(30) Dias, J. R. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 124.
(31) Cyvin, S. J.; Cyvin, B. N.; Brunvoll, J. *Chem. Phys. Lett.* **1987**, *140*, 124.
(32) Cyvin, S. J.; Cyvin, B. N.; Brunvoll, J. *Chem. Phys. Lett.* **1989**, *156*, 595.
(33) Chen, R. S.; Cyvin, S. J. *THEOCHEM* **1989**, *200*, 251.
(34) Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N. *THEOCHEM* **1990**, *205*, 361.
(35) Balaban, A. T.; Brunvoll, J.; Cioslowski, J.; Cyvin, B. N.; Cyvin, S. J.; Gutman, I.; He, W. C.; He, W. J.; Knop, J. V.; Kovačević, M.; Müller, W. R.; Szymanski, K.; Tošić, R.; Trinajstić, N. *Z. Naturforsch.* **1987**, *42A*, 863.
(36) He, W. J.; He, W. C.; Wang, Q. X.; Brunvoll, J.; Cyvin, S. J. *Z. Naturforsch.* **1988**, *43A*, 693.
(37) Dias[24] reported the numbers 17, 68, and 322 of the isomers $C_{40}H_{20}$, $C_{44}H_{22}$, and $C_{48}H_{24}$, respectively. These numbers deviate from those of Table I. Dias did not mention specifically the helicenic systems in this connection, but we find that even the inclusion of such systems does not explain the discrepancies. We have identified 1, 2, and 25 helicenic quasi-coronoids with h = 10, 11, and 12, respectively. In private communication in 1989, Dias admitted errors in his analysis.
(38) Knop, J. V.; Szymanski, K.; Jeričević, Ž.; Trinajstić, N. *MATCH* **1984**, *16*, 119.
(39) He, W. C.; He, W. J. *Theor. Chim. Acta* **1985**, *68*, 301.
(40) He, W. C.; He, W. J. *Tetrahedron* **1986**, *42*, 5291.
(41) Cioslowski, J. *J. Comput. Chem.* **1987**, *8*, 906.
(42) Nikolić, S.; Trinajstić, N.; Knop, J. V.; Müller, W. R.; Szymanski, K. *J. Math. Chem.*, in press.
(43) Brunvoll, J.; Cyvin, B. N.; Cyvin, S. J.; Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N. *THEOCHEM* **1990**, *207*, 131.
(44) Knop, J. V.; Szymanski, K.; Jeričević, Ž.; Trinajstić, N. *J. Comput. Chem.* **1983**, *4*, 23.
(45) Trinajstić, N.; Jeričević, Ž.; Knop, J. N.; Müller, W. R.; Szymanski, K. *Pure Appl. Chem.* **1983**, *55*, 379.
(46) Stojmenović, I.; Tošić, R.; Doroslovački, R. *Graph Theory*, Proceedings of the Sixth Yugoslav Seminar on Graph Theory, Dubrovnik, April 18-19, 1985; Tošić, R., Acketa, D., Petrović, V., Eds.; University of Novi Sad: Novi Sad, 1986; p 189.
(47) Trinajstić, N. Computerized generation and enumeration of polyhexes conducted by the Düsseldorf-Zagreb research group. Private communication, 1989.
(48) Cyvin, S. J.; Brunvoll, J.; Cyvin, B. N.; Bergan, J. L.; Brendsdal, E. *Symmetry*, in press.
(49) Brunvoll, J.; Cyvin, B. N.; Cyvin, S. J. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 171.
(50) Brunvoll, J.; Cyvin, S. J.; Cyvin, B. N. *J. Comput. Chem.* **1987**, *8*, 189.
(51) Cyvin, S. J.; Brunvoll, J.; Cyvin, B. N. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 236.
(52) Cyvin, S. J.; Brunvoll, J.; Cyvin, B. N. *Struct. Chem.*, in press.
(53) Randić, M. *J. Chem. Soc., Faraday Trans. 2* **1984**, *72*, 232.
(54) Cyvin, S. J.; Gutman, I. *MATCH* **1986**, *19*, 229.
(55) Gordon, M.; Davison, W. H. T. *J. Chem. Phys.* **1952**, *20*, 428.

Global Energy Minimization by Rotational Energy Embedding

GORDON M. CRIPPEN*

College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109

TIMOTHY F. HAVEL

Biophysics Research Division, University of Michigan, Ann Arbor, Michigan 48109

Received January 10, 1990

Given a sufficiently good empirical potential function for the internal energy of molecules, prediction of the preferred conformations is nearly impossible for large molecules because of the enormous number of local energy minima. Energy embedding has been a promising method for locating extremely good local minima, if not always the global minimum. The algorithm starts by locating a very good local minimum when the molecule is in a high-dimensional Euclidean space, and then it gradually projects down to three dimensions while allowing the molecule to relax its energy throughout the process. Now we present a variation on the method, called rotational energy embedding, where the descent into three dimensions is carried out by a sequence of internal rotations that are the multidimensional generalization of varying torsion angles in three dimensions. The new method avoids certain kinds of difficulties experienced by ordinary energy embedding and enables us to locate conformations very near the native for avian pancreatic polypeptide and apamin, given only their amino acid sequences and a suitable potential function.

INTRODUCTION

The fundamental problem in the conformational analysis of large molecules is that the conformers of physical interest correspond to the best few local energy minima out of a total number of minima that apparently increases exponentially with the size of the molecule. An increasingly popular approach is to rely on large but insufficient amounts of experimental data to restrict the geometric possibilities so much that sampling experimentally allowed conformations by distance geometry embedding or simulated annealing produces structures relatively near the correct state(s). Then energy refinement by constrained molecular dynamics can produce a relatively small number of low energy conformers satisfying all the experimental constraints. For a sampling of recent activity along these lines, see the references.[1-9] On the other hand, suppose we have very little experimental information beyond standard bond lengths, bond angles, and van der Waals radii. Then we are forced to rely almost exclusively on some given energy function to guide the search for the best conformations. If we treat the calculation of internal energy as a black box, where we put atomic coordinates in and we get an energy value and its gradient out, then global optimization methods hold little hope for finding global and nearly global energy minima without spending an exponentially increasing amount of computer time for larger and larger molecules.[10] One approach to escaping from this difficulty is to build into the global search procedure more a priori knowledge about the energy function. Our particular efforts along this line have concentrated on exploiting the large geometric component to the problem, which is made clearest when the energy function, E, is some sort of classical molecular mechanics potential. Such energies are calculated largely as a sum of atom–atom pairwise interactions, where the n atoms are treated as points engaged in isotropic interactions, and for each term there is often a unique, finite, optimal separation. Although we are of course interested in simulating molecules in ordinary three-dimensional space, $\mathcal{R}^3$, the *energy embedding* algorithm[11-14] first locates a good local minimum in $\mathcal{R}^{n-1}$, where there are few

0095-2338/90/1630-0222$02.50/0 © 1990 American Chemical Society

minima because most of the optimal interatomic separations can be achieved simultaneously. Then it seeks to gradually reduce the $(n - 1)$-dimensional conformation to a three-dimensional one while keeping the energy as low as possible. Our experience with the method is that it usually locates extremely good local minima in $\mathcal{R}^3$, but not necessarily *the* global minimum.

There are three major difficulties with energy embedding. One is that intrinsic torsional potentials are functions of the relative positions of four atoms defining the dihedral angle, and four atoms span at most a three-dimensional subspace, even in $\mathcal{R}^{n-1}$. As we have reported earlier,[14] that means that any procedure employing high-dimensional spaces still has large energy barriers for cis/trans peptide bond isomerism, and indeed, the four peptide bonds in cyclotetrasarcosyl[15] give rise to $2^4 = 16$ local minima in $\mathcal{R}^{n-1}$.

The second difficulty is that the path the algorithm takes while reducing the dimensionality is often difficult to follow, numerically speaking. In order to define the path, we first need some notation. Consider the energy, $E(\mathbf{x},\mathbf{y})$, to be a function of $\mathbf{x}$, the vector of the first three coordinates of all n atoms, and $\mathbf{y}$, the vector of coordinates 4 through $n - 1$ of all the atoms. Thus x has $3n$ components, and $\mathbf{y}$ has $n(n - 4)$ components. Then the "thickness" of the molecule outside of $\mathcal{R}^3$ can be measured by $\|\mathbf{y}\|^2$, which has an initial value in $\mathcal{R}^{n-1}$ denoted by $t_0 > 0$. Then the dimensionality reduction part of energy embedding can be described as

$$\text{minimize} \quad E(\mathbf{x},\mathbf{y}) \quad \text{subject to} \quad \sum_i y_i^2 = t \qquad \text{as } t = t_0 \text{ initially and } t \rightarrow 0 \tag{1}$$

This could in principle be solved as a constrained optimization problem by defining the Lagrangian

$$L(\mathbf{x},\mathbf{y},\lambda) = E(\mathbf{x},\mathbf{y}) + \lambda[\sum_i y_i^2 - t] \tag{2}$$

and solving the first order optimality conditions, namely $\nabla L = 0$, or more explicitly

$$\begin{gathered} \frac{\partial E}{\partial x_i} = 0 \\ \vdots \\ \frac{\partial E}{\partial y_i} + 2\lambda y_i = 0 \\ \vdots \\ \sum_i y_i^2 - t = 0 \end{gathered} \tag{3}$$

If we consider $\mathbf{x}$ and $\mathbf{y}$ to be functions of t, then implicit differentiation of this set of equations leads to a system of differential equations that can be solved for $\mathbf{x}(t)$ and $\mathbf{y}(t)$:

$$\begin{pmatrix} \frac{\partial^2 E}{\partial x_i \partial x_j} & \cdots & \frac{\partial^2 E}{\partial x_i \partial y_j} & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ \frac{\partial^2 E}{\partial y_i \partial x_j} & \cdots & \frac{\partial^2 E}{\partial y_i \partial y_j} + 2\lambda\delta_{ij} & \cdots & 2y_i \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 2y_j & \cdots & 0 \end{pmatrix} \begin{pmatrix} \frac{\mathrm{d}x_i}{\mathrm{d}t} \\ \vdots \\ \frac{\mathrm{d}y_i}{\mathrm{d}t} \\ \vdots \\ \frac{\mathrm{d}\lambda}{\mathrm{d}t} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 1 \end{pmatrix} \tag{4}$$

The matrix is just the Hessian of the Lagrangian, $\nabla^2 L$, where δ_{ij} is the Kronecker delta. It is computationally inefficient, but certainly possible to solve these differential equations along with the initial values $\mathbf{x}(t_0)$ and $\mathbf{y}(t_0)$ at the high-dimensional minimum to trace out the precise path taken in $\mathcal{R}^{n-1}$ by each atom as the molecule flattens out to lie entirely in $\mathcal{R}^3$. Initially the path can be computed in this way, since energy embedding offers initially a unique route away from the high-dimensional starting minimum,[14] given an initial choice of what three-dimensional subspace constitutes $\mathbf{x}$. The problem is that even in very small test cases, the Hessian becomes singular further along the path, so that solving the linear system for the derivatives becomes ill-conditioned. Apparently this is in the vicinity of a bifurcation in the path, a situation that is usually skipped over by the current energy embedding program, which takes much larger steps and therefore falls onto one or the other of the alternative paths in its more rapid approach to $\mathcal{R}^3$. More carefully tracing out the alternatives might lead to important local minima, but the computational cost would be quite high.

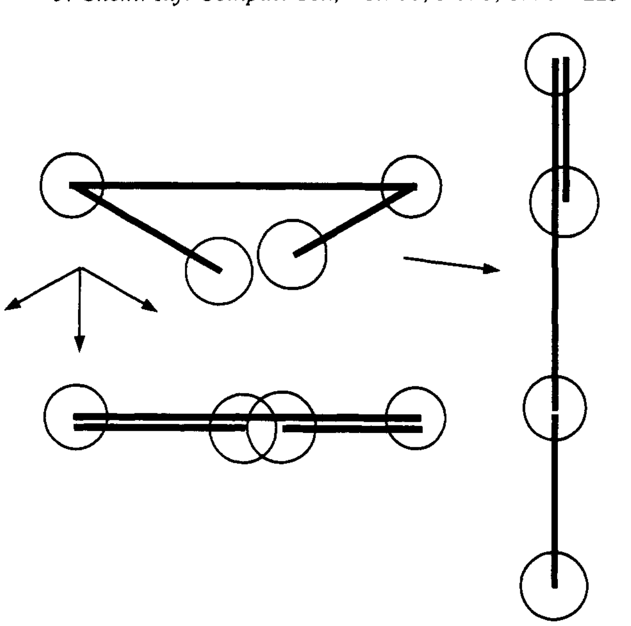

Figure 1. Different ways of projecting a simplified molecule from two to one dimension. Heavy lines indicate energetically important interactions, such as bond stretching. Circles show atomic van der Waals radii, such that overlapping circles indicate extremely unfavorable interactions. The axis to be reduced to zero, denoted by y in the text, is represented by arrows.

The third difficulty in energy embedding is that the dimensionality reduction is basically a modified projection, and sometimes all projections lead to a bad final conformation. Consider the very simple example shown in Figure 1, where $n = 4$, initially the molecule is in $\mathcal{R}^2$, and a final conformation in $\mathcal{R}^1$ is sought. The figure shows different choices of the one-dimensional subspace y to be projected out by arrows, and for each such choice, x is a line perpendicular to the arrow. Although the molecule will rotate as a whole and substantially change its conformation as $t \rightarrow 0$, for most choices of y (indicated by a fan of arrows), the result is the lower linear conformation, which suffers from bad van der Waals contacts because the strong bond stretching interactions have been largely preserved by reducing the two acute bond angles to zero. Some very carefully chosen orientations of y produce better conformations, such as the linear structure on the right, where the left acute bond angle has opened out to π, the molecule as a whole has rotated counterclockwise by roughly $\pi/2$, and the right acute angle has folded down to zero. However, there is no projection direction that yields the fully extended structure (not shown).

What seems to be needed is a variation of energy embedding that can more freely explore alternative ways of decreasing the dimensionality while overcoming rotational barriers. The new algorithm we call "rotational energy embedding", and it is described in detail in the next section. Finally, we illustrate its use on some small test problems.

METHODS

As in the previous versions of energy embedding, rotational energy embedding consists of first locating a local energy minimum in a high-dimensional space, and then transforming the conformation to one in $\mathcal{R}^3$ without increasing the energy too much in the process. We have made substantial methodological improvements on both parts.

The first step is to find a set of atomic Cartesian coordinates, $\mathbf{c}$, in $\mathcal{R}^{n-1}$ for the n atoms in the molecule such that the energy, $E(\mathbf{c})$, is minimal. Suppose E consists of a sum of pairwise interatomic interactions such that each term has a unique optimal separation, s_{ij}. Set up a matrix of squared optimal separations, $S = (s_{ij}^2)$, letting all diagonal elements $s_{ii}^2 = 0$. If two atoms repel each other at all distances, take that s_{ij} to be the triangle inequality limit:

$$s_{ij} = \min_k (s_{ik} + s_{kj}) \ \forall \ k = 1, ..., n \qquad (5)$$

If there is more than one local minimum for the interaction between two atoms, choose the distance corresponding to the deepest (most favorable) one. In other words, we construct an extremely low energy conformation in terms of distances by considering each two-body interaction individually, neglecting all other atoms in the molecule for the moment. Similarly, for a group of three atoms involving a bond angle bending term, or for four atoms involving bond angle bending and intrinsic torsional potentials, choose the optimal separations among the group to correspond to the lowest energy minimum of the group in isolation, insofar as possible.

If there is a set of coordinates $\mathbf{c} \in \mathcal{R}^{n-1}$ such that the squared interatomic distances exactly match S, then we say S is embeddable, and the conformation is the global energy minimum for that molecule in a space of any number of dimensions. However, in our experience with reasonable energy functions, S is generally not embeddable when $n > 5$. Thus the problem is to find a matrix of squared distances D that is "close" to S, but that is also embeddable in $\mathcal{R}^{n-1}$. One way is to simply apply the EMBED algorithm,[11] where the squared distances are converted to the corresponding metric matrix, that matrix is diagonalized, and the coordinates are calculated from all m_E eigenvectors corresponding to positive eigenvalues. Negative or zero eigenvalues result in axes where each atom has a zero coordinate. The resulting conformation is generally under some strain from incompatible demands on the distances and occupies only an m_E-dimensional subspace, where $m_E \approx n/2$. This set of coordinates $\mathbf{c} \in \mathcal{R}^{m_E}$ is an optimal match to S in that the metric matrix calculated from the coordinates is as close as possible in the matrix spectral sense to the metric matrix calculated from S, subject to the embeddability constraint that the coordinates' metric matrix can have no negative eigenvalues.[11,16]

Another approach to generating coordinates is the MAP algorithm of Glunt et al.[17] which finds D such that the Frobenius norm

$$\|S - D\| = \sum_{ij} (s_{ij}^2 - d_{ij}^2)^2 \qquad (6)$$

is minimized. Then D is converted to the corresponding metric matrix, which has only m_M positive eigenvalues and the rest zero, so that we directly obtain $\mathbf{c} \in \mathcal{R}^{m_M}$. We find the dimensionality of the coordinates generated by MAP tends to be lower than that resulting from EMBED: $m_M < m_E$. By either method, D does not optimally resemble S in an energy sense, and we must apply a local unconstrained minimization procedure, which stays in $\mathcal{R}^{m_M}$ for MAP coordinates, and which flattens out into an m_M-dimensional subspace for EMBED coordinates. Our experience has been that we reach the same conformation either way, or even if we make small coordinate perturbations before energy minimization so that the atoms initially span $\mathcal{R}^{n-1}$. This is the sort of behavior one would expect from the general theory of tensegrity frameworks,[11] because the molecule is under stress from the competition among some interactions striving to expand ($d_{ij} < s_{ij}$) and others striving to compress ($d_{ij} > s_{ij}$). It is not clear which of the two methods is computationally more efficient, since EMBED requires only one matrix diagonalization, compared to several for MAP, but MAP produces coordinates in $\mathcal{R}^{m_M}$ even before energy minimization, thus eliminating variables during minimization.

The second step in rotational energy embedding is to adjust the atomic coordinates so that $\mathbf{c} \in \mathcal{R}^3$ while increasing the energy as little as possible. The fundamental assumption is that the interaction energies between atoms tend to be more important when there are fewer covalent bonds separating them. Thus searching for low energy conformations by varying torsion angles is likely to be fruitful because only long-range interactions are varied, while bond lengths and vicinal bond angles remain fixed at their energetically optimal values. The logic is a little involved in order to treat branched chains in a numericlly stable fashion, but it reduces to a simple process in the example of a long unbranched carbon chain initially in three dimensions that we want to make planar. Just start at one end, noting that the first three atoms (numbered 1, 2, and 3) define a plane that will become the plane of the entire chain. Rotate about the 2–3 bond so as to bring atom 4 into this plane in either the cis or trans conformations, whichever is energetically preferable. Next rotate about the 3–4 bond to bring atom 5 into the plane, and so on down the chain. Note that cis or trans are the only two choices for every dihedral angle, because we want the final structure to be planar. Subsequently, the process could be continued to make the chain linear. Atoms 1 and 2 define a line, and changing the 1–2–3 bond angle to either 0 or π will bring atom 3 into that line. Then adjust the 2–3–4 bond angle, and so on until the end of the chain.

In order to explain the general dimensionality reduction process, we must first consider the molecule in question to be a tree graph, where the atoms are the nodes and the covalent bonds are the edges. If the molecule is cyclic, such as a polypeptide with disulfide crosslinks, each ring must be broken by deleting one bond for the purposes of defining the molecular tree, although the corresponding bond stretching terms and so on remain in the energy function. A rigid ring, such as benzene, could be retained as a local ring node in the tree having no internal degrees of freedom and moving as a rigid group, but we have not actually implemented such a feature. Following the usual terminology for tree graphs, we can choose one atom to be the "root", said to be located at the "top" of the tree, and then edges are said to point "down" from a "father" node to its "sons". Figure 2 shows an example of such a molecular graph, where atom 1 is the root, atom 3 is the father of atom 4, and atoms 7 and 8 are the sons of atom 6, for instance. In order to make this example easy to visualize, suppose the molecule is initially three-dimensional, and we want to reduce it to two dimensions. In the starting conformation we have illustrated, atoms 1–7 already lie in the plane, and the rest are out of plane in the third dimension. The objective is to define a sequence of rotations to bring the right half of the molecule into the plane, starting by bringing atom 8 into the plane. Choose atoms 6, 7, and 4 to define the plane because they are noncollinear and they are as close to atom 8 as possible in terms of graph distance, or number of bonds separating them from 8. Move the closest atom, 6, to the origin by a rigid translation of the whole molecule, then put one of the next closest atoms, 4, on the x-axis by a rigid rotation, followed by placing atom 7 in the xy-plane by another rigid rotation. The equations for accomplishing this, called a "canonical translation and rotation", are given below. Finally, rotate atoms 8–11 about the x-axis, such that atom 8 lies in the plane, either cis or trans to atom 7. (Here we employ something of a generalization of the chemical usage for cis

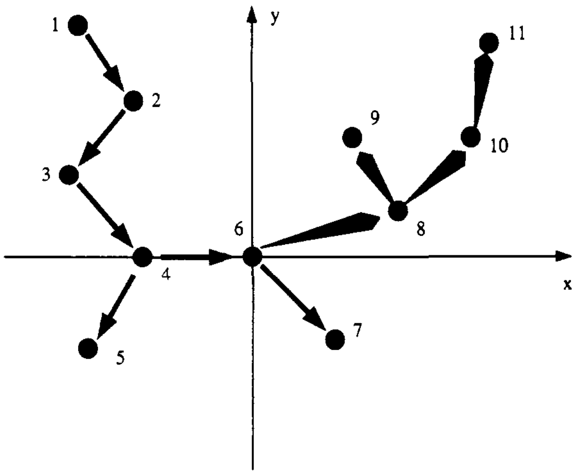

Figure 2. Example of a three-dimensional molecule being embedded in the plane. The heavy arrows are the (directed) covalent bonds defining the molecular tree, and the atoms are heavy dots marked by an arbitrary numbering scheme. Atoms 1–7 are in the xy plane, and atoms 8–11 are out of plane.

and trans to mean same and opposite side, respectively, of the x-axis in this case.) Continue on down the tree until all atoms lie in the plane.

The ability of this procedure to explore all possible assignments of cis and trans for the bonds enables it to discover which configurations are compatible with a low value of the energy. Similarly, when we use the analogous procedure to transform four-dimensional conformations into three-dimensional ones, we have the option of searching for different low energy diastereomers. Even in the absence of asymmetric carbons, the method can systematically search for low energy combinations of rotomeric states for bonds having energy minima separated by barriers that could not be crossed by oridinary energy minimization techniques. In general, rotational energy embedding carries out a broad search for low energy conformations each time the dimensionality is reduced by one. We believe this is the reason the method has proved so powerful.

What follows in a description of the general algorithm for going from a high-dimensional space to three dimensions. Each layer of nesting of loops or consequences of logical tests is indicated by another vertical bar. Otherwise the sequence of steps performed by the algorithm runs from top to bottom in the obvious way.

```
Start with the molecule consisting of $n$ atoms having coordinates in $\mathcal{R}^m$ where $3 < m \le n - 1$.
While $m > 3$...
    Starting at the root of the connectivity tree, find $m$ atoms in a breadth-first search whose affine span is an $m - 1$ dimensional subspace. Mark these atoms as "flat".
    While there are atoms still not marked as flat...
        Locate a nonflat atom $a_i$ having a flat father.
        This father is the first of $m$ atoms defining the rotation.
        In a breadth-first search both up and down the connectivity tree, locate a total of $m$ flat atoms, ordered in a list according to increasing graph distance from $a_i$.
        Perform the canonical rigid translation and rotation according to the list of $m$ flat atoms (eqs 8 and 9).
        Rotate $a_i$ to the cis and then trans conformation, moving with it all atoms below it in the tree (eq 11).
        Leave the molecule in the energetically better conformation.
            When $m < 6$, compare energies after performing a local minimization in each conformation by minimizing with respect to the $m - 1$ coordinates of each flat atom and the $m$ coordinates of the nonflat atoms.
        Mark $a_i$ as flat.
    A canonical rigid translation and rotation of the molecule should reveal that it now lies in $\mathcal{R}^{m-1}$.
    Minimize $E$ with respect to the $m - 1$ coordinates of each atom.
    Reduce $m$ by one.
Dimension of molecule is now 3. Quit.
```

(When we say above that the affine span of m atoms is an $m - 1$ dimensional subspace, we mean there is no lower dimensional subspace that contains these m atoms.)

To complete the description of our methods, we need to give the equations for rotations in high-dimensional spaces. The basic concept is that any rigid rotation in $\mathcal{R}^m$ can be viewed as a sequence of elementary rotations $R(i,j,\theta)$, each of which changes the ith and jth coordinates of each atom by a rotation through angle θ, but leaves all the rest unchanged. In other words, an elementary rotation matrix is just an $m \times m$ unit matrix except for the elements $r_{i,i}$, $r_{j,j}$, $r_{i,j}$, and $r_{j,i}$:

$$R(i,j,\theta) = \begin{pmatrix} 1 & 0 & \cdots & & & & \\ 0 & \ddots & & & & & \\ \vdots & & \cos\theta & \cdots & -\sin\theta & & \\ & & \vdots & \ddots & \vdots & & \\ & & \sin\theta & \cdots & \cos\theta & & \vdots \\ & & & & & \ddots & 0 \\ & & & & \cdots & 0 & 1 \end{pmatrix} \qquad (7)$$

Now the canonical transformation referred to in the algorithm above amounts to numbering the atoms 1, 2, ..., n, and wanting atom 1 at the origin, atom 2 along the first coordinate axis, atom 3 in the plane of the first two axes, etc. Let $\mathbf{c}_i = [c_{i,1}, c_{i,2}, ..., c_{i,m}]$ be the vector of coordinates of atom i, and $\mathbf{c}_i'$ be its new, transformed coordinates. Then the canonical transformation consists of first translating

$$\mathbf{c}_i' = (\mathbf{c}_i - \mathbf{c}_1) \ \forall\ i = 1, ..., n \qquad (8)$$

and then the nested sequence of rotations indicated by

$$\mathbf{c}_i' = R(k-1,j,\theta)\mathbf{c}_i \quad \{\{\{\forall\ i = 1, ..., n\}\ \forall\ j = k, ..., m\}\ \forall\ k = 2, ..., m\} \qquad (9)$$

where

$$\theta = \arctan\left(-\frac{c_{k,j}}{c_{k,k-1}}\right) \qquad (10)$$

The rotation angle is taken to be either the value θ calculated from eq 10 or $\theta + \pi$, chosen so that $c'_{k,k-1} > 0$. As an illustration of the use of eqs 8–10, suppose we have atoms 1, 2, 3, and 4 = n with arbitrary coordinates in 3 = m dimensions. Then eq 8 translates the molecule so that all coordinates of atom 1 are zero, but the other atoms have in general non-zero coordinates. Subsequently eq 9 starts with atom k = 2 and rotates all atoms in the 1–2 and 1–3 planes so that the second and third coordinates of atom 2 are zero, while its first coordinate may be non-zero. Then for atom k = 3, we rotate all atoms in the 2–3 plane so that the third coordinate of atom 3 is zeroed out. Atom 4 is left in general with three non-zero coordinates.

For the cis/trans rotations alluded to in the algorithm, the canonical transformation has arranged it so that we want $c'_{i,m}$

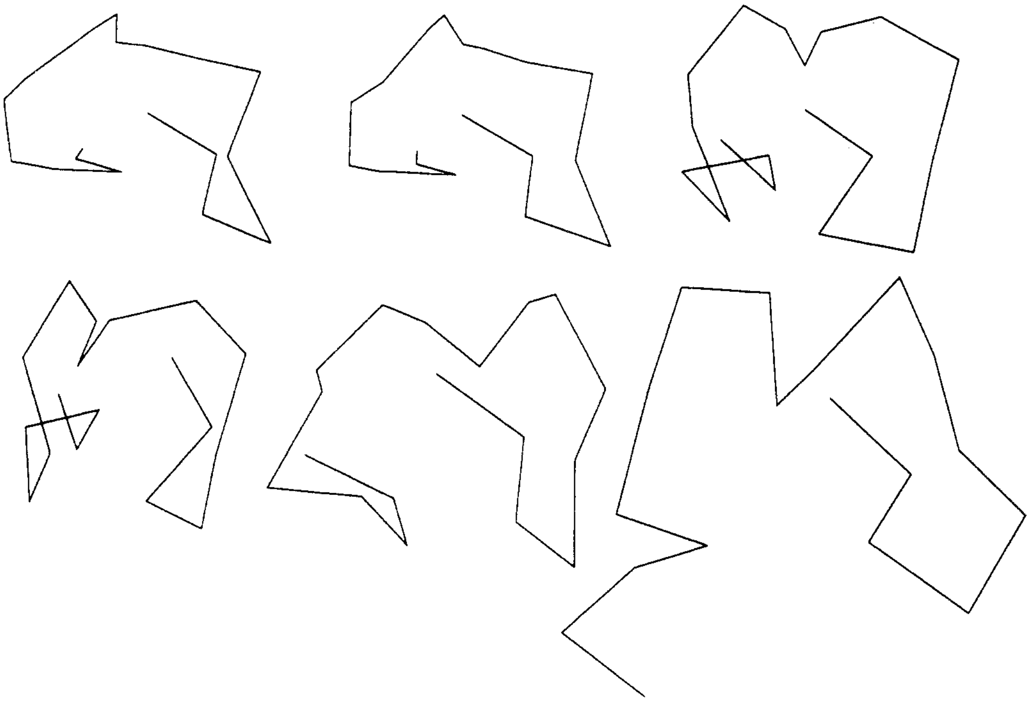

Figure 3. Best two-dimensional projections of apamin as it moves from 8 dimensions (upper left) to 6 (upper right) to 5 (lower left) to 3 (lower right) in steps of one dimension, according to the rotational energy embedding algorithm. Only the C^{α} tracings are shown.

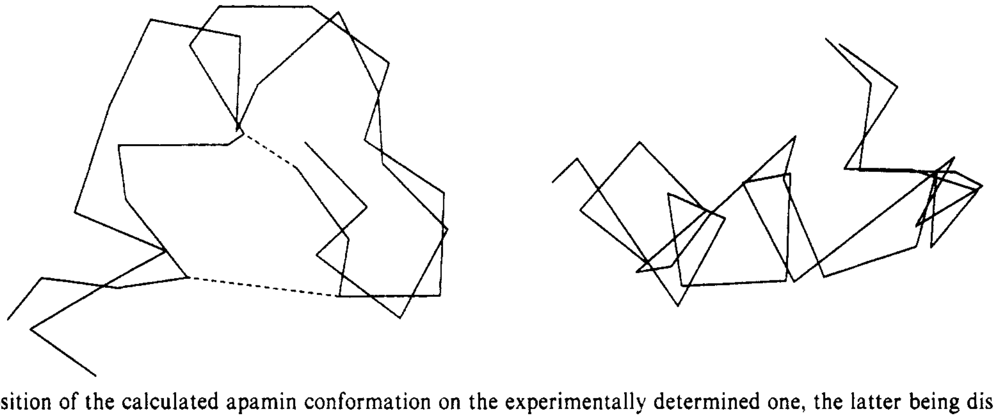

Figure 4. Superposition of the calculated apamin conformation on the experimentally determined one, the latter being distinguished by the dotted lines indicating the disulfide bridges. On the left is a frontal view of the dish-shaped structures, and on the right they have been rotated 90° about the horizontal axis to see the dish from the edge.

= 0 for the key ith atom, and its tree descendents will have the same rotation applied:

$$c' = R\left[m, m-1, \arctan\left(-\frac{c_{i,m}}{c_{i,m-1}}\right)\right]c \qquad (11)$$

where the cis and trans rotations correspond to choosing θ and $\theta + \pi$. In other words, the cis/trans rotation is just a special case of the canonical rotation, eq 9.

This algorithm has been coded in the C programming language, and runs under the Unix operating system. Those interested in the computer program should contact the authors. The execution time is generally better than for the earlier formulations of energy embedding, but it is still on the order of hours of CPU time for 60 points on a Sun 4 computer. The most time-consuming step is the repeated minimizations at the four and five-dimensional stages. If there are N dihedral angles in $\mathcal{R}^3$, then there are $2N - 3$ dihedral angles to manipulate in the last two dimensionality reduction passes, and two choices for each dihedral angle, yielding a total of $4N - 6$ local minimizations. If energy minimizations have a cost that varies with the square of the size of the molecule, then rotational energy embedding costs go up with the cube of the molecule's size.

RESULTS

The most significant application of rotational energy embedding has been in testing new potential functions for protein folding. We have already explained elsewhere[18] how to construct a potential function that mimics protein folding in the sense that the lowest local minimum we have been able to find is quite close (1.8 Å rms distance deviation) to the crystal structure of a small protein, avian pancreatic polypeptide[19] (data set 1ppt in the Bookhaven Protein Data Bank[20]), while the depths of all other minima are apparently substantially higher. This potential is otherwise not intended to be a good approximation to physical reality, since each amino acid residue is represented as a single point at the C^{α} atom, there is no explicit solvent, and there is no attempt to reproduce vibrational spectra. In any case, the objective was to establish as firmly as possible that the best local minimum lay near the native conformation, that other local minima were worse, and that just a knowledge of the potential function and the amino acid sequence was sufficient to locate this global, near-native minimum. Local minimization beginning at each of 737 likely starting conformations[18] produced a best conformation with potential value −406.9 arbitrary units and rms deviation from the native of only 1.84 Å. Other minimized conformations were above the −380 level and were further from the native.

Standard energy embedding with a variety of projection directions produced at best a minimized conformer having potential value −399.6 and rms 7.52 Å. Rotational energy embedding, on the other hand, found a minimized conformer with potential value −402.5 and rms from the native of 2.05 Å, while its rms to the best known minimum was only 1.64 Å. We conclude that although rotational energy embedding certainly does not always locate the global minimum, it enables us essentially to find the neighborhood of the crystal structure of 1ppt, whereas regular energy embedding could not.

A second test of the potential function and rotational energy embedding was the prediction of the conformation of apamin, an 18-residue peptide containing two disulfide bridges. Its conformation in aqueous solution has been determined to an accuracy in the backbone atoms of 1.34 Å by two-dimensional NMR.[21] We took as given the amino acid sequence and disulfide bridges of apamin, along with the potential function, which had been adjusted solely for 1ppt. The peptide in these calculations is represented as 18 points, one per residue, so one might expect an initial high-dimensional minimum in $\mathcal{R}^{17}$, but instead the internal stress reduced it to $\mathcal{R}^8$, where its potential was −211.19. Figure 3 shows a projection onto the two most important dimensions of apamin as it moves from $\mathcal{R}^8$ to $\mathcal{R}^3$. Although substantial conformational changes and a general expansion in the target three dimensions are clearly visible, there is little increase in energy until the last two steps. Rotational energy embedding produced a minimized conformer in $\mathcal{R}^3$ with potential value of −132.02 and rms from the experimentally determined structure of 1.85 Å. By way of comparison, local minimization starting from the NMR structure reaches −134.4 in potential by moving 1.65 Å rms away from the native. Figure 4 shows that in spite of substantial differences in secondary structure, the calculated and NMR conformations agree closely in overall chain path. It might seem that two disulfide bridges in only 18 residues would greatly constrain the range of conformational possibilities, thus rendering the 1.8-Å agreement trivial. As a check, we generated 30 structures of apamin by the EMBED algorithm, including 10 by a version with improved sampling,[22] giving as constraints only the 3.8-Å virtual bond lengths down the chain, 5.5 Å between the two pairs of Cys residues, and otherwise only a 4-Å default lower bound on interresidue distances. Surprisingly, we could generate structures having as much as 3.86 Å deviation from the native. We conclude that although the covalent structure of apamin allows a wide range of conformations, rotational energy embedding along with the potential function enable us to reach a conformer very close to the native.

ACKNOWLEDGMENT

We thank Dr. T. Hayden for supplying a listing of the computer program for MAP by W. Glunt, and we thank Dr. Wemmer for a listing of his apamin atomic coordinates. This work was supported by grants from the National Institutes of Health (GM37123, GM37708, and GM38221) and the National Science Foundation (DMB-8705006).

Registry No. Avian pancreatic polypeptide, 58591-52-9; apamin, 24345-16-2.

REFERENCES AND NOTES

(1) Holak, T. A.; Gondol, D.; Otlewski, J.; Wilusz, T. Determination of the Complete Three-Dimensional Structure of the Trypsin Inhibitor from Squash Seeds in Aqueous Solution by Nuclear Magnetic Resonance and a Combination of Distance Geometry and Dynamical Simulated Annealing. *J. Mol. Biol.* **1989**, *210*, 635–48.

(2) Summers, M. F.; South, T. L.; Kim, B.; Hare, D. R. High-Resolution Structure of an HIV Zinc Fingerlike Domain via a New NMR-Based Distance Geometry Approach. *Biochemistry* **1990**, *29*, 329–40.

(3) Endo, S.; Inooka, H.; Ishibashi, Y.,; Kitada, C.; Mizuta, E.; Fujino, M. Solution Conformation of Endothelin Determined by Nuclear Magnetic Resonance and Distance Geometry. *FEBS Lett.* **1989**, *257*, 149–54.

(4) Nerdal, W.; Hare, D. R.; Reid, B. R. Solution Structure of the *Eco*RI DNA Sequence: Refinement of NMR-Derived Distance Geometry Structures by NOESY Spectrum Back-Calculations. *Biochemistry* **1989**, *28*, 10008–21.

(5) Levy, R. M.; Bassolino, D. A.; Kitchen, D. B.; Pardi, A. Solution Structures of Proteins from NMR Data and Modeling: Alternative Folds for Neutrophil Peptide 5. *Biochemistry* **1989**, *28*, 9361–72.

(6) Kollman, P. A.; Grootenhuis, P. D. J.; Lopez, M. A. Computer Simulation Studies of Spherands, Crowns, and Porphyrins: Application of Computer Graphics, Distance Geometry, Molecular Mechanics, and Molecular Dynamics Approaches. *Pure Appl. Chem.* **1989**, *61*, 593–6.

(7) Lee, M. S.; Gippert, G. P.; Soman, K. V.; Case, D. A.; Wright, P. E. Three-Dimensional Solution Structure of a Single Zinc Finger DNA-Binding Domain. *Science* **1989**, *245*, 635–7.

(8) Crippen, G. M. Linearized Embedding: A New Metric Matrix Algorithm for Calculating Molecular Conformations Subject to Geometric Constraints. *J. Comput. Chem.* **1989**, *10*, 896–902.

(9) Lamerichs, R. M. J. N.; Padilla, A.; Boelens, R.; Kaptein, R.; Ottleben, G.; Rueterjans, H.; Granger-Schnarr, M.; Oertel, P.; Schnarr, M. The Amino-Terminal Domain of LexA Repressor Is α-Helical but Differs from Canonical Helix-Turn-Helix Proteins: A Two-Dimensional Proton NMR Study. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 6863–7.

(10) Crippen, G. M. Global Optimization and Polypeptide Conformation. *J. Comput. Phys.* **1975**, *18*, 224–31.

(11) Crippen, G. M.; Havel, T. F. Distance Geometry and Molecular Conformation. In *Chemometrics Research Studies Series*; Bawden, D., Ed.; Research Studies Press (Wiley): New York, 1988.

(12) Crippen, G. M. Conformational Analysis by Energy Embedding. *J. Comput. Chem.* **1982**, *3*, 471–6.

(13) Crippen, G. M. Energy Embedding of Trypsin Inhibitor. *Biopolymers* **1982**, *21*, 1933–43.

(14) Crippen, G. M. Why Energy Embedding Works. *J. Phys. Chem.* **1987**, *91*, 6341–3.

(15) Snow, M. E. Private communication, 1989.

(16) Crippen, G. M.; Havel, T. F. Stable Calculation of Coordinates from Distance Information. *Acta Crystallogr.* **1978**, *A34*, 282–4.

(17) Glunt, W.; Hayden, T. L.; Hong, S.; Wells, J. An Alternating Projection Algorithm for Computing the Nearest Euclidean Distance Matrix. *SIAM J. Matrix Anal. Appl.* **1989**, in press.

(18) Crippen, G. M.; Snow, M. E. A 1.8 Å Resolution Potential Function for Protein Folding. *Biopolymers* **1990**, in press.

(19) Glover, I.; Haneef, I.; Pitts, J.; Wood, S.; Moss, D.; Tickle, I.; Blundell, T. Conformational Flexibility in a Small Globular Hormone. X-ray Analysis of Avian Pancreatic Polypeptide at 0.98 Angstroms Resolution. *Biopolymers* **1983**, *22*, 293.

(20) 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. The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535–42.

(21) Pease, J. H. B.; Wemmer, D. E. Solution Structure of Apamin Determined by Nuclear Magnetic Resonance and Distance Geometry. *Biochemistry* **1988**, *27*, 8491–8.

(22) Havel, T. F. The Sampling Properties of Some Distance Geometry Algorithms Applied to Polypeptide Chains: A Study of 1830 Independently Computed Conformations. *Biopolymers* **1990**, in press.