

# 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/](http://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

structure annotation	HEADER	LYASE (CARBON-CARBON)					03-JUL-95			1DNP		
	TITLE	STRUCTURE OF DEOXYRIBODIPYRIMIDINE PHOTOLYASE										
	... ..											
	SOURCE	2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI										
amino acid field	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
cofactor filed	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

atom  
number

residue  
name

residue  
number

x, y, z coordinates

occupancy

temperature  
factor

atom  
type

atom  
name

polypeptide  
chain identifier