International Tables for Crystallography (2006). Vol. F, Section 24.5.2, pp. 675–677.

24.5. The Protein Data Bank, 1999–

BY H. M. BERMAN, J. WESTBROOK, Z. FENG, G. GILLILAND, T. N. BHAT, H. WEISSIG, I. N. SHINDYALOV AND P. E. BOURNE

24.5.1. Introduction

The Protein Data Bank (PDB) was established at Brookhaven National Laboratory (BNL) (Bernstein *et al.*, 1977) in 1971 as an archive for biological macromolecular crystal structures. In the beginning there were seven structures, and each year a handful more were deposited. In the 1980s the number of deposited structures began to increase dramatically. This was due to the improved technology for all aspects of the crystallographic process, the addition of structures determined by nuclear magnetic resonance (NMR) methods and changes in the community views about data sharing. By the early 1990s the majority of journals required a PDB accession code and at least one funding agency (National Institute of General Medical Sciences) adopted the guidelines published by the IUCr requiring data deposition for all structures.

The mode of access to PDB data has changed over the years as a result of improved technology, notably the availability of the World Wide Web (WWW) replacing distribution solely *via* magnetic media. Further, the need to analyse diverse data sets required the development of modern data-management systems.

Initial use of the PDB had been limited to a small group of experts involved in structural research. Today depositors to the PDB have varying expertise in the techniques of X-ray crystal-structure determination, NMR, cryoelectron microscopy and theoretical modelling. Users are a very diverse group of researchers in biology and chemistry, community scientists, educators and students at all levels. The tremendous influx of data soon to be fuelled by the structural genomics initiative, and the increased recognition of the value of the data toward understanding biological function, demand new ways to collect, organize and distribute the data.

The vision of the Research Collaboratory for Structural Bioinformatics (RCSB)* is to create a resource based on the most modern technology that would facilitate the use and analysis of structural data and thus create an enabling resource for biological research. In October 1998, the management of the PDB became the responsibility of the RCSB.† In this chapter, we describe the current procedures for deposition, processing and distribution of PDB data by the RCSB. We conclude with some current developments of the PDB.

24.5.2. Data acquisition and processing

A key component of creating the public archive of information is the efficient capture and curation of the data – data processing. Data processing consists of data deposition, annotation and validation. These steps are part of the fully documented and integrated data-processing system shown in Fig. 24.5.2.1.

In the present system (Fig. 24.5.2.2), data (atomic coordinates, structure factors and NMR restraints) may be submitted *via* e-mail or *via* the *AutoDep Input Tool* [*ADIT*: http://pdb.rutgers.edu/adit (Westbrook *et al.*, 1998)] developed by the RCSB. *ADIT*, which is also used to process the entries, is built on top of the mmCIF dictionary, which is an ontology of 1700 terms that define the macromolecular structure and the crystallographic experiment (Bourne *et al.*, 1997), and a data-processing program called *MAXIT* (Macromolecular Exchange and Input Tool; Feng, Hsieh *et al.*, 1998). This integrated system helps to ensure that the data that are deposited for an entry are consistent and error-free after annotation.

After a structure has been deposited using *ADIT*, a PDB identifier is sent to the author automatically and immediately (Fig. 24.5.2.1, step 1). This is the first stage in which information about the structure is loaded into the internal core database (see Section 24.5.3). The entry is then annotated by PDB staff using *ADIT*; several validation reports about the structure are produced. The completely annotated entry as it will appear in the PDB resource, together with the validation information, is sent back to the depositor (step 2). After reviewing the processed file, the author sends any revisions (step 3). Depending on the nature of these revisions, steps 2 and 3 may be repeated. Once approval is received from the author (step 4), the entry and the tables in the internal core database are ready for distribution.

All aspects of data processing, including communications with the author, are recorded and stored in the correspondence archive. This makes it possible for the PDB staff to retrieve information about any aspect of the deposition process and to monitor the efficiency of PDB operations closely.

Current status information including a list of authors, title and release category is stored for each entry in the core database and is made accessible for query *via* the WWW interface (http://www.rcsb.org/pdb/status.html). Entries before release are categorized as 'in processing' (PROC), 'in depositor review' (WAIT), 'to be held until publication' (HPUB) or 'on hold until a depositor specified date' (HOLD).

24.5.2.1. *Content of the data collected by the PDB*

All the data collected from depositors by the PDB are considered primary data. Primary data contain, in addition to the coordinates, general information required for all deposited structures and information specific to the method of structure determination. Table 24.5.2.1 contains the general information that the PDB collects for all structures as well as the additional information collected for those structures determined by X-ray methods. The additional items listed for the NMR structures are derived from the International Union of Pure and Applied Chemistry recommendations (Markley *et al.*, 1998) and will be implemented in the near future.

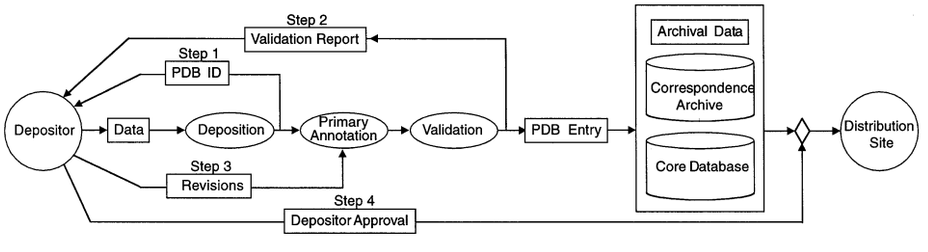

Fig. 24.5.2.1. The steps in PDB data processing. Ellipses represent actions and rectangles define content.

* The Research Collaboratory for Structural Bioinformatics (RCSB) is a consortium consisting of three institutions: Rutgers, The State University of New Jersey; San Diego Supercomputer Center, University of California, San Diego; and the National Institute of Standards and Technology.

† A call for proposals was issued by the National Science Foundation in 1998. The award was made to the RCSB after peer review of the proposals submitted.

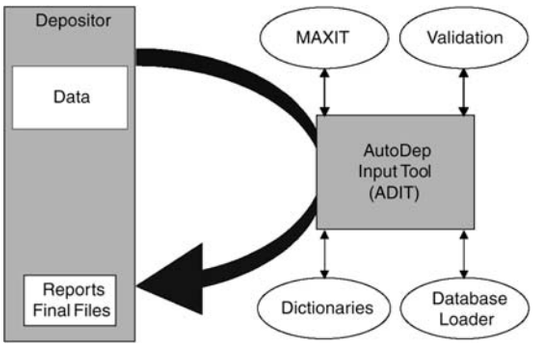

Fig. 24.5.2.2. The integrated tools of the PDB data-processing system.

The information content of data submitted by the depositor is likely to change as new methods for data collection, structure determination and refinement evolve and advance. In addition, the ways in which these data are captured is likely to change as the software for structure determination and refinement produce the necessary data items as part of their output. The data-input system for the PDB, *ADIT*, has been designed so as to incorporate these likely changes easily.

24.5.2.2. *Validation*

Validation refers to the procedure for assessing the quality of deposited atomic models (structure validation) and for assessing how well these models fit the experimental data (experimental validation). The PDB validates structures using accepted community standards as part of *ADIT*'s integrated data-processing system. All validation reports are communicated directly to the depositor. It is also possible to run these validation checks against structures that are not being deposited. A validation server (http://pdb.rutgers.edu/validate/) has been made available for this purpose.

Several types of checks are used in this process: *PROCHECK* (Laskowski *et al.*, 1993) is used for checking the structural features of proteins and *NUCheck* (Feng, Westbrook & Berman, 1998) is used for checking the structural features of nucleic acids. The information currently checked includes the following: bond lengths and bond angles, nomenclature, sequence, stereochemistry, torsion angles, ligand geometry, planarity of peptide bonds, intermolecular

Table 24.5.2.1. *Content of data in the PDB*

(*a*) Content of all depositions (X-ray and NMR)

Source – specifications such as genus, species, strain, or variant of gene (cloned or synthetic); expression vector and host, or description of method of chemical synthesis
Sequence – full sequence of all macromolecular components
Chemical structure of cofactors and prosthetic groups
Names of all components in structure
Qualitative description of characteristics of structure
Literature citations for the structure submitted
Three-dimensional coordinates

(*b*) Additional items for X-ray structure determinations

Temperature factors and occupancies assigned to each atom
Crystallization conditions, including pH, temperature, solvents, salts, methods
Crystal data, including the unit-cell dimensions and space group
Presence of noncrystallographic symmetry
Data-collection information describing the methods used to collect the diffraction data including instrument, wavelength, temperature and processing programs
Data-collection statistics including data coverage, R_{sym}, data above 1, 2, 3σ levels and resolution limits
Refinement information including R factor, resolution limits, number of reflections, method of refinement, σ cutoff, geometry r.m.s.d.
Structure factors – h, k, l, F_{obs}, $\sigma(F_{obs})$

(*c*) Additional items for NMR structure determinations

Model number for each coordinate set that is deposited and an indication if one should be designated as a representative, or an energy-minimized average model provided
Data-collection information describing the types of methods used, instrumentation, magnetic field strength, console, probe head, sample tube
Sample conditions, including solvent, macromolecule concentration ranges, concentration ranges of buffers, salts, antibacterial agents, other components, isotopic composition
Experimental conditions, including temperature, pH, pressure and oxidation state of structure determination and estimates of uncertainties in these values
Non-covalent heterogeneity of sample, including self-aggregation, partial isotope exchange, conformational heterogeneity resulting in slow chemical exchange
Chemical heterogeneity of the sample (*e.g.* evidence for deamidation or minor covalent species)
A list of NMR experiments used to determine the structure including those used to determine resonance assignments, NOE/ROE data, dynamical data, scalar coupling constants, and those used to infer hydrogen bonds and bound ligands. The relationship of these experiments to the constraint files are given explicitly
Constraint files used to derive the structure as described in task-force recommendations

Table 24.5.2.2. *Demographics of the released data in the PDB as of 14 September 1999*

Experimental technique	Molecule type				
	Proteins, peptides, and viruses	Protein–nucleic acid complexes	Nucleic acids	Carbohydrates and other	Total
X-ray diffraction and other	7946	390	439	14	8789
NMR	1365	53	270	4	1692
Theoretical modelling	202	16	15	0	233
Total	9513	459	724	18	10714

contacts, and positions of water molecules. In consultation with the community, other structure checks will be implemented over the next few years.

The experimental data are also checked. Currently, X-ray crystallographic data are validated and plans for checking NMR data are in progress. For X-ray crystallographic structures, the structure factors are validated using *SFCHECK* (Vaguine *et al.*, 1999). This program extracts the deposited *R* factor, resolution and model information, and then compares them with values calculated from coordinate and structure-factor files. It also calculates an overall *B* factor, coordinate errors, an effective resolution and completeness. The summary of the density correlation shift and *B* factor are reported for each residue. As specific procedures are developed for checking NMR structures against experimental data, they will be incorporated into the PDB validation procedures.

24.5.2.3. *NMR data*

The PDB staff recognize that NMR data need a special development effort. Historically these data have been retro-fitted into a PDB format defined around crystallographic information. As a first step towards improving this situation, the PDB carried out an extensive assessment of the current NMR holdings and presented the findings to a task force consisting of a cross section of NMR researchers. The PDB is working with this group, the BioMagResBank (BMRB; Ulrich *et al.*, 1989) and other members of the NMR community to develop an NMR data dictionary along with deposition and validation tools specific for NMR structures.

24.5.2.4. *Data-processing statistics*

Production processing of PDB entries by the RCSB began on 27 January 1999. As of 1 July 1999, when the RCSB became fully responsible for the PDB, approximately 80% of all structures submitted to the PDB are deposited *via ADIT* and processed by the RCSB. Another 20% are submitted *via AutoDep* to the European Bioinformatics Institute (EBI), who process these submissions and forward them to the PDB for archiving and distribution. The average time from deposition to the completion of data processing including author interactions is two weeks. The number of structures with a HOLD release status remains at about 20% of all submissions; 57% are held until publication (HPUB); and 23% are released immediately after processing.

Table 24.5.2.2 shows the breakdown of the types of structures in the PDB. As of 14 September 1999, the PDB contained 10 714 publicly accessible structures with another 1169 entries on hold (not shown). Of these, 8789 (82%) were determined by X-ray methods, 1692 (16%) were determined by NMR and 233 (2%) were theoretical models. Overall, 35% of the entries have deposited experimental data.

24.5.3. The PDB database resource

24.5.3.1. *The database architecture*

In recognition of the fact that no single architecture can fully express the information content of the PDB, an integrated system of heterogeneous databases and indices that store and organize the structural data has been created. At present there are five major components (Fig. 24.5.3.1):

(1) The core relational database managed by Sybase (Sybase Inc., 1995) provides the central physical storage for the primary experimental and coordinate data described in Table 24.5.2.1. The core PDB relational database contains all deposited information in a tabular form that can be accessed across any number of structures.

(2) The final curated data files (in PDB format) and data dictionaries are the archival data and are present as ASCII files in the ftp archive.

(3) The POM-based databases (Shindyalov & Bourne, 1997) consist of indexed objects containing native (*e.g.* atomic coordinates) and derived properties (*e.g.* calculated secondary-structure assignments and property profiles). Some properties require no derivation, for example, *B* factors; others must be derived, for example, exposure of each amino-acid residue (Lee & Richards, 1971) or Cα contact maps. Properties requiring significant computation time, such as structure neighbours (Shindyalov & Bourne, 1998), are pre-calculated when the database is incremented to save considerable user-access time.

(4) The Biological Macromolecule Crystallization Database (BMCD; Gilliland, 1988) is organized as a relational database within Sybase and contains three general categories of literature-derived information: macromolecular, crystal and summary data.

(5) The Netscape LDAP server is used to index the textual content of the PDB in a structured format and provides support for keyword searches.

In the current implementation, communication among databases has been accomplished using the common gateway interface (CGI). An integrated web interface dispatches a query to the appropriate database(s), which then executes the query. Each database returns the PDB identifiers that satisfy the query, and the CGI program

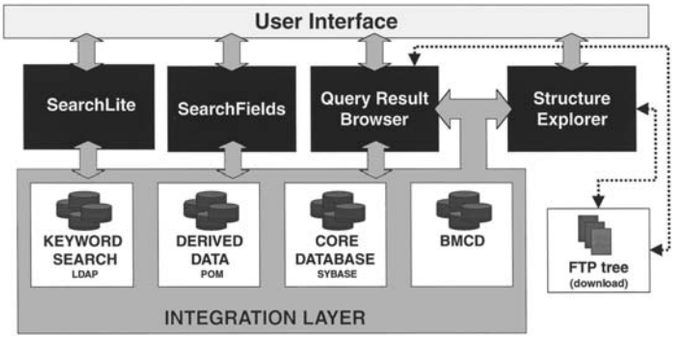

Fig. 24.5.3.1. The integrated query interface to the PDB.