Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

*REFMAC*5 dictionary: organization of prior chemical knowledge and guidelines for its use

Alexei A. Vagin, Roberto A. Steiner,‡ Andrey A. Lebedev, Liz Potterton, Stuart McNicholas, Fei Long and Garib N. Murshudov*

Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5YW, England

‡ Current address: IFOM – The FIRC Institute of Molecular Oncology, Via Adamello 16, 20139 Milano, Italy

Correspondence e-mail: garib@ysbl.york.ac.uk

Received 19 April 2004
Accepted 22 September 2004

One of the most important aspects of macromolecular structure refinement is the use of prior chemical knowledge. Bond lengths, bond angles and other chemical properties are used in restrained refinement as subsidiary conditions. This contribution describes the organization and some aspects of the use of the flexible and human/machine-readable dictionary of prior chemical knowledge used by the maximum-likelihood macromolecular-refinement program *REFMAC*5. The dictionary stores information about monomers which represent the constitutive building blocks of biological macromolecules (amino acids, nucleic acids and saccharides) and about numerous organic/inorganic compounds commonly found in macromolecular crystallography. It also describes the modifications the building blocks undergo as a result of chemical reactions and the links required for polymer formation. More than 2000 monomer entries, 100 modification entries and 200 link entries are currently available. Algorithms and tools for updating and adding new entries to the dictionary have also been developed and are presented here. In many cases, the *REFMAC*5 dictionary allows entirely automatic generation of restraints within *REFMAC*5 refinement runs.

1. Introduction

Macromolecular crystal structure analysis can be regarded as an application of Bayesian statistics. This type of statistical analysis is centred on the Bayes' theorem, which can be formulated as

$$P(\mathbf{x}; \mathbf{F}) = P(\mathbf{x})L(\mathbf{x}; \mathbf{F})$$

where $P(\mathbf{x}; \mathbf{F})$ is the probability distribution of the model's parameters $\mathbf{x}$ given the experimental data $\mathbf{F}$, $P(\mathbf{x})$ is the prior probability distribution of $\mathbf{x}$ and $(\mathbf{x}; \mathbf{F})$ is the likelihood of $\mathbf{x}$ given $\mathbf{F}$. A perspective on macromolecular crystallography formulated from the Bayesian viewpoint can be found in Bricogne (1997).

An essential and distinctive component of the Bayesian treatment and analysis of experimental data is the notion of prior knowledge. $P(\mathbf{x})$ embeds, in real space, the most important prior information available for macromolecular crystal structure analysis. This type of information can be loosely divided into two families: (i) the available three-dimensional structures of macromolecules deposited within the Protein Data Bank and (ii) the relative invariance of elementary chemical properties such as bond lengths, bond angles, chiral volumes and planes.

A very important, although heavily underused, source of prior information for macromolecular experimental techniques is the PDB (Bernstein *et al.*, 1977; Berman *et al.*, 2002). It is very likely that many features of newly determined

structures are already present within the PDB. This aspect of the utilization of the available information is growing rapidly and there are already some applications of it in such branches of crystal structure analysis as model building (Jones *et al.*, 1991) and density modification (Terwilliger, 2003). In the future, a greater utilization of this type of information can be envisaged. Careful analysis and statistically sensible use of this information will definitely enhance and extend the applicability of the currently available experimental techniques for macromolecular-structure analysis (*e.g.* crystallography). This information can also be used in other branches of computational macromolecular biology, such as three-dimensional structure prediction and homology modelling.

The importance of using known chemical properties such as bond lengths and bond angles as subsidiary conditions in macromolecular crystallographic refinement has been recognized for a long time (Waser, 1963; Diamond, 1972; Jack & Levitt, 1978; Konnert & Hendrickson, 1980). The primary justification for the use of these properties is that the experimental data alone are not sufficient to completely define the three-dimensional structure of macromolecules. Therefore, in order to extract information from the experiment while retaining chemical integrity it is necessary to use some prior chemical information. This is true virtually at all resolutions. Even when data at atomic or subatomic resolution are available, some restraints are needed to rationalize regions of the molecule which are too disordered. Moreover, the availability of data at very high resolutions encourages the analysis of chemical properties, such as charge densities, which are not considered at lower resolution. This in turn increases the number of parameters to be refined. To keep the observations-to-parameters ratio to a reasonable value, the use of prior chemical knowledge is required.

The Bayesian framework provides a natural platform for the incorporation of prior knowledge for data analysis. The purpose of the present contribution is to describe the design, organization and some practical aspects of the use of the dictionary of prior chemical information used by the maximum-likelihood macromolecular crystallographic refinement program *REFMAC*5 (Murshudov *et al.*, 1997) from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994).

2. Definitions

Monomer. A monomer indicates a chemical unit which, at least formally, can exist independently. For example, amino acids, nucleotides, monosaccharides and ligands are monomers.

Modification. A modification is a formalism that describes the result of changes brought about on a monomer by a chemical reaction. Examples of modification are the N-terminal methylation of amino acids and the methylation of pyranoses at the O1 position.

Link. A link is a formalism that embeds the information required to describe all changes and newly formed bonds occurring when two monomers undergo a chemical reaction that somehow joins them together. Examples of links are *trans*/*cis* peptide bonds, phosphodiester bonds and α1–4 glycosidic bonds.

Chirality. Chirality is a chemical concept that refers to the property of certain compounds of being non-superimposable on their mirror image. The type of chirality used in the *REFMAC*5 dictionary is similar to that used in SMILES strings (Weininger, 1988); that is, local chirality as opposed to absolute chirality. Unlike the CIP (Cahn *et al.*, 1966) and IUPAC (1979) conventions for chirality, local chirality is defined only by the immediate surrounding of an atom. Local tetrahedral chirality is the most common. It is usually present on C and N atoms with sp^3 hybridization when at least three non-H atoms are bound to them. Local tetrahedral chirality is defined by its sign. The sign can be either 'positive' or 'negative'. More complex local chiralities are present at metal centres.

Minimal description. Minimal description refers to the minimal information necessary to describe a monomer uniquely. It consists of the monomer name, the list of its atoms' identification codes and symbols, its bonds list and orders and optionally the chemical group to which it belongs (peptide, pyranose *etc.*). If required, the configuration of the monomer can be defined using chiralities.

Complete description. The complete description is a monomer description that contains all the information about its internal chemical structure. In addition to the items present in the minimal description, it also contains a tree representation of the monomer as well as its bond lengths, bond angles and torsion angles. When required, planes and chiral centres are also defined. Standard deviations are given for appropriate parameters.

3. General considerations on the design and storage of prior chemical information

In general, two main approaches have been used by various programs to organize prior chemical information. The first approach uses the concept of chemical atom types. The second approach is based on larger monomer fragments.

In the atom-type approach chemical elements are assigned to different atom types depending on their chemical environment. For example, C atoms characterized by different degrees of *sp*-hybridization or aromatic C atoms surrounded by different neighbours constitute different atom types. The most popular atom types are those used by the *AMBER* (Allinger, 1977; Pearlman *et al.*, 1995) and *CHARMM* (Brooks *et al.*, 1983) programs. All possible bond distances between pairs of atom types as well as all possible angles defined by three atom types and torsion angles defined for four atom types are tabulated. These values are used to define the initial bond lengths, angles, torsion angles and other chemical parameters of a compound. All these values are generally refined using some semi-empirical energy function (see Ponder & Case, 2003; Brooks *et al.*, 1983). Usually, such an energy function requires well defined point charges on each atom type.

Although the atom-type approach has been and is being successfully used in crystallographic refinement [*X-PLOR* (Brünger, 1992) and *CNS* (Brünger *et al.*, 1998)], it is exposed to a few potential problems. Firstly, for general cases, the number of possible atom types is enormously large and the number of possible bonds and angles is so huge that it becomes impractical to store all required information. To deal with this problem, simplifications are used in many cases (Allinger, 1977). Secondly, atom types and their parameters have been carefully analysed only in a limited number of cases. When less usual atom types are encountered, tabulated values might not be directly transferable. An additional problem of the atom-type approach arises in the case of metal coordination. A metal atom can have very different coordination properties while preserving the same formal oxidation state. As a result, metric properties related to the same metal-atom type are extremely hard to handle in an automatic manner using atom types.

The second approach used to encode prior chemical knowledge, the monomer approach, is particularly suited to the case of biological macromolecules (proteins, DNA/RNA, polysaccharides). It explicitly uses the fact that that these compounds are made up of repeating units. These building blocks (monomers), which can formally exist as independent entities, form larger molecules by undergoing reactions that link monomers together. In general, the linkage of monomers partly changes their nature; for example, by introducing or removing atoms. It can also generate new bonds, angles, torsion angles, planes and chiralities. The monomer approach handles in a natural way various changes affecting the monomers themselves. Many proteins are modified after translation or during their activity. Binding of carbohydrates to asparagine residues or phosphorylation of serines are two such examples. These modifications alter the characteristics of monomers whilst largely preserving their intimate nature. They can be handled by describing the change brought about on monomers in terms of atoms added to and removed from the original monomer.

Pioneering refinement programs such as *PROLSQ* (Konnert & Hendrickson, 1980) and *NUCLSQ* (Westhof *et al.*, 1988) used the monomer approach. However, in both programs links were hard-coded. In addition, *PROLSQ* dealt only with polypeptides, whereas *NUCLSQ* dealt only with nucleic acids. Extension to other polymers such as sugars or mixtures of different polymers was almost impossible and required substantial modifications to the code.

The full exploitation of the advantages of the monomer approach requires dynamic definition of links and modifications; for example, by the use of code-independent external data files. It also requires the availability of accurate complete descriptions of monomers, links and modifications. Ideally, a package that uses this dictionary design should have tools to add new entries and to update old ones.

4. *REFMAC*5 dictionary

The dictionary used by the program *REFMAC*5 has been designed according to flexibility criteria. It is largely based on the monomer approach described in the previous section and allows dynamic definition of links and modifications. It contains carefully analysed descriptions for most common monomers, modifications and links. When necessary, owing to the impossibility of storing all possible information, prior chemical knowledge is managed semi-automatically using the atom-type approach.

The *REFMAC*5 dictionary is written in an extended mmCIF format (Bourne *et al.*, 1997). This is based on the STAR style (Hall, 1991) and the CIF format (Hall *et al.*, 1991) used in small-molecule crystallography. The attractive side of the mmCIF format is that any data file based on it can easily be extended without affecting the functionalities of programs already using it.

4.1. General organization and current state of the dictionary

The *REFMAC*5 dictionary contains a list of monomers, modifications and links along with their descriptions. Monomer descriptions define the stereochemical parameters of independent compounds. Modifications and links encapsulate the changes brought about on them by chemical reactions. Modifications typically act on a single monomer, whilst links join monomers together.

The currently distributed version of the dictionary has entries for all amino acids as well as for many of their possible modifications, for all nucleic acids and some of their modifications and for most common sugars and their modifications. It also has entries for many organic and inorganic compounds frequently encountered when solving macromolecular structures. As some monomers have several well established common names, the dictionary contains a list of synonyms capable of handling them. The dictionary also contains frequently encountered links such as *trans*/*cis* and methylated peptide links, sugar–sugar and sugar–protein links, as well as DNA/RNA links.

More than 2000 monomer entries, 100 modification entries and 200 link entries are currently available. Such a large dictionary covers most common users' needs. A full list of monomers, modifications and links available within the *REFMAC*5 dictionary can be found at the web page http://www.ysbl.york.ac.uk/~alexei/dictionary.html.

The dictionary can be extended easily by users. Users can create and organize personal monomer entries as well as modifications and links. In case of conflict, a user's definitions always override those stored within the distributed dictionary.

At present, the dictionary is used mainly by the program *REFMAC*5 for restrained refinement. However, its organization is such that it can easily be used by other programs dealing with macromolecules; for example, the model-building program *COOT* (Emsley & Cowtan, 2004). Applications for molecular simulation and modelling that use the *REFMAC*5 dictionary are currently being developed.

4.2. Monomers

For a monomer to be completely defined, information must be available about its constitutent atom(s) and, if present,

```
#
data_comp_list
loop_
_chem_comp.id
_chem_comp.three_letter_code
_chem_comp.name
_chem_comp.group
_chem_comp.number_atoms_all
_chem_comp.number_atoms_nh
_chem_comp.desc_level
GLC-b-D  GLC 'beta_D_glucose       ' D-pyranose         24  12 .
#
data_comp_GLC-b-D
#
loop_
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.type_energy
_chem_comp_atom.partial_charge
 GLC-b-D       C1     C    CH1       0.000
 GLC-b-D       H1     H    HCH1      0.000
 ...
 ...
 GLC-b-D       HO6    H    HOH1      0.000
 GLC-b-D       O5     O    O2        0.000
loop_
_chem_comp_tree.comp_id
_chem_comp_tree.atom_id
_chem_comp_tree.atom_back
_chem_comp_tree.atom_forward
_chem_comp_tree.connect_type
 GLC-b-D  C1     n/a    C2     START
 GLC-b-D  H1     C1     .      .
 ...
 ...
 GLC-b-D  O6     C6     HO6    .
 GLC-b-D  HO6    O6     .      .
 GLC-b-D  O5     C5     .      END
 GLC-b-D  O5     C1     .    ADD
loop_
_chem_comp_bond.comp_id
_chem_comp_bond.atom_id_1
_chem_comp_bond.atom_id_2
_chem_comp_bond.type
_chem_comp_bond.value_dist
_chem_comp_bond.value_dist_esd
 GLC-b-D  O1     C1        single       1.410    0.020
 GLC-b-D  C2     C1        single       1.524    0.020
 ...
 ...
 GLC-b-D  HO6    O6        single       0.980    0.020
 GLC-b-D  C1     O5        single       1.410    0.020
loop_
_chem_comp_angle.comp_id
_chem_comp_angle.atom_id_1
_chem_comp_angle.atom_id_2
_chem_comp_angle.atom_id_3
_chem_comp_angle.value_angle
_chem_comp_angle.value_angle_esd
 GLC-b-D  H1     C1     O1      109.470    3.000
 GLC-b-D  O1     C1     C2      109.470    3.000
 ...
 ...
 GLC-b-D  C6     O6     HO6     109.470    3.000
 GLC-b-D  C5     O5     C1      111.800    3.000
loop_
_chem_comp_tor.comp_id
_chem_comp_tor.id
_chem_comp_tor.atom_id_1
_chem_comp_tor.atom_id_2
_chem_comp_tor.atom_id_3
_chem_comp_tor.atom_id_4
_chem_comp_tor.value_angle
_chem_comp_tor.value_angle_esd
_chem_comp_tor.period
 GLC-b-D  var_1    C1     C2     O2     HO2        0.000   20.000   1
 GLC-b-D  var_2    C1     C2     C3     C4       -50.095   20.000   3
 ...
 ...
 GLC-b-D  var_11   C5     O5     C1     C2       -55.889   20.000   3
 GLC-b-D  var_12   O5     C1     C2     C3        55.889   20.000   3
loop_
_chem_comp_chir.comp_id
_chem_comp_chir.id
_chem_comp_chir.atom_id_centre
_chem_comp_chir.atom_id_1
_chem_comp_chir.atom_id_2
_chem_comp_chir.atom_id_3
_chem_comp_chir.volume_sign
 GLC-b-D  chir_01  C5     C4     O5     C6        positiv
 GLC-b-D  chir_02  C4     C3     O4     C5        positiv
 GLC-b-D  chir_03  C3     C2     O3     C4        negativ
 GLC-b-D  chir_04  C2     C1     O2     C3        positiv
 GLC-b-D  chir_05  C1     O1     O5     C2        positiv
```

about its bonds, angles, torsion angles, planes and chiral centres. An example of a complete monomer description is given in Fig. 1.

Monomers are described by the following categories.

(i) General category. This category contains the short and full monomer names, the monomer three-letter PDB code and the group to which the monomer belongs to (peptide, DNA/RNA, pyranose, non-polymer). Group names are an important part of the monomer description, as they facilitate monomer handling. For example, if the monomer belongs to the group called 'peptide', then all links and modifications described for peptides can be applied to it. Moreover, the group type defines whether a monomer can belong to a chain (polypeptide, DNA/RNA or polysaccharide chains).

(ii) Atom category. This category lists atom and element names and their chemical types and charges. It may also contain Cartesian coordinates.

(iii) Tree category. This category describes the mathematical tree (acyclic graph) corresponding to the monomer chemical connectivity. It is used to generate coordinates. Missing atoms, *e.g.* H atoms, are restored using this tree.

(iv) Bond category. This category contains the list of bonded atoms, bond types and the ideal values of bond lengths and uncertainties associated with them. Alongside the atom category, this category defines completely the chemical structure of the monomer. In mathematical terms, such a structure is called a coloured graph. Edges are coloured by bond orders and vertices are coloured by chemical types.

(v) Angle category. This category contains the three-atom list of all possible angles in the monomer as well as their ideal values and associated uncertainties.

(vi) Torsion-angle category. This category contains the four-atom list of torsion angles, their types and names, their ideal values and associated uncertainties and their period. The latter value represents the number of energetic minima along the torsion angle. For example, χ angles along the $C^{\alpha}-C^{\beta}$ bond of glutamine residue have a period equal to three. A

Figure 1
Example of complete monomer description. This example shows the complete monomer description of the pyranose β-D-glucose. For compactness, most description categories given contain only a representative set of items. Missing items are represented by '. . . ' symbols. The first category (`_chem_comp`) is the general category. It contains the name of the monomer together with its long name and the name of the group to which it belongs to. In this category there is an indication of the level of monomer description. If this item has the value 'M' the entry has a minimal description. In this case the value is '.', which indicates a complete description. The second category (`_chem_comp_atom`) describes atoms with their names, element names, atom types and atom charges. It can also contain a monomer representation in Cartesian coordinates. The third category (`_chem_comp_tree`) is the acyclic graph description. Additional bonds are also present to indicate ring enclosure. These bonds have the label 'ADD'. The beginning and end of the tree are labelled 'START' and 'END', respectively. The fourth category (`_chem_comp_bond`) lists bonds together with their bond lengths, bond orders and uncertainties. Other categories present are for bond angles (`_chem_comp_angle`), torsion angles (`_chem_comp_tor`) and chiralities (`_chem_comp_chir`). When required, planes are indicated by the category (`_chem_comp_plane_atom`). The latter category is not present in this example as β-D-glucose does not need planarity restraints.

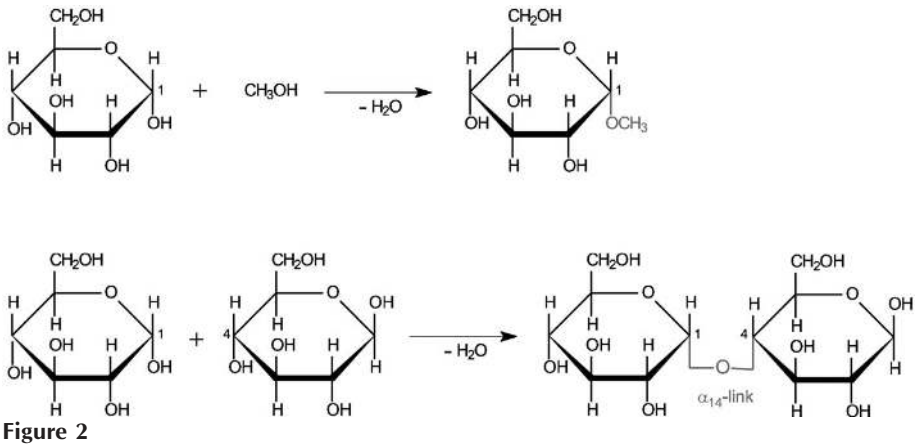


Figure 2
(*a*) Example of a sugar modification. The condensation of α-D-glucose with methanol gives methyl-α-D-glucoside. (*b*) Example of a sugar link. The disaccharide β-maltose is formed by condensation of α-D-glucose with β-D-glucose. The glycosidic bond is an α1–4 link.

```
data_mod_list
loop_
_chem_mod.id
_chem_mod.name
_chem_mod.comp_id
_chem_mod.group_id
O1MET  O1_metyl_of_sugar   .  pyranose

data_mod_O1MET

loop_
_chem_mod_atom.mod_id
_chem_mod_atom.function
_chem_mod_atom.atom_id
_chem_mod_atom.new_atom_id
_chem_mod_atom.new_type_symbol
_chem_mod_atom.new_type_energy
_chem_mod_atom.new_partial_charge
 O1MET    change   O1     .      .    O2      0.000
 O1MET    delete   HO1    .      .    .       0.000
 O1MET    add      .      HM3    H    HCH     0.000

loop_
_chem_mod_bond.mod_id
_chem_mod_bond.function
_chem_mod_bond.atom_id_1
_chem_mod_bond.atom_id_2
_chem_mod_bond.new_type
_chem_mod_bond.new_value_dist
_chem_mod_bond.new_value_dist_esd
 O1MET    add      O1      CM       single      1.420    0.020

 O1MET    add      CM      HM3      single      0.960    0.020
loop_
_chem_mod_angle.mod_id
_chem_mod_angle.function
_chem_mod_angle.atom_id_1
_chem_mod_angle.atom_id_2
_chem_mod_angle.atom_id_3
_chem_mod_angle.new_value_angle
_chem_mod_angle.new_value_angle_esd
 O1MET    add      C1      O1      CM      120.000    3.000
```

Figure 3
Example of a modification. This example describes methylation at the O1 position of pyranoses. See Fig. 2(*a*) for a graphical representation of this modification. The first (general) category (`_chem_mod`) reports the code for the modification 'O1MET' and describes whether the modification is to be applied to only a particular monomer or to a group of monomers. The 'O1MET' modification can be applied to all monomers belonging to the 'pyranose' group. The (`_chem_mod_atom`) category describes the list of all added, deleted or changed atoms. The following category (`_chem_mod_bond`) describes all added or deleted bonds. In a similar manner, the tree structure (`_chem_mod_tree`), bond angles (`_chem_mod_angle`), torsion angles (`_chem_mod_tor`), planes (`_chem_mod_plane_atom`) and chiralities (`_chem_mod_chir`) affected by the modification are handled.

torsion angle can be constant or variable. Constant torsion angles generally involve atoms belonging to the same plane or atoms along double bonds. Usually, these torsion angles have period equal to zero or one as they can only have a single value.

(vii) Plane category. This category contains the list of planes and of all atoms belonging to them.

(viii) Chirality category. This category contains the list of all chiral centres. For each chiral centre it also lists the central atom, the atoms bonded to it and the sign of the chiral volume. The current version of the dictionary allows undefined signs using 'both' or 'anomer' keywords. If the keyword 'both' is used the chirality of the monomer can change during restrained refinement. If the keyword 'anomer' is used the chirality is fixed and its sign is defined by the input coordinates. If for a monomer in a crystal there are two or more configurations, all of them can be handled simultaneously during refinement by assigning the keyword 'anomer' to each chiral centre.

Ideal values for bond lengths and bond angles for standard amino acids present in the dictionary have been taken from Engh & Huber (1991). Ideal values for bond lengths and angles for nucleic acids have been taken from Kennard & Taylor (1982). Ideal values for bond lengths and angles for most saccharides have been taken from Saenger (1983).

At present, about 1000 monomers of the 2000 available in the *REFMAC*5 dictionary are present with a complete description. The remaining monomers are present with a minimal description. Work is in progress to deliver in the shortest time possible a dictionary in which all entries are present with checked complete descriptions.

4.3. Modifications

A modification is a formalism which describes changes brought about on a single monomer by chemical reactions. An example of modification is shown in Fig. 2(*a*). Its dictionary description is given in Fig. 3. A modification allows atoms, bonds, angles, torsion angles, planes and chiral centres to be added to or deleted from monomers. The use of modifications greatly reduces the number of monomer descriptions that need to be stored and allows proper description of links between monomers, as some of them require monomers to first undergo modifications prior to linkage. Modifications can also be used for non-chemical changes on monomers such as changes in residue name. This is a convenient way of handling cases of multiple monomer names. In such cases the modification keyword is 'RENAME'. This keyword is also used to overcome the three-letter restriction imposed by the PDB convention.

4.4. Links

The link formalism allows the joining of monomers together. An example of a link is shown in Fig. 2(*b*). Its description is given in Fig. 4. Links can be considered to be the external counterpart of monomer descriptions. Whereas monomer descriptions give the internal structure of single chemical compounds, link descriptions define in detail the result of chemical reactions between monomers. Link descriptions contain information about the monomers or the group of monomers they act on as well as about the modifications these monomers should undergo prior to linkage. In the current version of the dictionary, a link can form only one bond. However, the introduction of several angles, torsion angles, planes and chiral centres is allowed.

```
#
data_link_list
loop_
_chem_link.id
_chem_link.comp_id_1
_chem_link.mod_id_1
_chem_link.group_comp_1
_chem_link.comp_id_2
_chem_link.mod_id_2
_chem_link.group_comp_2
_chem_link.name
ALPHA1-4 .    DEL-HO4  pyranose .    DEL-O1   pyranose glycosidic_bond_alpha1-4

data_link_ALPHA1-4
#
loop_
_chem_link_bond.link_id
_chem_link_bond.atom_1_comp_id
_chem_link_bond.atom_id_1
_chem_link_bond.atom_2_comp_id
_chem_link_bond.atom_id_2
_chem_link_bond.type
_chem_link_bond.value_dist
_chem_link_bond.value_dist_esd
 ALPHA1-4 1 O4      2 C1        single      1.439    0.020
loop_
_chem_link_angle.link_id
_chem_link_angle.atom_1_comp_id
_chem_link_angle.atom_id_1
_chem_link_angle.atom_2_comp_id
_chem_link_angle.atom_id_2
_chem_link_angle.atom_3_comp_id
_chem_link_angle.atom_id_3
_chem_link_angle.value_angle
_chem_link_angle.value_angle_esd
 ALPHA1-4 1 C4      1 O4      2 C1      108.700    3.000
 ALPHA1-4 1 O4      2 C1      2 O5      112.300    3.000
 ALPHA1-4 1 O4      2 C1      2 C2      109.470    3.000
 ALPHA1-4 1 O4      2 C1      2 H1      109.470    3.000
loop_
_chem_link_tor.link_id
_chem_link_tor.id
_chem_link_tor.atom_1_comp_id
_chem_link_tor.atom_id_1
_chem_link_tor.atom_2_comp_id
_chem_link_tor.atom_id_2
_chem_link_tor.atom_3_comp_id
_chem_link_tor.atom_id_3
_chem_link_tor.atom_4_comp_id
_chem_link_tor.atom_id_4
_chem_link_tor.value_angle
_chem_link_tor.value_angle_esd
_chem_link_tor.period
 ALPHA1-4 ALPHA_1  1 O4     2 C1     2 C2     2 C3        0.00  20.0 1
 ALPHA1-4 ALPHA_2  1 C4     1 O4     2 C1     2 C2        0.00  20.0 1
 ALPHA1-4 ALPHA_3  1 C3     1 C4     1 O4     2 C1        0.00  20.0 1
loop_
_chem_link_chir.link_id
_chem_link_chir.atom_centre_comp_id
_chem_link_chir.atom_id_centre
_chem_link_chir.atom_1_comp_id
_chem_link_chir.atom_id_1
_chem_link_chir.atom_2_comp_id
_chem_link_chir.atom_id_2
_chem_link_chir.atom_3_comp_id
_chem_link_chir.atom_id_3
_chem_link_chir.volume_sign
 ALPHA1-4   2 C1      1 O4      2 O5      2 C2      negativ
```

Figure 4
Example of a link. This example describes the α1–4 pyranose link. See Fig. 2(*b*) for a graphical representation of this link. The first (general) category (`_chem_link`) describes the name, the link identification code and the scope of this link. It also contains pointers to the modifications the monomers should undergo before the link can be applied. This link requires that the monomers belong to the 'pyranose' group and that the first and second monomers undergo DEL-HO4 and DEL-O1 modifications, respectively. The description for both these modification need to be available before the link can be applied. Other categories give the list of bonds (`_chem_link_bond`), bond angles (`_chem_link_angle`), torsion angles (`_chem_link_tor`), chiralities (`_chem_link_chir`) *etc*. with their 'ideal' values. Atom names in the link description are always given together with the monomer numbers they belong to.

4.5. Atom-type library

Although the *REFMAC*5 dictionary is largely based on monomers, it also contains an atom-type library. At present, it contains about 200 atom types. It includes all chemical elements as well as many atom types commonly encountered in chemistry. Each entry has information about the chemical element the atom type belongs to as well as about its van der Waals (VDW) and ionic radii. The atom-type library also contains information about possible bonds between atom types. For many pairs of atom types, bond orders and bond lengths are tabulated. Angles corresponding to some of the atom-type triplets are also listed. The atom-type library is in mmCIF format. Therefore, it can easily be updated and extended. A full list of all atom-type library entries can be found at the web page http://www.ysbl.york.ac.uk/~alexei/dictionary.html.

The bond lengths listed in the atom-type library have been taken from the *International Tables for Crystallography* (Allen *et al.*, 1992; Orpen *et al.*, 1992). VDW and ionic radii of atoms have been taken from various sources including Greenwood & Earnshaw (1989) and Cotton & Wilkinson (1972). Unfortunately, to our knowledge there is no single general reference for bond angles. Some of the angles have been taken from examples from the Cambridge Structural Database (Allen, 2002), while others have been derived using general information about atoms, *i.e.* their hybridization and the nature of the surrounding atoms.

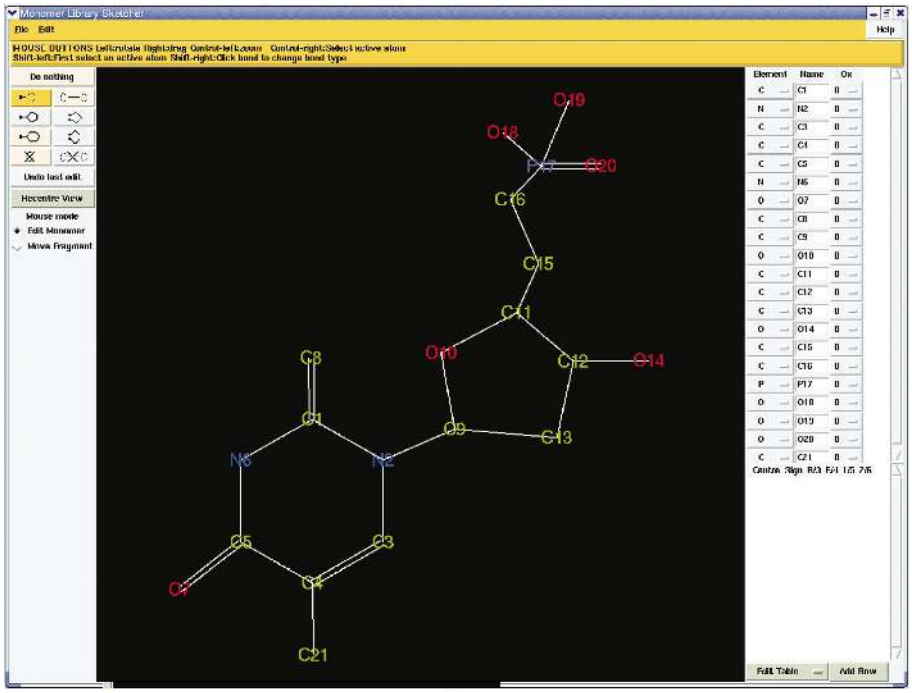
The atom-type library serves two main purposes: (i) it provides information about VDW and ionic radii as well as about atoms' hydrogen-bonding capability that is used to define non-bonding interactions in the course of refinement and (ii) it provides information about initial bond lengths and angles when new monomer entries are created.

5. New monomers

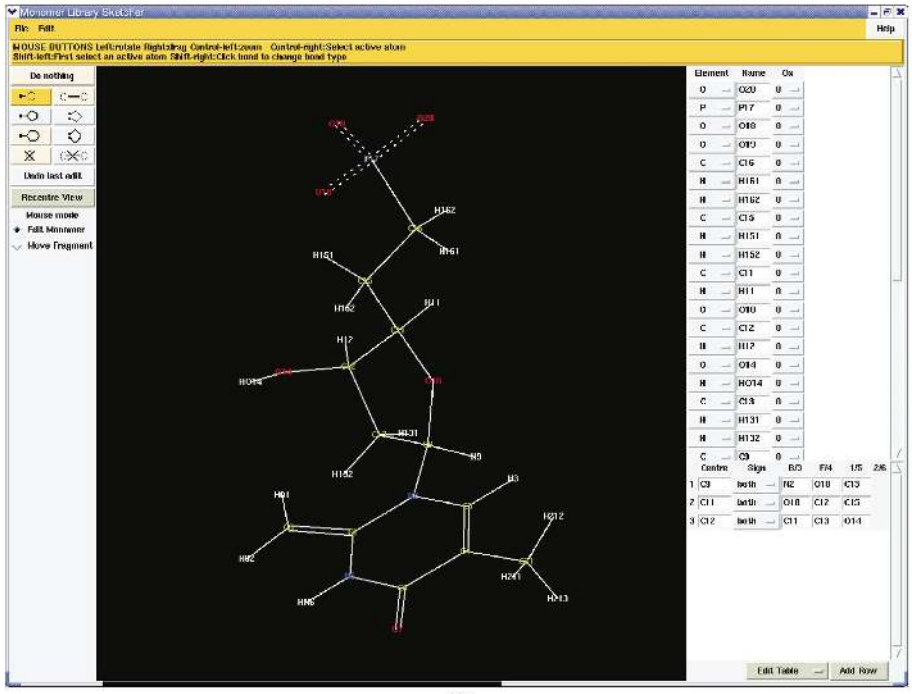
Generation of restraints for crystallographic refinement requires that all monomers present in the file which represent the structure to be refined are completely described.

In addition to the *REFMAC*5 dictionary, we have developed tools that allow the generation of complete monomer descriptions when these are not available. This is typically the case when new monomers, *i.e.* monomers not present in the dictionary, or when dictionary entries stored only with a minimal descriptions are encountered. The most usual new monomers are ligands that are not commonly found in macromolecular crystallographic analysis, which are of interest principally in the context of particular crystallographic projects. The program *LIBCHECK* is part of the *CCP*4

package and can be used to create complete monomer dictionary entries. *LIBCHECK* is best used through its graphical front-end *SKETCHER* (Potterton *et al.*, 2003), which is part of the *CCP*4*i* interface.



(*a*)



(*b*)

Figure 5
SKETCHER enables the user to draw a compound to specify all the information required by *LIBCHECK* to generate an complete structure description. *SKETCHER* can also invoke *REFMAC*5 to generate an idealized structure. (*a*) shows a finished basically two-dimensional sketch before running *LIBCHECK* and (*b*) shows the resultant structure idealized with *REFMAC*5. To create the sketch, the user can choose from the tool panel on the left to add a single atom or ring structure or to delete atoms or bonds. Appropriate mouse clicks in the main viewing window will select the active atom to which new fragments are bonded and place the new fragment or edit the bond type. New atoms are always initially carbon, but the element type and atom name can be edited in the table in the top right of the window. The atom chirality is listed and can be edited in the table in the bottom right of the window.

A complete monomer description can be generated once a minimal description is available. The algorithm used in this procedure is described in §5.3. Users have two main ways to create minimal descriptions for new monomers: (i) from their chemical structure and (ii) from their Cartesian coordinates.

5.1. Monomer description from its chemical structure

If the only reliable information available about a new monomer is its chemical structure, the best way to create its minimal description is with the aid of the program *SKETCHER*. The monomer is simply drawn specifying its constituting atoms and bonds (Fig. 5*a*). The information represented graphically is essentially the minimal monomer description. Using this information, *LIBCHECK* creates the complete monomer description using the algorithm described in §5.3. If desired, *REFMAC*5 can be invoked to idealize the structure. After the optimization, all information returns to *SKETCHER* and is displayed (Fig. 5*b*). The user can therefore check whether the desired description has been produced.

Although the primary purpose of *LIBCHECK* is to generate dictionary entries from which restraints for refinement can be created, this program can also be utilized to generate monomer Cartesian coordinates. This is particularly useful when a set of initial coordinates is required for model building. The program *LIBCHECK* can be invoked from the graphical interface *SKETCHER* to generate monomer Cartesian coordinates from the complete monomer description. As this description contains all necessary information, it is perfectly suited for coordinate generation. To this end, the *Z*-matrix representation of the monomer is used. This representation is commonly encountered in computational chemistry (Leach, 1997) and is closely related to the internal representation of monomers.

In general, the Z matrix contains $3N - 6$ parameters, where N is the number of atoms present in the monomer. This matrix does not define the orientation or rotation of the monomer. The first atom of the monomer is arbitrarily placed at the origin of the Cartesian coordinates system. For the second atom, only the bond length with respect to the first atom is stored. This atom can therefore be given any triplet of coordinates that satisfies the known bond length. For the third atom, the bond length with respect to the second atom and the bond angle formed with the first two atoms are given. Any position that satisfies these two parameters can be given to the third atom. Starting from the fourth atom, all atoms are defined by bond lengths, bond angles and torsion angles.

5.2. Monomer description from its Cartesian coordinates

If reliable Cartesian coordinates are available for a monomer, its minimal description can be extracted from them. To this end, the monomer-connectivity graph is first derived from atom coordinates. Two atoms are considered bonded if their distance is shorter than a certain threshold which is element-dependent. If atoms belong to the group (C, N, O, S, P, B, I, Cl) they are considered bonded if their distance is within 1.8 Å. If one atom belongs to the above list and the other is a heavy atom they are considered bonded if their distance is shorter than 2.3 Å. If atoms are heavy atoms they are considered bonded if their distance is within 2.8 Å. H atoms are considered bonded to any atom if their distance is shorter than 1.2 Å.

Bond orders and an approximation to atom types are then defined iteratively. The procedure is schematically as follows.

(i) Single bonds are tentatively defined using chemical knowledge criteria. For example, if a C atom is connected to four atoms all its bonds are classified as 'single' and the atom is considered sp^3-hybridized. Similar chemical considerations apply to other elements.

(ii) If all bond orders cannot be unambiguously assigned in step (i) the atom-type library is searched. All bonds involving atom pairs with element names matching those involved in the bond for which the bond order cannot be assigned are scanned. The library bonds whose bond lengths best compare with those with unknown bond orders are assigned.

(iii) Corrections are made for 'single' bonds adjacent to 'triple' bonds. 'Single' bonds adjacent to 'triple' bonds are usually shorter than other 'single' bonds. In order to avoid misclassification, corrections are made for this. If one of the atoms is either an *sp*-hybridized C or N atom the shortest bond is assumed to be 'triple' and the other bond is considered to be 'single'.

(iv) If an atom is a C atom and three bonds depart from it, if two of them are single and the third one is either 'delocalized' or 'aromatic' then the latter becomes 'double'.

Once all bond orders have been defined, a complete monomer description can be created according to the algorithm presented in the following section. It is important to stress that monomer descriptions should be derived from Cartesian coordinates only when these are of high quality. Small errors in bond lengths affect the derived bond orders and thus the final description of the monomer. Reliable coordinates for small molecules can be obtained from the Cambridge Structural Database (CSD; Allen, 2002) if

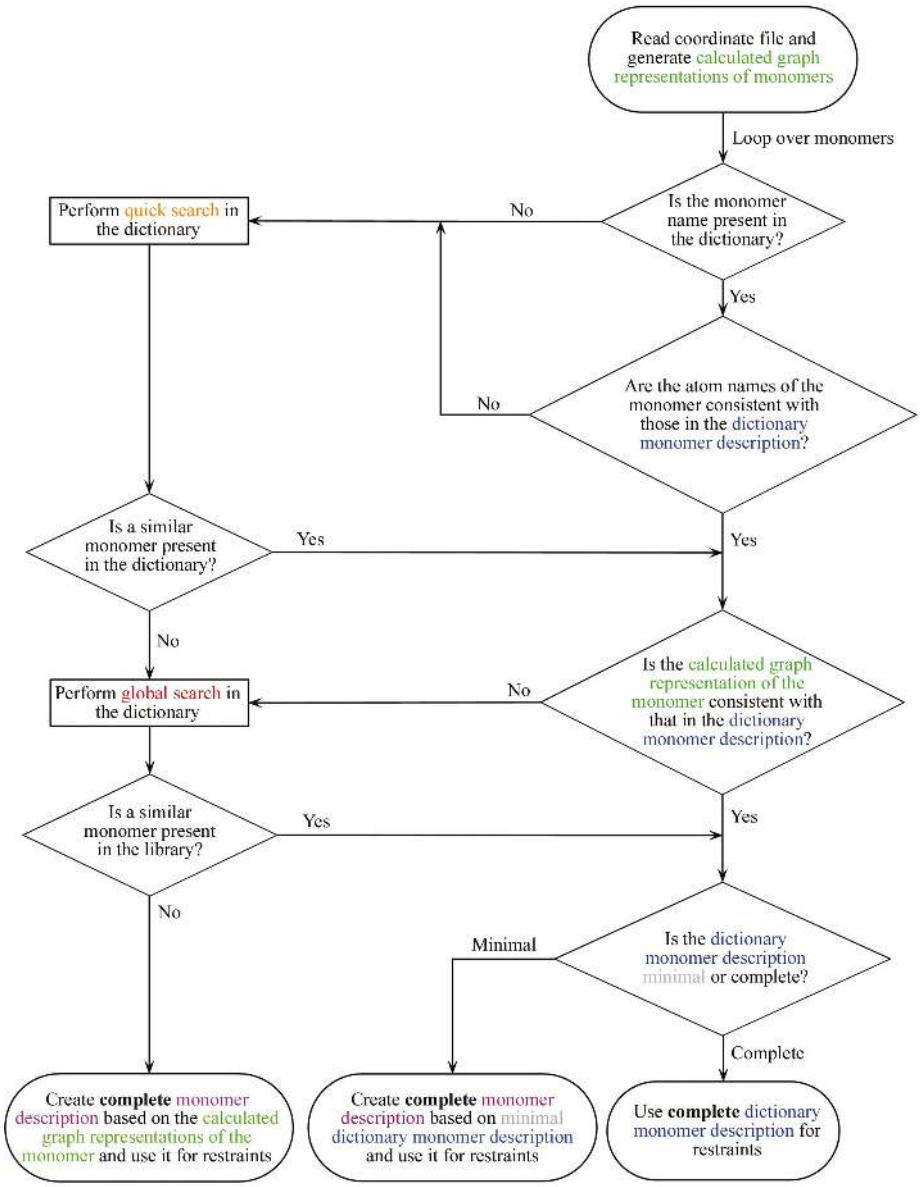


Figure 6
Flowchart representing the process of monomer recognition. See §6.1 for details.

their crystallographic structures have been determined. Alternatively, if only approximate coordinates are available these can be improved using molecular-mechanics or quantum-chemical geometry-optimization tools available within various packages.

5.3. Complete monomer description from minimal monomer description

Once a minimal monomer description is at hand, it can be used to derive the complete description. In brief, the algorithm used for this purpose is the following.

(i) Tentative chiralities of atoms are defined. If an atom has more than three bonds it is considered to be a potential chiral centre.

(ii) Rings are defined. Atoms belonging to a ring are flagged.

(iii) Chemical assumptions about O atoms are imposed. If two O atoms are bound to a C atom and the latter has three or fewer bonds whilst the O atoms have only one bond, then carbon–oxygen bonds are considered to be 'delocalized' bonds and O atoms are assumed to be negatively charged. For each oxygen, a formal charge of −0.5 is assigned. Similar assumptions are used for oxygen–phosphorus and oxygen–sulfur bonds.

(iv) Missing H atoms are added according to chemical valences, charges and bond orders. The number of missing H atoms is calculated using the formula H atoms = valence − total bond order + charge. 'Aromatic' and 'delocalized' bonds are considered to have a bond order of 1.5. If an element can have multiple valences the lowest valence compatible with the number of bonds is assumed. For example, if there is only one bond on an S atom then its valence is two; if there are three or four bonds then its valence is four. If more bonds are present, the valence is six.

(v) Atom types are refined. At this stage information such as the number of bonds, bond orders and number of H atoms present in the structure is known. All these properties are used to define atom types.

(vi) Chiralities are checked and added or removed if necessary. If an atom has either sp^2- or *sp*-hybridization it cannot be a chiral centre. If an atom has four bonds it can be a chiral centre. If there are chiralities defined in the the minimal description and they do not conflict with the atom degree of *sp*-hybridization their chirality signs are taken from it. If Cartesian coordinates are also available chirality signs are calculated from them. This ensures that the configuration of the monomer is consistent with that of the minimal description and/or coordinates.

(vii) Planes are defined and joined if necessary. Atoms with sp^2-hybridization are assumed to belong to a plane. If all atoms of a ring are sp^2-hybridized the ring is considered to be a flat ring and planes are joined together.

(viii) Planes are extended to add adjacent atoms. If an atom belongs to a plane, all its neighbours are also considered to belong to it.

(ix) A tree description of the monomer is created and the corresponding torsion angles are assigned.

(x) The complete description is created. Once atom types and bond orders for all bonds have been defined, bond lengths and bond angles are taken from the atom-type library. If accurate Cartesian coordinates are available target values can also be extracted from them.

5.4. Web resources

Two web resources can produce complete monomer descriptions compatible with *REFMAC*5. The first resource is hosted at the European Bioinformatics Institute (Golovin *et al.*, 2004) and can be accessed at the web address http://www.ebi.ac.uk/msd-srv/chempdb/cgi-bin/cgi.pl. The second is the program *PRODRG* (van Aalten *et al.*, 1996), which can be found at the web address http://davapc1.bioch.dundee.ac.uk/programs/prodrg/prodrg.html. Other programs, such as the *CORINA* suite (Sadowski *et al.*, 1994) available from http://www2.chemie.uni-erlangen.de/software/corina/index.html, can also give coordinates that can be used to create *REFMAC*5 monomer descriptions. *CORINA* can be used with help of the *CACTVS* (Ihlenfeldt *et al.*, 1994) interface.

6. Restraint generation for crystallographic refinement

Generation of restraints for the purpose of crystallographic refinement is an entirely automatic process conducted within *REFMAC*5 refinement runs when all monomers listed in a coordinate file are present in the dictionary with a complete description. This is the case when protein and nucleic acid structures as well as their complexes are refined. It is also the case when common sugars and organic/inorganic compounds are present in the structure. How to deal with ligands for which no dictionary entries are available and how to create complete descriptions for them has been described in §5

For restraints to be applied in refinement, the various monomers present in a coordinate file (typically in PDB format) need first to be recognized in the dictionary so that information can be taken from their complete descriptions; if minimal descriptions are available for some of them, complete descriptions can be created as described in §5.3.

6.1. Monomer recognition

A flowchart representing how *REFMAC*5 deals with monomer recognition is given in Fig. 6.

Atom coordinates corresponding to the various monomers present in the file are used to create a calculated graphical representation of the monomers. These monomers are matched against those present in the dictionary. Matching is carried out at the level of monomer and atom names and on the basis of connectivity. If some atoms are missing in the input file, the current implementation assumes an inaccuracy in the input file. These atoms are simply flagged as missing atoms. If requested they can be restored.

Failure in monomer matching activates a quick search of the whole dictionary. A sketchy description of the algorithm used in this search is as follows.

(i) The atoms present in the file monomer are counted and their types listed. When a monomer with the same property is found in the dictionary (active dictionary monomer) the search moves to step (ii).

(ii) The atom pairs in the file monomer separated by one up to seven bonds are evaluated. If the current active dictionary monomer is characterized by the same values the search moves to step (iii).

(iii) The number of C—C, C—O, C—N, N—O, N—N and O—O bonds in the file monomer are evaluated. If the current active dictionary monomer is characterized by the same values, the search moves to step (iv).

(iv) A graph-matching algorithm checks the isomorphism between the graph of the file monomer and that of the active dictionary monomer. If the graphs are isomorphous on the basis of atom correspondence, the file monomer is assigned.

In the current version of the algorithm the properties for the file monomer described above must be mirrored exactly by those of the dictionary monomer. In principle, softer criteria could also be used. This option is not yet available.

In case no dictionary monomer can be matched with the quick-search procedure, a more expensive global search can be performed. This search uses the Ullman's graph-matching algorithm (Ullman, 1976), which allows exact subgraph matching. If even the global search is unable to find a matching monomer in the dictionary the file monomer is treated as a 'new monomer' and its complete description is generated from its coordinates as described in §5.2.

```
0        1         2         3         4         5         6         7
12345678901234567890123456789012345678901234567890123456789012345678901234567890
LINK         SG  ACYS A 121                 SG  BCYS A 222                 SS
SSBOND   1 CYS A  298    CYS A  298                          4555
LINK         NE2  HIS A  71        2.055   ZN    ZN R   5                 HIS-ZN
LINK         C6  BBEN B   1                 O1  BMAF S   2                 BEN-MAF
LINK         OE2 AGLU A 320                 C1  AMAF S   2                 GLU-MAF
MODRES       MAN S    3  MAN-b-D                                          RENAME
MODRES       MAN S    4  MAN-b-D                                          RENAME
LINK              MAF S   2                    MAN S   3                 BETA1-4
LINK              MAN S   3                    MAN S   4                 BETA1-4
```

Figure 7
Example of a PDB header with links and modifications. *REFMAC*5 is able to read the header of a PDB coordinate file and to process the information needed for restraint generation. In this example lines 1–2 give column numbers which define the format of the various links and modifications. Note that links and modifications must strictly obey the format shown here. Lines 3–4 are examples of 'S' links. Line 5 is an example of metal–ligand link in which the target value is defined in the PDB file. Lines 6–11 list the links required to refine the structure given in Fig. 8. The dictionary file that describes the 'BEN-MAF' and 'GLU-MAF' links as well as the monomers is given as supplementary material. A more detailed description of the lines in this example is as follows. Lines 1–2, column numbers in the file; line 3, 'S' link between alternative conformations. The SG atom of Cys121 belonging to chain A (alternative conformation A) is bound to the SG atom of Cys222 of the same chain (alternative conformation B). Line 4, 'S' link between symmetry-related Cys residues (Cys298 of chain A). The 'SSBOND' record has the same meaning as in the PDB convention. '4555' is a symmetry indicator. When defining a link between symmetry-related residues, the symmetry operator does not have to be the correct one relating the two residues. It is simply a signal for the program to find the actual symmetry for the link. Line 5, 'HIS-ZN' link between the NE2 atom of His71 belonging to chain A and the Zn5 atom in chain R. The target distance is 2.055 Å. Line 6, 'BEN-MAF' link between the B conformation of the monomer BEN1 (atom C6, chain B) and the B conformation of the monomer MAF2 (O1, S). The 'BEN-MAF' link description is not amongst those defined in the dictionary. It needs to be defined by the user. This is best performed using an existing link as a template. Line 7, 'GLU-MAF' link between the A conformation of Glu320 (OE2, A) and A conformation of MAF2 (C1, S). Similarly to the 'BEN-MAF' link, the 'GLU-MAF' link needs to be defined by the user. Lines 8–9, 'RENAME' modifications. These explicitly state that residues MAN3 and MAN4 in chain S are β-D-mannose monomers. Lines 10–11, 'BETA1-4' links between MAF2 and MAN3 in chain S and between MAN2 and MAN4 in the same chain.

6.2. Monomer linkage

Monomers in the file are assigned to different chains depending on their group types (peptide, DNA/RNA, pyranose, non-polymer) and other criteria. In the case of polypeptides and DNA/RNA group types, chain-identification codes, consecutive residue numbering and optionally distance criteria are employed. If the latter criteria is used then consecutive monomers in the coordinate file are considered to be linked if their distance is shorter than a predefined value. For example, in the case of peptide monomers, monomers are considered to be linked if the distance between C_i and N_{i+1} is less than 1.728 Å. Distance criteria are useful when dealing with insertions and deletions. If none of the criteria above are fulfilled, then a 'gap' link is created. For saccharide-group types, distance and same chain-identification code criteria are used. Non-polymers cannot form chains.

Once chains have been identified, the program applies standard links according to the chain's group types. For example, for peptide chains *trans*/*cis*-peptide links are defined according to the ω angle. If the angle is between 80 and 280° the *trans* conformation is assumed. Similarly, standard links for polysaccharides and DNA/RNA are also defined.

Once chains have been defined, standard links such as disulfide bridges and sugar–protein links are automatically activated on the basis of distance criteria. If user-defined links are found in the header of the coordinate file, these are used. Distance criteria are also used to recognize potential links that are not part of the dictionary. For example, if two atoms are found to be very close, a potential link between them is proposed. By default, links determined in this way are not used unless explicitly requested.

After all standard links and modifications have been defined, other links, such as links between alternative conformations and symmetry-related atoms, are dealt with. By default, monomers in alternative conformation are assigned the same links. For example, if two cysteine residues (Cys1 and Cys2) are present in double conformation (A and B), by default both will be considered involved in SS links even if only one pair is engaged in a disulfide bridge according to distance criteria. 'S' links will be formed between the SG atoms of ACys1 and ACys2 and between the SG atoms of BCys1 and BCys2, respectively. The user has the possibility of modifying the default assignments, as shown in Fig. 7. The same figure shows an example of how links between between symmetry-related elements are defined.

Links between alternative conformation are a flexible tool that can handle complex situations. For example, in the crystal structure of β-mannanase in complex with 2,4-dinitrophenyl 2-deoxy-2-fluoro-β-mannotrioside the unhydrolysed substrate and a covalent intermediate are present simultaneously (Ducros *et al.*, 2002; Fig. 8). In the unhydrolysed form of the substrate the sugar moiety MAF is bound to the BEN compound, whereas in the covalent intermediate MAF is bound a glutamate residue. In the intermediate the CD—OE1 and CD—OE2 bonds of the glutamate are no longer equivalent. Once the descriptions of all links and modifications required have been defined, links between alternative conformations can handle this complicated type of situation (see Fig. 7).

7. Conclusions and future perspectives

A flexible machine/human-readable dictionary of monomers, links, modification and related items has been created and tested on a wide range of compounds. The dictionary is currently used for macromolecular restrained refinement by the program *REFMAC*5. It can also be used by other macromolecular programs such as model-building and macromolecular-modelling and simulation applications.

Flexibility in the organization of the dictionary allows researchers to add personal entries and to override existing descriptions. The most common crystallographic restraints are dealt with in an automatic manner. Complicated cases can also be handled with some user intervention.

Tools and algorithms have been developed to update and add new entry descriptions semi-automatically. If reliable coordinates are available, 'ideal' restraints can be extracted from them. When monomer descriptions are created from chemical structures, target values for restraints are taken from a built-in atom-type library. Currently, work to improve restraints using Cambridge Structural Database tools and quantum-chemical calculations is being considered.

Although addition of new entries is at present fairly automatic, the generation of links and modifications requires user intervention. Automatization of the latter process requires a database of possible reactions encountered in macromolecules. Link descriptions allow not only the definition of covalent bonds but also of other bonds. The automatic handling of Watson–Crick restraints between base pairs is currently under development. Future versions of the dictionary will contain tools to use popular computational chemical file formats such as SMILES (Weininger, 1988) and MDL MOLFILES (Dalby *et al.*, 1992).

The dictionary is distributed by CCP4 under the Part 0 licence that is LGPL-compatible. Programs and interface are available from CCP4 under the Part 2 licence. Neither programs nor dictionary nor algorithms have been patented in order to make sure that they are available to users as well as to the developer community.

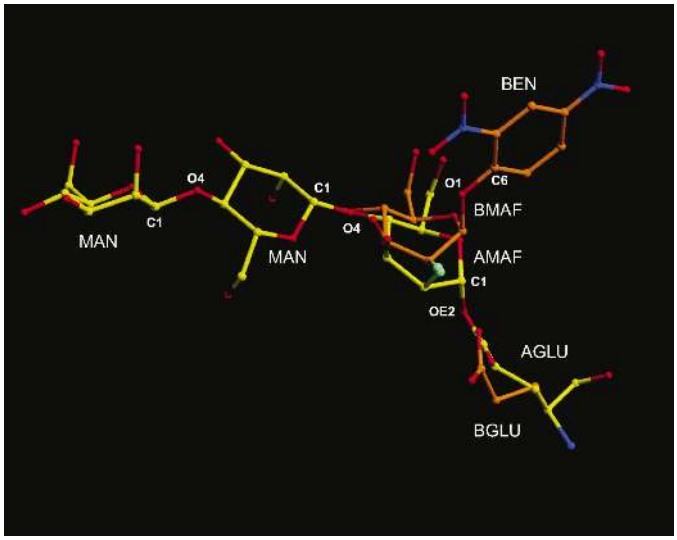


Figure 8
Example of complicated links involving residues in alternate conformations. In the crystal structure of β-mannanase in complex with 2,4-dinitrophenyl 2-deoxy-2-fluoro-β-mannotrioside the unhydrolysed substrate and a covalent intermediate are present simultaneously (Ducros *et al.*, 2002). In the unhydrolysed form the substrate the sugar moiety MAF is bound to BEN, whereas in the covalent intermediate MAF is bound a glutamate residue. The links and modifications needed to refine this complex are given in Fig. 7.

This work was supported by grants from the Wellcome Trust (GNM, RAS), BBSRC (AAV, AAL) and CCP4 (LP, SM, FL). We thank people from the YSBL, CCP4 staff and the user community for their continuous support and encouragement.

References

Aalten, D. van, Bywater, R., Findlay, J., Hendlich, M., Hooft, R. & Vriend, G. (1996). *J. Comput. Aided Mol. Des.* **10**, 255–262.
Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Allen, F., Kennard, O., Watson, D., Brammer, L., Orpen, A. & Taylor, R. (1992). *International Tables for Crystallography*, Vol. *C*, edited by A. J. C. Wilson, pp. 685–706. Dordrecht: Kluwer Academic Publishers.
Allinger, N. (1977). *J. Am. Chem. Soc.* **99**, 8127.
Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). *Acta Cryst.* D**58**, 899–907.
Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Bourne, P., Berman, H., McMahon, B., Watenpaugh, K., Westbrook, J. & Fitzgerald, P. (1997). *Methods Enzymol.* **277**, 571–590.
Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.
Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. & Karplus, M. (1983). *J. Comput. Chem.* **4**, 187–217.
Brünger, A. T. (1992). *X-PLOR Manual, Version 3.1.* Yale University, New Haven, USA.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, N., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Cahn, R., Ingold, C. & Prelog, V. (1966). *Angew. Chem.* **78**, 413–447.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cotton, F. & Wilkinson, G. (1972). *Advanced Inorganic Chemistry.* New York: Interscience.
Dalby, A., Nourse, J., Hounshell, D., Gushurst, A., Grier, D., Leland, B. & Laufer, J. (1992). *J. Chem. Inf. Comput. Sci.* **32**, 244–255.
Diamond, R. (1971). *Acta Cryst.* A**27**, 436–452.
Ducros, V., Sechel, D., Murshudov, G., Gilbert, H., Szabo, L., Stoll, D., Withers, S. & Davies, G. (2002). *Angew. Chem. Int. Ed.* **41**, 2824–2827.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Golovin, A. *et al.* (2004). *Nucleic Acids Res.* **32**, D211–D216.
Greenwood, N. & Earnshaw, A. (1989). *Chemistry of the Elements.* Oxford: Pergamon Press.
Hall, S. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 326–333.
Hall, S., Allen, A. & Brown, I. (1991). *Acta Cryst.* A**47**, 655–685.
Ihlenfeldt, W., Takahashi, Y., Abe, H. & Sasaki, S. (1994). *J. Chem. Inf. Comput. Sci.* **34**, 109–116.
IUPAC (1979). *Nomenclature of Organic Chemistry, Sections A, B, C, D, E, F and H.* Oxford: Pergamon Press.
Jack, A. & Levitt, M. (1978). *Acta Cryst.* A**34**, 931–935.
Jones, A. T., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kennard, O. & Taylor, R. (1982). *J. Am. Chem. Soc.* **104**, 3209–3212.
Konnert, J. & Hendrickson, W. (1980). *Acta Cryst.* A**36**, 344–350.
Leach, A. (1997). *Molecular Modelling: Principles and Applications.* Singapore: Longman.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Orpen, A., Brammer, L., Allen, F., Kennard, O., Watson, D. & Taylor, R. (1992). *International Tables for Crystallography*, Vol. *C*, edited by A. J. C. Wilson, pp. 707–791. Dordrecht: Kluwer Academic Publishers.
Pearlman, D., Case, D., Caldwell, J., Ross, W., Cheatham, T. III, DeBolt, S., Ferguson, D., Seibel, G. & Kollman, P. (1995). *Comput. Phys. Commun.* **91**, 1–41.
Ponder, J. & Case, D. (2003). *Adv. Protein Chem.* **66**, 27–85.
Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. (2003). *Acta Cryst.* D**59**, 1131–1137.
Sadowski, J., Gasteiger, J. & Klebe, G. (1994). *Chem. Inf. Comput. Sci.* **34**, 1000–1008.
Saenger, W. (1983). *Principles of Nucleic Acid Structure.* Berlin: Springer–Verlag.
Terwilliger, T. C. (2003). *Acta Cryst.* D**59**, 1688–1701.
Ullman, J. (1976). *J. Assoc. Comput. Mach.* **23**, 31–42.
Waser, J. (1963). *Acta Cryst.* **16**, 1091–1094.
Weininger, D. (1988). *J. Chem. Inf. Comput. Sci.* **28**, 31–36.
Westhof, E., Dumas, P. & Moras, D. (1988). *Acta Cryst.* A**44**, 112–123.