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.

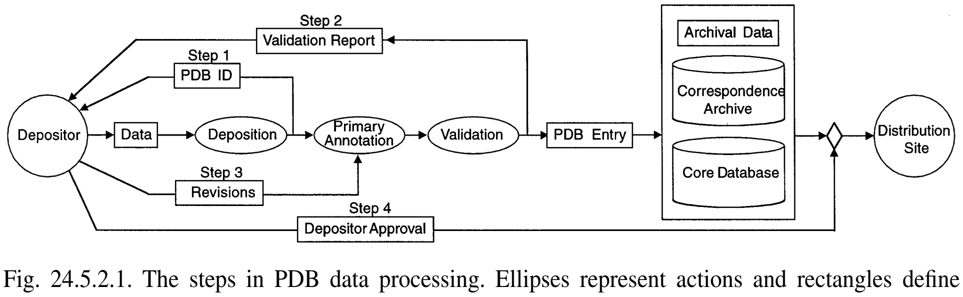

Fig. 24.5.2.1. The steps in PDB data processing. Ellipses represent actions and rectangles define content.

* 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.