SECTION FIVE

Structural Bioinformatics

CHAPTER TWELVE

Protein Structure Basics

Starting from this chapter and continuing through the next three chapters, we introduce the basics of protein structural bioinformatics. Proteins perform most essential biological and chemical functions in a cell. They play important roles in structural, enzymatic, transport, and regulatory functions. The protein functions are strictly determined by their structures. Therefore, protein structural bioinformatics is an essential element of bioinformatics. This chapter covers some basics of protein structures and associated databases, preparing the reader for discussions of more advanced topics of protein structural bioinformatics.

AMINO ACIDS

The building blocks of proteins are twenty naturally occurring amino acids, small molecules that contain a free amino group (NH_2) and a free carboxyl group (COOH). Both of these groups are linked to a central carbon (Cα), which is attached to a hydrogen and a side chain group (R) (Fig. 12.1). Amino acids differ only by the side chain R group. The chemical reactivities of the R groups determine the specific properties of the amino acids.

Amino acids can be grouped into several categories based on the chemical and physical properties of the side chains, such as size and affinity for water. According to these properties, the side chain groups can be divided into small, large, hydrophobic, and hydrophilic categories. Within the hydrophobic set of amino acids, they can be further divided into aliphatic and aromatic. *Aliphatic side chains* are linear hydrocarbon chains and *aromatic side chains* are cyclic rings. Within the hydrophilic set, amino acids can be subdivided into polar and charged. *Charged amino acids* can be either positively charged (basic) or negatively charged (acidic). Each of the twenty amino acids, their abbreviations, and main functional features once incorporated into a protein are listed in Table 12.1.

Of particular interest within the twenty amino acids are glycine and proline. Glycine, the smallest amino acid, has a hydrogen atom as the R group. It can therefore adopt more flexible conformations that are not possible for other amino acids. Proline is on the other extreme of flexibility. Its side chain forms a bond with its own backbone amino group, causing it to be cyclic. The cyclic conformation makes it very rigid, unable to occupy many of the main chain conformations adopted by other amino acids. In addition, certain amino acids are subject to modifications after

TABLE 12.1. Twenty Standard Amino Acids Grouped by Their Common Side-Chain Features

Amino Acid Group	Amino Acid Name	Three- and One-Letter Code	Main Functional Features
Small and nonpolar	Glycine Alanine Proline	Gly, G Ala, A Pro, P	Nonreactive in chemical reactions; Pro and Gly disrupt regular secondary structures
Small and polar	Cysteine Serine Threonine	Cys, C Ser, S Thr, T	Serving as posttranslational modification sites and participating in active sites of enzymes or binding metal
Large and polar	Glutamine Asparagine	Gln, Q Asn, N	Participating in hydrogen bonding or in enzyme active sites
Large and polar (basic)	Arginine Lysine Histidine	Arg, R Lys, K His, H	Found in the surface of globular proteins providing salt bridges; His participates in enzyme catalysis or metal binding
Large and polar (acidic)	Glutamate Aspartate	Glu, E Asp, D	Found in the surface of globular proteins providing salt bridges
Large and nonpolar (aliphatic)	Isoleucine Leucine Methionine Valine	Ile, I Leu, L Met, M Val, V	Nonreactive in chemical reactions; participating in hydrophobic interactions
Large and nonpolar (aromatic)	Phenylalanine Tyrosine Tryptophan	Phe, F Tyr, Y Trp, W	Providing sites for aromatic packing interactions; Tyr and Trp are weakly polar and can serve as sites for phosphorylation and hydrogen bonding

Note: Each amino acid is listed with its full name, three- and one-letter abbreviations, and main functional roles when serving as amino acid residues in a protein. Properties of some amino acid groups overlap.

a protein is translated in a cell. This is called *posttranslational modification*, and is discussed in more detail in Chapter 19.

PEPTIDE FORMATION

The peptide formation involes two amino acids covalently joined together between the carboxyl group of one amino acid and the amino group of another (Fig. 12.2). This

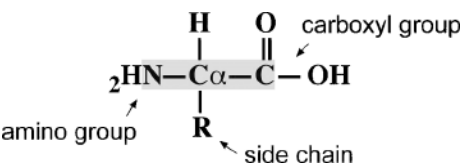


Figure 12.1: General structure of an amino acid. The main chain atoms are highlighted. The R group can be any of the twenty amino acid side chains.

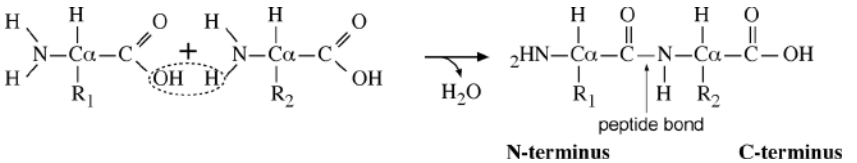


Figure 12.2: Condensation reaction between the carboxyl group of one amino acid and the amino group of another. The hydroxyl group of the carboxyl group and a hydrogen of the amino group are lost to give rise to a water molecule and a dipeptide.

reaction is a condensation reaction involving removal of elements of water from the two molecules. The resulting product is called a *dipeptide*. The newly formed covalent bond connecting the two amino acids is called a *peptide bond*. Once an amino acid is incorporated into a peptide, it becomes an amino acid residue. Multiple amino acids can be joined together to form a longer chain of amino acid polymer.

A linear polymer of more than fifty amino acid residues is referred to as a *polypeptide*. A polypeptide, also called a protein, has a well-defined three-dimensional arrangement. On the other hand, a polymer with fewer than fifty residues is usually called a peptide without a well-defined three-dimensional structure. The residues in a peptide or polypeptide are numbered beginning with the residue containing the amino group, referred to as the *N*-terminus, and ending with the residue containing the carboxyl group, known as the *C*-terminus (see Fig. 12.2). The actual sequence of amino acid residues in a polypeptide determines its ultimate structure and function.

The atoms involved in forming the peptide bond are referred to as the *backbone atoms*. They are the nitrogen of the amino group, the α carbon to which the side chain is attached and carbon of the carbonyl group.

DIHEDRAL ANGLES

A peptide bond is actually a partial double bond owing to shared electrons between O=C–N atoms. The rigid double bond structure forces atoms associated with the peptide bond to lie in the same plane, called the *peptide plane*. Because of the planar nature of the peptide bond and the size of the R groups, there are considerable restrictions on the rotational freedom by the two bonded pairs of atoms around the peptide bond. The angle of rotation about the bond is referred to as the *dihedral angle* (also called the *tortional angle*).

For a peptide unit, the atoms linked to the peptide bond can be moved to a certain extent by the rotation of two bonds flanking the peptide bond. This is measured by two dihedral angles (Fig. 12.3). One is the dihedral angle along the N–Cα bond, which is defined as phi (ϕ); and the other is the angle along the Cα–C bond, which is called psi (ψ). Various combinations of ϕ and ψ angles allow the proteins to fold in many different ways.

Ramachandran Plot

As mentioned, the rotation of ϕ and ψ is not completely free because of the planar nature of the peptide bond and the steric hindrance from the side chain R group.

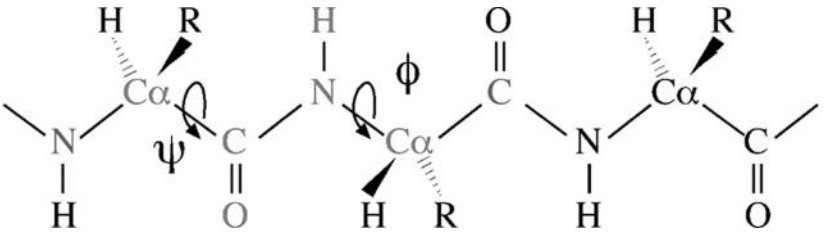


Figure 12.3: Definition of dihedral angles of ϕ and ψ. Six atoms around a peptide bond forming two peptide planes are colored in red. The ϕ angle is the rotation about the N–Cα bond, which is measured by the angle between a virtual plane formed by the C–N–Cα and the virtual plane by N–Cα–C (C in green). The ψ angle is the rotation about the Cα–C bond, which is measured by the angle between a virtual plane formed by the N–Cα–C (N in green) and the virtual plane by Cα–C–N (N in red) (see color plate section).

Consequently, there is only a limited range of peptide conformation. When ϕ and ψ angles of amino acids of a particular protein are plotted against each other, the resulting diagram is called a Ramachandran plot. This plot maps the entire conformational space of a peptide and shows sterically allowed and disallowed regions (Fig. 12.4). It can be very useful in evaluating the quality of protein models.

HIERARCHY

Protein structures can be organized into four levels of hierarchies with increasing complexity. These levels are primary structure, secondary structure, tertiary structure, and quaternary structure. A linear amino acid sequence of a protein is the primary structure. This is the simplest level with amino acid residues linked together through

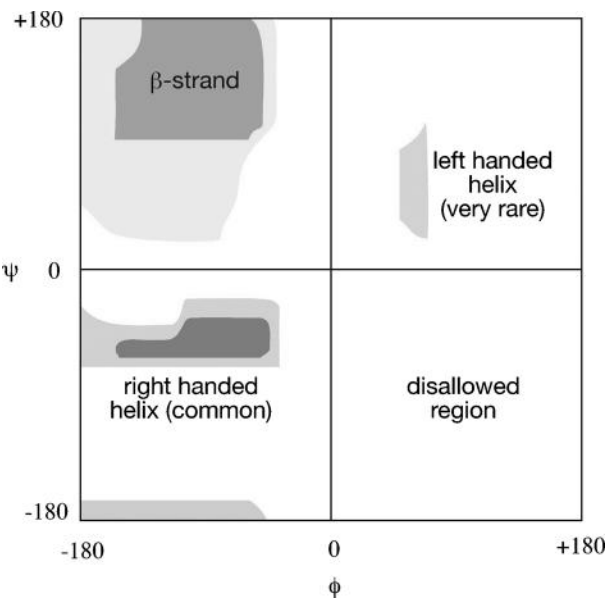


Figure 12.4: A Ramachandran plot with allowed values of ϕ and ψ in shaded areas. Regions favored by α-helices and β-strands (to be explained) are indicated.

peptide bonds. The next level up is the secondary structure, defined as the local conformation of a peptide chain. The secondary structure is characterized by highly regular and repeated arrangement of amino acid residues stabilized by hydrogen bonds between main chain atoms of the C=O group and the NH group of different residues. The level above the secondary structure is the tertiary structure, which is the three-dimensional arrangement of various secondary structural elements and connecting regions. The tertiary structure can be described as the complete three-dimensional assembly of all amino acids of a single polypeptide chain. Beyond the tertiary structure is the quaternary structure, which refers to the association of several polypeptide chains into a protein complex, which is maintained by noncovalent interactions. In such a complex, individual polypeptide chains are called *monomers* or *subunits*. Intermediate between secondary and tertiary structures, a level of supersecondary structure is often used, which is defined as two or three secondary structural elements forming a unique functional domain, a recurring structural pattern conserved in evolution.

Stabilizing Forces

Protein structures from secondary to quaternary are maintained by noncovalent forces. These include electrostatic interactions, van der Waals forces, and hydrogen bonding. Electrostatic interactions are a significant stabilizing force in a protein structure. They occur when excess negative charges in one region are neutralized by positive charges in another region. The result is the formation of salt bridges between oppositely charged residues. The electrostatic interactions can function within a relatively long range (15 Å).

Hydrogen bonds are a particular type of electrostatic interactions similar to dipole–dipole interactions involving hydrogen from one residue and oxygen from another. Hydrogen bonds can occur between main chain atoms as well as side chain atoms. Hydrogen from the hydrogen bond donor group such as the N–H group is slightly positively charged, whereas oxygen from the hydrogen bond acceptor group such as the C=O group is slightly negatively charged. When they come within a close distance (<3 Å), a partial bond is formed between them, resulting in a hydrogen bond. Hydrogen bonding patterns are a dominant factor in determining different types of protein secondary structures.

Van der Waals forces also contribute to the overall protein stability. These forces are instantaneous interactions between atoms when they become transient dipoles. A transient dipole can induce another transient dipole nearby. The dipoles of the two atoms can be reversed a moment later. The oscillating dipoles result in an attractive force. The van der Waals interactions are weaker than electrostatic and hydrogen bonds and thus only have a secondary effect on the protein structure.

In addition to these common stabilizing forces, disulfide bridges, which are covalent bonds between the sulfur atoms of the cysteine residue, are also important in maintaining some protein structures. For certain types of proteins that contain metal ions as prosthetic groups, noncovalent interactions between amino acid residues and the metal ions may play an important structural role.

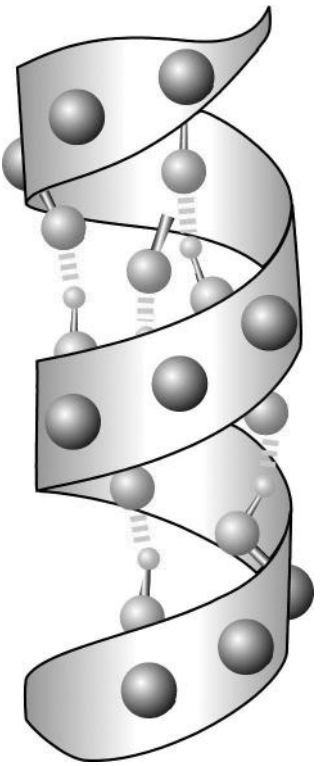


Figure 12.5: A ribbon diagram of an α-helix with main chain atoms (as grey balls) shown. Hydrogen bonds between the carbonyl oxygen (red) and the amino hydrogen (green) of two residues are shown in yellow dashed lines (see color plate section).

SECONDARY STRUCTURES

As mentioned, local structures of a protein with regular conformations are known as secondary structures. They are stabilized by hydrogen bonds formed between carbonyl oxygen and amino hydrogen of different amino acids. Chief elements of secondary structures are α-helices and β-sheets.

α-Helices

An α-helix has a main chain backbone conformation that resembles a corkscrew. Nearly all known α-helices are right handed, exhibiting a rightward spiral form. In such a helix, there are 3.6 amino acids per helical turn. The structure is stabilized by hydrogen bonds formed between the main chain atoms of residues i and $i + 4$. The hydrogen bonds are nearly parallel with the helical axis (Fig. 12.5). The average ϕ and ψ angles are 60° and 45°, respectively, and are distributed in a narrowly defined region in the lower left region of a Ramachandran plot (see Fig. 12.4). Hydrophobic residues of the helix tend to face inside and hydrophilic residues of the helix face outside. Thus, every third residue along the helix tends to be a hydrophobic residue. Ala, Gln, Leu, and Met are commonly found in an α-helix, but not Pro, Gly, and Tyr. These rules are useful in guiding the prediction of protein secondary structures.

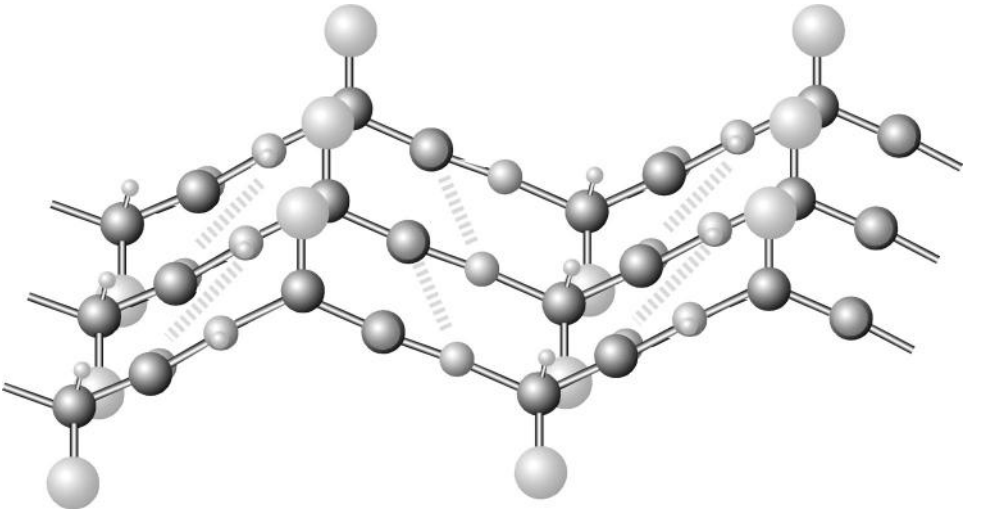


Figure 12.6: Side view of a parallel β-sheet. Hydrogen bonds between the carbonyl oxygen (red) and the amino hydrogen (green) of adjacent β-strands are shown in yellow dashed lines. R groups are shown as big balls in cyan and are positioned alternately on opposite sides of β-strands (see color plate section).

β-Sheets

A β-sheet is a fully extended configuration built up from several spatially adjacent regions of a polypeptide chain. Each region involved in forming the β-sheet is a β-strand. The β-strand conformation is pleated with main chain backbone zigzagging and side chains positioned alternately on opposite sides of the sheet. β-Strands are stabilized by hydrogen bonds between residues of adjacent strands (Fig. 12.6). β-strands near the surface of the protein tend to show an alternating pattern of hydrophobic and hydrophilic regions, whereas strands buried at the core of a protein are nearly all hydrophobic.

The β-strands can run in the same direction to form a parallel sheet or can run every other chain in reverse orientation to form an antiparallel sheet, or a mixture of both. The hydrogen bonding patterns are different in each configurations. The ϕ and ψ angles are also widely distributed in the upper left region in a Ramachandran plot (see Fig. 12.4). Because of the long-range nature of residues involved in this type of conformation, it is more difficult to predict β-sheets than α-helices.

Coils and Loops

There are also local structures that do not belong to regular secondary structures (α-helices and β-strands). The irregular structures are coils or loops. The loops are often characterized by sharp turns or hairpin-like structures. If the connecting regions are completely irregular, they belong to random coils. Residues in the loop or coil regions tend to be charged and polar and located on the surface of the protein structure. They are often the evolutionarily variable regions where mutations, deletions,

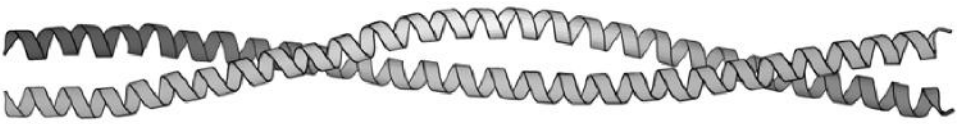


Figure 12.7: An α-helical coiled coil found in tropomyosin showing two helices wound around to form a helical bundle.

and insertions frequently occur. They can be functionally significant because these locations are often the active sites of proteins.

Coiled Coils

Coiled coils are a special type of supersecondary structure characterized by a bundle of two or more α-helices wrapping around each other (Fig. 12.7). The helices forming coiled coils have a unique pattern of hydrophobicity, which repeats every seven residues (five hydrophobic and two hydrophilic). More details on coiled coils and their structure prediction are discussed in Chapter 14.

TERTIARY STRUCTURES

The overall packing and arrangement of secondary structures form the tertiary structure of a protein. The tertiary structure can come in various forms but is generally classified as either globular or membrane proteins. The former exists in solvents through hydrophilic interactions with solvent molecules; the latter exists in membrane lipids and is stabilized through hydrophobic interactions with the lipid molecules.

Globular Proteins

Globular proteins are usually soluble and surrounded by water molecules. They tend to have an overall compact structure of spherical shape with polar or hydrophilic residues on the surface and hydrophobic residues in the core. Such an arrangement is energetically favorable because it minimizes contacts with water by hydrophobic residues in the core and maximizes interactions with water by surface polar and charged residues. Common examples of globular proteins are enzymes, myoglobins, cytokines, and protein hormones.

Integral Membrane Proteins

Membrane proteins exist in lipid bilayers of cell membranes. Because they are surrounded by lipids, the exterior of the proteins spanning the membrane must be very hydrophobic to be stable. Most typical transmembrane segments are α-helices. Occasionally, for some bacterial periplasmic membrane proteins, they are composed of β-strands. The loops connecting these segments sometimes lie in the aqueous phase, in which they can be entirely hydrophilic. Sometimes, they lie in the interface between the lipid and aqueous phases and are amphipathic in nature (containing polar residues facing the aqueous side and hydrophobic residues towards the lipid side). The amphipathic residues can also form helices which have a periodicity of

three or four residues. Common examples of membrane proteins are rhodopsins, cytochrome *c* oxidase, and ion channel proteins.

DETERMINATION OF PROTEIN THREE-DIMENSIONAL STRUCTURE

Protein three-dimensional structures are obtained using two popular experimental techniques, x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. The experimental procedures and relative merits of each method are discussed next.

X-ray Crystallography

In x-ray protein crystallography, proteins need to be grown into large crystals in which their positions are fixed in a repeated, ordered fashion. The protein crystals are then illuminated with an intense x-ray beam. The x-rays are deflected by the electron clouds surrounding the atoms in the crystal producing a regular pattern of diffraction. The diffraction pattern is composed of thousands of tiny spots recorded on a x-ray film. The diffraction pattern can be converted into an electron density map using a mathematical procedure known as Fourier transform. To interpret a three-dimensional structure from two-dimensional electron density maps requires solving the phases in the diffraction data. The phases refer to the relative timing of different diffraction waves hitting the detector. Knowing the phases can help to determine the relative positions of atoms in a crystal.

Phase solving can be carried out by two methods, molecular replacement, and multiple isomorphous replacement. Molecular replacement uses a homologous protein structure as template to derive an initial estimate of the phases. Multiple isomorphous replacement derives phases by comparing electron intensity changes in protein crystals containing heavy metal atoms and the ones without heavy metal atoms. The heavy atoms diffract x-rays with unusual intensities, which can serve as a marker for relative positions of atoms.

Once the phases are available, protein structures can be solved by modeling with amino acid residues that best fit the electron density map. The quality of the final model is measured by an R factor, which indicates how well the model reproduces the experimental electron intensity data. The R factor is expressed as a percentage of difference between theoretically reproduced diffraction data and experimentally determined diffraction data. R values can range from 0.0, which is complete agreement, to 0.59, which is complete disagreement. A major limitation of x-ray crystallography is whether suitable crystals of proteins of interest can be obtained.

Nuclear Magnetic Resonance Spectroscopy

NMR spectroscopy detects spinning patterns of atomic nuclei in a magnetic field. Protein samples are labeled with radioisotopes such as ^{13}C and ^{15}N. A radiofrequency radiation is used to induce transitions between nuclear spin states in a magnetic field. Interactions between spinning isotope pairs produce radio signal peaks that correlate with the distances between them. By interpreting the signals observed using NMR,

proximity between atoms can be determined. Knowledge of distances between all labeled atoms in a protein allows a protein model to be built that satisfies all the constraints. NMR determines protein structures in solution, which has the advantage of not requiring the crystallization process. However, the proteins in solution are mobile and vibrating, reflecting the dynamic behavior of proteins. For that reason, usually a number of slightly different models (twenty to forty) have to be constructed that satisfy all the NMR distance measurements. The NMR technique obviates the need of growing protein crystals and can solve structures relatively more quickly than x-ray crystallography. The major problem associated with using NMR is the current limit of protein size (<200 residues) that can be determined. Another problem is the requirement of heavy instrumentation.

PROTEIN STRUCTURE DATABASE

Once the structure of a particular protein is solved, a table of (x, y, z) coordinates representing the spatial position of each atom of the structure is created. The coordinate information is required to be deposited in the Protein Data Bank (PDB, www.rcsb.org/pdb/) as a condition of publication of a journal paper. PDB is a worldwide central repository of structural information of biological macromolecules and is currently managed by the Research Collaboratory for Structural Bioinformatics (RCSB). In addition, the PDB website provides a number of services for structure submission and data searching and retrieval. Through its web interface, called *Structure Explorer*, a user is able to read the summary information of a protein structure, view and download structure coordinate files, search for structure neighbors of a particular protein or access related research papers through links to the NCBI PubMed database.

There are currently more than 30,000 entries in the database with the number increasing at a dramatic rate in recent years owing to large-scale structural proteomics projects being carried out. Most of the database entries are structures of proteins. However, a small portion of the database is composed of nucleic acids, carbohydrates, and theoretical models. Most protein structures are determined by x-ray crystallography and a smaller number by NMR.

Although the total number of entries in PDB is large, most of the protein structures are redundant, namely, they are structures of the same protein determined under different conditions, at different resolutions, or associated with different ligands or with single residue mutations. Sometimes, structures from very closely related proteins are determined and deposited in PDB. A small number of well-studied proteins such as hemoglobins and myoglobins have hundreds of entries. Excluding the redundant entries, there are approximately 3,000 unique protein structures represented in the database. Among the unique protein structures, there are only a limited number of protein folds available (800) compared to ~1,000,000 unique protein sequences already known, suggesting that the protein structures are much more conserved. A protein fold is a particular topological arrangement of helices, strands, and loops. Protein classification by folds is discussed in Chapter 13.

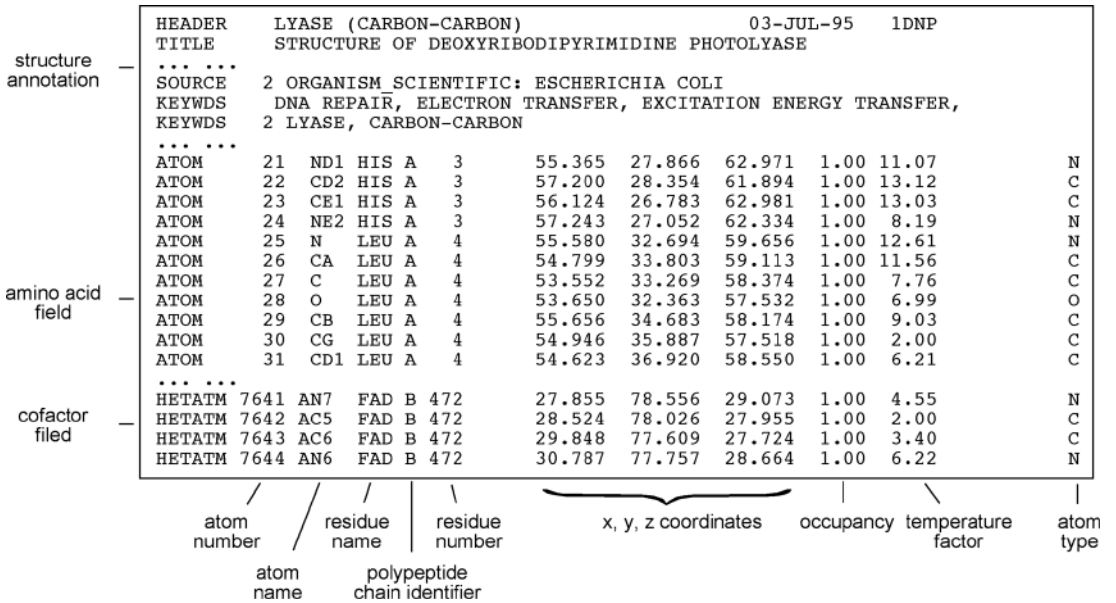


Figure 12.8: A partial PDB file of DNA photolyase (*boxed*) showing the header section and the coordinate section. The coordinate section is dissected based on individual fields.

PDB Format

A deposited set of protein coordinates becomes an entry in PDB. Each entry is given a unique code, PDBid, consisting of four characters of either letters A to Z or digits 0 to 9 such as 1LYZ and 4RCR. One can search a structure in PDB using the four-letter code or keywords related to its annotation. The identified structure can be viewed directly online or downloaded to a local computer for analysis. The PDB website provides options for retrieval, analysis, and direct viewing of macromolecular structures. The viewing can be still images or manipulable images through interactive viewing tools (see Chapter 13). It also provides links to protein structural classification results available in databases such as SCOP and CATH (see Chapter 13).

The data format in PDB was created in the early 1970s and has a rigid structure of 80 characters per line, including spaces. This format was initially designed to be compatible with FORTRAN programs. It consists of an explanatory header section followed by an atomic coordinate section (Fig. 12.8).

The header section provides an overview of the protein and the quality of the structure. It contains information about the name of the molecule, source organism, bibliographic reference, methods of structure determination, resolution, crystallographic parameters, protein sequence, cofactors, and description of structure types and locations and sometimes secondary structure information. In the structure coordinates section, there are a specified number of columns with predetermined contents. The ATOM part refers to protein atom information whereas the HETATM (for heteroatom group) part refers to atoms of cofactor or substrate molecules. Approximately ten columns of text and numbers are designated. They include information for the atom

number, atom name, residue name, polypeptide chain identifier, residue number, *x*, *y*, and *z* Cartesian coordinates, temperature factor, and occupancy factor. The last two parameters, occupancy and temperature factors, relate to disorders of atomic positions in crystals.

The PDB format has been in existence for more than three decades. It is fairly easy to read and simple to use. However, the format is not designed for computer extraction of information from the records. Certain restrictions in the format have significantly complicated its current use. For instance, in the PDB format, only Cartesian coordinates of atoms are given without bonding information. Information such as disulfide bonds has to be interpreted by viewing programs, some of which may fail to do so. In addition, the field width for atom number is limited to five characters, meaning that the maximum number of atoms per model is 99,999. The field width for polypeptide chains is only one character in width, meaning that no more than 26 chains can be used in a multisubunit protein model. This has made many large protein complexes such as ribosomes unable to be represented by a single PDB file. They have to be divided into multiple PDB files.

mmCIF and MMDB Formats

Significant limitations of the PDB format have allowed the development of new formats to handle increasingly complicated structure data. The most popular new formats include the macromolecular crystallographic information file (mmCIF) and the molecular modeling database (MMDB) file. Both formats are highly parsable by computer software, meaning that information in each field of a record can be retrieved separately. These new formats facilitate the retrieval and organization of information from database structures.

The mmCIF format is similar to the format for a relational database (see Chapter 2) in which a set of tables are used to organize database records. Each table or field of information is explicitly assigned by a tag and linked to other fields through a special syntax. An example of an mmCIF containing multiple fields is given below. As shown in Figure 12.9, a single line of description in the header section of PDB is divided into many lines or fields with each field having explicit assignment of item names and item values. Each field starts with an underscore character followed by category name and keyword description separated by a period. The annotation in Figure 12.9 shows that the data items belong to the category of "struct" or "database." Following a keyword tag, a short text string enclosed by quotation marks is used to assign values for the keyword. Using multiple fields with tags for the same information has the advantage of providing an explicit reference to each item in a data file and ensures a one-to-one relationship between item names and item values. By presenting the data item by item, the format provides much more flexibility for information storage and retrieval.

Another new format is the MMDB format developed by the NCBI to parse and sort pieces of information in PDB. The objective is to allow the information to be more easily integrated with GenBank and Medline through Entrez (see Chapter 2).

PDB HEADER PLANT SEED PROTEIN 11-OCT-91 1CBN

mmCIF

```
_struct.entry_id '1CBN'
_struct.title 'PLANT SEED PROTEIN'
_struct_keywords.entry_id '1CBN'
_struct_keywords.text 'plant seed protein'
_database_2.database_id 'PDB'
_database_2.database_code '1CBN'
_database_PDB_rev.rev_num 1
_database_PDB_rev.date_original '1991-10-11'
```

Figure 12.9: A comparison of PDB and mmCIF formats in two different boxes. To show the same header information in PDB, multiple fields are required in mmCIF to establish explicit relationships of item name and item values. The advantage of such format is easy parsing by computer software.

An MMDB file is written in the ASN.1 format (see Chapter 2), which has information in a record structured as a nested hierarchy. This allows faster retrieval than mmCIF and PDB. Furthermore, the MMDB format includes bond connectivity information for each molecule, called a "chemical graph," which is recorded in the ASN.1 file. The inclusion of the connectivity data allows easier drawing of structures.

SUMMARY

Proteins are considered workhorses in a cell and carry out most cellular functions. Knowledge of protein structure is essential to understand the behavior and functions of specific proteins. Proteins are polypeptides formed by joining amino acids together via peptide bonds. The folding of a polypeptide can be described by rotational angles around the main chain bonds such as ϕ and ψ angles. The degree of rotation depends on the preferred protein conformation. Allowable ϕ and ψ angles in a protein can be specified in a Ramachandran plot. There are four levels of protein structures, primary, secondary, tertiary, and quaternary. The primary structure is the sequence of amino acid residues. The secondary structure is the repeated main chain conformation, which includes α-helices and β-sheets. The tertiary structure is the overall three-dimensional conformation of a polypeptide chain. The quaternary structure is the complex arrangement of multiple polypeptide chains. Protein structures are stabilized by electrostatic interactions, hydrogen bonds, and van der Waals interactions. Proteins can be classified as being soluble globular proteins or integral membrane proteins, whose structures vary tremendously. Protein structures can be determined by x-ray crystallography and NMR spectroscopy. Both methods have advantages and disadvantages, but are clearly complementary. The solved structures are deposited in PDB, which uses a PDB format to describe structural details. However, the original PDB format has limited capacity and is difficult to be parsed by computer software.

To overcome the limitations, new formats such as mmCIF and MMDB have been developed.

FURTHER READING

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

Scheeff, E. D., and Fink, J. L. 2003. "Fundamentals of protein structure." In *Structural Bioinformatics*, edited by P. E. Bourne and H. Weissig, 15–39. Hoboken, NJ: Wiley-Liss.

Westbrook, J. D., and Fitzgerald, P. M. D. 2003. "The PDB format, mmCIF and other data formats." In *Structural Bioinformatics*, edited by P. E. Bourne and H. Weissig, 161–79. Hoboken, NJ: Wiley-Liss.

CHAPTER THIRTEEN

Protein Structure Visualization, Comparison, and Classification

Once a protein structure has been solved, the structure has to be presented in a three-dimensional view on the basis of the solved Cartesian coordinates. Before computer visualization software was developed, molecular structures were represented by physical models of metal wires, rods, and spheres. With the development of computer hardware and software technology, sophisticated computer graphics programs have been developed for visualizing and manipulating complicated three-dimensional structures. The computer graphics help to analyze and compare protein structures to gain insight to functions of the proteins.

PROTEIN STRUCTURAL VISUALIZATION

The main feature of computer visualization programs is interactivity, which allows users to visually manipulate the structural images through a graphical user interface. At the touch of a mouse button, a user can move, rotate, and zoom an atomic model on a computer screen in real time, or examine any portion of the structure in great detail, as well as draw it in various forms in different colors. Further manipulations can include changing the conformation of a structure by protein modeling or matching a ligand to an enzyme active site through docking exercises.

Because a Protein Data Bank (PDB) data file for a protein structure contains only x, y, and z coordinates of atoms (see Chapter 12), the most basic requirement for a visualization program is to build connectivity between atoms to make a view of a molecule. The visualization program should also be able to produce molecular structures in different styles, which include wire frames, balls and sticks, space-filling spheres, and ribbons (Fig. 13.1).

A wire-frame diagram is a line drawing representing bonds between atoms. The wire frame is the simplest form of model representation and is useful for localizing positions of specific residues in a protein structure, or for displaying a skeletal form of a structure when Cα atoms of each residue are connected. Balls and sticks are solid spheres and rods, representing atoms and bonds, respectively. These diagrams can also be used to represent the backbone of a structure. In a space-filling representation (or Corey, Pauling, and Koltan [CPK]), each atom is described using large solid spheres with radii corresponding to the van der Waals radii of the atoms. Ribbon diagrams use cylinders or spiral ribbons to represent α-helices and broad, flat arrows to represent β-strands. This type of representation is very attractive in that it allows easy

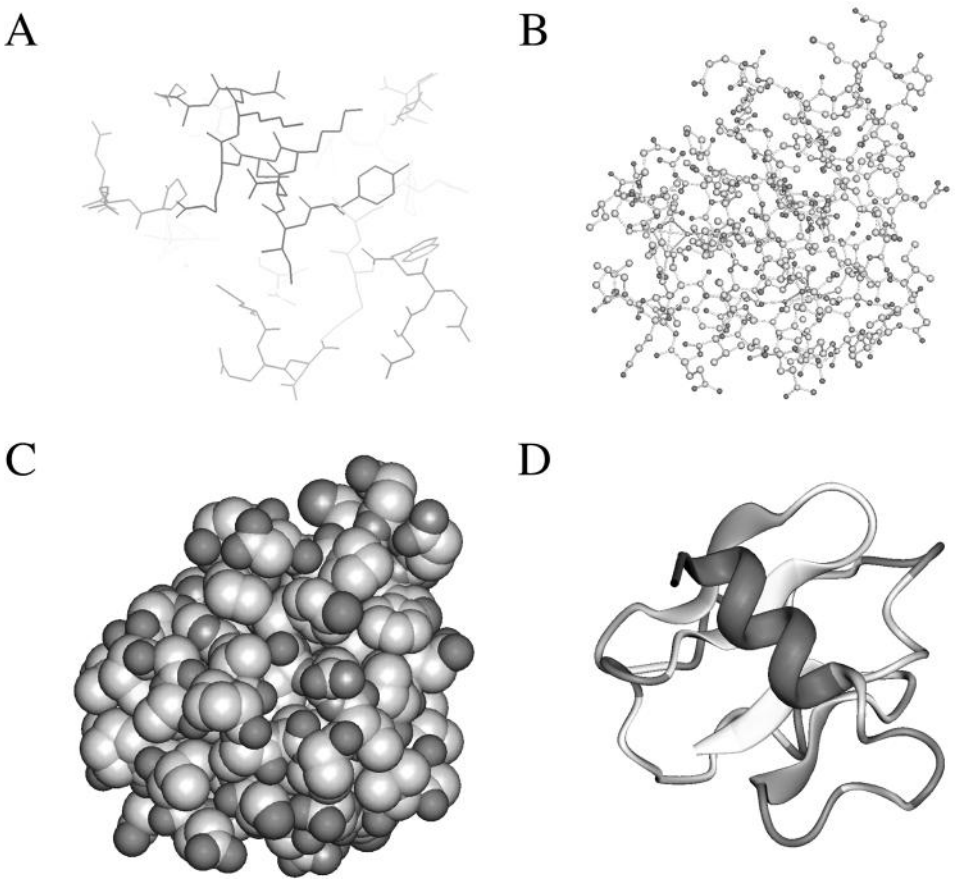


Figure 13.1: Examples of molecular structure visualization forms. **(A)** Wireframes. **(B)** Balls and sticks. **(C)** Space-filling spheres. **(D)** Ribbons (see color plate section).

identification of secondary structure elements and gives a clear view of the overall topology of the structure. The resulting images are also visually appealing.

Different representation styles can be used in combination to highlight a certain feature of a structure while deemphasizing the structures surrounding it. For example, a cofactor of an enzyme can be shown as space-filling spheres while the rest of the protein structure is shown as wire frames or ribbons. Some widely used and freely available software programs for molecular graphics are introduced next with examples of rendering provided in Figure 13.2.

RasMol (http://rutgers.rcsb.org/pdb/help-graphics.html#rasmol_download) is a command-line–based viewing program that calculates connectivity of a coordinate file and displays wireframe, cylinder, stick bonds, α-carbon trace, space-filling (CPK) spheres, and ribbons. It reads both PDB and mmCIF formats and can display a whole molecule or specific parts of it. It is available in multiple platforms: UNIX, Windows, and Mac. RasTop (www.geneinfinity.org/rastop/) is a new version of RasMol for Windows with a more enhanced user interface.

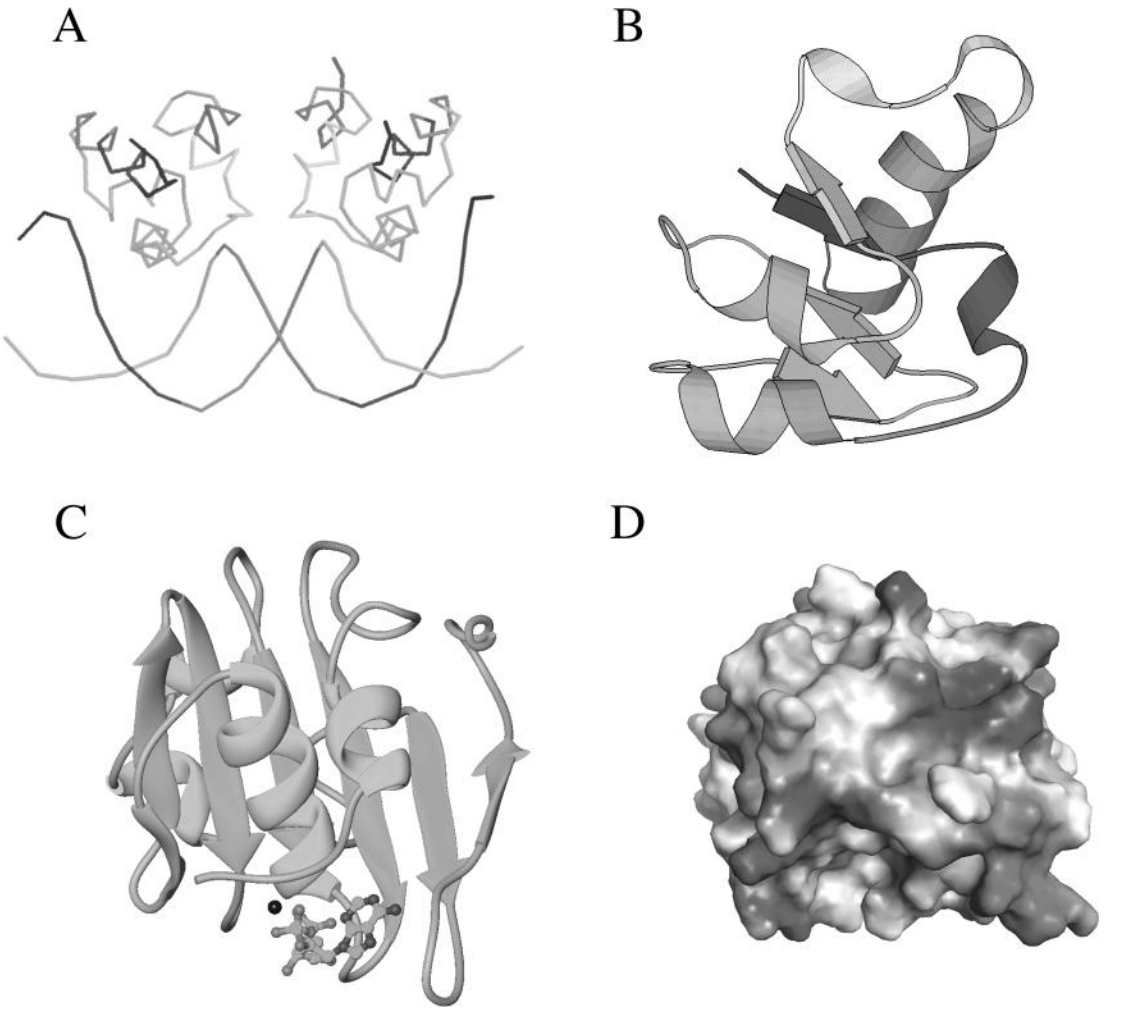


Figure 13.2: Examples of molecular graphic generated by **(A)** Rasmol, **(B)** Molscript, **(C)** Ribbons, and **(D)** Grasp (see color plate section).

Swiss-PDBViewer (www.expasy.ch/spdbv/) is a structure viewer for multiple platforms. It is essentially a Swiss-Army knife for structure visualization and modeling because it incorporates so many functions in a small shareware program. It is capable of structure visualization, analysis, and homology modeling. It allows display of multiple structures at the same time in different styles, by charge distribution, or by surface accessibility. It can measure distances, angles, and even mutate residues. In addition, it can calculate molecular surface, electrostatic potential, Ramachandran plot, and so on. The homology modeling part includes energy minimization and loop modeling.

Molscript (www.avatar.se/molscript/) is a UNIX program capable of generating wire-frame, space-filling, or ball-and-stick styles. In particular, secondary structure elements can be drawn with solid spirals and arrows representing α-helices

and β-strands, respectively. Visually appealing images can be generated that are of publication quality. The drawback is that the program is command-line–based and not very user friendly. A modified UNIX program called Bobscript (www.strubi.ox.ac.uk/bobscript/) is available with enhanced features.

Ribbons (http://sgce.cbse.uab.edu/ribbons/) another UNIX program similar to Molscript, generates ribbon diagrams depicting protein secondary structures. Aesthetically appealing images can be produced that are of publication quality. However, the program, which is also command-line-based, is extremely difficult to use.

Grasp (http://trantor.bioc.columbia.edu/grasp/) is a UNIX program that generates solid molecular surface images and uses a gradated coloring scheme to display electrostatic charges on the surface.

There are also a number of web-based visualization tools that use Java applets. These programs tend to have limited molecular display features and low-quality images. However, the advantage is that the user does not have to download, compile, and install the programs locally, but simply view the structures on a web browser using any kind of computer operating system. In fact, the PDB also attempts to simplify the database structure display for end users. It has incorporated a number of light-weight Java-based structure viewers in the PDB web site (see Chapter 12).

WebMol (www.cmpharm.ucsf.edu/cgi-bin/webmol.pl) is a web-based program built based on a modified RasMol code and thus shares many similarities with RasMol. It runs directly on a browser of any type as an applet and is able to display simple line drawing models of protein structures. It also has a feature of interactively displaying Ramachandran plots for structure model evaluation.

Chime (www.mdlchime.com/chime/) is a plug-in for web browsers; it is not a stand-alone program and has to be invoked in a web browser. The program is also derived from RasMol and allows interactive display of graphics of protein structures inside a web browser.

Cn3D (www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml) is a helper application for web browsers to display structures in the MMDB format from the NCBI's structural database. It can be used on- or offline as a stand-alone program. It is able to render three-dimensional molecular models and display secondary structure cartoons. The drawback is that it does not recognize the PDB format.

PROTEIN STRUCTURE COMPARISON

With the visualization and computer graphics tools available, it becomes easy to observe and compare protein structures. To compare protein structures is to analyze two or more protein structures for similarity. The comparative analysis often, but not always, involves the direct alignment and superimposition of structures in a three-dimensional space to reveal which part of structure is conserved and which part is different at the three-dimensional level.

This structure comparison is one of the fundamental techniques in protein structure analysis. The comparative approach is important in finding remote protein homologs. Because protein structures have a much higher degree of conservation than the sequences, proteins can share common structures even without sequence similarity. Thus, structure comparison can often reveal distant evolutionary relationships between proteins, which is not feasible using the sequence-based alignment approach alone. In addition, protein structure comparison is a prerequisite for protein structural classification into different fold classes. It is also useful in evaluating protein prediction methods by comparing theoretically predicted structures with experimentally determined ones.

One can always compare structures manually or by eye, which is often practiced. However, the best approach is to use computer algorithms to automate the task and thereby get more accurate results. Structure comparison algorithms all employ scoring schemes to measure structural similarities and to maximize the structural similarities measured using various criteria. The algorithmic approaches to comparing protein geometric properties can be divided into three categories: the first superposes protein structures by minimizing intermolecular distances; the second relies on measuring intramolecular distances of a structure; and the third includes algorithms that combine both intermolecular and intramolecular approaches.

Intermolecular Method

The intermolecular approach is normally applied to relatively similar structures. To compare and superpose two protein structures, one of the structures has to be moved with respect to the other in such a way that the two structures have a maximum overlap in a three-dimensional space. This procedure starts with identifying equivalent residues or atoms. After residue–residue correspondence is established, one of the structures is moved laterally and vertically toward the other structure, a process known as *translation,* to allow the two structures to be in the same location (or same coordinate frame). The structures are further rotated relative to each other around the three-dimensional axes, during which process the distances between equivalent positions are constantly measured (Fig. 13.3). The rotation continues until the shortest intermolecular distance is reached. At this point, an optimal superimposition of the two structures is reached. After superimposition, equivalent residue pairs can be identified, which helps to quantitate the fitting between the two structures.

An important measurement of the structure fit during superposition is the distance between equivalent positions on the protein structures. This requires using a least-square-fitting function called *root mean square deviation* (RMSD), which is the square root of the averaged sum of the squared differences of the atomic distances.

$$\mathrm{RMSD} = \sqrt{\sum_{i=1}^{N} D_i^2 / N} \qquad \text{(Eq. 13.1)}$$

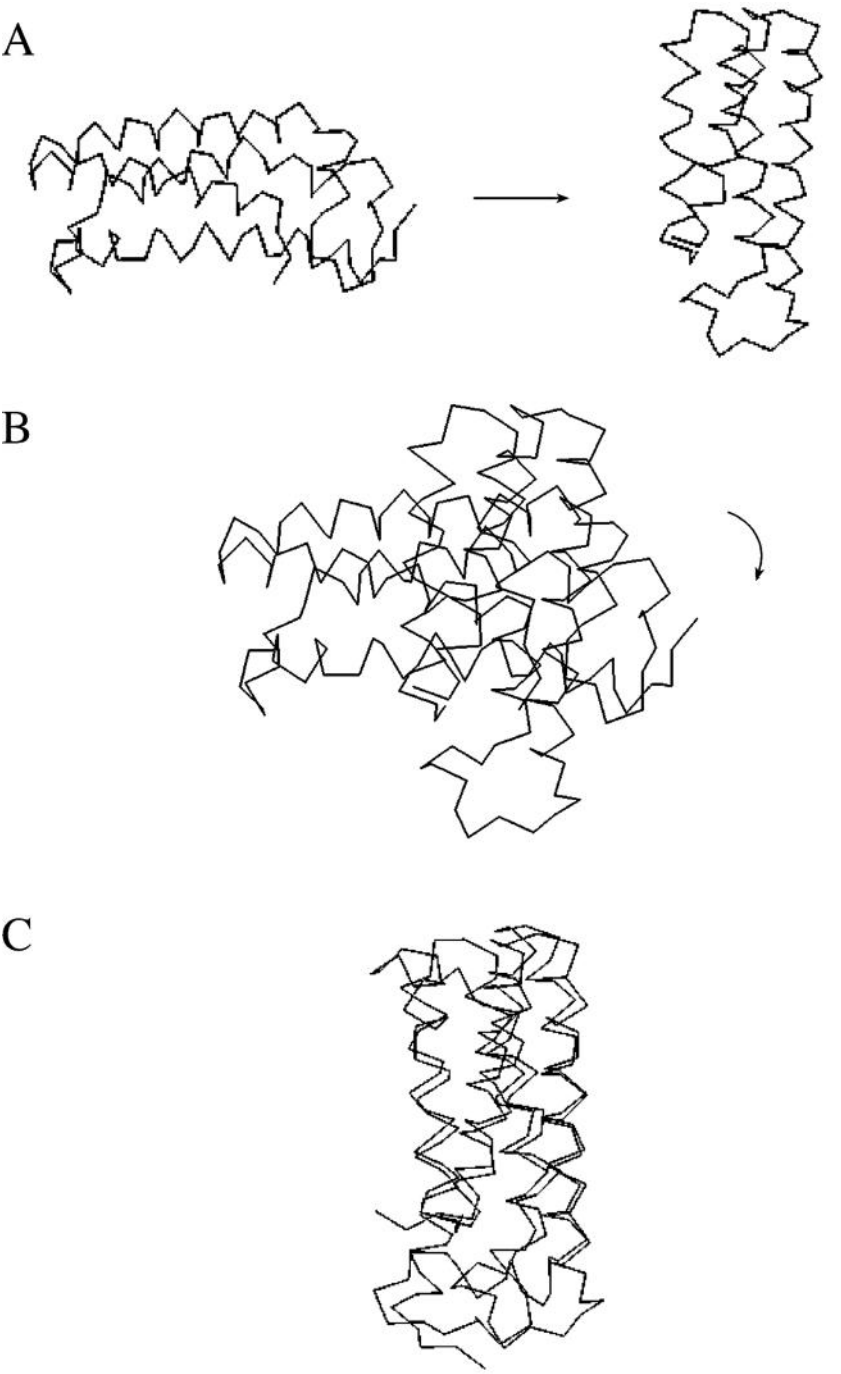


Figure 13.3: Simplified representation showing steps involved in the structure superposition of two protein molecules. **(A)** Two protein structures are positioned in different places in a three dimensional space. Equivalent positions are identified using a sequence based alignment approach. **(B)** To superimpose the two structures, the first step is to move one structure (*left*) relative to the other (*right*) through lateral and vertical movement, which is called translation. **(C)** The left structure is then rotated relative to the reference structure until such a point that the relative distances between equivalent positions are minimal.

where D is the distance between coordinate data points and N is the total number of corresponding residue pairs.

In practice, only the distances between Cα carbons of corresponding residues are measured. The goal of structural comparison is to achieve a minimum RMSD. However, the problem with RMSD is that it depends on the size of the proteins being compared. For the same degree of sequence identity, large proteins tend to have higher RMSD values than small proteins when an optimal alignment is reached. Recently, a logarithmic factor has been proposed to correct this size-dependency problem. This new measure is called RMSD_{100} and is determined by the following formula:

$$\text{RMSD}_{100} = \frac{\text{RMSD}}{-1.3 + 0.5\ln(\text{N})} \qquad \text{(Eq. 13.2)}$$

where N is the total number of corresponding atoms.

Although this corrected RMSD is more reliable than the raw RMSD for structure superposition, a low RMSD value by no means guarantees a correct alignment or an alignment with biological meaning. Careful scrutiny of the automatic alignment results is always recommended.

The most challenging part of using the intermolecular method is to identify equivalent residues in the first place, which often resorts to sequence alignment methods. Obviously, this restricts the usefulness of structural comparison between distant homologs.

A number of solutions have been proposed to compare more distantly related structures. One approach that has been proposed is to delete sequence variable regions outside secondary structure elements to reduce the search time required to find an optimum superposition. However, this method does not guarantee an optimal alignment. Another approach adopted by some researchers is to divide the proteins into small fragments (e.g., every six to nine residues). Matching of similar regions at the three-dimensional level is then done fragment by fragment. After finding the best fitting fragments, a joint superposition for the entire structure is performed. The third approach is termed *iterative optimization,* during which the two sequences are first aligned using dynamic programming. Identified equivalent residues are used to guide a first round of superposition. After superposition, more residues are identified to be in close proximity at the three-dimensional level and considered as equivalent residues. Based on the newly identified equivalent residues, a new round of superposition is generated to refine from the previous alignment. This procedure is repeated until the RMSD values cannot be further improved.

Intramolecular Method

The intramolecular approach relies on structural internal statistics and therefore does not depend on sequence similarity between the proteins to be compared. In addition, this method does not generate a physical superposition of structures, but instead provides a quantitative evaluation of the structural similarity between corresponding residue pairs.

The method works by generating a distance matrix between residues of the same protein. In comparing two protein structures, the distance matrices from the two structures are moved relative to each other to achieve maximum overlaps. By overlaying two distance matrices, similar intramolecular distance patterns representing similar structure folding regions can be identified. For the ease of comparison, each matrix is decomposed into smaller submatrices consisting of hexapeptide fragments. To maximize the similarity regions between two structures, a Monte Carlo procedure is used. By reducing three-dimensional information into two-dimensional information, this strategy identifies overall structural resemblances and common structure cores.

Combined Method

A recent development in structure comparison involves combining both inter- and intramolecular approaches. In the hybrid approach, corresponding residues can be identified using the intramolecular method. Subsequent structure superposition can be performed based on residue equivalent relationships. In addition to using RMSD as a measure during alignment, additional structural properties such as secondary structure types, torsion angles, accessibility, and local hydrogen bonding environment can be used. Dynamic programming is often employed to maximize overlaps in both inter- and intramolecular comparisons.

Multiple Structure Alignment

In addition to pairwise alignment, a number of algorithms can also perform multiple structure alignment. The alignment strategy is similar to the Clustal sequence alignment using a progressive approach (see Chapter 5). That is, all structures are first compared in a pairwise fashion. A distance matrix is developed based on structure similarity scores such as RMSD. This allows construction of a phylogenetic tree, which guides the subsequent clustering of the structures. The most similar two structures are then realigned. The aligned structures create a median structure that allows other structures to be progressively added for comparison based on the hierarchy described in the guide tree. When all the structures in the set are added, this eventually creates a multiple structure alignment. Several popular on-line structure comparison resources are discussed next.

DALI (www2.ebi.ac.uk/dali/) is a structure comparison web server that uses the intramolecular distance method. It works by maximizing the similarity of two distance graphs. The matrices are based on distances between all Cα atoms for each individual protein. Two distance matrices are overlaid and moved one relative to the other to identify most similar regions. DALI uses a statistical significance value called a Z-score to evaluate structural alignment. The Z-score is the number of standard deviations from the average score derived from the database background distribution. The higher the Z-score when comparing a pair of protein structures, the less likely the similarity

observed is a result of random chance. Empirically, a Z-score >4 indicates a significant level of structure similarity. The web server is at the same time a database that contains Z-scores of all precomputed structure pairs of proteins in PDB. The user can upload a structure to compare it with all known structures, or perform a pairwise comparison of two uploaded structures.

CE (Combinatorial Extension; http://cl.sdsc.edu/ce.html) is a web-based program that also uses the intramolecular distance approach. However, unlike DALI, a type of heuristics is used. In this method, every eight residues are treated as a single residue. The Cα distance matrices are constructed at the level of octameric "residues." In this way, the computational time required to search for the best alignment is considerably reduced, at the expense of alignment accuracy. CE also uses a Z-score as a measure of significance of an alignment. A Z-score >3.5 indicates a similar fold.

VAST (Vector Alignment Search Tool; www.ncbi.nlm.nih.gov:80/Structure/VAST/vast.shtml) is a web server that performs alignment using both the inter- and intramolecular approaches. The superposition is based on information of directionality of secondary structural elements (represented as vectors). Optimal alignment between two structures is defined by the highest degree of vector matches.

SSAP (www.biochem.ucl.ac.uk/cgi-bin/cath/GetSsapRasmol.pl) is a web server that uses an intramolecular distance–based method in which matrices are built based on the Cβ distances of all residue pairs. When comparing two different matrices, a dynamic programming approach is used to find the path of residue positions with optimal scores. The dynamic programming is applied at two levels, one at a lower level in which all residue pairs between the proteins are compared and another at an upper level in which subsequently identified equivalent residue pairs are processed to refine the matching positions. This process is known as double dynamic programming. An SSAP score is reported for structural similarity. A score above 70 indicates a good structural similarity.

STAMP (www.compbio.dundee.ac.uk/Software/Stamp/stamp.html) is a UNIX program that uses the intermolecular approach to generate protein structure alignment. The main feature is the use of iterative alignment based on dynamic programming to obtain the best superposition of two or more structures.

PROTEIN STRUCTURE CLASSIFICATION

One of the applications of protein structure comparison is structural classification. The ability to compare protein structures allows classification of the structure data and identification of relationships among structures. The reason to develop a protein structure classification system is to establish hierarchical relationships among protein structures and to provide a comprehensive and evolutionary view of known structures. Once a hierarchical classification system is established, a newly obtained protein structure can find its place in a proper category. As a result, its functions can be better understood based on association with other proteins. To date, several

systems have been developed, the two most popular being Structural Classification of Proteins (SCOP) and Class, Architecture, Topology and Homologous (CATH). The following introduces the basic steps in establishing the systems to classify proteins.

The first step in structure classification is to remove redundancy from databases. As mentioned in Chapter 12, among the tens of thousands of entries in PDB, the majority of the structures are redundant as they correspond to structures solved at different resolutions, or associated with different ligands or with single-residue mutations. The redundancy can be removed by selecting representatives through a sequence alignment–based approach. The second step is to separate structurally distinct domains within a structure. Because some proteins are composed of multiple domains, they must be subdivided before a sensible structural comparison can be carried out. This domain identification and separation can be done either manually or based on special algorithms for domain recognition. Once multidomain proteins are split into separate domains, structure comparison can be conducted at the domain level, either through manual inspection, or automated structural alignment, or a combination of both. The last step involves grouping proteins/domains of similar structures and clustering them based on different levels of resemblance in secondary structure composition and arrangement of the secondary structures in space.

As mentioned, the two most popular classification schemes are SCOP and CATH, both of which contain a number of hierarchical levels in their systems.

SCOP

SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) is a database for comparing and classifying protein structures. It is constructed almost entirely based on manual examination of protein structures. The proteins are grouped into hierarchies of classes, folds, superfamilies, and families. In the latest SCOP release version (v1.65, released December 2003), there are 7 classes, 800 folds, 1,294 superfamilies, and 2,327 families.

The SCOP families consist of proteins having high sequence identity (>30%). Thus, the proteins within a family clearly share close evolutionary relationships and normally have the same functionality. The protein structures at this level are also extremely similar. Superfamilies consist of families with similar structures, but weak sequence similarity. It is believed that members of the same superfamily share a common ancestral origin, although the relationships between families are considered distant. Folds consist of superfamilies with a common core structure, which is determined manually. This level describes similar overall secondary structures with similar orientation and connectivity between them. Members within the same fold do not always have evolutionary relationships. Some of the shared core structure may be a result of analogy. Classes consist of folds with similar core structures. This is at the highest level of the hierarchy, which distinguishes groups of proteins by secondary structure compositions such as all α, all β, α and β, and so on. Some classes are created based on general features such as membrane proteins, small proteins with few

secondary structures and irregular proteins. Folds within the same class are essentially randomly related in evolution.

CATH

CATH (www.biochem.ucl.ac.uk/bsm/cath_new/index.html) classifies proteins based on the automatic structural alignment program SSAP as well as manual comparison. Structural domain separation is carried out also as a combined effort of a human expert and computer programs. Individual domain structures are classified at five major levels: class, architecture, fold/topology, homologous superfamily, and homologous family. In the CATH release version 2.5.1 (January 2004), there are 4 classes, 37 architectures, 813 topologies, 1,467 homologous superfamilies, and 4,036 homologous families.

The definition for class in CATH is similar to that in SCOP, and is based on secondary structure content. Architecture is a unique level in CATH, intermediate between fold and class. This level describes the overall packing and arrangement of secondary structures independent of connectivity between the elements. The topology level is equivalent to the fold level in SCOP, which describes overall orientation of secondary structures and takes into account the sequence connectivity between the secondary structure elements. The homologous superfamily and homologous family levels are equivalent to the superfamily and family levels in SCOP with similar evolutionary definitions, respectively.

Comparison of SCOP and CATH

SCOP is almost entirely based on manual comparison of structures by human experts with no quantitative criteria to group proteins. It is argued that this approach offers some flexibility in recognizing distant structural relatives, because human brains may be more adept at recognizing slightly dissimilar structures that essentially have the same architecture. However, this reliance on human expertise also renders the method subjective. The exact boundaries between levels and groups are sometimes arbitrary.

CATH is a combination of manual curation and automated procedure, which makes the process less subjective. For example, in defining domains, CATH first relies on the consensus of three different algorithms to recognize domains. When the computer programs disagree, human intervention will take place. In addition, the extra Architecture level in CATH makes the structure classification more continuous. The drawback of the systems is that the fixed thresholds in structural comparison may make assignment less accurate.

Due to the differences in classification criteria, one might expect that there would be huge differences in classification results. In fact, the classification results from both systems are quite similar. Exhaustive analysis has shown that the results from the two systems converge at about 80% of the time. In other words, only about 20% of the structure fold assignments are different. Figure 13.4 shows two examples of agreement and disagreement based on classification by the two systems.

PDB code: 4tim

SCOP		CATH	
Class	Alpha and Beta (α/β)	*Class*	Alpha Beta
		Architecture	Barrel
Fold	TIM beta/alpha-barrel	*Topology*	TIM Barrel
Superfamily	Triosephosphate isomerase	*Homologous Superfamily*	Triosephosphate isomerase
Family	Triosephosphate isomerase	*Homologous Family*	Triosephosphate isomerase

PDB code: 1lys

SCOP		CATH	
Class	Alpha and Beta ($\alpha+\beta$)	*Class*	Mainly Alpha
		Architecture	Orthogonal Bundle
Fold	Lysozyme-like	*Topology*	Lysozyme
Superfamily	Lysozyme-like	*Homologous Superfamily*	Hydrolase (O-glycosyl)
Family	C-type lysozyme	*Homologous Family*	Hydrolase

Figure 13.4: Comparison of results of structure classification between SCOP and CATH. The classifications on the left is a case of overall agreement whereas the one on the right disagrees at the class level.

SUMMARY

A clear and concise visual representation of protein structures is the first step towards structural understanding. A number of visualization programs have been developed for that purpose. They include stand-alone programs for sophisticated manipulation of structures and light-weight web-based programs for simple structure viewing. Protein structure comparison allows recognition of distant evolutionary relationships among proteins and is helpful for structure classification and evaluation of protein structure prediction methods. The comparison algorithms fall into three categories: the intermolecular method, which involves transformation of atomic coordinates of structures to get optimal superimposition; the intramolecular method, which constructs an inter-residue distance matrix within a molecule and compares the matrix against that from a second molecule; and the combined method that uses both inter- and intramolecular approaches. Among all the structure comparison algorithms developed so far, DALI is most widely used. Protein structure classification is important for understanding protein structure, function and evolution. The most widely used classification schemes are SCOP and CATH. The two systems largely agree but differ somewhat. Each system has its own strengths and neither appears to be superior. It is thus advisable to compare the classification results from both systems in order to put a structure in the correct context.

FURTHER READING

Bourne, P. E., and Shindyalov, I. N. 2003. "Structure comparison and alignment." In *Structural Bioinformatics,* edited by P. E. Bourne and H. Weissig, 321–37. Hoboken, NJ: Wiley-Liss.

Carugo, O., and Pongor, S. 2002. Recent progress in protein 3D structure comparison. *Curr. Protein Pept. Sci.* 3:441–9.

Hadley, C., and Jones, D. T. 1999. A systematic comparison of protein structure classifications: SCOP, CATH, and FSSP. *Structure* 7:1099–112.

Jawad, Z., and Paoli, M. 2002. Novel sequences propel familiar folds. *Structure* 10:447–54.

Kinch, L. N., and Grishin, N. V. 2002. Evolution of protein structures and functions. *Curr. Opin. Struct. Biol*. 12:400–8.

Koehl, P. 2001. Protein structure similarities. *Curr. Opin. Struct. Biol.* 11:348–53.

Orengo, C. A., Pearl, F. M. G., and Thornton, J. M. 2003. "The CATH domain structure database." In *Structural Bioinformatics,* edited by P. E. Bourne and H. Weissig, 249–71. Hoboken, NJ: Wiley-Liss.

Ouzounis, C. A., Coulson, R. M., Enright, A. J., Kunin, V., and Pereira-Leal, J. B. 2003. Classification schemes for protein structure and function. *Nat. Rev. Genet*. 4:508–19.

Reddy, B. J. 2003. "Protein structure evolution and the scop database." In *Structural Bioinformatics,* edited by P. E. Bourne and H. Weissig, 239–48. Hoboken, NJ: Wiley-Liss.

Russell, R. B. 2002. Classification of protein folds. *Mol. Biotechnol*. 20:17–28.

Tate, J. 2003. "Molecular visualization." In *Structural Bioinformatics,* edited by P. E. Bourne and H. Weissig, 135–58. Hoboken, NJ: Wiley-Liss.

CHAPTER FOURTEEN

Protein Secondary Structure Prediction

Protein secondary structures are stable local conformations of a polypeptide chain. They are critically important in maintaining a protein three-dimensional structure. The highly regular and repeated structural elements include α-helices and β-sheets. It has been estimated that nearly 50% of residues of a protein fold into either α-helices and β-strands. As a review, an α-helix is a spiral-like structure with 3.6 amino acid residues per turn. The structure is stabilized by hydrogen bonds between residues i and $i+4$. Prolines normally do not occur in the middle of helical segments, but can be found at the end positions of α-helices (see Chapter 12). A β-sheet consists of two or more β-strands having an extended zigzag conformation. The structure is stabilized by hydrogen bonding between residues of adjacent strands, which actually may be long-range interactions at the primary structure level. β-Strands at the protein surface show an alternating pattern of hydrophobic and hydrophilic residues; buried strands tend to contain mainly hydrophobic residues.

Protein secondary structure prediction refers to the prediction of the conformational state of each amino acid residue of a protein sequence as one of the three possible states, namely, helices, strands, or coils, denoted as H, E, and C, respectively. The prediction is based on the fact that secondary structures have a regular arrangement of amino acids, stabilized by hydrogen bonding patterns. The structural regularity serves the foundation for prediction algorithms.

Predicting protein secondary structures has a number of applications. It can be useful for the classification of proteins and for the separation of protein domains and functional motifs. Secondary structures are much more conserved than sequences during evolution. As a result, correctly identifying secondary structure elements (SSE) can help to guide sequence alignment or improve existing sequence alignment of distantly related sequences. In addition, secondary structure prediction is an intermediate step in tertiary structure prediction as in threading analysis (see Chapter 15).

Because of significant structural differences between globular proteins and transmembrane proteins, they necessitate very different approaches to predicting respective secondary structure elements. Prediction methods for each of two types of proteins are discussed herein. In addition, prediction of supersecondary structures, such as coiled coils, is also described.

SECONDARY STRUCTURE PREDICTION FOR GLOBULAR PROTEINS

Protein secondary structure prediction with high accuracy is not a trivial ask. It remained a very difficult problem for decades. This is because protein secondary structure elements are context dependent. The formation of α-helices is determined by short-range interactions, whereas the formation of β-strands is strongly influenced by long-range interactions. Prediction for long-range interactions is theoretically difficult. After more than three decades of effort, prediction accuracies have only been improved from about 50% to about 75%.

The secondary structure prediction methods can be either ab initio based, which make use of single sequence information only, or homology based, which make use of multiple sequence alignment information. The ab initio methods, which belong to early generation methods, predict secondary structures based on statistical calculations of the residues of a single query sequence. The homology-based methods do not rely on statistics of residues of a single sequence, but on common secondary structural patterns conserved among multiple homologous sequences.

Ab Initio–Based Methods

This type of method predicts the secondary structure based on a single query sequence. It measures the relative propensity of each amino acid belonging to a certain secondary structure element. The propensity scores are derived from known crystal structures. Examples of ab initio prediction are the Chou–Fasman and Garnier, Osguthorpe, Robson (GOR) methods. The ab initio methods were developed in the 1970s when protein structural data were very limited. The statistics derived from the limited data sets can therefore be rather inaccurate. However, the methods are simple enough that they are often used to illustrate the basics of secondary structure prediction.

The Chou–Fasman algorithm (http://fasta.bioch.virginia.edu/fasta/chofas.htm) determines the propensity or intrinsic tendency of each residue to be in the helix, strand, and β-turn conformation using observed frequencies found in protein crystal structures (conformational values for coils are not considered). For example, it is known that alanine, glutamic acid, and methionine are commonly found in α-helices, whereas glycine and proline are much less likely to be found in such structures.

The calculation of residue propensity scores is simple. Suppose there are n residues in all known protein structures from which m residues are helical residues. The total number of alanine residues is y of which x are in helices. The propensity for alanine to be in helix is the ratio of the proportion of alanine in helices over the proportion of alanine in overall residue population (using the formula $[x/m]/[y/n]$). If the propensity for the residue equals 1.0 for helices ($P[\alpha\text{-helix}]$), it means that the residue has an equal chance of being found in helices or elsewhere. If the propensity ratio is less than 1, it indicates that the residue has less chance of being found in helices. If the propensity is larger than 1, the residue is more favored by helices. Based on this concept, Chou

TABLE 14.1. Relative Amino Acid Propensity Values for Secondary Structure Elements Used in the Chou–Fasman Method

Amino Acid	(α-Helix)	P (β-Strand)	P (Turn)
Alanine	1.42	0.83	0.66
Arginine	0.98	0.93	0.95
Asparagine	0.67	0.89	1.56
Aspartic acid	1.01	0.54	1.46
Cysteine	0.70	1.19	1.19
Glutamic acid	1.51	0.37	0.74
Glutamine	1.11	1.11	0.98
Glycine	0.57	0.75	1.56
Histidine	1.00	0.87	0.95
Isoleucine	1.08	1.60	0.47
Leucine	1.21	1.30	0.59
Lysine	1.14	0.74	1.01
Methionine	1.45	1.05	0.60
Phenylalanine	1.13	1.38	0.60
Proline	0.57	0.55	1.52
Serine	0.77	0.75	1.43
Threonine	0.83	1.19	0.96
Tryptophan	0.83	1.19	0.96
Tyrosine	0.69	1.47	1.14
Valine	1.06	1.70	0.50

and Fasman developed a scoring table listing relative propensities of each amino acid to be in an α-helix, a β-strand, or a β-turn (Table 14.1).

Prediction with the Chou–Fasman method works by scanning through a sequence with a certain window size to find regions with a stretch of contiguous residues each having a favored SSE score to make a prediction. For α-helices, the window size is six residues, if a region has four contiguous residues each having P(α-helix) > 1.0, it is predicted as an α-helix. The helical region is extended in both directions until the P(α-helix) score becomes smaller than 1.0. That defines the boundaries of the helix. For β-strands, scanning is done with a window size of five residues to search for a stretch of at least three favored β-strand residues. If both types of secondary structure predictions overlap in a certain region, a prediction is made based on the following criterion: if $\Sigma P(\alpha) > \Sigma P(\beta)$, it is declared as an α-helix; otherwise, a β-strand.

The GOR method (http://fasta.bioch.virginia.edu/fasta_www/garnier.htm) is also based on the "propensity" of each residue to be in one of the four conformational states, helix (H), strand (E), turn (T), and coil (C). However, instead of using the propensity value from a single residue to predict a conformational state, it takes short-range interactions of neighboring residues into account. It examines a window of every seventeen residues and sums up propensity scores for all residues for each of the four states resulting in four summed values. The highest scored state defines the conformational state for the center residue in the window (ninth position). The GOR method has

been shown to be more accurate than Chou–Fasman because it takes the neighboring effect of residues into consideration.

Both the Chou–Fasman and GOR methods, which are the first-generation methods developed in the 1970s, suffer from the fact that the prediction rules are somewhat arbitrary. They are based on single sequence statistics without clear relation to known protein-folding theories. The predictions solely rely on local sequence information and fail to take into account long range interactions. A Chou-Fasman–based prediction does not even consider the short-range environmental information. These reasons, combined with unreliable statistics derived from a very small structural database, limit the prediction accuracy of these methods to about 50%. This performance is considered dismal; any random prediction can have a 40% accuracy given the fact that, in globular proteins, the three-state distribution is 30% α-helix, 20% β-strands, and 50% coil.

Newer algorithms have since been developed to overcome some of these shortcomings. The improvements include more refined residue statistics based on a larger number of solved protein structures and the incorporation of more local residue interactions. Examples of the improved algorithms are GOR II, GOR III, GOR IV, and SOPM. These tools can be found at http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html. These are the second-generation prediction algorithms developed in the 1980s and early 1990s. They have improved accuracy over the first generation by about 10%. Although it is already significantly better than that by random prediction, the programs are still not reliable enough for routine application. Prediction errors mainly occur through missed β-strands and short-lengthed secondary structures for both helices and strands. Prediction of β-strands is still somewhat random. This may be attributed to the fact that long range interactions are not sufficiently taken into consideration in these algorithms.

Homology-Based Methods

The third generation of algorithms were developed in the late 1990s by making use of evolutionary information. This type of method combines the ab initio secondary structure prediction of individual sequences and alignment information from multiple homologous sequences (>35% identity). The idea behind this approach is that close protein homologs should adopt the same secondary and tertiary structure. When each individual sequence is predicted for secondary structure using a method similar to the GOR method, errors and variations may occur. However, evolutionary conservation dictates that there should be no major variations for their secondary structure elements. Therefore, by aligning multiple sequences, information of positional conservation is revealed. Because residues in the same aligned position are assumed to have the same secondary structure, any inconsistencies or errors in prediction of individual sequences can be corrected using a majority rule (Fig. 14.1). This homology-based method has helped improve the prediction accuracy by another 10% over the second-generation methods.

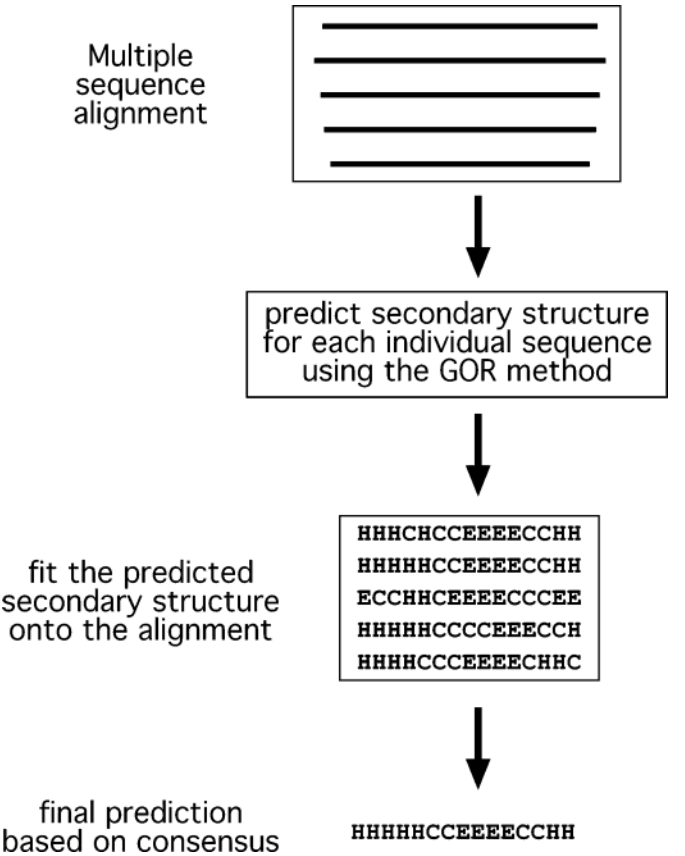


Figure 14.1: Schematic representation of secondary structure prediction using multiple sequence alignment information. Each individual sequence in the multiple alignment is subject to secondary structure prediction using the GOR method. If variations in predictions occur, they can be corrected by deriving a consensus of the secondary structure elements from the alignment.

Prediction with Neural Networks

The third-generation prediction algorithms also extensively apply sophisticated neural networks (see Chapter 8) to analyze substitution patterns in multiple sequence alignments. As a review, a *neural network* is a machine learning process that requires a structure of multiple layers of interconnected variables or nodes. In secondary structure prediction, the input is an amino acid sequence and the output is the probability of a residue to adopt a particular structure. Between input and output are many connected hidden layers where the machine learning takes place to adjust the mathematical weights of internal connections. The neural network has to be first trained by sequences with known structures so it can recognize the amino acid patterns and their relationships with known structures. During this process, the weight functions in hidden layers are optimized so they can relate input to output correctly. When the sufficiently trained network processes an unknown sequence, it applies the rules learned in training to recognize particular structural patterns.

When multiple sequence alignments and neural networks are combined, the result is further improved accuracy. In this situation, a neural network is trained not by a single sequence but by a sequence profile derived from the multiple sequence alignment. This combined approach has been shown to improve the accuracy to above 75%, which is a breakthrough in secondary structure prediction. The improvement mainly comes from enhanced secondary structure signals through consensus drawing. The following lists several frequently used third generation prediction algorithms available as web servers.

PHD (Profile network from Heidelberg; http://dodo.bioc.columbia.edu/predictprotein/submit_def.html) is a web-based program that combines neural network with multiple sequence alignment. It first performs a BLASTP of the query sequence against a nonredundant protein sequence database to find a set of homologous sequences, which are aligned with the MAXHOM program (a weighted dynamic programming algorithm performing global alignment). The resulting alignment in the form of a profile is fed into a neural network that contains three hidden layers. The first hidden layer makes raw prediction based on the multiple sequence alignment by sliding a window of thirteen positions. As in GOR, the prediction is made for the residue in the center of the window. The second layer refines the raw prediction by sliding a window of seventeen positions, which takes into account more flanking positions. This step makes adjustments and corrections of unfeasible predictions from the previous step. The third hidden layer is called the *jury network,* and contains networks trained in various ways. It makes final filtering by deleting extremely short helices (one or two residues long) and converting them into coils (Fig. 14.2). After the correction, the highest scored state defines the conformational state of the residue.

PSIPRED (http://bioinf.cs.ucl.ac.uk/psiform.html) is a web-based program that predicts protein secondary structures using a combination of evolutionary information and neural networks. The multiple sequence alignment is derived from a PSI-BLAST database search. A profile is extracted from the multiple sequence alignment generated from three rounds of automated PSI-BLAST. The profile is then used as input for a neural network prediction similar to that in PHD, but without the jury layer. To achieve higher accuracy, a unique filtering algorithm is implemented to filter out unrelated PSI-BLAST hits during profile construction.

SSpro (http://promoter.ics.uci.edu/BRNN-PRED/) is a web-based program that combines PSI-BLAST profiles with an advanced neural network, known as *bidirectional recurrent neural networks* (BRNNs). Traditional neural networks are unidirectional, feed-forward systems with the information flowing in one direction from input to output. BRNNs are unique in that the connections of layers are designed to be able to go backward. In this process, known as *back propagation,* the weights in hidden layers are repeatedly refined. In predicting secondary structure elements, the network uses the sequence profile as input and finds residue correlations by iteratively recycling the network (recursive network). The averaged output from the iterations is given as a final residue prediction. PROTER (http://distill.ucd.ie/porter/) is a recently developed program that uses similar BRNNs and has been shown to slightly outperform SSPRO.

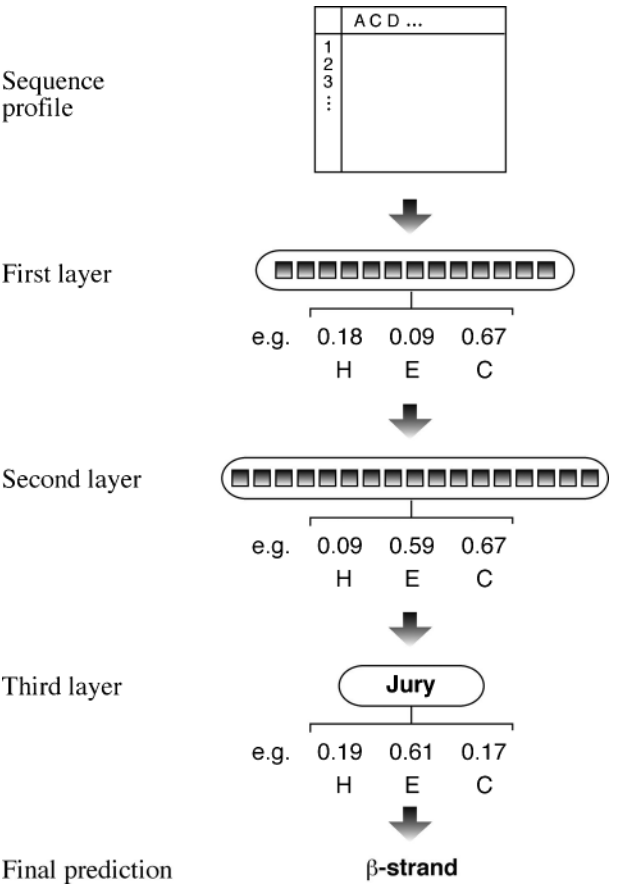


Figure 14.2: Schematic representation of secondary structure prediction in the PHD algorithm using neural networks. Multiple sequences derived from the BLAST search are used to compile a profile. The resulting profile is fed into a neural network, which contains three layers – two network layers and one jury layer. The first layer scans thirteen residues per window and makes a raw prediction, which is refined by the second layer, which scans seventeen residues per window. The third layer makes further adjustment to make a final prediction. Adjustment of prediction scores for one amino acid residue is shown.

PROF (Protein forecasting; www.aber.ac.uk/~phiwww/prof/) is an algorithm that combines PSI-BLAST profiles and a multistaged neural network, similar to that in PHD. In addition, it uses a linear discriminant function to discriminate between the three states.

HMMSTR (Hidden Markov model [HMM] for protein STRuctures; www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php) uses a branched and cyclic HMM to predict secondary structures. It first breaks down the query sequence into many very short segments (three to nine residues, called I-sites) and builds profiles based on a library of known structure motifs. It then assembles these local motifs into a supersecondary structure. It further uses an HMM with a unique topology linking many smaller HMMs into a highly branched multicyclic form. This is intended to better capture the recurrent local features of secondary structure based on multiple sequence alignment.

Prediction with Multiple Methods

Because no individual methods can always predict secondary structures correctly, it is desirable to combine predictions from multiple programs with the hope of further improving the accuracy. In fact, a number of web servers have been specifically dedicated to making predictions by drawing consensus from results by multiple programs. In many cases, the consensus-based prediction method has been shown to perform slightly better than any single method.

Jpred (www.compbio.dundee.ac.uk/~www-jpred/) combines the analysis results from six prediction algorithms, including PHD, PREDATOR, DSC, NNSSP, Jnet, and ZPred. The query sequence is first used to search databases with PSI-BLAST for three iterations. Redundant sequence hits are removed. The resulting sequence homologs are used to build a multiple alignment from which a profile is extracted. The profile information is submitted to the six prediction programs. If there is sufficient agreement among the prediction programs, the majority of the prediction is taken as the structure. Where there is no majority agreement in the prediction outputs, the PHD prediction is taken.

PredictProtein (www.embl-heidelberg.de/predictprotein/predictprotein.html) is another multiple prediction server that uses Jpred, PHD, PROF, and PSIPRED, among others. The difference is that the server does not run the individual programs but sends the query to other servers which e-mail the results to the user separately. It does not generate a consensus. It is up to the user to combine multiple prediction results and derive a consensus.

Comparison of Prediction Accuracy

An important issue in protein secondary structure prediction is estimation of the prediction accuracy. The most commonly used measure for cross-validation is known as a Q_3 score, based on the three-state classification, helix (H), strand (E), and coil (C). The score is a percentage of residues of a protein that are correctly predicted. It is normally derived from the average result obtained from the testing with many proteins with known structures. For secondary structure prediction, there are well-established benchmarks for such prediction evaluation. By using these benchmarks, accuracies for several third-generation prediction algorithms have been compiled (Table 14.2).

As shown in Table 14.2, some of these best prediction methods have reached an accuracy level around 79% in the three-state prediction. Common errors include the confusion of helices and strands, incorrect start and end positions of helices and strands, and missed or wrongly assigned secondary structure elements. If a prediction is consistently 79% accurate, that means on average 21% of the residues could be predicted incorrectly.

Because different secondary structure prediction programs tend to give varied results, to maximize the accuracy of prediction, it is recommended to use several most robust prediction methods (such as Porter, PROF, and SSPRO) and draw a consensus based on the majority rule. The aforementioned metaservers provide a convenient

TABLE 14.2. Comparison of Accuracy of Some of the State-of-the-Art Secondary Structure Prediction Tools

Methods	Q_3 (%)
Porter	79.0
SSPro2	78.0
PROF	77.0
PSIPRED	76.6
Pred2ary	75.9
Jpred2	75.2
PHDpsi	75.1
Predator	74.8
HMMSTR	74.3

Note: The Q_3 score is the three-state prediction accuracy for helix, strand, and coil.

way of achieving this goal. By using the combination approach, it is possible to reach an 80% accuracy. An accuracy of 80% is an important landmark because it is equivalent to some low-resolution experimental methods to determine protein secondary structures, such as circular dichroism and Fourier transform-induced spectroscopy.

SECONDARY STRUCTURE PREDICTION FOR TRANSMEMBRANE PROTEINS

Transmembrane proteins constitute up to 30% of all cellular proteins. They are responsible for performing a wide variety of important functions in a cell, such as signal transduction, cross-membrane transport, and energy conversion. The membrane proteins are also of tremendous biomedical importance, as they often serve as drug targets for pharmaceutical development.

There are two types of integral membrane proteins: α-helical type and β-barrel type. Most transmembrane proteins contain solely α-helices, which are found in the cytoplasmic membrane. A few membrane proteins consist of β-strands forming a β-barrel topology, a cylindrical structure composed of antiparallel β-sheets. They are normally found in the outer membrane of gram-negative bacteria.

The structures of this group of proteins, however, are notoriously difficult to resolve either by x-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. Consequently, for this group of proteins, prediction of the transmembrane secondary structural elements and their organization is particularly important. Fortunately, the prediction process is somewhat easier because of the hydrophobic environment of the lipid bilayers, which restricts the transmembrane segments to be hydrophobic as well. In principle, the secondary structure prediction programs developed for soluble proteins can apply to membrane proteins as well. However, they normally do not work well in reality because the extra hydrophobicity and length requirements distort the

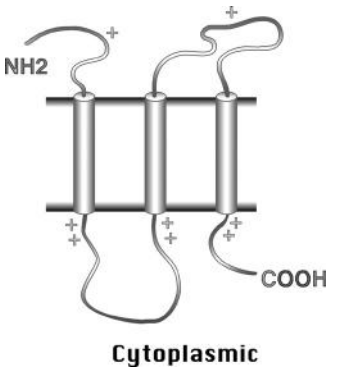


Figure 14.3: Schematic of the positive-inside rule for the orientation of membrane helices. The cylinders represent the transmembrane α–helices. There are relatively more positive charges near the helical anchor on the cytoplasmic side than on the periplasmic side.

statistical propensity of the residues. Thus, dedicated algorithms have to be used for transmembrane span predictions.

Prediction of Helical Membrane Proteins

For membrane proteins consisting of transmembrane α–helices, these transmembrane helices are predominantly hydrophobic with a specific distribution of positively charged residues. The α-helices generally run perpendicular to the membrane plane with an average length between seventeen and twenty-five residues. The hydrophobic helices are normally separated by hydrophilic loops with average lengths of fewer than sixty residues. The residues bordering the transmembrane spans are more positively charged. Another feature indicative of the presence of transmembrane segments is that residues at the cytosolic side near the hydrophobic anchor are more positively charged than those at the lumenal or periplasmic side. This is known as the *positive-inside rule* (Fig. 14.3), which allows the prediction of the orientation of the secondary structure elements. These rules form the basis for transmembrane prediction algorithms.

A number of algorithms for identifying transmembrane helices have been developed. The early algorithms based their prediction on hydrophobicity scales. They typically scan a window of seventeen to twenty-five residues and assign membrane spans based on hydrophobicity scores. Some are also able to determine the orientation of the membrane helices based on the positive-inside rule. However, predictions solely based on hydrophobicity profiles have high error rates. As with the third-generation predictions for globular proteins, applying evolutionary information with the help of neural networks or HMMs can improve the prediction accuracy significantly.

As mentioned, predicting transmembrane helices is relatively easy. The accuracy of some of the best predicting programs, such as TMHMM or HMMTOP, can exceed 70%. However, the presence of hydrophobic signal peptides can significantly compromise the prediction accuracy because the programs tend to confuse hydrophobic signal peptides with membrane helices. To minimize errors, the presence of signal peptides

can be detected using a number of specialized programs (see Chapter 18) and then manually excluded.

TMHMM (www.cbs.dtu.dk/services/TMHMM/) is a web-based program based on an HMM algorithm. It is trained to recognize transmembrane helical patterns based on a training set of 160 well-characterized helical membrane proteins. When a query sequence is scanned, the probability of having an α-helical domain is given. The orientation of the α-helices is predicted based on the positive-inside rule. The prediction output returns the number of transmembrane helices, the boundaries of the helices, and a graphical representation of the helices. This program can also be used to simply distinguish between globular proteins and membrane proteins.

Phobius (http://phobius.cgb.ki.se/index.html) is a web-based program designed to overcome false positives caused by the presence of signal peptides. The program incorporates distinct HMM models for signal peptides as well as transmembrane helices. After distinguishing the putative signal peptides from the rest of the query sequence, prediction is made on the remainder of the sequence. It has been shown that the prediction accuracy can be significantly improved compared to TMHMM (94% by Phobius compared to 70% by TMHMM). In addition to the normal prediction mode, the user can also define certain sequence regions as signal peptides or other nonmembrane sequences based on external knowledge. As a further step to improve accuracy, the user can perform the "poly prediction" with the PolyPhobius module, which searches the NCBI database for homologs of the query sequence. Prediction for the multiple homologous sequences help to derive a consensus prediction. However, this option is also more time consuming.

Prediction of β-Barrel Membrane Proteins

For membrane proteins with β-strands only, the β-strands forming the transmembrane segment are amphipathic in nature. They contain ten to twenty-two residues with every second residue being hydrophobic and facing the lipid bilayers whereas the other residues facing the pore of the β-barrel are more hydrophilic. Obviously, scanning a sequence by hydrophobicity does not reveal transmembrane β-strands. These programs for predicting transmembrane α-helices are not applicable for this unique type of membrane proteins. To predict the β-barrel type of membrane proteins, a small number of algorithms have been made available based on neural networks and related techniques.

TBBpred (www.imtech.res.in/raghava/tbbpred/) is a web server for predicting transmembrane β-barrel proteins. It uses a neural network approach to predict transmembrane β-barrel regions. The network is trained with the known structural information of a limited number of transmembrane β-barrel protein structures. The algorithm contains a single hidden layer with five nodes and a single output node. In addition to neural networks, the server can also predict using a support vector machine (SVM) approach, another type of statistical learning process. Similar to

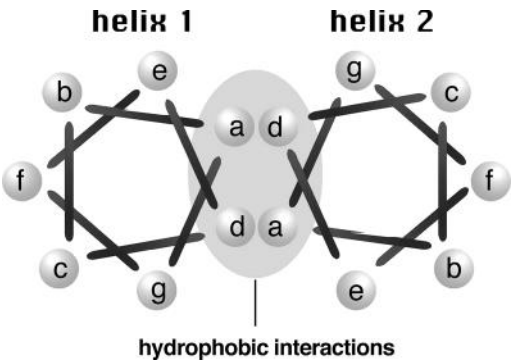


Figure 14.4: Cross-section view of a coiled coil structure. A coiled coil protein consisting of two interacting helical strands is viewed from top. The bars represent covalent bonds between amino acid residues. There is no covalent bond between residue *a* and *g*. The bar connecting the two actually means to connect the first residue of the next heptad. The coiled coil has a repeated seven residue motif in the form of *a-b-c-d-e-f-g*. The first and fourth positions (*a* and *d*) are hydrophobic, whose interactions with corresponding residues in another helix stabilize the structure. The positions *b*, *c*, *e*, *f*, *g* are hydrophilic and are exposed on the surface of the protein.

neural networks, in SVM the data are fed into kernels (similar to nodes), which are separated into different classes by a "hyperplane" (an abstract linear or nonlinear separator) according to a particular mathematical function. It has the advantage over neural networks in that it is faster to train and more resistant to noise. For more detailed information of SVM, see Chapter 19.

COILED COIL PREDICTION

Coiled coils are superhelical structures involving two to more interacting α-helices from the same or different proteins. The individual α-helices twist and wind around each other to form a coiled bundle structure. The coiled coil conformation is important in facilitating inter- or intraprotein interactions. Proteins possessing these structural domains are often involved in transcription regulation or in the maintenance of cytoskeletal integrity.

Coiled coils have an integral repeat of seven residues (heptads) which assume a side-chain packing geometry at facing residues (see Chapter 12). For every seven residues, the first and fourth are hydrophobic, facing the helical interface; the others are hydrophilic and exposed to the solvent (Fig. 14.4). The sequence periodicity forms the basis for designing algorithms to predict this important structural domain. As a result of the regular structural features, if the location of coiled coils can be predicted precisely, the three-dimensional structure for the coiled coil region can sometimes be built. The following lists several widely used programs for the specialized prediction.

Coils (www.ch.embnet.org/software/COILS_form.html) is a web-based program that detects coiled coil regions in proteins. It scans a window of fourteen, twenty-one, or twenty-eight residues and compares the sequence to a probability matrix

compiled from known parallel two-stranded coiled coils. By comparing the similarity scores, the program calculates the probability of the sequence to adopt a coiled coil conformation. The program is accurate for solvent-exposed, left-handed coiled coils, but less sensitive for other types of coiled coil structures, such as buried or right-handed coiled coils.

Multicoil (http://jura.wi.mit.edu/cgi-bin/multicoil/multicoil.pl) is a web-based program for predicting coiled coils. The scoring matrix is constructed based on a database of known two-stranded and three-stranded coiled coils. The program is more conservative than Coils. It has been recently used in several genome-wide studies to screen for protein–protein interactions mediated by coiled coil domains.

Leucine zipper domains are a special type of coiled coils found in transcription regulatory proteins. They contain two antiparallel α-helices held together by hydrophobic interactions of leucine residues. The heptad repeat pattern is L-X(6)-L-X(6)-L–X(6)-L. This repeat pattern alone can sometimes allow the domain detection, albeit with high rates of false positives. The reason for the high false-positive rates is that the condition of the sequence region being a coiled coil conformation is not satisfied. To address this problem, algorithms have been developed that take into account both leucine repeats and coiled coil conformation to give accurate prediction.

2ZIP (http://2zip.molgen.mpg.de/) is a web-based server that predicts leucine zippers. It combines searching of the characteristic leucine repeats with coiled coil prediction using an algorithm similar to Coils to yield accurate results.

SUMMARY

Protein secondary structure prediction has a long history and is defined by three generations of development. The first generation algorithms were ab initio based, examining residue propensities that fall in the three states: helices, strands, and coils. The propensities were derived from a very small structural database. The growing structural database and use of residue local environment information allowed the development of the second-generation algorithms. A major breakthrough came from the third-generation algorithms that make use of multiple sequence alignment information, which implicitly takes the long-range intraprotein interactions into consideration. In combination with neural networks and other sophisticated algorithms, prediction efficiency has been improved significantly. To achieve high accuracy in prediction, combining results from several top-performing third-generation algorithms is recommended. Predicting secondary structures for membrane proteins is more common than for globular proteins as crystal or NMR structures are extremely difficult to obtain for the former. The prediction of transmembrane segments (mainly α-helices) involves the use of hydrophobicity, neural networks, and evolutionary information. Coiled coils are a distinct type of supersecondary structure with regular periodicity of hydrophobic residues that can be predicted using specialized algorithms.

FURTHER READING

Edwards, Y. J., and Cottage, A. 2003. Bioinformatics methods to predict protein structure and function. A practical approach. *Mol. Biotechnol*. 23:139-66.

Heringa J. 2002. Computational methods for protein secondary structure prediction using multiple sequence alignments. *Curr. Protein Pept. Sci*. 1:273-301.

Lehnert, U., Xia, Y., Royce, T. E., Goh, C. S., Liu, Y., Senes, A., Yu. H., Zhang, Z. L., Engelman, D. M, and Gerstein M. 2004. Computational analysis of membrane proteins: Genomic occurrence, structure prediction and helix interactions. *Q. Rev. Biophys*. 37:121-46.

Möller, S., Croning, M. D. R., and Apweiler, R. 2001. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17:646-53.

Przybylski, D., and Rost, B. 2002. Alignments grow, secondary structure prediction improves. *Proteins* 46:197–205.

Rost B. 2001. Review: Protein secondary structure prediction continues to rise. *J. Struct. Biol.* 134:204–18.

———. 2003. Prediction in 1D: Secondary structure, membrane helices, and accessibility. In *Structural Bioinformatics,* edited by P. E. Bourne and H. Weissig, 559–87. Hoboken, NJ: Wiley–Liss.

CHAPTER FIFTEEN

Protein Tertiary Structure Prediction

One of the most important scientific achievements of the twentieth century was the discovery of the DNA double helical structure by Watson and Crick in 1953. Strictly speaking, the work was the result of a three-dimensional modeling conducted partly based on data obtained from x-ray diffraction of DNA and partly based on chemical bonding information established in stereochemistry. It was clear at the time that the x-ray data obtained by their colleague Rosalind Franklin were not sufficient to resolve the DNA structure. Watson and Crick conducted one of the first-known ab initio modeling of a biological macromolecule, which has subsequently been proven to be essentially correct. Their work provided great insight into the mechanism of genetic inheritance and paved the way for a revolution in modern biology. The example demonstrates that structural prediction is a powerful tool to understand the functions of biological macromolecules at the atomic level.

We now know that the DNA structure, a double helix, is rather invariable regardless of sequence variations. Although there is little need today to determine or model DNA structures of varying sequences, there is still a real need to model protein structures individually. This is because protein structures vary depending on the sequences. Another reason is the much slower rate of structure determination by x-ray crystallography or NMR spectroscopy compared to gene sequence generation from genomic studies. Consequently, the gap between protein sequence information and protein structural information is increasing rapidly. Protein structure prediction aims to reduce this sequence–structure gap.

In contrast to sequencing techniques, experimental methods to determine protein structures are time consuming and limited in their approach. Currently, it takes 1 to 3 years to solve a protein structure. Certain proteins, especially membrane proteins, are extremely difficult to solve by x-ray or NMR techniques. There are many important proteins for which the sequence information is available, but their three-dimensional structures remain unknown. The full understanding of the biological roles of these proteins requires knowledge of their structures. Hence, the lack of such information hinders many aspects of the analysis, ranging from protein function and ligand binding to mechanisms of enzyme catalysis. Therefore, it is often necessary to obtain approximate protein structures through computer modeling.

Having a computer-generated three-dimensional model of a protein of interest has many ramifications, assuming it is reasonably correct. It may be of use for the rational design of biochemical experiments, such as site-directed mutagenesis, protein stability, or functional analysis. In addition to serving as a theoretical guide to

design experiments for protein characterization, the model can help to rationalize the experimental results obtained with the protein of interest. In short, the modeling study helps to advance our understanding of protein functions.

METHODS

There are three computational approaches to protein three-dimensional structural modeling and prediction. They are homology modeling, threading, and ab initio prediction. The first two are knowledge-based methods; they predict protein structures based on knowledge of existing protein structural information in databases. Homology modeling builds an atomic model based on an experimentally determined structure that is closely related at the sequence level. Threading identifies proteins that are structurally similar, with or without detectable sequence similarities. The ab initio approach is simulation based and predicts structures based on physicochemical principles governing protein folding without the use of structural templates.

HOMOLOGY MODELING

As the name suggests, *homology modeling* predicts protein structures based on sequence homology with known structures. It is also known as *comparative modeling*. The principle behind it is that if two proteins share a high enough sequence similarity, they are likely to have very similar three-dimensional structures. If one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a high degree of confidence. Homology modeling produces an all-atom model based on alignment with template proteins.

The overall homology modeling procedure consists of six steps. The first step is template selection, which involves identification of homologous sequences in the protein structure database to be used as templates for modeling. The second step is alignment of the target and template sequences. The third step is to build a framework structure for the target protein consisting of main chain atoms. The fourth step of model building includes the addition and optimization of side chain atoms and loops. The fifth step is to refine and optimize the entire model according to energy criteria. The final step involves evaluating of the overall quality of the model obtained (Fig. 15.1). If necessary, alignment and model building are repeated until a satisfactory result is obtained.

Template Selection

The first step in protein structural modeling is to select appropriate structural templates. This forms the foundation for rest of the modeling process. The template selection involves searching the Protein Data Bank (PDB) for homologous proteins with determined structures. The search can be performed using a heuristic pairwise alignment search program such as BLAST or FASTA. However, the use of dynamic programming based search programs such as SSEARCH or ScanPS (see Chapter 4)

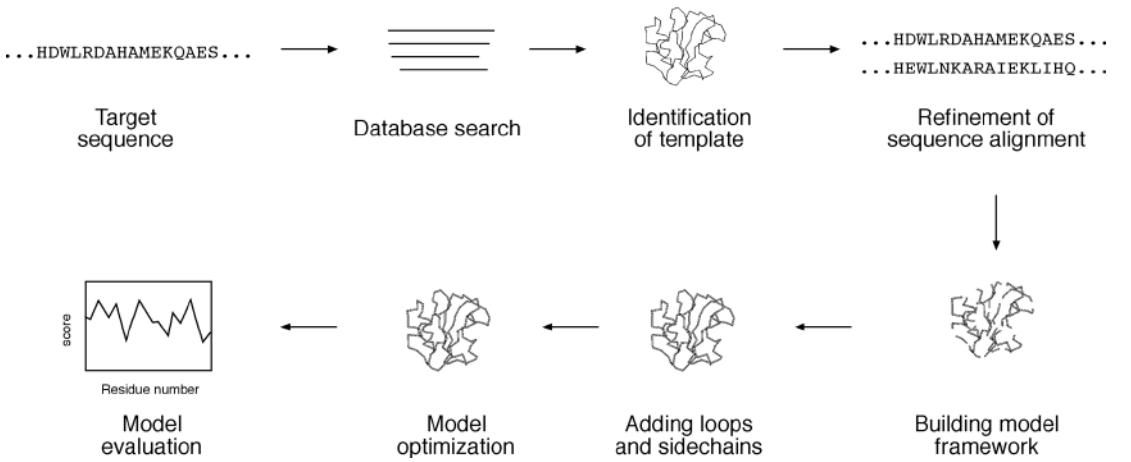


Figure 15.1: Flowchart showing steps involved in homology modeling.

can result in more sensitive search results. The relatively small size of the structural database means that the search time using the exhaustive method is still within reasonable limits, while giving a more sensitive result to ensure the best possible similarity hits.

As a rule of thumb, a database protein should have at least 30% sequence identity with the query sequence to be selected as template. Occasionally, a 20% identity level can be used as threshold as long as the identity of the sequence pair falls within the "safe zone" (see Chapter 3). Often, multiple database structures with significant similarity can be found as a result of the search. In that case, it is recommended that the structure(s) with the highest percentage identity, highest resolution, and the most appropriate cofactors is selected as a template. On the other hand, there may be a situation in which no highly similar sequences can be found in the structure database. In that instance, template selection can become difficult. Either a more sensitive profile-based PSI-BLAST method or a fold recognition method such threading can be used to identify distant homologs. Most likely, in such a scenario, only local similarities can be identified with distant homologs. Modeling can therefore only be done with the aligned domains of the target protein.

Sequence Alignment

Once the structure with the highest sequence similarity is identified as a template, the full-length sequences of the template and target proteins need to be realigned using refined alignment algorithms to obtain optimal alignment. This realignment is the most critical step in homology modeling, which directly affects the quality of the final model. This is because incorrect alignment at this stage leads to incorrect designation of homologous residues and therefore to incorrect structural models. Errors made in the alignment step cannot be corrected in the following modeling steps. Therefore, the best possible multiple alignment algorithms, such as Praline and T-Coffee (see Chapter 5), should be used for this purpose. Even alignment using the best alignment

program may not be error free and should be visually inspected to ensure that conserved key residues are correctly aligned. If necessary, manual refinement of the alignment should be carried out to improve alignment quality.

Backbone Model Building

Once optimal alignment is achieved, residues in the aligned regions of the target protein can assume a similar structure as the template proteins, meaning that the coordinates of the corresponding residues of the template proteins can be simply copied onto the target protein. If the two aligned residues are identical, coordinates of the side chain atoms are copied along with the main chain atoms. If the two residues differ, only the backbone atoms can be copied. The side chain atoms are rebuilt in a subsequent procedure.

In backbone modeling, it is simplest to use only one template structure. As mentioned, the structure with the best quality and highest resolution is normally chosen if multiple options are available. This structure tends to carry the fewest errors. Occasionally, multiple template structures are available for modeling. In this situation, the template structures have to be optimally aligned and superimposed before being used as templates in model building. One can either choose to use average coordinate values of the templates or the best parts from each of the templates to model.

Loop Modeling

In the sequence alignment for modeling, there are often regions caused by insertions and deletions producing gaps in sequence alignment. The gaps cannot be directly modeled, creating "holes" in the model. Closing the gaps requires loop modeling, which is a very difficult problem in homology modeling and is also a major source of error. Loop modeling can be considered a mini–protein modeling problem by itself. Unfortunately, there are no mature methods available that can model loops reliably. Currently, there are two main techniques used to approach the problem: the database searching method and the ab initio method.

The database method involves finding "spare parts" from known protein structures in a database that fit onto the two stem regions of the target protein. The stems are defined as the main chain atoms that precede and follow the loop to be modeled. The procedure begins by measuring the orientation and distance of the anchor regions in the stems and searching PDB for segments of the same length that also match the above endpoint conformation. Usually, many different alternative segments that fit the endpoints of the stems are available. The best loop can be selected based on sequence similarity as well as minimal steric clashes with the neighboring parts of the structure. The conformation of the best matching fragments is then copied onto the anchoring points of the stems (Fig. 15.2). The ab initio method generates many random loops and searches for the one that does not clash with nearby side chains and also has reasonably low energy and ϕ and ψ angles in the allowable regions in the Ramachandran plot.

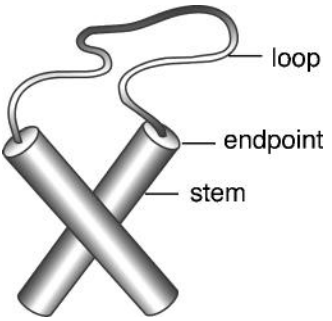


Figure 15.2: Schematic of loop modeling by fitting a loop structure onto the endpoints of existing stem structures represented by cylinders.

If the loops are relatively short (three to five residues), reasonably correct models can be built using either of the two methods. If the loops are longer, it is very difficult to achieve a reliable model. The following are specialized programs for loop modeling.

FREAD (www-cryst.bioc.cam.ac.uk/cgi-bin/coda/fread.cgi) is a web server that models loops using the database approach.

PETRA (www-cryst.bioc.cam.ac.uk/cgi-bin/coda/pet.cgi) is a web server that uses the ab initio method to model loops.

CODA (www-cryst.bioc.cam.ac.uk/~charlotte/Coda/search_coda.html) is a web server that uses a consensus method based on the prediction results from FREAD and PETRA. For loops of three to eight residues, it uses consensus conformation of both methods and for nine to thirty residues, it uses FREAD prediction only.

Side Chain Refinement

Once main chain atoms are built, the positions of side chains that are not modeled must be determined. Modeling side chain geometry is very important in evaluating protein–ligand interactions at active sites and protein–protein interactions at the contact interface.

A side chain can be built by searching every possible conformation at every torsion angle of the side chain to select the one that has the lowest interaction energy with neighboring atoms. However, this approach is computationally prohibitive in most cases. In fact, most current side chain prediction programs use the concept of *rotamers*, which are favored side chain torsion angles extracted from known protein crystal structures. A collection of preferred side chain conformations is a rotamer library in which the rotamers are ranked by their frequency of occurrence. Having a rotamer library reduces the computational time significantly because only a small number of favored torsion angles are examined. In prediction of side chain conformation, only the possible rotamers with the lowest interaction energy with nearby atoms are selected.

In many cases, even applying the rotamer library for every residue can be computationally too expensive. To reduce search time further, backbone conformation can be taken into account. It has been observed that there is a correlation of backbone conformations with certain rotamers. By using such correlations, many possible rotamers can be eliminated and the speed of conformational search can be much

improved. After adding the most frequently occurring rotamers, the conformations have to be further optimized to minimize steric overlaps with the rest of the model structure.

Most modeling packages incorporate the side chain refinement function. A specialized side chain modeling program that has reasonably good performance is SCWRL (sidechain placement with a rotamer library; www.fccc.edu/research/labs/dunbrack/scwrl/), a UNIX program that works by placing side chains on a backbone template according to preferences in the backbone-dependent rotamer library. It removes rotamers that have steric clashes with main chain atoms. The final, selected set of rotamers has minimal clashes with main chain atoms and other side chains.

Model Refinement Using Energy Function

In these loop modeling and side chain modeling steps, potential energy calculations are applied to improve the model. However, this does not guarantee that the entire raw homology model is free of structural irregularities such as unfavorable bond angles, bond lengths, or close atomic contacts. These kinds of structural irregularities can be corrected by applying the energy minimization procedure on the entire model, which moves the atoms in such a way that the overall conformation has the lowest energy potential. The goal of energy minimization is to relieve steric collisions and strains without significantly altering the overall structure.

However, energy minimization has to be used with caution because excessive energy minimization often moves residues away from their correct positions. Therefore, only limited energy minimization is recommended (a few hundred iterations) to remove major errors, such as short bond distances and close atomic clashes. Key conserved residues and those involved in cofactor binding have to be restrained if necessary during the process.

Another often used structure refinement procedure is molecular dynamic simulation. This practice is derived from the concern that energy minimization only moves atoms toward a local minimum without searching for all possible conformations, often resulting in a suboptimal structure. To search for a global minimum requires moving atoms uphill as well as downhill in a rough energy landscape. This requires thermodynamic calculations of the atoms. In this process, a protein molecule is "heated" or "cooled" to simulate the uphill and downhill molecular motions. Thus, it helps overcome energy hurdles that are inaccessible to energy minimization. It is hoped that this simulation follows the protein folding process and has a better chance at finding the true structure. A more realistic simulation can include water molecules surrounding the structure. This makes the process an even more computationally expensive procedure than energy minimization, however. Furthermore, it shares a similar weakness of energy minimization: a molecular structure can be "loosened up" such that it becomes less realistic. Much caution is therefore needed in using these molecular dynamic tools.

GROMOS (www.igc.ethz.ch/gromos/) is a UNIX program for molecular dynamic simulation. It is capable of performing energy minimization and thermodynamic

simulation of proteins, nucleic acids, and other biological macromolecules. The simulation can be done in vacuum or in solvents. A lightweight version of GROMOS has been incorporated in SwissPDB Viewer.

Model Evaluation

The final homology model has to be evaluated to make sure that the structural features of the model are consistent with the physicochemical rules. This involves checking anomalies in ϕ–ψ angles, bond lengths, close contacts, and so on. Another way of checking the quality of a protein model is to implicitly take these stereochemical properties into account. This is a method that detects errors by compiling statistical profiles of spatial features and interaction energy from experimentally determined structures. By comparing the statistical parameters with the constructed model, the method reveals which regions of a sequence appear to be folded normally and which regions do not. If structural irregularities are found, the region is considered to have errors and has to be further refined.

Procheck (www.biochem.ucl.ac.uk/∼roman/procheck/procheck.html) is a UNIX program that is able to check general physicochemical parameters such as ϕ–ψ angles, chirality, bond lengths, bond angles, and so on. The parameters of the model are used to compare with those compiled from well-defined, high-resolution structures. If the program detects unusual features, it highlights the regions that should be checked or refined further.

WHAT IF (www.cmbi.kun.nl:1100/WIWWWI/) is a comprehensive protein analysis server that validates a protein model for chemical correctness. It has many functions, including checking of planarity, collisions with symmetry axes (close contacts), proline puckering, anomalous bond angles, and bond lengths. It also allows the generation of Ramachandran plots as an assessment of the quality of the model.

ANOLEA (Atomic Non-Local Environment Assessment; http://protein.bio.puc.cl/cardex/servers/anolea/index.html) is a web server that uses the statistical evaluation approach. It performs energy calculations for atomic interactions in a protein chain and compares these interaction energy values with those compiled from a database of protein x-ray structures. If the energy terms of certain regions deviate significantly from those of the standard crystal structures, it defines them as unfavorable regions. An example of the output from the verification of a homology model is shown in Figure 15.3A. The threshold for unfavorable residues is normally set at 5.0. Residues with scores above 5.0 are considered regions with errors.

Verify3D (www.doe-mbi.ucla.edu/Services/Verify_3D/) is another server using the statistical approach. It uses a precomputed database containing eighteen environmental profiles based on secondary structures and solvent exposure, compiled from high-resolution protein structures. To assess the quality of a protein model, the secondary structure and solvent exposure propensity of each residue are calculated. If the parameters of a residue fall within one of the profiles, it receives a high score, otherwise a low score. The result is a two-dimensional graph illustrating the folding quality of each residue of the protein structure. A verification output of the above homology

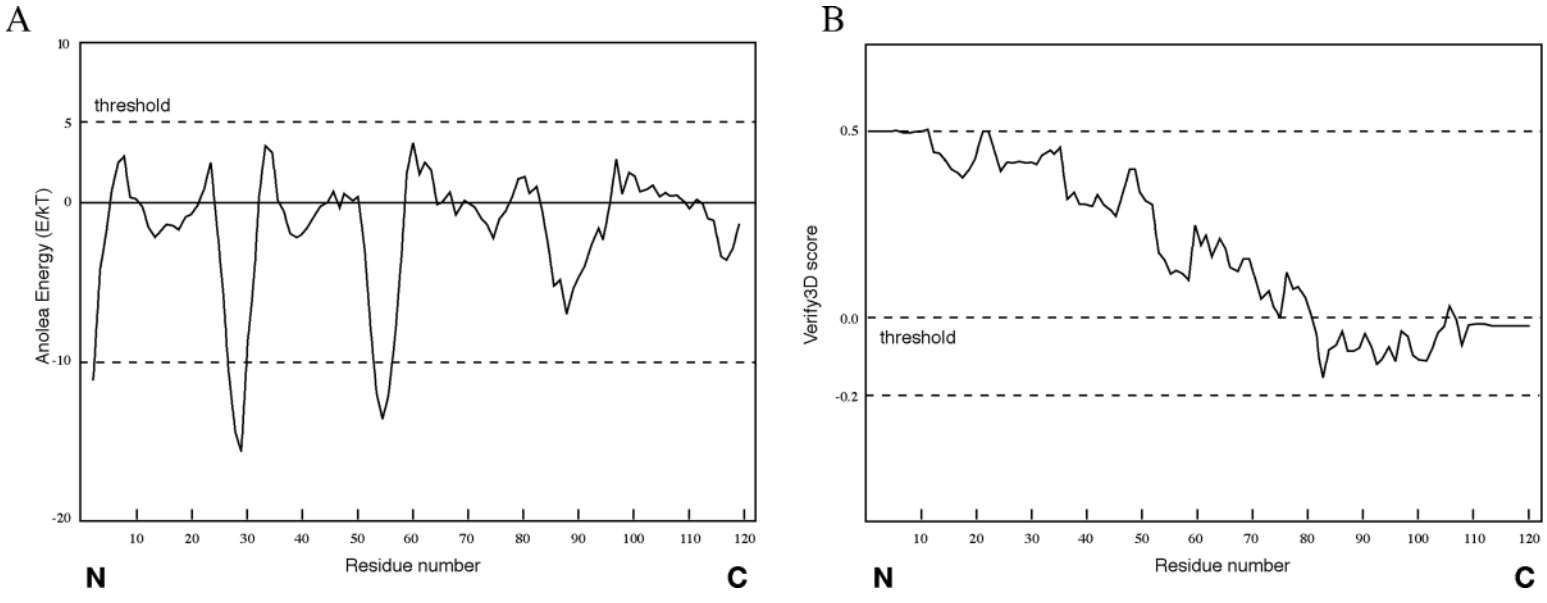


Figure 15.3: Example of protein model evaluation outputs by ANOLEA and Verify3D. The protein model was obtained from the Swiss model database (model code 1n5d). **(A)** The assessment result by the ANOLEA server. The threshold for unfavorable residues is normally set at 5.0. Residues with scores above 5.0 are considered regions with errors. **(B)** The assessment result by the Verify3D server. The threshold value is normally set at zero. The residues with the scores below zero are considered to have an unfavorable environment.

model is shown in Figure 15.3B. The threshold value is normally set at zero. Residues with scores below zero are considered to have an unfavorable environment.

The assessment results can be different using different verification programs. As shown in Figure 15.2, ANOLEA appears to be less stringent than Verify3D. Although the full-length protein chain of this model is declared favorable by ANOLEA, residues in the *C*-terminus of the protein are considered to be of low quality by Verify3D. Because no single method is clearly superior to any other, a good strategy is to use multiple verification methods and identify the consensus between them. It is also important to keep in mind that the evaluation tests performed by these programs only check the stereochemical correctness, regardless of the accuracy of the model, which may or may not have any biological meaning.

Comprehensive Modeling Programs

A number of comprehensive modeling programs are able to perform the complete procedure of homology modeling in an automated fashion. The automation requires assembling a pipeline that includes target selection, alignment, model generation, and model evaluation. Some freely available protein modeling programs and servers are listed.

Modeller (http://bioserv.cbs.cnrs.fr/HTML_BIO/frame_mod.html) is a web server for homology modeling. The user provides a predetermined sequence alignment of a template(s) and a target to allow the program to calculate a model containing all of the heavy atoms (nonhydrogen atoms). The program models the backbone using a homology-derived restraint method, which relies on multiple sequence alignment between target and template proteins to distinguish highly conserved residues from less conserved ones. Conserved residues are given high restraints in copying from the template structures. Less conserved residues, including loop residues, are given less or no restraints, so that their conformations can be built in a more or less ab initio fashion. The entire model is optimized by energy minimization and molecular dynamics procedures.

Swiss-Model (www.expasy.ch/swissmod/SWISS-MODEL.html) is an automated modeling server that allows a user to submit a sequence and to get back a structure automatically. The server constructs a model by automatic alignment (First Approach mode) or manual alignment (Optimize mode). In the First Approach mode, the user provides sequence input for modeling. The server performs alignment of the query with sequences in PDB using BLAST. After selection of suitable templates, a raw model is built. Refinement of the structure is done using GROMOS. Alternatively, the user can specify or upload structures as templates. The final model is sent to the user by e-mail. In the Optimize mode, the user constructs a sequence alignment in SwissPdbViewer and submits it to the server for model construction.

3D-JIGSAW (www.bmm.icnet.uk/servers/3djigsaw/) is a modeling server that works in either the automatic mode or the interactive mode. Its loop modeling relies on the database method. The interactive mode allows the user to edit alignments and select templates, loops, and side chains during modeling, whereas the automatic

mode allows no human intervention and models a submitted protein sequence if it has an identity >40% with known protein structures.

Homology Model Databases

The availability of automated modeling algorithms has allowed several research groups to use the fully automated procedure to carry out large-scale modeling projects. Protein models for entire sequence databases or entire translated genomes have been generated. Databases for modeled protein structures that include nearly one third of all known proteins have been established. They provide some useful information for understanding evolution of protein structures. The large databases can also aid in target selection for drug development. However, it has also been shown that the automated procedure is unable to model moderately distant protein homologs. Automated modeling tends to be less accurate than modeling that requires human intervention because of inappropriate template selection, suboptimal alignment, and difficulties in modeling loops and side chains.

ModBase (http://alto.compbio.ucsf.edu/modbase-cgi/index.cgi) is a database of protein models generated by the Modeller program. For most sequences that have been modeled, only partial sequences or domains that share strong similarities with templates are actually modeled.

3Dcrunch (www.expasy.ch/swissmod/SWISS-MODEL.html) is another database archiving results of large-scale homology modeling projects. Models of partial sequences from the Swiss-Prot database are derived using the Swiss-Model program.

THREADING AND FOLD RECOGNITION

As discussed in Chapters 12 and 13, there are only small number of protein folds available (<1,000), compared to millions of protein sequences. This means that protein structures tend to be more conserved than protein sequences. Consequently, many proteins can share a similar fold even in the absence of sequence similarities. This allowed the development of computational methods to predict protein structures beyond sequence similarities. To determine whether a protein sequence adopts a known three-dimensional structure fold relies on threading and fold recognition methods.

By definition, *threading* or *structural fold recognition* predicts the structural fold of an unknown protein sequence by fitting the sequence into a structural database and selecting the best-fitting fold. The comparison emphasizes matching of secondary structures, which are most evolutionarily conserved. Therefore, this approach can identify structurally similar proteins even without detectable sequence similarity.

The algorithms can be classified into two categories, pairwise energy based and profile based. The pairwise energy–based method was originally referred to as *threading* and the profile-based method was originally defined as *fold recognition*. However, the two terms are now often used interchangeably without distinction in the literature.

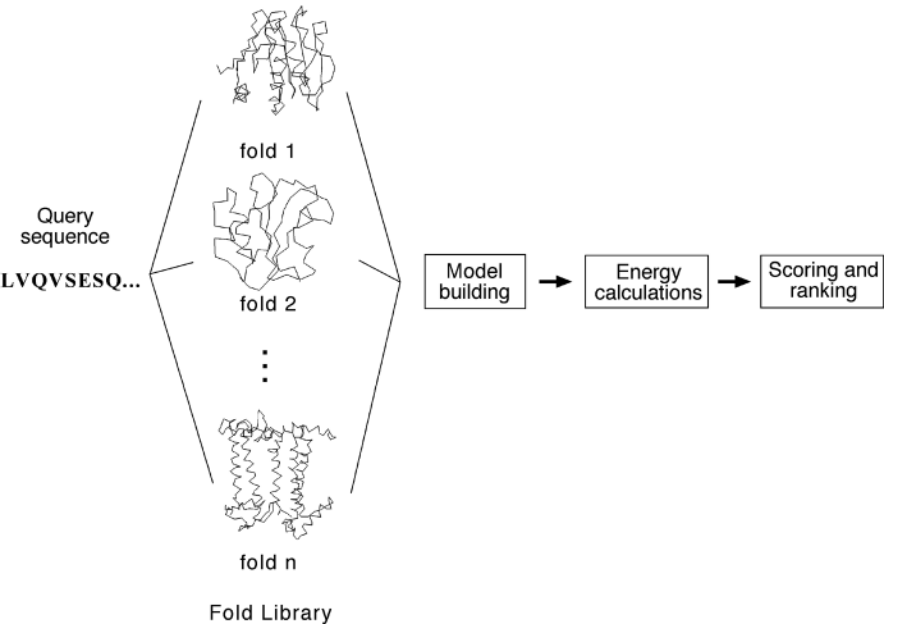


Figure 15.4: Outline of the threading method using the pairwise energy approach to predict protein structural folds from sequence. By fitting a structural fold library and assessing the energy terms of the resulting raw models, the best-fit structural fold can be selected.

Pairwise Energy Method

In the pairwise energy based method, a protein sequence is searched for in a structural fold database to find the best matching structural fold using energy-based criteria. The detailed procedure involves aligning the query sequence with each structural fold in a fold library. The alignment is performed essentially at the sequence profile level using dynamic programming or heuristic approaches. Local alignment is often adjusted to get lower energy and thus better fitting. The adjustment can be achieved using algorithms such as double-dynamic programming (see Chapter 14). The next step is to build a crude model for the target sequence by replacing aligned residues in the template structure with the corresponding residues in the query. The third step is to calculate the energy terms of the raw model, which include pairwise residue interaction energy, solvation energy, and hydrophobic energy. Finally, the models are ranked based on the energy terms to find the lowest energy fold that corresponds to the structurally most compatible fold (Fig. 15.4).

Profile Method

In the profile-based method, a profile is constructed for a group of related protein structures. The structural profile is generated by superimposition of the structures to expose corresponding residues. Statistical information from these aligned residues is then used to construct a profile. The profile contains scores that describe the propensity of each of the twenty amino acid residues to be at each profile position. The profile

scores contain information for secondary structural types, the degree of solvent exposure, polarity, and hydrophobicity of the amino acids. To predict the structural fold of an unknown query sequence, the query sequence is first predicted for its secondary structure, solvent accessibility, and polarity. The predicted information is then used for comparison with propensity profiles of known structural folds to find the fold that best represents the predicted profile.

Because threading and fold recognition detect structural homologs without completely relying on sequence similarities, they have been shown to be far more sensitive than PSI-BLAST in finding distant evolutionary relationships. In many cases, they can identify more than twice as many distant homologs than PSI-BLAST. However, this high sensitivity can also be their weakness because high sensitivity is often associated with low specificity. The predictions resulting from threading and fold recognition often come with very high rates of false positives. Therefore, much caution is required in accepting the prediction results.

Threading and fold recognition assess the compatibility of an amino acid sequence with a known structure in a fold library. If the protein fold to be predicted does not exist in the fold library, the method will fail. Another disadvantage compared to homology modeling lies in the fact that threading and fold recognition do not generate fully refined atomic models for the query sequences. This is because accurate alignment between distant homologs is difficult to achieve. Instead, threading and fold recognition procedures only provide a rough approximation of the overall topology of the native structure.

A number of threading and fold recognition programs are available using either or both prediction strategies. At present, no single algorithm is always able to provide reliable fold predictions. Some algorithms work well with some types of structures, but fail with others. It is a good practice to compare results from multiple programs for consistency and judge the correctness by using external knowledge.

3D-PSSM (www.bmm.icnet.uk/∼3dpssm/) is a web-based program that employs the structural profile method to identify protein folds. The profiles for each protein superfamily are constructed by combining multiple smaller profiles. First, protein structures in a superfamily based on the SCOP classification are superimposed and are used to construct a structural profile by incorporating secondary structures and solvent accessibility information for corresponding residues. In addition, each member in a protein structural superfamily has its own sequence-based PSI-BLAST profile computed. These sequence profiles are used in combination with the structure profile to form a large superfamily profile in which each position contains both sequence and structural information. For the query sequence, PSI-BLAST is performed to generate a sequence-based profile. PSI-PRED is used to predict its secondary structure. Both the sequence profile and predicted secondary structure are compared with the precomputed protein superfamily profiles, using a dynamic programming approach. The matching scores are calculated in terms of secondary structure, solvation energy, and sequence profiles and ranked to find the highest scored structure fold (Fig. 15.5).

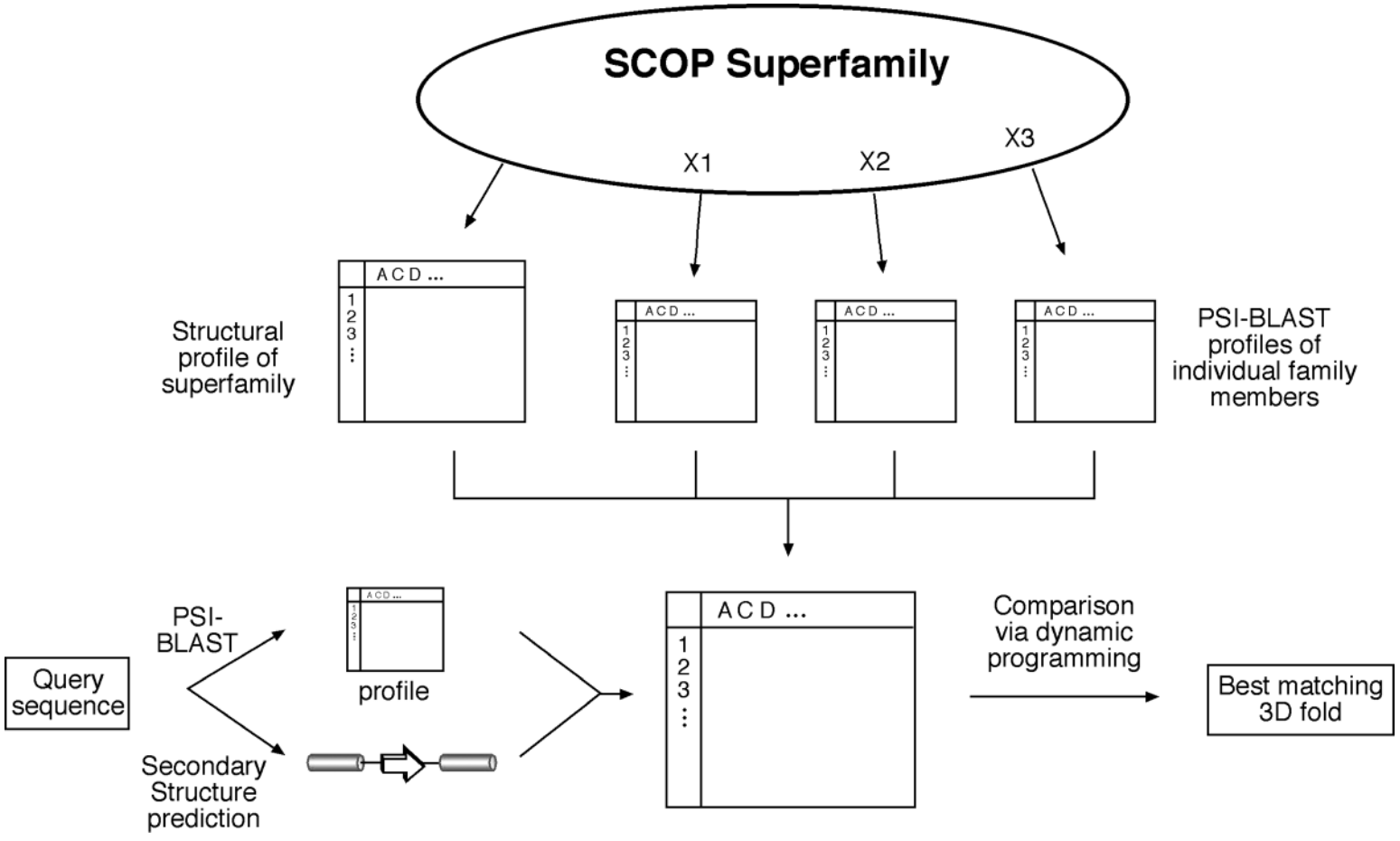


Figure 15.5: Schematic diagram of fold recognition by 3D-PSSM. A profile for protein structures in a SCOP superfamily is precomputed based on the structure profile of all members of the superfamily, as well as on PSI-BLAST sequence profiles of individual members of the superfamily. For the query sequence, a PSI-BLAST profile is constructed and its secondary structure information is predicted, which together are used to compare with the precomputed protein superfamily profile.

GenThreader (http://bioinf.cs.ucl.ac.uk/psipred/index.html) is a web-based program that uses a hybrid of the profile and pairwise energy methods. The initial step is similar to 3D-PSSM; the query protein sequence is subject to three rounds of PSI-BLAST. The resulting multiple sequence hits are used to generate a profile. Its secondary structure is predicted using PSIPRED. Both are used as input for threading computation based on a pairwise energy potential method. The threading results are evaluated using neural networks that combine energy potentials, sequence alignment scores, and length information to create a single score representing the relationship between the query and template proteins.

Fugue (www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html) is a profile-based fold recognition server. It has precomputed structural profiles compiled from multiple alignments of homologous structures, which take into account local structural environment such as secondary structure, solvent accessibility, and hydrogen bonding status. The query sequence (or a multiple sequence alignment if the user prefers) is used to scan the database of structural profiles. The comparison between the query and the structural profiles is done using global alignment or local alignment depending on sequence variability.

AB INITIO PROTEIN STRUCTURAL PREDICTION

Both homology and fold recognition approaches rely on the availability of template structures in the database to achieve predictions. If no correct structures exist in the database, the methods fail. However, proteins in nature fold on their own without checking what the structures of their homologs are in databases. Obviously, there is some information in the sequences that provides instruction for the proteins to "find" their native structures. Early biophysical studies have shown that most proteins fold spontaneously into a stable structure that has near minimum energy. This structural state is called the *native state*. This folding process appears to be nonrandom; however, its mechanism is poorly understood.

The limited knowledge of protein folding forms the basis of ab initio prediction. As the name suggests, the ab initio prediction method attempts to produce all-atom protein models based on sequence information alone without the aid of known protein structures. The perceived advantage of this method is that predictions are not restricted by known folds and that novel protein folds can be identified. However, because the physicochemical laws governing protein folding are not yet well understood, the energy functions used in the ab initio prediction are at present rather inaccurate. The folding problem remains one of the greatest challenges in bioinformatics today.

Current ab initio algorithms are not yet able to accurately simulate the protein-folding process. They work by using some type of heuristics. Because the native state of a protein structure is near energy minimum, the prediction programs are thus designed using the energy minimization principle. These algorithms search for every possible conformation to find the one with the lowest global energy. However,

searching for a fold with the absolute minimum energy may not be valid in reality. This contributes to one of the fundamental flaws of this approach. In addition, searching for all possible structural conformations is not yet computationally feasible. It has been estimated that, by using one of the world's fastest supercomputers (one trillion operations per second), it takes 10^{20} years to sample all possible conformations of a 40-residue protein. Therefore, some type of heuristics must be used to reduce the conformational space to be searched. Some recent ab initio methods combine fragment search and threading to yield a model of an unknown protein. The following web program is such an example using the hybrid approach.

Rosetta (www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php) is a web server that predicts protein three-dimensional conformations using the ab initio method. This in fact relies on a "mini-threading" method. The method first breaks down the query sequence into many very short segments (three to nine residues) and predicts the secondary structure of the small segments using a hidden Markov model–based program, HMMSTR (see Chapter 14). The segments with assigned secondary structures are subsequently assembled into a three-dimensional configuration. Through random combinations of the fragments, a large number of models are built and their overall energy potentials calculated. The conformation with the lowest global free energy is chosen as the best model.

It needs to be emphasized that up to now, ab initio prediction algorithms are far from mature. Their prediction accuracies are too low to be considered practically useful. Ab initio prediction of protein structures remains a fanciful goal for the future. However, with the current pace of high-throughput structural determination by the structural proteomics initiative, which aims to solve all protein folds within a decade, the time may soon come when there is little need to use the ab initio modeling approach because homology modeling and threading can provide much higher quality predictions for all possible protein folds. Regardless of the progress made in structural proteomics, exploration of protein structures using the ab initio prediction approach may still yield insight into the protein-folding process.

CASP

Discussion of protein structural prediction would not be complete without mentioning CASP (Critical Assessment of Techniques for Protein Structure Prediction). With so many protein structure prediction programs available, there is a need to know the reliability of the prediction methods. For that purpose, a common benchmark is needed to measure the accuracies of the prediction methods. To avoid letting programmers know the correct answer in the structure benchmarks in advance, already published protein structures cannot be used for testing the efficacy of new methodologies. Thus, a biannual international contest was initiated in 1994. It allows developers to predict unknown protein structures through blind testing so that the reliability of new prediction methods can be objectively evaluated. This is the experiment of CASP.

CASP contestants are given protein sequences whose structures have been solved by x-ray crystallography and NMR, but not yet published. Each contestant predicts the structures and submits the results to the CASP organizers before the structures are made publicly available. The results of the predictions are compared with the newly determined structures using structure alignment programs such as VAST, SARF, and DALI. In this way, new prediction methodologies can be evaluated without the possibility of bias. The predictions can be made at various levels of detail (secondary or tertiary structures) and in various categories (homology modeling, threading, ab initio). This experiment has been shown to provide valuable insight into the performance of prediction methods and has become the major driving force of development for protein structure prediction methods. For more information, the reader is recommended to visit the web site of the Protein Structure Prediction Center at http://predictioncenter.llnl.gov/.

SUMMARY

Protein structural prediction offers a theoretical alternative to experimental determination of structures. It is an efficient way to obtain structural information when experimental techniques are not successful. Computational prediction of protein structures is divided into three categories: homology modeling, threading, and ab initio prediction. Homology modeling, which is the most accurate prediction approach, derives models from close homologs. The process is simple in principle, but is more complicated in practice. It involves an elaborate procedure of template selection, sequence alignment correction, backbone generation, loop building, side chain modeling, model refinement, and model evaluation. Among these steps, sequence alignment is the most important step and loop modeling is the most difficult and error-prone step. Algorithms have been developed to automate the entire process and have been applied to a large-scale modeling work. However, the automated process tends to be less accurate than detailed manual modeling.

Another way to predict protein structures is through threading or fold recognition, which searches for a best fitting structure in a structural fold library by matching secondary structure and energy criteria. This approach is used when no suitable template structures can be found for homology-based modeling. The caveat is that this approach does not generate an actual model, but provide an essentially correct fold for the query protein. In addition, the protein fold of interest often does not exist in the fold library, in which case the method will fail.

The third prediction method – ab initio prediction – attempts to generate a structure without relying on templates, but by using physical rules only. It may be used when neither homology modeling nor threading can be applied. However, the ab initio approach so far has very limited success in getting correct structures. An objective evaluation platform, CASP, for protein structure prediction methodologies has been established to allow program developers to test the effectiveness of the algorithms.

FURTHER READING

Al-Lazikani, B., Jung, J., Xiang, Z., and Honig, B. 2001. Protein structure prediction. *Curr. Opin. Chem. Biol.* 5:51–6.

Baker, D., and Sali, A. 2001. Protein structure prediction and structural genomics. *Science*. 294:93–6.

Bonneau, R., and Baker, D. 2001. *Ab initio* protein structure prediction: Progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* 30:173–89.

Bourne, P. E. 2003. "CSAP and CAFASP experiments and their findings." In *Structural Bioinformatics*, edited by P. E. Bourne and H. Weissig, 501–7. Hoboken, NJ: Wiley-Liss.

Chivian, D., Robertson, T., Bonneau, R., and Baker, D. 2003. "*Ab initio* methods." In *Structural Bioinformatics*, edited by P. E. Bourne and H. Weissig, 547–56. Hoboken, NJ: Wiley-Liss.

Edwards, Y. J., and Cottage, A. 2003. Bioinformatics methods to predict protein structure and function. A practical approach. *Mol Biotechnol.* 23:139–66.

Fetrow, J. S., Giammona, A., Kolinski, A., and Skolnick, J. 2002. The protein folding problem: A biophysical enigma. *Curr. Pharm. Biotechnol.* 3:329–47.

Forster, M. J. 2000. Molecular modelling in structural biology. *Micron* 33:365–84.

Ginalski, K., Grishin, N. V., Godzik, A., and Rychlewski, L. 2005. Practical lessons from protein structure prediction. *Nucleic Acids Res.* 33:1874–91.

Godzik, A. 2003. "Fold recognition methods." In *Structural Bioinformatics*, edited by P. E. Bourne and H. Weissig, 525–46. Hoboken, NJ: Wiley-Liss.

Hardin, C., Pogorelov, T. V., and Luthey-Schulten, Z. 2002. *Ab initio* protein structure prediction. *Curr. Opin. Struct. Biol.* 12:176–81.

Krieger, E., Nabuurs, S. B., and Vriend, G. 2003. "Homology modeling." In *Structural Bioinformatics*, edited by P. E. Bourne and H. Weissig, 509–23. Hoboken, NJ: Wiley-Liss.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29:291–325.

Xu, D., Xu, Y., and Uberbacher, E. C. 2000. Computational tools for protein modeling. *Curr. Protein Pept. Sci.* 1:1–21.

CHAPTER SIXTEEN

RNA Structure Prediction

RNA is one of the three major types of biological macromolecules. Understanding the structures of RNA provides insights into the functions of this class of molecules. Detailed structural information about RNA has significant impact on understanding the mechanisms of a vast array of cellular processes such as gene expression, viral infection, and immunity. RNA structures can be experimentally determined using x-ray crystallography or NMR techniques (see Chapter 10). However, these approaches are extremely time consuming and expensive. As a result, computational prediction has become an attractive alternative. This chapter presents the basics of RNA structures and current algorithms for RNA structure prediction, with an emphasis on secondary structure prediction.

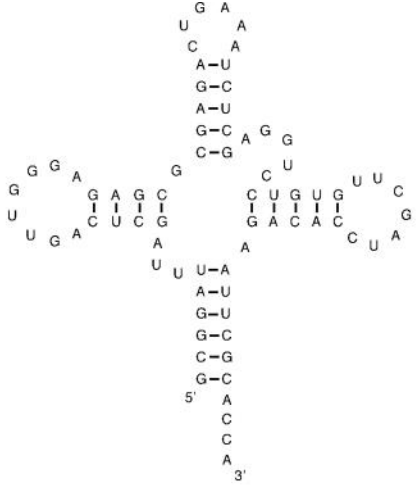
INTRODUCTION

It is known that RNA is a carrier of genetic information and exists in three main forms. They are messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA). Their main roles are as follows: mRNA is responsible for directing protein synthesis; rRNA provides structural scaffolding within ribosomes; and tRNA serves as a carrier of amino acids for polypeptide synthesis.

Recent advances in biochemistry and molecular biology have allowed the discovery of new functions of RNA molecules. For example, RNA has been shown to possess catalytic activity and is important for RNA splicing, processing, and editing. A class of small, noncoding RNA molecules, termed microRNA or miRNA, have recently been identified to regulate gene expression through interaction with mRNA molecules.

Unlike DNA, which is mainly double stranded, RNA is single stranded, although an RNA molecule can self-hybridize at certain regions to form partial double-stranded structures. Generally, mRNA is more or less linear and nonstructured, whereas rRNA and tRNA can only function by forming particular secondary and tertiary structures. Therefore, knowledge of the structures of these molecules is particularly important for understanding their functions. Difficulties in experimental determination of RNA structures make theoretical prediction a very desirable approach. In fact, computational-based analysis is a main tool in RNA-based drug design in pharmaceutical industry. In addition, knowledge of the secondary structures of rRNA is key for RNA-based phylogenetic analysis.

GCGGAUUUAGCUCAGUUGGGAGAGC
GCCAGACUGAAAUCUGGAGGUCCUG
UGUUCGAUCCACAGAAUUCGCACCA



Primary structure Secondary structure Tertiary structure

Figure 16.1: The primary, secondary, and tertiary structures of a tRNA molecule illustrating the three levels of RNA structural organization.

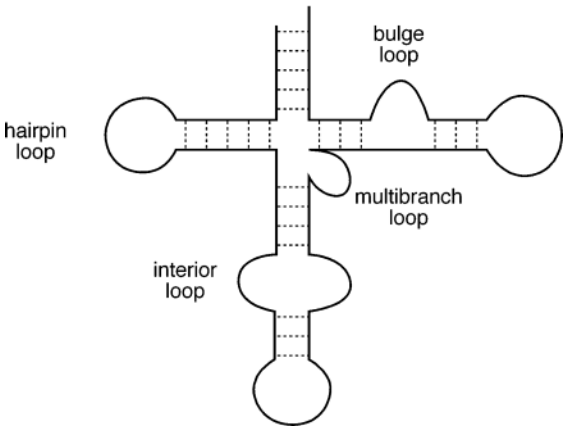


Figure 16.2: Schematic diagram of a hypothetical RNA molecular containing four basic types of RNA loops: a hairpin loop, bulge loop, interior loop, and multibranch loop. Dashed lines indicate base pairings in the helical regions of the molecule.

TYPES OF RNA STRUCTURES

RNA structures can be described at three levels as in proteins: primary, secondary, and tertiary. The primary structure is the linear sequence of RNA, consisting of four bases, adenine (A), cytosine (C), guanine (G), and uracil (U). The secondary structure refers to the planar representation that contains base-paired regions among single-stranded regions. The base pairing is mainly composed of traditional Watson–Crick base pairing, which is A–U and G–C. In addition to the canonical base pairing, there often exists noncanonical base pairing such as G and U base paring. The G–U base pair is less stable and normally occurs within a double-strand helix surrounded by Watson–Crick base pairs. Finally, the tertiary structure is the three-dimensional arrangement of bases of the RNA molecule. Examples of the three levels of RNA structural organization are illustrated in Figure 16.1.

Because the RNA tertiary structure is very difficult to predict, attention has been mainly focused on secondary structure prediction. It is therefore important to learn in more detail about RNA secondary structures. Based on the arrangement of helical base pairing in secondary structures, four main subtypes of secondary structures can be identified. They are hairpin loops, bulge loops, interior loops, and multibranch loops (Fig. 16.2).

The *hairpin loop* refers to a structure with two ends of a single-stranded region (loop) connecting a base-paired region (stem). The *bulge loop* refers to a single stranded region connecting two adjacent base-paired segments so that it "bubbles" out in the middle of a double helix on one side. The *interior loop* refers to two single-stranded regions on opposite strands connecting two adjacent base-paired segments. It can be said to "bubble" out on both sides in the middle of a double helical segment. The *multibranch loop*, also called *helical junctions*, refers to a loop that brings three or more base-paired segments in close vicinity forming a multifurcated structure.

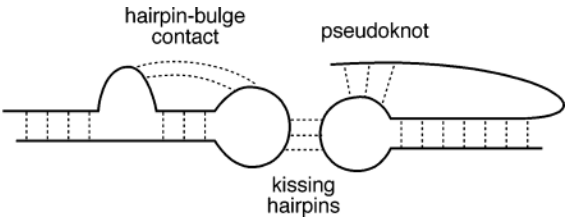


Figure 16.3: A hypothetical RNA structure containing a pseudoknot, kissing hairpin, and hairpin–bulge contact.

In addition to the traditional secondary structural elements, base pairing between loops of different secondary structural elements can result in a higher level of structures such as pseudoknots, kissing hairpins, and hairpin–bulge contact (Fig. 16.3). A *pseudoknot loop* refers to base pairing formed between loop residues within a hairpin loop and residues outside the hairpin loop. A *kissing hairpin* refers to a hydrogen bonded interaction formed between loop residues of two hairpin structures. The *hairpin–bulge* contact refers to interactions between loop residues of a hairpin loop and a bulge loop. This type of interaction forms supersecondary structures, which are relatively rare in real structures and thus are ignored by most conventional prediction algorithms.

RNA SECONDARY STRUCTURE PREDICTION METHODS

At present, there are essentially two types of method of RNA structure prediction. One is based on the calculation of the minimum free energy of the stable structure derived from a single RNA sequence. This can be considered an ab initio approach. The second is a comparative approach which infers structures based on an evolutionary comparison of multiple related RNA sequences.

AB INITIO APPROACH

This approach makes structural predictions based on a single RNA sequence. The rationale behind this method is that the structure of an RNA molecule is solely determined by its sequence. Thus, algorithms can be designed to search for a stable RNA structure with the lowest free energy. Generally, when a base pairing is formed, the energy of the molecule is lowered because of attractive interactions between the two strands. Thus, to search for a most stable structure, ab initio programs are designed to search for a structure with the maximum number of base pairs.

Free energy can be calculated based on parameters empirically derived for small molecules. G–C base pairs are more stable than A–U base pairs, which are more stable than G–U base pairs. It is also known that base-pair formation is not an independent event. The energy necessary to form individual base pairs is influenced by adjacent base pairs through helical stacking forces. This is known as *cooperativity* in helix formation. If a base pair is next to other base pairs, the base pairs tend to stabilize

each other through attractive stacking interactions between aromatic rings of the base pairs. The attractive interactions lead to even lower energy. Parameters for calculating the cooperativity of the base-pair formation have been determined and can be used for structure prediction.

However, if the base pair is adjacent to loops or bulges, the neighboring loops and bulges tend to destabilize the base-pair formation. This is because there is a loss of entropy when the ends of the helical structure are constrained by unpaired loop residues. The destabilizing force to a helical structure also depends on the types of loops nearby. Parameters for calculating different destabilizing energies have also been determined and can be used as penalties for secondary structure calculations.

The scoring scheme based on the combined stabilizing and destabilizing interactions forms the foundation of the ab initio RNA secondary structure prediction method. This method works by first finding all possible base-pairing patterns from a sequence and then calculating the total energy of a potential secondary structure by taking into account all the adjacent stabilizing and destabilizing forces. If there are multiple alternative secondary structures, the method finds the conformation with the lowest energy, meaning that it is energetically most favorable.

Dot Matrices

In searching for the lowest energy form, all possible base-pair patterns have to be examined. There are several methods for finding all the possible base-paired regions from a given nucleic acid sequence. The dot matrix method and the dynamic programming method introduced in Chapter 3 can be used in detecting self-complementary regions of a sequence. A simple dot matrix can find all possible base-paring patterns of an RNA sequence when one sequence is compared with itself (Fig. 16.4). In this case, dots are placed in the matrix to represent matching complementary bases instead of identical ones.

The diagonals perpendicular to the main diagonal represent regions that can self-hybridize to form double-stranded structure with traditional A–U and G–C base pairs. In reality, the pattern detection in a dot matrix is often obscured by high noise levels. As discussed in Chapter 3, one way to reduce the noise in the matrix is to select an appropriate window size of a minimum number of contiguous base matches. Normally, only a window size of four consecutive base matches is used. If the dot plot reveals more than one feasible structures, the lowest energy one is chosen.

Dynamic Programming

The use of a dot plot can be effective in finding a single secondary structure in a small molecule (see Fig. 16.4). However, if a large molecule contains multiple secondary structure segments, choosing a combination that is energetically most stable among a large number of possibilities can be a daunting task. To overcome the problem, a quantitative approach such as dynamic programming can be used to assemble a final structure with optimal base-paired regions. In this approach, an RNA sequence

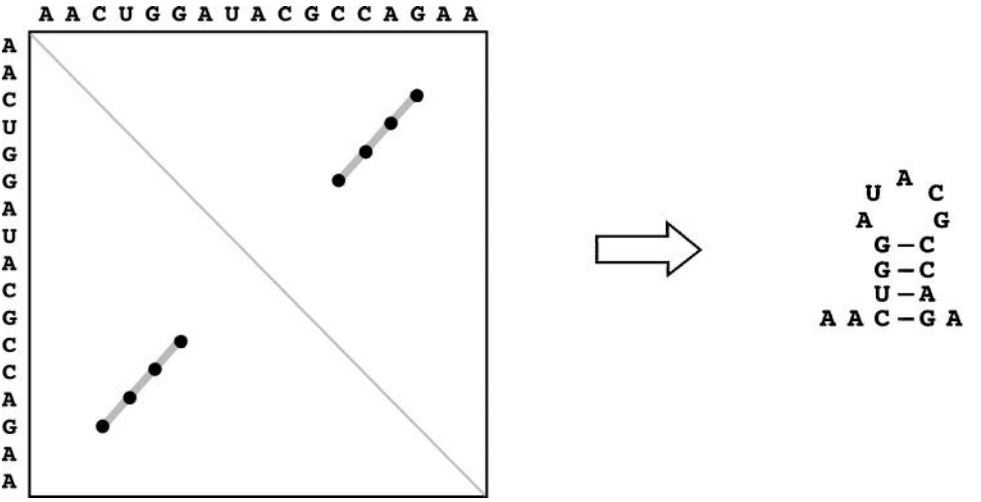


Figure 16.4: Example of a dot plot used for RNA secondary structure prediction. In this plot, an RNA sequence is compared with itself. Dots are placed for matching complementary bases when a window size of four nucleotide match is used. A main diagonal, which is perpendicular to the short diagonals, is placed for self-matching. Based on the dot plot, the predicted secondary structure for this sequence is shown on the right.

is compared with itself. A scoring scheme is applied to fill the matrix with match scores based on Watson–Crick base complementarity. Often, G–U base pairing and energy terms of the base pairing are also incorporated into the scoring process. A path with the maximal score within a scoring matrix after taking into account the entire sequence information represents the most probable secondary structure form.

The dynamic programming method produces one structure with a single best score. However, this is potentially a drawback of this approach because in reality an RNA may exist in multiple alternative forms with near minimum energy but not necessarily the one with maximum base pairs.

Partition Function

The problem of dynamic programming to select one single structure can be complemented by adding a probability distribution function, known as the *partition function*, which calculates a mathematical distribution of probable base pairs in a thermodynamic equilibrium. This function helps to select a number of suboptimal structures within a certain energy range. The following lists two well-known programs using the ab initio prediction method.

Mfold (www.bioinfo.rpi.edu/applications/mfold/) is a web-based program for RNA secondary structure prediction. It combines dynamic programming and thermodynamic calculations for identifying the most stable secondary structures with the lowest energy. It also produces dot plots coupled with energy terms. This method is reliable for short sequences, but becomes less accurate as the sequence length increases.

RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) is one of the web programs in the Vienna package. Unlike Mfold, which only examines the energy terms of

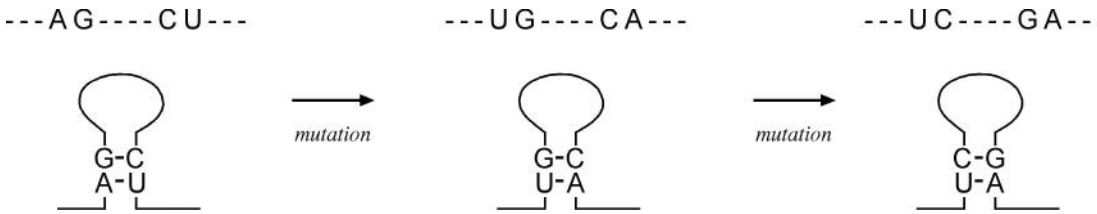


Figure 16.5: Example of covariation of residues among three homologous RNA sequences to maintain the stability of an existing secondary structure.

the optimal alignment in a dot plot, RNAfold extends the sequence alignment to the vicinity of the optimal diagonals to calculate thermodynamic stability of alternative structures. It further incorporates a partition function to select a number of statistically most probable structures. Based on both thermodynamic calculations and the partition function, a number of alternative structures that may be suboptimal are provided. The collection of the predicted structures may provide a better estimate of plausible foldings of an RNA molecule than the predictions by Mfold. Because of the much larger number of secondary structures to be computed, a more simplified energy rule has to be used to increase computational speed. Thus, the prediction results are not always guaranteed to be better than those predicted by Mfold.

COMPARATIVE APPROACH

The comparative approach uses multiple evolutionarily related RNA sequences to infer a consensus structure. This approach is based on the assumption that RNA sequences that deem to be homologous fold into the same secondary structure. By comparing related RNA sequences, an evolutionarily conserved secondary structure can be derived.

To distinguish the conserved secondary structure among multiple related RNA sequences, a concept of "covariation" is used. It is known that RNA functional motifs are structurally conserved. To maintain the secondary structures while the homologous sequences evolve, a mutation occurring in one position that is responsible for base pairing should be compensated for by a mutation in the corresponding base-pairing position so to maintain base pairing and the stability of the secondary structure (Fig. 16.5). This is the concept of *covariation*. Any lack of covariation can be deleterious to the RNA structure and functions. Based on this rule, algorithms can be written to search for the covariation patterns after a set of homologous RNA sequences are properly aligned. The detected correlated substitutions help to determine conserved base pairing in a secondary structure.

Another aspect of the comparative method is to select a common structure through consensus drawing. Because predicting secondary structures for each individual sequence may produce errors, by comparing all predicted structures of a group of aligned RNA sequences and drawing a consensus, the commonly adopted structure can be selected; many other possible structures can be eliminated in the process. The

comparative-based algorithms can be further divided into two categories based on the type of input data. One requires predefined alignment and the other does not.

Algorithms That Use Prealignment

This type of algorithm requires the user to provide a pairwise or multiple alignment as input. The sequence alignment can be obtained using standard alignment programs such as T-Coffee, PRRN, or Clustal (see Chapter 5). Based on the alignment input, the prediction programs compute structurally consistent mutational patterns such as covariation and derive a consensus structure common for all the sequences. In practice, the consensus structure prediction is often combined with thermodynamic calculations to improve accuracy.

This type of program is relatively successful for reasonably conserved sequences. The requirement for using this type of program is an appropriate set of homologous sequences that have to be similar enough to allow accurate alignment, but divergent enough to allow covariations to be detected. If this condition is not met, correct structures cannot be inferred. The method also depends on the quality of the input alignment. If there are errors in the alignment, covariation signals will not be detected. The selection of one single consensus structure is also a drawback because alternative and evolutionarily unconserved structures are not predicted. The following is an example of this type of program based on predefined aligned sequences.

RNAalifold (http://rna.tbi.univie.ac.at/cgi-bin/alifold.cgi) is a program in the Vienna package. It uses a multiple sequence alignment as input to analyze covariation patterns on the sequences. A scoring matrix is created that combines minimum free energy and covariation information. Dynamic programming is used to select the structure that has the minimum energy for the whole set of aligned RNA sequences.

Algorithms That Do Not Use Prealignment

This type of algorithm simultaneously aligns multiple input sequences and infers a consensus structure. The alignment is produced using dynamic programming with a scoring scheme that incorporates sequence similarity as well as energy terms. Because the full dynamic programming for multiple alignment is computationally too demanding, currently available programs limit the input to two sequences.

Foldalign (http://foldalign.kvl.dk/server/index.html) is a web-based program for RNA alignment and structure prediction. The user provides a pair of unaligned sequences. The program uses a combination of Clustal and dynamic programming with a scoring scheme that includes covariation information to construct the alignment. A commonly conserved structure for both sequences is subsequently derived based on the alignment. To reduce computational complexity, the program ignores multibranch loops and is only suitable for handling short RNA sequences.

Dynalign (http://rna.urmc.rochester.edu/) is a UNIX program with a free source code for downloading. The user again provides two input sequences. The program calculates the possible secondary structures of each using a method similar to Mfold.

By comparing multiple alternative structures from each sequence, a lowest energy structure common to both sequences is selected that serves as the basis for sequence alignment. The unique feature of this program is that it does not require sequence similarity and therefore can handle very divergent sequences. However, because of the computation complexity, the program only predicts small RNA sequences such as tRNA with reasonable accuracy.

PERFORMANCE EVALUATION

Rigorously evaluating the performance of RNA prediction programs has traditionally been hindered by the dearth of three-dimensional structural information for RNA. The availability of recently solved crystal structures of the entire ribosome provides a wealth of structural details relating to diverse types of RNA molecules. The high-resolution structural information can then be used as a benchmark for evaluating state-of-the-art RNA structure prediction programs in all categories.

If prediction accuracy can be represented using a single parameter such as the correlation coefficient, which takes into account both sensitivity and selectivity information (see Chapter 8), the ab initio–based programs score roughly 20% to 60% depending on the length of the sequences. Generally speaking, the programs perform better for shorter RNA sequences than for longer ones. For small RNA sequences, such as tRNA, some programs may be able to produce 70% accuracy. The major limitation for performance gains of this category appears to be dependence on energy parameters alone, which may not be sufficient to distinguish different structural possibilities of the same molecule.

Based on recent benchmark comparisons, the comparative-type algorithms can reach an accuracy range of 20% to 80%. The results depend on whether a program is prealignment dependent or not. Most of the superior performance comes from prealignment-dependent programs such as RNAalifold. The prealignment-independent programs fare much worse for predicting long sequences. For small RNA sequences such as tRNA, both subtypes can achieve very high accuracy (up to 100%). This illustrates that the comparative approach is consistently more accurate than the ab initio one.

SUMMARY

Detailed understanding of RNA structures is important for understanding the functional role of RNA in the cell. The demand for structural information about RNA has motivated the development of a large number of prediction algorithms. Current RNA structure prediction is predominantly focused on secondary structures owing to the difficulty in predicting tertiary structures. The secondary structure prediction methods can be classified as either ab initio or comparative. The ab initio method is based on energetic calculations from a single query sequence. However, the accuracy of the ab initio method is limited. The comparative approach, which requires multiple

sequences, is able to achieve better accuracy. However, the obvious drawback of the consensus approach is the requirement for a unique set of homologous sequences. Neither type of the prediction methods currently considers pseudoknots in the RNA structure because of the much greater computational complexity involved. To further increase prediction performance, the research and development should focus on alleviating some of the current drawbacks.

FURTHER READING

Doshi, K. J., Cannone, J. J., Cobaugh, C. W., and Gutell, R. R. 2004. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics* 5:105.

Doudna, J. A. 2000. Structural genomics of RNA. *Nat. Struct. Biol*. Suppl:954–6.

Gardner, P. P., and Giegerich, R. 2004. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* 5:140.

Gorodkin, J. Stricklin, S. L., and Stormo, G. D. 2001. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids. Res*. 10:2135–44.

Leontis, N. B., Stombaugh, J., and Westhof, E. 2002. Motif prediction in ribosomal RNAs lessons and prospects for automated motif prediction in homologous RNA molecules. *Biochimie* 84:961–73.

Major, F., and Griffey, R. 2001. Computational methods for RNA structure determination. *Curr. Opin. Struct. Biol*. 11:282–6.

Westhof, E., Auffinger, P., and Gaspin, C. 1997. "DNA and RNA structure prediction.: In: *DNA and Protein Sequence Analysis*, edited by M. J. Bishop and C. J. Rawlings, 255–78. Oxford, UK: IRL Press.