Genetica **106:** 149–158, 1999.

The protein data bank

Bridging the gap between the sequence and 3D structure world

J.L. Sussman[1,2], E.E. Abola[1], D. Lin[1], J. Jiang[1], N.O. Manning[1] & J. Prilusky[1,3]
[1]*Protein Data Bank, Biology Department, Building 463, Brookhaven National Laboratory, Upton, NY 11973-5000, USA;* [2]*Department of Structural Biology;* [3]*Bioinformatics Unit, Weizmann Institute of Science, Rehovot 76100, Israel*

Key words: X-ray, protein structure, databases, database management systems

Abstract

The protein data bank (PDB), at Brookhaven National Laboratory, is a database containing information on experimentally determined three-dimensional structures of proteins, nucleic acids, and other biological macromolecules, with approximately 9000 entries. The PDB has a 27-year history of service to a global community of researchers, educators, and students in a wide variety of scientific disciplines. Data are easily submitted via PDB's WWW-based tool AutoDep, in either PDB or mmCIF format, and are most conveniently examined via PDB's WWW-based tool 3DB Browser. Collaborative centers have been, and continue to be, established worldwide to assist in data deposition, archiving, and distribution.

Introduction

The protein data bank (PDB), at Brookhaven National Laboratory (BNL), is a database containing experimentally determined, three-dimensional structures of proteins, nucleic acids, and other biological macromolecules [1–3]. The PDB has a 27-year history of service to a global community of researchers, educators, and students in a wide variety of scientific disciplines. The archives contain atomic coordinates, bibliographic citations, primary and secondary structure information, crystallographic structure factors and NMR experimental data, as well as hyperlinks to many other scientific databases. Scientists around the world contribute structures to the PDB and use it on a daily basis. The common interest shared by this community is a need to access information that can relate the biological functions of macromolecules to their three-dimensional structures.

The PDB has introduced substantial enhancements to data deposition and management and user access in the past four years. A PDB Browser was first introduced for a PC as PDB-SHELL [4], then on UNIX systems as the PDB Browser [5, 6], and later via the Internet's World Wide Web (WWW), lets researchers search and retrieve information from the PDB faster and far more flexibly than the older printed indices. The WWW 3DB Browser [3, 7] has been upgraded and enhanced to meet the increasing needs of its user community. In parallel, PDB's new AutoDep facility lets researchers deposit their data quickly and accurately over the WWW directly to the PDB, at either the European Bioinformatics Institute (EBI), or at BNL. Data are then processed by the PDB staff at Brookhaven.

The PDB faces the constant challenge of keeping abreast of the ever-increasing amount of data it must store and provide to an ever-widening and diversified user community, while maintaining the highest standards of data integrity and reliability, and facilitating data retrieval, knowledge exploration, and hypothesis testing. Over the next few years, the PDB will be transformed from a simple data repository as at present into a more powerful, highly sophisticated knowledge-based system for archiving and accessing structural information that combines the advantages of object-oriented and relational database systems. So as not to interrupt current services, these changes have been introduced gradually, insulating users from drastic changes, and thus have provided both a high degree

of compatibility with existing software and a consistent user interface for casual browsers. Collaborative centers have been, and continue to be, established worldwide to assist in data deposition, archiving, and distribution.

Background and significance of the resource

The early years: 1971–1988

The PDB was established in 1971 by Dr Walter Hamilton, at the suggestion of members of the American Crystallographic Association (ACA) and participants at the 1971 Cold Spring Harbor Symposium, e.g., see, D.C. Phillips' remarks of how protein crystallography was 'Coming of Age' [8]. From the beginning, the PDB has operated with the continued support of the crystallographic community. The PDB has always been a truly international effort, initially with affiliated centers at Cambridge, UK; Melbourne, Australia; and Osaka, Japan. (These centers have subsequently been augmented by a number of on-line data providers, 42 at present; see the latest PDB Newsletter for a list). Data acquisition and dissemination, via tape media, was on a global scale from the outset, with a small staff that handled ∼25 structural depositions per year.

Introduction of the current PDB format in 1972 ensured that these data were readily accessible in a convenient and standard form, not only to crystallographers but also to biologists and chemists. This data format has evolved over the last twenty years into the *de facto* standard, serving as both input and output for literally hundreds of computer programs. It has proven to be quite flexible, and recently has been extended for applications that were not imaginable when it was first designed. For example, we have inserted HyperText links into PDB file headers, dynamically linking them to other databases throughout the world, via the World Wide Web (see URL http://www.pdb.bnl.gov/).

The data explosion: 1989–1992

Rapid developments in preparation of crystals of macromolecules and in experimental techniques for structure analysis and refinement have led to a revolution in Structural Biology. These factors have contributed significantly to an enormous increase in the number of laboratories performing structural studies of macromolecules to atomic resolution and the number of such studies per lab. Advances include:

- recombinant DNA techniques that permit almost any protein or nucleic acid to be produced in large amounts;
- rapid DNA (gene) sequencing techniques have made protein sequencing routine;
- better X-ray detectors;
- real-time interactive computer graphics systems, together with more automated methods for structure determination and refinement;
- synchrotron radiation, permit use of tiny crystals, multiple wavelength anomalous dispersion (MAD) phasing, and time-resolved studies via Laue techniques;
- NMR methods permitting structure determination of macromolecules in solution;
- electron microscopy (EM) techniques, for obtaining high-resolution structures.

These dramatic advances produced an abrupt transition from the linear growth of 15–25 new structures deposited per year in the PDB before 1987 to a rapid exponential growth reaching the current rate of about 10 submissions per day (see Figure 1).

In the same period, the proliferation and increasing power of computers, the introduction of relatively inexpensive interactive graphics, and growth of computer networks greatly increased the demand for access to PDB data in many diverse ways. The requirements of molecular biologists, rational drug designers, and others in academia and industry are often fundamentally different from those of crystallographers and computational chemists who had been the major PDB users since the 1970s. This presents a challenge for the PDB and has been addressed in a number of ways, see below.

PDB at present

Contents and access to the PDB archives

The archives contain atomic coordinates, bibliographic citations, primary and secondary structure information, as well as crystallographic structure factors and NMR experimental data. Annotations in the structure entries include amino acid or nucleotide sequences (with notes of any conflicts between the structure in the PDB and sequence databases), source organism from which the biological material was derived, references to papers, secondary structure, complexes with small molecules included

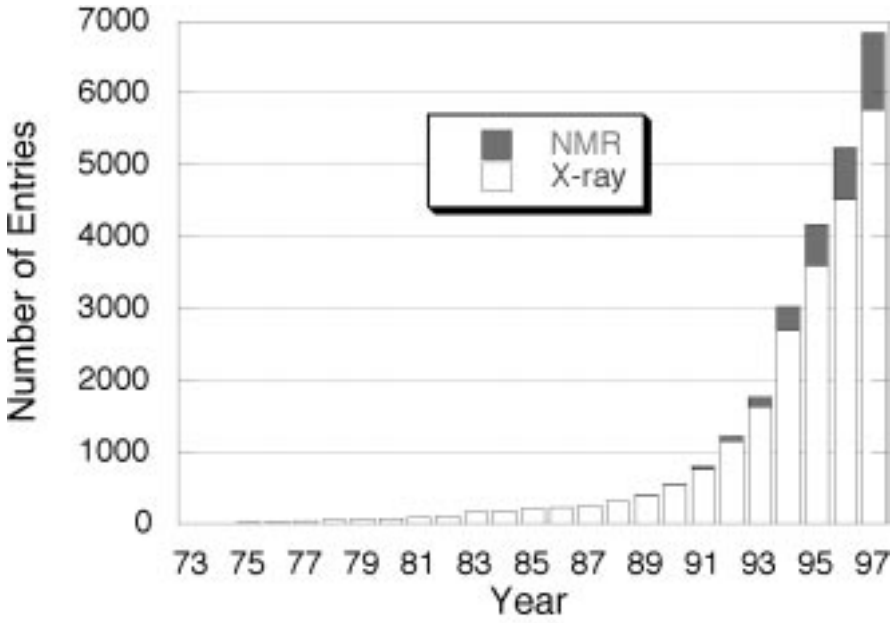

Figure 1. PDB coordinate entries available per year.

within the structure, etc. Third party annotations include images and movies of structures, pointers to other databases which contain information on the structural class or family of the particular structure; pointers to particular specialized databases (maintained by others) such as the Protein Kinase Resource (http://www.sdsc.edu/Kinases/pk_home.html), or the superfamily of alpha/beta hydrolases homologous to cholinesterase (http://meleze.ensam.inra.fr/cholinesterase/), and those that provide additional experimental information such as the BioMagResBank (BMRB) NMR structural database (http://www.bmrb.wisc.edu/) and other solution data, abstracts of articles, etc. Table 1 is a summary of the contents of PDB.

PDB entries are available on CD-ROM, which PC users can search using the PDB-SHELL browser [4] included on the CD-ROM. UNIX users can also search the CD-ROM if they download a copy of the browser software. The entries are also available over the Internet from Brookhaven and 17 mirror sites worldwide, listed in Table 2. They can be searched and retrieved *via* PDB's 3DB Browser [7], which is interfaced through Web browsers such as Netscape, Explorer, etc. Probably the best way to get a feeling for the 3DB Browser is just to try it. A simple example of its use is illustrated in Figure 2 in a search for a structure related to recent papers in *Nature* [9] and *Science* [10].

Table 1. PDB Archives: December 1998

Archive content	Number
Atomic coordinate entries	8856
Structure factor files	2320
NMR restraint files	487
Molecule type	
Proteins, peptides, and viruses	7857
Protein/nucleic acid complexes	363
Nucleic acids	624
Carbohydrates	12
Experimental technique	
Diffraction	7270
NMR	1387
Theoretical modeling	199

The 3DB Browser has a number of features that make it easy to access information found in PDB entries. Users can search according to any combination of fields such as compound name, experiment title, authors (depositors), biological source, journal references, date of deposition, and nature of small molecules (ligands and heterogens) complexed with the structure. Boolean operators allow highly complex search strings. Entries selected can be retrieved automatically, and the molecular structures can be

Table 2. PDB mirror sites: December 1998

Official PDB mirror site	URL
Argentina	
University of San Luis	http://pdb.unsl.edu.ar/
Australia	
Australian National Genomic Information Service, Sydney	http://molmod.angis.org.au/pdb/
The Walter and Eliza Hall Institute of Medical Research, Melbourne	http://pdb.wehi.edu.au/pdb/
Brazil	
ICB-UFMG, Instituto de Ciencias Biologicas, Universidade Federal de Minas Gerais,	http://www.pdb.ufmg.br/
China	
Institute of Physical Chemistry, Peking University, Beijing	http://www.ipc.pku.edu.cn/pdb
France	
Institut de Génétique Humaine, Montpellier	http://pdb.igh.cnrs.fr/
Germany	
GMD, German National Research Center for Information Technology, Sankt Augustin	http://pdb.gmd.de/
India	
Bioinformatics Centre, University of Pune	http://202.41.70.33/
Israel	
Weizmann Institute of Science, Rehovot	http://pdb.weizmann.ac.il/
Japan	
Institute of Protein Research, Osaka University	http://www2.protein.osaka-u.ac.jp/
Poland	
ICM - Interdisciplinary Centre for Modelling, Warsaw University	http://pdb.icm.edu.pl/
Taiwan	
National Tsing Hua University, HsinChu	http://pdb.life.nthu.edu.tw
United Kingdom	
Cambridge Crystallographic Data Centre, Cambridge	ttp://pdb.ccdc.cam.ac.uk/
EMBL Outstation, EBI, Hinxton, UK	http://www2.ebi.ac.uk/pdb
United States	
Bio Molecular Engineering Research Center, Boston University	http://www.pdb.bu.edu/
North Carolina Supercomputing Center, Research Triangle Park, North Carolina	http://pdb.ncsc.org/
University of Georgia, Athens, Georgia	http://pdb.bmb.uga.edu/
PDB at Brookhaven National Laboratory	http://www.pdb.bnl.gov/

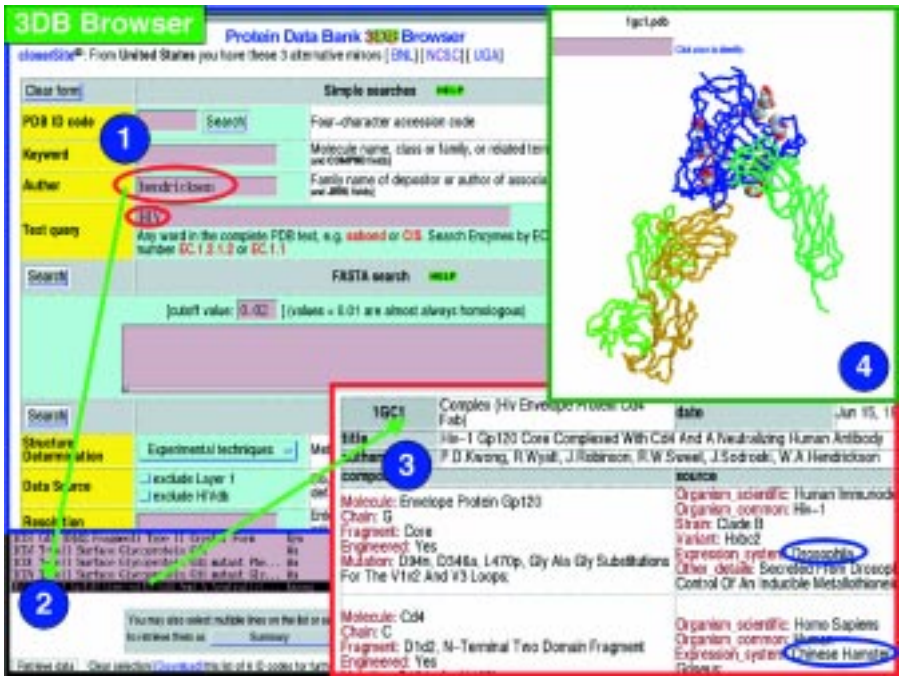

Figure 2. 3DB Browser as a tool to visualize recently published structures. (1) Search for author: Hendrickson; text query: HIV; (2) six hits obtained, 1GC1 highlighted; (3) 3DB Browser Atlas page. Ovals highlight the expression systems used for the different components in the multicomponent system; (4) Structure as visualized with MDL's Chemscape Chime™ plug in.

displayed using the public-domain molecular viewer RasMol [11], Netscape's Chemscape Chime™ plug-in, or a similar viewer. They also include Hyper-Text links to the SwissProt protein sequence database [12], BioMagResBank (BMRB) NMR structural database [13], the Enzyme Commission Database [14], PubMed access to the Medline database, and several other databases (see Table 3 for a list of Linked External Data Sources).

The main source of information for the 3DB Browser is the data from the PDB. This data is highly structured and most of the crystallographers are used to thinking of a piece of data from a PDB entry as belonging to a particular 'record' or 'field'. It makes sense to use these fields to constrain the search. Searching for 'rich' as a keyword has a different meaning than searching for the author Rich.

The simplest operation with the browser is to enter one or more words in the 'Text query' field and press the 'search' button. The browser engine will come back with those entries from the database that contain or are related to the provided words.

The symbol '∗' can be used as a wild card, to denote a sequence of any number (including 0) of arbitrary characters. Just add a star '∗' at the beginning or end of a word (or both) to 'extend' the search. For example, enter *tox* in the keyword field to retrieve those entries containing keywords like neurotoxic and toxin. Wild cards have no meaning in number-only fields, like resolution and date.

The Boolean operator AND is the default for 3DB Browser, and mandatory (you cannot change it) between fields (see Table 4). If you enter 'ATP' in the Associated group field and 'kinase' in the Keyword field, only those entries matching both constraints are returned. Inside a given field, you may apply Boolean logical operators at will to the words you enter. The available Boolean logical operators are AND, OR and NOT. The case is unimportant. The operator AND can be represented by '+' and the operator NOT represented by '−'. For example, 'zinc and (torpedo or snake)' in the Text query field will return those entries that contain either the word torpedo or the word snake, but only if the word zinc is also present. In addition, many specific records can be searched for regular expressions or numerical limits as shown in Table 4 and in the January and April PDB Newsletter (http://www.pdb.bnl.gov/pdb-docs/newsletter.html).

One of the main concerns for us, as database-interface developers, is the 'false negatives', that is,

Table 3. Linked external sources used by the 3DB Browser

Source name	Short description
BioMagResBank	Relational database for sequence-specific protein NMR data
BLOCKS	Database of conserved regions in groups of proteins
CATH	Protein structure classification
Dali/FSSP	Families of structurally similar proteins
EMBL	European molecular biology laboratory
Entrez	NCBI's documentation database
ENZYME	Enzyme nomenclature database
ESTHER	ESTerases and alpha/beta hydrolase enzymes and relatives
GenBank	NIH genetic sequence database
GDB	Genome data base
Kinase	Protein kinase database project
KineMage	Protein science's kinemage server
LPFC	Library of protein family cores
MacroMolecule	EBI's crystal macromolecule files
MMDB	Molecular modelling database
NDB	Nucleic acid database
OLDERADO	Core, domain and representative structure database
PDBOBS	Archive of obsolete PDB entries at SDSC
PDBREPORT	Structure verification reports for X-ray structures
PIR	Protein information resource
PROSITE	Dictionary of protein sites and patterns
ProtMotDB	Protein motions database
SCOP	Structural classification of proteins
SWISS-3DIMAGE	3D images of proteins and other biological macromolecules
SWISS-PROT	Annotated protein sequence database
TREMBL	Translation from EMBL

to not return data after a query, even when the data are available in the database. Frequently this happens because the user was unable to express the query in a way compatible with the search engine, or used words or keywords unknown to the search engine.

3DB Browser deals with this problem by incorporating several automatic and semi-automatic mechanisms to help the user in retrieving the requested data. The request from the user gets filtered and transformed by one or more of the following engines. At the end, the resulting query is the one used for the search (see Table 5).

Inside this section on understanding what the user looks for, we can include the improved search on the CRYST1 record using the short and extended Hermann–Mauguin symbols. You may enter either 'P 1 21 1' or 'P 21' in the Space group field and get the same result.

A search in 3DB Browser brings up a rich Atlas page summarizing additional knowledge related to the entry of interest. The links in this Atlas page carry you to the original sources of information. The number of external sources that 3DB searches and dynamically incorporates into the Atlas pages increases daily (Table 3).

The PDB has several mirror sites across the world. These sites have the same data and facilities as in the central PDB server. They are just closer to you, and, frequently, faster to access on the Internet. To help you know your neighborhood, the 3DB Browser incorporates 'closer-site', an automatic script that detects your location and offers alternative sites that are closer to you (in the network sense).

Internet access to the archives has become the primary mode of retrieving entries from the PDB. However, PDB continues to receive a considerable number of orders for our CD-ROM product. PDB anticipates that this will continue to be true for a variety of reasons. For example, network performance still remains poor in a number of locations, and these disks,

Table 4. Search fields of the 3DB Browser

Search field	PDB entry
Entry ID code	Four-character accession code
Keyword	Molecule name, class or family, or related term (HEADER, TITLE, KEYWDS and COMPND fields)
Author	Family name of depositor or author of associated publication (AUTHOR and JRNL fields)
Text query	Any word in the complete PDB text, excluding most field names
Experiment	Method of structure determination
FASTA search	FASTA search of the sequence
Resolution	A unique value or range of values, in Angstroms (REMARK 2 field)
Space group	Both extended and standard Hermann–Mauguin symbols (CRYST1 field)
Organism	Trivial name, systematic name or expression system (SOURCE field)
Date (lower)	Date entry was deposited or released
Date (upper)	Date entry was deposited or released
Associated group	Prosthetic group, metal ion, ligand or substrate, or its three letter PDB abbreviation (HET and HETNAM fields)
Chain size	A unique value or range of values

Table 5. Search engines used by the 3DB Browser

Engine	Example
American–British synonyms spelling search	'amoeba' and 'ameba' are equivalent; 'protease' is equivalent to 'proteinase' based on a dictionary built from the current PDB data, the spelling engine will produce words that are close to the entered one. As an example, entering 'imune' will offer 'immune' as a valid alternative.
Soundex search	Based on the soundex algorithm that approximates the sound of the word when spoken by an English speaker. Looking for author 'weich' will offer as alternatives: Weiss, Wess, Wyss

released quarterly, provide local access to the contents of the archive. With this software, all files in the PDB are stored locally and changes may be automatically updated on a daily basis by use of mirroring software distributed by the PDB.

Data deposition

Since its inception in 1971, the method followed by the PDB for entering and distributing information has paralleled the review and edit mode used by scientific journals. Currently, the author submits his/her data to the PDB, in mmCIF (http://ndbserver.rutgers.edu/NDB/mmcif/) or PDB format, via PDB's Web-based AutoDep facility (http://www.pdb.bnl.gov:8080) (see Figure 3). AutoDep then calls a suite of validation programs, whose output is returned via the WWW to the depositor within minutes of sending the data to the PDB. This has made it possible for authors to request that their data be 'released on publication' and has reduced the number of authors requesting that their data be held to less than 22% compared to over 75% just a year ago [15].

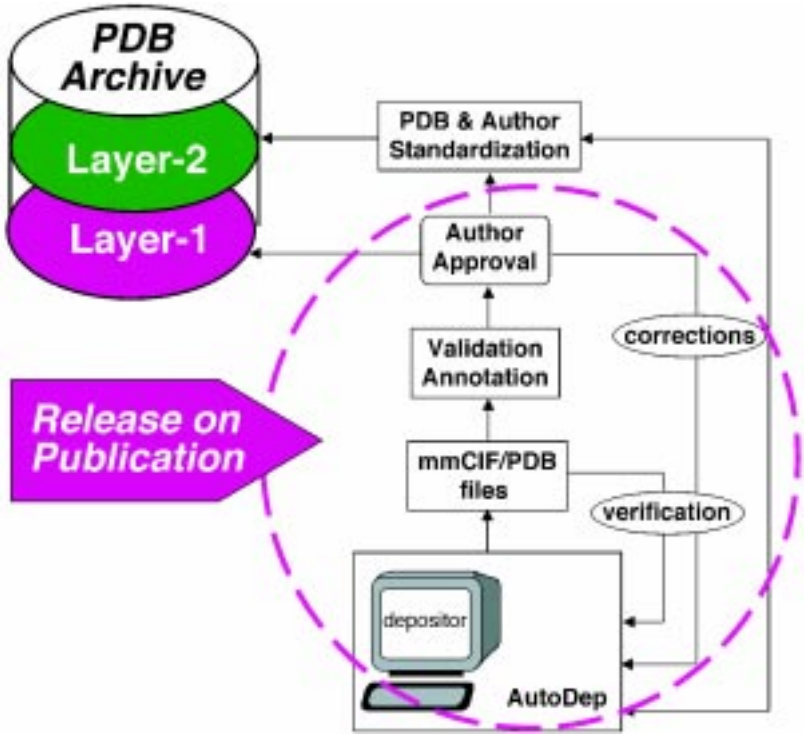

Figure 3. PDB WWW-based submission via AutoDep facilitates releasing the entries via a layered approach and making it possible to automatically release entries on publication, as indicated in the portion of the figure enclosed in a dashed circle.

Table 6. PDB's data validation

Source name	Short description
Stereochemistry	Bond distances & angles, Ramachandran plot (dihedral angles), planarity of groups, chirality
Bonded/non-bonded interactions	Crystal packing, unspecified inter- and intraresidue links
Crystallographic information	Matthews' coefficient, Z-value, cell transformation matrices
Noncrystallographic transformation	Validity of noncrystallographic symmetry
Primary sequence data	Discrepancies with sequence databases
Secondary structure	Generated automatically or visually checked
Heterogen groups	Identification, geometry and nomenclature
Miscellaneous checks	Solvent molecules outside the hydration sphere, syntax checks, internal data consistency checks

Based on these checks, authors may decide to give permission to release the entry immediately; to release it after up to a maximum one year hold; or go back and reexamine the structure in light of the output diagnostics before completing the submission procedure. The PDB ID code is issued only after the author gives release approval. The submitted data must include all mandatory information as described in the October 1997 PDB Newsletter (http://www.pdb.bnl.gov/pdb-docs/newsletter.html) and in the 'List of Items Mandatory for a Complete PDB Submission' (http://www.-pdb.bnl.gov/pdb-docs/mandatory_items.html). The data

Table 7. Key WWW sites related to 3D structure of biological macromolecules Browser's linked external data sources

Description	URL
PDB Homepage	http://www.pdb.bnl.gov/
3DB Browser	http://www.pdb.bnl.gov/pdb-bin/pdbmain
SwissProt database	http://www.expasy.ch/sprot/sprot-top.html
Entrez system	http://www3.ncbi.nlm.nih.gov/Entrez/
PubMed	http://www.ncbi.nlm.nih.gov/PubMed/
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/
DALI	http://croma.ebi.ac.uk/dali/
Nucleic acid database	http://ndbserver.rutgers.edu/
Pedro's BioMolecular research tools	http://www.public.iastate.edu/~pedro/research_tools. html
BioMagResBank	http://www.bmrb.wisc.edu/
Biological macromolecule crystallization DB	http://h178133.carb.nist.gov:4400/bmcd/bmcd.html
Archive of obsolete PDB entries	http://pdbobs.sdsc.edu/PDBobs.cgi

must also pass certain validation criteria as described in the January 1998 PDB Newsletter, and in the document 'Validation for Layered Release' (http://www.pdb.bnl.gov/pdb-docs/validation.html). Entries passing the validation criteria are released clearly identified as 'LAYER-1'. An associated file containing output diagnostics is also released.

Following this information, PDB staff processes the entry. The entry and the output of the validation suite are evaluated by a PDB scientific staff member, who completes the annotations and returns the entry to the author for comment and approval. Table 6 summarizes checks included in our current data validation suite. Corrections from the author are incorporated into the entry, which is reanalyzed and validated before being archived and released. Most of this work covers issues not now fully delegated to automatic software. The resulting entry, after author approval, replaces the LAYER-1 entry in the archive. We strongly believe that such thorough checking and annotation is essential for ensuring the long-term value of the data.

Examples of impact of the PDB

There are numerous examples in molecular biology, medicine and drug discovery where the PDB is playing an increasingly important role.

Possibly the best examples of the use of structural information used to help in the design of new drugs to combat disease is in the area of HIV infection. At present there are eight HIV proteins whose 3D structure have been determined, one of which is illustrated in Figure 2. These have aided in the design of several drugs that have as their targets one of these proteins.

Key sites related to PDB are given in Table 7.

Funding

The PDB is supported by a combination of Federal Government Agency funds and user fees. Support is provided by the U.S. National Science Foundation, the U.S. Public Health Service, National Institutes of Health, National Center for Research Resources, National Institutes of General Medical Sciences, National Library of Medicine, the U.S. Department of Energy and user fees.

References

1. Abola E.E., Bernstein F.C., Bryant S.H., Koetzle T.F. and Weng J.: Protein Data Bank. In: (Allen F.H., Bergerhoff G. and Sievers R. (eds), *Crystallographic Databases – Information Content, Software Systems, Scientific Applications.* Data Commission of the International Union of Crystallography, Bonn, 1987, pp. 107–132.

2. Abola E.E., Sussman J.L., Prilusky J. and Manning N.O.: Protein data bank archives of three-dimensional macromolecular structures, *Meth. Enzym.* **277** (1997): 556–571.
3. Sussman J.L., Lin D., Jiang J., Manning N.O., Prilusky J., Ritter O. and Abola E.E.: Protein data bank (PDB): Database of 3D structural information of biological macromolecules, *Acta Cryst.* **D54** (1998): 1078–1084.
4. Abola E.E.: PDB-SHELL. PDB Available anonymous FTP (1994): ftp.pdb.bnl.gov.
5. Peitsch M.C., Stampf D.R., Wells T.N.C. and Sussman J.L.: The Swiss 3D-image collection and Brookhaven protein data bank browser on the World-Wide Web, *TIBS* **20** (1995): 82–84.
6. Stampf D.R., Felder C.E. and Sussman J.L.: PDBBrowse – a graphics interface to the Brookhaven protein data bank. *Nature* **374** (1995): 572–574.
7. Sussman J.L.: Bridging the Gap. *Nature Struct. Biol.* **4** (1997): 517.
8. Phillips D.C.: Protein Crystallography 1971: Coming of Age. In: *Cold Spring Harbor Symposium of Quantitative Biology*, (eds) Cold Spring Harbor Laboratory, Cold Spring Harbor, (1971), pp. 589–592.
9. Kwong P.D., Wyatt R., Robinson J., Sweet R.W., Sodroski J. and Hendrickson W.A.: Structure of an HIV gp120 Envelope Glycoprotein in Complex with the CD4 Receptor and a Neutralizing Human Antibody, *Nature* **393** (1998): 648–659.
10. Rizzuto C.D., Wyatt R., Hernandez-Ramos N., Sun Y., Kwong P.D., Hendrickson W.A. and Sodroski J.: A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science* **280** (1998): 1949–1953.
11. Sayle R.A. and Milner-White E.J.: RASMOL: Biomolecular graphics for all, *TIBS* **20** (1995): 374–376.
12. Bairoch A. and Boeckmann B.: The SWISS-PROT Protein sequence data bank: current status, *Nucl. Acids Res.* **22** (1994): 3578–3580.
13. Seavey B.R., Farr E.A., Westler W.M. and Markley J.L.: A relational database for sequence-specific protein NMR data, *J. Biomol. NMR* **1** (1991): 217–236.
14. Bairoch A.: The ENZYME Data Bank, *Nucl. Acids Res.* **22** (1994): 3626–3627.
15. Sussman J.L.: Protein data bank deposits, *Science* **282** (1998): 1991.