Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2

Michael Nilges, Angela M. Gronenborn, Axel T. Brünger, and G. Marius Clore

Max-Planck Institut für Biochemie, D-8033 Martinsried bei München, FRG and Department of Molecular Biophysics and Biochemistry and the Howard Hughes Medical Institute, Yale University, New Haven, CT 06511, USA

Present address: 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

An automated method, based on the principle of simulated annealing, is presented for determining the three-dimensional structures of proteins on the basis of short (<5 Å) interproton distance data derived from nuclear Overhauser enhancement (NOE) measurements. The method makes use of Newton's equations of motion to increase temporarily the temperature of the system in order to search for the global minimum region of a target function comprising purely geometric restraints. These consist of interproton distances supplemented by bond lengths, bond angles, planes and soft van der Waals repulsion terms. The latter replace the dihedral, van der Waals, electrostatic and hydrogen-bonding potentials of the empirical energy function used in molecular dynamics simulations. The method presented involves the implementation of a number of innovations over our previous restrained molecular dynamics approach (Clore et al., 1986). These include the development of a new effective potential for the interproton distance restraints whose functional form is dependent on the magnitude of the difference between calculated and target values, and the design and implementation of robust and fully automatic protocol. The method is tested on three systems: the model system crambin (46 residues) using X-ray structure derived interproton distance restraints, and potato carboxypeptidase inhibitor (CPI; 39 residues) and barley serine proteinase inhibitor 2 (BSPI-2; 64 residues) using experimentally derived interproton distance restraints. Calculations were carried out starting from the extended strands which had atomic r.m.s. differences of 57, 38 and 33 Å with respect to the crystal structures of BSPI-2, crambin and CPI respectively. Unbiased sampling of the conformational space consistent with the restraints was achieved by varying the random number seed used to assign the initial velocities. This ensures that the different trajectories diverge during the early stages of the simulations and only converge later as more and more interproton distance restraints are satisfied. The average backbone atomic r.m.s. difference between the converged structures is 2.2 ± 0.3 Å for crambin (nine structures), 2.4 ± 0.3 Å for CPI (eight structures) and 2.5 ± 0.2 Å for BSPI-2 (five structures). The backbone atomic r.m.s. difference between the mean structures derived by averaging the coordinates of the converged structures and the corresponding X-ray structures is 1.2 Å for crambin, 1.6 Å for CPI and 1.7 Å for BSPI-2.

Key words: three-dimensional structure/crambin/CPI/BSPI-2/simulated annealing/distance restraints

Introduction

Determination of the three-dimensional structures of proteins from interproton distance data derived from NMR measurements presents a highly complex, non-linear optimization problem as the data are limited in their number, accuracy and range (<5 Å) and there are numerous false local minima along the convergence pathway. Over the past few years a number of methods with large radii of convergence have been developed to tackle this problem. These include distance geometry methods based on the metric matrix (Crippen and Havel, 1978; Kuntz et al., 1979; Havel et al., 1983; Havel and Wüthrich, 1984, 1985; Havel, 1986; Sippl and Scheraga, 1986), restrained least-squares minimization in torsion angle space with either a variable target function (Braun and Go, 1985) or a sequence of ellipsoids of constantly decreasing volumes, each of which contains the minimum of the target function (Billeter et al., 1987), and restrained molecular dynamics (Clore et al., 1985, 1986a; Kaptein et al., 1985; Brünger et al., 1986; Nilsson et al., 1986). The first three methods are based solely on the use of geometric restraints comprising interproton distances, bond lengths, bond angles, planes and soft van der Waals repulsion terms. In contrast, the restrained molecular dynamics method makes use of a full empirical energy function (comprising bonded and non-bonded interactions) supplemented by an effective potential term representing the experimental interproton distances. In terms of computational requirements, the first three methods are comparable and approximately five times faster than restrained molecular dynamics. Restrained molecular dynamics, on the other hand, has the advantage that the structures generated tend to be better in energetic terms than those obtained using the other methods, particularly with respect to the non-bonded interactions. It is for this reason that we have regularly employed restrained molecular dynamics either for the entire structure determination or to refine converged structures generated by the metric matrix distance geometry method (Clore et al., 1986b, 1987a,b,c,d,e). In our experience, restrained molecular dynamics refinement results not only in large improvements in the non-bonded contacts (i.e. van der Waals energy) as compared with the starting structures generated by distance geometry calculations, but also improves significantly the agreement with the experimental NMR data.

In this paper we present an alternative approach for the structure generation phase. This approach is based on the principle of simulated annealing (Kirkpatrick et al., 1983). As originally described, the simulated annealing method makes use of the Monte Carlo algorithm (Metropolis et al., 1953) to increase the temperature of the system temporarily in order to search for the global minimum region of the target function. In our application, we make use of Newton's equations of motion to achieve the same effect. The method, however, differs from conventional restrained molecular dynamics insofar that it is based purely on...
geometric restraints and the non-bonded terms of the target function are represented by a simple repulsion term (i.e. this replaces the dihedral, van der Waals, electrostatic and hydrogen-bonding potentials of the empirical energy function used in molecular dynamics). This has the advantage of significantly reducing the computational time requirements to a level comparable with the other methods mentioned above. In addition, two further innovations over our previous restrained molecular dynamics approach are implemented. The first involves the introduction of an effective nuclear Overhauser effect (NOE) potential whose functional form for a given interproton interaction is dependent on the magnitude of the difference between the calculated and target distance. This circumvents the need to classify the distances into short (|i−j| ≤ 5) and long (|i−j| > 5) range, a procedure which is somewhat arbitrary, and increases the likelihood of correct folding. The second involves the design and implementation of a protocol which is fully automatic and considerably more robust than those employed in our previously restrained molecular dynamics study on crambin. The novel method is illustrated using three examples. The first is the model system crambin with the same distance set derived from the crystal structure that we used before in our restrained molecular dynamics study (Brünger et al., 1986; Clore et al., 1986a) and in the comparison of the restrained molecular dynamics and metric matrix distance geometry methods (Clore et al., 1987d). The second and third examples are derived from our NMR work on potato carboxypeptidase inhibitor (CPI) and barley serine proteinase inhibitor 2 (BSPI-2) and make use of the same experimental interproton distance data that were employed to solve their structures by a combination of distance geometry and restrained molecular dynamics calculations (Clore et al., 1987d,e). These three particular examples were chosen as the proteins exhibit quite different structural features as well as different sizes: crambin (46 residues) is composed principally of two α-helices and a mini-antiparallel β-sheet (Hendrickson and Teeter, 1981); BSPI-2 is a predominantly β-sheet protein with a small α-helix and a large reactive site loop (McPhalen et al., 1985), and CPI has little or no regular secondary structure (Rees and Lipscomb, 1982).

Methods

All calculations were carried out on a CONVEX-C1XP computer using a modified version of the program XPLOR (Brünger et al., 1987a,b) which is derived from the program CHARMM (Brooks et al., 1983) and has been especially adapted for restrained molecular dynamics (e.g. Clore et al., 1985, 1986a; Brünger et al., 1986). Integration of the classical equations of motion was performed using a Verlet (1967) integration algorithm with initial velocities assigned to a Maxwellian distribution at an appropriate temperature. The time step of the integrator was 0.001 ps and the non-bonded contact list was updated every 0.008 ps. Displaying of trajectories was carried out using a modified version of the function network of FRODO (Jones, 1978) interfaced with XPLOR.

Results and discussion

The target function

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

\[
F_{\text{tot}} = F_{\text{covalent}} + F_{\text{repel}} + F_{\text{NOE}}
\]  (1)

\( F_{\text{tot}} \) in effect represents the potential energy of the system whose units in the present calculations are kcal/mol. These units are purely arbitrary. Thus, the simulated annealing procedure employed in this case involves the simultaneous integration of Newton’s equations of motion:

\[
\frac{d^2x_j}{dt^2} = -\frac{1}{m_j} \frac{\partial}{\partial x_j} F_{\text{tot}}(x_1, x_2, \ldots, x_n)
\]  (2)

for all \( n \) atoms in the system whose temperature is given by

\[
(T_{\text{temp}}) = \frac{2}{k_B (3\pi - 6)^{\frac{3}{2}}} \left( \sum_{i=1}^{n} m_i \omega_i^2 / 2 \right)
\]  (3)

\( F_{\text{covalent}} \) is the target function for maintaining correct bond lengths, angles and planes, and is given by

\[
F_{\text{covalent}} = \sum_{\text{bonds}} k_b (r - r_0)^2 + \sum_{\text{angles}} k_{\theta} (\theta - \theta_0)^2 + \sum_{\text{improper}} k_{\omega} (\omega - \omega_0)^2
\]  (4)

The values chosen for the force constants for the bond (\( k_b \)), angle (\( k_{\theta} \)) and improper torsions (\( k_{\omega} \)) are set to uniform high values to ensure near perfect stereochemistry throughout the calculations, namely 600 kcal/mol/Å², 500 kcal/mol/rad² and 500 kcal/mol/rad² respectively. (Note that the improper torsion terms serve to maintain planarity and chirality.)

\( F_{\text{repel}} \) is the target function used to prevent unduly close non-bonded contacts and is given by

\[
F_{\text{repel}} = \begin{cases} 0 & \text{if } r \geq s \cdot r_{\text{min}} \\ k_r (r^2 - r_{\text{min}}^2)^2 & \text{if } r < s \cdot r_{\text{min}} \end{cases}
\]  (5)

The values of \( r_{\text{min}} \) are the standard values of the van der Waals radii as represented by the Lennard–Jones potential used in the CHARMM empirical energy function (Brooks et al., 1983); \( s \) is a van der Waals radius scale factor, and \( k_r \) the van der Waals repulsion force constant. It should also be noted that the large reduction in the computational time required to evaluate \( F_{\text{repel}} \) compares with the usual full non-bonded interaction potential represented in the empirical energy function is due not only to the smaller number of terms that have to be calculated but also to a reduction in the number of pairs that have to be included in the non-bonded contact list. Thus, in the case of \( F_{\text{repel}} \) the non-bonded contact list comprises only all pairs up to 4.5 Å, compared with pairs up to 8 Å in the case of the full empirical non-bonded energy function.

\( F_{\text{NOE}} \) is the NOE-target function and is a complex term made up of three terms \( F_{\text{long}}, F_{\text{short}} \) and \( F_{\text{final}} \), whose functional form depends on the difference between the calculated (\( r_{ij} \)) and target (\( r_{ij}^0 \)) value of a particular interproton distance, as well as on the value of the variable force constant for the \( F_{\text{short}} \) term (see below). Our previous restrained molecular dynamics calculations have used biharmonic (Clore et al., 1985) and square well (Clore et al., 1986b) potentials. Such forms, however, give rise to severe problems in simulations that start from initial structures containing large violations between calculated and target distance values. It was for this reason, for example, that in our restrained dynamics model calculations on crambin starting from an extended strand, distances between residues separated by more than five residues in the sequence were initially excluded from the calculation and only introduced at a later stage once partial folding of the helices had occurred.
Converge to the correct global minimum region. If the polypeptide chain has folded incorrectly it can no longer form an extended strand) rather than from entirely random structures. At the end of each cycle of simulated annealing, and a special potential is used that places greater weight on smaller violations. A further requirement is that the value of the target function \( F \) should not exceed certain limits for technical reasons. This necessitates the automatic adjustment of the NOE force constants, as appropriate, during the course of the simulation.

The calculations have to start from unfolded structures (e.g. an extended strand) rather than from entirely random structures which may already be folded. The reason for this is that once the polypeptide chain has folded incorrectly it can no longer converge to the correct global minimum region. The random number seed, however, used for the assignment of the initial velocities is sufficient to ensure good sampling of the available conformational space consistent with the interproton distance data (see following section).

The NOE restraints are initially classified into two classes, \( \text{long} \) and \( \text{short} \), depending on the difference \( \Delta \text{viol} \) between the calculated \( r_{ij}^\text{calc} \) and upper limit of the target \( r_{ij}^\text{target} \) distances. Class \( \text{long} \) contains NOE restraints which are violated by more than \( r_{ij}^\text{target} \), class \( \text{short} \) contains all the others. In class \( \text{long} \) the target function is switched off (i.e. \( F_{\text{NOE}} = 0 \)). In class \( \text{short} \) the target function has the following functional form:

\[
F_{\text{short}} = \begin{cases} 
  k_b (\text{viol})^b & \text{if } r_{ij} > r_{ij}^\text{ target} \\
  0 & \text{if } r_{ij}^\text{ target} \leq r_{ij} \leq r_{ij}^b \\
  k_b (\text{viol})^b & \text{if } r_{ij} < r_{ij}^b 
\end{cases}
\]

The values of \( a \) and \( b \) are chosen such that \( F_{\text{short}} \) is continuous and differentiable at \( r_{switch} \). They are given by

\[
a = 5r_{switch}^4 - 2c\cdot r_{switch}^2 \\
b = -4r_{switch}^5 + c\cdot r_{switch}^2
\]

Calculational strategy

A simplified flowchart of the calculational strategy is shown in Figure 2. This potential form is designed to ensure that secondary structure elements defined by interproton distances between residues close together in the sequence are formed prior to the incorporation of NOE restraints between residues far apart in the sequence. The gradient of \( F_{\text{short}} \) is largest at \( r_{switch} + r_{ij}^b \) so that NOEs which are violated by about the value of \( r_{switch} \) experience the largest force. At the beginning of the simulation \( r_{switch} \) is set to a low value (\( \sim 3 \AA \)); by progressively increasing its value, the maximum of the driving force is shifted to larger violations. Thus once the formation of local secondary structures such as \( \alpha \)-helices has occurred, turns can be formed and tertiary structure folding can gradually take place. For the same reason, NOEs that are violated by more than \( r_{dist} \) are placed in a class \( \text{long} \) where they experience no force from the NOE restraints at all.

At the beginning of the calculations the hard sphere radii of the atoms are chosen as in the Lennard—Jones potential (i.e. \( s \) in equation 5 is set to 1.0). The calculations are initiated with 20 steps of Powell (1977) minimization to remove some bad non-bonded contacts. This is followed by Phase 1 of the simulated annealing protocol. The initial velocities at \( t = 0 \) ps are chosen from a Maxwellian distribution at 1000K. This temperature is chosen to ensure that local minima along the convergence pathway towards the global minimum region of the target function \( F_{\text{NOE}} \) can be overcome. Each cycle of annealing comprises 40 steps with a time step of 1 fs in which the non-bonded contact list is updated every eight steps and the velocities are rescaled to 1000K every 20 steps. After every cycle of annealing the force constant \( k_s \) for \( F_{\text{short}} \) is increased up to a maximum value of 15.8 kcal/mol/Å\(^2\) by multiplying its value by 10\(^{0.1}\). The value of \( k_{NOE} \) is then evaluated with the new value of \( k_s \), and if \( F_{\text{NOE}} \)
Fig. 2. Potential form of $F_{\text{short}}$ (equation 6) for $r_0 = 3.0 \text{ Å}$, $k_s = 1.0 \text{ kcal/mol/Å}^2$, $r^1 = 2.0 \text{ Å}$, $r^u = 3.5 \text{ Å}$, $r_{\text{switch}} = 1.5 \text{ Å}$ and $c = 0$. The positions of $r^1$, $r^u$ and $r_{\text{switch}} + r^u$ are indicated.

Fig. 3. Histogram showing the distribution of the number of NOE violations $N_{\text{viol}} (r_i > r_j)$ as a function of the value of the violation $(r_i - r_j)$ in the initial extended strand structures of CPI, crambin and BSPI-2. The Δ interval for counting $N_{\text{viol}}$ is 1 Å. Note that the vertical scale is logarithmic.

$> 2500 \text{ kcal/mol}$ $k_s$ is divided by $10^{1.1}$ until $F_{\text{NOE}} < 2500 \text{ kcal/mol}$. Once $k_s$ has reached its maximum value of $15.8 \text{ kcal/mol/Å}^2$, the NOE restraints are further reclassified between class short and class final if $\text{viol}_j$ is less than $r_{\text{dist}j}$. The value of $r_{\text{dist}j}$ chosen is 1.0 Å, and $F_{\text{final}}$ is a square-well function with a force constant $k_f$ of 60 kcal/mol/Å$^2$ which is never changed.

$$F_{\text{final}} = \begin{cases} k_f (\text{viol}_j)^4 & \text{if } r_{ij} > r_j^u \\ 0 & \text{if } r_j^1 \leq r_{ij} \leq r_j^u \\ k_f (\text{viol}_j)^4 & \text{if } r_{ij} < r_j^1 \end{cases}$$

Thus class final contains all the NOE restraints which have converged. In addition, the restraints in the long and short classes...
are counted and the smallest violation in class long is calculated. Three cases are distinguished: (i) if both the long and short classes are empty and \( F_{NOE} \) has a value below its target value (in this case 120 kcal/mol), global as well as local convergence has occurred, the Phase 1 calculation is stopped and the simulation proceeds directly to Phase 2 of the annealing protocol; (ii) if only the short class is empty and the value of \( F_{NOE} \) lies below its target value, 'local convergence' has been achieved, and the value of \( rdist^{\text{short}} \) for the reclassification between classes long and short is set to just above (0.02 Å) the smallest violations in class short; and (iii) if neither case (i) nor case (ii) applies then \( rdist^{\text{short}} \) is increased by 0.02 Å. Additionally, in cases (ii) and (iii) the value of \( rswitch \) is increased by 0.01 Å and the NOE restraints are reclassified between classes long and short.

The rationale behind the grouping of all 'converged NOE restraints' in class final is to ensure that once secondary structure elements have formed they are preserved and not disrupted at a later stage during the course of the simulation. This is required since there is no force other than the NOE restraints to stabilize such secondary structure elements, and the scale of the short potential has to be reduced drastically at times as longer-range NOEs are incorporated into \( F_{NOE} \). For this reason the force constant on the final potential is never reduced. To ensure that the reclassification only takes place once NOEs have really converged and to allow some rearrangement of the local structure, NOEs are only reclassified between the final and short classes when the force constant for \( F_{short} \) is at its maximum value.

After every 10 cycles of annealing, the velocities are partially rerandomized by adding the variation in the random number seed used to assign the initial velocities back to a temperature of 1000 K. This is done to slow down large-scale rigid body motions and to introduce a further random element into the protocol, in addition to that arising from the variation in the random number seed used to assign the initial velocities at the beginning of Phase 1. The random element arises insofar that a partial rerandomization of the velocities may change the direction of motion of the atoms in a non-deterministic manner.

The maximum number of cycles for Phase 1 was 250 for CPI and crambin, and 350 for BSPI-2 (the larger the protein, the more
M. Nilges et al.

Table II. NOE deviations and violations, deviations from ideality, van der Waals energies and radii of gyration

<table>
<thead>
<tr>
<th>Structure</th>
<th>NOE_m.s.</th>
<th>NOE_viol.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Å)</td>
<td>(Å)</td>
</tr>
<tr>
<td></td>
<td>NOEs</td>
<td>NOEs</td>
</tr>
<tr>
<td></td>
<td>(Å)</td>
<td>(Å)</td>
</tr>
<tr>
<td></td>
<td>Bonds (Å)</td>
<td>Angles (deg)</td>
</tr>
<tr>
<td>Crambin</td>
<td>249</td>
<td>642</td>
</tr>
<tr>
<td>No. of terms</td>
<td>35.3</td>
<td>173</td>
</tr>
<tr>
<td>Ini</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;SA&gt;</td>
<td>0.12 ± 0.11</td>
<td>0.6 ± 0.5</td>
</tr>
<tr>
<td>SA</td>
<td>0.03</td>
<td>0</td>
</tr>
<tr>
<td>(SA)r</td>
<td>0.09</td>
<td>1</td>
</tr>
<tr>
<td>&lt;DG&gt;</td>
<td>0.14 ± 0.04</td>
<td>1.9 ± 0.2</td>
</tr>
<tr>
<td>&lt;RD&gt;</td>
<td>0.08 ± 0.01</td>
<td>0.6 ± 1.3</td>
</tr>
<tr>
<td>X-ray</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>CPI</td>
<td>318</td>
<td>565</td>
</tr>
<tr>
<td>No. of terms</td>
<td>27.7</td>
<td>147</td>
</tr>
<tr>
<td>Ini</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;SA&gt;</td>
<td>0.10 ± 0.01</td>
<td>0.1 ± 0.4</td>
</tr>
<tr>
<td>SA</td>
<td>0.07</td>
<td>3</td>
</tr>
<tr>
<td>(SA)r</td>
<td>0.08</td>
<td>0</td>
</tr>
<tr>
<td>&lt;DG&gt;</td>
<td>0.14 ± 0.05</td>
<td>6.7 ± 4.5</td>
</tr>
<tr>
<td>&lt;RDDG&gt;</td>
<td>0.05 ± 0.01</td>
<td>0.09 ± 0.3</td>
</tr>
<tr>
<td>X-ray</td>
<td>0.41</td>
<td>24</td>
</tr>
<tr>
<td>BSPI-2</td>
<td>403</td>
<td>1069</td>
</tr>
<tr>
<td>No. of terms</td>
<td>43.7</td>
<td>148</td>
</tr>
<tr>
<td>Ini</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;SA&gt;</td>
<td>0.10 ± 0.02</td>
<td>1.0 ± 1.7</td>
</tr>
<tr>
<td>SA</td>
<td>0.10</td>
<td>2</td>
</tr>
<tr>
<td>(SA)r</td>
<td>0.08</td>
<td>0</td>
</tr>
<tr>
<td>&lt;DG&gt;</td>
<td>0.17 ± 0.02</td>
<td>13.2 ± 3.4</td>
</tr>
<tr>
<td>&lt;RDDG&gt;</td>
<td>0.06 ± 0.006</td>
<td>0.5 ± 0.7</td>
</tr>
<tr>
<td>X-ray</td>
<td>0.32</td>
<td>22</td>
</tr>
</tbody>
</table>

The notation of the structures is the same as that in Table I.

NOE_m.s. is the r.m.s. difference (r.m.s.d.) between the calculated (r_0) and target restraints, calculated with respect to the upper (r_0^u) and lower limits (r_0^l) such that

\[
\text{r.m.s.d.} = \begin{cases} 
\frac{1}{n} \sum_{i=1}^{n} (r_{0i} - r_{0i}^u)^2/n^{1/2} & \text{if } r_{0i} > r_{0i}^u \\
0 & \text{if } r_{0i}^l \leq r_{0i} \leq r_{0i}^u \\
\frac{1}{n} \sum_{i=1}^{n} (r_{0i} - r_{0i}^l)^2/n^{1/2} & \text{if } r_{0i} < r_{0i}^l 
\end{cases}
\]

NOE_viol is the number of violations for which r_{0i} > (r_{0i}^u + 0.5 Å). The interproton distance restraints for crambin, CPI and BSPI-2 are taken from Clore et al. (1986a, 1987d) and (1987e) respectively. In the case of crambin and CPI, the restraints include nine additional restraints for the three disulphide bridges present in these two proteins. Distances involving methyl and methylene protons are calculated using centre averaging with the same corrections to the upper limits of the target distances as that used in the essentially equivalent pseudo-atom representation (Wuthrich et al., 1983).

The van der Waals energy is calculated using the Lennard–Jones potential and parameters in the CHARMM empirical energy function (Brooks et al., 1983). Note that this energy term is not included in the target function (cf. equation 1) whose global minimum is simulated annealing. The only non-bonded contact term present in the target function is a hard-sphere repulsion term (cf. equation 5).

The van der Waals energy for the <DG> structures of CPI range from -67 to 4248 kcal/mol.

Cycles required. If at this stage there are still violations in class long, failure of convergence is presumed and the calculation comes to a complete halt. If, on the other hand, there are no violations in class long, the NOE restraints are once again reclassified between classes short and final at 10 Å in order to place all the NOEs into the final class. This is followed by Phase 2 of the annealing protocol which comprises 20 cycles of 100 steps of annealing at 300K. The velocities are rescaled to 300K after every cycle and the force constant kᵣ for the repulsion target function F_{rep} (cf. equation 5) is increased in steps of 0.2 from an initial value of 0.4 kcal/mol/Å² to a final value of 4 kcal/mol/Å². The values of the hard sphere atom radii are set to 0.8 times their Lennard–Jones values (i.e. s = 0.8 in equation 5). The resulting values are approximately the same as those used in the various distance geometry programs. The value of 4 kcal/mol/Å² for the force constant kᵣ was found to be sufficient to ensure that no close non-bonded contacts occur. Finally, Phase 2 is followed by 200 steps of restrained Powell minimization to complete the simulation.

Calculations on crambin, CPI and BSPI-2

The calculations on crambin, CPI and BSPI-2 (the 64-residue proteolytic fragment comprising residues 20–83) were carried out starting from an extended β-strand (r.m.s. atomic difference of ~38 Å, ~33 Å and ~58 Å from the respective crystal structures) using the same NOE distance data set that was employed in our previous studies (Clore et al., 1986a, 1987d,e).

In the case of crambin the NOE data set consisted of 240 interproton distances derived from the crystal structure (Hendrickson and Teeter, 1981), while for CPI and BSPI-2 they comprised...
Three-dimensional protein structures by simulated annealing

Fig. 4. Atomic r.m.s. distribution of the backbone (C, C', N, O) atoms of the <SA> structures about the mean structure SA (A, C and E), and atomic r.m.s. difference between the <SA> (★) and (SA)r (△) structures on the one hand and the corresponding X-ray structures on the other (B, D and F) for CPI (A and B), crambin (C and D) and BSPI-2 (E and F). The filled-in circles (★) represent the average r.m.s. difference between the <SA> structures and either the mean SA structure (A, C or E) or the X-ray structure (B, D and F), and the bars represent the standard deviations in these values. In the case of CPI, the <SA> structures are best fitted to residues 2-39 of the mean SA structure (A) and to residues 2-38 of the X-ray structure (B); in the case of BSPI-2 all the best fits are carried out with respect to residues 22-83.

309 and 403 interproton distances, respectively, derived from NOE measurements. The lower limit (r_{1}^0) for all the restraints was 1.8 Å, while the upper limits (r_{2}^0) were set to 2.7, 3.3 and 5 Å, corresponding to strong, medium and weak NOEs. Figure 3 shows the distribution of NOE violations in the initial structures revealing violations up to 88, 125 and 203 Å for CPI, crambin and BSPI-2 respectively. Note that crambin and BSPI-2 exhibit distinctive gaps in the distribution of the initial violations, while CPI shows a continuum of violations. As a result, class short is empty at several stages during the calculations in the case of crambin and BSPI-2, indicating that local convergence has occurred. For CPI, on the other hand, long-range NOE restraints
Fig. 5. (A) Best-fit superposition of the backbone (C, C\(^\alpha\), N) atoms of the nine converged SA structures of crambin; (B) best-fit superposition of the backbone (C, C\(^\alpha\), N, O) atoms of the (SA)r structure (thick lines) with the X-ray structure of crambin (thin lines). The three-picture stereo system used in this figure enables readers with both natural and cross-over stereo vision to view the images. For normal vision, select the left and centre images; for cross-over vision, use the centre and right images.

are taken into class short during almost the entire course of the calculations. In the case of distances involving methyl and methylenic protons, the NOE target function \( F_{\text{NOE}} \) was calculated using \(< r_c >\) centre averaging with the same corrections of the upper limits of the target distances used in the equivalent pseudo-atom representation (Wüthrich et al., 1983). An additional nine restraints were included for the three disulphide bonds present in crambin and CPI. (Note for each disulphide bridge there are three distance restraints, \( S_i-S_j \), \( S_i-C^\alpha_j \) and \( S_j-C^\alpha_i \), whose target values were set to 2.02 ± 0.02, 2.99 ± 0.5 and 2.99 ± 0.5 Å respectively.) These disulphide bridge restraints are treated in exactly the same manner as the interproton distance restraints.

A total of 13 calculations were carried out for crambin, 10 for CPI and 10 for BSPI-2, differing in the values of the random number seed used for the assignment of the velocities at \( t = 0 \) ps and for the partial rerandomization of velocities during the course of the simulations. Nine of the crambin calculations, eight of the CPI ones and five of the BSPI-2 ones converged to similar final structures with an average backbone (N, C\(^\alpha\), CO, O) atomic r.m.s. difference between them of 2.2 ± 0.3, 2.4 ± 0.3 and 2.5 ± 0.2 Å respectively (Table I), all of which satisfied the experimental restraints within the errors specified (Table II). This success rate is comparable in our experience with that obtained for these proteins with the metrix matrix distance geometry program DISGEO (Havel, 1986) and significantly higher than that obtained using the restrained molecular dynamics protocols used previously in our model crambin calculations (Clore et al., 1986a; Brünger et al., 1986; G.M.Clore, M.Nilges and A.T. Brünger, unpublished data). Typical computing times per simulation were ~ 1 h for CPI, ~ 1.5 h for crambin and ~4 h for BSPI-2 on a CONVEX-C1XP computer. Plots of atomic r.m.s. difference as a function of residue number between
Fig. 6. (A) Best-fit superposition of the backbone (C, C', N) atoms of the eight converged SA structures of CPI; (B) best-fit superposition (residues 2–38) of the backbone (C, C', N, O) atoms of the (SA)r structure (thick lines) with the X-ray structure of CPI (thin lines). The three-picture stereo system used in this figure enables readers with both natural and cross-over stereo vision to view the images. For normal vision, select the left and centre images; for cross-over vision, use the centre and right images.

the individual converged <SA> structures and the mean SA structure derived by averaging their coordinates are shown in Figure 4, and stereoviews of best-fit superpositions of the converged <SA> structures are shown in Figures 5–7.

From the atomic r.m.s. distribution of the <SA> structures (Table I) it is clear that the size of the conformational space sampled by simulated annealing is comparable with that sampled by restrained molecular dynamics and slightly larger than that sampled by metric matrix distance geometry. Although all the simulated annealing calculations start off from the same initial structure, it must be emphasized that varying the random number seed used in the assignment of the initial velocities ensures that different convergence pathways are followed such that the different trajectories do not possess any common intermediate structures. That is to say that during the initial stages of the simulation the different trajectories diverge. In the case of the crambin trajectories the maximum average and maximum absolute backbone atomic r.m.s. differences are 5.4 and 8.1 Å respectively. As the simulation proceeds, and more and more NOEs are satisfied, so convergence between the different trajectories gradually occurs. This is illustrated in Figure 8. One cannot expect the trajectories, however, to diverge to the extent that the distribution of the structures between the different trajectories would be totally random (with an expected mean backbone atomic r.m.s. difference of ~10 Å for a protein the size of crambin; Cohen and Sternberg, 1980). The reason for this is twofold. First, local convergence, driven by the short-range NOEs, occurs from the beginning of the calculations. Second, the structures have a tendency to stay extended in the absence of tertiary folding forces (i.e. the long-range NOEs) due to their intrinsic inertia (arising from the fact that the masses of the atoms enter explicitly into the calculations; cf. equation 2). Nevertheless, we feel that this does not introduce any significant bias into the end result, particularly as misfolding can also occur, and in our view it is equivalent to using a set of randomly chosen initial structures in static real space methods (Braun and Go, 1985; Billeter et al., 1987).

The non-bonded contacts in the converged structures are all good, as evidenced by negative values for van der Waals energy calculated using the CHARMM empirical energy function (Table II). Indeed they are comparable with those of the restrained molecular dynamics structures. Thus, our choice of a final van der Waals radius, a factor of 0.8 smaller than the one used to compute the Lennard–Jones van der Waals energy, is completely reasonable. Further, these results suggest that the converged <SA> structures do not require any further refinement by restrained molecular dynamics. In this respect, we note that the
non-bonded contacts in the metric matrix distance geometry structures are considerably poorer, as the van der Waals energies tend to be large and positive, and are only improved by additional restrained molecular dynamics refinement.

The converged <SA> structures are all reasonably close to the respective X-ray structures with an average backbone atomic r.m.s. difference of 2–2.5 Å (Table I). Averaging the structures results in mean structures that are close to their respective X-ray structure than any of the individual <SA> structures. The same is true of the metric matrix distance geometry and restrained molecular dynamics structures. Interestingly, the r.m.s. differences between the mean structures calculated by the three different methods are comparable with the difference between the individual mean structures and the X-ray structures. The average SA structures are clearly very bad both with respect to stereochemistry and non-bonded contacts (Table II). These are easily corrected by 1000 cycles of Powell restrained minimization with only minor accompanying atomic r.m.s. shifts to generate the structures (SA)r (see Table I). In this procedure the restraints force constant $k_F$ for the final NOE potential $F_{NOE}$ is kept constant at 60 kcal/mol/Å$^2$, the force constant $k_F$ for $F_{rep}$ is multiplied by two every 20 cycles from an initial value of 0.2 kcal/mol/Å$^2$ to a maximum value of 4 kcal/mol/Å$^2$, and the hard-sphere van der Waals radii are kept constant at 0.8 times their Lennard–Jones values. Best-fit superpositions of the (SA)r and X-ray structures are shown in Figures 5 (crambin), 6 (CPI) and 7 (BSPI-2).

Examination of the radii of gyration indicates that the <SA> structures, like the distance geometry structures, tend to be a little expanded relative to the X-ray structure, whereas the restrained dynamics structures tend to be compressed (Table II). This is due to the different representation of the van der Waals interactions used in the different methods (i.e., simple repulsion terms in the case of the simulated annealing and distance geometry calculations compared with a full Lennard–Jones potential with an attractive component in the case of the restrained molecular dynamics calculations).

**Concluding remarks**

In this paper we have shown that simulated annealing is an effective method of determining three-dimensional structures on the basis of interproton distance data. The present calculations indicate that it is comparable in speed with distance geometry calculations and significantly faster than restrained molecular dynamics calculations employing a full empirical energy function. This is largely due to the replacement of the non-bonded interaction potentials in the empirical energy function by a simple van der Waals repulsion term. In addition, the agreement with the experimental interproton distance restraints and the quality of the non-bonded contacts exhibited by the converged SA structures is comparable with that of structures obtained or refined by restrained molecular dynamics and significantly better than that of structures obtained by metric matrix distance geometry calculations alone (see Table II). Critical to the success of the
method is the protocol employed, in particular the way in which the NOE distances are partitioned between different functional forms.

At this stage we would not claim that the radius of convergence of the simulated annealing method is any larger than that of the various methods already published. Nevertheless, it forms a useful addition to the arsenal of tools available to the NMR spectroscopist interested in solving three-dimensional structures of proteins. This is particularly so as the convergence properties of the various methods are likely to be dependent on both the nature of the structure being solved and the extent of the experimental data at hand.

Acknowledgements
We thank Drs Flemming Poulson and Mogen Kjaer for making available their experimental data on BSPI-2. This work was supported by the Max-Planck Gesellschaft and Grant no. 321/4003/O3189O9A from the Bundesministerium für Forschung und Technologie (G.M.C. and A.M.G.).

References

Received on January 22, 1988, accepted on February 29, 1988