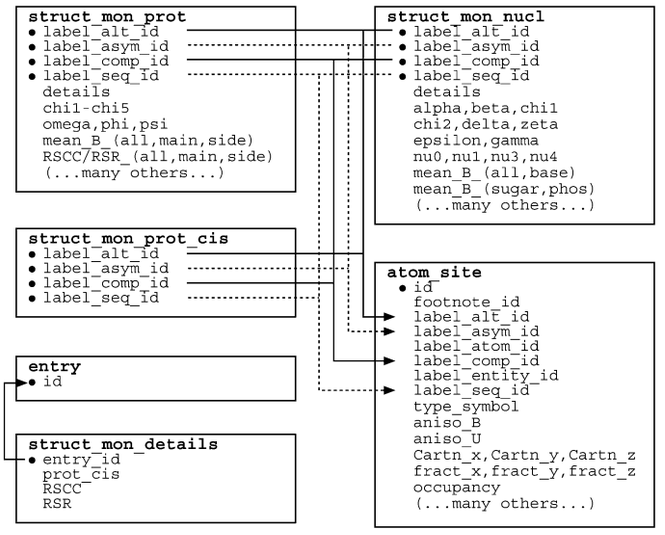

Fig. 3.6.7.11. The family of categories used to describe the structural features of monomers. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Example 3.6.7.11. *A hypothetical example of the structural features of a single protein residue described with data items in the* STRUCT_MON_PROT *category.*

```
_struct_mon_prot.label_comp_id          ARG
_struct_mon_prot.label_seq_id           35
_struct_mon_prot.label_asym_id          A
_struct_mon_prot.label_alt_id           .
_struct_mon_prot.chi1                 -67.9
_struct_mon_prot.chi2                -174.7
_struct_mon_prot.chi3                 -67.7
_struct_mon_prot.chi4                 -86.3
_struct_mon_prot.chi5                   4.2
_struct_mon_prot.RSCC_all               0.90
_struct_mon_prot.RSR_all                0.18
_struct_mon_prot.mean_B_all            30.0
_struct_mon_prot.mean_B_main           25.0
_struct_mon_prot.mean_B_side           35.1
_struct_mon_prot.omega                180.1
_struct_mon_prot.phi                  -60.3
_struct_mon_prot.psi                  -46.0
```

the mmCIF dictionary, it was found that the biological crystallography community felt that mmCIF should contain data items that allowed the local quality of the model to be recorded: these data items are found in the categories STRUCT_MON_DETAILS, STRUCT_MON_NUCL (for nucleotides), and STRUCT_MON_PROT and STRUCT_MON_PROT_CIS (for proteins). Using these categories, quantities that reflect the local quality of the structure, such as isotropic displacement factors, real-space *R* factors and real-space correlation coefficients, can be given at the monomer and submonomer levels.

In addition, these categories can be used to record the conformation of the structure at the monomer level by listing side-chain torsion angles. These values can be derived from the atom coordinate list, so it would not be common practice to include them in an mmCIF for archiving a structure unless it was to highlight conformations that deviate significantly from expected values (Engh & Huber, 1991). However, there are applications, such as comparative studies across a number of independent determinations of the same structure, where it would be useful to store torsion-angle information without having to recalculate it each time it is needed.

The relationships between the categories used to describe the structural features of monomers are shown in Fig. 3.6.7.11.

Three indicators of the quality of a structure at the local level are included in this version of the dictionary: the mean displacement (*B*) factor, the real-space correlation coefficient (Jones *et al.*, 1991) and the real-space *R* factor (Brändén & Jones, 1990). Other indicators are likely to be added as they become available. In the current version of the dictionary, these metrics can be given at the monomer level, or at the levels of main- and side-chain for proteins, or base, phosphate and sugar for nucleic acids (Altona & Sundaralingam, 1972).

The variables used when calculating real-space correlation coefficients and real-space *R* factors, such as the coefficients used to calculate the map being evaluated or the radii used for including points in a calculation, can be recorded using the data items `_struct_mon_details.RSC` and `_struct_mon_details.RSR`.

These data items are also provided for recording the full conformation of the macromolecule, using a full set of data items for the torsion angles of both proteins and nucleic acids. Although one could use these data items to describe the whole macromolecule, it is more likely that they would be used to highlight regions of the structure that deviate from expected values (Example 3.6.7.11). Deviations from expected values could imply inaccuracies in the model in poorly defined parts of the structure, but in some cases nonstandard torsion angles are found in very well defined regions and are essential to the proper configurations of active sites or ligand binding pockets.

A special case of nonstandard conformation is the occurrence of *cis* peptides in proteins. As the *cis* conformation occurs quite often, the category STRUCT_MON_PROT_CIS is provided so that an explicit list can be made of *cis* peptides. The related data item `_struct_mon_details.prot_cis` allows an author to specify how far a peptide torsion angle can deviate from the expected value of 0.0 and still be considered to be *cis*.

In these categories, properties are listed by residue rather than by individual atom. The only label components needed to identify the residue are `*_alt`, `*_asym`, `*_comp` and `*_seq`. If the author has provided an alternative labelling system, this can also be used. Since the analysis is by individual residue, there is no need to specify symmetry operations that might be needed to move one residue so that it is next to another.

3.6.7.5.5. *Noncrystallographic symmetry*

Data items in these categories are as follows:

(*a*) STRUCT_NCS_ENS

- `_struct_ncs_ens.id`

  `_struct_ncs_ens.details`

  `_struct_ncs_ens.point_group`

(*b*) STRUCT_NCS_ENS_GEN

- `_struct_ncs_ens_gen.dom_id_1`
  → `_struct_ncs_dom.id`
- `_struct_ncs_ens_gen.dom_id_2`
  → `_struct_ncs_dom.id`
- `_struct_ncs_ens_gen.ens_id`
  → `_struct_ncs_ens.id`
- `_struct_ncs_ens_gen.oper_id`
  → `_struct_ncs_oper.id`

(*c*) STRUCT_NCS_DOM

- `_struct_ncs_dom.id`

  `_struct_ncs_dom.details`

(*d*) STRUCT_NCS_DOM_LIM

- `_struct_ncs_dom_lim.beg_label_alt_id`
  → `_atom_sites_alt.id`
- `_struct_ncs_dom_lim.beg_label_asym_id`
  → `_atom_site.label_asym_id`
- `_struct_ncs_dom_lim.beg_label_comp_id`
  → `_atom_site.label_comp_id`

- **_struct_ncs_dom_lim.beg_label_seq_id**
  → **_atom_site.label_seq_id**
- **_struct_ncs_dom_lim.dom_id**
- **_struct_ncs_dom_lim.end_label_alt_id**
  → **_atom_sites_alt.id**
- **_struct_ncs_dom_lim.end_label_asym_id**
  → **_atom_site.label_asym_id**
- **_struct_ncs_dom_lim.end_label_comp_id**
  → **_atom_site.label_comp_id**
- **_struct_ncs_dom_lim.end_label_seq_id**
  → **_atom_site.label_seq_id**

_struct_ncs_dom_lim.beg_auth_asym_id
→ **_atom_site.auth_asym_id**
_struct_ncs_dom_lim.beg_auth_comp_id
→ **_atom_site.auth_comp_id**
_struct_ncs_dom_lim.beg_auth_seq_id
→ **_atom_site.auth_seq_id**
_struct_ncs_dom_lim.end_auth_asym_id
→ **_atom_site.auth_asym_id**
_struct_ncs_dom_lim.end_auth_comp_id
→ **_atom_site.auth_comp_id**
_struct_ncs_dom_lim.end_auth_seq_id
→ **_atom_site.auth_seq_id**

(*e*) STRUCT_NCS_OPER

- **_struct_ncs_oper.id**

_struct_ncs_oper.code
_struct_ncs_oper.details
_struct_ncs_oper.matrix[1][1]
_struct_ncs_oper.matrix[1][2]
_struct_ncs_oper.matrix[1][3]
_struct_ncs_oper.matrix[2][1]
_struct_ncs_oper.matrix[2][2]
_struct_ncs_oper.matrix[2][3]
_struct_ncs_oper.matrix[3][1]
_struct_ncs_oper.matrix[3][2]
_struct_ncs_oper.matrix[3][3]
_struct_ncs_oper.vector[1]
_struct_ncs_oper.vector[2]
_struct_ncs_oper.vector[3]

The bullet (•) indicates a category key. Where multiple items within a category are marked with a bullet, they must be taken together to form a compound key. The arrow (→) is a reference to a parent data item.

Biological macromolecular complexes may be built from domains related by symmetry transformations other than those arising from the crystal lattice symmetry. These domains are not necessarily discrete molecular entities: they may be composed of one or more segments of a single polypeptide or nucleic acid chain, of segments from more than one chain, or of small-molecule components of the structure. The categories above allow the distinct domains that participate in ensembles of structural elements related by noncrystallographic symmetry to be listed and described in detail. The relationships between categories used to describe noncrystallographic symmetry are shown in Fig. 3.6.7.12.

In the mmCIF model of noncrystallographic symmetry, the highest level of organization is the ensemble, which corresponds to the complete symmetry-related aggregate (*e.g.* tetramer, icosahedron). An identifier is given to the ensemble using the data item **_struct_ncs_ens.id**.

The symmetry-related elements within the ensemble are referred to as domains. The elements of structure that are to be considered part of the domain are specified using the data items in the STRUCT_NCS_DOM and STRUCT_NCS_DOM_LIM categories. By using the STRUCT_NCS_DOM_LIM data items appropriately, domains can be defined to include ranges of polypeptide chain or nucleic acid strand, bound ligands or cofactors, or even bound solvent molecules. Note that the category keys for STRUCT_NCS_DOM_LIM include the domain ID and the range specifiers. Thus a single domain may be composed of any number of ranges of elements.

Finally, the ensemble is generated from the domains using the rotation matrix and translation vector specified by data items in

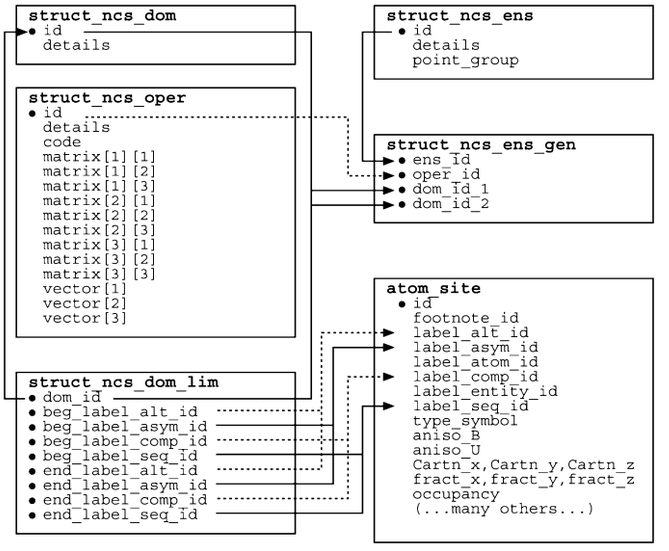

Fig. 3.6.7.12. The family of categories used to describe noncrystallographic symmetry. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

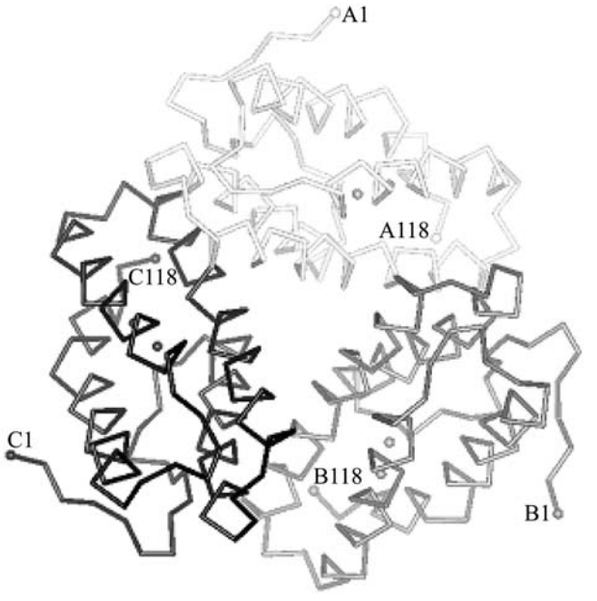

Fig. 3.6.7.13. Noncrystallographic symmetry in the structure of trimeric haemerythrin (PDB 1HR3) to be described with data items in the STRUCT_NCS_ENS, STRUCT_NCS_ENS_GEN, STRUCT_NCS_DOM and STRUCT_NCS_DOM_LIM categories.

the STRUCT_NCS_OPER category, which are referenced by the data items in the STRUCT_NCS_ENS_GEN category. There are data items appropriate for two common methods of describing noncrystallographic symmetry:

(1) In the first method, the coordinate list includes all copies of domains related by noncrystallographic symmetry and the aim is to describe the relationships between domains in the ensemble; in this case the data items in STRUCT_NCS_ENS_GEN specify a pair of domains and reference the appropriate operator in STRUCT_NCS_OPER. This method is indicated by giving the data item **_struct_ncs_oper.code** the value `given`.

(2) In the second method, the coordinate list contains only one copy of the domain and the aim is to generate the entire ensemble; in this case the data items in STRUCT_NCS_ENS_GEN

Example 3.6.7.12. *Noncrystallographic symmetry in the structure of trimeric haemerythrin (PDB 1HR3) described with data items in the* STRUCT_NCS_ENS, STRUCT_NCS_ENS_GEN, STRUCT_NCS_DOM *and* STRUCT_NCS_DOM_LIM *categories. For brevity, the data items in the* STRUCT_NCS_OPER *category are not shown.*

```
_struct_ncs_ens.id                    trimer
_struct_ncs_ens.point_group           3

loop_
_struct_ncs_ens_gen.ens_id
_struct_ncs_ens_gen.dom_id_1
_struct_ncs_ens_gen.dom_id_2
_struct_ncs_ens_gen.oper_id
 trimer  chain_A  chain_B  1
 trimer  chain_A  chain_C  2

loop_
_struct_ncs_dom.id
 chain_A    chain_B    chain_C

loop_
_struct_ncs_dom_lim.dom_id
_struct_ncs_dom_lim.beg_label_asym_id
_struct_ncs_dom_lim.beg_label_comp_id
_struct_ncs_dom_lim.beg_label_seq_id
_struct_ncs_dom_lim.beg_label_alt_id
_struct_ncs_dom_lim.end_label_asym_id
_struct_ncs_dom_lim.end_label_comp_id
_struct_ncs_dom_lim.end_label_seq_id
_struct_ncs_dom_lim.end_label_alt_id
 chain_A  A  ala  1  .  A  ala  118  .
 chain_B  B  ala  1  .  B  ala  118  .
 chain_C  C  ala  1  .  C  ala  118  .
```

specify a pair of domains and reference the appropriate operator in STRUCT_NCS_OPER, but now the data item **_struct_ncs_oper.code** is given the value `generate`.

Noncrystallographic symmetry in a trimeric molecule is shown in Fig. 3.6.7.13 and described in Example 3.6.7.12.

3.6.7.5.6. *External databases*

The data items in these categories are as follows:

(*a*) STRUCT_REF

```
• _struct_ref.id
  _struct_ref.biol_id
        → _struct_biol.id
  _struct_ref.db_code
  _struct_ref.db_name
  _struct_ref.details
  _struct_ref.entity_id
        → _entity.id
  _struct_ref.seq_align
  _struct_ref.seq_dif
```

(*b*) STRUCT_REF_SEQ

```
• _struct_ref_seq.align_id
  _struct_ref_seq.db_align_beg
  _struct_ref_seq.db_align_end
  _struct_ref_seq.details
  _struct_ref_seq.ref_id
        → _struct_ref.id
  _struct_ref_seq.seq_align_beg
        → _entity_poly_seq.num
  _struct_ref_seq.seq_align_end
        → _entity_poly_seq.num
```

(*c*) STRUCT_REF_SEQ_DIF

```
• _struct_ref_seq_dif.align_id
        → _struct_ref_seq.align_id
• _struct_ref_seq_dif.seq_num
        → _entity_poly_seq.num
  _struct_ref_seq_dif.db_mon_id
        → _chem_comp.id
  _struct_ref_seq_dif.details
  _struct_ref_seq_dif.mon_id
        → _chem_comp.id
```

The bullet (•) indicates a category key. Where multiple items within a category are marked with a bullet, they must be taken together to form a compound key. The arrow (→) is a reference to a parent data item.

Data items in the STRUCT_REF category allow the author of an mmCIF to provide references to information in external databases that is relevant to the entities or biological units described in the mmCIF. For example, the database entry for a protein or nucleic acid sequence could be referenced and any differences between the sequence of the macromolecule whose structure is reported in the mmCIF and the sequence of the related entry in the external database can be recorded. Alternatively, references to external database entries can be used to record the relationship of the structure reported in the mmCIF to structures already reported in the literature, for example by referring to previously determined structures of the same or a similar protein, or to a small-molecule structure determination of a bound inhibitor or cofactor. STRUCT_REF data items are not intended to be used to reference a database entry for the structure in the mmCIF itself (this would be the role of data items in the DATABASE_2 category), but it would not be formally incorrect to do so.

When the data items in these categories are used to provide references to external database entries describing the sequence of a polymer, data items from all three categories could be used. The value of the data item **_struct_ref.seq_align** is used to indicate whether the correspondence between the sequence of the entity or biological unit in the mmCIF and the sequence in the related external database entry is `complete` or `partial`. If the value is `partial`, the region (or regions) of the alignment may be identified using data items in the STRUCT_REF_SEQ category. Comments on the alignment may be given in **_struct_ref_seq.details** (Example 3.6.7.13).

The value of the data item **_struct_ref.seq_dif** is used to indicate whether the two sequences contain point differences. If the value is `yes`, the differences may be identified and annotated using data items in the STRUCT_REF_SEQ_DIF category. Comments on specific point differences may be recorded in **_struct_ref_seq_dif.details**.

Example 3.6.7.13. *The relationship of the sequence of the protein PDB 5HVP to a sequence in an external database described with data items in the* STRUCT_REF *and* STRUCT_REF_SEQ *categories.*

```
loop_
  _struct_ref.id
  _struct_ref.biol_id
  _struct_ref.entity_id
  _struct_ref.db_name
  _struct_ref.db_code
  _struct_ref.seq_align
  _struct_ref.seq_dif
 seq_pdb     1  .  PDB       5HVP       .          .
 seq_genbank .  1  GenBank   AAG30358   complete   yes

loop_
  _struct_ref_seq.align_id
  _struct_ref_seq.ref_id
  _struct_ref_seq.seq_align_beg
  _struct_ref_seq.seq_align_end
  _struct_ref_seq.db_align_beg
  _struct_ref_seq.db_align_end
  _struct_ref_seq.details
    align_seq_pdb_genbank seq_genbank 1 99 24 122
; The genbank reference is to the sequence of
  residues 1-376 of the viral pol 1 polypeptide;
  the protease is proteolytically released from
  this precursor during viral maturation.
;
```