J-CAMD 337

PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules*

D.M.F. van Aalten[a,]**, R. Bywater[b], J.B.C. Findlay[a], M. Hendlich[c], R.W.W. Hooft[d] and G. Vriend[d]

[a]*Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, U.K.*
[b]*Biostructure Department, NOVO NORDISK A/S, DK-2880 Bagsvaerd, Denmark*
[c]*Preclinical Research/Drug Design, Merck KGaA, D-64271 Darmstadt, Germany*
[d]*Biocomputing, EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany*

Received 15 January 1996
Accepted 12 February 1996

Keywords: Molecular dynamics; Graph theory; SMILES

Summary

A software package is described that operates on small molecules observed in the PDB collection of protein structures. Molecular topology files for many molecular modeling programs can be generated automatically. The three-dimensional coordinates of small molecules can be converted to molecular descriptor strings that encode them uniquely in order to enable small-molecule recognition, despite high variability in atom and molecule nomenclature. From this descriptor a plausible 3D structure can be regenerated using energy minimisation. Alternatively, an ensemble of structures can be generated using a distance-geometry-based algorithm.

Introduction

Many protein modeling and molecular dynamics packages used for the analysis of protein–ligand complexes, drug docking or drug design, rely on a detailed description of the ligand. This description, the 'ligand topology', contains information concerning bonds, bond angles, charges, etc. Construction of such topology files is mainly done by hand and is rather laborious. A computer program which would be able to automatically construct a topology from 3D coordinates would, for example, facilitate the study of interactions between a vast collection of drugs and a certain protein.

PDB files [1] are useful for studying small-molecule–protein interactions. Large numbers of complexes can be compared, but automatic comparison is hampered by the fact that the nomenclature of these molecules and atoms in these molecules is not always consistent. A method which describes a molecule in a unique way, independent of the order or names of the atoms of small molecules in a PDB file is thus required.

Here we describe a program, PRODRG, which generates topologies from 3D coordinates. The program is interfaced to the molecular modeling package WHAT IF [2] and the molecular dynamics package GROMOS [3], providing a fast route from loading a protein with a ligand in WHAT IF to simulating this complex in GROMOS. A topology in the MOL2 format [4] is also generated, allowing usage by a wide range of programs. PRODRG is also available as a stand-alone program. It provides the following options: (i) creation of WHAT IF/GROMOS/MOL2 topologies from 3D coordinates of small molecules; (ii) description of small molecules in a unique character string, enabling recognition from a set of coordinates; and (iii) regeneration of a plausible 3D structure from these unique molecular line descriptors, enabling the distribution of a structural database with WHAT IF in the form of these descriptors.

Methods

Two problems have to be solved in order to be able to generate topologies form small molecules described in PDB files: (1) PDB files generally do not contain proton

*The program PRODRG is freely available by writing to D.M.F. van Aalten. In addition, structures can be submitted to the WWW server http://swift.embl-heidelberg.de/prodrg_serv, which will return a MOLDES and three molecular topologies.

**To whom correspondence should be addressed.

0920-654X/$ 6.00 + 1.00 © 1996 ESCOM Science Publishers B.V.

coordinates; and (2) small molecules in PDB files are sometimes described with alternative names, and the ordering and nomenclature of atoms are not always unique.

To generate a molecular topology several tasks have to be performed: atomic connectivities and hybridizations need to be determined; protons need to be placed; and bond angles, torsion angles and charges have to be determined. A detailed description of every subtask is given below.

From 3D structure to atomic connections and hybridizations

Bonds between atoms are determined using a cutoff criterion, i.e. a bond exists when the distance between any two atoms is less than 2.0 Å (or 2.2 Å for bonds involving iron or S–S bonds). The bonds derived from the coordinates are verified using the PDB CONECT records, where present. As PDB files normally do not contain hydrogen atoms, the connectivity matrix only contains connections between heavy atoms. Hybridizations for these heavy atoms are determined using bonds, bond angles, and improper dihedral angles (the dihedral angle determined by an atom and three of its connected atoms [5]), using methods similar to those described in Refs. 6 and 7. The hybridization type is set by default to sp^3, but is set to sp^2 if one of the following conditions is met:

(i) the atom has three connections to heavy atoms and the absolute improper dihedral angle is smaller than 18°, that is, halfway between the optimal value for an sp^2 atom (0°) and that of an sp^3 atom (36°);

(ii) the atom has two connections to heavy atoms and the angle defined by the atom and its two connected atoms is larger than 114.5°, that is, halfway between the optimal value for an sp^3 atom (109°) and an sp^2 atom (120°);

(iii) the atom has only one connection to a heavy atom and the distance between the atom and its connected atom is less than X, where X depends on the type and hybridization of the connected atom. Distances X were taken from Ref. 8. Correct placement of protons (which are not generally present in PDB files) depends solely on the accuracy of these distances in the PDB file.

The criteria taken for atoms with three and two connections (18° and 114.5° cutoffs, respectively) are quite liberal. In theory, angles and improper dihedral angles of sp^2 atoms are close to 120° and 0°, respectively. However, we need to be able to derive the hybridizations from rather inaccurate coordinates. The protein–small-molecule complexes found in PDB [1] files are generally solved to a resolution of 1.5–2.5 Å. This causes significant uncertainty in properties like (dihedral) angles and therefore rather large deviations from optimal angles. The hybridizations generated for atoms with three connections to heavy atoms can be checked by verifying bond angles and bond lengths. Assignments for atoms with two connections to heavy atoms can be checked by verifying bond lengths. Assignments for atoms with a single connection however depend solely on the accuracy of the distance between the two atoms. As a result, the single connection atoms have the largest chance of being assigned a wrong hybridization type.

After determination of the atom types and hybridizations, it is possible to place hydrogen atoms on atoms with empty valences. The three topology formats supported require only polar hydrogens to be included in the connectivity matrix. Apolar hydrogens (i.e. those on carbon atoms only) are listed as such with their carbon atoms but do not have a separate entry in the connectivity matrix. Hydrogen atoms are added until all bond valences are satisfied, for example an sp^3 nitrogen atom with only one connection will get two polar hydrogens, forming an amine group. An sp^3 oxygen atom with only one connection will get one polar hydrogen, forming a hydroxyl group.

From connectivity matrix to molecular descriptor string

Every molecule can be described as a tree [9]. It is possible to describe such a tree on a single line using characters to describe atoms, connections, branches, etc. One example of such a line description is that used in the SMILES system [10]. This system is optimized for user interaction with a database system, i.e., the user is able to enter a description of the molecule using meaningful characters, such as C, N, O, etc. for atom types; –, = and # for single, double and triple bonds, respectively; and () to indicate branches. For a full description see Refs. 10–12. For example, isobutyric acid is described by CC(C)C(=O)O. We use a slightly different method, which is more specialized for generating molecular topologies and for reconstruction of 3D structures from these molecular descriptors (MOLDES).

In a MOLDES, every atom is represented by six fields. The first describes the atom type (see Table 1), the second describes the hybridization, the third gives the number of connections to atoms other than hydrogen atoms, and the fourth gives the number of connections to hydrogen atoms (both polar and apolar). The fifth and sixth field contain information concerning ring structure and chirality and will be described later. Using the first four fields, isobutyric acid could be described by: 9313 9331 9313 9230 6210 6311 1010. The MOLDES has certain advantages and disadvantages compared to the original SMILES notation [10]. The main disadvantage is that a MOLDES is much less human-readable then a SMILES string. From a SMILES string, a chemist can easily depict a 2D graph of the molecule, whereas this is more difficult from a MOLDES. However, the MOLDES was not designed with human-readability as a main aim. It was designed for being easy to process by a computer, possibly in the form of a searchable database. It contains

TABLE 1
INTEGER CODES USED TO DESCRIBE ATOM TYPES IN THE FIRST FIELD OF ATOMIC DESCRIPTORS IN A MOLDES

Code	Atom type
9	carbon
8	phosphorus
7	nitrogen
6	oxygen
5	sulphur
2	iron
1	hydrogen

Currently only a small set of atoms is supported, but this can, of course, easily be extended.

directly accessible information concerning the hybridization and number of connections, which facilitates building molecular topologies. For instance, an aliphatic CH_2 group is always defined as 9322 in MOLDES, whereas a SMILES string would have to be fully decoded first to reach this kind of information.

If every molecule would have a unique MOLDES that is independent of the order and nomenclature of atoms in the PDB file, it would be possible to read a small molecule from a PDB file, generate a MOLDES via the connectivity matrix (see above) and compare the MOLDES to a MOLDES database in order to identify the molecule. In principle, the order of the descriptions of the last three atoms in the MOLDES for isobutyric acid (see above) could also be different: 6311 1010 6210, which still describes isobutyric acid. To make a MOLDES unique, two ordering rules are introduced:

(1) the longest branch which is attached to an atom is listed first, then the second longest and so on. This rule already determines the order of the last three atomic descriptors in the isobutyric acid MOLDES to be 6311 1010 6210 instead of 6210 6311 1010, since the 'hydroxyl' branch is one atom longer than the 'carbonyl' branch;

(2) if there are branches of equal length, the value of the fields in the atomic descriptions in the MOLDES are compared. If, starting from the beginning, a field contains a higher value, the branch containing that field is listed first. For instance, if a certain atom has the two branches 9322 9313 and 9221 6210, the ethyl side chain will be listed first, since its first atom is an sp^3 atom. This comparison is rather simple, since the fields form 4-digit integers and the number 9322 is simply larger than 9221.

Using these two rules and a routine which recursively descends the molecular tree, every molecule can be defined by a unique string, independent of the order of the atoms in the connectivity matrix (and PDB file). The recursion is started by first searching the longest path connecting atoms in the whole molecule. Then, the branches of atoms are ordered using the rules described above. The same is done for branches in these branches and so on. The ordering system used for SMILES strings is much faster [10]. However, our algorithm is only intended for use with small molecules. This ordering system also ensures proper organization of GROMOS charge groups which is essential for assigning proper atomic charges (see below). Rings are treated as side chains, disconnected at the point where they rejoin the main path at an atom which is already described. For example, 1,4-dihydroxy cyclohexane is in MOLDES described by:

```
 1010 +6311 +9331 +9322 +9322B+9322 +9322 +9331X+6311 +1010 +
( H     O     CH    CH2   CH2   CH2   CH2   CH    O     H     )
```

The longest path in this structure runs from one hydroxyl hydrogen to the other via one side of the ring, the other side is described as a side chain to this longest path. The fifth field of each atomic descriptor describes the ring connections. A ring-closing bond, connecting the 'ring side chain' to an atom in the main chain is only listed once. This is the reason why the fifth atom has a 'B' in this field, indicating that it is bonded to the second unconnected ring atom, which is atom 8. This atom has an 'X' in the fifth field, indicating that the atom is involved in a ring connection, but that the actual connection has already been described. This method of describing ring connections has two advantages: (1) the identifiers 'B' and 'X' do not depend on the actual distance between them. This means that if the 'ring side chain' described by atoms 4 and 5 would have another side chain attached to it, the 'B' and 'X' ensure that the cyclohexane ring is still easily identifiable; (2) very complex ring systems can still be described using a single field dealing with ring connections per atom.

The sixth field in the atomic descriptor deals with chirality. A '+' means that there is either no chirality or that the improper dihedral angle constructed by the atom and the first three of its connected atoms, found by reading the MOLDES from left to right is positive. A '–' means that this angle is negative. Although this 'chirality' is not always compatible with the chemical definition of chirality, proper definition of stereochemistry is ensured by this system.

From molecular descriptor string to connectivity matrix

The information contained in a MOLDES is sufficient for reconstructing the connectivity matrix from which it was derived. This can be done by realizing the following:

(1) atom N is bonded to only one atom in the range 1–(N–1), except if it is involved in ring formation;

(2) unless atom N has just one connection, the atom N+1 is bonded to atom N;

(3) ring connections are dealt with using the fifth field in the atomic descriptor.

So, having arrived at atom N and searching for an atom to which it is connected, the MOLDES is checked from

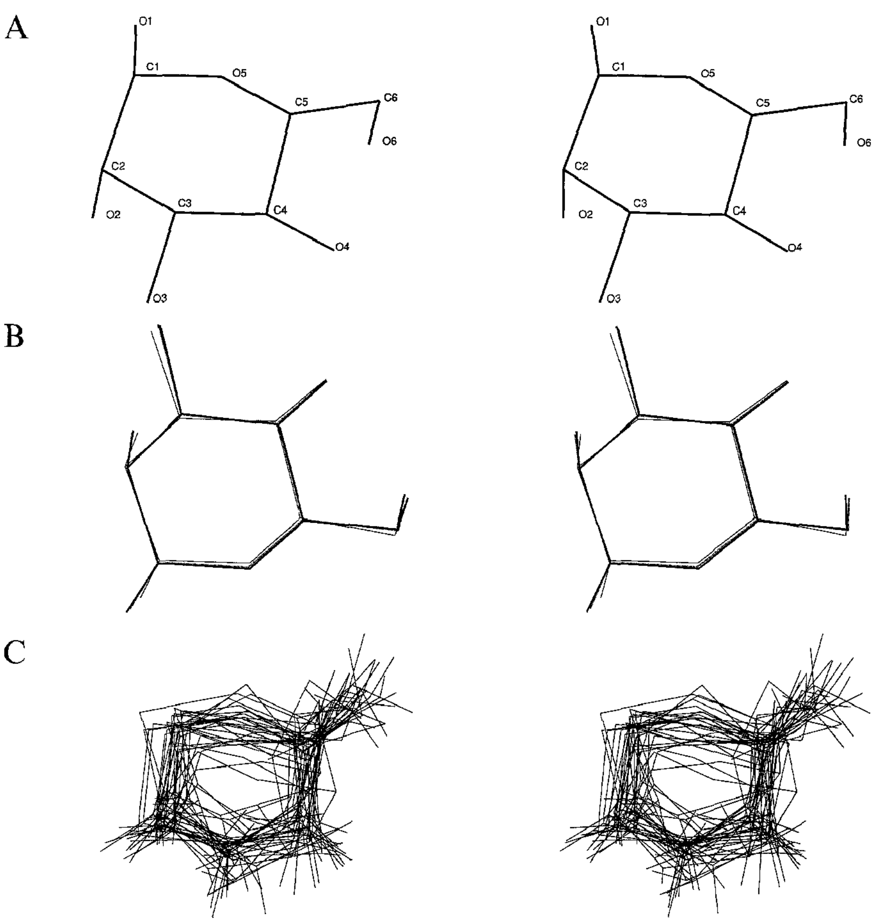

Fig. 1. (A) α-D-mannose PDB coordinates showing the atom names; (B) superposition of α-D-mannose PDB coordinates (thin solid line), PDB coordinates minimized using the PRODRG topology (thick dotted line) and coordinates resulting from the MOLDES → 3D approach (thick solid line); (C) superposition of 30 α-D-mannose structures generated with the DG approach.

atom N–1 to atom 1, until an atom with a free connection not involved in a ring is found, which by necessity is bonded to atom N. The same procedure is followed for atom N+1 and so on, until all atoms have been dealt with and the full connectivity matrix has been restored from the MOLDES.

From connectivity matrix to molecular topology

From the connectivity matrix a list of all bond angles, improper dihedral angles and a dihedral angle scan be reconstructed straightforwardly. Using this information, a molecular topology for WHAT IF [2] can be constructed. The WHAT IF topology record for a molecule consists of a list of atom names, bonds, angles, torsion angles and ideal values for bond lengths/angles. Included with every torsion angle is a list of atoms for which the coordinates should be recalculated when the torsion angle is changed. This facilitates interactive manipulation of flexible ligands. Finally, ideal atomic coordinates are listed together with the hydrogen-bonding characteristics of the atoms. Using the hybridization information and the number of connections to non-hydrogen and hydrogen atoms, it is possible to derive this information in an straightforward way. For example, an sp^3 oxygen atom connected to one hydrogen and one carbon atom can act both as donor and acceptor.

From the connectivity matrix, the list of hybridizations and the list of bonds, bond angles and (improper) dihedral angles, a molecular topology building block for GROMOS [3] can be constructed. The topology of a molecule is constructed using the GROMOS routine PROGMT, which builds the topology using topology building blocks (i.e. amino acids, heme, flavogroups, etc.). Adding the topology of a new small molecule to the existing file of topology building blocks has the advantage that the topology for the protein–small-molecule complex can be generated in one step (i.e. running the GROMOS routine PROGMT only once). In addition, a topology building block is reasonably human-readable, allowing modifications directly in the building block.

The first step in constructing a GROMOS topology for a small molecule is to determine the GROMOS atom type. The GROMOS'87 forcefield [3] works with a set of 37 types of atoms. All real atom types have to be mapped onto this set using rules based on three parameters: the periodic table atom type, the hybridization and the number of connections. For example, an sp^2 carbon in an aromatic ring will be assigned the atom type CR6*, the oxygen in a hydroxyl group is called OA and the nitrogen in an amine group is called NT. GROMOS uses the united-atom approach, i.e. apolar hydrogen atoms are included with the carbon atoms they are bonded to. So, for a methyl group, the atom type CH3 would be used. Polar hydrogens are treated separately. A MOLDES contains all this information and thus allows for easy determination of GROMOS atom types.

The second step is to derive information about bonds, bond angles and (improper) dihedral angles using the atom types. The GROMOS forcefield contains a list of possible combinations of atoms in bonds and angles. This list also contains information about optimal distances, optimal angles and force constants. Every combination of atoms has a unique number. For example, a bond with the bond code 14 implies a bond between an atom with type CH2 (methylene) and OA (oxygen in hydroxyl group) with an optimal distance of 1.43 Å and a force constant of 800 kJ mol^{-1} nm^{-2}. Based on these lists, bond codes, angle codes, improper dihedral angle codes and dihedral angle codes can be defined directly from the atom types. If a code does not exist for a certain combination of atoms in a bond or angle, the nearest plausible code is selected on the basis of atom names and hybridizations. In our implementation, the user is warned if such cases occur.

The third step involves determination of the charge groups. In GROMOS, groups of atoms are taken together to form a charge group with an integer charge. For example, an oxygen atom and a hydrogen atom forming a hydroxyl group would form a charge group together with the carbon atom to which the hydroxyl moiety is attached, with total charge 0.0, using atomic partial charges of +0.150 for the carbon, –0.548 for the oxygen and +0.398 for the hydrogen atom (see Ref. 3 for a full description of this approach). Charge groups and charges are determined using a database of often-occurring groups.

The fourth and last step is to find neighbours and second neighbours for each atom. In GROMOS, nonbonded interactions between neighbours and second neighbours are excluded, because they are implicitly incorporated in the bonded forces [3].

All information necessary to build a MOL2 topology file is included in the GROMOS topology. PRODRG

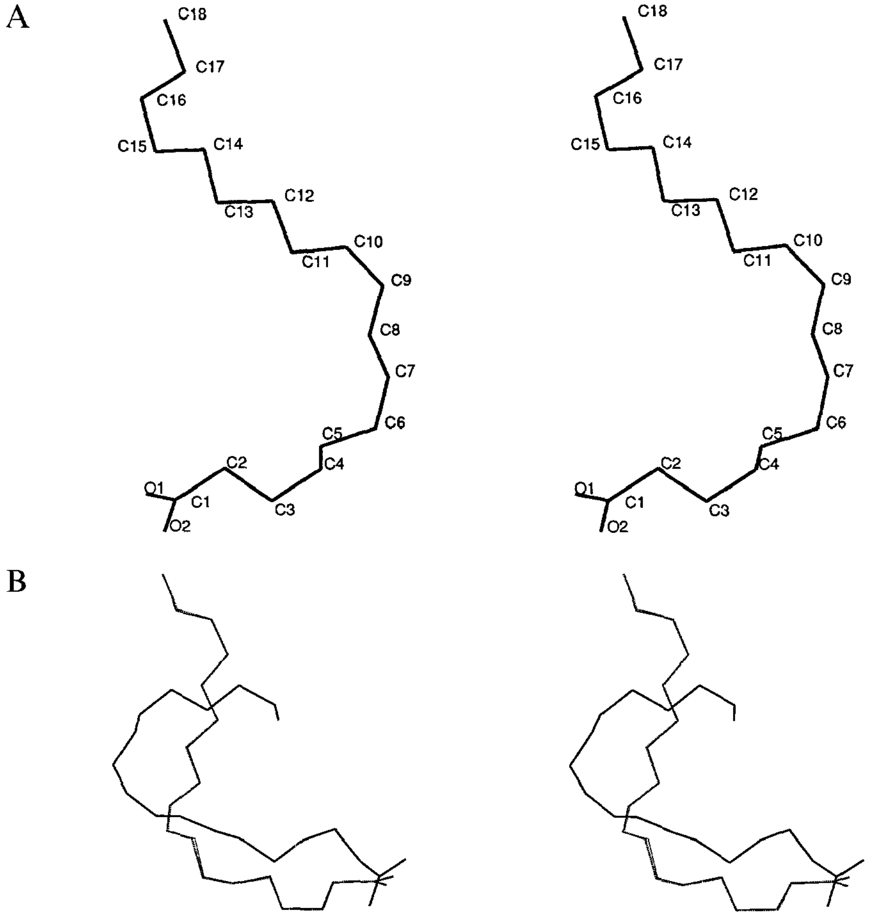

Fig. 2. (A) Oleic acid PDB coordinates showing the atom names; (B) superposition of oleic acid PDB coordinates (thin solid line), PDB coordinates minimized using the PRODRG topology (thick dotted line) and coordinates resulting from the MOLDES → 3D approach (thick solid line).

generates a molecular topology in MOL2 format using most of the data from the GROMOS topology, with the following changes:

(i) aromaticity needs to be explicitly defined in a MOL2 topology, which is derived using Hückel's rule;

(ii) the MOL2 topology format requires a more detailed bond-type definition, which is derived from the atom types and hybridizations.

From molecular topology to a 3D structure

We have implemented two approaches for generating a plausible 3D structure from a molecular topology. The first approach uses the GROMOS EM/MD procedure. The coordinates used are a set of random coordinates generated in a cube of 0.5 × 0.5 × 0.5 Å. SHAKE [13] cannot be used since initial bond lengths are incorrect. As will be shown below, proper bond lengths and bond angles are formed within a few EM steps, generating a plausible structure which is then further refined by a short MD simulation and a subsequent final EM. Using the sixth field of the atomic descriptors in the MOLDES, extra improper, dihedral restraints are applied to ensure proper chirality on atoms with four connections.

The second approach makes use of a distance-geometry (DG) algorithm similar to that described in [14]. A lower and an upper bound distance matrix are filled with 1–2 (bonds), 1–3 (angles), 1–4 (torsions) distances. In case of aromatic ring systems, some 1–5 and 1–6 distances can also be accurately specified. For small molecules (<20 atoms), about 50–75% of the matrix is defined at this stage. The matrix is then further filled in with random numbers and refined using triangular inequalities [15]. This matrix is used as input for the WHAT IF DG option and a set of plausible structures is generated.

Results

We tested the quality of our topology files and MOLDES strings by running cycles like: PDB coordinates → GROMOS topology → MD or DG → new coordinates, which should give a small rms between starting and final coordinates if the molecule has few internal degrees of freedom. Below, a few test cases are described in which our methods have been used on small molecules that are extracted from PDB files.

α-D-Mannose

The coordinates for α-D-mannose (Fig. 1A) were taken from the PDB entry 1APV [16], where it is bound to penicillopepsin. The MOLDES for α-D-mannose is:

```
 1010 +6311 +9331B-9331 +6311 +1010 +9331 +6311 +1010 +9331 -
( H     O     CH    CH    O     H     CH    O     H     CH

 6311 +1010 +6320 +9331X+9322 +6311 +1010 +
  O     H     O     CH    CH2   O     H     )
```

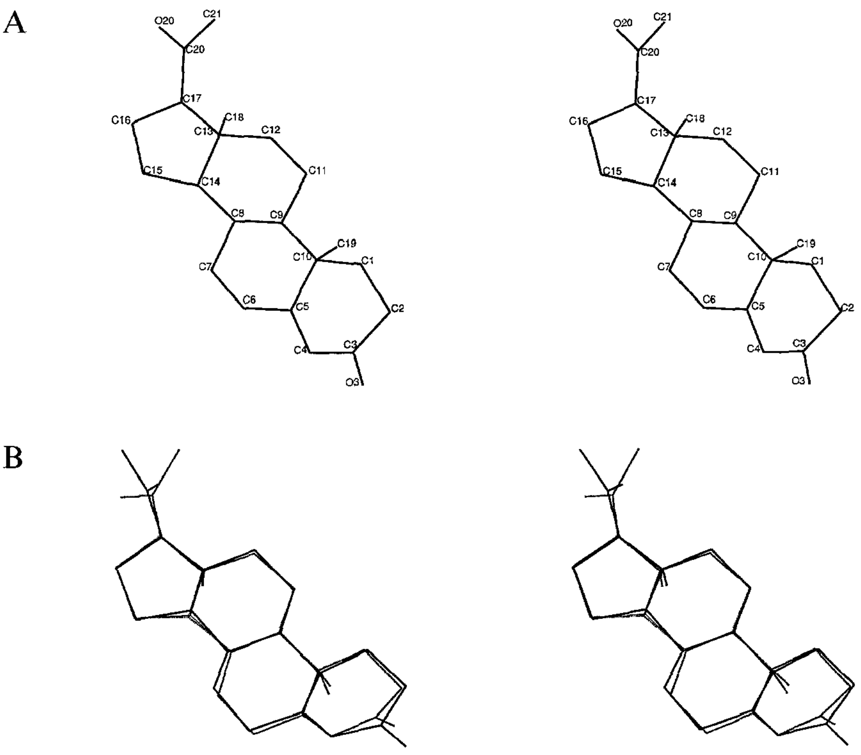

Fig. 3. (A) Progesterone PDB coordinates showing the atom names; (B) superposition of progesterone PDB coordinates (thin solid line), PDB coordinates minimized using the PRODRG topology (thick dotted line) and coordinates resulting from the MOLDES → 3D approach (thick solid line). Except around atom C14 the thin solid and the dotted lines coincide.

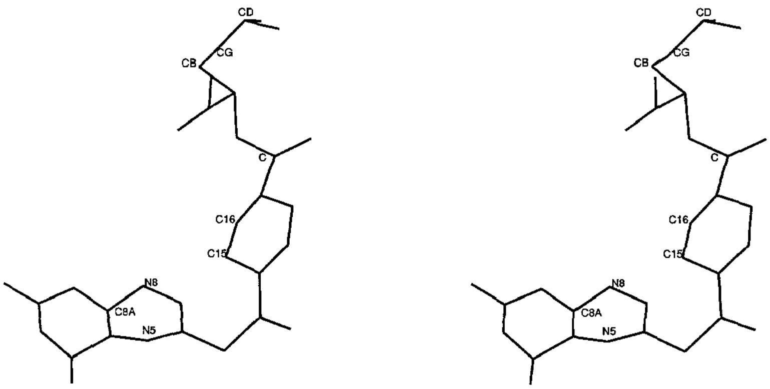

Fig. 4. Structure of methotrexate as found in the PDB entry 1TDR [19]. Atoms involved in groups which show large deviations from optimal behaviour are indicated.

All oxygen atoms are properly protonated to hydroxyl groups and all atoms are correctly assigned sp^3 hybridization. The field describing the chirality of the atoms proves to be useful, the MOLDES for β-D-mannose would contain a '+' instead of a '–' in the last field of the 4th carbon atom. The superposition of the PDB file structure, the energy-minimized PDB coordinates and the structure regenerated from the MOLDES of α-D-mannose using the EM/MD procedure shows that the three structures are very similar. The root mean square positional errors of the coordinates with respect to the initial (experimentally determined) coordinates (rms$_e$) are 0.12 Å and 0.16 Å, respectively (Fig. 1B). Distance geometry using the mannose topology yields a collection of structures with an average rms$_e$ of 0.43 (Fig. 1C).

Oleic acid

Oleic acid coordinates (see Fig. 2A) were taken from the PDB entry 1LID [17], where it is bound to the adipocyte-lipid-binding protein. Using the coordinates, the correct connectivity matrix is defined and all atoms are assigned sp^3 hybridization, except for atoms C9, C10 (which are involved in the double bond) and C1,O1,O2 of the unprotonated carboxyl group. The MOLDES for oleic acid is:

```
 9313 +9322 +9322 +9322 +9322 +9322 +9322 +9322 +9221 +9221 +
( CH3   CH2   CH2   CH2   CH2   CH2   CH2   CH2   CH    CH

 9322 +9322 +9322 +9322 +9322 +9322 +9322 +9230 +6210 +6210 +
  CH2   CH2   CH2   CH2   CH2   CH2   CH2   C     O     O    )
```

Figure 2B shows a superposition of the structure extracted from the PDB file. This structure is energy-minimized using the automatically generated molecular topology, and regenerated from the MOLDES using the EM-MD → 3D structure approach. The minimized PDB coordinates are close to their initial values (rms$_e$ of 0.06 Å), indicating the molecular topology is correct. The 3D structure regenerated from the MOLDES looks different (rms 2.8 Å), but this is to be expected, since oleic acid is a flexible molecule.

Progesterone

The PDB entry 1DBB [18] contains an immunoglobulin complexed with progesterone (Fig. 3A). PRODRG correctly recognizes the three double bonds and assigns proper hybridizations to all atoms. The ring structure is quite complex, but is nevertheless described in the MOLDES without any problems:

```
 9313 +9340F+9322 +9322 +9331E+9340D-9313 +9322 +9322 +9230 +
( CH2   C     CH2   CH2   CH    C     CH3   CH2   CH2   C

 6210 +9221 +9230X+9322 +9322 +9331X-9331X+9322 +9322 +9331A+
  O     CH    C     CH2   CH2   CH    CH    CH2   CH2   CH

 9230 +9313 +6210 +
  C     CH3   O    )
```

The longest path in the molecule almost includes the entire molecule, leaving C2 (Fig. 3A) with two unconnected bonds. One connection (F) is listed on atom C2 itself, the other (indicated by an 'A') is listed on an atom further on in the MOLDES. Here, again, the '+/–' designations enable conservation of proper stereochemistry. Figure 3B shows that progesterone is well-modeled by the molecular topology derived from the connectivity matrix. The structure regenerated from the MOLDES using the EM-MD approach gives an rms$_e$ of 0.5 Å (mainly caused by the rotation about the freely rotatable C17–C20 bond), and the energy-minimized coordinates deviate by only 0.04 Å from the X-ray coordinates.

Discussion

The accuracy of PRODRG is limited by the accuracy of the input coordinates. If these contain severe errors, wrong hybridization parameters are produced, resulting in incorrect placement of hydrogen atoms which in turn results in incorrect structures and topologies (and hence improper dynamic behaviour). However, the PRODRG facility to allow the user to interactively correct connections and hybridizations can be used to overcome the problems that result from errors in the coordinates.

An example of PRODRG failure, caused by poor coordinates is the molecule methotrexate (Fig. 4) in the PDB entry 1TDR [19]. The structure of this telluromethionyl dihydrofolate reductase and its bound ligand have been solved to 2.5 Å resolution. Carbon atoms involved in carbonyl and carboxyl groups have large deviations from their optimal improper dihedral angle. This problem is extreme for the carboxyl carbon CD, which has an improper dihedral angle of 53°. Atoms in the aromatic rings show very large deviations of ideal aromatic behaviour. For instance, the bond angle of atom C16 is 113°, which is closer to sp^3 than sp^2 behaviour. N5 and N8 also have large deviations, but stay within the limit of sp^2 behaviour. The improper dihedral angle of C8A is 19°, causing the atom to be recognized as an sp^3 CHR1R2R3 system. We intend to create a database of MOLDES strings that can be used for screening PDB entries for this kind of obvious errors.

Several parts of the PRODRG algorithm have been reported in the literature (e.g. Refs. 6,7 and 20). There are a few programs available which derive hybridizations and bond orders from high-resolution small-molecule structures [6,7]. There are many commercial programs available for 2D → 3D conversion [20]. PRODRG however, combines all these functionalities and is the first program to automatically generate GROMOS/WHAT IF topologies from PDB coordinates. The new 1D → 3D (MOLDES → 3D structure) approach using EM/MD is able to regenerate 100% of the structures (because it can start from random coordinates and use the topology to model the bonded forces in the molecules), compared to poorer performance of some of the commercially available programs [20].

The DG approach seems to work well for mannose. For less rigid or larger molecules, the DG structures show larger deviations. This is partly due to the fact that we have relatively few distances as input, and it is partly an inherent characteristic of DG methods that internal flexibility in molecules lead to a large variety in the obtained structures. Incorporation of additional knowledge and refinement procedures can reduce the errors, leaving a distribution of structures that represents the potential internal motions in the molecule.

Conclusions

The program PRODRG* described here proves to be useful in the following ways:

(1) WHAT IF, GROMOS and MOL2 topologies are constructed fully automatically. For a molecule such as progesterone, manually constructing a GROMOS molecular topology building block would take at least one day. This has now been reduced to 2 s;

(2) it encrypts every 3D structure into a unique MOLDES. This enables the construction of a database of MOLDES which can be searched when a PDB file is read, i.e.: when coordinates are read, a MOLDES is constructed, this MOLDES is compared with a MOLDES database and if a match is found correct names of the molecule and atoms can be determined, regardless of what is was called in the input file;

(3) a 3D structure can be generated from a MOLDES in a few seconds using an EM/MD procedure. In fact, a MOLDES database is a structural database. This is useful in the field of protein–small-molecule docking studies. Molecules can be extracted from the MOLDES database, their structure generated, and since a WHAT IF topology is made, flexibly docked into a protein using any of the docking programs interfaced to WHAT IF, or manually.

References

1 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.

2 Vriend, G., J. Mol. Graph., 8 (1990) 52.

3 Van Gunsteren, W.F. and Berendsen, H.J.C., BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen, Groningen, The Netherlands.

4 SYBYL Molecular Modelling Software, Tripos Associates, St. Louis, MO, 1991.

5 Van Gunsteren, W.F. and Berendsen, H.J.C., Angew. Chem., Int. Ed. Engl., 29 (1990) 992.

6 Meng, E.C. and Lewis, R.A., J. Comput. Chem., 12 (1991) 891.

7 Baber, J.C. and Hodgkin, E.E., J. Chem. Inf. Comput. Sci., 32 (1992) 401.

8 Weast, R.C. (Ed.) Handbook of Chemistry and Physics, Chemical Rubber Co., Cleveland, OH, 1964.

9 Kier, L.B. and Hall, L.H., Molecular Connectivity in Chemistry and Drug Research, Academic Press, New York, NY, 1976.

10 Weininger, D., J. Chem. Inf. Comput. Sci., 28 (1988) 31.

11 Weininger, D., Weininger, A. and Weininger, J.L., J. Chem. Inf. Comput. Sci., 29 (1989) 97.

12 Weininger, D., J. Chem. Inf. Comput. Sci., 30 (1990) 237.

13 Ryckaert, J.P., Cicotti, G. and Berendsen, H.J.C., J. Comput. Phys., 23 (1977) 327.

14 Crippen, G.M., J. Comput. Chem., 24 (1977) 96.

15 Converter; Biosym Techonologies Inc., San Diego, CA, 1992.

16 James, M.N.G., Sielecki, A.R., Hayakawa, K., Gelb, M.H., Biochemistry, 31 (1992) 3872.

17 Xu, Z.H., Bernlohr, D.A. and Banaszak, L.J., J. Biol. Chem., 268 (1993) 7874.

18 Arevalo, J.H., Stura, E.A., Taussig, M.J. and Wilson, I.A., J. Mol. Biol., 231 (1993) 103.

19 Boles, J.O., Lewinski, K., Kunkle, M.G., Hatada, M., Lebioda, L., Dunlap, R.J. and Odom, J.D., Acta Crystallogr. D., 51 (1995) 731.

20 Ricketts, E.M., Bradshwa, J., Hann, M., Hayes, F., Tanna, N. and Ricketts, D.M., J. Chem. Inf. Comput. Sci., 33 (1993) 905.