Cluster Distance Geometry of Polypeptide Chains

GORDON M. CRIPPEN
College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109-1065

Received 9 February 2004; Accepted 23 March 2004
DOI 10.1002/jcc.20056
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: Distance geometry has been a broadly useful tool for dealing with conformational calculations. Customarily each atom is represented as a point, constraints on the distances between some atoms are obtained from experimental or theoretical sources, and then a random sampling of conformations can be calculated that are consistent with the constraints. Although these methods can be applied to small proteins having on the order of 1000 atoms, for some purposes it is advantageous to view the problem at lower resolution. Here distance geometry is generalized to deal with distances between sets of points. In the end, much of the same techniques produce a sampling of different configurations of these sets of points subject to distance constraints, but now the radii of gyration of the different sets play an important role. A simple example is given of how the packing constraints for polypeptide chains combine with loose distance constraints to give good calculated protein conformers at a very low resolution.

 J Comput Chem 25: 1305–1312, 2004

Key words: distance geometry; protein conformation; excluded volume; conformational analysis; radius of gyration

Introduction

Distance geometry refers to a treatment of geometric problems that emphasizes Euclidean distances between points, rather than angles, and so forth.[1] This is sometimes a convenient way to deal with conformational problems, where the points correspond to atoms.[2–4] Currently the most frequent application is to calculate a sampling of sets of atomic coordinates given constraints on some of the interatomic distances derived from NMR experiments.[5–7] Other applications include protein homology modeling,[8,9] protein structure prediction,[10–12] and more abstract conformation[13] and sequence[14] spaces. A recurring task in all these applications is how to generate a set of coordinates for the points given at least some constraints on some of the interpoint distances. New methods for this continue to be developed,[15–17] each with its advantages and disadvantages, depending on the type of constraint set.

Here we consider a new sort of distance geometry problem where the points may be atoms or even whole amino acid residues, and we want to look at conformations at a very low resolution where the primary objects are clusters or subsets of the points. Much of the standard methodology for single points carries over in an analogous form for clusters, with the advantage of building in the space-filling features of real atoms or residues in a natural way. In order to make this correspondence clear, we first briefly outline the standard distance geometry methods, then show the equivalent procedures for clusters, and finally give a simple demonstration of calculating conformations for a protein given certain constraints.

Methods

Standard Distance Geometry

Suppose we have a set of n distinct points with a matrix of proposed squared Euclidean distances between them, $\mathbf{M} = (d_{ij}^2)$. Obviously $\mathbf{M}$ must be symmetric ($d_{ij}^2 = d_{ji}^2$), the diagonal must be zero ($d_{ii}^2 = 0$), and the elements must be non-negative ($d_{ij}^2 \geq 0$). There are in addition some nonobvious requirements that $\mathbf{M}$ must fulfill in order to correspond to a realizable arrangement of the points in three-dimensional space that is not confined to some planar subspace. These constraints are expressed in terms of $C(\mathbf{M}, k)$, the Cayley-Menger determinant involving the first k points:

$$C(\mathbf{M}, k) = \begin{vmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & d_{12}^2 & \cdots & d_{1k}^2 \\ 1 & d_{21}^2 & 0 & \cdots & d_{2k}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & d_{k1}^2 & d_{k2}^2 & \cdots & 0 \end{vmatrix} \quad (1)$$

Blumenthal's theorem[1] specialized to three dimensions and $n \geq 4$ requires that there be some ordering of the points such that $C(\mathbf{M},$

Correspondence to: G. M. Crippen; e-mail: gcrippen@umich.edu

Contract/grant sponsor: University of Michigan Bioinformatics Program.
Contract/grant sponsor: Howard Hughes Medical Institute.

2) > 0, $C(\mathbf{M}, 3) < 0$, $C(\mathbf{M}, 4) > 0$, and for any choice of additional fifth and sixth points that $C(\mathbf{M}, 5) = C(\mathbf{M}, 6) = 0$.

The constraints on the lower order determinants can readily be interpreted.[2,3] Because $C(\mathbf{M}, 2) = 2d_{12}^2$, the first inequality is trivially satisfied by $d_{12} > 0$. The second inequality involves factoring the determinant

$$C(\mathbf{M}, 3) = (d_{23} + d_{12} - d_{13})(d_{12} + d_{13} + d_{23})(d_{12} + d_{13} - d_{23}) \times (d_{12} - d_{13} - d_{23}) < 0 \quad (2)$$

which is satisfied if the triangle inequality, $d_{13} \leq d_{12} + d_{23}$, holds as a strict inequality for all three permutations of indices. The triangle inequality becomes an equality when the three points are colinear, so the requirement that $C(\mathbf{M}, 3) > 0$ means the first three points are not colinear. The algebra for four points is more complicated, but solving $C(\mathbf{M}, 4) = 0$ for d_{14}^2 (assuming fixed values of the other distances) gives two real, positive roots, and as long as d_{14}^2 lies strictly between these two bounds, we have $C(\mathbf{M}, 4) > 0$. This has been called the tetrangle inequality,[4] and satisfying the strict inequality implies the first four points are noncoplanar.

Another way to look at the Cayley-Menger determinants is in terms of coordinates of the points in the case that the constraints on the corresponding distances are satisfied and coordinates can consequently be found. Any two points i and j span a linear subspace that we can equip with a coordinate system consisting of an origin and an x-axis. Then the determinant for the ordered pair of points

$$C(\mathbf{M}, [i, j]) = 2\chi_{ij}^2 \quad (3)$$

where

$$\chi_{ij} = \begin{vmatrix} 1 & 1 \\ x_i & x_j \end{vmatrix} = x_j - x_i \quad (4)$$

is the orientation of the two ordered points relative to the coordinate system. That is, $\chi_{ij} > 0$ if going from x_i to x_j is the same direction as going from the origin to the positive x-axis, $\chi_{ij} < 0$ if the orientation is opposite, and $\chi_{ij} = 0$ if the two points coincide and hence have no orientation along the x-axis. Similarly, it is straightforward but tedious to verify that for an ordered set of three points, $[i, j, k]$, spanning two dimensions, the corresponding Cayley-Menger determinant is

$$C(\mathbf{M}, [i, j, k]) = -4\chi_{ijk}^2 \quad (5)$$

where the three-point orientation relative to the x- and y-axes of the coordinate system is

$$\chi_{ijk} = \begin{vmatrix} 1 & 1 & 1 \\ x_i & x_j & x_k \\ y_i & y_j & y_k \end{vmatrix} \quad (6)$$

and $\chi_{ijk} > 0$ when going from point i to j to k is counterclockwise in the xy-plane. When the three points are colinear, $\chi_{ijk} = 0$.

There are other useful quantities that follow directly from the distances without reference to coordinates. Given the matrix of squared distances, $\mathbf{M}$, one can directly calculate d_{io}^2, the squared distance from point i to the unweighted center of mass of all n points[18].

$$d_{io}^2 = n^{-1} \sum_j d_{ij}^2 - n^{-2} \sum_{j>k} d_{jk}^2 \quad (7)$$

Because the radius of gyration of the n points is $r_g = (n^{-1} \sum_i d_{io}^2)^{1/2}$, it follows from eq. (7) that

$$r_g^2 = n^{-2} \sum_{j>k} d_{jk}^2 \quad (8)$$

An important problem in distance geometry is finding coordinates for the points, if any can exist, that satisfy given bounds on the distances between some of the pairs of points. Thus for every distance we require that $l_{ij} \leq d_{ij} \leq u_{ij}$, and for some distances the upper and lower bounds may be significant constraints, while for others they may be the trivial $l_{ij} = 0$ and $u_{ij} = \infty$. Bound smoothing[3,4] is a process whereby the information from the tight bounds is spread around to all distances by lowering some upper bounds and raising some lower bounds. At the triangle inequality level, one simply repeatedly examines all triples of points and whenever

$$u_{ik} \leq u_{ij} + u_{jk} \quad (9)$$

is violated, u_{ik} is reduced to the right-hand side of the inequality. After no further reductions in upper bounds can be achieved, all triples of points are repeatedly checked for violations of

$$l_{ik} \geq l_{ij} - u_{jk} \quad (10)$$

in which case l_{ik} is increased to the right-hand side of the inequality.

From the smoothed bounds, one may pick distance values at random in the intervals $l_{ij} \leq d_{ij} \leq u_{ij}$, independently for each d_{ij}. However, the resulting distances may not necessarily obey the triangle inequality. In the process called metrization,[5] each time a particular d_{ij} is chosen, the corresponding bounds are tightened to $l_{ij} = d_{ij} = u_{ij}$, and smoothing of the other bounds by eqs. (9) and (10) may tighten them. When all the distances have been chosen in this way, they do obey the triangle inequality.

In order to determine coordinates from the selected distances, the metric matrix $\mathbf{G} = (g_{ij})$ is determined by

$$g_{ij} = \tfrac{1}{2}(d_{io}^2 + d_{jo}^2 - d_{ij}^2) \quad (11)$$

where the distances to the center of mass come from eq. (7). Then the $j = x, y, z$ coordinate vectors of the n points are determined by

$$\mathbf{c}_j = \lambda_j^{1/2} \mathbf{w}_j \quad (12)$$

from the three largest (positive) eigenvalues λ_j and corresponding eigenvectors w_j of **G**. If the other $n - 3$ eigenvalues are relatively large in magnitude, the resulting coordinates may need adjusting in order to obey the original bounds.

Cluster Distance Geometry

General Relationships

Suppose the objects of interest are not individual points, but rather sets of b points. So if n is a multiple of b, we can think of grouping **M** into $b \times b$ blocks to form an order n/b matrix $\mathbf{M}_b = (D_{IJ})$, where each element D_{IJ} is the sum of the b^2 interpoint squared distances, d_{ij}^2, within that block. It is still true that $\mathbf{M}_b$ is a symmetric matrix of non-negative elements, but now the diagonal elements are no longer necessarily zero. Instead, we see from eq. (8) that

$$D_{II} = 2b^2 r_{g,I}^2 \tag{13}$$

where $r_{g,I}^2$ is the squared radius of gyration of the Ith set of points.

The equivalent constraints on the Cayley-Menger determinants apparently hold. The two-block determinant is not as trivial as the two-point determinant:

$$C(\mathbf{M}_b, [I, J]) = 2D_{IJ} - D_{II} - D_{JJ} > 0 \tag{14}$$

It is easy enough to demonstrate that this condition holds when there are coordinates for the points. Consider the case $b = 2$, involving only four points labeled i, $i + 1$, j, and $j + 1$. Suppose these points have coordinates $[x_i, y_i, z_i]^T$, and so forth. Then

$$\begin{aligned} C(\mathbf{M}_2, [I, J]) &= 2(d_{ij}^2 + d_{i+1,j}^2 + d_{i,j+1}^2 + d_{i+1,j+1}^2 - d_{i,i+1}^2 - d_{j,j+1}^2) \\ &= 2(x_i - x_j)^2 + 2(x_{i+1} - x_j)^2 + 2(x_i - x_{j+1})^2 + 2(x_{i+1} - x_{j+1})^2 \\ &\quad - 2(x_i - x_{i+1})^2 - 2(x_j - x_{j+1})^2 + \text{similar for } y \text{ and } z \\ &= 2(x_i + x_{i+1} - x_j - x_{j+1})^2 + \text{similar for } y \text{ and } z = 8d_{\bar{I},\bar{J}}^2 \end{aligned} \tag{15}$$

where $\bar{I}$ and $\bar{J}$ denote the unweighted centers of mass (or centroids) of the two clusters. In general

$$C(\mathbf{M}_b, [I, J]) = 2b^2 d_{\bar{I},\bar{J}}^2 = 2b^2 \chi_{\bar{I}\bar{J}}^2 \geq 0 \tag{16}$$

where $\chi_{\bar{I}\bar{J}}$ is the one-dimensional orientation for the two centroids as in eq. (4), and $C(\mathbf{M}_b, [I, J]) = 0$ only when the two centroids coincide. Clearly for $b = 1$, this reduces to the standard distance geometry case.

Similarly, $C(\mathbf{M}_b, 3) < 0$ is widely observed to be always satisfied for configurations of points in three dimensions, but is not so simply interpreted in terms of something like the triangle inequality. In terms of coordinates in the plane, one can derive in analogy to eq. (15) that

$$C(\mathbf{M}_b, [I, J, K]) = -4b^2 \chi_{\bar{I}\bar{J}\bar{K}}^2 \leq 0 \tag{17}$$

and $C(\mathbf{M}_b, [I, J, K]) = 0$ when the corresponding three centroids are colinear. In terms of distances, like the treatment of the tetrangle inequality above, one can solve $C(\mathbf{M}_b, [I, J, K]) = 0$ for an upper and lower bound on, say, D_{IK}, given values for all the other matrix elements. This gives

$$D_{IJ} + D_{JK} - D_{JJ} - P^{1/2} \leq D_{IK} \leq D_{IJ} + D_{JK} - D_{JJ} + P^{1/2} \tag{18}$$

where

$$\begin{aligned} P &= (2D_{JK} - D_{JJ} - D_{KK})(2D_{IJ} - D_{II} - D_{JJ}) \\ &= C(\mathbf{M}_b, [J, K])C(\mathbf{M}_b, [I, J]) \end{aligned} \tag{19}$$

and the bounds in eq. (18) are always real because both factors in eq. (19) are positive, according to the requirement of eq. (14). In the $b = 1$ limit where $D_{II} = D_{JJ} = D_{KK} = 0$, this reduces to the usual triangle inequality limits of

$$\max(d_{jk} - d_{ij}, d_{ij} - d_{jk}) \leq d_{ik} \leq d_{ij} + d_{jk}. \tag{20}$$

The equivalent to triangle inequality level bound smoothing is somewhat more complicated than in standard distance geometry. For every pair of blocks there are bounds $L_{IJ} \leq D_{IJ} \leq U_{IJ}$ that may be trivial ($L_{IJ} = 0$ and $U_{IJ} = \infty$) for some or nontrivial given bounds for others. The diagonal elements may also have nontrivial bounds, rather than $d_{ii} = 0$ for a single point i. Thus the first step is to repeatedly scan all pairs of blocks looking for violations of

$$L_{IJ} \geq (L_{II} + L_{JJ})/2 \tag{21}$$

which would require raising L_{IJ} to the right-hand side of the inequality, in accord with eq. (14). Similarly, a less frequently occurring case is violations of

$$2U_{IJ} - L_{II} \geq U_{JJ} \tag{22}$$

which requires lowering U_{JJ} to the left-hand side of the inequality. The same holds for I and J exchanged.

After no further tightening of bounds can be achieved, the next step is to repeatedly examine all unordered triples of blocks I, J, and K for violations of

$$\begin{aligned} U_{IK} \leq U_{IJ} + U_{JK} - L_{JJ} \\ + [(2U_{JK} - L_{JJ} - L_{KK})(2U_{IJ} - L_{II} - L_{JJ})]^{1/2} \end{aligned} \tag{23}$$

which comes from maximizing the right-hand side of eq. (18). Note that that expression achieves its maximum subject to the constraints that $D_{IJ} \leq U_{IJ}$, $D_{JK} \leq U_{JK}$, $D_{II} \geq L_{II}$, $D_{KK} \geq L_{KK}$, and $D_{JJ} \geq L_{JJ}$. The constraints that $2D_{JK} - D_{JJ} - D_{KK} \geq 0$ and $2D_{IJ} - D_{II} - D_{JJ} \geq 0$ are not active at the maximum, and the two factors in the square root of eq. (23) are necessarily non-negative. Here U_{IK} is reduced to the right-hand side, very much like using eq. (9).

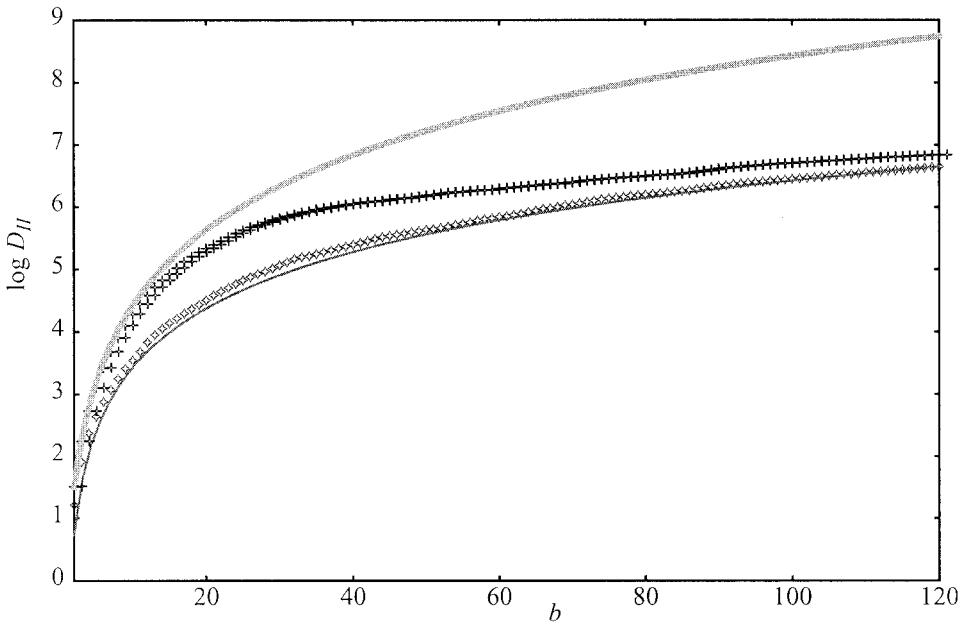

Figure 1. A semilog plot of the range of diagonal elements D_{II} as a function of block size b. Plotted values are those observed in a survey of 32 small protein crystal structures for log L_{II} (diamonds) and log U_{II} (crosses) compared to fitted upper bounds from eq. (30) (thick line) and fitted lower bounds from eq. (31) (thin line).

Eq. (23) is used to reduce off-diagonal upper bounds, just as in standard distance geometry, but it can also lead to reductions in U_{II}, U_{JJ}, and U_{KK}. For example, if substituting U_{II} for L_{II} in eq. (23) leads to a calculated $U_{IK} < L_{IK}$, then U_{II} must be reduced to

$$U_{II} \leq 2U_{IJ} - L_{JJ} - \frac{(L_{IK} - U_{IJ} - U_{JK} + L_{JJ})^2}{2U_{JK} - L_{JJ} - L_{KK}} \tag{24}$$

which comes from solving $C(\mathbf{M}_b, [I, J, K]) = 0$ for D_{II} when D_{IK}, D_{JJ}, D_{KK} are at their lower bounds, and D_{IJ} and D_{JK} are at their upper bounds. Swapping indices I and K in eq. (24) gives the condition for reducing U_{KK}. For reducing U_{JJ} in the equivalent situation, the limit is

$$U_{JJ} \leq \frac{(U_{IJ} - U_{JK})^2 + L_{IK}^2 + 2U_{IJ}(L_{KK} - L_{IK}) + 2U_{JK}(L_{II} - L_{IK}) - L_{II}L_{KK}}{L_{KK} + L_{II} - 2L_{IK}} \tag{25}$$

The last step is to repeatedly examine all unordered triples of blocks in order to raise L_{IK} by the equivalent of eq. (10). Referring to eq. (18), the question is whether the current value of L_{IK} is less than the minimal value of $D_{IJ} + D_{JK} - D_{JJ} - (C(\mathbf{M}_b, [I, J])C(\mathbf{M}_b, [J, K]))^{1/2}$ subject to $C(\mathbf{M}_b, [I, J]) \geq 0$, $C(\mathbf{M}_b, [J, K]) \geq 0$, and the upper and lower bounds on all five variables, D_{II}, D_{JJ}, D_{KK}, D_{IJ}, and D_{JK}. It is easy to verify that the left-hand side of eq. (18) can be expressed as

$$D_{IK} \geq \tfrac{1}{2}((\sqrt{C(\mathbf{M}_b, [I, J])} - \sqrt{C(\mathbf{M}_b, [J, K])})^2 + D_{II} + D_{KK}) \tag{26}$$

which suggests three cases depending on the allowed ranges of $C(\mathbf{M}_b, [I, J])$ and $C(\mathbf{M}_b, [J, K])$. The first case is when $\max C(\mathbf{M}_b, [I, J]) < \min C(\mathbf{M}_b, [J, K])$ as detected by $2U_{IJ} - L_{II} < 2L_{JK} - U_{KK}$ and $2L_{JK} - U_{KK} - L_{JJ} > 0$. Then the minimal value of D_{IK} is achieved at the combination of upper and lower bounds:

$$L_{IK} \geq U_{IJ} + L_{JK} - L_{JJ} - ((2U_{IJ} - L_{II} - L_{JJ})(2L_{JK} - L_{JJ} - L_{KK}))^{1/2} \tag{27}$$

which involves L_{KK}, not U_{KK}. The second case is when the allowed interval for $C(\mathbf{M}_b, [I, J])$ is strictly above that for $C(\mathbf{M}_b, [J, K])$. Simply exchanging indices I and K, one detects that $2U_{JK} - L_{KK} < 2L_{IJ} - U_{II}$ and $2L_{IJ} - U_{II} - L_{JJ} > 0$, resulting in

$$L_{IK} \geq L_{IJ} + U_{JK} - L_{JJ} - ((2L_{IJ} - L_{II} - L_{JJ})(2U_{JK} - L_{JJ} - L_{KK}))^{1/2}. \tag{28}$$

The third case is when the ranges overlap. The right-hand side of eq. (26) is minimized when they are equal, and it reduces to the two-block constraint that $L_{IK} \geq (L_{II} + L_{KK})/2$. In the special case of $b = 1$, the bounds on the diagonal elements are all zero, U_{IJ} becomes u_{ij}^2, and so forth, so that eq. (28) reduces to eq. (10), and eq. (27) reduces to the equivalent relation, $l_{ik} \geq l_{jk} - u_{ij}$.

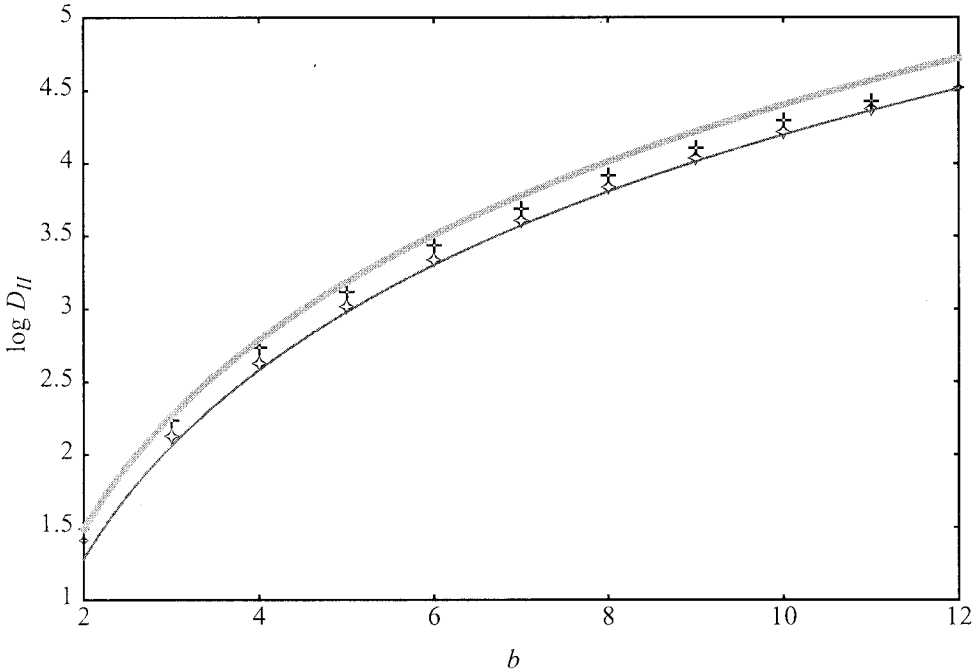

Figure 2. A semilog plot of the range of diagonal elements D_{II} as a function of block size b, when the corresponding protein residues are all part of a β-strand. Values observed in a survey of protein crystal structures for log L_{II} (diamonds) and log U_{II} (crosses) are compared to fitted upper (thick line) and fitted lower bounds (thin line) from eq. (34).

Protein Specific Relationships

Consider a low resolution representation of protein structure where the n points are the C^{α} atoms of the n amino acid residues in a single polypeptide chain, one point per residue. Because of the chemical structure of the chain and its space-filling atoms, there are lower and upper bounds on the cluster distances that are much tighter than the pure geometric constraints. For instance, sequentially adjacent C^{α} atoms linked with a *trans* peptide bond would have a fixed separation of 3.8 Å, assuming standard bond lengths and angles. On the other hand, the mathematical series

$$\sum_{i=1}^{b}\sum_{j=1}^{b}(i-j)^2 = \frac{b^2(b^2-1)}{6} \tag{29}$$

so the upper bound on the diagonal blocks $U_{II} \geq D_{II}$ is

$$U_{II} = (4.01\ \text{Å})^2\,\frac{b^2(b^2-1)}{6} \tag{30}$$

for a fully extended polypeptide chain. The scaling factor of 4.01 Å rather than 3.8 Å comes from a survey over crystal structures of small proteins (PDB codes 1A1X, 1ACF, 1BK2, 1BM8, 1BYW.A, 1C44.A, 1COA.I, 1CQY.A, 1DHN, 1DT4.A, 1ENH, 1EW4.A, 1G9O.A, 1HEY, 1I2T.A, 1JWO.A, 1MIL, 1MJC, 1OPS, 1PGB, 1PHT, 1PTF, 1QAU.A, 1TEN, 1TMY, 1TUL, 1UBI, 1VCC, 1WHI, 2IGD, 3IL8, and 9MSI.A) and takes into account some experimental variations from standard peptide geometry. Over this dataset, the bounds are tight for $b \leq 10$, but for larger blocks, finding such fully extended chain segments becomes unlikely, as shown in Figure 1.

For the lower bounds on the diagonal blocks, $L_{II} \leq D_{II}$, we know from eq. (13) that this corresponds to the minimal squared radius of gyration for a contiguous segment of b residues, and we know that the minimal $r_{g,I}$ varies linearly with $b^{1/3}$ from an earlier survey.[19] Once again surveying over the same 32 protein crystal structures as for the upper bound, a tight lower bound for $1 \leq b \leq 129$ is

$$L_{II} = 2b^2(3.17\ \text{Å})^2(b^{1/3}-1)^2 \tag{31}$$

as shown in Figure 1.

If all we know about a protein's structure is that it is a single, connected polypeptide chain, then the diagonal lower bounds are just L_{II} for all I from eq. (31), and the off-diagonal lower bounds L_{IJ} are that same value for all I and J by eq. (21). The default diagonal upper bounds, $U(b)$, come from eq. (30), but these can be used to build up the off-diagonal upper bounds. Note that $2U_{I,I+1} + U_{II} + U_{I+1,I+1} = U(2b)$ and $U_{II} = U_{I+1,I+1} =$

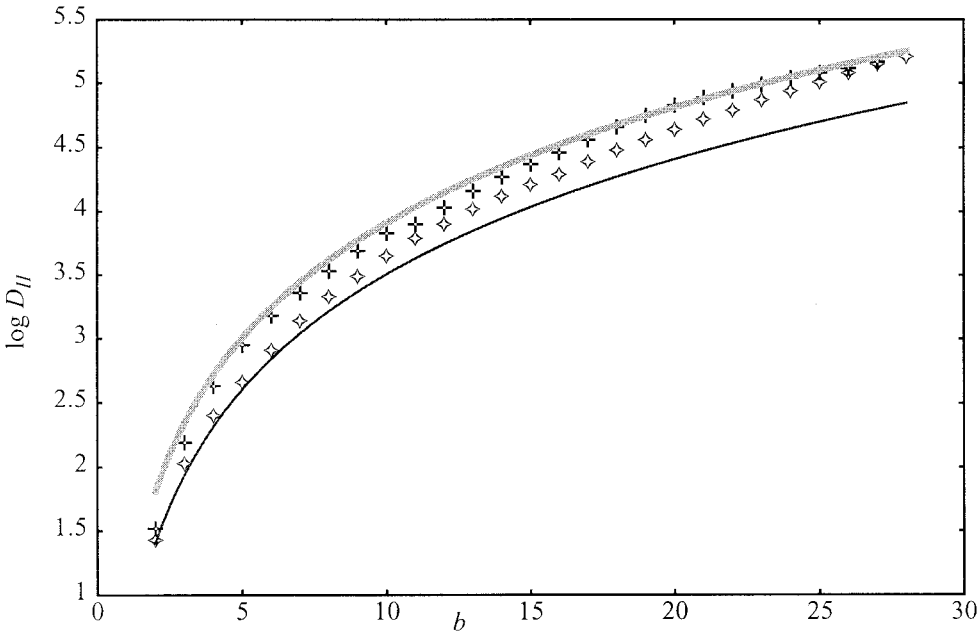

Figure 3. A semilog plot of the range of diagonal elements D_{II} as a function of block size b, when the corresponding protein residues are all part of an α-helix. Values observed in a survey of protein crystal structures for log L_{II} (diamonds) and log U_{II} (crosses) are compared to fitted upper (thick line) and fitted lower bounds (thin line) from eq. (35).

$U(b)$, so the first off-diagonal upper bounds are all determined. In general

$$U_{I,I+k} = U_{I+k,I} = \frac{1}{2}\left(U((1 + k)b) - \sum_{l,m} U_{l,m} \right) \qquad (32)$$

for $k = 1, \ldots, (n/b) - 1$, where the sum runs over the square submatrix $I \leq l$, $m \leq I + k$ except for the two corners, $(I, I + k)$ and $(I + k, I)$.

However, if any additional constraints have tightened some bounds more than those corresponding to a single polypeptide chain, the maximal packing density considerations can raise some off-diagonal lower bounds. Using the same notation as eq. (32), violations of

$$L_{I,I+k} = L_{I+k,I} \geq \frac{1}{2}\left(L((1 + k)b) - \sum_{l,m} U_{l,m} \right) \qquad (33)$$

require that $L_{I,I+k}$ and $L_{I+k,I}$ be raised to the value of the right-hand side.

A knowledge of secondary structure can give tighter bounds on the diagonal elements. If all the residues in a block of b residues are known to be part of a β-strand, then a survey over the same set of small proteins finds rather tight bounds:

$$(3.09\ \text{Å})^2 \frac{b^2(b^2 - 1)}{6} = L_{II} < U_{II} = (3.92\ \text{Å})^2 \frac{b^2(b^2 - 1)}{6} \qquad (34)$$

analogous to eq. (30), as shown in Figure 2. For an α-helical block, the dependence on b is of lower order, but not so low as in eq. (31). Cubic dependence fits well

$$(1.79\ \text{Å})^2 b^3 = L_{II} < U_{II} = (2.86\ \text{Å})^2 b^3 \qquad (35)$$

as shown in Figure 3.

Embedding

Finding coordinates for the centroids of the blocks from distance constraints is somewhat more elaborate than in standard distance geometry, but basically follows the same procedure. Initial bounds are either the trivial $L_{IJ} = 0$ and $U_{IJ} = \infty$, or in the case of a polypeptide chain use eqs. (30), (31), and (32). Additional information further tightens some bounds, such as knowledge of protein secondary structure via eqs. (34) and (35). Subsequent bound smoothing involves exhaustive application of two-block relations [eqs. (21) and (22), and for proteins eq. (33)] and three-block relations [eqs. (23), (24), (25), (27), and (28)]. Metrization works as before to choose random diagonal and off-diagonal D_{IJ} consistent with all applicable bound smoothing relations. Conceptually, this amounts to proposed radii of gyration for all the blocks, as well as distances between them. By eq. (16), the $(n/b) \times (n/b)$ matrix of proposed inter-centroid distances is calculated, which is then converted to coordinates as usual by eqs. (11) and (12). For very low resolution models, such as $(n/b) < 7$, the resulting coordinates often completely satisfy the original bounds without any further adjustment.

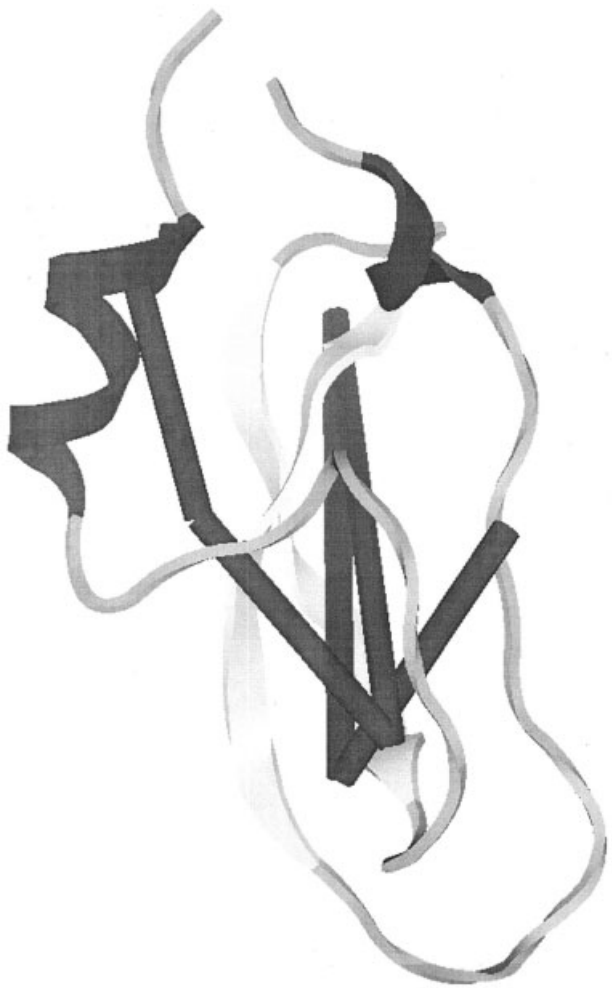

Figure 4. Backbone trace of BPTI (PDB entry 1G6X.A) less the first four residues (curved ribbon). The sequence of cylinders shows the positions of the centers of mass of the six sequential nine-residue segments.

Embedding Results

As a simple example of cluster distance geometry embedding applied to a protein, consider bovine pancreatic trypsin inhibitor (BPTI, PDB entry 1G6X.A). This consists of a single polypeptide chain of 58 residues. Deleting the N-terminal first four residues permits us to view the chain as a sequence of six clusters each consisting of nine C^{α} atoms or points. Figure 4 shows the C^{α} trace of the crystal structure and the greatly simplified trace of the six clusters, which nonetheless roughly outline the overall backbone path in space. For the sake of clarity, the radii of gyration of the six clusters are shown separately in Figure 5 as spheres centered at the cluster centers of mass and having the corresponding radii. Note how the C-terminal helix and the tight bend at the top of the illustration in Figure 4 correspond to small sphere in Figure 5, whereas the N-terminal extended strand on the right side has a large sphere.

In standard distance geometry, one could model this protein as 54 C^{α} points linked by virtual bonds and otherwise restricted by lower bounds between all points representing the self-avoiding character of the chain. If the only other constraints were upper bounds among points 1, 9, 18, 27, 36, 45, and 54 taken to be slightly greater than the distances in the crystal structure, then a wide variety of conformations would be possible, including great rearrangements of the general fold. For the cluster distance geometry treatment, we included the *a priori* constraints for polypeptide chains plus off-diagonal U_{IJ} that were 10% greater than those corresponding to the crystal structure. These explicit constraints plus the maximal packing density restrictions from eqs. (31) and (33) so greatly restrict the possible conformations that metrization sometimes has trouble finding a permitted set of D_{IJ} values. In terms of the root mean squared deviation (RMSD) after optimal superposition of the cluster centers of mass from the crystal structure versus those from embedding, RMSD values ranged from 3 to 6 Å. Figure 6 shows a calculated cluster structure superimposed on the crystal structure clusters where the RMSD in cluster centers of mass was 3.4 Å. Obviously the calculated structure is constrained to be no bigger than the native, but note how some large distances are indirectly enforced by the packing considerations. Clearly there is still some room for conformational variation, as shown in Figure 7 where the traces of several calculated

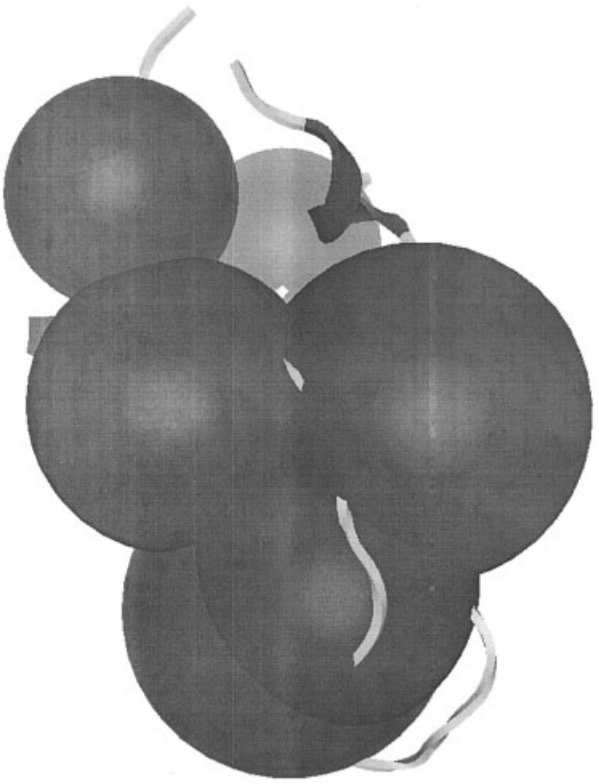

Figure 5. The radii of gyration of the six nine-residue segments of BPTI in the same view as the previous figure. Each segment is represented as a solid sphere centered at its center of mass having radius equal to the radius of gyration.

structures are shown superimposed on the crystal structure (not shown) in the same view as the previous figures. Overall, the RMSD of the corresponding cluster radii of gyration ranged from only 1.5 to 2.5 Å, which varies so little from Figure 5 that it would hardly be noticeable by eye.

Conclusion

Cluster distance geometry can be applied to a wide variety of geometric problems where a great reduction in resolution is desirable or necessary. Special problem-specific information can be built in, such as chain connectivity, secondary structure, and steric packing limitations, as shown for proteins. In particular, the packing constraints are very naturally incorporated, and their effect on other features is easily propagated. This is an effect that is difficult to include in standard distance geometry, which may make cluster distance geometry a useful approach even in instances that do not require a low resolution treatment.

Acknowledgments

All calculations were done in MOE using the SVL computer language.[20] Source code is available from the author on request.

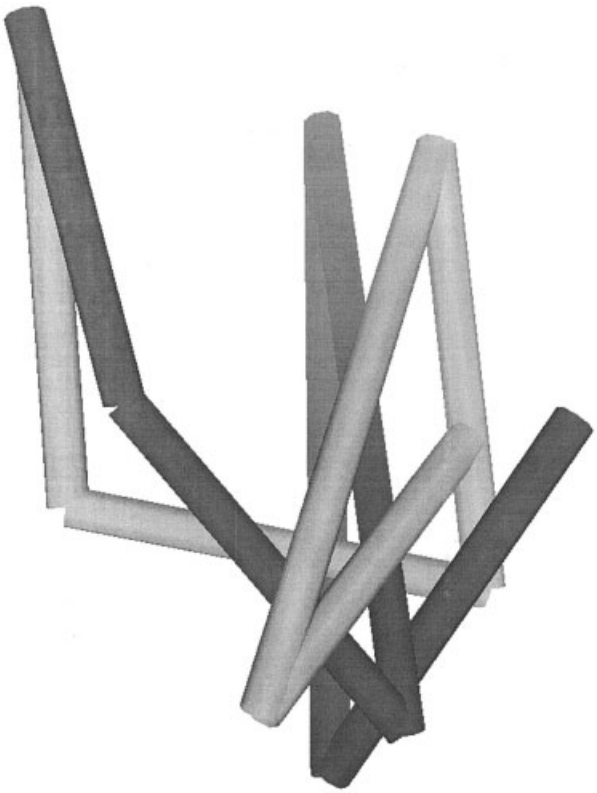

Figure 6. Superposition of the centers of mass of the BPTI segments for the native (string of darker cylinders extending slightly lower on the page) and calculated by embedding (other string of cylinders).

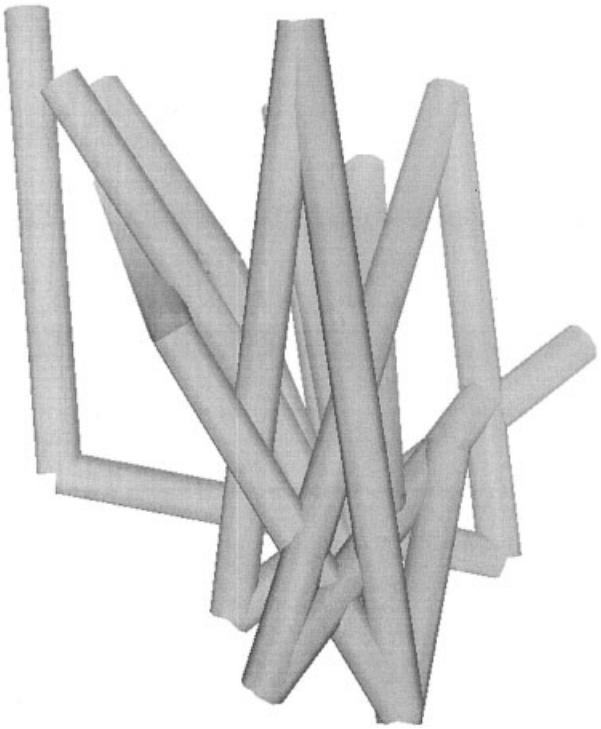

Figure 7. Superposition of the centers of mass for five calculated BPTI structures in the same view.

References

1. Blumenthal, L. M. Theory and Applications of Distance Geometry, 2nd edn.; Chelsea: New York, 1970, 98–105.
2. Crippen, G. M. J Comput Phys 1977, 24, 96.
3. Crippen, G. M. Distance Geometry and Conformational Calculations; Wiley: New York, 1981.
4. Crippen, G. M.; Havel, T. F. Distance Geometry and Molecular Conformation; Wiley: New York, 1988.
5. Havel, T. F.; Wüthrich, K. Bull Math Biol 1984, 46, 673.
6. Oezguen, N.; Adamian, L.; Xu, Y.; Rajarathnam, K.; Braun, W. J Biomol NMR 2002, 22, 249.
7. Wider, G.; Wüthrich, K. Curr Opin Struct Biol 1999, 9, 594.
8. Havel, T. F.; Snow, M. E. J Mol Biol 1991, 217, 1.
9. Combet, C.; Jambon, M.; Deleage, G.; Geourjon, C. Bioinformatics 2002, 18, 213.
10. Huang, E. S.; Samudrala, R.; Ponder, J. W. J Mol Biol 1999, 290, 267.
11. Xia, Y.; Huang, E. S.; Levitt, M.; Samudrala, R. J Mol Biol 2000, 300, 171.
12. Petersen, K.; Taylor, W. R. J Mol Biol 2003, 325, 1039.
13. Laboulais, C.; Ouali, M.; Le Bret, M.; Gabarro-Arpa, J. Proteins 2002, 47, 169.
14. Forster, M.; Heath, A.; Afzal, M. Bioinformatics 1999, 15, 89.
15. Glunt, W.; Hayden, T. Comput Chem 2001, 25, 223.
16. Williams, G. A.; Dugan, J. M.; Altman, R. B. J Comput Biol 2001, 8, 523.
17. Xu, H.; Izrailev, S.; Agrafiotis, D. K. J Chem Inf Comput Sci 2003, 43, 1186.
18. Crippen, G. M.; Havel, T. F. Acta Cryst A 1978, 34, 282.
19. Maiorov, V. N.; Crippen, G. M. J Mol Biol 1992, 227, 876.
20. Molecular Operating Environment (MOE), Chemical Computing Group, Inc. http://www.chemcomp.com.