Primary Protein Sequence Repositories

Protein information resource **(PIR)** at the NBRF (National Biomedical Research Foundation, USA), and **SWISS-PROT** at the SBI (Swiss Biotechnology Institute), Switzerland are protein sequence databases.

The **PIR-PSD** is a collaborative work between the **PIR**, **MIPS** (Munich Information Centre for Protein Sequences, Germany) and **JIPID** (Japan International Protein Information Database, Japan).

The **PIR-PSD** is now a comprehensive, non- redundant, expertly annotated, object relational DBMS. It is available at https://proteininformationresource.org under resource tab / menu.

A unique characteristic of the PIR-PSD is it's classification of protein sequences based on the super family concept. Sequence in PIR- PSD is also classified based on homology domain and sequence motifs. Homology domains may correspond to evolutionary building blocks, while sequence motifs represent functional sites or conserved regions. The classification approach allows a more complete understanding of sequence function structure relationship.

The primary structure database - PDB and CSD

PDB stands for Protein Databank. In spite of the name, PDB archive the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules, large peptides such as complexes of protein and nucleic acids. The database holds data derived from mainly three sources. Structure determined by X-ray crystallography form the large majority of the entries. This is followed by structures arrived at by NMR experiments. There are also structures obtained by molecular modelling. The data in the PDB is organized as flat files, one to a structure, which usually means that each file contain one molecule, or one molecular complex. It is now conserved by the Research Collaboratory for Structural Bioinformatics (**RCSB:- www.rcsb.org**).

The Cambridge Structural Database (CSD:- www.ccdc.cam.ac.uk) was originally a project of the University of Cambridge, which is set up to collect together the published three-dimensional structure of small organic molecules. This excludes proteins and medium sized nucleic acid fragments, but small peptides such as neuropeptides, and monomer and dimmers of nucleic acid finds a place in the CSD. Currently CSD holds crystal structures information for about 2.5 lakhs organic and metal organic compounds. All these crystal structures have been obtained using X-ray or neuron diffraction technique. For each entry in the CSD there are three distinct types of information stored. These are categorized as bibliographic information, chemical connectivity information and the three- dimensional coordinates. The annotation data field incorporates all of the bibliographic material for the particular entry and summarized the structural and experimental information for the crystal structure.

PROTEIN DATA BANK (PDB)

PDB stands for Protein Databank. In spite of the name, PDB archive the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules, large peptides such as complexes of protein and nucleic acids. The database holds data derived from mainly three sources. Structure determined by X-ray crystallography form the large majority of the entries. This is followed by structures arrived at by NMR experiments. There are also structures obtained by molecular modelling. It was established in 1971 at Brookhaven National Laboratory , It is now conserved by the Research Collaboratory for Structural Bioinformatics (**RCSB:- www.rcsb.org**).

Data Storage and Acquisition

The data for each structure is stored in a distinct file and hence the data is stored in flat file arrangement. Jmol, Pymol, and Rasmol and web browser plugins etc can be used to visualize pdb files.

File Format Description

The various sections of the PDB file are:

1. Title Section,
2. Primary Structure Section,
3. Heterogen Section,
4. Secondary Structure Section,
5. Connectivity Annotation Section,
6. Miscellaneous Features Section,
7. Crystallographic and Coordinate Transformation Section,
8. Coordinate Section,
9. Connectivity Section, and
10. Bookkeeping Section.

Selected Protein Data Bank Record Types

ATOM atomic coordinate record containing the x,y,z orthogonal Angstrom coordinates for atoms in standard residues (amino acids and nucleic acids).

HETATM atomic coordinate record containing the x,y,z orthogonal Angstrom coordinates for atoms in nonstandard residues. Nonstandard residues include inhibitors, cofactors, ions, and solvent. The only functional difference from ATOM records is that HETATM residues are by default not connected to other residues. Note that water residues should be in HETATM records.

TER indicates the end of a chain of residues. For example, a hemoglobin molecule consists of four subunit chains which are not connected. TER indicates the end of a chain and prevents the display of a connection to the next chain.

SSBOND defines disulfide bond linkages between cysteine residues.

HELIX indicates the location and type (right-handed alpha, *etc.*) of helices. One record per helix.

SHEET indicates the location, sense (anti-parallel, *etc.*) and registration with respect to the previous strand in the sheet (if any) of each strand in the model. One record per strand.

Record Type	Columns	Data
ATOM	1-4	‘‘ATOM’’
	7-11	Atom serial number
	13-16	Atom name
	17	Alternate location indicator
	18-20	Residue name
	22	Chain identifier
	23-26	Residue sequence number
	27	Code for insertions of residues
	31-38	X orthogonal Angstrom coordinate
	39-46	Y orthogonal Angstrom coordinate
	47-54	Z orthogonal Angstrom coordinate
	55-60	Occupancy
	61-66	Temperature factor
	73-76	Segment identifier (optional)
	77-78	Element symbol
	79-80	Charge (optional)
HETATM	1-6	‘‘HETATM’’
	7-80	same as ATOM records
TER	1-3	‘‘TER’’

It is necessary to choose the best possible crystallographic structure prior to embarking on a drug design project. This is because this structure serves as a starting point and template on which all successive steps are dependent.

One critical factor in crystallographic data selection is its resolution. Resolution implies the smallest distance within which atoms may be reliably distinguished.

The higher the resolution or the smaller the distance within which atoms may be reliably distinguished, the better is the crystallographic structure.

Resolutions ranging from 2-3.5Å are considered acceptable starting points for drug design projects.

About 85% of the models (entries) in the Protein Data Bank were determined by X-ray crystallography. (Most of the remaining 15% were determined by solution nuclear magnetic resonance.) Analysis of x-ray diffraction patterns from protein crystals produces an electron density map, into which an atomic model of the protein is fitted. Major errors sometimes occur when fitting models in to low-resolution electron density maps.

The **R value** is used to assess progress in the refinement of a model from X-ray crystallographic data, and can be used as one factor in evaluating the quality of a model.

R is a measure of error between the observed intensities from the diffraction pattern and the predicted intensities that are calculated from the model. R values of 0.20 or less are taken as evidence that the model is reliable.

(As a rule of thumb, models with R values substantially exceeding (resolution/10) should be treated with caution. Thus, if the resolution of a model is 2.5 Å, that model's R value should not exceed 0.25. Completely erroneous models (e.g. random models) give R values of 0.40 to 0.60.)

The value of Free R is the best clue as to whether major errors may be present in a published model.

Free R should not exceed the R value by more than 0.05; that is, if the R value is 0.20, free R should not significantly exceed 0.25. Free R values exceeding 0.40 raise serious doubts about the model

Examples of PDB Format (For Reference Only)

Fields following the temperature factor in ATOM and HETATM records are not shown in any of the examples.

Glucagon is a small protein of 29 amino acids in a single chain. The first residue is the amino- terminal amino acid, histidine, which is followed by a serine residue and then a glutamine. The coordinate information starts with:

```
ATOM      1  N   HIS     1      49.668  24.248  10.436  1.00 25.00
ATOM      2  CA  HIS     1      50.197  25.578  10.784  1.00 16.00
ATOM      3  C   HIS     1      49.169  26.701  10.917  1.00 16.00
ATOM      4  O   HIS     1      48.241  26.524  11.749  1.00 16.00
ATOM      5  CB  HIS     1      51.312  26.048   9.843  1.00 16.00
ATOM      6  CG  HIS     1      50.958  26.068   8.340  1.00 16.00
ATOM      7  ND1 HIS     1      49.636  26.144   7.860  1.00 16.00
ATOM      8  CD2 HIS     1      51.797  26.043   7.286  1.00 16.00
ATOM      9  CE1 HIS     1      49.691  26.152   6.454  1.00 17.00
ATOM     10  NE2 HIS     1      51.046  26.090   6.098  1.00 17.00
ATOM     11  N   SER     2      49.788  27.850  10.784  1.00 16.00
ATOM     12  CA  SER     2      49.138  29.147  10.620  1.00 15.00
ATOM     13  C   SER     2      47.713  29.006  10.110  1.00 15.00
ATOM     14  O   SER     2      46.740  29.251  10.864  1.00 15.00
ATOM     15  CB  SER     2      49.875  29.930   9.569  1.00 16.00
ATOM     16  OG  SER     2      49.145  31.057   9.176  1.00 19.00
ATOM     17  N   GLN     3      47.620  28.367   8.973  1.00 15.00
ATOM     18  CA  GLN     3      46.287  28.193   8.308  1.00 14.00
ATOM     19  C   GLN     3      45.406  27.172   8.963  1.00 14.00
```

Notice that each line or *record* begins with the record type, ATOM. The atom serial number is the next item in each record.

The atom name is the third item in the record. Notice that the first one or two characters of the atom name consists of the chemical symbol for the atom type. All the atom names beginning with ‘‘C’’ are carbon atoms; ‘‘N’’ indicates a nitrogen and ‘‘O’’ indicates oxygen. The next character is the remoteness indicator code, which is transliterated according to:

α	A
β	B
γ	G
δ	D
ε	E
ζ	Z
η	H

The last character of the atom name is a branch indicator, if required.

The next data field is the residue type. Notice that *each* record contains the residue type. In this example, the first residue in the chain is HIS (histidine) and the second residue is a SER (serine).

The next data field contains the residue sequence number. Notice that as the residue changes from histidine to serine, the residue number changes from ‘‘1’’ to ‘‘2.’’ Two like residues may be adjacent to one another, so the residue number is very important for distinguishing between them.

The next three data fields contain the X, Y, and Z coordinate values, respectively. The next data field is the occupancy. The final field shown is the temperature factor (B value).

The glucagon data file continues in this manner until the final residue is reached:

```
ATOM    239  N   THR    29       3.391  19.940  12.762  1.00 21.00
ATOM    240  CA  THR    29       2.014  19.761  13.283  1.00 21.00
ATOM    241  C   THR    29        .826  19.943  12.332  1.00 23.00
ATOM    242  O   THR    29        .932  19.600  11.133  1.00 30.00
ATOM    243  CB  THR    29       1.845  20.667  14.505  1.00 21.00
ATOM    244  OG1 THR    29       1.214  21.893  14.153  1.00 21.00
ATOM    245  CG2 THR    29       3.180  20.968  15.185  1.00 21.00
ATOM    246  OXT THR    29       -.317  20.109  12.824  1.00 25.00
TER     247      THR    29
```

Note that this residue includes the extra oxygen atom, ‘‘OXT,’’ on the terminal carboxyl group. The ‘‘TER’’ record terminates the amino acid chain.

A more complicated protein, fetal hemoglobin, consists of two amino acid chains (alpha and gamma) and two heme groups. The first ten lines of coordinates for this molecule are:

```
ATOM      1  N   VAL A   1       6.280  17.225   4.929  1.00  0.00
ATOM      2  CA  VAL A   1       6.948  18.508   4.671  1.00  0.00
ATOM      3  C   VAL A   1       8.436  18.338   4.977  1.00  0.00
ATOM      4  O   VAL A   1       8.813  17.657   5.941  1.00  0.00
ATOM      5  CB  VAL A   1       6.317  19.598   5.527  1.00  0.00
ATOM      6  CG1 VAL A   1       6.959  20.999   5.376  1.00  0.00
ATOM      7  CG2 VAL A   1       4.819  19.636   5.383  1.00  0.00
ATOM      8  N   LEU A   2       9.259  18.958   4.152  1.00  0.00
ATOM      9  CA  LEU A   2      10.715  18.872   4.330  1.00  0.00
ATOM     10  C   LEU A   2      11.156  20.058   5.187  1.00  0.00
```

This data file appears much the same as the file for glucagon, with the exception that the fifth data field now contains the single-character chain indicator. In this case, the chain indicator is ‘‘A,’’ denoting the alpha chain of the hemoglobin molecule. This field was simply blank in the glucagon example. At the end of chain A, the heme group records appear:

```
ATOM   1058  N   ARG A 141      -6.576  12.834 -10.275  1.00  0.00
ATOM   1059  CA  ARG A 141      -8.044  12.831 -10.214  1.00  0.00
ATOM   1060  C   ARG A 141      -8.186  14.096  -9.365  1.00  0.00
ATOM   1061  O   ARG A 141      -7.591  15.139  -9.671  1.00  0.00
ATOM   1062  CB  ARG A 141      -8.579  11.531  -9.580  1.00  0.00
ATOM   1063  CG  ARG A 141      -8.386  11.441  -8.054  1.00  0.00
ATOM   1064  CD  ARG A 141      -8.727  10.045  -7.568  1.00  0.00
ATOM   1065  NE  ARG A 141      -9.095  10.056  -6.143  1.00  0.00
ATOM   1066  CZ  ARG A 141      -9.268   8.931  -5.414  1.00  0.00
ATOM   1067  NH1 ARG A 141      -8.602   8.795  -4.282  1.00  0.00
ATOM   1068  NH2 ARG A 141     -10.097   7.962  -5.830  1.00  0.00
ATOM   1069  OXT ARG A 141      -8.973  13.984  -8.310  1.00  0.00
TER    1070      ARG A 141
HETATM 1071  FE  HEM A   1       8.133   8.321 -15.014  1.00  0.00
HETATM 1072  CHA HEM A   1       8.863   8.752 -18.417  1.00  0.00
HETATM 1073  CHB HEM A   1      10.362  10.946 -14.389  1.00  0.00
HETATM 1074  CHC HEM A   1       8.482   7.374 -11.743  1.00  0.00
HETATM 1075  CHD HEM A   1       6.982   5.180 -15.773  1.00  0.00
HETATM 1076  NA  HEM A   1       9.452   9.545 -16.178  1.00  0.00
```

The last residue in the alpha chain is an ‘‘ARG’’ (arginine). Again, the extra oxygen atom ‘‘OXT’’ appears in the terminal carboxyl group. The ‘‘TER’’ record indicates the end of the peptide chain. It is important to have ‘‘TER’’ records at the end of peptide chains so a bond is not drawn from the end of one chain to the start of another.

In the example above, the ‘‘TER’’ record is correct and should be present, but the molecule chain would still be terminated at that point even without a ‘‘TER’’ record, because ‘‘HETATM’’ residues are not connected to other residues or to each other. The heme group is a single residue made up of ‘‘HETATM’’ records.

At the end of the heme group associated with the alpha chain, the gamma chain begins:

```
HETATM 1109  CAD HEM A   1       7.582   6.731 -20.480  1.00  0.00
HETATM 1110  CBD HEM A   1       8.992   6.848 -20.968  1.00  0.00
HETATM 1111  CGD HEM A   1       8.998   6.529 -22.465  1.00  0.00
HETATM 1112  O1D HEM A   1       9.693   5.683 -22.895  1.00  0.00
HETATM 1113  O2D HEM A   1       8.276   7.153 -23.229  1.00  0.00
ATOM   1114  C   ACE G   0       7.896 -18.462  -1.908  1.00  0.00
ATOM   1115  O   ACE G   0       7.246 -18.839   -.922  1.00  0.00
ATOM   1116  CH3 ACE G   0       9.415 -18.301  -1.832  1.00  0.00
ATOM   1117  N   GLY G   1       7.354 -18.174  -3.077  1.00  0.00
ATOM   1118  CA  GLY G   1       5.904 -18.282  -3.283  1.00  0.00
ATOM   1119  C   GLY G   1       7.139 -19.112  -2.930  1.00  0.00
ATOM   1120  O   GLY G   1       7.026 -20.248  -2.448  1.00  0.00
ATOM   1121  N   HIS G   2       8.300 -18.533  -3.176  1.00  0.00
ATOM   1122  CA  HIS G   2       9.565 -19.224  -2.889  1.00  0.00
```

Here the ‘‘TER’’ card is implicit in the start of a new chain. The new chain identifier is ‘‘G.’’ The file continues in the same pattern as before until the entire gamma chain and its associated heme group have been specified.

The spacing of the data fields is crucial. If a data field does not apply, it should be left blank. For example, a protein which consists of a single amino acid chain has no chain identifier, and thus column 22 is blank.

From this example, it is apparent that Protein Data Bank format relies on the concept of *residues*. The rules for residues can be summarized as:

(1) All atoms within a single residue must have unique names. For example, residue ''VAL'' may have only one atom named ''CA.'' Other residues may also have a ''CA'' atom but not more than one ''CA'' may appear in ''VAL.''

(2) Residue names are a maximum of three characters long and uniquely identify the residue type. Thus, all residues of a given name in a file will be the same type of residue and have the same structure. Each occurrence of a particular residue in the Protein Data Bank file should have the same atoms with the same connectivity.

SWISS-PROT(http://www.expasy.ch/sprot) is a curated proteins sequence database which provides a high level of annotation. The data in each entry can considered separately as core data and annotation. The core data consists of the sequences entered in common single letter amino acid code, and the related references and bibliography. The taxonomy of the organism from which the sequence was obtained also forms part of this core information.

The annotation contains information on the function or functions of the protein, post-translational modification such as phosphorylation, functional and structural domains and sites, such as calcium binding regions, ATP-binding sites, zinc fingers, etc., known secondary structural features as for examples alpha helix, beta sheet, etc., the quaternary structure of the protein, similarities to other protein if any, and diseases that may rise due to different authors publishing different sequences for the same protein, or due to mutations in different strains often described as part of the annotation.

Lines of code in SWISS-PROT database:

Code	Expansion	Remarks
ID	Identification	Occurs at the beginning of the entry. Contains a unique name for the entry, plus information on the status of the entry. If it has been checked and conforms to SWISS-PROT standards, it is called STANDARD.
AC	Accession numbers	This is a stable way of identifying the entry. The name may change but not the AC. If the line has more than one number, it means that the entry was constituted by merging other entries.
DT	Date	There are three dates corresponding to the creation date of the entry and modification dates of the sequence and the annotation respectively
DE	Description	Lines that start with the identifier contain general description about the sequence.
GN	Gene name	The name of the gene (or genes) that codes for the protein
OS, OG,OC	Organism name, Organelle, Organism classification	The name and taxonomy of the organism, and information regarding the organelle containing the gene e.g. mitochondria or chloroplast, etc.
RN, RP,RX,RA RT,RL	Reference number, Position, comments, cross-reference, authors, title and location.	Bibliographic reference to the sequence. This includes information (following the code RP) on the extent of work carried out by the authors.
CC	Comments	These are free text comments that provide any relevant information pertaining to the entry.
DR	Database cross-reference	This line gives cross-references to other databases where information regarding this entry is also found. As for example to structural information for the protein in the PDB.
KW	Keywords	This line gives a list of keywords that can be used in indexes. Search programs very often simply go through such indices to identify required information
FT	Features Table	These lines describe regions or sites of interest in the sequence, e.g. post-transitional modifications, binding sites, enzyme active sites and local secondary structures
SQ	Sequence Header	This line indicates the beginning of the sequence data and gives a brief summary of its contents.

TrEMBL (part of uniport)
Translated EMBL is a computer-annotated protein sequence database that is released as a supplement to SWISS-PROT. It contains the translation of all coding sequences present in the EMBL Nucleotide database, which have not been fully annotated or Computationally analyzed. Thus it may contain the sequence of proteins that are never expressed and never actually identified in the organisms.

UNIPROT (https://www.uniprot.org)

It is a database of freely accessible protein sequences which contains high-quality data and functional information for the proteins. Many of the records have been obtained from genome sequencing projects. The information regarding the biological function of the protein has been extracted from the research literature.

European Bioinformatics Institute (EBI), Swiss Institute of Bioinformatics (SIB) and Protein Information Resource (PIR) constitute the UniProt consortium.

Each one of them is deeply engaged in protein database maintenance and annotation. It includes four core databases: UniProtKB, UniParc, UniRef, and UniMes.

UniProtKB

UniProt Knowledgebase (UniProtKB) is a protein database that is partially curated by experts. It includes three databases: Swiss-Prot, TrEMBL, and PIR-PSD. The former one contains reviewed and manually annotated records whereas the latter one comprises the un-reviewed and automatically annotated entries.