Journal of Biomolecular NMR, 2 (1992) 33–56
ESCOM

J-Bio NMR 049

Sampling and efficiency of metric matrix distance geometry: A novel partial metrization algorithm

John Kuszewski*, Michael Nilges and Axel T. Brünger**

The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, U.S.A.

Received 17 September 1991
Accepted 16 October 1991

Keywords: Distance geometry; Nuclear magnetic resonance; Three-dimensional structure; Simulated annealing

SUMMARY

In this paper, we present a reassessment of the sampling properties of the metric matrix distance geometry algorithm, which is in wide-spread use in the determination of three-dimensional structures from nuclear magnetic resonance (NMR) data. To this end, we compare the conformational space sampled by structures generated with a variety of metric matrix distance geometry protocols. As test systems we use an unconstrained polypeptide, and a small protein (rabbit neutrophil defensin peptide 5) for which only few tertiary distances had been derived from the NMR data, allowing several possible folds of the polypeptide chain. A process called 'metrization' in the preparation of a trial distance matrix has a very large effect on the sampling properties of the algorithm. It is shown that, depending on the metrization protocol used, metric matrix distance geometry can have very good sampling properties indeed, both for the unconstrained model system and the NMR-structure case. We show that the sampling properties are to a great degree determined by the way in which the first few distances are chosen within their bounds. Further, we present a new protocol ('partial metrization') that is computationally more efficient but has the same excellent sampling properties. This novel protocol has been implemented in an expanded new release of the program X-PLOR with distance geometry capabilities.

INTRODUCTION

In multidimensional nuclear magnetic resonance (NMR) spectroscopy of biological macromolecules in solution (Ernst et al., 1986; Wüthrich, 1986; Clore and Gronenborn, 1991), the number of independent distance and dihedral angle constraints that can be obtained from NOE and J-coupling experiments is usually less than the molecule's degrees of freedom. Therefore, the conformational sampling of the method used to determine structures consistent with those

*Present address: Department of Biology, The Johns Hopkins University, Charles and 34th Streets, Baltimore, MD 21218, U.S.A.
**To whom correspondence should be addressed.

constraints is of great practical importance. In fact, the spread of a family of independently-determined structures is often used to estimate the precision of NMR-derived structures (for example, Havel and Wüthrich, 1984). The fact that the sampling properties of the metric matrix distance geometry algorithm are not always optimal, especially in the case of extended polypeptides, has been of some concern for a while (Brünger et al., 1987; Nilges et al., 1988; Thomason and Kuntz, 1989). The poor sampling has been shown to be of particular significance in the determination of the structure of rabbit neutrophil defensin peptide 5 (NP-5), a system with few interproton distance NOE constraints (Pardi et al., 1988). Levy et al. (1989) used a Monte Carlo search in torsion angle space to find several NP-5 folding topologies of reasonable conformational energies which were consistent with the observed NOEs, while they obtained only a single fold with a distance geometry program.

A number of distance geometry programs are commonly used, among them DISGEO (Havel and Wüthrich, 1984), DSPACE (D. Hare and R. Morrison, unpublished data; Hare and Reid, 1986), and the UCSF program (Kuntz et al., 1979). Although based on the same theory, the three programs exhibit different sampling properties depending on what options are used. Using unconstrained polypeptides as model systems, Metzler et al. (1989) reported overly-extended conformations which are very similar to each other, as measured by their mean backbone root-mean-square differences (RMSDs) and end-to-end distances, calculated with the program DSPACE. Attempts to reproduce their results with the program DISGEO failed (Nilges, unpublished results). First, the calculated structures after embedding were much better in terms of their geometry than those reported by Metzler et al. (1989). Second, the nonuniform random number distribution that the standard version of DISGEO used to choose distances within their bounds had the curious effect that the unconstrained polypeptides typically had a helical shape; the radius and pitch of this helix did not vary greatly between structures. On the other hand, with a uniform random number distribution, the generated structures sampled the conformational space much better than Metzler et al. (1989) reported. The reason for this improved sampling of DISGEO is a process called metrization (Havel and Wüthrich, 1984) to obtain a trial distance matrix which is consistent at the triangle inequality level. However, since the commonly used 4-stage protocol for structure calculations with DISGEO includes a substructure generation stage which cannot employ metrization, the full effect of this improved sampling was generally not seen by researchers applying the program to NMR structure determination (in the model calculations mentioned above, the substructure stage was bypassed).

Havel (1990) compared the sampling of metric matrix distance geometry *with* metrization to that of the ellipsoid algorithm (Billeter et al., 1986) and variable target function searches in torsion angle space (Braun and Gō, 1985) for unconstrained polypeptides. He found that the metrization procedure used in DISGEO introduced a new sampling problem, since unconstrained polymer structures were significantly more compact than polymer theory predicted. Havel (1990) solved the problem by randomizing the sequence in which distances are chosen within their bounds, and by using a uniform random number distribution to choose distances within their

---

Abbreviations: RMSD, root-mean-square difference; SA, simulated annealing; NMR, nuclear magnetic resonance; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser effect spectroscopy; DG, distance geometry; NP-5, rabbit neutrophil defensin peptide 5.

bounds. Unconstrained alanine polymer structures calculated with this new 'random metrization' protocol appear to sample their available conformational space very well.

A question that is of great importance to the experimental NMR spectroscopist remains as yet to be explored, namely, how does metrization affect the quality and sampling of NMR-derived structures, rather than unconstrained model systems. In this work we therefore used the small protein rabbit neutrophil defensin peptide (NP-5) with NMR-derived interproton distance data (Pardi et al., 1988) as a test case. The conformation of NP-5 appears to be under-determined by the NMR data allowing several quite distinct structures. An ideal structure determination program should thus determine all possible structures which are compatible with the NMR data.

In addition, we address a drawback of using metrization, namely, that it is computationally more expensive than the more 'traditional' distance geometry approaches without metrization. In order to reduce these computational requirements, we made use of the following observation: the complete set of all possible interatom distances is redundant. Knowing the distances from each of four selected atoms to all other atoms in the molecule is, at least in principle, sufficient to determine the entire molecule's conformation in three dimensions (Schlitter, 1987; Crippen and Havel, 1988; Hadwiger and Fox, 1989). Here we illustrate how 'partial' metrization can be used to efficiently generate structures with low conformational energy that sample the conformational space well. Apart from NP-5 we also use an unconstrained alanine 30mer, in order to be able to compare our results with previous work.

METHODS

The metric matrix EMBED algorithm (Havel et al., 1983; Crippen and Havel, 1988) has been implemented in an expanded new release of the program X-PLOR (Brünger, 1990) which is available upon request from ATB. We refer to the new distance geometry capabilities as X-PLOR/dg. Since the theory of distance geometry has been extensively reviewed by Crippen and Havel (1988), we concentrate here only on the special features of the new implementation.

An overview of X-PLOR/dg

X-PLOR/dg translates covalent geometry from the X-PLOR parameter files into interatom distance constraints, and puts them together with the experimentally-derived distance constraints into a matrix of upper and lower bounds on the distances between all pairs of atoms in the system. These bounds are 'smoothed' by determining the bounds implied by triangle inequalities. The program then randomly picks actual distances between the upper and lower bounds and, optionally, performs 'metrization', that is, it resmooths the remaining distance bounds after each distance pick. The distances are then embedded into Cartesian space, followed by regularization and minimization. These steps are discussed in greater detail below and are summarized in Fig. 1.

Input data

X-PLOR/dg translates known bond lengths, bond angles, dihedral angles, planarity restraints, and van der Waals radii, together with distance ranges derived from NOE measurements and dihedral angle ranges derived from coupling constant measurements, into upper and lower bounds on the distances between the atoms involved using the equations derived by Crippen and Havel (1988). The equation relating upper and lower distance bounds between the 1,4 atoms in a dihed-

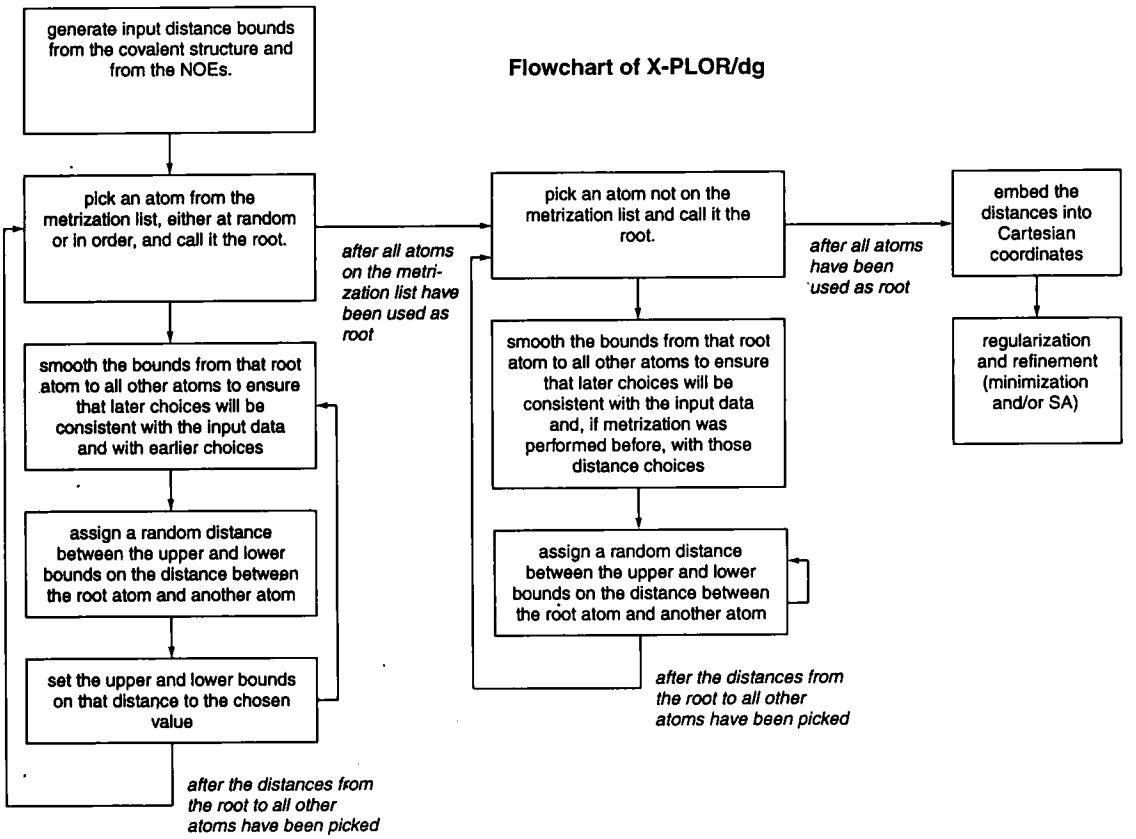


Fig. 1. The algorithm used in X-PLOR/dg to calculate structures via metric matrix distance geometry.

ral angle to the bounds on the angle φ derived by Crippen and Havel (1988) assumes that the distances among all atom pairs necessary to define the dihedral angle are precisely known. Since, in practice, there are only bounds known for these distances, reflecting the uncertainty of bond lengths and bond angles, X-PLOR/dg thus calculates the maximum and minimum 1,4 distance using all permutations of upper and lower bounds on φ and on the distances used to define the dihedral angle.

The covalent geometry is set using X-PLOR's parameter and topology files, rather than an additional library of standard residue coordinates, allowing use of the standard files for molecular dynamics, energy minimization (Brooks et al., 1983) or simulated annealing (Brünger and Karplus, 1991). However, in the course of this work, the bond angles and dihedral angles in the standard CHARMM parmallh6 parameter set had to be slightly modified to ensure geometric consistency for planar groups and aromatic rings (for example, the ideal value of the C-C-C angle inside the phenylalanine ring was changed from 109.5 to 120°).

The interproton distance and dihedral angle constraints are entered in using the existing X-PLOR facilities developed for NMR structure determination by simulated annealing. Thus, only one set of NMR data files needs to be created for both structure determination by distance geometry and refinement by simulated annealing.

Pseudoatoms

The bound smoothing involved in metrization can take up a substantial amount of computer time. Replacing methyl and methylene groups with single, large, 'pseudoatoms' (Wüthrich et al., 1983) can reduce the effective number of atoms in the system, thereby reducing the computer time needed to embed the structure. Unfortunately, some distance accuracy is lost because it requires corrections to be added to the measured NOE distances to and from these pseudoatoms (Wüthrich et al., 1983). In this study, pseudoatoms were not used to replace any groups, allowing more realistic packing in the final structures. NOE-derived distance constraints were treated with a method similar to that described by Pardi et al. (1988): the upper bounds on constraints to methyl groups were increased by the radius of the methyl group and reassigned to the methyl carbon. The upper bounds on distance constraints to pairs of nonstereospecifically assigned prochiral groups were increased by the distance between the two groups in order to be consistent with the work done by Pardi et al. (1988). Since NOEs involving aromatic rings could not be assigned to specific positions, a correction of 4.8 Å was added to all NOEs involving the protons on aromatic rings. Although we did not use them in this work, pseudoatoms could in principle be created in X-PLOR/dg by using special topology, parameter, and NOE input files.

Bound-smoothing and the shortest path algorithm

The input data usually define only a small fraction of the possible interatomic distances in the system. However, they imply upper and lower bounds on the distances between atom pairs not connected directly by the input data by means of triangle inequalities as shown by Havel et al. (1983). These implied bounds can be obtained by finding the shortest path between two points in a directed graph, where the directed graph represents all known or previously set distances. There are several algorithms which have been used to solve this well-studied computational problem, notably those of Dial et al. (1979) and Dijkstra (1959). Tarjan (1983) showed that the Dijkstra algorithm is theoretically the most efficient known and is quite amenable to parallel and vector-processing architectures. The Dijkstra algorithm has a complexity of $O(n^3)$. An implementation of Dijkstra's algorithm using the relaxed Fibonacci heaps described in Driscoll et al. (1988) would be expected to have a complexity of $O(n^2/\log(n))$, but it is less amenable to vector processing. Havel (program DISGEO, Havel and Wüthrich, 1984) suggested that the Dial algorithm (Dial et al., 1979) is much more efficient for systems with few known distances. However, we were able to match the performance of the Dial algorithm by implementing the Dijkstra algorithm on the tree of known distances. If the number of known distances is small, as is typically the case for NMR structure determinations, X-PLOR restricts execution of the innermost loops of the Dijkstra algorithm to the sparse tree of known distances. However, if metrization is employed, the number of known distances increases rapidly and at some point the overhead involved in restricting the loops to the known distances becomes noticeable. In this case, X-PLOR switches automatically to simpler loop structures that bypass the known distance tests. In our experience, this 'intelligent' implementation of the Dijkstra algorithm performs as well or better than the inplementation of the Dial algorithm in DISGEO.

Briefly, the Dijkstra algorithm finds the shortest path from a single atom (called the root) to all other atoms in the following way. A triangular matrix $D(atom, atom)$ is needed to provide the distance between two atoms. An array $P(atom)$ is needed to hold an increasingly accurate estimate of the length of the shortest path from the root to all other atoms. Finally, a logical array $F(atom)$

is needed to keep track of 'finished' atoms, which have had their root-to-self pathlength accurately determined. The value of D for each pair of atoms is initialized to the upper bound on that distance, if there is a constraint in the input data, otherwise to infinity. The pathlength P to each atom is initialized to infinity, except for the distance from the root to itself, which is obviously zero; and, initially, no atoms are 'finished', that is, F is *false* for all atoms. The following procedure is then repeated until all atoms are finished: the nonfinished atom with the minimum pathlength to the root is designated v (therefore, the root atom will always be the first atom to be designated v). The pathlength from the root to all other nonfinished atoms w is set to the minimum of its current pathlength and $D(root,v)+D(v,w)$. Atom v is then declared finished, and the loop begins again. When all atoms are finished, P contains the length of the shortest path from the root atom to all other atoms. This algorithm has been modified to calculate both upper and lower bounds and is implemented in X-PLOR/dg as described in Dress and Havel (1988).

A limitation of this simple bound smoothing algorithm is that the consistency of the distance matrix is only checked on the level of the triangle inequalities, which provide a necessary but not sufficient condition for successful embedding (Havel et al., 1983). Examining larger groups of atoms would produce tighter implied bounds, but using groups of four points in the bound smoothing process appears to be computationally too complex for practical use at the present time (Easthope and Havel, 1989; Kuszewski, unpublished).

Metrization

In order to 'embed' the structure; i.e., produce Cartesian coordinates from the distance constraints, actual distances between the upper and lower bounds must be chosen. This is done by assigning a random distance between the upper and lower bound on a given distance using one of several possible random distributions. Taking Havel's (1990) experiences into account, a uniform random number distribution has been chosen for our studies.

The bounds produced by the initial smoothing encompass all conformations which are consistent with the input data. A number of the distance geometry programs, such as older versions of DSPACE, pick the interatom distances independently of each other between the specified bounds. The resulting distances do therefore in general violate most triangle inequalities, and are even less likely to be consistent with a three-dimensional conformation. The embedding procedure, in best-fitting the coordinates to the distances, ends up producing distorted structures and exhibits very poor conformational sampling (see Results). The quality of the structures produced, as well as their conformational sampling, is expected to be improved by making the chosen distances more self-consistent.

A procedure, termed metrization, to ensure this self-consistency on the level of the triangle inequality has been developed (Havel and Wüthrich, 1984) and is available as an option in the program DISGEO. In it, a distance is chosen between the upper and lower bounds between two pairs of atoms. The bounds for this atom pair are set to the chosen value and the bounds matrix is resmoothed using the shortest path algorithm (see previous section) with this new information. The procedure is repeated for each atom pair, guaranteeing that each newly-chosen distance is consistent with the choices that have already been made. Thus, no overall initial bound-smoothing is required, since it is performed in the process of metrization (see Fig. 1).

If the distances from one atom to all the others are chosen one after another, successive calls to the bound-smoothing procedure can use the previous iteration's result as a very good starting

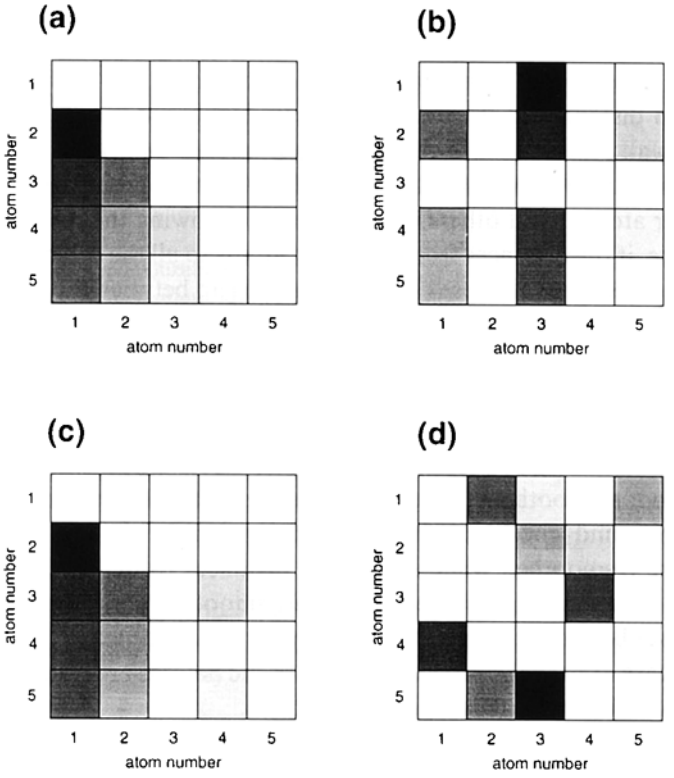


Fig. 2. The order in which interatomic distances are set during various complete metrization protocols. The darkness of a square indicates the point at which that interatomic distance is set (starting from black and going to white). The internal atom numbers run from the N-terminus to the C-terminus. Distances from each atom to itself are shown for clarity. Note that when the distance between atoms x and y is set, the y-to-x distance is also set, although this is not shown here for the sake of clarity. (a) Ordered metrization. The root atoms are chosen in order, and the distances from each root atom to the other atoms are also chosen in order. (b) Random metrization. The root atoms are chosen at random, but the distances from each root atom to the other atoms are chosen in order. (c) Combined metrization. In this case, the first two root atoms are chosen in order, and the remaining root atoms are chosen at random. Once again, the distances from each root atom to all other atoms are chosen in order. (d) An example of complete random metrization designed to forestall determination (and thereby increase the 'randomness' of the embedded structure) by avoiding the completion of columns or rows of the matrix for as long as possible. The root atom is changed after each distance is set.

point, allowing the resmoothing to be done very quickly. Only when the root atom changes does the bound-smoothing procedure take a significant amount of computer time. Both the present distance geometry package and the DISGEO program use this trick to reduce the bound-smoothing time during metrization.

As shown in the results, the order in which these root atoms are chosen has a great influence on the final conformation produced. The original implementation in DISGEO picks its root atoms beginning at one terminus and working across to the other (Fig. 2a), resulting in a skewed distribution of structures as was shown by Havel (1990). Randomizing the order in which the root atoms are picked (Fig. 2b) results in better conformational sampling (Havel, 1990). The implementation in X-PLOR/dg can use ordered metrization or random metrization.

Partial metrization

We propose here a novel metrization protocol to reduce the computer time while maintaining the good sampling properties of random metrization. Schlitter (1987) showed that the complete set of all interatom distances in the metric matrix is highly redundant. Knowing the distances from four atoms to all other atoms in the molecule is, at least in principle, sufficient to constrain the molecule into one three-dimensional conformation or its mirror image. That is, knowing the distances from four atoms to all others is equivalent to knowing the distances from all atoms to all others. Therefore, if the distances from four root atoms to all the others are set and the bounds matrices are resmoothed after all these choices, then the gap between the upper and lower bounds on all remaining interatom distances should be zero. In practice, this performance is not achieved, since consistency at the triangle inequality level is a necessary, but not sufficient, condition for three-dimensional embeddability (Havel et al., 1983). In the new protocol proposed here, which we refer to as 'partial metrization', the bounds matrices are resmoothed while choosing the distances from only a few root atoms, after which distances are chosen between their upper and lower bounds without resmoothing. Partial metrization protocols are described by the percentage of root atoms with bound-smoothing after each distance pick (e.g., '10% metrization' indicates that the bounds are resmoothed while setting the distances from 10% of the root atoms to all others, after which the remaining distance bounds are smoothed and distances are chosen between their upper and lower bounds without intervening resmoothing of the bounds). 'Four-atom metrization' is a special case of partial metrization in which the bounds are resmoothed only while picking the distances from four (selected) atoms to all others.

Embedding

The matrix of chosen interatomic distances is then converted to a metric matrix M where the metric matrix element M_{ij} is defined as

$$M_{ij} \equiv \tfrac{1}{2}(D(i,O)^2 + D(j,O)^2 - D(i,j)^2) \tag{1}$$

where $D(i,O)$ denotes the distance from atom i to the collective centroid of all atoms, and $D(i,j)$ the distance from atom i to atom j (Crippen and Havel, 1988). If the three largest eigenvalues of this matrix are all positive, then their corresponding eigenvectors give the Cartesian coordinates directly. If the three largest eigenvalues are not positive, then this set of trial distances cannot be embedded and new distances must be chosen. In the course of this work, the three largest eigenvalues have always been positive (data not shown).

Partial Embedding

Even using partial metrization, the computer time needed to embed relatively large structures can be considerable. Therefore, we have also implemented a partial embedding protocol within X-PLOR/dg. By performing initial bound smoothing on all atoms in the system but (partial) metrization and embedding on only a subset of them, the effective size of the embedding problem can be reduced considerably. Because the entire distance matrix is smoothed first, the distances chosen during metrization are guaranteed to be consistent with all the input information. The missing atoms are placed randomly around the embedded atoms after embedding but before regularization (Nilges et al., 1988). The effect of this protocol on the final structures' conformational sampling is discussed below.

TABLE I
MINIMIZATION PROTOCOL USED TO REFINE ALANINE 30MER AND NP-5 STRUCTURES[a]

Stage 1: 200 steps conjugate gradient minimization
- Bond term except disulphides (k_{bond} = 1000 kcal/mol Å^2)
- NOE term (k_{NOE} = 100 kcal/mol Å^2 if the interatom distance is outside the range d $-d_{minus}$ — d $+d_{plus}$, 0 otherwise)
- Repulsive nonbonded term except for atoms that are bonded to each other with van der Waals radii multiplied by 0.9 and k_{vdW} set to 4 kcal/mol Å^2

Stage 2: 200 steps conjugate gradient minimization
- Bond angles term added (k_{angle} = 500 kcal/mol rad^2), except for disulfides

Stage 3: 200 steps conjugate gradient minimization
- Improper dihedral term for chiral and planar groups added (k_{impr} = 500 kcal/mol rad^2)
- k_{angle} is reduced to 100 kcal/mol rad^2
- Only van der Waals interactions between atoms which are not bonded or share a common bonded atom to each other are included (radii are now standard size and k_{vdW} is decreased to 0.001 kcal/mol Å^2)

Stage 4: 300 steps conjugate gradient minimization
- k_{angle} is increased to 500 kcal/mol rad^2

Stage 5: 1000 steps conjugate gradient minimization
- Disulphide bond terms added ($k_{ss\text{-}bond}$ = 100 kcal/mol Å^2)
- van der Waals standard radii are multiplied by 0.7 and k_{vdW} is increased to 2 kcal/mol Å^2

Stage 6: 1000 steps conjugate gradient minimization
- k_{vdW} is increased to 4 kcal/mol Å^2

Stage 7: 1000 steps conjugate gradient minimization
- van der Waals standard radii are multiplied by 0.8

[a] Each step includes all the energy terms included in previous steps unless otherwise noted.

Scaling

The computation of the metric matrix allows the expected radius of gyration to be calculated before embedding. This is useful as the embedded structures' radii of gyration are often 10–20% smaller than the expected value (data not shown), presumably because of remaining inconsistencies with higher-order inequalities. Therefore, following a suggestion by T. Havel (personal communication), we scale the embedded coordinates by the ratio of the expected-to-actual radii of gyration. The scaling was expected to make subsequent energy minimization better behaved. As shown in the Results section this turned out to be unnecessary.

Regularization and refinement

The embedded three-dimensional structures have very poor geometry requiring extensive regularization. Special precautions have to be taken to avoid numerical instabilities during the regularization process. We have developed a multi-stage minimization protocol with a variable target function to regularize the structure (Table 1). In order to keep the final structures close to the embedded structure, the energies of the distance geometry structures were minimized using the

TABLE 2
SIMULATED ANNEALING PROTOCOL DEVELOPED BY NILGES ET AL. (1988)[a]

Stage 1: simulated annealing at 1000 K, 50 cycles of 75 fs molecular dynamics
- $k_{bond} = 1000$ kcal/mol Å^2
- $k_{NOE} = 100$ kcal/mol Å^2 if the interatom distance is outside the range $d-d_{minus} - d+d_{plus}$, 0 otherwise
- $k_{angle} = 500$ kcal/mol rad^2
- $k_{impr} = 500$ kcal/mol rad^2
- time step 1 fs
- van der Waals radii are at standard values
- van der Waals interactions are calculated only between atoms that are not bonded to each other or are bonded to a common third atom
- k_{vdW} starts at 0.001 kcal/mol Å^2 and is scaled each cycle by a factor of 1.125, ending at 0.25 kcal/mol Å^2
- $T_{bath} = 1000$ K

Stage 2: slow cooling to 300 K, 28 cycles of 50 fs molecular dynamics
- time step 1 fs
- van der Waals radii are scaled by 0.8
- k_{vdW} set to 2.0 kcal/mol Å^2
- T_{bath} starts at 1000 K and is reduced each cycle by 25 K, ending at 300 K

Stage 3: energy minimization
- 800 steps conjugate gradient minimization

[a] Each step includes all energy terms included in the previous step unless otherwise noted.

method by Powell (1977) implemented in X-PLOR. It should be noted that the minimization steps could be replaced by short molecular dynamics runs in order to increase the radius of convergence. The protocol uses a geometric force field (Nilges et al., 1988; Brünger, 1991) and the independent adjustments of individual energy terms is close in spirit to the simulated annealing protocol developed by Nilges et al. (1988). Since distance space cannot contain chirality information, this protocol has to be repeated for both mirror images of the embedded three-dimensional structure and the final structure is chosen to be the one with the the lower minimized energy. It should be noted that in a well-defined case, this energy-based chirality test could be replaced by an r.m.s. difference test to a reference structure with correct chirality. The regularized structures may require further refinement by simulated annealing. The protocol used for refinement of the NP-5 distance geometry structures is detailed in Table 2.

RESULTS AND DISCUSSION

Conformational sampling of unconstrained alanine 30mers

The effect of various metrization protocols on their structures' conformational sampling was investigated using unconstrained L-alanine 30mers. Distance geometry structures of the L-alanine 30mers were created using ordered metrization, random metrization, and no metrization. In addition, two sets of substructures were created by performing bound smoothing on the complete matrix, and metrization and embedding only for a subset of atoms. The first substructure consisted of the C, N, C^α and C^β atoms, the second only of the C^α atoms. The missing atoms were simply added for each residue in random positions around the C^α atoms. In each case, 100 struc-

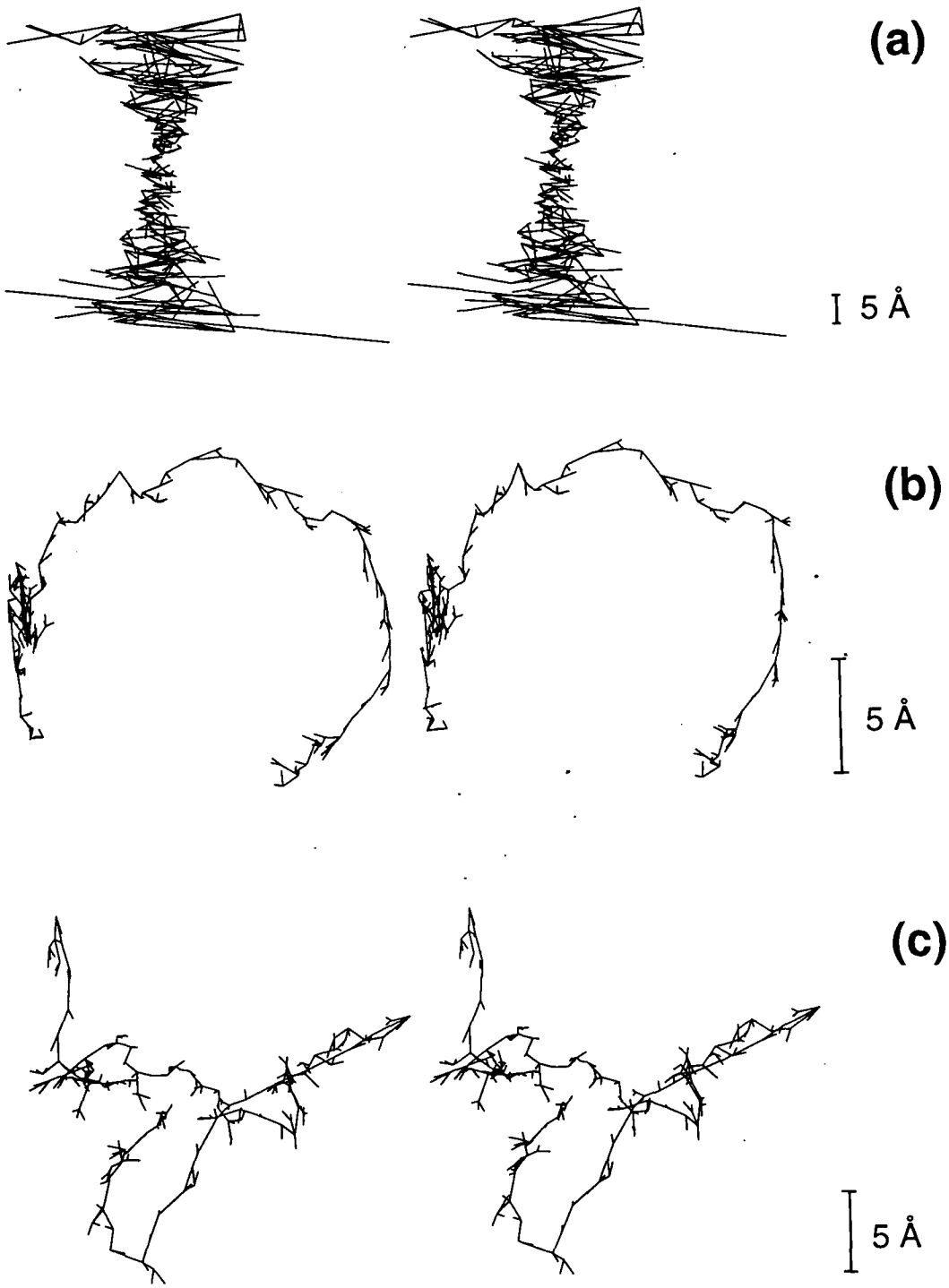


Fig. 3. Typical alanine 30mer structures embedded with various distance geometry protocols before energy minimization. (a) No metrization. (b) Complete ordered metrization. (c) Complete random metrization.

TABLE 3
COMPARISON OF UNCONSTRAINED POLYPEPTIDE BACKBONE RMSDs REPORTED IN METZLER ET AL. (1989), HAVEL (1990), AND THE PRESENT STUDY

	Metzler et al. (Lys-Glu) 12mer	Havel Ala 20mer	Havel Ala 40mer	Present study Ala 30mer
No metrization No minimization	9.55	4.0	–	14.15
No metrization With minimization	3.24	3.38	4.18	4.36
Ordered metrization With minimization	–	5.43	8.62	8.30
Random metrization With minimization	–	7.89	–	13.29
Random metrization With minimization C, N, C^{α}, C^{β} embedded	–	–	–	12.16
Random metrization With minimization C^{α} embedded	–	–	–	9.2
Simulated annealing	5.43	–	–	10.84

Note that Havel's values have been denormalized and take only C^{α} positions into account. The values given for Havel's random φ/ψ assignment are from his data set DM-1.

tures were generated using different initial random number seeds for the distance assignments prior to embedding and, in the case of the random metrization structures, for the selection of root atoms as well. The distance data used by the distance geometry algorithm comprised the geometry (bond length, angles, planarity, and van der Waals repulsion) of the polypeptide chain with the peptide bonds in the *trans* conformation. All these structures were regularized as described in Table 1. The SA structures were annealed and minimized as described in Table 2. To compare the sampling properties of distance geometry and simulated annealing, another set of 100 L-alanine 30mer structures was created by an SA protocol developed by Nilges et al. (1991) starting from an extended strand. Different choices of the initial random number seed resulted in different initial velocities, which produced a large variation of the SA structures.

Figure 3 shows typical embedded distance geometry alanine 30mer structures before minimization. Without metrization (Fig. 3a), our structures are very distorted and appear to be similar to those published in Metzler et al. (1989). Ordered metrization starting from the N-terminus (Fig. 3b) creates embedded structures that are too compact at the N-terminus and too elongated at the C-terminus. Embedded structures produced by random metrization (Fig. 3c) do not have that problem, as was suggested by Havel (1990). The conformational sampling properties of the dis-

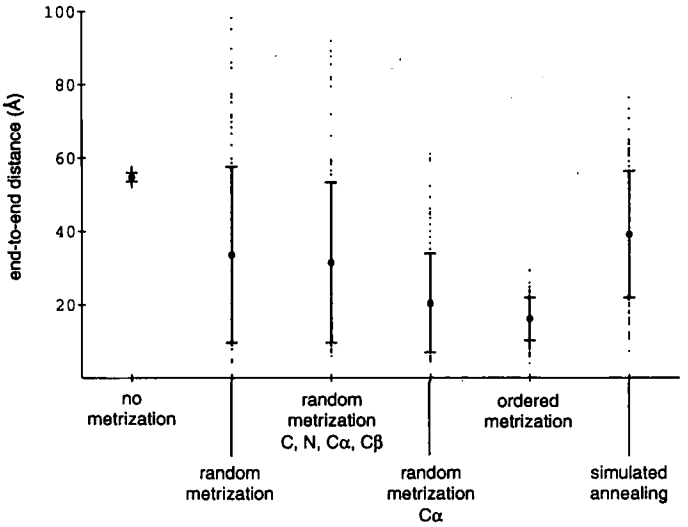


Fig. 4. End-to-end distances of structures produced with various protocols. The end-to-end distance of each structure is shown by a dot. Bars show the mean and standard deviation of the end-to-end distances produced by each protocol. (1) no metrization + conjugate gradient minimization. (2) random 100% metrization + conjugate gradient minimization. (3) random 100% metrization of a substructure (C, N, C^{α} and C^{β} atoms) + conjugate gradient minimization. (4) random 100% metrization of a substructure (C^{α} atoms) + conjugate gradient minimization. (5) ordered 100% metrization + conjugate gradient minimization. (6) simulated annealing using the protocol developed by Nilges et al. (1991). For (3) and (4), the missing atoms were added in ramdom positions before minimization. Note that the end-to-end distance for a fully-extended alanine 30mer (i.e., $\phi = -139°$, $\psi = 135°$ for all residues) is 100.3 Å.

tance geometry and SA protocols are shown in Table 3 and in Figs. 4 and 5. Table 3 shows the mean backbone RMSD; Fig. 4 gives the distribution of the end-to-end distances; and Fig. 5 shows Ramachandran plots for the simulated annealing and the three different metrization protocols.

The structures embedded without metrization have poor sampling in Cartesian space, as shown by their small backbone RMSDs (Table 3) and tight distribution of end-to-end distances (Fig. 4), and in φ/ψ space, as shown by their heavy overpopulation of backbone conformations near the center of the Ramachandran plot (Fig. 5c).

The ordered metrization structures tend to be too compact at their N-termini, presumably because the N-terminal region contains the initial root atoms for ordered metrization. This is carried through to the minimized structures, as shown by their short end-to-end distances (Fig. 4). This also explains their small backbone RMSDs after minimization (Table 3), even though their sampling in φ/ψ space, judging from their Ramachandran plots (Fig. 5b), is very broad.

Random metrization produces a wide range of structures, as shown by their end-to-end distances (Fig. 4). The presence of a relatively large number of greatly extended structures helps to explain the slight overpopulation of points near the extended limits of the Ramachandran plot (Fig. 5a) and their very large backbone RMSDs (Table 3). Partial embedding of substructures after random metrization reduces the conformational sampling (Table 3 and Fig. 4). This effect is most pronounced for the substructures that consist only of the C^{α} atoms. However, even in this case the sampling is better than for ordered metrization.

Simulated annealing starting from an extended strand ($\varphi = -139°, \psi = 135°$) produces struc-

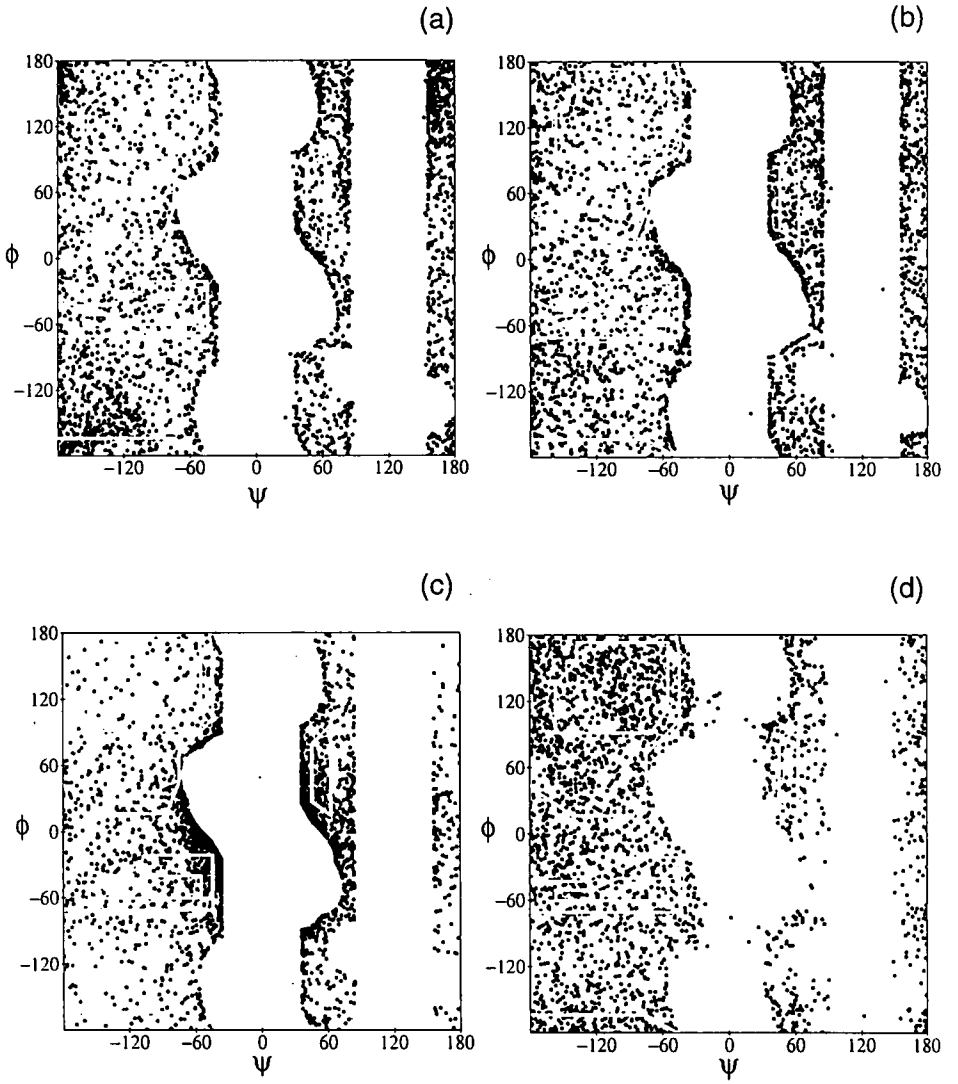


Fig. 5. Ramachandran plots for structures created with various protocols. 100 alanine 30mer conformations were calculated with each of the protocols described below and their Ramachandran plots superimposed. (a) Random 100% metrization + conjugate gradient minimization. (b) Ordered 100% metrization + conjugate gradient minimization. (c) No metrization + conjugate gradient minimization. (d) Simulated annealing using the protocol from Nilges et al. (1991).

tures whose average end-to-end distance is about the same as in the case of random metrization, but it does not produce the extremely extended or compact structures that random metrization can (Fig. 4). The Ramachandran plot (Fig. 5d) shows fewer structures existing in the regions with positive φ than those for the distance geometry calculations. The allowed regions are more uniformly sampled than in any of the other Ramachandran plots.

A comparison of these results and those reported in Metzler et al. (1989) and Havel (1990) is given in Table 3. Note that the very large backbone RMSDs among nonminimized structures

TABLE 4
NP-5 STRUCTURES CALCULATED WITH DISTANCE GEOMETRY

(a) Percentage of successful NP-5 structures among 200 distance geometry calculations. Successful is defined as no NOE violations being greater than 0.5 Å and the deviations of bond lengths and bond angles from ideality being less than 0.015 Å and 3°, respectively (using the parameters defined in Table 1).

	Minimization only	+ Simulated annealing
No metrization	5%	76%
Ordered 100% metrization	16%	54%
Random 100% metrization	24%	56%

(b) Percentage of the successful NP-5 structures that are within 4 Å of the standard fold or the best pseudo-mirror-image fold MC-6 described in Levy et al. (1989). All successful NP-5 structures that are more than 4 Å away from either fold are grouped together under 'other'.

	Minimization only			+ Simulated annealing		
	Standard	Pseudo-mirror-image	Other	Standard	Pseudo-mirror-image	Other
No metrization	100%	0%	0%	95%	2%	3%
Ordered 100% metrization	48%	10%	42%	66%	5%	29%
Random 100% metrization	60%	10%	30%	68%	12%	21%

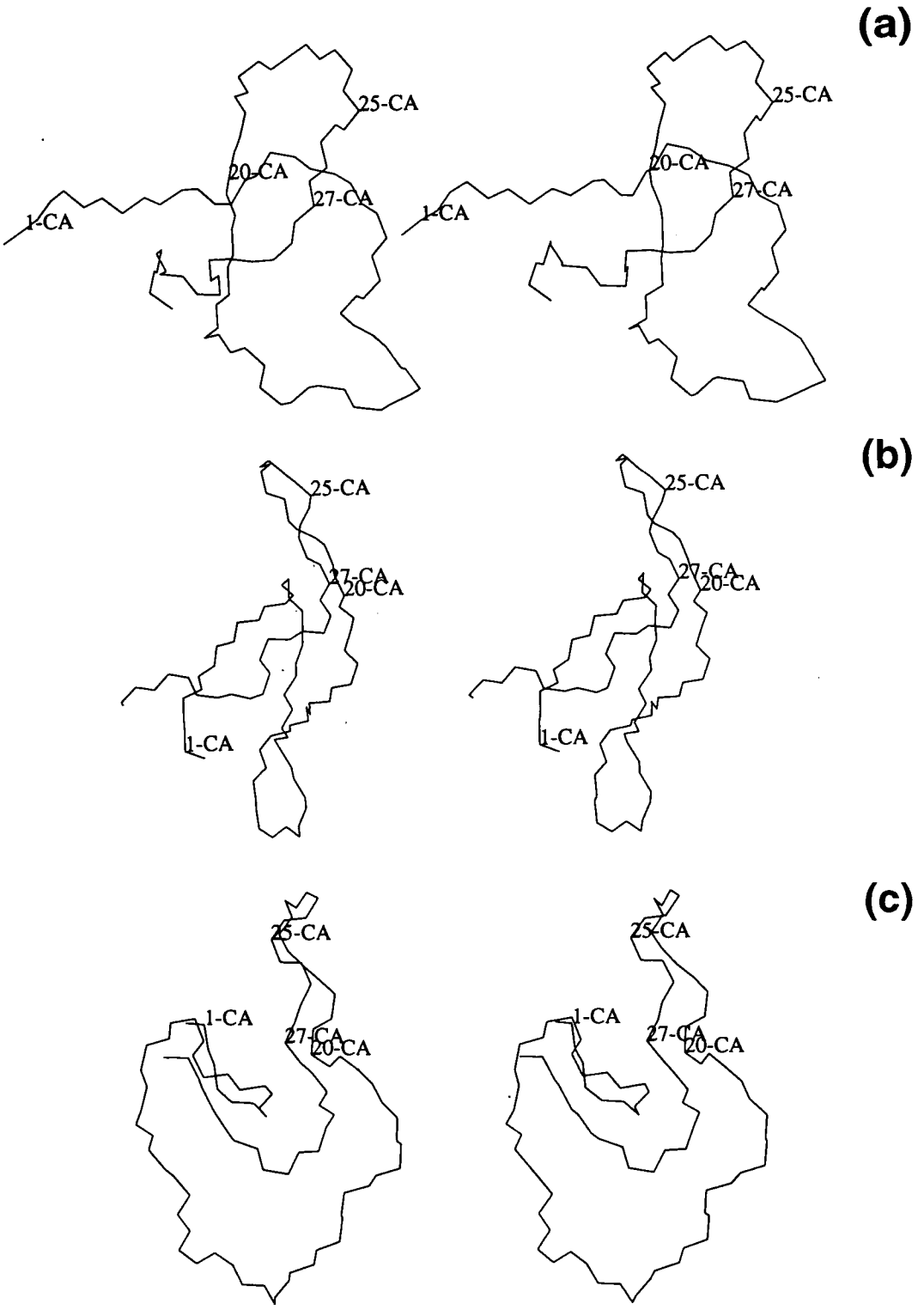
without metrization found by Metzler et al. (1989) are confirmed here and, in particular, the small C^{α} RMSD for nonminimized structures reported by Havel (1990) remains an unexplained curiosity.

To test the effect of scaling the structures (see Methods section), 100 alanine 30mers were embedded with complete ordered metrization and another 100 structures with complete random metrization. The structures were not scaled and were subsequently minimized using the same protocol as the scaled embedded structures. Their backbone RMSDs, Ramachandran plots, and distribution of their end-to-end distances were not significantly different from those with scaling (data not shown).

Conformational sampling of NP-5

In order to examine the sampling properties of distance geometry in solution NMR structure determination, structures of NP-5 were calculated using the NOE distance constraints from Pardi et al. (1988). The 93 NOEs from set A of Pardi et al., only 43 of which are nonsequential, were used, along with actual disulphide bonds between residues 3 and 31, 5 and 20, and 10 and 30. 200 NP-5 structures were embedded using complete ordered metrization and 200 structures with complete random metrization. Another 200 structures were embedded without metrization. These were minimized as described in Table 1 and subjected to a short SA protocol as described in Table 2.

The success rates of several protocols are summarized in Table 4a. The success rate of a structure-determination protocol is defined here as the number of structures in each set with no NOE-


(a)
25-CA
20-CA
27-CA
1-CA
25-CA
20-CA
27-CA
1-CA
(b)
25-CA
27-CA
20-CA
1-CA
25-CA
27-CA
20-CA
1-CA
(c)
25-CA
1-CA
27-CA
20-CA
25-CA
1-CA
27-CA
20-CA

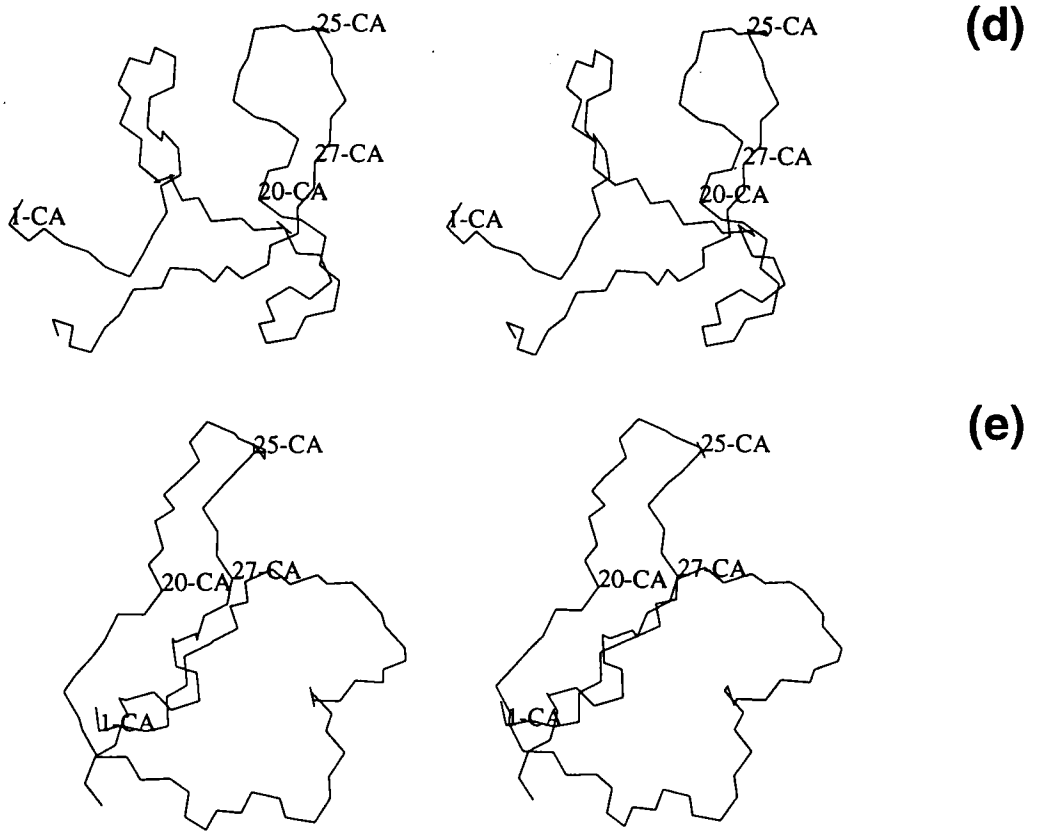


Fig. 6. NP-5 structures produced by distance geometry. The structures shown are among those successfully embedded (as defined in Table 4a) and they keep their fold virtually unaltered throughout the simulated annealing refinement. (a) The standard fold of NP-5 produced with complete random metrization and conjugate gradient minimization. (b) The pseudo-mirror-image fold of NP-5 produced with complete random metrization and conjugate gradient minimization. (c) A new NP-5 fold similar to the standard fold but with the 1–20 loop rotated up and the C-terminus caught behind the 1–20 loop. Produced with complete ordered metrization and conjugate gradient minimization. (d) A new NP-5 fold similar to the pseudo-mirror-image fold but with the 1–20 loop rotated up and away from the viewer. Produced with complete random metrization and conjugate gradient minimization. (e) A new NP-5 fold similar to the pseudo-mirror-image fold but with the 1–20 loop curved around toward the termini before entering the β sheet. Produced with complete random metrization and conjugate gradient minimization.

derived distance constraint violation greater than 0.5 Å and low deviations of the geometry from ideality (as defined in Table 1). Note that simulated annealing refinement of the minimized structures can double the success rate of conjugate gradient minimization alone.

The family of successful distance geometry structures falls into several clusters (Table 4b), where a cluster is defined as a collection of structures with an r.m.s. difference of backbone atoms of no more than 4.0 Å from the mean. The largest cluster (Fig. 6a) corresponds to the standard NP-5 fold first reported by Pardi et al. (1988). The second largest cluster (Fig. 6b) corresponds to the best pseudo-mirror-image conformation described by Levy et al. (1989). However, we observed a number of structures that fall into neither conformational cluster. These additional struc-

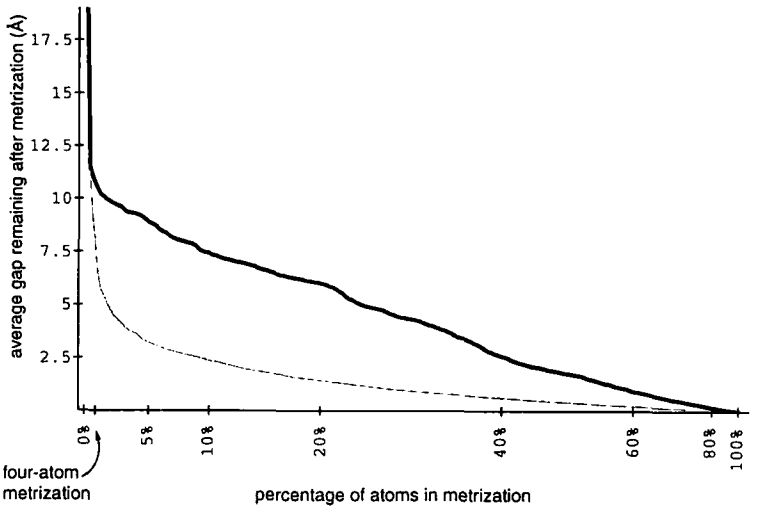


Fig. 7. Average gap left between upper and lower bounds on interatomic distances left unset after partial metrization vs. percentage of root atoms used for metrization. The thick line shows ordered metrization; the thin line shows random metrization.

tures exhibit low conformational energies and good agreement with the NOE distance constraints. For instance, Fig. 6c shows a variation on the standard fold in which the bulk of the N-terminal region is rotated by 90° relative to the β-sheet, and the C-terminus is caught behind the N-terminal loop. Figures 6d and e show similar variations on the pseudo-mirror-image structure, with the N-terminal region rotated in different ways relative to the rest of the structure.

Partial metrization and the quality of the embedded structures

In order to examine the effect of different metrization percentages on the quality of the structures obtained, NP-5 structures were calculated with partial metrization increasing from no atoms to all 472, one at a time. This process was repeated for both ordered and random metrization using the same set of random numbers for distance assignments. As shown in Fig. 7, the average gap between the upper and lower bounds on the nonmetrized distances drops extremely quickly while performing metrization on the first few atoms. It might be possible to lower the remaining gap even more quickly with judicious choices of early root atoms. The NOE template analysis presented by Hempel and Brown (1989) can identify well-defined domains of a protein from the NOESY data. Choosing early root atoms to be in different domains may help ensure that the maximum amount of information is imparted with minimal metrization.

In ordered metrization, the sets of distances from any of the first four root atoms to most of the others are very similar (since they are coming from four nearby atoms), and thus setting the distances relating to these first four atoms does not contribute much more information than setting the distances from one atom to all the others (considering the uncertainties in the distances). In random metrization, the four root atoms are scattered throughout the molecule, so the early distance choices contribute more information to determine the structure, explaining random metrization's ability to lower the gap more quickly.

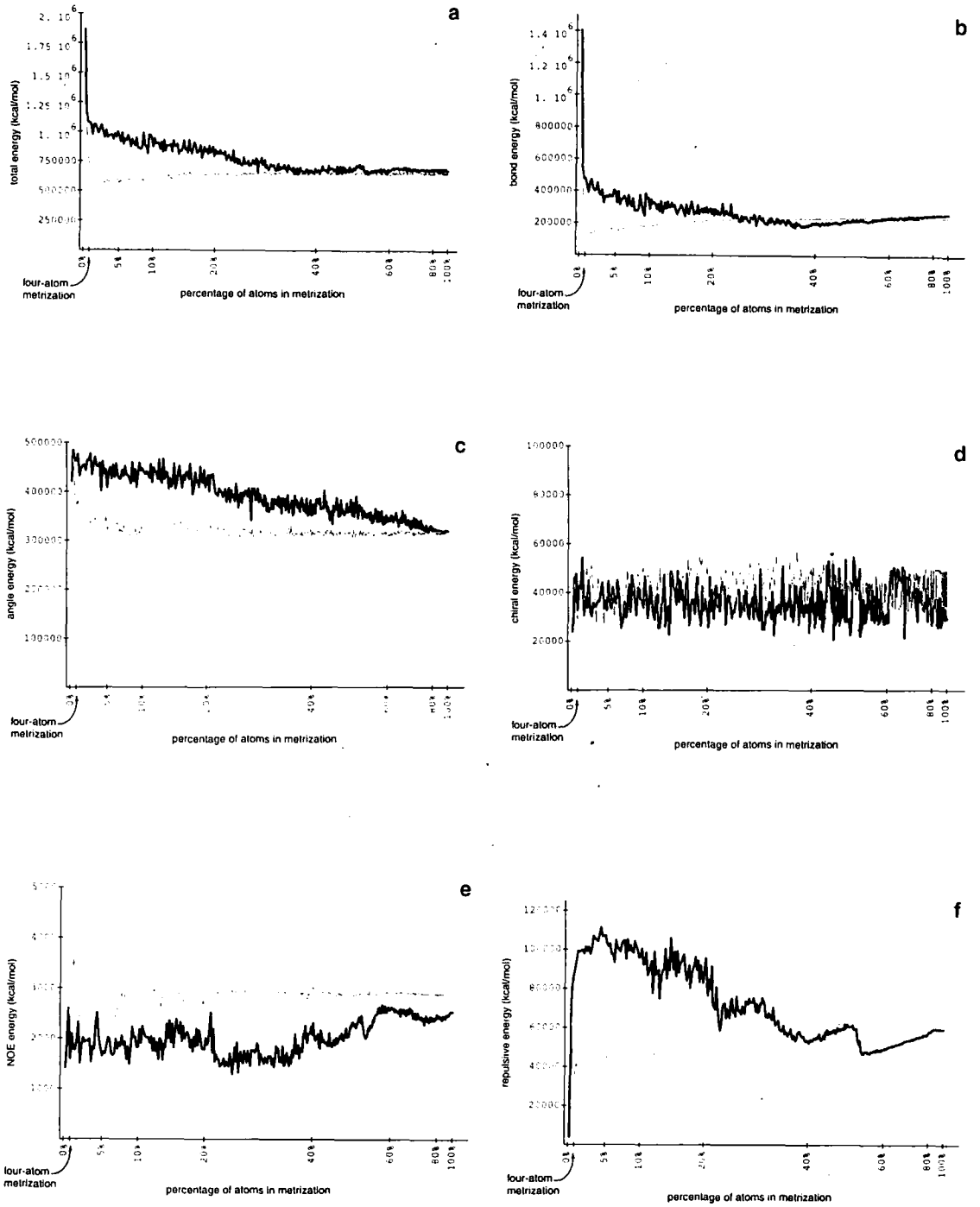


Fig. 8. Conformational energies of embedded structures vs. percentage of root atoms used for metrization. The thick lines show increasing ordered metrization; the thin lines show increasing random metrization. (a) Total energy (consisting of bond, bond angle, planarity, chirality, repulsion, and NOE energy terms). (b) Bond energy. (c) Angle energy. (d) Chiral energy. (e) NOE energy. (f) Repulsion energy.

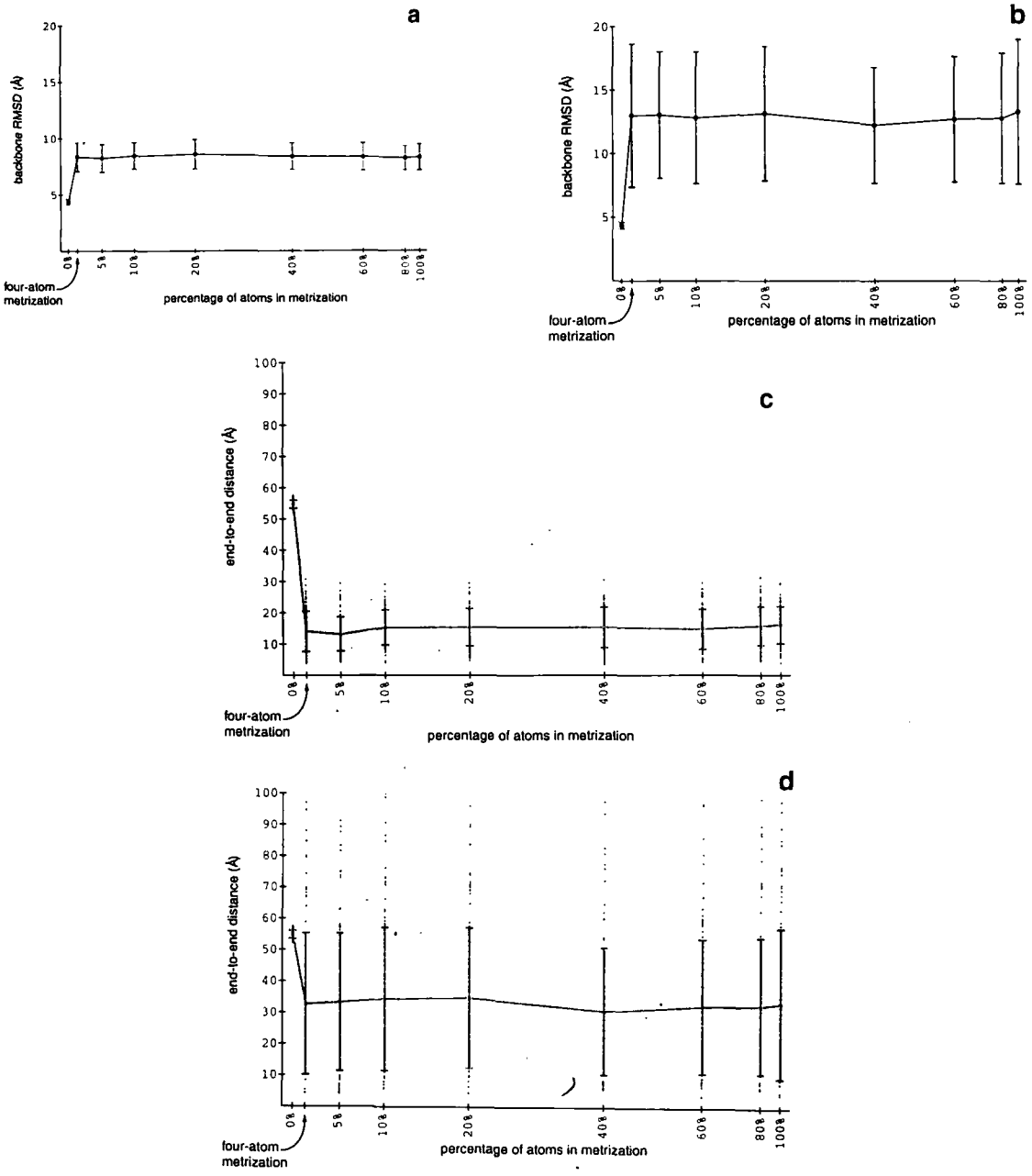


Fig. 9. Effect of increased metrization on conformational sampling of unconstrained alanine 30mers. (a) Backbone RMSD of embedded and minimized structures for increasing ordered metrization. Bars show the standard deviation of the backbone RMSD for each set of structures. (b) Backbone RMSD of embedded and minimized structures for increasing random metrization. Bars show the standard deviation of the backbone RMSD for each set of structures. (c) End-to-end distances of embedded and minimized structures for increasing ordered metrization. Bars show the mean and standard deviation of the end-to-end distances of the structures in each group. (d) End-to-end distances of embedded and minimized structures for increasing random metrization. Bars show the mean and standard deviation of the end-to-end distances of the structures in each group.

As shown in Fig. 8, the conformational energies of structures before minimization drop quickly after performing metrization on the first few root atoms. Increasing the number of atoms involved in random metrization lowers the conformational energies much more quickly than in ordered metrization (Fig. 8). Note that the chiral energy (Fig. 8d) does not change significantly with increasing ordered or random metrization, since chirality information cannot be represented in distance space.

Partial metrization and the sampling of the embedded structures

The effect of increasing the number of atoms involved in partial metrization on the conformational sampling of the structures produced was examined by calculating 100 structures at four-atom partial metrization and at 5%, 10%, 20%, 40%, 60% and 80% partial metrization for both ordered as well as random root atom sequence. As is shown in Fig. 9, increasing metrization beyond the first four root atoms has little effect on the conformational sampling after minimization, either for ordered or for random metrization. No significant differences were found in the Ramachandran plots (not shown) of these ensembles of structures with different degrees of metrization.

In order to test more stringently the influence of the first four root atoms used in metrization on the conformational sampling, 100 alanine 30mers were calculated with ordered metrization for the first four atoms and random metrization thereafter. Another 100 structures were calculated with random metrization for the first four atoms and ordered metrization for the rest. As shown in Fig. 10, the positions of the first four root atoms virtually determine the distribution of the end-to-end distances, i.e., ordered metrization for the first four points produces the same end-to-end distances as complete ordered metrization, and random metrization for the first four points produces the same end-to-end distances as complete random metrization. The average backbone RMSD and Ramachandran plots for the structures calculated with combined metrization were

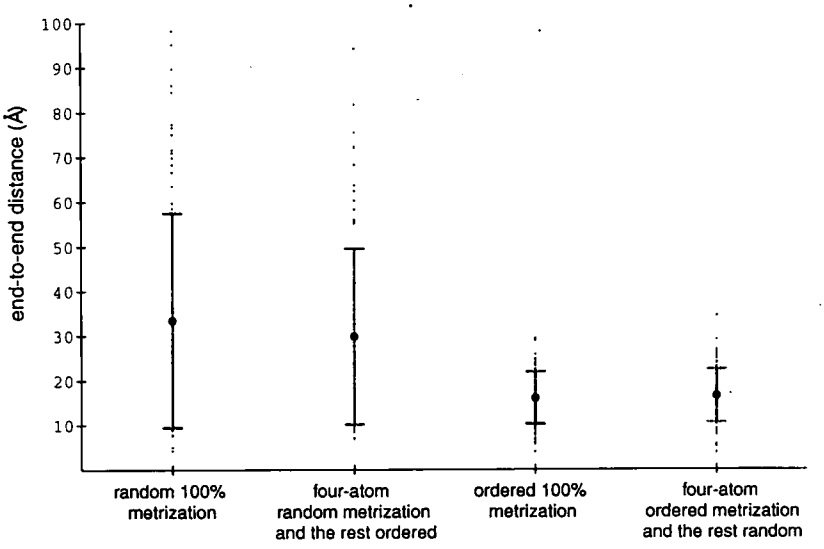


Fig. 10. Effect of combined metrization on conformational sampling of alanine 30mers. The end-to-end distance of each structure is shown by a dot. The bars show the mean and standard deviation of the structures calculated with each metrization protocol.

also not significantly different from those for structures calculated with the full metrization protocol that uses the same type of metrization as the one in the first stage of combined metrization (data not shown).

From Figs. 9 and 10 it is clear that the embedded structure's conformation is nearly completely determined after the distances from four atoms to all the others have been set. It might be possible to improve distance geometry's sampling even further by choosing interatomic distances in complete random order (Fig. 2d) that avoids setting the distances from four points to all the others for as long as possible. However, this would eliminate the performance gains made by starting most calls to the shortest-path routine with the same root as the previous call.

Effect of problem size on CPU time

In order to quantify the actual performance of X-PLOR/dg, alanine polymers with sizes varying from 100 to 1000 atoms were embedded after various metrization protocols. The actual CPU times for the metrizations and embeddings on a Convex 210 are shown in Fig. 11. To give a rough comparison of these protocols' performance to that of simulated annealing, the same alanine polymers were subjected to a simulated annealing protocol similar to that described in Nilges et al. (1991), starting from random backbone torsion angles. For the comparison between the CPU times it should be taken into account that the reported CPU times for the distance geometry calculations are for metrization/bound-smoothing and embedding only, i.e., without regularization and refinement. The CPU time for the minimization protocol described in Table 1 is also shown in the figure. An annealing refinement such as the one described in Nilges et al. (1988) takes about twice the CPU time of the minimization scheme. The CPU time for de novo simulated annealing and minimization has a nearly-linear behavior while the CPU time for distance geometry exhibits a progressive increase when the number of atoms is increased. The large difference between the times for no metrization and partial metrization is due to the fact that the complexity of the bound smoothing algorithm is dependent on the number of known distances in the molecule. It rises

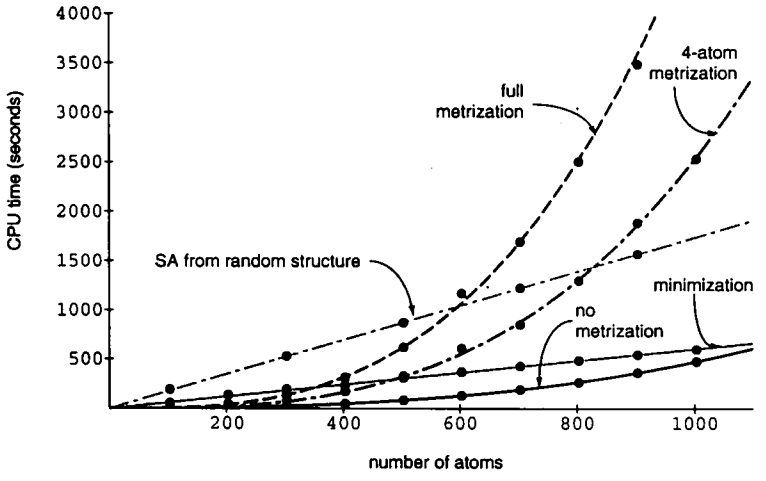


Fig. 11. CPU time for embedding various molecules vs. number of atoms in the molecule. The times are on a Convex 210.

steeply with the number of experimental distance bounds, and also with the number of distances set during the metrization. The plot indicates that in order to use distance geometry efficiently for larger systems one has to avoid embedding all the atoms, and work with substructures. The remaining atomic positions can be generated and refined as described previously (Nilges et al., 1988).

CONCLUDING REMARKS

Performing partial random metrization during the construction of a trial distance matrix before embedding into three-dimensional space has two effects. First, the conformational energies of embedded structures are significantly better than those produced without metrization. Second, the sampling of the conformational space is drastically improved, especially through randomization of the root atom order. In fact, the structures calculated by the partial random metrization protocol are as different from each other as those produced by molecular dynamics-based simulated annealing. This appears to apply to unconstrained polypeptides as well as to proteins constrained by NMR interproton distance information. While computationally the partial random metrization approach is not as efficient as inferior nonmetrization approaches, the excellent sampling properties make it a useful tool not only for NMR structure determination, but also for model building purposes. It should be noted however, that the CPU (Fig. 11) and memory requirements of distance geometry inherently limit the application of this method to small-to-medium-size molecules up to a few thousand atoms on present computer architectures.

ACKNOWLEDGEMENTS

MN acknowledges useful discussions with Dr. Timothy Havel. Support from the National Science Foundation (ATB, Grant DIR-9021975) is gratefully acknowledged.

REFERENCES

Billeter, M., Havel, T.F. and Wüthrich, K. (1986) *J. Comput. Chem.*, **8**, 132–141.

Braun, W. and Gō, N. (1985) *J. Mol. Biol.*, **186**, 611–626.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) *J. Comp. Chem.*, **4**, 187–217.

Brünger, A.T. (1991) In *Topics in Molecular Biology* (Ed., Goodfellow, J.M.) Macmillan Press Ltd., London, pp. 137–178.

Brünger, A.T. and Karplus, M. (1991) *Acc. of Chem. Res.*, **24**, 54–61.

Brünger, A.T., Clore, G.M., Gronenborn, A.M. and Karplus, M. (1987) *Protein Eng.*, **1**, 399–406.

Brünger, A.T. (1990) X-PLOR software manual version 2.1., New Haven, Yale University.

Clore, G.M. and Gronenborn, A.M. (1991) *Science*, **252**, 1390–1399.

Crippen, G.M. and Havel, T.F. (1988) *Distance Geometry and Molecular Conformation*, Research Studies Press, Taunton, Somerset, England.

Dial, R., Glover, F., Karney, D. and Klingman, D. (1979) *Networks*, **9**, 215–248.

Dijkstra, E.W. (1959) *Numer. Math.*, **1**, 269–271.

Dress, A.W.M. and Havel, T.F. (1988) *Discrete Appl. Math.*, **19**, 129–144.

Driscoll, J.R., Gabow, H.N., Shrairman, R. and Tarjan, R.E. (1988) *Comm. of the ACM*, **31**, 1343–1354.

Easthope, P.L. and Havel, T.F. (1989) *Bull. Math. Bio.*, **51**, 173–194.

Ernst, R.R., Bodenhausen, G. and Wokaun, A. (1986) *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, Clarendon Press, Oxford.

Hadwiger, M.A. and Fox, G.E. (1989) *J. Biomol. Struct. Dyn.*, **7**, 749–771.
Hare, D.R. and Reid, B.R. (1986) *Biochemistry*, **25**, 5341–5350.
Havel, T.F., Kuntz, I.D. and Crippen, G.M. (1983) *Bull. Math. Bio.*, **45**, 665–720; (1985) *errata in Bull. Math. Bio.*, **47**, 157.
Havel, T.F. and Wüthrich, K. (1984) *Bull. Math. Bio.*, **46**, 673–698.
Havel, T.F. (1990) *Biopolymers*, **29**, 1565–1585.
Hempel, J.C. and Brown, F.K. (1989) *J. Am. Chem. Soc.*, **111**, 7323–7327.
Kuntz, I.D., Crippen, G.M. and Kollman, P.A. (1979) *Biopolymers*, **18**, 939–957.
Levy, R.M., Bassolino, D.A., Kitchen, D.B. and Pardi, A. (1989) *Biochemistry*, **28**, 9361–9372.
Metzler, W.J., Hare, D.R. and Pardi, A. (1989) *Biochemistry*, **28**, 7045–7052.
Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988) *FEBS Lett.*, **229**, 317–324.
Nilges, M., Kuszewski, J. and Brünger, A.T. (1991) In *Computational Aspects of the Study of Biological Macromolecules* (Ed., Hoch, J.C.) Plenum Press, New York.
Pardi, A., Hare, D.R., Selsted, M.E., Morrison, R.D., Bassolino, D.A. and Bach, A.C. (1988) *J. Mol. Biol.*, **201**, 625–636.
Powell, M.J.D. (1977) *Mathematical Programming*, **12**, 241–254.
Schlitter, J. (1987) *J. Appl. Math. Physics (ZAMP)*, **38**, 1–9.
Tarjan, R.E. (1983) *Data Structures and Network Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia.
Thomason, J.F. and Kuntz, I.D. (1989) *J. Cell Biochem., Suppl.* **13A**, no. 37.
Wüthrich, K., Billeter, M. and Braun, W. (1983) *J. Mol. Biol.*, **169**, 949–961.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York.