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.

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. It also provides links to protein structural classification results available in databases such as SCOP and CATH. 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. 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.

Disadvantages of PDB format

1. The format is not designed for computer extraction of information from the records.
2. 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.
3. 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 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. 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 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. An MMDB file is written in the ASN.1 format, 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.