

Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules

JOEL L. SUSSMAN,^{a,b*} DAWEI LIN,^a JIANSHEG JIANG,^a NANCY O. MANNING,^a JAIME PRILUSKY,^c OTTO RITTER^{a,d} AND ENRIQUE E. ABOLA^a

^aBiology Department, Building 463, Brookhaven National Laboratory, Upton, NY 11973-5000, USA, ^bDepartment of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel, ^cBioinformatics Unit, Weizmann Institute of Science, Rehovot 76100, Israel, and ^dDepartment of Molecular Biophysics, German Cancer Research Center, 69120 Heidelberg, Germany. E-mail: jls@bnl.gov

(Received 24 April 1998; accepted 9 July 1998)

Abstract

The Protein Data Bank (PDB) at Brookhaven National Laboratory, is a database containing experimentally determined three-dimensional structures of proteins, nucleic acids and other biological macromolecules, with approximately 8000 entries. Data are easily submitted via PDB's WWW-based tool *AutoDep*, in either mmCIF or PDB format, and are most conveniently examined via PDB's WWW-based tool *3DB Browser*.

1. 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 (Abola *et al.*, 1987, 1997; Bernstein *et al.*, 1977). The PDB has a 26-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, citations, primary and secondary structure information, crystallographic structure 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. The PDB browser, first introduced on PC and UNIX systems and later via the World Wide Web (WWW), allows researchers to search and retrieve information from the PDB faster and far more flexibly than the older printed indices. The *3DB Browser* (Sussman, 1997) has been upgraded and enhanced to meet the increasing needs of its user community. In parallel, PDB's new *AutoDep* facility allows researchers to 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.

2. Background and significance of the resource

2.1. 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* (Phillips, 1971). 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 complete 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 20 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 recently inserted HyperText links into PDB file headers, dynamically linking them to other databases throughout the world, via the WWW (see URL <http://www.pdb.bnl.gov/>).

2.2. The data explosion: 1989–1992

Rapid developments in the 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 laboratory. Advances include: (1) recombinant DNA techniques that permit almost any protein or nucleic acid to be produced in large amounts; (2) rapid protein and DNA (gene) sequencing techniques that have made protein sequencing routine; (3) better X-ray detectors; (4) real-time interactive computer graphics systems, together with more automated methods for structure determination and refinement; (5) synchrotron radiation, allowing the use of extremely tiny crystals, multiple-wavelength anomalous dispersion (MAD) phasing, and time-resolved studies *via* Laue techniques; (6) NMR methods permitting structure determination of macromolecules in solution; and (7) electron microscopy (EM) techniques, for obtaining high-resolution structures.

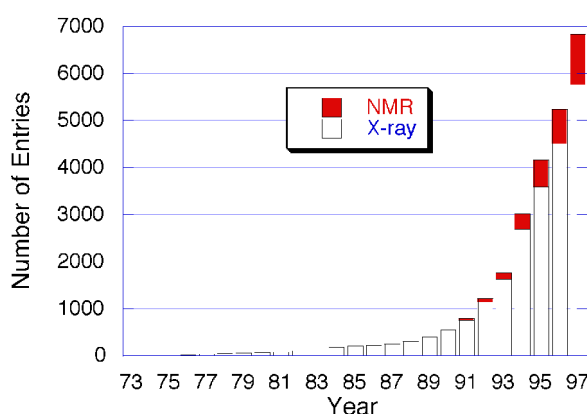


Fig. 1. PDB coordinate entries available per year

Table 1. *PDB archive contents as of June 1998*

Released atomic coordinate entries	7864
Structure-factor files	1972
NMR restraint files	429
Molecule type	
Proteins, peptides and viruses	6917
Protein/nucleic acid complexes	318
Nucleic acids	556
Carbohydrates	12
Others	1
Experimental technique	
Diffraction	6437
NMR	1240
Theoretical modeling	187

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 50 submissions per week (see Fig. 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.

3. PDB at present

3.1. 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 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) esther (<http://www.ensam.inra.fr/cholinesterase/>) or Archive of Obsolete PDB Entries databases (<http://pdboobs.sdsc.edu/PDBObs.cgi>) and those that provide additional experimental information such as the BioMagResBank (BMRB) NMR structural

Table 2. *PDB mirror sites June 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/
Israel	
Weizmann Institute of Science, Rehovot	http://pdb.weizmann.ac.il/
Poland	
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	http://pdb.ccdc.cam.ac.uk/
EMBL Outstation, EBI, Hinxton, UK	http://www2.ebi.ac.uk/pdb
United States	
North Carolina Supercomputing Center, Research Triangle Park, North Carolina	http://pdb.ncsc.org/
University of Georgia, Athens, Georgia, USA	http://pdb.bmb.uga.edu/
PDB at Brookhaven National Laboratory	http://www.pdb.bnl.gov/

database (<http://www.bmrb.wisc.edu/>) and other solution data, abstracts of articles, *etc.*

Table 1 is a summary of the contents of PDB. Present plans are to keep abreast of the deposition rate within a timeline of three months or less from receipt of an entry to final archiving. This includes the time spent in careful checking by the PDB professional staff as well as a period for the depositor to double check the processed entry.

PDB entries are available on CD-ROM, which PC users can search using the *PDB-SHELL* browser. UNIX users can also search the CD if they download a copy of the browser software. The entries are also available over the Internet from Brookhaven and 14 mirror sites worldwide, listed in Table 2. They can be searched and retrieved *via* the Internet browser (Peitsch *et al.*, 1995; Stampf *et al.*, 1995) and now the *3DB Browser* (Sussman, 1997), that is interfaced through WWW browsers such as Netscape, Explorer *etc.*, as illustrated in Fig. 2. All these search methods provide direct access to the molecular viewing program *RasMol* (Sayle & Milner-White, 1995).

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 such fields as compound name, experiment title, authors (depositors), biological source, journal references, date of deposition, and nature of small molecules (heterogens) complexed with the structure. Boolean operators

allow highly complex search strings. Entries selected can be retrieved automatically, and the molecular structures can be displayed using the public domain molecular viewer *RasMol* (Sayle & Milner-White, 1995), Netscape's Chemscape Chime plug-in, or a similar viewer. They also include HyperText links to the SwissProt protein sequences database (Bairoch & Boeckmann, 1994) BioMagResBank (BMRB) NMR structural database (Seavey *et al.*, 1991), the Enzyme Commission Database (Bairoch, 1994), PubMed access to the Medline database, and several other databases (see Table 3 for a list of linked external data sources). 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, 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.

3.2. 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 WWW-based *AutoDep* facility (<http://www.pdb.bnl.gov:8080>) (Fig. 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.

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](http://www.pdb.bnl.gov/pdb-docs/mandatory_items.html)). The data 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, PDB staff process the entry as was performed previously. The entry and the output of the validation suite are then evaluated by a PDB scientific staff member, who completes the annotations and returns the entry to the author for comment and approval. Table 4 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

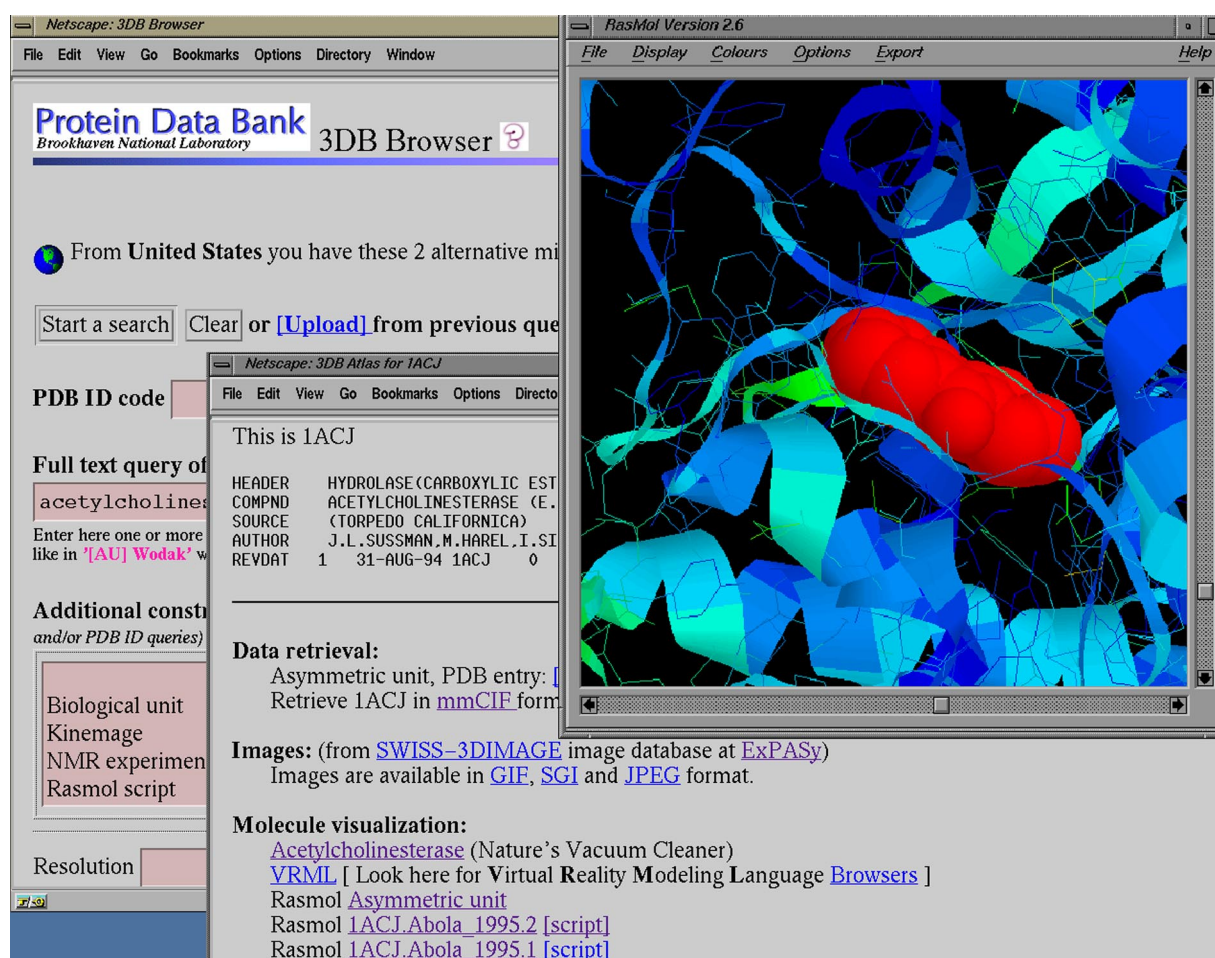


Fig. 2. The WWW 3DB Browser in action (Peitsch *et al.*, 1995; Stampf *et al.*, 1995; Sussman, 1997). On the left is the browse screen, with windows to enter search strings. In the upper right, the selected entries, *i.e.*, acetylcholinesterase (1ACJ) (Harel *et al.*, 1993) displayed as a ribbon diagram with *RasMol* (Sayle & Milner-White, 1995). On the lower right, the text of the 1ACJ entry is shown with the blue text indicating a HyperLink to other databases, including the SwissProt protein sequence database (Bairoch & Boeckmann, 1994)

Table 3. *3DB Browser's linked external data sources*

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 sequence database
Entrez	NCBI's documentation database
ENZYME	Enzyme nomenclature database
ESTHER	ESTerases and alpha/beta hydrolase enzymes and relatives database
GenBank	NIH genetic sequence database
GDB	Genome database
Kinase	Protein kinase database project
KineMage	Protein Science's Kinemage server
LPFC	Library of protein family cores
MacroMolecule	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
PubMed	Medline bibliographic database
SCOP	Structural Classification of Proteins
Swiss 3D-Image	Three-dimensional images of proteins and other biological macromolecules
SwissProt	Annotated protein sequence database
TREMBL	Translation from EMBL sequence database

approval, will be equivalent to the traditional PDB entry and will be designated LAYER-2. We strongly believe

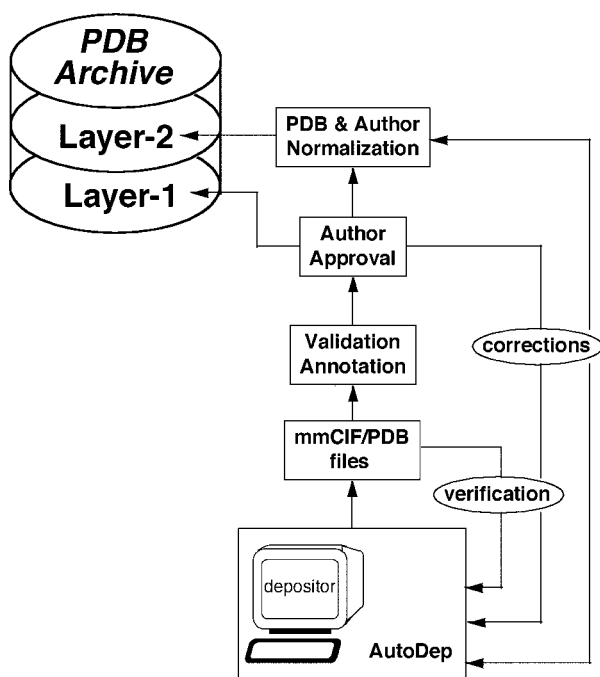


Fig. 3. PDB WWW-based submission, via *AutoDep*, releasing the entries via a layered approach.

that such thorough checking and annotation is essential for ensuring the long-term value of the data.

Originally data flow was a manual system, designed for a staff of one to two scientists, and a deposition rate of about 25–50 entries per year. One person processed an entry from submission through its release. By the late 1980s, when the first steps at automation were being introduced, running the validation programs took about 4 h per entry. Today, the same step, which is highly automated and includes a vastly improved set of validation programs, takes about 15 min. Graphical viewing of data, a useful and powerful annotating and checking tool, has been available to processors since 1992.

Ideally, PDB would like the entire deposition process to be automatic. However, certain kinds of problems continue to require manual intervention and processing. The most troublesome areas remain those involving handling of heterogens (small molecules complexed with the structure), resolving crystal packing issues, representing molecules with non-crystallographic symmetry, and resolving conflicts between the submitted amino-acid sequence and that found in the sequence databases. Publications and other references are sometimes consulted to verify factual information such as crystal data, biological details, reference information, *etc.* Processing programs, although much improved over those used in 1991, still allow errors to pass undetected through the system, requiring a visual check of all entries. We are striving to expand the *AutoDep* suite of

Table 4. *PDB's data validation checks*

Class	What is checked
Stereochemistry	Bond distances & angles, Ramachandran plot (dihedral angles), planarity of groups, chirality
Bonded/non-bonded interactions	Crystal packing, unspecified inter- and intrasidue 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
Heterogen groups	Identification, geometry and nomenclature
Miscellaneous checks	Solvent molecules outside the hydration sphere, syntax checks, internal data consistency checks

deposition and validation programs to accommodate the somewhat conflicting desires of both depositors and users, while ensuring that the archives maintain the highest standard of accuracy. This includes acquiring software from collaborators to address deficiencies that both we and our users have identified.

3.3. Funding

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

of Medicine, the US Department of Energy and user fees.

4. 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 already seven HIV proteins whose three-dimensional structures have been determined, see Fig. 4. These have aided in the design of several drugs that have as their targets one of these proteins.

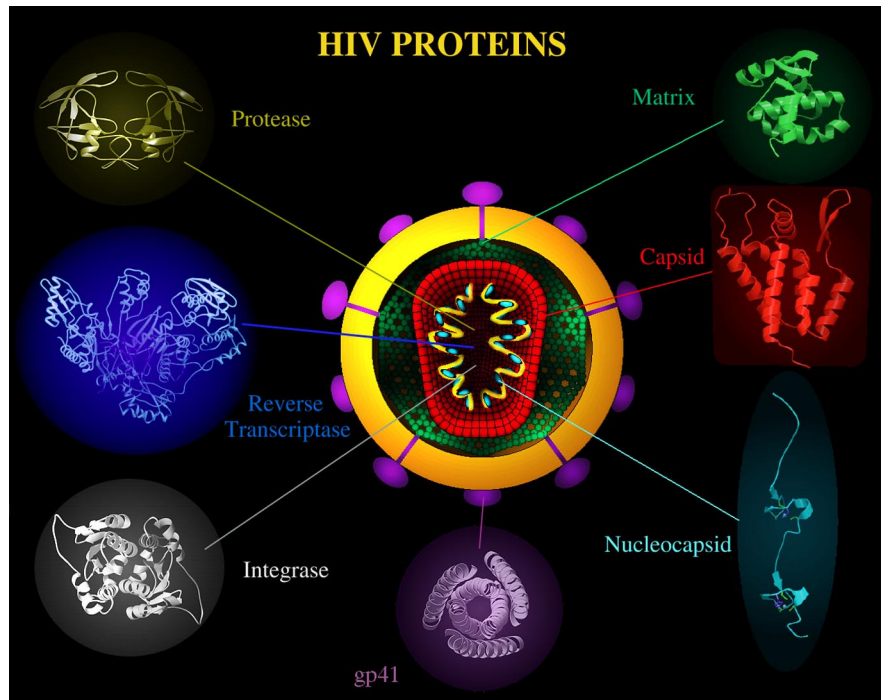


Fig. 4. Schematic representation of the human immunodeficiency virus, surrounded by the structures of individual proteins solved using either crystallography or NMR. Figure prepared by Dr Jacek Lubkowski, National Cancer Institute, Frederick Cancer Research and Development Center, as a modification of a figure made by Michael Summers, University of Maryland Baltimore County.

Table 5. Key WWW sites related to three-dimensional structures of biological macromolecules

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 BioMolec Research Tools	http://www.public.iastate.edu/~pedro/research_tools.html
BioMagResBank	http://www.bmrb.wisc.edu/
Biological Macromolecule Crystallization Database and the NASA Archive for Protein Crystal Growth Data	http://ibm4.carb.nist.gov:4400/
Archive of Obsolete PDB Entries	http://pdobobs.sdsc.edu/PDOBobs.cgi

5. Future plans: PDB to 3DB

A new database, 3DB-Base, is being developed at the PDB. Collaborative international centers are also being established to assist in data deposition, archiving, and distribution, including the European Bioinformatics Institute (EBI), Osaka University, Weizmann Institute of Science and the BioMagResBank (BMRB) at the University of Wisconsin.

Converting the PDB to 3DB involves changes in every aspect of current operations. The new system relies on a relational database system for data management and archiving using the Object-Protocol Model (OPM) tools (<http://gizmo.lbl.gov/opm.html>) (Chen & Markowitz, 1995). This development effort attempts to address the needs of the diverse user community served by the PDB. The system is being designed with the expectation that it will be federated with other biological databases. Our hope is that this system will allow complex queries to be submitted to the 3DB, parts of which may need to be sent automatically to other databases for processing, and return a composite answer. In addition to providing users with a powerful environment for complex *ad-hoc* queries, 3DB-Base will also facilitate management of the growing archive, which is expected to contain over 30 000 structural reports by the year 2000. It will fully support the new IUCr archival format mmCIF for deposition and queries. This work is being performed as a collaboration among the following groups: The Protein Data Bank, Brookhaven National Laboratory; European Bioinformatics Institute (EBI); Cambridge Crystallographic Data Centre (CCDC); Bioinformatics Unit, Weizmann Institute of Science; BioMagResBank, University of Wisconsin (BMRB); OPM Data Management Tools Project, Lawrence Berkeley National Laboratory; and Gene Logic Inc., Berkeley, CA.

6. Related databases

See Table 5 for key WWW sites related to three-dimensional structures of biological macromolecules.

References

- Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *Crystallographic Databases – Information Content, Software Systems, Scientific Applications*, edited by F. H. Allen, G. Bergerhoff & R. Sievers, pp. 107–132. Bonn: IUCr.
- Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). *Methods Enzymol.* **277**, 556–571.
- Bairoch, A. (1994). *Nucleic Acids Res.* **22**, 3626–3627.
- Bairoch, A. & Boeckmann, B. (1994). *Nucleic Acids Res.* **22**, 3578–3580.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
- Chen, I. A. & Markowitz, V. M. (1995). *Information Sys.* **20**, 393–418.
- Harel, M., Schalk, I., Ehret-Sabatier, L., Bouet, F., Goeldner, M., Hirth, C., Axelsen, P., Silman, I. & Sussman, J. L. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 9031–9035.
- Peitsch, M. C., Stampf, D. R., Wells, T. N. C. & Sussman, J. L. (1995). *Trends Biol. Sci.* **20**, 82–84.
- Phillips, D. C. (1971). *Cold Spring Harbor Symp. Quant. Biol.* pp. 589–592.
- Sayle, R. A. & Milner-White, E. J. (1995). *Trends Biol. Sci.* **20**, 374–376.
- Seavey, B. R., Farr, E. A., Westler, W. M. & Markley, J. L. (1991). *J. Biomol. NMR*, **1**, 217–236.
- Stampf, D. R., Felder, C. E. & Sussman, J. L. (1995). *Nature (London)*, **374**, 572–574.
- Sussman, J. L. (1997). *Nature Struct. Biol.* **4**, 517.