PDBBrowse — a graphics interface to the Brookhaven Protein Data Bank

David R. Stampf, Clifford E. Felder and Joel L. Sussman

With a few keystrokes and mouse button clicks on a UNIX workstation, the entire Protein Data Bank archive is now available for searching and displaying the three-dimensional protein structures contained within.

THE Brookhaven Protein Data Bank (PDB) has developed an X-windows-based interactive browser for searching PDB's database of three-dimensional (3D) structures of biological macromolecules. It greatly enhances the PDB's printed index listings and various *ad hoc* search protocols that have been developed for finding PDB entries. Its highly modular and flexible design allows for rapid modification to meet changing needs. Multiple search strings covering various search fields, corresponding to the different PDB entry header record types (such as compound, header, author, or biological source), are supported, using boolean 'and', 'or' and 'not' operators. The selected entry files can be retrieved automatically, and the molecular structures can be displayed using the public-domain X-based molecular viewer RasMol (or a similar viewer).

PDB background

The PDB archives the experimental findings describing, to atomic resolution, the 3D structures of more than 3,000 proteins, nucleic acids and other biological macromolecules. Each one of these data entries is an annotated ASCII text file, identified by a unique 4-character ID-CODE. The average size of a typical entry is over 4,000 lines or 300 kb, and the entire database, currently being updated about every three weeks, is doubling in size every two years.

This amount of data is beginning to present a significant problem for those wishing to search the entire collection. For example, a simple *ad hoc* author search on an MIPS R3000 SGI workstation at 33 MHz, using the UNIX 'head' and 'grep' commands, took 10 minutes of clock time, compared with 12 seconds with the browser. Such a search becomes prohibitively slow if the collection is mounted on multiple CDs. Even searching index listings, both PDB-supplied and those created by its users, which investigators have relied upon until now, is becoming too inefficient and unwieldy. Typically such indices cover only a subset of possible search fields, such as compound and author name, which makes it impractical to search other data fields (for example, by heterogens or experiment type) or combinations of search strings.

To remedy this situation, the PDB now provides a browser utility that allows a complete text search of the PDB entries online in a friendly, X-windows-based environment. All components of this browser have been written using public domain products, which are freely available to users, as is the source of the browser.

FIG. 2 The PDB browser in action. Above is one of the selected entries, for example, acetylcholinesterase (1ACE) was retrieved into the molecular viewer RasMol. At right 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.

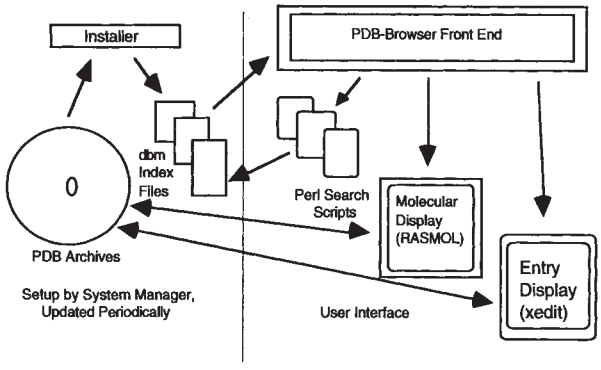


FIG. 1 Modular design of the PDB browser: The index files are written in UNIX hashed dbm database format, and the scripts in Perl[3] command interpreter language.

A number of other tools have been developed to aid in searching the PDB. Some of these are commercial products, which have search capabilities not provided by the PDB browser, for example, IDITIS (marketed by Oxford Molecular, Mountain View, California), MacImdad (marketed by Molecular Applications Group, Palo Alto, California) and others. In contrast, the PDBBrowse, as well as a number of search tools developed in other laboratories, are in the public domain. One such utility is SCOP (Structural Classification of Proteins), which has recently been developed by A. Murzin, S. Brenner, T. Hubbard and C. Chothia at the MRC Laboratory of Molecular Biology and Centre for Protein Engineering, Cambridge, UK[1]. It provides a detailed and comprehensive description of the structural and evolutionary relationships of proteins whose 3D structures have been determined and is available on the World Wide Web (WWW)[2], (URL:http://scop.mrclmb.cam.ac.uk/scop or ttp://ncbi.him.hih.gov/repository/scop/index.html). Although there is some overlap with the PDBBrowse, SCOP is structured completely differently, and thus complements rather than duplicates PDBBrowse.

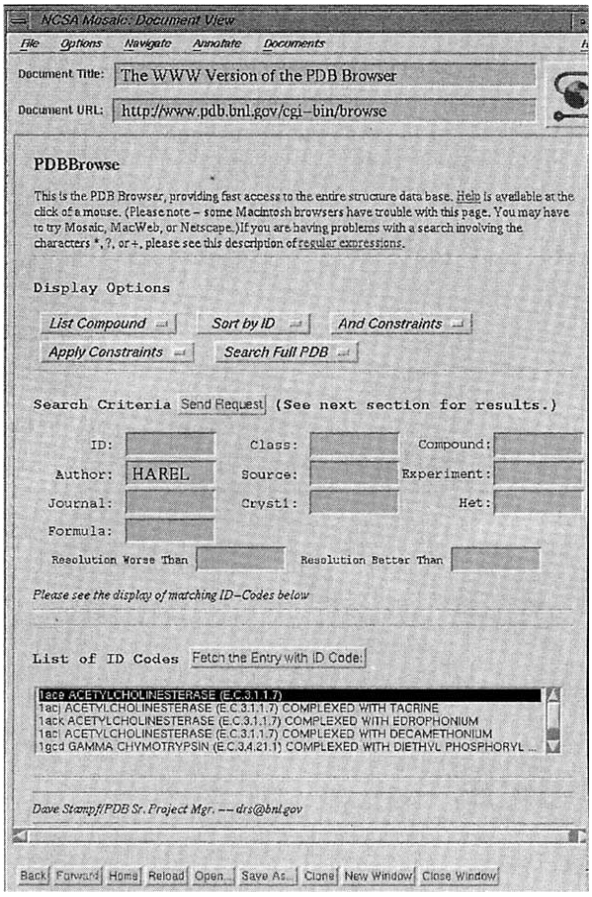


FIG. 3 The PDBBrowse showing the path to find acetylcholinesterase (1ACE)[7].

Modular architecture

The browser utility, consisting mostly of a series of Perl[3] scripts, is constructed in a modular fashion, as shown in Fig. 1. First, the installation Makefile invokes a Perl script to create the different index files, one for each PDB header record type, in UNIX hashed dbm format. Second, to use the browser, one starts up its friendly, window front end, called simply 'browse', one version of which runs under the public-domain command language and windowing management utility 'tcl/tk'[4]. The modular design has made it easy to modify the current tcl/tk front end with an 'html' form in order to support a second front-end, accessible via the WWW (URL: http://www.pdb.bnl.gov)[5]. The actual searches are done using a set of Perl scripts called by the front end, that search the index files. (Users without access to an X-windows or WWW interface can run searches by invoking these scripts directly.) The selected entries can then be retrieved and displayed by user-specified utility programs called from the front end.

With the modular design of the browser, each component is independent of the others, so if the format of the PDB collection were to be modified or changed drastically, only the scripts creating the index files would need to be changed. Additional indices covering new search fields can be added easily at any time, as can new methods for viewing or displaying the retrieved data.

With this arrangement, every time the PDB collection at a given installation is updated, the index files also have to be updated. This can be accomplished quite easily, however, by downloading or installing the updated index files that are provided with the PDB distribution. (Nevertheless, the file location index, which is unique to each site, must be updated locally by typing the command 'hash.pl'.) This capability allows those sites that choose to have their PDB holdings updated automatically over the network, using the 'mirror' application, to have the index files updated automatically as well, and in a completely transparent way.

Usage

As shown in Figs 2 and 3, the top portion of the browser front end allows the user to enter search strings into one or more data fields that correspond to the different record types in the header portion of PDB entries. These include, but are not limited to, the chemical compound, the source of the protein, its functional classification, the experimental technique, heterogen molecules present in the structure, author names, crystallographic data, resolution, and remarks. One can combine search strings using .AND. or .OR. logical combinations, using the set of boolean logical switches located below the search fields. Subsequent searches can be performed on the entries selected. (Running a completely new search requires clearing all fields, via the Edit menu.) A online summary of each entry selected by the search is displayed at the bottom, in a format that the user can modify using the 'List Format' menu on the top. The user can then select any of these entries to retrieve and display via the standard X-windows editor 'xedit', (or a different editor selected by the user in environment variable $EDITOR), and display the 3D molecular picture via the public-domain molecular viewing application 'RasMol'[6] (or a different molecular display program specified by the user in environment variable $MOLVIEW).

For particular applications, it is quite easy for the end user to modify the browser, to add custom search scripts and include the corresponding filters and front-end entry windows. For example, a Perl script (simhead.pl) was written that defines a set of equivalence classes on the PDB, based upon the functional classification of the entry. Using this custom script, the browser located all 136 entries with classifications matching those found in the HAREL search illustrated in Fig. 3. The elapsed time for this search was under 4 seconds. One could also envision custom scripts that constrain the search by the volume or shape of the molecule. Everyone is invited to contribute custom scripts to the PDB for possible inclusion in future releases of the browser.

Installation information

The PDB browser may be obtained over the Internet by anonymous ftp to ftp.pdb.bnl.gov, gopher to gopher.pdb.bnl.gov, or WWW to http://www.pdb.bnl.gov/. It is also located on the PDB's CD-ROM releases of July 1994 and later. Source codes for the complete browser package, as well as distribution releases for the public-domain products Perl[3], tcl/tk[4] and RasMol[6], will all be found under directory /pub/pdbbrowse, each in its own separate subdirectory. Before downloading and installing, be sure to examine file 'ReadMeFirst' for complete instructions. For optimal performance, the complete PDB distribution should be mounted locally to the system in use. However, having a remote mount to a disk or CD on another workstation is also quite satisfactory. If neither is possible, the program can access the PDB on another distribution site by using ftp automatically. The time necessary to install the entire package (Perl, Tcl/tk or Mosaic, RasMol and the browser itself) should be under 3 hours. In case of difficulties, you are invited to send electronic mail to the browser's author, David Stampf, stampf@bnl.gov.

Future activities

By the time this article is published, we expect to have released a version of the browser that uses index files at the PDB via a network connection. Thereafter, as updates to the PDB are released by any method of distribution, these index files will also be updated automatically with them.

This browser was inspired by a similar program for MS DOS/MS Windows, called pdbshell, that was written by Enrique Abola and Eugene Ko. That program is based on the Foxpro database program and is also available on the PDB file server and CD-ROM under the directory /pub/pdbshell. A version of that browser for Macintosh is currently under development, as RasMol has recently been released for the Macintosh. □

The PDB is supported by the US National Science Foundation, the US Public Health Service, the US National Institutes of Health

(National Center for Research Resources, National Institute of General Medical Sciences and the National Library of Medicine) and the US Department of Energy and user fees. We wish to thank Enrique Abola and the members of PDB for their help with this work. D.R.S. is in the Department of Chemistry, Brookhaven National Laboratory, Upton, New York 11973 USA. C.E.F. is in the Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100 Israel. J.L.S. holds a joint appointment with the Department of Structural Biology at the Weizmann Institute of Science and Brookhaven National Laboratory. For more information, fill in reader service number 100.

1. Barton, G. J. *TIBS* **19**, 554–555 (1994).
2. Schatz, B. R. & Hardin, J. B. *Science* **265**, 895–901 (1994).
3. Schwartz, R. L. & Wall, L. in *Programming Perl* (O'Reilly & Associates, Sebastopolo, California 1992).
4. Ousterhout, J. in *Tcl and the Tk Toolkit* (Addison-Wesley, New York, 1994).
5. Peitsch, M. C., Stampf, D. R., Wells, T. N. C. & Sussman, J. L. *TIBS* **20**, 82–84 (1995).
6. Sayle, R. (Glaxo Research & Development, Middlesex UK, 1994).
7. Sussman, J. L. *et al. Science* **253**, 872–879 (1991).

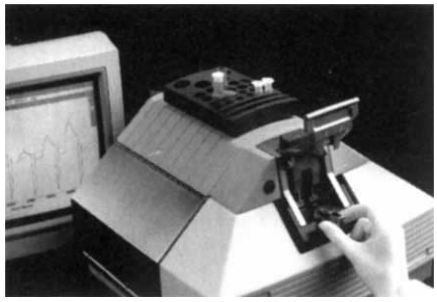
Experimental biology

Featured this week in product review are gel analysis systems for molecular biology research, cultureware for neuronal cells. an optical biosensor system and a molecular weight analyser.

INCSTAR has introduced **rabbit antiserum to calcitonin gene-related peptide** (CGRP), which is useful for immunohistochemical research applications in the human spinal cord as well as that of a number of animal species (*Reader Service No. 101*). Immunohistochemical studies have demonstrated that CGRP is widely distributed in the central and peripheral nervous systems. The antiserum against this neuro-peptide was generated in rabbits to synthetic rat α-CGRP (1–37) and coupled to bovine thyroglobulin with glutaraldehyde. The specificity of the antiserum for immunohistochemistry was examined by soluble pre-adsorption with the peptides in question at the final concentration of 10^{-5} M. The manufacturer states that immuno-labelling was eliminated by pre-adsorption with rat α-CGRP and partially eliminated by pre-adsorption with rat β-CGRP, whereas pre-adsorption with the following peptides resulted in loss of immunostaining: rat amylin, rat adrenomedulin, calcitonin, neurotensin, somatostatin, substance P, leucine enkephalin, methionine enkephaline, VIP, CCK-8, vasopressin and neuropeptide Y.

MaxPlax packaging extract from Epicentre Technologies utilizes a new restriction-free packaging strain to achieve what the manufacturer states are extremely **high packaging effciencies for λDNA** ($1–3 \times 10^9$ PFU/μg DNA) (*Reader Service No. 102*). This packaging is also designed to enhance packaging of λ-DNA bearing the mammalian methylation pattern. The product is supplied as predispensed single-tube reactions (five or ten extracts per kit). MaxPlax is designed to be an easy-to-use reagent with applications in constructing cDNA libraries, genomic cloning of highly modified DNA into λ-phage or cosmid vectors and rescuing λ-shuttle vectors from transgenic animals.

Fisons Applied Sensor Technology states that epitope mapping of proteins can be dramatically simplified and accelerated



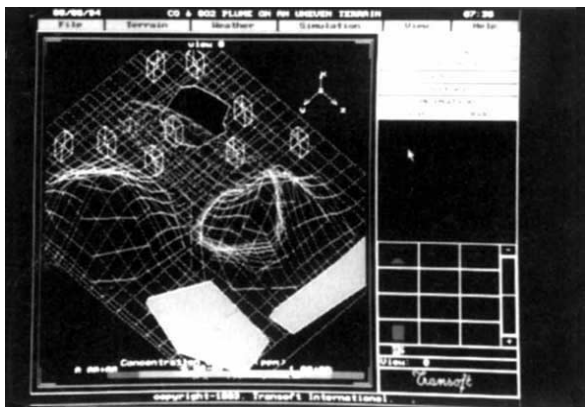
IAsys optical biosensor sensor — Fisons.

using the IAsys **optical biosensor system** (*Reader Service No. 103*). Such mapping, combined with pre-existing information, can provide insights into the antigen's structure and function. In addition, it allows monoclonal antibodies to be categorized by the epitope they recognize. The system studies interactions between biomolecules using a resonant-mirror sensor, integrated into a 200 μl stirred cuvette. This obviates the need to purify and label the monoclonal antibodies, or to carry out time-consuming conventional immunoassays. The antigen is immobilizes on the IAsys biosensor surface and crude, unlabelled, hybridoma supernatants are added sequentially, followed by the regeneration of the antigen.

A new formulation of **[^{35}S]methionine** that can be stored at 4 °C without any affect on performance is now available from Amersham Life Science (*Reader Service No. 104*). Redivue L-[^{35}S]methionine is designed to be used in all experiments where a standard formulation is used to label proteins by either *in vivo* or *in vitro* methods. [^{35}S]methionine normally requires storage at −70 °C, but the new Redivue formulation avoids repeated freeze/thaw cycles and sub-aliquoting. The intense red dye of Redivue products is clearly visible in cell culture media or wheat germ lysate, even at 1 in 50 dilutions, allowing simple determination of which reaction mixes contain the label. Redivue L-[^{35}S]methionine is available at a specific activity of >37 Tbq mmol^{-1} in a radioactive concentration of 370 Mbq ml^{-1}.

Research Biochemicals carries a range of **research-grade dopamine agonists** (*Reader Service No. 105*). New compounds include Quinelorane, a selective dopamine agonist for the D_2-like receptor family with a partial ergoline structure, and Quinpirole, which has a greater affinity for D_2 dopamine receptors and greater *in vivo* potency. A frozen aqueous suspension of membranes prepared from CCL1.3 mouse fibroblasts transfected to express the human D_3 dopamine receptor is also available.

Transoft now offers Fluidyn software for **simulating the spread of various types of pollution** from spills, leaks and explosions into the atmosphere, rivers, coastal regions and inland areas so users can then develop an effective plan of action (*Reader Service No. 106*). The software is designed to represent critical



Fluidyn software from Transoft International.