CIF APPLICATIONS

Authors of any software that reads, writes or validates CIF data are invited to contribute to this series. Authors should state clearly when submitting a manuscript to a Co-editor that the paper should be included as part of the CIF Applications series. An appropriate series number will be assigned by the Editorial Office.

J. Appl. Cryst. (1999). **32**, 125–133

CIF Applications. XI. *A La Mode*: a ligand and monomer object data environment. I. Automated construction of mmCIF monomer and ligand models

Les Clowney, John D. Westbrook* and Helen M. Berman *at Nucleic Acid Database Project, Department of Chemistry, Rutgers, The State University of New Jersey, USA. E-mail: jwest@ndb.rutgers.edu*

(*Received 9 December 1997; accepted 3 April 1998*)

Abstract

The macromolecular Crystallographic Information File (mmCIF) dictionary [Fitzgerald *et al.* (1997). *The Macromolecular Crystallographic Information File Dictionary*, http://ndbserver.rutgers.edu/mmcif] provides a comprehensive description of chemical components used as models in the crystallographic refinement of macromolecular structures. A new ligand and monomer object data environment named *A La Mode* is described for building chemical-component models in the mmCIF representation from surveys of high-resolution small-molecule crystal structures. Examples of the application of this system are presented for an intercalating drug component and for a nucleotide unit constructed from independent base, sugar and phosphate components.

1. Introduction

In the crystallographic determination of macromolecular structure at less than atomic resolution, standard values and uncertainties of bond distances and bond angles are required to supplement the experimental X-ray data. The importance of the accuracy of these standard geometric quantities has been well demonstrated for the refinement of both protein (Engh & Huber, 1991) and nucleic acid (Parkinson *et al.*, 1996) macromolecular systems. In previous work, dictionaries of nucleic acid standard geometries have been reported for nitrogenous bases (Clowney *et al.*, 1996) and sugar and phosphate constituents (Gelbin *et al.*, 1996). Using these standard geometries and their associated uncertainties, new parameters for the refinement of nucleic-acid-containing structures were developed (Parkinson *et al.*, 1996). These standard geometries were developed by performing surveys of high-resolution small-molecule crystal structures in the Cambridge Structural Database (CSD) (Allen *et al.*, 1979) and the Nucleic Acid Database (NDB) (Berman *et al.*, 1992). The experience gained in performing these surveys in a largely manual fashion has led to the development of a ligand and monomer object data environment (*A La Mode*). *A La Mode* automates many of the laborious tasks in building dictionaries of geometrical standards including: database query construction, integration of database survey results, accurate book-keeping of survey results, analysis and comparison of covalent geometry and stereochemistry, and the assembly of complex model structures from the results of multiple database surveys.

A La Mode has been designed to start with a minimum topological description of a ligand or monomer component and to perform the tasks required to construct the mmCIF component description from a database of high-resolution small-molecule crystal structures. The starting description of chemical connectivity is used to construct a query for the CSD. The pool of structures which satisfy this chemical description is then filtered by *A La Mode* according to bonding type, covalent geometry, stereochemistry, experimental or statistical criteria in order to select the particular structures used to construct the component model. Linking and merging operations are provided to permit the construction of complex models from a collection of smaller model components. *A La Mode* provides a variety of tools to analyze, display and compare the geometrical features of model structures. The system permits detailed analysis of statistical correlation among geometrical features within a model and between different models.

The process of model construction and all associated model-building tasks are fully documented within an mmCIF data model using a small set of extensions (Clowney & Westbrook, 1997*a*) to the mmCIF dictionary. The functionality of *A La Mode* is presented to the user as a small set of Unix commands (Clowney & Westbrook, 1997*b*) and these commands may be scripted together to perform complex operations. A WWW query interface has also been developed to provide a flexible means of selecting models from a database of model objects.

2. The mmCIF representation of chemical components

The mmCIF representation of macromolecular structure uniformly describes monomer units, ligands and solvent molecules as a set of chemical components. *A La Mode* takes advantage of the detailed description of chemical components provided in the mmCIF dictionary (Fitzgerald *et al.*, 1993) as a means of representing dictionaries of standard geometrical features. This representation, which includes valence geometry, stereochemistry and linkage geometry, is also tightly integrated into the mmCIF description of crystallographic refinement and macromolecular structure. By providing explicit data items for the complete geometrical descriptions of chemical components (monomers, ligands or solvent), this representation makes it possible to include the detailed features of the refinement model in the description of the crystallographic experiment. In addition, the mmCIF representation provides a standard format in which dictionaries of geometrical standards can be expressed, facilitating both the interchange and reuse of this information.

The mmCIF description for chemical components is shown schematically in Fig. 1. In this figure, each mmCIF category and its associated data items are enclosed in boxes. Categories are a basic organizational feature of the mmCIF dictionary and they are used to collect groups of related data items. An mmCIF category is essentially a table in which each data item is a column. The data items which determine the uniqueness of each table row are identified as key data items. In the figure these items are preceded by solid circles. The hierarchical nature of the mmCIF chemical-component description is illustrated by the connections drawn between data items which are members of multiple categories. The arrows on the connections between related data items are drawn pointing towards the parent definition of the data item. For instance, the component identifier `_chem_comp.id` defined in category `chem_comp` is the parent definition of this identifier, and is also referenced in each of the other categories describing chemical components. Similarly, the atom identifiers which define bonds, angles and torsions are all related to the parent definition of the atom identifier in the `chem_comp_atom` category. Collectively, the categories in Fig. 1 describe the atom and component nomenclature, covalent geometry, stereochemistry, bond types and three-dimensional coordinates of a component model. Data items are also provided in these categories to hold estimates of the uncertainties in each of the geometrical quantities.

The mmCIF description of covalent linkages between chemical components is illustrated in Fig. 2. The collection of categories in this figure describes the geometrical features required to specify the linkages between components according to component type. The mmCIF dictionary enumerates the component types which describe linkages occurring in the most common polymer systems (*e.g.* D-peptide linking, DNA linking, RNA linking). Complete definitions and examples of each of these data items can be found in the mmCIF dictionary (Fitzgerald *et al.*, 1997).

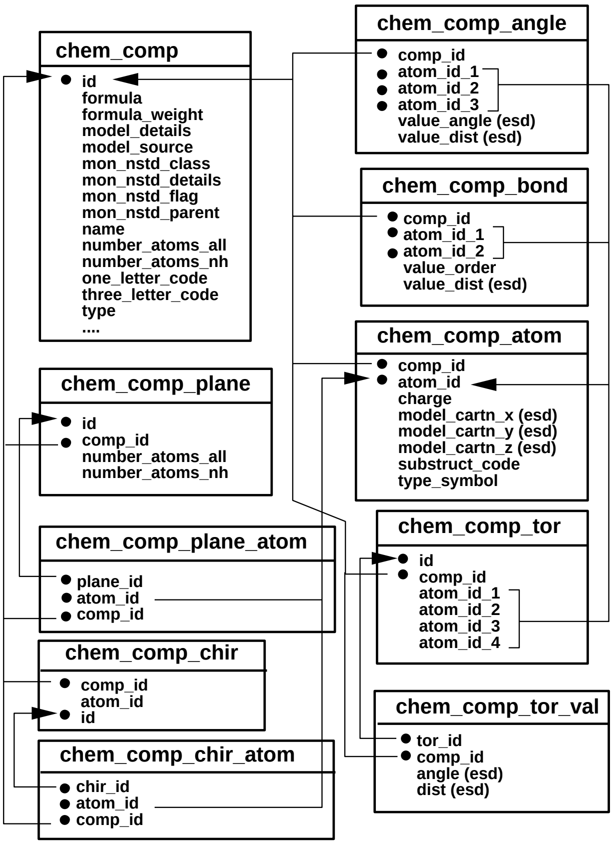


Fig. 1. A schematic representation of the mmCIF categories describing chemical components. In this figure, boxes enclose the mmCIF categories. The upper portion of each box identifies the category and the lower portion contains a list of the data items within the category. Data items that are preceded by a solid circle are the key data items which uniquely identify each row of data within the category. For those data items which occur in multiple categories, the arrows point at the location of the parent definition of the data item. Complete definitions and examples of each of these data items can be found in the mmCIF dictionary (Fitzgerald *et al.*, 1997).

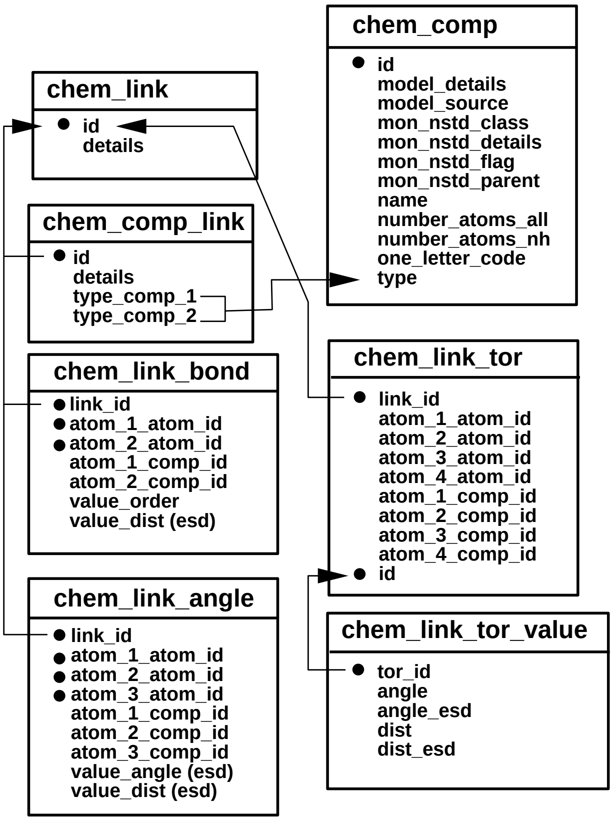


Fig. 2. A schematic representation of the mmCIF categories describing linkages between chemical components. In this figure, boxes enclose the mmCIF categories. The upper portion of each box identifies the category and the lower portion contains a list of the data items within the category. Data items that are preceded by a solid circle are the key data items which uniquely identify each row of data within the category. For those data items which occur in multiple categories the arrows point at the location of the parent definition of the data item. Complete definitions and examples of each of these data items can be found in the mmCIF dictionary (Fitzgerald *et al.*, 1997).

Fig. 3 contains an abbreviated example of the mmCIF component data file for the ligand component adriamycin. The example illustrates the simple appearance of the comprehensive mmCIF structural description. Fig. 4 shows an example of the data file representing the linkage between a ribose sugar and a purine base. This example highlights the application of the `chem_comp_link` category family in the description of linkages which can be used between any pair of components of the appropriate component type.

3. The model-building system

A functional diagram of the model-building system employed by *A La Mode* is shown in Fig. 5. The model-building functions of *A La Mode* were created to facilitate the construction of component models from available high-resolution crystal structures and to minimize the manual intervention required in model construction and analysis, where possible. *A La Mode* was designed to build fully populated mmCIF component descriptions and to supplement this mmCIF data representation with all of the internal information (Clowney & Westbrook, 1997*a*) required to document completely the model-building process.

The minimum input to the model-building system consists of a list of atom names and an associated list of bonds. This information can be input directly, it can be derived from a list of atoms and Cartesian coordinates, or it can be derived from a simple molecular input line entry system (SMILES) string (Weininger *et al.*, 1989; James *et al.*, 1996). A component identifier along with this minimal chemical description are stored in the mmCIF categories `chem_comp`, `chem_comp_atom` and `chem_comp_bond`. This sparsely populated mmCIF data

```
###############
## CHEM_COMP ##
###############

_chem_comp.id                     DM2
_chem_comp.name                   adriamycin
_chem_comp.model_details
;
This component fragment constructed from
CSD crystal structures with the heavy
atom topology of daunomycin.
;
_chem_comp.formula               'C27 H30 N1 O11'
_chem_comp.formula_weight         544.52
_chem_comp.number_atoms_all       69
_chem_comp.number_atoms_nh        39
_chem_comp.type                   non-polymer

####################
## CHEM_COMP_ATOM ##
####################

loop_
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.model_cartn_x
_chem_comp_atom.model_cartn_y
_chem_comp_atom.model_cartn_z
_chem_comp_atom.charge
DM2  C1  C  -1.890  0.874  -4.439  0
DM2  C2  C  -3.027  1.401  -3.720  0
DM2  C3  C  -3.146  2.759  -3.180  0
DM2  C4  C  -2.127  3.676  -3.900  0
#   ...... Abbreviated ......

####################
## CHEM_COMP_BOND ##
####################

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.value_dist
_chem_comp_bond.value_dist_esd
DM2  C1  C2  arom  1.3899  0.0320
DM2  C2  C3  arom  1.3862  0.0442
DM2  C3  C4  arom  1.4204  0.0440
#   ...... Abbreviated ......
```

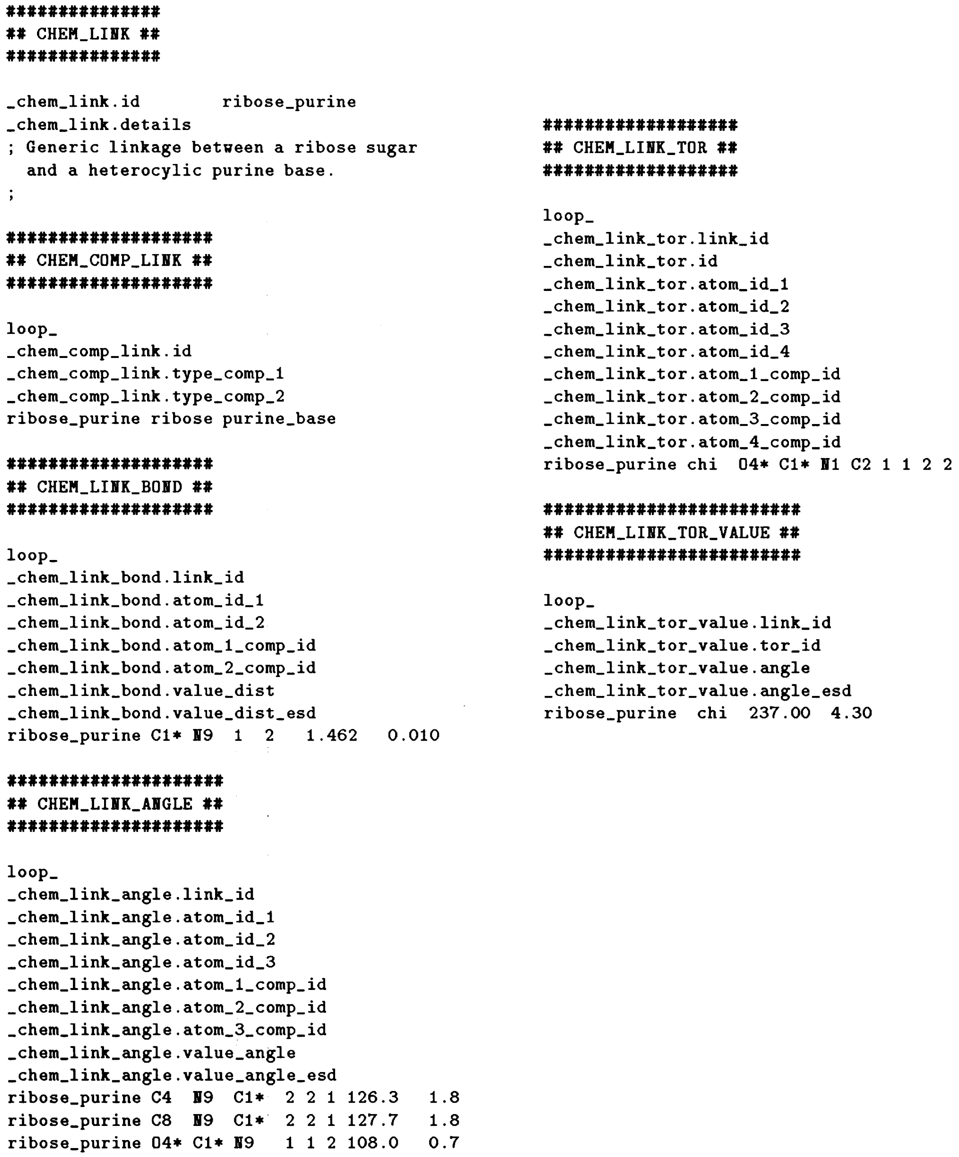
```
#####################
## CHEM_COMP_ANGLE ##
#####################

loop_
_chem_comp_angle.comp_id
_chem_comp_angle.atom_id_1
_chem_comp_angle.atom_id_2
_chem_comp_angle.atom_id_3
_chem_comp_angle.value_angle
_chem_comp_angle.value_angle_esd
_chem_comp_angle.value_dist
_chem_comp_angle.value_dist_esd
DM2  C1  C2  C3  121.73  2.4908  2.4244  0.0580
DM2  C2  C3  C4  118.86  3.2715  2.4149  0.0409
#   ...... Abbreviated ......

###################
## CHEM_COMP_TOR ##
###################

loop_
_chem_comp_tor.id
_chem_comp_tor.comp_id
_chem_comp_tor.atom_id_1
_chem_comp_tor.atom_id_2
_chem_comp_tor.atom_id_3
_chem_comp_tor.atom_id_4
C1-C2-C3-C4  DM2  C1  C2  C3  C4
#   ...... Abbreviated ......

#########################
## CHEM_COMP_TOR_VALUE ##
#########################
loop_
_chem_comp_tor_value.tor_id
_chem_comp_tor_value.comp_id
_chem_comp_tor_value.angle
_chem_comp_tor_value.angle_esd
_chem_comp_tor_value.dist
_chem_comp_tor_value.dist_esd
C1-C2-C3-C4  DM2  0.83  6.08  2.8044  0.0345
#   ...... Abbreviated ......
```

Fig. 3. An abbreviated example of the mmCIF description of the drug ligand adriamycin.

file serves as the *A La Mode* input file and this file also serves as the repository for information generated at each step in the model-construction process.

The chemical information in the *A La Mode* input file is used to construct an initial survey query of a structural database. The *A La Mode* system has been designed to isolate database-specific aspects of the query process into separate modules. To date, query modules have been developed for the NDB and CSD database systems; however, *A La Mode* could easily be extended to accommodate other database query interfaces.

An input file and a Unix script illustrating a simple survey for a target ligand adriamycin are shown in Fig. 6. The input description in this example consists of some identifying information and formal charge in the `chem_comp` category, the list of heavy atoms in the `chem_comp_atom` category and the list of bonded heavy atoms in the `chem_comp_bond` category. The values for coordinates and bond distances are intentionally omitted in the input file. The `survey.sh` script file reads the skeleton component description in the `input.cif` file, and constructs and runs a survey query for the CSD database using this geometrical description. The arguments to the search command, `alamode_csd`, indicate that H atoms are missing in the input structure description and that the selected structures should be sorted into classes according to bond order. If structures are found by the survey query, the script analyzes the selected structures by computing histograms of the distributions of each geometrical feature and statistical diagnostics



```
################
## CHEM_LINK ##
################

_chem_link.id         ribose_purine
_chem_link.details
; Generic linkage between a ribose sugar
  and a heterocylic purine base.
;

#####################
## CHEM_COMP_LINK ##
#####################

loop_
_chem_comp_link.id
_chem_comp_link.type_comp_1
_chem_comp_link.type_comp_2
ribose_purine ribose purine_base

#####################
## CHEM_LINK_BOND ##
#####################

loop_
_chem_link_bond.link_id
_chem_link_bond.atom_id_1
_chem_link_bond.atom_id_2
_chem_link_bond.atom_1_comp_id
_chem_link_bond.atom_2_comp_id
_chem_link_bond.value_dist
_chem_link_bond.value_dist_esd
ribose_purine C1* N9  1  2   1.462   0.010

######################
## CHEM_LINK_ANGLE ##
######################

loop_
_chem_link_angle.link_id
_chem_link_angle.atom_id_1
_chem_link_angle.atom_id_2
_chem_link_angle.atom_id_3
_chem_link_angle.atom_1_comp_id
_chem_link_angle.atom_2_comp_id
_chem_link_angle.atom_3_comp_id
_chem_link_angle.value_angle
_chem_link_angle.value_angle_esd
ribose_purine C4  N9  C1*  2 2 1 126.3   1.8
ribose_purine C8  N9  C1*  2 2 1 127.7   1.8
ribose_purine O4* C1* N9   1 1 2 108.0   0.7
```

```
###################
## CHEM_LINK_TOR ##
###################

loop_
_chem_link_tor.link_id
_chem_link_tor.id
_chem_link_tor.atom_id_1
_chem_link_tor.atom_id_2
_chem_link_tor.atom_id_3
_chem_link_tor.atom_id_4
_chem_link_tor.atom_1_comp_id
_chem_link_tor.atom_2_comp_id
_chem_link_tor.atom_3_comp_id
_chem_link_tor.atom_4_comp_id
ribose_purine chi  O4* C1* N1 C2 1 1 2 2

#########################
## CHEM_LINK_TOR_VALUE ##
#########################

loop_
_chem_link_tor_value.link_id
_chem_link_tor_value.tor_id
_chem_link_tor_value.angle
_chem_link_tor_value.angle_esd
ribose_purine  chi  237.00  4.30
```

Fig. 4. An abbreviated example of the mmCIF description of the linkage between chemical components. In this example, the linkage between a ribose sugar and a purine base is illustrated. The component types used in this example are extensions to the standard set of component types provided in the mmCIF dictionary. These extensions, which are defined in the *A La Mode* local dictionary (Clowney & Westbrook, 1997*a*), were required in order to describe the conformational diversity of nucleotide systems.

about each distribution. The output file for the survey query, model.cif, contains a populated mmCIF component description similar to that shown in Fig. 3. Bond distances, bond angles and torsion angles are computed for the selected structures by the alamode_csd program. If angles and torsions are not specified in the input file, then they are derived from the input list of bonds.

The database survey process is adaptive and provides the automatic adjustment of query constraints according to the composition of the survey result. This facility is illustrated in Fig. 5 for the simple case of relaxing a query constraint (*e.g.* *R* factor or the maximum average error in carbon–carbon bond length) in order to obtain an acceptable survey size. This adaptive procedure can also be used to search systematically for the largest fragment of a target component in a database.

Once the initial database survey has been performed, the pool of structures obtained from the survey can be further refined using a variety of *A La Mode* selection filters. Filters can be used to select groups of structures from the survey pool with particular geometrical features, bonding types, stereochemical features, experimental features or statistical properties. For example, geometric filters can be used to select a particular torsional conformation, or stereochemical filters may be used to reject structures with a particular chirality. *A La Mode* implements these selection filters in a pipeline fashion. Each filter accepts a component model object as input and returns a component model object as output. Using this architecture, very complicated selections can be performed by cascading an appropriate sequence of selection filters.

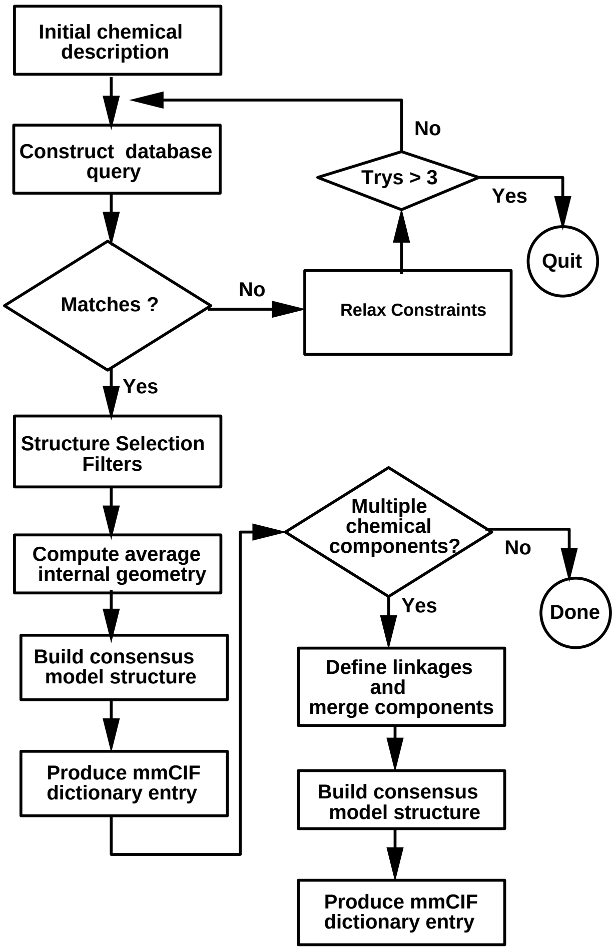

Initial chemical description
Construct database query
No
Trys > 3
Yes
Quit
Matches ?
No
Relax Constraints
Yes
Structure Selection Filters
Multiple chemical components?
No
Compute average internal geometry
Yes
Done
Build consensus model structure
Define linkages and merge components
Produce mmCIF dictionary entry
Build consensus model structure
Produce mmCIF dictionary entry

Fig. 5. A schematic diagram describing the *A La Mode* model-building procedure.

At any point after the initial database survey, *A La Mode* provides analysis methods which report the statistical features of the distribution of any geometrical or stereochemical quantity. The distribution of any feature may also be displayed as a histogram. *A La Mode* also computes and tabulates comparative statistics for the distributions of geometrical features in different models. This functionality greatly facilitates the detailed analysis of relationships which may exist among the geometrical quantities. These analytical features are essential for dealing with the conformational diversity typical of both nucleotide and many ligand systems. Such systems may exhibit multimodal torsional distributions and correlated geometries within flexible ring systems. In these complicated cases, the analysis tools provided by *A La Mode* assist in the detection of the clusters of related geometrical features and this information can be used by the geometrical filters to focus a structure selection to a particular subset.

A La Mode has been designed to work with a number of data sources including: CSD, NDB and PDB (Protein Data Bank) (Bernstein *et al.*, 1977). Not all of these sources of data provide the same sophistication in search and analysis tools. Where the CSD provides robust structural query and analysis features, the macromolecular data in the PDB exists as a collection of flat files. Although it would be possible to perform the majority of the *A La Mode* statistical analysis using tools provided with the CSD system, it would not be possible to obtain this analysis from all the macromolecular data sources. In order to deal with this diversity of data sources, *A La Mode* has reused much of the structural analysis software orginally developed to support the NDB. The resulting design permits a uniform level of search and analysis on high-resolution small-molecule structures, macromolecular monomers and ligands bound in macromolecules.

Perhaps the most important new functionality provided by *A La Mode* is those features which facilitate the construction of complex models from sets of smaller components. In many cases, it may not be desirable or possible to construct a component model directly from the survey of small-molecule crystal structures. In such cases, a composite model may be constructed from a collection of smaller models obtained from different database surveys. *A La Mode* builds composite models by merging groups of models or model fragments into the composite model. Typically, the geometrical features of a linkage are modeled within a separate component which incorporates the desired chemical environment at the linkage termini. The merging features of *A La Mode* allow for the extraction of the model linkage geometry and the insertion of the linkage geometry into the composite model, overwriting any features duplicated by the linkage in the composite model. Linkages between subcomponents in a composite model can be documented either in the mmCIF chem_comp_link categories or absorbed as regular features in the chem_comp bond, angle and torsion categories. The latter approach is useful when a composite model component may be used subsequently as a monomer in a polymeric structure.

After the desired simple or composite model structure has been assembled, the final component model can be constructed by populating the remaining data items in the mmCIF component description. This involves computing formula information, the mean and standard deviation for each geometrical feature in the sample set of the model and a consensus set of Cartesian coordinates. The consensus structure is the set of Cartesian coordinates that best fits the average internal geometry determined for the model structure. The consensus coordinates are produced using an iterative simplex procedure (Dantzig *et al.*, 1955) which minimizes the error between the target average model geometry and the internal geometry predicted for the consensus coordinate set.

The mmCIF components that are produced by the *A La Mode* model builder are stored either as mmCIF data files or as objects which can be managed by the *A La Mode* persistent object manager, *CIFOBJ* (Schirripa & Westbrook, 1996). The latter object manager supports the searchable interface to the *A La Mode* database of model objects.

4. *A La Mode* database interface

The *A La Mode* database of component models provides a WWW form interface from which a user may search the database by component identifier, chemical name, formula or SMILES string. Search targets may include wild-card characters or more complicated patterns described by regular expressions (IEEE, 1991). Searches are performed directly on the object database and the search results are linked to an HTML atlas of ligand and monomer models. The HTML presentation permits the user to navigate the molecular graphics, distribution histograms, tabular summaries of geometrical features, and structural comparisons which are compiled for each model. The following sections provide some examples of this *A La Mode* atlas.

4.1. *The A La Mode nucleotide model atlas*

Fig. 7 shows the nucleotide atlas entry for 2′-deoxyadenosine-5′-phosphate. This page provides a top-level description of the model and presents the user with a chemical name, charge, a description of the particular conformation modeled, formula data for the full component (with H atoms) as well as the modeled component, a chemical diagram, a SMILES string and a table of components from which the composite model was built. The *A La Mode* atlas entries for the particular sugar, base and phosphate models from which this nucleotide was constructed can be accessed from links within this latter table. The mmCIF data file for this nucleotide can also be accessed from the link beside the component name on this page.

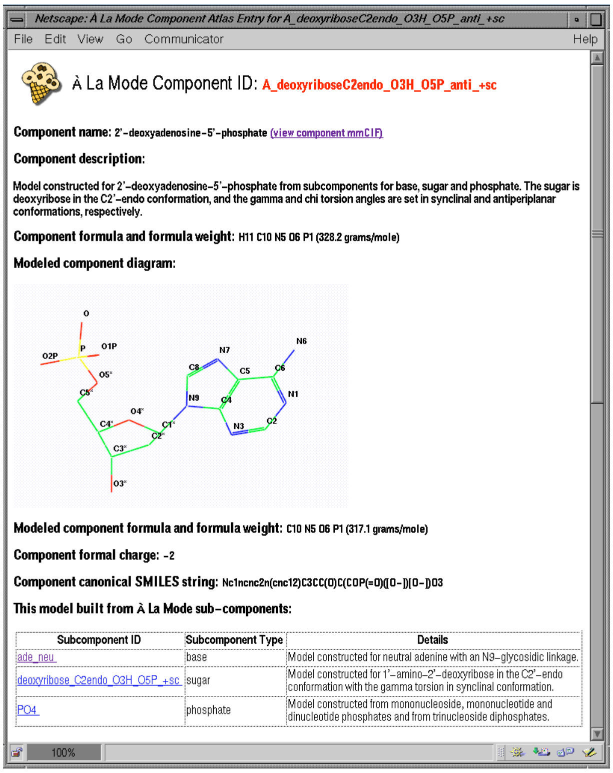
```
data_DM2
#
# File: input.cif
#       A La Mode Survey Query Input
#       for ligand target adriamycin

################
## CHEM_COMP ##
################

_chem_comp.id                     'DM2'
_chem_comp.name                   'ADRIAMYCIN'
_chem_comp.type                   non-polymer
_chem_comp.ndb_formal_charge      0

#####################
## CHEM_COMP_ATOM ##
#####################

loop_
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.model_cartn_x
_chem_comp_atom.model_cartn_y
_chem_comp_atom.model_cartn_z
     DM2 'C1'        C  .  .  .
     DM2 'C2'        C  .  .  .
     DM2 'C20'       C  .  .  .
     DM2 'C3'        C  .  .  .
     DM2 'C4'        C  .  .  .
     DM2 'C5'        C  .  .  .
     DM2 'O4'        O  .  .  .
     DM2 'C21'       C  .  .  .
     DM2 'C6'        C  .  .  .
#      .... Abbreviated  ....
```

```
#####################
## CHEM_COMP_BOND ##
#####################

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.value_dist
_chem_comp_bond.value_dist_esd
     DM2 'C1'     'C2'      .  .  .
     DM2 'C1'     'C20'     .  .  .
     DM2 'C2'     'C3'      .  .  .
     DM2 'C3'     'C4'      .  .  .
     DM2 'C4'     'C5'      .  .  .
     DM2 'C4'     'O4'      .  .  .
     DM2 'O4'     'C21'     .  .  .

#      .... Abbreviated ....


#
#
# File: survey.sh
#       Script file to run adriamycin survey query.
#
alamode_csd --geo input.cif --order -s --no_H
        --derive --sample_size
--max_hits 200 > model.cif

if [ -f sample_size ]; then
 alamode_extract --histograms
  --title "Adriamycin Results" plot2d < model.cif
 alamode_extract --outliers  < model.cif  > outly.out
 alamode_extract --normality < model.cif  > stats.out
 alamode_extract --extrema   < model.cif  > extrm.out
fi
#
```

Fig. 6. An abbreviated *A La Mode* input file, `input.cif`, illustrating the minimum topological description required to initiate a database survey query. The input file is followed by a script file, `survey.sh`, containing the commands to perform and analyze the CSD database survey.

This example illustrates a composite model for a particular conformation of a nucleotide which has been constructed from independent sugar, base and phosphate components. The first step in building this nucleotide model was to survey the appropriate databases for representative structures for each of these three subcomponents. *A La Mode* is designed to construct survey queries from an mmCIF-encoded set of topological constraints. Using the CSD query construction model, *A La Mode* takes particular advantage of the powerful two-dimensional structure selection features provided by the CSD program *Quest* (Allen *et al.*, 1979). For instance, it is possible to specify for each atom a hydrogen count, or prohibit heavy-atom attachment at a particular atomic site.

The second step in building the nucleotide model involves determining average values and uncertainties for the covalent geometry which produces the desired conformation. Associated with this step is a systematic determination of the geometrical features and the chemical topology which may be correlated with the particular choice of conformation. *A La Mode* automates the pairwise comparisons of distributions of geometrical features that are necessary to make this determination. An exhaustive comparison of the geometrical features of two models obtained from independent surveys can be performed with a single *A La Mode* command.


Netscape: À La Mode Component Atlas Entry for A_deoxyriboseC2endo_O3H_O5P_anti_+sc
File
Edit
View
Go
Communicator
Help
100%

À La Mode Component ID: A_deoxyriboseC2endo_O3H_O5P_anti_+sc

Component name: 2'–deoxyadenosine–5'–phosphate (view component mmCIF)

Component description:

Model constructed for 2'–deoxyadenosine–5'–phosphate from subcomponents for base, sugar and phosphate. The sugar is deoxyribose in the C2'–endo conformation, and the gamma and chi torsion angles are set in synclinal and antiperiplanar conformations, respectively.

Component formula and formula weight: H11 C10 N5 O6 P1 (328.2 grams/mole)

Modeled component diagram:

Modeled component formula and formula weight: C10 N5 O6 P1 (317.1 grams/mole)

Component formal charge: -2

Component canonical SMILES string: Nc1ncnc2n(cnc12)C3CC(O)C(COP(=O)([O-])[O-])O3

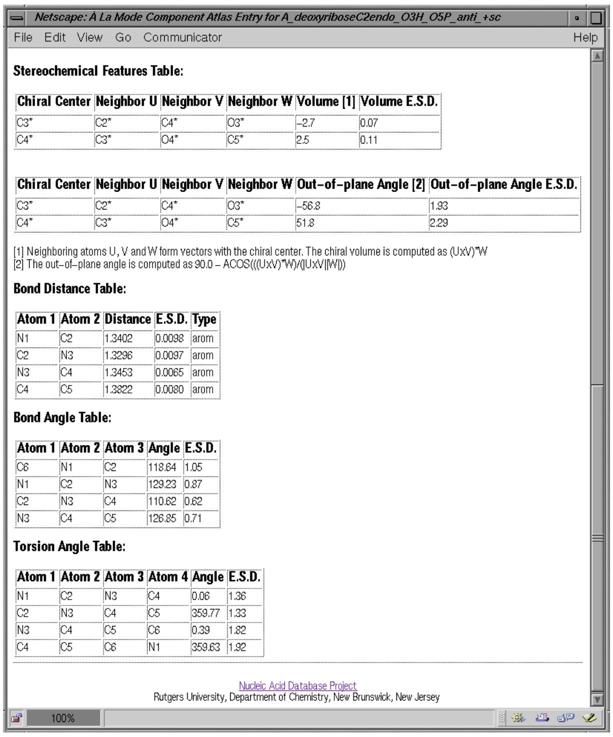
This model built from À La Mode sub–components:

Subcomponent ID	Subcomponent Type	Details
ade_neu	base	Model constructed for neutral adenine with an N9–glycosidic linkage.
deoxyribose_C2endo_O3H_O5P_+sc	sugar	Model constructed for 1'–amino–2'–deoxyribose in the C2'–endo conformation with the gamma torsion in synclinal conformation.
PO4	phosphate	Model constructed from mononucleoside, mononucleotide and dinucleotide phosphates and from trinucleoside diphosphates.

Fig. 7. An example *A La Mode* nucleotide atlas page for the 2′-deoxyadenosine-5′-phosphate. This portion of the atlas entry shows the conformational description, a chemical diagram, formula and formula weight, SMILES (Weininger *et al.*, 1989; James *et al.*, 1996) string and a table of the model components from which the nucleotide model was constructed. The names of each subcomponent are linked to a similar atlas entry which provides detailed information about the subcomponent. This includes the details of the database surveys from which the subcomponents were constructed.

The final step in the construction of the nucleotide model is the assembly of sugar, base and phosphate models. This step also requires the insertion of linking geometry and the geometrical settings for the target conformation. The combined description of the internal coordinates of the model is then used to construct the Cartesian coordinates for the composite nucleotide model. We illustrate the model-building process with a specific example, 2′-deoxyribose-5′-phosphate, in Fig. 7. In this example, the candidate structures for the base component were obtained from a survey of adenine structures in the CSD, unsubstituted except at N9. Candidate structures for the sugar component were obtained from a survey of 2′-deoxyribose in the CSD with an attached purine or pyrimidine at C1′. No substitutions were permitted in the sugar ring or at C5′ and substituents at O5′ were rejected if they formed a cycle with the sugar. CSD structures with an *R* factor greater than 0.08, containing heavy-metal atoms, or with carbon–carbon bond distance errors greater than 0.02 Å were rejected from the survey. The phosphate model component was derived from a survey of high-resolution mononucleoside, mononucleotide and dinucleotide phosphates, and trinucleotide diphosphate structures in the NDB.

To obtain the desired nucleotide conformation, *A La Mode* filters were used to select only those sugar structures from the CSD survey in C2′-*endo* conformation. From a detailed analysis of conformational correlation it was determined that the χ and γ torsion angles could be better modeled from a broader range of structures. The χ torsion angle in this model


Netscape: À La Mode Component Atlas Entry for A_deoxyriboseC2endo_O3H_O5P_anti_+sc
File
Edit
View
Go
Communicator
Help
100%

Stereochemical Features Table:

Chiral Center	Neighbor U	Neighbor V	Neighbor W	Volume [1]	Volume E.S.D.
C3*	C2*	C4*	O3*	–2.7	0.07
C4*	C3*	O4*	C5*	2.5	0.11

Chiral Center	Neighbor U	Neighbor V	Neighbor W	Out–of–plane Angle [2]	Out–of–plane Angle E.S.D.
C3*	C2*	C4*	O3*	–56.8	1.93
C4*	C3*	O4*	C5*	51.8	2.29

[1] Neighboring atoms U, V and W form vectors with the chiral center. The chiral volume is computed as (UxV)*W
[2] The out–of–plane angle is computed as 90.0 – ACOS(((UxV)*W)/(|UxV||W|))

Bond Distance Table:

Atom 1	Atom 2	Distance	E.S.D.	Type
N1	C2	1.3402	0.0098	arom
C2	N3	1.3296	0.0097	arom
N3	C4	1.3453	0.0065	arom
C4	C5	1.3822	0.0080	arom

Bond Angle Table:

Atom 1	Atom 2	Atom 3	Angle	E.S.D.
C6	N1	C2	118.64	1.05
N1	C2	N3	129.23	0.87
C2	N3	C4	110.62	0.62
N3	C4	C5	126.85	0.71

Torsion Angle Table:

Atom 1	Atom 2	Atom 3	Atom 4	Angle	E.S.D.
N1	C2	N3	C4	0.06	1.36
C2	N3	C4	C5	359.77	1.33
N3	C4	C5	C6	0.39	1.82
C4	C5	C6	N1	359.63	1.92

Nucleic Acid Database Project
Rutgers University, Department of Chemistry, New Brunswick, New Jersey

Fig. 8. An example *A La Mode* nucleotide-atlas page for the 2′-deoxyadenosine-5′-phosphate. This portion of the atlas entry details the stereochemical features and the covalent geometry of the nucleotide model.

was determined from an independent survey of C2′-*endo* ribose/2′-deoxyribose sugars and purine bases in synclinal conformation. The average antiperiplanar value for the γ torsion used in this model was obtained from an independent survey of all ribose/2′-deoxyribose sugars. *A La Mode* extraction and insertion filters were used to calculate and extract these torsion values and to insert them into the final composite model.

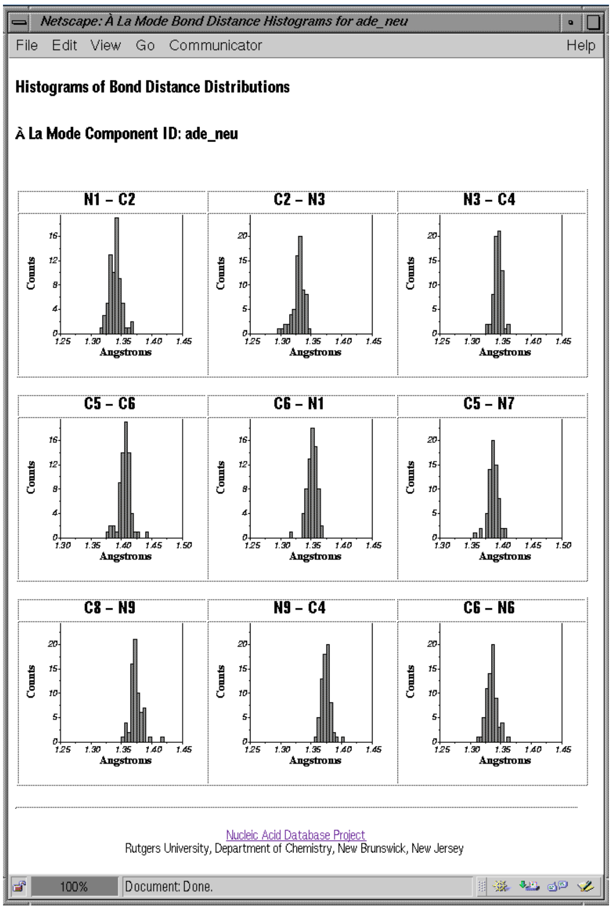
At the end of the model-building process *A La Mode* constructs the HTML atlas summary shown in Fig. 7 as well as a geometrical summary shown in abbreviated form in Fig. 8. A histogram of the distribution of each of the features in the geometrical summary is also produced. Fig. 9 shows one page of the distributions of bond distances, which is linked to the table of bond distances and uncertainties.

4.2. *The A La Mode drug/ligand model atlas*

The previous example illustrated the full functionality of the *A La Mode* system in the integration of multiple-component models and geometrical features to construct and analyze a composite model for a nucleotide. Although this same approach can be applied in building any model system, it is also useful to have a procedure for building and analyzing potential model structures which does not involve this complicated decomposition process.

Accordingly, *A La Mode* provides a completely automated procedure which, starting from a minimal chemical description of a ligand, produces a geometrical summary and comparison of the candidate model structures and a comparison of each candidate model with any bound example of the ligand.

A La Mode first constructs a broad database survey using only the topological description of the ligand. If bonding type or hydrogen counts are not available, *A La Mode* will sort out each class or structure that differs in chemical bonding. If bonding information is available, then it can be used in constructing the survey query. The pool of structures which result from this survey query can be used directly to construct a ligand model; however, in most cases this will result in neither an accurate nor a chemically representative model. Instead of building the model structure at this stage, *A La Mode* produces pairwise comparisons of the candidate structures, comparisons of the candidate and bound structures, and pairwise comparisons of the bound structures. Collectively, these comparisons can be used to evaluate if a subset of the candidate structures can be appropriately averaged to build a model structure, or if it is more appropriate to select individual candidate structures as models.



Histograms of Bond Distance Distributions

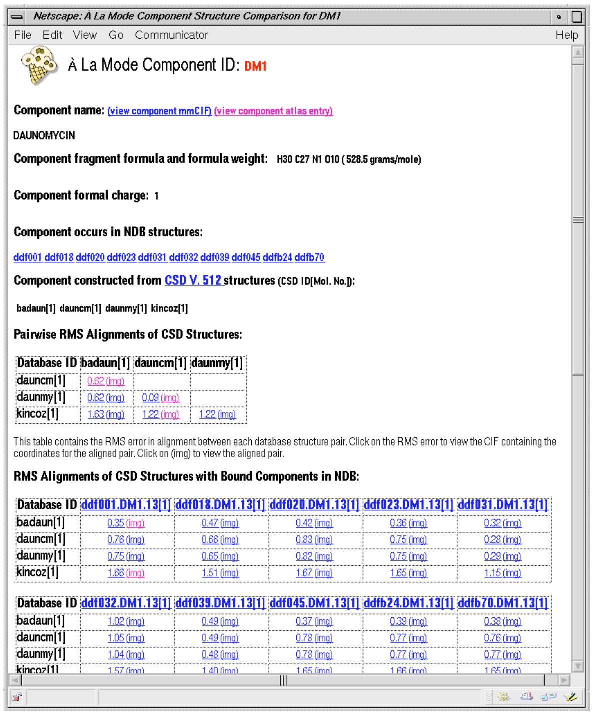
À La Mode Component ID: ade_neu

Nucleic Acid Database Project
Rutgers University, Department of Chemistry, New Brunswick, New Jersey

Document: Done.

Fig. 9. Example *A La Mode* nucleotide-atlas page for the 2′-deoxyadenosine-5′-phosphate. This portion of the atlas-page entry shows a selection of the distributions of bond distances for this nucleotide model.

Fig. 10 shows an abbreviated example of a ligand structure-comparison page from the *A La Mode* drug-atlas entry for


Netscape: À La Mode Component Structure Comparison for DM1
File Edit View Go Communicator Help

À La Mode Component ID: DM1

Component name: (view component mmCIF) (view component atlas entry)

DAUNOMYCIN

Component fragment formula and formula weight: H30 C27 N1 O10 (528.5 grams/mole)

Component formal charge: 1

Component occurs in NDB structures:

ddf001 ddf018 ddf020 ddf023 ddf031 ddf032 ddf039 ddf045 ddfb24 ddfb70

Component constructed from CSD V. 512 structures (CSD ID[Mol. No.]):

badaun[1] dauncm[1] daunmy[1] kincoz[1]

Pairwise RMS Alignments of CSD Structures:

Database ID	badaun[1]	dauncm[1]	daunmy[1]
dauncm[1]	0.62 (img)		
daunmy[1]	0.62 (img)	0.09 (img)	
kincoz[1]	1.63 (img)	1.22 (img)	1.22 (img)

This table contains the RMS error in alignment between each database structure pair. Click on the RMS error to view the CIF containing the coordinates for the aligned pair. Click on (img) to view the aligned pair.

RMS Alignments of CSD Structures with Bound Components in NDB:

Database ID	ddf001.DM1.13[1]	ddf018.DM1.13[1]	ddf020.DM1.13[1]	ddf023.DM1.13[1]	ddf031.DM1.13[1]
badaun[1]	0.35 (img)	0.47 (img)	0.42 (img)	0.36 (img)	0.32 (img)
dauncm[1]	0.76 (img)	0.66 (img)	0.83 (img)	0.75 (img)	0.28 (img)
daunmy[1]	0.75 (img)	0.65 (img)	0.82 (img)	0.75 (img)	0.29 (img)
kincoz[1]	1.66 (img)	1.51 (img)	1.67 (img)	1.65 (img)	1.15 (img)

Database ID	ddf032.DM1.13[1]	ddf039.DM1.13[1]	ddf045.DM1.13[1]	ddfb24.DM1.13[1]	ddfb70.DM1.13[1]
badaun[1]	1.02 (img)	0.49 (img)	0.37 (img)	0.39 (img)	0.38 (img)
dauncm[1]	1.05 (img)	0.49 (img)	0.78 (img)	0.77 (img)	0.76 (img)
daunmy[1]	1.04 (img)	0.48 (img)	0.78 (img)	0.77 (img)	0.77 (img)
kincoz[1]	1.57 (img)	1.40 (img)	1.65 (img)	1.66 (img)	1.65 (img)

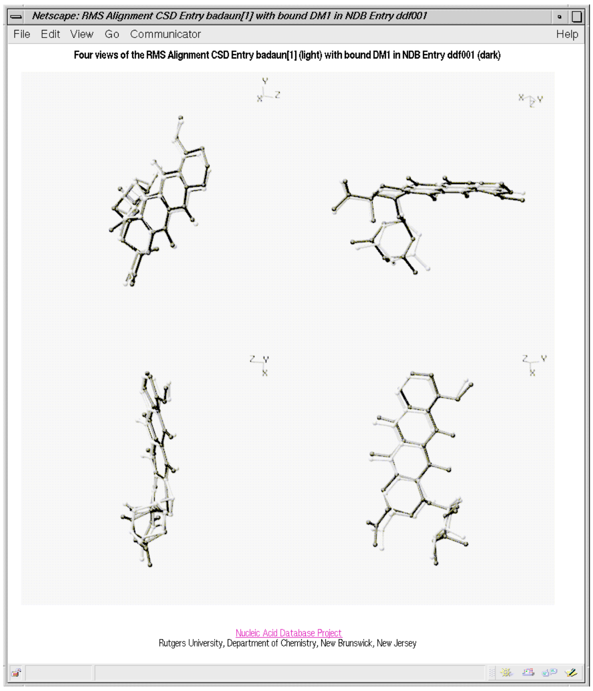
Fig. 10. Example *A La Mode* drug-atlas page for daunomycin. This portion of the atlas-page entry shows the formula and formula weight of the modeled component, the list of NDB entries in which the component is found, the list of CSD entries which were found as candidate models, tables of r.m.s. comparisons among the CSD candidate structures and between each candidate CSD structure, and each example of the bound ligand in the NDB.

daunomycin. In addition to drug name, formula and charge, this page tabulates the list of CSD identifier codes for the candidate structures, the list of NDB identifier codes in which this drug ligand has been reported, a table comparing the candidate structures and a table comparing the candidate structures with each bound form of the drug.

Each element of the comparison table contains the error in r.m.s. alignment, a link to the coordinate set pair in the aligned frame and a link to four static views of the alignment. An example of the presentation of the graphical views of the r.m.s. alignment between a daunomycin structure found in the CSD survey [CSD code BADAUN (Angiuli *et al.*, 1971)] and the bound drug found in the NDB [NDB code DDF001 (Wang *et al.*, 1987)] is shown in Fig. 11. The pairwise comparisons of the bound drug forms are similarly presented but not shown in the preceding figures.

5. Limitations

A La Mode is designed to assist a knowledgeable user in the construction of dictionaries of standard geometries for monomer and ligand components. The selection of a representative subset of structures from a database survey, or the choice of subcomponents to be used to build a complex model is ultimately the responsibility of the user. These decisions may require detailed understanding of the chemistry of the monomer or ligand system. *A La Mode* provides the tools and diagnostics to help the user make these difficult decisions efficiently.


Netscape: RMS Alignment CSD Entry badaun[1] with bound DM1 in NDB Entry ddf001
File
Edit
View
Go
Communicator
Help
Y
X
Z
X
Y
Z
Z
Y
X
Z
Y
X

Four views of the RMS Alignment CSD Entry badaun[1] (light) with bound DM1 in NDB Entry ddf001 (dark)

Nucleic Acid Database Project
Rutgers University, Department of Chemistry, New Brunswick, New Jersey

Fig. 11. Example *A La Mode* drug-atlas page for daunomycin. This portion of the atlas-page entry shows four views of the r.m.s. alignment of the CSD daunomycin structure BADAUN (Angiuli *et al.*, 1971) and the bound form of the drug in NDB entry DDF001 (Wang *et al.*, 1987).

6. Availability

The searchable HTML interface to the database of nucleotide model components created by *A La Mode* is available at http://ndbserver.rutgers.edu/alamode/. A database of drug ligands which bind to oligonucleotides will be available shortly. Documentation and other *A La Mode* resources will be made available at this site.

We gratefully acknowledge Dr Frank Allen of the Cambridge Crystallographic Data Centre for his comments on this manuscript and the CCDC for providing the invaluable database resource of small-molecule crystal structures. We would like to thank Kyle Burkhardt and Glen Charydczak for their help in testing and debugging many of the functions of the *A La Mode* system. This work was supported by a grant from NSF for the Nucleic Acid Database.

References

Allen, F. H., Bellard, S., Brice, M. D., Cartright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *Acta Cryst.* B**35**, 2331–2339.

Angiuli, R., Foresti, E., Sanseverino, L. R. D., Isaacs, N. W., Kennard, O., Motherwell, W. D. S., Wampler, D. L. & Arcamone, F. (1971). *Nature (London) New Biol.* **234**, 78–80.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J. D., Gelbin, A., Demeny, T., Hsieh, S., Srinivasan, A. R. & Schneider, B. (1992). *Biophys. J.* **63**, 751–759.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 509–518.

Clowney, L. & Westbrook, J. D. (1997*a*). *A La Mode. Extensions to the mmCIF Dictionary*. Rutgers University, New Brunswick, NJ, USA, http://ndbserver.rutgers.edu/alamode.

Clowney, L. & Westbrook, J. D. (1997*b*). *A La Mode. Reference Manual*. Technical Report NDB-241, Rutgers University, New Brunswick, NJ, USA, http://ndbserver.rutgers.edu/alamode/.

Dantzig, G. B., Orden, A. & Wolfe, P. (1955). *J. Math.* **5**, 183–195.

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

Fitzgerald, P., Berman, H. M., Bourne, P., McMahon, B., Watenpaugh, K. & Westbrook, J. D. (1997). *The Macromolecular Crystallographic Information File Dictionary*, http://ndbserver.rutgers.edu/mmcif.

Fitzgerald, P. M. D., Berman, H. M., Bourne, P. E. & Watenpaugh, K. (1993). *The Macromolecular CIF Dictionary*. ACA Annual Meeting, Albuquerque, New Mexico, USA. No. D008.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 519–528.

IEEE (1991). *Draft Standard for Information Technology: Portable Operating System Interface* (*POSIX*). Part 2: *Shells and Utilities*. IEEE P1003.2 Draft 11.2.

James, C. A., Weininger, D., Delany, J. & Scofield, J. (1996). *Daylight Toolkit Programmer's Guide*. Daylight Chemical Information Systems, Inc., Mission Viejo, CA, USA.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brunger, A. & Berman, H. M. (1996). *Acta Cryst.* D**52**, 57–64.

Schirripa, S. & Westbrook, J. D. (1996). *CIFOBJ. A Class Library of mmCIF Access Tools*. Reference guide, CIFOBJ V. 1.01. Technical Report NDB-269, Rutgers University, New Brunswick, NJ, USA, http://ndbserver.rutgers.edu/mmcif/software/.

Wang, A. H. J., Ughetto, G., Quigley, G. J. & Rich, A. (1987). *Biochemistry*, **26**, 1152–1163.

Weininger, D., Weininger, A. & Weininger, J. L. (1989). *J. Chem. Inf. Comput. Sci.* **29**, 97–101.