Knowledge Based Method for Building Molecular Models

Zhuming Ai* and Yu Wei

Laboratory of Molecular and Biomolecular Electronics, Southeast University, Nanjing, 210018, People's Republic of China

Received November 19, 1992

This paper presents a method for the generation of three-dimensional molecular models automatically using a knowledge base. The method builds the vast majority of molecular structures using a combined distance geometry and joining approach based on a knowledge base. The results are accurate enough to be used in computer-aided molecular design, and the method has no restriction such as those in template methods. The knowledge base can be set up from experimental data such as that in a crystallographic database.

INTRODUCTION

The three-dimensional (3D) geometry of a molecule is a very important property for studying the behavior of the molecule and its interactions with other molecules. The construction of 3D molecular models is the first step in theoretical calculations, such as quantum chemistry and molecular mechanics, for calculating the dipole moment, surface area, charge distribution, volume, steric congestion, and many other attributes. The known 3D molecular structures are obtained from experiments. At present, techniques such as X-ray diffraction and NMR can accurately determine molecular structure. However experimental measurement requires that the compound be synthesized. In computer-aided molecular design (CAMD), hypothetical molecules whose structures have not been determined experimentally are often studied. So non-experimental methods to build molecular models are of importance in CAMD.

A number of programs have been developed for constructing molecular models. Energy minimization by molecular mechanics or by quantum chemistry can calculate the molecular structure. But such programs still need an initial starting structure geometry. Interactive molecular graphics programs are often used to build 3D molecular models.[1] However, the quality of the results has a close relationship with the chemical knowledge of the operator, and the construction of closed ring systems is a difficult task. CAMSEQ uses templates to build the rings, and this method has been widely used to avoid the problem in building rings.[2] Because of the limitation of the template library (the target must be found in the library in order to get a result), there are restrictions in building macro-rings. CONCORD is a quite general program that obtains bond lengths from a table and bond angles and torsion angles partly from rules and partly from a minimization process. This method has been used to construct a large database of 3D coordinates from connection tables.[3]

Crippen has developed a distance geometry method that needs no template to construct molecular models, although the result is quite crude especially when the molecule contains a long chain.[4] Recently distance geometry has been used in molecular conformational searching.[5,6] Although in general distance geometry has low accuracy, it is a powerful tool for building macro-rings when the characteristics of macro-rings are taken into account, as shown in this paper.

Wipke and co-workers have proposed an automatic method for building molecular models using reasoning by analogy.[7] Their program AIMB can generate models without minimization. The program uses a knowledge base that requires a significant amount of disk space. AIMB is based on the template method, so the rings in the molecule to be built can not contain more than 15 atoms.

In the following sections, a molecular model building method will be proposed on the basis of a knowledge base of molecular fragmentary structure. First, we describe how the knowledge base is constructed in our approach. Then the method of building molecular models will be discussed. Finally, we illustrate the operation of the approach in a number of examples.

KNOWLEDGE BASE

There exist some molecular model building programs that use a database. CONCORD uses a simple database containing averaged bond lengths and bond angles. It is currently one of the best available methods for generating small-molecule 3D structures interactively. AIMB uses the Cambridge Crystallographic Database, which is a library of completely defined chemical structures with 3D coordination, as its knowledge base to build molecular models. To analogize any small part in a molecule, the effect of the long range interactions from the atoms in the entire molecule has to be considered, and such effects of the long range interactions constitute the majority of the information in the knowledge base.

From the principles of quantum chemistry, the 3D structure of a small part in a molecule was determined mainly by the atoms in the part and those near that part. Atoms far away from the part have relatively little contribution to the formation of the part. Because the bonded interactions are much greater than the nonbonded interactions in a molecule, the bond lengths and bond angles are determined almost totally by the nearby atoms with electron orbit overlapping. Then part of the structure in a molecule (the bond lengths, bond angles, and torsion angles between an atom and the nearby atoms) can be determined using molecular fragmentary structure information.

In the above viewpoint, there is redundant information in a knowledge base such as the one used by AIMB. If long distance interaction can be omitted, large amounts of information in such a knowledge base can be discarded.

In our knowledge base, a record is a molecular fragmentary structure or a conformational unit that includes the bond lengths, bond angles, and torsion angles of a center atom in a specific environment. The conformational unit includes a central atom and the bond length, bond angle, and torsion

* To whom correspondence should be addressed.

0095-2338/93/1633-0635$04.00/0

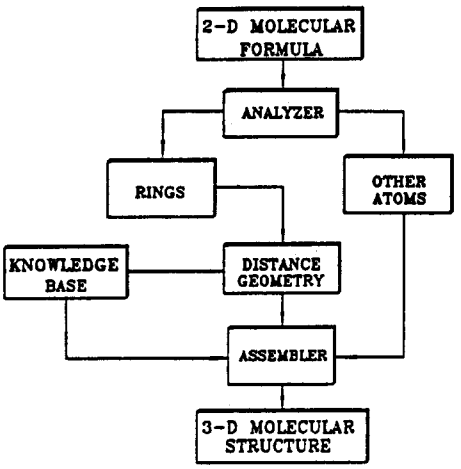

Figure 1. Diagram of building molecular models using knowledge base.

angle between the central atom and the nearby atoms. Environmental information is also included to distinguish the unit from other units with the same central atom but different nearby atoms, and thus different bond lengths, bond angles, and torsion angles.

For each central atom, the environmental information includes atomic number, degree of connection, bond type, equivalent atom, and ring code. The degree of connection is the number of atoms connected to this atom. The bond type is the type for each connection, such as single, double, triple, or π bond. The equivalent atom is used to represent the atoms connected to the central atom through a specific bond. It is defined as the sum of the atomic numbers of the atoms connected to the atom through the bond, or it is the "weight" of the bond. The long range interactions are represented by the equivalent atoms. The ring code denotes if the atom is part of a ring. If it is in a ring, the number of the members in the ring is also included. Bond angles are represented by the vectors with unit length showing the direction of the bonds. Torsion angles are those between the central atom and the "weightiest" atom connected to each bond, and also represented by vectors.

Such a knowledge base is small and easy to manage. It can be used in a desk top computer. To speed up the searching process, an index file can be generated, from which the position of a desired bond type in an atom can be directly reached.

Although the long range interactions are specifically discarded in a large scale, their effects on local geometry still remain in some degree. The field of equivalent atom in the knowledge base is used for this purpose. Anyway the less information about long range interactions may therefore tend to produce extended conformations. However the results are good enough to be used in CAMD.

METHOD FOR BUILDING THE MODELS

The method for building the 3D models of a molecular structure is based on joining. The molecule is divided into conformational units. There are two kinds of conformational units. One is an acyclic unit that contains only one central atom in the acyclic part of the molecule; the other is a cyclic unit, that is, a ring system. When a molecule is subdivided into conformational units, the molecule can be treated as a tree when rings are processed as nodes. Then the strategy for building a molecular model is shown in Figure 1.

(1) Analysis of the Molecular Formula. The molecular formula is represented as atomic numbers, connection tables, bond types, and the numbers of hydrogen connected to the atoms. The first stage of building a molecular model is to subdivide the molecular formula into conformational units. For the acyclic units, the algorithm is simple: one atom defines one unit. To solve the problem of ring construction, ring systems are treated as units, cyclic units. So the rings should be detected. A cyclic unit is defined as a group of one or more connected rings. If there is more than one ring in the cyclic unit, each ring must share at least two mutual atoms with another ring in the unit. Then fused and bridged ring systems are treated as one unit, while spiro rings are treated as more then one unit because a ring only shares one mutual atom with another ring.

A graph searching algorithm is used to find out the smallest set of smallest rings in the molecule. If there are more than two atoms shared in two rings, they are merged into one, and then all the cyclic units are determined.

(2) Method for Building Rings Automatically. For a ring, if the fragmentary structures of all atoms are determined, the ring can be determined. For the closure of rings, to build the ring's model from molecular fragmentary structure interactively is not an easy task. Using distance geometry, the problem can be solved naturally.

The fundamental problem of distance geometry is as follows: given the distance constraints that define a molecule, find the conformations that satisfy them.[8] Interatomic distances are represented by the distance matrix **D**. Its elements, d_{ij}, are the upper and lower bounds of atomic distances:

$$d_{ij} = \begin{cases} u_{ij}, & \text{upper bound of distance between atoms } i \text{ and } j \\ l_{ij}, & \text{lower bound of distance between atoms } i \text{ and } j \\ 0, & i = j \end{cases}$$

Because usually the information in the knowledge base is not sufficient to get all $u_{ij} = l_{ji}$, to determine the conformation precisely is difficult from the viewpoint of mathematics. So distance geometry has less accuracy, especially in building long chain molecular models. However in building rings the atomic distances have more restrictions than those in the chains; the upper and lower bounds in the distance matrix can be much closer. So distance geometry can get more accurate results in building rings.

Using the bond lengths and bond angles in the knowledge base, the 1–2 and 1–3 atomic distance in the rings can be determined. With triangle inequality and tetrangle inequality, the other bounds can be smoothed. The properties of rings made by using the bounds have more restrictions than that of the chains. Then the conformation can be calculated using distance geometry. Usually the conformation is not within the restriction of the distance matrix, so an error function representing the difference between molecular structure and the distance matrix should be built up and optimized. The properties of rings should be presented in the function. After the error function was optimized, a relatively accurate ring model was built.

The error function E should satisfy the following necessary conditions: (i) $E \geq 0$; (ii) $E = 0$ for all the molecular conformations if their atomic distances satisfy the restrictions of the distance matrix.

There are many forms of such a function E, for example,

$$E = \sum_{j<i} \begin{cases} (d_{ij}^2 - u_{ji}^2)^2, & \text{if } d_{ij} > u_{ji} \\ (d_{ij}^2 - l_{ij}^2)^2, & \text{if } d_{ij} < l_{ij} \\ 0, & \text{others} \end{cases} \quad (1)$$

From the viewpoint of mathematics, if $E = 0$, a correct model

is built up. However there are two special situations. One is that there is more than one conformation satisfying the distance matrix with their $E = 0$, but the function E has no ability to confirm which one is the best. The other is that there is no conformation satisfying the distance matrix, or there is no set of points in 3D space in which the interpoint distances satisfy the restriction of the distance matrix. Distance geometry cannot solve the last problem, but in practice we expect a conformation that is near the restriction of the matrix. Although a conformation can be obtained when E is optimized, it cannot be confirmed if it is the best one from the viewpoint of chemistry.

The error function must not only satisfy the necessary condition but also represent chemistry concepts. Molecular mechanics has provided theories as well as force field functions. Although the functions are not suitable to be used directly as the error functions, for at least they do not satisfy the two necessary conditions, quite similar functions can be induced as follows:

$$E = \begin{cases} E_e, & \text{if } d_{ij} > u_{ij} \text{ or } d_{ij} < l_{ji} \\ 0, & \text{others} \end{cases} \quad (2)$$

where

$$E_e = \sum \begin{cases} k_1(d_{ij} - u_{ij})^2, & \text{for 1–2 interactions} \quad (3) \\ k_2(d_{ij} - u_{ij})^2, & \text{for 1–3 interactions} \quad (4) \\ k_3[(u_{ij}/d_{ij})^{12} - 2(u_{ij}/d_{ij})^6], & \text{others} \quad (5) \end{cases}$$

Obviously this function E satisfies the necessary conditions of error functions in distance geometry; that is, only when the atomic distances are out of the range of the distance matrix do they contribute to the error.

This error function represents the concept of molecular mechanics. Expressions 3 and 5 are very similar to the 1–2 and 1–4 interaction force field functions in molecular mechanics, respectively. The 1–3 interaction force field function in molecular mechanics is

$$V_\theta = (1/2)k_\theta(\theta - \theta_0)^2$$

To simplify the calculation, expression 4 is obtained in which $(\theta - \theta_0)^2$ is replaced by $(d_{ij} - u_{ij})^2$, and the constant k varies correspondingly.

The constants k_1, k_2, and k_3 are quite different from their counterparts in molecular mechanics. Their values are not so critical as those in molecular mechanics because of the effect of the distance matrix. The values actually show the ratio of contributions to the error function from the bond lengths, bond angles, and torsion angles, respectively. The values can be adjusted by the user. Although the optimization of the constants remains a problem, the values from simply trying are good enough.

The problem of stereochemistry is not concerned in the error function. In most situations it is not harmful since we have developed a separate program to generate symmetric structures.

With this error function, which combines the chemistry concept in it, an optimized molecular conformation can be obtained by the optimization of the function. The conformation is either the optimized one in the conformations satisfying the distance matrix or the nearest one from the restriction of the matrix when there is no conformation satisfying the matrix.

(3) Molecular "Tree" Construction. When all the rings are built, the molecule can be joined together as a tree if a ring is treated as a virtual atom or a node, and a joining method can be used.

When the joining process reaches a ring or it grows from a ring, it needs an algorithm to generate the coordinates of the atoms connected to the ring. This is due to the fact that the ring is built up by distance geometry, and the positions of the atoms nearby do not exactly match the molecular fragmentary structures from the knowledge base. So the atoms connected to rings cannot be directly generated by simply transition and rotation. Here an error minimization method is used.

For the atom under processing, suppose three nearby atoms in the ring (if less than three, augment 0s) need to be matched, which composes matrix **A**. Matrix **B** contains corresponding atoms in the knowledge base; it should be transmitted to match matrix **A**. Then an error matrix can be calculated

$$\mathbf{E} = \mathbf{BT} - \mathbf{A} = \begin{bmatrix} e_{00} & e_{01} & e_{02} \\ e_{10} & e_{11} & e_{12} \\ e_{20} & e_{21} & e_{22} \end{bmatrix}$$

where **T** is a 3×3 rotation matrix with three independent variables x_1, x_2 and x_3. For simplicity, suppose **A** has been transmitted to the origin, and it is easy to be transmitted back to the previous position.

Now the error function which needs to be minimized can be generated

$$F(x_1,x_2,x_3) = \sum_{i,j=0}^{2} e_{ij}^2$$

For minimizing the error function F, a Lagrange multiplier method can be used.[9]

RESULTS AND DISCUSSION

Using the approach discussed above, with a very simple molecular fragmentary structure knowledge base, some molecular models have been constructed.

The first molecular model to be built is a crown ether. It is a macro-ring with 21 atoms. Template methods such as AIMB proposed by Wipke look at the knowledge base to see whether there is a ring template containing 21 atoms. Because the knowledge base used in AIMB only contains the ring templates including up to 15 atoms, this model exceeds the range of Wipke's method.

The root mean square (RMS) of the difference between the molecular model built by our program and that determined by experiment has been used to measure the accuracy of the result:

$$\mathrm{RMS} = \{\{\sum_i [(x_{ei} - x_{bi})^2 + (y_{ei} - y_{bi})^2 + (z_{ei} - z_{bi})^2]\}/N\}^{1/2}$$

where x_{ei}, y_{ei}, and z_{ei} are coordinates determined using X-ray diffraction[10] and x_{bi}, y_{bi}, and z_{bi} are built up coordinates. The molecular structure has been build up as solid lines in Figure 2 with RMS = 0.743.

We find that the structure near the small ring is better; this is due to the fact that in this part there are more restrictions in the distance matrix. We have used the error function, eq 1, to build the model; one of the results is shown in Figure 2. The result varies greatly from the different initial data of the distance matrix, and the conformation is not extended as the result from eq 2.

Figure 3 is another example; it is the phosporinane oxide built by our program and that from experiment.[11]

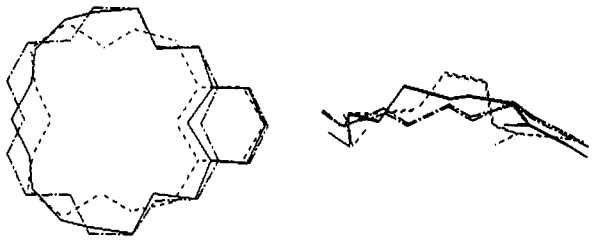

Figure 2. Results of our approach and experiments: (- -) result from eq 1; (—) result from eq 2; (– • –) result from experiment.

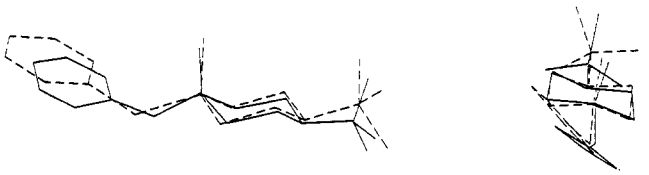

Figure 3. Phosporinane oxide built up model (dashed lines) and crystal structure (solid lines).

The comparison shows that the molecular models built by our program and the experiment results are well matched. It needs to be noted that the results presented here are based on a very simple molecular fragmentary structure knowledge base that only contains the standard bond lengths and bond angles. With more knowledge in the base, better results are expected.

CONCLUSIONS

The approach for building molecular models automatically has been proposed in this paper. The result is accurate enough to be used as the initial starting data for the theoretical methods such as molecular mechanics and quantum chemistry, or other needs of CAMD. The method overcomes the difficulties in building rings, and the closed rings can always be achieved. The knowledge base needed in the method is small, and the method has no restriction as found in the template method. The method can be used in a desk top computer.

REFERENCES AND NOTES

(1) Ai, Z.; Li, G.; Wei, Y. MOLMO: Interactive Molecular Graphics on a Personal Computer. *Huaxue Tongbao* **1991**, *8*, 59–62.
(2) Potenzone, R., Jr.; Canicchi, E.; Weintraub, H. J. R.; Hopfinger, A. J. Molecular Mechanics and the CAMSEQ Processor. *Comput. Chem.* **1977**, *1*, 187–194.
(3) Rusinko, A.; Sheridan, R. P.; Nilakantan, R.; Haraki, K. S.; Bauman, N.; Venkataraghavan, R. Using CONCORD to Construct a Large Database of Three-Dimensional Coordinates from Connection Tables. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251–255.
(4) Crippen, G. M. A Novel Approach to Calculation of Conformation: Distance Geometry. *J. Comput. Phys.* **1977**, *24*, 96–107.
(5) Leach, A. R.; Smellie, A. S. A Combined Model-Building and Distance Geometry Approach to Automated Conformational Analysis and Search. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 379–385.
(6) Wenger, J. C.; Smith, D. H. Deriving Three-Dimensional Representations of Molecular Structure from Connection Tables Augmented with Configuration Designations Using Distance Geometry. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 29–34.
(7) Wipke, W. T.; Hahn, M. A. Analogy and Intelligence in Model Building. In *Artificial Intelligence Applications in Chemistry*; Pierce, T., *et al.*, Eds.; ACS Symposium Series No. 306; American Chemical Society: Washington, DC, 1986; pp 136–146.
(8) Crippen, G. M.; Harvel, T. F. *Distance Geometry and Molecular Conformation*; Research Studies Press: Taunton, England, 1988.
(9) Bertsekas, D. P. *Constrained Optimization and Lagrange Multiplier Methods*; Academic Press: New York, 1982.
(10) van Staveren, C. J.; Aarts, V. M. L. J.; *et al. J. Am. Chem. Soc.* **1986**, *108*.
(11) Haque, M.; Ahmed, J.; Horne, W. *Acta. Crystallogr.* **1986**, *C42*, 99, 5271–5276.