International Tables for Crystallography (2006). Vol. F, Chapter 24.1, pp. 649–656.

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

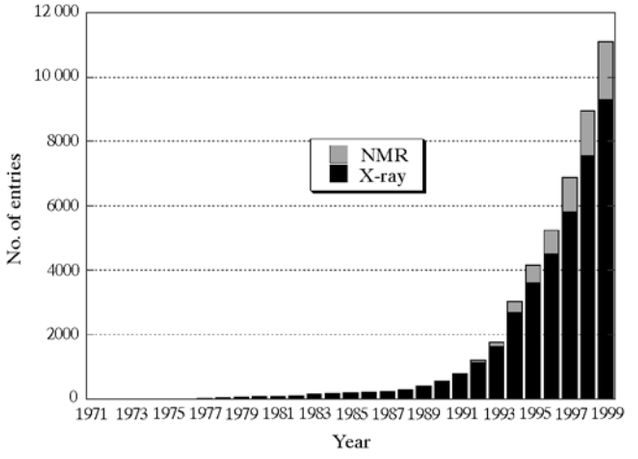

Fig. 24.1.2.1. PDB coordinate entries available per year.

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

databases (see Table 24.1.3.3 for a list of linked external data sources).

The main source of information for the 3*DB Browser* is the data from the PDB. These data are highly structured, and most crystallographers usually consider a datum 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 from 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 words provided.

The symbol '*' can be used as a wild card to denote a sequence of any number (including 0) of arbitrary characters. Just add an asterisk, '*', 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 3*DB Browser* and is mandatory (it cannot be changed) between fields (see Table 24.1.3.4). If 'ATP' is entered in the Associated group field and 'kinase' in the Keyword field, only those entries matching both constraints are returned. Inside a given field, Boolean logical operators may be applied at will to the words entered. 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 can be 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 24.1.3.4 [see *Protein Data Bank Quarterly Newsletter* (1998), **83**, pp. 3–5, *The 'Intelligent' Search Engine Behind the 3DB Browser*^{TM}, and *Protein Data Bank Quarterly Newsletter* (1998), **84**, pp. 3–4, *3DB Browser*^{TM}: *Tips, Questions and Answers* at http://www.rcsb.org/pdb/newsletter.html).

One of the main concerns for us, as database-interface developers, is the 'false negative', that is, the failure to 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.

3*DB Browser* deals with this problem by incorporating several automatic and semi-automatic mechanisms to help the user retrieve the requested data. The request from the user gets filtered and transformed by one or more engines. At the end, the resulting query is the one used for the search (see Table 24.1.3.5).

A search in 3*DB Browser* brings up a rich Atlas page, summarizing additional information about the entry of interest. The links in this Atlas page carry one to the original sources of information. The number of external sources that 3*DB Browser* searches and dynamically incorporates into the Atlas pages increases daily (Table 24.1.3.3).

The PDB's WWW server is the major tool used to access the three-dimensional macromolecular structural information archived at the PDB. Thousands of times a day, scientists, students and other users around the world visit the PDB to browse through and access these data. In order to meet the need for rapid access worldwide, a global network of 17 official mirror sites has been established. To help orient the user, 3*DB Browser* incorporates *CloserSite* (see http://pdb.weizmann.ac.il/pdb-docs/closerSite.html), an automatic script that detects one's location and offers closer alternative sites (in the network sense).

The information on the PDB's web server changes frequently. New information is generated on a daily basis. Synchronizing the PDB and its mirror sites to provide exactly the same services while requiring minimum human involvement is a necessary but nontrivial task.

A protocol for the automatic mirroring of the web sites was developed at BNL based on ftp mirroring technology. This protocol has been used successfully by PDB and its mirror sites for approximately two years.

Fig. 24.1.3.2 outlines the web mirroring protocol, which consists of the following five steps.

(1) Develop and test HTML pages and common-gateway-interface (CGI) codes on the development server in a special source-code control area.

(2) Copy the working code and HTML pages to a read-only area.

(3) Mirror the updated information onto an internal test server that uses its own directory tree, distinct from that used for development. This internal server simulates the production environment under controlled conditions. For example, we verify that updated files are mirrored properly and that relative HTML links work.

(4) Copy the files outside the firewall to an account accessible only to the mirror sites.

(5) Activate the mirror software to transfer the updated files to the PDB web server. Official mirror-site servers are updated automatically by their own mirroring procedures.

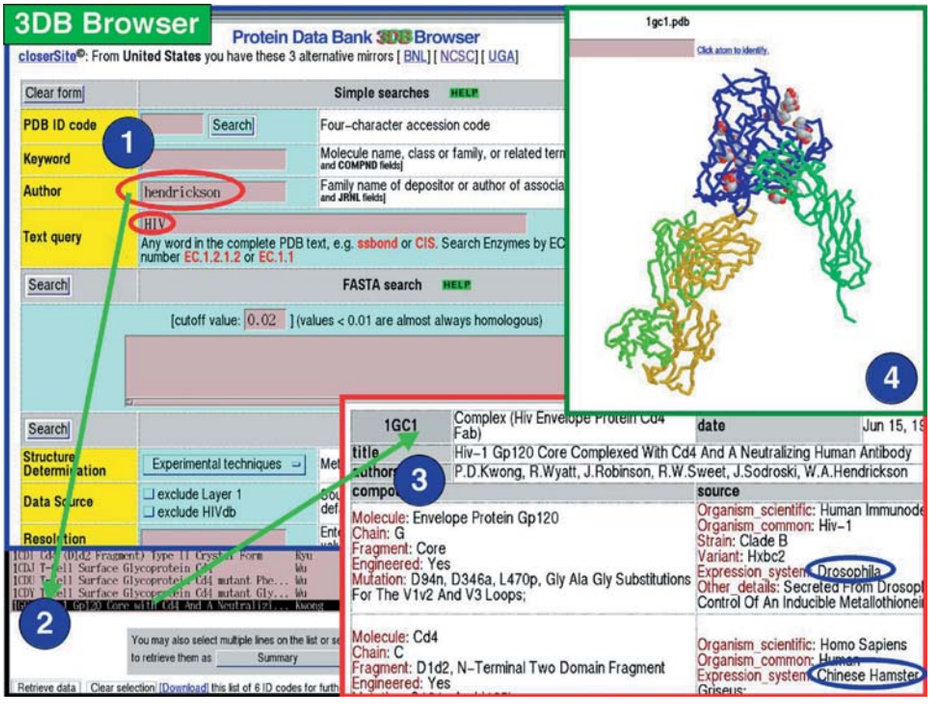

Fig. 24.1.3.1. 3*DB Browser* as a tool to visualize recently published structures. (1) Search for author: Hendrickson; text query: HIV. (2) Six hits obtained, PDB ID Code 1GC1 highlighted. (3) 3*DB Browser* Atlas page. Ovals highlight the expression systems used for the different components in the multicomponent system. (4) Structure as visualized with MDL's *Chemscape Chime* plug-in.

Table 24.1.3.3. *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 Data Base
Kinase	Protein Kinase Database Project
KineMage	Protein Science's *Kinemage* server
LPFC	Library of Protein Family Cores
MacroMolecule	Crystal MacroMolecule files at the EBI
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	3D images of proteins and other biological macromolecules
SwissProt	Annotated protein sequence database
TREMBL	Translation from EMBL sequence database

Table 24.1.3.4. *Search fields of 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 Å (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, substrate, or its three-letter PDB abbreviation (HET and HETNAM fields)
Chain size	A unique value or range of values

Table 24.1.3.5. *Search engines used by 3DB Browser*

Engine	Example
American–British	'Amoeba' and 'ameba' are equivalent
Synonyms	'Protease' is equivalent to 'proteinase'
Spelling search	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 and Wyss

Special steps are taken to isolate files, thus obviating problems associated with the existence of files and directories not related to PDB web activities. HTML documents are stored under the directory /pdb-docs/, and executables are stored under the directory /pdb-bin/. In addition, index files and local configuration files are stored in the directory /PDB-support/.

Specific areas on the http server are dedicated to PDB web activities. All the HTML pages and CGI scripts are in the /pdb-docs/ and /pdb-bin/ directories, respectively. There are also index files and local configuration files in /PDB-support/. This avoids confusing PDB applications with other applications on the same server, which would complicate the mirror procedure.

Relative links are used in all the HTML pages and the HTML pages generated by the scripts. For example, to create a hyperlink to *3DB Browser* in the file named index.html, 3DB Browser is used instead of <a href="http://

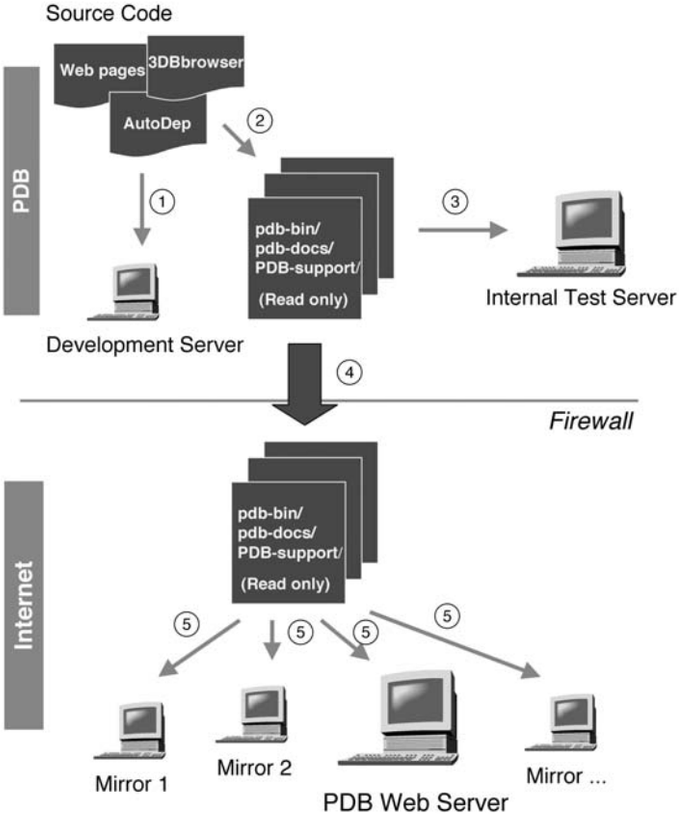

Fig. 24.1.3.2. Schematic diagram of the PDB WWW mirror system.

www.pdb.bnl.gov/pdb-bin/pdbmain”>3DB Browser. The advantage of relative links is that pages copied to the mirror sites' machines will point to local resources without having to be edited locally. This is one of the key points in automating the web mirror procedure. To make relative links work properly, the mirror sites maintain a local configuration file. The configuration file reflects the local directory tree and available resources. The PDB provides a generic template, and mirror sites modify it according to their setup. This configuration file is excluded from the automatic mirroring procedure to avoid being overwritten by the original template file. Changes to the configuration files are sent to mirrors by e-mail one week in advance, to be included manually.

To avoid duplication and allow easy maintenance of the resources, PDB's web and ftp servers share some files. All mirror sites support both web and ftp servers. When a hyperlink points to a file on the ftp server, a server side include (SSI) script is used to access the local ftp server of each mirror site. Its function is to use configuration variables to generate a path to the local file dynamically.

HTML pages and CGI scripts are put into a read-only account available to official mirror sites. Mirror sites use the ftp mirror tool mirror.pl (ftp://sunsite.org.uk/packages/mirror/) to mirror the updated information from this account. For security reasons, this account is not an anonymous ftp account, but requires a password for access. In addition, this account can only be accessed by ftp. This process can be made as a cron job to automate the update procedures fully. Although the procedure is automatic, an e-mail message is sent to mirror sites for update verification. For details on the PDB mirror system, see *Protein Data Bank Quarterly Newsletter* (1999), **87**, pp. 3–5, *PDB World Wide Web Mirroring System* at http://www.rcsb.org/pdb/newsletter.html).

Web access to the archives has become the primary mode of retrieving entries from the PDB. However, the PDB continues to receive a considerable number of orders for our CD-ROM product. The PDB anticipates that this will continue to be so 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. PDB files may first be copied from the CD-ROM to a local disk, and then incremental updates can easily be made using mirroring software.

24.1.3.2. *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 their data to the PDB, in mmCIF (http://ndbserver.rutgers.edu/NDB/mmcif/) or PDB format, *via* the PDB's web-based *AutoDep* facility (Lin *et al.*, 2000; http://autodep.ebi.ac.uk) (see Fig. 24.1.3.3). *AutoDep* then calls a suite of validation programs, whose output is returned *via* the web 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 (Sussman, 1998).

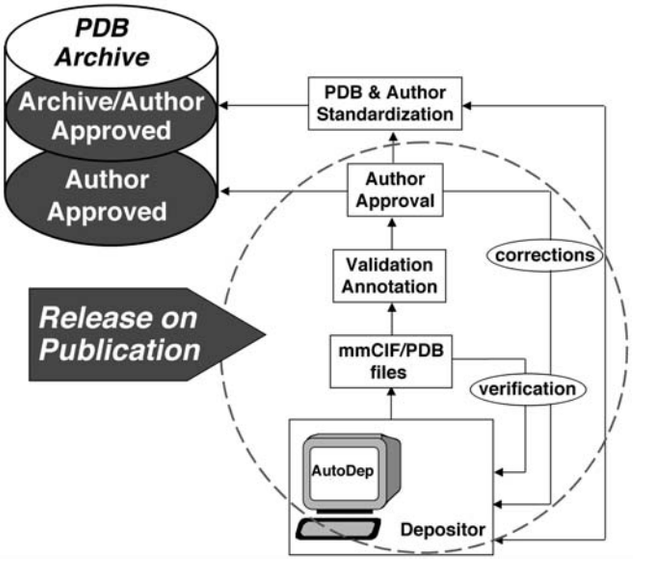

Fig. 24.1.3.3. PDB WWW-based submission *via AutoDep* facilitates releasing the entries *via* a layered approach, making it possible to release entries automatically on publication, as indicated in the portion of the figure enclosed in a dashed circle.

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 to go back and re-examine 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 [see *Protein Data Bank Quarterly Newsletter* (1987), **82**, pp. 2–3, *Proposed Mandatory Items* at http://www.rcsb.org/pdb/newsletter.html and in the *List of Items Mandatory for a Complete PDB Submission* at http://pdb.rutgers.edu/~adbnl/pdb-docs/mandatory_items.html). The data must also pass certain validation criteria (see *Validation for Layered Release* at http://pdb.rutgers.edu/~adbnl/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. The entry and the output of the validation suite are evaluated by a PDB scientific staff

Table 24.1.3.6. *PDB data-validation checks*

Class	What is checked
Stereochemistry	Bond distances and 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

Table 24.1.3.7. *PDB structure-factor submissions, as of November 1998*

Year	No. of X-ray structure submissions	No. of structure-factor submissions (%)
1994	804	205 (25.0)
1995	963	343 (36.0)
1996	1124	546 (49.0)
1997	1484	932 (62.8)
1998	1616	868 (53.7)
Total	5991	2894 (48.3)

member, who completes the annotations and returns the entry to the author for comment and approval. Table 24.1.3.6 summarizes the checks included in our current data-validation suite. Corrections from the author are incorporated into the entry, which is reanalysed 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.

The PDB has long made available the experimental data that were used to determine the three-dimensional structures in the database. In recent years, more and more depositors and users of the PDB have come to appreciate the importance of reliable access to such fundamental data. The deposition of the experimental data, along with the coordinates, is essential for the following reasons.

(1) Rigorous validation of the structure-determination results can only be carried out using both atomic parameters and experimental structure-factor amplitudes.

(2) Archiving of these data will ensure their preservation and continued accessibility.

Whether or not to require that the experimental data be deposited concomitantly with the structure data has recently been hotly discussed in the scientific press (Baker *et al.*, 1996) and on the internet (*EBI/MSD Draft Consultative Document for Deposition of Structure Factors*, http://msd.ebi.ac.uk/sf/sf.html).

At present, more than 50% of the X-ray diffraction submissions are being deposited with their associated structure factors (see Table 24.1.3.7), compared with 25% four years ago. This increase is probably partly due to the ease of uploading the files *via* our web-based submission tool, *AutoDep*, which is available at the EBI (http://autodep.ebi.ac.uk). The PDB strongly encourages all researchers to deposit their structure factors at the time of coordinate submission. Furthermore, we actively encourage journals to require their submission as a prerequisite for publication [see *Protein Data Bank Quarterly Newsletter* (1996), **75**, p. 1, *What's New at the PDB* at http://www.rcsb.org/pdb/newsletter.html).

In order to facilitate the use of deposited structure factors, we at the PDB, together with a number of macromolecular crystallographers and the IUCr Working Group on Macromolecular CIF, developed a standard interchange format for structure factors [*PDB Structure Factor mmCIF* at http://ndb-mirror-2.rutgers.edu/NDB/ftp/PDB/structure_factors/cifSF_dictionary; *Protein Data Bank Quarterly Newsletter* (1995), **74**, p. 1, *What's New at the PDB* at http://www.rcsb.org/pdb/newsletter.html]. This standard is the mmCIF format, *i.e.*, the IUCr-developed macromolecular Crystallographic Information File. It was chosen for its simplicity of design and for being clearly self-defining. The format is also easy to expand as new crystallographic experimental methods or concepts are developed, by simply adding additional tokens. The entire mmCIF crystallographic dictionary (http://ndb.rutgers.edu/NDB/mmcif) has recently been ratified by the IUCr's Committee for the Maintenance of the CIF Standard (COMCIFS).

The PDB has written a program to quickly and easily convert structure factors, as output by the most frequently used crystallographic programs, into mmCIF format. This tool, which also converts binary CCP4 MTZ files, will be accessible through the *AutoDep* program following final testing. MTZ files, which are useful in individual laboratories, are not appropriate for archival purposes. This is because particular groups arbitrarily attach different labels to the MTZ columns.

During the past year, the PDB has converted virtually all the old structure-factor files to this standard format and is keeping up-to-date on all new submissions. As of November 1998, there are about 2000 structure-factor files released in structure-factor mmCIF format (Jiang *et al.*, 1999; PDB mmCIF structure-factor files can be found at ftp://ftp.rcsb.org/pub/data/structures/divided/structure_factors/), with about an additional 1300 'on hold'. The current IUCr policy states that 'The IUCr also urges crystallographers to use their influence to ensure that all journals that publish articles on macromolecular three-dimensional structure require the deposition of both atomic parameters and structure-factor amplitudes.' and 'Authors are urged to release the atomic parameters and structure-factor amplitudes immediately after the publication date. This should be the normal practice. They can, however, request a delay of up to six months in the release of the atomic parameter data and the structure-factor amplitudes.' (Commission on Biological Macromolecules, 2000). The structure factors are also available *via* 3*DB Browser* (http://pdb-browsers.ebi.ac.uk/pdb-bin/pdbmain or http://bioinfo.weizmann.ac.il:8500/oca-bin/ocamain). This can be seen on the browser's Atlas page for each structure.

The ready availability of structure-factor files in a standard format has made it possible for any scientist to validate a structure in the PDB *versus* its experimentally observed data. There are now some excellent tools available for this, such as the Uppsala Electron Density Server (http://alpha2.bmc.uu.se/valid/density/form1.html) and the program *SFCHECK* (http://www.iucr.org/iucr-top/comm/ccom/School96/pdf/sw.pdf). The PDB has also observed that one of the most popular uses for these stored structure factors is for the crystallographer who did the experiment to be able to retrieve their own misplaced data.

24.1.4. Examples of the impact of the PDB

There are numerous examples in molecular biology, medicine and drug discovery where the PDB is playing an increasingly important role, as can be seen in the many sites related to the PDB (see Table 24.1.4.1).

One key example is the impact that structural information is having on the design of new drugs to combat diseases such as AIDS. At present, the three-dimensional structures of eight HIV proteins have been determined, one of which is illustrated in Fig. 24.1.3.1. These three-dimensional structures have aided researchers in the design of several drugs that have one of these proteins as their targets. Other examples can be seen in our basic understanding of the immune system (Madden *et al.*, 1993), Fig. 24.1.4.1, and the interaction between proteins and DNA (Schultz *et al.*, 1991), Fig. 24.1.4.2.

The PDB is a major international resource used by scientists, educators and students throughout the world. During the past few years, we at the PDB, in collaboration with many others, have greatly enhanced this resource into a powerful user-friendly tool for bridging the gap between the three-dimensional structure and the genome worlds (Sussman, 1997). Some examples follow.

(1) The PDB's *AutoDep* procedure (Lin *et al.*, 2000) has made

Table 24.1.4.1. *Key web sites related to three-dimensional structures of biological macromolecules*

Description	URL
PDB home page	http://www.rcsb.org
3*DB Browser*	http://pdb-browsers.ebi.ac.uk/pdb-bin/pdbmain or http://bioinfo.weizmann.ac.il:8500/oca-bin/ocamain
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://www2.ebi.ac.uk/dali/
Nucleic Acid Database	http://ndbserver.rutgers.edu/
BioMagResBank	http://www.bmrb.wisc.edu/
Biological Macromolecule Crystallization Database and the NASA Archive for Protein Crystal Growth Data	http://wwwbmcd.nist.gov:8080/bmcd/bmcd.html
Archive of obsolete PDB entries	http://pdbobs.sdsc.edu/PDBobs.cgi
EBI PDB submission site	http://autodep.ebi.ac.uk

deposition of structural data much easier. More importantly, the data are much richer in information content and more accurately checked before release. *AutoDep* has also made uploading coordinates, structure factors and NMR restraint files very simple for the depositors.

(2) Results of the layered-release protocol have exceeded our best expectations, with the number of new entries being requested to be 'on hold' now down to only about 20% (and still decreasing), compared with well over 75% just a year ago (Sussman, 1998).

(3) The PDB is now receiving structure factors for a very high percentage of the structures determined by X-ray crystallography (Jiang *et al.*, 1999).

(4) There is now a close interaction between the PDB and most journals relevant to structural studies to ensure coordinate deposition in the PDB (and release) as a prerequisite for acceptance of manuscripts, as seen in editorials in several prominent scientific journals (Bloom, 1998; Cambell, 1998; Editorial Board, 1998).

Numerous close interactions and/or collaborations with scientists from around the world have yielded beneficial results for the entire community. This has resulted in the PDB becoming a truly international endeavour. Some examples follow.

(1) The first remote PDB deposition site has been established in Europe at the EBI (http://autodep.ebi.ac.uk).

(2) Improvement in handling of ligands and het groups for both deposition and retrieval of information has been achieved using programs developed by M. Hendlich (University of Marburg, Germany) and the CCDC (Cambridge, England).

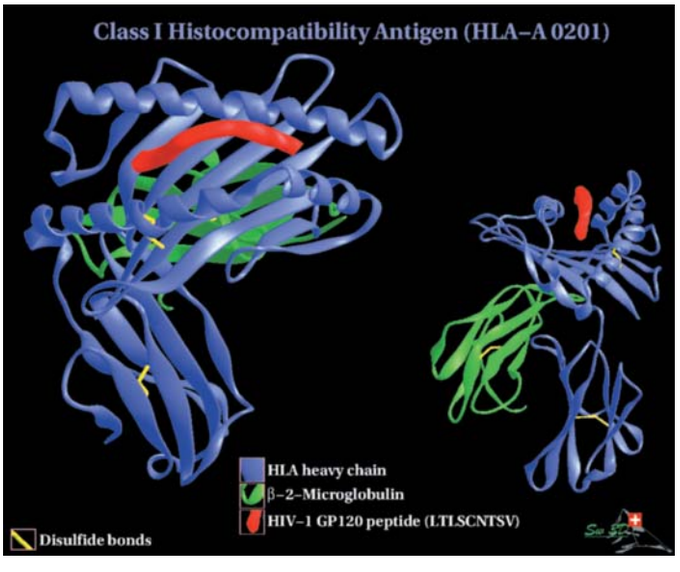

Fig. 24.1.4.1. Crystal structure of a complex of a peptide from an HIV-1 protein bound to the human class I MHC molecule HLA-A2 (Madden *et al.*, 1993), PDB ID code 1HHG, as illustrated in the SwissProt images available on the WWW (Peitsch *et al.*, 1995).

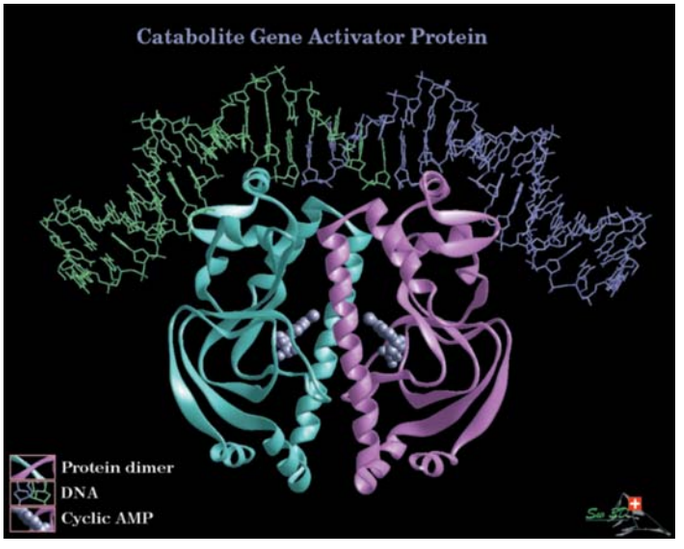

Fig. 24.1.4.2. Crystal structure, at 3 Å resolution, of the *E. coli* catabolite gene activator protein (CAP) complexed with a 30-base-pair DNA sequence. It shows that the DNA is bent by 90°. This bend results almost entirely from two 40° kinks that occur between TG/CA base pairs at positions 5 and 6 on each side of the dyad axis of the complex (Schultz *et al.*, 1991), PDB ID code 1CGP, as illustrated in the SwissProt images available on the WWW (Peitsch *et al.*, 1995).

(3) Tools to improve access and examination of three-dimensional structural information, such as *PDB Lite* and *Noncovalent Bond Finder* (E. Martz, University of Massachusetts, USA) have been developed.

(4) The user-friendly way of accessing the PDB *via* 3*DB Browser* (developed in close collaboration with Dr Jaime Prilusky, Bioinformatics Unit, Weizmann Institute of Science, Israel) has already become the standard for several online journals pointing to the PDB Atlas pages of structures.

(5) There is close interaction with the BioMagResBank (BMRB, University of Wisconsin) for the handling of NMR structural data.

(6) The fact that industrially determined three-dimensional structures are now being deposited with the PDB, even without publication, has been made possible *via* the close collaboration between the PDB and the HIV Protease Database (developed by Alexander Wlodawer at NCI, Frederick, MD, USA, and Jiri Vondrasek at IOCB, Prague, Czech Republic: see http://www.ncifcrf.gov/CRYS/HIVdb).

(7) The 17 official mirror sites in 13 countries around the world now provide easy and fast local access to the PDB web pages and database archives.

Acknowledgements

This work has been carried out by a most dedicated and talented staff at the PDB, including Frances Bernstein, Betty Deroski, Arthur Forman, Sabrina Hargrove, Mariya Kobiashvili, Pat Langdon, Michael Libeson, John McCarthy, Christine Metz, Otto Ritter, Regina Shea, Janet Sikora, Lu Sun, Subramanyam Swaminathan and Dejun Xue. In addition, John Rose (University of Georgia), Mia Raves (Utrecht University), Simone Botti, Meir Edelman, Clifford Felder, Kurt Giles, Harry Greenblatt, Gitay Kryger, Michal Harel, Marilyn Safran, Israel Silman, Vladimir Sobolev (Weizmann Institute of Science), Kim Henrick (EBI), Gert Vriend (EMBL-Heidelberg), Barry Honig (Columbia University) and Axel Brünger (Yale University) have provided invaluable support throughout the years. The PDB Advisory Board and the BNL administration together with the BNL Chemistry and Biology Departments have been an invaluable resource over the years. We wish to express our great appreciation and respect for the members of this team, who have constantly shown enormous initiative and professional capability in all their endeavours.

according to major consensus. Table 24.5.9.1 indicates how to access these resources.

24.5.10. Conclusion

These are exciting and challenging times to be responsible for the collection, curation and distribution of macromolecular structure data. Since the RCSB assumed responsibility for data deposition in February 1999, the number of depositions has averaged approximately 50 a week. However, with the advent of a number of structure genomics initiatives worldwide, this number is likely to increase. We estimate that the PDB, which at writing contains approximately 10 500 structures, could triple or quadruple in size over the next five years. This presents a challenge of timely distribution while maintaining high quality. The PDB's approach of using modern data-management practices should permit us to accommodate a large data influx.

The maintenance and further development of the PDB are community efforts. The willingness of others to share ideas, software and data provides a depth to the resource not obtainable otherwise. Some of these efforts are acknowledged below. New input is constantly being sought and the PDB invites comments at any time by e-mail to info@rcsb.org.

Acknowledgements

The continuing support of Ken Breslauer (Rutgers), John Rumble (NIST) and Sid Karin (SDSC) is gratefully acknowledged. Current collaborators contributing to the future development of the PDB are the BioMagResBank, the Cambridge Crystallographic Data Centre, the HIV Protease Database Group, The Institute for Protein Research, Osaka University, The National Center for Biotechnology Information, the ReLiBase developers, the Swiss Institute for Bioinformatics/Glaxo and the European Bioinformatics Institute.

The cooperation of the BNL PDB staff is also gratefully acknowledged.

Parts of this chapter have appeared in *Nucleic Acids Research* (Berman *et al.*, 2000) and are reproduced here with permission of Oxford University Press.

This work is supported by grants from the National Science Foundation, the Office of Biology and Environmental Research at the Department of Energy, and two units of the National Institutes of Health: the National Institute of General Medical Sciences and the National Library of Medicine.

References

24.1

Abola, E. E. (1994). *PDB-SHELL*. Available at ftp://pdb.bmc.uu.se/pub/databases/pdb/pdb_software/pdbshell/.

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *Protein Data Bank*. In *Crystallographic databases – information content, software systems, scientific applications*, edited by F. H. Allen, G. Bergerhoff & R. Sievers, pp. 107–132. Bonn: International Union of Crystallography.

Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). *Protein Data Bank archives of three-dimensional macromolecular structures*. *Methods Enzymol*. **277**, 556–571.

Bairoch, A. (1994). *The ENZYME data bank*. *Nucleic Acids Res*. **22**, 3626–3627.

Bairoch, A. & Boeckmann, B. (1994). *The SWISS-PROT protein sequence data bank: current status*. *Nucleic Acids Res.* **22**, 3578–3580.

Baker, E. N., Blundell, T. L., Vijayan, M., Dodson, E., Dodson, G., Gilliland, G. L. & Sussman, J. L. (1996). *Crystallographic data deposition*. *Nature (London)*, **379**, 202.

Bloom, F. E. (1998). *Policy change*. *Science*, **281**, 175.

Cambell, P. (1998). *New policy for structure data*. *Nature (London)*, **394**, 105.

Commission on Biological Macromolecules (2000). *Guidelines for the deposition and release of macromolecular coordinate and experimental data*. *Acta Cryst*. D**56**, 2.

Editorial Board (1998). *New policy on release of structural coordinates*. *Proc. Natl Acad. Sci. USA*, **95**, iii.

Jiang, J., Abola, E. & Sussman, J. L. (1999). *Deposition of structure factors at the Protein Data Bank*. *Acta Cryst.* D**55**, 4.

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J. & Hendrickson, W. A. (1998). *Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody*. *Nature (London)*, **393**, 648–659.

Lin, D., Manning, N. O., Jiang, J., Abola, E. E., Stampf, D., Prilusky, J. & Sussman, J. L. (2000). *AutoDep: a web-based system for deposition and validation of macromolecular structural information*. *Acta Cryst.* D**56**, 828–841.

Madden, D. R., Garboczi, D. N. & Wiley, D. C. (1993). *The antigenic identity of peptide–MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2*. *Cell*, **75**, 693–708.

Peitsch, M. C., Stampf, D. R., Wells, T. N. C. & Sussman, J. L. (1995). *The Swiss 3D-image collection and Brookhaven Protein Data Bank browser on the World-Wide Web*. *Trends Biochem. Sci.* **20**, 82–84.

Phillips, D. C. (1971). *Protein crystallography 1971: coming of age*. *Cold Spring Harbor Symp. Quant. Biol.* pp. 589–592.

Rizzuto, C. D., Wyatt, R., Hernandez-Ramos, N., Sun, Y., Kwong, P. D., Hendrickson, W. A. & Sodroski, J. (1998). *A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding*. *Science*, **280**, 1949–1953.

Sayle, R. A. & Milner-White, E. J. (1995). *RASMOL: biomolecular graphics for all*. *Trends Biochem. Sci.* **20**, 374–376.

Schultz, S. C., Shields, G. C. & Steitz, T. A. (1991). *Crystal structure of a CAP–DNA complex: the DNA is bent by 90 degrees*. *Science*, **253**, 1001–1007.

Seavey, B. R., Farr, E. A., Westler, W. M. & Markley, J. L. (1991). *A relational database for sequence-specific protein NMR data*. *J. Biomol. Nucl. Magn. Reson.* **1**, 217–236.

Stampf, D. R., Felder, C. E. & Sussman, J. L. (1995). *PDBBrowser – a graphics interface to the Brookhaven Protein Data Bank*. *Nature (London)*, **374**, 572–574.

Sussman, J. L. (1997). *Bridging the gap*. *Nature Struct. Biol.* **4**, 517.

Sussman, J. L. (1998). *Protein Data Bank deposits*. *Science*, **282**, 1991.

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O. & Abola, E. E. (1998). *Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules*. *Acta Cryst.* D**54**, 1078–1084.

24.2

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information*. *Acta Cryst.* B**35**, 2331–2339.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. & Schneider, B. (1992). *The Nucleic Acid Database – a comprehensive relational database of three-dimensional structures of nucleic acids*. *Biophys. J.* **63**, 751–759.