Limited Sampling of Conformational Space by the Distance Geometry Algorithm: Implications for Structures Generated from NMR Data[†]

William J. Metzler,[‡] Dennis R. Hare,[§] and Arthur Pardi*,[‡]

Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, Colorado 80309-0215, and Hare Research, Inc., 14810 216th Avenue, N.E., Woodinville, Washington 98072

Received March 1, 1989; Revised Manuscript Received April 26, 1989

ABSTRACT: Calculations with a metric matrix distance geometry algorithm were performed that show that the standard implementation of the algorithm generally samples a very limited region of conformational space. This problem is most severe when only a small amount of distance information is used as input for the algorithm. Control calculations were performed on linear peptides, disulfide-linked peptides, and a double-stranded DNA decamer where only distances defining the covalent structures of the molecules (as well as the hydrogen bonds for the base pairs in the DNA) were included as input. Since the distance geometry algorithm is commonly used to generate structures of biopolymers from distance data obtained from NMR experiments, simulations were performed on the small globular protein basic pancreatic trypsin inhibitor (BPTI) that mimic calculations performed with actual NMR data. The results on BPTI and on the control peptides indicate that the standard implementation of the algorithm has two main problems: first, that it generates extended structures; second, that it has a tendency to consistently produce similar structures instead of sampling all structures consistent with the input distance information. These results also show that use of a simple root-mean-square deviation for evaluating the quality of the structures generated from NMR data may not be generally appropriate. The main sources of these problems are identified, and our results indicate that the problems are not a fundamental property of the distance geometry algorithm but arise from the implementations presently used to generate structures from NMR data. Several possible methods for alleviating these problems are discussed.

Recent progress in two-dimensional nuclear magnetic resonance (NMR)[1] spectroscopy and computational techniques has made it quite routine to determine the structures of small biopolymers in solution [see Wüthrich (1986, 1989) for reviews]. A frequently used procedure for generating structures from NMR distance data is application of a metric matrix distance geometry algorithm (Crippen, 1977; Havel et al., 1983). The distance geometry algorithm is popular since it efficiently produces structures consistent with the input distance data. Some alternate procedures for generating structures from NMR distance data are minimization in torsion angle space using a variable target function (Braun & Go, 1985; Braun, 1987) and a simulated annealing algorithm (Nigles et al., 1988a). Studies have shown that these methods have similar efficiencies to metric matrix distance geometry techniques (Wagner et al., 1987; Nigles et al., 1988a). Other procedures such as restrained molecular dynamics (Kaptein et al., 1985; Brünger et al., 1986), restrained molecular mechanics (Fesik et al., 1986; Holak et al., 1987), and restrained Monte Carlo (Bassolino et al., 1988) techniques presently appear to be less efficient at generating starting structures consistent with the input distance information, and for this reason a distance geometry algorithm is commonly used to generate starting structures for restrained molecular dynamics or Monte Carlo calculations (Nigles et al., 1988a; Bassolino et al., 1988; Clore et al., 1987a,b).

The distance information obtained from NMR experiments is generally not sufficient to define uniquely the conformation of a molecule with a known covalent structure. The reasons for this are that the longest proton-proton distance that can be measured from NMR data is 4.5-5.0 Å and that the distances can only be determined with precisions ranging from ±0.2-1.0 Å (Wüthrich, 1986). Therefore, an important question to address when solution structures are being determined from NMR distance data is how well the computational method samples conformational space for all conformational classes consistent with the distance data. Recent studies have suggested that metric matrix distance geometry technique searches a more restricted region of conformational space than restrained molecular dynamics calculations or minimization with a variable target function (Wagner et al., 1987; de Vlieg et al., 1988; Nigles et al., 1988a), but these workers did not identify the source of the limited sampling. Their studies were also performed on systems with a large amount of distance information, and therefore, the global conformation of the molecule was well-defined. In this work, distance geometry calculations have been performed on a number of oligopeptides, an oligonucleotide, and the small protein bovine pancreatic trypsin inhibitor, BPTI, and the results show that the standard metric matrix distance geometry algorithm searches a severely limited region of conformational space consistent with the input distance data when there is limited distance information defining the conformation of the molecule. These studies also point out problems associated with determining how well a structure is actually defined from

[†] This work was supported in part by the Searle Scholars Program of the Chicago Community Trust (85-C110), NIH GM35801, and NIH A127026 to A.P., NIH GM35620 to D.R.H., and NRSA GM13075 to W.J.M.

[*] To whom correspondence should be addressed.

[‡] University of Colorado at Boulder.

[§] Hare Research, Inc.

[1] Abbreviations: NMR, nuclear magnetic resonance; 2D, two dimensional; BPTI, basic pancreatic trypsin inhibitor; rms, root mean square.

0006-2960/89/0428-7045$01.50/0

a given set of distance data. The source of the limited sampling by the distance geometry algorithm, as well as suggestions for improving the distance geometry calculations, is discussed.

Materials and Methods

The distance geometry calculations were performed with the FORTRAN program DSPACE (Hare Research, Inc.). The method used to generate structures with DSPACE is similar to that previously described (Hare & Reid, 1986; Patel et al., 1987; Nerdal et al., 1988). In this work a distance bounds matrix was created on the basis of the covalent structure of the molecule and any simulated experimental distance constraints. The input for the bounds matrix is in the form of upper and lower distance bounds, and this matrix is then subjected to a smoothing procedure based on triangle inequalities (Crippen, 1981). Random distances between the upper and lower bounds are chosen for all pair of atoms in the molecule, and an embedding procedure is employed to reduce the problem from *n*-dimensional distance space (where *n* is the number of atoms) to three-dimensional distance space (Havel et al., 1979). These embedded structures provided the starting point for refinement in the DSPACE program where a variety of nonlinear optimization routines, including conjugate gradient minimization, randomization, and a simulated annealing type algorithm (Patel et al., 1987; Nerdal et al., 1988), were used to extract a structure from local minima and to reduce the distance violations between the structure and the input bounds matrix.

For the structure calculations on the linear and disulfide-linked oligopeptides, four sets of structures (each with 10 structures per set) were generated which will be referred to as embedded, refined, annealed, and random structures. The embedded structures were created by the standard metric matrix embed procedure (Havel et al., 1979; Havel & Wüthrich, 1984). These structures were then refined with a conjugate gradient minimization which reduced the large bond and distance violations of the embedded structures. A total of 512 iterations of conjugate gradient refinement were generally sufficient to ensure that there were no bond or distance violations greater than 0.1 Å in these refined structures. The simulated annealing type algorithm in the DSPACE program (Nerdal et al., 1988) was used to produce the annealed structures. Refined structures were heated by application of 256 iterations of undamped annealing (with a maximum allowed velocity of 0.1 Å/s), "cooled" with 32 iterations of conjugate gradient refinement, and then subjected to 256 iterations of damped annealing. A final 128 iterations of conjugate gradient refinement ensured that there were no distance violations greater than 0.1 Å in these structures. Random structures were produced with the FORTRAN program IMPACT, kindly provided by Dr. R. M. Levy. Starting with a fully refined structure, ϕ and ψ backbone torsion angles were changed to random values between 0 and 360°. These structures were then energy minimized for 100 cycles to eliminate unfavorable van der Waals interactions except for the disulfide-linked peptide, which was refined with the DSPACE program to generate the correct disulfide bond length.

An additional set of oligopeptide structures were generated with the submatrix embed procedure (Havel & Wüthrich, 1984) recently described by Kuntz and co-workers (Thomason & Kuntz, 1989). In this procedure, structures were produced by an embed that uses only a small subset of the total number of atoms. For the calculations presented here, all the α-carbons in the polypeptide backbone were used for the submatrix embed. Structures embedded in this way were then refined with 2000 iterations of conjugate gradient refinement and are referred to as "submatrix" embedded structures throughout the text.

The non-self-complementary DNA decamer 5′-d-(CGTCACGCGC)-3′ was used as a model system for the control calculations on DNA oligonucleotides. In order to generate a double-stranded structure for the DNA, the three CG and two AT intra base pair hydrogen-bonding distances were given upper and lower bounds of 1.8 and 1.9 Å, respectively. Embedded structures for the DNA decamer were generated by standard methods (Patel et al., 1987; Nerdal et al., 1988), and these structures were refined by 1000–2000 rounds of conjugate gradient minimization until there were no violations greater than 0.1 Å.

The rms distance deviations for each set of structures (Nyburg, 1973) were calculated with the program IMPACT. All calculations were performed on an Alliant FX-8 computer or a Sun 4/260 computer at the University of Colorado.

Results and Discussion

Sampling of Conformational Space by the Embed Procedure. A set of linear peptides $(\text{Lys-Glu})_n$ with $n = 3, 6, 12$, and 24 was used as a control to test the extent to which the metric matrix distance geometry program samples conformational space. Distance geometry structures were generated from a bounds matrix that contained no additional distance information other than that required to define the covalent structure of the molecules. Given these limited distance constraints, any conformation with correct bond angles and bond lengths that does not have unfavorable van der Waals interactions will be consistent with the input bounds matrix. One would therefore anticipate that a set of random structures would be generated by the embed procedure. Figure 1a shows the superposition of the backbone atoms for the embedded starting structures of the linear peptide $(\text{Lys-Glu})_{12}$, Figure 1b shows a superposition of the backbone atoms for the refined structures, in which a simple conjugate gradient minimization of the embedded structures was performed until there were no distance violations greater than 0.1 Å, and Figure 1c shows the random structures, produced with the program IMPACT as described under Materials and Methods. It is clear that the embed procedure does not randomly sample conformational space consistent with the input distance data but instead samples only a very limited region of conformational space consistent with these distances. Furthermore, the metric matrix embed procedure tends to repeatedly generate structures that are extended relative to the input distance constraints. The limitations of the metric matrix embed procedure in sampling conformational space were quantified by determining the rms distance deviations for the linear and disulfide-linked peptides. Table I shows the large differences between the rms deviations among the structures produced by embedding and refinement with the distance geometry program, and the rms deviation among the random structures. The rms deviations for the refined structures are much smaller than those of the random structures, indicating that a severely restricted region of conformational space is sampled. Although the rms deviations for the embedded starting structures are similar to those of the random structures, these are artificially high due to the large bond and distance violations at the ends of the polypeptide chains of the embedded structures, as seen in Figure 1a. Such large distance violations are not present in the peptide with a disulfide linkage, $\text{Cys-(Lys-Arg)}_5\text{-Cys}$ (structure not shown), and these violations appear to be an end effect of the embed in the linear peptides.

Figure 1 indicates that the distance geometry algorithm is producing a set of very extended structures consistent with the

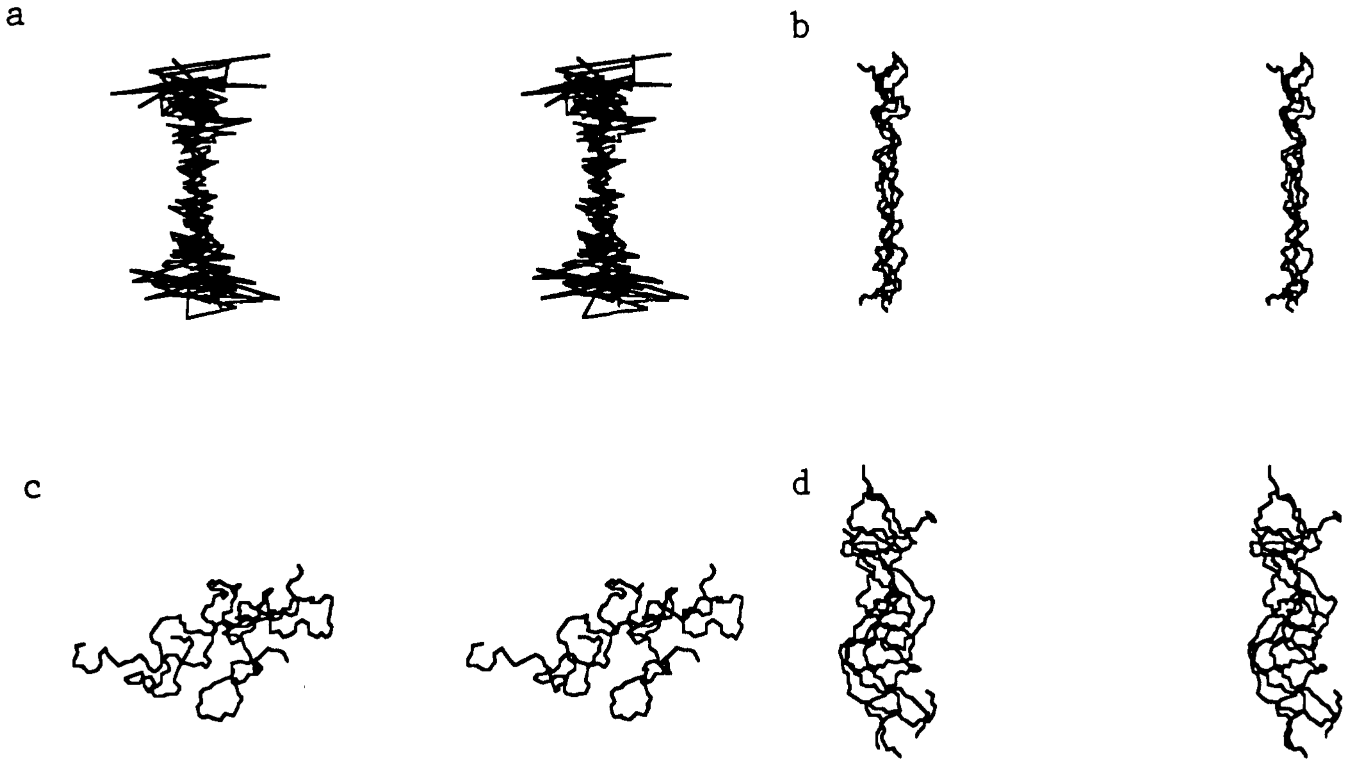

FIGURE 1: Stereoview of the superposition of the backbone (N, C$^\alpha$, and C) atoms of the oligopeptide (Lys–Glu)$_{12}$: (a) embedded structures, (b) refined structures, (c) random structures, and (d) annealed structures. The structures were generated as described under Materials and Methods. For clarity, a randomly chosen set of four of the ten structures generated for each class were shown in the figure.

Table I: Comparison of Control Structures

		average rms deviation (Å)				
molecule	no. of amino acids	embedded[a]	refined[b]	annealed[c]	random[d]	submatrix embedded[e]
(Lys-Glu)$_3$	6	2.75	1.97	2.28	2.31	2.41
(Lys-Glu)$_6$	12	5.08	2.53	2.74	4.48	3.73
(Lys-Glu)$_{12}$	24	9.55	3.24	5.43	7.94	6.84
(Lys-Glu)$_{24}$	48	17.77	3.77	8.15	12.52	8.48
BPTI[f]	58	23.99	3.90	8.50	13.19	8.63
(Lys-Glu)$_C$[g]	12	3.04	2.03	2.83	4.05	2.98

[a] Structures generated with the embed procedure. [b] Structures generated by conjugate gradient refinement of the embedded structures (see Materials and Methods). [c] Structures generated by a simulated annealing type procedure of the refined structures (see Materials and Methods). [d] Random structures. [e] Structures generated by conjugate gradient refinement of the submatrix embedded structures (see Materials and Methods). [f] BPTI with distance constraints which only define the amino acid sequence. [g] $\overline{\text{Cys-(Lys-Glu)}_5\text{-Cys}}$.

Table II: Extendedness of Oligopeptide Structures

		average measured $d_{1,n}$ (Å)[a]				
molecule	upper bound	embedded	refined	random	annealed	submatrix embedded
(Lys-Glu)$_3$	18.2	12.0 (2.2)	13.2 (4.0)	12.0 (7.3)	11.0 (9.0)	10.1 (8.7)
(Lys-Glu)$_6$	40.0	26.5 (7.8)	29.6 (3.9)	19.9 (15.3)	20.7 (12.2)	16.3 (12.4)
(Lys-Glu)$_{12}$	83.5	50.2 (10.5)	52.5 (5.8)	26.1 (24.9)	37.8 (24.4)	20.6 (21.7)
(Lys-Glu)$_{24}$	170.7	108.1 (21)	101.7 (4.1)	33.4 (49.4)	61.7 (61.3)	26.3 (15.9)

		% upper bound[b]				
molecule	upper bound	embedded	refined	random	annealed	submatrix embedded
(Lys-Glu)$_3$	18.2	66.2	72.7	66.2	60.2	55.4
(Lys-Glu)$_6$	40.0	66.3	73.9	49.7	51.8	40.8
(Lys-Glu)$_{12}$	83.5	60.1	63.0	31.2	45.3	24.7
(Lys-Glu)$_{24}$	170.7	63.3	59.6	19.6	36.1	15.4

[a] The average distance from Cα(1) to Cα(n) in the structure with the range of the difference between largest and smallest distances in each set given in parentheses (see text). [b] Percent upper bound = (average distance/upper bound) × 100.

input distances. To quantity this observation, Table II summarizes data on the largest distances observed in these peptide structures and gives what percentage these distances are of the upper bounds for this pair of atoms. The largest possible distance for a linear oligopeptide will obviously occur between residues on opposite ends of the peptide chain. Therefore, we have chosen to monitor the variation among the calculated structures by measuring the distance between the C$^\alpha$ of residue 1 and the C$^\alpha$ of residue n, where n is the number of amino acids in the oligopeptide. This distance will be referred to as $d_{1,n}$. For the embedded and refined structures, we see that $d_{1,n}$ consistently averages over 60% of the upper bounds for this distance. This is in contrast to what is observed for the random structures, where $d_{1,n}$ drops dramatically with increasing size

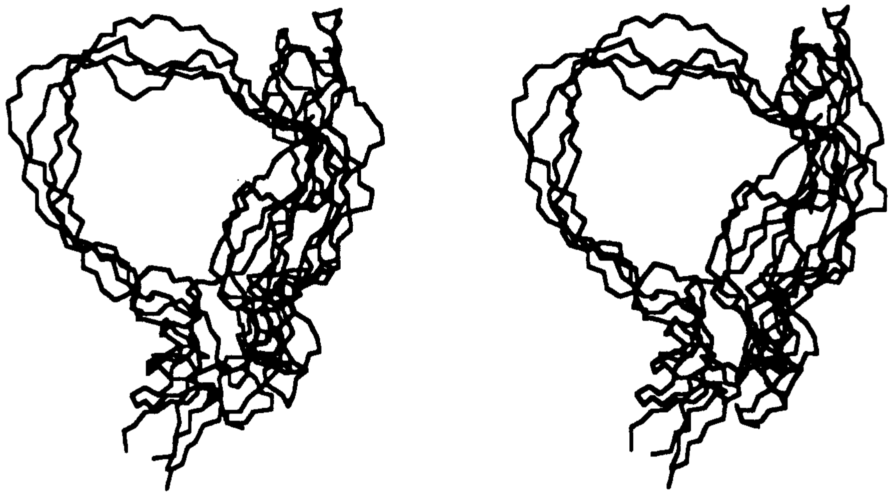

FIGURE 2: Stereoview of the superposition of the backbone (N, C^α, and C) atoms of four of the ten BPTI structures generated with the partial distance data set (see text).

of an oligopeptide to a value of 19.6% for the peptide (Lys-Glu)$_{24}$. This clearly illustrates that the embedded and refined structures are much more extended than random structures. Also listed in Table II is the range of values for $d_{1,n}$ in each set of structures. For the random structures, this range is quite large, reflecting the large variation of conformations among these structures. This is not the case for the refined structures. Here, the range is consistently small, reflecting both the extended conformation and the poor sampling of conformational space by the embed procedure. Although the observed ranges for $d_{1,n}$ for the embedded structures might appear to represent some conformational variation among the structures, these values are again artificially high due to the large bond angle and distance violations at the ends of the oligopeptide chain. The range of values of $d_{1,n}$ for the embedded structures are much more similar to those observed for the refined structures when distances on the ends of the oligopeptides are removed from the analysis (data not shown).

Reproducing Known Structures with the Distance Geometry Algorithm. The studies on the oligopeptides discussed above are being used as controls to determine the limitations of standard metric matrix distance geometry calculations in sampling conformational space, but they do not mimic the situation normally observed in calculating structures of biopolymers from NMR data. Therefore, simulations on BPTI were performed to see how the limited sampling of the embed procedure will affect the results of calculations that mimic experimental data.

Previous simulations using BPTI (Havel & Wüthrich, 1983, 1985) have demonstrated that distance geometry techniques can accurately determine the structure of a molecule when a sufficient set of "experimental"-type distances constraints are used as input. The problem is in defining what constitutes a sufficient set of distances. To address this question, a target structure was generated by regularization (Pardi et al., 1988) of the BPTI crystal structure (Deisenhofer & Steigemann, 1975) to produce a structure with essentially the same conformation as the crystal structure but with standard bond lengths and bond angles. Four sets of calculations were performed on BPTI which varied in the distance information used as input for the distance geometry program. These four calculations had input distance consisting of (1) only distances defining the covalent structure with no disulfide linkages, (2) the distances in (1) and the three disulfide linkages, (3) the distance in (2) as well as distance constraints measured from the regularized target structure that mimic those observed in actual NMR experiments (Wagner et al., 1987), and (4) all distances in (3) except any simulated experimental-type distances involving residues 39–48; the structures from these four sets of calculations will be referred to as the linear, disulfide, full distance, and partial distance BPTI structures, respectively.

Table III: rms Deviations for BPTI Structures (Å)[a]

data set	embedded	refined	target
linear	23.99	3.90	33.30
disulfide	9.28	2.09	13.36
full	2.21	2.65	3.06
partial	2.28	3.29	7.08

[a] For the embedded and refined structures the rms deviation is calculated among each respective set of 10 structures. For the target structure, the rms deviation is calculated for the best fit of each of the 10 structures of a given set to the target structure. See text for definitions of the data sets.

Table III lists the average rms distance deviations for the embedded and fully refined distance geometry structures for these four sets of calculations. As expected, the linear and disulfide BPTI structures have similar properties to the linear and disulfide-linked oligopeptides discussed above. Comparison of the rms deviations for the partial distance and full distance BPTI structures with those of the linear and disulfide BPTI structures indicates that there is very little difference between the rms deviations of the structures that contain a large number of simulated experimental-type distance constraints and those that contain no such distance constraints. Thus by the criteria of rms deviation, the linear and disulfide BPTI structures are as well-defined by the distance constraints as the partial distance and full distance BPTI structures. Table III also lists the average rms deviations from the target BPTI structure for the four sets of simulated structures. Although the rms deviations for the individual sets of structures are similar, there is no correlation between the rms distance deviations for each set of structures and the rms deviations from the target structure. Figure 2 shows a stereoview of a superposition of the partial distance BPTI structures, and the extended loop involving residues 39–48 is readily seen. From Table III it is clear that linear, disulfide, and partial distance BPTI structures do not accurately reproduce the target structure. However, if there is a reasonable amount of distance information included in the input data, such as for the full distance structures, the metric matrix distance geometry procedure quite accurately reproduces the target structure. A critical point to be made here is that the rms distance devia-

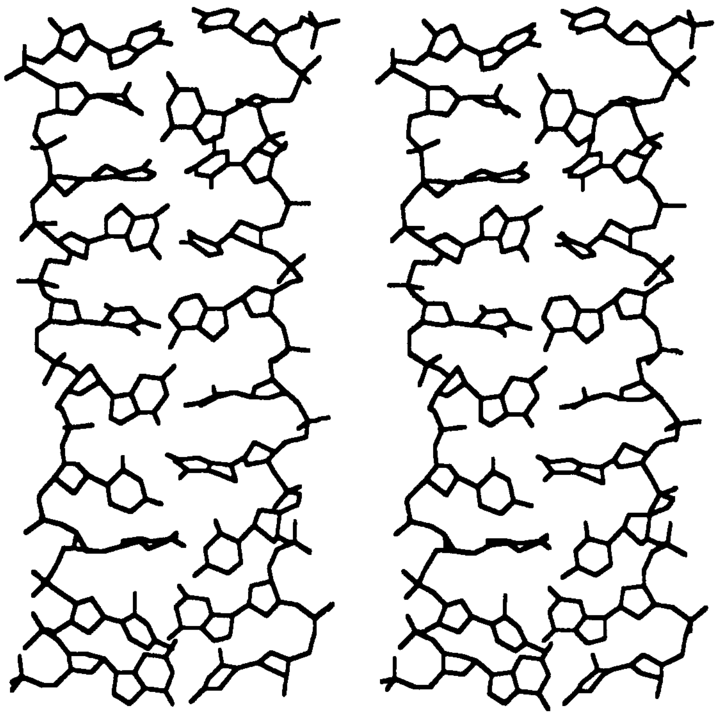

FIGURE 3: Stereoview of the non-hydrogen atoms of the double-stranded DNA decamer 5′-d(CGTCACGCGC)-3′ generated by embedding and refinement with the distance geometry algorithm. Only distances that define the covalent structure and the Watson–Crick hydrogen-bonding interactions were used as input distance information for these calculations (see text).

tions for a set of distance geometry structures cannot always be used to determine how well a structure is determined from the input distance data. Large rms distance deviations clearly indicate a poorly defined structure, but small rms deviations *do not* imply that the structure is well-defined.

Distance Geometry with Simulated Annealing. Addition of a simulated annealing type algorithm to the distance geometry procedure, as has been done in the DSPACE program (Nerdal et al., 1988), can help to alleviate some of the conformational space sampling problem in the distance geometry algorithm. Figure 1d shows the effect that application of several rounds of simulated annealing has on the conformations of the refined structures (Figure 1b). Much more variability now exists in the backbone torsion angles of the annealed structures with this variability evidenced by the larger values for the rms deviation of the annealed structures as compared with those of the refined structures (Table I). Similar results are observed in Table II, where the annealed structures are much less extended and show a wider range of values for $d_{1,n}$. However as illustrated in Tables I and II, the simulated annealed structures generated here still do not sample the complete range of conformational space available to the random structures.

Control Calculations for DNA Oligomers. Distance geometry calculations are frequently used to generate solution structures of DNA oligomers from NMR data (Patel et al., 1987; Reid, 1987). We have recently described control calculations to see how precisely and accurately the solution structure of a DNA oligomer can be reproduced from a known structure with distance data that mimic that observed in NMR experiments (Pardi et al., 1988). Here we describe control calculations where only those distances required to define the covalent structure and the hydrogen bonds in the Watson–Crick base pairs are used as input for the distance geometry algorithm. Figure 3 shows a stereoview of the non-hydrogen atoms for one of the refined structures of the DNA decamer. This DNA oligomer forms a ladder-like structure with no helical twist. The rms deviation for all non-hydrogen atoms among the 10 refined DNA structures was 3.4 Å, which is slightly higher than that observed in the distance geometry simulations (Pardi et al., 1988) but still shows that the algorithm is not sampling conformational space consistent with the input distance data.

Previous NMR studies have indicated a significant degree of underwinding for the solution structures of DNA oligomers (Patel et al., 1987). These studies also employed a distance geometry algorithm, and it is possible that this underwinding arises at least partially from the tendency of the distance geometry algorithm to produce extended structures consistent with the input distance information. For DNA double helices a more extended helix will have a smaller helical twist and therefore would be underwound with the extreme being the ladder-like structures shown in Figure 3.

Origin of the Tendency for the Distance Geometry Algorithm To Produce Extended Structures. Table II shows that the standard metric matrix distance geometry algorithm produces much more extended structures for a molecule than structures in random conformations. This problem arises primarily from the choice of the distribution used to pick distances between the upper and lower bounds in the distance matrix. The step before the embeddding procedure in the distance geometry algorithm involves picking a random distance between the upper and lower bounds for each pair of atoms in the smoothed distance bounds matrix (Havel et al., 1979, 1983). This selection assumes that in an ensemble of structures with random conformations there is a uniform distribution of distances between the upper and lower bounds for each pair of atoms in the molecule. Polymer theory shows that this is an incorrect distribution for freely rotating chains and the correct distribution of distances will be skewed toward the lower bonds (Cantor & Schimmel, 1980; Flory, 1969), as observed for the random structures in Table II. This skewing arises simply because there is only a single conformation where the distance between two pairs of atoms far apart in the primary structure will be equal to the upper bound, but there will be many conformations with distances equal to the lower bound. This leads to the results shown in Table II where $d_{1,n}$ is 19% of the upper bound for $(\text{Lys-Glu})_{24}$ for the random structures whereas a random selection of distances between the upper and lower bounds would give a value for $d_{1,n}$ of approximately 50% of the upper bound. Thus the extendedness observed in the embed and refined structures can be corrected by using a more appropriate distribution of distances. Havel et al. (1983) previously noted that the uniform distribution may not be appropriate when there are fairly large differences between the upper and lower bounds and suggested the use of distributions that mimic random structures. We are presently testing various distributions including ones that have a random distribution of distances for small upper bounds (<5 Å) and distributions that mimic those produced by freely rotating chains for larger upper bounds.

Another source of extended structures arises from the lack of energy terms in the distance geometry algorithm. The van der Waals interaction in the DSPACE program is accounted for by a simple hard-sphere model, usually corresponding to the lower bound between atoms, and thus there is no attractive interaction. With no attractive force to help bring atoms together, the structures produced by the distance geometry algorithm will tend to be more expanded than structures produced by algorithms that incorporate a Lennard-Jones-type potential. However, such a potential is quite short range, and although it could account for a slightly expanded structure, it does not readily account for the highly extended structures observed on the linear peptides shown in Figure 1a and Table II.

Origins and Possible Solutions for the Limited Sampling of the Embed Procedure. The results discussed in the previous sections clearly show that the embed procedure of the metric matrix distance geometry algorithm does not adequately sample conformational space. We were concerned as to whether this problem arose from the implementation of the embed algorithm in the DSPACE program or whether it was a fundamental property of the standard embed procedure. To test this question, embeds of short linear peptides were performed with an independent metric matrix distance geometry program, a recent version of a distance geometry program from Professor I. D. Kuntz's group (UCSF). The results with this program were essentially the same as those shown in Table I and Figure 1. Therefore, it is clear that the tendency to produce similar (and extended) structures consistent with the input distance matrix is a general property of the standard embed procedure.

Analysis of the embed procedure on the control calculations with the linear peptides indicates that the sampling problem arises primarily from the type of distance bounds matrix produced by NMR data. The elements in the bounds matrix consist of upper and lower distance bounds between all pair of atoms in the molecule. In the limit of little distance information, besides that used to define the covalent structure of a molecule, most elements in the smoothed distance bounds matrix have large differences between their upper and lower bound. The first step in the embed procedure is to pick a random value between the upper and lower bound for each element in the distance matrix, and since random values are chosen, there will be many very large inconsistencies between pairs of distances. It is this averaging of a very large number of inconsistent distances that leads to the results given in Table II, where the observed average distance after the embed is approximately 60% of the upper bound with very little variance among the structures. Thus the sampling problem in the distance geometry algorithm should be significantly reduced by application of the following procedures: (i) production of a very self-consistent distance matrix for the embed procedure and (ii) generation of an ensemble of different self-consistent distance matrixes that span the range of distances consistent with the input data.

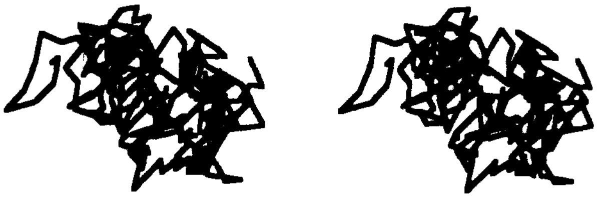

FIGURE 4: Stereoview of the superposition of the backbone (N, C^α, and C) atoms of four of the ten submatrix embedded structures (see text) of the oligopeptide $(\text{Lys-Glu})_{12}$. The structures are shown on the same scale as in Figure 1.

Results from Kuntz and co-workers indicate that appropriate application of submatrix embedding (Havel & Wüthrich, 1984) appears to reduce significantly the sampling problem in the distance geometry algorithm (Thomason & Kuntz, 1989). This submatrix embedding has been incorporated into the DSPACE algorithm, and the results are shown in Tables I and II. The average rms deviations for submatrix embedded structures of the control peptides given in Table I are larger than the rms deviations for the annealed structures but still slightly less than the values for the random structures. However, the extendedness of the structures produced by the submatrix embed procedure closely parallels the random structures with these submatrix embedded structures actually being on the average 20% less extended than the random structures. Figure 4 shows a stereoview of the superposition of the backbone atoms for a set of four of the submatrix embedded structures. Comparison of these structures with the structures in Figure 1 indicates that the submatrix embedding procedure significantly improves the sampling of conformational space by the distance geometry algorithm. In the submatrix embedding performed here, all the C^α atoms in the peptides were used as the subset of atoms. Kuntz and co-workers are more extensively investigating how variations in the number of atoms used in the submatrix embed affect the sampling of the distance geometry algorithm (I. D. Kuntz, personal communication). We note that submatrix embedding is not a direct solution to the sampling problem in the distance geometry algorithm, because it does not produce a self-consistent data matrix for the embed. However, the results presented here, along with those of Thomason and Kuntz (1989), clearly indicate this procedure leads to a great improvement in the sampling of conformation space as compared with a standard embed.

The embed procedure involves diagonalization of the distance matrix, but as discussed above, there are often a huge number of inconsistent distances in this matrix. Therefore, another possible solution to the sampling problem would be to perform the diagonalization by heavily weighting those elements that contain atoms which are directly connected by a chemical bond or an experimental NOE. Thus the embed would be produced from a much more self-consistent data set. We are presently trying to implement this procedure in the DSPACE program to see how it affects the limited sampling problem.

One method for truly sampling all conformations consistent with the input distances is to perform a grid search-type procedure in angle space. This simple search procedure may be useful on very small molecules, but it is not efficient enough for larger systems. Marshall and co-workers have developed search algorithms that can truncate the systematic search if whole regions of conformational space are ruled out by the input distance constraints (Marshall et al., 1989). The time required for this type of search decreases tremendously as one adds more distance constraints to the system. Therefore this may be a viable approach for generating structures of peptides or proteins that have a large number of long-range distance constraints. Systematic search procedures can also be used to study the conformations of small regions of a molecule where the local structure is defined mainly by local distances. This method may prove useful for probing conformational space in nucleic acid double helices where the structure of the molecule is determined solely from distances involving local (nearest-neighbor) interactions.

Conclusions

A fundamental property of the present procedures for generating structures of molecules from NMR data is that there is no direct criteria for determining if the structure is overdetermined from the experimental data. Therefore, one must include a step in the structure determination process that tests how well the structure is defined from the input data. A statistical approach is normally used for this step and generally involves analysis of the differences among an ensemble of structures. For this statistical approach to be meaningful, one must know that the structure generation procedure randomly samples the parameters of interest.

This work describes control calculations that test the sampling properties of the distance geometry algorithm. Control calculations were performed with only those distances required to define the covalent structure of the molecule. Thus if an algorithm adequatley sampled conformational space, it would generate structures with essentially random conformations. The results discussed here clearly show that the present implementations of the metric matrix distance geometry algorithm do not adequately sample conformational space. Our results also strongly indicate that this is a problem with the present implementations of the embed procedure in the distance geometry algorithm and is not a fundamental property of the algorithm itself. The limited sampling problem seems to arise primarily from the way the distance matrix is generated for the embed step in the distance geometry algorithm. In the limit of little input distance information, the standard procedure produces a very non-self-consistent data matrix with a distribution of distances skewed toward longer values. We are presently evaluating various methods for producing self-consistent data matrixes with more realistic distribution of distances which should lead to improved sampling by the distance geometry algorithm.

Comparative studies have indicated that other computational techniques such as restrained molecular dynamics, simulated annealing, or minimization in torsional angle space (Havel & Wüthrich, 1985; Nigles et al., 1988b; Brünger et al., 1987; de Vlieg et al. 1988) have a tendency to search a less limited region of conformational space than does the distance geometry algorithm. These studies concentrated on the limited sampling for local regions of the molecule and seemed to assume there was always sufficient distance information to correctly define global conformation of the molecule. However, the limited sampling in the distance geometry algorithm is not restricted to local conformations and will adversely affect the global conformation of the molecule if there is not enough distance information. Thus care should be taken when hybrid methods such as distance geometry and restrained molecular dynamics, simulated annealing, or Monte Carlo (Nigles et al., 1988b; Clore et al., 1987a,b; Bassolino et al., 1988) are used to generate global conformations of molecules in the limit of insufficient input distance data.

Although the distance geometry algorithm accurately and precisely reproduces known structures, given a sufficient amount of distance information (Havel & Wüthrich, 1985; Havel et al., 1979), a problem arises for unknown structures where it is generally not possible to define what is a "sufficient" number of distances. The results presented here show that the standard implementation of the distance geometry algorithm gives very misleading results if there is not enough input distance information because it produces structures that incorrectly appear to be very well-defined (even in the limit of no noncovalent distance information). The results of these calculations also show that use of rms deviations for judging the quality of the structures produced by the distance geometry algorithm can be very misleading.

Added in Proof

After acceptance of our manuscript, we became aware of a complementary manuscript on the sampling properties of the metric matrix distance geometry algorithm (Havel, 1989). Significantly improved sampling has been achieved by application of a procedure where the trial distances are not chosen independently thus leading to a very self-consistent distance matrix for the embed algorithm.

Acknowledgments

We thank R. D. Morrison, X. L. Zhang, A. J. Wand, J. F. Thomason, P. L. Weber, C. Oshiro, R. M. Levy, and I. D. Kuntz for helpful discussions and C. Oshiro for performing distance geometry calculations for us.

References

Bassolino, D. A., Hirata, F., Kitchen, D., Pardi, A., & Levy, R. M. (1988) *Int. J. Supercomput. Appl. 2*, 41–61.

Braun, W. (1987) *Q. Rev. Biophys. 19*, 115–157.

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

Brünger, A. T., Clore, G. M., Gronenborn, A. M., & Karplus, M. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 3801–3805.

Cantor, C. R., & Schimmel, P. R. (1980) in *Biophysical Chemistry*, pp 992–1010, Freeman, New York.

Clore, G. M., Gronenborn, A. M., Nilges, M., Sukumaran, D. K., & Zarbock, J. (1987a) *EMBO J. 6*, 1833–1842.

Clore, G. M., Sukumaran, D. K., Nilges, M., Zarbock, J., & Gronenborn, A. M. (1987b) *EMBO J. 6*, 529–537.

Crippen, G. M. (1977) *J. Comp. Phys. 26*, 449–452.

Crippen, G. M. (1981) in *Distance and Conformational Calculations*, Research Studies Press, Wiley, Chichester, England.

Deisenhofer, J., & Steigemann, W. (1975) *Acta Crystallogr., Sect. B 31*, 231–250.

de Vlieg, J., Scheek, R. M., Vangunsteren, W. F., Berendsen, H. J. C., Kaptein, R., & Thomason, J. (1988) *Proteins 3*, 209–218.

Fesik, S. W., O'Donnell, T. J., Gampe, R. T., Jr., & Olejniczak, E. T. (1986) *J. Am. Chem. Soc. 108*, 3165–3170.

Flory, P. J. (1969) in *The Statistical Mechanics of Chain Molecules*, Wiley, New York.

Hare, D. R., & Reid, B. R. (1986) *Biochemistry 25*, 5341–5350.

Havel, T. F. (1989) *Biopolymers* (submitted for publication).

Havel, T. F., & Wüthrich, K. (1984) *Bull. Math. Biol. 46*, 673–698.

Havel, T. F., & Wüthrich, K. (1985) *J. Mol. Biol. 182*, 281–284.

Havel, T. F., Kuntz, I. D., & Crippen, G. M. (1979) *Biopolymers 18*, 73-81.

Havel, T. F., Kuntz, I. D., & Crippen, G. M. (1983) *Bull. Math. Biol. 45*, 665-720.

Holak, T. A., Prestegard, J. H., & Forman, J. D. (1987) *Biochemistry 26*, 4652-4660.

Kaptein, R., Zuiderweg, E. R. P., Scheek, R. M., Boelens, R., & van Gunsteren, W. F. (1985) *J. Mol. Biol. 182*, 179-182.

Marshall, G. R., Beusen, D. D., Iijima, H., Karasek, S. F., Shands, B., & Dammkoeler, R. A. (1989) *J. Cell. Biochem. Suppl. 13A*, 23.

Nerdal, W., Hare, D. R., & Reid, B. R. (1988) *J. Mol. Biol. 201*, 717-739.

Nilges, M., Clore, G. M., & Gronenborn, A. M. (1988a) *FEBS Lett. 229*, 317-324.

Nilges, M., Clore, G. M., & Gronenborn, A. M. (1988b) *FEBS Lett. 229*, 129-136.

Nyburg, S. C. (1974) *Acta Crystallogr., Sect. B. 30*, 251-253.

Pardi, A., Hare, D. R., & Wang, C. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 8785-8789.

Patel, D. J., Shapiro, L., & Hare, D. (1987) *Annu. Rev. Biophys. Biophys. Chem. 16*, 423-454.

Reid, B. R. (1987) *Q. Rev. Biophys. 20*, 1-34.

Thomason, J. F., & Kuntz, I. D. (1989) *J. Cell. Biochem. Suppl. 13A*, 37.

Wagner, G., Braun, W., Havel, T. F., Schaumann, T., Go, N., & Wüthrich, K. (1987) *J. Mol. Biol. 196*, 611-639.

Wüthrich, K. (1986) in *NMR of Proteins and Nucleic Acids*, Wiley, New York.

Wüthrich, K. (1989) *Science 243*, 45-50.

Vibrational Spectroscopy of Bacteriorhodopsin Mutants: Chromophore Isomerization Perturbs Tryptophan-86[†,‡]

Kenneth J. Rothschild,[*,§] Daniel Gray,[§] Tatsushi Mogi,[‖,⊥] Thomas Marti,[‖] Mark S. Braiman,[§,#]
Lawrence J. Stern,[‖] and H. Gobind Khorana[‖]

Physics Department and Program in Cellular Biophysics, Boston University, Boston, Massachusetts 02215, and Departments of Chemistry and Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received January 30, 1989; Revised Manuscript Received May 2, 1989

ABSTRACT: Fourier transform infrared difference spectra have been obtained for the bR → K and bR → M photoreactions of bacteriorhodopsin mutants with Phe replacements for Trp residues 10, 12, 80, 86, 138, 182, and 189 and Cys replacements for Trp residues 137 and 138. None of the tryptophan mutations caused a significant shift in the retinylidene C=C or C—C stretching frequencies of the light-adapted bR$_{570}$ state. Since these frequencies are known to be strongly correlated with the visible absorption maximum of the chromophore, it is concluded that none of the tryptophan residues are essential for forming a normal bR$_{570}$ chromophore. However, a 742-cm^{-1} negative peak attributed previously to the perturbation of a tryptophan residue during the bR → K photoreaction was found to be absent in the bR → K and bR → M difference spectra of the Trp-86 mutant. On this basis, we conclude that the structure or environment of Trp-86 is altered during the bR → K photoreaction. All of the other Trp → Phe mutants exhibited this band, although its frequency was altered in the Trp-189 → Phe mutant. In addition, the Trp-182 → Phe mutant exhibited much reduced formation of normal photoproducts relative to the other mutants, as well as peaks indicative of the presence of additional chromophore conformations. A model of bR is discussed in which Trp-86, Trp-182, and Trp-189 form part of a retinal binding pocket. One likely function of these tryptophan groups is to provide the structural constraints needed to prevent chromophore photoisomerization other than at the C_{13}=C_{14} double bond.

Understanding light-driven proton transport in bacteriorhodopsin (bR)[1] remains an important problem in biology and biophysics. This 26 000-dalton protein found in the purple membrane (PM) of *Halobacterium halobium* contains a retinylidene chromophore. Due to bR's relative simplicity, it has been extensively investigated by using a wide variety of biochemical and biophysical techniques (Stoeckenius & Bogomolni, 1982). Recent success in the isolation of site-directed mutant forms of bR (Nassal et al., 1987; Braiman et al., 1987) has made it possible to study the role of individual residues in the bR proton pump mechanism (Hackett et al., 1987; Mogi et al., 1987, 1988, 1989; Khorana, 1988; Ahl et al., 1988).

Several studies have focused on the possible involvement of tryptophan residues in bR proton pumping and color regulation [e.g., Maggiora and Schowen (1977)]. Fluorescence and NMR techniques have detected tryptophan residues interacting

[†] Supported by grants from NSF (PCM-8110992), NIH (GM-28289-06), and ONR (N00014-82-K-0668) to H.G.K. and from NSF (DMB-8509587) and ONR (N00014-88-K-0464) to K.J.R. M.S.B. is a Lucille P. Markey Scholar, and this work was supported in part by a grant from the Lucille P. Markey Charitable Trust. Th.M. is the recipient of a fellowship from the Swiss National Science Foundation.

[‡] This is paper 3 in a series titled "Vibrational Spectroscopy of Bacteriorhodopsin Mutants". Papers 1 and 2 in this series are Braiman et al. (1988a,b).

[*] Address correspondence to this author.

[§] Boston University.

[‖] Massachusetts Institute of Technology.

[⊥] Present address: Department of Biology, Faculty of Science, University of Tokyo, Hongo, Tokyo 113, Japan.

[#] Present address: Department of Biochemistry, University of Virginia Medical School, Charlottesville, VA 22908.

[1] Abbreviations: PM, purple membrane; bR, bacteriorhodopsin; bO, bacterioopsin; hR, halorhodopsin; FTIR, Fourier transform infrared; au, absorbance units; HOOP, hydrogen out-of-plane.

0006-2960/89/0428-7052$01.50/0