J Biomol NMR (2015) 63:21–37
DOI 10.1007/s10858-015-9959-y

ARTICLE

NMR structure calculation for all small molecule ligands and non-standard residues from the PDB Chemical Component Dictionary

Emel Maden Yilmaz[1] · Peter Güntert[1,2,3]

Received: 28 May 2015 / Accepted: 22 June 2015 / Published online: 30 June 2015

Abstract An algorithm, CYLIB, is presented for converting molecular topology descriptions from the PDB Chemical Component Dictionary into CYANA residue library entries. The CYANA structure calculation algorithm uses torsion angle molecular dynamics for the efficient computation of three-dimensional structures from NMR-derived restraints. For this, the molecules have to be represented in torsion angle space with rotations around covalent single bonds as the only degrees of freedom. The molecule must be given a tree structure of torsion angles connecting rigid units composed of one or several atoms with fixed relative positions. Setting up CYANA residue library entries therefore involves, besides straightforward format conversion, the non-trivial step of defining a suitable tree structure of torsion angles, and to re-order the atoms in a way that is compatible with this tree structure. This can be done manually for small numbers of ligands but the process is time-consuming and error-prone. An automated method is necessary in order to handle the large number of different potential ligand molecules to be studied in drug design projects. Here, we present an algorithm for this purpose, and show that CYANA structure calculations can be performed with almost all small molecule ligands and non-standard amino acid residues in the PDB Chemical Component Dictionary.

Keywords Drug design · Protein–ligand complex · Non-standard amino acid · Molecular topology · Structure calculation · CYANA

Introduction

Proteins comprise besides the 20 standard amino acids a variety of other building blocks and interact with a large number of low molecular weight ligands, and many more potential ligands are considered in drug design. NMR spectroscopy is excellently suited for studying protein–ligand interactions and for discovering high-affinity ligands for proteins (Arkin and Wells 2004), e.g. by "SAR by NMR" (Shuker et al. 1996) and related approaches. NMR has also been used to determine the three-dimensional (3D) structures of peptides and proteins that include a variety of non-standard amino acids, for instance the immunosuppressant cyclosporine A (Kallen et al. 1991; Weber et al. 1991) or the cytotoxic channel-forming non-ribosomal protein polytheonamide B (Hamada et al. 2010). The 3D structures of these molecular systems can be calculated on the basis of conformational restraints from NMR experiments.

NMR structure determination is composed of several steps: sample preparation, NMR spectroscopy, resonance assignments, collection of conformational restraints, structure calculation and structure refinement (Güntert 2009; Wüthrich 1986). Since the beginnings of NMR protein structure determination research, it was proposed that the complete procedure could be automated, which

✉ Peter Güntert
guentert@em.uni-frankfurt.de

[1] Center for Biomolecular Magnetic Resonance, Institute of Biophysical Chemistry, Goethe University Frankfurt am Main, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany

[2] Laboratory of Physical Chemistry, ETH Zürich, Zurich, Switzerland

[3] Graduate School of Science, Tokyo Metropolitan University, Hachioji, Tokyo, Japan

would dramatically speed up structure determination by NMR and make the method more objective (López-Méndez and Güntert 2006). There exists a collection of computational methods that were developed in order to automate specific parts of protein NMR structure determinations. The spectra analysis starts with peak picking, i.e. identifying the NMR signals in multidimensional spectra. Many methods for automated peak picking have been developed. They rely on rule-based feature recognition, neural networks, Bayesian networks, anti-phase fine structure pattern detection, e.g. (Alipanahi et al. 2009; Klukowski et al. 2015; Koradi et al. 1998). There are also algorithms to automate the chemical shift assignment step of NMR structure determination (Guerry and Herrmann 2011), e.g. (Bahrami et al. 2009; Bartels et al. 1997; Schmidt and Güntert 2012), as well as methods to achieve automated nuclear Overhauser effect (NOE) assignments such as ARIA (Bardiaux et al. 2012), AutoStructure (Huang et al. 2006), CANDID (Herrmann et al. 2002), and CYANA (Güntert and Buchner 2015). All these approaches were developed for work with proteins but can in principle also be applied to molecular systems containing arbitrary molecules.

A technical challenge in this respect is that the programs for NMR resonance assignment and structure calculation must be able to handle non-standard amino acids and arbitrary low molecular weight ligands. This poses a new demand on the CYANA software platform (Güntert 2009) that has so far been used extensively with proteins and nucleic acids, but only for a limited number of other molecules. CYANA structure calculations are performed in torsion angle space, i.e. with rotations about covalent bonds as the only degrees of freedom. To extend the use of CYANA to arbitrary molecules, the algorithm has to be extended by a new method to automatically generate the necessary residue library (topology) entries for these compounds, including the automated identification of a tree structure of torsion angles, which is required for the highly efficient torsion angle space molecular dynamics algorithm (Jain et al. 1993) that is used in CYANA for the structure calculations by simulated annealing (Güntert et al. 1997).

In recent years, the availability of structural information about biologically relevant molecules has grown rapidly. This data is stored in several structural databases, including the RCSB Protein Data Bank (PDB) (Berman et al. 2000; Westbrook et al. 2015) for proteins and nucleic acids with their ligands, and the Cambridge Structural Database (Allen et al. 1979) for other molecules. Such databases contain experimentally determined structures of low molecular weight molecules, including potential drug candidates, in their own format and chemical component representation scheme. They do not provide explicitly the torsion angle information that is required for NMR assignment and structure calculations using the torsion angle dynamics algorithm in CYANA. Thus, it is important to bridge the gap between the chemical component databases and CYANA such that the preparation of the chemical components for CYANA calculations can be done automatically without any manual work, which can be cumbersome, tedious and error-prone.

In order to use, without extensive manual work, arbitrary molecules from molecular databases in CYANA, sophisticated software tools are needed to generate the chemical component library entries. The CORINA software (Sadowski et al. 1994) can generate 3D coordinates for small- and medium-sized, typically drug-like molecules. This software supports many chemical file formats such as SD/RDfile, SMILES, SYBYL MOL/MOL2, MacroModel, Maestro, PDB, CIF or CTX, but not directly the format for torsion angle space calculations by CYANA. A software that can handle CYANA's format is Wit!P (http://www.biochem-caflisch.uzh.ch/download/). Partial conversions of 3D coordinates into CYANA residue library format are also provided by the molecular graphics program MOLMOL (Koradi et al. 1996). Nevertheless, none of these programs truly automates the complete conversion process. In general, considerable manual modifications are needed in order to obtain a working CYANA residue library entry. In contrast, the CYLIB program that we present in this paper is able to generate CYANA residue library entries fully automatically.

CYLIB uses the structural data of a molecule in the PDB Chemical Component Dictionary database for generating a corresponding entry in the CYANA residue library, such that the molecule can be used in NMR assignment and structure calculations with CYANA. The PDB Chemical Component Dictionary (http://www.wwpdb.org/data/ccd) is as a reference file describing all residue and small molecule components found in PDB entries (Westbrook et al. 2015). This dictionary contains detailed chemical descriptions for standard and modified amino acids/nucleotides, small molecule ligands, and solvent molecules. Each chemical component definition includes descriptions of chemical properties such as stereochemical assignments, chemical descriptors, systematic chemical names, and idealized coordinates generated by the CORINA software (Sadowski et al. 1994).

The present work addresses two main application areas. One application of CYLIB lies in pharmaceutical industry where NMR is a powerful method to study the interactions of drug candidates with target proteins in drug design. Since CYANA can be used for calculating the three-dimensional structure of complexes of proteins with ligands that are available in the residue database, CYLIB will simplify NMR studies in drug design by enabling the use of CYANA without the need for manual preparations of

CYANA residue library entries. The other main application of CYLIB is the structure calculation of peptides and proteins containing non-standard amino acids.

Algorithm

The goal of the CYLIB algorithm is to convert the mmCIF entries of the PDB Chemical Component Dictionary into CYANA format, such that they can be used directly in CYANA calculations without further manual adaptions.

Input and output files

The PDB Chemical Component Dictionary (Westbrook et al. 2015) is stored in mmCIF (macromolecular crystallographic information file) format (Bourne et al. 1997). A PDB Chemical Component Dictionary entry should not be confused with a "normal" macromolecular structure file, i.e. a PDB file that contains a protein 3D structure, although both can be stored in mmCIF format. As an example, the PDB Chemical Component Dictionary entry for the non-standard amino acid *N*-methyl-D-asparagine (PDB Chemical Component Dictionary code MND), which occurs in the cytotoxic channel-forming non-ribosomal protein polytheonamide B (Hamada et al. 2010), is shown in Fig. 1. The data that is relevant for CYLIB comprises the atom names, atom attributes (element type, aromatic flag, etc.), 3D coordinates, and covalent bonds. The MND entry in Fig. 1 is a non-standard amino acid. However, CYLIB can also handle almost any arbitrary (non-amino acid) molecule.

As its result, the CYLIB algorithm generates an output file containing the CYANA residue library entry that corresponds to the input mmCIF file. The resulting CYANA residue library entry for MND is shown in Fig. 2. The entry starts with a header line that specifies the entry name, the number of torsion angles, the number of atoms, the index of the first atom that belongs to the residue (those with lower indices belong to the preceding residue; see "Overlap atoms" section below), and the index of the last atom that belongs to the residue (those with higher indices belong to the next residue).

CYANA residue library entries can be stored in two equivalent formats. In the first, "traditional" format (Fig. 2a), atoms in torsion angle definitions and covalent connectivities are identified by atom numbers that correspond to those in the first column of the list of atoms. In the second, name-based format (Fig. 2b), the atoms are identified by their names. Since for the "overlap atoms" (comprising the backbone atoms C, O, N in amino acids; see below) it is possible to have two atoms with the same name in the atom list, an atom name may be preceded by a minus sign '−' to indicate that it refers to an atom of the preceding residue, or a plus sign '+' to indicate that it refers to an atom of the following residue.

Preceding the list of atom names, types, coordinates, and covalent connectivities, the residue library entry contains definitions of the rotatable torsion angles. Torsion angle definitions consist of a running index, the torsion angle name, three numbers that are irrelevant for CYANA (present for compatibility with other programs), and either four or five atom pointers. The first four atoms are those that define the torsion angle value, i.e. the torsion angle rotates the bond between the second and the third atom, and its value is 0 if the first and fourth atom are in *cis* position with respect to the rotatable bond. The first atom that is rotated by a change of the torsion angle is the one in the list of atoms that follows the third atom in the torsion angle definition. The fifth atom, if present, specifies the last atom that is rotated by a change of the torsion angle. This is the case for side-chain torsion angles. Backbone torsion angles have no fifth atom (i.e. the number 0 in Fig. 2a) in the torsion angle definition. They rotate the entire rest of the molecule, i.e. all atoms in the list of atoms that follow the third atom in the torsion angle definition as well as all atoms of all subsequent residues in the sequence.

Tree structure of torsion angles

The fast algorithm for torsion angle dynamics in CYANA requires that the molecule be represented as a tree structure consisting of a base and n rigid bodies, which are connected by n rotatable bonds (Güntert et al. 1997). The degrees of freedom are exclusively torsion angles, i.e. rotations about single bonds. Each rigid body is made up of one or more atoms for which the relative positions are invariable. The tree structure starts from the base, which is located at the N-terminus of the polypeptide chain, and terminates with "leaves" at the ends of the side-chains and at the C-terminus. The rigid bodies are numbered from 0 to n. The base has the number 0. Each other rigid body, with a number $k > 0$, has a single nearest neighbor with number $p(k) < k$ in the direction toward the base. The torsion angle between the rigid bodies $p(k)$ and k is denoted by θ_k. The conformation of the molecule is uniquely specified by the values of all torsion angles, $(\theta_1, \ldots, \theta_n)$.

To consistently implement the tree structure of the molecule, the torsion angles and atoms must be ordered such that two conditions are always fulfilled: (1) A change of a torsion angle must not affect the positions of the first, second, third, or fourth atom in any preceding torsion angle definition. (2) The set of atoms whose positions will be affected by a change of a torsion angle consists of all atoms following the third atom in the torsion angle definition up to the fifth atom in the torsion angle definition (or the end of the main chain for backbone torsion angles).

Fig. 1 The PDB Chemical Component Dictionary entry MND (*N*-methyl-D-asparagine) in mmCIF format. For brevity, some lines that are irrelevant for CYLIB have been omitted

```
data_MND
#
_chem_comp.id                                    MND
_chem_comp.name                                  N-methyl-D-asparagine
_chem_comp.type                                  "D-peptide linking"
_chem_comp.pdbx_type                             ATOMP
_chem_comp.formula                               "C5 H10 N2 O3"
_chem_comp.pdbx_formal_charge                    0
_chem_comp.pdbx_initial_date                     2008-06-02
_chem_comp.pdbx_modified_date                    2011-06-04
_chem_comp.pdbx_release_status                   REL
_chem_comp.formula_weight                        146.144
_chem_comp.three_letter_code                     MND
_chem_comp.pdbx_model_coordinates_missing_flag   N
_chem_comp.pdbx_ideal_coordinates_details        Corina
_chem_comp.pdbx_ideal_coordinates_missing_flag   N
_chem_comp.pdbx_model_coordinates_db_code        2RPL
_chem_comp.pdbx_processing_site                  PDBJ
#
loop_
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.alt_atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.charge
_chem_comp_atom.pdbx_align
_chem_comp_atom.pdbx_aromatic_flag
_chem_comp_atom.pdbx_leaving_atom_flag
_chem_comp_atom.pdbx_stereo_config
_chem_comp_atom.model_Cartn_x
_chem_comp_atom.model_Cartn_y
_chem_comp_atom.model_Cartn_z
_chem_comp_atom.pdbx_model_Cartn_x_ideal
_chem_comp_atom.pdbx_model_Cartn_y_ideal
_chem_comp_atom.pdbx_model_Cartn_z_ideal
_chem_comp_atom.pdbx_component_atom_id
_chem_comp_atom.pdbx_component_comp_id
_chem_comp_atom.pdbx_ordinal
MND N    N    N 0 1 N N N 8.200 1.610  -0.920 -0.787 1.699  -0.097 N    MND 1
MND CA   CA   C 0 1 N N R 8.323 1.545  -2.363 -0.893 0.301  0.341  CA   MND 2
MND CB   CB   C 0 1 N N N 7.978 2.897  -3.040 0.155  -0.542 -0.388 CB   MND 3
MND CG   CG   C 0 1 N N N 7.111 3.926  -2.291 1.534  -0.098 0.026  CG   MND 4
MND OD1  OD1  O 0 1 N N N 6.570 3.701  -1.210 1.666  0.804  0.826  OD1  MND 5
MND ND2  ND2  N 0 1 N N N 7.233 5.204  -2.674 2.621  -0.704 -0.492 ND2  MND 6
MND CE2  CE2  C 0 1 N N N 6.323 6.272  -2.283 3.962  -0.272 -0.090 CE2  MND 7
MND C    C    C 0 1 N N N 7.463 0.403  -2.892 -2.270 -0.223 0.022  C    MND 8
MND O    O    O 0 1 N N N 6.246 0.431  -2.716 -2.948 0.331  -0.811 O    MND 9
MND OXT  OXT  O 0 1 N Y N 8.076 -0.651 -3.447 -2.742 -1.303 0.663  OXT  MND 10
MND H    H    H 0 1 N N N 8.431 0.721  -0.525 -0.942 1.780  -1.090 H    MND 11
MND H2   H2   H 0 1 N Y N 8.824 2.305  -0.563 -1.425 2.288  0.417  H2   MND 12
MND HA   HA   H 0 1 N N N 9.374 1.343  -2.618 -0.722 0.243  1.416  HA   MND 13
MND HB2  HB2  H 0 1 N N N 8.939 3.394  -3.241 0.019  -1.593 -0.131 HB2  MND 14
MND HB3  HB3  H 0 1 N N N 7.431 2.648  -3.962 0.040  -0.414 -1.464 HB3  MND 15
MND HD2  HD2  H 0 1 N N N 8.001 5.441  -3.269 2.516  -1.425 -1.132 HD2  MND 16
MND HE21 HE21 H 0 0 N N N 6.653 7.219  -2.735 4.098  0.779  -0.347 HE21 MND 17
MND HE22 HE22 H 0 0 N N N 5.307 6.034  -2.631 4.077  -0.400 0.987  HE22 MND 18
MND HE23 HE23 H 0 0 N N N 6.322 6.370  -1.187 4.708  -0.873 -0.609 HE23 MND 19
MND HXT  HXT  H 0 1 N Y N 7.432 -1.317 -3.659 -3.630 -1.602 0.425  HXT  MND 20
#
loop_
_chem_comp_bond.comp_id
_chem_comp_bond.atom_id_1
_chem_comp_bond.atom_id_2
_chem_comp_bond.value_order
_chem_comp_bond.pdbx_aromatic_flag
_chem_comp_bond.pdbx_stereo_config
_chem_comp_bond.pdbx_ordinal
MND N   CA   SING N N 1
MND CA  CB   SING N N 2
MND CA  C    SING N N 3
MND CB  CG   SING N N 4
MND CG  OD1  DOUB N N 5
MND CG  ND2  SING N N 6
MND ND2 CE2  SING N N 7
MND C   O    DOUB N N 8
MND C   OXT  SING N N 9
MND N   H    SING N N 10
MND N   H2   SING N N 11
MND CA  HA   SING N N 12
MND CB  HB2  SING N N 13
MND CB  HB3  SING N N 14
MND ND2 HD2  SING N N 15
MND CE2 HE21 SING N N 16
MND CE2 HE22 SING N N 17
MND CE2 HE23 SING N N 18
MND OXT HXT  SING N N 19
```

a

```
RESIDUE   MND      7   22    3   21
   1 OMEGA    0    0    0.0000    2     1     3     4     0
   2 PHI      0    0    0.0000    1     3     5    20     0
   3 CHI1     0    0    0.0000    3     5     6     7    18
   4 CHI2     0    0    0.0000    5     6     7     8    15
   5 CHI3     0    0    0.0000    6     7     9    10    15
   6 CHI4     0    0    0.0000    7     9    10    11    14
   7 PSI      0    0    0.0000    3     5    20    22     0
   1 C    C_BYL    0    0.0000    0.3312    2.3979    0.0689    2    3    0    0    0
   2 O    O_BYL    0    0.0000    1.3414    1.9311    0.5929    1    0    0    0    0
   3 N    N_AMI    0    0.0000   -0.7870    1.6990   -0.0970    1    4    5    0    0
   4 H    H_AMI    0    0.0000   -1.5596    2.1268   -0.5219    3    0    0    0    0
   5 CA   C_ALI    0    0.0000   -0.8930    0.3010    0.3410    3    6   19   20    0
   6 CB   C_ALI    0    0.0000    0.1550   -0.5420   -0.3880    5    7   16   17    0
   7 CG   C_BYL    0    0.0000    1.5340   -0.0980    0.0260    6    8    9    0    0
   8 OD1  O_BYL    0    0.0000    1.6660    0.8040    0.8260    7    0    0    0    0
   9 ND2  N_AMO    0    0.0000    2.6210   -0.7040   -0.4920    7   10   15    0    0
  10 CE2  C_ALI    0    0.0000    3.9620   -0.2720   -0.0900    9   11   12   13    0
  11 HE21 H_ALI    0    0.0000    4.0980    0.7790   -0.3470   10    0    0    0   14
  12 HE22 H_ALI    0    0.0000    4.0770   -0.4000    0.9870   10    0    0    0   14
  13 HE23 H_ALI    0    0.0000    4.7080   -0.8730   -0.6090   10    0    0    0   14
  14 QE2  PSEUD    0    0.0000    4.2943   -0.1647    0.0103    0    0    0    0    0
  15 HD2  H_AMI    0    0.0000    2.5160   -1.4250   -1.1320    9    0    0    0    0
  16 HB2  H_ALI    0    0.0000    0.0190   -1.5930   -0.1310    6    0    0    0   18
  17 HB3  H_ALI    0    0.0000    0.0400   -0.4140   -1.4640    6    0    0    0   18
  18 QB   PSEUD    0    0.0000    0.0295   -1.0035   -0.7975    0    0    0    0    0
  19 HA   H_ALI    0    0.0000   -0.7220    0.2430    1.4160    5    0    0    0    0
  20 C    C_BYL    0    0.0000   -2.2700   -0.2230    0.0220    5   21   22    0    0
  21 O    O_BYL    0    0.0000   -3.1060    0.4948   -0.5246   20    0    0    0    0
  22 N    N_AMI    0    0.0000   -2.5117   -1.4842    0.3643   20    0    0    0    0
```

b

```
RESIDUE   MND      7   22    3   21
   1 OMEGA    0    0    0.0000 -O    -C    N     H
   2 PHI      0    0    0.0000 -C    N     CA    C
   3 CHI1     0    0    0.0000 N     CA    CB    CG    QB
   4 CHI2     0    0    0.0000 CA    CB    CG    OD1   HD2
   5 CHI3     0    0    0.0000 CB    CG    ND2   CE2   HD2
   6 CHI4     0    0    0.0000 CG    ND2   CE2   HE21  QE2
   7 PSI      0    0    0.0000 N     CA    C     +N
   1 C    C_BYL    0    0.0000    0.0000    0.0000    0.0000 -O    N
   2 O    O_BYL    0    0.0000   -0.6699    0.0000   -1.0316 -C
   3 N    N_AMI    0    0.0000    1.3290    0.0000    0.0000 -C    H     CA
   4 H    H_AMI    0    0.0000    1.8071   -0.0000    0.8555  N
   5 CA   C_ALI    0    0.0000    2.0987    0.0000   -1.2510  N     CB    HA    C
   6 CB   C_ALI    0    0.0000    1.7513   -1.2492   -2.0628  CA    CG    HB2   HB3
   7 CG   C_BYL    0    0.0000    0.3059   -1.1892   -2.4840  CB    OD1   ND2
   8 OD1  O_BYL    0    0.0000   -0.3794   -0.2402   -2.1665  CG
   9 ND2  N_AMO    0    0.0000   -0.2253   -2.1898   -3.2146  CG    CE2   HD2
  10 CE2  C_ALI    0    0.0000   -1.6310   -2.1319   -3.6237  ND2   HE21  HE22  HE23
  11 HE21 H_ALI    0    0.0000   -2.2660   -2.0854   -2.7385  CE2   -     -     -     QE2
  12 HE22 H_ALI    0    0.0000   -1.7948   -1.2435   -4.2349  CE2   -     -     -     QE2
  13 HE23 H_ALI    0    0.0000   -1.8778   -3.0215   -4.2023  CE2   -     -     -     QE2
  14 QE2  PSEUD    0    0.0000   -1.9795   -2.1168   -3.7252
  15 HD2  H_AMI    0    0.0000    0.3220   -2.9490   -3.4684  ND2
  16 HB2  H_ALI    0    0.0000    2.3863   -1.2957   -2.9481  CB    -     -     -     QB
  17 HB3  H_ALI    0    0.0000    1.9150   -2.1367   -1.4521  CB    -     -     -     QB
  18 QB   PSEUD    0    0.0000    2.1507   -1.7162   -2.2001
  19 HA   H_ALI    0    0.0000    1.8512    0.8898   -1.8300  CA
  20 C    C_BYL    0    0.0000    3.5726    0.0000   -0.9348  CA    O     +N
  21 O    O_BYL    0    0.0000    3.9668   -0.0000    0.2304  C
  22 N    N_AMI    0    0.0000    4.3965   -0.0000   -1.9776  C
```

Fig. 2 The CYANA residue library entry MND that was produced by CYLIB from the PDB Chemical Component Dictionary entry shown in Fig. 1. **a** Format using numeric atom pointers in torsion angle definitions and covalent connectivities. **b** The same entry in the CYANA format using atom names in torsion angle definitions and covalent connectivities

Program parameters

The CYLIB program has the following command line options, by which the user can modify the way that the algorithm will work:

-f *file* Read *file* as an input PDB Chemical Component Dictionary mmCIF file

-aa The molecule is an amino acid, i.e. overlap atoms will be added

-n Add overlap atoms only to the N-terminus of the molecule

-c Add overlap atoms only to the C-terminus of the molecule

-fba *atom* Take the given *atom* as the first atom of the backbone

-lba *atom* Take the given *atom* as the last atom of the backbone

-sc Treat all rings as rigid

-nic Use non-ideal Cartesian coordinates. (PDB Chemical Component Dictionary files contain two sets of coordinates, "ideal" and "non-ideal")

-o *file* Write the output CYANA residue entry to *file*
Default is the name of the input file, but with extension '.lib'

-np Do not add pseudo atoms to the structure

-info Print details of the running program to the screen

-debug Print extensive details and variable values to the screen

-help Print this list of program options to the screen

Implementation

CYLIB is implemented in the Fortran programming language in order to be compatible with the CYANA software. An object oriented programming approach was followed in the implementation by encapsulating data types and subroutines. A Unified Modeling Language (UML) class diagram of the algorithm is given in Fig. 3. Five classes are used for the mmCIF chemical component entry: ChemicalComponent, ChemicalComponentAtom, ChemicalComponentBond, ChemicalComponentDescriptor, and ChemicalComponentIdentifier. The values of an mmCIF chemical component entry are read from a file and populate the corresponding attributes of these classes. Similarly, there are two classes to represent the CYANA structure of the same molecule, i.e. the CyanaTorsionAngle and CyanaAtom classes. CyanaResidue is the class that collects all data for a CYANA residue entry by using the composition method. These classes contain the standard structure description of the mmCIF and CYANA formats.

Overview

To obtain a new CYANA residue library entry from a PDB chemical component dictionary entry, the CYLIB program has to resolve, in addition to straightforward file format conversion, the following non-trivial issues:

1. Overlap atoms: For amino acid-type residues that form part of a polypeptide chain, find or create three atoms at the N- and C-terminus that are required by CYANA to link the residue to its neighbors. This step is not necessary for individual molecules.
2. Backbone identification: Identify the backbone of the molecule. In case of a non-standard amino acid, the peptide backbone must be found; for other molecules the choice is in general not unique.
3. Ring structures: Identify rigid and flexible ring structures in the molecule. The former are kept rigid by not defining any torsion angles within them, whereas the latter can be made flexible by defining torsion angles within them (see below).
4. Atom order: Sort the atoms such that a tree structure of torsion angles can be imposed on the molecule.
5. Pseudo atoms: Add pseudo atoms, which are used as reference points for experimental NMR restraints in CYANA.
6. Cartesian coordinates: Choose which Cartesian coordinates to use from the PDB chemical component dictionary entry, and calculate the coordinates for overlap atoms and pseudo atoms that have been added to the entry.
7. Torsion angle definitions: Define the torsion angles such that the molecule obtains a consistent tree structure.
8. Atom types: Choose the correct CYANA atom types for all atoms in the molecule.

In the following sections these steps are presented in more detail.

Overlap atoms

The CYANA residue library defines individual residues. In the library they are not explicitly bound to other molecule/residues. To build a chain-like macromolecule such as a protein or DNA/RNA, its constituent residues are connected according to a given, specific sequence. To enable fast molecular dynamics simulation in torsion angle space in CYANA, the entire molecular system, which may comprise several molecules, must be represented as a single tree structure with the torsion angles as the only degrees of freedom (see above). Therefore, multiple molecules, e.g. a protein–ligand complex or a multimeric protein, are formally connected using sterically "invisible" linker residues into a single chain.

To link a residue to its predecessor in the sequence, three atoms of the residue are superimposed onto the corresponding three atoms of the preceding residue. To make this possible, interior residues in a chain contain three atoms, the so-called overlap atoms, twice. For amino acid residues, the overlap atoms are the atoms C, O, N of the

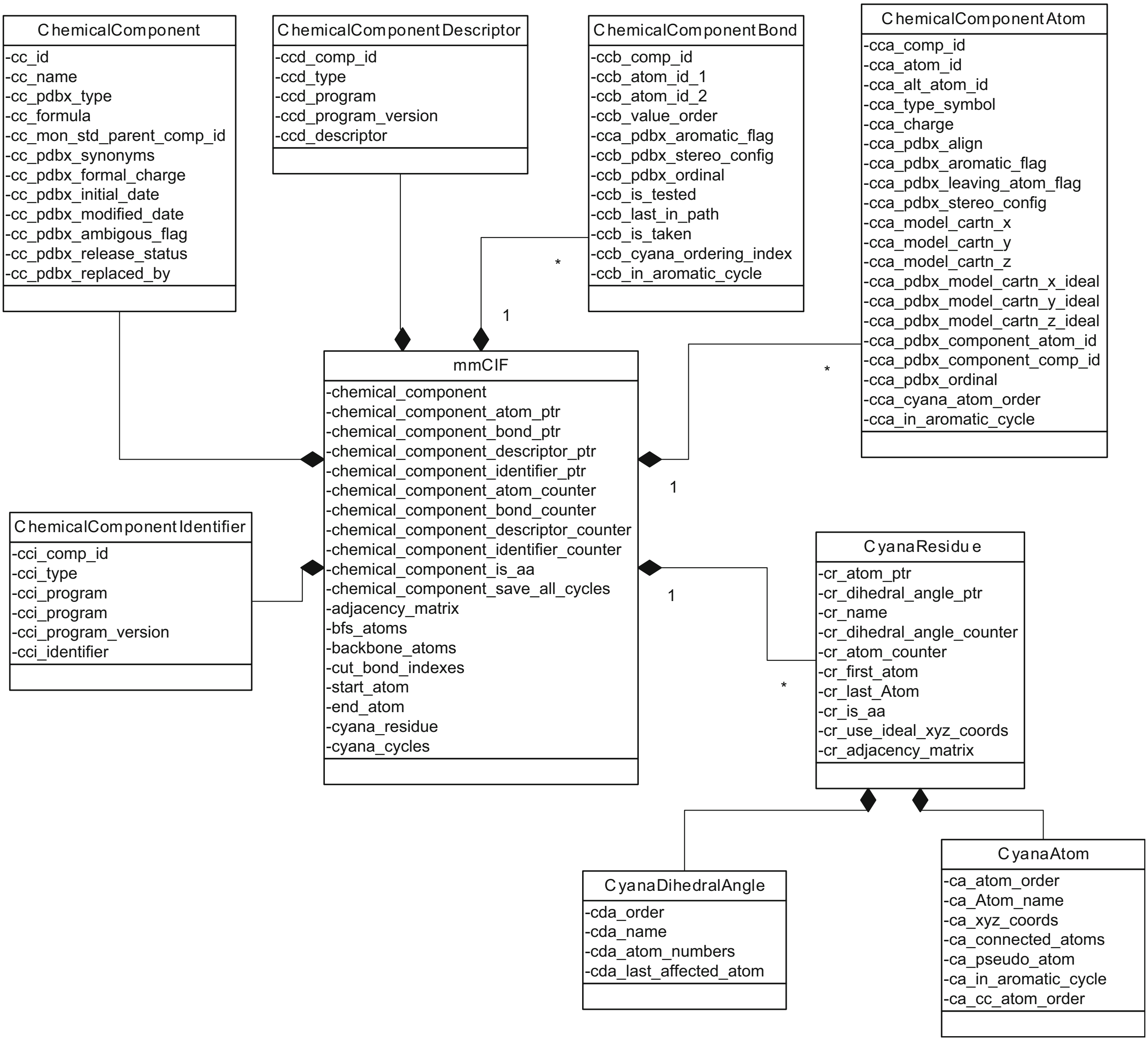

Fig. 3 Unified Modelling Language (UML) class diagram of the CYLIB program

backbone peptide group, which are present in a CYANA residue library entry at the N-terminal end (to link to the preceding residue) and at the C-terminal end of the residue (to link to the following residue). The N-terminal overlap atoms are always the first three atoms of the residue library entry, whereas the C-terminal overlap atoms can occur anywhere after the first three atoms in the list of atoms.

For instance, assuming that an amino acid-like residue to be used at sequence position i in a protein is going to be added to the CYANA residue library, the carbon (C) and oxygen (O) atoms of the backbone carboxyl group of residue $i-1$ and the nitrogen (N) atom of the backbone amide group of residue $i+1$ are needed in order to covalently link the residue to other residues in CYANA (Fig. 4). However, these additionally needed atoms are in general not present in the input data from the PDB chemical component dictionary entry, which, on the other hand often contains a second amide hydrogen (e.g. HXT) and a second carbonyl oxygen (e.g. OXT). During the conversion the first hydrogen (HXT) atom of the residue i is removed and the C and O atoms of the residue $i-1$ are added to the N-terminus of the residue i. Likewise, the last OXT atom of the carboxyl group of the residue i is removed and the N atom of the residue $i+1$ is added.

Using the current version of CYANA it is no longer necessary to have explicit overlap atoms for molecules that are not covalently bound to its neighbors. Instead, the first three atoms of the molecule are implicitly taken as the overlap atoms with the preceding linker residue. The names of the three first atoms of the molecule do not have to

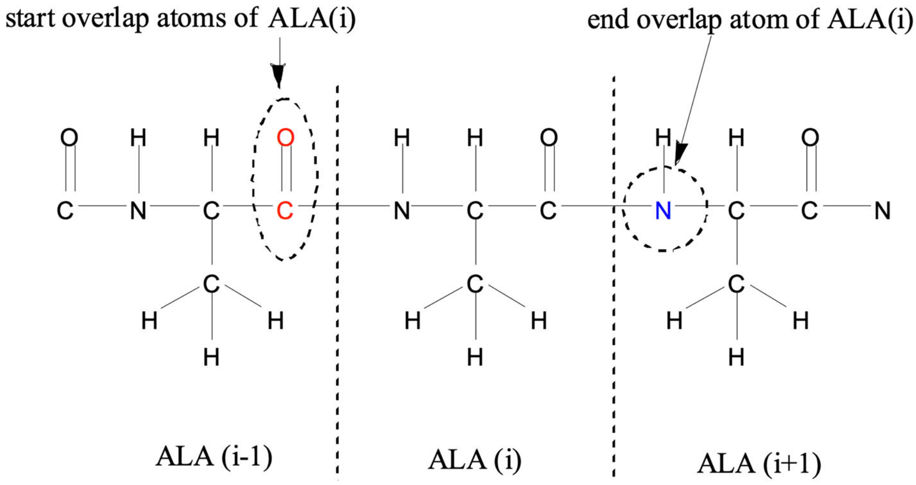

Fig. 4 Start (*red*) and end (*blue*) overlap atoms of an Ala residue at position *i* in a polypeptide chain

match with names of the dummy atoms of the preceding linker residue.

Depending on the input options (see above), the overlap atoms are added to both termini of amino acid-like residues (option -aa; for residues in the interior of a polypeptide chain), only to the N-terminus (option −n; for residues at the C-terminus of a polypeptide chain), only to the C-terminus (-c; for residues at the N-terminus of a polypeptide chain), or not at all (default; for individual molecules).

Backbone identification

The tree structure of torsion angles consists of a backbone that runs through all residues/molecules in the molecular system under consideration and (in general short) side-chains that are attached to the backbone. The torsion angle definitions and the ordering of the atoms depend on the backbone of the molecule, so the backbone has to be determined or chosen first. Two different methods have been implemented in CYLIB for determining the backbone: The first method works without receiving any backbone-related information from the user. This function works for small molecules; however, suboptimal results may be obtained for larger molecules. Therefore, a second approach was developed, for which the start and end atoms of the backbone within the residue/molecule are received from the user and the algorithm determines the backbone as the shortest path of covalent bonds between these two atoms.

The chemical component is represented by an undirected graph, where the vertices are the atoms and the edges are the bonds between these atoms. Hydrogen atoms have only one covalent bond and can therefore not occur in the backbone of the molecule. They are removed from the graph. The graph is saved in a two-dimensional “adjacency matrix”.

A breadth-first search algorithm (Cormen et al. 1990) is applied to this matrix. The pseudo code of this algorithm is given in Fig. 5. It takes the adjacency matrix as input and creates a tree structure as output. The root vertex is the start atom of the backbone. At the beginning of the algorithm, all vertices (except the root vertex) are initialized with these values: color = white, distance = ∞, ancestor = none (lines 1–4 in Fig. 5). The root vertex is initialized with color = gray, distance = 0, and ancestor = none (lines 5–7 in Fig. 5) and enqueued into the *Q* queue. Then the main part of the algorithm starts, which is given in lines 10–18 of the pseudo code. The idea is to start with one vertex, which is the root in our algorithm, and build a tree by expanding all of the edges of an already existing subtree. This process starts with pulling one entry, vertex *u*, from *Q*. For each vertex *v* that is connected to the vertex *u*, the distance of *v* is increased by one and the ancestor of *v* is set to *u* if the color of *u* is white. After these steps, the vertex *v* is enqueued into *Q*. If all of the connected vertices of *u* are examined, then its color is set to black. This process is repeated until the queue is empty. The result of the algorithm is a tree, in which each vertex has its ancestor and its distance from the root of the tree. These distances are the shortest distances between the vertices and the root of the tree structure. The path between the root *s* of the tree, and any vertex *v* that is reachable from the root can be found by using the ancestor value of the vertex *v*. The first edge of the path is the edge between the vertex *v* and its ancestor *a*. After that, the edge between the vertex *a* and its ancestor is added to the path. This process continues until the ancestor is the root of the tree. The resulting path is the shortest path between the vertex *v* and the root *s*.

Ring structures

The chemical properties of the bonds in a molecular ring are different from the ones that do not belong to a ring. For example, if there is an aromatic bond in a ring, then the whole ring is handled as a rigid structure without internal

```
BFS(G, s)

1  for each vertex $u \in V[G] - \{s\}$
2      do color[u] $\leftarrow$ WHITE
3         d[u] $\leftarrow \infty$
4         $\pi$[u] $\leftarrow$ NIL
5  color[s] $\leftarrow$ GRAY
6  d[s] $\leftarrow$ 0
7  $\pi$[s] $\leftarrow$ NIL
8  Q $\leftarrow \emptyset$
9  ENQUEUE(Q, s)
10 while Q $\neq \emptyset$
11     do u $\leftarrow$ DEQUEUE(Q)
12        for each v $\in$ Adj[u]
13            do if color[v] = WHITE
14                then color[v] $\leftarrow$ GRAY
15                     d[v] $\leftarrow$ d[u] + 1
16                     $\pi$[v] $\leftarrow$ u
17                     ENQUEUE(Q, v)
18        color[u] $\leftarrow$ BLACK
```

Fig. 5 Pseudo code of the breadth-first search algorithm that is used in the CYLIB program. The variables used in the pseudo code are: *V*[*G*], vertices set of the graph *G*; *d*[*u*], distance of the vertex *u* from the source vertex; $\pi[u]$, the ancestor of the vertex *u*; Adj[*u*], adjacent vertices of the vertex *u*; color[*u*]: the color of the vertex *u*; *Q*, first-in first-out queue structure

degrees of freedom. It is thus necessary to detect the ring structures in the molecule.

In order to find the rings in the molecule, the aforementioned breadth first search algorithm is used again. The coloring functionality of this algorithm is used to detect whether the current vertex has already been visited. Line 13 of the pseudo code in Fig. 5 shows the color of the vertex, which will be analyzed at that step. If the color of this vertex is gray, then this vertex has been analyzed already by a different path, which means that there are at least two different ways leading to the same vertex from the root of the tree. Hence the vertex forms part of a ring structure.

The whole graph is analyzed with this approach and the ring structures are detected as a result. Atoms in ring structures are then treated specially when establishing the correct atom order in the next step.

Atom order

The most important challenge in the creation of new CYANA residue library entries is to produce results that are compatible with the CYANA tree structure of torsion angles (Güntert et al. 1991, 1997). There are five atom indices (or names) in the torsion angle declarations of CYANA residue library entries. These atom indices are important because they represent the atoms whose positions will be affected by a rotation in that torsion angle according to the two aforementioned rules: (1) A rotation of a torsion angle must not affect the positions of the first, second, third and fourth atoms in any preceding torsion angle definition. (2) A rotation of a torsion angle must change only the atoms whose indices are between the third and the fifth atom of the torsion angle definition. For the backbone torsion angles, the fifth atom is absent in the declaration indicating that all atoms until the end of the main chain will be affected by a rotation of a backbone torsion angle. CYLIB must change the order of the atoms of the chemical component in order to fulfill these rules.

Before applying these rules, the rings of the molecule are identified. If the ring contains aromatic bonds (as defined in the covalent bond information of the PDB Chemical Component Dictionary entry), then the ring is treated as a rigid structure. Otherwise it is a potentially flexible ring whose bonds can be rotated, unless the user has explicitly chosen to keep all ring structures rigid with the -sc command line option (see above). In order to allow rotatable bonds in a ring in a way that is compatible with the, in principle, linear tree structure of the molecule in CYANA, one bond of the ring is temporarily removed (see Fig. 8 below for an example). This bond will be closed by distance restraints during the CYANA calculation. To decide which bond of a flexible ring should be removed, the atoms of a ring that belong to the backbone of the molecule are examined first. If the ring involves at least two backbone atoms, then the last atom of this ring on the backbone, atom *a*, is determined, and the ring atom that does not lie on the backbone, but has a bond to the atom *a* is selected as atom *b*. If atom *b* has at least one bond that does not belong to the ring structure, then the bond between atom *a* and atom *b* is removed. Otherwise, the bond between atom *b* and the neighboring non-backbone atom is removed. If only one atom of the ring belongs to the backbone, then a bond of the ring opposite to that atom is removed. In the present version of CYLIB a ring is treated as rigid as soon as it contains at least one aromatic bond. This may result in certain rings being kept rigid that contain in fact also rotatable bonds. In a future version of the algorithm, this restriction may be lifted.

After these steps have been completed, the ordering of the atoms is achieved by using a stack data structure. The pseudo code of this method is given in Fig. 6.

Pseudo atoms

Pseudo atoms are used to represent groups of hydrogen atoms or methyl groups that are connected to the same heavy (non-hydrogen) atom, or the two aromatic hydrogens at symmetric positions on an aromatic ring, e.g. in the amino acid residues Phe and Tyr (Fig. 7). They must be placed at the center of the positions of the atoms they represent. Within CYANA, NMR restraints may either be referred to the pseudo atom position, or the pseudo atoms may be used merely as references for a group of equivalent atoms and restraints involving a pseudo atom will be expanded into ambiguous distance restraints within CYANA structure calculations. Pseudo atoms that directly represent hydrogen atoms are called first-level pseudo atoms (Fig. 7a). There are also second-level pseudo atoms that represent multiple first-level pseudo atoms, and thus indirectly a larger group of hydrogens (Fig. 7b). For instance, in the amino acid Val there are two first-level pseudo atoms to represent the two isopropyl methyl groups, and one second-level pseudo atom that represents the two first-level pseudo atoms, i.e. both methyl groups. The whole graph is examined and first- and second-level pseudo atoms are added to the structure, wherever applicable.

Cartesian coordinates

The PDB Chemical Component Dictionary entries contain two sets of Cartesian coordinates for the atoms (Westbrook et al. 2015): experimental model coordinates, which are taken from one of the PDB macromolecular structure files that contains an experimentally determined structure of the compound, and computed ideal coordinates, which have, in most cases, been determined with the CORINA software (Sadowski et al. 1994). CYLIB uses the ideal coordinates, if available, unless the user selects to use the experimental model coordinates by setting the -nic command line option (see above). The coordinates of overlap atoms and pseudo atoms that are not present in the input PDB Chemical Component Dictionary entry have to be determined.

```
Order_Cyana_Atoms (G)

1  i ← 1
2  for each atom a ∈ backbone(G)
3      order[a] ← i
4      i ++
5      S ← {a}
6      if there ∃ atom b, bonded to a
7          order[b] ← i
8          i ++
9          S ← {b}
10         a ← b
11     else
12         b ← S.pull
13         goto line number 6
14     end if
15 end for
```

Fig. 6 Pseudo code of the algorithm implemented in CYLIB for ordering the atoms according to the CYANA tree structure of torsion angles

Given three atoms at positions $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, the position $\boldsymbol{d}$ of a fourth atom that is attached to the atom at position $\boldsymbol{c}$ with given bond length $l = |\boldsymbol{d} - \boldsymbol{c}|$, bond angle τ (defined by the points $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{d}$), and torsion angle ϕ (defined by the points $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{d}$) can be calculated as follows:

$$\boldsymbol{d} = \boldsymbol{c} + (1 - \cos\phi)(\boldsymbol{e} \cdot \boldsymbol{f})\boldsymbol{e} + \cos\phi \boldsymbol{f} + \sin\phi\,(\boldsymbol{e} \times \boldsymbol{f})$$

The auxiliary three-dimensional vectors $\boldsymbol{e}$ and $\boldsymbol{f}$ are given by

$$\boldsymbol{e} = \frac{\boldsymbol{c} - \boldsymbol{b}}{|\boldsymbol{c} - \boldsymbol{b}|}$$

$$\boldsymbol{f} = -l\cos\tau\,\boldsymbol{e} + l\sin\tau\,\frac{(\boldsymbol{e} \times (\boldsymbol{a} - \boldsymbol{b})) \times \boldsymbol{e}}{|(\boldsymbol{e} \times (\boldsymbol{a} - \boldsymbol{b})) \times \boldsymbol{e}|}$$

To attach the first overlap atom at the N-terminus, i.e. the backbone carbonyl carbon C of the preceding residue, the atoms $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ are, respectively, C, CA, N of the current residue, $l = 1.329$ Å, $\tau = 121.6°$, and $\phi = \phi_0 + 180°$, where ϕ_0 denotes the value of the torsion angle formed by the atoms H, N, CA, C. To attach the second overlap atom, i.e. the backbone carbonyl oxygen O of the preceding residue, the atoms $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ are CA, N of the current residue and the C attached in previous step, $l = 1.230$ Å, $\tau = 120.8°$, and $\phi = 0°$. To attach the last overlap atom at the C-terminus, i.e. the backbone amide nitrogen N of the next residue, the atoms $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ are, respectively, N, CA, C of the current residue, $l = 1.329$ Å, $\tau = 116.2°$, and $\phi = \phi_0 + 180°$, where ϕ_0 denotes the value of the torsion angle formed by the atoms N, CA, C, O.

The coordinates of pseudo atoms are set to the average of the coordinates of the atoms that they represent.

Torsion angle definitions

Torsion angle definitions in CYANA describe which bonds can be rotated and which atoms are affected by a change of the torsion angle. CYANA uses torsion angles as the degrees of freedom of the system, e.g. for target function minimization and molecular dynamics simulation. The torsion angles are crucial for the structure determination algorithm of CYANA but not defined in the PDB Chemical Component Dictionary. Hence, CYLIB must create the torsion angle definitions by analyzing the connectivity graph of the molecule and the information on covalent bond types.

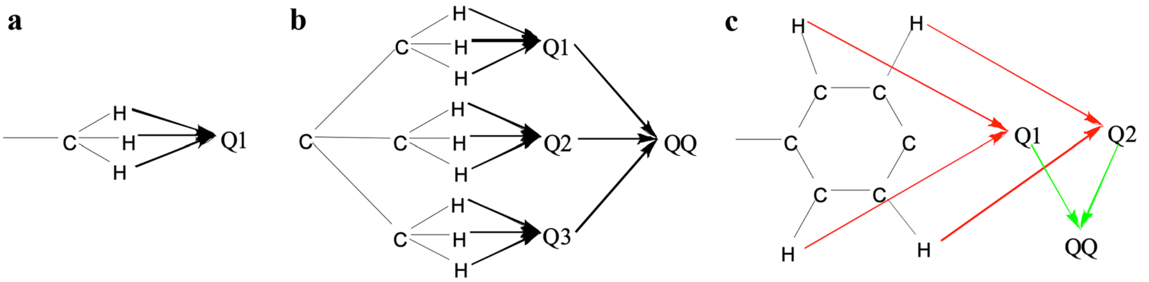

Fig. 7 Examples of pseudo atoms. **a** First layer pseudo atom (Q1) for a methyl group. **b** Second layer pseudo atom (QQ) representing multiple equivalent methyl groups. Each methyl group is represented by first layer pseudo atom (Q1, Q2, Q3). **c** First-(Q1, Q2) and second-level (QQ) pseudo atoms for a symmetric six-membered aromatic ring. The actual geometric position of the pseudo atoms is always in the center of the atoms that they represent

First, the non-rotatable covalent bonds are detected. Bonds with at least one of the following properties cannot be rotated: (1) Bonds involving a hydrogen atom. (2) Bonds in aromatic rings. (3) Double and triple bonds. All other covalent bonds are defined as rotatable bonds. Metal coordination is given as "single bonds" in most PDB Chemical Component Dictionary entries, and handled as such in the present version of CYLIB.

The definition of a torsion angle in the CYANA residue library format is composed of its order number, torsion angle name, four atoms defining the torsion angle, and a fifth atom, which is the last atom whose position is affected by a rotation of the torsion angle (Fig. 2). The definition of the last affected atom depends on the location of the torsion angle in the molecule, i.e. whether it is in the backbone or in a side-chain. If the torsion angle is in the backbone, then all atoms following the third atom of the torsion angle definition will be affected by a rotation of the torsion angle. The last affected atom of a backbone torsion angle is set to zero to distinguish it from the side-chain torsion angles. On the other hand, a rotation of a side-chain torsion angle will change the location of all atoms following the third atom of the torsion angle definition until the explicitly specified last affected atom.

The torsion angles of the backbone are given the names PHIm, where $m = 1, 2, \ldots$ is an integer counter. Similarly the torsion angles of the side-chain(s) have the names CHIn, where $n = 1, 2, \ldots$ in another integer counter. There are special naming conventions for amino acid-like residues: For them, the first backbone torsion angle is called OMEGA, the second backbone torsion angle is called PHI, and the last backbone torsion angle is called PSI.

As explained above, one of the covalent bonds in an aliphatic ring may be removed in order to incorporate a flexible ring into the tree structure. After defining the rotatable torsion angles of the molecule, these bonds are restored in the list of covalent connectivities of the atoms.

Atom types

The atom types of the molecules are assigned depending on the atoms' neighboring atoms and bonds. Pseudo atoms have the atom type PSEUD. The atom types of the real atoms are set according to the first of the following rules that applies:

1. For hydrogen atoms: H_AMI, H_OXY, or H_SUL, if the atom has a bond to a nitrogen, oxygen, or sulphur atom, respectively. H_ARO, if the atom has an aromatic bond to a carbon atom. H_ALI, if the atom has a non-aromatic bond to a carbon atom. H_XXX, otherwise.
2. For carbon atoms: C_ALI, if it has four bonds. C_BYL, if it has three non-aromatic bonds. C_ARO if it has three bonds and at least one of them is aromatic. C_XXX, otherwise.
3. For nitrogen atoms: N_AMI, if it is located in the backbone. N_AMO, otherwise.
4. For oxygen atoms: O_BYL, if it has one bond. O_HYD, if it has two single bonds, and at least one of them is to a hydrogen. O_EST, if it has two single bonds, and none of them is to a hydrogen. O_XXX, otherwise.
5. For sulphur atoms: S_OXY, if it has one bond. S_RED, if it has two bonds. S_XXX, otherwise
6. For phosphorus, fluorine, chlorine, bromine, iodine, and metal atoms: P_ALI, FLUOR, CHLOR, BROM, IOD, METAL, respectively.
7. Otherwise: X_XXX.

It should be noted that CYANA does not use a "full" physical force field with Lenard–Jones and electrostatic potentials, and, because the program works in torsion angle space with fixed covalent geometry, it does not need energy terms to maintain the covalent geometry. Therefore, atom types are used in CYANA only to specify the chemical element, the atom radius for the steric repulsion, and the hydrogen bonding capabilities of an atom. Only atoms that differ in one of these properties have to be distinguished by a unique atom type. Therefore, the number of atom types needed in CYANA is significantly lower than in conventional molecular dynamics simulation programs.

Results and discussion

The CYLIB algorithm was applied to all entries in the PDB Chemical Component Dictionary in order to generate the corresponding CYANA residue library entries. As examples, we present in detail the conversion of a non-standard amino acid residue and of a general molecule, as well as statistics and a discussion on the conversion and use in CYANA structure calculations of all entries in the PDB Chemical Component Dictionary.

The non-standard amino acid pyrrolysine

In this section, the basic steps for the conversion of the non-standard amino acid pyrrolysine (PYH), which occurs, for instance, in the PDB macromolecular structure file 2ZCE (Yanagisawa et al. 2008), is briefly explained. The input mmCIF file PYH.cif is extracted from the PDB Chemical Component Dictionary. The CYLIB program is called with the command `cylib --aa PYH`.

Since the molecule is treated as an amino acid by the -aa option, the program assumes that the start and the end atoms of the backbone are called N and C. The algorithm adds the overlap atoms to the structure, determines the backbone of the molecule, and examines the structure for molecular rings. It identifies one aliphatic ring, and accordingly removes one of the bonds of this ring in order to make it flexible within the CYANA tree structure of torsion angles. The resulting structure of the molecule is shown in Fig. 8a. The software orders the atoms to be compatible with the tree structure and adds pseudo atoms to the structure. Figure 8b shows the final order of the atoms. The software then calculates the Cartesian coordinates of the overlap and pseudo atoms, and defines the rotatable torsion angles of the structure (Fig. 8c). The temporarily removed bond in the aliphatic ring is restored into the covalent connectivities. Finally, the CYANA residue library entry is saved in a file.

The TNF-alpha converting enzyme (TACE) inhibitor JMV 390

As an example for the conversion of a general (not amino acid) molecule, this section presents the conversion steps for the TNF-alpha converting enzyme (TACE) inhibitor JMV 390 (*N*-[(2*R*)-2-benzyl-4-(hydroxyamino)-4-oxobutanoyl]-L-isoleucyl-L-leucine), which occurs in the PDB macromolecular structure file 2FV9 (Ingram et al. 2006), and is available as PDB Chemical Component Dictionary entry 002. The CYLIB program is called with the command `cylib -fba C1 -lba C23 002`.

Several intermediate steps and the final result of the conversion with CYLIB are shown in Fig. 9. According to the user-specified command line parameters above, the first backbone atom is C1 and the last backbone atom is C23. The software determines the backbone as the shortest path between these atoms (Fig. 9a) and orders the atoms so that they will be consistent with the CYANA torsion angle tree structure (Fig. 9b). The ring in the structure is aromatic, and will therefore be kept rigid. Pseudo atoms are added to the structure (Fig. 9c). Then the bonds are analyzed and the Cartesian coordinates and covalent connectivities of the atoms are set. The rotatable torsion angles are defined (Fig. 9d). Finally, the program determines the atom types and writes the CYANA residue library entry into an output file.

Conversion of the entire PDB Chemical Component Dictionary into a CYANA residue library

The CYLIB algorithm was applied to convert all entries in PDB Chemical Component Dictionary into corresponding CYANA residue library entries. On April 16, 2015, the PDB Chemical Component Dictionary contained in total 19,706 molecules including standard and non-standard amino acids, small molecule ligands and solvent molecules (Westbrook et al. 2015). The entire PDB Chemical Component Dictionary can be downloaded from http://www.wwpdb.org/data/ccd as a single components.cif file that contains all of these entries one after another in mmCIF format (Bourne et al. 1997). For better handling, we split this file into individual mmCIF files such that each file comprises one molecule. Each of these individual files was submitted to CYLIB for conversion into a CYANA residue library entry file. CYLIB yielded a CYANA residue library entry file for 18,516 (94.0 %) out of all 19,706 input mmCIF files (Table 1). Some entries cannot be converted because the result could not be represented in the CYANA residue library format. This includes 204 entries that do not contain any covalent bonds (e.g. single metal ions; these can, however, be represented by the existing METAL entry in the standard CYANA residue library), 99 entries with atoms that have more than 4 covalent bonds, and 5 entries that do not represent a single molecule (multiple unconnected fragments/molecules). Some other input files may contain other, more complex inconsistencies that preclude a successful conversion. For the remaining entries, the

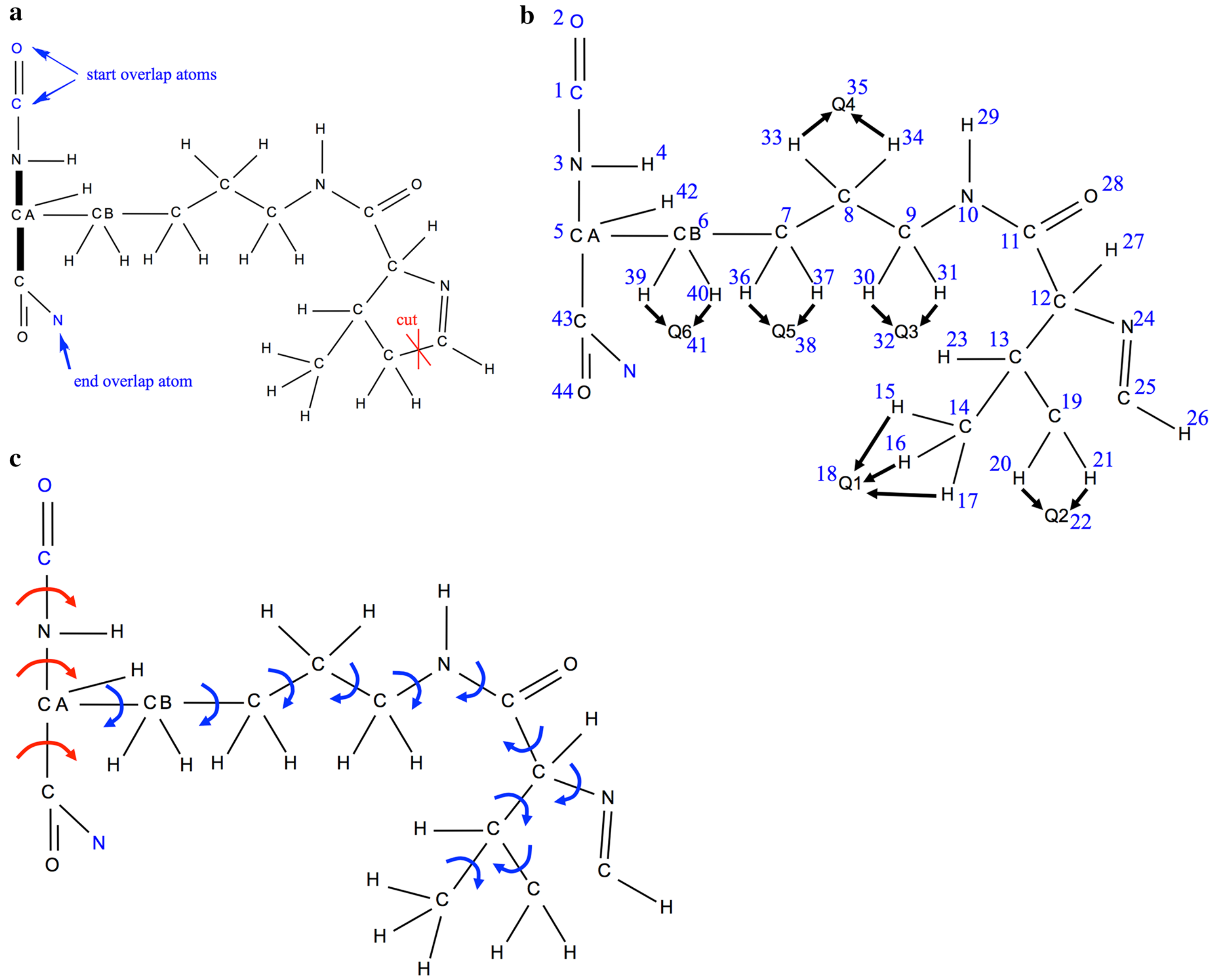

Fig. 8 Steps in the conversion of the PDB Chemical Component Dictionary entry PYH (the non-standard amino acid pyrrolysine) into a CYANA residue library entry. **a** Overlap atoms (*blue*), backbone identification (*thick black bonds*), and "cut" of aliphatic ring (*red*) to enable a linear (branched) tree structure of torsion angles. **b** Atom order (*blue numbers*) after the addition of pseudo atoms. **c** Rotatable torsion angles in the backbone (*red*) and in the side-chain (*blue*)

program could not accomplish the conversion although the input files do not contain obvious inconsistencies. The conversion with CYLIB may have failed for example because of the complexity of (especially aliphatic, flexible) ring structures. It should in principle be possible to handle many of these cases by future further development of the CYLIB algorithm.

Structure calculations with the CYANA residue library entries created by CYLIB

In order to evaluate the success of this conversion, we performed CYANA structure calculations with each successfully converted compound. To this end a CYANA sequence file containing the compound was created, and, after loading the corresponding residue library file, read into CYANA. A random structure of the molecule was generated by setting all rotatable torsion angles to random values. This initial structure was then subjected to a minimization of the CYANA target function using a maximal number of 1000 conjugate gradient minimization (Güntert et al. 1991) steps. Since no experimental NMR data is available for these compounds, the target function comprised only terms for the steric repulsion. Finally, the structure was saved as a PDB coordinate file. Conjugate gradient minimization was chosen for this test instead of the more efficient torsion angle dynamics algorithm because the former is more susceptible to inconsistencies in the tree structure of torsion angles that affect the calculation of the target function or its gradient with respect to torsion angles. Thus, incorrect residue library entries are likely to lead to premature termination of the conjugate

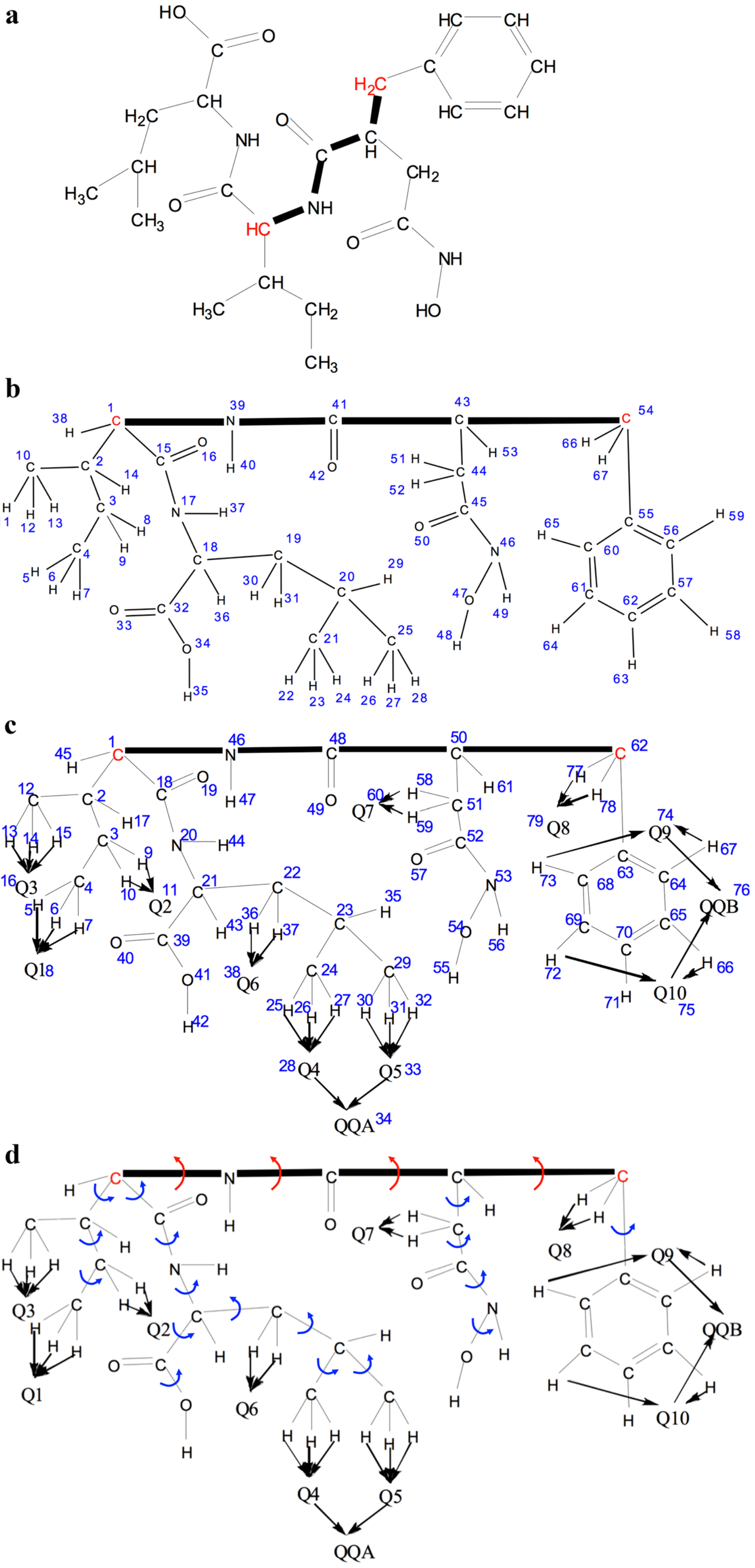

Fig. 9 Steps in the conversion of the PDB Chemical Component Dictionary entry 002 (*N*-[(2R)-2-benzyl-4-(hydroxyamino)-4-oxobutanoyl]-L-isoleucyl-L-leucine) into a CYANA residue library entry. **a** Backbone identification (*thick black bonds*) according to user-specified first and last backbone atoms (*red*). **b** Atom order (*blue numbers*) before the addition of pseudo atoms. The backbone is shown as the horizontal sequence of bonds. **c** Atom order (*blue numbers*) after the addition of pseudo atoms. **d** Rotatable torsion angles in the backbone (*red*) and in the side-chains (*blue*)

Table 1 Application of CYLIB to all PDB Chemical Component Dictionary entries

Quantity	Number	Percentage
All PDB Chemical Components Dictionary entries	19,706	100.0
Entries with no covalent bonds (e.g. single metal ions)	204	1.0
Entries with atoms having more than four covalent bonds	99	0.5
Entries that are not one connected molecule	5	0.03
CYANA residue library entries produced by CYLIB	18,516	94.0
Entries with completed CYANA structure calculations	18,037	91.5

gradient minimizer long before a (local) minimum, where the norm of the gradient is below a small tolerance value, has been reached. The CYANA commands to perform this test structure calculation are

```
cyanalib
read lib $f.lib append
read seq $f.seq
random
minimize 1000
write pdb $f.pdb
```

The cyanalib command reads the standard CYANA residue library, which contains the atom type definitions (and the standard amino acid and nucleotide residues). $f denotes the name of the current compound from the PDB Chemical Component Dictionary, possibly prefixed by ‘A’ if the name starts with a number because CYANA residue names must start with a letter. $f.lib is the name of the CYANA residue library file generated by CYLIB, $f.seq the name of the CYANA sequence file (that contains only one “residue” with the name of the current compound), and $f.pdb the name of the output structure file in PDB format.

The results of the test calculations are summarized in Table 1. 18,516 residue library entries were created fully automatically by CYLIB and tested in CYANA structure calculations. For 18,037 of these compounds the test structure calculation could be completed successfully by writing the output PDB structure file. This means that 91.5 % of the PDB Chemical Component Dictionary entries can be used in CYANA structure calculations without any further manual work. Some of the residue library files generated by CYLIB could not be read by CYANA. The most common reasons for this were: In 255 cases the molecule had no rotatable torsion angles (CYANA requires at least one rotatable torsion angle in the entire molecular system; this is only a problem if the molecule is calculated alone, as in this test), in 164 cases there was an inconsistency of pseudo atom pointers, and in 54 cases duplicate pseudo atom names were present. The conjugate gradient minimization stopped prematurely only in 5 cases. In 3 of these 5 cases, coordinate values were missing in the input mmCIF file for some of the atoms. Thus the CYLIB program produced in almost all cases CYANA residue library entries that can be used for CYANA structure calculations.

Conclusions

In this paper we have presented the CYLIB algorithm that automatically generates residue library entries for CYANA structure calculations. This algorithm is capable of converting any molecule definition in the PDB Chemical Component Dictionary into a CYANA residue library entry. These residues represent information about the atoms of the structure, the chemical bonds and the rotatable torsion angles, and are compatible with the optimized tree structure of the torsion angles in the CYANA program. By using the conversion software, externally maintained residue entries, like the ones of the PDB, can now be used easily in CYANA structure calculations.

The most important consequence of this work is that it greatly expands the range of application of the CYANA software package. Before this work started, only the 20 standard amino acid residues and the standard nucleotides of DNA/RNA were included in the standard CYANA residue library file, which means that only these residues could be used in CYANA without further manual processing work. If one needed to use another molecule in CYANA, then the user had to create the corresponding entry for the CYANA residue library manually, which is a cumbersome and potentially error-prone task, especially for complex molecules. Two main application areas can be envisaged for the CYLIB algorithm: First, the interactions of drug candidates with the target proteins can be examined readily with CYANA by using automatically produced residue library entries for the drug candidate molecules. Secondly, NMR structure calculations of peptides and proteins containing non-standard amino acids can be accomplished efficiently in CYANA.

While CYLIB achieves the automated generation of residue library entries for CYANA in the large majority of cases, some future enhancements could be implemented in order to improve the usability of this program. As presented in the previous chapter, a small number of PDB Chemical Component Dictionary entries could not be converted and some of the generated residue library entries could not be used in CYANA structure calculations. Further investigation of these exceptional cases is likely to lead to enhancements of the CYLIB algorithm that will allow their successful conversion. Another promising approach to extend the applicability of CYLIB and thus CYANA will be the use of other sources of input than the PDB Chemical Component Dictionary. In the implementation of the CYLIB algorithm a focus was put on separating the input steps, e.g. parsing the PDB Chemical Component Dictionary mmCIF format, from the conversion and the output. This separation of functionality will simplify the task of supporting other compound databases such as, for instance, the Cambridge Structural Database (Allen et al. 1979) or proprietary databases in the pharmaceutical industry, all of which use their own data representation formats. It should be straightforward to extend the program for reading other formats that provide the same information as the PDB Chemical Component Dictionary. On the other hand, there are also input formats containing less information, e.g. regarding the single/double/aromatic character of chemical bonds. The extension of CYLIB to such formats will require more effort to derive the missing information from the coordinates or connectivities.

Acknowledgments We gratefully acknowledge financial support by the Lichtenberg program of the Volkswagen Foundation and a Grant-in-Aid for Scientific Research of the Japan Society for the Promotion of Science (JSPS).

References

Alipanahi B, Gao X, Karakoc E, Donaldson L, Li M (2009) PICKY: a novel SVD-based NMR spectra peak picking method. Bioinformatics 25:i268–i275

Allen FH, Bellard S, Brice MD, Cartwright BA, Doubleday A, Higgs H, Hummelink T, Hummelink-Peters BG, Kennard O, Motherwell WDS, Rodgers JR, Watson DG (1979) The Cambridge crystallographic data centre: computer-based search, retrieval, analysis and display of information. Acta Crystallogr Sect B Struct Commun 35:2331–2339

Arkin MR, Wells JA (2004) Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. Nat Rev Drug Discov 3:301–317

Bahrami A, Assadi AH, Markley JL, Eghbalnia HR (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. PLoS Comp Biol 5:e1000307

Bardiaux B, Malliavin T, Nilges M (2012) ARIA for solution and solid-state NMR. Meth Mol Biol 831:453–483

Bartels C, Güntert P, Billeter M, Wüthrich K (1997) GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. J Comput Chem 18:139–149

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucl Acids Res 28:235–242

Bourne PE, Berman HM, McMahon B, Watenpaugh KD, Westbrook JD, Fitzgerald PMD (1997) Macromolecular crystallographic information file. Methods Enzymol 277:571–590

Cormen TH, Leiserson CE, Rivest RL (1990) Introduction to algorithms. MIT Press, Cambridge

Guerry P, Herrmann T (2011) Advances in automated NMR protein structure determination. Q Rev Biophys 44:257–309

Güntert P (2009) Automated structure determination from NMR spectra. Eur Biophys J 38:129–143

Güntert P, Buchner L (2015) Combined automated NOE assignment and structure calculation with CYANA. J Biomol NMR. doi:10.1007/s10858-015-9924-9

Güntert P, Braun W, Wüthrich K (1991) Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. J Mol Biol 217:517–530

Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 273:283–298

Hamada T, Matsunaga S, Fujiwara M, Fujita K, Hirota H, Schmucki R, Güntert P, Fusetani N (2010) Solution structure of polytheonamide B, a highly cytotoxic non-ribosomal polypeptide from marine sponge. J Am Chem Soc 132:12941–12945

Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319:209–227

Huang YJ, Tejero R, Powers R, Montelione GT (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. Proteins 62:587–603

Ingram RN, Orth P, Strickland CL, Le HV, Madison V, Beyer BM (2006) Stabilization of the autoproteolysis of TNF-alpha converting enzyme (TACE) results in a novel crystal form suitable for structure-based drug design studies. Protein Eng Des Sel 19:155–161

Jain A, Vaidehi N, Rodriguez G (1993) A fast recursive algorithm for molecular dynamics simulation. J Comput Phys 106:258–268

Kallen J, Spitzfaden C, Zurini MGM, Wider G, Widmer H, Wüthrich K, Walkinshaw MD (1991) Structure of human cyclophilin and its binding site for cyclosporin A determined by X-ray crystallography and NMR spectroscopy. Nature 353:276–279

Klukowski P, Walczak MJ, Gonczarek A, Boudet J, Wider G (2015) Computer vision—based automated peak picking applied to protein NMR spectra. Bioinformatics. doi:10.1093/bioinformatics/btv318

Koradi R, Billeter M, Wüthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 14:51–55

Koradi R, Billeter M, Engeli M, Güntert P, Wüthrich K (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. J Magn Reson 135:288–297

López-Méndez B, Güntert P (2006) Automated protein structure determination from NMR spectra. J Am Chem Soc 128:13112–13122

Sadowski J, Gasteiger J, Klebe G (1994) Comparison of automatic three-dimensional model builders using 639 X-ray structures. J Chem Inf Comput Sci 34:1000–1008

Schmidt E, Güntert P (2012) A new algorithm for reliable and general NMR resonance assignment. J Am Chem Soc 134:12817–12829

Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Discovering high-affinity ligands for proteins: SAR by NMR. Science 274:1531–1534

Weber C, Wider G, von Freyberg B, Traber R, Braun W, Widmer H, Wüthrich K (1991) NMR structure of cyclosporin A bound to cyclophilin in aqueous solution. Biochemistry 30:6563–6574

Westbrook JD, Shao C, Feng Z, Zhuravleva M, Velankar S, Young J (2015) The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. Bioinformatics 31:1274–1278

Wüthrich K (1986) NMR of proteins and nucleic acids. Wiley, New York

Yanagisawa T, Ishii R, Fukunaga R, Kobayashi T, Sakamoto K, Yokoyama S (2008) Crystallographic studies on multiple conformational states of active-site loops in pyrrolysyl-tRNA synthetase. J Mol Biol 378:634–652