5.5. The use of mmCIF architecture for PDB data management

BY J. D. WESTBROOK, H. YANG, Z. FENG AND H. M. BERMAN

5.5.1. Introduction

The Protein Data Bank (PDB) is an archive for macromolecular structures (Bernstein *et al.*, 1977; Berman *et al.*, 2000) and a major component of a global resource for macromolecular structural science (Berman *et al.*, 2003). The scale of its data handling operations is large, and depends on the effective exploitation of the latest developments in the science and technology of informatics. A significant component of its data storage and retrieval strategy is the management of structural data in mmCIF format with appropriate extensions.

Over its 30-year history, the PDB archive has grown from seven entries in 1973 to a collection of over 30 000 structures as of May 2005. The growth in the size of the archive has been accompanied by increases in both data content and in the structural complexity of individual entries. As the PDB has grown, there has been a significant broadening of its user community. In response to this change, the role of the PDB has expanded from being simply a provider of structure data files to providing a key information resource for the structural biology community.

Looking forward, an acceleration in the growth of the PDB archive is anticipated owing to developments in high-throughput structural determination methodologies and worldwide structural genomics efforts. To support the continued growth and evolution of the PDB archive, a framework is required that supports automation and scalability, and that can adapt to changes in both data content and delivery technology.

At the core of the PDB informatics infrastructure is an ontology of data definitions which electronically encode domain information in the form of precise definitions, examples and controlled vocabularies. In addition to domain information, data definitions also encode information such as data type, data relationships, range restrictions and presentation units.

The software-accessible PDB exchange data dictionary (Appendix 3.6.2) is the key part of the PDB informatics infrastructure. The exchange dictionary is an extension of the macromolecular Crystallographic Information File (mmCIF) data dictionary (Bourne *et al.*, 1997). The dictionary provides the foundation for software tools which exchange and validate data, create and load databases, translate data formats, and serve application program interfaces. The components of the informatics infrastructure developed by the PDB are being used to build a data pipeline to support high-throughput structure determination.

5.5.2. Representing macromolecular structure data

Macromolecular structure data have historically been represented in a simple record-oriented format developed by the PDB; this format has been widely used in structural and computational biology.

```
CRYST1  129.230   60.440   56.630  90.00 119.05  90.00 C 1 2 1       4
ATOM      1  N   ASP A   1      23.482  -0.621  -1.419  1.00 35.27           N
ATOM      2  CA  ASP A   1      24.897  -0.728  -1.885  1.00 32.46           C
ATOM      3  C   ASP A   1      25.573   0.515  -1.339  1.00 28.22           C
ATOM      4  O   ASP A   1      24.918   1.359  -0.744  1.00 29.11           O
ATOM      5  CB  ASP A   1      24.976  -0.729  -3.427  1.00 38.24           C
```

Fig. 5.5.2.1. Excerpt of records from a PDB data file.

While this PDB format has in general been adequate for representing coordinate data, it has proved less satisfactory for the description of related information such as chemical and biological features and experimental methodology. To provide a more rigorous data encoding that includes all of this related information, the Protein Data Bank has in recent years adopted a comprehensive ontology of structure and experiment based on the content of the mmCIF data dictionary.

5.5.2.1. PDB format

For the past 30 years, the PDB has served as the single central repository for macromolecular structure data. The data format used to store archival entries in the PDB is a column-oriented data format resembling many data formats developed to accommodate the limitations of paper punched-card technology (see Chapter 1.1). An example of the data format is shown in Fig. 5.5.2.1.

Many of the data records in this format are prefixed with a record tag (*e.g.* CRYST1, ATOM) followed by individual items of data. The specifications for the records in this data format are described informally by Callaway *et al.* (1996). In addition to the labelled records as in Fig. 5.5.2.1, many data records in the PDB format are presented as unstructured or only semi-structured remark records.

5.5.2.2. Ontology representation of macromolecular structure data

In 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) assumed the management responsibilities for the PDB. One important outcome was the change in the underlying data representation used to process PDB data. The PDB now collects and processes data using a data representation based on a comprehensive ontology of macromolecular structure and experiment: the PDB exchange data dictionary. This representation is an extension of the mmCIF data dictionary, now the standard data representation for experimentally determined three-dimensional macromolecular structures. The dictionary and data files based on this data ontology (Westbrook & Bourne, 2000) are expressed using Self-defining Text Archival and Retrieval (STAR) syntax (Chapter 2.1).

Although the mmCIF dictionary was developed within the crystallographic community, the metadata model employed by mmCIF is quite general and has been adopted by other application domains including NMR, molecular modelling and molecular recognition (dictionaries are available at http://mmcif.pdb.org/). Within the crystallographic community, metadata dictionaries have also been

Affiliations: JOHN D. WESTBROOK, HUANWANG YANG, ZUKANG FENG and HELEN M. BERMAN, Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers, The State University of New Jersey, Department of Chemistry and Chemical Biology, 610 Taylor Road, Piscataway, NJ 08854-8087, USA.

references