International Tables for Crystallography (2006). Vol. F, Section 24.1.2, pp. 649–650.

24. CRYSTALLOGRAPHIC DATABASES

24.1. The Protein Data Bank at Brookhaven

BY J. L. SUSSMAN, D. LIN, J. JIANG, N. O. MANNING, J. PRILUSKY AND E. E. ABOLA

24.1.1. 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; Sussman *et al.*, 1998). 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, ligand 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 the 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 over the past five years. A PDB browser was first introduced for a PC as *PDB-SHELL* (Abola, 1994), then on UNIX systems as *PDB Browser* (Peitsch *et al.*, 1995; Stampf *et al.*, 1995), and later *via* the World Wide Web (WWW). It permits researchers to search and retrieve information from the PDB faster and far more flexibly than from the older printed indices. The WWW 3*DB Browser* (Sussman, 1997; Sussman *et al.*, 1998) has been upgraded and enhanced to meet the increasing needs of its user community. In parallel, the PDB's *AutoDep* facility [see *Protein Data Bank Quarterly Newsletter* (1998), **85**, p. 3, *Release of AutoDep 2.1* at http://www.rcsb.org/pdb/newsletter.html] lets researchers deposit their data quickly and accurately over the WWW directly to the PDB, either at 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 past few years, the PDB has been transformed from a simple data repository into a powerful, highly sophisticated knowledge-based system for archiving and accessing structural information. 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 centres have been, and continue to be, established worldwide to assist in data deposition, archiving and distribution.

As of 1 July 1999, the operation of the PDB in the United States is being transferred from BNL to the Research Collaboratory for Structural Bioinformatics (RCSB). The RCSB (http://www.rcsb.org/), a consortium composed of Rutgers, the State University of New Jersey; the University of California at San Diego; and the National Institute of Standards and Technology (NIST), has received a five-year award from the National Science Foundation (NSF), the Department of Energy (DOE) and two units of the National Institutes of Health: the National Institute of General Medical Sciences (NIGMS) and the National Library of Medicine (NLM).

24.1.2. Background and significance of the resource

24.1.2.1. *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 centres at Cambridge, England, Melbourne, Australia, and Osaka, Japan. These centres have subsequently been augmented by a number of online data providers, 41 at present [see *Protein Data Bank Quarterly Newsletter* (1999), **87**, p. 12, *Affiliated centers and mirror sites* at http://www.rcsb.org/pdb/newsletter.html). Data acquisition and dissemination, *via* tape media, were on a global scale from the outset, with a small staff handling about 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 has recently been extended for applications unimaginable 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 http://www.rcsb.org).

24.1.2.2. *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 laboratory. Advances include:

(1) recombinant DNA techniques that permit almost any protein or nucleic acid to be produced in large amounts;

(2) rapid 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, permitting use of 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.

These dramatic advances produced an abrupt transition from the linear growth of 15–25 new structures deposited per year in the

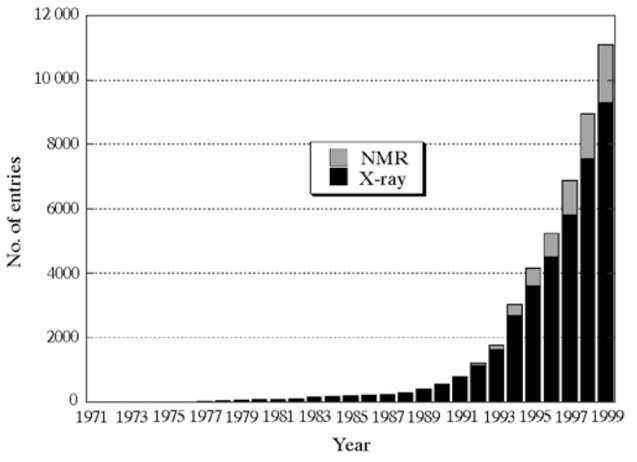

Fig. 24.1.2.1. PDB coordinate entries available per year.

Table 24.1.3.1. *PDB archive contents as of May 1999*

9862	Atomic coordinate entries
2768	Structure-factor files
560	NMR restraint files
Molecule type:	
8754	Proteins, peptides and viruses
415	Protein/nucleic acid complexes
681	Nucleic acids
12	Carbohydrates
Experimental technique:	
8103	Diffraction
1544	NMR
215	Theoretical modelling

PDB before 1987 to a rapid exponential growth reaching the current rate of about ten submissions per day (see Fig. 24.1.2.1).

In the same period, the proliferation and increasing power of computers, the introduction of relatively inexpensive interactive graphics, and the 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 the crystallographers and computational chemists who have 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).

24.1.3. The PDB in 1999

24.1.3.1. *Contents and access to the PDB archives*

The archives contain atomic coordinates, bibliographic citations, primary- and secondary-structure information, 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 organisms from which the biological materials were derived, descriptions of the experiments, secondary structures, complexes with small molecules included within the structures, references to papers *etc*. Third-party annotations include images and movies of structures; pointers to specialized databases (maintained by others), such as the Protein Kinase Resource (http://www.sdsc.edu/Kinases/pk_home.html) and ESTHER (ESTerases and α/β Hydrolase Enzymes and Relatives) (http://www.ensam.inra.fr/cholinesterase/), and pointers to databases that provide additional experimental information, such as the BioMagResBank (BMRB) NMR structural database (http://www.bmrb.wisc.edu/). Table 24.1.3.1 gives a summary of the contents of the PDB archives.

PDB entries are available on CD-ROM, which PC users can search using the *PDB-SHELL* browser included (Abola, 1994). UNIX users can also search the CD-ROM if they download a copy of the browser software. The entries are also available over the WWW from Brookhaven and 17 mirror sites worldwide (Table 24.1.3.2). They can be searched and retrieved *via* the PDB's 3*DB Browser* (Sussman, 1997), which is interfaced through web browsers such as Netscape Communicator and Internet Explorer. Probably the best way to get a feeling for 3*DB Browser* is just to try it. A simple example of its use is illustrated in Fig. 24.1.3.1 in a search for a structure related to recent papers in *Nature* (Kwong *et al.*, 1998) and *Science* (Rizzuto *et al.*, 1998).

3*DB 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 displayed using the public-domain molecular viewer *RasMol* (Sayle & Milner-White, 1995), MDL's *Chemscape Chime* plug-in, or a similar viewer. The entries also include HyperText links to the SwissProt protein-sequence database (Bairoch & Boeckmann, 1994), the BioMagResBank (BMRB) NMR structural database (Seavey *et al.*, 1991), the Enzyme Commission Database (Bairoch, 1994), PubMed access to the Medline database, and several other

Table 24.1.3.2. *PDB mirror sites as of May 1999*

Official PDB mirror sites
Argentina: University of San Luis
Australia: Australian National Genomic Information Service, Sydney; The Walter and Eliza Hall Institute of Medical Research, Melbourne
Brazil: ICB-UFMG, Inst. de Ciencias Biologicas, Univ. Federal de Minas Gerais
China: Institute of Physical Chemistry, Peking University, Beijing
France: Institut de Génétique Humaine, Montpellier
Germany: GMD, German National Research Center for Information Technology, Sankt Augustin
India: Bioinformatics Centre, University of Pune
Israel: Weizmann Institute of Science, Rehovot
Japan: Institute of Protein Research, Osaka University
Poland: ICM - Interdisciplinary Centre for Modelling, Warsaw University
Taiwan: National Tsing Hua University, HsinChu
United Kingdom: Cambridge Crystallographic Data Centre, Cambridge; EMBL Outstation, EBI, Hinxton
United States: Bio Molecular Engineering Research Center, Boston University; North Carolina Supercomputing Center, Research Triangle Park; University of Georgia, Athens, Georgia; PDB at Brookhaven National Laboratory

references