International Tables for Crystallography (2006). Vol. G, Section 3.6.3, pp. 145–147.

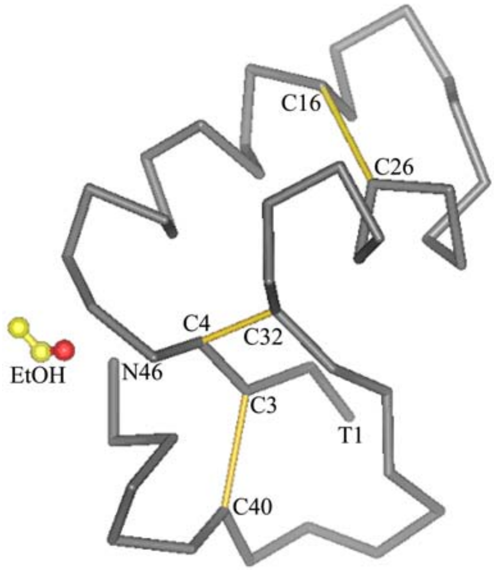

Fig. 3.6.3.1. A representation of crambin (PDB 3CNR) with a co-crystallized ethanol molecule.

(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     .
```

Example 3.6.3.3. *Identification of the biological function of the components of the crambin structure.*

```
_struct_biol.id              crambin_1
_struct_biol.details
; The function of this protein is unknown and
  therefore the biological unit is assumed to be
  the single polypeptide chain without
  co-crystallization factors i.e. ethanol.
;
_struct_biol_gen.biol_id     crambin_1
_struct_biol_gen.asym_id     chain_a
_struct_biol_gen.symmetry    1_555
```

α-helix, β-strand or turn in Example 3.6.3.4 refer to the chemical components of the structural unit labelled chain_a at the given locations in the sequence (*e.g.* helix H1 runs from the isoleucine at position number 7 to the proline at position number 19 in the amino-acid sequence).

Interactions between different parts of the structure are described using data items in the STRUCT_CONN and related categories. In Example 3.6.3.5, some of the disulfide bridges and intramolecular hydrogen bonds are reported. As with the secondary structural elements, the partners in the links are identified by complex labels that include the chemical component involved, the object within the asymmetric unit that is under consideration, the position in the amino-acid (or nucleotide) sequence and the individual atom.

The objects identified at the highest level of the description of the structure are arbitrary. To discover their chemical identity, one needs to consult the ENTITY category group. As indicated above, each separate chemical species in the crystal should be specified in the entity table. Chemical entities are classified as polymer, non-polymer or water. Non-polymeric molecules, such as the co-crystallized ethanol in this example, are described as distinct chemical components using data items in the CHEM_COMP family of categories. Polymeric molecules are described using data items in the ENTITY_POLY family of categories.

In Example 3.6.3.6, the natural source for crambin is described, the overall features of the polypeptide chain are listed and the component parts (in effect the amino-acid sequence) are tabulated. Note that sequence heterogeneity is described by allowing a sequence number to be correlated with more than one monomer identifier (in the example, sequence number 22 is assigned both to proline and serine, while 25 is assigned to both leucine and isoleucine). Sequence heterogeneity can be defined by assigning suitable labels in the ATOM_SITE list.

Example 3.6.3.4. *Description of the secondary structure of crambin.*

```
loop_
  _struct_conf.id
  _struct_conf.conf_type_id
  _struct_conf.beg_label_comp_id
  _struct_conf.beg_label_asym_id
  _struct_conf.beg_label_seq_id
  _struct_conf.end_label_comp_id
  _struct_conf.end_label_asym_id
  _struct_conf.end_label_seq_id
  _struct_conf.details
    H1 HELX_RH_AL_P ILE chain_a  7 PRO chain_a 19
                                  'HELX-RH3T 17-19'
    H2 HELX_RH_AL_P GLU chain_a 23 THR chain_a 30
                                  'Alpha-N start'
    S1 STRN_P       CYS chain_a 32 ILE chain_a 35 .
    S2 STRN_P       THR chain_a  1 CYS chain_a  4 .
    S3 STRN_P       ASN chain_a 46 ASN chain_a 46 .
    S4 STRN_P       THR chain_a 39 PRO chain_a 41 .
    T1 TURN-TY1_P   ARG chain_a 17 GLY chain_a 20 .
    T2 TURN-TY1_P   PRO chain_a 41 TYR chain_a 44 .
```

Example 3.6.3.5. *Interactions between parts of the crambin structure.*

```
loop_
  _struct_conn.id
  _struct_conn.conn_type_id
  _struct_conn.ptnr1_label_comp_id
  _struct_conn.ptnr1_label_asym_id
  _struct_conn.ptnr1_label_seq_id
  _struct_conn.ptnr1_label_atom_id
  _struct_conn.ptnr1_role
  _struct_conn.ptnr1_symmetry
  _struct_conn.ptnr2_label_comp_id
  _struct_conn.ptnr2_label_asym_id
  _struct_conn.ptnr2_label_seq_id
  _struct_conn.ptnr2_label_atom_id
  _struct_conn.ptnr2_role
  _struct_conn.ptnr2_symmetry
  _struct_conn.details
    SS1 disulf CYS chain_a  3 S  .        1_555
               CYS chain_a 40 S  .        1_555 .
    SS2 disulf CYS chain_a  4 S  .        1_555
               CYS chain_a 32 S  .        1_555 .
    HB1 hydrog SER chain_a  6 OG positive 1_555
               LEU chain_a  8 O  negative 1_556 .
    HB2 hydrog ARG chain_a 17 N  positive 1_555
               ASP chain_a 43 O  negative 1_554 .
```

The individual amino acids in the protein sequence of Example 3.6.3.6 are labelled by the data item `_entity_poly_seq.mon_id`; this refers to the separate chemical components listed in the CHEM_COMP family of categories (Example 3.6.3.7). As mentioned above, entries in these categories may be individual monomeric species within the crystal structure, or they may be amino acids or nucleotide bases that form the macromolecular polymer. In most cases, the entries recorded in these categories will be summaries of chemical information for standard amino acids and nucleotides, or references to external libraries of standard data for these. However, the categories contain enough data items to describe modified residues or co-crystallization factors in full if necessary.

At the most detailed level, the individual atom sites are described with data items in the ATOM category group, as shown for crambin in Example 3.6.3.8. A few points about this

Example 3.6.3.6. *Description of the crambin polypeptide.*

```
_entity_name_com.entity_id      A
_entity_name_com.name           crambin
_entity_src_nat.entity_id       A
_entity_src_nat.common_name     'Abyssinian Cabbage'
_entity_src_nat.genus           Crambe
_entity_src_nat.species         abyssinica
_entity_src_nat.details         ?
_entity_poly.entity_id          A
_entity_poly.type               polypeptide(L)
_entity_poly.nstd_chirality     no
_entity_poly.nstd_linkage       no
_entity_poly.nstd_monomer       no
_entity_poly.type_details
 'Sequence heterogeneity at residues 22 and 25'
loop_
  _entity_poly_seq.entity_id
  _entity_poly_seq.num
  _entity_poly_seq.mon_id
    A    1   THR     A    2   THR
# - - abbreviated - - -
    A   22   PRO     A   22   SER
    A   23   GLU     A   24   ALA
    A   25   LEU     A   25   ILE
# - - abbreviated - - -
    A   47   ALA     A   48   ASN
```

Example 3.6.3.7. *Separate chemical components forming the crambin polypeptide.*

```
loop_
_chem_comp.id
_chem_comp.mon_nstd_flag
_chem_comp.formula
_chem_comp.name
  ethanol .   ’C2 H6 O1’        "ethanol"
  ALA     yes ’C3 H7 N1 O2’     "alanine"
  ARG     yes ’C6 H14 N4 O2’    "arginine"
  ASN     yes ’C4 H8 N2 O3’     "asparagine"
  ASP     yes ’C4 H7 N1 O4’     "aspartic acid"
  CYS     yes ’C3 H7 N1 O2 S1’  "cysteine"
  GLU     yes ’C5 H9 N1 O4’     "glutamic acid"
  GLY     yes ’C2 H5 N1 O2’     "glycine"
  ILE     yes ’C6 H13 N1 O2’    "isoleucine"
  LEU     yes ’C6 H13 N1 O2’    "leucine"
  PHE     yes ’C9 H11 N1 O2’    "phenylalanine"
  PRO     yes ’C5 H9 N1 O2’     "proline"
  SER     yes ’C3 H7 N1 O3’     "serine"
  THR     yes ’C4 H9 N1 O3’     "threonine"
  TYR     yes ’C9 H11 N1 O3’    "tyrosine"
  VAL     yes ’C5 H11 N1 O2’    "valine"
```

example should be noted. The composite labelling of each site includes a pointer to the description of the parent molecule as a specific object in the asymmetric unit (`_atom_site.label_asym_id`) and to the relevant monomeric building block of which the atom is a member (`_atom_site.label_comp_id`). The label component `_atom_site.label_alt_id` indicates alternative conformations in which an atom site may be found. For example, the atom sites numbered 3 and 4 are alternative locations for the α-carbon of the terminal residue. It may be deduced from the occupancies that the alternative conformations A and B are modelled with 80% and 20% occupancy, respectively, but this can be stated explicitly using the ATOM_SITES_ALT category. The sequence heterogeneity at residue 22 is shown by the presence of pointers to proline and serine, and the occupancy factors show that proline and serine are present in the ratio 60 to 40. There is also an alternative conformation within the serine at residue 22, split equally across two sites.

3.6.4. Content of the macromolecular CIF dictionary

Because it is derived from the core CIF dictionary, the mmCIF dictionary shares the same general structure as outlined in Chapter 3.2. However, DDL2 permits the formal assignment of categories to *category groups*. Table 3.6.4.1 lists the major category groups in the mmCIF dictionary (a full list is given in Appendix 3.6.1 and at the beginning of Chapter 4.5).

Small capitals are used for the names of category groups and individual categories in this volume, but the identifiers in the dictionary are actually lower-case strings.

The ordering of category groups in the remainder of this chapter follows the thematic scheme of Table 3.1.10.1. The discussion proceeds under the headings *Experimental measurements* (Section 3.6.5), *Analysis* (Section 3.6.6), *Atomicity, chemistry and structure* (Section 3.6.7), *Publication* (Section 3.6.8) and *File metadata* (Section 3.6.9).

Certain conventions of style and layout have been followed to summarize the large amount of information in the mmCIF dictionary and to help the reader navigate their way through this chapter. Appendix 3.6.1 is an overview of the mmCIF dictionary structure by category and lists all the categories with the number of the section in which they are discussed. This acts as an index between the alphabetical ordering within the dictionary and the thematic ordering of this chapter. Each thematic section lists the categories discussed in that section. Within each subsection, the data names within the relevant categories are listed. Category keys, pointers to parent data items and aliases to data items in the core CIF dictionary are indicated. For each category, the data item (or set of data items that must be considered together) that forms the category key is marked by a bullet (•) and listed first; the other data names follow in alphabetical order.

Example 3.6.3.8. *Partial listing of the atomic coordinates of crambin.*

```
loop_
_atom_site.label_seq_id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_alt_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.footnote_id
_atom_site.label_entity_id
_atom_site.id
1  N  N    THR chain_a A  16.864  14.059   3.442
   0.80  6.22  .   A    1
1  N  N    THR chain_a B  17.633  14.126   4.146
   0.20  8.40  .   A    2
1  C  CA   THR chain_a A  16.868  12.814   4.233
   0.80  4.45  .   A    3
1  C  CA   THR chain_a B  17.282  12.671   4.355
   0.20  7.82  .   A    4
1  C  C    THR chain_a .  15.583  12.775   4.990
   1.00  4.39  .   A    5
1  O  O    THR chain_a .  15.112  13.824   5.431
   1.00  7.04  .   A    6
1  C  CB   THR chain_a A  18.060  12.807   5.200
   0.80  5.42  .   A    7
1  C  CB   THR chain_a B  18.202  11.709   5.108
   0.20 11.07  .   A    8
1  O  OG1  THR chain_a A  19.233  12.892   4.380
   0.80  7.87  .   A    9
1  O  OG1  THR chain_a B  17.662  10.381   4.831
   0.20 14.39  .   A   10
1  C  CG2  THR chain_a A  18.117  11.578   6.092
   0.80  6.88  .   A   11
1  C  CG2  THR chain_a B  17.973  11.955   6.599
   0.20 19.74  .   A   12
# - - abbreviated - - -
22 N  N    PRO chain_a .   4.909  12.659  -3.127
   0.60  3.03  .   A  352
22 C  CA   PRO chain_a .   6.035  13.459  -2.622
   0.60  3.04  .   A  353
22 C  C    PRO chain_a .   6.362  13.139  -1.174
   0.60  3.08  .   A  354
22 O  O    PRO chain_a .   5.473  12.959  -0.323
   0.60  3.67  .   A  355
22 C  CB   PRO chain_a .   5.528  14.895  -2.825
   0.60  4.19  .   A  356
22 C  CG   PRO chain_a .   4.614  14.846  -4.059
   0.60  3.91  .   A  357
22 C  CD   PRO chain_a .   3.904  13.493  -3.885
   0.60  3.25  .   A  358
22 N  N    SER chain_a .   4.909  12.659  -3.127
   0.40  3.03  .   A  366
22 C  CA   SER chain_a .   6.035  13.459  -2.622
   0.40  3.04  .   A  367
22 C  C    SER chain_a .   6.362  13.139  -1.174
   0.40  3.08  .   A  368
22 O  O    SER chain_a .   5.473  12.959  -0.323
   0.40  3.67  .   A  369
22 C  CB   SER chain_a .   5.644  14.934  -2.679
   0.40  3.96  .   A  370
22 O  OG   SER chain_a C   4.712  15.250  -1.677
   0.20  3.53  .   A  371
22 O  OG   SER chain_a D   6.688  15.800  -2.315
   0.20  7.09  .   A  372
```

For measured or derived numerical quantities that should be specified with a standard uncertainty (in older terminology, an estimated standard deviation), the core dictionary uses the DDL1

references