the first three terms represent the energy associated with the deviation of bond lengths (bl), bond angles (ba), and torsion angles (ta) from equilibrium; E_{vdw} and E_{es} are the van der Waals and electrostatic interactions between nonbonded atoms, respectively; and E_{hb} is the hydrogen bond energy.

Because of the marginal stability of the folded conformation over the unfolded state, results are crucially dependent on the accuracy of these potentials. Much effort has been devoted to the development of force fields (27). The procedures commonly used to minimize the energy functions are described in Ref. (28). The values for the various parameters were either experimentally derived or theoretically estimated from values for small molecules obtained in a manner that cannot readily be used for macromolecules owing to effects such as solvation, which are difficult to estimate. Thus, this approach has met with limited success in the prediction of structures. However, it has proved to be useful in obtaining good structures in cases where approximate models are already available, such as by X-ray diffraction, nmr, or molecular modeling (qv). This method has also been successfully used in studies on substrate binding and the effect of amino acid mutations on protein stability.

Attempts have also been made at predicting the secondary structure of proteins from the propensities for residues to occur in the α-helix or the β-sheet (29). However, the assignment of secondary structure for a residue only has an average accuracy of ~60%. A better success rate (70%) is achieved when multiple-aligned sequences having high sequence similarity are available.

The conformation an amino acid adopts depends on the residues in its neighborhood. This was made clear from a study of identical pentapeptides in unrelated proteins which were observed to adopt the same conformation only 20% of the time (30). The effect that residues neighboring in sequence or in space have on the conformation is not well understood.

Even when the secondary structure of a protein is known, there are a large number of ways that this structure can be packed together. Studies dealing with the identification of the topological constraints in the packing of helices and sheets have revealed certain patterns, but as of this writing accurate prediction is not possible.

9.1. Comparative Modeling. Given the limited success of prediction schemes, much effort has been turned to comparative modeling, ie, for a given sequence, identifying an approximately correct fold from among the different possible folds observed in nature. This technique is comparatively easy when homologous proteins having a known structure can be identified, but if no such protein is available it becomes a daunting task.

A review of the methods used in homology modeling is available (31). The various steps are as follows: (*1*) align the new sequence with its homologues of known structure imposing the constraint of no deletion/insertion in helical or strand segments; (*2*) model the backbone atoms of the conserved regions, ie, buried helices and strands and functionally important residues, from the mean positions in the homologues; (*3*) model the turns and loops connecting helices and strands, made difficult owing to conformational flexibility (approaches include database search of all known structures or, from energetic considerations, using the so-called loop closure method (32); (*4*) model the side-chain atoms based on rotamer libraries obtained from observed structures; and (*5*) refine the model by energy minimization.

9.2. Nonhomologous Extension Modeling.

In the case of protein sequences that have no clear sequence homologues with known structure, the problem becomes more challenging. One approach is to identify only the tertiary structural class to which a protein sequence belongs rather than its detailed structure. In some cases, this is done from an analysis of the difference in properties of amino acids in the various classes (33). More recently a method has been developed which, instead of estimating the likelihood of a single or short segment of amino acids adopting a particular folded structure, estimates it for the entire sequence (34). This holds promise because any relevant structural information about the amino acids can be included in the calculations to increase the accuracy of prediction.

A related approach seeks to determine the compatibility of a given sequence by threading it through a structure. To each residue position in the structure an environment is associated which may be characterized by its secondary structure, level of exposure to solvent, nature of preferred amino acids in its neighborhood, etc. A score is assigned to each amino acid in these positions. All possible alignments of the sequence having the structure are made and the alignment having the optimal total score is taken as the most compatible structure (Fig. 4). Several research groups are engaged in threading studies using different score functions. Only limited success has been achieved.

Simplified models for proteins are being used to predict their structure and the folding process. One is the lattice model, where proteins are represented as self-avoiding flexible chains on lattices, and the lattice sites are occupied by the different residues (35). When only hydrophobic interactions are considered and the residues are either hydrophobic or hydrophilic, simulations have shown that, as in proteins, the structures with optimum energy are compact and few in number. An additional component, hydrogen bonding, has to be invoked to obtain structures similar to the secondary structures observed in Nature. However, not much progress has been made in recent years as it has been realized that this is indeed an oversimplification of the folding process.

10. Practical Applications

Study of the structure and function of proteins is possible because of the revolution in molecular biology, which has enabled the cloning of genes and the expression of their products in large quantities. Great strides have also been made in the area of peptide synthesis, improvements in experimental techniques for structure determination, and computational speed. Areas of molecular and structural biology are rapidly expanding (see PROTEIN ENGINEERING).

10.1. Development of Drugs.

The vast majority of drugs available were derived from large-scale screenings. In many cases little is known about the target protein or other macromolecule, or the mode of action. The type of molecule for use as a drug is inferred from the biochemical mechanism of a disease. This information is then used to screen databanks containing thousands of compounds. Modifications to the leading candidates are then made and tested further for efficacy. It can take from 6 to 12 years for a drug to come to the market (36).

Rational drug design seeks to decrease this time lag and the resulting cost of development by rejecting unsuitable compounds. The design of drugs from first principles is not possible, but progress has been made. The first step is to obtain the quantitative structure–activity relationships (QSAR) between the biological activity of the protein and its chemical properties, particularly in its active site. The leading candidates are then examined further. The task is made simpler if the structural information of the protein is known, particularly in its complexed state. The increasing computing power and the resulting advance in the development of molecular modeling techniques have made the dream of *de novo* design of novel molecules a distinct possibility (37). Good reviews on molecular modeling (qv) techniques are available (38).

The success of the Human Genome Project has opened up a number of avenues in medicine. Genomic data is being mined to identify novel protein targets against which new medications can be developed (39). The international Hap-Map project seeks to understand common patterns of genetic variation from the large database of single nucleotide polymorphisms (SNP) present in the human population. The field of pharmacogenomics, still in its infancy, seeks to use this information to improve therapeutic efficacy and reduce drug toxicity with the ultimate target of developing drugs that are tailor-made for each individual.

10.2. New and Improved Biomaterials

Protein Computers. The membrane protein bacteriorhodopsin holds great promise as a memory component in future computers. This protein has the property of adopting different states in response to varying optical wavelengths. Its transition rates are very rapid. Bacteriorhodopsin could be used both in the processor and storage, making a computer smaller, faster, and more economical than semiconductor devices (40). However, over the past decade, not much progress has been made in this direction.

Biomechanical Machines. The mechanical properties of fibrous polypeptides could be put to use for the commercial production of fibers (qv) that are more elastic and resilient than available synthetics (see SILK). The biochemical properties of proteins could also be harnessed for the conversion of mechanical energy to chemical energy (41).

Enzymes for Extreme Conditions. The possibility of using enzymes from extremophiles, which thrive in oil wells, hot temperatures, freezing conditions, etc, is being explored for the removal of environmental contaminants and survival at extreme temperatures.

Conversion of Biomass for Fuels. Many predictions for protein uses in the 1970s have already been fulfilled (42). For example, biomass is being used for fuel. Many predictions were again made in 1995 (43).

BIBLIOGRAPHY

"Proteins" in *ECT* 1st ed., Vol. 11, pp. 226–248, by H. B. Vickery, The Connecticut Agricultural Station; in *ECT* 2nd ed., Vol. 16, pp. 610–640, by M. V. Tracey, Commonwealth Scientific and Industrial Research Organization, Australia; in *ECT* 3rd ed., Vol. 19, pp. 314–341, by P. L. Pellett, University of Massachusetts; in *ECT* 4th ed., Vol. 20, pp. 428–446, by Raman Nambudripad, Beth Israel Hospital, and Temple F. Smith, Boston University; "Proteins" in *ECT* (online), posting date: December 4, 2000, by Raman Nambudripad, Beth Israel Hospital, and Temple F. Smith, Boston University.

1. C. B. Anfinsen, *Science* **181**, 223 (1973).
2. R. Scopes, *Protein Purification: Principles and Practice*, 2nd ed., Springer-Verlag, Berlin, 1987.
3. F. Sanger, *Adv. Prot. Chem.* **7**, 1 (1952).
4. NCBI website: http://www.ncbi.nlm.nih.gov/Genomes/
5. A. Bairoch and B. Boeckmann, *Nucleic Acids Res.* **20**, 2019 (1992).
6. G. N. Ramachandran, C. Ramakrishnan and V. Sasisekharan, *J. Mol. Biol.* **7**, 95 (1963).
7. P. L. Privalov, *Adv. Prot. Chem.* **33**, 167 (1979).
8. C. Branden and J. Tooze, *Introduction to Protein Structure*, Garland Publishers, Inc., New York, 1991.
9. J. Li and co-workers, *Structure* **3**, 541 (1995).
10. P. L. Privalov, *Physical Basis of the Stability of the Folded Conformations of Proteins: Protein Folding*, Freeman, New York, 1992; K. A. Dill, *Biochemistry* **29**, 7133 (1990).
11. O. B. Ptitsyn, *Adv. Prot. Chem.* **47**, 83 (1995).
12. A. R. Dinner and co-workers, *Trends Biochem. Sci.* **25**, 331 (2000).
13. F. U. Hartl and M. Hayer-Hartl, *Science* **295**, 1852 (2002).
14. A. Smith, *Nature (London)* **426**, 883 (2003).
15. R. E. Dickerson and I. Geis, *The Structure and Action of Proteins*, Benjamin/Cummings, 1969.
16. W. E. Meador, A. R. Means, and F. A. Quiocho, *Science* **257**, 1251 (1992).
17. A. Fersht, *Enzyme Structure and Mechanism*, 2nd ed., W. H. Freeman, New York, 1985.
18. J. L. Kim, D. B. Nikolov, and S. K. Burley, *Nature (London)* **365**, 520 (1993).
19. H. Kitano, *Nature* **420**, 206 (2002); B. R. Jasny and L. B. Ray, *Science* **301**, 1863 (2003).
20. T. F. Smith and M. S. Waterman, *Adv. Appl. Math.* **2**, 482 (1981); S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* **48**, 443 (1970).
21. Rat Genome Sequencing Project Consortium, *Nature (London)* **428**, 493 (2004).
22. H. M. Berman and co-workers, *Nucleic Acids Res.* **28**, 235 (2000).
23. C. A. Orengo, D. T. Jones, and J. M. Thornton, *Nature (London)* **372**, 631 (1994).
24. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, *PJ. Mol. Biol.* **247**, 536 (1995).
25. C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry, Part II, Techniques for the Study of Biological Structure and Function*, W. H. Freeman, New York, 1980.
26. PDB Website – www.rcsb.org/pdb/holdings.html
27. U. Buckert and N. L. Allinger, *Molecular Mechanics*, American Chemical Society, Washington, D.C., 1982; T. A. Clark, *A Handbook of Computational Chemistry*, John Wiley & Sons. Inc., New York, 1985.
28. D. H. J. Mackay, in G. D. Fasman, ed., *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum Press, New York, 1989, p. 317.

29. P. Y. Chou and G. D. Fasman, *Ann. Rev. Biochem.* **47**, 251 (1978).
30. W. Kabsch and C. Sander, *Proc. Natl. Acad. Sci. USA* **81**, 1075 (1984).
31. M. S. Johnson, N. Srinivasan, R. Sowdhamini, and T. L. Blundell, *Crit. Rev. Biochem. Mol. Biol.* **29**, 1 (1994).
32. Q. Zheng, R. Rosenfeld, S. Vadja, and C. Delisi, *J. Comp. Chem.* **14**, 556 (1993).
33. K. Nishikawa, Y. Kubota, and T. Ooi, *J. Biochem.* **94**, 981 (1983); C. Zhang and K. Chou, *Protein Sci.* **1**, 401 (1992); I. Dubchak, S. R. Holbrook, and S-H. Kim, *Proteins* (1993).
34. C. Stultz, J. White, and T. F. Smith, *Protein Sci.* **2**, 305 (1993).
35. K. A. Dill, *Protein Sci.* **4**, 561 (1995).
36. J. A. Dimasi, N. R. Bryant, and L. Lasagna, *Clin. Pharmacol. Ther.* **50**, 471 (1991).
37. J. E. Brody, The New York Times, C1 (Nov. 7, 1995).
38. N. C. Cohen, J. M. Blaney, C. Humblet, P. Gund, and D. C. Barry, *J. Med. Chem.* **33**, 883 (1990). P. M. Dean, *Molecular Foundations of Drug-Receptor Interaction*, Cambridge University Press, Cambridge, U. K., 1987.
39. G. Gunter, *Nature (London)* **429**, 439 (2004).
40. R. R. Birge, *Computer* **25**, 56 (1992).
41. D. W. Urry, in C. G. Gebelin, ed., *Biotechnological Polymers: Medical, Pharmaceutical and Industrial Applications*, Technomic Publishers, Lancaster, Pa., 1993.
42. P. Handler, *Biology and the Future of Man*, Oxford University Press, New York, 1970.
43. D. E. Koshland, Jr., *Science* **267**, 1609 (1995).

Raman Nambudripad
Newton, MA

Temple F. Smith
Boston University

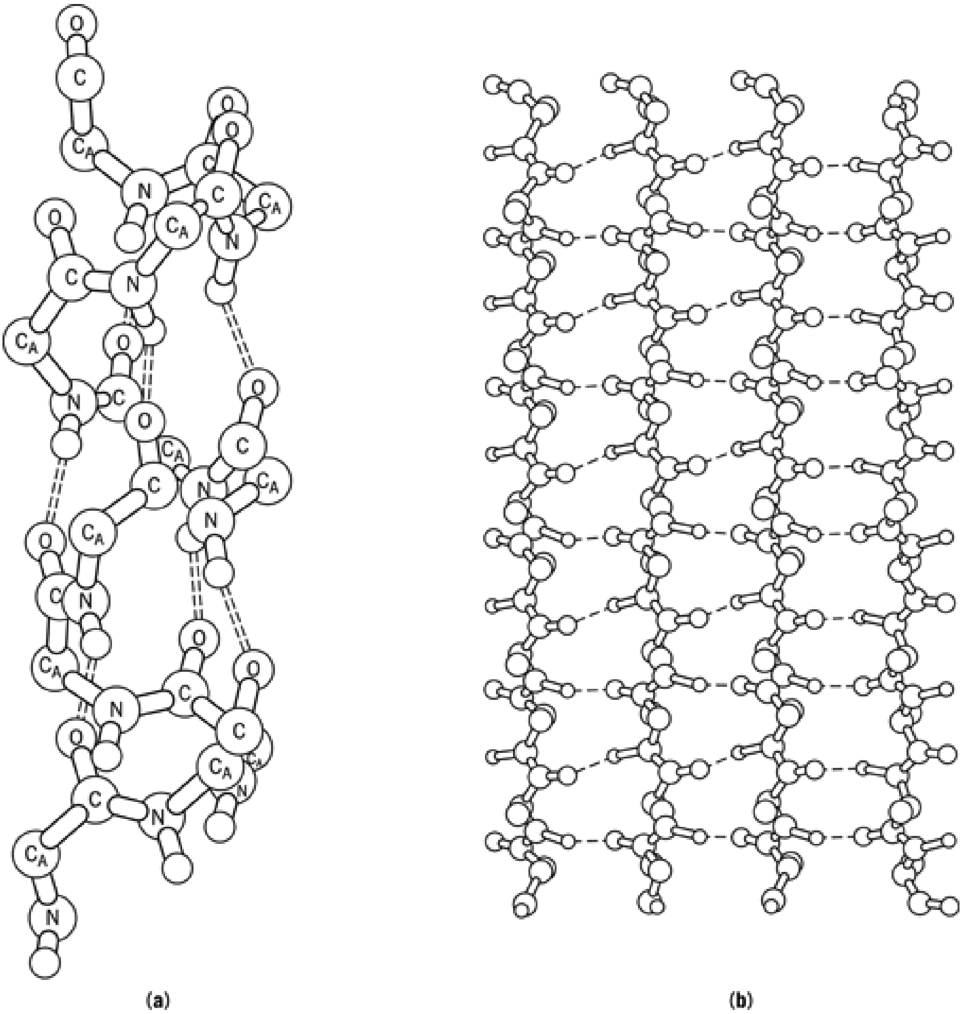


Fig. 1. The two principal elements of secondary structure in proteins. (**a**) The α-helix stabilized by hydrogen bonds between the backbone of residue *i* and *i*+4. There are 3.6 residues/turn of helix and an axial translation of 150 pm/residue. The parameter C_A represents the carbon connected to the amino acid side chain, R. (**b**) the β-sheet showing the hydrogen-bonding pattern between neighboring extended β-strands. Successive residues along the chain point alternately up and down the plane of the paper.

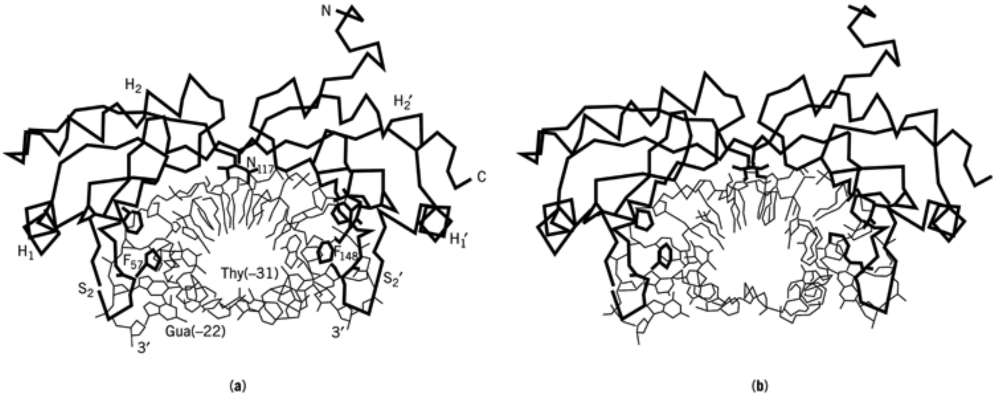


Fig. 2. Two views of the structure of the TATA-box binding protein (TBP)−DNA complex, where TATA is the nucleotide sequence thymine-adenosine-thymine-adenosine. The helices and strands of the protein (—) can be clearly seen. The DNA (—) is partially unwound upon binding to the protein (17).

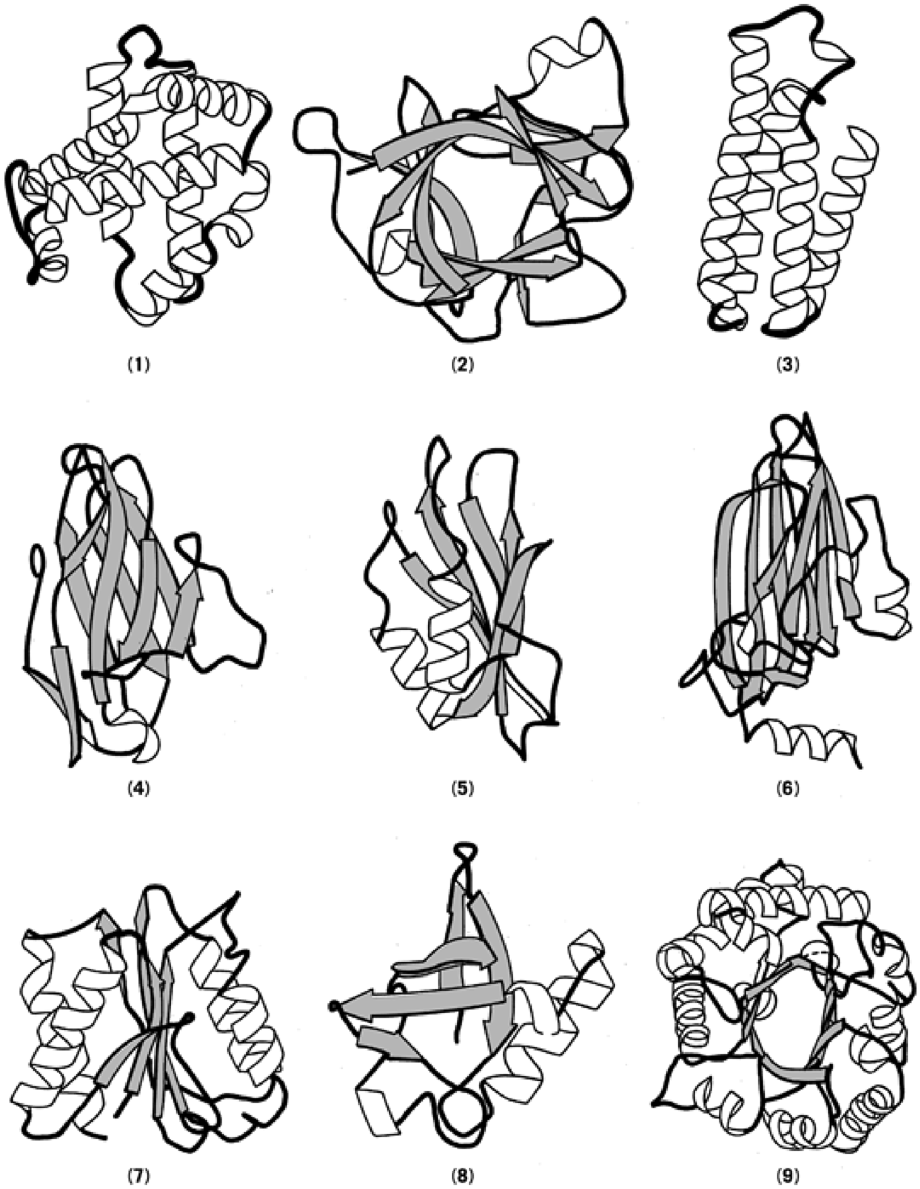


Fig. 3. Representation of the nine principal folds that recur in protein structures, where the codes of the representative proteins taken from the Brookhaven Protein Data Bank (PDB) codes (22) are given in parentheses (23): (**1**) globin (1THB); (**2**) trefoil (1ILB); (**3**) up−down (256B); (**4**) immunoglobulin folds (2RHE); (**5**) α/β-sandwich (1APS); (**6**) jelly roll (2STV); (**7**) doubly wound (4FXN); (**8**) UB α/β-roll (1UBQ); and (**9**) TIM barrel (7TIM).

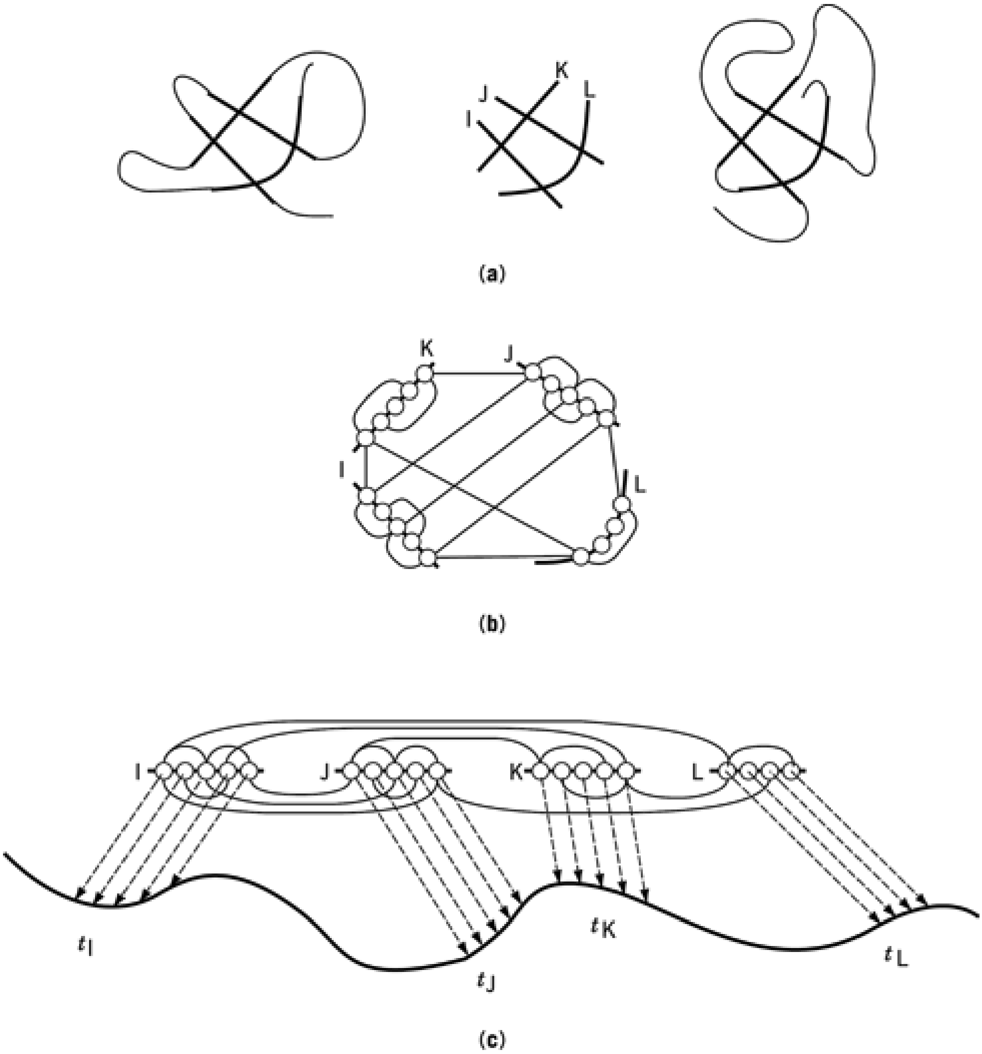


Fig. 4. Schematic of the protein threading problem, where (—) I–L represent particular core segments and (—) corresponds to variable loop segments: (**a**) the variable loop regions from two proteins are removed leaving a single core; (**b**) the interacting residues in the core are shown connected; and (**c**) one possible threading, *t*, of the sequence through the core where the dashed arrows represent structural positions occupied by the amino acids in the sequence.