

Bioinformatics: Life Sciences on Your Computer

Structure Databases, Structure Viewers

This is supplemental reading for the video lecture on the PDB Database.

Protein structure databases

Here are some major protein databases:

PDB: <http://www.pdb.org/pdb/home/home.do>

PDBe: <http://www.ebi.ac.uk/pdbe/>

MMDB: <http://www.ncbi.nlm.nih.gov/sites/entrez?db=structure>

PDBSum: <http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/>

Target Track: <http://sbkb.org/tt/>

Protein structure formats

The PDB format for protein structure is the oldest format, dating back to the early 1970s. It was designed to be compatible with FORTRAN programming. The idea is to describe the three dimensional coordinated of every atom in the protein structure.

The mmCIF format is newer format and designed for a relational database. The information storage is better.

MMDB is the NCBI format for structure information. Basically it is a PDB file translated into a format that works with ASN.1, which is integral to NCBI databases.

In July 2007, the PDB database introduced PDBML, a new XML format for PDB that is said to be more user-friendly.

PDB file

Below is an example of a PDB file. Note the amino acid, the atom and the xyz coordinates of every atom. With a PDB file, the user can reconstruct a structure with that information, Visualizing programs can recreate the entire protein model from this PDB file.

```

HEADER      TRANSFERASE                      18-DEC-08   2ZWV
TITLE      CRYSTAL STRUCTURE OF THERMUS THERMOPHILUS 16S rRNA
TITLE      2 METHYLTRANSFERASE RSMC (TTHA0533)
(-)
COMPND     2 MOLECULE: PROBABLE RIBOSOMAL RNA SMALL SUBUNIT
COMPND     3 METHYLTRANSFERASE;
(-)
SOURCE     2 ORGANISM_SCIENTIFIC: THERMUS THERMOPHILUS;
SOURCE     3 ORGANISM_TAXID: 300852;
(-)
AUTHOR     H.WANG, M.KAWAZOE, T.KAMINISHI, A.TATSUGUCHI, C.NAOE, T.TERADA,
AUTHOR     2 M.SHIROUZU, C.TAKEMOTO, S.YOKOYAMA, RIKEN STRUCTURAL
(-)
REMARK 200 EXPERIMENTAL DETAILS
REMARK 200 EXPERIMENT TYPE                : X-RAY DIFFRACTION
REMARK 200 DATE OF DATA COLLECTION       : 03-MAR-07
REMARK 200 TEMPERATURE (KELVIN)          : 100
REMARK 200 PH                             : 8.9
REMARK 200 NUMBER OF CRYSTALS USED        : 1
(-)
ATOM       1  N   SER A  2      45.623  19.967  23.987  1.00 52.48      N
ATOM       2  CA  SER A  2      46.685  19.984  22.942  1.00 50.33      C
ATOM       3  C   SER A  2      46.300  20.908  21.788  1.00 51.12      C
ATOM       4  O   SER A  2      46.924  21.952  21.577  1.00 52.96      O
ATOM       5  CB  SER A  2      46.920  18.565  22.417  1.00 47.98      C
ATOM       6  OG  SER A  2      45.715  17.994  21.938  1.00 49.64      O
ATOM       7  N   LEU A  3      45.267  20.521  21.045  1.00 44.14      N
ATOM       8  CA  LEU A  3      44.886  21.319  19.917  1.00 39.44      C
ATOM       9  C   LEU A  3      44.067  22.570  20.356  1.00 41.34      C
ATOM      10  O   LEU A  3      43.262  22.543  21.306  1.00 45.49      O
(-)
ATOM      2855 C   GLY A 374      36.607  38.031  32.464  1.00 35.87      C
ATOM      2856 O   GLY A 374      36.908  38.301  33.645  1.00 38.56      O
TER        2857 GLY A 374
HETATM    2858 N   SAH A 376      24.272  14.684  22.476  1.00 25.40      N
HETATM    2859 CA  SAH A 376      23.776  13.382  22.005  1.00 24.17      C
HETATM    2860 CB  SAH A 376      24.908  12.612  21.323  1.00 27.40      C
(-)

```

--image from <http://lib.bioinfo.pl/files/courses/materials/141/Fig2.png>

PDB browsing

The [PDB database](http://www.rcsb.org/pdb) has 101,046 structures as of June 17, 2014. Of those structures, 89,522 are crystal structures, meaning that X-ray crystallography was used to solve the 3-D structures. The PDB website allows for keyword or author searches. The browse capability has many options, some summarized below:

Browse tab	Example
Biological process	Proteins associated with locomotion
Cell component	Lysosomal proteins
Molecular function	Electron carrier proteins
EC Numbers	3.2.2.1 – purine nucleosidases
Transporter classification	1.A.11 – ammonia transporter family proteins
Source organisms	Proteins from <i>Synechococcus elongatus</i>
Genome location	Proteins from <i>Equus caballus</i> (horse)
MeSH	Proteins involved in avian influenza
SCOP	Alpha helical proteins
CATH	8 Propellor beta proteins
Protein modification	Proteins with an amidated residue

Discussion topic: Enter accession number 1AEW. Note that PDB has mixed numbers and letters, including 0 and O and 1 and I. Be careful. Examine the record in detail and discuss.

PDB Searching

The image below shows what comes up if the term "virus" is used as a search term. Useful suggestions include taxonomy (proteins from specific viruses), molecule (essentially where "virus" appears in the title) and text (where "virus" appears as text in the record). All results are links--clicking that link takes you to those specific records, which can then be narrowed further in a similar manner.

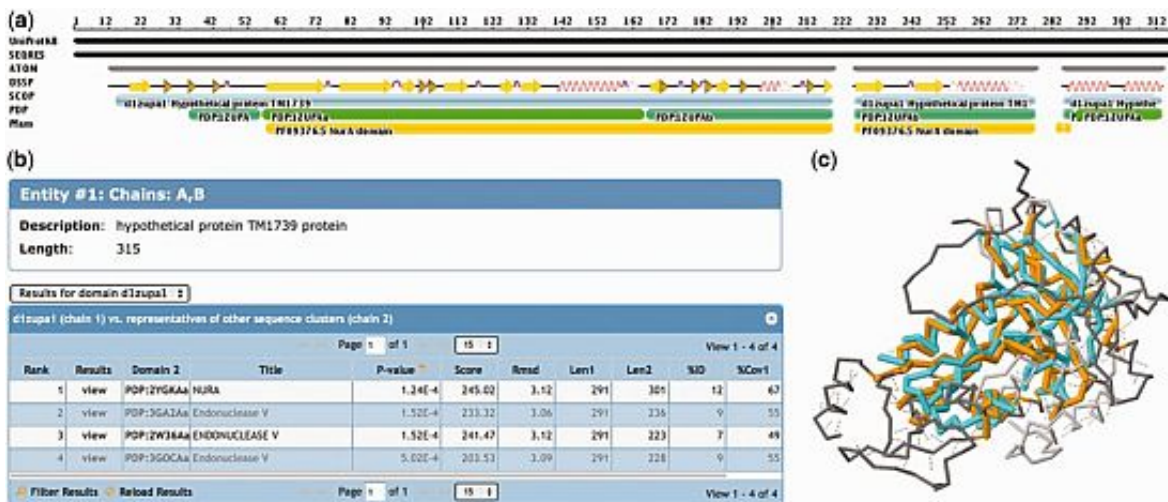
The screenshot shows the PDB search interface with the search term "virus" entered. The results are organized into several categories:

- Retrieve:**
 - Virus Structures:**
 - Adenovirus [virus]
 - Tobacco Mosaic Virus
 - Simian Virus 40
 - Dengue Virus
- Molecule of the Month:** (Empty)
- Molecule Name:**
 - Hepatitis Delta virus ribozyme (8)
 - Virus-like particle (1)
 - Hepatitis B virus receptor ... (4)
 - Hepatitis A virus cellular ... (1)
 - Subgroup ... sarcoma virus receptor ... (2)
 - Hepatitis B virus X-interacting ... (2)
- Organism:**
 - Human immunodeficiency virus 1 (992)
 - Influenza A virus (271)
 - Hepatitis C virus (216)
 - Saccharomyces cerevisiae virus L-A (L1) (1)
 - Rice black ... dwarf virus (1)
 - Chlorella virus (1)
- Taxonomy:**
 - Eukaryota (42168)
 - Bacteria (30666)
 - Viruses (5000)
 - Unassigned (3419)
 - Archaea (3262)
 - Other (456)
- PDB Text:**
 - virus
 - dirus
- Journal:**
 - Virus Res
 - Virus Res.
 - Viruses
- Structural Domains:**
 - Human immunodeficiency virus type ... (246)
 - Vaccinia Virus protein ... (374)
 - Multimerization ... sendai virus [SCOP] (1)
 - Hepatitis C Virus Capsid ... (1)
 - Multimerization ... sendai virus [SCOP] (1)
 - Cricket Paralysis Virus, Vp4 ... (1)
- Ontology Terms:**
 - regulation ... response to virus by virus ... (389)
 - Brome mosaic virus [Genome ... (2)
 - B04.820 ... Influenza A virus [MeSH ... (233)
 - response to virus (GO ... (270)
 - Acidianus two-tailed virus [Genome ... (1)
 - G06.590.875.780: Virus Replication ... (334)

Each category has a "More" link and a "Find all" link. A "close" button is located at the bottom right of the search results panel.

--image from Rose et al. (2013). *Nucleic Acids Research*. 41(D1):D475-D482.

The following sample graphic was added to all PDB records for 2013. You can see that it shows graphic secondary structure information, Pfam HMM matches and SCOP classification.



--image from Rose et al. (2013). *Nucleic Acids Research*. 41(D1):D475-D482.

NCBI Structure Database

The MMDB database is also known as the structure database at NCBI. Structure records appear first at PDB and later at NCBI because the PDB database accepts data in the PDB file format, the standard for protein structures. However, since NCBI hosts a large relational database based on ASN.1, there is a lag caused by the necessity for NCBI to convert files from the PDB format to files that are compatible with ASN.1.

NCBI has the 3-D structure viewer Cn3D available for download and you can [download the latest version](#). More recently, NCBI structure files are also viewable with RasMol, a popular structure viewer.

Protein visualization

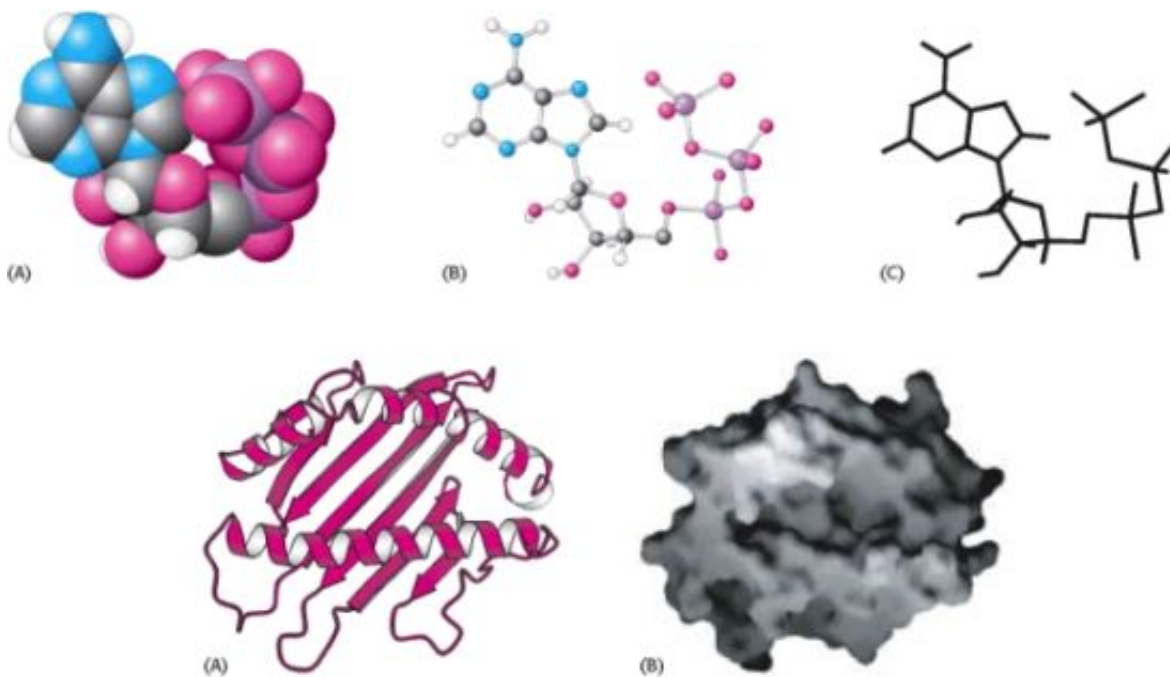
There are many ways to represent protein structures. Some of the more common ones are listed below.

Space-filling spheres: Corey, Pauling, Koltan (CPK), Van der waals radii

Balls and Sticks: Spheres for atoms, lines for bonds

Wireframes: Line drawing of bonds, atoms not shown

Ribbons: Coils for alpha-helix, arrows for beta-strands



--image from Berg et al. *Biochemistry*, 5th Edition.

The ribbon structure is a secondary structure representation; probably most common representation of proteins. A ribbon-like coil depicts an alpha-helix while a flat arrow represents beta-strand. The direction of the arrow is from N-terminus toward C-terminus.

Downloadable viewers

RasMol: <http://www.rasmol.org/>

DeepView: <http://spdbv.vital-it.ch/>

Molscript: <http://www.avatar.se/molscript/>

Grasp: http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:GRASP

Cn3d: <http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml> (MMDB files only)

Web-based viewers

JMol: <http://jmol.sourceforge.net/> (replaced WebMol at PDB)

Kiosk: at PDB

Chime: <http://www.umass.edu/microbio/chime/> (phased out--replaced by Jmol)

Broad classification schemes

Once structures are solved, they can be placed in databases where structures are classified according to structure type. Those databases can be manually curated or automatically curated into databases such as SCOP and CATH.