A Brief Guide to the Protein Data Bank

Isaac Kunen

March 6, 2000

1 Introduction

The Protein Data Bank (PDB) is an archive for the structural data of macromolecules, primarily proteins. As of February 29, 2000, the PDB contained 11798 structures, of which 10468 were proteins. The remaining structures were for other macromolecules such as nucleic acids, carbohydrates, etc.

The PDB is currently maintained by the Research Collaboratory for Structural Bioinformatics and can be accessed at `http://www.rcsb.org/pdb/`. This web site allows for researchers to access current sequence and structure data for proteins, and download that information in an easily machine readable form that facilitates automated processing of the data.

In this document we provide a brief introduction to the PDB for those who would wish to use the data in the repository. We do not touch on how to submit new data to the database. We provide some pointers to additional information on these and other PDB related topics at the end of this paper.

2 Searching the Database

The Protein Data Bank provides several mechanisms to locate a record of interest. If one already knows the four character PDB identifier for a protein, it can be entered directly to find the entry. Given that the PDB identifier for the protein Rubisco in *Spinacia Oleracea* is 1AA1, one could simply retrieve that record.

If one does not already know the identifier, one can perform a search using the “SearchLite” form. This search allows for a keyword search across any of the data stored about the molecules. For example, we can search for “Rubisco” and the first entry returned is the entry with PDB identifier 1AA1. In addition if we know that Rubisco is a carboxylase and an oxygenase, we can perform a search for “oxygenase and carboxylase”, which will also locate our protein.

Continuing with the example, we can narrow our search by specifying which field the data must match in. If we knew that the particular sequence we were looking for was published by Taylor and Anderson, then we could perform the search “author: taylor and author: anderson”, which also yields our protein, Rubisco.

In addition to the SearchLite search, the PDB also lets one perform a “SearchFields” search, which allows for greater flexibility in specifying search criteria. This also allows for some filtering of data. For example, one could restrict the results so that only glycoproteins whose structure was found using NMR were returned.

3 Viewing and Downloading Structures

When one first “explores” a structure, the PDB presents the user with a summary page, which provides some general information about the entry such as the authors, the organism the molecule is from, etc. From here the user can elect to view other information on the molecule.

One can choose to view the three-dimensional structure of the molecule either interactively in one of several interactive 3D formats (VRML, Rasmol, FirstGlance, Protein Explorer, or in a Java applet) or as a static image.

For more information on structural neighbors, the PDB provide links to other tools on the WWW that perform this search. The PDB will automatically forward the identifying information on the sequence in question to the other tool.

The PDB maintains tables of geometric and physical data on the molecules in the database. These data include molecular weights, secondary structures, dihedral angles, and bond length information.

Finally, the PDB allows the user to view and download the complete datafile for each molecule. The file can be viewed either in its entirety, or as simply the "header". The header view does not show the atom coordinates, but does preserve all other information. Additionally, the entire datafile can be downloaded.

4 Understanding the PDB Data Files

The PDB datafiles are rather complex, but are in a format that is very amenable to computerized processing. Each line in a PDB entry begins with a keyword that identifies what type of data the line contains. Each piece of data in the line is guaranteed to fall on certain columns of the line, which aids in machine readability

For example, the keyword "AUTHOR" indicates that the line contains the names of people who discovered the structure. A "REMARK" line is one that is simply a comment and does not adhere to any particular structure. A "SEQRES" line lists part of the sequence data for a chain. As an example, this is a line from the datafile for Rubisco (PDB identifier 1AA1):

```
SEQRES 1 B 475 MET SER PRO GLN THR GLU THR LYS ALA SER VAL GLY PHE
```

The line indicates that this is a sequence line, and that it is the first line for this particular sequence. This is a sequence for the subchain named "B", and that subchain is 475 residues in length. Finally, the first 13 residues are listed.

Further information can be found on the PDB web page.

5 Further Resources

For more information about the PDB, we recommend the following WWW pages:

- http://www.rcsb.org/pdb/
  The Protein Data Bank (PDB)

- http://www.rcsb.org/pdb/downloading.html
  A simple guide to downloading structure files from the PDB

- http://www.rcsb.org/pdb/query_tut.html
  A tutorial on querying the PDB

- http://www.rcsb.org/pdb/docs/format/pdbguide2.2/guide2.2_frame.html
  The full description of the PDB file format

- http://www3.oup.co.uk:80/nar/Volume_28/Issue_01/html/gkd090_gml.html
  A paper describing the goals and workings of the PDB (Also available in PDF format at http://www.rcsb.org/images/gkd090_gml.pdf)

- http://www.rcsb.org/pdb/holdings.html
  Current statistics on the holdings of the PDB

- http://pdb.rutgers.edu/ Information on how to submit data to the PDB