Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Automated production of small-molecule dictionaries for use in crystallographic refinements

Richard B. Greaves,* Alexei A. Vagin and Eleanor J. Dodson

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

Correspondence e-mail: greaves@yorvic.york.ac.uk

Received 8 February 1999
Accepted 4 May 1999

Many macromolecules are now being studied crystallographically in complexes with a range of ligands and other associated molecules. It is necessary to have templates describing the expected geometry of such molecules before refinement and model building can be carried out. This paper describes a method for generating templates beginning from the *SMILES* description of the molecule, the final format of the molecular template being based on the mmCIF definitions for chemical composition. Additionally, the program *SMILE2DICT*, which converts the *SMILES* string to a more extended format, is described. The description details the input required, the output produced and how the program relates to attempts to automate the procedure of model building for crystallographic refinement. Examples of input to and output from the program are given.

1. Introduction

The process of solving a macromolecular structure rests heavily on prior knowledge of the stereochemistry of similar molecules. Both at the stage when structures are being built into density maps and during refinement, it is necessary to have templates for the expected structure. Accurate examples are available for the peptide backbone and amino acids, for nucleotides and for other common components of macromolecules. These have been garnered from statistical analysis of the many small organic molecule crystal structures deposited in the Cambridge Structural Database (Allen *et al.*, 1991). (The deposited macromolecular structures in the Protein Data Bank are not usually sufficiently well defined to provide good templates.) Engh & Huber (1991) codified the geometric information for amino acids into dictionaries for *X-PLOR* (Brünger, 1992), and this has since been adapted for other refinement and model-building programs.

Increasingly, structural studies are carried out on proteins complexed with a variety of ligands, and there is a great need for some simple way of codifying the expected geometry of these moieties and generating a suitable three-dimensional template from the chemical connectivity alone. The atoms present and their chirality, plus their connectivity, are usually known and from this, plus a knowledge of expected bond distances and angles, a set of standardized model coordinates can be assigned. It is our aim to provide tools which allow reliable construction of model coordinates for a small molecule.

One approach is to build the molecule of choice in a graphics-based sketching program such as *ChemDraw* (Klein, 1995) or the Molecular Editor in *Quanta*98 (Molecular Simulations Inc., 1998), which assemble the molecule from 'building blocks' such as benzene rings and methylene groups. These can be pieced together to build coordinates, which in turn can be used to generate the standard 'dictionary' files required by the various refinement and model-building programs which need to know atom types and have values for bond distances and angles, chiralities, planar groups and so on. However, organic chemists increasingly work with *SMILES* strings and many databases now exploit this formalism.

There has been no consensus on output dictionary format, but we are utilizing the definitions of the mmCIF dictionary, already accepted by the IUCr and PDB. The geometries of valine models derived from the Engh and Huber parameters and from *SMILE2DICT* are shown in Fig. 1 for comparison. The largest difference is that for the CA—CB bond, but this is within the statistically acceptable 'three-e.s.d.' range. Engh and Huber results are based on the analysis of many peptide structures, whereas the *LIBCHECK* parameters (Vagin *et al.*, 1998) which are used by *SMILE2DICT* in dictionary compilation, are derived from the geometry of all small organic molecules based on defined atom types.

2. *SMILE2DICT*

SMILES (simplified molecular input line entry specification) (Weininger, 1998) is a compact method of expressing molecular connectivity widely used by organic chemists. The strings do not require great computing resources to parse or to store. They simplify pattern matching and are in common use in conjunction with databases.

The *SMILES* syntax is extremely simple, defining connectivity and a range of chemical atom types. Simple linear chains are written as strings of letters such as CCO, which is ethanol (where each letter represents a 'heavy' atom). Upper-case letters denote aliphatic atoms and lower-case letters represent aromatic atoms. Ring closures are denoted by a numeric suffix *e.g.* c1ccccc1 is benzene. Branching is indicated by brackets, *e.g.* c1ccc(C)cc1 is toluene. The branched group can be quite complex and may involve further branches, *e.g.* in valine, N[C@@H](C(C)C)C(=O)O, where '=' denotes a double bond and '@@' a chiral atom. (see Figs. 2*a*–2*d*)

In *SMILES*, chirality is indicated by a chiral specification ('@' or '@@') written as an atomic property following the atomic symbol of the chiral atom. If a tetrahedral centre is not specified as chiral in the *SMILES* string, then its chirality is implicitly undefined.

Looking at the chiral centre from the direction of the previous atom bonded to it (within the context of *SMILES*), @ means that the next three groups linked to the chiral centre are listed anticlockwise, @@ means that they are listed clockwise.

In the valine example, the linked groups are N, H, $CH(CH_3)_2$ and COOH. Looking down the N—C$^{\alpha}$ bond we see these groups distributed as H, $CH(CH_3)_2$ and COOH clockwise. A similar designation could be employed to mark the improper dihedral needed to preserve configuration at CB, *i.e.* N[C@@H](C@(C)C)C(=O)O would be the full designation for valine.

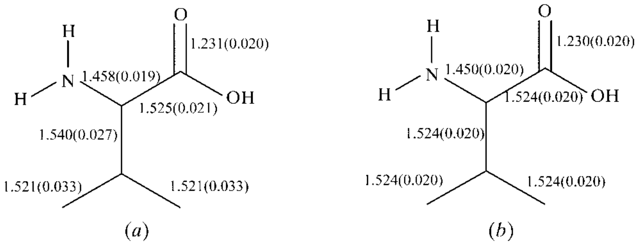

Figure 1
(*a*) Engh and Huber bond lengths, (*b*) bond lengths from SMILE2DICT for valine (with e.s.d. values in parentheses).

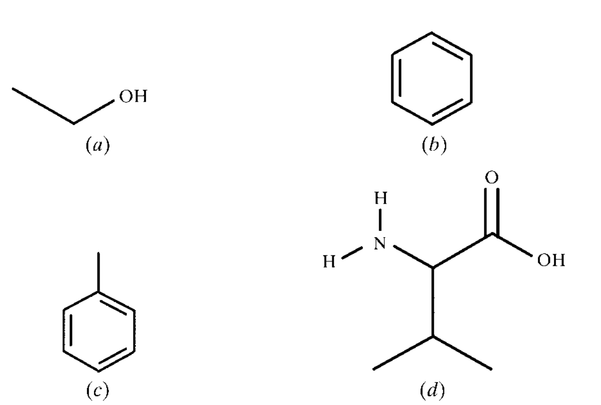

Figure 2
(*a*) Ethanol, CCO; (*b*) benzene, c1ccccc1; (*c*) toluene, c1ccc(C)cc1; (*d*) valine, N[C@@H](C(C)C)C(=O)O.

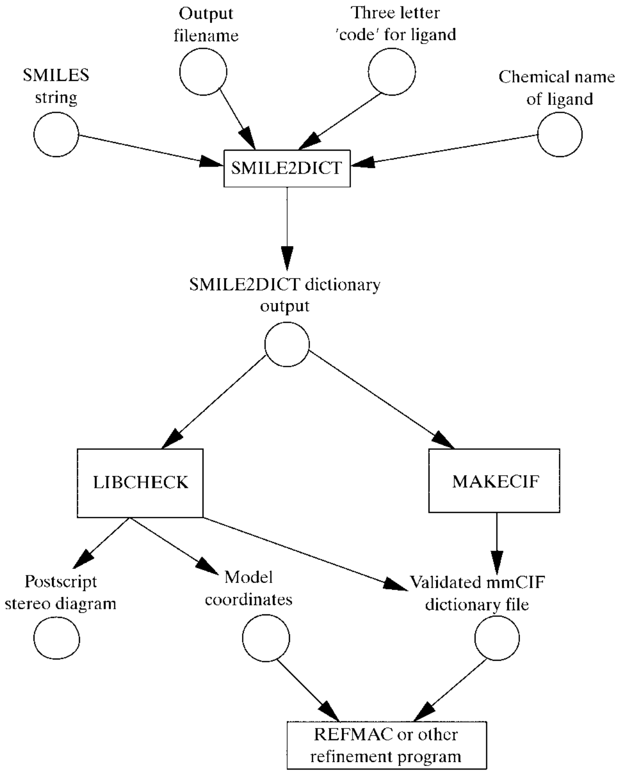
Output filename
Three letter 'code' for ligand
SMILES string
Chemical name of ligand
SMILE2DICT
SMILE2DICT dictionary output
LIBCHECK
MAKECIF
Postscript stereo diagram
Model coordinates
Validated mmCIF dictionary file
REFMAC or other refinement program

Figure 3
Flow diagram illustrating the relationship between *SMILE2DICT*, *LIBCHECK* and *MAKECIF*.

Thus, a wide range of the chemical atom types encountered in typical organic small molecules can be specified in a *SMILES* string. These include sp^3, sp^2 and sp^1 C atoms, sp^3 and sp^2 N atoms, sp^3 and sp^2 O atoms, as well as S and P atoms. In addition, the program *SMILE2DICT* recognizes and assigns further atom types appropriately, *e.g.* ring atoms. A summary of these atom types is presented in Table 1.

Whilst this is seemingly quite limited, and more sophisticated chemistry is needed to describe more complex molecules, it is sufficient to represent the level of information known for many macromolecular ligands.

For use within refinement and model-building programs a more complete description of the molecule is required, for example the identity and configuration of chiral atoms, planar atoms in the system *etc.*

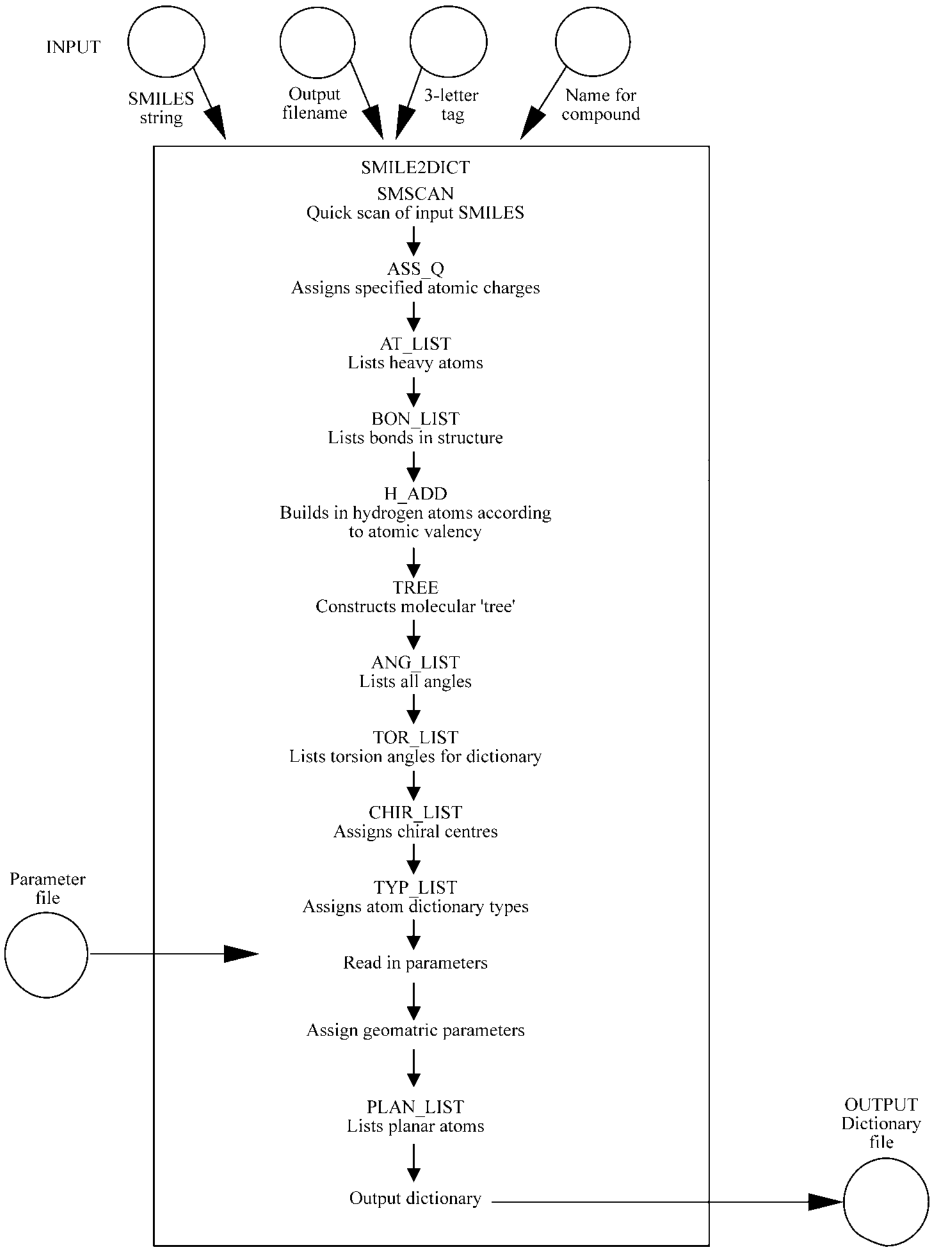

Figure 4
Program flow diagram for *SMILE2DICT*.

3. Program description

SMILE2DICT is a Fortran77 program which produces input for two further programs *LIBCHECK* and *MAKECIF* (Vagin *et al.*, 1998), which will be described in more detail elsewhere. The aim of the software is to be able to produce a mmCIF-compatible dictionary file from the input small-molecule *SMILES*. The relationship of *SMILE2DICT* to *LIBCHECK* and *MAKECIF* is illustrated in Fig. 3.

The input to the program is a *SMILES* formula, stored in a file as a character string. The user is asked to supply a name for his output dictionary, a chemical name for his compound, a three-letter code (such as the residue identification in PDB-format coordinate files) for the molecule and lastly the name of the file which contains the *SMILES* string. The structure of the program is shown in Fig. 4.

The program then runs a preliminary scan of the *SMILES* input to verify whether the brackets within it are balanced. Unbalanced parentheses would indicate an error, since this situation could not arise in a properly written *SMILES*. The next step is to list all the atoms present in the molecule. By default, atom names are assigned by counting along the *SMILES* input and numbering the atoms sequentially. This will mean that the output dictionary will contain atom names other than those used in the literature concerning that compound. The user is of course able to rename the atoms in his initial model coordinates file by editing the output dictionary.

The program next produces a preliminary bond list which is specified by the *SMILES* input. H atoms are added at the appropriate places according to atomic valency, and a final bond list and connectivity 'tree' are calculated. The final step is to list all the bond and torsion angles, to assign atom types based on the connectivity, to assign geometrical parameters and list the planar atoms. All the above are written to the output file. Missing parameters trigger warnings to the user which

Table 1
The atom types generated by *SMILE2DICT*.

Atom type	Chemical type	Example and notes
CSP	sp^1 C atom	Any triple-bonded C atom
C	sp^2 C atom	sp^2 without H atom, *e.g.* carbonyl C
C1	sp^2 C atom	sp^2 with 1 H atom
C2	sp^2 C atom	sp^2 with 2 H atoms, *e.g.* ethene
CR15	sp^2 C atom	C in 5-membered ring, 1 H atom
CR16	sp^2 C atom	C in 6-membered ring, 1 H atom
CR6	sp^2 C atom	C in 6-membered ring, no H atoms
CR56	sp^2 C atom	C between 5- and 6-membered rings
CR66	sp^2 C atom	C between 6-membered rings
CR5	sp^2 C atom	C in 5-membered ring, no H atoms
CH1	sp^3 C atom	With 1 H atom
CH2	sp^3 C atom	With 2 H atoms
CH3	sp^3 C atom	With 3 H atoms
CT	sp^3 C atom	Tetrahedral C without H atom
HCHx	Aliphatic H	H on CHx atom (aliphatic)
HCRx	Aromatic H	H on CRx atom (aromatic)
HNCx	Aliphatic H	H on NCx atom
HNHx	Aliphatic H	H on NHx atom
HNRx	Aromatic H	H on NRx atom
HOHx	Aliphatic H	Hydrogen on OHx atom
HSH1	Aliphatic H	H on SH1 atom
NS	sp^1 N atom	Triply bonded N, no H atom
NS1	sp^1 N atom	Triply bonded N, 1 H atom
N	sp^2 N atom	Without H atom, *e.g.* N of Pro
NC1	sp^2 N atom	Charged N with 1 H atom
NC2	sp^2 N atom	Charged N with 2 H atoms
NH1	sp^2 N atom	With 1 H atom, *e.g.* N of Ala
NH2	sp^2 N atom	With 2 H atoms, *e.g.* amide
NR15	sp^2 N atom	N in 5-membered ring, 1 H atom
NR16	sp^2 N atom	N in 6-membered ring, 1 H atom
NR5	sp^2 N atom	N in 5-membered ring, no H atom
NR6	sp^2 N atom	N in 6-membered ring, no H atom
NT	sp^3 N atom	Without H atom
NT1	sp^3 N atom	With 1 H atom
NT2	sp^3 N atom	With 2 H atoms
NT3	sp^3 N atom	With 3 H atoms
O	sp^2 O atom	With no net charge, *e.g.* carbonyl O atom
OC	sp^2 O atom	Charged O atom, *e.g.* carboxyl O atom
OP	sp^2 O atom	O bonded to phosphorus
OS	sp^2 O atom	O bonded to sulfur
OB	sp^2 O atom	O bonded to boron
O2	sp^3 O atom	O bonded to two atoms, *e.g.* ribose
OC2	sp^3 O atom	Charged O bonded to two atoms
OH1	sp^3 O atom	Alcohol O atom
OH2	sp^3 O atom	Tetrahedral O atom, *e.g.* O in water
P	P	Phosphorus
S	S	Sulfur without H atom
SH1	S	Sulfur with H atom, *e.g.* SG of Cys

appear on screen, but not in the output file. For these bonds, angles or torsions, a default value will appear in the output (0.000 for undefined bond lengths and angles and 999.0 for undefined torsions). Parameters for these internal coordinates need to be assigned as their absence from the dictionary will cause run-time errors in *LIBCHECK*. The integrity of the *LIBCHECK* output is of the utmost importance, as it is this that will be used by *REFMAC* (Murshudov *et al.*, 1997) in the refinement procedure.

Although *SMILE2DICT* will recognize chiral centres where they are specified within the *SMILES* input, it cannot check whether the correct configuration has been assigned or not, nor can it predict a chiral volume for a given centre. It will not recognize unlabelled chiral centres, nor can it assign prochiral atoms. Similarly, geometric (*cis–trans*) isomerism is largely

```
data_comp_list

loop_
_chem_comp.id
_chem_comp.one_letter_code
_chem_comp.name
_chem_comp.type
_chem_comp.number_atoms_all
_chem_comp.number_atoms_nh
_chem_comp.desc_level
BNZ x 'benzene                    UNKNOWN    12    6
.
#
data_comp_BNZ
#
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
 BNZ      C1    C    CR16       0.000
 BNZ      C2    C    CR16       0.000
 ...      ..    .    ....       .....
loop_
_chem_comp_tree.comp_id
_chem_comp_tree.atom_id
_chem_comp_tree.atom_back
_chem_comp_tree.back_type
_chem_comp_tree.atom_forward
_chem_comp_tree.connect_type
 BNZ      C1    .    .        C2    START
 BNZ      C2    C1   .        C3    .
 ...      ..    ..   .        ..    .
 BNZ      C6    C5   .        .     END
 ...      ..    ..   .        .     .
 BNZ      C6    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
 BNZ         C1   C2      coval       1.390   0.000
 ...         ..   ..      .....       .....   .....
_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
 BNZ         C1   C2   C3    120.000    0.000
 ...         ..   ..   ..    .......    .....
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
 BNZ    tor1   C6   C1   C2   C3     0.000    0.000  2
 ...    ....   ..   ..   ..   ..     .....    .....  .
loop_
_chem_comp_plane_atom.comp_id
_chem_comp_plane_atom.plane_id
_chem_comp_plane_atom.atom_id
_chem_comp_plane_atom.dist_esd
 BNZ   plan      C6      0.020
 ...   ....      ..      .....
```

Figure 5
Output dictionary.

ignored. A final limitation on the use of *SMILE2DICT* arises for metal-ion coordination compounds, the interpretation of which is presently beyond the capabilities of *SMILE2DICT*.

The output dictionary has the format shown in Fig. 5, in line with the dictionary format for mmCIF. Here, atom types CR16 are aromatic planar C atoms with one hydrogen bonded to them and HCR6 are simply the H atoms on a CR16. All the atom types are taken from Alexei Vagin's *LIBCHECK* dictionary.

The output can then be input to *MAKECIF* to ensure that the dictionary is totally compliant with mmCIF protocols and to *LIBCHECK*, which will produce an initial set of coordinates for the molecule and a labelled PostScript stereodiagram of the user's ligand.[1]

The program will be made publicly available through *CCP*4 and be released alongside *MAKECIF*.

The authors would like to thank the European Union, who funded this research (contract number BIO4-CT96-0189) through the Framework IV programme, and the BBSRC, who have funded the computational infrastructure of the Structural Biology Centre at York.

References

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Brünger, A. T. (1992). *X-PLOR Version* 3.1 *Manual.* New Haven, Connecticut: Yale University Press.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Klein, F. M. (1995). *J. Chem. Inf. Comput. Sci.* **35**, 166–167.

Molecular Simulations Inc. (1998). *QUANTA*98. 9685 Scranton Road, San Diego, CA 92121-3752, USA.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Vagin, A. A., Murshudov, G. N. & Strokopytov, B. V. (1998). *J. Appl. Cryst.* **31**, 98–102.

Weininger, D. (1988). *J. Chem. Inf. Comput. Sci.* **28**, 31–36.

[1] An example run on coumaroyl CoA is available from the IUCr electronic archive (Reference: li0328). Services for accessing these data are described at the back of this issue.