# Section II – Understanding the Protein Data Bank

The focus of Section II of the MSOE Center for BioMolecular Modeling Jmol Training Guide is to learn about the Protein Data Bank, the worldwide repository for large crystal structure files. Section II will emphasize using Jmol in a more sophisticated manner. Mastering the material in this section will prepare you for both designing Rapid Prototyped Models and competing in the Protein Modeling Event developed by the CBM for Science Olympiad.
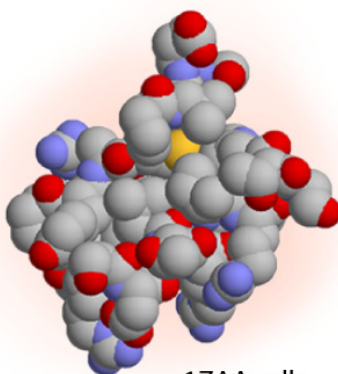
At the end of this section, you should be familiar with the RCSB Protein Data Bank and PDB files, including:

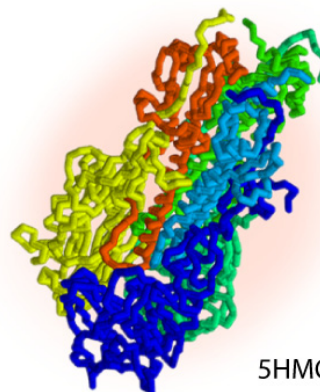**Section 2:** Understanding the Protein Data Bank

    2-1 - Using the RCSB Protein Data Bank
    2-2 - Searching the Protein Data Bank
    2-3 - Inside a PDB File
    2-4 - Referencing the PDB file
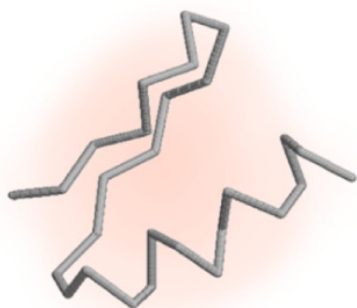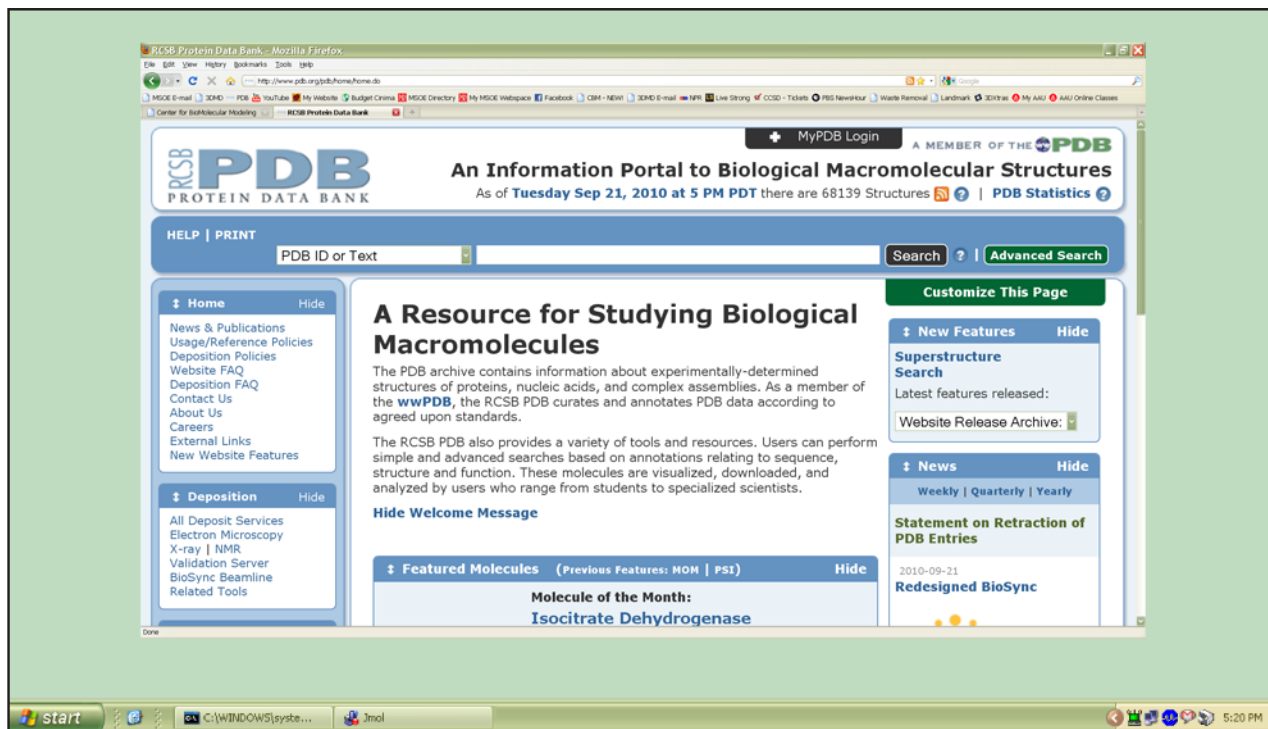

1ZAA.pdb

1ZAA.pdb

5HMG.pdb

2-1 - Using the RCSB Protein Data Bank

- The RCSB PDB (http://www.pdb.org) is the largest worldwide repository for the processing and distribution of 3-D structure data of large molecules of proteins and nucleic acids.

- PDB.org has several regularly updated features, as well as some interesting interviews and newsletters. All of these resources can be accessed from the PDB homepage.

  - **Molecule of the Month -** David Goodsell explores the structure/function relationship of a new protein each month.
    http://www.pdb.org/pdb/static.do?p=education_discussion/molecule_of_the_month/current_month.html

  - **PDB Newsletter –** This monthly newsletter includes updates from the PDB as well as articles about how people are utilizing the PDB for various projects. There have been several references to the Science Olympiad Protein Modeling Event in the recent newsletters.
    http://www.pdb.org/pdb/static.do?p=general_information/news_publications/index.html

  - **Educational Resources –** Background information on using the PDB, as well as lesson plans, tutorials, and current events are posted here.
    http://www.pdb.org/pdbstatic/education_discussion/educational_resources/education_flyer.pdf

- **Structure Summary Page -** Every PDB file in the Protein Data Bank has its own page with information about that PDB file. These pages include:

- o **PDB Structure ID Number -** This 4 letter/number ID is a unique identifier that is assigned to the crystal data file upon deposition into the database.

- o **Title -** Title of the PDB file

- o **Authors -** These are the researchers who were involved with the crystallization of the molecule.

    ⭐ The senior author or principal investigator is usually the last author in science publications. The first author is the one who did the most research related to the article.

- o **Primary Citation -** The journal article that accompanies the PDB file; excellent research resource for understanding the function of the molecule.

- o **History of Deposition and Release -** The date that the molecule was deposited into the PDB and the date the information was released to the public.

  o **Method of Structure Determination -** The method that was used to obtain the structural data (ex: NMR, X-ray diffraction).

  o **Resolution –** This indicates how accurately the position of each atom was determined. The smaller the number, the better the data.

  o **Molecular Description -** This indicates the number of chains within the molecule and the chain identity; for example in the hemoglobin file (1A3N.pdb), chains A and C are the alpha-globin molecules and chains B and D are the beta-globin molecules.

  o **Functional Class -** What type of molecule is it? (Ex: a toxin, an enzyme)

  o **Source of the Molecule -** From which species was the molecule isolated? (human, bacterium, virus, mouse)

  o **Chemical Component -** Heterologous groups that were cocrystallized with the molecule are listed here. Not all PDB files will have this section.

    ▪ The 2-3 letter identifier used to designate the chemical components contained within the file listed are recognized by Jmol.

    ▪ For example, if NAG (N-acetyl-glucosamine) is contained within the crystal, Jmol recognizes "NAG" for this file. If you enter the command "select NAG" Jmol would recognize the atoms within that chemical component of the PDB file.

- **Sequence Details -** Just above the Title there is a series of tabs - the third of which is "Sequence". This section of the PDB file page provides specific sequence information as well as secondary structure information for the molecule. You can identify the alpha helices or beta sheets as well as the amino/carboxy termini, which are the first and last amino acids of the protein, respectively.

2-2 - Searching the Protein Data Bank

- **PDB Search -** One of the key features of the RCSB Protein Data Bank is the ability to search the database for structures. You can search using key words or authors by entering these terms in the search box, located near the top center of every www.pdb.org page. If you already know the PDB structure ID number, you can enter that number in the search field. After you have entered the search terms in the field, hit enter or click on the "search" button to the right of the search field.
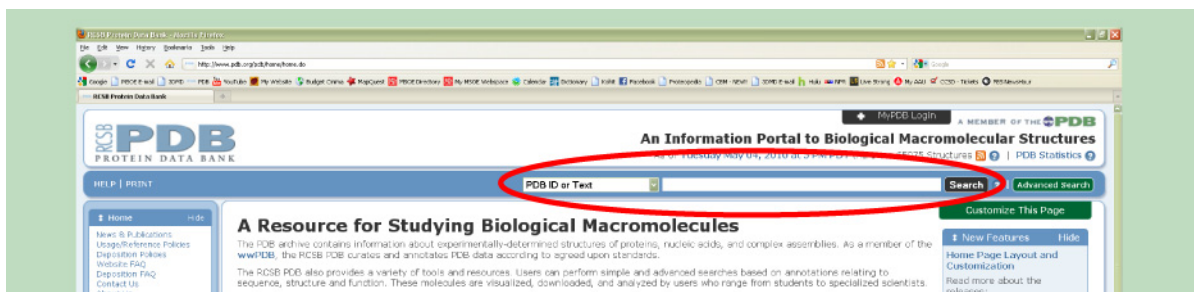
- Often more than one PDB file exists for a specific type of protein - for example, there are over four hundred PDB file entries for "hemoglobin". It is often a good idea to use this structure summary page to answer the following questions about a searched PDB file so that you can select the best possible file.

    o Who are the authors of the PDB file?

    o In which journal was the primary citation published?

    o On what date was the file deposited into the PDB?

    o How many chains are in this file?

    o Are there any heterologous groups within this PDB file? If so, which ones?

    o From what source was this molecule isolated?

    o What is the resolution of the structure?

- **Jmol Preview Window -** Every PDB file can be previewed using the online Jmol Preview Window. To view this preview, simply click the "View in Jmol" button that is located directly below the molecule image on each Structure Summary Page.

  ⭐ Note: You will not be able to design a model for printing using this preview version of Jmol.
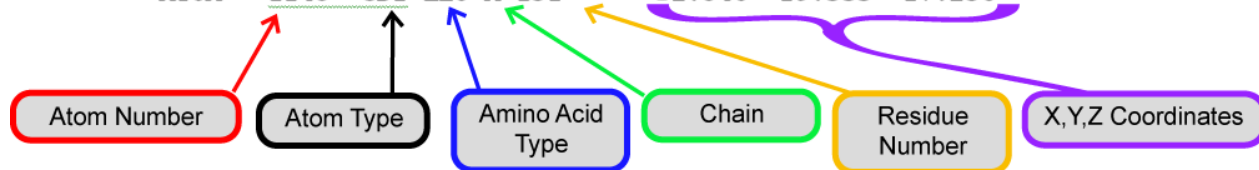
2-3 - Inside a PDB File

- A PDB file is the file format for the X,Y,Z coordinates for each atom in the molecular structure. Molecular Visualization programs like Jmol translate these X,Y,Z coordinates into

interactive 3-dimensional images, but you can also examine the file directly by opening it in any number of text editing programs such as Microsoft Word or Wordpad.

- Most PDB Files will begin with written information, termed the "Header", about the structure, the lab that determined it, and the techniques used in the lab. Any comments like this will always start with the word "REMARK" and are ignored by Jmol. The information used to create the Structure Summary page of the RCSB PDB for each PDB file is included in theHeader.

- The main portion of a PDB file will be the X,Y,Z coordinates. Although there are several similar versions of the nomenclature used to record these coordinates, the basic information remains the same. Below is a sample set of coordinates from a PDB file that has been opened with text editing software.

```
ATOM    1132   NH1  ARG A 149      31.814 -31.597   16.995
ATOM    1133   NH2  ARG A 149      32.203 -32.934   18.816
ATOM    1134   N    ASN A 150      29.346 -24.359   18.812
ATOM    1135   CA   ASN A 150      28.480 -23.190   18.933
ATOM    1136   C    ASN A 150      28.606 -22.168   17.808
ATOM    1137   O    ASN A 150      27.803 -21.276   17.678
ATOM    1138   CB   ASN A 150      28.732 -22.524   20.282
ATOM    1139   CG   ASN A 150      28.284 -23.389   21.447
ATOM    1140   OD1  ASN A 150      27.205 -23.981   21.430
ATOM    1141   ND2  ASN A 150      29.110 -23.463   22.466
ATOM    1142   N    LEU A 151      29.629 -22.313   16.996
ATOM    1143   CA   LEU A 151      29.868 -21.415   15.894
ATOM    1144   C    LEU A 151      29.953 -22.205   14.597
ATOM    1145   O    LEU A 151      30.149 -23.422   14.614
ATOM    1146   CB   LEU A 151      31.208 -20.735   16.100
ATOM    1147   CG   LEU A 151      31.436 -19.884   17.337
ATOM    1148   CD1  LEU A 151      32.846 -19.333   17.256
```

Atom Number | Atom Type | Amino Acid Type | Chain | Residue Number | X,Y,Z Coordinates

---

2-4 - Referencing the PDB

PDB Search - When you use a PDB file to create an image or a model, you must cite the PDB file in your reference list. A PDB structure entry should be cited with its PDB ID and primary reference. For example:

PDB ID: 1QYS. Kuhlman, B., Dantas, G. Ireton, G.C., Varani, G., Stoddard, B.C., Baker, D. (2003). Design of a Novel Globular Protein Fold with Atomic Level Accuracy. Science 302; 1364-1368.