ELSEVIER

Conformational analysis using distance geometry methods

David C. Spellmeyer, Alex K. Wong, Michael J. Bower, and Jeffrey M. Blaney

Chiron Corporation, Emeryville, California, USA

Distance geometry methods have been used extensively to build models of molecules of various sizes, including small molecules, peptides, and proteins. These methods are often overlooked as tools for conformational analysis, even though they often perform as well as other conformational sampling methods. We have implemented two new distance geometry approaches in the DGEOM95 package. In the first new method, the traditional embedding algorithm is replaced with a procedure that generates random 4D coordinates for each atom, followed by refinement of these coordinates into 3D using the distance geometry error function. The conformational sampling produced by this method is comparable to that obtained with partial metrization, and superior to that obtained with the original embedding procedure. In the second method, a molecular dynamics step is included in the refinement stage. Although this method can be applied to any embedding algorithm, substantial improvements in sampling are seen primarily with the original embedding algorithm. © *1997 by Elsevier Science Inc.*

Keywords: distance geometry, conformational analysis, solvent boxes, DGDYN, random coordinates, cycloheptadecane, Met-enkephalin

INTRODUCTION

One of the major challenges in molecular modeling is to find a general, fast, reliable method for determining the global minimum energy conformation of a molecule, regardless of the size of the system under consideration. A variety of methods exist as solutions to this problem. Among these are distance geometry techniques. Although distance geometry (DG) methods were developed as a general model-building tool,[1] they are well suited for conformational searching, often performing as well as, or better than, other conformational sampling methods.[2–4] Distance geometry methods, however, are most commonly applied not to model building and not to conformational analysis, but rather to nuclear magnetic resonance (NMR) structural determination.[2,5] This is but a small subset of the applications in which DG is applicable. The lack of applications of DG methods in conformational analysis is in part due to the perception that DG methods are applicable only with NMR data and in part because one DG method performed poorly in one conformational analysis application.[6] In this article, two new DG methods are introduced, and several methods are compared. The examples are meant to highlight that DG methods perform well in conformational analysis applications and are computationally inexpensive.

An excellent review of conformational search methods has been presented, and will not be presented here.[7] However, a brief introduction to sampling procedures is appropriate.

Conformational analysis methods generally fall into two classes: deterministic and stochastic. In deterministic methods, conformational space is searched exhaustively. Often, this takes the form of a complete torsion search of all the rotatable bonds in a molecule. The time required for this type of search scales as the exponent of the number of rotatable bonds. While deterministic methods are guaranteed to succeed only if the granularity of the search is fine enough, it is often difficult to define the granularity needed to successfully complete the search. For instance, it may be possible to perform a complete search in torsion space if one limits the number of rotamers per bond to a small number, although the global minimum might be missed. The same search can become computationally infeasible if the number of scanned rotamers increases. The practical consideration of CPU time limits the choice of search complexity one can perform, and one must resort to coarse granularity or to a stochastic method.

Stochastic methods are designed to search only a subset of the conformational space and to produce results that include only the lowest energy conformations of a molecule. Some stochastic methods are based on algorithms that limit or pare the conformational search tree used in the deterministic methods.[6,8–15] Other methods use molecular dynamics,[16] Monte Carlo,[17,18] genetic algorithms,[19] random perturbations to the coordinates,[6,20] or a combination[21] to locate, sample, and optimize local minima.

Address reprint requests to: David C. Spellmeyer, Chiron Corporation, 4560 Horton Street, Emeryville, California 94608.

Paper submitted to Electronic Conference of the Molecular Graphics and Modelling Society, October 1996.

1093-3263/97/$17.00
PII S1093-3263(97)00014-3

Stochastic methods are not guaranteed to converge to the same set of low-energy conformations as produced with a deterministic method. This makes evaluating the "goodness" of stochastic methods problematic. Often, one uses the location of the global minimum as a "goodness" criterion for how effective a stochastic method is. However, this can be misleading. One can generate a large number of conformations that are very close structurally to the global minimum, but miss the lowest energy state because of a slight deviation in a torsion angle. This does not mean that the method used is poor.

Several different metrics are used in this article, including energetic criteria, geometric criteria, convergence ability, and generation of novel conformations. In general, one must evaluate a given stochastic method for its ability to perform well on several examples and one must use that information to evaluate the appropriateness of the method for a particular problem.

Distance geometry methods, like Monte Carlo (MC), molecular dynamics (MD) methods, and genetic algorithm (GA) search methods, are stochastic methods. Both the MD and MC approaches use one conformation as a starting point and generate a set of solutions, all of which stem from that first structure.

Unlike MC and MD methods, DG and GA methods create independently generated random samples in conformational space. Genetic algorithm methods require a scheme to encode all conformational transformations into a bit string that is then manipulated in order to locate minima. In DG methods, conformational space is searched by generating a large number of independent solutions within the constraints of the model. Each structure is then evaluated for inclusion in the final ensemble of low-energy conformations, usually by application of a force field or other energetic evaluation.

Distance geometry methods were first introduced to chemical applications in the late 1970s when Crippen and Havel described the original metric matrix method.[1] Several improvements to this method have been introduced since then, including improved sampling about dihedral angles,[4] full metrization,[22,23] and partial metrization.[24] These methods all are intended to provide better initial trial coordinates from the bounds matrix and to help ensure reasonable three-dimensional coordinates would result after error refinement.[25,26] New DG "error" functions have also been introduced, which improve the convergence of the refinement stages.[25,27,28]

In this article, two new sampling methods are introduced. In the first method, the initial guess for the Cartesian coordinates of a molecule are chosen completely at random. Refinement proceeds as usual against the DG error function. In the second, a dynamics calculation is performed as part of the DG error function refinement step. It is hoped that introduction of the dynamics routines will improve the sampling efficiency of DGEOM95.[29]

In this article, we examine the effects of using different coordinates generation and the effect of including a dynamics calculation as part of the refinement process on several examples. Because distance geometry is useful as a model-building tool, it is important to evaluate how new sampling techniques perform on both simple as well as complex systems. The six examples we have chosen to evaluate the conformational searching capabilities of DG methods are as follows:

1. A small organic molecule[6]
2. A long polypeptide[26]
3. A modest-sized peptide
4. The solution of a protein NMR structure[5]
5. The generation of models of a DNA bisintercalator[30]
6. The generation of solvent boxes of TIP4P water[31] and liquid *n*-pentane

Several metrics are applied to these examples. Each method demonstrates both strengths and weaknesses.

BACKGROUND

The use of distance geometry methods for general model building has been reviewed and will not be repeated here.[2] However, it might be useful to provide some background on how DG methods work and on the differences in coordinate generation.

Distance geometry is a general method for converting a set of distance ranges (or "bounds") into a set of Cartesian coordinates that are consistent with those bounds. Any molecular system can be described as the set of minimum and maximum interatomic distances between all pairs of atoms in the molecule. The complete conformational space of the molecule is contained within this space. In distance geometry, a matrix is defined as the set of minimum and maximum distances, and then used to create a series of conformers that are consistent with those distances.

The process of defining the distance matrix and chirality constraints needed to complete a distance geometry calculation is detailed elsewhere.[2,7] A brief summary is presented here. The distance matrix is constructed by analyzing all of the pairwise distances in a molecule. The lower and upper bounds for bonded atoms (1,2 distances) are set to the bond distance. Angles (1,3 distances) are converted to specific distances, which are entered in exactly the same way as 1,2 distances. The interatomic distances for 1,4-nonbonded atoms in nonrotatable bonds can be calculated and entered as exact distances. However, most interatomic distances cannot be calculated exactly. For instance, the 1,4 distance for a bond with free rotation can be defined as a minimum distance (in the syn form) and a maximum distance (in the anti form). These minimum and maximum distances are entered as the lower and upper bounds, respectively. For atoms that are 1,5 nonbonded and greater (i.e., atoms that are separated by three or more bonds), it is more difficult to calculate possible lower and upper bounds. By default, the upper bounds are set to a large number (999 Å) and the lower bounds are set to the sum of the van der Waals radii.

Other specific constraints can be added, such as nonbonded interatomic distances (e.g., nOcs), groups that are to remain rigid, chirality, etc. The matrix is then "smoothed" using triangle inequality rules (for any three atoms, *A, B,* and *C,* the distance *AC* must be less than or equal to the distance *AB* plus the distance *BC*). This matrix then defines the set of all possible conformations a molecule can adopt. An example of the bounds matrix for *n*-pentane is shown in Table 1.

One must now convert this distance bounds matrix into Cartesian coordinates. Significant effort has gone into development of a variety of methods for this conversion. Crippen and Havel demonstrated that one can convert an exact distance matrix into a single set of three-dimensional coordinates.[1] The process of converting an inexact distance matrix (one with some upper and lower bounds that are not equal) is more challenging.

Table 1. Upper and lower bounds matrix for *n*-pentane after bounds smoothing

	Atom				
Atom	1	2	3	4	5
1	0.0	1.531	2.701	4.164	5.402
2	1.531	0.0	1.551	2.779	4.165
3	2.701	1.551	0.0	1.551	2.701
4	3.452	2.779	1.551	0.0	1.531
5	3.300	3.543	2.701	1.531	0.0

In Crippen and Havel's original method, each pairwise distance is set to a random value between the upper and lower bounds.[1] The three-dimensional (3D) structure represented by this matrix can be approximated by solving for the three largest eigenvalues of this matrix. These are used as a starting point for further optimization against the "error function." This optimization is needed to further refine the structure, since the choice of random interatomic distances often results in poor 3D structures. It has been shown that two stages of optimization can be used to produce the desired 3D structures. The first stage of optimization takes place in four dimensions.[2] This allows the atoms in the molecules to pass "through" each other in three dimensions. The second minimization brings the four-dimensional structure into three dimensions.

The sampling produced by the original method has been shown to favor extended conformations.[4,26] Therefore, a good deal of effort has been expended to improve the initial guess of the interatomic distances and produce better structures. Peishoff and Dixon introduced a method to sample more evenly about torsions.[4] Havel and co-workers introduced a method, called metrization, in which all of the pairwise interatomic distances are selected so that triangle inequality is obeyed.[22,23] This method produces very good convergence to 3D coordinates and has good sampling characteristics, but it is computationally very expensive. Kuszewski et al. introduced a modification of this procedure in which only $4N$ distances are required to satisfy the triangle inequality prior to embedding and coordinate refinement.[24] They rationalized that one could describe a system nearly completely by fixing distances to only four atoms within the system. In partial metrization, four atoms are chosen at random, and all of the pairwise distances involving these atoms are chosen so that triangle inequality is obeyed. The remaining distances are chosen randomly. The structures are then refined against the DG error function as described above. Dixon has implemented the method used in DGEOM95, the program employed for the calculations herein.[29]

Figure 1 shows the trial coordinates of benzene for the original sampling and partial metrization methods. The ring system with the partial metrization method is flat and one set of the para-protons is well positioned. The ring system generated as the initial coordinates using the original sampling method is visibly puckered.

While these conformers do resemble, more or less, the final structure of benzene, for many molecules this is not the case. For example, Figure 2 shows the initial coordinates for trimethoprim as generated with the original sampling and partial metrization methods. These are compared with the final coordinates after refinement against the DG error function. As

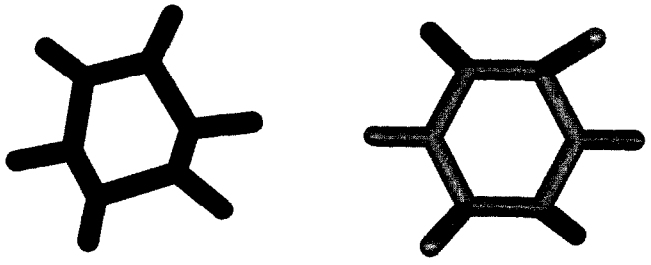

Figure 1. The initial coordinates of benzene generated using the original sampling method (left) and the partial metrization method (right).

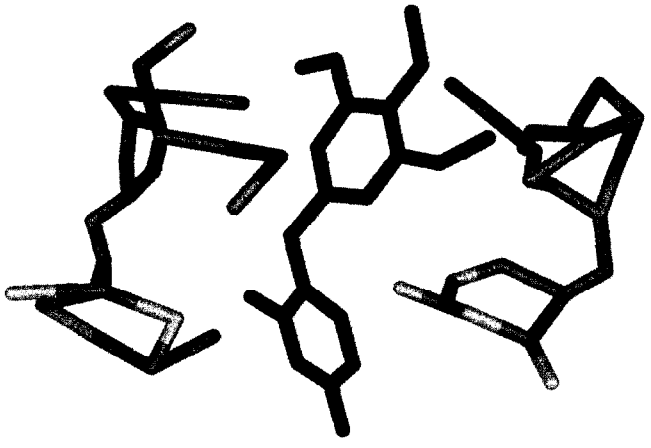

Figure 2. The initial coordinates of trimethoprim generated using the original sampling method (left) and the partial metrization method (right). Middle: An example of final refined coordinates.

Figure 2 shows, the coordinates generated by either method resemble the final structure, but still require a great deal of refinement.

The fact that such rough initial coordinates refined so well prompted us to investigate the use of randomly chosen 4D coordinates as the initial coordinates. In this approach, the 4D coordinates are set by choosing a random number between -1.0 and 1.0 and multiplied by the maximum upper bound distance in the molecule. Conceptually, the maximum upper bounds (maxupper) defines a box with sides of length twice the

size of "maxupper" in which all the coordinates must lie. The compression to 3D takes place during the error refinement stage of DGEOM95. It is typically in the second minimization that having a coordinate with a fourth dimension component is penalized using a harmonic weighting. As the structure minimizes, the 4D structure refines smoothly to 3D. Surprisingly, the structures refine quite well and the sampling produced by this method is quite good. Empirical results have shown that sampling is dependent on the size of the "box," or "scaling factor." We will show that different-sized boxes can be used to achieve different sampling results.

Figure 3 shows this method for generation of a conformation of benzene. The box length is defined as twice the distance of the maximum upper bound (the distance between two para-related protons). The starting position of each atom is generated as an independent random coordinate. These are shown in the box on the left. These coordinates are rapidly minimized against the DG error function to produce the conformer on the right. The carbons and associated protons are shown shaded consistently in both views.

Figure 4 shows a 3D projection of the initial 4D coordinates chosen as the initial coordinate set for trimethoprim using the random coordinate method. The final conformation is shown in the middle of the box. The box size used for coordinate generation for trimethoprim is substantially larger than in the benzene example because of the larger maximum upper bounds in trimethoprim. The initial coordinates span much of this box. The coordinates are unrecognizable as trimethoprim. Many of the bonds are extremely long and many of the angles are near zero. To refine this structure, atoms will have to be pulled "through" each other and rings must flatten, while the atoms are being pulled together. Nonetheless, these coordinates do refine and the refined structure is geometrically reasonable. The random coordinate methods used here are part of the DGEOM95 distribution.[29]

We have also examined the effect of incorporating a molecular dynamics-like calculation between the first and second minimization steps in DGEOM95. This approach can be used for any of the sampling methods described above. The distance geometry dynamics (DGDYN) approach is most similar to a reduced-coordinate molecular dynamics approach. Unfortunately, the DGEOM95 error function contains no value of energy, and is defined in terms of $Å^4$ plus $Å^3$, so "time" and "mass" and "temperature" are all meaningless, and a true

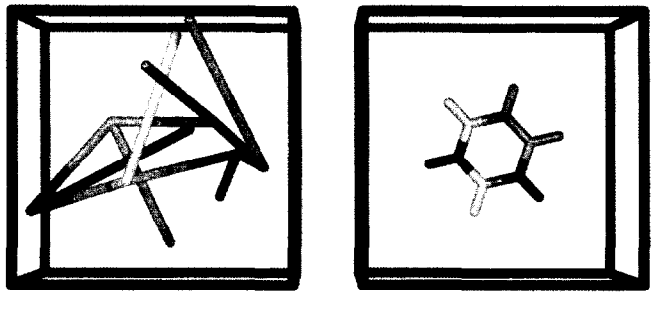

Figure 3. The initial coordinates generated using random coordinates (left) and the final refined coordinates (right). The fourth dimensional coordinate is generated but is not shown.

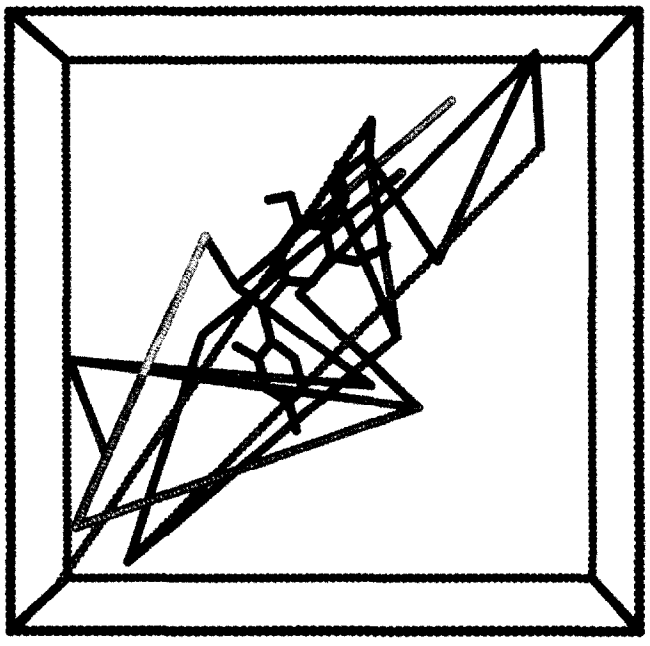

Figure 4. The initial coordinates of trimethoprim generated using random coordinates (in grayscale) with scaling box shown, and an example of final refined coordinates (in black). Again, 4D coordinates are generated, although only 3D coordinates are depicted.

dynamics calculation cannot be defined. Therefore, an empirically derived calculation is performed, either in three dimensions or in four (as determined by a user-defined parameter). A velocity Verlet method[32] is used for the integrator and stochastic collisions[33] are used for the temperature scaling. This method has not been implemented with periodic boundary conditions, and so no pressure coupling has been added. These algorithms borrow heavily from the excellent presentation by Allen and Tildesley,[34] and are derivatives of the methods developed and used in SPASMS.[35] The stochastic collisions are implemented by assigning random velocities from a Boltzmann distribution about a specified "temperature" at a specified "time" interval. We have empirically chosen a set of time step sizes and temperatures that seems appropriate. The masses for all the atoms in the system are defined to be of unit value.

Figure 5 shows the result of applying 1 000 steps of DGDYN to the initial original sampling-generated coordinates for trimethoprim. After 1 000 steps of DGDYN, the coordinates resemble the final coordinates; the bond lengths are more reasonable and the rings have flattened out. Because there is excess "energy" in the molecule, this is not a fully refined structure. Several examples are included to evaluate the effect of inclusion of the DGDYN step on sampling characteristics.

Unfortunately, the empirical nature of this approach and the unusual error function sometimes allow the system under study to exhibit unpredictable and erratic behavior. A good deal of effort has been made to mitigate these occurrences, such as automatic adjustment of the time step size, the temperature of the simulation, the frequency of the temperature coupling, etc., but they do still occur from time to time. Therefore, one must fine-tune the input parameters for each system under investi-

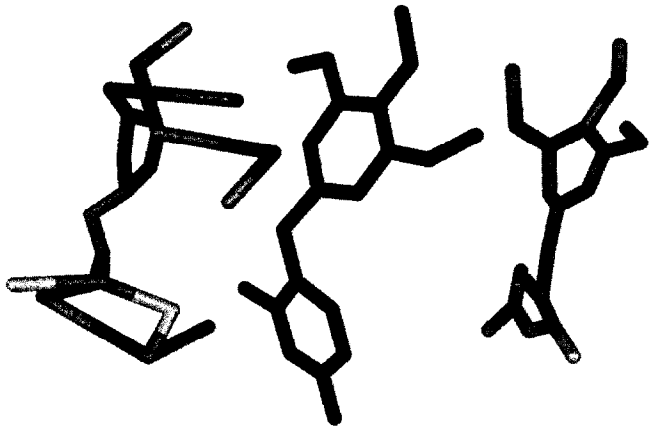

Figure 5. The initial coordinates of trimethoprim generated using the original sampling method (left), after 1 000 DGDYN steps (right), and an example of a refined structure (middle).

gation. Although this makes the DGDYN method more complicated to use than the standard methods—it can't be used as a "black box"—the results can be worth the additional effort. The DGDYN used here is part of the DGEOM95 distribution.[29]

Finally, we have implemented periodic boundary conditions for generating solvent boxes to be used in statistical mechanical simulations. In this approach, one inputs the molecules used to define the molecular system to be studied, including as many solute molecules or solvent molecules as desired. The user also specified a box size, which must be cubic in this first implementation. The lower bounds between all molecules are set to the van der Waals radii of the atoms, ensuring that no molecules interpenetrate another. The upper bounds are set to the square root of 3 times the input box size. The random coordinate method is used to generate a starting set of coordinates, which are then imaged into the periodic box. The distance geometry refinement function is then applied with the periodic boundary conditions in place. The entire system is then refined to a low-error starting point. The most time-consuming part of this calculation is the triangle bounds smoothing, since this is an order N^3 process, where N is the number of atoms. However, once this part of the process has been completed, several different solutions can be generated with approximately order N^2 time (the error function refinement). Thus, one can generate independent starting points for molecular dynamics or Monte Carlo simulations with this approach. This code is not part of the DGEOM95 package.[29] However, the authors will gladly make available all changes needed to incorporate periodic boundaries into DGEOM95.

We have applied this method to the generation of approximately 40 boxes of TIP4P water and 20 boxes of liquid pentane. These boxes are then subjected to molecular mechanics minimization and several picoseconds of molecular dynamics each. We believe that this method represents an alternative to current approaches of generating solvent boxes and solvated systems. However, the calculation scales poorly, and clearly cannot be applied to all systems.

RESULTS AND DISCUSSION

Cycloheptadecane

Perhaps one of the most studied system for conformational analysis is cycloheptadecane ($C_{17}H_{34}$—referred to as C_{17}). Saunders et al. have compared a variety of approaches for searching the conformational space of this flexible ring system.[6] This system was chosen because the ring system is relatively large, because it is cyclic, because no one method can enumerate and calculate all the minima, and because it is a hydrocarbon, for which the force field used was thought to be well suited.

In their original study,[6] conformations of C_{17} were generated and then minimized using MM2.[36] The comparison of several different methods showed that 262 distinct minima exist that are within 3.0 kcal/mol of the global minimum. Several methods were compared, including systematic and random conformational searches in both internal coordinates and in Cartesian spaces, distance geometry, and molecular dynamics. This study also showed that while every method did find the global minimum, no one method found all of these conformations. Some methods performed better than others. Since this first study, several other methods have been tested for their ability to find the lowest energy conformations of C_{17}.[4,37–39]

Distance geometry calculations were included in this study and were shown to be inefficient as compared to the other methods. The authors did find that embedding both carbons and hydrogens tended to produce more low-energy structures. Peishoff and Dixon have shown that a substantial improvement in these results is obtained when one includes torsion sampling and eliminates distance correlation calculations.[4] In this case, significantly more low-energy conformations of C_{17} are generated in the same amount of computer time as in the original work. In fact, this study showed that the DG methods compare favorably to the other methods, and outperformed several of them.

For this evaluation, four different DG calculations were performed. In each method, 10 000 conformations of C_{17} were generated (with hydrogens), followed by minimization with MM2[36] or MM3[40] as implemented in the Macromodel Batchmin program.[41] Hydrogens were included in the conformation generation step and eclipsed conformations of bonds were allowed in all methods. The first method is the original distance selection with torsion sampling and without distance correlation. This is similar to that employed by Peishoff and Dixon.[4] In the second method, partial metrization was used to generate the starting conformations. In the third, random coordinates were used, while in the fourth, random coordinates with the DGDYN approach was used. In all cases, all conformations generated were kept, rather than using the default option of rejecting conformations that have already been sampled. The results of these calculations are shown in Table 2. Two conformations were considered identical if all dihedral angles in one conformer were within 5° of the corresponding dihedral angles in the second conformer. This is one of the criteria used in the original work.[6]

All starting structures were converted from the DGEOM95 output format to the Macromodel input format using Babel.[42] In each case, a handful of structures was not converted. However, the minimum number of structures converted is more than

Table 2. Summary of conformation generation with a variety of distance geometry methods

	Random coordinate generation	Partial metrization	Random coordinates with DGDYN	Original sampling method
Number of conformations generated	10 000	10 000	10 000	10 000
Number of conformers minimized	9 841	9 586	9 780	9 868
Number of unique conformers with MM3 minimization	5 909	5 497	6 080	4 899
Number unique conformers with MM2 minimization	6 469	6 153	6 681	5 430
Conformers within 3 kcal/mol of minimum with MM3	98	117	125	126
Conformers within 3 kcal/mol of minimum with MM2	172	192	209	242

9 500. Thus, we believe that we have enough conformations in each method to make comparisons valid.

Of course, one would like to eliminate conformers that produce redundant minimizations. It is quite difficult to predict which minimum a conformation is likely to minimize to directly from the DGEOM output. Therefore, we are required to minimize with MM2 and then eliminate conformations after the fact. In this case, one would like the DGEOM method that produces the largest number of unique conformations. The random coordinate method did most well at locating unique conformations, with more than 60% of the 9 841 conformations minimized proving to be unique. The only method to do better than this was the random coordinates combined with DGDYN, which generated 62% unique structures. Partial metrization was comparable to these other two methods, at 57%, while the original sampling method did most poorly, at about 50%.

Shenkin and McDonald have shown that Macromodel can locate 132 conformations of C_{17} within 3.0 kcal/mol of the global minimum with MM3.[39] All four of the methods studied here were able to locate the global energy minimum. Random coordinates with DGDYN located four conformers within 1.0 kcal/mol of the global minimum. The random coordinate method located three, while partial metrization and the original sampling method each located two. The random coordinate method, however, did most poorly at locating the largest number of the 132 presumed lowest energy structures, locating only 98 low-energy conformations. Partial metrization located 117 unique conformations within 3.0 kcal/mol of the global minimum. The random coordinate with DGDYN located 125 unique low-energy conformers, and the original sampling method located 126. Interestingly, we missed 19 conformations that Shenkin and McDonald located,[39] while we found 12 conformers that they did not report. Thus, there are at least 142 low-energy conformers of C_{17}, and no one method has succeeded in locating them all. Table 3 presents the number of conformers missed with our method as compared to the Macromodel search.

The results for DG-generated conformers of C_{17} are largely similar when MM2 is used to minimize the DGEOM95-generated structures (see Table 2). All four methods were able to locate the global minimum. In all cases, a larger number of unique conformations was located, and a larger number of structures within 3 kcal/mol of the minimum is located. The largest number of low-energy conformers was located with the original sampling method. We were able to locate 242 conformers with 3 kcal/mol of the global minimum, whereas Peishoff and Dixon located 223, and Saunders et al. located 172 with DGEOM. This method located the smallest number of unique conformers; only 5 430 of the 9 868 starting conformations were unique. The random coordinate method produced 172 low-energy conformations, while locating 6 469 unique conformations. The partial metrization method located 192 low-energy conformations, and 6 153 unique conformations. Finally, the random coordinate method with DGDYN produced the largest number of unique conformers (6 681) and 209 low-energy conformations. Our results show that distance geometry methods are comparable to other search methods reported for the C_{17} system.

Table 3. Number of unique low-energy conformers located with distance geometry methods and with Macromodel

	Unique conformers with DG method	Unique conformers with macromodel
Original sampling	4	8
Partial metrization	6	20
Random coordinate generation	8	36
Random coordinates with DGDYN	12	19

Interestingly, the lowest energy conformer of C_{17} displays an expanded ring, rather than a collapsed ring. The fact that the original sampling method tends to produce lower energy conformers than the other methods is probably a result of the tendency of this method to produce extended conformations. In a ring system, this preference leads to (1) more expanded ring conformers (i.e., conformers in which the atoms of the ring are as distant from each other as possible) and (2) less unique conformers, since the method tends to sample only in one area of conformation space. Other methods do a better job of sampling both extended and more compact structures, as is described later in this article. The random coordinate and partial metrization methods produce rings that are more compressed than the original sampling method.

The molecular mechanics minimizations for this particular problem required substantial amounts of computer time to complete, as much as 2 days of CPU time per 10 000-conformer run (on a Silicon Graphics R4400 under AIX 5.2). Rarely does one want to expend so much computer time on a molecular system. Rather, one typically would like to sample reasonably the conformational profile of a system. Figure 6 shows the energy profiles of the 10 000 conformers and the first 1 000 trials from several of the DGEOM runs. In each of the methods, the energy distribution profiles of the 1 000 conformer sets resemble the final results. The 1 000 trial runs would provide the desired distribution, but might not locate the same number of low-energy conformations. However, in each case, the lowest energy conformer was located within 1 000 trials.

The ability to generate unique conformations can be assessed in a different manner. At the beginning of a DGEOM run, it is quite likely that each independent conformation generation will result in a unique structure. As the total number of conformations increases, it is more likely that a conformation will be revisited. Once a large number of conformations has been generated, it is likely that we will produce more and more redundant conformations. In the graphs shown in Figure 7, we plot a running average of the number of redundant independent trials that are sampled over 500 independent trials. In this study, we do not take into account the symmetry-related conformations, since DGEOM95 is not coded to eliminate them. Rather, we consider redundant any conformation with an RMSD less than 0.5 Å from a conformation already generated.

In the case of the original sampling method, it is clear that about 20 to 30 trials are redundant per 500 trials over the first 2 000 conformations. This rises rapidly to about 70 redundant structures at 4 000 trials, and continues to rise rapidly. At 10 000 trials, about 150 of every 500 are redundant. Inclusion of the symmetry-related conformers would increase the rejection rate, of course, since the full analysis showed that the original sampling method produced 50% redundancy.

The remaining methods clearly perform better at generating unique conformations. Partial metrization generates about 60 to 70 redundant trials for every 500 trials. This remains relatively constant throughout the simulation. It appears that there might be an upward trend as there are more and more conformers generated. The random coordinate method shows even better sampling, approaching 50 redundant trials for every 500 trials only after 6 000 trials have been done. There is a more definite upward trend, but it is clear that there is no difficulty in generating nonredundant conformations even after 10 000 conformations have been generated. Adding DGDYN to the random coordinate method improves the sampling marginally. Here, there seems to be no upward trend in the graph (see Figure 7). We have not investigated the number of trials needed to find an upward trend in this graph.

For cycloheptadecane, all methods tested were able to locate the global minimum. No one method was able to locate all of the low-energy minima with either MM2 or MM3 optimization. The original sampling method performed well at locating

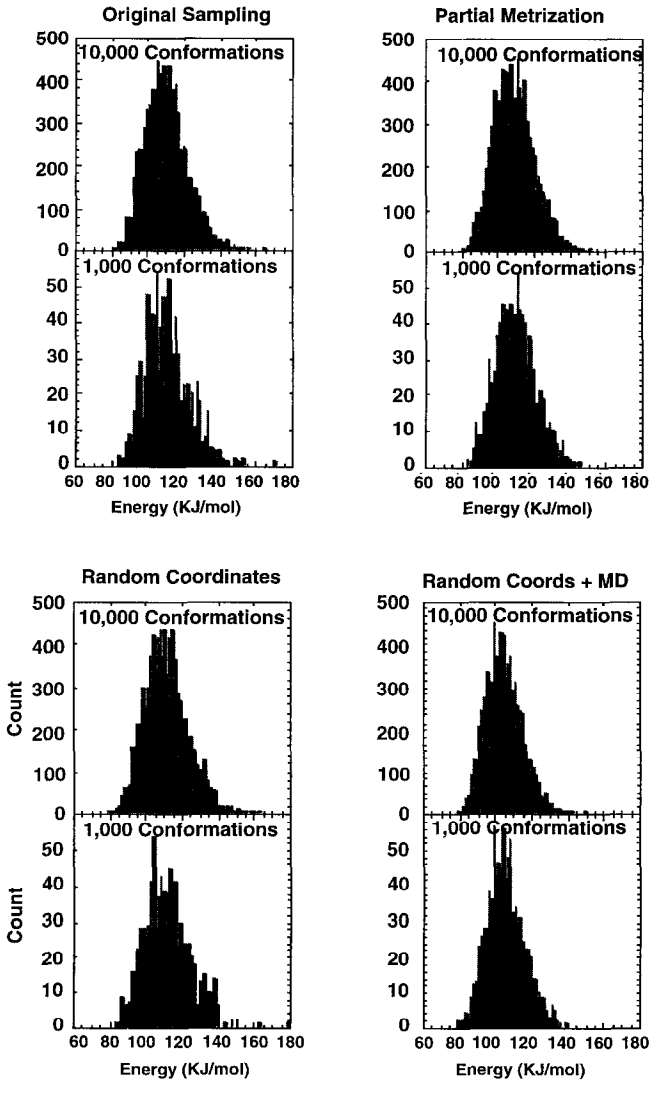

Figure 6. Comparison of the energy distributions for all 10 000 conformations (top, each panel) and the first 1 000 conformations from each DGEOM run (bottom, each panel).

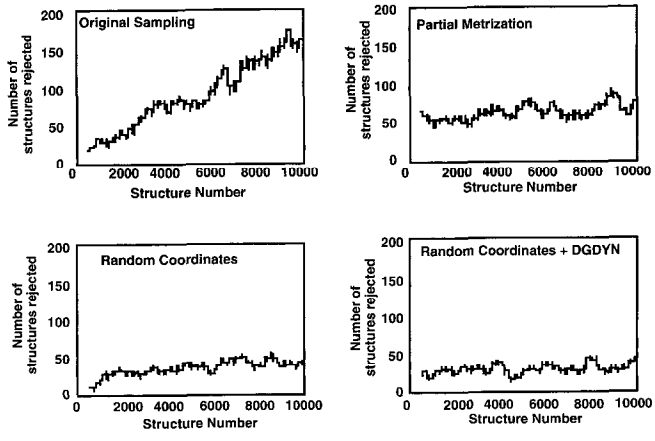

Figure 7. Running average (more than 500 conformers) of the number of trials rejected in an attempt to locate a unique conformation of C_{17}. A value of 50 would indicate that 10% of trials are rejected (50 of 500 trials).

low-energy conformers, but performed poorly at finding unique conformers. The random coordinate method performed poorly at locating all of the low-energy conformers, but performed well at locating unique conformers. The partial metrization method performed somewhere between the two in both measures. Adding DGDYN to the random coordinate method provided a good number of low-energy conformers, and was excellent at locating unique conformers. We located no new low-energy conformers using MM2, but did locate several new conformers using MM3.

Polymer simulation of a long polypeptide: Ala_{20}

Havel has applied the distance geometry procedure to the study of long polypeptide chains,[26] and showed that the typical embedding algorithms tend to produce extended chain conformations. Havel also showed that a metrization method improved the sampling considerably. In this study, we evaluate the end-to-end distance of Ala_{20} as a measure of the sampling characteristics of each of the distance geometry conformation generation methods. There are very few constraints on this molecule. The bonds, angles, and 1,4-nonbonded distances are well defined. The amide bonds are required to be trans and planar. The chirality of each residue is enforced (to the "L" configuration). However, none of these is a long-distance constraint. The lower bound for atoms that are 1,5-nonbonded (or more) are set to the van der Waals distance, and the upper bound is defined as the sum of the bond distances in the path from atom *i* to atom *j*. Because energetics are not part of the distance geometry equation, we expect to see an even sampling of the end-to-end distances ranging from contact distance to fully extended.

An input file of Ala_{20} was constructed from a single alanine residue, using the model-building features of PSSHOW.[43] Default input values for DGEOM were used for all conformation generation methods. Any conformer that displayed an RMSD of less than 0.5 Å for all heavy atoms was rejected. A total of 500 conformations was generated for each of the following methods: (1) original distance selection with torsion sampling, (2) protocol 1 with DGDYN, (3) metrization, (4) metrization with DGDYN, (5) random coordinates with a scale factor equal to 1, (6) protocol 5 with DGDYN, (7) random coordinates with a scale factor equal to 2, and (8) protocol 5 with DGDYN. The distance from the N terminus to the C terminus was measured with PSSHOW. Figures 8–13 and 15–16 show the histograms of the end-to-end distances for each of the coordinate generation and refinement methods. No attempts have been made to identify secondary structural features from these studies.

A fully extended conformation of Ala_{20} has a head-to-tail distance of 65 Å. Of course, a compact structure has a head-to-tail distance of about 3 Å (the sum of the van der Waals distance of the terminal atoms).

Figure 8 shows the original sampling method with torsion sampling generates conformations of Ala_{20} that have end-to-end distances averaging about 42 Å, with none greater than 50 Å and only a few less than 35 Å. This indicates that the overall conformational profile for Ala_{20} is not evenly sampled. Figure 9 shows that addition to 1 000 steps of DGDYN produces a broader spread in the end-to-end distances, with a much lower average value (31 vs. 42 Å). The sampling is definitely improved toward the shorter distances, but the longer distances are not being sampled at all.

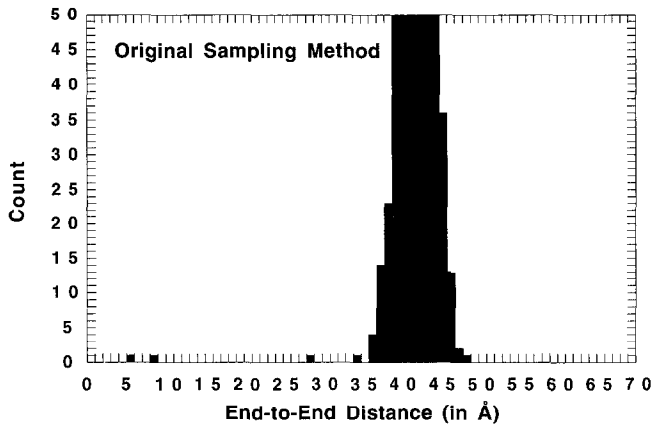

Figure 8. The head-to-tail distance for Ala_{20}. A total of 500 trial conformers was generated with the original sampling method.

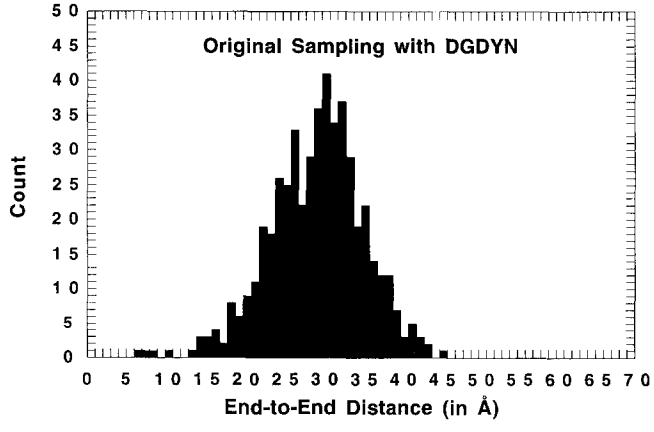

Figure 9. The head-to-tail distance for Ala_{20}. A total of 500 trial conformers was generated with the original sampling method, as in Figure 3, followed by 1 000 steps of DGDYN.

The partial metrization method produces conformers that are skewed toward shorter distances (Figure 10). There is a maximum in the end-to-end distance distribution at about 12–15 Å. We do see conformations with end-to-end distances ranging from the minimum to the maximum distances. Including DGDYN with this method produces an end-to-end distance distribution that shows somewhat more intermediate distances (Figure 11). The average has shifted from about 10 Å to somewhere nearer 20 Å. Some conformers with short end-to-end distances are generated, but no conformers with an extended geometry are generated.

The random coordinate method with a scale factor of 1 produces conformations with end-to-end distances that range from contact distance to about 45 Å (Figure 12). The distribution shows a maximum around 17 Å. This suggests a tendency to sample more compact conformers than the original sampling method, but somewhat more open conformers than the partial metrization. This distribution is not altered much by the inclusion of DGDYN (Figure 13).

Because Ala_{20} is unconstrained, we might expect to see an

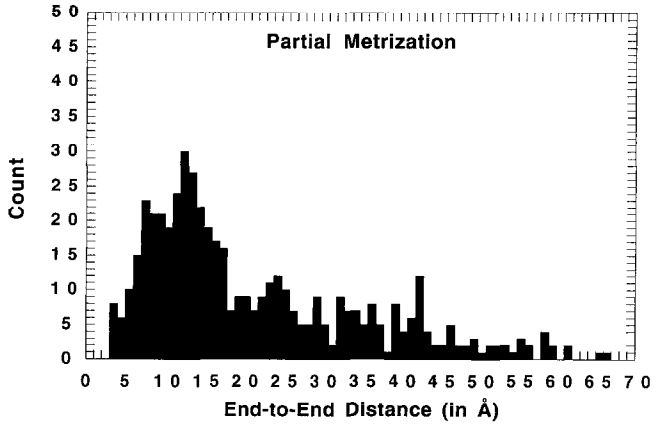

Figure 10. The head-to-tail distance for Ala_{20}. A total of 500 trial conformers was generated with partial metrization.

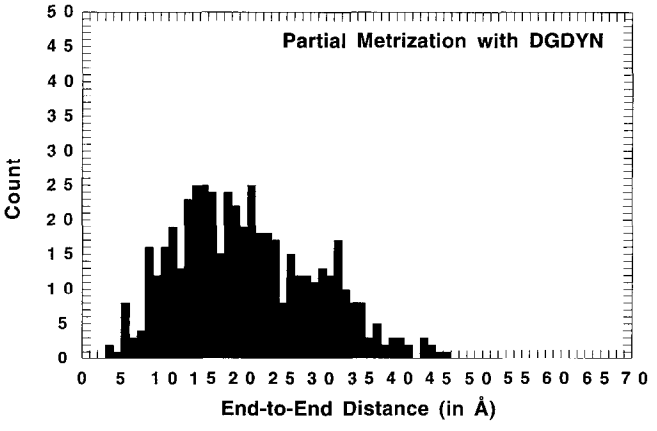

Figure 11. The head-to-tail distance for Ala_{20}. A total of 500 trial conformers was generated with partial metrization followed by 1 000 steps of DGDYN.

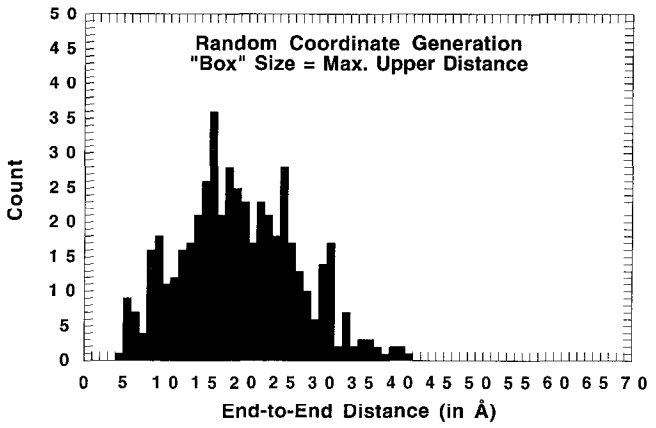

Figure 12. The head-to-tail distance for Ala_{20}. A total of 500 trial conformers was generated with random coordinate DGEOM. A scaling factor of 1.0 was used (i.e., each of the 4D coordinates for each atom was chosen in a range from about −60 to +60 Å). All conformers were refined against the DGEOM95 error function.

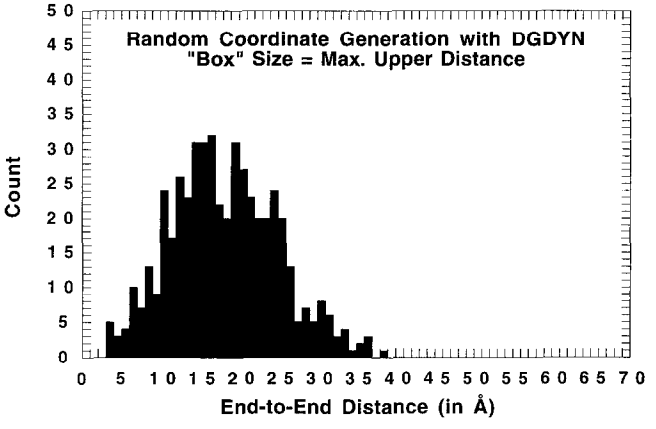

Figure 13. The head-to-tail distance for Ala_{20}. A total of 500 trial conformers was generated with random coordinate DGEOM as in Figure 7, followed by 1 000 steps of DGDYN.

even distribution in the end-to-end distance profile. However, this is not observed in the random coordinate method. This can be shown to be a result of the coordinate choice. As we described above (see Background), the coordinates are chosen from a uniform random distribution in four dimensions ranging from zero to one and multiplied by the maximum distance allowed for the molecule. A second random number is used for each coordinate to determine if the value of the coordinate is positive or negative. Figure 14 shows the interatomic distances sampled for 1 000 trials of only two atoms in three dimensions. The random coordinate selection is performed within a range from −1.0 to 1.0, so the maximum distance between the two atoms would be 2.82 Å. While each of the six coordinates is chosen independently from a uniform distribution ranging from −1.0 to 1.0, the histogram of distances for 1 000 trials shows a Gaussian distribution about the average distance of about 1.4 Å. Likewise, the coordinates for all of the atoms in Ala_{20} are chosen independently. Each interatomic distance is then expected to begin the refinement process at about the average of the corner-to-corner distance for the box bounding the coordinate selection. Refinement of the coordinates would most likely shorten these distances, resulting in conformers that, on average, show nonuniform sampling biased toward the shorter distances.

Choosing the random coordinates with a scaling factor of two (i.e., selecting each coordinate from the range −120 to +120 Å) reduces this sampling bias (Figure 15). The random coordinate method with a scaling factor equal to two produces the most evenly distributed end-to-end distances of all the methods used. This profile looks like a Gaussian profile with a much larger standard deviation about an average that is nearer to half of the maximum interatomic distance in the molecule. As was seen in all other methods, Figure 16 shows that inclusion of the DGDYN step with this random coordinate method produces conformers with more compact end-to-end distances (about 25 Å), and no conformers with end-to-end distances greater than 45 Å.

For this particular application, DGDYN produces conform-

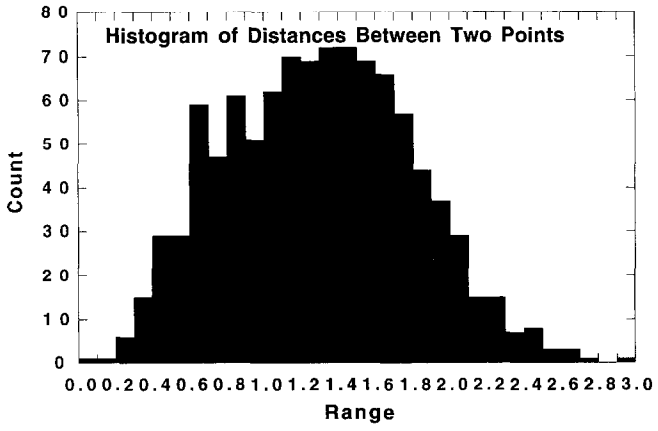

Figure 14. A histogram of the interatomic distances generated for 1 000 trials with random coordinates selected between −1 and +1.

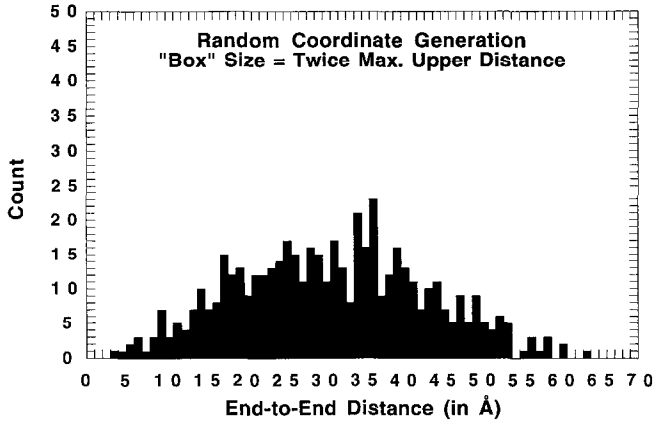

Figure 15. The head-to-tail distance for Ala_{20}. A total of 500 trial conformers was generated with random coordinate DGEOM. A scaling factor of 2.0 was used (i.e., each of the 4D coordinates for each atom was chosen in a range from about −120 to +120 Å). All conformers were refined against the DGEOM95 error function.

ers that sample relatively shorter distances at the expense of the longer distances. Three of the four sampling methods appear to be very similar, with Gaussian distributions about an end-to-end distance of about 20 Å. In turn, these are all fairly similar to the profile of end-to-end distances obtained with the random coordinate method with a "box" size set to the maximum upper bound distance. The fourth method, the original method, produces an end-to-end distance profile that is similarly shaped, but with an average displaced to somewhat longer distances.

[Met^5]-Enkephalin

Nayeem et al. have compared the ability of the simulated annealing and Monte Carlo methods to converge on the global minimum structure for the pentapeptide [Met^5]-enkephalin.[44] The global minimum was determined previously by Purisima and Scheraga, using an energy-based distance geometry ap-

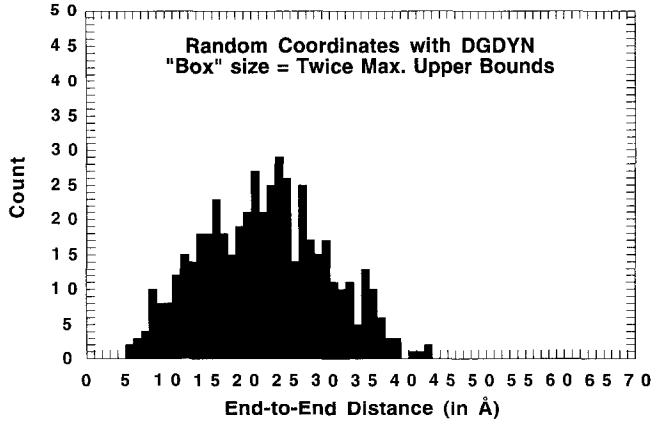

Figure 16. The head-to-tail distance for Ala_{20}. A total of 500 trial conformers was generated with random coordinate DGEOM as in Figure 10, followed by 1 000 steps of DG-DYN.

proach.[45] Although [Met^5]-enkephalin has only 77 atoms, there are an estimated 10^{11} different possible conformations, assuming 3 different minima for each rotatable bond. Of course, not all of these are going to lead to a reasonable, low-energy structure, but a systematic search requires that all of these be scanned. Ideally, we would be able to reproduce this global minimum with a relatively few trial conformations.

A total of 500 conformations of [Met^5]-enkephalin was generated with the same eight DG protocols used in the study of Ala_{20}. Each of these was minimized with SPASMS,[35] using the AMBER force field,[46] with a distance-dependent dielectric of 4. The results of these minimizations are shown in Figures 17–26.

The energy of Scheraga's conformer[44,45] proved to be

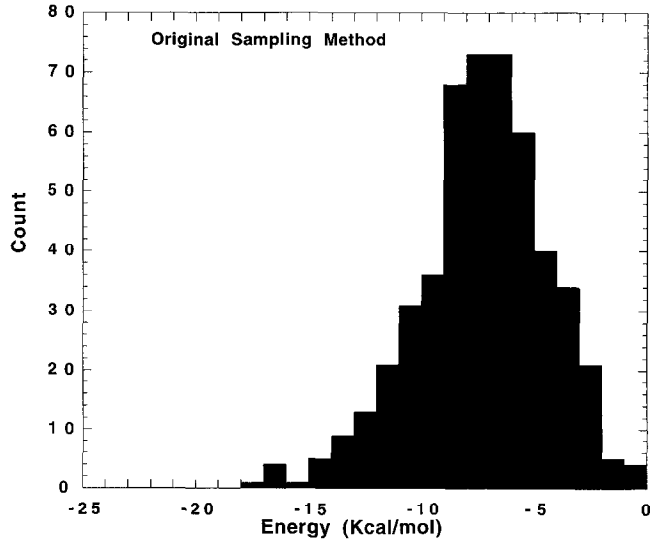

Figure 17. Histogram of energies (kcal/mol) of 500 conformers of [Met^5]-enkephalin generated with the original sampling method.

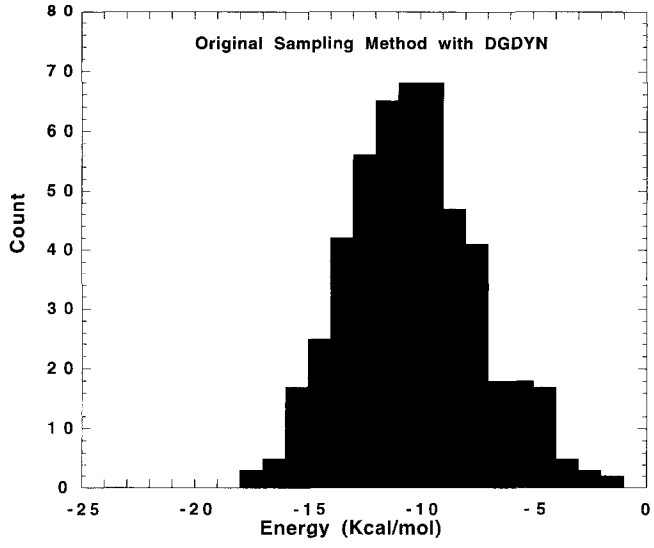

Figure 18. Histogram of energies (kcal/mol) of 500 conformers of [Met^5]-enkephalin generated with the original sampling method followed by 1 000 steps of DGDYN.

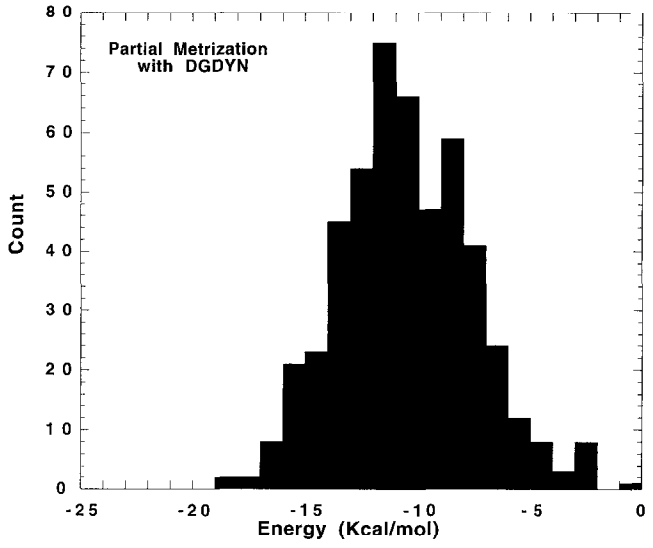

Figure 20. Histogram of energies (kcal/mol) of 500 conformers of [Met^5]-enkephalin generated with the partial metrization DG method followed by 1 000 steps of DGDYN.

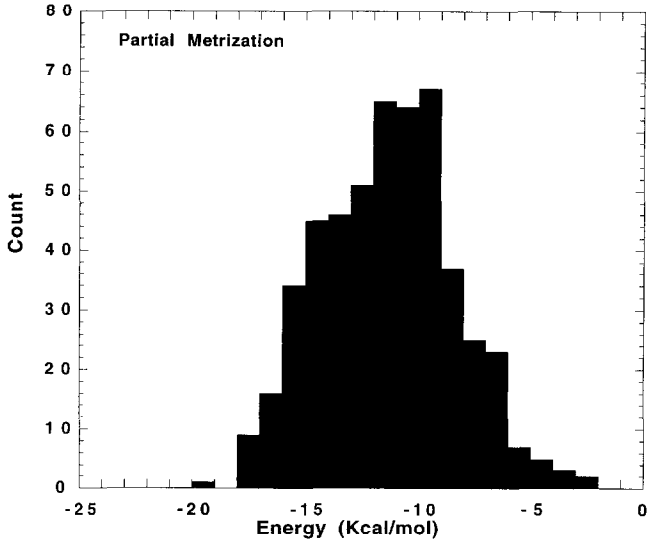

Figure 19. Histogram of energies (kcal/mol) of 500 conformers of [Met^5]-enkephalin generated with the partial metrization DG method.

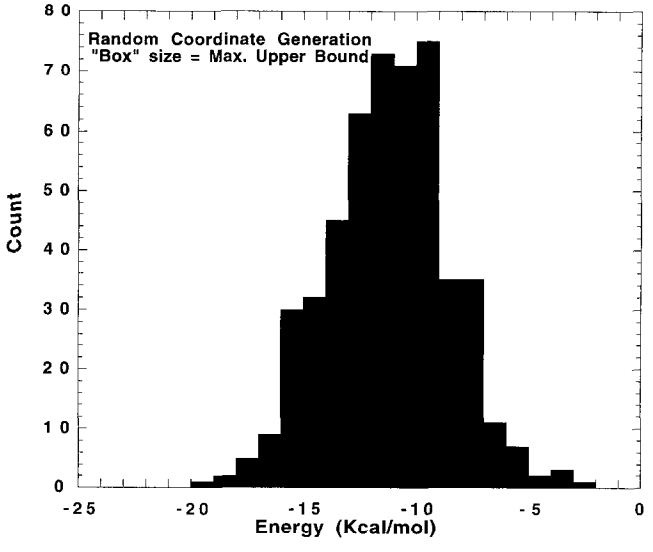

Figure 21. Histogram of energies (kcal/mol) of 500 conformers of [Met^5]-enkephalin generated with the random coordinate DG with a scale factor of 1.0.

the lowest in energy of all conformations studied. We will refer to this conformer as the global minimum, although we have not performed the systematic search required to verify this claim. The energy of the global minimum is −21.06 kcal/mol.

In this case, we use an energetic criterion to evaluate the sampling methods. For this example, all methods produce very similar energy histograms. The conformers range in energy from −19.93 to 0.85 kcal/mol. The standard deviation for all methods is about 3.0 kcal/mol. The original sampling method is the only method with a visibly different histogram, with an average that is substantially higher than those of the other methods. Thus, this method would not be preferred for this type of calculation.

All methods failed to locate the global minimum. However, several methods produced conformers within 2 kcal/mol of the global minimum. The random coordinate method (scale factor of 2) with DGDYN produced the conformer (see Figure 25) with the energy closest to the global minimum (−19.93 kcal/mol). All but one method (the original sampling method) produced conformers within 2 Å RMSD of the global minimum. The partial metrization method produced the conformer (see Figure 26) with the lowest RMSD (computed for all nonhydrogen atoms) from the global minimum.

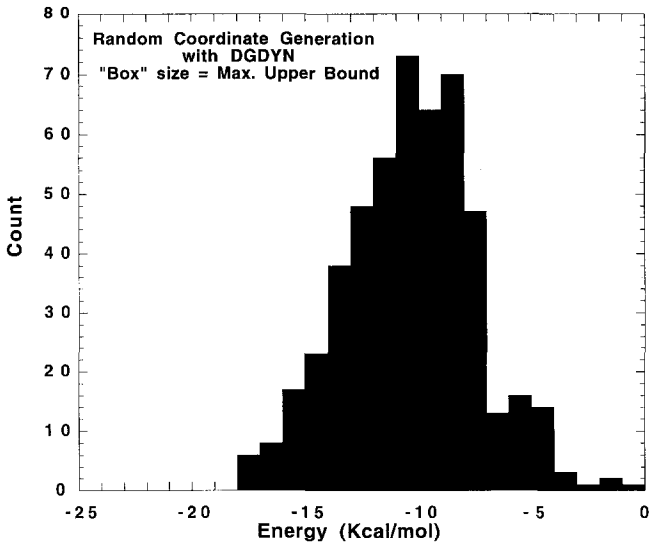

Figure 22. Histogram of energies (kcal/mol) of 500 conformers of [Met^5]-enkephalin generated with the random coordinate DG with a scale factor of 1.0 followed by 1 000 steps of DGDYN.

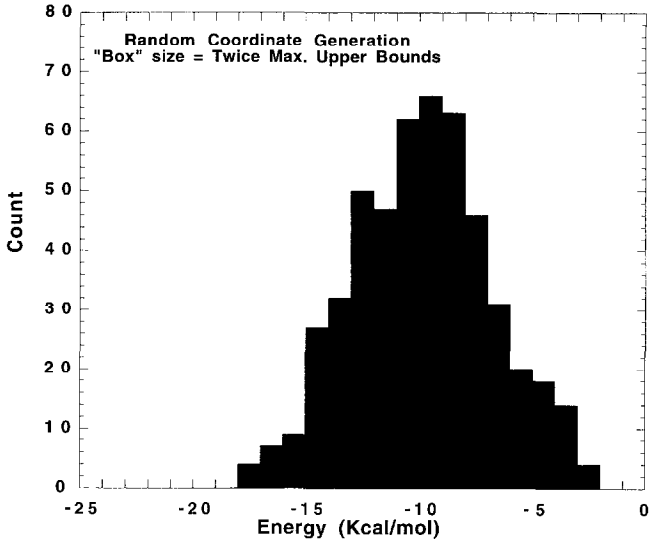

Figure 23. Histogram of energies (kcal/mol) of 500 conformers of [Met^5]-enkephalin generated with the random coordinate DG with a scale factor of 2.0.

A DNA bisintercalator

Blaney and Dixon have used distance geometry to build models[2] of a macrocyclic bisacridine (SDM) complex with DNA based on initial work by Veal et al.[30] This study provides a challenging conformational analysis problem.

Blaney and Dixon were able to show that with a few constraints, models of the SDM–DNA complex could be built quickly.[2] The SDM monomer was built with Sketcher[47] and the DNA coordinates were taken from the X-ray crystal structure (1D32 from the Brookhaven Protein Databank)[48] of the

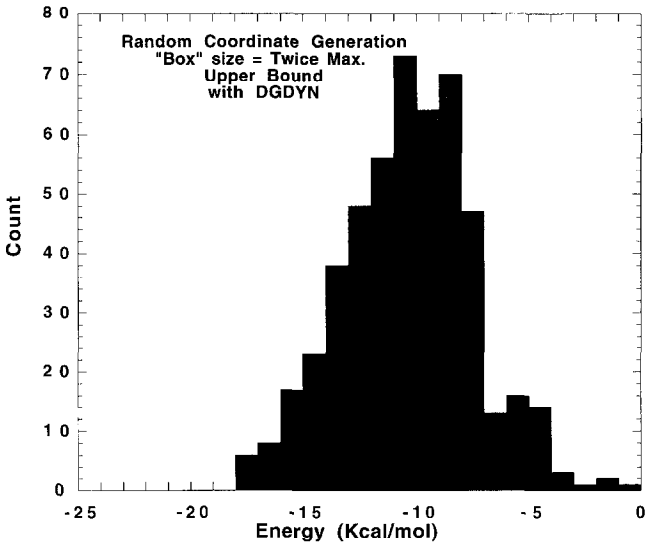

Figure 24. Histogram of energies (kcal/mol) of 500 conformers of [Met^5]-enkephalin generated with the random coordinate DG with a scale factor of 2.0 followed by 1 000 steps of DGDYN.

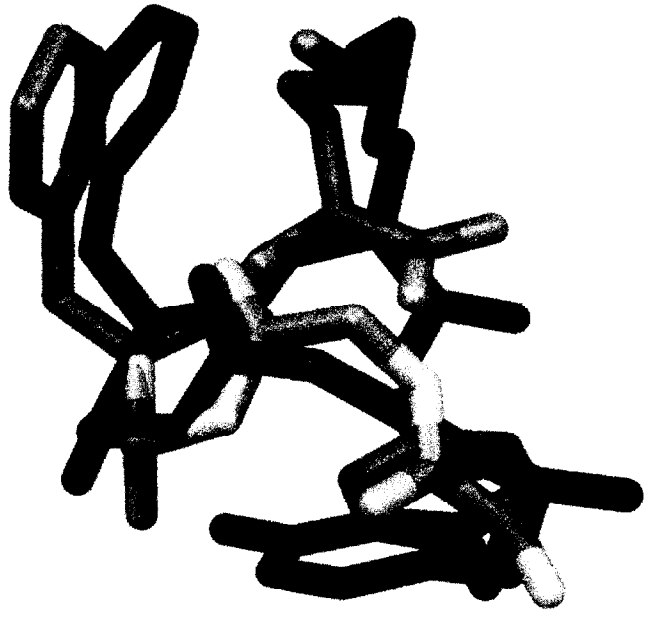

Figure 25. Lowest energy conformer of [Met^5]-enkephalin found with random coordinate generation and scale factor of 2 with 1 000 steps of DGDYN. Global minimum is shown in black.

ditercalinium-d(CGCG)$_2$ complex. They used distance geometry to:

1. Build the cyclic dimer of the macrocycle (monomer shown in Figure 27)
2. Keep the DNA conformationally rigid
3. Set the SDM–DNA lower bounds to van der Waals distances

Figure 26. [Met^5]-enkephalin conformer with lowest RMSD to global minimum (in black). This conformer was found with partial metrization. RMSD = 1.350 Å.

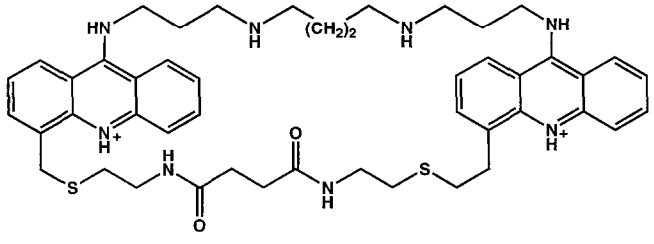

Figure 27. Structure of SDM bisacridine.

4. Dock the central ring of each acridine near the center of each intercalation site on the DNA

They were able to show that some modest constraints, an arbitrary SDM monomer conformation, and the DNA coordinates were sufficient to produce several families of binding. These can be divided into three structurally reasonable groups.

1. Conformations in which the SDM linkers bind in opposite grooves of DNA (Figure 28).
2. Conformations in which both SDM linkers bind in the major groove of DNA (Figure 29).
3. Conformations in which both SDM linkers bind in the minor groove of DNA (Figure 30).

In addition, Blaney and Dixon identified at least one structure in which one of the linkers wrapped around the outside of the phosphate backbone, although this is an unlikely conformation.[2]

For the purposes of this study, we are interested in the ability of each of the methods to populate each of these three binding modes. We will compare the resulting populations, the number

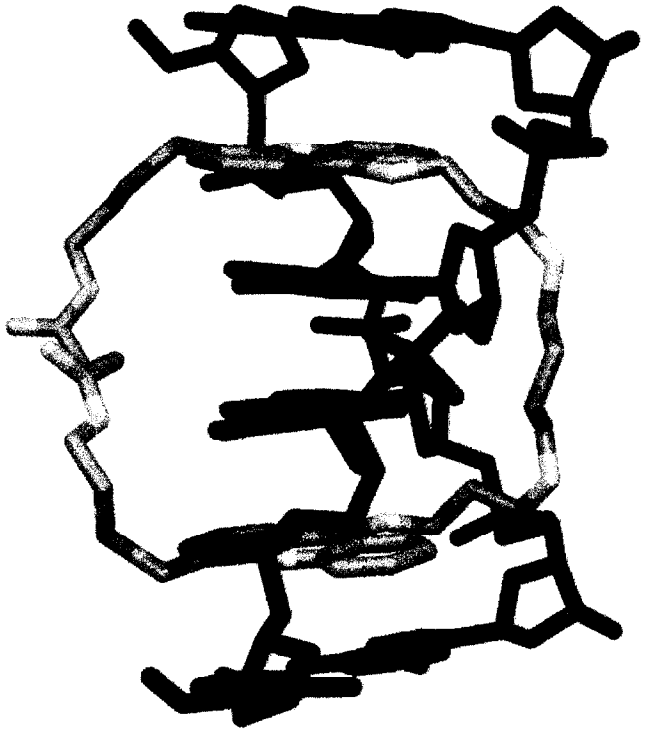

Figure 28. Conformations in which the SDM linkers bind in opposite grooves of DNA.

of trials that succeed out of a total of 50 initial trials, and the amount of computer time required to run each of the calculations.

The input conformations and the constraints files described by Blaney and Dixon (and provided as part of the DGEOM95 distribution package) were used with no changes. The results from 50 trials for these five methods are shown in Table 4:

1. The original sampling with torsion sampling
2. Inclusion of 1 000 steps of DGDYN with the original sampling method
3. Partial metrization
4. Inclusion of 1 000 steps of DGDYN with the partial metrization method
5. The random coordinate generation with a scale factor of 2

Since the original sampling method tends to populate extended conformers, one would expect the linker arms to be generated far apart from each other. The result would be a tendency to place the linkers on opposite sides of the DNA. Indeed, of the 34 solutions from a total of 50 initial trials generated with this method, only 3 were found in which the linker arms were in the same groove of DNA. This method required 2 532 s to complete, running on an IBM RS-6000/580.

Inclusion of 1 000 steps of DGDYN decreased the number of converged solutions to 26 and required more computer time (5 410 s). Only one of the conformers placed both linker arms in the major groove, and no conformers placed both linker arms the minor groove. In this DGDYN simulation, one initial trial was lost because the dynamics routine produced an infinity in the kinetic term, resulting in total loss of the coordinates.

Partial metrization provided the fewest number of converged

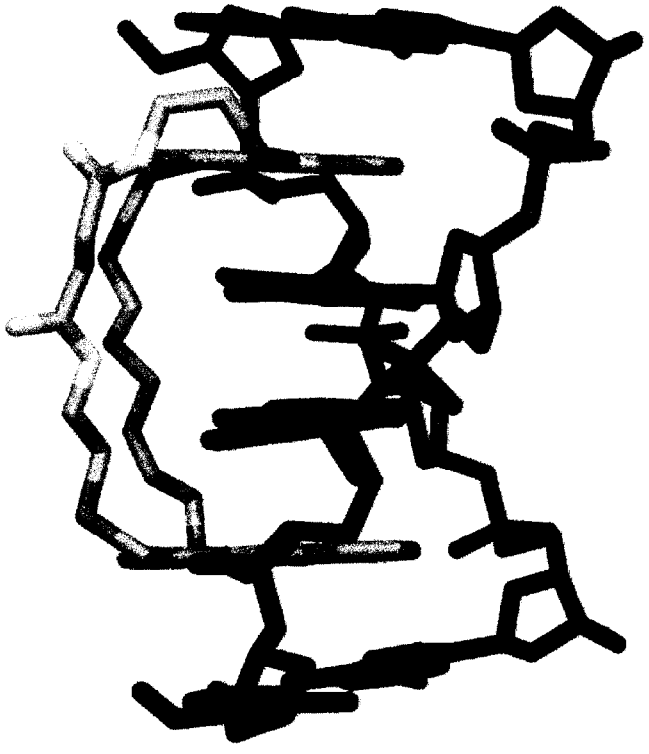

Figure 29. Conformations in which both SDM linkers bind in the major groove of DNA.

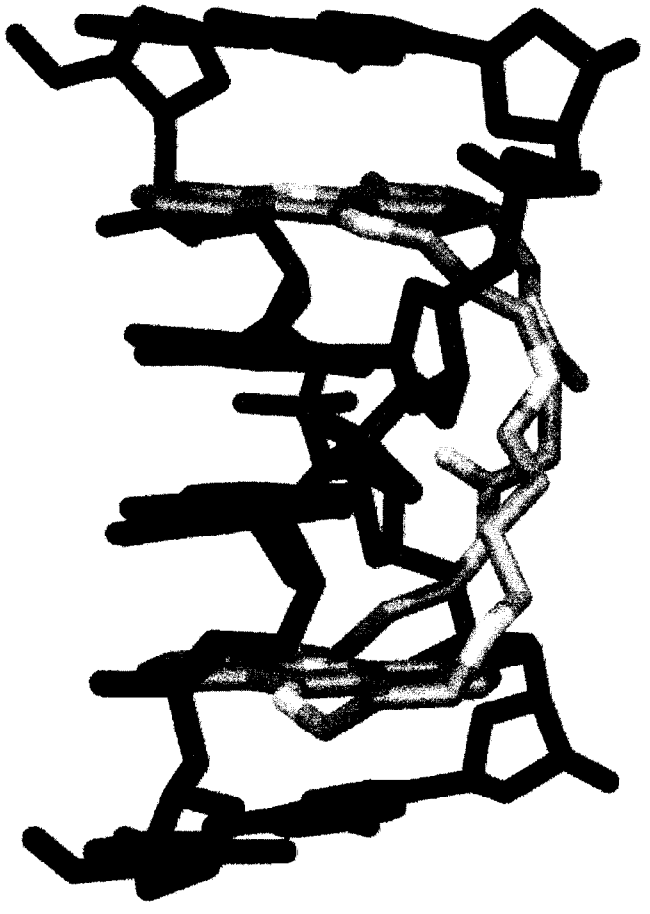

Figure 30. Conformations in which both SDM linkers bind in the minor groove of DNA.

solutions (13), and the most with both linker arms in either the major (2) or minor (7) grooves. This method required 11 694 s of computer time. Thus the efficiency in producing converged structures is much lower than the original sampling method. However, the improved sampling might be worth the additional computer time.

Addition of 1 000 DGDYN steps to the partial metrization method produced 24 conformers out of a trial of 50. One was lost to a failure in the dynamics calculation. Of these, 17 are conformers in which the linker arms are displayed in different grooves of the DNA. Six are now found with both linker arms in the major groove, and 1 is found with both linker arms in the minor groove. This sampling profile is very similar to that seen in the random coordinate method and the original sampling method with or without DGDYN. It is fairly expensive in terms of computer time, too, requiring 15 906 s to complete 50 trials.

The random coordinate method produced fewer conformers (28) than the original sampling method, in about 50% more computer time (4 068 s). However, this method showed a tendency to populate conformations in which the SDM linker arms were in the major groove of DNA, finding 6 such conformers. Thus the conformational efficiency and the ability to populate all three conformer classes make this an attractive method for this type of problem.

Pancreatic trypsin inhibitor: A difficult NMR problem set

Since distance geometry is applied widely to NMR data sets, it is important to evaluate the ability of the sampling methods to solve an NMR structure. In this case, we have chosen a fairly difficult NMR data set originally published by Havel.[5] A total of 1 020 nOes are used in the solution of the NMR structure of bovine pancreatic trypsin inhibitor (BPTI). In addition, BPTI contains nearly 900 atoms, making this a very large system for study.

The maximum upper bounds for each type of nOe were set in the same manner as in Havel[5]: "strong" is assigned a maximum upper bound of 2.5 Å, "medium" is assigned 3.0 Å, and "weak" is assigned 4.5 Å. Several of these nOes are to unspecified protons in methylene or methyl groups. In these cases, a dummy atom was created and held at the center of the two (methylene groups) or three protons (methyl groups) and the nOe was assigned to the dummy atom.

Four simulations with this data were run. The original sampling method, partial metrization, and two random coordinate generation methods were all used to solve the NMR structure with the nOes in place. The CPU time required to refine 10 trials was compiled with each method.

The results are somewhat surprising. The original sampling method was able to solve the NMR structure relatively easily, with all 10 trials converging to solutions. This required 4.1 h on an IBM RS-6000/580. The partial metrization approach as implemented in DGEOM95[29] failed to converge any of the trials and required 36.6 h to complete. The random coordinate method with a scaling factor of 2 also failed to converge any of

Table 4. Comparison of coordinate generation methods for the generation of models of the SDM–DNA complex

Method used	Number of conformers generated	Number of conformers with SDM linkers in opposite grooves of DNA	Number of conformers with both SDM linkers in major groove of DNA	Number of conformers with both SDM linkers in minor groove of DNA	Computer time required for simulation
Original sampling	34	31	1	2	2 532
Original sampling with 1 000 steps of DGDYN	26	25	1	0	5 410
Partial metrization	13	4	2	7	11 694
Partial metrization with DGDYN	24	17	6	1	15 906
Random coordinate generation with scale factor 2	28	20	6	2	4 068

the 10 trials. However, running the random coordinates with a scaling factor of 30 produced 3 converged structures from 10 trials in a total of 15.7 h. An attempt to increase the scaling factor to 50 also converged approximately 30% of the structures (a total of 10 of 33 initial trials converged).

Thus, in this particular case, one cannot do better than the original sampling method. The partial metrization results are somewhat surprising. In an attempt to determine if the partial metrization method could reach a solution, DGEOM95 was allowed to generate structures until one converged. After 50 initial random trials, one converged structure was generated and the simulation terminated. Because the computer time required is rather substantial, and because the original sampling method performed so well, we chose not to study the effect of DGDYN on this system.

We should note that the Havel and Snow full metrization method[23] and the Kuszewski et al. partial metrization methods[24] are both routinely applied successfully to the solution of protein NMR. We have included the PTI example only for comparison of the ability of DGEOM95 to handle a wide range of problems. While several variants of the embedding algorithms do produce correct structures for this data set, it is likely that the general package will not be competitive with the specialized programs for solving protein structures.

Liquid water

We implemented periodic boundary conditions in DGEOM95[29] with the intention to create solvent boxes for molecular dynamics or Monte Carlo simulations. The first application of this procedure was to create a box of 216 TIP4P water[31] molecules. The three-atom H–O–H framework for the TIP4P molecules was used to create the water box with DGEOM95; the fourth point is created as part of the molecular dynamics calculation in SPASMS.[35] The lower bounds for each intermolecular interaction were set to the van der Waals radius of water. A box 20 Å on a side was used for the periodic imaging. The upper bounds were set to match the length of the diagonal of this box. Random coordinate DGEOM was used to create 40 boxes of TIP4P waters. Each resulting configuration represents an independent trial configuration.

The parameter files used for the SPASMS studies were set up using the AMBER 3.0A suite of programs.[49] Each configuration was converted from a pdb file to a coordinate input file for use in SPASMS. Each was subjected to 10 steps of conjugate gradient optimization (as AMBER TIP3P water,[46] since the SETTLE constraints[50] cannot be applied to minimizations), and 2 ps of NVT molecular dynamics with SPASMS. Each was then subjected to 40 ps of NPT molecular dynamics. For the averaging runs, a step size of 1.5 fs was used, with residue-based cutoffs of 8.5 Å, and residue-based switching functions were applied from 7.75 to 8.0 Å. Nonbonded updates were performed every 10 steps. Stochastic collisions with a reference temperature of 300 K were performed every 1 000 steps. SETTLE was performed on all bonds during the simulation.[50]

Close nonbonded contacts in which a charge term became infinite caused 7 of the original 40 boxes to fail during molecular mechanical minimization. These could likely have been saved through the use of a large van der Waals radius on the protons during minimization. However, this study was intended as an example, so no effort was expended to rescue these seven runs.

The average, maximum, and minimum energies for the remaining 33 simulations are shown in Figure 31, while the volume fluctuations are shown in Figure 32. Although a great deal of statistical analysis might be performed on these simulations, it is the intent of this work to demonstrate that the DGEOM-generated boxes can be used in a molecular dynamics simulation. Therefore, it is of interest to show that all of the boxes converge on the experimental density of about 30 $Å^3$ per water molecule. Indeed, within 20 ps, the volume has converged. For the last 100 ps of the simulation, the volume fluctuates about an average of 6 462.7 ± 16.0 $Å^3$ for the box, or 29.92 $Å^3$ per water molecule. The experimental value is 29.92 $Å^3$ per molecule, in agreement with the value reported in the original TIP4P work.[31]

DGEOM95 could likely be optimized to handle boxes of solvent. In particular, it should be possible to increase the efficiency of the convergence of the initial boxes and computational speed. In this study, only 40 of 74 trials met the DGEOM95 convergence criteria, and required a total of 7.1 h

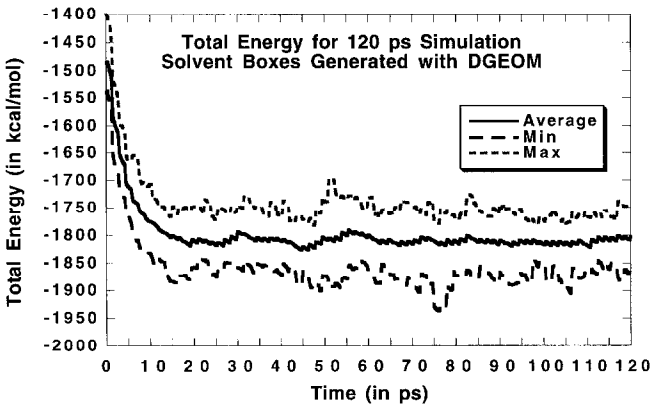

Figure 31. Total energy for 120-ps simulations of 216 TIP4P waters.

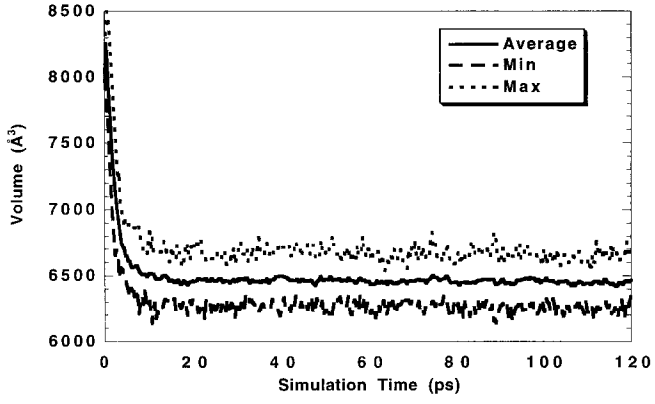

Figure 32. Volume fluctuations for 120-ps simulations of 216 TIP4P waters.

of computer time to generate. Each SPASMS minimization required approximately 6 s to complete, the equilibration steps required approximately 300 s to complete, and the 120-ps simulations require approximately 4.7 h to complete. All times reported are for an IBM RS-6000/580 or 370 with 64 MB of memory, running AIX 3.2.4.

Liquid pentane

The usefulness of the DGEOM95 solvent box approach can also be demonstrated with more flexible ligands. In this example, 10 boxes of 125 liquid *n*-pentane molecules were generated using the periodic imaging-based random coordinate DGEOM method. The triangle bounds smoothing method in DGEOM95 scales as the cube of the number of atoms, so the solvent boxes were generated with only the 5 carbon atoms as part of the DGEOM step, rather than the total complement of 17 atoms. A periodic box of 30 Å per side was used. The van der Waals radius of carbon was set to 1.60 Å. A total of 20 boxes was generated, requiring 3.5 h of CPU time on the same machine as used above. Only 10 of these were used in the remaining study because of the immense amount of computer time required to complete the molecular dynamics study.

These 10 boxes of united atom pentane were converted to all-atom pentane using the general model-building features of PSSHOW.[43] They were converted to the AMBER 3.0A[49] coordinate input files and subjected to four steps in the molecular mechanical study with SPASMS. The carbon and hydrogen nonbonded parameters are those described in Cornell et al.,[51] as are the partial atomic charges. The first step was a 10-step minimization with periodic boundary conditions, employing a 12-Å cut-off. The box size was set to 30 Å, and the molecules were imaged on the central carbon using a residue-based pair list. A 0.1-ps NVT simulation with a time step of 0.5 fs with stochastic collisions every step (temperature of 300 K) was used to remove remaining high-energy contacts. A 30-ps NPT simulation with a time step of 1.5 fs, stochastic collisions[33] every 1.5 ps, a reference temperature of 300 K, and RATTLE[52] applied to all bonds was used as the equilibration and averaging run. A 3-ps NPT simulation was used to collect the dihedral angle samples. CPU times for the first two steps were approximately 15 min each, the 30-ps simulation required 24.5 h, and the dihedral sampling run required 2.5 h. In total, more than 11 days of CPU time was required to complete these simulations.

The average total energy for the 30-ps simulation is shown in Figure 33, and the total volume fluctuations are shown in Figure 34. These are substantially noisier than those observed in the water simulations, primarily due to three factors. First, the simulations are substantially shorter, and the system might not have reached equilibrium even after 30 ps. Second, there are only 125 pentane molecules in this study. Smaller system sizes tend to produce noisier energy profiles. Third, there are only 10 simulations, rather than 33. Again, this would lead to a more noisy system. Nonetheless, it does appear that these boxes have reached equilibrium at about 15 ps of simulation time. The average volume of all 10 averaging runs is 199.9 (±11.4) Å^3 per molecule. This compares quite favorably with 192.8 Å^3 reported for *n*-pentane.[53] Figure 35 shows the theoretical population density from the gas-phase energy distribution for *n*-pentane. Liquid hydrocarbons tend to display similar profiles in the gas phase, as in the liquid.[53] To show that valuable information can be obtained from these simulations, we have chosen to plot the dihedral angle distribution obtained from these simulations. Those are shown in Figure 36. Note that the contouring for the simulations is somewhat noisier than

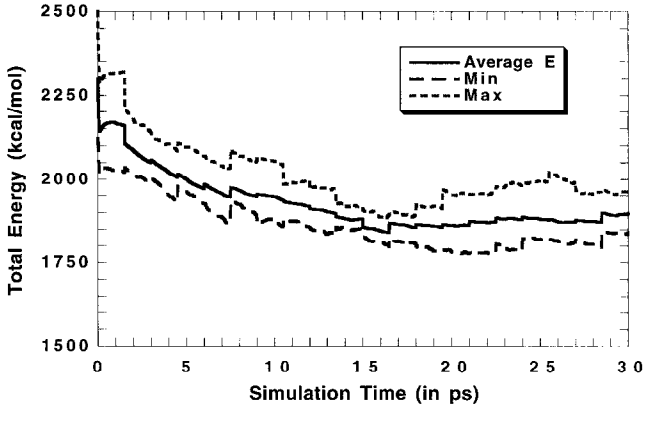

Figure 33. Total energy of ten 30-ps NPT simulations of liquid n*-pentane (125 molecules; all atom) using coordinates generated with distance geometry.*

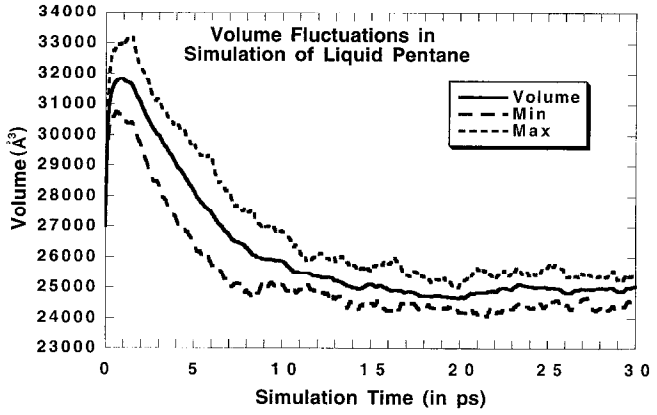

Figure 34. Volume fluctuations of ten 30-ps NPT simulations of liquid n*-pentane (125 molecules; all atom) using coordinates generated with distance geometry.*

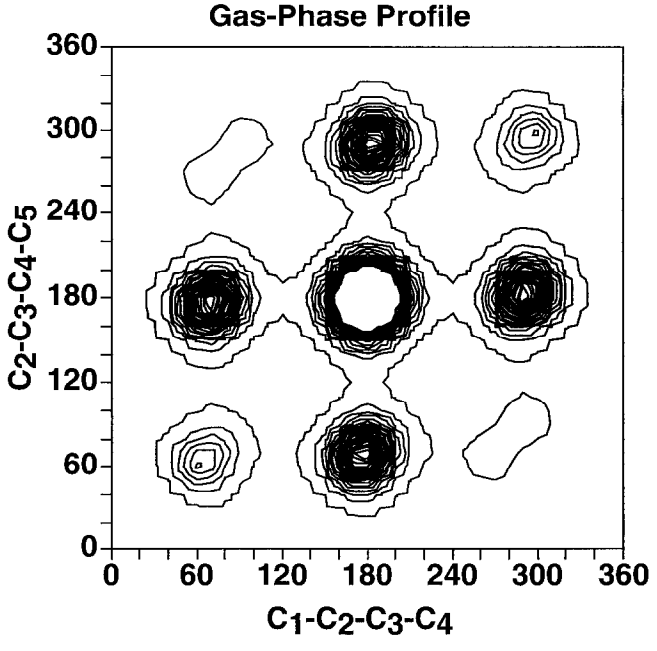

Figure 35. Theoretical dihedral angle distribution of n*-pentane generated from gas-phase energy profile.*

the theoretical values. However, it is clear that the energy distributions are very similar. A similar profile could certainly be obtained from a single, longer simulation.

CONCLUSIONS

We have presented the application of a variety of distance geometry methods to the conformational analysis of several examples. One new coordinate generation method (the random coordinate approach) and one new sampling method (DGDYN) have been applied and compared to the original sampling method and partial metrization approaches implemented in DGEOM.

The original sampling method performed well in solving the NMR structure of BPTI, but fared less well in other examples.

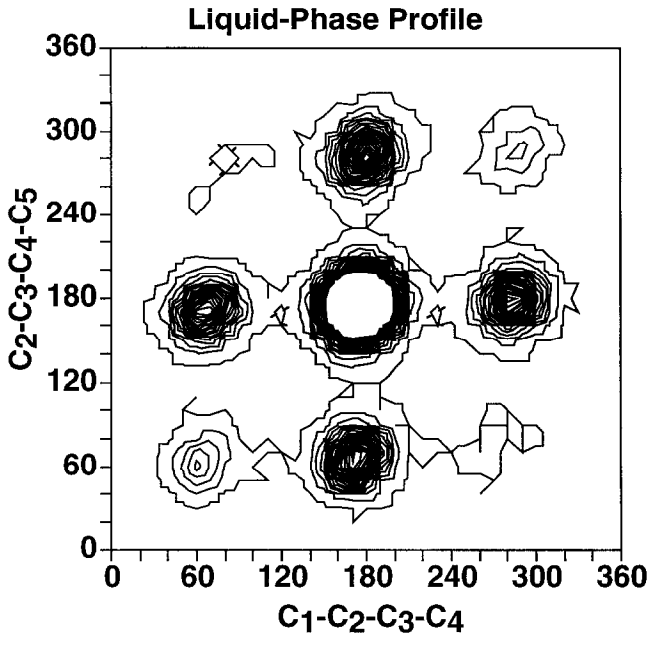

Figure 36. Dihedral angle distribution of liquid n*-pentane generated from 10 NPT simulations. Values for dihedral angles were sampled over the last 3 ps for all 10 simulations.*

It generated the largest number of solutions of the bisintercalator, but did not do well in the generation of unique conformers. It was able to generate the largest number of low-energy conformers of C_{17} but generated the fewest number of unique conformers. It performed very poorly in generating Ala_{20} conformers, producing extended conformers only. It also performed poorly in generating low-energy conformers of [Met^5]-enkephalin. In all cases, this method produced conformers in less CPU time than any other method. Thus, this method would be appropriate for generating structures from NMR data sets.

Partial metrization performed well in generating low-energy conformers of C_{17} and [Met^5]-enkephalin. This method produced the largest number of short end-to-end distances in Ala_{20}. It generated the largest number of minor groove solutions to the bisintercalator example, but generated the fewest number of solutions overall. The partial metrization method implemented in DGEOM95 was unable to generate any conformations of BPTI that were compatible with the NMR constraints. This is more computationally expensive than both original sampling and random coordinate generation methods.

The random coordinate generation method performed well in generating unique conformers in C_{17}, but did not locate many of the lowest energy conformers. With a scale factor of 1, this method sampled Ala_{20} conformers with shorter end-to-end distances. Increasing the scale factor to 2 produced the most uniform distribution of end-to-end distances. This method produced fewer solutions than original sampling in the bisintercalator example, but produced more major groove binders. It required modest amounts of additional CPU time than the original sampling method, but substantially less than partial metrization. The random coordinate methods performed as well as the other methods for [Met^5]-enkephalin. The performance on the NMR structure of BPTI was intermediate be-

tween those of the original sampling and partial metrization methods.

Inclusion of DGDYN improved the sampling and the location of low-energy conformers when combined with the random coordinate method in the C_{17} example. This method was able to generate the largest number of unique C_{17} conformers. Inclusion of DGDYN with any of the coordinate generation methods in Ala_{20} produced conformers with shortened end-to-end distances. Sampling was slightly poorer with the inclusion of DGDYN with the original sampling method, but slightly better when combined with the partial metrization method in the bisintercalator example.

None of the methods was able to locate the global energy minimum in [Met^5]-enkephalin within 500 random trials. However, all methods—except the original sampling—were able to locate conformers within 2 kcal/mol or 2.0-Å RMSD of the minimum. This is often sufficiently good for most applications of stochastic conformational analysis.

We have demonstrated that one can use distance geometry to produce boxes of either water or pentane. The input is relatively simple, consisting of only a single monomer (or the solute and one solvent monomer). The result is a nonbiased box that can readily be equilibrated for statistical mechanical simulations.

Surprisingly, random coordinate generation performs competitively with several variants of embedding algorithms on a range of model-building problems. This technique is very simple, general, and works well on simple conformational analysis problems and complicated models of intermolecular interactions. It usually provides comparable sampling in less computer time than partial metrization. It is not competitive with specialized embedding and sampling algorithms for solving protein structures from NMR data sets, although it does produce correct solutions. The best random coordinate method appears to use a box size of double the largest distance bound. We recommend this as a standard default method for most modeling problems in distance geometry.

REFERENCES

1 Crippen, G.M. and Havel, T.F. Stable calculation of coordinates from distance information. *Acta Crystallogr.* 1978, **A34,** 282–284

2 Blaney, J.M. and Dixon, J.S. In: *Reviews in Computational Chemistry* (K.B. Lipkowitz and D.B. Boyd, eds.), Vol. 5. VCH, New York, 1993, pp. 299–332

3 Peishoff, C.E., Dixon, J.S., and Kopple, K.D. Application of the distance geometry algorithm to cyclic oligopeptide conformation searches. *Biopolymers* 1990, **30,** 45–56

4 Peishoff, C.E. and Dixon, J.S. Improvements to the distance geometry algorithm for conformational sampling of cyclic structures. *J. Comput. Chem.* 1992, **13,** 565–569

5 Havel, T.F. and Wütrich, K. An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. *J. Mol. Biol.* 1985, **182,** 281–294

6 Saunders, M., Houk, K.N., Wu, Y.-D., Still, W.C., Lipton, M., Chang, G., and Guida, W.C. Conformations of cycloheptadecane: A comparison of methods for conformational searching. *J. Am. Chem. Soc.* 1990, **112,** 1419–1427

7 Leach, A.R. In: *Reviews in Computational Chemistry* (K.B. Lipkowitz and D.B. Boyd, eds.), Vol. 2. VCH, New York, 1991, pp. 1–55

8 Lipton, M. and Still, W.C. The multiple minimum problem in molecular modeling: Tree searching internal coordinate conformational space. *J. Comput. Chem.* 1988, **9,** 343–355

9 Go, N. and Scheraga, H.A. Ring closure and local conformational deformations of chain molecules. *Macromolecules* 1970, **3,** 178

10 Bruccoleri, R.E. and Karplus, M. Chain closure with bond angle variations. *Macromolecules* 1985, **18,** 2767–2773

11 Bruccoleri, R.E. and Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* 1987, **26,** 137–168

12 Moult, J. and James, M.N.G. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins Struct. Funct. Genet.* 1986, **1,** 146–163

13 Motoc, I., Dammkoehler, R.A., Mayer, D., and Labanowski, J. Three-dimensional quantitative structure–activity relationships. I. General approach to the pharmacophore model validation. *Quant. Struct. Activity Relation* 1986, **5,** 99–105

14 Motoc, I., Dammkoehler, R.A., and Marshall, G.R. In: *Mathematical and Computational Concepts in Chemistry* (N. Trinajstic, ed.). Ellis Horwood, Chichester, 1986, pp. 222–251

15 Dammkoehler, R.A., Karasek, S.F., Shands, E.F.B., and Marshall. G.R. Constrained search of conformational hyperspace. *J. Comput. Aided Mol. Des.* 1989, **3,** 3–21

16 Wilson, S.R. and Cui, W. In: *The Protein Folding Problem and Tertiary Structure Prediction* (J.K. Merz and S. LeGrand, eds.). Birkhauser, Boston, 1994, pp. 43–70

17 Chang, G., Still, W.C., and Guida, W.C. An internal coordinate Monte Carlo method searching conformational space. *J. Am. Chem. Soc.* 1989 **111,** 4379

18 Senderowitz, H., Guarnieri, F., and Still, W.C. A smart Monte Carlo technique for free energy simulations of multiconformational molecules: Direct calculations of the conformational populations of organic molecules. *J. Am. Chem. Soc.* 1995, **117,** 8211–8219

19 Judson, R.S., Colvin, M.E., Meza, J.C. Huffer, A., and Gutierrez, D. Do Intelligent configuration search techniques outperform random search for large molecules? *Int. J. Quantum Chem.* 1992, **44,** 277–290

20 Ferguson, D.M. and Raber, D.J. A new approach to probing conformational space with molecular mechanics: Random incremental pulse search. *J. Am. Chem. Soc.* 1989, **111,** 4371–4378

21 Guarnieri, F. and Still, W.C. A rapidly convergent simulation method: Mixed Monte Carlo/stochastic dynamics. *J. Comput. Chem.* 1994, **15,** 1302–1310

22 Havel, T. and Wuthrich, K. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular H–H proximities in solution. *Bull. Math. Biol.* 1984, **46,** 673–698

23 Havel, T.F. and Snow, M.E. A new method for building protein conformations from sequence alignments with homologues of known structure. *J. Mol. Biol.* 1991, **217,** 1–7

24 Kuszewski, J., Nilges, M., and Brunger, A.T. Sampling and efficiency of metric matrix distance geometry—a novel partial metrization algorithm. *J. Biomol. Nucl. Magn. Reson.* 1992, **2,** 33–56
25 Havel, T.F. An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog. Biophys. Mol. Biol.* 1991, **56,** 43–78
26 Havel, T.F. The sampling properties of some distance geometry algorithms applied to unconstrained polypeptide chains: A study of 1830 independently computer conformations. *Biopolymers* 1990, **29,** 1565–1585
27 Crippen, G.M. *Distance Geometry and Conformational Calculations,* Vol. 1. Research Studies Press (Wiley), New York, 1981
28 Crippen, G.M. and Havel, T.F. *Distance Geometry and Molecular Conformation.* Research Studies Press (Wiley), New York, 1988
29 Blaney, J.M., Crippen, G.M., Dearing, A., Dixon, J.S., and Spellmeyer, D.C. DGEOM95. Available from the Quantum Chemistry Program Exchange, Bloomington, Indiana, 1995
30 Veal, J.M., Li, Y., Zimmerman, S.C., Lamberson, C.R., Cory, M., Zon, G., and Wilson, W.D. Interaction of a macrocyclic bisacridine with DNA. *Biochemistry* 1990, **29,** 10918–10927
31 Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 1983, **79,** 926–935
32 Swope, W.C., Andersen, H.C., Berens, P.H., and Wilson, K.R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* 1982, **76,** 637–649
33 Andersen, H.C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* 1980, **72,** 2384–2393
34 Allen, M.P. and Tildesley, D.J. *Computer Simulations of Liquids,* Oxford Science Publications, New York, 1987
35 Spellmeyer, D.C., Swope, W.C., Evensen, E.-R., and Ferguson, D.M. SPASMS. Available from Department of Pharmaceutical Chemistry, University of California, San Francisco, California, 1992
36 Allinger, N.L. Conformational analysis. 130. MM2: A hydrocarbon force field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc.* 1977, **99,** 8127–8134
37 Weinberg, N. and Wolfe, S. A comprehensive approach to the conformational analysis of cyclic compounds. *J. Am. Chem. Soc.* 1994, **116,** 9860–9868
38 Goto, H. and Osawa, E. Further developments in the algorithm for generating cyclic conformers. Test with cycloheptadecane. *Tetrahedron Lett.* 1992, **33,** 1343–1346
39 Shenkin, P.S. and McDonald, D.Q. Cluster analysis of molecular conformations. *J. Comput. Chem.* 1994, **15,** 899–916
40 Allinger, N.L., Yuh, Y.H., and Lii, J.H. Molecular Mechanics. The MM3 force field for hydrocarbons. 1. *J. Am. Chem. Soc.* 1989, **111,** 8551–8566
41 Still, W.C., MacPherson, L.J., Harada, T., Callahan, J.F., and Rheingold, A.L. Macromodel. *Tetrahedron* 1984, **40,** 2775
42 Walters, P. and Stahl, M. Babel, version 1.06. University of Arizona, Tuscon, Arizona, 1994
43 Dearing, A. and Swanson, E. PSSHOW, version 1.9. Seattle, Washington, 1994
44 Nayeem, A., Vila, J., and Scheraga, H.A. A comparative study of the simulated annealing and Monte Carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [Met]-enkephalin. *J. Comput. Chem.* 1991, **12,** 594–605
45 Purisima, E.O. and Scheraga, H.A. An approach to the multiple-minima problem in protein folding by relaxing dimensionality. Tests on enkephalin. *J. Mol. Biol.* 1987, **196,** 697
46 Weiner, S.J., Kollman, P.A., Nguyen, D.T., and Case, D.A. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* 1986, **7,** 230–252
47 Biosym Technologies. *Sketcher.* Biosym Technologies, San Diego, California, 1992
48 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 1977, **112,** 535–542
49 Singh, U.C., Weiner, P.K., Caldwell, J., and Kollman, P.A. AMBER 3.0A. Available from Department of Pharmaceutical Chemistry, University of California, San Francisco, California, 1989
50 Miyamoto, S. and Kollman, P.A. SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* 1992, **13,** 952–962
51 Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M.J., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. A second generation force-field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc.* 1995, **117,** 5179–5197
52 Andersen, H.C. Rattle: a "velocity" version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* 1983, **52,** 24–34
53 Jorgensen, W.L., Madura, J.D., and Swenson, C.J. Optimized intermolecular potential functions for liquid hydrocarbons. *J. Am. Chem. Soc.* 1984, **106,** 6638–6646