Journal of Biomolecular NMR, 3 (1993) 19–40
ESCOM

J-Bio NMR 099

Intrinsic nature of the three-dimensional structure of proteins as determined by distance geometry with good sampling properties

Takahisa Nakai, Akinori Kidera and Haruki Nakamura*

Protein Engineering Research Institute, 6-2-3, Furuedai, Suita, Osaka, 565, Japan

Received 12 August 1992
Accepted 12 October 1992

Keywords: Distance geometry; Protein structure; Simulated annealing; Sampling properties

SUMMARY

A protocol for distance geometry calculation is shown to have excellent sampling properties in the determination of three-dimensional structures of proteins from nuclear magnetic resonance (NMR) data. This protocol uses a simulated annealing optimization employing mass-weighted molecular dynamics in four-dimensional space (Havel, T.F. (1991) *Prog. Biophys. Mol. Biol.*, **56**, 43–78). It attains an extremely large radius of convergence, allowing a random coil conformation to be used as the initial estimate for the succeeding optimization process. Computations are performed with four systems of simulated distance data as tests of the protocol, using an unconstrained L-alanine 30mer and three different types of proteins, bovine pancreatic trypsin inhibitor, the α-amylase inhibitor Tendamistat, and the N-terminal domain of the 434-repressor. The test of the unconstrained polypeptide confirms that the sampled conformational space is that of the statistical random coil. In the larger and more complicated systems of the three proteins, the protocol gives complete convergence of the optimization without any trace of initial structure dependence. As a result of an exhaustive conformational sampling by the protocol, the intrinsic nature of the structures generated with distance restraints derived from NMR data has been revealed. When the sampled structures are compared with the corresponding X-ray structures, we find that the averages of the sampled structures always show a certain pattern of discrepancy from the X-ray structure. This discrepancy is due to the short distance nature of the distance restraints, and correlates with the characteristic shape of the protein molecule.

INTRODUCTION

Protein structures in solution have been determined from distance and torsion angle restraints obtained by multi-dimensional nuclear magnetic resonance (NMR) experiments (Wüthrich, 1986; Braun, 1987; Kaptein et al., 1988; Clore and Gronenborn, 1991). There are several methods of

*To whom correspondence should be addressed.

Abbreviations: r.m.s.d., root-mean-square deviation; MD, molecular dynamics; NMR, nuclear magnetic resonance; NOE, nuclear Overhauser enhancement; BPTI, bovine pancreatic trypsin inhibitor.

determining three-dimensional (3D) structures; the metric matrix algorithm in Cartesian space (Crippen, 1981; Havel et al., 1983; Havel and Wüthrich, 1984; Crippen and Havel, 1988; Kuntz et al., 1989; Havel, 1991), the variable target function algorithm in the torsion angle space (Braun and Gō, 1985), and the dynamical simulated annealing protocol (Nilges et al., 1988, Clore and Gronenborn, 1989). Each method produces a protein structure as an ensemble of conformations, all of which satisfy the experimental restraints. The reliability of the distance geometry structures should be evaluated by the sampling properties of an algorithm, i.e., whether the sampled structures cover all the allowed conformational space and whether they have an unbiased distribution.

The sampling properties of these algorithms have been investigated (Havel and Wüthrich, 1985; Wagner et al., 1987; Nilges et al., 1988; Metzler et al., 1989; Havel, 1990, 1991; Kuszewski et al., 1992; Liu et al., 1992) and it has been found that the original metric matrix algorithm (Crippen, 1981; Havel et al., 1983) searches a rather limited and biased conformational space. The poor sampling properties appear most distinctly in unconstrained polypeptide chains (Metzler et al., 1989; Havel, 1990; Kuszewski et al., 1992) where sampled structures assume largely extended conformations with small root-mean-square deviations (r.m.s.d.).

Usually a trace of the initial structure remains even after the optimization process. The original metric matrix algorithm tends to generate biased initial structures. This problem has been solved by a procedure called randomized metrization (Havel, 1990; Kuszewski et al., 1992), which affords a sufficiently wide variety of initial structures. In unconstrained L-alanine polymers, randomized metrization produces an ensemble of conformations whose distribution is almost equivalent to that of self-avoiding random coils.

When an optimization method is powerful enough to give a large radius of convergence, a random coil conformation can be used as the initial estimate for the succeeding optimization process. The variable target function algorithm and the dynamic simulated annealing protocol contain optimization procedures with large radii of convergence. Among them, an optimization technique proposed by Havel (1991), which uses mass-weighted molecular dynamics (MD) in four-dimensional (4D) space, is extremely powerful. In this method, a random coil can be used as the initial structure to avoid biased sampling.

This article investigates the sampling properties of a distance geometry calculation that uses mass-weighted MD in 4D space. For this purpose, we prepared two simulated systems, an unconstrained L-alanine 30mer and a small globular protein, bovine pancreatic trypsin inhibitor (BPTI). These two systems have been used intensively as tests of the distance geometry algorithm and have yielded valuable information about the sampling properties (Havel and Wüthrich, 1985; Havel, 1990, 1991; Oshiro et al., 1991; Kuszewski et al., 1992). In the latter case, NMR data are simulated from the X-ray structure (Wlodawer et al., 1984) in the Protein Data Bank (Bernstein et al., 1977). The sampling properties can be clearly observed in the comparison of the sampled structures with the X-ray structure. In the Results section, the analyses of the sampling properties are described.

Until now, the sampling problem has prevented us from discussing whether the short proton–proton distances observed in the NMR data are sufficient to determine the 3D structure of a protein in solution. What would occur within the sampled structures in a situation where the long distance information was lacking? In the discussion, we attempt to answer this question, using two different types of globular proteins: the α-amylase inhibitor Tendamistat (Pflugrath et al., 1986) and the N-terminal domain of the 434-repressor (Mondragon et al., 1989).

METHODS

(1) Distance geometry program EMBOSS

Distance geometry calculations were performed with the program system EMBOSS (EMBedding and Optimization of biomolecular Structures on Supercomputers). The program system consists of three modules, each having the following function: (1) preparation of a distance-bound matrix from NMR data; (2) construction of the initial coordinates for the following optimization, either by generating a random coil conformation or by the embedding algorithm (Crippen, 1981; Crippen and Havel, 1988; Kuntz et al., 1989; Havel, 1991); and (3) optimization of the structure by simulated annealing. In the embedding process, the randomized metrization algorithm (Havel, 1990; Kuszewski et al., 1992) is implemented to enhance conformational sampling. Simulated annealing is performed by mass-weighted MD in 4D space (Havel, 1991), and attains an extremely large radius of convergence. The program is highly vectorized to give optimal performance on a supercomputer. High-speed computation allows one to sample a sufficiently large number of conformations in macromolecular systems. Inquiries about the availability and distribution of the EMBOSS program should be addressed to the authors.

(i) Preparation of a distance-bound matrix

Distance bounds are prepared from NMR data together with 1-2, 1-3, and 1-4 distances implicated from the covalent geometry of the molecule. The parameters of the covalent geometry used are those of AMBER (Weiner et al., 1986), after modification for the pseudo-structure (Wüthrich et al., 1983). Distance bounds are in the form of upper and lower values. All the undefined upper distance bounds are initially set to a large value (999.0 Å), and the lower distance bounds, set without experimental information, are the sum of the van der Waals radii (Kuntz et al., 1989) of the corresponding atoms. The distance-bound matrix thus prepared is used in the embedding and optimization procedures.

(ii) Embedding algorithm

The conventional metric matrix method (Crippen, 1981; Havel et al., 1983; Kuntz et al., 1989) and randomized metrization (Havel, 1990; Kuszewski et al., 1992) are available in EMBOSS.

Before embedding the distance matrix in 3D space, undefined distance bounds are refined by bound smoothing based on triangle inequalities (Havel et al., 1983). For this purpose, EMBOSS adopts Floyd's shortest-path algorithm (Aho et al., 1983; Dress and Havel, 1988), which is suitable for vectorization. In the conventional method, trial distances are generated randomly between the refined upper and lower distance bounds, with a uniform distribution. We call this procedure *no metrization* in this paper.

However, the conventional method (*no metrization*) described above has been found to cause biased and insufficient sampling (Wagner et al., 1987; Nilges et al., 1988; Metzler et al., 1989). Havel (1990) developed randomized metrization for bound smoothing in order to improve the sampling properties of the embedding algorithm. Since *no metrization* chooses trial distances independently of the other distances, the resultant distances have no guarantee of satisfying triangle inequalities. In fact, metrization ensures the self-consistency of all triangle inequalities by the following procedures: (1) a pair of atoms is randomly chosen and the distance between them

is set to a value between the upper and lower bounds; (2) these bounds are then revised to be equal to the trial distance; (3) using this new distance information, all distance bounds are smoothed again. These procedures are repeated until all distances are set. EMBOSS performs randomized metrization based on Dijkstra's algorithm (Aho et al., 1983). Here, we simply call this *metrization.*

The trial distance matrix is transformed into the metric matrix and then embedded in 4D space (Crippen and Havel, 1978, 1988; Crippen, 1981), to yield the starting structure for the optimization. EMBOSS also generates the starting structure by choosing torsion angles randomly. Hereinafter, this is referred to as *random coil.*

(iii) Penalty functions for optimization

The penalty function E_{tot} for simulated annealing in 4D space does not include any conformational energies, but it is made up of the distance restraint, E_{dist} (Kuntz et al., 1989), the chirality restraints, E_{chi} (Havel et al., 1983; Havel, 1991) and E_{cms}, and the restraint for the 4D coordinates, E_{4D} (Kuntz et al., 1989).

$$E_{tot} = E_{dist} + E_{chi} + E_{cms} + E_{4D} \tag{1}$$

The distance restraint term, E_{dist}, is given by

$$E_{dist} = \begin{cases} \sum k_{dist}\left(\dfrac{B_{ij}^{u2} - D_{ij}^2}{B_{ij}^{u2}}\right)^2, & \text{if } B_{ij}^u < D_{ij}; \\ 0, & \text{if } B_{ij}^l \le D_{ij} \le B_{ij}^u; \\ \sum k_{dist}\left(\dfrac{B_{ij}^{l2} - D_{ij}^2}{B_{ij}^{l2}}\right)^2, & \text{if } B_{ij}^l > D_{ij} \end{cases} \tag{2}$$

where the summation is over all of the lower distance bounds, B_{ij}^l, and the upper bounds, B_{ij}^u, defined explicitly by either the NMR data or the covalent geometry. The distance of the model structure, D_{ij}, is defined in 4D space, $D_{ij}^2=(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2+(w_i-w_j)^2$, where w_i and w_j are the 4D coordinates.

The chirality restraints contain two kinds of penalty functions, E_{chi} and E_{cms}. The former is the chiral volume restraint to be applied to local chiral centers (e.g., α-carbon atoms) and planar groups (e.g., peptide bonds and aromatic rings) (Havel et al., 1983; Havel, 1991),

$$E_{chi} = \begin{cases} \sum k_{chi}(C_{ijkl}^u - F_{ijkl})^2, & \text{if } C_{ijkl}^u < F_{ijkl}; \\ 0, & \text{if } C_{ijkl}^l \le F_{ijkl} \le C_{ijkl}^u; \\ \sum k_{chi}(C_{ijkl}^u - F_{ijkl})^2, & \text{if } C_{ijkl}^l > F_{ijkl} \end{cases} \tag{3}$$

where C_{ijkl}^l and C_{ijkl}^u are the lower and upper limits of the chiral volume for a tetrahedron (i, j, k, l), respectively, and F_{ijkl} is the associate chiral volume calculated from the structure. It is noted that E_{chi} is evaluated in 3D space described by the coordinates x_i, y_i, z_i, because the chiral volume is calculated from a triple product defined only in 3D space.

The penalty function, E_{cms}, is the shape restraint to place an asymmetric atom at the center of a tetrahedron given by

$$E_{cms} = \Sigma\ k_{cms}\left(\mathbf{r}_c - \frac{\mathbf{r}_1+\mathbf{r}_2+\mathbf{r}_3+\mathbf{r}_4}{4}\right)^2 \quad (4)$$

where $\mathbf{r}_c$ is the coordinate of the asymmetric atom and $\mathbf{r}_1,...,\mathbf{r}_4$, are those of the four atoms defining the tetrahedron. To make it consistent with E_{chi}, E_{cms} is also calculated in 3D space described by the coordinates (x_i, y_i, z_i). The reason for the incorporation of E_{cms} is to maintain not only the chiral volume but also the shape of a correct tetrahedron. The restraint function, $E_{chi}+E_{cms}$, allows almost all the chiral centers to be correctly maintained during the course of the optimization.

The weight parameters, k_{dist}, k_{chi}, and k_{cms}, are 5.0, 0.1, and 0.1 kcal/mol, respectively, throughout all stages of the optimization.

At the final stage of the optimization, the system in 4D space should be recovered to 3D space by reducing E_{4D} to zero.

$$E_{4D} = \Sigma\ k_{4D}w_i^2 \quad (5)$$

This can be achieved by increasing the value of k_{4D} from the initial value of 0.05 kcal/mol to the final value of 5.0 kcal/mol throughout the final minimization stage.

(iv) Protocol of simulated annealing

The conformational space is searched by using the combination of two devices, the 4D space and the mass-weight MD (Havel, 1991). Increasing dimensionality flattens the potential surface to reduce the local minima (Crippen, 1982; Purisima and Scheraga, 1987), and the large mass weights (1000 Da for all atoms in this work) make a long-time simulation possible (step size = 50 fs in this work). EMBOSS implements the temperature-regulated MD algorithm (Berendsen et al., 1984) to control the temperature in the simulated annealing process. The protocol for the optimization is given in Table 1. The initial minimization stage is followed by a heating stage of

TABLE 1
PROTOCOL OF SIMULATED ANNEALING OPTIMIZATION

Stage 1:	500-steps conjugate gradient minimization k_{dist} = 5.0, k_{chi} = k_{cms} = 0.1, k_{4D} = 0.05
Stage 2:	7000-steps molecular dynamics at 300 K k_{dist} = 5.0, k_{chi} = k_{cms} = 0.1, k_{4D} = 0.05 atomic mass 1000 Da, step size 50 fs, coupling time constant[a] 0.5 ps
Stage 3:	3000-steps slow-cooling molecular dynamics to 10 K k_{dist} = 5.0, k_{chi} = k_{cms} = 0.1, k_{4D} = 0.05 atomic mass 1000 Da, step size 50 fs, coupling time constant 0.5 ps, cooling rate 20 K/100 steps
Stage 4:	1000-steps conjugate gradient minimization k_{dist} = 5.0, k_{chi} = k_{cms} = 0.1, k_{4D} = 0.05

[a] Coupling time constant is used in the temperature-regulated molecular dynamics algorithm (Berendsen et al., 1984).

7000-step MD (350 ps) at 300 K. The system is then gradually cooled down to 10 K in 3000-step MD (150 ps). Finally, the weight, k_{4D}, of E_{4D} is increased to compress the fourth coordinate, w_i, to zero (Eq. 5).

(2) Simulation of the NMR distance data

In order to examine the sampling properties of the protocols implemented in EMBOSS, we used several sets of distance restraints with different qualities simulated from the X-ray structure of BPTI, 5PTI (Wlodawer et al., 1984) of the Protein Data Bank (Bernstein et al., 1977). With these simulated data sets, it is possible to compare the calculated structures, using the X-ray structure as the solution to the problem.

Prior to calculation of the distances between hydrogen atoms, the X-ray structure, including all hydrogen atoms, is regularized to the AMBER standard geometry (Weiner et al., 1986) by energy minimization with the molecular simulation program PRESTO (Morikami et al., 1992). The definitions of the simulated distances are the same as those of Havel and Wüthrich (1985). A pair

TABLE 2
SIMULATED NMR DATA SETS FOR THE STRUCTURAL CALCULATIONS

Data set[a]	Total number[b]	Sequential[c]		Medium range[d]		Long range[e]	
		Number	Upper bound[f]	Number	Upper bound	Number	Upper bound
I	533(136)	43	2.5	14	2.5	24	2.5
		33	3.0	28	3.0	62	3.0
		39	4.0	132	4.0	158	4.0
II	393(130)	43	2.5	14	2.5	24	2.5
		33	3.0	18	3.0	39	3.0
		39	4.0	76	4.0	107	4.0
III	393(130)	43	2.5	108	5.0	170	5.0
		33	3.0				
		39	4.0				
IV	393(130)	43	2.5	108	4.0	170	4.0
		33	3.0				
		39	4.0				
IX	171(62)	43	2.5	32	4.0	63	4.0
		33	3.0				

[a] Data sets are obtained from the crystal structure of BPTI (5PTI) (Wlodawer et al., 1984) according to the definition described by Havel and Wüthrich (1985). The names I–IX of the data sets correspond to those of Havel and Wüthrich (1985).

[b] Total number indicates the total number of distance restraints for each data set. The number in parentheses indicates the number of distance restraints between backbone protons.

[c] Sequential includes the distances, $d_{\alpha N}$, d_{NN}, and $d_{\beta N}$, which are, respectively, the distances between the $C^{\alpha}H$ of residue i and the HN of residue i+1, between HN_i and HN_{i+1} and between $C^{\beta}H_i$ and HN_{i+1}.

[d] Medium range comprises distances between protons located within a pentapeptide segment ($|i-j|<5$).

[e] Long range comprises distances between protons in the residues separated by at least four intervening residues ($|i-j|\geq 5$).

[f] Upper bounds of the distances shorter than 2.5 Å are set to 2.5 Å, while those of the distances from 2.5 to 3.0 Å are set to 3.0 Å, and those of the distances from 3.0 to 4.0 Å are set to 4.0 Å. All lower bounds are set to the sum of the van der Waals radii of the corresponding atoms.

of proton atoms belonging to NH, CH, CH_2, CH_3, and an aromatic group in different residues is chosen if the distance between the two is less than 4.0 Å. For CH_2, CH_3, and an aromatic group, a pseudoatom representation (Wüthrich et al., 1983) is adopted. The shortest distance between a pair of these hydrogen atoms is chosen and imposed upon their corresponding pseudoatoms. Table 2 summarizes the numbers and the classifications of the distance restraints. Data set I is the most precise, and includes the complete list of 533 inter-residue proton-proton distances derived from the X-ray structure. Data set II mimics an ordinary level of experimental NMR data. Data sets III and IV have the same number of distance restraints as data set II, but have different upper limits for the medium- and long-range distances. Comparison between II, III, and IV reveals that the precision of the distances has an effect on the sampling properties. Data set IX mimics the NOE data at an early stage of an NMR experiment. The numbers of data sets I–IV and IX correspond to those of Havel and Wüthrich (1985).

RESULTS

The quality of distance geometry calculation can be examined by considering the sampling properties: whether it covers all the allowed conformational space and whether it is unbiased. We investigated the sampling properties of the protocols described in the Methods section with two systems of simulated distance data, unconstrained L-alanine 30mers (Metzler et al., 1989; Havel, 1990; Kuszewski et al., 1992) and BPTI (Havel and Wüthrich, 1985; Havel, 1991; Oshiro et al., 1991). The former has no distance restraints, except for those derived from the covalent geometry, and forms a self-avoiding polymer chain that provides a stringent test of the maximum conformational search. The latter simulates the more realistic case of an ordinary NMR experiment. We applied six protocols, using three kinds of initial structures, by (1) *no metrization*, (2) *metrization*, and (3) *random coil*. Two optimization methods were used: conjugate gradient minimization (MIN) in 4D space and the simulated annealing (SA) in 4D space, that is, *no metrization*+MIN, *no metrization*+SA, *metrization*+MIN, *metrization*+SA, *random coil*+MIN, and *random coil*+SA. All computations and data analyses were performed on a FACOM VP2600 supercomputer or a VAX8810 computer. The CPU time to calculate a BPTI structure was 61 s, 154 s, 3094 s, and 149 s for *no metrization*+MIN, *no metrization*+SA, *metrization*+SA, and *random coil*+SA, respectively, on the FACOM VP2600 (a theoretical peak performance of 5 GFLOPS).

(1) Unconstrained L-alanine 30mers

Each of the above six protocols generated 100 structures from the information of the 1-2, 1-3, and 1-4 distances and the chirality. In the case of the alanine 30mers, no differences in the convergence level of the optimization were found among the various protocols, irrespective of the method used to obtain the initial structure or of the method of optimization used; the success rates (Kuszewski et al., 1992) were greater than 0.97 for all protocols. The system is so simple that any structure may easily attain its near global minimum. Removal of the influences of the convergence level is important to avoid complications when comparing optimized structures.

In Fig. 1, the results are compiled in the form of the end-to-end distance, the distance between Ala^1-C^α and Ala^{30}-C^α. The distributions obtained by MIN (*no metrization*+MIN and *metrization*+MIN) as the optimization method are basically the same as those reported by Kuszewski et

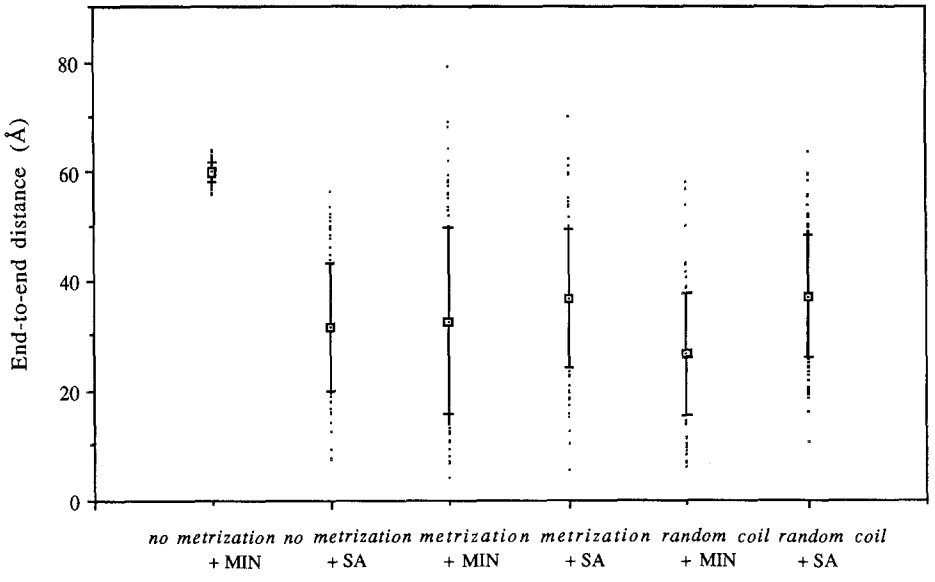


Fig. 1. End-to-end distances of structures generated by the six protocols. Three types of starting conformations (*no metrization, metrization,* and *random coil*) and two optimization methods (conjugate gradient minimization (MIN) and the simulated annealing (SA)) were used. The distances are calculated between the Ala^{1}-C^{α} and Ala^{30}-C^{α} atoms. The end-to-end distances are shown by dots. Boxes and bars show the means and standard deviations, respectively.

al. (1992). The conventional protocol (*no metrization*+MIN) gave only extended structures with a narrow range of end-to-end distances, as found by several investigators (Metzler et al., 1989; Havel, 1990; Kuszewski et al., 1992). The randomized metrization procedure, *metrization*+MIN, improved these poor sampling properties to give a wide distribution of end-to-end distances. All three protocols using simulated annealing, including *no metrization*+SA, gave broad distributions that were basically identical to each other. This means that a large thermal agitation in the stage 2 of the SA (Table 1) searches a vast area of conformational space to erase the dependence of the optimized structures on the initial structure.

(2) Bovine pancreatic trypsin inhibitor (BPTI)

Three-dimensional structures of BPTI were calculated by the four protocols, using data set II of Table 2: 200 structures by *no metrization*+MIN, 100 structures by *no metrization*+SA, 100 structures by *metrization*+SA, and 100 structures by *random coil*+SA. In Table 3 the success rates of the four protocols are listed. All of them were very high, with rates above 87%. The best 50 structures with the 50 smallest distance violations were selected for the statistics of each protocol. The other two protocols, *metrization*+MIN and *random coil*+MIN, gave success rates of almost zero, although they generated alanine 30mers successfully. This is because the initial structures obtained by the *metrization* and *random coil* protocols deviate too much from the X-ray coordinates to be optimized by the minimization method; the r.m.s.d. values between the initial and the X-ray structures were 7.9±1.3 Å and 14.0±3.0 Å for *metrization* and *random coil*, respectively. On

TABLE 3
STATISTICS OF STRUCTURES GENERATED BY VARIOUS PROTOCOLS

Protocols[a]	Deviations from ideal geometry				No. of 4D violations > 0.001 Å[c]	No. of violations > 0.2 Å[d]	Max. violation (Å)[e]	Max. contact (Å)[f]	Success rate[g]
	Bond ($\times 10^3$Å)	Angle (°)	ω-Torsion[b] (°)	Chiral volume ($\times 10^2$Å^3)					
No metrization + MIN	3.53 ± 0.32	0.97 ± 0.12	7.65 ± 0.82	2.06 ± 0.24	0	0.4 ± 0.6	0.42	0.15	0.90
No metrization + SA	2.97 ± 0.03	0.74 ± 0.02	5.65 ± 0.44	1.81 ± 0.02	0	0	0.05	0.02	0.96
Metrization + SA	2.97 ± 0.04	0.74 ± 0.02	5.66 ± 0.44	1.82 ± 0.02	0	0	0.07	0.01	0.87
Random coil + SA	2.99 ± 0.04	0.75 ± 0.03	5.76 ± 0.37	1.82 ± 0.02	0	0	0.05	0.02	0.90

[a] The names of the protocols are as follows: *no metrization* + MIN uses the initial structures generated by the conventional embedding algorithm, *no metrization*, and the conjugate gradient minimization in 4D space (MIN). *No metrization* + SA comprises *no metrization* and the simulated annealing (SA) in 4D space described in Table 1. *Metrization* + SA uses the initial structures generated by the randomized metrization and the SA optimization. *Random coil* + SA generates random coil for the initial structures and optimizes them by SA. The best 50 structures of BPTI giving the 50 smallest distance violations are analyzed for each protocol.

[b] The ω-torsion before the four proline residues were excluded from the calculation.

[c] The average number of the structures that have a fourth coordinate, w_i, (described in the Methods section) greater than 0.001 Å.

[d] The average number of the distances that violate the given restraints by more than 0.2 Å.

[e] The maximum violation in the upper and lower distance bounds.

[f] The maximum contact is the largest distance violation in the lower bounds for all non-bonded atom pairs.

[g] Success rate is defined by the ratio of the structures in which there is no distance violation greater than 0.5 Å, the deviations of bond lengths and angles from the ideal are less than 0.015 Å and 3°, respectively, and the maximum value of the fourth dimensional coordinate is less than 0.001 Å. The success rate of *no metrization* + MIN was calculated for 200 initial structures. In each of the other three protocols, the success rate was calculated for 100 initial structures.

the other hand, the conventional method, *no metrization*, gave small r.m.s.d. values of 3.0±0.1 Å, which are local minima that can be surmounted even by MIN. Therefore, we excluded the statistics of *metrization*+MIN and *random coil*+MIN.

The statistics for the convergence of the four protocols, the deviations from ideal geometry and the distance violations, are summarized in Table 3. The three protocols using SA converged almost completely. The structures satisfied the distance restraints and deviated only slightly from the ideal covalent geometry, with no van der Waals collision. The convergence of *no metrization*+MIN was actually sufficient, although the statistics were slightly worse than those of the SA protocols.

Comparisons of the optimized structures are summarized in Table 4 in the form of the mean backbone r.m.s.d. The r.m.s.d. value (DG vs. DG) for each protocol listed in Table 4a was calculated by averaging the r.m.s.d. values for all 50 × 49/2 pairs of sampled structures, and

TABLE 4

BACKBONE RMSD AND RADIUS OF GYRATION OF BPTI STRUCTURES GENERATED BY VARIOUS PROTOCOLS

(a) Backbone r.m.s.d. values and radii of gyrations of the individual structures generated by four protocols

Protocols	*No metrization* + MIN	*No metrization* + SA	*Metrization* + SA	*Random coil* + SA
Backbone r.m.s.d. values (Å)[a]				
DG vs. $\overline{DG}$	0.78 ± 0.15 (0.78)[b]	0.79 ± 0.13 (0.79)	0.83 ± 0.16 (0.83)	0.81 ± 0.14 (0.81)
DG vs. DG	1.11 ± 0.18	1.13 ± 0.20	1.19 ± 0.23	1.15 ± 0.20
$\overline{DG}$ vs. X-ray	1.35 ± 0.11	1.20 ± 0.20	1.19 ± 0.22	1.17 ± 0.22
$\overline{DG}$ vs. X-ray	1.10 (1.10)[c]	0.92 (0.90)	0.86 (0.85)	0.86 (0.84)
Radius of gyration, R_G (Å)[d]	10.31 ± 0.11	10.59 ± 0.09	10.60 ± 0.08	10.59 ± 0.09

(b) Backbone r.m.s.d. values between the different protocols[e]

	No metrization + MIN	*No metrization* + SA	*Metrization* + SA	*Random coil* + SA
No metrization + MIN		0.66	0.68	0.61
No metrization + SA	1.30		0.14	0.16
Metrization + SA	1.33	1.16		0.15
Random coil + SA	1.28	1.14	1.17	

[a] The notation of the structures is as follows. DG is a set of the best 50 structures generated by each protocol. $\overline{DG}$ is the mean structure obtained by averaging the coordinates of the 50 structures after superposition of the backbone atoms. X-ray is the crystal structure of BPTI (Wlodawer et al., 1984). The values of DG vs. $\overline{DG}$ and DG vs. X-ray are the means of 50 r.m.s.d. values. The value of DG vs. DG is the mean of 50 × (50−1)/2 r.m.s.d. values.

[b] Mean ± standard deviation. The value calculated by (DG vs. DG) × $[(n-1)/2n)]^{1/2}$ is given in parentheses. The coincidence of these values with DG vs. $\overline{DG}$ indicates that the superposition procedure defining $\overline{DG}$ is appropriately done.

[c] The value defined by $[(\text{DG vs. X-ray})^2-(\text{DG vs. } \overline{DG})^2]^{1/2}$ is given in parentheses.

[d] The radius of gyration, R_G is calculated for the backbone atoms (N,C^{α},C).

[e] Lower off-diagonals are the averages of the r.m.s.d. values (DG vs. DG) between two sampled structures generated by different protocols. Upper off-diagonals are the r.m.s.d. values ($\overline{DG}$ vs. $\overline{DG}$) between the mean structures generated by different protocols (see text for details).

represents the radius of distribution produced by each protocol. The r.m.s.d. values (DG vs. DG) between two protocols, which are the numbers in the lower off-diagonal in Table 4b, are the average r.m.s.d. values for all 50 × 50 pairs. Each structure of a pair was produced by each different protocol. This r.m.s.d. value represents the distance between two ensembles. We found that the two kinds of r.m.s.d. (DG vs. DG) values in Tables 4a and b were almost equivalent in the three protocols that used SA in the optimization process. This means that the distributions of the ensembles generated by the three SA protocols overlap with each other and that these protocols give almost equivalent ensembles, irrespective of the way in which the initial structures were generated. This can be seen more clearly in the small r.m.s.d. values ($\overline{DG}$ vs. $\overline{DG}$) between the mean structures produced by the three protocols, which are listed in the upper off-diagonal of Table 4b. However, the structures calculated by the conventional protocol, *no metrization*+MIN, were outside the distribution obtained by the three protocols with SA.

This is more explicitly demonstrated in the 2D representation of the distributions in Fig. 2. This figure was obtained by principal component analysis of a 201×201 matrix, whose elements were the r.m.s.d. values between any pair of the 201 structures generated by the four protocols, including the X-ray structure. Three protocols with SA gave overlapping clusters, all of which contained the X-ray structure. On the contrary, *no metrization*+MIN tended to cluster outside the distributions of the three SA protocols.

As in the case of L-alanine 30mers, the more complicated and larger system of BPTI showed that the protocol of SA in 4D space searches a sufficiently large conformational space to determine structures without any trace of initial structure dependence, whereas the conventional protocol, *no metrization*+MIN, gave a poor sampling. This can be seen in Fig. 3, which illustrates

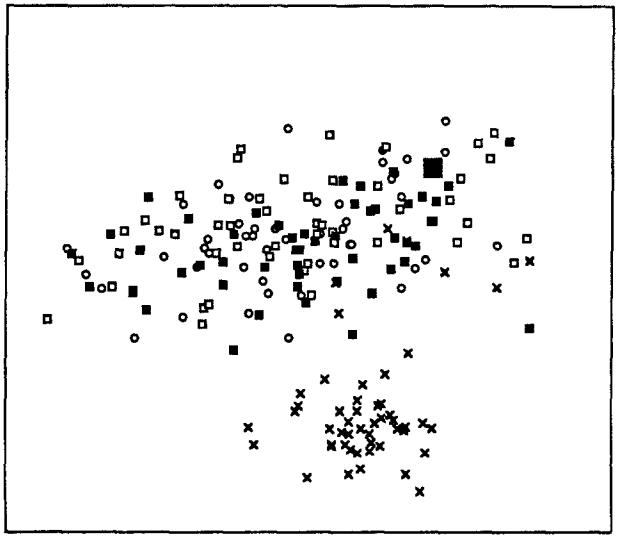


Fig. 2. Two-dimensional representation of the 201 BPTI structures generated with various protocols, calculated by principal component analysis. Large filled square (■) indicates the X-ray structure. The 50 structures generated by each of the four protocols are plotted as *random coil* + SA (■), *metrization* + SA (□, *no metrization* + SA (○), and *no metrization* + MIN (X).

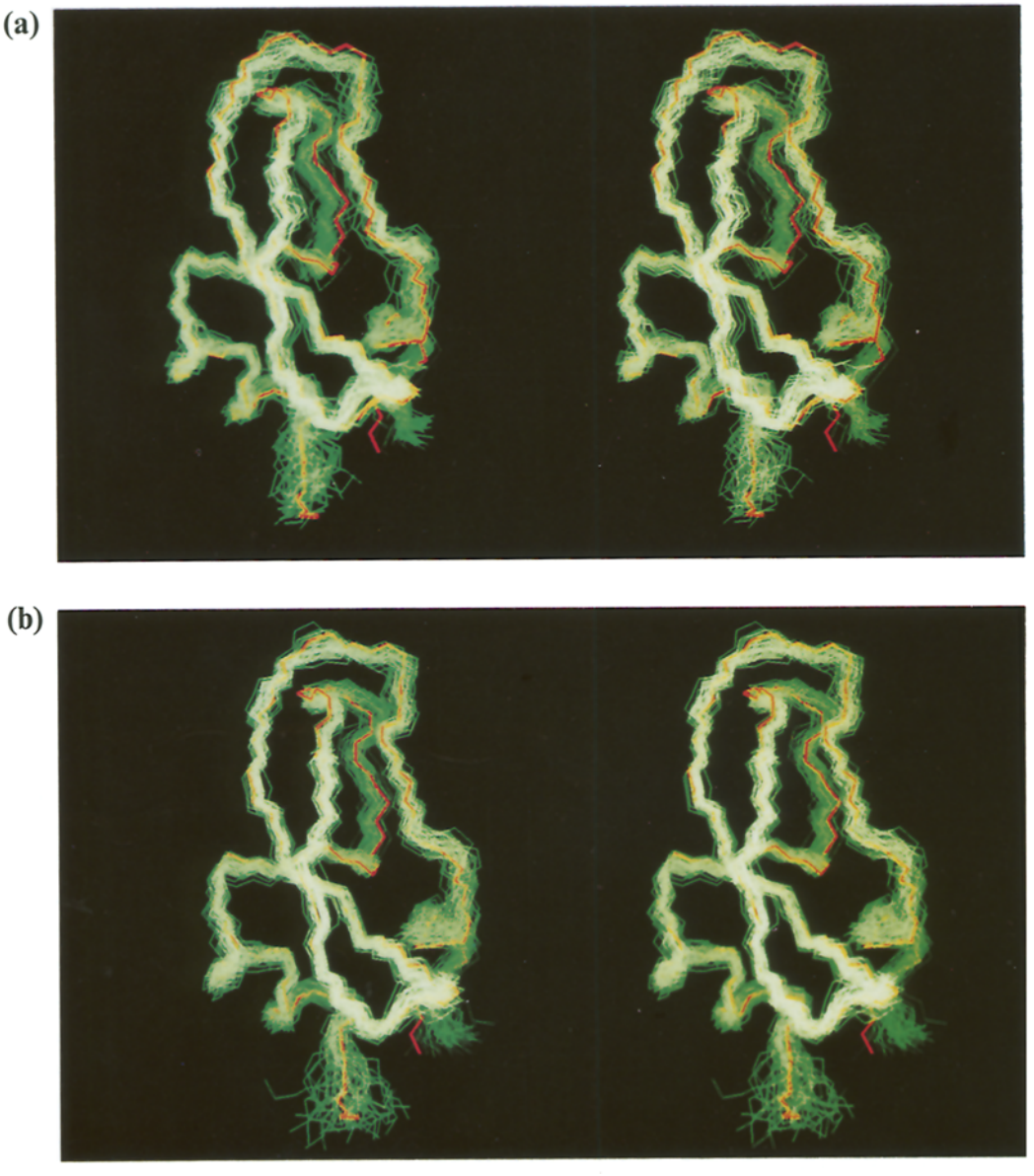


Fig. 3. Stereoview of the backbone atoms (N, C^{α}, and C) of the 50 BPTI structures (green lines) calculated from the simulated data set II, superposed onto the X-ray structure (red line). (a) The structures gneerated by *no metrization* + MIN. (b) The structures generated by *random coil* + SA.

the backbone conformations of the structures generated by (a) *no metrization*+MIN and (b) *random coil*+SA. The backbone atoms of the 50 structures have been superposed on that of the X-ray structure. It is apparent that the 50 structures generated by *random coil*+SA surround the

X-ray structure, whereas the structures generated by *no metrization*+MIN show a distinct difference in the loop structure (top of Fig. 3) from that of the X-ray structure. As seen from the statistics of Table 4, there was no marked difference between the superpositions of the backbone atoms generated by other protocols with SA and that of *random coil*+SA. The simple minimization (MIN) cannot escape from the biased initial structures given by the conventional embedding procedure, *no metrization*. It is noted that *no metrization* generated similar coordinates, that is, the r.m.s.d. value for any two structures immediately after embedding was only 2.0±0.1 Å, while *metrization* and *random coil* generated a wide distribution of initial structures, with r.m.s.d. values of 6.6±1.0 Å and 12.9±2.4 Å, respectively. A broad distribution of initial structures and a powerful optimization method are necessary conditions for good sampling.

In order to investigate the effect of the precision of the distance bounds on the sampling properties, we generated structures of BPTI by the protocol of *random coil*+SA, using the five different data sets with different precisions listed in Table 2. The r.m.s.d. values of the structures produced with the various data sets are listed in Table 5. Comparison of data sets I, II, and IX shows that the average of the sampled structures got closer to the X-ray structure with increasing precision of the distance data. An increase in the precision of the upper distance bounds, from 5.0 to 4.0 Å (from III to IV), was more effective than a classification with the uniform 4.0 Å distance bounds into 2.5, 3.0, and 4.0 Å (from IV to II). These basic features are the same as those observed by Havel (1991).

DISCUSSION

(1) Sampling properties by the protocols

In L-alanine 30mers (Fig. 1) and BPTI (Fig. 2), it has been confirmed that the SA protocols sample a very wide conformational space.

When conformational sampling is adequate, it would be expected that the average structure of the ensemble converges to the X-ray structure, from which the distance data were derived. That is, the r.m.s.d. values ($\overline{DG}$ vs. X-ray) would be much smaller than the corresponding r.m.s.d. values (DG vs. $\overline{DG}$) in Table 4. However, the r.m.s.d. values ($\overline{DG}$ vs. X-ray) are almost equal to the r.m.s.d. values (DG vs. $\overline{DG}$) even after the SA protocol. In other words, since the r.m.s.d. (DG vs. $\overline{DG}$) of a protocol refers to the radius (1σ level, σ is the standard deviation) of the distribution of the sampled structures, the X-ray structure does not stay at the center of the distribution. Even though all the other statistics indicate that the three SA protocols have good sampling properties, a small deviation remains in the final structures. In the case of BPTI, the conventional protocol, *no metrization*+MIN, gave a similar size of the ensemble as those of the SA protocols (r.m.s.d. values of DG vs. DG), but the ensemble average had a larger deviation from the X-ray structure (r.m.s.d. values of $\overline{DG}$ vs. X-ray). For the various distance data sets shown in Table 5, the r.m.s.d. values ($\overline{DG}$ vs. X-ray) were also found to be almost equal to the r.m.s.d. values (DG vs. $\overline{DG}$) after the SA protocol. These findings will be discussed in more detail in the next section.

(2) Discrepancy between the X-ray structure and the calculated structures

The r.m.s.d. values (X-ray vs. $\overline{DG}$) in Table 4 indicate the discrepancy between the average of

TABLE 5
BACKBONE RMSD AND RADIUS OF GYRATION OF BPTI STRUCTURES PRODUCED BY VARIOUS DISTANCE DATA SETS

	Data set[a]				
	I	II	III	IV	IX
Backbone r.m.s.d. values (Å)[b]					
DG vs. $\overline{DG}$	0.69 ± 0.13 (0.69)	0.81 ± 0.14 (0.81)	1.38 ± 0.24 (1.38)	0.92 ± 0.17 (0.92)	2.04 ± 0.62 (2.06)
DG vs. DG	0.98 ± 0.20	1.15 ± 0.20	1.97 ± 0.34	1.31 ± 0.24	2.94 ± 0.80
$\overline{DG}$ vs. X-ray	0.96 ± 0.18	1.17 ± 0.22	1.80 ± 0.27	1.29 ± 0.23	2.77 ± 0.64
$\overline{DG}$ vs. X-ray	0.67 (0.67)	0.86 (0.84)	1.17 (1.23)	0.92 (0.90)	1.88 (1.87)
Radius of gyration, R_G (Å)[c]	10.54 ± 0.08	10.59 ± 0.09	10.88 ± 0.12	10.61 ± 0.10	11.00 ± 0.19

[a] The names of the data sets are defined in Table 1.
[b] The notation of the structures is the same as that in Table 4. The best 50 structures giving the smallest distance violations were analyzed for each data set.
[c] The radius of gyration, R_G is calculated for the backbone atoms. The value of R_G for the X-ray structure is 10.58.

the sampled structures and the X-ray structure from which the distance bounds were derived. In this section, we discuss the details of this finding. The structures shown in Fig. 3 always contain some ambiguity in their superposition, which would hinder a detailed structural comparison. As a measure of the structural difference without superposition, the average difference distance, $\langle \Delta d_{ij} \rangle$, between a pair of C^{α} atoms is adopted here,

$$\langle \Delta d_{ij} \rangle = \langle d_{ij} - d_{ij}^{c} \rangle = \langle d_{ij} \rangle - d_{ij}^{c} \tag{6}$$

where the averaging $\langle \quad \rangle$ is over the 50 sampled structures, d_{ij} is the distance between the C^{α} atoms of residues i and j of a sampled structure, and d_{ij}^{c} is the corresponding distance in the X-ray structure. A positive value of $\langle \Delta d_{ij} \rangle$ means that this part of the average structure expands more than that of the X-ray structure.

Figure 4 shows the average difference distance matrices $\langle \Delta d_{ij} \rangle$ of the sampled structures of BPTI. These matrices explicitly indicate which part of the sampled structure deviates from the X-ray structure. Figs. 4a and b are those for distance data set II (Table 2) obtained by the two different protocols, *no metrization*+MIN and *random coil*+SA. Similar patterns are seen in the two matrices: negative signs (contraction) of 1–10 with 10–15 and 35–40, and positive signs (expansion) of 25–30 with 1–10, 40–45, and 50–55. Table 6a summarizes the correlation coefficients between the two average difference distance matrices together with the average differences $\langle \Delta d \rangle$ defined by

$$\langle \Delta d \rangle = \frac{2}{n(n-1)} \sum_{i<j} \langle \Delta d_{ij} \rangle \tag{7}$$

where n is the number of residues. The quantity $\langle \Delta d \rangle$ corresponds to the r.m.s.d. (X-ray vs. $\overline{DG}$) of Table 4, and decreases significantly after the SA protocol. However, the correlation coefficients with *no metrization*+MIN remain about 0.8 even after the SA protocol. The SA protocols allow the ensemble average to approximate the X-ray structure more closely, but the pattern of discrepancy from the X-ray structure is not altered. The same tendencies are found in the matrices calculated with the distance bounds of different precision by *random coil*+SA, as shown in Figs. 4b, c, and d. Increasing the precision of the distance bounds makes $\langle \Delta d \rangle$ smaller. However, similar patterns are found again in Figs. 4b, c, and d for data sets II, I, and III, respectively, and the high correlation coefficients among them are listed in Table 6b.

The use of the SA protocol with high-precision distance bounds ensures a good sampling in the sense of small $\langle \Delta d \rangle$ or r.m.s.d. (X-ray vs. $\overline{DG}$) values. However, a characteristic pattern of discrepancy always remains in a structure calculated by any protocol with distance bounds of any precision level. It is notable that this pattern is basically the same as that of the conventional method, *no metrization*+MIN, which is considered to give biased sampling. Since the SA protocol completely removes the possibility of biased sampling caused by dependence on the initial guess, the deviation appearing in the difference distance matrices should be attributed to the inherent nature of the given distance bounds.

To see whether the characteristic pattern shown in Figs. 4a–d originates in the given distance bounds, we calculated an average difference distance matrix for the distances obtained immediately after *metrization*. The distances after *metrization* satisfy the given distance bounds and triangle inequalities without the influence of either embedding or optimization. Most of the

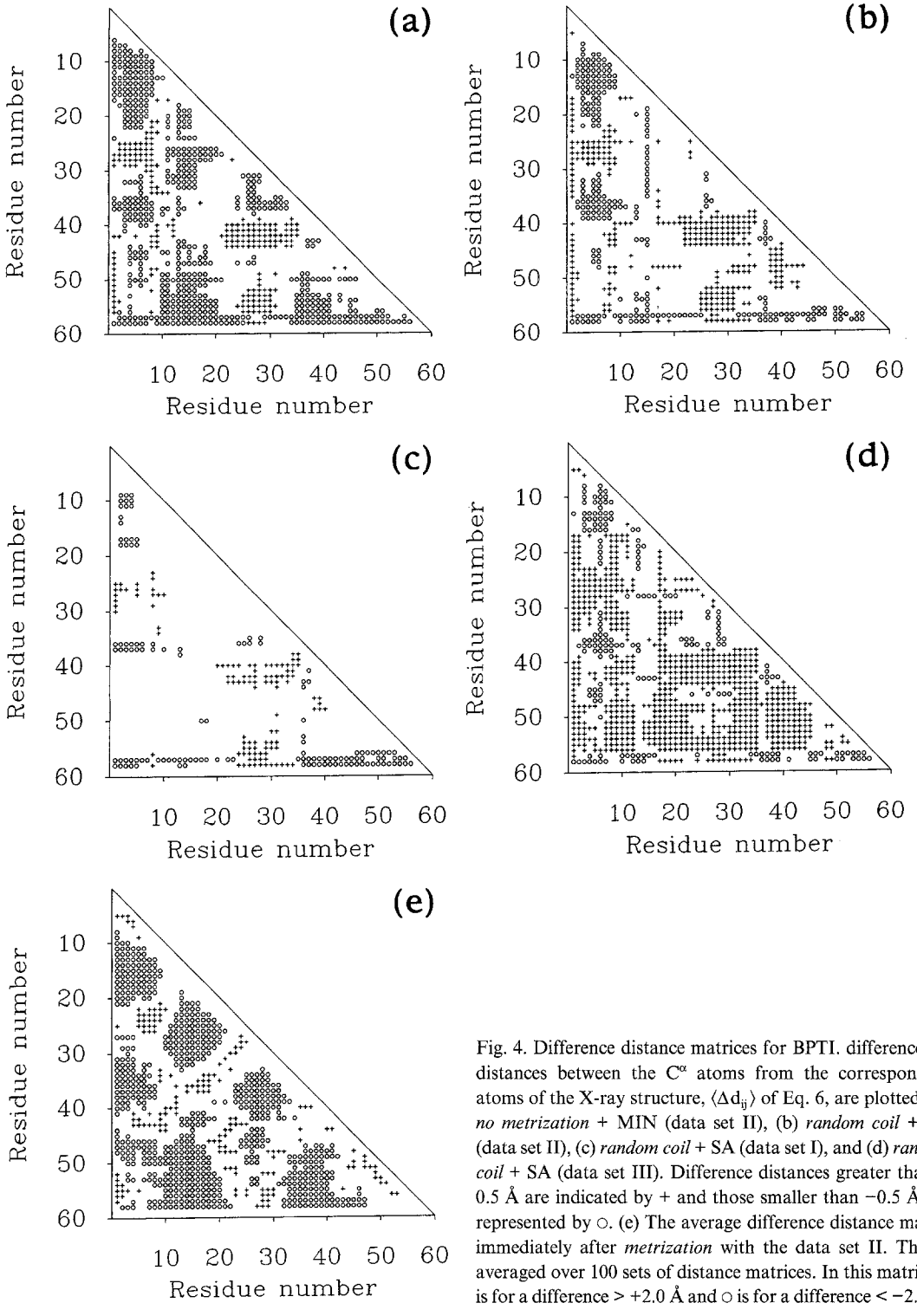


Fig. 4. Difference distance matrices for BPTI. differences in distances between the C^{α} atoms from the corresponding atoms of the X-ray structure, $\langle \Delta d_{ij} \rangle$ of Eq. 6, are plotted. (a) *no metrization* + MIN (data set II), (b) *random coil* + SA (data set II), (c) *random coil* + SA (data set I), and (d) *random coil* + SA (data set III). Difference distances greater than + 0.5 Å are indicated by + and those smaller than −0.5 Å are represented by ○. (e) The average difference distance matrix immediately after *metrization* with the data set II. This is averaged over 100 sets of distance matrices. In this matrix, + is for a difference > +2.0 Å and ○ is for a difference < −2.0 Å.

typical features found in Figs. 4a–d are reproduced in the average difference distance matrix in Fig. 4e, which is calculated by averaging 100 sets of distance matrices immediately after *metrization* to data set II (the correlation coefficient with the matrix of Fig. 4a is 0.52). This similarity strongly suggests that the structural deviation found in the difference distance matrices is an accurate reflection of the inherent nature of the distance bounds to the optimized structures. We found basically the same pattern in the distance matrices just after the conventional bound smoothing, *no metrization*, and the correlation coefficient between the average difference distance matrices given by *metrization* and by *no metrization* is 0.92.

In conclusion, the discrepancy with the X-ray structure found in difference distance matrices after the SA protocol is not caused by poor sampling of the protocol, but instead represents the inherent nature of the given distance bounds. In other words, the pattern found in Figs. 4a–d is characteristic of a 3D structure generated only from the observable short distance restraints less than 4–5 Å, with a certain range of error and without long distances. This nature of the distance bounds, which included only the observable short distances, is inherent in the NMR data.

TABLE 6
COMPARISON OF AVERAGE DIFFERENCE DISTANCE MATRICES[a]

(a) Comparison of the protocols

	Correlation coefficients			
	No metrization + MIN	*No metrization* + SA	*Metrization* + SA	*Random coil* + SA
No metrization + SA	0.79	–		
Metrization + SA	0.77	0.99	–	
Random coil + SA	0.82	0.98	0.98	–
		Average differences $\langle \Delta r \rangle$(Å)[b]		
	−0.27	0.05	0.06	0.05

(b) Comparison of the data sets

	Correlation coefficients				
Data set[c]	I	II	III	IV	IX
II	0.71	–			
III	0.64	0.82	–		
IV	0.70	0.94	0.89	–	
IX	0.58	0.75	0.78	0.74	–
		Average difference $\langle \Delta r \rangle$(Å)			
	−0.03	0.05	0.46	−0.09	0.62

[a] The average difference distance matrix is defined by Eq. 6.
[b] The average difference $\langle \Delta r \rangle$ is defined by Eq. 7.
[c] The names of data sets are the same as those of Table 5.

(3) Nature of the structural discrepancies

We now discuss the properties of the 3D structure generated by the distance geometry with the observable short distances using three examples, BPTI (Wlodawer et al., 1984), the α-amylase inhibitor Tendamistat (Pflugrath et al., 1986), and the N-terminal domain of the 434-repressor (Mondragon et al., 1989). As in the above computations, 100 structures were generated by the protocol of *random coil*+SA with distance data set II for each of the three proteins. Among them, the best 50 structures with the 50 smallest distance violations were selected. These three proteins are classified into completely different structural classes: small disulfide bond-rich protein (BPTI), a β-protein (Tendamistat), and an α-protein (434-repressor). Our purpose in this section is to find the relationship between the two patterns: the pattern of discrepancy from the X-ray structure found in the average difference distance matrices and the pattern of distance restraints reflecting the structural class.

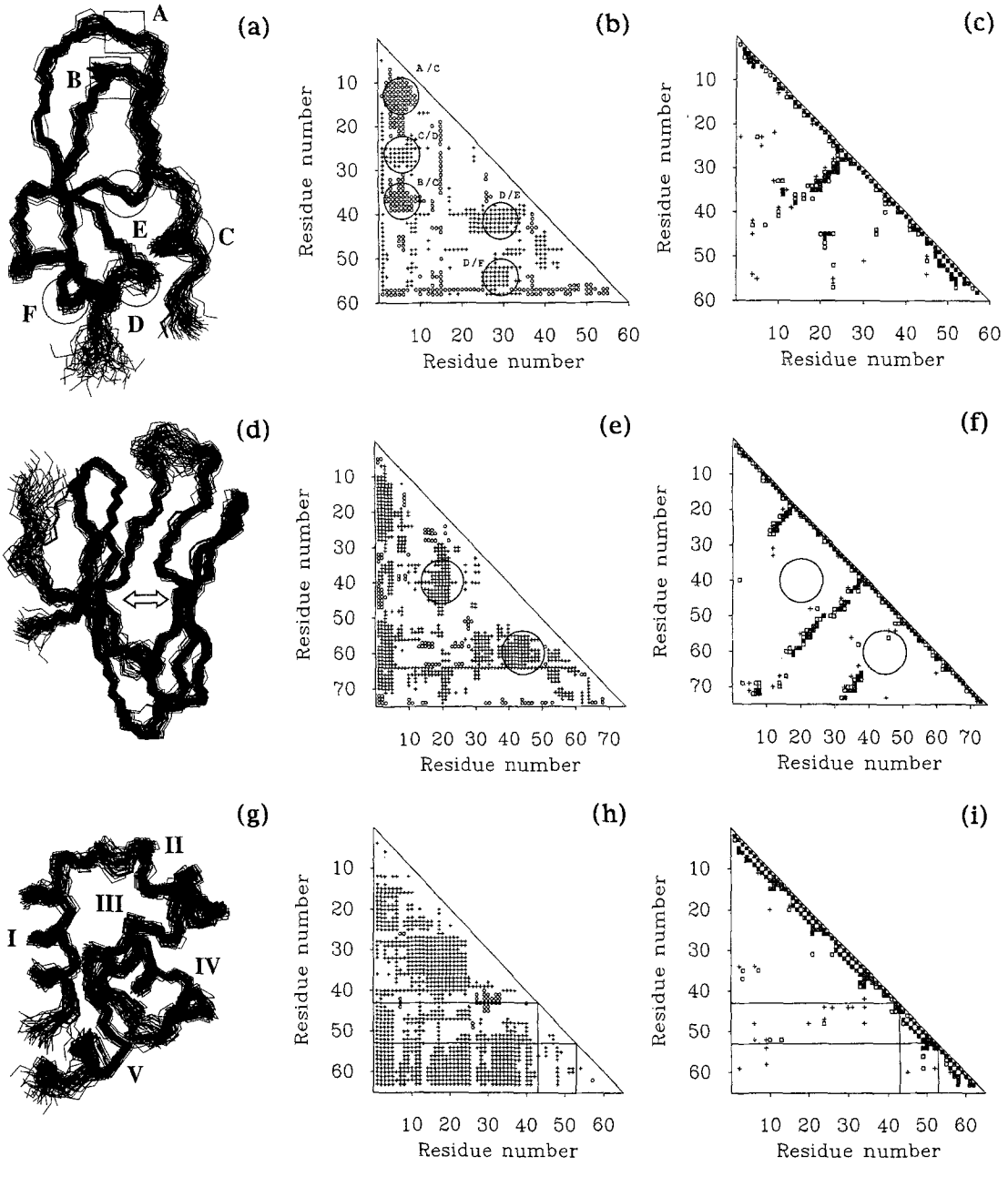
Figure 5c shows the distance restraints map for BPTI, where the location of the input upper bounds is given in the form of a triangular matrix. As shown in Figs. 5a and c, BPTI has only three secondary structure units. Two small helices, at the N- and C-termini (a 3_{10} helix of the residues 3–6 and an α-helix of the residues 48–55), are characterized by distance restraints between i and i+3 (or i+4). One β-structure exists at the center of the molecule (residues 18–24 and 29–35), with restraints that form a line perpendicular to the diagonal in Fig. 5c. The other parts are less restrained and are rather flexible.

In Fig. 5b, the distances between the C^{α} atom pairs are correctly determined when restraints are given between the two backbone hydrogens (HN and H^{α}) of the corresponding residues. Deviations from the X-ray structure are found in the distances of the residue pairs without restraints. We calculated the averages of the singular values for the sampled structures of BPTI, which correspond to the three radii of the ellipsoid approximating the backbone conformations of BPTI. These averages are listed in Table 7. The largest principal axis shrinks while the other two axes expand, that is, the calculated structures tend to be more spherical. This tendency was already reported by Oshiro et al. (1991). This can be seen in the difference distance matrix. Expansion is found in the distances among the residues, marked by circles in the superposed structures in Fig. 5a, that are along the second or third principal axes. Contraction is revealed in the distances between the residues, marked by squares and circles, that are parallel to the first principal axis. These regions are represented by circles in Fig. 5b.

The β-protein, Tendamistat (Fig. 5d), consists of two β-sheets (the first sheet comprises residues 12–16, 20–25, and 52–58, and the second sheet comprises residues 31–37, 46–48, and 67–72).

→

Fig. 5. (a) The 50 backbone conformations of BPTI generated by *random coil* + SA with data set II, superposed onto the X-ray structure. The distances among the residues marked by circles (C, D, E, and F) are more expanded than those of the X-ray structure. In contrast, the distances between the residues indicated by squares (A and B) and circles (C, D, E, and F) are shorter than those of the X-ray structure. (b) The average difference distance matrix of BPTI with a threshold of ±0.5 Å. The regions deviating from the X-ray structure are shown by circles. Regions A–F are associated with those in Fig. 5a. (c) The distance restraints of data set II for BPTI: side chain-side chain (+), side chain-backbone (□), and backbone-backbone (⧈). The plot shows 393 inter-residual distance restraints. (d) The 50 backbone conformations of Tendamistat, generated in the same way as BPTI, with 431 distance restraints, superposed onto the X-ray structure. The backbone



atoms in residues 1–6 and 73–78 are excluded from the superposition calculation. The arrow shows the direction perpendicular to the two β-sheets. (e) The average difference distance matrix of Tendamistat, with a threshold of ±0.5 Å. The regions shown by circles indicate the inter-sheet difference distances. (f) The simulated distance restraints for Tendamistat. The meanings of the symbols are the same as for BPTI. The same circles as in Fig. 5e are also indicated. (g) The 50 backbone conformations of the 434-repressor, generated in the same way as BPTI, with 371 distance restraints, superposed onto the X-ray structure. I–V are the sequential names of the five α-helices. (h) The average difference distance matrix of 434-repressor, with a threshold of ± 0.5 Å. (i) The simulated distance restraints for 434-repressor. The meanings of the symbols are the same for BPTI. For Figs. 5g and 5i, residues 43–53 around helix IV are enclosed by lines.

TABLE 7
BACKBONE RMSD AND RADIUS OF GYRATION OF THREE DIFFERENT PROTEINS

	Proteins		
	BPTI	Tendamistat	434-repressor
Backbone r.m.s.d. values (Å)[a]			
DG vs. $\overline{DG}$	0.81 ± 0.14	0.66 ± 0.13	0.75 ± 0.12
DG vs. DG	1.15 ± 0.20	0.94 ± 0.16	1.07 ± 0.17
DG vs. X-ray	1.17 ± 0.11	0.94 ± 0.12	1.09 ± 0.13
$\overline{DG}$ vs. X-ray	0.86	0.67	0.79
R_G and singular values (Å)[b]			
R_G	10.59 ± 0.09 (1.00)	10.81 ± 0.08 (1.02)	10.40 ± 0.08 (1.04)
⟨1⟩	112.4 ± 1.6 (0.98)	121.6 ± 1.6 (1.01)	93.8 ± 1.3 (1.04)
⟨2⟩	63.3 ± 1.1 (1.03)	74.0 ± 1.3 (1.05)	82.6 ± 1.3 (1.06)
⟨3⟩	53.6 ± 1.3 (1.05)	53.4 ± 0.8 (1.01)	69.4 ± 1.4 (1.03)

[a] The notation of the structures is the same as in Table 4. DG comprises the best 50 structures of each protein. $\overline{DG}$ is the mean structure obtained by averaging the coordinates of the 50 structures after superposition of the backbone atoms of residues 1–58, 7–72, and 1–63 for BPTI (Wlodawer et al., 1984), Tendamistat (Pflugrath et al., 1986), and 434-repressor (Mondragon et al., 1989), respectively. X-ray is the crystal structure for each protein.

[b] R_G is the radius of gyration of the backbone atoms (N,C^{α},C). Singular values ⟨1⟩, ⟨2⟩ and ⟨3⟩ are for the first, second, and third principal axes, respectively. Singular values are related to R_G by the equation, $R_G = [(\langle 1\rangle^2+\langle 2\rangle^2+\langle 3\rangle^2)/(\text{number of atoms})]^{1/2}$. The ratio to the value of X-ray is given in parentheses.

The distances within each sheet are precisely determined by the restraints between the main chain hydrogen atoms that define the β-structure (the three long lines perpendicular to the diagonal in Fig. 5f). However, there are fewer restraints between the two β-sheets than within the same β-sheet. As a result of this pattern of distance restraints, the inter-sheet distances of the sampled structures are more expanded than that of the X-ray structure, as shown in Fig. 5e. The large second singular value corresponds to this expansion (Table 7), because the second principal axis coincides with the direction perpendicular to the two β-sheets.

The results of the computation for the 434-repressor are summarized in Figs. 5g–i. This α-protein is almost spherical and is formed by five α-helices (residues 2–12, 17–24, 28–35, 45–51, and 56–61). Each α-helix has a well-defined structure due to the strong local restraints between the residues i and i+3 and between i and i+4. However, the distance restraints between the two α-helices are limited in number and exist only between side-chain hydrogen atoms, so that the packing of the α-helices is rather loose. This pattern of distance restraints is reflected in the large value of the radius of gyration in Table 7. The calculated structure expands in all three directions, as shown in the three singular values. The difference distance matrix also exhibits an overall expansion. However, the 4th α-helix (helix IV), which has many inter-helix restraints, is located at the correct position.

The above observations for the three proteins clarify what happens in the distance geometry calculation. From usual NMR experiments, the structural information is given only by the short distances, with a certain range of error, between neighboring atom pairs. This distance data lacks

the direct information of the distances between distant atom pairs. Our findings are summarized as follows: (1) expansion occurs within the portions of the structure lacking distance restraints, and in particular, within the distances between a pair of secondary structure elements; (2) an elongated molecule tends to become more spherical.

CONCLUSION

In this paper, it is shown that a simple protocol using SA in 4D space determines protein structures with a high convergence rate and searches a sufficiently large conformational space to erase the initial structure dependence. It is applicable to many kinds of proteins, and to various qualities of distance data, including simulated data and real NMR data (Nakai et al., 1992; Ogata et al., 1992).

Using the SA protocol, which has excellent sampling properties, we analyzed the discrepancies between the sampled structures and the X-ray structures of several proteins, from which the distance data are derived. Due to the short distance nature of the given distance information, the average of the sampled structures deviated from the corresponding X-ray structure to some extent for each protein. These discrepancies correlate with the overall shape of the protein molecule. In data sets of medium quality used in this work, there is no guarantee that the X-ray structure should be found at the center of the distribution of the ensemble. Distance data of sufficiently high precision are now available from advanced NMR experiments. With such high-quality information, discrepancies should be reduced to a low level.

ACKNOWLEDGEMENTS

We are grateful to S. Sato and K. Morikami for helpful discussions.

REFERENCES

Aho, A., Hopcroft, J. and Ullman, J. (1983) *Data Structures and Algorithms*, Addison-Wesley, Reading, MA.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A. and Haak, J.R. (1984) *J. Chem. Phys.*, **81**, 3684–3690.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535–542.

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

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

Clore, G.M. and Gronenborn, A.M. (1989) *CRC Crit. Rev. Biochem. Mol. Biol.*, **24**, 479–564.

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

Crippen, G.M. (1981) *Distance Geometry and Conformational Calculations*, Research Studies Press, John Wiley, New York, NY.

Crippen, G.M. (1982) *J. Comp. Chem.*, **3**, 471–476.

Crippen, G.M. and Havel, T.F. (1978) *Acta Crystallogr.*, **A34**, 282–284.

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

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

Havel, T.F. (1990) *Biopolymers*, **29**, 1565–1585.

Havel, T.F. (1991) *Prog. Biophys. Mol. Biol.*, **56**, 43–78.

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

Havel, T.F. and Wüthrich, K. (1985) *J. Mol. Biol.*, **182**, 281–294.
Havel, T.F., Kuntz, I.D. and Crippen, G.M. (1983) *Bull. Math. Biol.*, **45**, 665–720.
Kaptein, R., Boelens, R., Scheek, R.M. and van Gunsteren, W.F. (1988) *Biochemistry*, **27**, 5389–5395.
Kuntz, I.D., Thomason, J.T. and Oshiro, C.M. (1989) *J. Biomol. NMR*, **2**, 33–56.
Kuszewski, J., Nilges, M. and Brünger, A.T. (1989) *Methods Enzymol. B*, **177**, 159–203.
Liu, Y., Zhao, D., Altman, R. and Jardetzky, O. (1992) *J. Biomol. NMR*, **2**, 373–388.
Metzler, W.J., Hare, D.R. and Pardi, A. (1989) *Biochemistry*, **28**, 7045–7052.
Mondragon, A., Subbiah, S., Almo, S.C., Drottar, M. and Harrison, S.C. (1989) *J. Mol. Biol.*, **205**, 189–200.
Morikami, K., Nakai, T., Kidera, A., Saito, S. and Nakamura, H. (1992) *Comput. Chem.*, **16**, 243–248.
Nakai, T., Yoshikawa, W., Nakamura, H. and Yoshida, H. (1992) *Eur. J. Biochem.*, **208**, 41–51.
Nilges, M., Clore, G.M. and Grenenborn, A.M. (1988) *FEBS Lett.*, **229**, 317–324.
Ogata, K., Hojo, H., Aimoto, S., Nakai, T., Nakamura, H., Sarai, A., Ishii, S. and Nishimura, Y. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 6428–6432.
Oshiro, C.M., Thomason, J.T. and Kuntz, I.D. (1991) *Biopolymers*, **31**, 1049–1064.
Pflugrath, J.W., Wiegand, G. and Huber, R. (1986) *J. Mol. Biol.*, **189**, 383–386.
Purisima, E.O. and Scheraga, H.A. (1987) *J. Mol. Biol.*, **196**, 697–709.
Wagner, G., Braun, W., Havel, T.F., Shaumann, T., Gō, N. and Wüthrich, K. (1987) *J. Mol. Biol.*, **196**, 611–639.
Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A. (1986) *J. Comp. Chem.*, **7**, 230–252.
Wlodawer, A., Walter, J., Huber, R. and Sjolin, L. (1984) *J. Mol. Biol.*, **180**, 301–331.
Wüthrich, W., Billeter, M. and Braun, W. (1983) *J. Mol. Biol.*, **169**, 949–961.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.