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 Collaborator 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 data base 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.

PDB Format

A deposited set of protein coordinates becomes an entry in PDB. Each entry is given a unique code, PDB id, 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 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. 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