Journal of Mathematical Chemistry 6(1991)307–324

Review

CHEMICAL DISTANCE GEOMETRY: CURRENT REALIZATION AND FUTURE PROJECTION

Gordon M. CRIPPEN
College of Pharmacy, University of Michigan, Ann Arbor, MI 48109, USA

Received 26 November 1990

Abstract

Since the 1988 monograph "Distance Geometry and Molecular Conformation" by Crippen and Havel, there have been significant changes in the application of distance geometry to problems of chemical interest. This review attempts to outline what the current state of the art is, in both the underlying mathematical methods and chemical applications, and to indicate future developments. Rather than go into details concerning algorithms and theorems, the emphasis is on defining the kinds of problems we can solve or would like to, and then guiding the interested reader to the recent literature. Special emphasis is given to the problem of determining macromolecular conformation in solution by NMR, including energy functions, and dealing with conformational flexibility.

1. Definitions

1.1. DISTANCE GEOMETRY

As far as pure mathematics is concerned, topology is the study of bijective and bicontinuous mappings; then *distance geometry* or *metric topology* is the subset of topology where the space involved has a "distance" defined and the mappings are also distance preserving. This has been studied by a long series of mathematicians, dating back at least to the mid-nineteenth century work of Cayley, and continuing up to the present [1]. In more homely terms, classical geometry concentrates on points, lines, and angles, while distance geometry focuses on distances between points. The connection with chemistry is that if the points represent atoms, then the interatomic distances are often directly related to experimental results and important contributions to the energy of the system. Traditionally, computational chemistry has described molecular geometry in terms of either three-dimensional Cartesian coordinates or internal coordinates (bond lengths, vicinal bond angles, and dihedral angles), whichever parameterization was most convenient to the problem at hand. In distance geometry, we instead take the interatomic distances as the fundamental coordinates of molecules, exploit their

close relationships to the experimental evidence and internal energies, and then worry about converting to atomic Cartesian coordinates only later.

1.2. FUNDAMENTAL PROBLEM OF DISTANCE GEOMETRY

The challenge in this business is that last little detail: how do we convert distances to Cartesian coordinates? A rigorous presentation of all the factors involved takes up a substantial part of our recent book [2]*, so here I will simply enumerate the important cases that arise in practical problems. The term *embedding* here refers to mapping some set of constraints expressed primarily in terms of distances among a set of points or atoms into Cartesian coordinates for all points.

(i) There may be no embedding possible. For example, if we denote the distance between points i and j by d_{ij}, then the triangle inequality

$$d_{ij} \leq d_{ik} + d_{kj} \tag{1}$$

must be obeyed for all triples of points. If the constraints imply a violation of the triangle inequality for some triplet, then embedding fails for any dimension of target space $\mathbb{R}^n$. In general, there are several kinds of conditions that the distances among all small subsets of points must obey. Ideally, we would like to be able to trace back the failure to some small set of mutually incompatible constraints, but except for violations of the triangle inequality and the tetrangle inequality (a four-point relation), this remains an unsolved problem for two reasons. First, the failure to embed may be marginal, and the numerical methods used tend to spread the error over many subsets of points, as in least-squares fitting. Second, there may be many different constraints or subsets of constraints, each with their respective subsets of points, such that removing any one of the subsets of constraints will permit embedding.

(ii) There may be no solution in $\mathbb{R}^3$, but there is one in higher dimensions. Or more rarely, it may happen that one seeks an embedding in $\mathbb{R}^3$, but the result instead lies in a lower-dimensional subspace. Once again, there are multiple causes, and tracing back is generally difficult.

(iii) There may be a unique solution, meaning of course that the resulting coordinates are unique up to a rigid rotation and translation. This almost never happens, and in particular, for more than three points being embedded in $\mathbb{R}^3$, distance constraints alone are insufficient. In addition, there must be chirality constraints, which relate the relative handedness of quartets of points in $\mathbb{R}^3$ and can be generalized to any number of dimensions. As with distance constraints, there are certain relationships among specified chiralities and between chiralities and distance constraints that must be obeyed in order to embed. In chemical applications, this is usually not a

*Affectionately referred to by some as "The Green Death".

major problem because the absolute orientation of the relatively few asymmetric centers (and the signs of dihedral angles) are generally known and mutually compatible.

(iv) There may be a range of solutions. Experimental data are never exact, but more importantly, they are generally never plentiful enough to determine a unique molecular conformation. Suppose the constraints are enough to confine the molecule to some finite, connected region of its conformation space. It is a topic of current debate as to how best to explore this region in terms of sets of allowed Cartesian coordinates for the atoms. Should the sampling be uniform? What is uniform? Should the sampling concentrate on the extremes? See the discussion in the NMR section. There is no known convenient but accurate representation of the (probably nonconvex, high genus) allowed conformation space that affords easy production of atomic coordinates.

(v) There are disjoint solution sets. This disturbing situation arises even in systems as simple as cyclohexane, typically with relatively tight sets of constraints. It implies that even if several different successful embeddings are known, they may not necessarily be at all similar to yet undiscovered embeddings, nor can the known structures be smoothly perturbed to reach them.

Having seen the major possible outcomes of attempting to embed, we can work backwards to a relatively general statement of what has been referred to as the *fundamental problem of distance geometry*. Suppose we have some upper and/or lower bounds on some of the interatomic distances

$$0 \leq l_{ij} \leq d_{ij} \leq u_{ij} \leq \infty, \tag{2}$$

where the bounds may be tight ($l_{ij} \lesssim u_{ij}$), loose ($l_{ij} \ll u_{ij}$), or nonexistent ($l_{ij} = 0$ and $u_{ij} = \infty$). In addition, suppose there are some restrictions on the values of chirality χ that may be attained for some of the ordered quartets of atoms $[a_i, a_j, a_k, a_l]$,

$$\chi_{ijkl} \in S_{ijkl} \subset \{-1, +1, 0\}, \tag{3}$$

where the chirality values of -1, $+1$, and 0 correspond to the chemical ideas of S, R, and coplanar. Then the fundamental problem is to decide whether embedding is possible, given these constraints. If not, one would like to know which subsets of the constraints are mutually incompatible. A similar analysis would be desirable if embedding necessarily led to conformations in dimensions higher or lower than 3. If embedding in $\mathbb{R}^3$ is possible, we want to characterize the allowed conformations as completely as possible in scientifically useful ways. In particular, it would be useful to have a representative of every major type of conformation, information about the extremes of the allowed conformation space, and a uniform sampling of all possibilities.

1.3. VARIATIONS ON THE PROBLEM

The fundamental problem is very hard, both theoretically and in practice. Clearly, one approach is to redefine or restrict the problem so that it is easier to treat. Taken to extremes, one runs the risk of solving a problem of no practical interest, but the optimists among us can at least learn from such studies something about the structure of the fundamental problem and attempt to generalize the methods that worked for restricted problems.

Instead of trying to embed $\mathbb{R}^3$, the fundamental problem becomes greatly simplified in one dimension. The idea of chirality reduces to a question of ordering along the real number line for pairs of atoms. The algorithm by Dress and Havel [3] begins with any set of distance equalities and inequalities and relative ordering information. From this, it enumerates all possible orderings of the atoms along the line, if any, such that an embedding is possible. For each of these orderings, it is easy to produce a satisfactory set of atomic coordinates. The algorithm is certainly feasible for as many as ten atoms, but the computer time required goes up faster than a polynomial function of the problem size, as would be expected from results on just distance equality constraints [4]. In higher dimensions, the generalization of ordering along the line is called a *chirotope* [2,5], and one can enumerate them, but they become much more numerous in two and three dimensions.

Experimental distance information is never exact, and the usual treatment is to fit these constraints in a least-squares sense. In the original development of the EMBED algorithm, I had rejected this approach on the grounds that experiments usually give information concerning only a few of the total number of atom pairs, and even for these few, one usually knows only widely separated upper and lower distance bounds, rather than an estimated best distance value and standard deviation. Nonetheless, we shall see there are computational situations where we have a full matrix of preferred distances, and a least-squares approach is useful. Suppose we are given a set of *trial distances* t_{ij} for all pairs of atoms, and we seek a set of coordinates $\boldsymbol{c}_i \in \mathbb{R}^m$ for $i = 1, \ldots, n$ such that the calculated distances $d_{ij} = \| \boldsymbol{c}_i - \boldsymbol{c}_j \|$ agree with the trial distances in a least-squares sense. Another way to say this is that the objective is to minimize the square of the Frobenius norm of the difference between the trial and calculated distance matrices. Now, Glunt et al. [6] have devised an algorithm, called "modified alternating projections" or MAP, that solves this problem. For n points, the dimensionality of the resulting coordinates can be as high as $m = n - 1$, but even for $n \gg 3$, it often turns out that MAP produces coordinates such that $n - 1 > m > 3$. Furthermore, the answer is unique. Even though the coordinates are generally not in $\mathbb{R}^3$, we shall see in the section on energy optimization that this technique is very useful. The same group have subsequently shown that proceeding from the MAP result in $\mathbb{R}^m$ down to $\mathbb{R}^3$ destroys the uniqueness of the solution, but the multiple solutions can be characterized as local minima of a function, and these local minima all lie on a sphere [7]. This is a remarkable result that promises interesting future developments.

Suppose the only constraints are exact distances between some pairs of points. There are two reasons this problem is so attractive: engineers are very concerned about the rigidity of a structure built up of bars having fixed lengths, and the rigidity of such a "bar and joint framework" is almost always independent of the values given to the distance constraints. This means we can draw on a large body of work aimed toward deciding whether a given set of distance equality constraints determines a unique molecular conformation, and the tests are relatively rapid searches for certain graph properties [8,2]. Here, the graph has nodes identified with the atoms and an edge between two nodes whenever the corresponding atoms have a given fixed distance. Probably one of most significant theoretical results is that finding embeddings for such graphs is strongly NP-hard [4], and hence the more general fundamental problem is at least this bad. Characterization of the motions of a flexible bar and joint framework is less well developed, but Hendrickson [9] has shown that flexings almost always move a joint (=atom) in a closed loop diffeomorphic to a circle. This is certainly familiar in chemistry in that molecular conformations change by spinning around single bonds, or cyclic molecules undergo pseudorotation. Hendrickson also exhibits a whole class of frameworks that are locally rigid but have multiple ways they can be constructed in space. These are examples of the fact that local rigidity does not imply that the embedding is globally unique. This corresponds to the unhappy case of disjoint regions of conformation space being the solutions to even such a restricted version of the distance geometry problem. As a matter of practical application, one can construct algorithms for embedding using distance equality constraints (see below), but almost the only experimental source for such constraints are the relatively stiff bond lengths and vicinal bond angles. The additional equalities required to give the molecule any sort of distinctive conformation must come from choosing values for other atom pairs from within experimentally determined ranges, or from choosing values that would lead to low-energy conformations. The other problem is that most of the useful results in this area have the caveat "almost all", meaning that there are special values of distance constraints where the theorems may fail. Molecules viewed at atomic resolution have many special symmetries, triples of collinear atoms, quadruples of coplanar atoms, etc. that would have zero probability of occurring if interatomic distances were chosen at random. If the points are instead whole groups of atoms and the constraints are applied to macromolecules built out of such groups, this sort of trouble is unlikely.

2. Methods

2.1. NAIVE APPROACHES

Brutality works. It is now feasible to attack the distance geometry problem by means that were unthinkable only a few years ago, due to advances in computer power and availability. For example, one can adjust dihedral angles by interactive

computer graphics until the geometric constraints are met or the investigator loses patience, whichever comes first. This is not a recommendable approach for a thorough search over several degrees of freedom, but it is routinely used as a way to escape from a set of unsatisfied constraints by visually inspecting the unsatisfactory conformation and making a large, concerted manipulation by hand, so that residual violations can be subsequently removed automatically by smooth, continuous perturbations. The automatic methods amount to defining a penalty function and minimizing it with some standard general program for unconstrained minimization. The penalty function consists typically of a sum of terms, one for each constraint, each term constructed to be equal to zero only if its constraint is satisfied and having monotonically increasing value as the constraint is violated by a greater and greater margin. If the only constraints are upper bounds on distances, one can construct a penalty function with only a single minimum value region, and a minimization algorithm can always find the solution. The minute distance equalities or lower bounds are introduced, the penalty function has a nonconvex feasible region and multiple minima, the majority of which have penalty function values greater than zero. The question becomes how to leave these infeasible local minima without always resorting to human intervention.

If the problem is small enough, picking random starting points for the minimization may suffice. If not, there are innumerable, more-or-less global search algorithms to try. Of these, the most popular one among the NMR community is constrained molecular dynamics. Starting at some initial conformation that has some relatively minor constraint violations, one carries out a molecular dynamics simulation on the molecule where the potential is the weighted sum of the usual empirical energy function and the penalty function. Thermal motions send the molecule over small potential barriers, and the molecule tends toward a conformation that better satisfies the geometric constraints and has a relatively good internal energy. The drawback is that if the starting point is not very good, molecular dynamics is a hopelessly expensive way to seek solutions. The search is better than local minimization, but still not very broad. If at the end of a search, there are still substantial constraint violations, it may be ambiguous whether the constraints themselves are mutually contradictory, or whether they correspond to conformations with very high energy.

Greater potential barriers can be surmounted and wider regions explored by running molecular dynamics at high temperature. Then, in order to converge on conformations having low potential values, i.e. that satisfy the geometric constraints, it is necessary to gradually lower the temperature. This *simulated annealing* procedure is guaranteed to converge on the global minimum with probability one if the cooling is carried out "slowly enough" [10]. Reasonable cooling schedules can be found empirically such that starting coordinates approximately satisfying the constraints can be refined to much better solutions [11,12]. In fact, it is possible to reach a solution from arbitrary random starting coordinates [13], but it is much more efficient to begin with coordinates from the EMBED algorithm, and most determinations of

protein conformation from NMR data that employ simulated annealing are done that way.

All of these penalty function methods suffer from concentrating on exhibiting an embedding if one can be found, but in case of failure, the outcome is not clear. Perhaps a longer search would find a solution. Perhaps the remaining small violations can be completely erased by extending the search, or perhaps they represent a genuine incompatibility. Even in the case of large residual violations, it is generally difficult to identify which subset of the constraints is mutually incompatible, because the violations tend to spread among all the terms.

2.2. EXACT DISTANCE CONSTRAINTS

Given a sparse set of exact distances between pairs of atoms, the computer program ABBIE by Hendrickson [9] finds an embedding in $\mathbb{R}^3$. It first analyzes the corresponding constraints graph to detect rigid subgraphs which are either small enough to embed immediately by minimizing the penalty function, or must be broken into smaller rigid subgraphs. Then, the embedded subgraphs or their mirror images are rejoined as rigid groups until finally the whole molecule is built up. The procedure seems to depend on there being enough fixed distances to uniquely determine the conformation, although of course most distances could be unspecified. Chirality constraints are not employed except for realizing that either an embedded subgraph or its mirror image must be rejoined to other parts of the molecule in order to satisfy all distance constraints. Unfortunately, most chemical applications do not have this type of constraint set, but at least the approach can in principle detect subsets of incompatible constraints, due to its divide-and-conquer strategy.

Hadwiger and Fox [14] have been developing a related means of building up a set of coordinates for a molecule, starting with small subsets of atoms having many distance equality constraints among them. Although they have not yet presented an algorithm, the idea seems to be a matter of piecing together small rigid groups or their mirror images, possibly branching on the two alternatives at each step, so as to broadly search the set of allowed conformations. Sometimes there will be given chiral constraints that dictate which of the two joinings are allowed, but usually both alternatives will be permissible. If there are insufficient constraints to consolidate two pieces, sufficient intergroup distance values could be chosen at random. In the process, there would be some provision to detect constraint incompatibilities, such as triangle inequality violations. The attractive feature is the relatively comprehensive search of conformation space associated with exploring the tree of alternative joinings of groups, but it remains to be seen how well the approach compares to EMBED. In particular, many branches of the search tree could be dead ends when there are severe steric constraints, i.e. important lower bounds on interatomic distances.

2.3. EMBED ALGORITHM

The standard algorithm for solving the full distance geometry problem is called EMBED, and it has been described in abundant detail [2]. For the purposes of this discussion, it suffices to note its distinguishing features. At the outset, the emphasis is on the distance inequality bounds, and much effort goes into using the sparse set of constraints to raise the lower bounds and lower the upper bounds for interatomic distances that have not been specified at all. Subsequently, sets of *trial distances* are selected by some random process from the respective allowed ranges. From here on, all distances are treated on an equal footing, regardless of whether they stem from some equality constraint or were selected from a large range. The advantage is that there is no favoritism of one part of the molecule over another, and errors do not propagate as they might in a procedure that built up the molecules from some small subset of atoms. Each set of trial distances is converted into a *trial metric matrix*, where the ijth element is the scalar product of the vectors from the center of mass to the ith and jth atoms. *Trial coordinates* can be easily calculated from the three largest eigenvalues and corresponding eigenvectors of the metric matrix. As we have already discussed, the problem addressed by Hayden and coworkers seeks the trial coordinates whose calculated distance matrix best agrees with the trial distances in the Frobenius norm, whereas here the trial coordinates correspond to a calculated metric matrix that best agrees with the trial metric matrix in the spectral sense. The trial coordinates may agree with the trial distances in some broad sense that makes subsequent refinement easy, but generally, to the eye they bear little resemblance to the desired molecule. Finally, the trial coordinates are used as the starting point in a minimization of a penalty function based on the given constraints. Only the penalty function includes any information about desired chiralities. The resulting *refined coordinates* either agree with the input constraints very closely or the structure is rejected. For difficult sets of constraints, the attrition rate can be appreciable, depending on the refinement techniques used. Conformation space is explored by producing a series of sets of refined coordinates derived from different random trial distances, the justification being that the interaction of the constraints can be extremely complicated, so that some sort of directed exploration or summary of the allowed conformations is not yet practical. (The complete analysis of the conformation space of cyclohexane seems to be the most complicated example to date.)

EMBED has been programmed several times by different people, incorporating different features that affect its convenience of use and efficiency, as explained in a recent review [15]. It is most readily available from QCPE (Quantum Chemistry Program Exchange) as Havel's DISGEO program (written in PASCAL, tailored for NMR studies of protein conformation) and Blaney's DGEOM (written in FORTRAN for more general molecular distance geometry problems). Commercial sources include Smellie's CONSTRICTOR (Oxford Molecular, Ltd.) and Hare's DSPACE (Hare Research).

The most advanced implementation of the general EMBED scheme is the DG-II program, written in C by Tim Havel. In addition to the usual bound smoothing with

the triangle inequality, it offers optional tetrangle bound smoothing. For typical NMR data on proteins, this is not important, but for small molecules having some large lower bounds on distances, the contraction of allowed distance ranges can be significant [16]. DG-II employs the best known method of selecting random trial distances: the atoms are randomly ordered, and then the trial distances are chosen from the (permuted) matrix of ranges using *metrization*, an algorithm that contracts the ranges of yet unspecified distances such that the resulting sets of trial distances are consistent with the triangle inequality [2]. As usual, most of the computational effort goes into refining the trial coordinates for each random structure generated. In order to facilitate the process, DG-II first starts with coordinates $\mathbb{R}^4$, derived as usual from the trial distances, and then reduces the constraint violations by simulated annealing, while compressing the structure into $\mathbb{R}^3$. At this point, most of the residual violations are minor deformations of local structure, such as non-planar benzene rings, and further improvement by simulated annealing or local minimization with respect to Cartesian coordinates is slow. Consequently, the next step for problems with exact local distance constraints is to fit the current atomic coordinates to a "regularized" conformation having exactly the desired bond lengths, bond angles, planarities, etc. The fitting procedure also takes into account any dihedral angle constraints. From here, the refinement proceeds by minimizing the penalty function with respect to dihedral angles, starting at the regularized conformation.

Clearly, DG-II has several adjustable parameters that affect performance, such as a simulated annealing schedule, convergence criteria at various steps, etc. Since the convergence versus failure to converge for a given trial structure to a satisfactory refined structure can depend on variations in parameters at the level of machine accuracy, the performance of the algorithm can only be described on an average case basis. Let the test molecule be pancreatic trypsin inhibitor (BPTI), a small, stable protein having 58 residues, an accurately determined crystal structure, and a structure in aqueous solution studied by NMR. The test set of constraints consists of standard bond lengths, bond angles, Van der Waals radii, and residue chiralities, plus 500 distance constraints of the type one would derive from NMR NOE measurements, but in this case taken from the crystal structure atomic coordinates. That way, there is no question about geometric infeasibility of the constraint set. Even though the molecule contains around 700 atoms, producing each refined structure took only about four hours on a Sun Sparcstation 1, most of that time being spent in the simulated annealing stage. Only about 20% of the trial coordinate sets failed to converge, due to some incorrect chain crossing that would require a large, concerted conformational change to correct. The other 80% reached a maximum distance violation of 0.3 Å, about five times better than earlier computer programs. The breadth of sampling in this relatively constrained example also increased in that there was a 1.5 Å rms deviation among backbone conformations, compared to only 1 Å previously.

Selecting trial distances has been done in the past by a variety of random and not-so-random processes, leading to concern that EMBED's sampling of the allowed conformations is biased [17,18]. The most difficult part of this question is deciding

what unbiased sampling should look like. For rigid valence geometry, the assumption underlying standard polymer statistics is that uniform sampling means uniform sampling of dihedral angles. For poly-L-alanine subject to only local geometric constraints (Flory θ conditions), one can calculate the mean-square end-to-end distance directly from standard polymer theory. Given the same constraints, DG-II produces a random sampling of conformers having the same mean-square end-to-end distance, as well as a fairly uniform dihedral angle distribution [19].

2.4. ENERGY OPTIMIZATION

If EMBED or any other algorithm is able to generate a (large) sample of conformers consistent with a given list of geometric constraints, most chemists would be more interested in those having the lowest energies, as calculated by some standard empirical molecular mechanics potential. Clearly, this is the overwhelming trend in the determination of macromolecular conformation by NMR, where almost invariably, structures produced by EMBED are subjected to a round of molecular dynamics. The task, therefore, amounts to optimizing a nonlinear function of the atomic coordinates subject to a collection of nonlinear constraints. Before diving into applications, it is worthwhile to stand back and enumerate the three possible scenarios.

(1) The constraints may have no feasible region, so the issue of energy minimization does not even arise.

(2) There are one or more disjoint, probably nonconvex, feasible regions of various dimensionality as far as the constraints are concerned. Within the interior of one of these regions, there may be one or more local energy minima. I will refer to these as *unconstrained minima.*

(3) A local energy minimum may lie outside a feasible region such that there arise one or more points along the boundary of the region that are *constrained minima.* The Kuhn–Tucker optimality conditions apply here [2]; in particular, the energy gradient is not zero, and there are nonzero Lagrange multipliers associated with at least some of the constraints, each corresponding to the force that the constraint exerts on the solution to keep it from leaving the feasible region and proceeding toward the local minimum outside the region.

A survey of recent papers on the determination of solution conformation of macromolecules by NMR clearly shows that the preferred method is to use EMBED to find conformers in agreement with the geometric constraints derived from experiment, and then to improve their energy by molecular dynamics. If the starting structure is near enough to a deep unconstrained minimum (case 2, above), molecular dynamics will be attracted to it, possibly cross small intervening energy barriers to reach it, and the geometric constraints will still be obeyed. Energy minimization is simple in that an algorithm will almost always eventually converge to some nearby local minimum and stay there, but molecular dynamics is always somewhat uncertain, because a

longer simulation may produce a rare but important conformational change in the unpredictable future. Typical experience seems to be that the molecular dynamics trajectory moves out of the feasible region relatively soon and is unlikely to return. The standard remedy is to add the penalty function to the potential energy function, balanced by some weighting factor. For a large weight on the penalty terms, this produces an approximation to the constrained minimum (case 3, above) while still skipping over minor local minima. Resorting to such "constrained molecular dynamics" is a disturbing admission that the empirical energy function and/or the formulation of the simulation (e.g. temperature, pressure, and solvation) do not agree with experiment. I am not aware of any study resolving this discrepancy that so many researchers gloss over! Perhaps the NMR experiments are being overinterpreted (see NMR section); perhaps the energy functions need to be adjusted; perhaps the treatment of the macromolecules's environment, such as solvation, is so inadequate that an otherwise correct set of energy parameters leads the simulation astray; or perhaps the trajectory would eventually lead back to a feasible region, but the simulation was halted prematurely.

Although the initial stages of the EMBED algorithm deal strictly with constraints on interatomic distances, the refinement stage can be adapted to treat many different problems, among them being the geometrically constrained optimization of an energy function. Augmented Lagrangian functions [20] provide a general, numerically stable method for locating constrained local minima of any objective function, such as an empirical energy function. The results depend, of course, on the starting point in such situations, but one simply uses EMBED to produce a series of different trial coordinate sets as a random sampling of starting points. Using an augmented Lagrangian produces a more accurate constrained minimum than minimizing (or running molecular dynamics with) the weighted sum of the objective and penalty functions. Furthermore, it produces the Lagrange multipliers for the various constraints at the solution, thus revealing which constraints oppose the energy function in case 3, or which constraints are in conflict with each other in case 1. So far, this approach has not been tried with a standard energy function as the objective, or with a typical set of constraints for a small protein.

Instead of adding on energetic considerations at only the last stage of EMBED, *energy embedding* attempts to include information about the energy function at earlier stages. The basic approach [21–23] as reviewed in ref. [2] is that one begins in a high-dimensional space, where there are actually fewer local energy minima, relatively easily locates an energy minima, and then proceeds toward $\mathbb{R}^3$ in such a way as to raise the energy as little as possible. Recent studies have shed new light on the start of the procedure in the high-dimensional space. If there are strong intrinsic torsional terms in the energy function, such as for rotating about a peptide bond, then there will be local minima corresponding to the different stable states of these torsions even in high dimensions [24]. Even if all the terms are pairwise interatomic interactions with a unique energetically optimal separation implied for each pair of atoms, this set of trial distances is not in general embeddable in any dimension. For n atoms, one can either start at an arbitrary conformation in $\mathbb{R}^{n-1}$ and minimize the energy,

or one can use the MAP algorithm of Hayden's group [6] to find an embeddable least-squares approximation to the trial distances. Either way, the final conformation lies in an m-dimensional subspace, where $m \approx n/3$, and m is the same by either procedure [25]. Then there are various ways to proceed toward a three-dimensional structure, such as driving the trial distance matrix toward embeddability [26,27] or smoothly following the path from the high-dimensional starting point toward a three-dimensional conformation as defined by a set of differential equations that keep the energy minimal while reducing a continuous dimensionality parameter [25]. However, we find that it is much more efficient and effective to successively reduce the dimension one unit by setting the high-dimensional analogues of dihedral angles to different combinations of cis and trans values, keeping the energetically more favorable alternatives [25]. This *rotational energy embedding* procedure has located remarkably low-energy minima in test cases of about 40 interacting particles much more quickly than other global search methods can, but it does not necessarily find the global minimum.

3. Determination of conformation by NMR

3.1. CURRENT METHODS

The EMBED algorithm can be used to solve a wide variety of geometric problems arising in drug design [28,29] and the general determination of molecular conformation from a variety of experimental and theoretical information [30]. However, it is most frequently employed in the determination of the conformation of small proteins and oligonucleic acids in solution. In fact, it occupies a position analogous to that of direct Fourier methods in X-ray crystallography. The efficiency of EMBED, or any other algorithm, depends in part on the special kinds and quantity of input available from present-day NMR. In the case of a typical small protein, the vast majority of constraints are lower bounds on the distance between almost all pairs of atoms, due to their Van der Waals radii. This information is nonspecific, in that it does not restrict the allowed set of conformations to any particular region, but since globular proteins are quite closely packed, a large fraction of the total conformation space is eliminated. Next come the *holonomic* constraints that are known *a priori* from the crystal structures of related small molecules: standard bond lengths, vicinal bond angles, chirality of asymmetric centers, planarity of conjugated rings, etc. These constraints at least restrict the molecule to have at most some finite diameter, but otherwise the allowed region still consists of a complicated subspace of the space of all atomic Cartesian coordinates. This subspace corresponds to different choices of the dihedral angles, and there is still a large portion of the Cartesian coordinate space that is pervaded by it, in that for every set of coordinates corresponding to only mild violation of the holonomic constraints, there is a point in the subspace nearby. In other words, a protein with unspecified dihedral angles is still pretty floppy. The next most plentiful category of constraints at last come from the NMR experiments, either

as some restrictions on a few of the dihedral angles or mostly as short upper bounds on distances between hydrogen atoms which are separated by relatively few covalent bonds. Usually, these constraints are denoted as *short-range NOEs*. Without going into the physics of NMR spectroscopy (see refs. [31–33]), suffice it to say that one can detect some, but not all, instances where two hydrogen atoms are within about 5 Å of each other by observing a "nuclear Overhauser effect" (NOE). NOEs can be observed whether they are short range (the atoms are linked by a chain of only a few bonds) or long range (distant in amino acid sequence but not through space), but most NOEs are short range. For various technical reasons, NOEs are difficult to quantitate, so the most precise interpretations that are justifiable tend to break the NOEs into two or three classes, such as $d_{ij} < 3$ Å versus $3 < d_{ij} < 5$ Å. In particular, failure to observe an NOE between a given pair of atoms does not imply $d_{ij} > 5$ Å. Even with these seemingly loose kinds of constraints, the relatively plentiful short-range NOEs, combined with the holonomic constraints, are often enough to strongly restrict the path of the polypeptide chain over a several-residue segment. Analogous to the idea of persistence length in polymer theory, however, the chain direction at the end of longer segments tends to lose any correlation with the chain direction at the beginning, if you think about generating an ensemble of segments given only these constraints. In other words, moderately precise determination of the local chain geometry does not greatly restrict the overall protein folding possibilities. Finally, there are generally enough long-range NOEs to force the large-scale folding pattern, even if the polypeptide is treated locally as a freely jointed chain. In combination with all the foregoing classes of constraints, a small protein typically has its backbone conformation determined to within an average of 1 or 2 Å, where the tightly packed secondary structural elements in the core are more tightly constrained, and exterior loops and chain termini either actually have more freedom to move, or at least the experimental evidence does not greatly restrict their allowed positions. Typically, one calculates a large sampling of conformations using EMBED, and all successful refined coordinate sets cluster around each other to this sort of precision. The nagging worry is that there may be another distant, disconnected region of conformation space that also satisfies the constraints, but the random sampling happened to miss it. For instance, it is possible to satisfy a large set of simulated NOEs, disulfide bridges, hydrogen bonds, and holonomic constraints for BPTI by giving the polypeptide chain an overall conformation that is the mirror image of the crystal structure, and then making small adjustments to the left-handed α-helices, etc. to compensate for the still correct L-amino acids [34]!

As I have already explained, there are a number of different ways to solve the distance geometry problem for such sets of constraints. The EMBED algorithm tends to have a high rate of success, even though it does not really take into account the preponderance of local constraints over the relatively few long-range constraints, but rather treats all interatomic distances equally. The main alternatives are Braun and Go's DISMAN algorithm [35], which does exploit the heavy local constraints, and simulated annealing and constrained molecular dynamics. There is a tendency in the

field to use "distance geometry" as a buzz-word to denote any computer program designed to find conformations subject to constraints on distances. If we use that term for these alternative approaches, even though they never deal with the matrix of interatomic distances as the primary variables for describing conformation, then we will have to call all molecular energy minimization and molecular dynamics programs "distance geometry", too.

In the early stages of EMBED, it is essential to interpret the NMR experiments in terms of bounds on interatomic distances. However, we have already seen several examples of how the final refinement stage is much more flexible. Several laboratories are implementing *back-calculation* refinement procedures, where the trial atomic coordinates in $\mathbb{R}^3$ are varied so that the NMR spectra calculated from them agree optimally with the original spectra [36,37]. This is clearly a sensible thing to do, although the complexity of the computer programs and the cost of running them are increased. An essential part of this line of inquiry will be to detect and avoid local optima in the refinement where the calculated conformation fails to agree with the NMR within experimental accuracy.

So far, I have discussed only NMR studies on macromolecules in solution. In solid-state NMR, the sample is not necessarily crystalline, but there is at least uniaxial alignment of the molecules with respect to the external magnetic field. Then one can experimentally determine the angle between various covalent bond vectors and the magnetic field vector, although there may be as much as a fourfold ambiguity in the value. Brenneman and Cross [38] have shown how to combine this information with holonomic constraints to determine possible values of dihedral angles in polypeptides. So far, they have only shown how this works on simulated data sets, but there is no reason to suspect it will have trouble with real experimental data. Internally, the method seems rather complicated, largely because the ambiguity in the experimental angle values forces one into a combinatorial search. Issues of insufficient, inaccurate, and contradictory data still need to be explored.

Central to Brenneman and Cross' approach is the metric matrix, where the entries are the scalar products between pairs of vectors attached to various parts of the molecule. Embeddability in $\mathbb{R}^3$ requires that the metric matrix as a whole, or any submatrix of it, have rank no more than 3, which is precisely how EMBED converts the trial distances to trial coordinates. This is very similar to *linearized embedding* [39,40], where the metric matrix, rather than the distance matrix, plays a central role. In a molecule with fixed local geometry (fixed bond lengths, bond angles, planar rings, asymmetric centers, but free dihedral angles about rotatable bonds), one can view the molecule as a collection of mutually rigid groups of atoms linked together by the rotatable bonds. Instead of describing its conformation in terms of torsion angles, the linearized representation sets up a local coordinate system within each rigid group, so that the position of any atom is given by a linear polynomial with fixed scalar coefficients independent of conformation, where the variables are the unit vectors that are the axes of all the local coordinate systems. A particular conformation is specified by choosing certain relative orientations for all the unit vectors, subject

to the embeddability condition that the rank of the metric matrix formed by these unit vectors be 3 (or less). Producing the linearized representation for a molecule is a moderately complicated process, but then all the local geometry is automatically built in, there are relatively few conformational variables, and the embeddability condition is simple. Distance constraints can also be expressed as bounds on linear combinations of metric matrix elements. Initial experience with the method on challenging test cases involving real experimental data [40] indicate that the method works at least as well as EMBED on these problems, and it may eventually work better on some classes of problems.

3.2. FUTURE ISSUES

The big issues in the future of distance geometry, particularly in connection with NMR, hinge on defining exactly what we want out of it. The first question is: do we want to know what a certain collection of experiments has told us about a molecule's conformation, or do we want to find the theoretically preferred conformer(s) out of all those allowed by the experimental evidence? Many arguments would be eliminated if people would simply confess which of these two goals they are aiming for. Agreed, the second goal has a lot of precedent in science, for example, in the refinement programs used in solving macromolecular crystal structures. Nevertheless, I have several objections.

(1) I would like to know the full *range* of conformers (if any) compatible with a given body of experimental data. If the range is broad, either the real molecule moves through the range, or one can suggest further experiments aimed at determining the conformation more precisely. If the range is narrow, and it disagrees with my favorite potential energy function, I know I have to fix the potential function.

(2) Apparently current empirical potential energy functions do not agree well with the experimentally determined structures, when these are narrowly determined. Therefore, using energy to select preferred conformers in the widely determined case seems to be a dubious practice.

(3) Proclaiming an energetically optimal and geometrically feasible conformer to be "the predicted/determined" structure is a misrepresentation at best. It is not a prediction or model building purely from low-level information, such as covalent structure and an energy function. Neither is it a definitive experimental determination because, especially with some computational methods, it is difficult to say which features came from the experimental data and which came from theory.

The other main issue I see is that of conformational flexibility. Although globular proteins can have rather well-defined conformation, in that one can solve protein X-ray crystal structures to high accuracy using essentially a static model, clearly there

are important dynamic features. NMR has been rightly heralded as a method capable of exploring such flexibility in some detail. So far, however, distance geometry methods have been applied to static models of protein structure, where all observed NOEs, etc., correspond to a set of geometric constraints that must be satisfied simultaneously and at all times. This leads to a sometimes distressing interaction between the computational chemist, who finds that mathematically the given constraints are mutually inconsistent, and the NMR spectroscopist, who insists his assignments are correct and the NOEs are reproducibly observed. If instead of calculating the feasibility or infeasibility of inequality constraints, one can turn to an optimal (but perhaps not complete) agreement between experiment and calculated static model. This leads to a problem well known in X-ray crystallography, where the optimal static model substantially disagrees with *all* the conformations available to the real molecule. For example, suppose the real molecule has a freely rotatable bond such that some group of atoms can lie anywhere on a possibly large circle with equal probability, when averaged over time and space in the macroscopic sample. The least-squares fit to this places the group at the center of the circle, far from any real position. In NMR, the key fact is that an NOE may be observed between two atoms undergoing motions rapid on the NMR time scale, such that they are close only a substantial fraction of the time.

So far, there is no very elegant theoretical description of large-scale, constrained molecular motions. On the practical front, Kim and Prestegard [41] have produced a greatly improved fit to the constraints for a small protein by using a two-state model. In general, one could imagine trying this when the given full set of constraints is geometrically infeasible, but usually there would be multiple ways to divide the constraints into two overlapping subsets such that each is feasible. Alternatively, Torda et al. [42] have set up a pseudopotential for a molecular dynamics simulation, such that two atoms involved in an NOE are drawn together only if their distance has been too great over a substantial fraction of the recent past in the simulation. This allows the molecule to switch back and forth between one or more alternative states, satisfying only various subsets of the NOE constraints at any one instant, but satisfying them all over the time average, even if the total set of constraints is geometrically infeasible. The advantage is that the simulation automatically takes care of deciding which alternative states are necessary. The possible disadvantages are that it is not known whether this method will always yield a solution when the data can be explained in terms of a small number of alternative states; and secondly, the method may be able to fit any collection of constraints whatsoever. The third problem is that the result of a successful calculation is an entire molecular dynamics trajectory, which may be difficult to summarize. This leads back to the need for a convenient representation for large, concerted, molecular motions. Clearly, applied distance geometry is a lively field with many problems outstanding, and in need of all the clever ideas anyone can contribute.

Acknowledgements

This work was supported by grants from the National Institutes of Health (GM37123 and DA06746).

References

[1] L.M. Blumenthal, *Theory and Applications of Distance Geometry* (Chelsea Publ. Co., Bronx, New York, 1970).
[2] G.M. Crippen and T.F. Havel, *Distance Geometry and Molecular Conformation*, in: *Chemometrics Research Studies Series*, ed. D. Bawden, Research Studies Press (Wiley, New York, 1988).
[3] A.W.M. Dress and T.F. Havel, SIAM J. Discr. Math. (1991), in press.
[4] J.B. Saxe, Embeddability of weighted graphs in k-space is strongly NP-hard, in: *17th Allerton Conf. on Communication, Control and Computing* (1979), pp. 480–489.
[5] J.E. Goodman and R. Pollack, SIAM J. Comput. 12(1983)484–507.
[6] W. Glunt, T.L. Hayden, S. Hong and J. Wells, SIAM J. Matrix Anal. Appl. 11(1990)589–600.
[7] W. Glunt, T.L. Hayden and W.-M. Liu, Bull. Math. Biol. (1991), in press.
[8] T. Tay and W. Whitely, Structural Topology 9(1984)31–38.
[9] B.A. Hendrickson, The molecule problem: Determining conformation from pairwise distances, Technical Report 90-1159, Department of Computer Science, Cornell University, Ithaca, NY (1990).
[10] H.J. Kushner, SIAM J. Appl. Math. 47(1987)169–185.
[11] M. Nilges, G.M. Clore and A.M. Gronenborn, FEBS Lett. 229(1988)317–324.
[12] T.A. Holak, M. Nilges and H. Oschkinat, FEBS Lett. 242(1989)218–224.
[13] M. Nilges, G.M. Clore and A.M. Gronenborn, FEBS Lett. 239(1988)129–136.
[14] M.A. Hadwiger and G.E. Fox, J. Biomol. Struct. and Dynamics 7(1989)749–771.
[15] I.D. Kuntz, J.F. Thomason and C.M. Oshiro, Methods Enzymol. 177 (Nucl. Magn. Reson., Pt. B) (1989)159–204.
[16] P.L. Easthope and T.F. Havel, Bull. Math. Biol. 51(1989)173–194.
[17] W.J. Metzler, D.R. Hare and A. Pardi, Biochem. 28(1989)7045–7052.
[18] R.M. Levy, D.A. Bassolino, D.B. Kitchen and A. Pardi, Biochem. 28(1989)9361–9372.
[19] T.F. Havel, Biopolymers 29(1990)1565–1585.
[20] G.M. Crippen, A.S. Smellie and J.W. Peng, J. Chem. Inf. Comp. Sci. 28(1988)125–128.
[21] G.M. Crippen, J. Comp. Chem. 3(1982)471–476.
[22] G.M. Crippen, Biopolymers 21(1982)1933–1943.
[23] G.M. Crippen, J. Phys. Chem. 91(1987)6341–6343.
[24] M.E. Snow and G.M. Crippen, Int. J. Peptide Protein Res. (1991), in press.
[25] G.M. Crippen and T. Havel, J. Chem. Inf. Comp. Sci. 30(1990)222–227.
[26] E.O.Purisima and H.A. Scheraga, Proc. Natl. Acad. Sci. USA 83(1986)2782–2786.
[27] E.O. Purisima and H.A. Scheraga, J. Mol. Biol. 196(1987)697–709.
[28] W.C. Ripka, W.J. Sipio and J.M. Blaney, Lect. Heterocycl. Chem. 9(1987)95–104.
[29] R.K. Gordon, E. Breuer, F.N. Padilla, R.M. Smejkal and P.K. Chiang, Mol. Pharmacol. 36(1989) 766–772.
[30] P.A. Kollman, P.D.J. Grootenhuis and M.A. Lopez, Pure Appl. Chem. 61(1989)593–596.
[31] K. Wüthrich, *NMR of Proteins and Nucleic Acids* (Wiley, New York, 1986).
[32] K. Wüthrich, Science 243(1989)45–50.
[33] K. Wüthrich, Acc. Chem. Res. 22(1989)36–44.
[34] T.F. Havel, private communication (1984).
[35] M. Billeter, T. Schaumann, W. Braun and K. Wüthrich, Biopolymers 29(1990)695–706.

[36] R. Boelens, T.M.G. Koning and R. Kaptein, J. Mol. Struct. 173(1988]299–311.
[37] M.F. Summers, T.L. South, B. Kim and D.R. Hare, Biochem. 29(1990)329–340.
[38] M.T. Brenneman and T.A. Cross, J. Chem. Phys. 92(1990)1483–1494.
[39] G.M. Crippen, J. Comput. Chem. 10(1989)896–902.
[40] H.I. Mosberg, K. Sobczyk-Kojiro, P. Subramanian, G.M. Crippen, K. Ramalingam and R.W. Woodard, J. Amer. Chem. Soc. 112(1990)822–829.
[41] Y. Kim and J.H. Prestegard, Biochem. 28(1989)8792–8797.
[42] A.E. Torda, R.M. Scheek and W.F. van Gunsteren, J. Mol. Biol. 214(1990)223–235.