Advanced Automatic Generation of 3D Molecular Structures

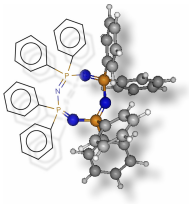
Gábor Imre, Adrián Kalászi, Imre Jákli, Ödön Farkas. Laboratory of Chemical Informatics, Institute of Chemistry,
Eötvös Loránd University, 1/A Pázmány Péter sét., Budapest H-1117 Hungary, farkas@chem.elte.hu

Introduction

Numerous theoretical methods in the field of computational chemistry fall back on the availability of 3D structures of compounds. Determining molecular structure without human interaction is an essential component of several techniques, like QSAR, 3D pharmacophore analysis, reaction prediction, etc. Moreover, current computational tools used for structure determination, including force-fields and quantum chemical methods, require a complete set of initial 3D coordinates. The efficiency of 3D structure based HTS (high throughput screening) tools also can be enhanced by employing conformational analysis to yield multiple valid structures.

Our approach utilizes a composition of several methods ranging from pure rule based[1], multi dimensional distance geometry method[2] to stored substructure lookup features in a flexible software framework. The actual implementation is a highly portable JAVA software, which fits in a broad scale of applications: it can be used in small web drawing applets[3] as well as a standalone database processing component.

The coordinate determination process is characteristically a "divide and conquer" approach: the structure is composed of fragments, which are joined together. From the available fragment conformers, the conformers of the joined structures can be generated during the fuse step. The fragment conformers are generated either through further fragmentation or with an elemental structure/conformer prediction method, consequently the conformational analysis is an inherent part of the building process (in contrast with methods proceeding from 3D initial structures[4]). The novelty of our approach lies in the diversity of the utilized elemental methods and the arisen scalability options.



Molecular modeling, 3D QSAR/QSPR, etc. needs starting molecular structures in 3D, virtual synthesis also needs 3D validation of structures.

Figure 1. Why automatic 3D coordinate generation is important?

Integration with ChemAxon products

Our implementation is integrated into ChemAxon product portfolio as a calculator plugin, so its users can access it several ways. Single low energy conformer generation and force-field based energy minimizations are also accessible independently from the calculator plugin.

GUI integration: 3D structure and conformation generation can be accessed from the Marvin Sketch and View applications and applets. Apart from the plugin interface (*Tools* ▷ *Conformation* ▷ *Conformers*) coordinate generation can be invoked from *Edit* ▷ *Clean* ▷ *3D* submenu, pressing CTRL-3 or by opening a 3D viewer or MarvinSpace window from 2D drawing mode.

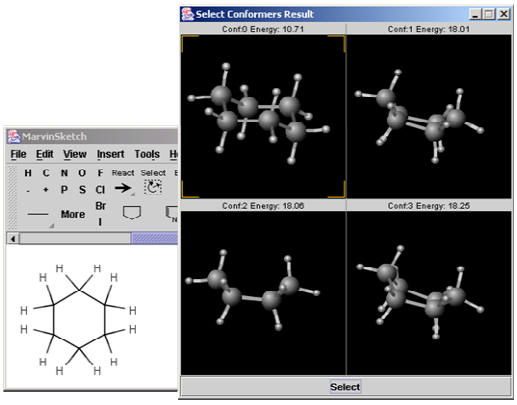


Figure 2. Example of GUI integration: Displaying the output of Conformation/Conformers calculator plugin

Command line interface: Batch processing of multiple structures can be automated through the provided command line tools, where the fine tuning possibility is also present. 3D structure generation functionality is integrated into `molconvert`.

```
molconvert sdf -3:"S{fine}E" 0D.smi > 3D.sdf
```

Figure 3. Using `molconvert` command line tool to generate 3D structure. This example uses fine coordinate generation and stores calculated energy. For help, type „`molconvert -H3D`"

Conformer generation can be accessed through the calculator plugin, which also can be called from command line:

```
cxcalc conformers -m 250 -s true test.sdf
```

Figure 4. Using `cxcalc` to calculate conformers. For help, type „`cxcalc conformers -h`"

API integration: Custom applications may utilize the 3D coordinate generation and conformer analysis functionality through the public API. Fine tuning of the cleaning process can be done by passing additional parameters. Both Molecule.clean() method and calculator plugin interface ConformerPlugin can be used.

```
// read input molecule
MolImporter mi = new MolImporter("test.mol");
Molecule mol = mi.read(); mi.close();
// create plugin
ConformerPlugin plugin = new ConformerPlugin();
// set target molecule
plugin.setInputMolecule(mol);
// set parameters and run calculation
plugin.setMaxNumberOfConformers(400);
plugin.setTimelimit(900);
plugin.run();
// get and process results
Molecule[] conformers = plugin.getConformers();
for (int i = 0; i < plugin.getConformerCount(); ++i) {
   Molecule m = conformers[i];
   // do something with the conformer ...
}
```

Figure 5. Using plugin interface in JAVA code

First presented at 1st European Chemistry Congress, Budapest, Hungary
27-31 August, 2006

ChemAxon

The method in a nutshell

The utilized divide-and-conquer approach builds the given structure from smaller fragments. Fragments are substructures of the original molecule with substituted H atoms on cut bonds. The build process is represented internally as an object tree, where each tree node represents a (partial) conformational analysis of the specific fragment (H-substituted substructure) and also a method (fragment-fragment fuse or direct fragment build/retrieve) which generates the analysis.

Conformational analysis (even when generating only one conformer) is done through a demand-driven model: a build request is passed to the root node of the build tree. Every fragment-fragment fuse node tries to fulfill the build request using fragment conformers generated by associated sub-trees. Once the fuse possibilities are exhausted an additional build request will be passed down to the involved fragments.

The suitable fragment decomposition utilizes several heuristics. The topological and geometrical equivalences should be properly treated during the fuse phase (e.g. identifying the possible fuse alignments), also, recognizing equivalent substructures may substantially accelerate the process. The task is performed by Substructure3DSearch (see below).

The actual positioning of the fusing fragments is done by quaternion fit (see below).

Primary fragment conformer generation

Build via single atom fuses: The spatial alignment of the atoms are determined by traversing the structural graph atom by atom. In each step the possible orientations are identified for the actual atom. This approach is relatively fast and accurate for small sized structures.

The Clean3D functionality in previous versions of Marvin (prior 4.1) relied entirely on this method and it is currently used as elementary fragment builder.

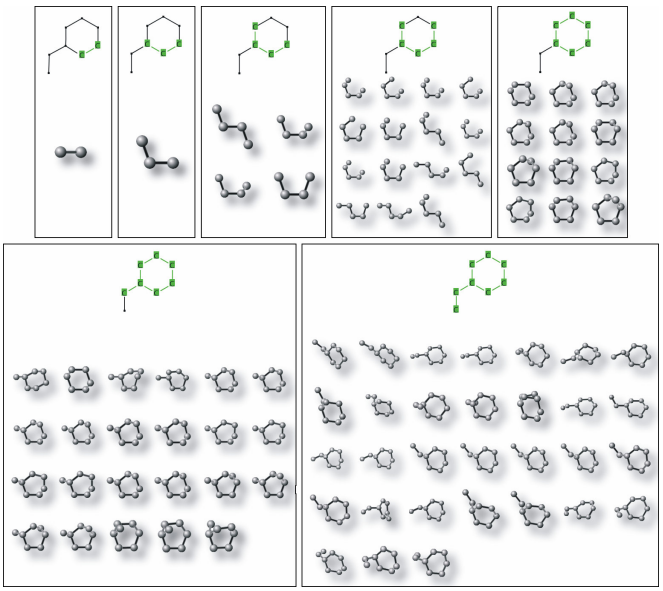


Figure 6. Building a fragment using single atom fuses. Note that fragment conformers are enumerated through the building process.

This method places atoms in a step-by-step manner and determines multiple energetically favorable structures (conformers) for fragments. If the resulting conformer count is greater than a predefined limit, some of them will be ignored in the further process. The conformer count limit affects the scalability and the accuracy of the coordinate generation process.

In each step the base fragment is extended with a connected atom. Starting from multiple fragment conformers, multiple possible atom orientations are determined for each of them. Several limiting heuristics are developed to balance conformational diversity and conformer count.

As stated above the method itself can yield valuable conformers for the majority of the structures, however, its scales poorly with the structure size. Since this method is efficient for small (ring) fragments, it is adequate using it as the primary source of primitive fragments.

In the current implementation this method is encapsulated as a build tree leaf.

Fragment database lookup: The coordinate generation process will be further accelerated by using a fragment conformer database. Implementing an adaptive cache is also in progress.

These scheduled features are expected to have further remarkable impact on the performance without affecting reliability.

Direct build using generalized Minkowski metric: Our initial approach[2] to the 3D coordinate generation problem performs fairly well for compact structures having hard tensions. The currently separately implemented technique is a suitable option for the most problematic types of small fragments. Integration into the above mentioned build tree based hierarchy is scheduled.

The development was initiated by a novel distance geometry type method[1]. It can generate useful 3D coordinates even for extremely stretched structures. The method generates higher dimensional coordinates for any given set of interatomic distances. Distance criteria can be established from topology. Atom-atom distance “wishes” mainly comes from estimated or determined internal coordinates (bond lengths, bond angles, dihedral angles). These local assumptions about the 3D geometry may contain inconsistency, therefore, 3D coordinates satisfying all of the criteria may not exist.

In a non Euclidean, Minkowski-like space all internal distance requirements can be satisfied, using a special metric tensor, **W**:

$$\mathbf{W} = \begin{bmatrix} \pm 1 & 0 & & 0 \\ 0 & \pm 1 & & 0 \\ & & \ddots & \vdots \\ 0 & 0 & \cdots & \pm 1 \end{bmatrix} = \begin{bmatrix} w_1 & 0 & & 0 \\ 0 & w_2 & & 0 \\ & & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

Figure 7. Metric tensor used

Accordingly, the norm of a vector (square of “distance”, metrid) is $d^2(\mathbf{a}) = \mathbf{a}^T\mathbf{W}\mathbf{a}$

This definition induces the presence of singular directions: in these directions a vector with non-zero coordinates have 0 metrid.

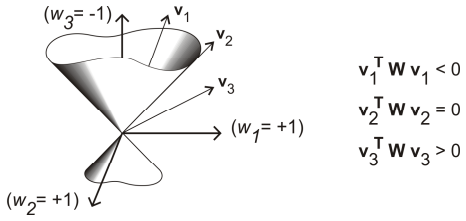
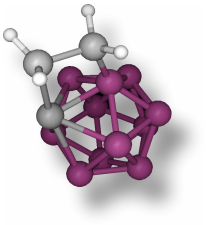


Figure 8. The illustration shows a 3D Minkowski space with metric tensor (1,1,-1). Vectors lying on the depicted cone surface has 0 metrid values, vectors originating form the origin and pointing inside the dual cone has negative metrid values.

For N nodes, arbitrary distance matrix can be satisfied in at most N-1 dimensions. A straight algorithm has been constructed, which can assign such coordinates for a point to satisfy the distances to previously placed points.

After assigning the Minkowski coordinates, geometry optimization is used to reduce dimensionality. The optimization usually destroys some of the established distances, however, with the aid of a proper force-field, the resulting structure is a low energy, valid conformer. The main attribute of the applied force-field is the slight forces pointing from over-3D extra dimensions to zero, which collapses the structure into 3D. For keeping the structure valid, a special molecular mechanics force-field is responsible. Classical force fields (like Dreiding) can be extended to multiple dimensions or a pseudo force field based on the original distances can be constructed. Extending a real-world force-field is a simple task considering that the used energy components can be represented in an at most 3D dimensional subspace.



This method - as expected - can produce valid coordinates for structures with heavy tensions, but the process is slow, since the total number of starting variables to optimize is proportional to the square of the atom count.

Although the efficiency of this approach as an only method of universal coordinate generation is questionable, it can support building 'problematic' fragments of input structures.

Figure 9. Application domain of Minkowski-based direct fragment building

Structure analysis and manipulation tools used

Molecular mechanics force field: A flexible interface connects the molecular mechanics force-fields to the software, allowing extension to the multidimensional Minkowski space. The Dreiding[5] force-field is currently available, implementation of other force-fields is in progress.

```
molconvert sdf -3:"c2E" 3D.sdf > 3D-energy.sdf
```

Figure 10. Invoking only Dreiding energy calculation on 3D structures without coordinate generation. The SDF property „`Energy`" will store the calculated energy value. For help, type „`molconvert -H3D`"

Numerical optimization: Local energy minima related to the generated conformations are determined via a special subspace-Hessian based optimization method[6]. Optimization is also applied to resolve slight atom-atom proximity or bond length problems found in fused fragments.

Adjustable optimization criteria help balancing between optimization step count and accuracy.

```
molconvert sdf -3:"c2o1L2" 3D.sdf > 3D-optimized.sdf
```

Figure 11. Invoking geometry optimization on 3D structures without coordinate generation using „strict" optimization criteria. For help, type „`molconvert -H3D`"

Quaternion fit (JQuatFit): JQuatFit is based on the work of Hamilton[7]. It can fit two molecular structures via a non-iteraive, linear scaling, extremely fast method.

Used for fitting common atoms in equivalence check and fusing fragments.

Substructure3DSearch: It is based on the substructure search implemented by ChemAxon. Simplified for fast exact match (using graph invariants); Extended with: geometry matching (using quaternion fit) to separate conformers; high/low priority matching for selecting suitable fuse positions; geometry constrained topological matching for fragment re-use. It can also quickly distinguish conformers with optional diversity limit

Molecular dynamics: Due to symmetry considerations, it is possible to erroneously identify a resulting structure (an arbitrary critical point of the potential energy surface) as a minimum. MD is used to resolve such problems and also gives a chance to manage serious bond length or proximity problems.

Molecular dynamics calculations are also available as a stand-alone plugin.

Hyperfine: The optional post processing stage of the conformational analysis invokes several molecular dynamic / geometry optimization cycles on each generated conformer to eliminate the invalid local energy minima.

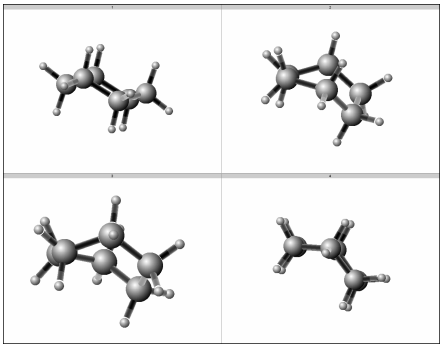


Figure 12. Invalid local energy minimum found during the conformational analysis of cyclohexane (calculation made using the very strict optimization limit)

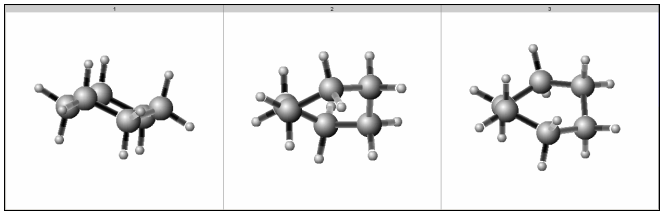


Figure 13. Using hyperfine eliminates the symmetrical boat conformer.

Custom diversity: According to the default behavior of the conformational analysis all of the resulting conformers are reported. However, it is possible to define a minimal diversity limit concerning the resulting structures. The use of this option leads to an even map of the explored conformational space using fewer, representative structures.

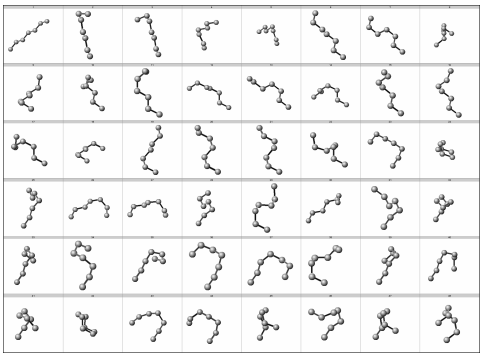


Figure 14. Generating all conformers of heptane results 48 different conformers (H atoms are removed after the coordinate generation process in order to enhance visibility).

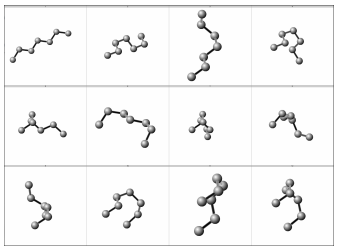


Figure 15. Using diversity limit 1.0 (RMSD in Å; including H atoms for best map) results 12 conformers.

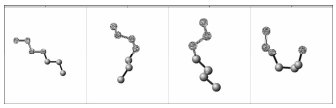


Figure 16. Using limit 1.2 results 4 conformers

Results, performance

The present method is capable of generating valid low energy conformers for a wide range of input structures. The latest version was tested on the NCI (National Cancer Institute) open database of 250251 structures (August 2000 version).

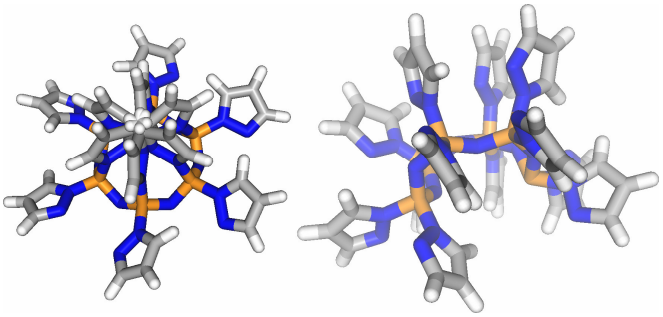


Figure 17. Comparison with Corina: 3D coordinates generated for a structure[8] from NCI. Left: coordinates generated with Corina Online Demonstration[9] showing multiple atom overlaps. Right: coordinates generated by our method integrated into Marvin.

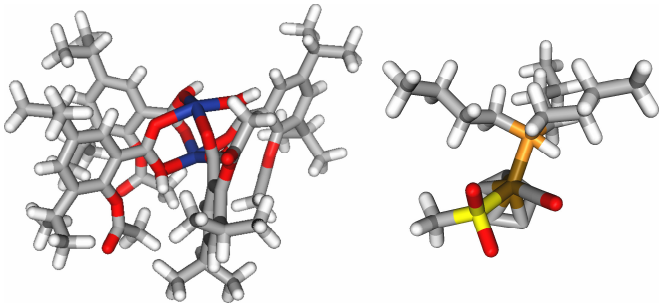


Figure 18. Another two structures[10,11] from the NCI data set. Coordinates generated by our method.

The coordinate generation primarily failed for 193 structures, that is 99.92% conversion rate[12]. The average conversion time was about 0.65 s per structure.

References, Notes

1. J. Sadowski and J. Gasteiger, "*From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders*", Chem. Rev., *93*, 2567-2581 (**1993**)
2. G. Imre, G. Veress, A. Volford and Ö. Farkas, "*Molecules from the Minkowski Space: An approach to building 3D molecular structures*", J. Mol. Struct. (Theochem), *666-667*, 51-59 (**2003**)
3. `http://www.chemaxon.com/marvin`
4. J. Weiser, M. C. Holthausen, L. Fitjer, "*HUNTER: A Conformational Search Program for Acyclic to Polycyclic Molecules with Special Emphasis on Stereochemistry*", J. Comput. Chem., *18*, 1265-1281 (**1997**)
5 S. L. Mayo, B. D. Olafson, W. A. Goddard III., J. Phys. Chem., 1990, 94, 8897-8909
6 Ö. Farkas, H. B. Schlegel, "*Geometry optimization methods for modelling large molecules*", J. Mol. Struct. (Theochem), *666-667*, 31-39 (**2003**)
7 `http://en.wikipedia.org/wiki/Quaternion`
8. `c1:c:n:n(:c1)P2(=NP(=NP(=NP(=NP(=N2)(n3:c:c:c:n3)n4:c:c:c:n4)(n5:c:c:c:n5)n6:c:c:c:n6)(n7:c:c:c:n7)n8:c:c:c:n8)(n9:c:c:c:n9)n%10:c:c:c:n%10)(n%11:c:c:c:n%11)n%12:c:c:c:n%12)n%13:c:c:c:n%13`
9. `http://www.mol-net.de/online_demos/corina_demo.html`
10. `CC(C)C1=CC(C2=[O+][Cu@-3]34[OH+]C(=[O+][Cu@@-3]([OH+]2)([OH+]C(=[O+]3)C5=C(OC(C)=O)C(=CC(=C5)C(C)C)C(C)C)[O+]=C([OH+]4)C6=C(OC(C)=O)C(=CC(=C6)C(C)C)C(C)C)C7=C(OC(C)=O)C(=CC(=C7)C(C)C)C(C)C)=C(OC(C)=O)C(=C1)C(C)C`
11. `CCCCP(CCCC)(CCCC)[Fe]1234(C#[O+])(C5C1=C2C3=C45)S(C)(=O)=O`
12. Test results on the same test set for the forthcoming release of the coordinate generation are: Failed for 0 structures, that is 100% conversion rate.