Quarterly Reviews of Biophysics 19, 3/4 (1987), pp. 115–157
Printed in Great Britain

Distance geometry and related methods for protein structure determination from NMR data

WERNER BRAUN

Institut für Molekularbiologie u. Biophysik, Eidgenössische Technische Hochschule, Zürich – Hönggerberg, CH-8093 Zürich, Switzerland

1. INTRODUCTION

The method of choice to reveal the conformation of protein molecules in atomic detail has been X-ray single-crystal analysis. Since the first structural analysis of diffraction patterns, computer calculations have been an important tool in these studies (Blundell & Johnson, 1976). As is described by Sheldrick (1985), it has been taken for granted that a necessary first step in the determination of a protein structure would be writing computer programs to fit structure factors. In contrast the combined use of the structural analysis of NMR data and computer calculations has been quite limited. An early attempt of such structural calculations was the quantitative determination of mononucleotide conformations in solution using lanthanide ion shifts (Barry *et al.* 1971).

The reason for the lack of a close connexion between data and structural analysis is the absence of a direct relation between NMR data and spatial structure as in the case of the X-ray diffraction pattern. The relation between chemical shifts and structure is complex and still not fully understood (Wüthrich, 1986). The ring current shift can be interpreted only in cases when the structure is already known by some other method. Adding lanthanide ions to induce the paramagnetic shifts (Barry *et al.* 1971) might influence the molecular conformation and can only be used in special cases. Vicinal coupling constants (Karplus, 1959, 1963) and nuclear Overhauser effects (Noggle & Schirmer, 1971) have a direct geometric meaning but problems such as the inherent flexibility of the molecules, spin diffusion and the short-range character of both data types made it doubtful that these geometric data allow it to deduce the spatial structure of a protein directly from the experimental data without any *a priori* knowledge of the structure (Jardetzky & Roberts, 1981).

A second reason for the lack of direct methods is the difficult computational problem of calculating tertiary protein structures that are compatible with the given experimental data and the stereochemical constraints. This problem is due to the inaccuracy and the short-range character of the geometric constraints from the vicinal coupling constants and the NOE data.

The short-range character of these two data types is inherently different. In the case of the vicinal coupling constants, the information on the torsion angles is of short range relative to the covalent structure, so it is straightforward to characterize a consistent local conformation in terms of torsional angles. However, the accumulation of local errors along the polypeptide chain prevents us from deducing from this a reliable rough model for the global polypeptide fold.

In contrast, NOE data are information on short spatial distances. In proteins only proton–proton spins separated by *c.* 5 Å or less give rise to a detectable NOE signal. The dense packing of protein structures found in the X-ray crystal structures (Richards, 1974) should give a reasonably large number of short contacts between protons separated far along the polypeptide chain. The calculational problem is then to convert this information from the distance space into the 3-dimensional cartesian space.

Most of the methods originally applied were of the indirect type. In this

approach one first proposes one or several models for the polypeptide structure from model building or energy minimization calculations. Each model is then checked for consistency with the data. In case the deviations are significantly larger than the expected experimental errors, the model is discarded (Leach *et al.* 1977; Jones *et al.* 1978; Bothner-By & Johner, 1978; Krishna *et al.* 1978).

In this review only the direct computational approach of polypeptide and protein structure determination from NMR data will be described and several computational tools will be discussed.

A survey will be given of the theoretical aspect of the metric matrix approach. As the mathematical theorems of this approach have been reviewed in some detail (Crippen, 1981; Havel *et al.* 1983), I will describe those features of the method which have proven particularly useful in practice and will try to formulate open problems that should be solved if one wants to proceed along these lines.

A second method, the variable target function method (Braun & Gō, 1985), has been recently successfully applied to determine the tertiary structure of several polypeptides (Kobayashi *et al.* 1985; Ohkubo *et al.* 1986) and proteins (Braun *et al.* 1986; Kline *et al.* 1986; Wagner *et al.* 1987) from NMR data sets. The basic principles will be reviewed, current applications described and future developments sketched.

Restrained molecular dynamics (Kaptein *et al.* 1985; Brünger *et al.* 1986) is a third avenue converting NMR data sets into 3-dimensional structures. Existing computer programs for MD calculations (van Gunsteren & Berendsen, 1982; Brooks *et al.* 1983) have been modified to calculate protein structures satisfying the NMR distance constraints. Scope and limits of this method will be described and compared to the above-mentioned methods.

A survey of the application of these methods to the calculation of protein structures from NMR data will be given. References to work with oligopeptides will be made if it is relevant to the development of methods for the determination of protein structures.

Computer graphics methods (Zuiderweg *et al.* 1984; Billeter *et al.* 1985) are of great help to get a first impression of which parts of the molecule are already restricted by the data and are useful in the analysis of computed structures. They do not yet represent a computer solution of the problems *per se*. The Artificial Intelligence approach PROTEAN (Jardetzky *et al.* 1986) is not an algorithmic computational tool but rather a system of different computer programs operating on different levels, symbolic inference, heuristic reasoning and numerical calculations. It seems to be an attempt to integrate in a computerized way some of the described algorithmic tools. Both methods therefore fall outside the scope of this review.

Calculation of 3-dimensional structures is, however, only one aspect of the direct computational method. The development of parameters to judge the quality of the calculated structures and questions concerning the significance of the structures obtained are equally important.

2. GEOMETRIC CONSTRAINTS

2.1 *Distance constraints*

Before we can proceed to formulate the mathematical problem which is to be solved in the direct method of protein structure determination from NMR data, we have to characterize the geometric constraints available from the experiments.

The most useful quantities derived from NOE data are the cross-relaxation rates σ_{ij} between two proton spins i and j. These quantities can be obtained in a first approximation directly from the NOE cross peaks observed in 1-dimensional (Wagner & Wüthrich; 1979) or 2-dimensional NOE experiments (Jeener *et al.* 1979; Anil Kumar *et al.* 1980; Macura & Ernst, 1980) if one measures with short mixing times (Anil Kumar *et al.* 1981). Recently a more rigorous approach including multispin effects has been proposed to derive σ_{ij} from the 2-D NOE maps (Keepers & James, 1984; Olejniczak *et al.* 1986). We shall show that very accurate NOE data are not required in the first cycle of the tertiary structure determination of proteins by the direct method, because these data are only used to estimate the upper limit of distances; therefore we are not concerned about the best experimental techniques for measuring σ_{ij} experimentally and the accuracy of the measurement, but we have to discuss the different models of their geometric interpretation.

The cross-relaxation rates σ_{ij} are given by

$$\sigma_{ij} = f(\tau_{ij}) \langle r_{ij}^{-6} \rangle, \tag{2.1.1}$$

where τ_{ij} is the distance between spins i and j, and $f(\tau_{ij})$ is a function of the correlation time τ_{ij} for the reorientation of the vector connecting the two spins, and the bracket $\langle\ \rangle$ denotes averaging over the ensemble of molecular structures interconverting in thermal equilibrium.

In a rigid protein structure the correlation time τ_{ij} between all the different pairs of protons would be identical and equal to the correlation time τ_R for the overall tumbling of the molecule. Also the thermal averaging would be trivial and equation (2.1.1) could be used to calculate unknown distances r_{ij} from a set of known distances r_{kl} by

$$r_{ij} = r_{kl} \left[\frac{\sigma_{kl}}{\sigma_{ij}} \right]^{\frac{1}{6}}. \tag{2.1.2}$$

This approach has been used in the spatial characterization of the haem methionine binding mode of ferrocytochrome *c* (Senn *et al.* 1984) and has been found to be particularly useful in the structural interpretation of NOE data for oligonucleotides (Clore & Gronenborn, 1985).

In a more realistic approach the inherent flexibility of protein structures can be taken into account. As described in Braun *et al.* (1981), the ratio of an effective cross-relaxation rate in a flexible protein compared to a calibration cross-relaxation rate between spins with a fixed, known distance can be estimated by a function

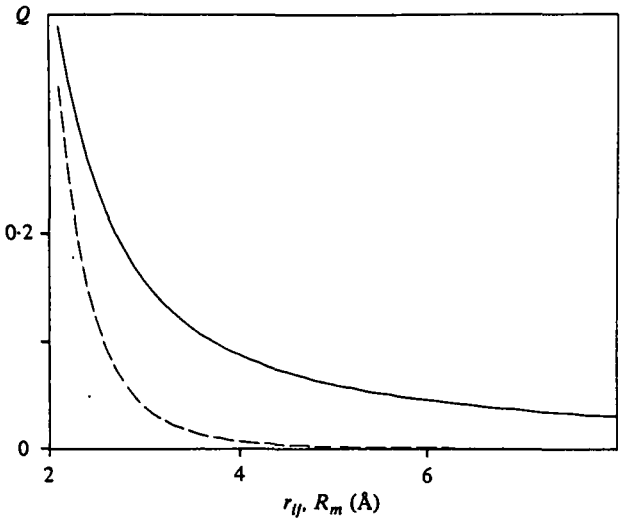

Fig. 1. Comparison of the cross-relaxation rates as a function of ^1H–^1H distances in a flexible (——) and rigid protein structure (--). For the flexible protein structure, the ratio of the cross-relaxation rates between two protons i and j relative to two protons with fixed, known distances (methylene protons) is estimated as a function $Q(R_m)$ of the maximum distance between i and j, by uniform averaging the interatomic distance between the van der Waals contact of 2 Å and R_m. The estimation was done in such a way that the correct result for Q should be below the solid line under the assumptions described in the text. Measuring Q therefore allows a rather conservative estimate of the upper limit of the distance. (Reproduced from Braun *et al.* 1981.)

of the maximal distance R_m. The 'maximal' distance is generally defined as the distance up to which a significant fraction, e.g. 95 % of the population, is occupied:

$$\frac{\sigma_{ij}}{\sigma_0} \leqslant Q(R_m). \tag{2.1.3}$$

The derivation of equation (2.1.3) is based on two arguments. The first is that in macromolecular systems the sign of $f(\tau)$ is negative and the inherent flexibility of the angular dependence in addition to the overall tumbling can only reduce the NOE effect:

$$\frac{f(\tau_{ij})}{f(\tau_R)} \leqslant 1. \tag{2.1.4}$$

The second argument assumes that the density distribution of the proton–proton distances behaves well in the sense that the maximum distance R_m and the maximal value of the density distribution ρ_{max} are anticorrelated, i.e. if R_m gets large, ρ_{max} gets small. This assumption is valid for frequently occurring distributions such as the Maxwellian, Lorentzian or Gaussian distributions, but it excludes cases such as a two-state model with two delta distributions at a small and a large distance. The average value $\langle r^{-6} \rangle$ clearly is not affected much by the maximum distance for this distance distribution. Such cases might exist in protein structures in

solution. But they seem to be not the statistically dominant cases for proton–proton distances in proteins; otherwise the proposed direct method would not work at all. However, by doing distance geometry calculations we sometimes obtain evidence for averaging processes over at least two conformations (see, for example, the example of the α-amylase inhibitor in section 4.2).

The ratio (2.1.3) can be estimated as follows:

$$\frac{\sigma_{ij}}{\sigma_0} = \frac{f(\tau_{ij})}{f(\tau_R)} r_0^6 \int_{r_m}^{\infty} \rho_{ij}(r)\, r^{-6} dr$$

$$\leqslant r_0^6 [\rho_{\max} \tfrac{1}{5}(r_m^{-5} - R_m^{-5}) + R_m^{-6}], \qquad (2.1.5)$$

where r_m is the minimal distance available, usually the sum of the van der Waals radii. When R_m gets large, the right-hand side gets small under our assumption. This function of R_m on the right-hand side can now be used to estimate for a measured ratio of the cross-relaxation rates an upper limit for the proton–proton distance.

A specific model, the uniform averaging model, for calculating $Q(R_m)$ is given in Fig. 1. This simple model might be replaced by models available from statistical analysis of molecular dynamic calculations (Olejniczak *et al.* 1984) or Monte Carlo simulations. Even if it is not possible to characterize all types of proton–proton distance distributions in proteins by one general model, certain features of a statistical analysis of molecular dynamics calculations could be used, e.g. the observation that distances between proton spins separated by only a few torsion angles show less variations than long-range distances.

The uniform averaging model has been used in Braun *et al.* (1983) to determine the distance constraints for protons separated by at most three torsion angles about single bonds differently from those for protons separated by more than three torsion angles. In the first case the rigid model was applied with four classes of distance limits: 2·4, 2·7, 3·1 and 4·0 Å. In the second case the uniform averaging model was applied with the same levels of intensities and mixing times but loosened upper limits. In subsequent protein-structure determinations, a similar scheme for the translation of NOE cross-peaks into upper limit distance constraints was used (Williamson *et al.* 1985; Kline *et al.* 1986; Braun *et al.* 1986; Wagner *et al.* 1987).

The main conclusion is that NMR data in proteins give upper-limit distance constraints or imprecise distance information with errors comparable to the size of the distances itself. On the other hand, the number of distance constraints is much larger than the number of degrees of freedom. The distance constraints provide us with a large network of restrictions. This fact converts the problem into a computationally difficult class, which cannot be solved by a fast algorithmus (Saxe, 1979). This computational problem is comparable in complexity to the protein folding problem.

2.2 *Dihedral angles constraints*

Vicinal proton–proton coupling is another source of useful geometric information. The dependence of the vicinal coupling constant between two protons H^1 and H^2 on the dihedral angle ϕ is given by a Karplus type equation (Karplus, 1959, 1963):

$$J_{H^1H^2}(\phi) = A + B\cos\phi + C\cos 2\phi. \qquad (2.2.1)$$

The parameters A, B and C for the vicinal coupling constants $^3J_{\alpha NH}$ and $^3J_{\alpha\beta}$ for polypeptides have been empirically determined by a best-fit procedure for the measured vicinal coupling constants for systems where also a highly refined X-ray structure was available. Numerous attempts have been done along these lines to determine the 'best' set of parameters (cf. De Marco *et al.* 1978*a*, *b*). All of these calibrations of course assume that the solution structure of a protein used for calibration is highly rigid and is the same as the X-ray structure. Because of this basic drawback it is advisable to use geometric information from the measured coupling constants only when it is insensitive to variations in the parameters used.

In the future, NMR structures of small globular proteins might be used for calibrating the parameters of the Karplus curve. Pardi *et al.* (1984) used the X-ray structure of BPTI (Walter & Huber, 1983) to calibrate the parameters of the amide proton-C^α proton coupling constant $^3J_{\alpha NH}$. To get a rough estimate of the influence on the calibration of taking either structure, differences in the dihedral angles between the X-ray and a representative NMR structure of BPTI (Wagner *et al.* 1987) were calculated. The DISMAN structure 1 of BPTI (see Table 1) was used as a reference structure for the family of NMR structures. The mean deviation of the ϕ backbone angles of this structure compared to the ϕ angles of the X-ray structure for those 46 amino acid residues which have been used in the calibration study amounts to 24°. All residues for which $^3J_{\alpha NH}$ coupling constants of 36 or 68 °C were measured have been included in this comparison except the carboxy terminal Ala-58. This mean deviation corresponds roughly to the scatter of the experimental data points around the best-fit theoretical curve, fig. 3 in Pardi *et al.* (1984).

But even if all the parameters A, B and C were exactly known, flexibility of the molecule prevents us using equation (2.2.1) in a straightforward way in the direct determination of polypeptide or protein conformations. The measured values of the coupling constants are averaged over the ensemble of equilibrium conformations. This fact requires that we use only the extreme values of the vicinal coupling constants for structural interpretation, because for these extreme values averaging should not have a major effect. But using only the extreme values of the Karplus curve of the vicinal coupling constant leads to a rather large inaccuracy in the dihedral angle obtained from the measured coupling constant.

The fact that averaging processes can only diminish the extreme cases has also been demonstrated by Nagayama & Wüthrich (1981) in a two-dimensional representation of the two $^3J_{\alpha\beta}$ coupling constants of the methylene protons whose dihedral angles are correlated by 120°. They distinguished three limiting cases for the fluctuations of the χ^1 angle: the fully rigid case where the experimentally

measured values are at the extreme boundary values, the case of small (30°) fluctuations around a single rotameric conformation where the experimental data point is near the boundary values, and the case of rapid exchange between at least two rotamers.

All these considerations on the flexibility of the molecule lead to a similar conclusion as to which type of dihedral angle constraints can be expected in a direct-method approach. As in the case of distance constraints, the experiments define an allowed interval for dihedral angles and the problem consists of finding all molecular conformers with dihedral angles in these allowed intervals.

3. THEORY

3.1 *Formulation of the mathematical problem*

Molecular conformations compatible with the NMR data are characterized by allowed ranges of geometric quantities such as dihedral angles or distances. The basic question in the determination of protein conformations is to characterize the conformation space compatible with these constraints. The result of such a characterization does not consist of a single structure satisfying the experimental data best but rather of a set of structures where each structure should be considered as a particular representation of the allowed conformation space. Systematic grid search calculations through all possible conformations can be done for small oligopeptides (Smith & Veber, 1986) but is not feasible for protein-structure determination. Parameters used to characterize the extent of the conformation space are the average root-mean-square distances (r.m.s. D) between pairs of structures (McLachlan, 1979) for a subset of atoms or for all atoms, standard deviations of dihedral angles and stereoviews of superpositions of structures.

The relation between the inaccuracies with which individual geometric quantities are known and the r.m.s. D values which characterize the restriction of the whole set of restrictions is by no means trivial and is sometimes surprising. A striking result of this non-trivial relation has been shown by Havel *et al.* (1979) in a simplified model of protein conformation. Each residue is represented by its C^{α} position. For each pair of C^{α}-atoms the contact of two residues is defined if the C^{α}–C^{α} distance is less than 10 Å. Then it was shown that all conformations having the same contact–noncontact scheme as the globular X-ray conformation are restricted to about 1 Å r.m.s. D value around the X-ray conformation.

The combined effect of qualitative distance information can have a quite dramatic effect on possible structures. This relation was further analysed by Wako & Scheraga (1981) in a statistical analysis of these calculations.

Distance information of this type is generally not available from present NMR techniques and it is unlikely to obtain in the future especially good information on long distances of the order of the radius of gyration. Even so, these results gave some hope that the inclusion of the packing restriction in an all-atom model together with a fine net of short proton–proton distance constraints is enough to define the globular fold of the protein. The tools to tackle this question were developed from a variety of different approaches and are described in the following sections. Having these tools and a good set of distance constraints, it actually could

be shown that the hope was justified. A clear test for this hypothesis was presented in the structural determination of the X-ray single crystal and the NMR structure of α-amylase inhibitor where for the first time an independent structural analysis of an unknown structure of a globular protein by both methods was done (Pflugrath *et al.* 1986; Kline *et al.* 1986).

3.2 *Metric matrix methods*

In the case in which all distances between all pairs of atoms are known exactly, distances can be converted into cartesian coordinates by an elegant use (Crippen & Havel, 1978; Crippen, 1981) of the matrix

$$G_{ij} = \mathbf{r}_i \cdot \mathbf{r}_j, \tag{3.2.1}$$

where $\mathbf{r}_i$ denotes the cartesian coordinates of atom i and $\cdot$ the dot product. Thus G_{ij} is a $N \times N$ matrix. The matrix elements of the metric matrix determine the coordinates uniquely except for a rotation and inversion. The relation is simply given by diagonalization of the matrix:

$$G_{ij} = \sum_{\alpha}^{N} \lambda_\alpha E_{i,\alpha} E_{j,\alpha}. \tag{3.2.2}$$

This relation can be seen by proving the two important properties of the metric matrix. The metric matrix is positive semidefinite and has rank 3. This means that all eigenvalues of the metric matrix are greater than or equal to zero and at most three eigenvalues are different from zero. This can be derived from the quadratic form of the metric matrix:

$$\sum_{i,j}^{N} g_{ij} z_i z_j = \left(\sum_{i}^{N} z_i \mathbf{r}_i\right) \cdot \left(\sum_{i}^{N} z_i \mathbf{r}_i\right) \geqslant 0. \tag{3.2.3}$$

If the quadratic form is zero one obtains a 3-dimensional vector equation or three linear equations in the N variables z_i:

$$\sum_{i}^{N} z_i \mathbf{r}_i = 0. \tag{3.2.4}$$

Therefore there are at least $N-3$ linear independent non-trivial solutions, which means that the metric matrix has at most only three eigenvalues different from zero. In the general eigenvector decomposition equation (3.2.2) of a metric matrix corresponding to a set of 3-dimensional coordinates all but three terms vanish and a comparison of (3.2.1) and (3.2.2) leads to

$$r_i^\alpha = \sqrt{(\lambda_\alpha)}\, E_{i,\alpha}. \tag{3.2.5}$$

The metric matrix can be calculated directly from the distances. This makes this quantity important for practical use:

$$G_{ii} = \frac{1}{N} \sum_{j}^{N} D_{ij}^2 - \frac{1}{2N^2} \sum_{j,k}^{N} D_{jk}^2, \tag{3.2.6}$$

$$G_{ij} = \tfrac{1}{2}(G_{ii} + G_{jj} - D_{ij}^2). \tag{3.2.7}$$

In the first of these equations for the diagonal term it is implicitly assumed that the structure is centred to the origin. Another choice would be setting one particular atom usually numbered o at the origin:

$$G_{ii} = \mathbf{r}_i \cdot \mathbf{r}_i = D^2_{0,i}. \tag{3.2.8}$$

The set of equations (3.2.1)–(3.2.8) gives a simple and direct relation between distances and coordinates. A detailed mathematical description of these equations can be found in Havel *et al.* (1983). The equations are derived on the assumption that all distances are known exactly, i.e. in the case of a complete and correct distance matrix. In practical applications these basic assumptions are almost never fulfilled in typical NMR data sets, but there is a hope that the assumptions still represent a useful approximation (Braun *et al.* 1981; Crippen *et al.* 1981).

As we have seen in the previous section, distance information is given in the form of an interval,

$$L_{ij} \leqslant D_{ij} \leqslant U_{ij}, \tag{3.2.9}$$

and usually only for a small subset of all possible atom pairs. In the case of the α-amylase inhibitor (Kline *et al.* 1986) there are about 500 distance constraints from NMR data. These constraints are to be complemented with about 1500 constraints for bond lengths and bond-angle constraints. But this is a small number compared to all possible atom pair distances, which must be known for all 827 atoms in the pseudo-atom representation (Wüthrich *et al.* 1983) to generate a full metric matrix.

Initial distances are chosen at random between the limits given in (3.2.9). This usually leads to a distance matrix not embeddable in three 3 dimensions, i.e. the metric matrix calculated from the distances is not positive semidefinite with rank 3. This means there are no coordinates in 3 dimensions with the same distances as the randomly chosen distances. In practice the approximation using the three greatest eigenvalues in equation (3.2.5) to calculate the coordinates is usually done.

In our experience, with a large system of say $N \geqslant 50$ (i.e. even a short polypeptide chain with all atoms included would be large in this sense) the three eigenvalues with greatest absolute value are not always positive. This is partially related to the fact that the randomly chosen distances within bounds satisfying triangle inequalities do not necessarily satisfy the triangle inequalities among themselves. This is especially true if the bounds are loose. As a simple example, let us assume that the upper and lower bounds for the 3 distances of a triangle are 10 and 2 Å, respectively. Then the direct and the inverse triangle inequality for the upper and lower bounds are satisfied. However, choosing the three distances at random independently within the allowed range might lead to three distances not consistent with the triangle inequality (e.g. 10, 2 and 2 Å).

Crippen *et al.* (1981) calculated correlation coefficients between the three distances of a triangle imposed by the upper and lower bounds. These were then used in correlating random choices of the initial distances. The probability of violating the triangle inequality is thereby reduced but not to zero. A different

solution to correct the triangle inequalities exactly for the distances within the given bounds was proposed by Braun *et al.* (1981). In this approach specific changes for those distances violating the triangle inequality were derived, such that the new distances are still within the allowed range and satisfy the triangle inequality. Higher-order inequalities (Havel *et al.* 1983) restricting further upper and lower bounds have also been derived but it is not clear if they are of practical use because of the enormous amount of computing time.

The connexion between diagonalization and coordinates has also been recognized and used in an abstract way by regarding r_i *as a* $3N$ dimensional vector of a conformation i and calculating (3.2.1) for a set of M conformations. In this case G is a $M \times M$ matrix. Two-dimensional projected coordinates of each conformation i were then used in a two-dimensional graphical representation of the set of conformations in the refinement studies of X-ray protein structure determination (Diamond, 1974) and analysis of molecular-dynamics calculations (Levitt, 1983).

For use in the tertiary structure determination of a polypeptide chain it is not sufficient to have an embed algorithm; one also has to combine the standard geometry of individual amino acids (e.g. ECEPP geometric parameters) in a library with the embed procedure to extract the distance constraints which define the stereo chemistry.

This was first done in the calculation of micelle-bound glucagon (Braun *et al.* 1981). A new FORTRAN computer program, based on the metric matrix approach of Crippen & Havel (1978), was written to model the individual amino acids by distances. By relying on pure geometrical principles and on simplified representations of protein structures, the distance geometry algorithm (Crippen, 1977, 1981) was designed to circumvent the local minimum problem of empirical energy minimization (Nemethy & Scheraga, 1977). We intended, however, to use it as some sort of model building algorithm for polypeptide chains, where the experimental distance information could easily be included as an additional constraint. In designing the program we had to define the standard bond lengths and bond angles by distances. This was done by interfacing the metric matrix approach with the standard amino acid library of ECEPP (Momany *et al.* 1975). The only necessary input information for the chemical structure consisted then of the amino acid sequence. All relevant distance information was automatically read from the ECEPP library.

The basic EMBED algorithm was extended by using new triangle inequalities for the distances (see above) to get an improved set of initial distances in the initial embedding. Truncation of the metric matrix to the space spanned by the eigenvectors with the three greatest eigenvalues was replaced by a gradual contraction using a convex combination of old and new distances derived from the approximate coordinates. This procedure led to an improved set of initial coordinates for the refinement.

Test calculations also showed that individual chirality terms must be added to the error function in the refinement procedure to get the correct chirality of the asymmetric C^{α} carbon atoms of the amino acid residues and the chirality of the C^{β} atom of Thr and Ile. This requirement extended the original scheme of the

metric matrix approach from a pure distance geometry problem towards a refinement problem.

An extension of the approach along these lines was done in the program DISGEO (Havel & Wüthrich, 1984), where several new features were included to make this approach workable also for small proteins in the pseudo-atom representation. Embedding is done in two steps, where first the conformation of a substructure consisting only of a subset of a third of the atoms in the complete structure is calculated and then the distances extracted from the calculated substructures are relaxed somewhat and included as additional constraints for the embedding of the complete structure. Even in the pseudo-atom representation, the N^2 memory demand for a small protein is a major problem in this approach. This was solved by an efficient storage of incomplete distance information and use of disk storage in cases which are not time-critical. Because of sequential disk access, the computation of triangle inequality limits on all distances from the given set of input distance bounds is a highly non-trivial problem and has been solved by implementing a special shortest-path algorithm. In addition to the chirality constraints to fix the chirality of the L amino acids, further chirality constraints were included to force planarity of the peptide planes and the aromatic rings. Chirality constraints were also used to restrict torsion angles around a single bond where the ambiguity of the relation between chirality and torsion angles (one chirality value typically corresponds to two torsion angles) is resolved by using several dihedral angles around the same chemical bond (Havel & Wüthrich, 1985).

3.3 *Future developments*

The basic EMBED algorithm proposed first by Crippen & Havel (1978) does not include any term dealing with the chirality constraints. In our first approach to use this method to model the individual amino acid residues by distance information, L and D amino acids could not be distinguished by distance information because a point inversion of a 3-dimensional structure converts a L into a D configuration leaving the distances invariant. Therefore an *ad hoc* procedure was proposed (Braun *et al.* 1981; Crippen *et al.* 1981) to include the chirality constraints in the refinement procedure where a target function of the type discussed in the next section is minimized. Even though this approach worked in practice to some extent, it left some burden to the final refinement procedure. An algebraic embed procedure including chirality constraints still needs to be developed.

The basic metric matrix approach has two drawbacks if the system is a typical small protein of the size which can now be studied by 2-D NMR experiments. For this size the truncation to the three largest eigenvalues of the metric matrix represents a poor approximation. The memory demand is quadratical in N, the number of atoms. Even on a virtual memory system this represents a potential limitation because of paging. It seems that the redundancy of the distance information has not yet been exploited fully to generate a complete, but small subset of distances with a size linear in N carrying the necessary 3-dimensional information.

Current research (Sippl & Scheraga, 1985, 1986; Schlitter, 1986) is concerned with the correction or prediction of the undefined distances such that the complete distance matrix is embeddable. Theorems on necessary and sufficient conditions on the embeddability (Blumenthal, 1970) are not of much help because in typical cases the conditions are violated, and the way to satisfy these conditions is then done by a nonlinear best-fit procedure, i.e. that feature one wanted to avoid at the outset of this approach. The proposed approach used Caley–Menger coordinates to fill out a complete distance matrix from a sparse, incompletely defined distance matrix. This was done by a suitable simplification of the Caley–Menger determinants such that they can be calculated with minimal effort. However, it has not been demonstrated that this procedure can be used in practice for a molecule of the size of a protein.

3.4 *Variable target function method*

As we have seen in the previous section, the mathematical problem of the general embedding problem is not yet solved. There are mathematical indications that there might not be a reasonable algorithm to solve it in a reasonable computer time for large system, because it was shown that this problem is *NP* hard, i.e. the number of operations needed to solve that general problem is for any algorithm not bounded by a polynomial in N (Saxe, 1979).

Practical experience gained with the NMR data of glucagon (Braun *et al.* 1981, 1983) and preliminary data on the NMR data of BPTI using a simplified two-point representation for each residue, where one point represents the C^{α} atom and the other point the side-chain of the residue, suggested that, for larger systems, one certainly cannot avoid the use of nonlinear optimization at some phase of the algorithm. Then it is a natural idea to try the nonlinear optimization method from the outset in a straightforward way.

In the general frame of a nonlinear optimization scheme one should try to keep the number of independent variables as small as possible (Fletcher, 1980). An obvious choice are then the torsion angles as independent variables. The problem of local minima (Nemethy & Scheraga, 1977; Gō & Scheraga, 1978) seems sometimes easier to be solved by artificially enlarging the number of degrees of freedom (Crippen, 1977; Purisima & Scheraga, 1986), but reduction of the higher dimensional structures to the 3-dimensional space usually poses additional problems. So it is not yet clear if the local minima problem is solved by the above-mentioned procedures or only translated to a different problem. A real solution of the local minima problem in protein-structure calculations is not just an *ad hoc* algorithm proposed to avoid the local minima, but also a realistic picture of the nature of local minima. This is still missing.

To have a program at hand to treat proteins of the size that can be studied by the present NMR techniques, a new algorithm was implemented into a FORTRAN computer program DISMAN (Braun & Gō, 1985). This program tries to make best use of the available data. Knowledge of stereo chemical data such as standard bond lengths, bond angles and the repulsive core radii have to be added to the pure experimental data from NMR to start structural elucidation. This is done by

generating the atomic coordinates from the dihedral angles and changing the dihedral angles in such a way that a target function becomes zero for a structure which fulfils all distance constraints. The target function is a measure of how good the distance constraints are fulfilled. There are many ways to construct such target functions. A typical form of the target function is given by

$$T = \sum_{i<j} [\theta(D_{ij} - U_{ij})(D_{ij} - U_{ij}) + \theta(L_{ij} - D_{ij})(L_{ij} - D_{ij})]. \qquad (3.4.1)$$

The function $\theta(x)$, which is 0 for $x \leqslant 0$ and 1 for $x > 0$, is used to sum up all distance violations. The summation over the atompairs i and j is, of course, only over those pairs where there are constraints. The function T is 0 for a solution, is positive for all conformations not satisfying the constraints perfectly and increases as the distance constraints violations are getting worse. Usually the target functions are variations of the type (3.4.1) that only the square of distances is used because of efficient computation and such that they are also continuously differentiable at the boundaries $D_{ij} = L_{ij}$ and $D_{ij} = U_{ij}$. This can be done by taking some powers of the distance violations (see, for example, Braun *et al.* 1981). Summing up only terms which represent distance violations can be done without any discontinuity of the derivative. This approach yields a clear correspondence between a solution and $T = 0$, which is a definite advantage of distance geometry calculations over energy minimization where such *a priori* knowledge of the global minimum is missing. This advantage is lost if one uses the approach of Marion *et al.* (1986) of including terms when the distance constraints are satisfied.

The variable target function approach is not entirely new, because computation in torsion angle space has been done extensively in energy minimization studies of proteins (Burgess & Scheraga, 1975; Meirovitch & Scheraga, 1981; Levitt, 1982, 1983). Two new features are an efficient way to calculate gradient information of the target function (Noguti & Gō, 1983; Abe *et al.* 1984) and the way the target function in (3.4.1) is minimized. Our strategy of minimizing is similar to the strategy of Ooi *et al.* (1978) in the regularization studies of proteins.

The method of variable target functions means that one does not try to minimize T at once but rather to minimize gradually a series of functions which approximate T. More specifically, for a polypeptide chain of n residues the target functions $T_{k,l}$ ($k = 1, 2, \ldots, n$ and $l = 1, 2, \ldots, n$) only include those terms of the form as in (3.4.1) for atom pairs belonging to residues with difference of their sequence numbers less than k if the upper or lower limits are from NMR data or less than l if the lower limits are the sum of repulsive core radii. The strategy consists in first minimizing $T_{k,l}$ with small values of k and l and then gradually increasing k and l up to n. The final solution of the problem consists of one or several conformations having zero values for $T_{n,n} = T$. The exact definition of the terms used in DISMAN can be found in (Braun & Gō, 1985). In case of an overdetermined problem the best conformation consistent with the input distance information and stereo-chemical criteria is the one which gives the global minimum of the target function.

This strategy was shown to be effective if good distance information of a short range nature is available. In the case of artificial distance data with exact

proton-proton distances less than 5 Å, the polypeptide backbone structure of BPTI could be almost exactly regenerated with r.m.s. D values of 0·03 to 0·14 Å. The calculations started with ten randomly chosen initial conformations differing from each other and from the final structures by about 15 Å (Braun & Gō, 1985). In practice this good distance information is certainly not available; however, good distance information of a short-range nature is always obtained in the process of sequential resonance assignments (Wagner & Wüthrich, 1982*a*) and measurements of spin–spin coupling constants can be used directly to restrict the allowed torsion angles.

Exact characterization of good distance information which can guide the conformation from correct short-range to medium or long-range conformations is missing. More extensive numerical experience is certainly needed. Some heuristic ideas of describing the success of the method are as follows. Short-and long-range distance constraints impose different types of restrictions on the polypeptide conformation. Once short-range distance constraints are fulfilled, the polypeptide chain keeps a large amount of 'flexibility' for those conformational changes maintaining the short-range distance information. Small local changes can give rise to drastic global changes. So these small changes can be used to satisfy the long-range distance constraints.

In DISMAN certainly, only a specific aspect of the variable target function method is implemented. In future a cybernetic choice of the target function using feedback methods could give improved results. Now, information on the success or failure of a certain run is not used in the calculation of further conformations. Also, a combination with Monte Carlo Methods of escaping local minima might be a possibility to improve the performance of the method.

The second device implemented in DISMAN is a method of fast calculation of the gradient of the target function. A different scheme has been proposed independently by Levitt (1983). It is therefore of some interest to see the relation between these schemes. Both methods can be applied either to an empirical energy function or to the target function of the type (3.4.1). To make the comparison transparent we describe it in the notation of Abe *et al.* (1984) for numbering the atoms and the dihedral angles. Atom indices are denoted by Greek letters $\alpha, \beta, \gamma, \ldots$ and torsion angle indices by $a, b, c, \ldots$.

The basic idea in the first scheme of efficient calculation of the gradient was introduced by Noguti & Gō (1983). It relies on the 'factorization' of the terms in the gradient into quantities dependent on torsion angles and quantities dependent on individual atoms and of the grouping of all atoms in the molecule into units V_a attached to each rotatable bond. Each unit consists of one or more atoms. Units are defined by the property that there are no rotatable bonds within them and therefore the relative positions of atoms within each of them remain fixed for any conformational changes of the molecule. The gradient of a function E, which is a sum over pairwise distance dependent potentials $\phi_{\alpha\beta}(|\mathbf{r}_\alpha - \mathbf{r}_\beta|)$ is given by

$$\frac{\partial E}{\partial \theta_a} = -\mathbf{e}_a \cdot \mathbf{F}_a - (\mathbf{e}_a \wedge \mathbf{r}_{\epsilon(a)}) \cdot \mathbf{G}_a, \qquad (3.4.2)$$

where $\mathbf{F}_a$ and $\mathbf{G}_a$ are calculated via simple recurrent equations:

$$\mathbf{F}_a = \mathbf{f}_a + \sum_{k=1}^{p(a)} \mathbf{F}_{s(k,a)}, \tag{3.4.3}$$

$$\mathbf{G}_a = \mathbf{g}_a + \sum_{k=1}^{p(a)} \mathbf{G}_{s(k,a)}. \tag{3.4.4}$$

$$\mathbf{f}_a = \sum_{\substack{\alpha \in V_a \\ \beta}} \frac{\phi'_{\alpha\beta}}{|\mathbf{r}_\alpha - \mathbf{r}_\beta|} (\mathbf{r}_\alpha \wedge \mathbf{r}_\beta), \tag{3.4.5}$$

$$\mathbf{g}_a = \sum_{\substack{\alpha \in V_a \\ \beta}} \frac{\phi'_{\alpha\beta}}{|\mathbf{r}_\alpha - \mathbf{r}_\beta|} (\mathbf{r}_\alpha - \mathbf{r}_\beta). \tag{3.4.6}$$

The indices $s(k, a)$ $k = 1, 2, \ldots, p(a)$ describe the hierarchical tree structure of a branched polymer. It gives for each rotatable bond a the indices of the torsional angles branching from bond a and $p(a)$ counts the branches at a. The order of the torsion angles is chosen appropriately (Abe *et al.* 1984).

Because of the grouping of the atoms in no-overlapping units, all auxiliary quantities $\mathbf{f}_a$ and $\mathbf{g}_a$ can be calculated in N^2 number of operations and the result stored appropriately. The summation of the recurrent equations and the calculation of the individual components of the gradient (3.4.2) is then done in a second phase where the calculational effort is on the order of m, the number of torsional angles.

In the second scheme (Levitt, 1983) first derivatives are calculated with respect to all cartesian coordinates and stored.

$$\frac{\partial E}{\partial \mathbf{r}_\alpha} = \sum_\beta \frac{\phi'_{\alpha\beta}}{|\mathbf{r}_\alpha - \mathbf{r}_\beta|} (\mathbf{r}_\alpha - \mathbf{r}_\beta). \tag{3.4.7}$$

This can be done in N^2 number of operations. In a second step these derivatives are transformed to torsion angle space:

$$\frac{\partial E}{\partial \theta_a} = \sum_\alpha \frac{\partial E}{\partial \mathbf{r}_\alpha} \cdot [\mathbf{e}_a \wedge (\mathbf{r}_\alpha - \mathbf{r}_{\varepsilon(a)})]. \tag{3.4.8}$$

This second step needs only $N \times m$ number of operations.

Both methods are mathematical equivalent because of the following relation:

$$(\mathbf{r}_\alpha - \mathbf{r}_\beta) \cdot [\mathbf{e}_a \wedge (\mathbf{r}_\alpha - \mathbf{r}_{\varepsilon(a)})] = -\mathbf{e}_a \cdot (\mathbf{r}_\alpha \wedge \mathbf{r}_\beta) - (\mathbf{e}_a \wedge \mathbf{r}_{\varepsilon(a)}) \cdot (\mathbf{r}_\alpha - \mathbf{r}_\beta). \tag{3.4.9}$$

Both methods have roughly the same efficiency and are a factor N faster than straightforward analytical (Pottle *et al.* 1980) or numerical methods. The first method needs less memory space because the recurrent equations can be implemented using a stack method. In practice, however, this is not a crucial advantage. This is important, however, for the second derivative method. The first method has been implemented to calculate analytical second-order derivatives in the normal mode analysis (Gō *et al.* 1983) where other methods use analytical first- and numerical second-order derivatives (Levitt *et al.* 1985). The first method can also easily extended to systems consisting of two or more molecules (Braun *et al.*

1984), which is important for studying enzyme-substrate systems. In this extension the rotation and translation parameters (translation vectors and Eulerian angles for rotation) can be embedded in a natural way in the hierarchical tree structure and are used formally as torsion angles.

Explicit restrictions on torsional angles from spin–spin coupling constants (Karplus, 1959; DeMarco *et al.* 1978*a*; *b*) can be implemented easily. This is done in DISMAN following the same philosophy used in constructing the target function from distance constraints. For each restricted torsion angle to an allowed region, i.e. to a region compatible with the NMR data, the target function is defined as zero within the allowed region, has continuous first derivatives at both region boundaries and increases smoothly with the amount of deviation from the allowed region. Some care has to be taken, because the 3-dimensional coordinates are 2π periodic functions of the torsion angles. Also the evaluation of the function should be fast. In DISMAN a fourth-order polynomial with continuous first derivatives and 2π periodicity is constructed.

If $[\theta_L, \theta_R]$ denotes the allowed region where all angles have been scaled to radians, we can always achieve the condition

$$0 \leqslant \theta_R - \theta_L \leqslant 2\pi. \tag{3.4.10}$$

This means in practice that if we want to restrict the angle near 180°, we have to choose the left and right limits of the angle interval in degrees below and above 180°. The allowed region can also be characterized by its mean value m and width w where

$$m = \tfrac{1}{2}(\theta_R + \theta_L); \quad w = \tfrac{1}{2}(\theta_R - \theta_L). \tag{3.4.11}$$

The deviation $\Delta\theta$ of the current torsion angle from its mean value is also 2π-periodic and can be always shifted in the interval $[0, 2\pi]$. The allowed region splits by this shift in two separate regions, $[0, w]$ and $[2\pi - w, 2\pi]$; and the forbidden region therefore remains in one contiguous segment $[w, 2\pi - w]$. This transformation of the allowed regions in the variable θ and in the deviation from the mean value m, $\Delta\theta$, are schematically sketched in Fig. 2 by the hatched areas. By transforming the deviation $\Delta\theta$ by

$$\widetilde{\Delta\theta} = \frac{\Delta\theta - \pi}{\pi - w} \tag{3.4.12}$$

the forbidden region is shifted to $[-1, +1]$ and the target function is defined as

$$T(\widetilde{\Delta\theta}) = \begin{cases} 0 & \text{if } |\widetilde{\Delta\theta}| \geqslant 1; \\ (\widetilde{\Delta\theta}^2 - 1)^2 & \text{otherwise.} \end{cases} \tag{3.4.13}$$

T and its first derivative are continuous also at the boundary $\widetilde{\Delta\theta} = \pm 1$. The singular point of the transformation (3.4.12) occurs at $w = \pi$, i.e. there is no restriction on the original torsional angle $\theta_R = \theta_L + 2\pi$. This singular point and a small neighbourhood can therefore be always avoided. The practical interesting point of fixing the angle to one particular value by $\theta_R = \theta_L$ means $w = 0$ and poses no particular singularity problems.

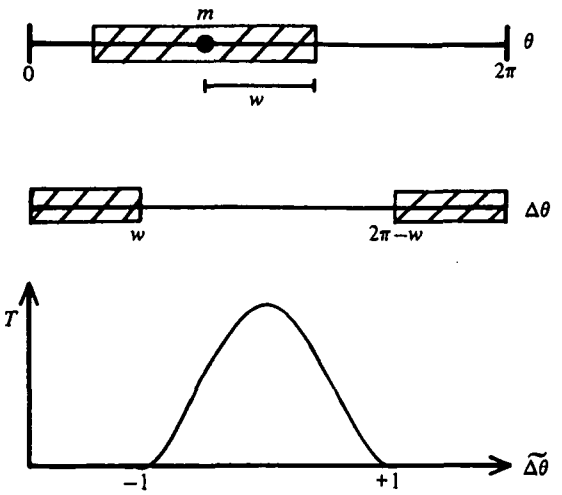

Fig. 2. Graphical representation of the transformations used in the definition of the target function for torsion-angle restrictions. The letters m and w denote the midpoint and the width of the allowed interval for θ which is drawn as the hatched segment at the top line. The deviation $\Delta\theta$ of the torsion angle from the mean value m is shifted into the interval $[0, 2\pi]$ and the allowed area for it is therefore split into the two hatched segments indicated in the middle line. At the bottom the graph of the target function T is schematically sketched as a function of $\widetilde{\Delta\theta}$.

The procedure can also be easily extended to several allowed regions by first ordering the allowed intervals $[\theta^i_L, \theta^i_R]$ $(i = 1, 2, \ldots)$ and shifting each forbidden interval $[\theta^i_R, \theta^{i+1}_L]$ by a slightly modified equation (3.4.12) into $[-1, +1]$ and apply (3.4.13). This implementation is of some practical importance in the stereospecific resonance assignment.

The variable target function method certainly has some resemblance to restrained energy minimization in torsion angle space (Levitt, 1983). Both methods use torsion angles as independent variables and are minimizing functions of a similar type in the torsion angle space. The basic difference is the fact that in the variable target function method a series of target functions $T_{k,l}$ are minimized rather than a certain pseudo energy function. This approaches the local minima problem in a quite different way from modification of the infinite repulsion energy terms by finite models for overlap atoms ('soft atoms'). This means that in the stage of the minimization of $T_{k,l}$ all atom pairs whose residue numbers differ by more than l can freely penetrate each other, whereas there is still some barrier in the soft atom model. Also the restraints or constraints are brought differently into play by the two methods.

However, apart from these more technical differences, the main difference is the philosophy of the approach. We are mainly concerned with the direct structural consequences of the pure NMR data with the least amount of additional assumptions. These are stereo chemical data like bond lengths, bond angles and van der Waals or repulsive core radii. The relevance of a structure found by restrained energy minimization (this method has not yet been applied to a typical artificial or experimental NMR data set for proteins) is difficult to judge. Is it mainly

determined by the energy or by the restraining terms? The fact that empirical energy calculations have not yet succeeded in predicting the correct global fold of proteins in solution indicates that the additional energy terms are mainly an additional burden to the calculation without guiding the minimizer to the correct global fold. Of course the structures found by the variable target function methods might have quite unrealistic high energies. So these structures should be further treated with an energy program which will change the conformation minimally but will reduce the energy drastically. A suitable tool for this seems to be the Newton–Raphson minimizer (unpublished results), which can find the next local minima most efficiently. Escaping the local minimum is in my opinion an undesirable property at that stage. The need to avoid any unnecessary burden in the calculation comes from the requirement to explore the vast conformation space as largely as possible. Improving the efficiency of the present algorithms, adapting them to the next generation of supercomputers and avoiding any computational load not dictated by the experimental data might then allow them to generate statistically significant ensembles of structures. The number of solutions found today in typical application (see Section 4) are of course only case studies.

3.5 *Restrained molecular dynamics*

Restrained molecular dynamics (MD) has been shown to be an additional valuable tool in elucidating the molecular conformations compatible with NMR data (Kaptein *et al.* 1985; Clore *et al.* 1985; Brünger *et al.* 1986). Existing computer programs for MD calculations (van Gunsteren & Berendsen, 1982; Brooks *et al.* 1983) have been modified to allow inclusion of the NMR data. This is done by adding a pseudo pair potential of the form

$$U_{\mathrm{NOE}} = \begin{cases} \frac{1}{2}c_1(r_{ij}-r_{ij}^0)^2 & \text{if} \quad r_{ij} > r_{ij}^0 \\ \frac{1}{2}c_2(r_{ij}-r_{ij}^0)^2 & \text{if} \quad r_{ij} < r_{ij}^0. \end{cases} \tag{3.5.1}$$

to the potential function used in the free dynamics. The target distances r_{ij}^0 are estimated from the NOE intensity cross-peaks using the r^{-6} dependence. The force constants c_1 and c_2 are chosen in such a way that if the deviations of the actual distances r_{ij} from their target distances r_{ij}^0 are equal to the estimated errors of r_{ij}^0 the pseudo energy U_{NOE} increases by $\frac{1}{2}kT$. These additional harmonic pseudo forces act like 'strings' between those atom pairs constrained by the NMR data and drive the molecular conformations towards conformations compatible with the NMR data. In some calculations (Kaptein *et al.* 1985) c_2 is also set to zero. In that way only the positive evidence of a NOE cross-peak is taken into account.

The numerical integration schemes used to solve Newton's equations of motions are described in the papers mentioned above and will not be repeated here. Our primary interest is: in what respect do these programs use NMR data and what can one learn from these calculations?

Restrained MD calculations of protein conformations using NMR data have been done with two different aims. One aim was the refinement (Kaptein *et al.* 1985) of a model built structure of the lac repressor head piece (Zuiderweg *et al.*

1984) that crudely satisfies the NOE distance constraints. The global fold of the polypeptide chain, in this specific case the relative position and orientation of three α-helices, is already determined by the model built structure. Restrained MD is then used to decrease both the potential energy and the pseudo energy U_{NOE} arising from the NOE distance constraints and to study time-dependent effects (local correlation time and time-averages of geometric quantities related to NOE data) of the trajectories of the modified dynamics by (3.5.1).

In a more ambitious use of restrained MD (Clore *et al.* 1985; Brünger *et al.* 1986) the aim is similar to that in the two previously described methods, to establish the global fold of the polypeptide chain purely on the basis of the NMR distance constraints and independently of the initial conformations. Starting conformations were chosen with either extended or helical segments that could be assigned by typical NOE pattern (Wüthrich *et al.* 1984). Initial and final structures differ by r.m.s. D values of the size of the molecule. Applications of this method to real experimental NMR data have not yet been reported. In the case of an artificial distance data set for the protein crambin (Brünger *et al.* 1986), folding of the structure by this method was successful with distance constraints which can be obtained in principle by NMR. The way the distance constraints were included in the calculations is similar to the strategy used in the variable target function method. First, the molecular dynamics calculations were done by including only short-range distance constraints for 2 and 5 ps and by starting from a completely extended conformation. After 500 cycles of conjugent gradient energy minimization with the short-range constraints, several phases of restrained molecular dynamics with all distance constraints and increasing weight factors for them were done.

3.6 *Analysis of structures*

The considerations on the internal flexibility of proteins in Section 2.1 and 2.2 explained that the central problem for the NMR structure determinations of proteins is not a nonlinear best-fit problem of the structure to a number of measured parameters. It is the characterization of the conformation space of all structures compatible with all constraints. Several parameters have been used to quantify this aspect of the structure determination.

First the structures should be consistent with the NMR constraints. In the ideal case there should be no constraints violations at all. In practice statistics of the number and size of the residual violations should be presented to judge the quality of the obtained structures. The sum of distance violations divided by the number of distance constraints (the average distance violation) would be a quantity in [Å] which could be used to compare the quality of calculations with different data sets and different proteins. It is roughly the equivalent of the R-factor used in the X-ray structure determination.

Using only a small subset of the cross-peaks of the NOESY spectra, i.e. a small subset of distance constraints, it is in general quite easy to generate structures with small average-distance violation. But then the variations of the calculated

structures might be large. Quantities measuring the variations of the structures are root-mean-square distances (r.m.s. D) between pairs of structures for a subset of atoms or for all atoms (see McLachlan, 1979, for definition, history of use and a fast way of calculating r.m.s. D values). Depending on the subset of the atoms used in the calculation for the r.m.s. D, the value is a measure of local or global conformational variations. Important quantities are the r.m.s. D values for the backbone structure (BB) or the restricted side-chain representation (Braun *et al.* 1983). Variations of the local conformations comparing two conformations can also be quantified by averages of the differences of torsion angles over all residues for the torsion angles ϕ, ψ or χ^1, DHAD values (Havel & Wüthrich, 1985). The extent of the conformation space of all structures is measured by using the average values over all pairs of conformations. More sophisticated parameters such as the 'volume' of the allowed conformation space are difficult to estimate from the few calculated structures.

Additional methods to visualize the variations are stereo views of optimal superposed structures. They usually show quite clearly those parts of the NMR structures least constrained by the data.

A further method to judge the quality of NMR structures is the calculation of NOESY spectra from the ensemble of calculated structures. Comparing them with the experimental NOESY spectra represents an objective test of the extent of the used distance constraints. In the case of the experimental data sets of metallothionein (see Section 4.2.7), preliminary methods were encouraging (unpublished results).

4. APPLICATIONS

Applications of the programs to calculate structures compatible with distance constraints have been done with two different types of distance constraints data sets.

With the first type, the distance constraints were extracted from known X-ray structures. In the calculations with these simulated distance constraints sets (Havel & Wüthrich, 1985; Braun & Gō, 1985; Brünger *et al.* 1986) one first wants to demonstrate that, for a sufficient complete distance constraints data set, the calculated structures converge to the structure from which the distances were extracted. Also one wants to explore the theoretical structural consequences of distance data sets to set guidelines for the experimental work. What structural features can be expected in a typical experimental data set? Which additional data can significantly improve the structures? Present calculations already indicate where improvements of the structures can be expected.

A systematic exploration and a theoretical analysis of the correlation between distance constraints data sets and their structural restrictions is still missing but the described computational tools in Section 3 above, together with the availability of supercomputers, can give us in the next few years the necessary data base to understand in more detail the restrictions imposed by short proton–proton distances on the protein conformations. This question, besides of being important

in the structure determination in solution by NMR data, also has some relevance for the prediction of the tertiary structure of proteins by empirical energy calculations (Nemethy & Scheraga, 1977), because the main part of the non-bonded interactions is of short-distance range compared to the radius of gyration of a typical globular protein.

The second type consists of the real experimental NMR data sets. In these distance constraints data sets it is not *a priori* clear that a solution exists at all; and if there is a solution whether it is a unique solution. Calculations of protein conformations compatible with the NMR data, besides giving an objective and quantitive measure of the restriction of the NMR data, are also an independent check of the sequence-specific resonance assignment (Dubs *et al.* 1979; Wagner & Wüthrich 1982*a*).

4.1 *Simulated data sets*

Calculations of protein conformations with simulated data sets have so far been reported for bovine pancreatic trypsin inhibitor BPTI (Havel & Wüthrich, 1984; 1985; Braun & Gō, 1985) and crambin (Brünger *et al.* 1986). The number of calculated structures per given data set in all of these studies were rather limited (typically three structures), so the reported r.m.s. D values have to be taken with some care.

The three methods reported here use very different approaches to fold a protein under the influence of the distance constraints. In all three cases the distance constraints data sets were chosen from those short-distance constraints that can be expected to be determined by NMR methods (i.e. shorter than 4 or 5 Å). The results obtained clearly demonstrate that the global features of a structure calculated from present NMR data sets are reliable, despite the fact that all experimentally accessible distances are short compared to the diameter of the molecules. This observation contradicts popular criticism (see, for example, Schmidt & Kuntz, 1984). However, this does not mean that every experimental distance constraints data set leads to uniquely defined global structure. In each application of one of the described methods it is necessary to show the structural restriction of the data.

As an illustrative example of the type of calculations, the results of the DISMAN calculations with several distance constraints data sets of BPTI are presented (Braun & Gō, 1985). In test calculations of this type one has to separate the influence of the starting conformations from the influence of the data set. The approach taken in this study was therefore first to generate ten structures by choosing the variable dihedral angles randomly. The r.m.s. D values comparing all pairs of initial structures ranged from around 8 to 22 Å. A typical example of one of these initial structures is shown in Fig. 3. This set of initial structures was then used as starting conformation with several data sets. By using the most stringent data set (EX5), where all exact short proton–proton distances less than 5 Å have been used as constraints, the polypeptide backbone of BPTI was nearly exactly regenerated with r.m.s. D values ranging from 0·01 to 0·15 Å by starting

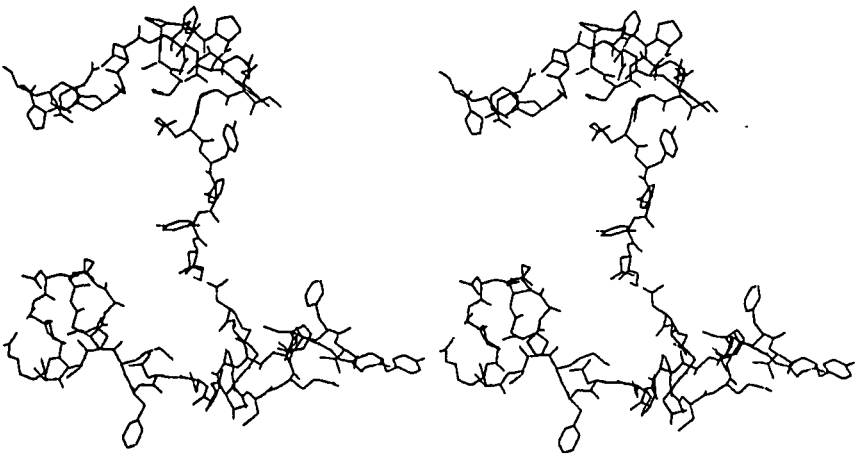

Fig. 3. Stereo view of one of the ten initial random structures of BPTI used in the calculations of DISMAN with the simulated data sets. For clarity only the heavy atoms are shown. In the calculations all hydrogen atoms are included.

from the ten randomly chosen initial structures. This again shows that short proton–proton distances are potentially a quite powerful source of information for restricting the global polypeptide fold.

Data sets of the type experimentally available by the present NMR techniques are the data sets AL5 and AU5 defined below. Both data sets consists of short- and long-range constraints. Short-range distance constraints were defined as constraints between the protons NH, H^{α} and H^{β}, which belong to residues separated sequentially by 2 or less intervening residues. Intraresidue constraints were deliberately excluded, because the stereospecific assignment of the methylene protons (a tedious and difficult procedure) for all residues is required. These distance types were put into six classes from 2 to 5 Å with an interval of 0·5 Å. The upper and lower bound distance constraints were defined as the upper and lower bounds of the class to which it belongs. In AU5 only the upper bounds for the short-range constraints were used, in AL5 upper and lower bounds were included. For the long-range constraints upper limit distance constraints were set in both data sets to 5 Å if the corresponding proton–proton distance were less than 5 Å. So the two data sets differ in the short-range data sets where in AL5 also lower limits have been included. Since quantification of lower limits from the NOE data is difficult because of the inherent flexibility of the protein, we want to study the theoretical implications for the structure by this data set.

Both data sets were used in calculations with the same initial random structures. Because of restricted computer time we had to choose three among the ten previously generated initial conformations. The average r.m.s. *D* values comparing the calculated structures with the BPTI X-ray structure for the data set AL5 were 1·4 and 2·3 Å for backbone and all atoms, respectively, and 1·5 and 2·5 Å for AU5. The result indicates that the differences in the restrictions of structures between the data sets AL5 and AU5 is not significant. In Fig 4 the heavy atom representation of calculated structures with AL5 (A) and AU5 (B) superposed to the BPTI X-ray

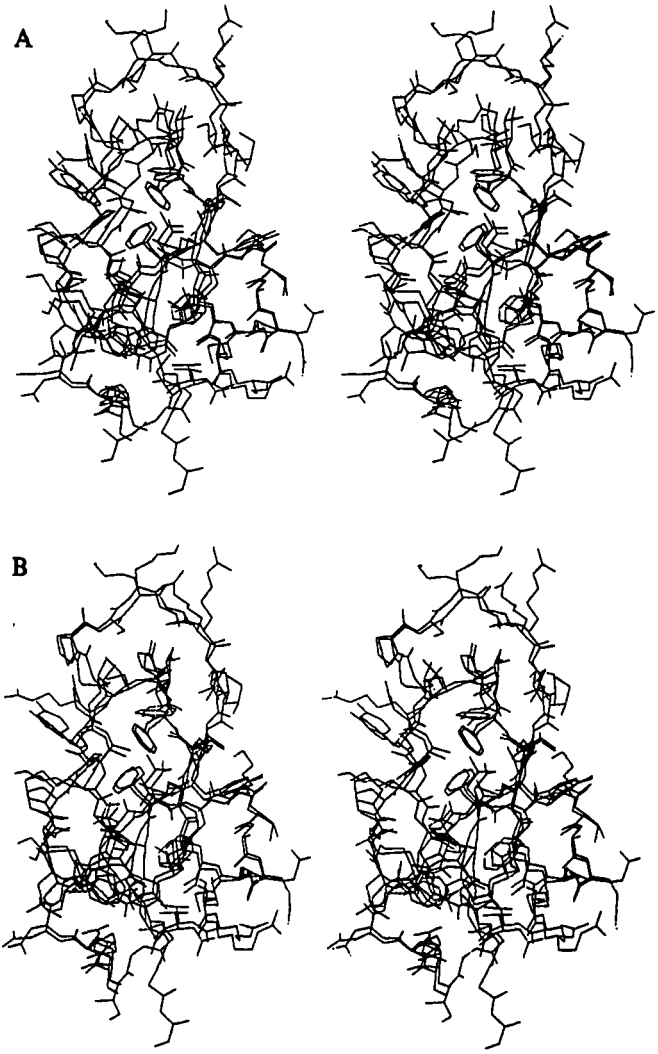

Fig. 4. Stereo view of the heavy atom structures calculated by DISMAN with the distance constraints data set AL_5 (Fig. 4A) and AU_5 (Fig. 4B). In both cases all heavy atoms of the calculated structures were best fit to the heavy atom of the X-ray structure of BPTI. Starting conformation for both structures is the structure shown in Fig. 3.

structure is shown. Both structures were calculated starting from the initial conformation of Fig. 3. Also in this figure, the coincidence of the calculated structures with the BPTI X-ray structure is similar for both data sets. Remarkably well defined are the side-chain conformations of the interior side-chain, especially the orientation of the aromatic ring planes, even though the long-range distance constraints were chosen in both data sets with a rather loose upper bound of 5 Å. This indicates that part of this restriction is certainly due to the packing constraints in the interior of the globular protein (Richards, 1974).

For simulated data sets, the three methods have not yet been tested with exactly the same data sets. So it is too early to judge the sampling property of each method.

In the calculation with the experimental BPTI data, care was taken that the two programs DISGEO and DISMAN used exactly the same data sets. So the results obtained from that study give indications on the sampling property of the two programs.

4.2 *Experimental data sets*

Only work dealing with calculation of protein structures from experimental NMR data sets is included in the following list. Papers dealing with small polypeptide chains have been included as well if the work can be considered as an immediate precursor in the development of methods for the structure determination of proteins.

4.2.1 *Micelle-bound glucagon*

This study (Braun *et al.* 1981, 1983) of the polypeptide hormone glucagon bound to perdeuterated dodecylphosphocholine micelles (MB-glucagon) was the first systematic application of the combined use of distance geometry calculations and NOE data to determine the secondary and tertiary structure of a polypeptide. The uniform averaging model (see Section 2.1) proposed in this work was an attempt to include effects of the internal flexibility of the molecule. The partial success of a detailed atomic structure determination of this molecule from a set of semi-quantitative NOE measurements and the existence of a computer program to generate structures from the experimental NOE data stimulated improvements in recording NOESY spectra and motivated the collection of rather large distance constraints data sets of proteins (Arseniev *et al.* 1984; Zuiderweg *et al.* 1984; Williamson *et al.* 1985; Braun *et al.* 1986; Kline *et al.* 1986; Wagner *et al.* 1986).

In the second paper (Braun *et al.* 1983) a more detailed study containing a larger input constraint data set was used to calculate structures. Also a first step was taken towards a more realistic estimate of the upper-limit distance constraints by a combination of a rigid and flexible model. Calculations were done for 4 overlapping segments of 11 residues because NOE data in this system were only available between those residues which differ in their residue number by at most 4 residues and the memory restrictions of the metric matrix approach limited us to about that size of polypeptides.

In this paper we also introduced several parameters to evaluate the quality of the obtained structures. The variations of the structures were quantified by the r.m.s. D values of different atom representations. Important are the r.m.s. D values for the backbone structure (BB) and the restricted side-chain representation (SR) where only atoms of complete side-chains were included in the best-fit superposition for those residues with NOEs involving the peripheral hydrogen atoms. Both values for the four segments showed a good negative correlation with the number of distance constraints within the segments. Smaller r.m.s. D values prevail for those segments with a larger set distance constraints. The backbone r.m.s. D values for the best-defined segments were of the order of 1 Å, thus

indicating that the atom positions in these fragments are determined nearly within the limiting uncertainty expected from the thermal fluctuations of the polypeptide chain (Karplus & McCammon, 1981).

Besides being a study to develop methods for the determination of the tertiary structure of polypeptides and proteins, this study also bears on the comparison of the polypeptide conformation of glucagon in different environments. Especially interesting is the comparison of the structures adopted by glucagon in single crystals (Sasaki *et al.* 1975) and of MB-glucagon in a lipid–water interphase. The individual molecules of the trimer in the X-ray single crystal structure are mainly α-helical with hydrophobic contacts between the monomers. Comparing the individual segments of the MB-glucagon with the corresponding segments of the monomers in the X-ray structure showed an increasing similarity of both structures towards the carboxy terminal end with decreasing r.m.s. D values of 3·5 Å for segment 10–20, 2·1 Å for segment 17–27 and 1·6 Å for segment 19–29 for backbone atoms.

4.2.2 *Micelle-bound melittin*

Detailed information on the conformation of membrane-bound polypeptides is still not available by any technique. It is therefore of great interest to study suitable model systems to understand the effect of the polypeptide on the lipid organization and the conformation of the polypeptide chain in such an environment. This type of system, a polypeptide chain incorporated in a detergent micelle, is not in the reach of the X-ray diffraction method. Recently, ^{1}H-NMR studies of gramicidin A incorporated into sodium dodecyl-d_{25} sulphate micelles gave a detailed picture of the transmembrane ion channel (Arseniev *et al.* 1985).

The conformation of melittin bound to the same micelle system (Brown *et al.* 1982) as the MB glucagon was studied by using the same interpretation of semiquantitative NOE information as distance constraints as in Braun *et al.* (1981).

A new technical aspect in this study was the use of distance geometry calculations to make sequence-specific resonance assignments. Mellitin has in the peptide segments 16–24 six residues which could, with the experiments available at the time, not be sequence-specifically assigned: two isoleucines at sequence positions 17 and 20, two lysines at 21 and 23 and two arginines at 22 and 24. There are a total of eight possibilities for the sequence-specific assignments. Distance geometry calculations clearly showed that only one assignment of Ile17 and Ile20 is compatible with all the other NMR information and a three-dimensional stereochemically reasonable structure.

The new technique of sequence-specific resonance assignments (Wüthrich *et al.* 1982; Billeter *et al.* 1982; Wagner & Wüthrich, 1982*a*; van de Ven *et al.* 1984; Weber *et al.* 1985; Wüthrich, 1986) certainly will give an assignment of the H^{α}, H^{β} and NH protons based entirely on NMR evidence for a large majority of the amino acid residues. The combination of NMR assignments and stereochemical considerations from the calculated three-dimensional structure certainly will

improve the reliability of the assignments and might be especially useful for stereospecific assignments in proteins of the two β-methylene protons and the two methyl groups of Val and Leu. This type of calculation usually requires only considerations of short polypeptide segments and is therefore very cheap. The technique of first using only a subset of NMR information that is entirely based on NMR evidence and then using distance-geometry calculations for further assignments of resonances based on the ensemble of calculated structures, is not a circular argument, but rather a boot-strap technique. It is similar to the routinely used strategy in the X-ray structure analysis of proteins of calculating structure factors from a preliminary model, and combining them with the multiple isomorphous replacement phases to a new phase set which is then used in a calculation of a new electron density map. The refined structure is then obtained by fitting it to this new density map.

4.2.3 *Insectotoxin I_5A*

Spatial structure information on short insectotoxins of 35–36 amino acid residues is very scarce, so Arseniev *et al.* (1984) applied the above-described direct method to study the solution conformation of the short scorpion insectotoxin I_5A. The overall spatial structure of determination of I_5A was done with the same program previously applied to glucagon. The program was modified for a pseudo atomic representation of the molecule, where each of the amino acid residues is represented by two spherical pseudoatoms α for the backbone NH-C$^{\alpha}$H-CO fragment and β for the side chain. Van der Waals radii for these pseudo atoms were defined such that the spherical volumes of these groups of atoms coincide with the van der Waals volumes of the same atomic groups as given by Richards (1974).

In addition to NMR data, dense-packing considerations were included in the calculations. The average r.m.s. D values between the ensemble of 15 calculated structures, all of which are compatible with the NMR constraints, were 2·1 Å for all α and β pseudo atoms, indicating that the overall global fold was well defined by the data. The work does not give indications of how much the dense-packing considerations contributed to this result. Perfect packing of the interior side-chains, at least in the sense that they are comparable to the packing of highly refined X-ray structures, can be obtained in principle with experimental NMR data set alone, as is shown by the calculations with the simulated data sets AL_5 and AU_5 (see Fig. 4A and B).

4.2.4 *Lac repressor headpiece*

The Lac repressor headpiece of *E. coli* was the first protein for which tertiary structure determination by a combination of model building step and restrained molecular dynamics from NMR data were reported (Kaptein *et al.* 1985). First, three α helical regions were identified from typical short-range NOE patterns (Billeter *et al.* 1982 ; Zuiderweg *et al.* 1983; Wüthrich *et al.* 1984), then the spatial arrangement of the three α-helices was determined by a model building procedure

from inter-helical long range NOE's (Zuiderweg *et al.* 1984). The helical segments were thereby treated as rigid building blocks. This second model building step was done manually with both a mechanical model and a computer-graphics program (Billeter *et al.* 1985). The third refinement step consisted of a restrained energy minimization calculation and a restrained molecular dynamics calculation (Kaptein *et al.* 1985), where the potential energy function was modified by an additional term of the type (3.5.1). In the third step the sum of inter-proton distance-constraints violations dropped from 54·2 to 10·5 Å for a total number of 159 NOE distance constraints, giving a final average violation of 0·066 Å. This value is of the same order of magnitude as those obtained from two recent distance-geometry calculations based on experimental distance constraints: α-amylase inhibitor with a final average violation of 0·025 Å (Kline *et al.* 1986) and BPTI (Wagner *et al.* 1987) with values ranging between 0·007 and 0·04 Å for the best five structures calculated by DISMAN and DISGEO. The potential energy reduced orders of magnitude to a value (−1882 kJ/mol) reasonable for a protein of this size. The magnitude and type of the structural changes from the initial to final refinement step were not reported.

The results presented certainly show that protein structures with reasonable energy values and compatible with the NOE distance constraints can be obtained by this method. The physical significance of the fluctuations during the time course of restrained dynamical calculations is less clear, because the width of the pseudo potential (3.5.1), and therefore the thermal fluctuations related to it, were essentially set *ad hoc* to 1 Å. The value of the dynamical trajectories in recalculating the NOESY spectrum, which would be of great general interest, is thereby reduced. In addition, the physical time for which trajectories can be calculated in reasonable computer time (100 ps) is quite short on the NMR time scale. All dynamical processes occurring with a rate faster than 10^{-3}/s contribute to the averaged values measured by NMR experiments and cannot be time-resolved. The capability of the dynamics calculations of escaping local minima does not lead to structures that agree significantly better with the experimental data than those generated by comparable experimental data sets treated with distance-geometry calculations. A combination of distance geometry and free molecular dynamics calculations seems to be more promising in this respect.

4.2.5 *Proteinase inhibitor IIA*

The structure determination of proteinase inhibitor IIA from bull seminal plasma (BUSI) (57 amino acid residues) (Williamson *et al.* 1985) followed the same procedures as was outlined in the structure determination of MB-glucagon. A total of 202 distance constraints derived from NOESY cross-peaks were interpreted as follows. For the intraresidue and sequential connectivities between amide, C^{α} and C^{β} protons with at most three intervening torsion angles, the relative cross-peak heights in the NOESY spectra recorded with a mixing time of 80 ms were calibrated on the basis of known sequential distances (Billeter *et al.* 1982) to three different classes of upper-limit distance constraints of 2·5, 3·0 and 4·0 Å. For all

long-range connectivities between amide and/or C^{α} protons an upper limit of 4·0 Å and for those cross-peaks involving also side-chain protons an upper limit distance constraints of 5·0 Å was assumed. This distinction assumes a somewhat higher overall flexibility for the amino acid side-chains compared to the backbone segments.

Additional geometric constraints included the distance constraints to define three disulphide bonds. These disulphide constraints were inferred from a qualitative analysis of those NOE crosspeaks near in the covalent structure to potential S–S bridges and from a comparison to the disulphide linkages of other proteins homologous to BUSI. Hydrogen-bond patterns were established on the basis of observed NOE cross-peaks and slow amide proton exchange rates (Wagner & Wüthrich, 1982*b*) in those regular secondary structures which could be assigned by the pattern of typical NOE cross-peak for regular secondary structures (Wüthrich *et al.* 1984). The extreme values of the spin–spin coupling constants $^{3}J_{HN_{\alpha}}$ and $^{3}J_{\alpha\beta}$, i.e. the large ($\geqslant$ 8·0 Hz) and small ($\leqslant$ 5·5 Hz) values, were used for constraining the ϕ and χ^{1} torsion angles.

Calculations were done with the program DISGEO (Havel & Wüthrich, 1984) using two different distance constraints: a first set N, where only the NOE distance constraints were used, and a second set C, where in addition to N the above-mentioned additional constraints were included. Five structures were calculated from each set, starting from ten different initial distance sets for the substructures. The average for the r.m.s. *D* values were 1·9 Å for the N and 2·1 Å for the C data sets for the C^{α} atoms and 3·0 and 3·4 Å for all heavy atoms. The reason for this increase of the r.m.s. *D* values by going from a less to a more constraining data set is discussed in Williamson *et al.* (1985).

This result and that in fig. 6 in Williamson *et al.* (1985) clearly show that the global chain fold is well defined by the NMR data set. The fact that this chain fold and the chain fold of two homologous proteins, the porcine pancreatic secretary inhibitor (Bolognesi *et al.* 1982) and the third domain of the Japanese quail ovomucoid (Papamokos *et al.* 1982), for which X-ray structures were available, closely coincide, again confirm that the short-distance constraints from NMR data sets determine the overall tertiary dimension.

4.2.6 *DNA binding helix F of the cyclic AMP receptor protein of* E. coli

The determination of the solution conformation of the DNA binding helix F of the cyclic AMP receptor protein of *E. coli* with a combined use of ^{1}HNMR and restrained molecular dynamics (Clore *et al.* 1985) can be considered as a pilot study for the question if restrained molecular dynamics can determine the tertiary structure of polypeptides or proteins directly from the NMR data or is it only useful in the refinement procedure (Kaptein *et al.* 1985).

To test this approach, three different initial structures, an α-helix, a β strand and a 3_{10}-helix, were selected. The constraints obtained from an interpretation of the NMR spectra as discussed above exhibited a pattern typical for α-helices (Wüthrich *et al.* 1984). In all three cases, convergence to an α-helical structure was achieved with a r.m.s. *D* value less than 2 Å for the backbone atoms.

The CPU time needed to achieve this result on a CRAY 1 computer is rather high if corresponding calculations would be done with the other two described methods. The calculations with the simulated data set of crambin (Brünger *et al.* 1986) showed that at least small proteins should be in the reach of this approach in practice with the availability of a supercomputer.

4.2.7 *Metallothionein-2*

Metallothioneins are small metal- and cystein-rich proteins. Metal storage or heavy metal detoxification were proposed as physiological functions for these proteins (Kägi & Nordberg, 1979). The proteins from mammalian sources contain about 60 amino acid residues; among them are 20 cysteines. These cysteines can bind up to seven metal ions.

NMR evidence identified two metal clusters of three and four metal ions by ^{113}Cd–^{113}Cd decoupling experiments (Armitage & Otvos, 1982). Heteronuclear ^{1}H–^{113}Cd COSY experiments were used to correlate the chemical shift of distinct, individual ^{113}Cd ions with the ^{1}H chemical shift of metal-bound cysteinyl residues (Otvos *et al.* 1985; Live *et al.* 1985; Frey *et al.* 1985) of rabbit liver Cd_7^{2+}-metallothionein-2.

Preliminary calculations with the DISMAN program to determine the global polypeptide fold of this protein with the NMR data described below were promising (Braun *et al.* 1986). This study also showed that the existence of some regular secondary structural elements which could be identified by typical NOE patterns (Wüthrich *et al.* 1984) is not necessary for protein structure determination by NMR data. At the beginning of this study no X-ray structure of any metallothionein was available. At present a 2·3 Å-resolution X-ray structure of rat liver Cd, Zn metallothionein (Furey *et al.* 1986) can be used to study similarities and differences between single crystal and solution structures.

The NMR information was translated into distance constraints as input for the DISMAN program in the following way. The metal cystein connectivities, combined with the amino acid position of the sequence specific assigned cysteines (Neuhaus *et al.* 1985), provide a network of 54 distance constraints for the 3-metal and 76 distance constraints for the four-metal cluster. A tetrahedral arrangement of four sulphur atoms around each metal ion, with a cadmium–sulphur distance of 2·6 Å, was assumed. This assumption is based on several spectroscopic experiments (Vašak & Kägi, 1983). The seven Cd ions were thereby represented by 20 pseudo atoms covalently linked to the sulphur atom of each of 20 metal bound cysteines. Equivalent Cd reference points were then forced to coincide by distance constraints and in the final structures the real Cd ion positions were calculated as the average positions of the equivalent Cd reference points. NOESY spectra recorded with mixing times of 150 ms provided further information on additional short proton–proton contacts.

Large-scale calculations are currently under way with a more comprehensive set of ^{1}H–^{1}H distance constraints from NOESY spectra recorded with different mixing times and numerous torsion-angle constraints for ϕ and χ^1 by combined

evaluation of intraresidue, sequential NOE's and the vicinal coupling constants $^3J_{\alpha NH}$ and $^3J_{\alpha\beta}$ for rabbit and rat liver Cd_7 MT2. The extensive calculations also should give a larger ensemble of final structures to improve the statistical significance of the r.m.s. *D* values. The NMR structures obtained should give a firm basis for comparing them to the X-ray structure. It is known there are a quite large number of different metal-cysteine coordinations found by the two techniques (Furey *et al.* 1986; Wagner *et al.* 1985).

4.2.8 *α-Amylase inhibitor*

α-Amylase inhibitor from *Streptomyces tendae*, tendamistat binds tightly to and inhibits specifically mammalian α-amylases (Vertesy *et al.* 1984). The polypeptide chain consists of 74 amino acid residues, among them 4 Cys residues forming two disulphide bridges (Cysll–Cys27, Cys45–Cys73). An extensive set of constraints from NMR experiments could be assembled and used in the structure determination (Kline *et al.* 1986). These constraints include 401 distance constraints from nuclear Overhauser effects, 168 distance constraints from hydrogen bonds and disulphide bridges, and 50 torsion-angle constraints from measurements of spin–spin coupling constants. Calculations with the DISMAN program (Braun & Gō, 1985) provided four structures compatible with the NMR constraints. The initial structures were randomly chosen with no assumption on the existence of regular secondary structures.

Parameters to judge the quality of the obtained structures are the residual violations of the NMR constraints. The average violation of the distance constraints per constraint dropped from the range 10–15 Å for the initial structures to about 0·025 Å for the final structures, indicating that all the final structures were in agreement with the NMR constraints and were entirely determined by the NMR data and not by the initial conformations.

A second numeric quantity is the r.m.s. *D* value for the backbone atoms; averaged over all pairwise comparisons between the final structures; this was 1·6 Å for residues 6–73. This compares with 15·2 Å for the average value of the backbone atoms in the same residue region of the initial conformation.

Other more qualitative criteria are the dense packing in the interior of globular proteins and handedness properties as found to be typical for globular proteins (Richardson, 1981). One is the right-handed twist of the β-sheet structure, which could be checked by the signed distance map (Braun, 1983). In Fig. 5 of one of the four structures, a signed distance map is plotted, and the typical handedness as found in other antiparallel β structures is shown by the negative signs for the broad bands running perpendicular to the diagonal.

On the proposal of R. Huber (MPI Munich), the structural analysis of α-amylase inhibitor was done independently and in parallel by X-ray diffraction methods and NMR techniques. This should provide an objective test for the reliability of the NMR structure determination and to make sure that differences and similarities between NMR and X-ray structures are not biased by the exchange of structural information between the two groups using the two different methods. In the future

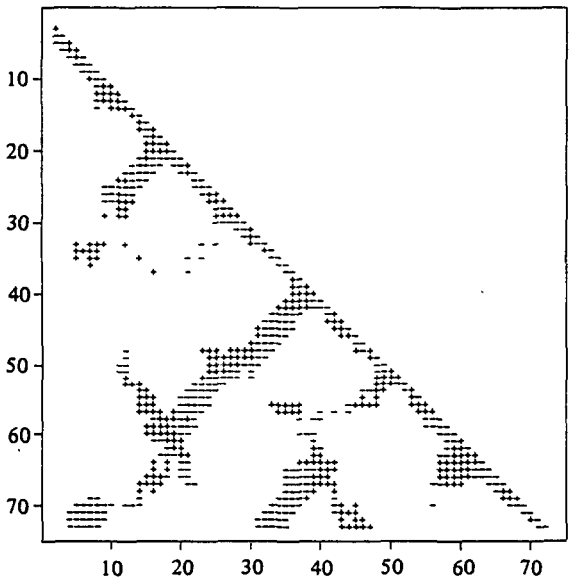

Fig. 5. Signed distance map (Braun, 1983) of one of the four NMR structures of α-amylase inhibitor Tendamistat (Kline *et al.* 1986). For each C^{α}–C^{α} atom pair with residue numbers i and j a plus or minus sign is plotted according to the handedness of the two tangent vectors to the polypeptide chain fold at residues i and j and the vector connecting residues i and j if the C^{α}–C^{α} distance is less than 10 Å.

a combination of both techniques and the optimal use of the structural results of either method might be used to speed up tertiary structure determination of proteins.

The result of the comparison of the global polypeptide fold of the NMR structures (Kline *et al.* 1986) and of the X-ray structure (Pflugrath *et al.* 1986) of α amylase inhibitor is shown in Fig. 6. The NMR structures (Fig. 6*A*) and the X-ray structure (Fig. 6B) closely coincide. The structures form a Greek-key β-barrel with the topology $+1, +3, -1, -1, +3$. The coincidence of the fold of the polypeptide chain as found by the two methods is especially close in those segments that are narrowly confined by NMR data. These segments include the six strands of the two β-sheets. Deviations are pronounced at less-well-defined regions such as the segment from residues 62 to 66. The side-chain conformations of the segment Trp18-Arg19-Tyr20, which are presumably involved in binding to the amylase, can be described as a sandwich structure found by both methods (see Figs. 7 and 6B). This is a remarkable result because the residues of this segment form part of the surface.

4.2.9 *Basic pancreatic trypsin inhibitor*

Over several years an extensive set of distance and torsion-angle constraints have been compiled for basic pancreatic trypsin inhibitor (BPTI), the favourite test protein for several experimental and theoretical studies which include the distance

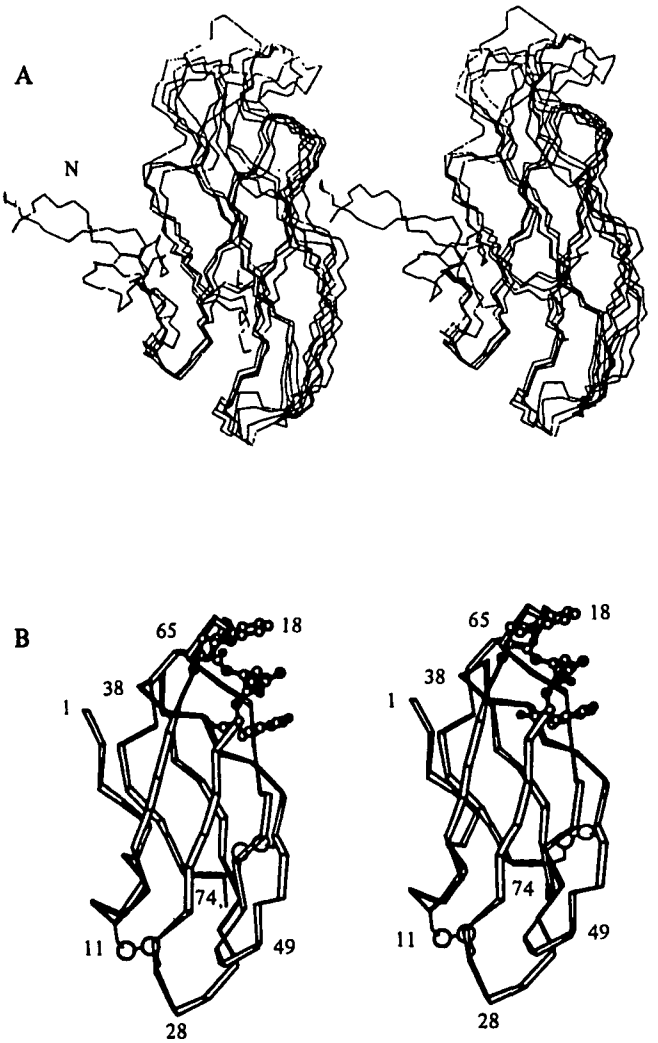

Fig. 6. Comparison of the polypeptide fold of α-amylase inhibitor Tendamistat as determined by NMR (Kline *et al.* 1986) and X-ray techniques (Pflugrath *et al.* 1986). (A) Stereo view of the backbone of four structures calculated by DISMAN from NMR data. The structures were superimposed to minimize the r.m.s. *D* value of the backbone atoms. The residue range 62–66 was eliminated from the best fit to better align the constrained areas (reproduced with permission from Kline *et al.* 1986). (B) Stereo ribbon diagrams of the C^{α} chain trace of the X-ray structure (Pflugrath *et al.* 1986). The triplet Trp18-Arg19-Tyr20 is shown in ball and stick form, where open circles are carbon atoms and filled circles are nitrogen and oxygen atoms. Both views have been chosen such that the 3 C^{α} positions of Ser 17, Lys 34 and Asn 25 in (A) and (B) have the same orientation.

geometry calculations with simulated NMR data sets (Havel & Wüthrich, 1984; 1985; Braun & Gō, 1985). BPTI also played a pilot role in the development of several of the methods described with the structure determinations of other proteins.

The experimental NMR data sets have now been used to study the performance

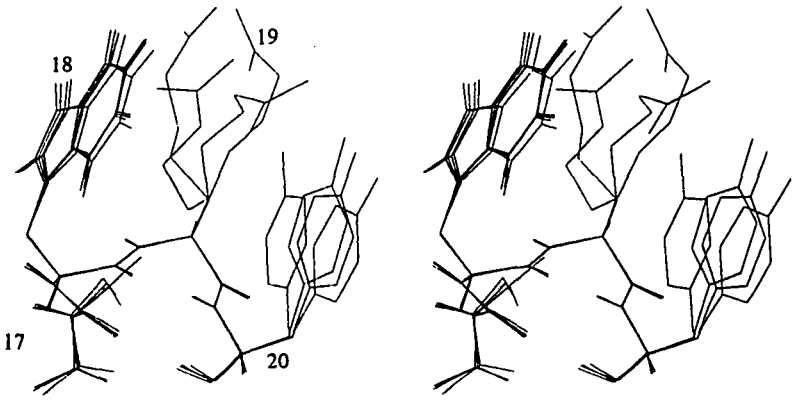

Fig. 7. Stereo view taken from the four NMR structures in Fig. 6(A), showing the complete residues Ser 17, Trp 18, Arg19 and Tyr20. The backbone atom were best fit to one structure. For the backbone and the indole ring all atoms are shown, for the rest of the structure the hydrogen atoms have been omitted.

of the two different programs, DISGEO and DISMAN, on exactly the same data set and to determine program-specific and data-specific influences on the calculated structures (Wagner *et al.* 1987). The NMR structures obtained by two different programs should also be a reliable basis for comparing the highly refined X-ray structure (Walter & Huber, 1983) in single crystals to the solution structure.

In Table 1 the five best structures according to the sum of final distance violations of both programs are compared for their final errors. It shows that the quality of convergence obtained with the two programs is nearly the same. No structures were found where all input constraints were exactly satisfied but the final distance- and torsion-angle violations are tolerable if we consider the accuracy of the estimations of these data. This also showed that there are no severe inconsistencies in the NMR data set and the problem of averaging over two very different states (see Section 2.1) is certainly not a major one in the direct determination of the tertiary protein structures.

In Table 2 the backbone and side-chain conformations of the NMR structures were compared to the X-ray structure. Both programs give quite similar values if one compares the solution to the X-ray structure, e.g. 2·4 and 2·3 Å for backbone atoms for DISMAN and DISGEO, respectively, and 2·7 Å for both programs if in addition the constrained side-chain atoms were included. Side-chains were considered as constrained when there was at least one distance constraint on a side-chain atom. In the case of the long residues Lys and Arg the constraint must be beyond the β proton. Differences of the behaviour of the two programs can be see by calculating the r.m.s. D values within the set of structures calculated by each program. The variability among all DISGEO structures is significantly smaller than among the DISMAN structures. This is particularly pronounced in the average r.m.s. D values of the backbone and non-constrained side-chains (4·2 versus 2·4 Å for DISMAN and DISGEO).

The study also again showed that the global polypeptide fold could be reliably

Table 1. *Statistics of the constraints violations for the BPTI solution conformations computed from the n.m.r. data using* DISMAN *and* DISGEO *(adapted from Wagner* et al. 1987)

Run	NMR distance constraints* Sum of violations (Å)	0·1–0·3	0·3–0·5	> 0·5	Non-bonded contacts† 0·1–0·3	0·3–0·5	> 0·5	Torsion angles‡ > 5°
DISMAN								
1	1·6	2	1	0	0	0	0	1
2	3·1	4	2	0	0	0	0	1
3	3·4	10	1	0	3	1	0	1
4	5·2	16	2	0	5	0	0	1
5	7·9	13	3	3 (1·3)	8	0	0	1
DISGEO								
1	2·7	7	0	0	2	0	0	0
2	2·7	7	0	0	4	0	0	0
3	4·9	10	4	0	5	0	0	1
4	9·0	19	3	2 (1·1)	5	0	0	1
5	9·4	12	5	6 (0·7)	13	0	0	2

* Number of violations of the NOE distance constraints, H-bond constraints, salt-bridge constraints and disulphide bridge constraints. The numbers in parentheses indicate the maximum distance violations.

† Bad non-bonded contacts were checked against the values of the repulsive core radii as given in table 1 of Braun & Gō (1985).

‡ Violations of constraints for the torsion angles ϕ and χ^1. The torsion-angle violation greater than 5° in the five DISMAN structures are all of the order of around 10° for the χ^1 torsion angle of GLU 49 and arise from a too-stringent salt-bridge constraint. In the DISGEO structures this too-stringent constraint is accommodated by distortions of the covalent structure.

Table 2. *Variations of side-chain conformations compared to variations of the polypeptide backbone conformations among the ten BPTI structures determined from n.m.r. data (adapted from Wagner* et al. 1987)

	Average r.m.s. D values and standard deviations in (Å)		
	Backbone*	Backbone and constrained side-chains†	Backbone and non constrained side chains‡
X-ray *v.* DISMAN	2·4 ± 0·5	2·7 ± 0·5	3·7 ± 0·5
X-ray *v.* DISGEO	2·3 ± 0·3	2·7 ± 0·2	3·5 ± 0·3
DISMAN *v.* DISGEO	2·9 ± 0·6	3·3 ± 0·4	4·1 ± 0·6
DISMAN *v.* DISMAN	2·9 ± 0·7	3·2 ± 0·5	4·2 ± 0·7
DISGEO *v.* DISGEO	1·6 ± 0·4	2·3 ± 0·3	2·4 ± 0·3

* The r.m.s. D values were calculated including all the backbone atoms N, C^{α} and C^{β} of the residues 1-58. Averages were taken over the five DISMAN or the five DISGEO structures, respectively.

† Backbone atoms of all residues and the heavy atoms of the constrained side-chains, i.e. P2, F4–E7, P9–T11, C14, A16, I18–A25, A27, L29–C30, T32–Y35, C38, A40, N43–F45, S47–M52, T54–C55. The residues Lys and Arg are only included in this list when distance constraints to the side-chain atoms were beyond the β-protons.

‡ Backbone atoms of all residues and the heavy atoms of the side-chains not listed under the footnote†.

determined by the NMR data whereas the local structures can still be improved significantly. The average standard deviation for backbone dihedral angles averaged over all residues is still quite high in both programs (around 60°). On well-defined segments the individual deviations are, however, in the range of 20–30°. This result is not a problem of convergence but a property of the NMR data sets.

5. SUMMARY

Computational tools have been developed in the last few years, to allow a direct determination of protein structures from NMR data. Numerical calculations with simulated and experimental NMR constraints for distances and torsional angles show that data sets available with present NMR techniques carry enough information to determine reliably the global fold of a small protein. The maximum size of a protein for which the direct method can be applied is not limited by the computational tools but rather by the resolution of the two-dimensional spectra. A general estimate of the maximum size would be a molecular weight of about 10000 (Markley *et al.* 1984), but parts of larger proteins might be accessible with the method.

Effort for improvement of the NMR structures should be concentrated more on the local conformation rather than the global features. The r.m.s. D values for

variations of the polypeptide backbone fold are on the order of 1·5–2 Å for several of the studied proteins, indicating that the global structure is well determined by the present NMR data and their interpretation. The local structures are sometimes rather poor, with standard deviations for the backbone torsion angles of about 50°. Possible improvements would be stereospecific resonance assignments of individual methylene protons and individual assignments of the methyl groups of the branched side-chains. Accurate estimates of the short-range NOE distance constraints by calibrating the distance constraints, including segmental flexibility effects, and combined use of distance geometry, energy minimization and molecular dynamics calculations, are further tools for improving the structures.

6. ACKNOWLEDGEMENTS

I would like to thank Prof. K. Wüthrich for fruitful discussions, Prof. W. Steinmetz for critical reading of the manuscript and Mrs E. H. Hunziker for help in the preparation of the illustrations.

7. REFERENCES

Abe, H., Braun, W., Noguti, T. & Gō, N. (1984). Rapid calculation of first and second derivatives of conformational energy with respect to dihedral angles for proteins. General recurrent equations. *Computers & Chemistry* **8**, 239–247.

Anil Kumar, Ernst, R. R. & Wüthrich, K. (1980). A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. *Biochem. biophys. Res. Commun.* **95**, 1–6.

Anil Kumar, Wagner, G., Ernst, R. R. & Wüthrich, K. (1981). Buildup rates of the nuclear Overhauser effect measured by two-dimensional proton magnetic resonance spectroscopy: implications for studies of protein conformation. *J. Am. chem. Soc.* **103**, 3654–3658.

Armitage, I. M. & Otvos, J. D. (1982). In *Biological Magnetic Resonance* (ed. L. J. Berliner and J. Reuben). New York: Plenum Press.

Arseniev, A. S., Barsukov, I. L., Bystrov, V. F., Lomize, A. L. & Ovchinnikov, Yu. A. (1985). ^{1}H-NMR study of gramicidin A transmembrane ion channel. Head-to-head right-handed, single-stranded helices. *FEBS Lett.* **186**, 168–174.

Arseniev, A. S., Kondakov, V. L., Maiorov, V. N. & Bystrov, V. F. (1984). NMR solution spatial structure of short scorpion insectoxin I_5A. *FEBS Lett.* **165**, 57–62.

Barry, C. D., North, A. C. T., Glasel, J. A., Williams, R. J. P. & Xavier, A. V. (1971). Quantitative determination of mononucleotide conformations in solution using lanthanide ion shift and broadening NMR probes. *Nature* **232**, 236–245.

Billeter, M., Braun, W. & Wüthrich, K. (1982). Sequential resonance assignments in protein ^{1}H nuclear magnetic resonance spectra. Computation of sterically allowed proton–proton distances in single crystal protein conformations. *J. molec. Biol.* **155**, 321–346.

Billeter, M., Engeli, M. & Wüthrich, K. (1985). Interactive program for investigation of protein structures based on ^{1}H NMR experiments. *J. molec. Graph.* **3**, 79–83.

BLUMENTHAL, L. M. (1970). *Theory and Applications of Distance Geometry*. New York: Chelsea.

BLUNDELL, T. L. & JOHNSON, L. N. (1976). *Protein Crystallography*, New York: Academic Press.

BOLOGNESI, M., GATTI, G., MENEGATTI, E., GUARNERI, M., MARQUART, M., PAPAMOKOS, E. & HUBER, R. (1982). Three-dimensional structure of the complex between pancreatic secretary trypsin inhibitor (Kazal-type) and trypsinogen at 1·8 Å resolution. Structure solution, crystallographic refinement and preliminary structural interpretation. *J. molec. Biol.* **162**, 839–868.

BOTHNER-BY, A. A. & JOHNER, P. E. (1978). Specificity of interproton nuclear Overhauser effects in gramicidin S dissolved in deuterated ethylene glycol. *Biophys. J.* **24**, 779–790.

BRAUN, W. (1983). Representation of short- and long-range handedness in protein structures by signed distance maps. *J. molec. Biol.* **163**, 613–621.

BRAUN, W., BÖSCH, C., BROWN, L. R., GŌ, N. & WÜTHRICH, K. (1981). Combined use of proton-proton Overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations. Application to micelle-bound glucagon. *Biochim. biophys. Acta* **667**, 377–396.

BRAUN, W. & GŌ, N. (1985). Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. *J. molec. Biol.* **186**, 611–626.

BRAUN, W., WAGNER, G., WÖRGÖTTER, E., VASAK, M., KÄGI, J. H. R. & WÜTHRICH, K. (1986). Polypeptide fold in the two metal clusters of Metallothionein-2 by nuclear magnetic resonance and distance geometry. *J. molec. Biol.* **187**, 125–129.

BRAUN, W., WIDER, G., LEE, K. H. & WÜTHRICH, K. (1983). Conformation of glucagon in a lipid–water interphase by ^{1}H nuclear magnetic resonance. *J. molec. Biol.* **169**, 921–948.

BRAUN, W., YOSHIOKI, S. & GŌ, N. (1984). Formulation of static and dynamic conformational analysis of biopolymers systems consisting of two or more molecules. *J. Phys. Soc. Japan* **53**, 3269–3275.

BROOKS, B. R., BRUCCOLERI, R. E., OLAFSON, B. D., STATES, D. J., SWAMINATHAN, S. & KARPLUS, M. (1983). CHARMM: a program for macromolecular energy, minimization and dynamics calculations *J. comput. Chem.* **4**, 187–217.

BROWN, L. R., BRAUN, W., ANIL KUMAR & WÜTHRICH, K. (1982). High resolution nuclear magnetic resonance studies of the conformation and orientation of melittin bound to a lipid-water interface. *Biophys. J.* **37**, 319–328.

BRÜNGER, A. T., CLORE, G. M., GRONENBORN, A. M. & KARPLUS, M. (1986).Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: Application to crambin. *Proc. natn. Acad. Sci. U.S.A.* **83**, 3801–3805.

BURGESS, W. A. & SCHERAGA, H. A. (1975). Assessment of some problems associated with prediction of the three-dimensional structure of a protein from its amino-acid sequence. *Proc. natn. Acad. Sci U.S.A.* **72**, 1221–1225.

CLORE, G. M. & GRONENBORN, A. M. (1985). Assessment of errors involved in the determination of interproton distance ratios and distances by means of one- and two-dimensional NOE measurements. *J. magn. Reson.* **61**, 158–164.

GLORE, G. M., GRONENBORN, A. M., BRÜNGER, A. T. & KARPLUS, M. (1985). Solution conformation of a heptadecapeptide comprising the DNA binding helix F of the cyclic AMP receptor protein of Escherichia coli. Combined use of ^{1}H nuclear magnetic resonance and restrained molecular dynamics. *J. molec. Biol.* **186**, 435–455.

CRIPPEN, G. M. (1977). A novel approach to the calculation of conformation: Distance Geometry. *J. comp. Phys.* **26**, 449–452.

CRIPPEN, G. M. (1981). Distance geometry and conformational calculations. In *Chemometrics Research Studies Series*, vol. 1 (ed. D. Bawden). New York: Research Studies Press.

CRIPPEN, G. M. & HAVEL, T. F. (1978). Stable calculations of coordinates from distance information. *Acta Crystallogr.* A **34**, 282–284.

CRIPPEN, G. M., OPPENHEIMER, N. & CONOLLY, M. (1981). Distance geometry analysis of the N.M.R. evidence on the solution conformation of bleomycin. *Int. J. Peptide Protein Res.* **17**, 156–169.

DE MARCO, A., LLINAS, M. & WÜTHRICH, K. (1978*a*). Analysis of the ^{1}H-NMR spectra of ferrichrome peptides. I. The non-amide protons. *Biopolymers* **17**, 617–636.

DE MARCO, A., LLINAS, M. & WÜTHRICH, K. (1978*b*). Analysis of the ^{1}H-NMR spectra of ferrichrome peptides. II. The amide resonances. *Biopolymers* **17**, 637–650.

DIAMOND, R. (1974). Real-space refinement of the structure of hen egg-white lysozyme. *J. molec. Biol.* **82**, 371–391.

DUBS, A., WAGNER, G. & WÜTHRICH, K. (1979). Individual assignments of amide proton resonances in the proton NMR spectrum of the basic pancreatic trypsin inhibitor. *Biochim. biophys. Acta* **577**, 177–194.

FLETCHER, R. (1980). *Practical Methods of Optimization: Unconstrained Optimization.* New York: Wiley.

FREY, M. H., WAGNER, G., VASAK, M., SORENSEN, O. W., NEUHAUS, D., WÖRGÖTTER, E., KÄGI, J. H. R., ERNST, R. R., WÜTHRICH, K. (1985). Polypeptide-metal cluster connectivities in metallothionein 2 by novel ^{1}H–^{113}Cd heteronuclear two-dimensional NMR experiments. *J. Am. chem. Soc.* **107**, 6847–6851.

FUREY, W. F., ROBBINS, A. H., CLANCY, L. L., WINGE, D. R., WANG, B. C. & STOUT, C. D. (1986). Crystal structure of Cd, Zn metallothionein. *Science* **231**, 704–710.

GŌ, N., NOGUTI, T. & NISHIKAWA, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. natn. Acad. Sci. U.S.A.* **80**, 3696–3700.

GŌ, N. & SCHERAGA, H. A. (1978). Calculation of the conformation of cyclohexaglycyl. 2. Application of a Monte-Carlo method. *Macromolecules* **11**, 552–559.

HAVEL, T. F., CRIPPEN, G. M. & KUNTZ, I. D. (1979). Effects of distance constraints on macromolecular conformation. II Simulation of experimental results and theoretical predictions. *Biopolymers* **18**, 73–81.

HAVEL, T. F., KUNTZ, I. W. & CRIPPEN, G. M. (1983). The theory and practice of distance geometry. *Bull. math. Biol.* **45**, 665–720.

HAVEL, T. F. & WÜTHRICH, K. (1984). A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular ^{1}H–^{1}H proximities in solution. *Bull. math. Biol.* **46**, 673–698.

HAVEL, T. F. & WÜTHRICH, K. (1985). An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. *J. molec. Biol.* **182**, 281–294.

JARDETZTKY, O., LANE, A., LEFEVRE, J.-F., LICHTARGE, O., HAYES-ROTH, B. & BUCHANAN, B. (1986). In *NMR in the Life Sciences*, New York: Plenum Press.

JARDETZKY, O. & ROBERTS, G. C. K. (1981). *NMR in Molecular Biology*, New York: Academic Press.

JEENER, J., MEIER, B. H., BACHMANN, P. & ERNST, R. R. (1979). Investigation of

exchange processes by two-dimensional NMR spectroscopy. *J. chem. Phys.* **71**, 4546–4553.

JONES, C. R., SIKAKANA, C. T., HEHIR, S., KUO, M. & GIBBONS, W. A. (1978). The quantitation of nuclear Overhauser effect methods for total conformational analysis of peptides in solution. Application to gramicidin S. *Biophys. J.* **24**, 815–832.

KÄGI, J. H. R. & NORDBERG, M. (1979). *Metallothionein*, Basel: Birkhäuser-Verlag.

KAPTEIN, R., ZUIDERWEG, E. R. P., SCHEEK, R. M., BOELENS, R. & VAN GUNSTEREN, W. F. (1985). A protein structure from nuclear magnetic resonance data. Lac repressor headpiece. *J. molec. Biol.* **182**, 179–182.

KARPLUS, M. (1959). Contact electron-spin coupling of nuclear magnetic moments. *J. Chem. Phys.* **30**, 11–15.

KARPLUS, M. (1963). Vicinal proton coupling in nuclear magnetic resonance. *J. Am. chem. Soc.* **85**, 2870–2871.

KARPLUS, M. & MCCAMMON, J. A. (1981). The internal dynamics of globular proteins. *C.R.C. Crit. Rev. Biochem.* **9**, 293–349.

KEEPERS, J. W. & JAMES, T. L. (1984). A theoretical study of distance determinations from NMR. Two-dimensional nuclear Overhauser effect spectra. *J. magn. Reson.* **57**, 404–426.

KLINE, A. D., BRAUN, W. & WÜTHRICH, K. (1986). Studies by ^{1}H nuclear magnetic resonance and distance geometry of the solution conformation of tendamistat an α-amylase inhibitor. *J. molec. Biol.* **189**, 377–382.

KOBAYASHI, Y., OHKUBO, T., KYOGOKU, Y., NISHIUCHI, Y., SAKAK-IBARA, S., BRAUN, W. & GŌ, N. (1985). Conformational analysis of conotoxin and its analogue by NMR measurements and distance geometry algorithm. *Proc 9th Am. Peptide Symp.* (ed. K. D. Kopple and C. M. Deber). Pierce Chem. Comp., Rockford (in the Press.)

KRISHNA, N. R., AGRESTI, D. G., GLICKSON, J. D. & WALTER, R. (1978). Solution conformation of peptides by the intramolecular nuclear Overhauser effect experiment. Study of Valinomycin-K. *Biophys. J.* **24**, 791–814.

LEACH, S. J., NÉMETHY, G. & SCHERAGA, H. A. (1977). Use of proton nuclear Overhauser effects for the determination of the conformations of amino acid residues in oligopeptides. *Biochem. biophys. Res. Commun.* **75**, 207–215.

LEVITT, M. (1982). Protein conformation, dynamics, and folding by computer simulation. *Ann. Rev. Biophys. Bioeng.* **11**, 251–271.

LEVITT, M. (1983). Protein folding by restrained energy minimization and molecular dynamics. *J. molec. Biol.* **170**, 723–764.

LEVITT, M., SANDER, C. & STERN, P. S. (1985). Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. molec. Biol.* **181**, 423–447.

LIVE, D., ARMITAGE, I. M., DALGARNO, D. C. & COWBURN, D. (1985). Two-dimensional ^{1}H–^{113}Cd chemical shift correlation maps by ^{1}H-detected multiple quantum NMR in metal complexes and metalloproteins. *J. Am. chem. Soc.* **107**, 1775–1777.

MACURA, S. & ERNST, R. R. (1980). Elucidation of cross relaxation in liquids by two-dimensional N.M.R. spectroscopy. *Mol. Phys.* **41**, 95–117.

MARION, D., GENEST, M., CAILLE, A., PEYPOUX, F., MICHEL, G. & PTAK, M. (1986). Conformational study of bacterial lipopeptides: refinement of the structure of iturin A in solution by two-dimensional ^{1}H-NMR and energy calculations. *Biopolymers* **25**, 153–170.

MARKLEY, J. L., WESTLER, W. M., CHAN, T.-M., KOJIRO, C. L. & ULRICH, E. L. (1984). Two-dimensional NMR approaches to the study of protein structure and function. *Proc. 74th Ann. Meeting Am. Soc. Biol. Chem., San Francisco* **43**, 2648–2656.

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

Meirovitch, H. & Scheraga, H. A. (1981). Introduction of short-range restrictions in a protein-folding algorithm involving a long-range geometrical restriction and short-, medium- and long-range interactions. *Proc. natn. Acad. Sci. U.S.A.* **78**, 6584–6587.

Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975). Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J. phys. Chem.* **79**, 2361–2381.

Nagayama, K. & Wüthrich, K. (1981). Structural interpretation of vicinal proton–proton coupling constants $^{3}J_{H^{\alpha}H^{\beta}}$ in the basic pancreatic trypsin inhibitor measured by two-dimensional J-resolved NMR spectroscopy. *Eur. J. Biochem.* **115**, 653–657.

Nemethy, G. & Scheraga, H. A. (1977). Protein folding. *Q. Rev. Biophys.* **10**, 239–352.

Neuhaus, D., Wagner, G., Vašak, M., Kägi, J. H. R. & Wüthrich, K. (1985). Systematic application of high-resolution, phase-sensitive two-dimensional ^{1}H-NMR techniques for the identification of the amino-acid–proton spin systems in proteins. *Eur. J. Biochem.* **151**, 257–273.

Noggle, J. H. & Schirmer, R. E. (1971). *The Nuclear Overhauser Effect*. New York: Academic Press.

Noguti, T. & Gō, N. (1983). A method of rapid calculation of a second derivative matrix of conformational energy for large molecules. *J. Phys. Soc. (Japan)* **52**, 3685–3690.

Ohkubo, T., Kobayashi, Y., Shimonishi, Y., Kyogoku, Y., Braun, W. & Gō, N. (1986). A conformational study of polypeptides in solution by ^{1}H-NMR and distance geometry. *Biopolymers* **25** (S), 123–134.

Olejniczak, E. T., Dobson, C. M., Karplus, M. & Levy, R. M. (1984). Motional averaging of proton nuclear Overhauser effects in proteins. Predictions from a molecular dynamics simulation of lysozyme. *J. Am. Chem. Soc.* **106**, 1923–1930.

Olejniczak, E. T., Gampe, R. T. & Fesik, S. W. (1986). Accounting for spin diffusion in the analysis of 2D NOE data. *J. magn. Reson.* **67**, 28–41.

Ooi, T., Nishikawa, K., Oobatake, M. & Scheraga, H. A. (1978). Flexibility of bovine pancreatic trypsin inhibitor. *Biochim. biophys. Acta* **536**, 390–405.

Otvos, J. D., Engeseth, H. R. & Wehrli, S. (1985). Multiple-quantum ^{113}Cd–^{1}H correlation spectroscopy as a probe of metal coordination environments in metalloproteins *J. magn. Reson.* **61**, 579–584.

Papamokos, E., Weber, E., Bode, W., Huber, R., Empie, M. W., Kato, I. & Laskowski, M. (1982). Crystallographic refinement of Japanese quail ovomucoid, a Kazal-type inhibitor, and model building studies of complexes with serine proteases. *J. molec. Biol.* **158**, 515–537.

Pardi, A., Billeter, M. & Wüthrich, K. (1984). Calibration of the angular dependence of the amide proton–C^{α} proton coupling constants, $^{3}J_{HN\alpha}$, in a globular protein. *J. molec. Biol.* **180**, 741–751.

Pflugrath, J. W., Wiegand, G., Huber & Vértesy, L. (1986). Crystal structure determination, refinement and the molecular model of the α-amylase inhibitor Hoe-467A. *J. molec. Biol.* **189**, 383–386.

Pottle, C., Pottle, M. S., Tuttle, R. W., Kinch, R. J. & Scheraga, H. A. (1980). Conformational analysis of proteins: algorithms and data structures for array processing. *J. comp. Chem.* **1**, 46–58.

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

RICHARDS, F. M. (1974). The interpretation of protein structures: total volume, group volume distribution and packing density. *J. molec. Biol.* **82**, 1–14.

RICHARDSON, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.* **34**, 167–335.

SASAKI, K., DOCKERILL, S., ADAMIAK, D. A., TICKLE, I. J. & BLUNDELL, T. (1975). X-ray analysis of glucagon and its relationship to receptor binding. *Nature* (*Lond.*) **257**, 751–757.

SAXE, J. B. (1979). Embeddability of weighted graphs in k-space is strongly NP-hard. *Proc. 17th Allerton Conf. Communication, Control and Computing*, pp. 480–489.

SCHLITTER, J. (1986). Calculation of coordinates from incomplete and incorrect distance data. *J. appl. Math. Phys.* (in the Press).

SCHMIDT, P. G. & KUNTZ, I. D. (1984). Distance measurements in spin-labeled lysozyme. *Biochemistry* **23**, 4261–4266.

SENN, H., BILLETER, M. & WÜTHRICH, K. (1984). The spatial structure of the axially bound methionine in solution conformations of horse ferrocytochrome c and *Pseudomonas aeruginosa* ferrocytochrome c 551 by ^{1}H NMR. *Eur. biophys. J.* **11**, 3–15.

SHELDRICK, G. M. (1985). Computing aspects of crystal structure determination. *J. molec. Struct.* **130**, 9–16.

SIPPL, M. J. & SCHERAGA, H. A. (1985). Solution of the embedding problem and decomposition of symmetric matrices. *Proc. natn. Acad. Sci. U.S.A.* **82**, 2197–2201.

SIPPL, M. J. & SCHERAGA, H. A. (1986). Caley-Menger coordinates. *Proc. natn. Acad. Sci. U.S.A.* **83**, 2283–2287.

SMITH, G. M. & VEBER, D. F. (1986). Computer-aided, systematic search of peptide conformations constrained by NMR data. *Biochem. Biophys. Res. Comm.* **134**, 907–914.

VAN DE VEN, F. J. M., DE BRUIN, S. H. & HILBERS, C. W. (1984). Two-dimensional Fourier transform ^{1}H NMR studies of ribosomal protein E-L30. *FEBS Letters* **169**, 107–111.

VAN GUNSTEREN, W. F. & BERENDSEN, H. J. C. (1982). Molecular dynamics: perspective for complex systems. *Biochem. Soc. Trans.* **10**, 301–305.

VAŠAK, M. & KÄGI, J. H. R. (1983). In *Metal Ions in Biological Systems* (Sigel, H. ed.) Marcel Dekker, New York.

VERTESY, L., OEDING, V., BENDER, R., ZEPF, K. & NESEMANN, G. (1984). Tendamistat (HOE 467), a tight-binding α-amylase inhibitor from Streptomyces tendae 4158. *Eur. J. Biochem.* **141**, 505–512.

WAGNER, G., BRAUN, W., HAVEL, T. F., SCHAUMANN, T., GŌ, N. & WÜTHRICH, K. (1987). Protein structures in solution by nuclear magnetic resonance and distance geometry: the polypeptide fold of the basic pancreatic trypsin inhibitor determined using two different algorithms, DISGEO and DISMAN. (Submitted.)

WAGNER, G., FREY, M. H., NEUHAUS, D., WÖRGÖTTER, E., BRAUN, W., VAŠAK, M., KÄGI, J. H. R. & WÜTHRICH, K. (1985). In *Proc. 2nd Int. meeting on Metallothionein*, Birkhäuser Verlag, Basel.

WAGNER, G. & WÜTHRICH, K. (1979). Truncated driven nuclear Overhauser effect (TOE). A new technique for studies of selective ^{1}H–^{1}H Overhauser effects in the presence of spin diffusion. *J. magn. Reson.* **33**, 675–680.

WAGNER, G. & WÜTHRICH, K. (1982*a*). Sequential resonance assignments in protein ^{1}H NMR spectra: basic pancreatic trypsin inhibitor. *J. molec. Biol.* **155**, 347–366.

WAGNER, G. & WÜTHRICH, K. (1982*b*). Amide proton exchange and surface conformation

of the basic pancreatic trypsin inhibitor in solution. Studies with two-dimensional nuclear magnetic resonance. *J. molec. Biol.* **160**, 343–361.

WAKO, H. & SCHERAGA, H. A. (1981). On the use of distance constraints to fold a protein. *Macromolecules* **14**, 961–969.

WALTER, J. & HUBER, R. (1983). Pancreatic trypsin inhibitor. A new crystal form and its analysis. *J. molec. biol.* **167**, 911–917.

WEBER, P. L., WEMMER, D. E. & REID, B. R. (1985). ^{1}H NMR studies of the λ Cro repressor. 2. Sequential resonance assignments of the ^{1}H NMR spectrum. *Biochemistry* **24**, 4553–4562.

WILLIAMSON, M. P., HAVEL, T. F. & WÜTHRICH, K. (1985). Solution conformation of proteinase inhibitor IIA from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry. *J. molec. Biol.* **182**, 295–315.

WÜTHRICH, K. (1986). *NMR of Proteins and Nucleid Acids*. New York: Wiley.

WÜTHRICH, K., BILLETER, M. & BRAUN, W. (1983). Pseudo-structures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton–proton distance constraints with nuclear magnetic resonance. *J. molec. Biol.* **169**, 949–961.

WÜTHRICH, K., BILLETER, M. & BRAUN, W. (1984). Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton–proton distances. *J. molec. Biol.* **180**, 715–740.

WÜTHRICH, K., WIDER, G., WAGNER, G., BRAUN, W. (1982). Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *J. molec. Biol.* **155**, 311–319.

ZUIDERWEG, E. R. P., BILLETER, M., BOELENS, R., SCHEEK, R. M., WÜTHRICH, K. & KAPTEIN, R. (1984). Spatial arrangement of the three α helices in the solution conformation of *E. coli* lac repressor DNA-binding domain. *FEBS Lett.* **174**, 243–247.

ZUIDERWEG, E. R. P., KAPTEIN, R. & WÜTHRICH, K. (1983). Secondary structure of the *lac* repressor DNA-binding domain by two-dimensional ^{1}H nuclear magnetic resonance in solution. *Proc. natn. Acad. Sci. U.S.A.* **80**, 5837–5841.