Linearized Embedding: A New Metric Matrix Algorithm for Calculating Molecular Conformations Subject to Geometric Constraints

Gordon M. Crippen

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

Received 16 September 1988; accepted 6 February 1989

There are many methods in the literature for calculating conformations of a molecule subject to geometric constraints, such as those derived from two-dimensional NMR experiments. One of the most general ones is the EMBED algorithm, based on distance geometry, where all constraints except chirality are converted into upper and lower bounds on interatomic distances. Here we propose a variation on this where the molecule is assumed to have fixed bond lengths, vicinal bond angles and chiral centers; and these holonomic constraints are enforced separately from the experimental constraints by being built into the mathematical structure of the problem. The advantages of this approach are: (1) for molecules having large rigid groups of atoms, there are substantially fewer variables in the problem than all the atomic coordinates; (2) rigid groups achieve in the end more accurate local geometry (e.g., planar aromatic rings are truly planar, chiral centers always have their correct absolute chirality); (3) it is easier to detect inconsistencies between the holonomic and the experimental constraints; and (4) when generating a random sampling of conformers consistent with all constraints, the probability of achieving satisfactory structures tends to be greater.

INTRODUCTION

The molecular embedding problem consists of finding one or more sets of atomic coordinates such that a given list of geometric constraints is satisfied. For instance, suppose a protein has been investigated by two-dimensional NMR, so that we have upper bounds on the distances between perhaps a hundred assigned pairs of protons. From whatever source, we will call these the **experimental** constraints. In addition, one generally assumes that such a molecule is under no great strain, so that all bond lengths and bond angles are known from standard values taken from X-ray crystallographic studies on small molecules. The absolute handedness of all chiral centers is also known from the covalent structure of the molecule. We will refer to these numerous a priori constraints as **holonomic**. The third constraint category consists of the very numerous interatomic lower bounds on distances due to assuming that atoms separated by more than two bonds interact as hard spheres, as given by their standard **van der Waals** radii. Clearly it is possible in such a problem involving hundreds of atoms and thousands of constraints to have no conformations satisfying them all because of mutual inconsistencies among some of the constraints, or perhaps there will be only one correct conformation, or even a whole family of allowed conformers. The goal of the embedding algorithm is to find out which of these three possible outcomes applies, and in the first case to identify the source of conflict, while in the last case to explore the range of possibilities. As a practical matter, a conformation is said to satisfy the constraints if none are violated more than a given amount. For example, in this study we require distance bounds to be obeyed within a 0.5 Å tolerance.

Out of the many approaches to this problem, the standard and very general algorithm is EMBED,[1,2] frequently used for the determination of conformations of small proteins in solution by NMR by applying the specially adapted program DISGEO.[3] In broad terms, the algorithm consists of the following steps:

1. Convert all experimental, holonomic, and van der Waals constraints into upper

Journal of Computational Chemistry, Vol. 10, No. 7, 896–902 (1989)
© 1989 by John Wiley & Sons, Inc.
CCC 0192-8651/89/070896-07$04.00

and/or lower distance bounds on some, if not all, of the interatomic distances.
2. "Smooth" the distance bounds by the triangle inequality so that there is some upper and some lower bound on *all* interatomic distances.
3. Choose independent random values for each distance within its respective range.
4. Convert this trial set of distances to the corresponding metric matrix and find the closest corresponding three-dimensional metric matrix.
5. Calculate trial coordinates from the three-dimensional metric matrix and adjust them by minimizing an error function that has a penalty term for each of the constraints given in step 1.
6. Go back to step 3 until enough successful conformations have been generated.

Step 2 is important enough in what follows to deserve special comment. For any pair of atoms having no given upper distance bound we initially assign a large positive number; any missing lower bounds are given initially as zero. Then any time there are atoms i, j, and k such that $u_{ij} > u_{ik} + u_{kj}$, then one can reduce u_{ij} to the sum on the right-hand side. This is done repeatedly until no upper bound can be lowered further. There is a similar procedure for raising some of the lower bounds, once again involving triples of atoms. This reasoning at the triangle inequality level is vital for spreading out the information given about some distances to form conservative conclusions about all distances. Further tightening of the bounds can be achieved by invoking the tetrangle inequality, which involves four atoms at a time and corresponds in some instances to finding the *cis* and *trans* distance bounds for two atoms joined by three bonds. Tetrangle inequality bound smoothing is much more expensive than the triangle inequality procedure, so it is generally not performed.[2,4]

In practice EMBED works well in most situations because almost all constraints can be adequately represented as distance bounds. The main exception is chirality, which is only incorporated in step 5 as a term that is positive as long as there is a violation. Thus when there are multiple chiral centers having distance constraints linking them, this local minimization step tends to stick in minima with residual errors. Second, planar aromatic ring systems tend to be slightly puckered, even when the final error value is small, because visually disturbing deviations from planarity actually correspond to very small violations in interatomic distance constraints. Third, if there are n atoms in the molecule, then step 4 deals with $n \times n$ matrices, and step 5 is a local minimization in $3n$ variables. This seems wasteful when positioning a large aromatic ring system that may involve a dozen atoms but only six degrees of freedom, since the ring is a rigid body. Fourth, while there are generally a very large number of holonomic constraints, these are generally mutually consistent, but when inconsistencies are signalled in step 2 or 5, it is often difficult to tell whether the experimental constraints are themselves mutually inconsistent or whether they conflict with the holonomic constraints. In this article we will attempt to improve on some of these points by building into the mathematical description of the molecule all the holonomic constraints from the very outset.

METHODS

The molecule is represented in a "linearized" form very similar to that described in our earlier work.[5] First, view the molecule as a tree graph, where the nodes are the atoms and the edges are the covalent bonds (neglecting one bond on every ring for the time being so as to avoid cycles in the graph). Choose as its root one of the atoms with the smallest maximal distance to the other atoms, in the graph theory sense, technically called a center of the tree graph. Since the coordinates of the other atoms will be expressed relative to the root atom, choosing the center helps keep roundoff errors low. In Figure 1, C1 is the root.

Next, set up a collection of local coordinate systems to define the position of every atom. The location of the root atom is simply given by some arbitrary vector, $\mathbf{w}$, relative to an external frame of reference. At the root atom, define a local right-handed orthogonal coordinate system by letting $\mathbf{u}_1$ be the unit vector with origin at C1 along the C1—C4 bond, $\mathbf{u}_2$ is orthogonal to $\mathbf{u}_1$ in the C1—C4—H3 plane, and $\mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2$. This differs slightly from our earlier work,[5] where the coordinate axes were always par-

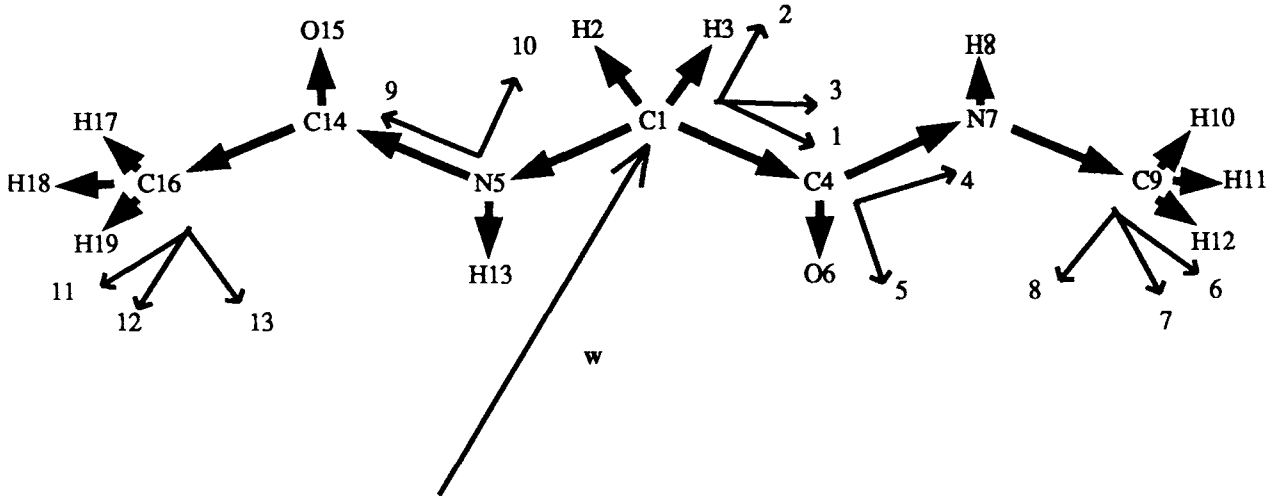

Figure 1. The linearized representation of *N*-acetylglycine-*N'*-methylamide. Atoms are labelled with their atomic symbol and a unique number. Bonds are drawn with heavy arrows indicating descent in the tree. The overall translation vector is **w**, and the unit vectors $\mathbf{u}_1, \ldots, \mathbf{u}_{13}$ are simply indicated by their subscripts. Local coordinate systems of three unit vectors are always right handed.

allel to bonds and therefore in general not orthogonal. Now we can describe the positions of atoms H2, H3, C4, and N5, but that of N7 is affected by the rotatable C1—C4 bond. Therefore we establish yet another system centered on C4, but if we assume fixed planar peptide bonds, only $\mathbf{u}_4$ and $\mathbf{u}_5$ are necessary for the rigid group of atoms {C1, C4, O6, N7, H8, C9}. Proceeding in this fashion all over the molecular tree, we employ altogether $n_u = 13$ unit vectors to account for 19 atoms in this example. Then the position of the ith atom, $\mathbf{a}_i$, is given by

$$\mathbf{a}_i = \mathbf{w} + \sum_{j=1}^{n_u} \alpha_{ij}\mathbf{u}_j \qquad (1)$$

where the αs are coefficients that depend only on which atom and which unit vector are involved, not on the conformation.

As the molecule changes conformation, the values of the unit vectors will change, but certain relationships among them must be preserved. For all **dependent** unit vectors $\mathbf{u}_k$,

$$\mathbf{u}_k = \mathbf{u}_i \times \mathbf{u}_j \qquad (2)$$

where $\mathbf{u}_i$ and $\mathbf{u}_j$ are its two **defining** unit vectors. For instance in Figure 1, unit vector 3 depends on 1 and 2, 8 depends on 6 and 7, and 13 depends on 11 and 12. For all unit vectors there are normalization, orthogonality, and other geometric constraints of the form

$$l_{ij} \leq \mathbf{u}_i \cdot \mathbf{u}_j \leq m_{ij} \qquad (3)$$

where $l_{ij} = m_{ij} = 1$, whenever $i = j$ (normalization), $= 0$ for $i \neq j$ and both are in the same local system (orthogonality), equal to some other constant in some situations (for example $\mathbf{u}_1 \cdot \mathbf{u}_4$), and otherwise $-1 \leq l_{ij} \leq m_{ij} \leq 1$. For a small molecule, the limits on all unit vector inner products can be set by an exhaustive search of all combinations of torsion angles for the rotatable bonds, but for the larger molecule used as the test case in this study, we simply set $l_{ij} = -1$ and $m_{ij} = 1$. Letting the dependent unit vectors be determined by eq. (2), the embedding problem reduces to determining the components of the independent unit vectors subject to the constraint equalities and inequalities of eq. (3). Note that this automatically fixes correct bond lengths, bond angles, and chirality.

A complete solution to the embedding problem posed in the Introduction also needs to treat distance constraints. If atom positions $\mathbf{a}_k$ and $\mathbf{a}_l$ are given in terms of the unit vectors according to eq. (1), then the squared distance between them is simply

$$d_{kl}^2 = \sum_j (\alpha_{kj} - \alpha_{lj})^2 + \sum_i \sum_j (\alpha_{ki} - \alpha_{li}) \times (\alpha_{kj} - \alpha_{lj})\mathbf{u}_i \cdot \mathbf{u}_j \qquad (4)$$

which is a linear function of the inner products. This means that a necessary (but not entirely sufficient) condition that the experiment constraints on the (squared) distances be consistent with the holonomic constraints, is that the equalities and inequalities from eq. (3) added to the inequalities derived from eq. (4) have a feasible solution, regarding the unit vector inner products as the n_u^2 variables.

This is a standard task for linear programming, and can be readily solved by the simplex algorithm, "phase one." (Any standard textbook on linear programming[6] explains in detail what phase one of the simplex algorithm is and how the calculation can be set up and carried out.) The reason the feasibility check is not a sufficient condition is that the inner products are not really fully independent variables. They must certainly satisfy eq. (2), which essentially maintains correct chirality for each asymmetric center, but also for the independent unit vectors, the metric matrix

$$\mathbf{G} = (g_{ij}) = (\mathbf{u}_i \cdot \mathbf{u}_j) \tag{5}$$

must have rank 3 in order to have a conformation in three dimensions.[2]

We have not rigorously pursued the idea of checking the constraints by linear programming because the test is insufficient, although apparently very good. In EMBED step 3 (see Introduction) trial distances are either chosen completely at random within their allowed ranges, or after one trial distance is chosen, the allowed ranges of those distances still to be determined are reduced by deductions on the level of the triangle inequality. Here the analogous procedure would be to solve two linear programs, both involving all inner product inequalities and equalities as constraints and all variable but so far undetermined inner products as the variables. The first program would seek to maximize the one inner product whose value we wish to choose, and the other linear program would minimize it. Then choose a trial value of that inner product from the range thus determined. This restricts the feasible region of the subsequent linear programs, but cannot eliminate it. Then go on to select a value for the next variable inner product, and so on, until all have been chosen. The idea is very appealing, but it is not clear what the trade-off is in improved performance vs. increased computer time.

There is one last sort of constraint that must be introduced. Referring to the example in Figure 1, C1 is supposed to lie in the C4—O6—N7—H8—C9 plane, but it is inadequately constrained to do so. In this case $\mathbf{u}_1 \cdot \mathbf{u}_4$ has a fixed value because $\mathbf{u}_1$ happened to be defined to lie along the C1—C4 bond and $\mathbf{u}_4$ lies along the C4—N7 bond. That still leaves C1 free to lie anywhere on a circle described by spinning about the C4—N7 axis. Its correct position can be fixed by requiring the C1—C4—O6 angle to have the correct value. Actually we fix the $i - j - k$ angle by introducing a term

$$((\mathbf{a}_i - \mathbf{a}_j) \cdot (\mathbf{a}_k - \mathbf{a}_j) - c)^2 \tag{6}$$

into the error function, where c is the required value for the inner product expression as determined from any conformation of the molecule. Such angle constraints are added for every pair of covalently adjacent local coordinate systems.

Finally, the algorithm for linearized embedding may be summarized as follows:

1. Begin with the atomic coordinates of the molecule in any conformation having correct covalent geometry but arbitrary dihedral angles for rotatable bonds. The covalent bonding is also known, as well as which bonds are rotatable. One bond from each cycle is formally deleted, and in the case of flexible rings, the bond length and bond angle constraints for proper closure are noted.
2. Construct the linearized representation of the molecule by rerooting the connectivity tree at a tree center, and then recursively set up local coordinate systems in a depth-first tree traversal, initiating a new system after crossing a rotatable bond. Each system introduces one, two, or three new unit vectors, depending on the dimensionality of the corresponding rigid group of atoms. In the case of three vectors, note the dependency relationship given in eq. (2).
3. Determine the constraints on all inner products between pairs of unit vectors without doing an exhaustive conformational scan, as explained after eq. (3). Also set up angle constraints as in eq. (6) between adjacent coordinate systems.
4. Create a full set of constraints consisting of the dependency relations, the inner product bounds, angle constraints, van der Waals lower bounds on interatomic distances, and any experimentally determined upper and lower bounds on distances.
5. Estimate the maximal and minimal value each constrained squared distance could

attain by substituting into eq. (4) the maximal and minimal values of each inner product (depending on the sign of each term's coefficient). When applicable, raise lower distance bounds and lower upper bounds beyond the levels stipulated by the original van der Waals and experimental constraints.

6. Still using eq. (4), the current distance bounds may imply that some of the inner product lower bounds must be raised or upper bounds must be lowered. For example, solve eq. (4) for one of the inner products and then substitute in the extremal allowed values of the distance and the other inner products so as to maximize the one side of the equation. If this value is less than the current upper bound for the chosen inner product, the bound must be lowered.
7. Iterate steps 5 and 6 until a consistent set of bounds on distances and inner products has been reached. This is a useful check on the interrelation of experimental and holonomic constraints. Failure to reach consistency or producing a lower bound higher than its corresponding upper bound is an indication of some sort of inconsistency among the constraints.
8. Set up a trial metric matrix involving only the independent unit vectors as in eq. (5) by choosing each entry independently at random (with uniform distribution) between its upper and lower limits. Just as in step 4 of EMBED, find the nearest rank three approximation to the metric matrix.
9. Calculate from this rank three matrix the coordinates for the independent unit vectors, just as in step 5 of EMBED. From this it is straightforward to calculate the dependent unit vectors via eq. (2), and then the atomic coordinates by eq. (1).
10. This trial set of coordinates does not in general satisfy all the constraints, so one must locally minimize an error function, as in step 5 of EMBED. The error function has the customary distance bound terms from the van der Waals and experimental constraints, and in addition, inner product bound terms involving all unit vectors (whether independent or dependent) and angle constraints. The error function is viewed as a function of the x, y, and z coordinates of all independent unit vectors.
11. Additional random conformers may be generated by going back to step 8.

For the sake of clarity, we have omitted many computational details in the description of this algorithm. The computer programs, corresponding to steps 1–11, are written in the C++ language without very much attention to execution speed, and run on a Sun 3 computer under the Unix operating system. It is difficult to see how they might be translated into fortran. Those interested in obtaining the code should contact the author

RESULTS

When applied to very small test cases, such as ethane, the results are spectacular compared to EMBED's performance. The bound smoothing step corresponds roughly to tetrangle inequality reasoning, so that correct bounds are deduced for the distances between hydrogens on opposite methyl groups and for the equivalent inner product bounds. The minimization of residual errors in step 10 is just not required. The chirality of both methyl groups is always correct (thinking of each hydrogen as uniquely labelled). Of course, one hardly needs fancy embedding algorithms to generate ethane conformations having required interhydrogen distances.

Our main test of the method was on [D-Pen^2, D-Pen^5]-enkephalin (DPDPE), i.e., oxidized Tyr-D-Pen-Gly-Phe-D-Pen. Since penicillamine is just β,β-dimethylcysteine, this is a very sterically hindered cyclic pentapeptide. Experimental constraints derive from the NMR studies of Mosberg and coworkers.[7,8] The issue here is not whether these experimental constraints are scientifically correct or even what they are (see reference 6 for that), but rather how does the performance of the linearized embedding algorithm compare with that of EMBED.

Steps 1, 2, and 3 are computationally quick but intricate. We assumed that all peptide bonds should be fixed in the planar trans configuration. The pentapeptide is represented as 63 atoms, including α-hydrogens, amino hydrogens, and pseudo-atoms repre-

senting the centers of the two aromatic rings and the centers of mass for the β-methyl hydrogens in the penicillamine residues. There were 18 proton pairs having restricted distance ranges, as derived from NOEs. (Where the assignment was ambiguous, say as to which phenylalanine aromatic hydrogen, the corresponding pseudoatom was used.) Only 40 unit vectors were needed to fully describe the molecule, and of those 33 are independent. In step 4, for the purposes of comparison, both EMBED and the linearized approach were given exactly the same holonomic, experimental, and van der Waals constraints, except of course the two programs express these in different terms. EMBED had on the order of 2000 distance constraints plus a chirality constraint for each asymmetric carbon and several extra degenerate chirality constraints to enforce the planarity of aromatic rings and peptide groups. The new algorithm, however, expressed the holonomic constraints in the way atoms were defined in terms of unit vectors, and by means of angle constraints. There were no chirality constraints at all, of course.

Both programs eventually work out a full set of upper and lower bounds between all atom pairs from the experimental, van der Waals, and holonomic constraints. Neither reported any inconsistencies at this point, although no *perfectly* satisfactory conformer (i.e., zero value of the error function) has ever been found for this particular test case. Steps 5 and 6 of the new algorithm, however, reported eight redundancies in the experimental constraints, given the holonomic ones. For example, the conservatively interpreted NMR results led to requiring the distance between Pen[5] α-H and the pseudoatom on the proto-R β-methyl to lie between 2.02 and 3.90 Å. The holonomic constraints conclude that the range is really only 2.49 to 3.74 Å, thus declaring two experimental distance bounds to be redundant. Additionally, there were a number of van der Waals constraints between atoms separated by three bonds that were lower distance bound values smaller than those deduced from the covalent structure.

Steps 8, 9, and 10 of the linearized algorithm were repeated 20 times, while the equivalent steps 3, 4, and 5 of EMBED were performed 408 times. The success rate noted in Table I shows that the new algorithm handles this very taxing test case with significantly lower attrition. "Success" in both algorithms was taken to be that no distance constraint was violated by more than 0.5 Å, but one must realize that otherwise the two algorithms place their errors in different categories. EMBED might have a single bond length off by 0.5 Å, although this did not actually occur for successful conformations, whereas bond distortions in the linearized approach arise from errors in the inner product bounds, which were generally less than 20% in successful structures. Angle constraints were satisfied to better than 10°. EMBED might refine to structures having slightly nonplanar aromatic rings, but the new method by construction must have all these atoms coplanar because there are only two unit vectors defining their positions. On a problem like this one involving only 63 atoms, both methods spend the majority of their time in the optimization of their respective error functions, using conjugate gradients in both programs, as it so happens. The two error functions are of similar complexity, the chirality terms of EMBED being replaced by the nearly equivalent angle terms and distances involving dependent unit vectors. In its present implementation, the linearized computer program requires an average of 1.1 hours per attempted structure on a Sun 3/160 computer with a floating point accelerator. This is about three times as much CPU time as for EMBED, thus nearly negating its attrition advantage. No doubt this could be improved. It is intriguing (but not statistically significant) that the linearized embedded structures are much more likely to refine to

Table I. Comparison of results for solving the embedding problem in the case of DPDPE using the standard EMBED and the new linearized algorithms.

	Number of variables[a]	Success rate[b]	Redundancies noted
EMBED	189	1/14	no
Linearized embedding	99	1/5	yes

[a]The number of variables used in minimizing the error function.

[b]Average number of conformations generated having no distance violation greater than 0.5 Å, compared to the total number of conformations produced.

very low energy minima upon subsequent energy minimization, than those structures produced by EMBED.

One might ask why neither method ever finds a completely satisfactory conformation of DPDPE. The answer appears to be that there are slight conflicts between the experimental and holonomic constraints on the one hand and the van der Waals constraints on the other. Building a space filling CPK model of the peptide is nearly impossible because of steric crowding, and computer graphics using standard van der Waals radii shows only the smallest of holes in the middle of the tetrapeptide loop formed by the disulfide bridge between the two penicillamine residues. The linearized method with all constraints found 20 conformations having final refined error function values of 71.8 or greater, of which four were accepted as satisfactory. Eliminating the van der Waals constraints and running for another 20 tries produced values as low as 2.3, by way of comparison. Trivially, when the experimental constraints and disulfide bond closure constraints were also deleted, the structures generated either had zero error, or a local coordinate system at an α-carbon was stuck in the wrong orientation relative to the adjacent peptide group.

We have not carried out a systematic experimental sampling of the dependence of the linearized embedding algorithm execution time versus the size of the problem. The number of variables increases roughly linearly with the number of atoms for "ordinary" organic molecules. For generating small numbers of conformers, the bound smoothing step is the most time consuming part of the calculation, and the associated computational cost apparently scales with the cube of the number of atoms for typical sets of distance constraints. For small molecules and large numbers of generated structures, the bottleneck is the minimization of the error function, the cost of which goes up with the square of the number of atoms. Keep in mind, however, that the size of the problem can be measured in different ways, and the time required for these calculations can depend sensitively on the interactions between constraints. For example, when there are few constraints, additional constraints typically slow down EMBED, but eventually when there are very many constraints, adding yet more of them tends to speed the calculation! Since linearized embedding deals with roughly half the number of variables that EMBED does, one might hope for an optimized program running four- to eight-times faster, although we certainly have not yet achieved that.

So far, we have not explored the ability of the linearized approach to search out the allowed conformation space of a molecule, and whether its sampling of the space differs from that of EMBED or other methods. Work is under way to carefully define and examine this important question. From our limited experience so far, we can only say the sampling appears to be about as broad, inasmuch as substantially different backbone and sidechain conformations were seen in the four successful structures.

As it stands, we would recommend the linearized embedding algorithm for problems where the covalent geometry is assumed to be rigid and where there are relatively large rigid groups of atoms, such as aromatic ring systems. Collinearity and coplanarity constraints are elegantly enforced, and one can often gain insight into the incompatibilities between experimental and holonomic constraints.

This work was supported by grants from the National Institutes of Health (GM37123) and the National Science Foundation (DMB-8705006).

References

1. G.M. Crippen, "Distance Geometry and Conformational Calculations," in D. Bawden, Ed., *Chemometrics Research Studies Series,* Volume 1, Research Studies Press (Wiley), New York, 1981.
2. G.M. Crippen and T.F. Havel, "Distance Geometry and Molecular Conformation," in D. Bawden, Ed., *Chemometrics Research Studies Series.* Research Studies Press (Wiley), New York, 1988.
3. T.F. Havel, *QCPE Bulletin,* **6**(1), 1986.
4. P. Easthope and T. Havel, *Bull. Math. Biol.,* **51,** 1989.
5. G.M. Crippen, *J. Comp. Chem.,* **8,** 943, 1987.
6. K.G. Murty, *Linear and Combinatorial Programming,* John Wiley and Sons, New York, 1976.
7. H. Mosberg, P. Subramanian, K. Sobczyk, G. Crippen, K. Ramalingam, and R. Woodard, "Combined Use of Stereospecific Deuteration, NMR, and Distance Geometry for Conformational Analysis of [D-Pen^2, D-Pen^5]Enkephalin," in *10th American Peptide Symposium,* St. Louis, May 1987.
8. H. Mosberg, P. Subramanian, K. Sobczyk, G. Crippen, K. Ramalingam, and R. Woodard, (in press).