FACULTY OF ENGINEERING AND TECHNOLOGY

Department of Biotechnology

3D proteins structure file formats: PDB, CIF, MMDB

The **Protein Data Bank (pdb) file format** is a textual file format describing the three-dimensional structures of molecules held in the Protein Data Bank. The pdb format accordingly provides for description and annotation of protein and nucleic acid structures including atomic coordinates, secondary structure assignments, as well as atomic connectivity. In addition experimental metadata are stored. PDB format is the legacy file format for the Protein Data Bank which now keeps data on biological macromolecules in the newer mmCIF file format.

The PDB file format was invented in 1976 as a human-readable file that would allow researchers to exchange protein coordinates through a database system. Its fixed-column width format is limited to 80 columns, which was based on the width of the computer punch cards that were previously used to exchange the coordinates. Through the years the file format has undergone many changes and revisions

STRUCTURE FILE FORMATS

PDB

The PDB file format is column oriented, like that of the punched cards used by early FORTRAN programmers. The exact file format specification is available through the PDB Web site. Most software developed by structural scientists is written in FORTRAN, whereas the rest of the bioinformatics world has adopted other languages, such as those based on C. PDB files are often a paradox: they look rather easy to parse, but they have a few nasty surprises, as already alluded to in this chapter. To the uninitiated, the most obvious problem is that the information about biopolymer bonds is missing, obliging one to program in the rules of chemistry, clues to the identity of each atom given by the naming conventions of PDB, and robust exception handling. PDB parsing software often needs lists of synonyms and tables of exceptions to correctly interpret the information. However this chapter is not intended to be a manual of how to construct a PDB parser.

Two newer chemical-based formats have emerged: mmCIF (MacroMolecular Chemical Interchange Format) and MMDB (Molecular Modeling Database Format). Both of these file formats are attempts to modernize PDB information. Both start by using data description languages, which are consistently machine parsable. The data description languages use "tag value" pairs, which are like variable names and values used in a programming language. In both cases, the format specification is composed in a machine-readable form, and there is software that uses this format specification document to validate incoming streams of data. Both file formats are populated from PDB file data using the strategy of alignment-based reconstruction of the implicit ATOM and HETATM chemical graphs with the explicit SEQRES chemical graphs, together with extensive validation, which is recorded in the file. As a result, both of these file formats are superior for integrating with biomolecular sequence databases over PDB format data files, and their use in future software is encouraged.

Crystallographic Information Framework

The International Union of Crystallography is the sponsor of the **Crystallographic Information Framework**, a standard for information interchange in crystallography.

The acronym CIF is used both for the *Crystallographic Information File*, the data exchange standard file format of Hall, Allen & Brown (1991) and for the *Crystallographic Information Framework*, a broader system of exchange protocols based on data dictionaries and relational rules expressible in different machine-readable manifestations, including, but not restricted to, Crystallographic Information File and XML.

XYZ

CIF was developed by the IUCr Working Party on Crystallographic Information in an effort sponsored by the IUCr Commission on Crystallographic Data and the IUCr Commission on Journals, and was adopted in 1990 as a standard file structure for the archiving and distribution of crystallographic information. It is now well established and is in regular use for reporting crystal structure determinations to *Acta Crystallographica* and other journals. It is often cited as a model example of integrating data and textual information for data-centric scientific communication.

XYZ

In 2006 the importance of CIF and the value of its accompanying web-based service for the validation of structural data, *checkCIF*, were recognised by the Award for Publishing Innovation of the Association of Learned and Professional Society Publishers (ALPSP). In their report, the judges 'were impressed with the way in which CIF and checkCIF are easily accessible and have served to make critical crystallographic data more consistently reliable and accessible at all stages of the information chain, from authors, reviewers and editors through to readers and researchers. In doing so, the system takes away the donkeywork from ensuring that the results of scientific research are trustworthy without detracting from the value of human judgement in the research and publication process'.

This part of the IUCr web site provides comprehensive documentation and software resources for users and developers of CIF

software. https://www.iucr.org/resources/cif

mmCIF

The mmCIF (Bourne et al., 1995) file format was originally intended to be a biopolymer extension of the CIF (Chemical Interchange Format; Hall et al., 1991) familiar to small-molecule crystallographers and is based on a subset of the STAR syntax (Hall et al., 1991). CIF software for parsing and validating format specifications is not forward-compatible with mmCIF, since these have different implementations for the STAR syntax. The underlying data organization in an mmCIF record is a set of relational tables. The mmCIF project refers to their format specification as the mmCIF dictionary, kept on the Web at the Nucleic Acids Database site. The mmCIF dictionary is a large document containing specifications for holding the information stored in PDB files as well as many other data items derivable from the primary coordinate data, such as bond angles. The mmCIF data specification gives this data a consistent interface, which has been used to implement the NDB Protein Finder, a Web-based query format in a relational database style, and is also used as the basis for the new RCSB software systems.

Validating an incoming stream of data against the large mmCIF dictionary entails significant computational time; hence, mmCIF is probably destined to be an archival and advanced query format. Software libraries for reading mmCIF tables into relational tables into memory in FORTRAN and C are available.

MMDB

The MMDB file format is specified by means of the ASN.1 data description language (Rose, 1990), which is used in a variety of other settings, surprisingly enough including applications in telecommunications and automotive manufacturing. Because the US National Library of Medicine also uses ASN.1 data specifications for sequence and bibliographic information, the MMDB format borrows certain elements from other data specifications, such as the parts used in describing bibliographic references cited in the data record. ASN.1 files can appear as human-readable text files or as a variety of binary and packed binary files that can be decoded by any hardware platform. The MMDB standard residue dictionary is a lookup table of information about the chemical graphs of standard biopolymer residue types. The MMDB format specification is kept inside the NCBI toolkit distribution, but a browser is available over the Web for a quick look. The MMDB ASN.1 specification is much more compact and has fewer data items than the mmCIF dictionary, avoiding derivable data altogether.

In contrast to the relational table design of mmCIF, the MMDB data records are structured as hierarchical records. In terms of performance, ASN.1-formatted MMDB files provide for much faster input and output than do mmCIF or PDB records. Their nested hierarchy requires fewer validation steps at load time than the relational scheme in mmCIF or in the PDB file format; hence, ASN.1 files are ideal for three-dimensional structure database browsing.

A complete application programming interface is available for MMDB as part of the NCBI toolkit, containing a wide variety of C code libraries and applications. Both an ASN.1 input/output programming interface layer and a molecular computing layer (MMDB-API) are present in the NCBI toolkit. The NCBI toolkit supports x86 and alpha-based Windows' platforms, Macintosh 68K and PowerPC CPUs, and a wide variety of UNIX platforms. The three-dimensional structure database viewer (Cn3D) is an MMDB-API-based application with source code included in the NCBI toolkit.

XYZ

abc

XYZ

abc

XYZ

abc

XYZ

abc

MCQs

1. A
2. A
3. A
4. A
5. A
6. A
7. A
8. A
9. A
10. A