International Tables for Crystallography (2006). Vol. G, Section 3.6.2, pp. 144–145.

3.6. Classification and use of macromolecular data

By P. M. D. Fitzgerald, J. D. Westbrook, P. E. Bourne, B. McMahon, K. D. Watenpaugh and H. M. Berman
with Appendix 3.6.2 by J. D. Westbrook, K. Henrick, E. L. Ulrich and H. M. Berman

3.6.1. Introduction

As described in Chapter 1.1, the macromolecular crystallographic information file (mmCIF) dictionary (Fitzgerald *et al.*, 1996; Bourne *et al.*, 1997) was initially commissioned as an extension to the core CIF dictionary (Hall *et al.*, 1991), with the intention of adding data names suitable for a full description of a macromolecular crystallographic experiment and its results. However, the need to specify relationships between the data items describing different components of a complex macromolecular structure led to the development of a richer dictionary definition language (DDL2). The data names were then defined according to the DDL2 formalism. For consistency, the existing core dictionary data items were also recast in the DDL2 formalism. Since no other DDL2 applications were envisaged at that time, the core items were then embedded in the mmCIF dictionary as a subset of the complete dictionary. The current release of the mmCIF dictionary described in this chapter includes all the data items in version 2.3.1 of the core dictionary. The mmCIF dictionary is not routinely updated to match additions to the core dictionary, but it is expected that when new versions of the mmCIF dictionary are released to meet the requirements of the macromolecular community, the most recent version of the core dictionary will be incorporated in the new mmCIF dictionary as part of the revision.

The resulting stand-alone dictionary is very large and is described in detail in this chapter. The philosophy behind the design of the dictionary is discussed in Section 3.6.2 and an example of its use is given and discussed in Section 3.2.3. The contents of the dictionary are then described in the remainder of the chapter, starting at Section 3.6.4. The discussion follows the sequence of Table 3.1.10.1: experimental measurements, analysis, structure, publication and file metadata are considered in turn. The discussion of individual categories may be found by using the overview of the dictionary structure given in Appendix 3.6.1.

The data names in the mmCIF dictionary derived from the core CIF dictionary differ from their DDL1 counterparts in that a full stop (.) is used to designate explicitly the category to which the data name belongs, *e.g.* **`_cell.length_a`** is used in place of **`_cell_length_a`**. Sometimes the mmCIF counterpart of a core data name may have a different form, for example to enforce the rule in DDL2 that the category name is the initial part of any data name within that category. This convention is generally observed in DDL1, but is not mandatory. Formally, the corresponding DDL1 core data name is obtained from the **`_item_aliases.alias_name`** attribute of the definition. The provision of a formal alias for all data names derived from the core dictionary allows a DDL2-compliant parser to read and interpret a data file constructed according to the DDL1 dictionary described in Chapter 3.2. Achieving this compatibility with CIFs built using DDL1 dictionaries was a very important goal in the design of DDL2 and the mmCIF dictionary.

In this chapter, categories and individual data names that correspond to matching entries in the core dictionary are not discussed in detail unless they are used in a different way in mmCIF. Chapter 3.2 should therefore be read first for a description of the categories common to both the core and mmCIF dictionaries. This chapter concentrates on the categories specific to mmCIF. Formal differences between mmCIF categories and core CIF categories are also summarized.

3.6.2. Considerations underlying the design of the dictionary

From the outset, mmCIF was envisaged as a providing a more detailed description of macromolecular structures than the existing Protein Data Bank (PDB) format (Chapter 1.1). A number of considerations guided the development of version 1 of the mmCIF dictionary. These included:

(i) Every field of every PDB record type should be represented by an mmCIF data item if the PDB field is important for describing the structure, the experiment that was conducted in determining the structure or the revision history of the entry. It is important to note that it is straightforward to convert an mmCIF data file to a PDB file without loss of information, since all the information is parsable. It is not possible, however, to automate completely the conversion of a PDB file to an mmCIF, since many mmCIF data items are either not present in the PDB file or are present in PDB REMARK records that in some cases cannot be parsed. The contents of PDB REMARK records are maintained as separate data items within mmCIF so as to preserve all the information, even if the information is not parsable.

(ii) Data items should be defined so that all the information given in the materials and methods section of an article describing the structure can be referenced. This includes major features of the crystal, the diffraction experiment, the phasing calculations and the refinement.

(iii) Data items should be provided for describing the biologically active molecule and any important structural subcomponents.

(iv) It should be possible to represent atom positions using either orthogonal ångström or fractional coordinates.

(v) Data items should be provided for describing the initial experimental reflection data, including all the data sets used in the phasing of the structure, and the final processed data.

(vi) Crystallographic and noncrystallographic symmetry should be described.

Affiliations: Paula M. D. Fitzgerald, Merck Research Laboratories, Rahway, New Jersey, USA; John D. Westbrook, 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, New Jersey, USA; Phillip E. Bourne, Research Collaboratory for Structural Bioinformatics, San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0537, USA; Brian McMahon, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England; Keith D. Watenpaugh, retired; formerly Structural, Analytical and Medicinal Chemistry, Pharmacia Corporation, Kalamazoo, Michigan, USA; 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, New Jersey, USA; Kim Henrick, EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England; Eldon L. Ulrich, Department of Biochemistry, University of Wisconsin Madison, 433 Babcock Drive, Madison, WI 53706-1544, USA.

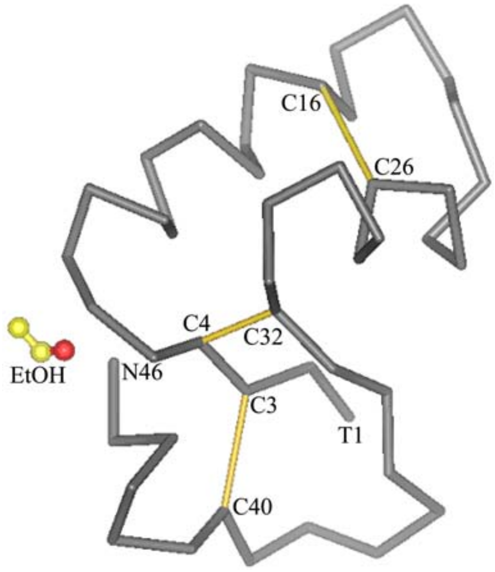

Fig. 3.6.3.1. A representation of crambin (PDB 3CNR) with a co-crystallized ethanol molecule.

(vii) Data items should be present for describing the characteristics and geometry of canonical and non-canonical amino acids, nucleotides, sugars and ligand groups.

(viii) Data items should be provided that permit a detailed description of the chemistry of the component parts of the macromolecule to be given.

(ix) Data items should be present that provide specific pointers from elements of the structure (*e.g.* the sequence, bound inhibitors) to appropriate entries in publicly available databases.

(x) Data items should be present that provide meaningful three-dimensional views of the structure so as to highlight functional and structural aspects of the macromolecule.

(xi) Data items specific to an NMR experiment or modelling study would not in general be included in version 1. However, data items that summarize the features of an ensemble of structures and permit a description of each member of the ensemble to be given should be available.

(xii) A comprehensive set of data items for providing a higher-order structure description (for example, to cover supersecondary structure and functional classification) was considered to be beyond the scope of version 1.

Based on the above, the first version of the mmCIF dictionary with approximately 1700 data items (including those data items taken from the core CIF dictionary) was developed and officially approved in October 1997. Subsequent revisions have increased the number of data items to over 2000. It is not expected that all the data items will be present in every mmCIF data file. Instead, the goal was to provide a wide range of data items from which users can select those that best suit the structure they wish to describe.

3.6.3. Overview of the mmCIF data model

The solution and refinement of a macromolecular structure is complex and often difficult, as there are a large number of atoms in a typical macromolecule, the molecular conformation can be complex and it can be difficult to model included solvent molecules. However, even when a satisfactory structural model has been derived, describing the structure can be a considerable challenge. Using diagrams can help, but two-dimensional projections are often inadequate for illustrating important features and a complete understanding of the three-dimensional structure of a macromolecule can often only be reached by using interactive molecular graphics software.

Example 3.6.3.1. *Specification of the three distinct components of the crambin structure.*

```
loop_
  _struct_asym.id
  _struct_asym.entity_id
  _struct_asym.details
   chain_a A       'single polypeptide chain'
   ethanol ethanol 'cocrystallized ethanol molecule'
   water   HOH     .
```

The mmCIF dictionary provides several ways for describing the structure. The PUBL categories can be used to record text describing the structure. The complete list of atomic coordinates may be used as input for visualization programs that allow a range of wire-frame, stick, space-filling, ribbon or cartoon representations to be generated based upon inbuilt heuristics and user interaction. However, most importantly, the mmCIF approach also offers a large collection of categories which are designed to provide descriptions of the structure at different levels of detail, and the relationships between data items in different categories permit the function of an individual atom site at any particular level of detail to be traced.

Before beginning the detailed description of the full mmCIF dictionary, it is helpful to demonstrate how it is used to describe the structure of a biological macromolecule. Fig. 3.6.3.1 shows the small protein crambin, which is a single polypeptide chain of 48 residues. The molecule co-crystallizes with a molecule of ethanol, although this is not thought to have any biological effect. Almost a quarter of the residues have side chains that adopt alternative conformations, and there is sequence heterogeneity at positions 22 (Pro/Ser) and 25 (Leu/Ile). Three disulfide links stabilize the structure.

The highest level of the description of the structure uses data items from the STRUCT category group. The crystallographic asymmetric unit contains one protein molecule, one co-crystallization ethanol molecule and a water solvent molecule. These are described with data items from the STRUCT_ASYM category (Example 3.6.3.1).

Each entry in this list assigns a label to a discrete component of the asymmetric unit and associates it with an entry in the entity list that defines each distinct chemical species in the crystal (Example 3.6.3.2).

The biological functions of the components of the crystal structure are described using data items in the STRUCT_BIOL and related categories. For crambin, the biological function is still unknown (see Example 3.6.3.3). This example also shows how the biological unit is generated from specific discrete objects in the asymmetric unit. In this case the relationship is trivial, but it will often be much more complex.

The secondary structure of the protein is described using data items in the STRUCT_CONF category (and in the STRUCT_SHEET category where relevant). The beginning and end labels for each

Example 3.6.3.2. *Specification of the distinct chemical entities in the crambin structure.*

```
loop_
  _entity.id
  _entity.type
  _entity.formula_weight
  _entity.src_method
   A        polymer      4716     natural
   ethanol  non-polymer    52     synthetic
   HOH      water          18     .
```