

Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms

Circumventing problems associated with folding

Michael Nilges, G. Marius Clore and Angela M. Gronenborn

Max-Planck-Institut für Biochemie, D-8033 Martinsried bei München, FRG and Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA

Received 16 August 1988

A new real space method, based on the principles of simulated annealing, is presented for determining protein structures on the basis of interproton distance restraints derived from NMR data. The method circumvents the folding problem associated with all real space methods described to date, by starting from a completely random array of atoms and introducing the force constants for the covalent, interproton distance and repulsive van der Waals terms in the target function appropriately. The system is simulated at high temperature by solving Newton's equations of motion. As the values of all force constants are very low during the early stages of the simulation, energy barriers between different folds of the protein can be overcome, and the global minimum of the target function is reliably located. Further, because the atoms are initially only weakly coupled, they can move essentially independently to satisfy the restraints. The method is illustrated using two examples of small proteins, namely crambin (46 residues) and potato carboxypeptidase inhibitor (39 residues).

NMR; Protein structure; Interproton distance; Dynamical simulated annealing

1. INTRODUCTION

Over the last few years a number of computational methods for determining three-dimensional structures of proteins from interproton distance data obtained by two-dimensional NMR spectroscopy, in particular nuclear Overhauser enhancement (NOE) measurements, have been put forward [1–13]. As recent publications show [12–17], there is still considerable interest in developing new methods and improving existing

ones. The calculations, which may be carried out in either real space [1–7] or n -dimensional distance space [8–11], involve locating the global minimum of a target function which is made up of stereochemical and experimental restraints and has many false local minima. A method especially adapted to this type of nonlinear optimization problem is simulated annealing [18], which has been used in areas as diverse as electrical circuit design [18] and X-ray crystallographic refinement [19].

Two different strategies involving the application of simulated annealing for protein structure determination from NMR data have recently been proposed [16,17]. The first is a hybrid of the metric matrix distance geometry and simulated annealing methods [16]: substructures which contain only about a third of the atoms and have approximately the correct fold, are first generated by projection

Correspondence address: G.M. Clore and A.M. Gronenborn, Laboratory of Chemical Physics, Building 2, Room 123, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA

Abbreviations: NMR, nuclear magnetic resonance; NOE, nuclear Overhauser effect; RMS, root mean square

from n -dimensional distance space employing a metric matrix distance geometry algorithm (without checking the triangle inequalities), and are subsequently used as starting structures for the simulated annealing calculation. By this means, the folding problem inherent in all real space methods proposed to date is avoided, and the concept of simulated annealing is mainly used to improve the local conformation of the substructure. In the second approach, simulated annealing is used to fold an extended strand [17]. This involves using a soft asymptotic NOE potential and varying the NOE-target function such that interproton distances between protons far apart in the sequence are gradually incorporated into the calculation. This is similar in spirit to the variable target function algorithm of Braun and Go [1] and is designed to avoid incorrect folding of the polypeptide chain.

In this paper, we describe a new real space method based on the principles of simulated annealing which circumvents the folding problem completely, and sets out to obtain correctly folded structures starting out from a completely random array of atoms. The method is illustrated with two examples: the model system crambin (46 residues), which has been used in several of our previous studies [4,5,16,17], with interproton distance data derived from the crystal structure [20]; and the potato carboxypeptidase inhibitor CPI (39 residues), which has been investigated by NMR [21] and for which an independent crystal structure is available [22].

2. CALCULATIONAL STRATEGY

The total target function F_{tot} for which the global minimum region is searched is made up of the following terms:

$$F_{\text{tot}} = F_{\text{covalent}} + F_{\text{repel}} + F_{\text{NOE}} \quad (1)$$

The system is simulated at a temperature T by solving Newton's equations of motion using the molecular dynamics program XPLOR (Brünger, A.T., unpublished; [4,19]). This is in contrast to the original applications of simulated annealing [18] which made use of the Metropolis algorithm [23]. F_{tot} represents the effective potential energy in the dynamics calculation and the temperature is related to the kinetic energy of the system (see eqns 2 and 3 in [16]).

F_{covalent} maintains correct bond lengths, angles, chirality and planes, and is given by:

$$F_{\text{covalent}} = \sum_{\text{bonds}} k_b(r - r_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{impropers}} k_\phi(\phi - \phi_0)^2 + \sum_{\omega} k_\omega(1 + \cos\omega) \quad (2)$$

The force constants of the potential terms for bonds, angles, impropers (which serve to maintain planarity and chirality) and the peptide bond dihedral angle ω are set to uniform values which are varied during the calculation (see fig.1). All other dihedral angle force constants are set to 0 as the dihedral potential at rotatable bonds is effectively a non-bonded interaction. Disulfide bridges are introduced as bonds in their own right from the start of the calculations. (Note that the same would apply to a cyclic peptide bond.)

The non-bonded interactions are represented by a simple van der Waals repulsion term with a variable force constant k_{rep} [16,17]:

$$F_{\text{repel}} = \begin{cases} 0 & , \text{ if } r \geq s \cdot r_{\text{min}} \\ k_{\text{rep}}(s^2 r_{\text{min}}^2 - r^2)^2 & , \text{ if } r < s \cdot r_{\text{min}} \end{cases} \quad (3)$$

The values of r_{min} are the standard values of the van der Waals radii as represented by the Lennard-Jones potential used in the CHARMM empirical energy function [24]. s is set to 1.0 in the structure determination phase of the present calculations and to 0.825 in the second (cooling) phase. The resulting radii in the second phase are similar to the radii used in the various distance geometry programs [1,2,9].

The NOE distance restraints are represented by a square-well potential with a variable force constant k_{NOE} [25]:

$$F_{\text{NOE}} = \begin{cases} k_{\text{NOE}}(r_{ij} - r_{ij}^u)^2 & , \text{ if } r_{ij} \geq r_{ij}^u \\ 0 & , \text{ if } r_{ij}^l < r_{ij} < r_{ij}^u \\ k_{\text{NOE}}(r_{ij} - r_{ij}^l)^2 & , \text{ if } r_{ij} \leq r_{ij}^l \end{cases} \quad (4)$$

where r_{ij}^u and r_{ij}^l are the upper and lower limits of the target distance restraints, and r_{ij} the calculated values.

The calculational strategy is relatively straightforward. First, a starting conformation with a random array of atoms is generated by assigning random values to the x , y and z coordinates of the atoms according to a Gaussian distribution with a standard deviation of 1.0 Å centred about the coordinate origin. (Note that the exact form of the distribution is of no relevance.) One of the starting conformations for crambin is shown in fig.2A. The starting structure can be envisaged as a very high temperature conformation of the system. Very close non-bonded contacts are first removed by a few cycles of Powell [26] minimization with all force constants set to very low values (0.001 kcal·mol⁻¹·Å⁻² for the bond and NOE terms, 0.001 kcal·mol⁻¹·Å⁻⁴ for the F_{repel} term and 0.001 kcal·mol⁻¹·rad⁻² for the angular terms). The force constants for F_{covalent} and F_{NOE} are then set to values such that the initial potential energy F_{tot} is approximately equal to the kinetic energy at 1000 K. All force constants are set to identical values apart from the repulsion force constant k_{rep} which is set to 0.001 kcal·mol⁻¹·Å⁻⁴. Initial velocities are assigned according to a Maxwell distribution at 1000 K and all masses are set to uniform values (10 a.u.).

The first phase of the calculation consists of ~45 cycles of dynamical simulated annealing, each comprising 1 ps dynamics with a time step of 1 fs at 1000 K. The exact number of cycles depends on the initial values of the force constants (see above). The velocities are rescaled whenever the temperature is lower than 500 K or higher than 1250 K. At the beginning of each cycle, the force constants (covalent and NOE) are increased by

multiplying them by a factor of 1.25, up to maximum values of $100 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$ for k_{NOE} , $500 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$ for k_b , and $500 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{rad}^{-2}$ for the angle, dihedral and improper terms. The repulsion force constant k_{rep} is left unchanged until the bond force constant k_b reaches a value of

$100 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$, at which time k_{rep} is increased by multiplying its value by a factor of 1.5 at the beginning of each subsequent cycle, up to a maximum value of $0.25 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-4}$. Although the value of k_{rep} is very low during the early stages of the protocol, it is sufficient to define

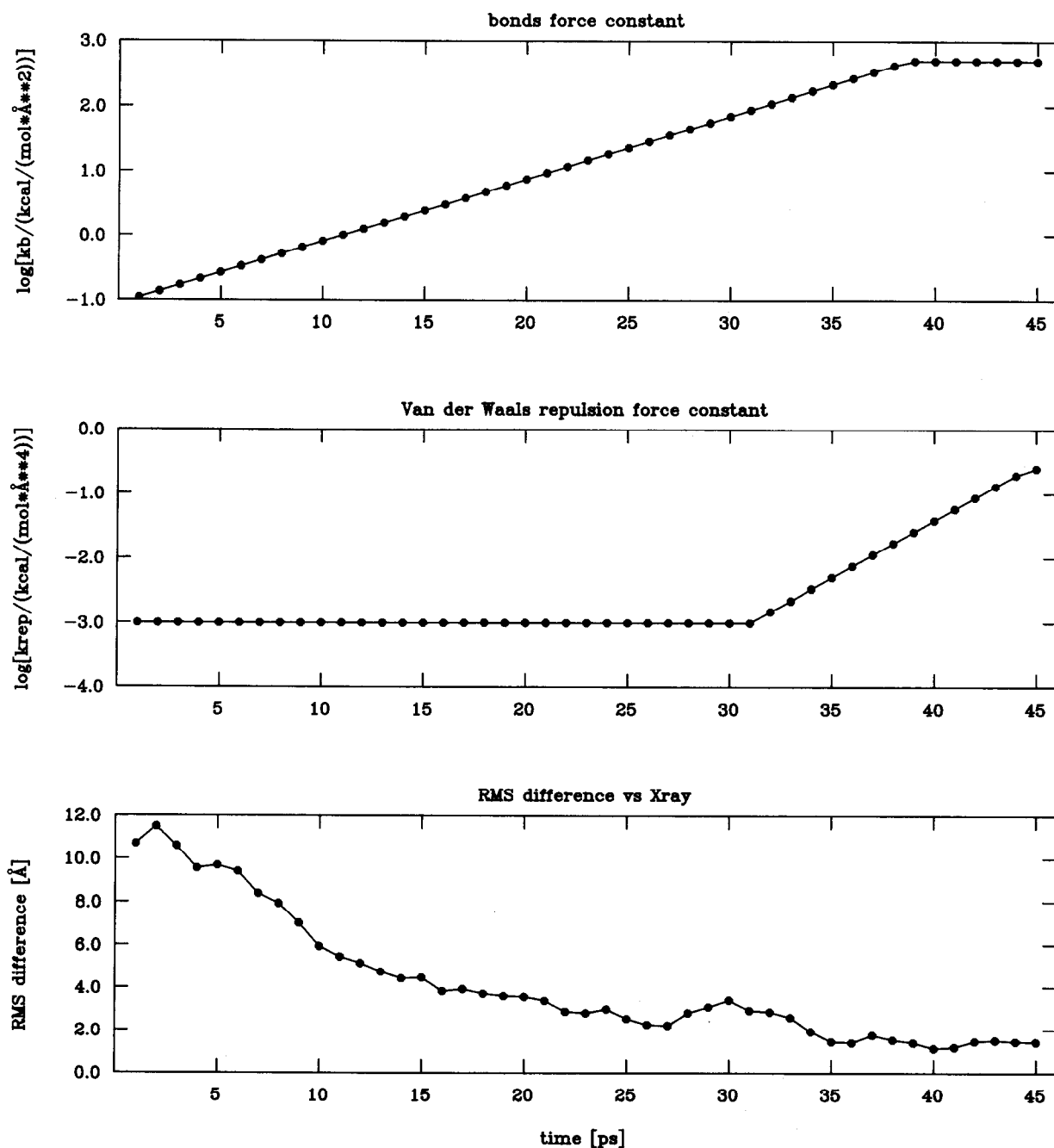


Fig.1. Time dependence of the bond force constant k_b , the van der Waals repulsion force constant k_{rep} and the backbone (N, C α , C, O) atomic RMS difference versus the crystal structure during the course of a dynamical simulated annealing calculation on crambin.

Table 1
Structural statistics

	⟨SA⟩ structures ^a			
	Crambin		CPI	
NOE _{RMS} (Å) ^b	0.036 ± 0.007		0.098 ± 0.013	
Non-bonded contacts				
F_{repel} (kcal·mol ⁻¹) ^c	34.2 ± 19.1		25.6 ± 9.8	
$E_{\text{L-J}}$ (kcal·mol ⁻¹) ^d	-127.1 ± 35.2		-84.3 ± 36.9	
Deviations from idealized geometry				
Bonds (Å)	0.020 ± 0.006		0.017 ± 0.001	
Angles (°)	2.008 ± 0.085		2.457 ± 0.010	
Impropers (°) ^e	1.540 ± 0.037		1.433 ± 0.036	

^a There are seven final converged SA structures each for crambin and CPI

^b The RMS deviations from the interproton distance restraints are calculated with respect to the upper and lower limits of the distance restraints [5]

^c The values for the van der Waals repulsion term F_{repel} (eqn 3) are calculated with a force constant of 4 kcal·mol⁻¹·Å⁻⁴ and with the hard sphere van der Waals radii set to 0.825 times the standard values used in the CHARMM empirical energy function [24]

^d $E_{\text{L-J}}$ is the Lennard-Jones van der Waals energy calculated using the CHARMM empirical energy function [24]

^e The improper terms serve to maintain planarity and appropriate chirality

and maintain the global structure while allowing atoms to get very close to each other and even move through each other to improve the structure locally. As the values of all force constants are very low during the early stages of the simulation, energy barriers between different folds of the protein can be overcome, and the global minimum of the target function is reliably located. Because the atoms are initially only weakly coupled, they can move essentially independently to satisfy the restraints, thereby avoiding problems associated with folding.

The path of the calculations is illustrated in figs 1 and 2 for a typical crambin trajectory. Fig.1 shows the increase in the bond and repulsion force constants, together with the backbone (C, C^α, N, O) atomic RMS difference versus the crystal structure of crambin, as a function of time. Fig.2B shows snapshots of the 'trajectory' every 4 ps, and fig.2C shows the best fit superposition of the final structure with the X-ray structure. Note that the global features of the structure begin to emerge at remarkably low values of the force constants. The atoms first arrange themselves roughly in their correct global position, prior to the evolution of local structural elements.

Varying force constants rather than the temperature is mainly a matter of convenience, as no variable time step integrator is required when the calculations are carried out at a constant temperature. While the two methods are not exactly equivalent, the former has the additional advantage that the force constants can be easily varied at different rates which improves the efficiency of the method [17]. For a potential that is harmonic in the coordinates, scaling of a force constant by a factor c corresponds to simultaneously scaling the temperature by $1/c$ and the time by $1/\sqrt{c}$.

In stage 2 of the protocol, which consists of 20 cycles of 0.1 ps dynamics, the parameters of the van der Waals repulsion term F_{repel} are set to their final values ($s = 0.825$, $k_{\text{rep}} = 4$ kcal·mol⁻¹·Å⁻⁴), and the temperature is cooled down to

Table 2
Atomic RMS differences^a

	Atomic RMS difference (Å)			
	Crambin		CPI (residues 2–39) ^b	
	Backbone atoms	All atoms	Backbone atoms	All atoms
⟨SA⟩ vs $\overline{\text{SA}}$	0.86 ± 0.07	1.17 ± 0.12	1.05 ± 0.28	1.63 ± 0.31
⟨SA⟩ vs X-ray	1.23 ± 0.17	1.73 ± 0.18	1.87 ± 0.33	2.88 ± 0.38
SA vs X-ray	0.90	1.27	1.55	2.38

^a The notation of the structures is as follows: ⟨SA⟩ are the seven converged structures produced by dynamical simulated annealing starting from different random arrays of atoms; and $\overline{\text{SA}}$ is the mean structure obtained by averaging the coordinates of the ⟨SA⟩ structures best fitted to each other. The X-ray structures of crambin and CPI are from [20] and [22], respectively

^b Best fitting in the case of CPI was performed with respect to residues 2–39 as no NOEs involving residue 1 were observed [21]. In comparisons with the X-ray structure of CPI, best fitting was carried out with respect to residues 2–38, as residue 39 is cleaved from the protein in the CPI-carboxypeptidase complex from which the X-ray structure of CPI is derived [22]

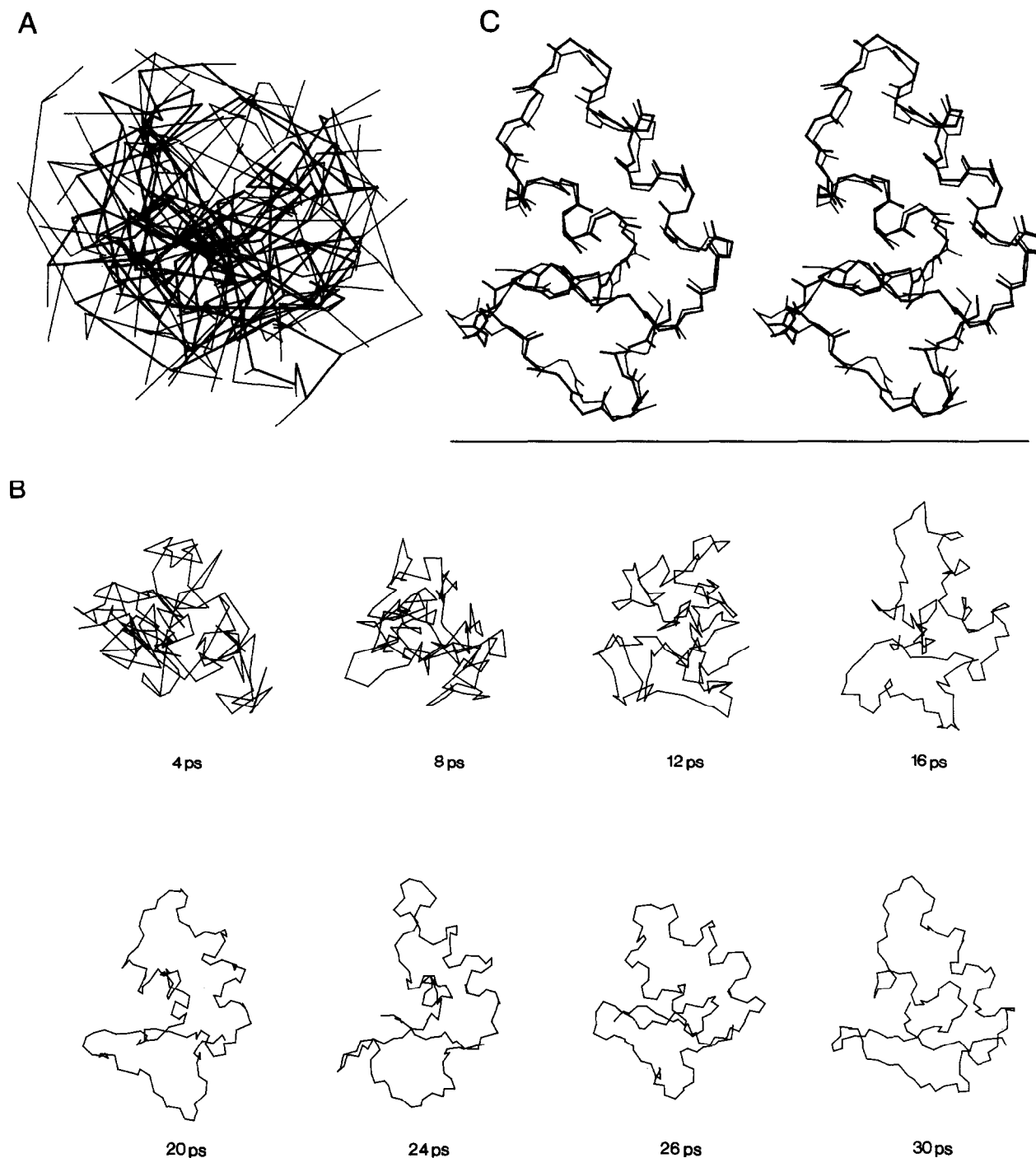


Fig.2. Path of a dynamical simulated annealing calculation on crambin. (A) Initial structure comprising a random array of atoms with the backbone and side chain bonds shown as thick and thin lines, respectively; (B) snapshots at 4 ps intervals during the first phase of the dynamical simulated annealing calculation (only the N, C α and C backbone atoms are shown); and (C) best fit superposition of the backbone (N, C α , C, O) atoms of a final dynamical simulated annealing structure (thick lines) resulting from the trajectory shown in (B) on the X-ray structure of crambin (thin lines).

300 K. The NOE force constant is set to $50 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$. These relative values of the final force constants have been found to ensure that the experimental restraints are satisfied within the error bounds, nearly perfect stereochemistry is achieved and no unduly close non-bonded contacts appear in the structure (see [17,18] and table 2). Stage 2 is followed by 200 cycles of restrained Powell minimization.

3. RESULTS ON CRAMBIN AND CPI

The results of the calculations on crambin and CPI are summarized in tables 1 and 2. The sets of NOE restraints were identical to those employed in the previous calculations [4,5,16,17,21]. In the case of crambin the NOE data set consisted of 240 interproton distances derived from the crystal structure [20]; for CPI it comprised 309 interproton distances derived from NOE measurements [21]. The distances were classified into three distance ranges, 1.8–2.7, 1.8–3.3, and 1.8–5.0 Å, corresponding to strong, medium and weak NOEs [25,27]. Distances to methyl, methylene and aromatic ring protons that were not assigned stereospecifically were calculated with respect to the average position of these protons and the upper limits of the corresponding restraints were corrected appropriately as described [28]. Disulfide bridges, of which there are three for both crambin and CPI, were treated as normal bonds from the start of the calculations. 10 structures were calculated for both crambin and CPI. In the case of crambin all calculations converged to the correct polypeptide fold, while in the case of CPI 9 of the 10 calculations converged. All structures which had a positive Lennard-Jones van der Waals energy (evaluated with the CHARMM empirical energy function [24]) were rejected (two for CPI, three for crambin). Note that this energy term is not used during the calculations but only serves as an independent check at the end of the calculation. In the five structures with positive Lennard-Jones energies, the poor non-bonded contacts always occurred at the disulfide bridges.

The quality of the structures as regards deviations from ideal stereochemistry, van der Waals contacts and NOE RMS differences is comparable to those previously published (see table 1; [16,17]). Atomic RMS distributions and RMS differences with respect to the crystal structures are also of similar size (table 2). Some of the calculated struc-

tures for the model crambin calculations show very good agreement with the crystal structure with backbone atomic RMS difference values as low as $\sim 0.9 \text{ \AA}$ (see for example fig.2C).

To test further the convergence power of the protocol we performed one calculation for crambin taking the mirror image of the crystal structure as the initial conformation (i.e. a conformation containing only D-amino acids). In this starting conformation, all terms in the total target function, with a single exception are ideally satisfied. The only term which discerns this starting conformation from the correct solution is the improper term maintaining the chirality at the tetrahedral carbon atoms. Thus a mirror image of the true solution represents a deep false minimum, and, as the whole structure has to be inverted, very high energy barriers have to be overcome to reach the correct global fold. The calculation, however, converged with a final backbone atomic RMS difference of 1.3 \AA with respect to the X-ray structure.

4. CONCLUDING REMARKS

In this paper, we have presented a new approach for determining three-dimensional structures from interproton distance data by simulated annealing. By starting from a completely random array of atoms and introducing the force constants for the various terms in the target function appropriately, the folding problem associated with all real space methods described to date is efficiently avoided. The key to the method lies in reducing the force constants of all terms in the target function to values such that the barriers between different folds are of the order of the kinetic energy of the system. Thus, the atoms can move virtually independently of each other to satisfy the restraints. The quality of the structures generated using this approach is comparable to those reported previously using dynamical simulated annealing [16,17] and significantly better than those generated by metric matrix distance geometry alone [29]. In addition, the success rate of the present calculations is very high. At the same time, this method has the good sampling properties of simulated annealing and restrained dynamics calculations. The protocol should perform satisfactorily even if there are few or no long range

distances as, for example, in extended structures for which metric matrix distance geometry methods, in our experience, perform poorly [30].

The protocol described here also offers an easy and straightforward approach, similar to that suggested by Pardi et al. [31], for dealing with the problem of stereospecific assignments at prochiral centers. When two separate signals are observed for two β methylene protons which cannot be specifically assigned, they may be assigned arbitrarily to β_1 and β_2 ; the corresponding tetrahedral improper for the prochiral center is left out, and the assignment which best satisfies the NOE data is automatically selected during the course of the calculation. If only one signal is observed, the pseudo-atom representation [28] is used. As the two protons are equivalent in this case, inclusion of a tetrahedral improper is not necessary. For methylene proton resonances that are stereospecifically assigned, an improper term for the prochiral center is introduced. In contrast to the method suggested in [31], the force constants need not be lowered and raised again specifically at the prochiral centers, but they are varied together with all other force constants.

While the CPU times needed for the calculations (several hours on a CONVEX C1-XP or a VAX 8550) are of course larger than those of our hybrid distance geometry-dynamical simulated annealing method [16], they are by no means prohibitive, and still considerably shorter than for a restrained molecular dynamics calculation employing a full empirical energy function [3–6]. We note, however, that, as the non-bonded cutoff radius is only 4 Å, computational times rise linearly with the number of atoms. As a result, the time requirements for the dynamical simulated annealing method become increasingly favourable with increasing size of the protein, relative to other distance geometry methods [1,2,7–11].

Acknowledgements: We thank Dr T.A. Holak for useful discussions. Part of this work was supported by the Max-Planck Gesellschaft and Grant no.321/4003/0318909A from the Bundesministerium für Forschung und Technologie (G.M.C. and A.M.G.).

REFERENCES

- [1] Braun, W. and Go, N. (1985) *J. Mol. Biol.* 186, 611–626.
- [2] Billeter, M., Havel, T.F. and Wüthrich, K. (1987) *J. Comput. Chem.* 8, 132–141.
- [3] Clore, G.M., Gronenborn, A.M., Brünger, A.T. and Karplus, M. (1985) *J. Mol. Biol.* 186, 435–455.
- [4] Brünger, A.T., Clore, G.M., Gronenborn, A.M. and Karplus, M. (1986) *Proc. Natl. Acad. Sci. USA* 83, 3801–3805.
- [5] Clore, G.M., Brünger, A.T., Karplus, M. and Gronenborn, A.M. (1986) *J. Mol. Biol.* 191, 523–551.
- [6] Kaptein, R., Zuiderweg, E.R.P., Scheek, R.M., Boelens, R. and Van Gunsteren, W.F. (1985) *J. Mol. Biol.* 182, 179–182.
- [7] Holak, T.A., Prestegard, J.H. and Forman, J.D. (1987) *Biochemistry* 26, 4652–4660.
- [8] Crippen, G.M. and Havel, T.F. (1978) *Acta Crystallogr. Sect. A* 34, 282–284.
- [9] Kuntz, I.D., Crippen, G.M. and Kollman, P.A. (1979) *Biopolymers* 18, 939–957.
- [10] Havel, T.F. and Wüthrich, K. (1984) *Bull. Math. Biol.* 46, 673–698.
- [11] Havel, T.F. and Wüthrich, K. (1985) *J. Mol. Biol.* 182, 281–294.
- [12] Kraulis, P.J. and Jones, T.A. (1987) *Proteins* 2, 188–201.
- [13] Vázquez, M. and Scheraga, H.A. (1988) *J. Biomol. Struct. Dyn.* 5, 705–755.
- [14] Vázquez, M. and Scheraga, H.A. (1988) *J. Biomol. Struct. Dyn.* 5, 757–784.
- [15] Sherman, S.A., Andrianov, A.M. and Akhrem, A.A. (1988) *J. Biomol. Struct. Dyn.* 5, 785–795.
- [16] Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988) *FEBS Lett.* 229, 317–324.
- [17] Nilges, M., Clore, G.M., Brünger, A.T. and Gronenborn, A.M. (1988) *Protein Engineering* 2, 27–38.
- [18] Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983) *Science* 220, 671–680.
- [19] Brünger, A.T., Kuriyan, J. and Karplus, M. (1987) *Science* 235, 458–460.
- [20] Hendrickson, W.A. and Teeter, M.M. (1981) *Nature* 290, 107–112.
- [21] Clore, G.M., Gronenborn, A.M., Nilges, M. and Ryan, C.A. (1987) *Biochemistry* 26, 8012–8023.
- [22] Rees, D.C. and Lipscomb, W.N. (1982) *J. Mol. Biol.* 160, 475–498.
- [23] Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, A. and Teller, E. (1953) *J. Chem. Phys.* 21, 1087–1092.
- [24] Brooks, B.R., Brucoleri, R.E., Olafson, B.D., States, D.J., Swaminatham, S. and Karplus, M. (1983) *J. Comput. Chem.* 4, 187–217.
- [25] Clore, G.M., Nilges, M., Sukumaran, D.K., Brünger, A.T., Karplus, M. and Gronenborn, A.M. (1986) *EMBO J.* 5, 2729–2735.
- [26] Powell, M.J.D. (1977) *Math. Program.* 12, 241–254.
- [27] Williamson, M.P., Havel, T.F. and Wüthrich, K. (1985) *J. Mol. Biol.* 182, 295–315.
- [28] Wüthrich, K., Billeter, M. and Braun, W. (1983) *J. Mol. Biol.* 169, 949–961.

- [29] Clore, G.M., Nilges, M., Brünger, A.T., Karplus, M. and Gronenborn, A.M. (1987) FEBS Lett. 213, 269–277.
- [30] Brünger, A.T., Clore, G.M., Gronenborn, A.M. and Karplus, M. (1987) Protein Eng. 1, 399–406.
- [31] Pardi, A., Hare, D.R., Selsted, M.E., Morrison, R.D., Bassolino, D.A. and Bach, A.C., ii (1988) J. Mol. Biol. 201, 625–636.