The sampling properties of some distance geometry algorithms applied to unconstrained polypeptide chains: A study of 1830 independently computed conformations

ARTICLE *in* BIOPOLYMERS · OCTOBER 1990

Impact Factor: 2.39 · DOI: 10.1002/bip.360291207 · Source: PubMed

1 AUTHOR:

Timothy F. Havel
Energy Compression Inc.
120 PUBLICATIONS **7,020** CITATIONS

SEE PROFILE

The Sampling Properties of Some Distance Geometry Algorithms Applied to Unconstrained Polypeptide Chains: A Study of 1830 Independently Computed Conformations

TIMOTHY F. HAVEL

Division of Biophysics, University of Michigan, 2200 Bonisteel Boulevard, Ann Arbor, Michigan 48109

SYNOPSIS

In this paper we study the statistical geometry of ensembles of poly(L-alanine) conformations computed by several different distance geometry algorithms. Since basic theory only permits us to predict the statistical properties of such ensembles a priori when the distance constraints have a very simple form, the only constraints used for these calculations are those necessary to obtain reasonable bond lengths and angles, together with a lack of short- and long-range atomic overlaps. The geometric properties studied include the squared end-to-end distance and radius of gyration of the computed conformations, in addition to the usual rms coordinate and ϕ/ψ angle deviations among these conformations. The distance geometry algorithms evaluated include several variations of the well-known embed algorithm, together with optimizations of the torsion angles using the ellipsoid and variable target function algorithms.

The conclusions may be summarized as follows: First, the distribution with which the trial distances are chosen in most implementations of the embed algorithm is not appropriate when no long-range upper bounds on the distances are present, because it leads to unjustifiably expanded conformations. Second, chosing the trial distances independently of one another leads to a lack of variation in the degree of expansion, which in turn produces a relatively low rms square coordinate difference among the members of the ensemble. Third, when short-range steric constraints are present, torsion angle optimizations that start from conformations obtained by choosing their ϕ/ψ angles randomly with a uniform distribution between $-180°$ and $+180°$ do not converge to conformations whose angles are uniformly distributed over the sterically allowed regions of the ϕ/ψ plane.

Finally, in an appendix we show how the sampling obtained with the embed algorithm can be substantially improved upon by the proper application of existing methodology.

INTRODUCTION

The computation of biomolecular conformation from two-dimensional nmr data is usually accomplished by means of what are known as distance geometry algorithms[1-4] followed by various refinement procedures based on potential energy functions[5-8] (see Refs. 9 and 10 for reviews). Some comparisons of the various procedures available have been attempted,[5,11] but at this time the methodology is evolving at such a rapid pace as to render these sorts of comparisons obsolete almost before they are in press. One general observation that has been made, however, is that distance geometry algorithms that generate their starting conformations by "distance space" methods, i.e., the embed algorithm[12,13] and its many variations, produce ensembles of conformations that are relatively expanded and whose rms coordinate deviations (RMSD) are substantially smaller than those produced by "torsion space" refinements. The subjective impression of similarity one gets from looking at the resultant superimposed conformations has lead several investigators to conclude that the embed algorithm intrinsically produces a "biased" sampling of conformation space.

The purpose of this paper is to report some com-

CCC 0006-3525/90/12-131565-21 $04.00
Biopolymers, Vol. 29, 1565–1585 (1990)

putational experiments we have performed in order to evaluate the situation in an objective way, to determine the reasons behind it, and to show how they can be largely eliminated. For our evaluation, we make use of the statistical properties of polypeptide chains in the random coil state that have been most extensively studied in the past,[14] in addition to the usual RMSD. Since the statistical behavior of most of the geometric properties of polymer chains can be calculated from basic theory only with very simple types of distance constraints, the only constraints present in these calculations are those necessary to achieve a lack of unacceptable short and/or long-range atomic overlaps, in addition to reasonable bond lengths and angles. By calculating these same geometric properties for ensembles of polyalanine as computed by both torsion space as well as distance space algorithms, we are able to quantitatively access the nature of the sampling obtained with the various alternatives, at least with such simplified constraints. In an appendix, we describe some simple modifications to the "classical" embed algorithm and show that they can significantly improve its sampling properties.

After the computations reported in the main part of this paper were completed, we learned that a paper on the sampling properties of distance space methods with biopolymers had been submitted.[15] The authors of that paper concentrate upon the average RMSD with respect to the main-chain N, C^{α}, and C′ atoms in their computed structures, together with the average unsquared end-to-end distances. Their observations for these quantities are basically in accord with those reported here, and they have also shown that similar results are obtained with nucleic acids. They did not study the sampling obtained with torsion space methods as well as several established improvements upon the basic distance space approach, however, and their conclusions are based upon the observation of a relatively limited variety of statistical properties, which also differ from those for which theoretical predictions are currently available.

GEOMETRIC PROPERTIES AND THEIR STATISTICS

The statistics of two different kinds of geometrical properties were evaluated for each ensemble of computed conformations. The first type are intrinsic properties of single conformations within the ensembles, while the second type are measures of the dissimilarity between pairs of conformations within the ensembles. In this section we describe these geometric properties in detail, along with what is known about their statistical behavior.

Chain Dimensions

The theory of polymer statistical mechanics, as developed over the last 50 years in large part by Paul Flory and co-workers,[16,17] concentrates on those geometric properties of polymer chains that are amenable to theoretical calculation as well as to experimental verification. Predominant among these are the mean square distance between the ends of the polymer chain

$$\langle (\mathbf{r}_1 - r_N)^2 \rangle = \int (\mathbf{r}_1 - \mathbf{r}_N)^2 f(\mathbf{r}_1, \ldots, \mathbf{r}_N) \times d\mathbf{r}_1 \cdots d\mathbf{r}_N$$

where $f(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ is the configurational probability density function[18] and the mean square radius of gyration

$$\langle R_G^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} \int (\mathbf{r}_i - \mathbf{r}_0)^2 f(\mathbf{r}_1, \ldots, \mathbf{r}_N) \times d\mathbf{r}_1 \cdots d\mathbf{r}_N$$

where $\mathbf{r}_0 = N^{-1} \sum_{i=1}^{N} \mathbf{r}_i$ is the centroid of the configuration (see Refs. 14 and 19 for accounts).

Since $f = Z^{-1}\exp(-E/kT)$ is given by a Boltzman distribution in the configuration dependent potential energy E, the exact values of these quantities of course depend upon this energy function, and in particular upon the strength of the interactions of the polymer chain with itself relative to the strength of its interactions with the surrounding solvent. The asymptotic dependence of the mean square end-to-end distance upon the chain length $N-1$, however, is known in three important limiting cases:

1. In poor solvents in which the attractive interactions of the chain with itself are predominant, polymers exist in a "molten globule" state in which simple dimensional arguments show that $\langle (\mathbf{r}_1 - \mathbf{r}_N)^2 \rangle \sim N^{\frac{2}{3}}$ (where "$\sim$" means "proportional to").
2. In good solvents in which the repulsive interactions of the chain with itself predominate, it has been shown that $\langle (\mathbf{r}_1 - \mathbf{r}_N)^2 \rangle \sim N^{\frac{6}{5}}$.[20,21] This expansion of the coil is known as the *excluded volume effect.*

3. In a Θ *solvent* in which repulsive and attractive forces are exactly balanced, the coil behaves like a random walk through space subject to the constraints imposed by the "short-range" interactions, i.e., chain connectivity, bond angle rigidity, etc. In this case the end-to-end vector has an asymptotic Gaussian distribution whose expected mean square value is $\langle (\mathbf{r}_1 - \mathbf{r}_N)^2 \rangle \sim N$.

Finally, in the limit as $N \to \infty$, the mean square end-to-end distance and the mean square radius of gyration are related simply by $\langle (\mathbf{r}_1 - \mathbf{r}_N)^2 \rangle = 6 \cdot \langle R_G^2 \rangle$. See Ref. 22 for a recent discussion.

The constants of proportionality as well as the exact values of the end-to-end distance and/or radius of gyration for finite values of N can be computed from the basic theory only in case 3, in which the long-range interactions can be neglected.[14,23] These calculations are also open to direct experimental verification, since the radius of gyration can be determined by scattering experiments and Θ conditions can be obtained by adjusting solvent mixtures until the second osmotic virial coefficient vanishes. Surprisingly, the calculations are actually easier for polypeptides than they are for most polymers.[24,25] This is a consequence of the fact that, due to the large separation between adjacent amino acid residues imposed by *trans* peptide bonds, the average total energy due to short-range interactions can be written as a sum of contributions depending only on the conformational state of each residue.

In particular, for all-*trans* polyglycine and polyalanine chains whose conformational states can be characterized by means of their ϕ and ψ angles alone,[26]

$$E_{\text{short}}(\phi_1, \psi_1, \ldots, \phi_N, \psi_N) = \sum_{i=1}^{N} e(\phi_i, \psi_i)$$

where e is the potential energy of a single residue in the chain expressed as a function of its ϕ and ψ angles. Thus, the average state of each residue can be used to compute the average dimensions of the polypeptide, regarded as a chain of virtual bonds between α-carbons. Extensive studies on polyglycine, *D* and *L* polyalanine as well as copolymers thereof have been carried out by Brant, Miller, and Flory.[25-29] In order to obtain a dimensionless quantity less dependent on chain length, in this work they have concentrated on the *characteristic ratio* obtained by dividing the mean square end-to-end distance by the value expected for an unconstrained freely jointed random chain of links equal to the virtual bond length of 3.80 Å, i.e.,

$$\sigma^2(N) = \frac{\langle (\mathbf{r}_1 - \mathbf{r}_N)^2 \rangle}{3.80^2(N-1)}$$

Similarly, we report the squared radius of gyration normalized by the value expected for a freely jointed chain

$$\tau^2(N) = \frac{6 \cdot \langle R_G^2 \rangle}{3.80^2(N-1)}$$

If $e = 0$, i.e., if one has a uniform distribution of ϕ and ψ angles, then the limit $\sigma^2(\infty) = 1.93$ Å.[25] Due to its symmetry, this value is close to the experimental value of $\sigma^2(\infty)$ for polyglycine as well as to the value calculated when the short-range potential e includes all the usual torsional, van der Waals and electrostatic interactions.[28,29] The situation for poly(L-alanine) is drastically different, where both the calculated and observed values of $\sigma^2(\infty)$ are 9 ± 1. Moreover, unlike polyglycine, the calculated value depends strongly upon the dielectric constant, and declines to only 3 when electrostatic terms are neglected completely.[29] Earlier results[25] indicate that $\sigma^2(\infty)$ is only moderately affected by the neglect of the attractive van der Waals interactions, and is about 4 when both electrostatic and attractive van der Waals interactions are ignored. This last case is probably the one most directly comparable to the geometric calculations reported here. Finally, we note that Brant and Flory's results imply that the asymptotic limit of the characteristic ratio σ^2 is attained substantially faster than the corresponding ratio τ^2 for mean square radius of gyration, and is essentially constant when the number of residues $N \geq 40$.

In this work we also evaluate the second moments or standard deviations of the squared end-to-end distances and squared radii of gyration. For a more detailed examination of the distribution of the end-to-end vectors, one cannot use a laboratory frame because the vector averages to zero when the orientation of the molecule as a whole is random. Therefore one uses a frame whose x axis is directed along the first virtual bond, whose y axis lies in the plane of the first two virtual bonds and such that the second virtual bond has a positive y component, and whose z axis complete a right-hand frame. Unlike the mean square end-to-end distance, the end-to-end vector in this molecule frame, known as the *persistence vector,* has a finite average asymptotic

value,[30] which we attempted to estimate for our ensembles. The detailed distribution of persistence vectors and their squares is relatively difficult to calculate from first principles even in the absence of long-range interactions,[31] although they can be estimated by Monte Carlo techniques.[32-34] The size of the ensembles that could be generated with the methods described in this paper, however, were not large enough for us to obtain a meaningful estimate of the distribution of the persistence vector.

Dissimilarity Measures

The most commonly used measure of dissimilarity between two different conformations of the same biopolymer is the RMSD, which appears to have been introduced by Ref. 35. This is given by

$$\mathrm{RMSD} = \min_{\boldsymbol{R}} \left(\frac{1}{N} \sum_{i}^{N} \| \boldsymbol{R} \cdot \mathbf{p}_i - \mathbf{p}'_i \|^2 \right)^{1/2}$$

where N is the total number of atoms, $\mathbf{p}_i$ and $\mathbf{p}'_i$ denote the atomic coordinates of the ith atom of each of the two structures in a common center-of-mass coordinate system, and the minimum is taken over all 3×3 rotation matrices $\boldsymbol{R}$. Although this measure is always nonnegative and is zero if and only if the conformations are identical, the present author knows of no proof that it has even the most rudimentary property that is generally regarded as necessary for a well-behaved dissimilarity measure, namely that it fulfills the triangle inequality. Nevertheless, because of its sensitivity to the overall handedness of the conformation, it is generally preferred to the distance matrix error

$$\mathrm{DME} = \left(\frac{2}{N(N-1)} \sum_{i<j}^{N} (d_{ij} - d'_{ij})^2 \right)^{1/2}$$

which not only obeys the triangle inequality but is actually an $N(N-1)/2$-dimensional Euclidean distance.[13] As long as the mirror image of $\mathbf{p}$ with least RMSD is used, these two measures appear to have a good linear correlation.[36] In this paper, we consider only the RMSD calculated with respect to the C^α atom positions alone, using the method of McLachlan.[37]

Another measure of dissimilarity between pairs of conformations, which appears to first have been used in Ref. 38, is the root mean square difference between their torsion angles:

$$\mathrm{DHAD} = \left(\frac{1}{M} \sum_{i}^{M} (\vartheta - \vartheta'_i)^2 \right)^{1/2}$$

where M is the number of torsion angles in question, and by "difference" we mean the size in degrees of the smaller of the two bond rotations that could be used to make the ϑ angles equal. This measure can be shown to obey the triangle inequality. In this paper, all DHADs reported are computed with respect to the ϕ and ψ angles of the polypeptide conformations being compared. Assuming ideal covalent geometry, a zero DHAD between two polypeptide conformations implies a zero C^α RMSD difference, but the converse is false since there exists a two-dimensional infinity of polypeptide backbone conformations consistent with any feasible choice of C^α coordinates.[39] However, even if the RMSD is computed between all N, C^α, and C′ backbone atoms, it has been shown that these two measures are at best only weakly correlated.[11,38] Indeed, it has long been known that small changes in the dihedral angles of a long chain can lead to drastic changes in the Cartesian coordinates, whereas the above-cited distance geometry calculations have shown that conformations whose Cartesian coordinates are very similar can have substantial differences in their dihedral angles. Plainly, there is more than one way in which two conformations can "differ" from one another!

Although the expected value of the RMSD is difficult to calculate from an assumed distribution on the underlying conformation space, theoretical analysis leads to the conclusion that for a flexible random chain in the absence of long-range interactions it should be asymptotically equal to $L\sqrt{N-1}$, where $N-1$ is the number of bonds in the chain and L is their "effective" length[40]. For a freely jointed chain of 3.8 Å virtual bonds between consecutive α-carbons, the effective bond length has been found by Monte Carlo studies to be 1.34 Å. In an actual polypeptide chain wherein the virtual bond angles must lie between ca. 90° and 140°, this constant would be significantly larger, and has been estimated to be about 2 Å. The Monte Carlo studies also indicate that the ratio $\rho(N) = \mathrm{RMSD}/(1.34\sqrt{N-1})$ is already at 80% of its limiting value when the chain length is only 10, and that the actual distribution of RMSD is strongly skewed towards lower values.

The average DHAD between a pair of conformations obtained by making a uniform random selection of their ϕ and ψ angles is relatively easy to compute. Since we may without loss of generality use the angle selected in the first conformation as our orgin, we get

$$\left(\frac{1}{180} \int_0^{180} \vartheta^2 d\vartheta \right)^{1/2} = (180^2/3)^{1/2} = 103.9°$$

The average in polypeptide random coils should be lower than this, because even in a random coil the angles are not uniformly distributed but are concentrated in sterically allowed regions of the Ramachandran map.[26] Nevertheless, we can standardize our DHAD by dividing through by this value, i.e., $\delta(N) = \mathrm{DHAD}/103.9$.

Finally, we note that in our own empirical estimates of the mean RMSD and DHAD differences, we took the average over all $\binom{n}{2}$ pairs of conformations within each n-member ensemble, rather than dividing the ensemble up into $n/2$ disjoint pairs and averaging over just these pairs. The correlation between the differences for two pairs that have a conformation in common (e.g., $\{a, b\}$ and $\{b, c\}$) will affect the standard deviations of the differences, of course, but the law of large numbers tells us that the averages still constitute unbiased estimates of the means, and the averages are all that we report below.

DISTANCE GEOMETRY CALCULATIONS

In order to maximize the relevance of this study to ongoing protein structure determinations, all of the computer programs implementing the various distance geometry algorithm variants evaluated in this study were taken "off the shelf" and used without modification. These distance geometry algorithm variants were obtained from two fundamentally different approaches to computing polypeptide conformations consistent with distance constraints. The first set of variants are based on the *embed algorithm*,[12,13] which is classed as a "distance space" method because it uses an estimate of the distances in a possible conformation of the molecule in order to obtain a starting structure for refinement by means of a matrix method known as *embedding*. The refinement itself proceeds via a routine numerical minimization of the distance constraint violations, usually using the conjugate gradient method, and with the atomic Cartesian coordinates as the variables.

The second set of variants are torsion space methods that attempt to perform a global optimization using the torsion angles as their variables. Here two different optimization algorithms were used: the *ellipsoid algorithm*[1] and the *variable target function algorithm*.[3] In both cases, the starting structure for the optimization is obtained from an estimate of the torsion angles in a possible conformation of the molecule, and the optimization consists of a minimizing the distance constraint violations with respect to the torsion angles. The ellipsoid algorithm is a general purpose global optimization algorithm that proceeds by generating a sequence of ellipsoids of decreasing volume, each of which contains an optimum solution.[41] The variable target function method, on the other hand, was developed specifically for the purpose of computing protein conformations from distance constraints,[3] and proceeds by minimizing the violations of distance constraints between pairs of atoms that are separated by successively longer segments of polypeptide chain. As in the embed algorithm, these successive minimizations are accomplished by means of standard descent algorithms. Although the ellipsoid and variable target function algorithms differ substantially from one another, we shall see that the statistical behavior of the resultant conformations is generally dominated by the distribution of starting structures.

Because it is a relatively simple structure whose statistical geometry has been studied under a variety of assumed probability distributions (see above references), chains of poly(L-alanine) were used for these calculations. For each set of distance constraints and each distance geometry program variant tested, we calculated ensembles of 50, 25, and 15 conformations for poly(L-alanine) chains 20, 40, and 60 residues in length, respectively.

Distance Constraints

To simplify the calculations, the methyl group of each alanine residue was replaced by a single "pseudoatom" located at the centroid of its three hydrogen atoms (with the exception of those calculations made using the DISMAN program,[3] which implements the variable target function method, as will be described below). In all cases the constraints included the bond lengths, the geminal distances that define the bond angles, and the distances across the peptide bond that fix it in a *trans* configuration (explicitly in the distance space methods, implicitly in the torsion space methods). In addition, the distance space methods used "chirality constraints" to maintain the correct configuration at the α-carbons as well as good covalent geometry overall (as described in Refs. 2 and 12).

The sets of distance constraints tested were as follows:

A. No constraints at all beyond those necessary to fix the covalent geometry of the molecule, as described above. In this case, for reasons given in the discussion section, one expects to obtain a uniform sampling in torsion angle space.

B. The covalent geometry constraints together with lower bounds on the "short-range" nonbonded and nongeminal distances between pairs of atoms that were either in the same residue or else in one or both of the immediately adjacent peptide bonds (as described in Ref. 25), which were obtained as the sum of suitable hard-sphere radii. This situation corresponds roughly to what would be expected in a Θ solvent where only short-range interactions have any net effect.
C. The covalent geometry constraints together with lower bounds between all nonbonded, nongeminal pairs derived from the same hard-sphere radii as used in constraint set B. These constraints should correspond roughly to what would be expected in a good high dielectric solvent in which the only effective interaction is steric repulsion.

The radii were chosen so as to obtain lower bounds that all lay between the Ramachandran normal and extreme limits[26] (see Table I), except with the DISMAN program, where they were chosen so as to give a van der Waals energy at contact of 3 kcal/mole[3] (see below).

Table I Hard Sphere Radii (Å) Used in All Computations

Atom Type	DISMAN	DISGEO and ELLIPSE
H	0.95	0.95
O	1.30	1.35
N	1.20	1.35
C	1.40	1.45
Me	—	1.70

Calculations with the Embed Algorithm

An implementation of the embed algorithm known as the DISGEO program was used for these computations,[2] which has been extensively used in the calculation of protein structure from two-dimensional nmr data.[8,42-44] This program employs several specialized techniques for improving convergence in calculations on large proteins, and it was of particular interest to determine the effect of these upon the sampling properties of the program.

The first of these techniques involves the calculation of intermediate "substructures," containing an evenly distributed subset of the atoms in the complete molecule, from which complete structures are calculated by means of a second embedding and refinement.[2] Since the substructures necessarily take account of only a subset of the information available for the complete molecule, one would not expect the complete structures calculated from them to be distributed in the same way as they would be if all the information was used together in a single embedding. In particular, the atoms of the substructure are treated differently from the rest of the atoms in the complete structure (by being embedded first), and hence one would expect that there would be some differences in the way in which they tend to be positioned by this process.

Another important technique, known as "metrization," is used to estimate interatomic distances in such a manner that the resulting "trial distances" (from which the starting coordinates are calculated) obey the triangle inequality.[2] It is known that this technique does not produce a uniform sampling with respect to the "natural" (i.e., Lebesgue[45]) probability distribution on the set of all such metrics, and in addition, the result depends upon the order in which the trial distances are selected in the course of the metrization procedure. Because it made it relatively easy to implement, the DISGEO program uses a fixed order in which distances from the N-terminal atoms are selected first, and it is to be expected that the conformation of the N-terminal end of the molecule will not be distributed in the same way as the C-terminal end as a consequence.

To test these expectations, the following embed algorithm variants were used:

1. Substructures were calculated without metrization, and complete structures were calculated from these with metrization. This is the usual procedure employed with the DISGEO program when computing the conformations of proteins from nmr data.[2]
2. Complete structures were calculated directly from the constraints present in the given data via a single embedding using neither metrization nor intervening substructures. This is essentially the "classical" embed algorithm[13] in which we chose the trial distances as independent random variables from within the triangle inequality smoothed bounds.
3. Complete structures were calculated directly from the constraints present in the given data via a single embedding using metrization but with no intervening substructures.

In every case, the magnitude of the maximal hard-sphere violations after refinement were almost always less than 0.1 Å, and better than 95% of the violations were less than 0.05 Å. The bond lengths were usually within 0.01 Å of their standard values, and the bond angles within $\pm 5°$.

At this point it should be mentioned that the random number generator used by the DISGEO program is based upon the multiplicative congruential method,[46] and is not one of the best. Because the triangle inequality smoothed distance bounds generally encompase a range that is much wider than the range of distances that actually can occur in solution (see Ref. 12 for further discussion), eight of these random numbers are generated, summed, and divided by eight in order to obtain a roughly Gaussian distribution within the bounds on the individual trial distances. Given this fact, together with the complicated sequence of transformations to which the trial distances are subjected in the course of embedding, it is unlikely at best that significant differences would have been obtained by employing a better random number generator. In order to be able to see the differences in the results obtained with each of the constraints A, B, and C above, the same random number generator seed was used with each, while different seeds were used with different program variants.

Calculations with Torsion Space Methods

A new implementation of the ellipsoid algorithm, henceforth called the ELLIPSE program, was used for these calculations. It differs from that reported in Ref. 1 primarily in the way in which it handles the constraints. Rather than looking for a single significantly violated dihedral angle, upper distance bound, or hard-sphere constraint, the program sums the violations of each of these three classes of constraints. If the sum of the dihedral angle violations exceeds a given threshold, the gradient of this sum is computed and used in the next step of the refinement. Otherwise, if the sum of the distance bound violations exceeds a second threshold, the gradient of this sum is added to the gradient of the dihedral angle violations and the combined gradient used in the next step. Otherwise, if the sum of the hard-sphere violations exceeds a third threshold, we use the gradient of this sum plus the gradient of the sum of the torsion angle and distance-bound violations. Finally, if all three summed violations are below their thresholds, all three thresholds are lowered by some given fraction, and the process restarted from there. This new procedure has not been thoroughly compared to the previous one, but the results obtained so far indicate that it is at least as good. In any case, none of the constraint sets evaluated here (A–C above) involve any explicit torsion angle constraints.

Starting conformations for the refinement were generated according to two different strategies:

1. In the first, the ϕ and ψ angles were chosen with a uniform distribution in the interval $[-180°, +180°]$.
2. In the second, the ϕ and ψ angles were chosen with a uniform distribution within the extreme limits given by Ramachandran and Sasisekharan.[26]

The random number generator used to generate these starting conformations was the standard UNIX™ library function "random." In both cases, statistics were computed for ensembles of conformations generated by these two procedures as well as for the result of refining such ensembles vs the constraint sets B and C above. Conformations were only accepted if the hard-sphere constraints used were satisfied about as well as they were in the conformations obtained from the DISGEO program, and of course the covalent geometry was perfectly standard. Because of the high attrition rate experienced in case 1 above, it was not practical to perform these calculations on the 40- and 60-residue polyalanine chains.

The variable target function calculations were made using the DISMAN program,[3] which has also been used extensively for the calculation of protein conformations from nmr data.[47–49] As with the DISGEO and ELLIPSE calculations, the only constraints used were the lower bounds needed to avoid hard-sphere overlaps, and the calculations were performed in complete accord with the established procedures for this program.[11] Because of the way this program is written, however, it was not practical to use it in a fashion perfectly analogous to that described above for DISGEO and ELLIPSE. In the first place, a full methyl group had to be used for these calculations rather than a united methyl pseudoatom. In the second, the hard-sphere radii used by the program are built into it, and hence could not be made exactly equal to those used by DISGEO and ELLIPSE. Finally, it was not possible to input specific lower bound constraints, which made it impossible to take account of only those hard-sphere overlaps classified as "short range" by Brant and Flory.[25] Fortunately, since the first stage of the variable target function method looks at only those hard-sphere overlaps between pairs of atoms in adjacent amino acid residues, it was possible to selectively account for the majority of short-range pairs simply by terminating the optimization after this stage. These small differences in procedure certainly had some small effects upon the results obtained, but these effects are predictable so that with a little care the results obtained with the DISMAN

program remain directly comparable to those obtained with the other programs.

RESULTS

The results of our computations are summarized in Tables II–IV, wherein the ensemble designations "DG-n([ABC])," "EL-n([ABC]),"and "DM-n([ABC])" stand for constraint set A, B, or C of variant n = 1, 2, or 3 of the DISGEO, ELLIPSE, and DISMAN programs, respectively (the numbers and letters here all correspond with those used in the previous section). In the 20-residue case, the averages of ρ = RMSD/$(1.34\sqrt{N-1})$, δ = DHAD/103.9, as well as the characteristic ratio σ^2, the normalized squared radius of gyration τ^2, and their standard deviations were computed over subsets of the total ensembles of sizes 10, 20, 30, 40, and 50 (complete definitions of these terms may be found in the geometric properties section). From 20 on, the values were usually equal to within 0.1, indicating that our averages have thoroughly converged. At the same time, the fluctuations in the components of the persistence vector X, Y, and Z remained substantial, especially in the Z component; hence we have rounded their values to the nearest Ångstrom, since greater precision is certainly not significant.

Looking at the values of σ^2 and τ^2 for 20-poly(L-alanine) (Table II), we see that the conformations computed by the embed algorithm with and without intervening substructures (DG-1 and -2) tend to be substantially more expanded than those computed by torsion space methods (EL-1, EL-2, and DM-1). The most compact conformations, however, were obtained from the embed algorithm when metrization was used (DG-3). An equally pronounced difference may also be observed in the percent standard deviations Δ of σ^2 and τ^2, which are much lower for the embed algorithm than they are for the torsion space optimizations, especially when no metrization is employed.

On comparing these numbers with the corresponding numbers obtained for the 40- and 60-residue chains (Tables III and IV), we see that although the values of σ^2 and τ^2 are comparable in the ensembles EL-1, EL-2, and DM-1 (as they are for a free jointed chain), for the ensembles DG-1 and DG-2 their ratio appears to be about 2 in the limit as the number of residues $N \rightarrow \infty$. It may also be seen that both σ^2 and τ^2 appear (on the basis of the three points available) to increase linearly with N in all cases. When metrization is used, (DG-3) τ^2 actually exceeds σ^2 substantially, and moreover, their values appear to depend only weakly upon N in these ensembles. The lack of monotonicity in the values of

Table II[a] Average Geometric Properties of Ensembles of 20-Poly(L-Alanine)

Ensemble	ρ	δ	$\sigma^2 \pm \Delta\%$	$\tau^2 \pm \Delta\%$	[X, Y, Z]
DG-1(A)	0.60	0.96	4.35 ± 10	2.58 ± 4	[17, 15, −2]
DG-1(B)	0.62	0.97	4.65 ± 9	2.77 ± 5	[23, 12, 1]
DG-1(C)	0.63	0.99	5.21 ± 9	3.28 ± 4	[20, 16, −2]
DG-2(A)	0.55	0.92	4.70 ± 7	2.66 ± 4	[23, 16, −3]
DG-2(B)	0.58	0.97	4.91 ± 8	2.83 ± 5	[22, 17, −4]
DG-2(C)	0.58	0.98	5.46 ± 8	3.31 ± 4	[22, 17, −4]
DG-3(A)	0.87	0.97	0.54 ± 40	1.13 ± 12	[0, 0, −1]
DG-3(B)	0.89	0.98	0.52 ± 36	1.20 ± 13	[0, 0, −1]
DG-3(C)	0.93	0.99	0.55 ± 45	1.27 ± 12	[0, 0, 0]
EL-1(A)	1.20	1.00	1.64 ± 64	1.84 ± 39	[5, 5, −2]
EL-1(B)	1.19	0.99	1.80 ± 77	1.87 ± 48	[2, 7, 2]
EL-1(C)	1.09	0.99	2.24 ± 63	2.11 ± 31	[6, 8, −2]
EL-2(A)	1.19	0.88	3.16 ± 59	2.61 ± 36	[10, 7, 1]
EL-2(B)	1.24	0.88	2.41 ± 82	2.23 ± 44	[9, 7, 0]
EL-3(C)	1.15	0.87	2.88 ± 56	2.57 ± 33	[12, 4, 0]
DM-1(A)	1.15	0.99	1.95 ± 66	1.87 ± 40	[8, 6, 5]
DM-1(B)	1.19	0.99	1.85 ± 70	1.82 ± 41	[7, 4, 2]
DM-1(C)	1.14	1.00	2.19 ± 67	2.03 ± 37	[8, 3, 3]

[a] The ensemble designators XY-n ([ABC]) are described at the beginning of the results section; ρ is the normalized RMSD, δ is the normalized dihedral angle difference, σ^2 is the characteristic ratio, τ^2 is the normalized radius of gyration squared, and [X, Y, Z] is the persistence vector (see geometric properties section).

Table III[a] Average Geometric Properties of Ensembles of 40-Poly(L-Alanine)

Ensemble	ρ	δ	$\sigma^2 \pm \Delta\%$	$\tau^2 \pm \Delta\%$	$[X, Y, Z]$
DG-1(A)	0.52	0.96	8.48 ± 6	4.61 ± 3	[26, 26, 6]
DG-1(B)	0.54	0.97	8.71 ± 7	4.77 ± 3	[31, 22, −20]
DG-1(C)	0.54	0.96	9.27 ± 5	5.11 ± 3	[33, 25, −7]
DG-2(A)	0.45	0.91	9.03 ± 4	4.76 ± 2	[40, 34, −2]
DG-2(B)	0.49	0.97	9.10 ± 4	4.84 ± 2	[40, 36, 2]
DG-2(C)	0.50	0.97	9.66 ± 5	5.25 ± 2	[37, 37, 7]
DG-3(A)	0.84	0.95	0.45 ± 43	1.08 ± 10	[1, 2, −1]
DG-3(B)	0.91	0.98	0.45 ± 41	1.15 ± 11	[0, 0, 0]
DG-3(C)	1.03	1.00	0.45 ± 59	1.28 ± 10	[2, 0, −1]
DM-1(A)	1.25	1.00	1.60 ± 80	1.84 ± 41	[7, 4, 3]
DM-1(B)	1.23	1.00	2.03 ± 79	1.93 ± 54	[10, 5, 4]
DM-1(C)	1.21	1.00	2.21 ± 58	2.26 ± 38	[5, 6, 5]

[a] All symbols are defined as in Table II.

σ^2 and τ^2 with increasing N that is observed in DM-1 is presumably due to statistical fluctuations in the smaller ensembles.

The components of the persistence vectors are relatively difficult to interpret, but do at least give some information about shape in addition to overall size. In almost all cases the X and Y components are nonnegative, the only exception being the 60-residue DISMAN calculation DM-1 (which may be a statistical fluctuation). This is in accord with previous calculations.[34] Even in the 50 conformation ensembles computed for $N = 20$, the Z component could change substantially on taking out only 10 conformations, indicating that it had not converged to a well-defined average value.

Turning now to the RMSDs observed for 20-poly(L-alanine), we see that the RMSD of ensembles DG-1 and DG-2 runs about 60% of what one would expect for a freely jointed chain of the corresponding length (i.e. $\rho \approx 0.6$), and increases to about 90% of this extreme when metrization is also used (DG-3). There is also a slight tendency for ρ to increase on refinement with respect to the steric constraints [DG-n(A), (B), and (C)]. The ensembles obtained from the torsion space optimizations (EL-1, EL-2, and DM-1), on the other hand, all have $\rho \approx 1.2$. Thus previous observations that torsion space optimizations produce ensembles with a higher RMSD is confirmed. On comparison with the corresponding values of ρ in Tables III and IV, we also see that these differences become more pronounced as the length of the chain increases.

Oddly enough, the very large RMSD reported in Ref. 15 among the *unrefined* starting conformations

Table IV[a] Average Geometric Properties of Ensembles of 60-Poly(L-Alanine)

Ensemble	ρ	δ	$\sigma^2 \pm \Delta\%$	$\tau^2 \pm \Delta\%$	$[X, Y, Z]$
DG-1(A)	0.45	0.96	12.90 ± 4	6.80 ± 2	[33, 34, 36]
DG-1(B)	0.48	0.97	13.00 ± 6	6.87 ± 2	[16, 54, 5]
DG-1(C)	0.50	0.98	13.56 ± 6	7.30 ± 3	[14, 42, 23]
DG-2(A)	0.41	0.92	13.45 ± 3	6.95 ± 1	[60, 32, 3]
DG-2(B)	0.44	0.97	13.51 ± 2	7.01 ± 1	[56, 31, −7]
DG-2(C)	0.45	0.99	14.16 ± 2	7.39 ± 1	[69, 35, 3]
DG-3(A)	0.89	0.92	0.45 ± 35	1.05 ± 10	[2, 2, −1]
DG-3(B)	0.92	0.95	0.45 ± 41	1.10 ± 10	[4, 3, −2]
DG-3(C)	1.09	0.99	0.45 ± 56	1.25 ± 10	[−1, 3, −5]
DM-1(A)	1.36	1.00	2.18 ± 68	2.18 ± 53	[4, 8, −13]
DM-1(B)	1.30	0.99	1.23 ± 77	1.69 ± 53	[−4, 6, −1]
DM-1(C)	1.38	1.00	1.93 ± 113	2.15 ± 53	[−4, 2, −12]

[a] All symbols are defined as in Table II.

produced by the embed algorithm in the absence of substructures and metrization was substantially less in the corresponding calculations reported here (i.e., runs DG-2). In our case the RMSD among these starting conformations was never more than a factor of two higher than the RMSD after refinement (when metrization was used, i.e. DG-3, the RMSD before refinement actually tended to be slightly *less* than it was afterward). A direct comparison of our results with theirs is rendered a bit difficult by the fact that the lengths of our polymer chains are not the same, and by their use of all main-chain atoms to compute the RMSD instead of just the C^{α} atoms as we have done here. The RMSD of 17.8 Å that they report for a 24-mer of the copolymer $(\text{Lys-Glu})_n$ still seems surprising in view of our value of only 4.0 Å for the C^{α} atoms of the starting conformations in the DG-2 runs with 20-poly(L-alanine) (data not shown in tables). This discrepancy may have something to do with our use of a Gaussian distribution for our trial distances, instead of a uniform distribution. It may also in some way be connected to the presence of the long side chains in their copolymer.

In no case save possibly EL-2 is the overall scatter in the ϕ/ψ angles far from being random ($\delta = 1$). A more demanding test can be obtained by examining the distribution of the residues in the ϕ/ψ plane (see Figures 1 and 2). Here we can see that with the embed algorithm and no metrization (DG-1 and DG-2), there is a tendency for those conformations computed with no steric constraints present to cluster in the center of the ϕ/ψ plane, and that although refinement with respect to the steric constraints pushes them out of this region, it does not spread them uniformly over the sterically allowed regions of the map. This clustering was substantially less in the unrefined starting conformations (data not shown), but was nevertheless noticeable after refinement even when no chirality constraints were used. The presence of significant numbers of residues in partially disallowed regions of the ϕ/ψ plane, which may be observed in DG-1, 2, and 3, is due to the flexibility in the covalent structure which is present in those calculations.

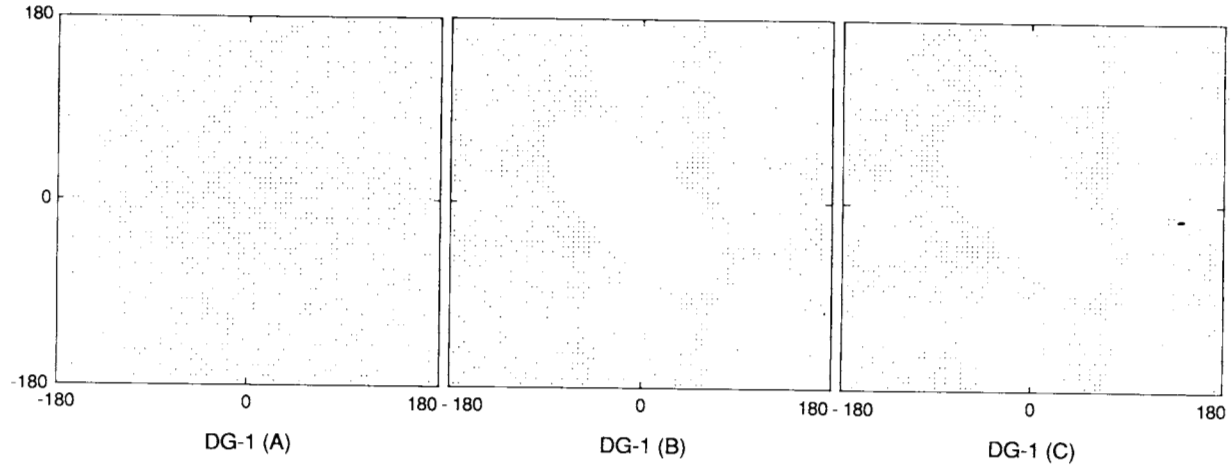
In the torsion angle optimizations, of course, the scatter in the ϕ/ψ angles is uniform by construction when no steric constraints are present [Figure 2, EL-1(A), EL-2(A), and DM-1(A)]. When such constraints are used, however, the residues are no longer distributed uniformly over sterically allowed regions of the ϕ/ψ plane, but tend to be denser near the borders of these regions [EL-1, -2, and DM-1, (B) and (C)]. In addition, the relatively small left and right-handed helical regions are more densely populated than is the relatively large β-sheet region. These tendencies are observed regardless of the optimization algorithm used. Scatter plots were also prepared for the 40- and 60-residue computations, but since these did not differ in any essential way from the 20-residue runs they have not been included here.

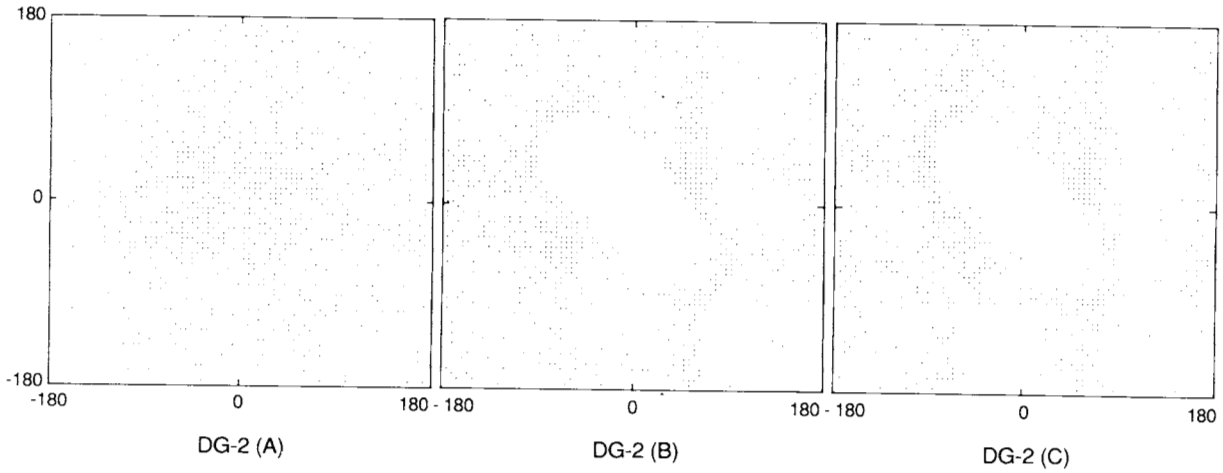
One other thing that perhaps needs some clarification is that in EL-2(B) many residues occur with ψ in the partially forbidden region of $+60°$, where the β-methyl overlaps the carbonyl oxygen of the succeeding residue. This occurs because the starting conformations, when selected from within the extreme limits of the Ramachandran diagram, did not have to be changed by more than 11° in DHAD before the total violation fell below the desired threshold (although the RMSD between the starting and final conformations was sometimes as large as 8!). Probably, these small residual violations could have been eliminated by using a larger starting ellipsoid and correspondingly larger amount of computer time.

DISCUSSION

One way to model geometric constraints is to use an "energy" function that is zero for any conformation that satisfies the constraints and infinite elsewhere. The fact that the distribution of states in an isolated physical system is that which maximizes its entropy[50] then gives us a well-defined probability density function on conformation space that adds no information to the system beyond that intrinsically expressed by the distance and chirality constraints themselves. For the purposes of obtaining an understanding of the consequences of the constraints for the overall geometry of the molecule, the optimal strategy is probably to sample randomly with respect to this *thermodynamic* distribution. At least it is the only probability density that is justified by the available geometric data alone, and whenever we use the term *random* below, we mean random relative to this density function.

Unfortunately, no analytic formula or even numerical method for estimating this probability density at any given point of conformation space is currently available. One can call upon Bernouli's "principle of insufficient reason" to argue that, when expressed as a function of the torsion angles, the probability density will be constant within the region of conformation space allowed by the distance and chirality constraints, and zero elsewhere. It is somewhat surprising that at the time of writing the author knows of no proof that this actually characterizes the thermodynamic density function. It is not hard to show, however, that for a freely jointed two-di-





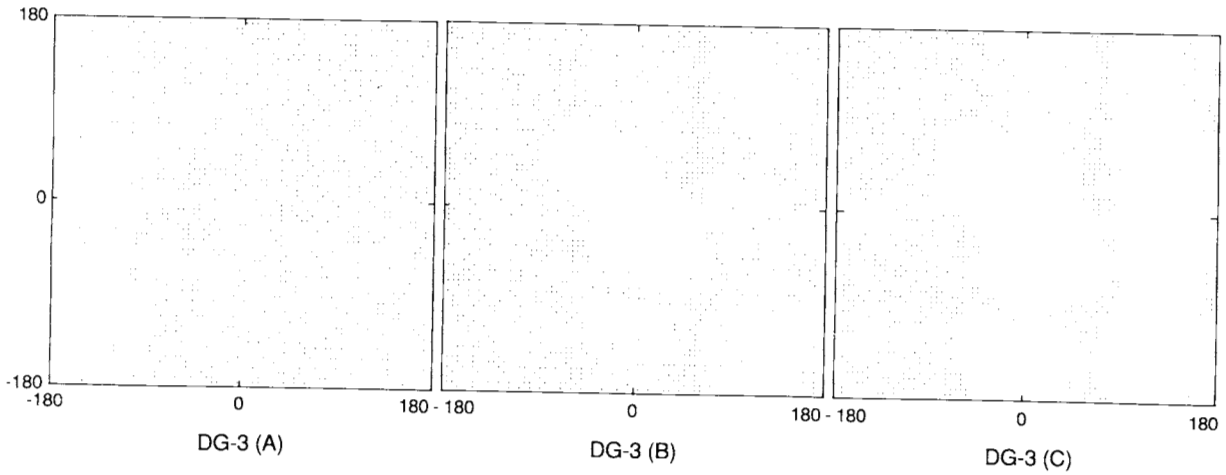


Figure 1 Scatter plots showing the distribution of residues in the ϕ/ψ plane in each of the ensembles of fifty 20-poly(L-alanine) conformations computed by distance space methods. The vertical axis corresponds to ψ, the horizontal to ϕ. DG-1 refers to the ensembles computed by embedding initial substructures and using metrization to embed the final structures (see the distance geometry calculations section for an explanation of terms). DG-2 refers to the ensembles computed using the "classical" embed algorithm involving neither substructures nor metrization. DG-3 refers to the ensembles computed using metrization but no substructures. The labels A, B, and C refer to the data sets consisting of the distance and chirality constraints that follow from the covalent structure of 20-poly(L-alanine) (A), together with short-range radii (B), and both short- and long-range radii (C), as described in the distance geometry calculations section.

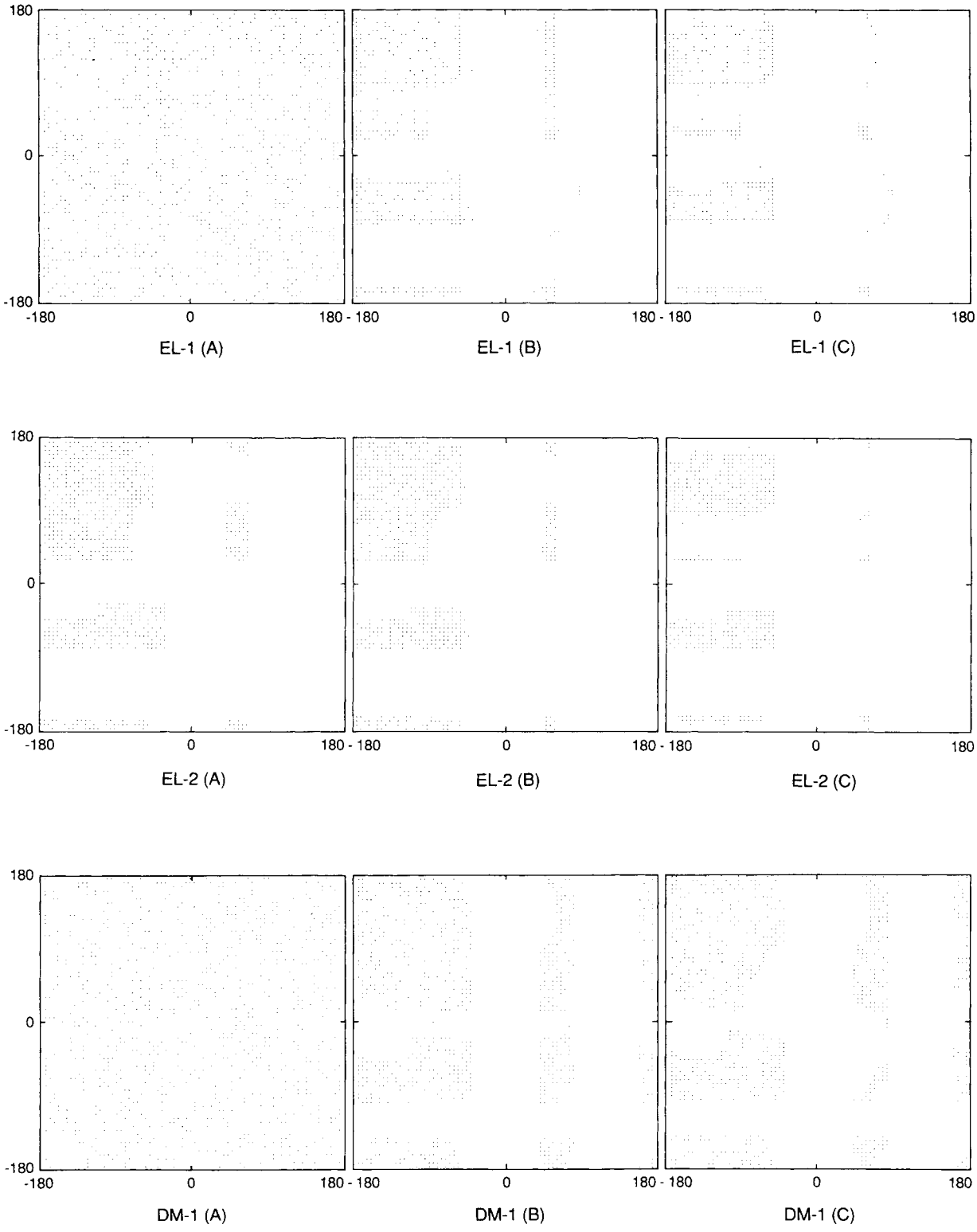


Figure 2 Scatter plots showing the distribution of residues in the ϕ/ψ plane in each of the ensembles of fifty 20-poly(L-alanine) conformations computed by torsion space methods. The vertical axis corresponds to ψ, the horizontal to ϕ. EL-1 refers to the ensembles computed by refinement, using the ellipsoid algorithm, of conformations obtained from a uniform random selection of their ϕ/ψ angles within the interval $[-180°, +180°]$. EL-2 refers to the ensembles computed by the refinement with the ellipsoid algorithm of conformations obtained by selecting their ϕ/ψ angles randomly from within Ramachandran's extreme limits. DM-1 refers to those ensembles computed with the DISMAN program starting from conformations whose ϕ/ψ angles were selected with a uniform random distribution within $[-180°, +180°]$. The labels A, B, and C have the same meanings that they have in Figure 1.

mensional chain, the maximum entropy density is constant when expressed as a function of the angles between successive links of the chain, and it would be at least surprising if the analogous result did not hold in three dimensions. For the purposes of the ensuing discussion, we shall therefore assume that this is the case.

Because of the complicated nature of the feasible region, the average geometric properties of a molecule with respect to this probability distribution can be calculated only with very simple types of distance constraints like those considered in this paper. Although these results doubtless give some indication of how well and in what ways the various distance geometry algorithms available succeed in sampling according to the thermodynamic distribution, we stress that it would be dangerous to extrapolate too far toward much more complicated and extensive sets of constraints like those available from two-dimensional nmr studies.

One thing that is shown clearly by this study is that the average of no single geometric property, including the commonly used difference measures, can be taken as a measure of "randomness." In the past, we have used the average distance matrix error DME[51] and RMSD[38] among conformations as an ad hoc method of estimating the relative sizes of the regions of conformation space that are spanned by ensembles of conformations computed by distance geometry algorithms. The appropriateness of such measures has been amply demonstrated by the fact that when one eliminates some of the data used to compute such ensembles, the average RMSD usually increases. Nevertheless, the fact that the RMSD correlates poorly with the DHAD, which is an equally qualified measure of "size," is already enough to show that it should not be taken too literally. It is also trivial to construct nonrandom ensembles whose RMSD is very high, thus showing that it is not a rigorous measure of randomness per se. As long as no rigorous measure of randomness exists that can also be computed in a practical way, however, the RMSD of an ensemble will remain quite useful as a means of getting an initial idea of what is going on. This should of course be followed up by a more careful study using as many other kinds of measures as possible in order to decide exactly what the available data are really telling us about the conformation, as we have done in this paper.

Keeping these limitations in mind, we can make the following conclusions concerning the sampling obtained from the embed algorithm and its variations. First, with respect to the simple constraints used here, the conformations produced by the "classical" embed algorithm as well as the version using intermediate "substructures" (i.e., DG-1 and DG-2) are undoubtedly more extended than can be justified by the thermodynamic distribution alone. This was only to be expected, simply because the individual trial distances are chosen between the lower and upper triangle inequality limits[12] with a roughly Gaussian distribution whose mean is one half the sum of these limits. When no long-range upper bounds are present, the upper triangle limit on the end-to-end distance necessarily exceeds its maximum possible value, which is proportional to the number of monomers in the chain, whereas the lower triangle limit stays constant at the sum of the hard sphere radii employed. It follows that the rms end-to-end distance in the computed conformations will also be proportional to the length of the chain rather than to the square root thereof. This argument of course also holds when a uniform distribution is used to select the trial distances.

As will be shown in the appendix, the degree of expansion of the conformations obtained from the embed algorithm is easy to control. A more serious problem is presented by the fact that the RMSD of the ensembles DG-1 and DG-2 appears to be too low. This in turn is a consequence of the lack of *variation* in the degree of expansion of the computed conformations, rather than of the expansion itself. In fact, the RMSD can be decomposed into the difference between the squares of the radii of gyration of the two conformations being compared and the correlation between their coordinates[40] so that (for a given degree of correlation) one would expect expanded structures to have a *greater* RMSD. The standard deviation in $\sigma^2(\infty)$ expected for a freely jointed chain can easily be calculated from its known probability distribution[14] to be $\sqrt{2/3} = 82\%$ of its mean value.

This lack of variation is probably an "averaging" phenomena inherent in the way in which embedding operates, by best fitting a set of inconsistent trial distances. As also noted in Ref. 15, it should therefore be possible to increase the RMSD in the computed ensembles by reducing the level of inconsistency present in the trial distances, *provided* that this can be done without introducing new (and possibly worse) biases in the result. The increased RMSD obtained on choosing the trial distances by metrization (DG-3) supports this view. Unfortunately, the implementation of the metrization algorithm used in these calculations computes the trial distances with a fixed order proceeding from the N to the C end of the chain. As a consequence, the end-to-end squared distances of the conformations tend to be relatively small, and examination of the computed conformations by computer graphics also

reveals that the N-terminal ends of these conformations are more compact than their C-terminal ends.

Turning now to the torsion angle optimizations, we observe that although the conformations usually expanded on refining them with respect to the hard-sphere lower bounds, comparison of σ^2 and τ^2 for EL-2 with those for EL-1 and DM-1, (B) and (C), indicates that the latter ensembles are still too compact. Indeed, since Brant and Flory's results imply that $\sigma^2(\infty)$ should be about 4 under the conditions of this study,[25] it would appear that all of the ensembles obtained by torsion space methods are a little on the compact side. This is probably because the starting conformations do not take account of the long-range steric constraints, and the refinements do not change the conformations by more than is necessary to eliminate them. Fortunately, this tendency does not appear to be large, and the variations in the degree of compactness are roughly of order what one would expect from the available theory. As stressed above, the "correct" degree of expansion when long-range upper bounds on the distances are present (as obtained from nmr spectroscopy) is unknown, both now and in the foreseeable future.

For the reasons stated above, we regard the distribution of residues in the ϕ/ψ plane as a more discriminating measure of thermodynamic randomness than their end-to-end squared distances or radii of gyration. Here it appears that the embed algorithm samples fairly well. There is a slight tendency for the residues to cluster toward the center of the ϕ/ψ diagram when no short- or long-range hard-sphere lower bounds are present. Given that the conformations involved are extended, this observation is somewhat surprising, but the tendency is not very strong, and apparently the *trans* configurations of our peptide bonds together with the many more or less *trans* dihedral angles that still occur is sufficient to give the overall degree of extendedness obtained. It should also be noted that, from a practical point of view, it is more important to sample widely in torsion space than it is in Cartesian space. This is because the difficulty of making further adjustments to the conformations by molecular mechanics or dynamics procedures is much more strongly affected by the size of the change in torsion angles (DHAD) required than it is by the change in coordinates (RMSD). It has been shown that molecular dynamics procedures are in fact capable increasing the spatial scatter in protein conformations obtained from the embed algorithm, as well as making the ϕ/ψ distribution closer to what is found in protein crystal structures.[6,52]

The sampling in the ϕ/ψ plane of the torsion space algorithms (EL-1 and DM-1) when no steric constraints are present is of course perfect by definition. It may therefore at first seem surprising that this statement fails to hold after refinement. Regardless of whether the refinement is done by the ellipsoid algorithm or variable target function technique, there is a tendency for the ϕ/ψ angles to cluster on the boundaries of the sterically allowed regions of the Ramachandran map. This is easily understood, since these refinements move the conformation until the constraints are just satisfied, and then leave them there at the boundaries. In addition, both helical regions of the ϕ/ψ plane tend to be populated more densely than the β region, probably because the size of the "basins" in the hypersurface of the error or penalty functions used by these refinements are not simply proportional to the size of the feasible regions that they contain. These undesirable tendencies are alleviated substantially by choosing the initial distribution of residues randomly from within the sterically allowed region of the Ramachandran diagram, rather than uniformly over the entire ϕ/ψ plane (EL-2).

It is worth emphasizing once again that the statistical geometry of the polypeptide backbone is very similar in the computations DG-1 and DG-2. This is because the refinement step does not drastically change the coordinates of the conformations, i.e., their distribution is dominated by the distribution of starting conformations. The backbone conformation is fully determined by the conformation of the substructure (which includes its N and C atoms), and comes out much the same whether or not the rest of the structure is embedded at the same time. It has been observed, however, that with a very small substructure consisting of only the C^α atoms, the main-chain variability appears to be improved.[15,53] This is probably due to the fact that the trial distances among such a small number of atoms can never be too drastically inconsistent, and does not appear to be a good, general method of improving main-chain variability. Moreover, in polypeptides with long side chains, the computation of complete structures from intermediate substructures has been observed to place underconstrained side chains in similar positions most of the time.[11] Thus, when convergence is not a serious problem and the available data are not sufficient to determine the conformation with reasonable precision so that sampling, especially of side-chain conformations, becomes important, intermediate substructures should not be used.

Fortunately, the above observations also suggest a number of ways in which the sampling obtained

with the distance geometry algorithms studied here can be brought closer to the thermodynamic ideal. When the embed algorithm is used without metrization, and no long-range distance constraints are present, it appears that distribution with which the individual trial distances are selected will have to be biased toward lower values if the correct degree of expansion is to be obtained. This has never been done before simply because experimental information concerning the "expandedness" of globular proteins is usually not available, at least from nmr spectroscopy. In protein structure determinations by nmr, however, it often turns out that there are substantial segments of the polypeptide chain for which little data are available, and in such cases it appears that more "reasonable" guesses at the conformations of these segments could be obtained by using an appropriately biased distribution (although such guesses are still not likely to be right!). It also appears that the sampling is improved in many ways when metrization is used, although as previously noted the implementation of this procedure used in the DISGEO program introduces other unjustifiable tendencies. In the appendix, we demonstrate that these problems can be easily overcome without making any major change in the method.

In the case of torsion space refinements, we have shown that a step toward our goal of thermodynamic sampling may be obtained simply by choosing the starting ϕ and ψ angles randomly within Ramachandran's extreme limits (provided, of course, that no explicit information to the contrary is available). If the distance constraints dictate otherwise, it should still be possible for the refinement to converge to regions outside the Ramachandran extreme limits. In other cases, it may happen that the resulting bias causes us to miss finding some unusual conformations consistent with the constraints. Another approach to eliminating clustering in the ϕ/ψ plane that might be less prone to missing conformations would be to alter the weights given to the various terms of the error function used, but this would also appear to be relatively difficult to put into practice.

Although a number of investigators are working on methods of exhaustively searching conformation space,[54-56] none of these methods can presently be applied to problems of the sizes that distance geometry methods can handle without using either heuristic rules or grid sizes that also give rise to the same danger of missing important conformations. For the time being as well as, we believe, a number of years to come, such dangers will remain for all practical intents and purposes unavoidable. The intent behind distance geometry calculations is in this sense similar to that of simulated annealing,[57] in that rather than attempting to solve a difficult problem completely we simply attempt to find a number of reasonably good solutions from which we can subsequently choose the "best" on the basis of more complex criteria. Until the goal of thermodynamic sampling of the space of conformations consistent with the experimentally available distance constraints can be achieved, practitioners of distance geometry will simply have to make a somewhat subjective judgement concerning which biases are to be preferred for the problem at hand. Fortunately, the data that are now available from nmr spectroscopy are often good enough to determine an essentially unique conformation, at which point the correct result will be found regardless of bias.

APPENDIX

Some Ways to Improve the Sampling Obtained with the Embed Algorithm

Both the foregoing pages as well as Ref. 15 hint at a variety of ways in which the sampling obtained with the embed algorithm can be substantially improved upon, at least in the limiting of case of no long-range constraints beyond the hard-sphere radii of the atoms. We felt it necessary, however, to delay the publication of this paper until we could prove by concrete demonstration that improvements we proposed would actually work. In this appendix, we present a brief account of how these improvements were obtained, together with the results of the demonstration itself. A more detailed paper describing these and other improvements to the embed algorithm is in preparation by the author.

Basically, there were two more or less separate problems to be solved:

1. the average degree of expansion of the computed conformations had to be made to agree with the predictions of Flory's theory;[14]

and

2. the variations in the degree of expansion had to be increased in accord with Flory's theory so that, in particular, the RMSD among the computed conformations agreed with McLachlan's predictions.[40]

It is noteworthy that both of these problems could be solved without making any fundamental extensions to the embed algorithm as implemented in the DISGEO program.

An Improved Distribution for the Trial Distances

Because the mean square radius of gyration is given in terms of the mean square distances by

$$\langle R_G^2 \rangle = \left\langle \frac{1}{N^2} \sum_{i<j} d_{ij}^2 \right\rangle = \frac{1}{N^2} \sum_{i<j} \langle d_{ij}^2 \rangle$$

(see Refs. 13 and 14), one approach to the first goal is simply to alter the distribution by which the (squared) *trial* distances are chosen so as to achieve the desired value of $\langle R_G^2 \rangle$. Since these random trial distances are essentially never compatible with a three-dimensional structure, however, they must be transformed by the metric matrix procedure that forms the heart of the embed algorithm into a set of "best-fit" three-dimensional distances.[12,58]

Geometrically, this transformation corresponds to the orthogonal projection of a higher dimensional structure onto a three-dimensional subspace. Since projection in a Euclidean space necessarily decreases (or at least does not increase) the values of all of the distances, it may at first seem that the conformations obtained from trial distances that give the correct radius of gyration according to the above formula will usually wind up being too compact. As is explained in Chapter 3 of Ref. 12, however, these higher dimensional structures are actually contained in a non-Euclidean space wherein projection can either increase or decrease the distances. In practice, therefore, the radii of gyration of the projected three-dimensional structures obtained by this procedure are usually not drastically different from those predicted from the trial distances directly, and we can empirically adjust the distribution of the (squared) trial distances to that the desired radius of gyration is attained.

Although Flory's theory predicts that the individual squared distances will have a Gaussian distribution in the event that no long-range distance constraints are present, as discussed above the situation when such constraints are present is much less well understood. Moreover, in the determination of protein structure from two-dimensional nmr data, the constraints contain numerous upper bounds on the long-range distances whose values are actually *less* than the averages predicted by Flory's theory. Thus in such cases the naive use of the distributions that apply to freely jointed chains will actually result in a further tendency toward *expansion* rather than compactness!

The principle of maximum entropy states that when all we know about a random variable t is its mean value, the least biased distribution to use is that which gives an exponential density function $f(t) \sim \exp(-\lambda \cdot t)$.[50] Other reasons why we prefer this distribution include the fact that it is "memoryless."[45] As will be seen presently, that is important because in the course of metrization the lower and upper limits between which we sample the values of the individual distances are constantly changing, and this property ensures that the "biases" in the distances toward smaller values remain the same as we change these limits. Finally, it turns out to be a relatively simple matter to generate random numbers with this distribution between given limits and with any desired mean therein.

The desired mean value of the distribution between the lower and upper limits of each *squared* distance d_{ij}^2 is obtained by making a suitable choice of λ_{ij}. In our case we have used the following formula to predict the mean value of the squared distances:

$$\overline{d_{ij}^2} = (l_{ij} + \alpha \cdot (u_{ij} - l_{ij})) \cdot (u_{ij} - \alpha \cdot (u_{ij} - l_{ij}))$$

for some α between 0 and $\frac{1}{2}$. Observe that when $\alpha = \frac{1}{2}$, the root mean square distance $(\overline{d_{ij}^2})^{1/2}$ is thus the arithmetic mean of the lower and upper distance limits l_{ij} and u_{ij}. On the other hand, when $\alpha = 0$ the root mean square distance $(\overline{d_{ij}^2})^{1/2}$ is equal to geometric mean of its limits. Because the geometric mean of two numbers never exceeds their arithmetic mean, it follows easily that $\overline{d_{ij}^2}$ is an increasing function of α, i.e., the smaller we make α the smaller we can expect our squared trial distances to be.

As noted in the discussion section, when l_{ij} and u_{ij} are the triangle inequality limits implied by the bond lengths of a freely jointed linear chain, then l_{ij} is the sum of the hard-sphere radii assigned to the atoms, which is roughly independent of i and j, while u_{ij} is the sum of the lengths of the bounds connecting i and j. It follows that when $\alpha = 0$ the mean square value of the distance will be proportional to $|i - j|$, in accord with Flory's theory for a non-self-avoiding chain. In practice, we have found that $\alpha = 0$ results in ensembles that are too compact, in that for example $l_{ij} = 0$ then implies that $\overline{d_{ij}^2} = 0$ regardless of the value of u_{ij}. In the calculations reported below, we have simply set $\alpha = 0.1$.

A Randomized Metrization Procedure

Metrization is a algorithm by which trial distances that satisfy the triangle inequality can be randomly chosen within any prescribed self-consistent bounds on their values. It was invented several years ago by the author in the course of implementing the DISGEO program;[2] a proof of its correctness may be found in Ref. 59. Because the triangle inequality $d_{ij} \leq d_{ik} + d_{kj}$ is a well-known[12,60] necessary condition

for a symmetric matrix $[d_{ij}]$ of nonnegative real numbers to be the distances among a collection of points in three-dimensional space, the starting conformations obtained by the metric matrix method[58] from such distances would be expected to have a lower initial error and to converge to structures with a lower final error on refinement than those obtained by selecting the trial distances independently, as has generally been done in the past.[13,51] Moreover, for reasons given in the discussion section above, it is reasonable to expect that better sampling will be obtained when metrization is used, although it is not clear a priori how much better. Perhaps because it is relatively difficult to program the metrization algorithm in an efficient way, to our knowledge it has yet to be incorporated into any of the many other available implementations of the embed algorithm besides the DISGEO program.

Nevertheless, the metrization procedure that DISGEO uses leads to other sampling problems. Briefly, in order to simplify the programming it was necessary to choose the trial distances in the course of metrization in a certain fixed order, so that all the distances involving the atom numbered "1" are chosen first, then all those distances involving the atom numbered "2" (except for d_{12}, of course), and so on up through the last distance $d_{N-1,N}$. Because each distance selected reduces the range of values from which the remaining distances can be selected so as to obtain consistency with the triangle inequality, this means that the resultant distributions on the distances selected toward the end of the procedure are nonuniform. The net effect is that the resultant final conformations tend to be more compact at their N than at their C terminal ends.

In order to eliminate this obviously undesirable tendency, we have now implemented a metrization procedure in which the numbering of the atoms is permuted randomly before beginning the calculation. Although the trial distances that are chosen last in the permuted order are doubtless still nonuniform, which distances are affected this way is now different for each new structure calculated so that only the joint distribution of all the distances together is affected. The overall distribution, in turn, can be adjusted by an appropriate choice of exponent λ_{ij} so that the mean square values of the distances are in accord with Flory's theory for a freely jointed chain. While unjustifiable correlations between the resultant trial distances doubtless remain, there is no longer any well-defined relation between these correlations and the covalent structure of the molecule (or for that matter, any available experimental data). In any case, we doubt that a computationally tractable means of eliminating these correlations will be found in the near future, and the computational results given below indicate that the improvements in the sampling properties of the embed algorithm which are obtained from randomized metrization are already quite substantial.

Computational Results and Discussion

In order to evaluate the effects of the above measures upon the sampling properties of the embed algorithm, we combined them in all possible ways, i.e.,

DG-4. The trial distances were chosen without metrization but with an exponential density from within the triangle inequality limits determined by the constraints.

DG-5. The trial distances were chosen with both randomized metrization and an exponential density from within the triangle in-

Table A.1[a] Average Geometric Properties of Ensembles of 20-Poly(L-Alanine) Computed With Exponential Density and Randomized Metrization

Ensemble	ρ	δ	$\sigma^2 \pm \Delta\%$	$\tau^2 \pm \Delta\%$	$[X, Y, Z]$
DG-4(A)	0.65	0.90	1.97 ± 14	1.17 ± 5	[8, 10, 1]
DG-4(B)	0.69	0.94	2.02 ± 14	1.22 ± 6	[9, 8, 0]
DG-4(C)	0.70	0.97	3.13 ± 15	2.14 ± 6	[4, 14, 1]
DG-5(A)	0.94	0.96	0.56 ± 72	0.80 ± 41	[3, 2, 0]
DG-5(B)	0.96	0.96	0.57 ± 78	0.84 ± 40	[4, 3, 2]
DG-5(C)	1.03	1.00	0.85 ± 81	1.38 ± 29	[3, 3, 1]
DG-6(A)	1.39	0.98	2.17 ± 113	2.18 ± 73	[7, 3, 0]
DG-6(B)	1.40	0.99	2.21 ± 102	2.19 ± 72	[7, 5, −3]
DG-6(C)	1.35	0.97	2.32 ± 107	2.44 ± 63	[6, 6, −3]

[a] All symbols are defined as in Table II of the main part of the paper.

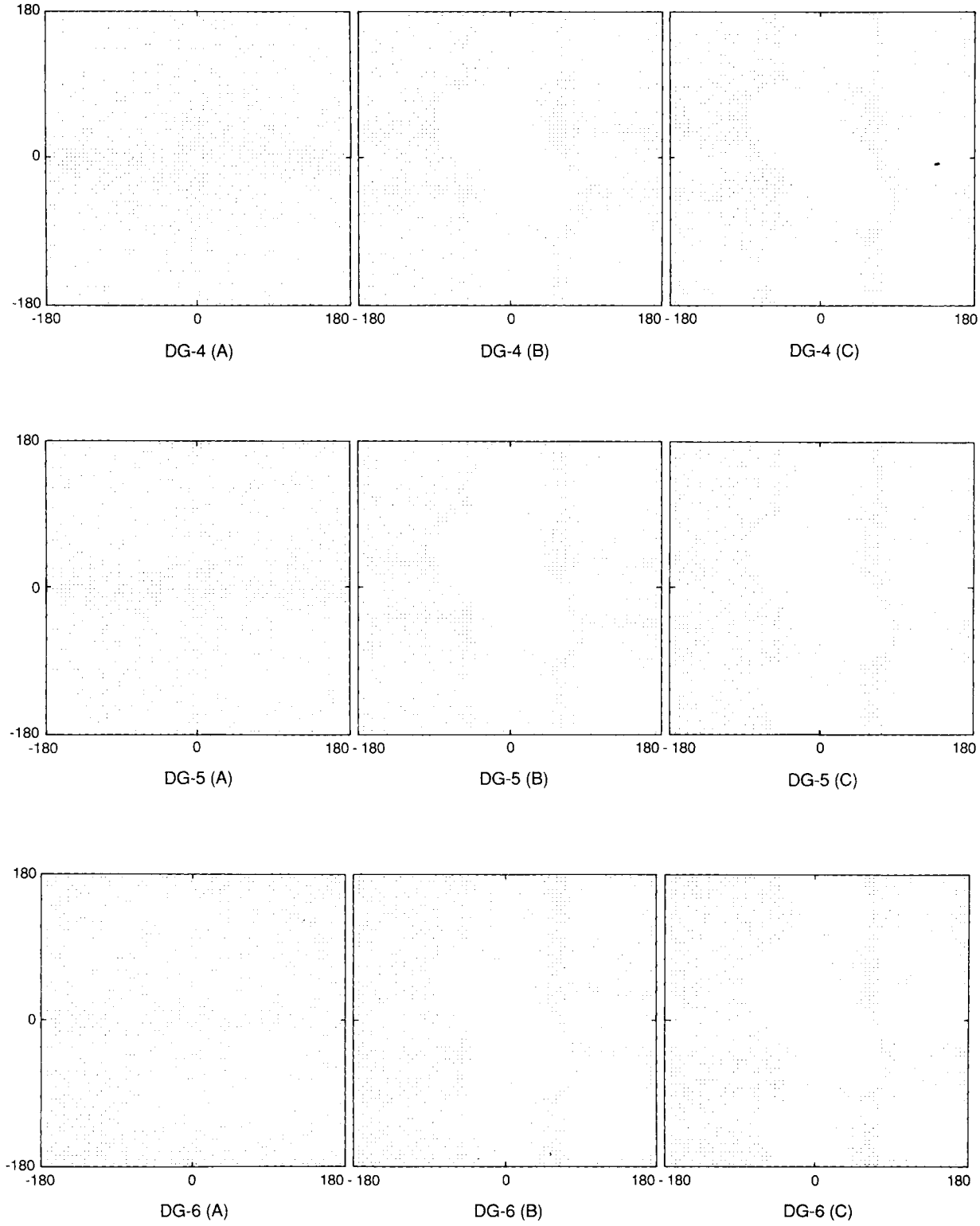


Figure A1 Scatter plots showing the distribution of residues in the ϕ/ψ plane in each of the ensembles of fifty 20-poly(L-alanine) conformations computed by the modified distance space methods evaluated in the appendix. The vertical axis corresponds to ψ, the horizontal to ϕ. DG-4 refers to the ensembles obtained by refinement of conformations embedded from trial distances selected independently with an exponential density from within their triangle inequality limits. DG-5 refers to the ensembles obtained by refinement of conformations embedded from trial distances selected by randomized metrization and with an exponential density. DG-6 refers to the ensembles obtained by refinement of conformations embedded from trial distances selected by randomized metrization but with a uniform density. The labels A, B, and C have the same meanings that they have in Figure 1.

equality limits determined by the constraints.

DG-6. The trial distances were chosen with randomized metrization but with a uniform distribution from within the triangle inequality limits determined by the constraints.

The random number generator used for these computations was the UNIX™ function "random." With each of these three variations on the embed algorithm, we computed ensembles of fifty 20-poly(L-alanine) conformations each, first with no lower bound constraints, then with short-range lower bounds only, and finally with both short and long-range lower bounds, i.e., the constraint sets denoted as A, B, and C in the main paper.

The results of these computations are given in Table A1, which lists the same geometric parameters used in the main part of the paper, together with Figure A1, which shows scatter plots of the amino acid residues in the ϕ/ψ plane for each of these computations. The most important results of these calculations may be summarized as follows.

When the exponential density was used for the selection of the trial distances but these were selected without metrization (DG-4), it will be observed that the desired degree of compactness has been approximately obtained. A value for $\sigma^2(20)$ of 3 is exactly what one would expect from Brant and Flory's results for a self-avoiding chain of this length.[25] The variations in the compactness remain too low, however, and this in turn gives us about the same RMSD as was obtained in runs DG-1 and DG-2. Moreover, the tendency to cluster in the middle of the ϕ/ψ diagram is markedly more noticeable in these compact conformations, especially with respect to the vertical ψ axis.

On selecting the trial distances using the randomized metrization procedure in addition to an exponential density (DG-5), the variations in compactness increase substantially and with them, the RMSD likewise increases. Thus our earlier expectation that decreasing the level of inconsistency present in the trial distances will alleviate this problem has been validated. At the same time, the low values of σ^2 and τ^2 show that these ensembles are now too compact, and the clustering in their ϕ/ψ plots remains. Observe, however, that when the long-range hard-sphere constraints are included in the calculations [DG-5(C)] the ensembles become significantly more expanded, and moreover, the clustering in the ϕ/ψ plots becomes distinctly less pronounced. This shows, among other things, that any strategy aimed at improving sampling must take into account the types of constraints to which it will be applied, and that one must be ready to modify that strategy as the circumstances demand.

Finally, when randomized metrization is used together with a simple uniform distribution to select the trial distances (DG-6), the approximately correct degree of compactness is again obtained. We add that there was only a slight tendency for the radius of gyration, as computed directly from the trial distances by means of the formula given above, to decrease on projection into three dimensions, and that it also did not change much on subsequent refinement. Thus taking account of the triangle inequality alone appears to be sufficient to endow our "random" distances with at least some of the statistical properties that Euclidean distances have in "random" conformations. Perhaps more importantly, the variations in compactness are now extremely large, leading to an RMSD that actually exceeds that obtained by *any* of the distance geometry programs studied in this paper, including torsion space methods. Last but not least, the scatter in the ϕ/ψ plots is, if not perfectly uniform, at least not far less so than the scatter obtained with the torsion space methods, which generate their starting conformations explicitly by choosing their angles with a uniform distribution.

Let it be clearly understood that we do not claim that even the ensembles DG-6 (A), (B), and (C) are truly random, in the thermodynamic sense of the word used here. Doubtless, as our understanding of the relation between the underlying (assumed!) probability density on the accessible conformation space and the statistics of the geometric properties of those conformations improves, flaws will be found by those who look carefully. We also stress once again that we presently have no objective criteria for what "random" means when long-range distance constraints are present, as is the case in protein structure determinations from nmr data. Nevertheless, it appears reasonably certain that when the randomized metrization technique described above is employed, distance space methods will produce conformations that are sufficiently diverse to make it at least unlikely that any observable strong similarities between all members of a large ensemble are due to chance alone. This is perhaps what is most important to the majority of readers of this paper.

The author thanks M. Clore, G. Crippen, S. Hyberts, G. Montelione, R. Scheek, and G. Wagner for their help and/or inspiration in carrying out these studies. I also wish to

thank W. Metzler, D. Hare, and A. Pardi for making a copy of their paper available in advance of its publication. This work was supported by the National Institutes of Health grant R01 GM-38221.

The improvements to the embed algorithm described in the appendix were developed and implemented with support from NIH grant GM-37708.

REFERENCES

1. Billeter, M., Havel, T. F. & Wüthrich, K. (1986) *J. Comput. Chem.* **8,** 132–141.
2. Havel, T. F. & Wüthrich, K. (1984) *Bull. Math. Biol.* **46,** 673–698.
3. Braun, W. & Gō, N. (1985) *J. Mol. Biol.* **186,** 611–626.
4. Nerdal, W., Hare, D. R. & Reid, B. R. (1988) *J. Mol. Biol.* **201,** 717–739.
5. Clore, G. M., Nilges, M., Brunger, A. T., Karplus, M. & Gronenborn, A. M. (1987) *FEBS Lett.* **213,** 269–277.
6. de Vlieg, J., Scheek, M., van Gunsteren, W. F., Berendsen, H. J. C., Kaptein, R. & Thomason, J. (1988) *Proteins Struct. Funct. Genet.* **3,** 209–218.
7. Nilges, M., Clore, G. M. & Gronenborn, A. M. (1988) *FEBS Lett.* **229,** 317–324.
8. Williamson, M. P., Havel, T. F. & Wüthrich, K. (1985) *J. Mol. Biol.* **182,** 295–315.
9. Braun, W. (1987) *Quart. Rev. Biophys.* **19,** 115–157.
10. Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids.* John Wiley & Sons, New York.
11. Wagner, G., Braun, W., Havel, T. F., Schauman, T., Gō, N. & Wüthrich, K. (1987) *J. Mol. Biol.* **196,** 611–639.
12. Crippen, G. M. & Havel, T. F. (1988) *Distance Geometry and Molecular Conformation,* Research Studies Press, Letchworth, UK (Publisher in the U.S. is John Wiley & Sons, New York).
13. Havel, T. F., Kuntz, I. D. & Crippen, G. M. (1983) *Bull. Math. Biol.* **45,** 665–720.
14. Flory, P. J. (1969) *The Statistical Mechanics of Chain Molecules,* Wiley Interscience, New York.
15. Metzler, W. J., Hare, D. R. & Pardi, A. (1989) *Biochemistry* **28,** 7045–7052.
16. Flory, P. J. (1980) *Pure Appl. Chem.* **52,** 241–252.
17. Flory, P. J. (1985) In Mandelkern, L., Mark, J.-E., Suter, U. W. & Yoon, D. Y., Eds., *Selected Works of Paul Flory, vol. 1,* Stanford University Press, Stanford, CA.
18. McQuarrie, D. A. (1976) *Statistical Mechanics,* Harper & Row, New York.
19. Doi, M. & Edwards, S. F. (1986) *The Theory of Polymer Dynamics, Vol. 73, International Series on Monographs on Physics,* Claredon Press, Oxford, UK.
20. Flory, P. J. (1949) *J. Chem. Phys.* **17,** 303–310.
21. Flory, P. J. & Fisk, S. (1953) *J. Chem. Phys.* **44,** 2243–2248.
22. Williams, C. & Brochard, F. (1981) *Ann. Rev. Phys. Chem.* **32,** 433–451.
23. Flory, P. J. (1974) *Macromolecules* **7,** 381–392.
24. Brant, D. A. & Flory, P. J. (1965) *J. Am. Chem. Soc.* **87,** 2788–2791.
25. Brant, D. A. & Flory, P. J. (1965) *J. Am. Chem. Soc.* **87,** 2791–2800.
26. Ramachandran, G. N. & Sasisekharan, V. (1963) *Adv. Protein Chem.* **68,** 284–438.
27. Miller, W. G. & Flory, P. J. (1966) *J. Mol. Biol.* **15,** 298–314.
28. Miller, W. G., Brant, D. A. & Flory, P. J. (1967) *J. Mol. Biol.* **23,** 67–80.
29. Brant, D. A., Miller, W. G. & Flory, P. J. (1967) *J. Mol. Biol.* **23,** 47–65.
30. Flory, P. J. (1973) *Proc. Natl. Acad. Sci.* **70,** 1819–1823.
31. Flory, P. J. & Yoon, D. Y. (1975) *J. Chem. Phys.* **61,** 5358–5365.
32. Hesselink, F. Th. (1974) *Biophys. Chem.* **2,** 76–81.
33. Premilat, S. & Hermans, J., (1973) *J. Chem. Phys.* **59,** 2602–2612.
34. Conrad, J. C. & Flory, P. J. (1976) *Macromolecules* **9,** 41–47.
35. Rao, S. T. & Rossmann, M. G. (1973) *J. Mol. Biol.* **76,** 241–256.
36. Cohen, F. E. & Sternberg, M. J. E. (1980) *J. Mol. Biol.* **138,** 321–333.
37. McLachlan, A. D. (1979) *J. Mol. Biol.* **128,** 49–79. (see Appendix for method of computing RMSD).
38. Havel, T. F. & Wüthrich, K. (1985) *J. Mol. Biol.* **182,** 281–294.
39. Purisima, E. O. & Scherga, H. A. (1984) *Biopolymers* **23,** 1207–1224.
40. McLachlan, A. D. (1984) *Biopolymers* **23,** 1325–1331.
41. Ecker, J. G. & Kupferschmidt, M. (1985) *SIAM J. Control Optim.* **23,** 657–674.
42. Clore, G. M., Sukumaran, D. K., Nilges, M. & Gronenborn, A. M. (1987) *Biochemistry* **26,** 1732–1745.
43. Clore, G. M., Sukumaran, D. K., Nilges, M., Zarbock, J. & Gronenborn, A. M. (1987) *EMBO J.* **6,** 529–537.
44. Moore, J. M., Case, D. A., Chazin, W. J. Gippert, G. P., Havel, T. F., Powls, R. & Wright, P. E. (1988) *Science* **240,** 314–317.
45. Feller, W. (1971) *An Introduction to Probability Theory and Its Applications, volume II,* Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York.
46. Hammersley, J. M. & Handscomb, D. C. (1964) *Monte Carlo Methods,* John Wiley & Sons, NY.
47. Kline, A. D., Braun, W. & Wüthrich, K. (1988) *J. Mol. Biol.* **204,** 675–724.
48. Cooke, R. M., Wilkinson, A. J., Baron, M., Pastore, A., Tappin, M. J., Cambell, I. D., Gregory, H. & Sheard, B. (1987) *Nature* **327,** 339–341.
49. Schultz, P., Wörgötter, E., Braun, W., Wagner, G., Vašák, M., Kägi, J. H. R. & Wüthrich, K. (1988) *J. Mol. Biol.* **203,** 251–268.

50. Jaynes, E. T. (1979) In Levine, R. D. & Tribus, M., Eds., *The Maximum Entropy Formalism.* MIT Press, Cambridge, MA.
51. Havel, T. F., Kuntz, I. D. & Crippen, G. M. (1979) *Biopolymers* **18,** 73–82.
52. Gronenborn, A. M. & Clore, G. M. (1989) *Biochemistry* **28,** 5978–5984.
53. Thomason, J. F. & Kuntz, I. D. (1989) *J. Cell. Biochem. Suppl.* **13A,** 37.
54. Lichtarge, O., Cornelius, C. W., Buchanan, B. G. & Jardetzky, O. (1987) *Proteins* **2,** 340–358.
55. Bruccoleri, R. E. & Karplus, M. (1987) *Biopolymers* **26,** 137–168.
56. Lipton, M. & Clark Still, W. (1988) *J. Comput. Chem.* **9,** 343–355.
57. Aarts, E. & Korst, J. (1989) *Simulated Annealing and Boltzman Machines,* John Wiley & Sons, Ltd., Chichester, UK.
58. Crippen, G. M. & Havel, T. F. (1978) *Acta Cryst.* **A34,** 282–284.
59. Dress, A. W. M. & Havel, T. F. (1988) *Discrete Applied Math.* **19,** 129–144.
60. Blumenthal, L. (1953) *Theory and Applications of Distance Geometry,* Cambridge Unversity Press, Cambridge, UK (Reprinted by the Chelsea Publishing Co., 1970).

Received April 12, 1989
Accepted October 10, 1989