J. Mol. Biol. (1995) **251**, 308–326

JMB

Global Fold Determination from a Small Number of Distance Restraints

A. Aszódi[1]*, M. J. Gradwell[2] and W. R. Taylor[1]

[1]*Division of Mathematical Biology* [2]*Division of Molecular Structure, National Institute for Medical Research The Ridgeway, Mill Hill London, NW7 1AA, UK*

**Corresponding author*

We have designed a distance geometry-based method for obtaining the tertiary fold of a protein from a limited number of structure-specific distance restraints and the secondary structure assignment. Interresidue distances were predicted from patterns of conserved hydrophobic amino acids deduced from multiple alignments. A simple model chain representing the protein was then folded by projecting its distance matrix into Euclidean spaces with gradually decreasing dimensionality until a final three-dimensional embedding was achieved. Tangled conformations produced by the projection steps were eliminated using a novel filtering algorithm. Information on various aspects of protein structure such as accessibility and chirality was incorporated into the conformation refinement, increasing the robustness of the algorithm. The method successfully identified the correct folds of three small proteins from a small number of restraints, indicating that it could serve as a useful computational tool in protein structure determination from NMR data.

Keywords: protein folding; distance geometry; structure determination; distance prediction; molecular modelling

Introduction

Most interactions governing the formation of native protein structures can be coded in terms of interresidue or interatomic distances, providing a set of internal coordinates that can be processed conveniently by distance geometry methods in simulations. On the experimental side, the impressive development of NMR spectroscopy now enables the determination of the molecular conformation of small proteins in solution. Structural information from NMR experiments comes in the form of distance restraints (i.e. acceptable interatomic distance ranges flanked by lower and upper bounds), usually estimated from 2D NOE spectra. Plausible conformations that conform to these distance restraints can be generated by a wide variety of computational tools (Kuntz *et al*., 1989; Havel, 1991; James, 1994), most of which fall into one of the broad categories of restrained molecular dynamics and distance geometry methods. The latter are natural candidates for the NMR structure determination problem, since they enable the simultaneous satisfaction of all distance restraints in addition to the holonomic constraints (bond lengths, angles, etc.) imposed upon the structure and generate model conformations that are compatible with the information obtained from the experiments. Frequently a hybrid approach is chosen whereby the distance geometry program generates a family of plausible structures that can then be refined by molecular dynamics simulations.

The route from the raw NMR spectra to the Brookhaven databank is often tortuous: the distance restraints may be insufficient, too lax or inconsistent. Robust algorithms are needed that can produce adequate global folds even if the restraints are not tight or there are just a few of them. A theoretical analysis of the minimum number of restraints necessary for a successful structure determination, based on statistical mechanics, was carried out by Gutin & Shakhnovich (1994). Smith-Brown *et al*. (1993) folded various protein chains by a Monte Carlo method guided by a small amount of distance restraints, while Hoch & Stern (1992) compared a restrained molecular dynamics optimisation with the classic distance geometry protocol (Crippen & Havel, 1988). Their results suggest that it is possible to design reasonably robust methods that can locate the correct fold using a small amount of distance information. Similar conclusions have been drawn when the problem has been approached from a

Abbreviations used: 2D and 3D, two- and three-dimensional; NOE, nuclear Overhauser effect; PDB, Protein Data Bank.

0022–2836/95/320308–19 $08.00/0

structure prediction viewpoint (Taylor, 1993; Dandekar & Argos, 1994).

We have shown that model polypeptide chains can be folded into compact three-dimensional conformations possessing a hydrophobic core and secondary structure using distance geometry techniques (Aszódi & Taylor, 1994a,b). The model chains were random 1:1 copolymers of ''hydrophobic'' and ''hydrophilic'' monomers, and the interresidue distances were set according to the hydrophobicity of the residue pairs. Although the chains folded into protein-like conformations, they could not be regarded as models of any particular protein. A logical extension of these studies was to make the simulations more realistic by modelling the polypeptide chain more accurately and by supplying external distance restraints to the program. In order to increase the robustness of the algorithm, a considerable amount of background information about various aspects of protein structure such as hydrophobic packing, interresidue distance distribution, general topological properties and chirality was incorporated.

Algorithms

The algorithm presented in this work is the extension of the gradual projection approach implemented in the program DRAGON-2 (Aszódi & Taylor, 1994b) and the core algorithm remained essentially the same. To avoid repetition, only the additions and improvements to the method are described below.

Strategy

The objective of our calculations was to simulate the determination of the structures of small monomeric proteins by NMR. The following pieces of information were assumed to be available.

(1) The sequence of the protein.
(2) A set of sequences homologous to that of the protein to be modelled.
(3) A full assignment of secondary structure derived possibly from short-range NOEs.
(4) A set of long-range distance restraints in the form of lower and upper distance bounds between amino acid side-chains.

The sequence and secondary structure assignment information was obtained from the Protein Data Bank (Bernstein *et al*., 1977). First, a multiple sequence alignment was constructed that provided information about the conservation of individual amino acid residues. The conservation information, coupled with hydrophobicity data, was in turn used to predict the distances between those residues for which no extra distance data were available. These predicted distances, together with simulated NOE-derived restraints, were submitted to the gradual projection algorithm. The number of simulated NOE restraints was varied to determine the robustness of the approach. For each restraint set, 25 model structures were generated. The same model chains, complete with secondary structure assignment and distance restraint data were also submitted to the program X-PLOR (Brünger, 1992) and the resulting model conformations served as controls.

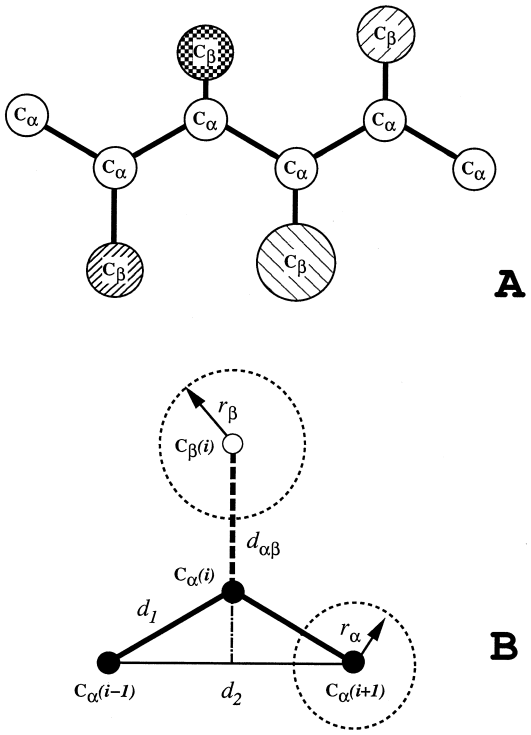


Figure 1. The geometry of the model polypeptide chain. A, The backbone is built of C^{α} atoms and the side-chains are represented by pseudo-C^{β} atoms corresponding to the 20 natural amino acids. The different chemical identities are indicated by the different sizes and patterns of the pseudo-C^{β} atoms. B, the *i*th pseudo-C^{β} atom (open circle) lay in the plane of the $(i-1)$th, *i*th and $(i+1)$th C^{α} atoms (filled circles). The distance $d_{\alpha\beta}$ between a C^{α} atom and its corresponding pseudo-C^{β} atom as well as the van der Waals radii of the pseudo-C^{β} atoms r_{β} depended on the amino acid type of the monomer (see Table 1).

The model chain

The model chain was represented as a linear heteropolymer of the 20 natural amino acids. The geometric and physicochemical properties of the amino acid monomers were described by a set of attributes as follows.

Chain geometry

The polypeptide chain was modelled by an α-carbon backbone and single C^{β} atoms (Figure 1A) representing amino acid side-chains (''lollipop model'', Levitt, 1976). The positions of C^{β} atoms were determined by the backbone conformation: the position of the *i*th dummy C^{β} atom was obtained by reflecting the midpoint between the $(i-1)$th and $(i+1)$th C^{α} atoms through the *i*th C^{α} atom (Figure 1B) as described (Aszódi & Taylor, 1994b). However,

Table 1. Properties of the model amino acids

Monomer properties			
Amino acid	r_β (Å)	$d_{\alpha\beta}$ (Å)	H
Ala	1.76	1.53	1.73
Cys	2.03	2.06	0.84
Asp	2.23	2.47	0.03
Glu	2.48	3.11	0.01
Phe	2.79	3.41	1.48
Gly	0.00	0.00	1.27
His	2.61	3.15	0.06
Ile	2.60	2.33	3.46
Lys	2.67	3.44	0.03
Leu	2.60	2.61	2.56
Met	2.61	2.96	0.86
Asn	2.35	2.45	0.01
Pro	2.21	1.88	0.18
Gln	2.57	3.07	0.03
Arg	2.88	4.11	0.00
Ser	1.94	1.89	0.49
Thr	2.25	1.95	0.59
Val	2.38	1.97	2.46
Trp	3.07	3.87	0.74
Tyr	2.88	3.76	0.59

r_β is the van der Waals radius of the pseudo-C^β side-chain atom. $d_{\alpha\beta}$ is the distance between the C^α and pseudo-C^β atoms. H is the hydrophobicity of the residue according to Levitt.

for each amino acid monomer, the distance between the C^α and C^β atoms ($d_{\alpha\beta}$) now corresponded to the average C^α to centroid distances observed in native proteins (Table 1). Volume exclusion was modelled by centering hard van der Waals spheres on the atoms. For C^α atoms, the van der Waals radius was set to $r_\alpha = 2.0$ Å, regardless of amino acid type. For the C^β atoms, the van der Waals radii r_β were set so that the van der Waals sphere had the same volume as the corresponding side-chain type in native proteins (Table 1).

Other attributes

In addition to the geometry data described above, each monomer in the chain had attributes describing hydrophobicity and conservation. Levitt's hydrophobicity scale (Levitt, 1978) was chosen from the many scales available (Table 1), but the program can accommodate any scale provided the values are shifted so that all of them are non-negative. Conservation of each residue in the model sequence was calculated from the multiple alignment as described below.

Chirality

In all distance geometry-based methods chirality needs special treatment, since the distance matrices of a point set and its mirror image are identical. Also, objects that were chiral in N-dimensional spaces become achiral in $N + 1$ dimensions so that handedness consistency cannot be guaranteed in the gradual projection method (Aszódi & Taylor, 1994b). While the ultimate source of structural asymmetry in proteins is the consistent handedness of the amino acids, our model chains were built from symmetric monomers and the correct 3D chirality was imposed upon them in a separate refinement stage. The handedness of helices and the asymmetric twist of β-sheets were adjusted through the torsion angle about the main-chain H bonds as described (Aszódi & Taylor, 1994b).

Interresidue distances

Although the model chain monomers were composed of two atoms, the position of the fake C^β atoms were determined by the C^α backbone conformation and the C^α to centroid separation data, therefore the model chain could fully be described by the C^α–C^α distances. These distances, henceforth referred to as interresidue distances, fell in the following categories.

(1) ''Hard'' distances: these were determined by the protein chain geometry, e.g. the virtual C^α–C^α bond lengths and distances between residues within the same secondary structural element and were assumed to be accurately known.

(2) ''Experimental'' distances: these were distance restraints specific to the target molecule and were assumed to have been supplied by experimental measurements such as NMR spectroscopy, usually with less accuracy than the hard distances.

(3) ''Soft'' distances: this large category contained all the remaining distances, of which little more was known than that they fell between the lower and upper limits determined by the bump distances and the estimated diameter of the molecule, respectively.

Each category required different treatment, as outlined below.

Hard distances

One of the few distances that can be predicted with any certainty is the virtual C^α bond length $d_1 = 3.8$ Å. The separation between second neighbours is more variable; in the present study these were set to an average value of $d_2 = 6.0$ Å (Figure 1B) but allowing for a moderate wobble of the C^α virtual bond angles (Aszódi & Taylor, 1994a).

Since the secondary structure assignment was assumed to be known, the C^α distances derived from ideal helical and sheet conformations (Pauling & Corey, 1951a,b; Wako & Scheraga, 1982) were incorporated into the matrix of desired distances. Most of these values are fairly constant due to the strict geometric requirements imposed upon the protein chain by backbone hydrogen-bonding.

Simulated NOE distance restraints

Simulated long-range distance restraints were obtained from the PDB structures of the proteins to be modelled. (Long-range here refers to sequential rather than spatial separation: in general, only distances between residues separated by at least five other residues in the sequence were considered.) All

pairwise distances between these amino acid side-chain centroids were calculated and those shorter than a threshold of 5.0 Å were selected. These selected distances were converted to lower and upper distance bounds by subtracting and adding 2.0 Å, respectively. In order to obtain an abundant set of restraints, the threshold was set to 7.0 Å. To simulate the more realistic experimental case when the number of long-range NOEs may be rather small, a variable fraction of randomly chosen restraints were gradually removed from the basic 5.0 Å threshold set to produce more and more sparse sets, following Hoch & Stern (1992). In general, one set was constructed for a given number of restraints. The sensitivity of the method to the choice of restraints was tested by generating 20 different random datasets containing the same number of restraints for each target structure.

Soft distances

In previous versions of DRAGON, these virtually unknown distances were adjusted so that pairs of ''hydrophobic'' residues should come close together. This crude model of the hydrophobic effect was significantly refined in the present program. Interresidue distances were predicted from the pairwise conserved hydrophobicity score (Taylor, 1991), based on the assumption that pairs of residues that are both conserved and hydrophobic, are likely to be close together in the core of the molecule.

Sequences homologous to the target sequences were obtained by searching the SwissProt database (Bairoch & Boeckmann, 1991) and a final set that originated from a wide range of evolutionarily distant organisms showing only a modest level of homology to the model sequence was selected by hand. Multiple alignments of these sequences were generated by the MULTAL algorithm (Taylor, 1988) as implemented in the CAMELEON package (Version 3.0C, Oxford Molecular) using the default parameter setting.

The conservation g_i at the ith position of an alignment of N sequences was measured by the average of the pairwise similarity scores of all amino acids in the position:

$$g_i = \frac{2}{N(N-1)} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} M(R_{ij}, R_{ik}) \qquad (1)$$

where R_{ij} is the type of the amino acid in the jth sequence in alignment position i and $M(\cdot, \cdot)$ is an entry in the PAM250 amino acid similarity matrix (Dayhoff *et al*., 1978). The similarity matrix was scaled so that all entries were $\geqslant 0$ by subtracting the smallest entry from all the others. The similarity score between a gap and any amino acid was always set to zero. The conservation value was normalised by the maximal score encountered among the amino acid pairs:

$$c_i = \frac{g_i}{max_{j<k} M(R_{ij}, R_{ik})} \qquad (2)$$

ensuring that $0 \leqslant c_i \leqslant 1$. The pairwise hydrophobic packing score could then be defined as:

$$h_{ij} = c_i H_i + c_j H_j \qquad (3)$$

where c_i is the conservation and H_i is the average Levitt hydrophobicity (Levitt, 1978) of the amino acids in the ith sequential position. The hydrophobic scores were converted into predicted distances *via* the transformation:

$$d_{ij} = -p_1 h_{ij}^{p_2} + p_3 \qquad (4)$$

where the three parameters $p_1, p_2, p_3 > 0$ were estimated by non-linear regression so that the distribution of the predicted distances matched the observed interresidue distance distribution in a representative set of small monomeric proteins (Aszódi & Taylor, 1995).

Distance adjustments

For hard and soft distance data, the actual values in the distance matrix were ''massaged'' towards the corresponding desired values by a convex function (Aszódi & Taylor, 1994a). The extent of the adjustment was regulated by ''strictness'' values, which were close to 1 for hard distances and close to 0 for soft distances. In energetic terms, this adjustment strategy is equivalent to replacing the detailed description of the interresidue potentials by their corresponding minima.

Adjustment of ''experimental'' distances was carried out so that if an interresidue distance was shorter than the lower simulated NOE limit then it was increased, if it was longer than the upper limit then it was reduced. No adjustment was performed if the distance in question fell between the two limits. This strategy corresponds to a potential well with a flat bottom.

For all interresidue distances, a minimal and maximal separation can be calculated; the minimal value, representing steric repulsion, was determined by the sum of the appropriate van der Waals radii, the maximal value by the expected maximal separation for two residues linked by a fully extended chain or by the estimated diameter of the molecule (Aszódi & Taylor, 1994a), whichever the smaller. These minimal and maximal limits were applied to all residue pairs as hard restraints.

Residue accessibility

Accessibility was calculated by the ''cone'' algorithm (Aszódi & Taylor, 1994b) with minor modifications, the most important being that the cone for the kth residue was required to contain only those residues closer than 8.0 Å. This local approach improved the estimation of burial for residues situated near crevices on the protein surface and resulted in a considerable increase in computing speed.

In previous versions of DRAGON, residue burial adjustment was based upon a simple logic: it was undesirable for hydrophobic residues to be on the

surface, and equally undesirable for hydrophilics to reside in the interior. Since in the present study the residues had different chemical identities corresponding to the 20 natural amino acids with 20 different hydrophobicity values, a more detailed treatment was necessary. To this end the distribution of conic accessibility values for all 20 amino acids in a reference set of protein structures were approximated by histograms. For each amino acid type, the central 80% portion of its accessibility distribution was considered acceptable and if a residue in the model had an accessibility value within this range then no adjustment was performed. If a residue had an accessibility value that fell into the lower 10% of the accessibility distribution corresponding to its amino acid type, then that residue was considered too exposed and it was moved towards the centre of the molecule. Similarly, residues that were more buried than 90% of the native residues of the same amino acid type were moved towards the surface. The extent of the adjustment in both cases was graded according to the actual position in the accessibility distribution curve; wild outliers were adjusted more strictly.

Tangles

The various distance geometry approaches based on the projection algorithm just generate point coordinates from interpoint distances and have no information about the connectivity of the polypeptide chain. As a result, tangled conformations are frequently observed. Such structures are very few, if not completely absent, among naturally occurring folded polypeptides (Connolly *et al*., 1980) and therefore tangled conformations produced by the distance geometry algorithm must be filtered out. To this end a simple and fast heuristic was devised that can correct most tangled conformations.

First, the polypeptide chain was divided into segments according to its secondary structure. The basic idea of tangle filtering was that different segments were not allowed to penetrate each other. To this end, chain segments were represented with suitably chosen sets of tetrahedra and penetration was detected if another segment intersected at least one of these tetrahedra.

The concept of containment detection is easier to understand in a unidimensional example (Figure 2A). If the point P_{in} is between the two points R_0 and R_1, then its position vector $\vec{P}_{in}$ can be expressed as a linear combination of the position vectors $\vec{R}_0$ and $\vec{R}_1$:

$$\vec{P}_{in} = s_0\vec{R}_0 + s_1\vec{R}_1 \tag{5}$$

so that the coefficients s_0 and s_1 obey the following relations:

$$s_0 + s_1 = 1, \quad 0 \leqslant s_0, s_1 \leqslant 1 \tag{6}$$

The two points R_0 and R_1 define a unidimensional simplex. By analogy, the point P_{in} is lying within a tetrahedron (the 3D simplex) defined by the four

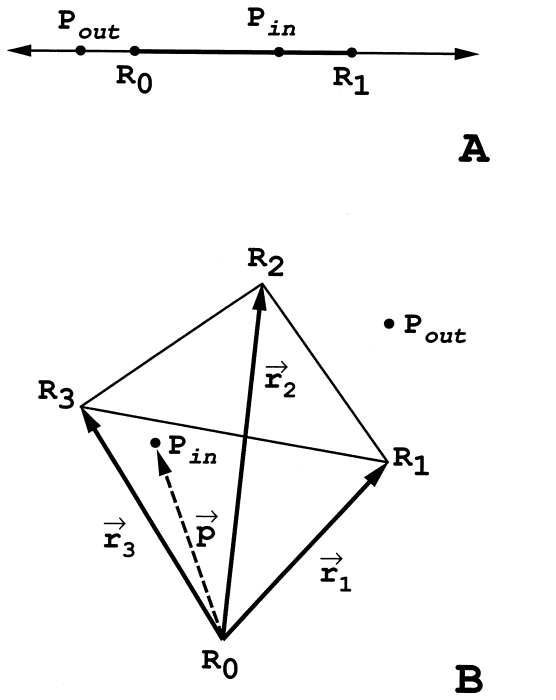


Figure 2. The tetrahedral containment algorithm. A, The unidimensional case. B, The three-dimensional case. See the text for explanation and details.

points R_0, R_1, R_2, R_3 if the coefficients s_i, $i = 0, \ldots, 3$ in the linear combination:

$$\vec{P}_{in} = \sum_{i=0}^{3} s_i\vec{R}_i \tag{7}$$

satisfy the following relations:

$$\sum_{i=0}^{3} s_i = 1, \quad \forall s_i : 0 \leqslant s_i \leqslant 1 \tag{8}$$

The coefficients can be found as follows. By subtracting $\vec{R}_0$ from both sides of equation (7) and expressing $s_0 = s_1 + s_2 + s_3$ the equation:

$$\vec{p} = s_1\vec{r}_1 + s_2\vec{r}_2 + s_3\vec{r}_3 \tag{9}$$

is obtained. In other words, the vector $\vec{s} = (s_1, s_2, s_3)$ corresponds to the position vector $\vec{p}$ of P in a local coordinate system with origin R_0 and (non-orthogonal) base vectors $\vec{r}_1$, $\vec{r}_2$, $\vec{r}_3$ (Figure 2B). Equation (9) can be recast in matrix form:

$$\boldsymbol{R}\cdot\vec{s} = \vec{p} \tag{10}$$

where the columns of the matrix **R** are the base vectors:

$$\boldsymbol{R} = [\vec{r}_1|\vec{r}_2|\vec{r}_3] \tag{11}$$

Equation (10) can be solved by singular value decomposition (Rózsa, 1991), which automatically finds the 3D subspace spanned by the tetrahedron even in $D > 3$ hyperspaces where the matrix **R** has D rows and three columns. For a proper tetrahedron, **R** has a rank $\rho(\boldsymbol{R}) = 3$. If the four points do not span a tetrahedron because they lie in a $D < 3$D subspace,

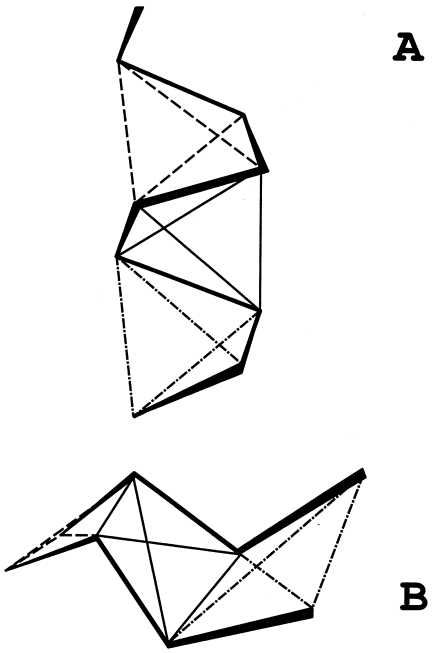


Figure 3. Sets of tetrahedra superimposed on secondary structures. The thick lines symbolise the C^{α}-backbones, the various thin lines represent the tetrahedra. A, Helices. Only every second tetrahedron is shown for clarity. B, Sheets.

this condition is indicated by the decomposition algorithm returning a rank $\rho(\boldsymbol{R}) < 3$.

Having found the coefficient vector $\hat{s} = (s_1, s_2, s_3)$, the missing coefficient s_0 can now be obtained by simply observing that $s_0 = s_1 + s_2 + s_3$. If all coefficients are in the range 0. . . 1, then the point P is inside the tetrahedron. To decide whether a portion of a line between two points P_A and P_B is contained by a tetrahedron even if the points themselves are outside, we observe that every point P on the P_A:P_B segment can be expressed as the linear combination:

$$P = \lambda P_A + (1 - \lambda)P_B, \quad 0 \leqslant \lambda \leqslant 1 \tag{12}$$

see equation (5), and the tetrahedral coefficients, s_i for P can be obtained from those of P_A and P_B (a_i and b_i, respectively) because:

$$P = \sum_{i=0}^{3} s_i \hat{R}_i = \sum_{i=0}^{3} (\lambda\alpha_i + (1 - \lambda)b_i)\hat{R}_i \tag{13}$$

If there is a range of $0 \leqslant \lambda_{min} \leqslant \lambda_{max} \leqslant 1$ such that all coefficients s_i satisfy the conditions in equation 8 then the segment P_A:P_B intersects the tetrahedron.

The chain segments were represented by tetrahedra as follows. Helices were made up by tetrahedra defined by the points $(i, i + 1, i + 2, i + 3)$, $(i + 1, i + 2, i + 3, i + 4)$, . . . where i is the index of the first C^{α} atom in the helix (Figure 3A). Sheets were treated as sets of non-overlapping tetrahedra made up by the four C^{α} atoms surrounding each main-chain H bond (Figure 3B). Finally, no tetrahedra were put on coil segments. Tangle detection was then carried out following each projection step by testing all tetrahedra in helices and sheets for containing pieces of the chain from all other segments. If two segments were found to intersect then both were moved away from each other as rigid bodies. The procedure was repeated until no more intersections were found.

† The *B*-factors of the tendamistat structure (3AIT) were simulated from RMS deviations between a set of NMR structures.

Model quality assessment

Rigid-body structural alignment

Model structures were compared with the known PDB structures by weighted rigid-body rotation using McLachlan's algorithm (McLachlan, 1979). Only the backbone C^{α} atoms were used in the comparison. The weights w_i were derived from the *B*-factors B_i in the PDB entries†, so that flexible portions of the target structure with large *B*-factors were given less weight:

$$w_i = \frac{B_{max} - B_i}{B_{max} - B_{min}} \tag{14}$$

where B_{min} and B_{max} were the smallest and largest *B*-factor in the known structure, respectively. The similarities were then expressed as C^{α} coordinates RMS deviations.

Precision and accuracy

Typically, the analysis of NMR-derived data leads to a collection of plausible structures. Since the native structure is unknown, direct assessment of modelling accuracy *via* RMS deviations is impossible. Accuracy may instead be determined by comparing the simulated NMR spectra calculated from the models with the experimental spectrum in a manner analogous to the *R*-factor determination in X-ray crystallography (James, 1994). In our simulation study, we had the advantage of knowing the target structures and therefore accuracy was monitored through RMS deviations as described above.

The precision, on the other hand, is easy to determine; a useful measure could be the average RMS deviation of the individual models from their average. Whereas a high level of precision (i.e. a set of very similar model structures) does not necessarily mean that the native conformation was found, a low level of precision usually indicates that the modelling algorithm did not perform adequately. The RMS C^{α} backbone deviations between the individual models and their average was determined and the average of these RMS values was used as an indication of modelling precision.

Topology comparison

Model structure topologies were compared with the corresponding native topologies by visual inspection. The target and model backbone coordi-

nates were smoothed by repeatedly scanning a moving average window along the chain, and the smoothed models aligned to the target were displayed and inspected using a simple molecular graphics program developed in our laboratory. Tangle checks were performed on the unsmoothed model backbones.

The model folds were classified into correct, slightly incorrect, incorrect and mirror image topologies. Folds in which just one secondary structure element was misplaced were regarded as slightly incorrect, two or more misplaced segments were classified as incorrect topologies. Topologies that were mirror images of the native topology formed a special group.

X-PLOR controls

The performance of the DRAGON algorithm was compared with that of X-PLOR Version 3.1 (Brünger, 1992), a popular tool frequently used for NMR structure determination. The model residue and parameter files were written for X-PLOR so that all data corresponded to the conventions used by DRAGON as closely as possible.

Chain geometry

A general carbon atom type was defined for representing backbone C^{α} atoms with a van der Waals radius of 2.0 Å. The monomer side-chains were described by 20 pseudo-C^{β} atoms for each amino acid type using the same van der Waals radii and C^{α}–C^{β} distances as in DRAGON (Table 1). The coplanarity of the atoms in the monomers was maintained by defining the appropriate improper axes of rotation. The virtual C^{α} bond angles and the C^{α}–C^{α}–C^{β} angles were allowed to vary around their mean values as in DRAGON. This was achieved by incorporating the equivalent distance ranges $C^{\alpha}(i-1)$–$C^{\alpha}(i+1)$, $C^{\alpha}(i-1)$–$C^{\beta}(i)$ and $C^{\beta}(i)$–$C^{\alpha}(i+1)$ as constraints throughout the calculations. The force constants were 1000 kcal / mol per Å^2 for the pseudobonds and 500 kcal / mol per rad^2 for the pseudoimproper rotation axes and the distance ranges restraining the pseudobond angles. These values were the same for all monomers and were kept constant throughout the simulations.

Secondary structure geometry was maintained as follows. For helices, the $C^{\alpha}(i)$–$C^{\alpha}(i+3)$ distances were restrained with an upper limit of 6.9 Å. In order to obtain right-handed helices, the virtual C^{α} bond torsional angles were also loosely constrained to $48(\pm 20)°$. β-Sheet geometry was imposed by restraining the direct and adjacent cross-strand C^{α} distances with target values of 5.1 Å and 6.1 Å, respectively. The distances between equivalent carbon atoms two strands away from each other on a β-sheet were also constrained with target values of 10 Å. These constraints were enforced by a soft, asymptotic square-well function (Nilges *et al*., 1988) with a constant force constant of 50 kcal / mol per Å^2.

Simulated annealing protocol

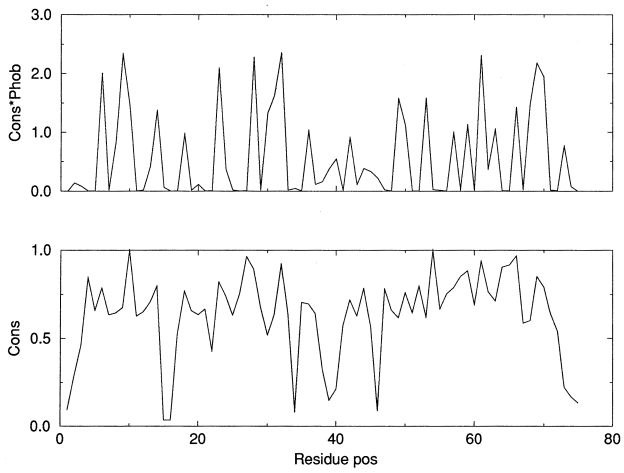
For each subset of long-range simulated NOE distance restraints, a family of 25 conformations was calculated. The NOE restraints were applied to the pseudo-C^{β} atoms as in DRAGON-3, and were enforced by the same square-well potential as described above for the α-helices and β-sheets. Non-bonded interactions were represented by a quartic hard-sphere potential. A simulated annealing protocol, based on a method by Nilges *et al*. (1988), was used, starting from extended conformations. During dynamics, temperature coupling (Berendsen *et al*., 1984) was used and bond stability was imposed using the SHAKE method (Rykaert *et al*., 1977; van Gunsteren & Berendsen, 1977). Initially the force constant for the non-bonded interaction potential was 0.002 kcal / mol per Å^2 with the atomic radii set to 1.2 times the values used in DRAGON. The slope of the asymptote for the distance restraint function was 0.1. Using these values, 50 cycles of restrained Powell minimization (Powell, 1977) were performed. The atomic velocities were scaled to a Maxwellian population at 4000 K and 80 ps of molecular dynamics was calculated, with a target temperature of 4000 K. The slope of the asymptote for the distance constraint potential was increased to 1.0 and a further 40 ps of high-temperature dynamics was calculated. The system was then cooled to a simulated temperature of 100 K in 25 K steps, and for each target temperature ~0.39 ps of restrained dynamics was calculated. At each cycle, the sizes of the atomic radii relative to the values used in DRAGON were scaled down from 1.2 and the force constant for the non-bonded potential was scaled up from 0.002 exponentially such that the final values were 1.0 and 4.0, respectively. Finally, 1000 cycles of Powell minimization were performed.

Implementation

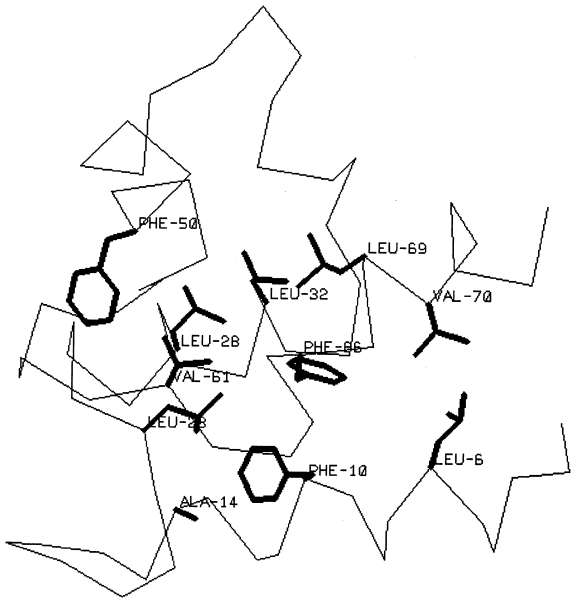
DRAGON-3 was implemented as an ANSI C program by A.A. The models presented here were produced by running DRAGON-3 in batch mode on various SGI computers. Execution time for a 3ICB model (75 residues) was about two minutes on an SGI Challenge S equipped with a R4400 / 150 MHz processor. X-PLOR was run on a Sun Sparcstation 10 model 41, and a 3ICB model with 86 long-range restraints was generated in about 35 minutes. The corresponding execution time for DRAGON on the same machine was about four minutes. The plots were generated using the ACE / gr program by Paul J. Turner, the protein cartoons were drawn by the RasMol package (author Roger Sayle).

Data

For the purposes of this study, a set of well-resolved small monomeric proteins was needed that represented the major structural classes. The target structures (Brookhaven codes are given in parentheses) chosen for analysis were bovine



A



B

Figure 4. Distribution of conserved hydrophobic residues in the bovine vitamin D-dependent Ca^{2+}-binding protein (3ICB). A, Conservation (lower graph) and conserved hydrophobicity (upper graph) profiles obtained from the multiple alignment. B, Position of hydrophobic side-chains with a conservation score higher than 0.75.

vitamin D-dependent Ca^{2+}-binding protein (3ICB), tendamistat (3AIT) and thioredoxin (2TRX).

Bovine vitamin D-dependent Ca^{2+}-binding protein (3ICB)

This protein represented the all-α structural class. It is 75 residues long and contains four α-helices and a small irregular helix. The X-ray structure was determined at 2.3 Å resolution (Szebenyi & Moffat, 1986). The following sequences were aligned with the 3ICB sequence (PIR access codes in parentheses): troponin C from chick (P02588) and from the Japanese horseshoe crab (P15159), *Caenorhabditis elegans* calmodulin-like protein (P04630), β-parvalbumin from *Xenopus laevis* (P05940) and from *Latimeria chalmersii* (P02623), human S100-L protein (P29034) and rat calpactin I light chain (P05943).

Tendamistat (3AIT)

This 74-residue long protein, an α-amylase inhibitor from *Streptomyces tendae*, is a sandwich of two, three-strand antiparallel β-sheets, and it contains two disulphide bonds. The structure of the PDB entry was determined by constrained energy minimisation from solution NMR measurements using the AMBER 3.0 protocol (Billeter *et al*., 1990). The sequence database search provided a set of relatively close homologues to the *S. tendae* sequence and another set of very distant, spurious matches. The multiple alignment was therefore constructed from various α-amylase inhibitor sequences from other *Streptomyces* species (PIR access codes P01093, P07512, P09921, P20078 and P20596).

Thioredoxin (2TRX)

Thioredoxin is an electron transport protein from *Escherichia coli*. The PDB entry contains two chemically identical chains in the unit cell designated A and B. Since parts of chain B are disordered, chain A was chosen as the target structure. Thioredoxin is 108 residues long and contains a five-strand mixed β-sheet in the core, shielded by five helices in a fold similar to that of flavodoxin (4FXN). The structure was determined at a resolution of 1.68 Å (Katti *et al*., 1990). The multiple alignment was constructed from the following protein sequences (PIR access codes in parentheses): thioredoxin from *Anabaena* (P20857), *Saccharomyces cerevisiae* (P22217), *Aspergillus nidulans* (P29429) and *Pisum sativum* (P29450), protein S–S isomerase (EC 5.3.4.1) precursor from mouse (P09103) and from man (P07237), bloodstream-specific protein 2 precursor from *Trypanosoma brucei* (P12865) and a sequence described as a ''probable ERP72 protein homolog'' from *C. elegans*.

Results

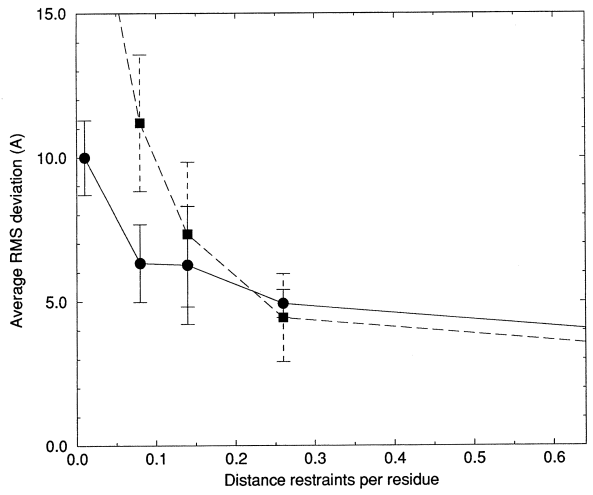
Bovine vitamin D-dependent Ca^{2+}-binding protein

Conserved hydrophobicity

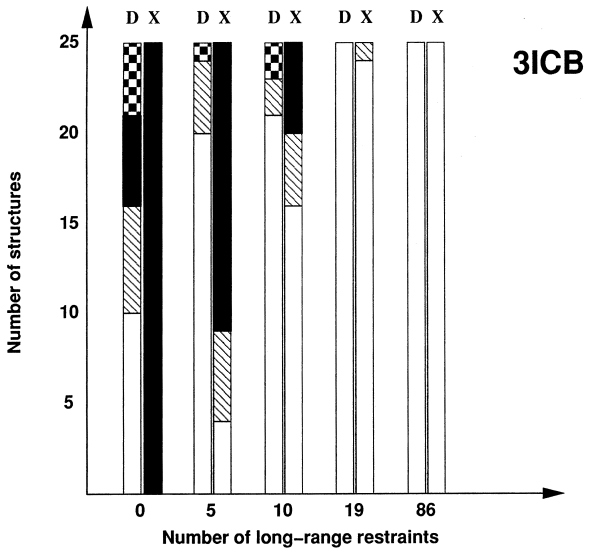
The conserved hydrophobicity score deduced from the multiple alignment highlighted the conserved components of the hydrophobic core (Figure 4). With the exception of Leu23 and Val61, all hydrophobic residues with a conservation higher than 0.75 were in helical regions.

Model quality dependence on the number of restraints

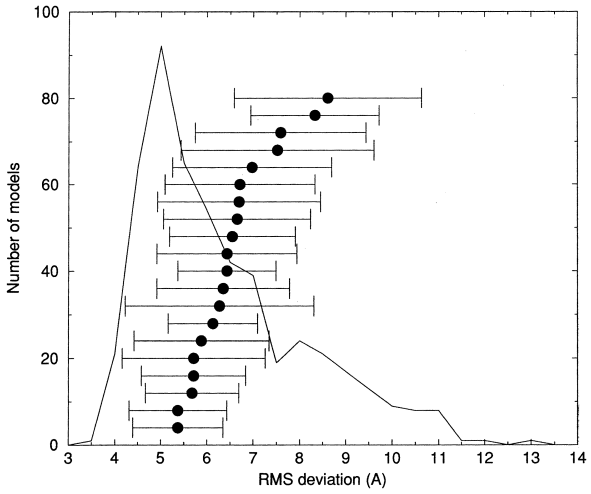
Four sets of simulated NOE restraints were defined for the models, containing 86, 19, 10 and 5



A



B



C

Figure 5. Simulation results for the bovine vitamin D-dependent Ca^{2+}-binding protein (3ICB). A, Average RMS deviation of the models from the PDB structure as a function of the number of long-range distance restraints per residue. Circles and squares represent the models generated by DRAGON-3 and X-PLOR, respectively. B, Results of the visual topology checks. Open, hatched and filled bars symbolise correct, slightly incorrect and seriously incorrect topologies, respectively. Chequered bars symbolise mirror images. For every restraint set, the left and right bars represent DRAGON (D) and X-PLOR (X) models, respectively. C, Effect of the choice of restraints. The average RMS deviations for 20 datasets are depicted as filled circles (the abscissae are the RMS values and the points are shifted vertically for clarity). The error bars correspond to the standard deviations of the RMS values within the datasets. The pooled distribution of the model RMS deviations is plotted in the background.

long-range restraints, respectively. An additional dataset containing no restraints served as an internal control.

With 86 long-range restraints, both DRAGON and X-PLOR correctly identified the native fold, the average RMS deviation being slightly better for the X-PLOR structures than for the DRAGON structures. As the number of restraints decreased, DRAGON performed increasingly better than X-PLOR. With 19 restraints, the average RMS of the models produced by the two programs were about the same, X-PLOR still having a slight advantage. With ten or five restraints, however, the DRAGON structures still had essentially correct topologies (save occasional misplaced helices), while the X-PLOR models became increasingly inaccurate (Figure 5A and B). Without any long-range restraints, X-PLOR failed to produce compact structures at all, which rendered the comparison meaningless (Table 2).

Although the average RMS of the DRAGON structures was as high as 10 Å in the no-restraints case, the program still correctly identified the native topology in 40% of the runs and always produced compact, globular structures (Figure 6).

Table 2. RMS deviation of model structures from target 3ICB

Number of restraints	Restraints per residue	RMS (Å)		S.D. (Å)	
		DRAGON	X-PLOR	DRAGON	X-PLOR
86	1.14	2.88	**2.39**	0.21	0.17
19	0.25	4.93	4.44	**0.48**	1.52
10	0.13	6.26	7.33	**2.04**	2.51
5	0.07	**6.32**	11.2	**1.64**	2.38
0	0.00	**10.0**	21.3	**1.50**	3.20

Averages and standard deviations of the individual RMS values are tabulated as a function of the number of simulated NOE restraints for models produced by DRAGON and X-PLOR. Entries significantly ($P < 0.01$) lower than their counterparts are set in boldface.

Model quality dependence on the choice of restraints

The sensitivity to the choice of the distances was tested on 20 different datasets each containing ten restraints. For each dataset, 25 models were generated by DRAGON. The pooled distribution of the RMS values of these 500 models was unimodal with a maximum at 5.0 Å, the overall average RMS value was 6.55($\pm$1.76)Å. The average RMS deviations of the individual datasets spanned a range from 5.37 Å to 8.61 Å (Figure 5C).

Due to the much higher CPU time requirements of the X-PLOR calculations, five representative datasets were chosen among the 20 datasets described above and 25 models were generated from these by X-PLOR. For the same dataset, DRAGON-3 performed consistently better than X-PLOR, as indicated by comparison of the corresponding RMS averages (Table 3).

Tendamistat

Conserved hydrophobicity

The hydrophobic core of tendamistat is not so well defined as that of 3ICB (Figure 7). The edge strands 52–57 in sheet 1 and 67–73 in sheet 2 possess no hydrophobic residues with a conservation score higher than 0.75, while the remaining strands contain mainly smaller hydrophobic residues. Two conserved hydrophobic residues, Trp18 and Ala28 occupy exposed positions, while the flexible N-terminal tail of the molecule contains two more conserved hydrophobics, Val4 and Ala8.

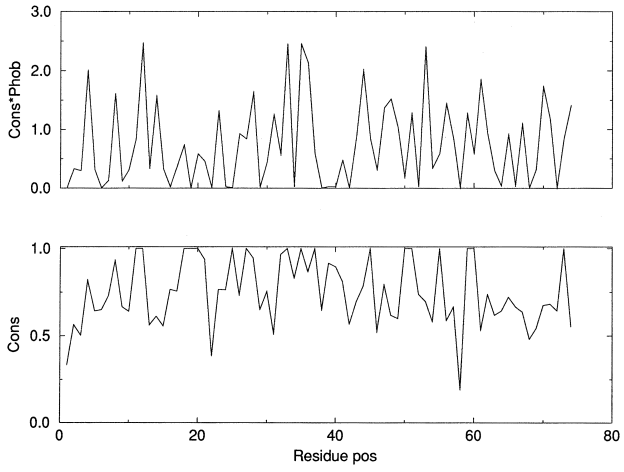
Table 3. Comparison of representative ten-restraint datasets for 3ICB

RMS (Å)		S.D. (Å)	
DRAGON	X-PLOR	DRAGON	X-PLOR
5.37	7.14	**1.06**	3.03
5.88	8.16	**1.46**	2.87
6.43	7.76	**1.06**	2.50
6.71	8.40	1.62	1.72
8.33	12.1	**1.39**	2.54

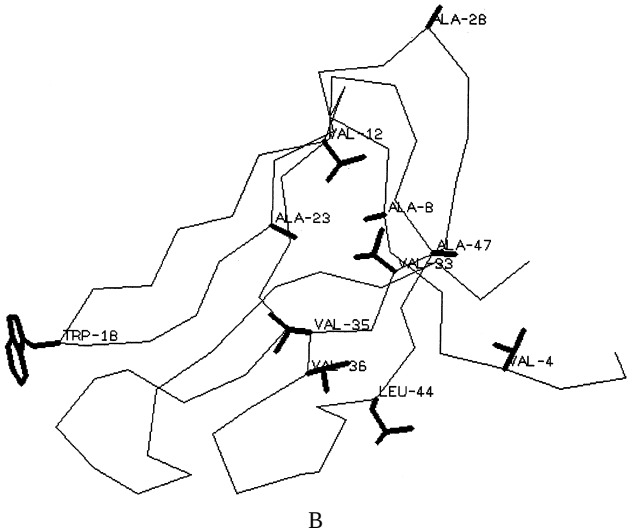
Averages and standard deviations of the individual RMS values are tabulated as a function of the number of simulated NOE restraints for models produced by DRAGON and X-PLOR. Entries significantly ($P < 0.01$) lower than their counterparts are set in boldface.

Model quality dependence on the number of restraints

The tendamistat datasets contained 120, 46, 10, 5 and 0 long-range simulated NOE restraints. For this



A



B

Figure 7. Distribution of conserved hydrophobic residues in tendamistat (3AIT). Symbols are as in Figure 4. Note the exposed hydrophobic residues Trp18 and Ala28.

small β-sandwich protein, DRAGON always performed significantly better than X-PLOR, the latter producing numerous mirror-image topologies and heavily tangled structures (Figure 8A and B). The DRAGON models were essentially correct down to the five long-range restraint case, where packing became loose but the native topology was still preserved (Figure 9). Slightly incorrect topologies (5 out of 25) were observed only without any long-range restraints (Table 4).

From a modelling point of view, the structure of tendamistat has a few inconvenient features. The N-terminal end of the molecule is rather mobile, which is reflected in the *B*-factor entries in the PDB file (obtained from the RMS deviations between a set of NMR structures). The presence of conserved apolar residues occupying exposed positions on the surface, which is not uncommon among small proteins involved in non-covalent protein-protein interactions such as enzyme inhibition, can confuse DRAGON's hydrophobic core-building heuristics, which forces exposed hydrophobic residues towards the centre of the molecule if no extra information is available. With just five long-range distance restraints DRAGON therefore was also unable to pack the β-sheets together.

Model quality dependence on the choice of restraints

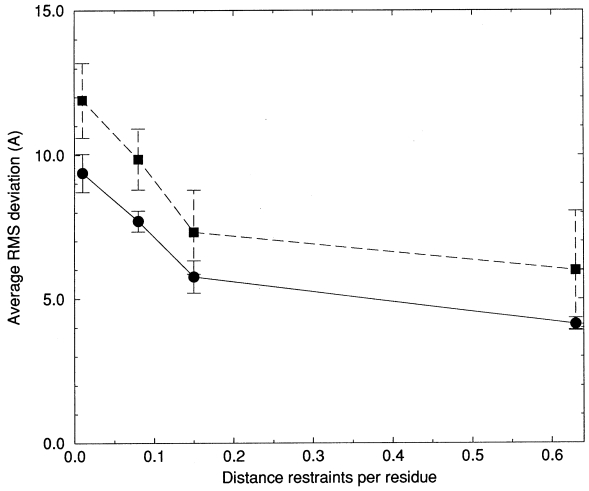
The assessment was performed in the same way as for 3ICB models (see above). The 20 distance sets, each containing ten restraints, were submitted to DRAGON and 25 models were generated for each set. The pooled distribution of the RMS deviations of these 500 models was bimodal, with maxima at approximately 6.0 and 9.0 Å. The overall average RMS value was 7.49($\pm$1.52)Å. The average RMS deviations of the individual datasets spanned a range from 5.24 Å to 9.31 Å (Figure 8C).

In the majority of cases DRAGON again performed significantly better than X-PLOR when the two methods were compared using five representative datasets (Table 5).

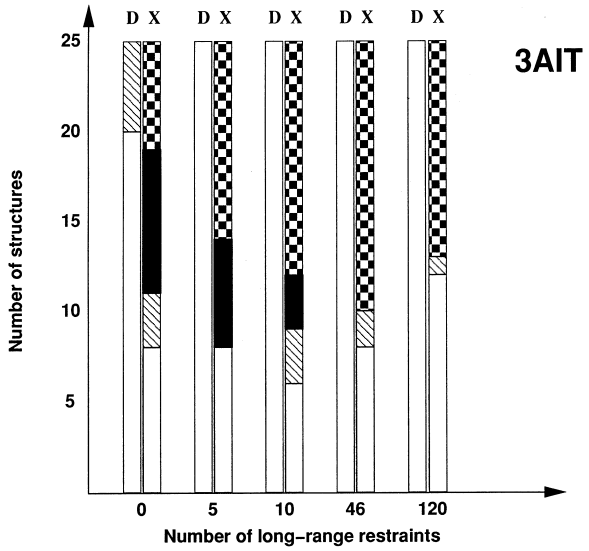
Thioredoxin

Conserved hydrophobicity

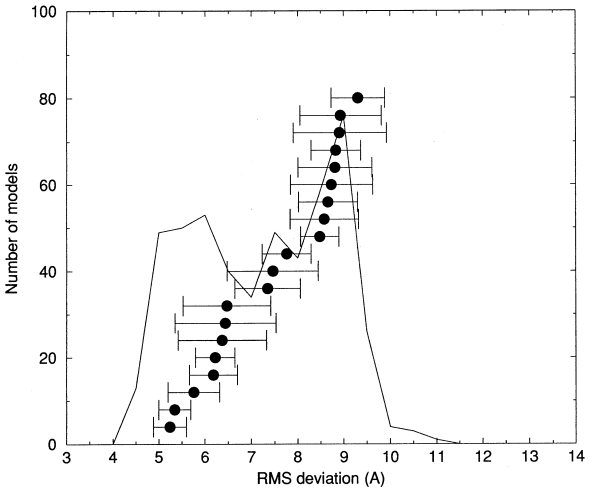
Thioredoxin has the most complex structure among the proteins modelled in this study. It contains a five-strand mixed β-sheet in the core, shielded by helices and represented a challenge to both programs. The key conserved hydrophobic residues distribute evenly in the core (Figure 10), but the overall level of conservation was somewhat lower than for the two other target proteins, probably due to the larger number of sequences (12) in the multiple alignment. Therefore only 11 hydrophobic residues had conservation scores higher than 0.75. One of these, Trp31, occupies an exposed position at the beginning of the second helix (residues 32 to 49).



A



B



C

Figure 8. Simulation results for tendamistat (3AIT). Symbols are as in Figure 5.

Table 4. RMS deviation of model structures from target 3AIT

Number of restraints	Restraints per residue	RMS (Å)		S.D. (Å)	
		DRAGON	X-PLOR	DRAGON	X-PLOR
120	1.62	**3.72**	5.29	**0.16**	2.19
46	0.62	**4.12**	5.99	**0.22**	2.06
10	0.14	**5.76**	7.32	**0.56**	1.46
5	0.07	**7.70**	9.85	**0.36**	1.06
0	0.00	**9.38**	11.9	**0.66**	1.30

Data and symbols are as in Table 2.

Model quality dependence on the number of restraints

The NOE distance dataset contained 161, 48, 30, 20 and 0 simulated restraints. Starting from the two largest restraint sets, DRAGON always found the correct topology and the average RMS of the models were good. X-PLOR, on the other hand, produced several mirror-image topologies even for the largest dataset and the structures were occasionally tangled as well, giving rise to significantly higher average RMS values. When the number of restraints was reduced to 30, DRAGON produced two distinct subsets of solutions: 12 models out of 25 were just as good as those obtained from the large dataset runs, possessing correct topologies with RMS values around 4.5 Å, while the other half were less satisfactory, with incorrect topologies (Figure 11A and B). At this stage, the models produced by X-PLOR were comparable with those of DRAGON, with the exception that the number of mirror images and other seriously incorrect topologies was higher (Figure 12). On further reduction of the number of restraints, both programs produced models with high RMS values and about one-third of the models had correct topologies (Table 6).

Model quality dependence on the choice of restraints

The most interesting behaviour was expected at the ''crossover point'' where two distinct families of models were produced from a 30-restraint set. The 20 distance sets, each containing 30 restraints, were therefore submitted to DRAGON and for each set 25 models were generated. The pooled distribution of the RMS deviations of the 500 models was multimodal, containing four distinct peaks at 4.5 Å, 6.5 Å, 9 Å and 10.5 Å, respectively. The overall average RMS value was 7.97($\pm$2.35)Å. The average RMS deviations of the individual datasets spanned a range from 5.3 Å to 10.1 Å (Figure 11C). The standard deviations of RMS values were high for most datasets, indicating that a considerable range of the overall distribution was sampled, producing subsets of high and low-quality models from the same restraints.

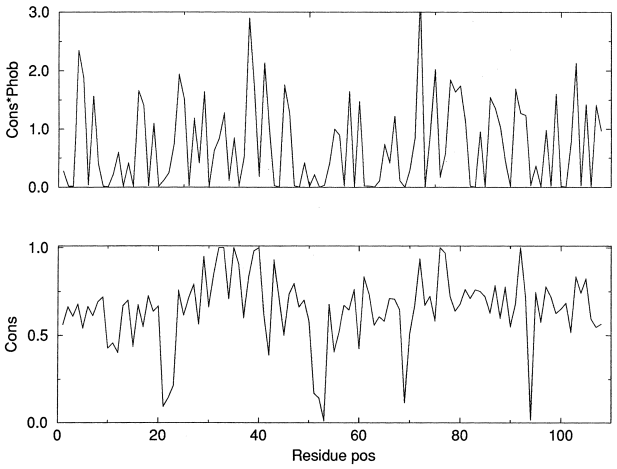
DRAGON and X-PLOR performed similarly when compared using representative 30-restraint datasets. In four out of five cases, there was no significant difference between the average RMS values (Table 7).

With only a small number of restraints, DRAGON

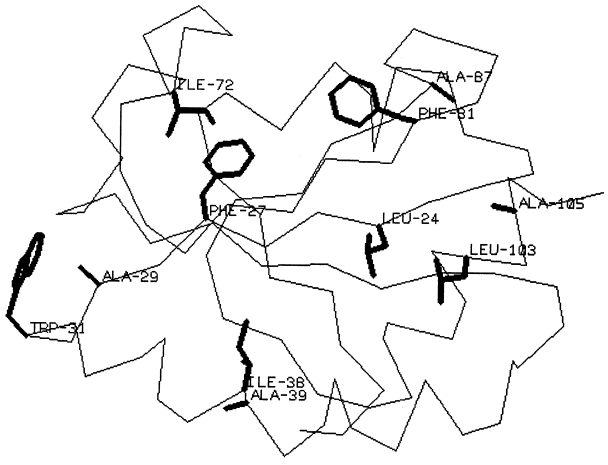
Table 5. Comparison of representative ten-restraint datasets for 3AIT

RMS (Å)		S.D. (Å)	
DRAGON	X-PLOR	DRAGON	X-PLOR
5.24	6.45	**0.36**	1.68
6.22	7.61	**0.43**	1.44
7.35	8.19	**0.70**	1.57
8.58	10.6	**0.74**	1.20
8.83	**7.79**	**0.54**	0.87

Data and symbols are as in Table 3.



A



B

Figure 10. Distribution of conserved hydrophobic residues in thioredoxin (2TRX). Symbols are as in Figure 4. Note the exposed hydrophobic residues Trp18 and Ala28.

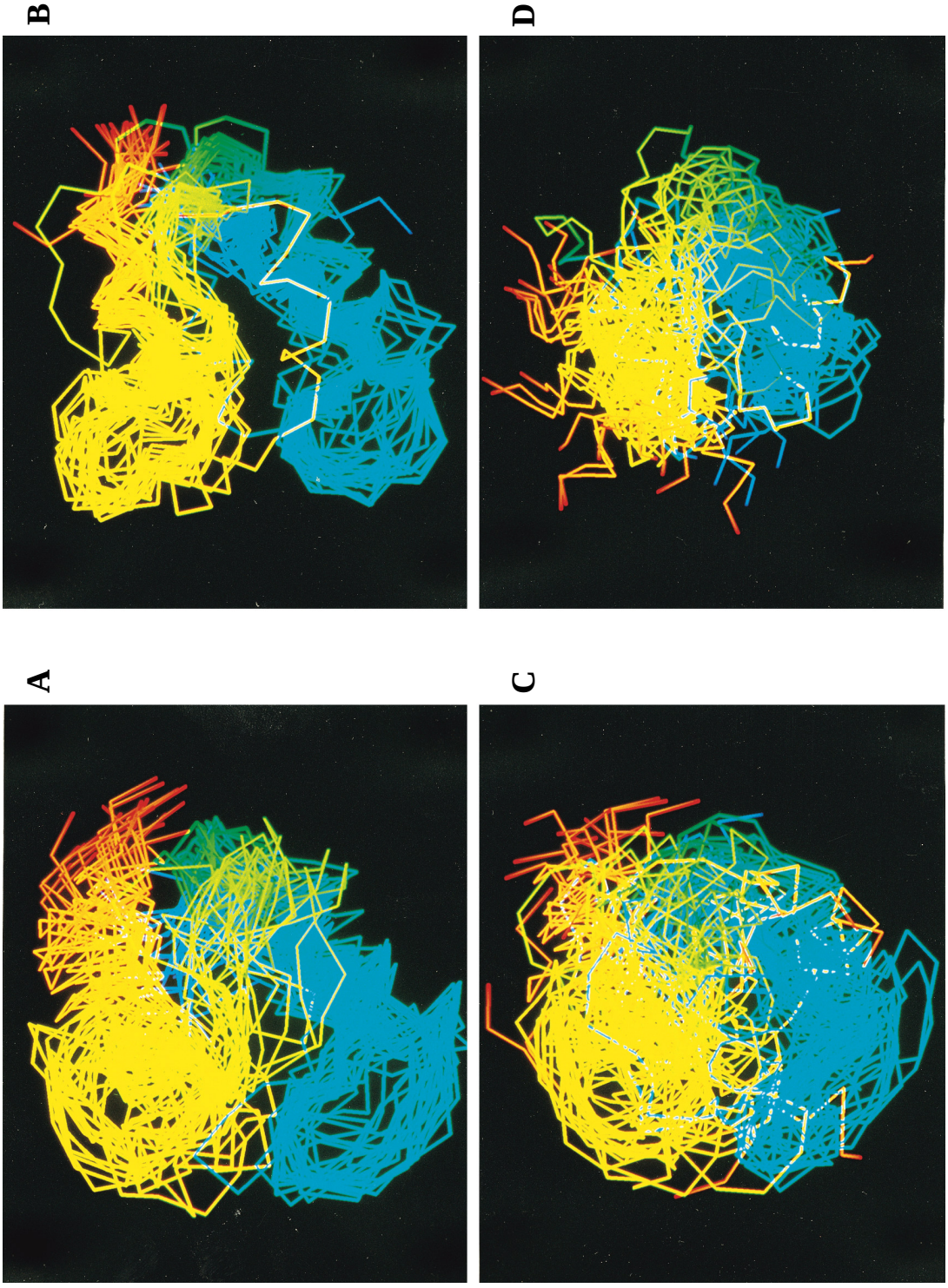


Figure 6. Model C^{α} backbones aligned to the native CABP structure 3ICP. Residue positions are colour-coded from blue (N terminus) to red (C terminus) in a smooth range. A and B, Models generated from 19 long-range restraints by DRAGON and X-PLOR, respectively. X-PLOR produced better structures in this case, except one outlier, which had a slightly incorrect topology. The DRAGON models were less accurate but all 25 models had correct topologies. C and D, Models generated from five long-range restraints by DRAGON and X-PLOR. Of the 25 DRAGON models, 20 still had correct topologies, whereas 21 X-PLOR models were incorrect. Note that the X-PLOR model images were scaled down to fit in the screen.

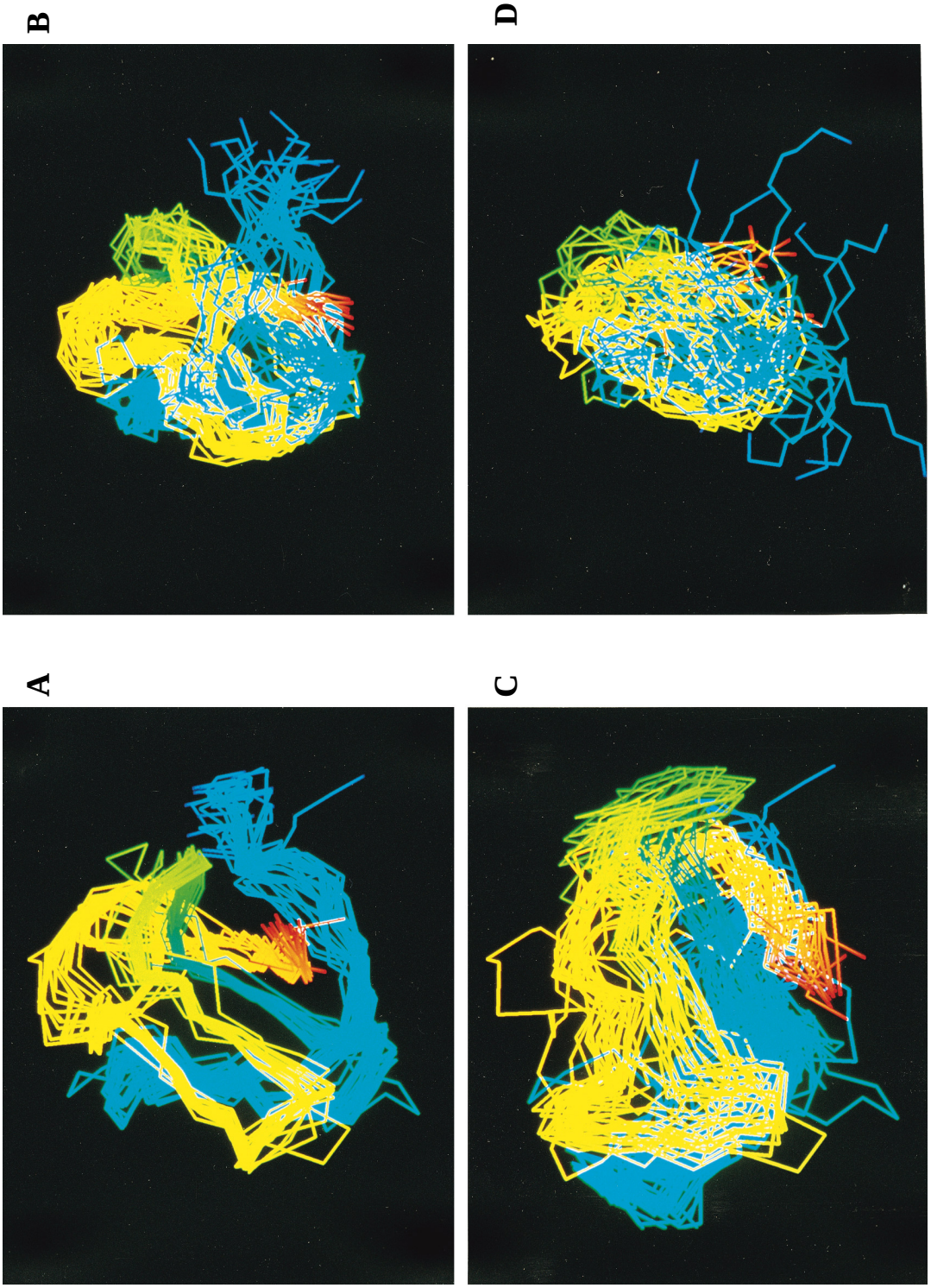


Figure 9. Model C^{α} backbones aligned with the native tendamistat structure 3AIT. Colour codes as in Figure 5. A and B, Models generated from 46 long-range restraints by DRAGON and X-PLOR, respectively. DRAGON produced markedly better results, whereas the X-PLOR models were more variable. C and D, Models generated from five long-range restraints by DRAGON and X-PLOR. All DRAGON models still had correct topologies, while the majority of the X-PLOR models became incorrect.

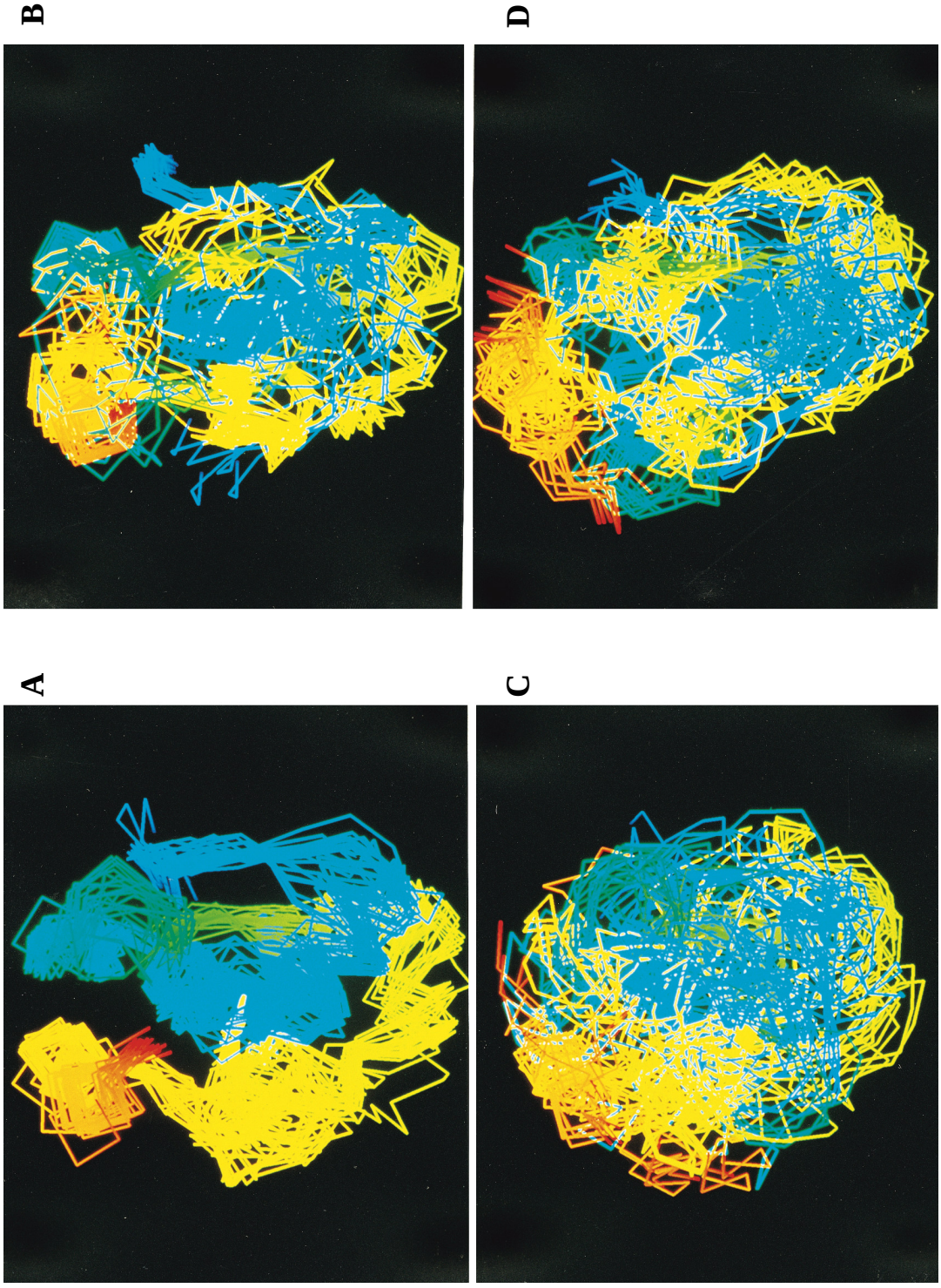
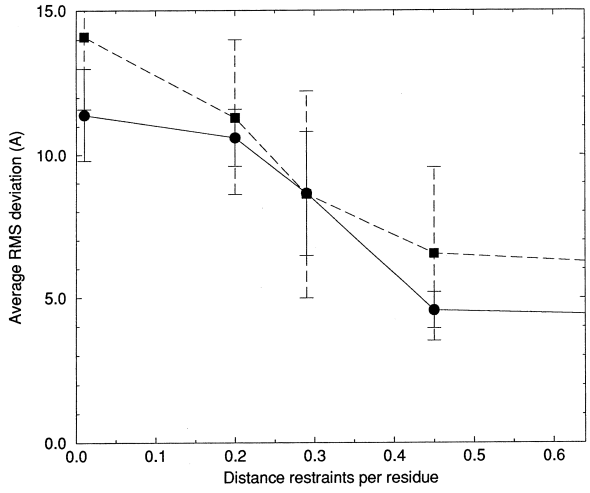
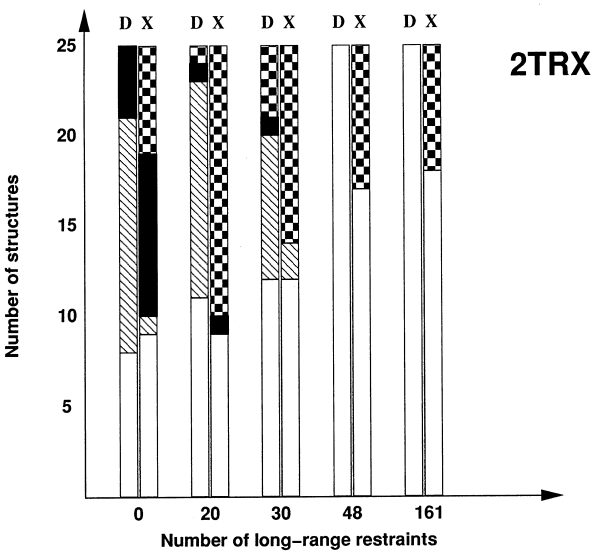


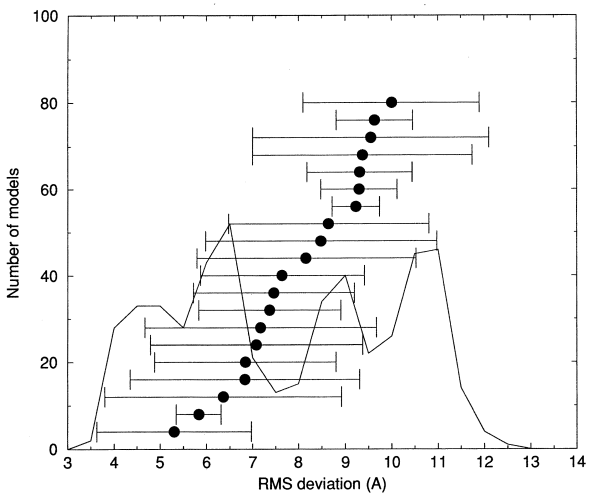
Figure 12. Model C^{α} backbones aligned with the native thioredoxin structure 2TRX (A-chain). Colour codes as in Figure 5. A and B, Models generated from 48 long-range restraints by DRAGON and X-PLOR, respectively. The DRAGON models were correct in all cases, while the X-PLOR produced a number of incorrect results. C and D, Models generated from 20 long-range restraints by DRAGON and X-PLOR. Both methods located the correct topologies in about ten cases out of 25 but the overall model quality was poor.



A



B



C

Figure 11. Simulation results for thioredoxin (2TRX). Symbols are as in Figure 5.

performed significantly better, because X-PLOR failed to pack the helices against the central sheet and therefore the RMS deviations were much larger than those obtained for the compact DRAGON structures. It was clear, however, that neither program could produce useful results if less than 20 long-range restraints were available. Comparison with the results obtained from the simulations of 3ICB and 3AIT suggests that the minimal number of long-range restraints necessary for correct fold identification is a non-linear function of chain length.

Discussion

Accuracy

In most cases, DRAGON produced more accurate results than X-PLOR, especially when the number of distance restraints was small. It should be kept in mind that in a real NMR structure determination problem the maximal accuracy is limited by the quality of the experimental data. This aspect was modelled by establishing an artificial ± 2 Å range for all simulated restraints, which of course may be variable when real experimental distance estimates are used. Also, an important simplification in the present study was that side-chains were replaced by a single pseudo-C^{β} atom and the restraints were imposed on these whereas in reality the NOE-derived restraints apply to the H atoms. It is possible, however, to convert individual inter-H distance restraints into side-chain centroid distance restraints. These restraints can be used by DRAGON to generate a set of starting structures, which in turn can be refined by another program such as X-PLOR using the original H–H distance restraints applied to a full-atom representation of the molecule.

In the absence of detailed structural information, usually a number of different folds are generated. The gradual projection algorithm of DRAGON was often capable of finding the correct topology even under unfavourable circumstances due to the ease with which large structural rearrangements can be made in high-dimensional spaces. This correlates with the observation that local minima can be avoided by performing energy minimisation in four or more dimensions (Crippen, 1982; Purisima & Scheraga, 1986).

Precision and sampling

The precision of the DRAGON models was almost always significantly better than those produced by X-PLOR, judged by the RMS deviation values of the individual models measured with respect to the corresponding average structures (Table 8).

Precision, however, cannot be evaluated separately from the sampling properties of the algorithm in question. While it is practically impossible to perform an exhaustive search in protein conformational spaces, good sampling at least should ensure that a reliable, unbiased estimate is obtained.

The sampling properties of distance geometry-

Table 6. RMS deviation of model structures from target 2TRX

Number of restraints	Restraints per residue	RMS (Å)		S.D. (Å)	
		DRAGON	X-PLOR	DRAGON	X-PLOR
161	1.49	**3.86**	4.96	**0.13**	2.55
48	0.44	**4.57**	6.54	**0.62**	3.02
30	0.28	8.64	8.61	**2.17**	3.62
20	0.19	10.6	11.3	**1.0**	2.7
0	0.00	**11.4**	14.1	**1.6**	2.5

Data and symbols are as in Table 2.

based techniques were assessed in detail by Havel (1991). The ''randomized metrization'' method used in his programs, which coupled the selection of uniformly distributed trial distances with triangle inequality smoothing, provided very good results. An efficient version of this algorithm called ''partial metrization'' was implemented by Kuszewski *et al*. (1992). A similar approach was used in the DRAGON simulations, which started with a random initial distance matrix but the triangle inequality smoothing was performed indirectly, through adjusting the elements of the metric matrix in a self-consistent loop (Aszódi & Taylor, 1994a). This smoothed distance matrix, subsequently projected into a high-dimensional Euclidean space, corresponded to a much greater structural variability than can be obtained from starting a random coil conformation in 3D. Our previous studies with the predecessors of the present algorithm (Aszódi & Taylor, 1994b) showed that in the absence of specific distance restraints a wide variety of structures was generated. Taking into account the low standard deviation of model RMS values obtained in the present study, it can be concluded that the algorithm delivered high precision and good sampling at the same time.

Robustness

In the present context, the robustness of an algorithm means the ability to produce satisfactory results even if the input data are scarce or unreliable. For an NMR structure determination problem, the lack of interresidue distance data could be compensated for by applying various heuristics based on our general knowledge about proteins. DRAGON employs a wide range of such heuristics to achieve its goal; the most important of these are the modelling of the hydrophobic effect, the elimination of tangled conformations and the chirality checks.

Table 7. Comparison of representative ten-restraint datasets for 2TRX

RMS (Å)		S.D. (Å)	
DRAGON	X-PLOR	DRAGON	X-PLOR
5.30	6.48	**1.67**	3.07
6.84	6.03	1.96	2.28
7.46	8.56	**1.74**	3.67
8.48	10.3	2.50	2.73
9.37	10.2	2.37	2.63

Data and symbols are as in Table 3.

Compactness and the hydrophobic effect

The hydrophobic effect was modelled by reducing the distances between hydrophobic residues and by constantly monitoring the accessibility of every side-chain in the molecule. The construction of the hydrophobic core was directed by the identification of the conserved hydrophobic residues. When no external distance restraints were available, this approach still succeeded in producing compact structures, often with the correct topology, whereas X-PLOR usually generated loose tangles ''floating'' randomly. It must be noted, however, that DRAGON was more successful with α-helices, which often provide clear structural orientation in the form of their hydrophobic moments.

The projection into spaces of gradually decreasing dimensionality also facilitated the formation of compact structures, since this operation effectively compresses the point set being projected. Various checks were applied to ensure that the correct point density was maintained through the series of projections (Aszódi & Taylor, 1994b).

Tangles

Another, often overlooked, problem in simulations is tangling. Tangles are very rare if not totally absent

Table 8. RMS deviation of model structures from their respective average structure

Number of restraints	Restraints per residue	RMS (Å)	
		DRAGON	X-PLOR
3ICB			
86	1.14	0.97	0.90
19	0.25	2.48	2.59
10	0.13	**4.03**	5.43
5	0.07	**4.62**	7.00
0	0.00	**7.69**	10.7
3AIT			
120	1.62	**1.69**	4.00
46	0.62	**1.73**	4.36
10	0.14	**2.87**	5.78
5	0.07	**2.61**	5.35
0	0.00	**3.38**	5.72
2TRX			
161	1.49	**1.54**	4.79
48	0.44	**1.92**	5.71
30	0.28	8.06	**6.62**
20	0.19	**4.20**	7.60
0	0.00	**6.57**	8.87

Bold figures indicate entries significantly lower than their counterparts.

in native polypeptide folds, yet they frequently occur in models. Distance geometry techniques that rely on metric matrix projection are particularly liable to generate this kind of artifact, especially when larger molecules are modelled, which can fold into more complex patterns. Tangles may trap the chain in an incorrect conformation and thereby hinder the convergence of the algorithm, rendering it less robust. Tangles are difficult to detect by computer because of the lack of a reliable mathematical definition. (Knots, which can exactly be defined in topological terms, occur in closed curves only, while protein backbones, in general, are open curves.) The heuristic employed in DRAGON-3 relied on the backbone hydrogen-bond topology of secondary structures and therefore cannot be considered universal: in its present implementation, it cannot detect tangles in a chain with no secondary structure at all. (Although the present algorithm can be extended, further complexity would incur a severe performance penalty.) Most tangles were nevertheless successfully eliminated in the simulations, indicating that an advantageous trade-off was achieved between generality and practicality.

Chirality

Chirality presents a persistent problem to all distance geometry-based techniques, since the handedness information is not contained in the distance matrix and therefore even a perfect set of distance restraints corresponds to the correct topology, and its mirror image. Fortunately, incorrect mirror images can be filtered out at the tertiary structural level since the correct chirality of the secondary structural elements are known. If the handedness of helices and sheets are carefully monitored as is done in the DRAGON algorithm, then the mirror image topology will be an energetically slightly unfavourable diastereomer, rather than an indistinguishable enantiomer, of the correct topology.

Spatial distribution of distance restraints

The distribution of the interresidue distance restraints within the molecule determines the quality of the simulation to a considerable degree, as indicated by our results obtained from randomly chosen distance sets containing the same number of restraints. Although theoretically it might be possible to find the best arrangement of a given number of restraints for a particular structure, in practice the freedom of choice usually does not exist. Our simulations suggest that restraints between secondary structure elements alone are, in general, not sufficient to determine the overall fold correctly. Restraints between coils are also important, as indicated by the tendamistat simulations where the long, flexible chain segments showed a definite tendency to intercalate between the sheets, thus giving rise to incorrect models.

Comparison with other methods

Comparing the performance of DRAGON-3 with other methods is difficult, because of the differences in the published simulation protocols. For example, Smith-Brown *et al*. (1993) impose simulated C^{α}–C^{α} restraints with a constant 2 Å difference between the lower and upper bounds onto a polyglycine back-bone. On the other hand, Hoch & Stern (1992) use a lollipop chain similar to ours, but the distance restraints are applied to the C^{α} and pseudo-C^{β} atoms as well, making the interpretation of the results more difficult (see above). Ideally, one should compare all available methods using the same protocol, which is, unfortunately, impractical. We therefore decided to compare DRAGON-3 with X-PLOR, a widely used and very flexible tool. The simplified chain representation of DRAGON-3 was transferred easily to the X-PLOR protocol, enabling us to make meaningful comparisons.

The task given to both algorithms was by no means trivial. The distance restraint sets were designed to be inaccurate; the lower and upper distance bounds were separated by 4 Å, spanning a range twice as much as the study described by Smith-Brown *et al*. (1993). Also, the simulated NOE restraints were applied to the pseudo-C^{β} atoms, while accuracy was tested by C^{α} coordinate RMS deviations; thus, the restraints had only an indirect effect on the backbone conformations.

The results indicated that DRAGON had a definite advantage in situations where robustness became important, due to the extra background information derived from the conserved hydrophobicity patterns, and the detangling procedure. DRAGON also has the appealing property that it always converges to a compact three-dimensional conformation.

Conclusion

We have presented here a distance geometry-based approach that was designed to incorporate a general background knowledge about proteins in order to produce acceptable results even when specific structural information is scarce. The aim of our work has been to show that topologically correct solutions to sparse distance data are more easily obtained when general properties of proteins are incorporated, giving rise to compact, tangle-free conformations with well-defined hydrophobic cores. The method of attaining these attributes in the final model is secondary and, although we preferred a distance geometry-based approach, similar results might well have been attained through optimisation in Euclidean space. However, we did not simply encode these general heuristics in one of the several current refinement methods, as their orientation to detailed atomic representation is not ideally suited to the broad (topological) emphasis we desired. We regard our method as suitable for generating an ensemble of good starting conformations that might then be further refined by a program such as X-PLOR. A molecular modelling tool based on this

tandem approach should help the experimentalist obtaining more reliable structures from NMR data.

Acknowledgements

The authors thank Dr J. Feeney and Dr A. Šali for their helpful comments and suggestions.

References

Aszódi, A. & Taylor, W. R. (1994a). Folding polypeptide alpha-carbon backbones by distance geometry methods. *Biopolymers*, **34**, 489–505.

Aszódi, A. & Taylor, W. R. (1994b). Secondary structure formation in model polypeptide chains. *Protein Eng.* **7**, 633–644.

Aszódi, A. & Taylor, W. R. (1995). Estimating polypeptide α-carbon distances from multiple sequence alignments. *J. Math. Chem.* **17**, 167–184.

Bairoch, A. & Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* **19**, 2247–2249.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A. & Haak, J. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Billeter, M., Schaumann, T., Braun, W. & Wüthrich, K. (1990). Restrained energy refinement with two different algorithms and force fields of the structure of the alpha-amylase inhibitor tendamistat determined by NMR in solution. *Biopolymers*, **29**, 695–706.

Brünger, A. T. (1992). *X-PLOR Manual*. Yale University, New Haven, CT.

Connolly, M. L., Kuntz, I. D. & Crippen, G. M. (1980). Linked and threaded loops in proteins. *Biopolymers*, **19**, 1167–1182.

Crippen, G. M. (1982). Conformational analysis by energy embedding. *J. Comp. Chem.* **3**, 471–476.

Crippen, G. M. & Havel, T. F. (1988). *Distance Geometry and Molecular Conformation. Chemometrics Research Studies Press, Wiley, New York.*

Dandekar, T. & Argos, P. (1994). Folding the main-chain of small proteins with the genetic algorithm. *J. Mol. Biol.* **236**, 844–861.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 345–352, Nat. Biomed. Res. Foundation, Washington DC.

Gutin, A. M. & Shakhnovich, E. I. (1994). Statistical mechanics of polymers with distance constraints. *J. Chem. Phys.* **100**, 5290–5293.

Havel, T. F. (1991). An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog. Biophys. Mol. Biol.* **56**, 43–78.

Hoch, J. C. & Stern, A. S. (1992). A method for determining overall protein fold from NMR distance restraints. *J. Biomol. NMR*, **2**, 535–543.

James, T. L. (1994). Computational strategies pertinent to NMR solution structure determination. *Curr. Opin. Struct. Biol.* **4**, 275–284.

Katti, S. K., LeMaster, D. M. & Eklund, H. (1990). Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *J. Mol. Biol.* **212**, 167–184.

Kuntz, I. D., Thomason, J. F. & Oshiro, C. M. (1989). Distance geometry. *Methods Enzymol.* **177**, 159–204.

Kuszewski, J., Nilges, M. & Brünger, A. T. (1992). Sampling and efficiency of metric matrix distance geometry: a novel partial metrization algorithm. *J. Biomol. NMR*, **2**, 33–56.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.

Levitt, M. (1978). Conformational preferences of amino acids in globular proteins. *Biochemistry*, **17**, 4277–4285.

McLachlan, A. D. (1979). Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128**, 49–79.

Nilges, M., Gronenborn, A. M., Brünger, A. T. & Clore, G. M. (1988). Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase and barley serine protease inhibitor 2. *Protein Eng.* **2**, 27–38.

Pauling, L. & Corey, R. B. (1951a). The pleated sheet, a new configuration of polypeptide chains. *Proc. Natl Acad. Sci. USA.* **37**, 251–256.

Pauling, L. & Corey, R. B. (1951b). The structure of synthetic polypeptides. *Proc. Natl Acad. Sci. USA*, **37**, 241–250.

Powell, M. J. D. (1977). Restart procedures for the Conjugate Gradient Method. *Math. Progr.* **12**, 241–254.

Purisima, E. O. & Scheraga, H. A. (1986). An approach to the multiple-minima problem by relaxing dimensionality. *Proc. Natl Acad. Sci. USA*, **83**, 2782–2786.

Rózsa, P. (1991). *Lineáris algebra és alkalmazásai*. Tankönyvkiadó, Budapest (in Hungarian).

Rykaert, J. P., Ciccotti, G. & Berendsen, H. J. C. (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comp. Phys.* **23**, 327–341.

Smith-Brown, M. J., Kominos, D. & Levy, R. M. (1993). Global folding of proteins using a limited number of distance constraints. *Protein Eng.* **6**, 605–614.

Szebenyi, D. M. E. & Moffat, K. (1986). The refined structure of vitamin D-dependent calcium-binding protein from bovine intestine. Molecular details, ion binding and implications for the structure of other calcium-binding proteins. *J. Biol. Chem.* **162**, 8761–8777.

Taylor, W. R. (1988). A flexible method to align large numbers of biological sequences. *J. Mol. Evol.* **28**, 161–169.

Taylor, W. R. (1991). Towards protein tertiary fold prediction using distance and motif constraints. *Protein Eng.* **4**, 853–870.

Taylor, W. R. (1993). Protein fold refinement: building models from idealised folds using motif constraints and multiple sequence data. *Protein Eng.* **6**, 593–604.

van Gunsteren, W. F. & Berendsen, H. J. C. (1977). Algorithms for macromolecular dynamics and constrained dynamics. *Mol. Phys.* **74**, 1311–1327.

Wako, H. & Scheraga, H. A. (1982). Distance-constraint approach to protein folding. I. Statistical analysis of protein conformations in terms of distances between residues. *J. Protein Chem.* **1**, 5–45.

Edited by F. Cohen

(Received 19 December 1994; accepted in revised form 23 May 1995)