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 ($C\alpha$), 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.

R
n
↓
Alkyl

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

Hydrophobic ⇔ water fearing

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

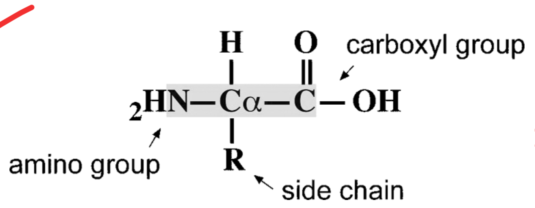
Amino Acid Group	Amino Acid Name	Three- and One-Letter Code	Main Functional Features
Small and nonpolar	Glycine Alanine Proline	Gly, G Ala, A Pro, P	Nonreactive in chemical reactions; Pro and Gly disrupt regular secondary structures
Small and polar	Cysteine Serine Threonine	Cys, C Ser, S Thr, T	Serving 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

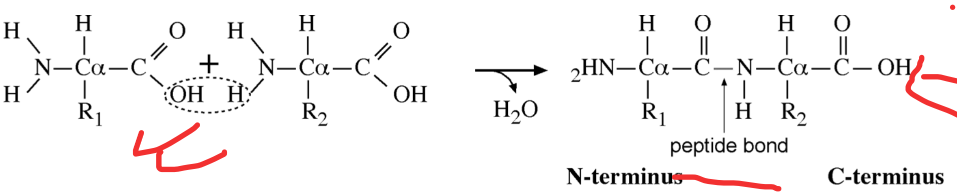
The peptide formation involes two amino acids covalently joined together between the carboxyl group of one amino acid and the amino group of another (Fig. 12.2). This



Amino acid

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.

water removed.



dipeptide

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.

Amino-acid oxide

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.

Imp

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.

busy and unstability and conjusted

— up

down

x-ax is

(peptide plane)

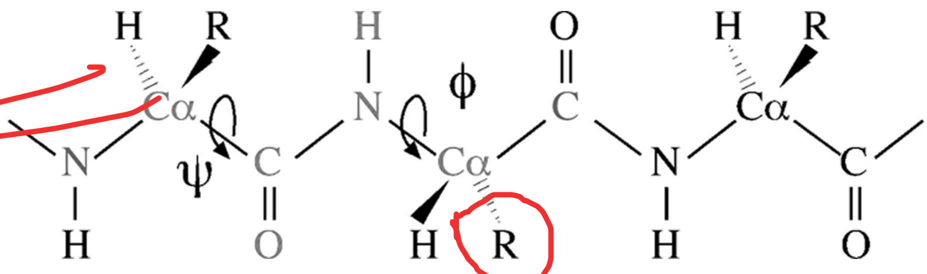
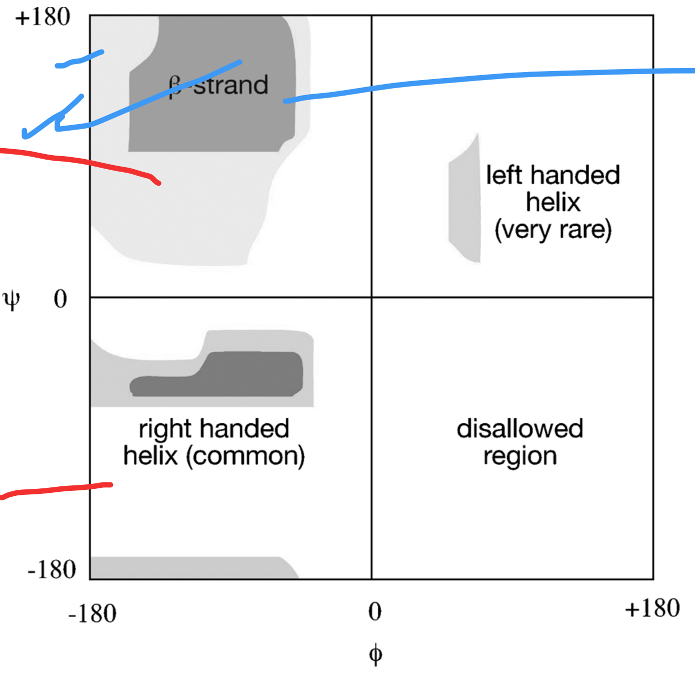


Figure 12.3: Definition of dihedral angles of ϕ and ψ. Six atoms around a peptide bond forming two peptide planes are colored in red. The ϕ angle is the rotation about the N–Cα bond, which is measured by the angle between a virtual plane formed by the C–N–Cα and the virtual plane by N–Cα–C (C in green).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



β-sheet

ψ = near 180

φ = -ve
near -180

α helix

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.

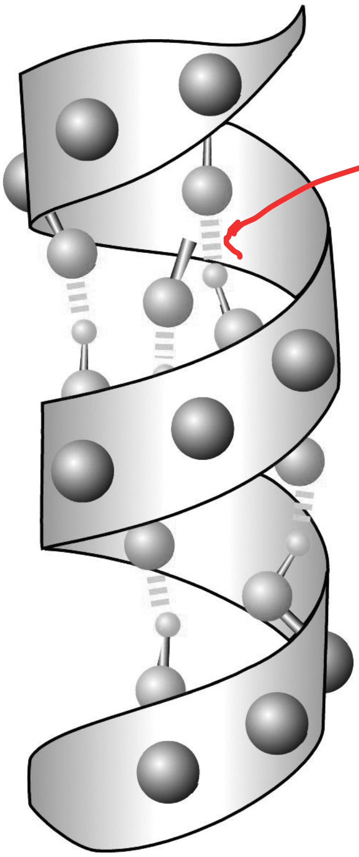
Stabilizing Forces

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.



Salt bridge

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).

IMP

1 α helix
2 β -sheets
3 Coiled coils

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.

a pattern to Soum's Biology

Helical struct

look slowly

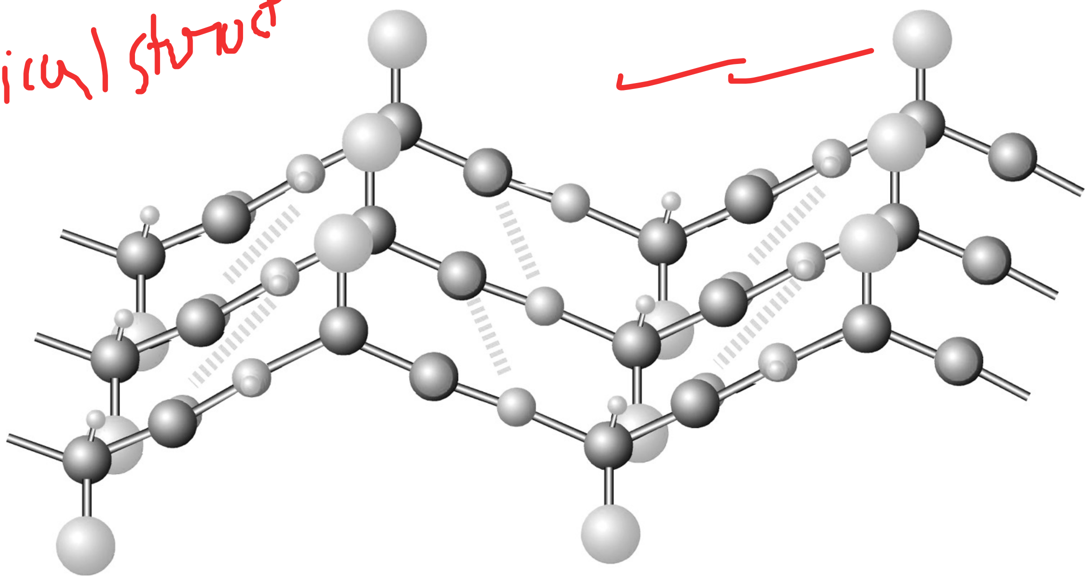


Figure 12.6: Side view of a parallel β-sheet. Hydrogen bonds between the carbonyl oxygen (red) and the amino hydrogen (green) of adjacent β-strands are shown in yellow dashed lines. R groups are shown as big balls in cyan and are positioned alternately on opposite sides of β-strands (see color plate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

β-strand is like primary structure

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.

non-helical struct

non-helical

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.

βsheet →

β-strand

parallel and anti-parallel β-strand

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,

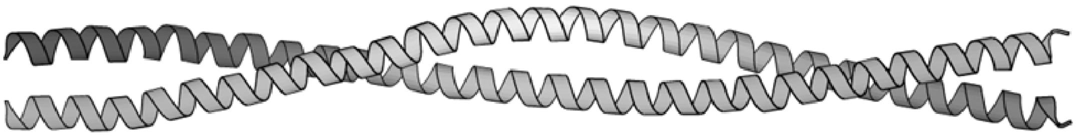


Figure 12.7: An α-helical coiled coil found in tropomyosin showing two helices wound around to form a helical bundle.

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

Coiled Coils

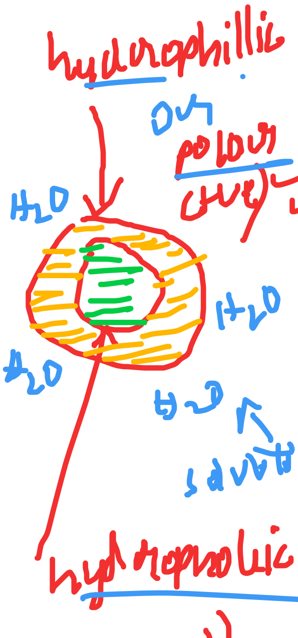
Coiled coils are a special type of supersecondary structure characterized by a bundle of two or more α-helices wrapping around each other (Fig. 12.7). The helices forming coiled coils have a unique pattern of hydrophobicity, which repeats every seven residues (five hydrophobic and two hydrophilic). More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

→ globular protein

→ integral membrane protein

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.

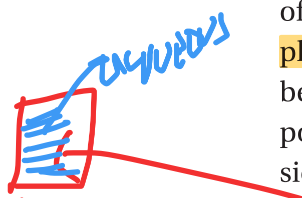


Globular Proteins

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



hydrophobic
(lipids)
(stable) α helices

hydrophillic

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

Imp

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

Imp

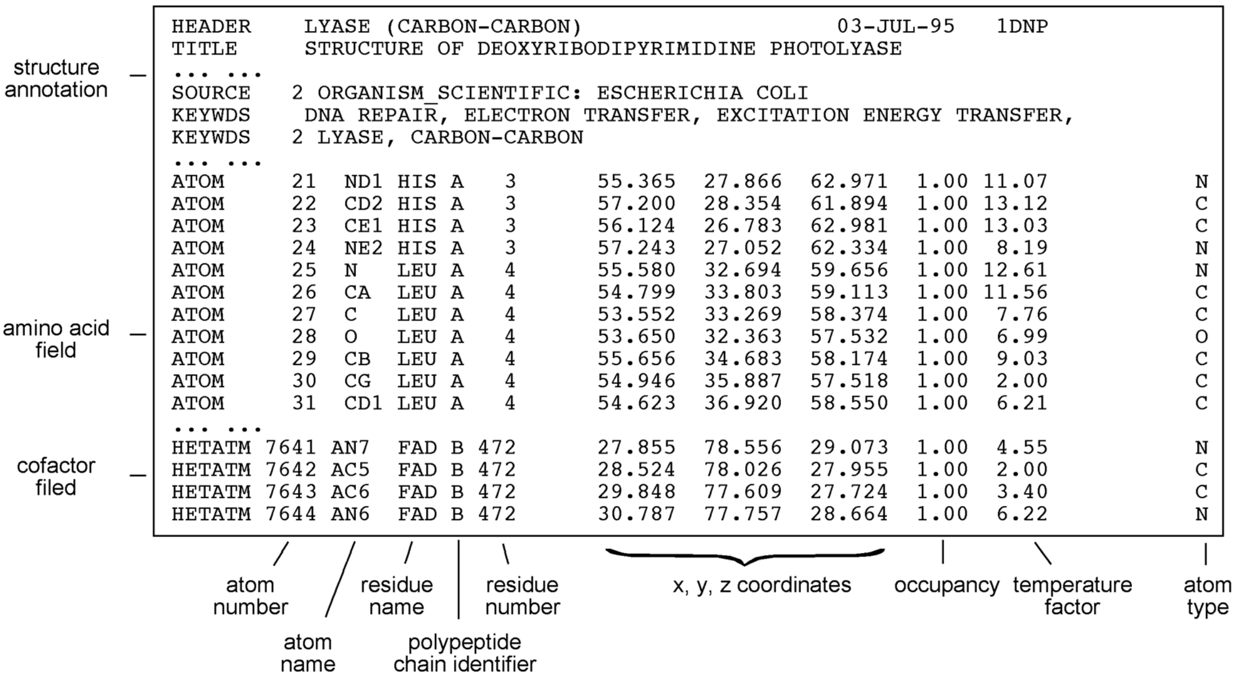
NMR spectroscopy detects spinning patterns of atomic nuclei in a magnetic field. Protein samples are labeled with radioisotopes such as ^{13}C and ^{15}N. A radiofrequency radiation is used to induce transitions between nuclear spin states in a magnetic field. Interactions between spinning isotope pairs 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.



```
HEADER    LYASE (CARBON-CARBON)                        03-JUL-95   1DNP
TITLE     STRUCTURE OF DEOXYRIBODIPYRIMIDINE PHOTOLYASE
... ...
SOURCE   2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI
KEYWDS    DNA REPAIR, ELECTRON TRANSFER, EXCITATION ENERGY TRANSFER,
KEYWDS   2 LYASE, CARBON-CARBON
... ...
ATOM     21  ND1 HIS A   3      55.365  27.866  62.971  1.00 11.07           N
ATOM     22  CD2 HIS A   3      57.200  28.354  61.894  1.00 13.12           C
ATOM     23  CE1 HIS A   3      56.124  26.783  62.981  1.00 13.03           C
ATOM     24  NE2 HIS A   3      57.243  27.052  62.334  1.00  8.19           N
ATOM     25  N   LEU A   4      55.580  32.694  59.656  1.00 12.61           N
ATOM     26  CA  LEU A   4      54.799  33.803  59.113  1.00 11.56           C
ATOM     27  C   LEU A   4      53.552  33.269  58.374  1.00  7.76           C
ATOM     28  O   LEU A   4      53.650  32.363  57.532  1.00  6.99           O
ATOM     29  CB  LEU A   4      55.656  34.683  58.174  1.00  9.03           C
ATOM     30  CG  LEU A   4      54.946  35.887  57.518  1.00  2.00           C
ATOM     31  CD1 LEU A   4      54.623  36.920  58.550  1.00  6.21           C
... ...
HETATM 7641 AN7  FAD B 472      27.855  78.556  29.073  1.00  4.55           N
HETATM 7642 AC5  FAD B 472      28.524  78.026  27.955  1.00  2.00           C
HETATM 7643 AC6  FAD B 472      29.848  77.609  27.724  1.00  3.40           C
HETATM 7644 AN6  FAD B 472      30.787  77.757  28.664  1.00  6.22           N
```

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.

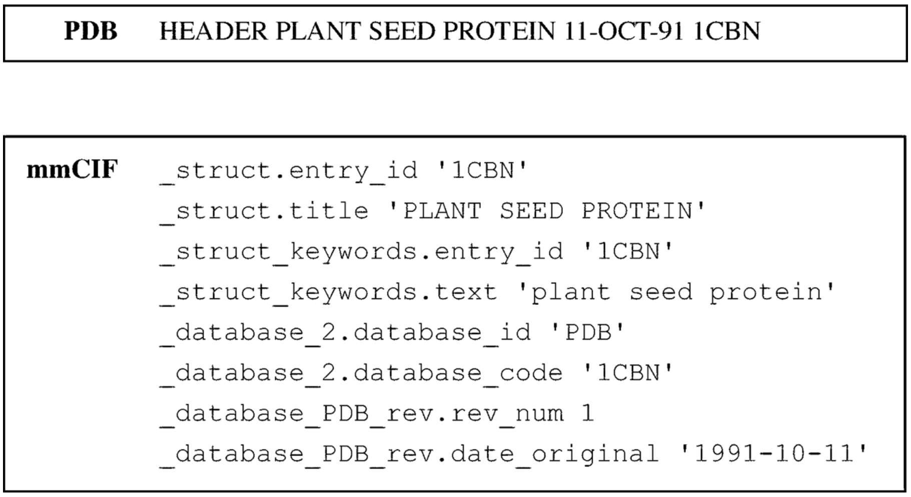
The PDB format has been in existence for more than three decades. It is fairly easy to read and simple to use. However, the format is not designed for computer extraction of information from the records. Certain restrictions in the format have significantly complicated its current use. For instance, 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.