Protein Data Bank

Quarterly Newsletter

Release #71 *January 1995*

January 1995 PDB Release

3091 full-release atomic coordinate entries (174 new additions)

- 2869 proteins, enzymes, and viruses
- 212 nucleic acids
- 10 carbohydrates

368 structure factor entries
31 NMR experimental entries

The total size of the atomic coordinate entry database is 1003 Mbytes uncompressed.

Table of Contents

page

The latest version of the Electronic Deposition Form should be obtained from the FTP /**pub** directory before depositing data.

What is New at the PDB

During December 1994, the number of entries in the PDB passed the 3,000 mark and has now reached 3,091. The number of entries continues to rise exponentially (Fig. 1) with a 75 percent increase in the size of the PDB in 1994 alone. In parallel, the number of accesses to PDB over the Internet has been increasing as shown in the number of FTP downloads per month (Fig. 2). This is equivalent to one download in each minute of every day – and it doesn't include the number of entries downloaded *via* the WWW!

To make network access easier, the PDB Browser has been modified so that it now runs on WWW client viewers such as Mosaic and is accessible through the PDB's HTML home page (**http://www.pdb.bnl.gov/**). This is described in the article "PDB Announces a WWW Version of the PDB Browser". See also M.C. Peitsch, T.N.C. Wells, D.R. Stampf, and J.L. Sussman, *TIBS* **20**, 82-84 (1995).

In parallel with our efforts at Brookhaven, a number of outside groups have created and developed tools that add enormously to the value of the data in the PDB, by examining them from different points of view. One such project is SCOP: a Structural Classification of Proteins Database for the Investigation of Sequences and Structures developed at the MRC Laboratory of Molecular Biology and the Cambridge Centre for Protein Engineering. This project is further described in a separate article. We encourage groups who have been developing programs or tools to extract various kinds of information from the PDB for their own requirements, or which may be of general use, to send a description of this to the PDB (newsletter@pdb.pdb.bnl.gov). We will try to provide brief articles in this Newsletter, post them on our Listserver, and where appropriate, insert hyperlinks in our Mosaic HTML home page.

PDB intends to introduce significant changes to the format of the ATOM/HETATM records. Please pay particular attention to the article "PDB Proposes Changes to ATOM/HETATM Records" describing these changes – your input is valued.

Users who are interested in receiving future printed copies of the PDB Newsletter should note that our mailing list is being re-initialized following the distribution of this issue. Please see the following article for additional information.

– Joel L. Sussman

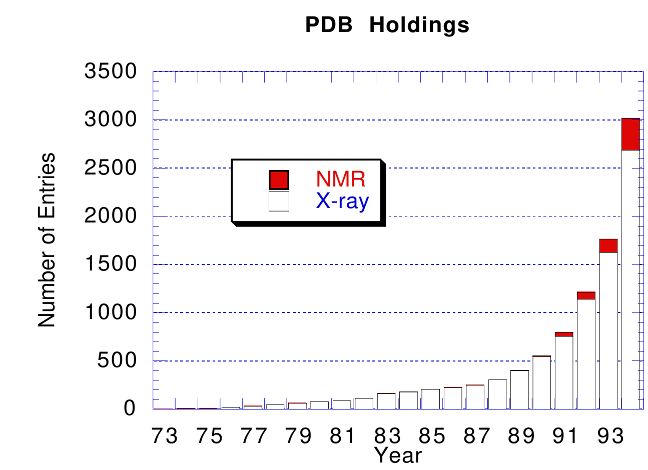

Fig. 1

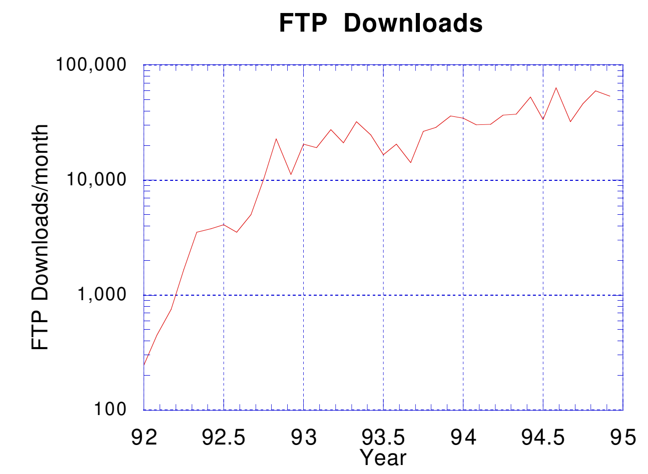

Fig. 2

Newsletter Distribution Changes

In view of the ease of retrieving the Newsletter *via* Internet (FTP, Gopher, and WWW from the subdirectory /**newsletter** in PostScript and ASCII formats), the distribution of large numbers of printed Newsletters through the postal system seems quite wasteful, especially in this era of tight budgets. Therefore, the PDB is re-initializing the Newsletter mailing list following the distribution of this issue.

Current subscribers, as well as new subscribers, who wish to receive printed copies of future Newsletters must contact us immediately. A Newsletter Request Form, similar to the one shown toward the back of this Newsletter issue, is available in the file **news_mailinglist** from the FTP /**pub** directory. This form may be completed and returned electronically (send_news@pdb.pdb.bnl.gov) or *via* the postal system (PDB Newsletter Mailing List, Chemistry Department, Building 555, Brookhaven National Laboratory, Upton, NY 11973 USA).

PDB Proposes Changes to ATOM/HETATM Records

PDB is planning a change in format of the ATOM and HETATM records. This modification addresses needs brought to our attention by many users. If acceptable, this change will take place in approximately two to three months. Along with other changes to our format, as discussed in the article "Revised PDB Format Description", these changes are part of our current efforts to produce PDB entries using CIF.

PDB will introduce this format change in such a way as to minimize negative impact on existing software. Numerous applications rely on the current coordinate record format, and we wish to give ample notice of this change. We need input from the community regarding this issue, so please examine the proposal carefully and send us your comments (abola1@bnl.gov).

PDB proposes to use columns 73 - 76 to identify specific segments of the molecule, and columns 77 - 80 to provide element information. Currently, columns 71 - 80 contain the entry's ID code and line number. If elimination of these data presents a problem to any programs, we need to be informed.

Columns 73 - 76 will contain the segment id (SEGID), which will identify specific segments of molecules. The segment can consist of a complete chain or a portion of a chain. The importance of this new field can be readily understood if one considers an antibody structure having two molecules in the asymmetric unit. Since each chain must have a unique chain identifier, the two heavy chains and two light chains cannot currently be labeled to indicate their nature. SEGID's of CH, VH1, VH2, VH3, CL, and VL would clearly identify regions of the chains and the relationship between them. Users of X-PLOR will be familiar with SEGID as used in the refinement application of X-PLOR.

SEGID is defined as a string of at most four (4) alphanumeric characters, left justified, and can include a white space, *e.g.*, CH86, A 1, NASE.

Columns 77 - 78 will contain the atom's element symbol, right justified, and columns 79 - 80 will indicate any charge on the atom, *e.g.*, MN2+, O1-, H. In the past, hydrogen naming sometimes conflicted with IUPAC conventions. For example, we have not been able to label a hydrogen HG11, but as 1HG1 in order for it not to be confused with mercury. After adopting the format change, HG11 will be allowed in columns 13 - 16, and hydrogen will be clearly identified in columns 77 - 78, thus columns 13 - 16 will continue to be used to uniquely identify each atom.

Please send your comments on these proposed changes to ATOM and HETATM records to Enrique Abola (abola1@bnl.gov). Again, if acceptable to our users, these changes will be implemented in approximately two to three months.

PDB Release Policy

To clarify PDB's policy regarding on-hold entries, the following is included in the acknowledgment letter sent to depositors upon data acceptance:

> *PDB follows the IUCr guidelines which state that coordinates may be held (before release) no longer than one (1) year and structure factors may be held no longer than four (4) years from the date of publication. PDB has chosen to apply the same guidelines to NMR restraints data, allowing a maximum hold of four (4) years. Requests that the PDB delay release of your data (put it on hold) should be submitted at the time of the initial deposition. PDB cannot consider hold requests received more than one week after the date of this acknowledgment letter. A one-time extension of up to six (6) months due to delay in publication can be requested in writing. In no case will a coordinate data set be held for longer than eighteen (18) months from the date of deposition. Deposition of data constitutes your acceptance of PDB's Release Policy.*
>
> *Please note that information on the status of all entries, including those on hold, is available to the public. The PDB Pending and Waiting list, updated daily, is available through anonymous FTP and is searchable using WWW and Gopher. This file gives the status, including hold expiration date if the entry is on hold, of every pending entry. In addition, the full list of on-hold entries, with their hold expiration dates, is available via FTP, WWW, and Gopher. See the file named /**pub/on_hold.list**.*

Revised PDB Format Description

The revised PDB Format Description for Atomic Coordinate Entries has been recently released. The purpose of this document is to completely describe the contents of PDB coordinate entry files. Several changes are being introduced to PDB files to make them more explicit to the human reader and more easily computer-parseable. Additionally, once the macromolecular CIF has been adopted, these changes will pave the way for conversion to CIF.

An important enhancement to the current PDB format is in ATOM/HETATM records (see article "PDB Proposes Changes to ATOM/HETATM Records"). Additions to PDB files include new record types, such as TITLE, CAVEAT, KEYWRD, CISPRO, MODRES, DBREF, and SEQADV; introduction of keyword:value pairs in certain records such as COMPND, SOURCE, and REMARK 3; further detailing of the heterogen groups with the new records HETNAM, HETSYN, and HETSIT (list of residues in very close proximity to a given heterogen); the deprecation of footnotes; and restructuring of some REMARK records to make them more machine-accessible.

This revised Format Description will be helpful to several communities. It will assist depositors in preparing entries for deposition, guide software and information resource developers, and help users of the PDB understand the contents of coordinate entries. Ultimately, this document and the enhanced entry format will facilitate the conversion of PDB files into CIF.

This document is available in several formats (HTML, plain text, and PostScript) from FTP, Gopher, and WWW in the /**pub** directory. A more advanced HyperText version is being designed for later release.

A convenient way to peruse the Format Description is to use Mosaic or another WWW interface. Access the PDB home page by opening the URL **http://www.pdb.bnl.gov**/. Move to "PDB Format Description" where you'll see several choices, including:

- **Table of Contents**
  Advances you to the pertinent page of the document.
- **The Format Description Document**
  Brings up the entire document.
- **Keyword Search of The Format Description**
  Allows you to locate a specific word within the document.

A sample page from the revised PDB Format Description follows:

```
CRYST1

Overview:

The CRYST1 records present the unit cell parameters, space
group, and Z value. This record is present even if the struc-
ture was not determined by crystallographic means, in which
case it simply defines a unit cube.

Record Format:

  COLUMNS      DATA TYPE      FIELD       DEFINITION

   1 -  6     Record name    "CRYST1"
   7 - 15     Real           a           a (Angstroms)
  16 - 24     Real           b           b (Angstroms)
  25 - 33     Real           c           c (Angstroms)
  34 - 40     Real           alpha       alpha (degrees)
  41 - 47     Real           beta        beta (degrees)
  48 - 54     Real           gamma       gamma (degrees)
  56 - 66     LString(11)    sGroup      Space group
  67 - 70     Integer        z           Z-value
```

```
Details:

  - If the coordinate entry describes a structure determined by
    a technique other than crystallography, CRYST1 will contain
    a=b=c=1.0, alpha=beta=gamma=90, space group P 1, and z=1.

  - The Z-value is the number of polymeric chains in a unit
    cell. In the case of heteropolymers, Z is the number of
    occurrences of the most populous chain.

  - As an example, given two chains A and B, each with a dif-
    ferent sequence, and the space group P 2 that has 2 equi-
    points in the standard unit cell, the following table gives
    the correct Z-value.

        Asymmetric Unit Content           Z-value

                 A                          2
                 AA                         4
                 AB                         2
                 AAB                        4
                 AABB                       4

Verification/Validation/Value Control Authority:

The given space group and Z-values are checked during processing
for correctness and internal consistency. The calculated SCALE
is compared to that supplied by the depositor. Packing is also
computed, and close contacts of symmetry-related molecules are
diagnosed.

Relationships to Other Record Types:

The unit cell parameters are used to calculate SCALE. If the
EXPDTA record is NMR, FIBER DIFFRACTION, or THEORETICAL
MODEL, the CRYST1 record is predefined as a=b=c=1.0, alpha=
beta=gamma=90, space group P 1, and z=1. In these cases, an
explanatory REMARK will also appear in the entry.

Deposition Form Section and Prompt:

CRYSTAL (CRYST1)
a (A)   :
b (A)   :
c (A)   :
alpha (deg)    :
beta  (deg)    :
gamma (deg)    :
space group    :
space group number      :

Z-value :

Text to explain unusual unit-cell data         :
Symmetry operations for non-standard setting   :

CIF Equivalent:

From the core CIF Dictionary:
-----------------------------

data_cell_[]
;       Data items in the _cell_ category record details about
        the crystallographic cell parameters.
;

data_cell_angle_
        loop_ _name       '_cell_angle_alpha'
                          '_cell_angle_beta'
                          '_cell_angle_gamma'
data_cell_formula_units_Z
data_cell_length_
        loop_ _name       '_cell_length_a'
                          '_cell_length_b'
                          '_cell_length_c'
data_cell_special_details
data_cell_volume
    _definition

Examples:
         1         2         3         4         5         6         7
1234567890123456789012345678901234567890123456789012345678901234567890
CRYST1   52.000   58.600   61.900  90.00  90.00  90.00 P 21 21 21    8
CRYST1    1.000    1.000    1.000  90.00  90.00  90.00 P 1           1
======================================================================
```

PDB Announces a WWW Version of the PDB Browser

Previous Newsletter articles described a GUI-based browser utility running under tcl/tk for searching the PDB archives. This browser provided the ability to search text portions of entries using arbitrary regular expressions. Additionally, the browser incorporated graphical tools such as RASMOL and MidasPlus to view selected molecules. The browser was written in a modular fashion, permitting replacement of the search, front-end, or display mechanisms.

Recently, a WWW version of the browser was made available to the user community (**http://www.pdb.bnl.gov/cgi-bin/browse**). This replaces the front end of the browser with popular WWW viewer programs such as Mosaic or Netscape. The WWW front end provides most of the functionality of the original browser while adding the following benefits:

- All searching and access is over the network. One need not install Perl, tcl, and tk. (But on general principles, you should!)

- The archive being searched is the up-to-date PDB (or a mirrored copy!).

- Hypertext links are provided in the PDB file display to the Enzyme Data Bank as well as to the sequence databases.

- Hypertext links are provided to use the default display program (*e.g.*, RasMol) and pictures from M. Peitsch.

- PDB access is provided for PC, Macintosh, and Unix computers.

A sample session is shown in Fig. 3. The primary drawback is that the search of the remark fields has been deleted due to the time necessary to complete the searches and the network timeouts that resulted.

Short-term plans include replacing search scripts with a SYBASE database engine, increasing linkages to sequence and reference databases, and taking advantage of video and audio capabilities of WWW. Exporting of the server source code to PDBSA Affiliated Centers, permitting the user community to take advantage of the best network links available, is taking place.

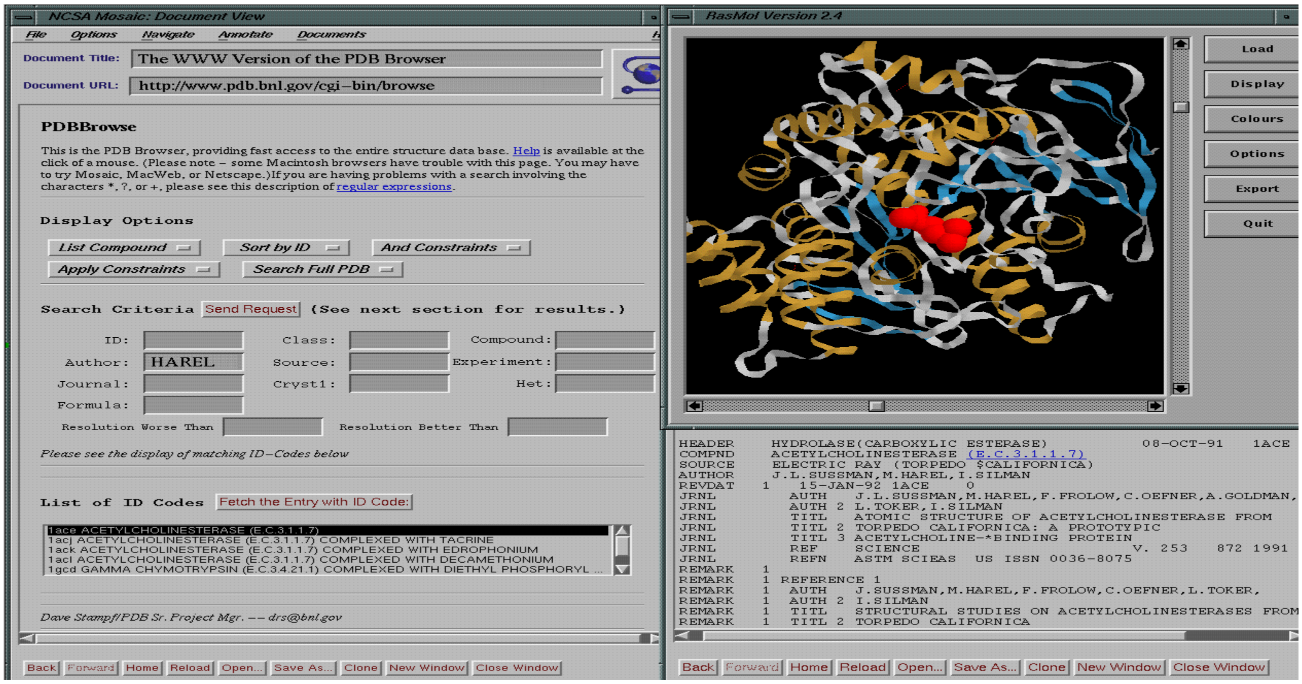

Fig.3 – The Brookhaven PDB-Browser in action [M.C. Peitsch, T.N.C. Wells, D.R. Stampf, and J.L. Sussman, *TIBS* **20**, 82-84 (1995); D.R. Stampf, C.E. Felder, and J.L. Sussman, *Nature* (1995) *in press*]. On the left is the browse screen, with windows to enter search strings in the appropriate categories. The 'Display Options' add Boolean logic to multiple search questions. The 'Search Full PDB' provides an option to search the entire PDB or to just search the list from a previous query. At the bottom is the list of selected entries. In the upper right, one of the selected entries was retrieved into the molecular viewer *RasMol* [R. Sayle, Glaxo Research & Development, Middlesex UK, 1994]. It is shown as a ribbon diagram with a model of its substrate in its active site, represented as a space filling model. On the lower right, the text of the entry is shown.

NMR Depositors

PDB would like to remind depositors of NMR entries including multiple models that all models should be presented in a common aligned orientation corresponding to the alignment shown in any related publications.

Use Care When Depositing Data

Occasionally a new set of coordinates received at PDB contains errors in the data. For example, we may find an atom distant from the rest of its residue. Sometimes we find that errors have been introduced by a depositor doing last-minute hand-editing of the data. We recommend that depositors do a final check of each coordinate file immediately before sending it to the PDB.

In addition, authors should carefully check the proposed entry that we send them, and consider all points mentioned in our accompanying letter.

Viewing the molecule on a graphics terminal also may help in detecting errors.

SCOP: a Structural Classification of Proteins Database for the Investigation of Sequences and Structures

This article was written by Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia, MRC Laboratory of Molecular Biology and Cambridge Centre for Protein Engineering, Hills Road, Cambridge CB2 2QH, England. It describes a database that complements the PDB and should be of interest to the user community.

Presently, the PDB contains 3091 entries and the number is increasing by about seventy-five per month. To facilitate the understanding of, and access to, this information, we have constructed the Structural Classification of Proteins (SCOP) database. This database provides a detailed and comprehensive description of the structural and evolutionary relationships of proteins whose three-dimensional structures have been determined. It includes all proteins in the current version of the PDB and many proteins whose structures have been published but whose coordinates are not available from the PDB.

The classification of the proteins, or the individual domains in the case of large proteins, is on hierarchical levels that embody the following evolutionary and structural relationships:

- **Family**
  Proteins are clustered together into families on the basis of clear evidence for their having a common evolutionary origin.

- **Superfamily**
  Families whose proteins have low sequence identities but whose structures and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies.

- **Common Fold**
  Superfamilies and families are defined as having a common fold if their proteins have same major secondary structures in the same arrangement and with the same topological connections.

- **Class**
  For the convenience of users, the different folds have been grouped into classes. Most of the folds are assigned to the All-alpha, All-beta, alpha and beta, or alpha plus beta classes.

Each entry (for which coordinates are available) has links to images of the structure, interactive molecular viewers, the atomic coordinates, sequence data, homologues, and MEDLINE abstracts. Two major searching facilities are currently available. Homology searching permits users to enter a sequence and obtain a list of structures which have significant levels of sequence similarity. Keyword searching finds matches from both the text of the SCOP database and the headers of PDB structure files.

The SCOP database is available as a set of tightly coupled hypertext pages on WWW. This allows it to be accessed by any machine on the Internet (including Macintoshes, PCs, and workstations) using free WWW reader programs, such as Mosaic. Once such a program has been started, it is necessary only to open URL **http://scop.mrc-lmb.cam.ac.uk/scop/** to obtain the home page of the database.

The SCOP database was originally created as a tool for understanding protein evolution through sequence-structure relationships and determining if new sequences and new structures are related to previously known protein structures. In a more general way, the highest levels of classification provide an overview of the diversity of protein structures now known, and would be appropriate both for researchers and students. The specific lower levels should be helpful for comparing individual structures with their evolutionary and structurally-related counterparts. In addition, the search capabilities with their easy access to data and images make SCOP a powerful general-purpose interface to the PDB.

For comments or additional information, please contact Steven Brenner (scop@www.bio.cam.ac.uk).

The EBI NetNews Filtering Service

This article was written by Jose R. Valverde and Rob Harper, European Bioinformatics Institute, an EMBL Outstation, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, England. It describes one way to access Usenet News that may be of interest to users who are overwhelmed with the information available.

Usenet News is possibly the most popular tool used today by researchers for efficient communication. It provides fast dissemination of news, fast on-line help, and a free forum for scientific discussion. However, the newsgroups either tend to be too specialized for interdisciplinary research (thus requiring a scientist to follow many groups), or they address too broad an interest (which results in information overload).

The NetNews Filtering Service provides an easy way in which researchers can filter newsgroups automatically so that they will only receive those articles in which they are interested. A user can define one or several independent profiles that describe his or her fields of interest by means of keywords. The server will then periodically check the profiles against all the newsgroups in Bionet, EMBnet, and Sci and select only those articles that best reflect the interests of the user. The filter makes a preliminary review and selects a small number of articles, and, if the user finds any of these especially interesting, the server can be instructed to mail the full contents.

The EBI NetNews Filtering Service can be accessed in either of two ways: giving commands by e-mail or through an easy graphical user interface using a WWW client (*e.g.*, NCSA Mosaic). In the first case a user can start by sending an e-mail message to netnews@ebi.ac.uk with no subject and a body consisting of a single line with the word “**help**”. Working with Mosaic or another WWW browser, a user can select the URL **http://www.ebi.ac.uk** which will connect to our server and then select “Documentation Software and Services” from the main menu. In both cases the user can get on-line help, look at examples, and make tests before subscribing.

For comments or additional information, please contact Jose R. Valverde (Jose.Valverde@ebi.ac.uk).

MakeMolS

This article was written by Liisa Holm, European Molecular Biology Laboratory, Heidelberg, Germany. It describes a tool that may be of interest to a wide cross section of PDB users.

MakeMolS is a simple tool to facilitate the generation of input scripts to MolScript [P. Kraulis, *J. Appl. Cryst.* **24**, 946-950 (1991)], a popular program for creating molecular graphics in PostScript form. MolScript supports a variety of representations of protein structures, including Jane Richardson-type schematic ribbon drawings to highlight secondary structure elements. The sole functionality of MakeMolS is to read the secondary structure definitions from a DSSP file [W. Kabsch and C. Sander, Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **22**, 2577-2637 (1983)] and write them out in the form of MolScript instructions for a ribbon drawing as illustrated in Fig. 4 below. The DSSP program and database are available from the URL **http://www.sander.embl-heidelberg.de/dssp/** and from anonymous FTP (**ftp.embl-heidelberg.de**) in the following directories: **/pub/databases/protein_extras/dssp** and **/pub/software/unix/dssp**. The Fortran source code of MakeMolS may be obtained from the PDB FTP directory **/pub/program_tape/makemols.f**. The user may further modify the scripts in a text editor to exploit the full spectrum of options provided by MolScript.

For comments or additional information, please contact Liisa Holm (HOLM@embl-heidelberg.de).

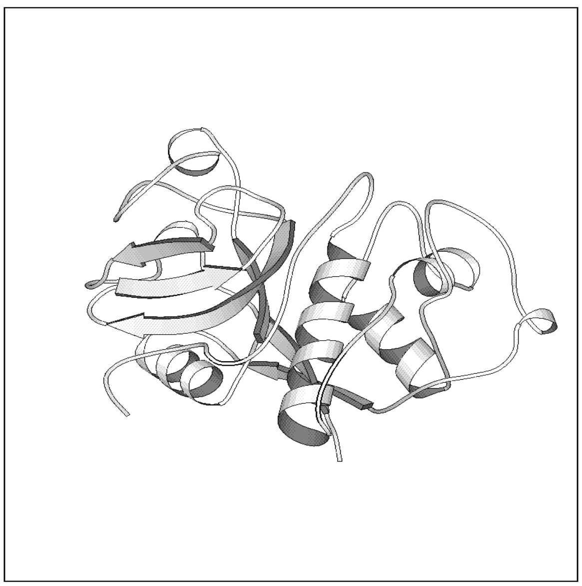

Fig.4 – Ribbon drawing of papaya fruit latex papain [PDB entry 1ppn, R.W. Pickersgill, G.W. Harris, and E. Garman, *Acta Cryst.* **B48**, 59 (1992)] by the MolScript program using a script automatically generated by MakeMolS.

ZINC - Galvanizing CIF to Work with Unix

The July 1994 Newsletter described the basic structure of a CIF (Crystallographic Information File), the standard that will probably become the interchange format of the future for the PDB due to increased amounts of information that it can contain. It was also pointed out that those who are accustomed to working with the PDB format in a Unix environment (with grep, awk, perl, diff, etc.) will not be able to use those skills in dealing directly with CIF.

This article describes a format called ZINC (Zinc Is Not CIF) which is fully accessible to Unix tools, and a number of utilities that allow a CIF to be converted into a ZINC and back again, as well as versions of some familiar Unix tools (grep, diff) and some surprising new tools (zincSubset and zincNl) that should make access to CIF much easier.

What is the Problem with CIF in a Unix Environment?

CIF defines a format that is generally at odds with Unix tools.

- Many Unix tools are line oriented – they expect related information to be on a single line whereas CIF allows and encourages the use of multiple lines, both by limiting the line length and in the definition of loops.

- Many Unix tools break the lines into fields based on a separator character. Both PDB and CIF formats work against this, the former by having column-based fields, the latter by encouraging the liberal use of white space (across line boundaries).

- A number of Unix tools treat the information in files as being position dependent (diff, head, tail, etc.). The PDB file format adapted neatly to this, but CIF allows a much wider variation in the placement of data.

– ZINC

Zinc is not an interchange format as CIF is, but rather a piping format, *i.e.*, a format that makes the contents of a CIF accessible to Unix utilities. Each data line of a ZINC file consists of five tab-separated fields:

```
block    name   index    value    loop-id
```

The first field is the name of the CIF data block (the data_ prefix is omitted) and is repeated on each line where appropriate. The second field is the name of the data item. The third field is an index specifier, which is empty for non-looped data, and is a zero-based index for looped data. The fourth field is the data item itself. For multiple-line CIF data, new line characters are replaced by the two characters \n. The backslash character becomes the escape character throughout the ZINC format. The fifth field is a loop identifier. Comments that appear in a CIF are associated with the previous token and are also represented in the ZINC format.

– Existing Tools

A number of tools have been developed to support the user community in using the ZINC format and to access the information contained in a CIF. Most are simple and allow users to modify the code to tackle new problems. Source is provided for all.

- **cifZinc**, obviously the first required tool, converts an existing CIF into a ZINC. This is a C program that converts even the largest CIF in a few seconds.

- **zincCif**, the next most important tool, takes a ZINC and creates a pretty-printed CIF. Very often the pipeline:

  ```
  cifZinc old.cif | zincCif > new.cif
  ```

  will produce a better looking CIF than the original.

- **zincGrep**, greps a ZINC (or a CIF file specified as a command-line argument) for a regular expression and returns the block name, data name, index, and value of the match. This is the single most requested tool for dealing with CIF.

- **cifdiff** is a four-line C-shell script that takes two CIFs and determines the differences between them. It will handle CIFs that have been rearranged, and even loops with rearranged columns, and provide only the real differences.

- **zb**, a small (< 200 lines) tcl/tk program that provides a simple GUI front end to a ZINC or CIF, allows users to browse through the contents. Multiple files as well as multiple data blocks can be viewed simultaneously on any X terminal. zb recognizes command-line argument file names in the form *.cif as being a CIF and converts it to a ZINC automatically.

- **zincSubset** is another C-Shell script that is very short but very useful. It allows users to generate a custom subset of any ZINC (CIF), simply by listing the data blocks and data names that he or she wishes to include. For example, if you wanted to extract only the names and definitions from the mmCIF dictionary, you would create a file (*e.g.*, defs) with two lines that appear as:

  _name
  _definition

  (preceded and followed by tabs) and run the command:

  ```
  zincSubset defs mmcif94 | zincCif
  ```

- **zincNl**, a perl script, takes a ZINC file and creates a Fortran compatible namelist file allowing easy access to any CIF by Fortran programs without the need for extensive I/O libraries or reprogramming. As with zb above, it will automatically convert a CIF to a ZINC. It may be used in the following pipeline which extracts the coordinates from a CIF and presents them to a Fortran program *via* the namelist mechanism:

  ```
  zincSubset coords datafile | zincNl | myfortran
  ```

  (where coords is a file that simply has three lines with x, y, and z surrounded by tabs).

– A Sample CIF and ZINC

The following CIF illustrates most aspects involved in translating to a ZINC:

```
#
#     A simple CIF
#

data_object

#
#     polygon
#
  _name
;
triangle
;
  loop_
      _x _y
      0.0      0.0
      1.0      0.0
      0.0      1.0

      _num_sides 3
```

In ZINC, this would appear as:

```
           (          0    #
           (          1    #    A simple CIF
           (          2    #
object     (          3    #
object     (          4    #    polygon
object     (          5    #
object     _name           ;\ntriangle\n;
object     _x         0    0.0                 _x
object     _y         0    0.0                 _x
object     _x         1    1.0                 _x
object     _y         1    0.0                 _x
object     _x         2    0.0                 _x
object     _y         2    1.0                 _x
object     _num_sides      3
```

Note that comments “belong” to a data block and are represented with an open parenthesis for the data name. Initially, the data block name is defined to be the null string.

– Code

The formal definitions of ZINC and the above mentioned programs are available from PDB using FTP, Gopher, or WWW (**ftp://pub/other-software/Zinc**). Please give it a try. All are invited to submit their favorite scripts that use ZINC. Any comments should be directed to Dave Stampf (drs@bnl.gov).

Mirroring

Earlier this year, PDB started updating the FTP server with fully-released entries on a frequent basis – typically every few weeks. PDB users often ask how they can keep their local archives up to date without having to continually check the current holdings of the PDB. Fortunately, a public domain package called *mirror* exists that does exactly this. This program is a perl script that runs on your system and periodically creates an FTP connection to the PDB, determines the difference between your local archive and the PDB (including deletions!), and performs all necessary tasks needed to make them equivalent. We have used this program with the Weizmann Institute of Science in Israel, Turku University in Finland, the European Molecular Biology Laboratory in Germany, and the European Bioinformatics Institute in England with great success even over very poor network connections. We encourage the members of our user community who maintain complete local archives to do the same.

If you wish to try this, here is what you need to do:

0) **Think:** Decide what your local archive will look like. Will it be a virtual image of the PDB? Will it only hold compressed files? Will it use the all_files in one directory scheme, or the 2-character directory scheme? Your space limitations and your usage patterns will determine what is best for you.

1) **Prepare:** Set up your local archive from a 1994 version of the PDB CD-ROM. Be sure the dates on the files match those on the CD. In order to copy a directory on the CD to a local disk without changing the dates, use the following tar pipeline:

```
mkdir /usr/distr      or where you want the files to go

cd /CDROM/distr       or where you are copying the files
                      from

tar cf - . | (cd /usr/distr; tar xvf -)
```

2) **Get mirror:** Copy the mirror software from PDB or elsewhere (/**pub**/**other-software**/**Mirror**). There is nothing to compile, but you must have already installed perl and have run the h2ph perl script. Install the mirror program in one of the standard locations.

3) **Configure:** Adapt a copy of the mirror.defaults configuration file. A sample of this file is also on the PDB FTP server.

4) **TEST!!!** If you run mirror -n it will tell you what it will do without doing it. If you are about to transfer a gigabyte over a megabit line, you may wish to reconsider.

5) When you are satisfied, make the mirror run as part of your crontab and forget about it. Your local archive will be kept up to date “automagically”!

As the mirror README says, “Objects in the mirror are closer than they appear!”

Access to PDB

– World Wide Web (WWW)

PDB has a World Wide Web (WWW) server on the computer system **www.pdb.bnl.gov** (130.199.144.1). This server is accessible using the document URL **http://www.pdb.bnl.gov/**.

Besides including links to the PDB FTP and Gopher servers, the WWW server includes links to many other useful databases and information servers.

– Gopher

PDB has a Gopher server on the system **gopher.pdb.bnl.gov** (130.199.144.1). This server is accessible using a Gopher client connecting to the following link:

```
Name   =   Protein Data Bank FTP server
Type   =   1
Host   =   gopher.pdb.bnl.gov
Port   =   70
Path   =   1/
```

As a Gopher client, you may navigate through a hierarchy of directories and documents or ask an index server to return a list of all documents that contain one or more specified words. For instance, you can choose “The PDB Anonymous FTP” after reaching PDB’s Gopher server in order to search and download the same information and coordinate files as through FTP. Alternatively, you can select “An (almost) full-text search of the PDB Bibliographic Headers” in order to search PDB using any keyword**.**

– FTP

PDB has an anonymous FTP account on the computer system **ftp.pdb.bnl.gov** (Internet address 130.199.144.1). Files may be transferred to and from this system using anonymous as the FTP user name and your e-mail address as the password. Besides downloading entries, data files, and documentation, it is possible to upload any files that you may wish to send to PDB, into the special directory /**new_uploads**. Those using VMS may need to place quotes around file names.

– Listserv

PDB has a mailing list devoted to discussions concerning its operation, contents, and access procedures.

To subscribe, send e-mail to listserv@pdb.pdb.bnl.gov with the one-line message: **subscribe PDB-L Firstname Lastname**.

To find out what can be done with this mailing list, send e-mail to the same address (listserv@pdb.pdb.bnl.gov) with the one-line message: **help**.

To send a message to all PDB-L subscribers, e-mail the message to: **PDB-L@pdb.pdb.bnl.gov**.

BROOKHAVEN ORDER FORM

Name of User	______________________	Date	______________
Organization	______________________	Phone	______________
Address	______________________	Fax	______________
	______________________	E-mail	______________
	______________________		

- ***Price is valid through September 30, 1995***
- ***Price shown is per release - there are four releases per year***
- ***Facsimile and phone orders are not acceptable***

The Protein Data Bank ***MUST*** receive all three of the following items before shipment can be completed (please send all required items together *via* postal mail — facsimile and phone orders are ***NOT*** acceptable):

1. Completed order form;
2. Mailing label indicating exact shipping address;
3. Payment (using one of the two options below):

☞ Check payable to Brookhaven National Laboratory in U.S. dollars and drawn on a U.S. bank. Foreign checks are not accepted and will be returned.

☞ Original purchase order payable to Brookhaven National Laboratory. After your order is processed, you will be invoiced by Brookhaven National Laboratory.

A wire transfer is acceptable only ***AFTER*** we have received an original purchase order from your organization and you have been invoiced by Brookhaven. After receiving Brookhaven's invoice, your bank may send a wire transfer to:

Bank name	:	Morgan Guaranty Trust Company of New York
Account name	:	Brookhaven National Laboratory
Account number	:	076-51-912

Please send all three required items together *via* postal mail to:

Protein Data Bank Orders
Chemistry Department, Building 555
Brookhaven National Laboratory
P.O. Box 5000
Upton, NY 11973-5000
USA

Protein Data Bank CD-ROM — ISO 9660 Format *$300.00*

(tax and shipping charges not applicable)

NEWSLETTER REQUEST FORM

Current subscribers, as well as new subscribers, who wish to receive printed copies of future Newsletters must contact us immediately. A form similar to the one below is available in the file **news_mailinglist** from the FTP /**pub** directory. It is preferable that this form be retrieved, completed, and returned electronically to:

send_news@pdb.pdb.bnl.gov

If Internet capabilities are unavailable, please complete this form and send it by postal mail to: PDB Newsletter Mailing List, Chemistry Department, Building 555, Brookhaven National Laboratory, Upton, NY 11973 USA.

```
Please send me future printed copies of the Newsletter:

prefix:__________ first name:_________________________________ middle name:___________________

last name:_________________________________________________________________ suffix:__________

professional_title:____________________

department_1:________________________          department_2:__________________________

institute_1:_________________________          institute_2:___________________________

po_box:______________________________  bldg:_____   street_address:_______________________

city:________________________________          state:_________________________________

postal_code:_________________________          country:_______________________________

telephone_number:____________________          fax_number:____________________________

network_address_1:___________________          network_address_2:_____________________
```

AFFILIATED CENTERS

Twenty-two affiliated centers offer DATAPRTP information for distribution. These centers are members of the Protein Data Bank Service Association (PDBSA). Centers designated with an asterisk(*) may distribute DATAPRTP information both on-line and on magnetic or optical media; those without an asterisk are on-line distributors only.

BMERC
BioMolecular Engineering Research Center
College of Engineering, Boston University
Boston, Massachusetts
Kathleen Klose (617-353-7123)
klose@darwin.bu.edu

*BIOSYM
BIOSYM Technologies, Inc.
San Diego, California
Laurel Frey (619-546-5509)
rcenter@biosym.com or laurel@biosym.com

CAN/SND
Canadian Scientific Numeric Data Base Service
Ottawa, Ontario, Canada
Roger Gough (613-993-3294)
cansnd@vm.nrc.ca

CAOS/CAMM
Dutch National Facility for Computer Assisted Chemistry
Nijmegen, The Netherlands
Jan Noordik (+1 31-80-653386)
noordik@caos.caos.kun.nl

*CCDC
Cambridge Crystallographic Data Centre
Cambridge, United Kingdom
David Watson (+1 44-223-336394)
dgw1@chemcrys.cam.ac.uk

CSC
CSC Scientific Computing Ltd.
Espoo, Finland
Heikki Lehvaslaiho (+1 358-0-457-2076)
heikki.lehvaslaiho@csc.fi

CINECA
NE Italy Interuniversity Computing Center
Casalecchio di Reno (BO), Italy
Laura Setti (+1 39-51-6599478)
asltc0@icineca.cineca.it

ICGEB
International Centre for Genetic Engineering and Biotechnology
Trieste, Italy
Sandor Pongor (+1 39-40-3757300)
pongor@icgeb.trieste.it

EMBL
European Molecular Biology Laboratory
Heidelberg, Germany
Hans Doebbeling (+1 49-6221-387-247)
hans.doebbeling@embl-heidelberg.de

INN
Israeli National Node
Weizmann Institute of Science
Rehovot, Israel
Leon Esterman (+1 972-8-343934)
lsestern@weizmann.weizmann.ac.il

*JAICI
Japan Association for International Chemical Information
Tokyo, Japan
Hideaki Chihara (+1 81-3-5978-3608)

*MAG
Molecular Applications Group
Palo Alto, California
Hilary Jensen (415-473-3039)
hilary@suerte.mag.com

*MSI
Molecular Simulations Inc.
Burlington, Massachusetts
Lance J. Ransom Wright (617-229-9800)
lance@msi.com

NCHC
National Center for High-Performance Computing
Hsinchu, Taiwan, ROC
Jyh-Shyong Ho (+1 886-35-776085; ex: 342)
c00jsh00@nchc.gov.tw

NCSA
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Champaign, Illinois
Patricia Carlson (217-244-0768)
pcarlson@ncsa.uiuc.edu

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, Maryland
Stephen Bryant (301-496-2475)
bryant@ncbi.nlm.nih.gov

*OML
Oxford Molecular Ltd.
Oxford, United Kingdom
Steve Gardner (+1 44-865-784600)
steve@gardner.demon.co.uk

*Osaka University
Institute for Protein Research
Osaka, Japan
Yoshiki Matsuura (+1 81-6-879-8605)
matsuura@protein.osaka-u.ac.jp

Pittsburgh Supercomputing Center
Pittsburgh, Pennsylvania
Hugh Nicholas (412-268-4960)
nicholas@cpwpsca.psc.edu

SDSC
San Diego Supercomputer Center
San Diego, California
Lynn Ten Eyck (619-534-8189)
teneyckl@sdsc.edu

SEQNET
Daresbury Laboratory
Warrington, United Kingdom
User Interface Group (+1 44-925-603351)
uig@daresbury.ac.uk

*Tripos
Tripos, Inc.
St. Louis, Missouri
Akbar Nayeem (314-647-1099; ex: 3224)
akbar@tripos.com

Protein Data Bank
Chemistry Department, Bldg. 555
Brookhaven National Laboratory
P.O. Box 5000
Upton, NY 11973-5000
USA

Statement of Support

PDB is supported by a combination of Federal Government Agency funds (work supported by the U.S. National Science Foundation; the U.S. Public Health Service, National Institutes of Health, National Center for Research Resources, National Institute of General Medical Sciences, and National Library of Medicine; and the U.S. Department of Energy under contract DE-AC02-76CH00016) and user fees.

To Contact PDB

Telephone .... +1 516-282-3629
Facsimile... +1 516-282-5751

Internet:
pdb@bnl.gov............................general correspondence
orders@pdb.pdb.bnl.gov ....order information
sysadmin@pdb.pdb.bnl.gov .....network services
listserv@pdb.pdb.bnl.gov ....Listserver subscriptions
pdb-l@pdb.pdb.bnl.gov.............Listserver postings
errata@pdb.pdb.bnl.gov ....entry error reporting

Please include your name, postal mailing address, e-mail address, facsimile number, and telephone number in all correspondence.

PDB Staff

Joel L. Sussman, Head
David R. Stampf, Sr. Project Mgr.
Enrique E. Abola, Science Coordinator

Frances C. Bernstein
Judith A. Callaway
Minette Cummings
Betty R. Deroski
Pamela A. Esposito
Arthur Forman
Patricia A. Langdon
Michael D. Libeson
Nancy O. Manning
John E. McCarthy
Regina K. Shea
John G. Skora
Karen E. Smith
Dejun Xue

First Class