Journal of Biomolecular NMR, 2 (1992) 535–543
ESCOM

J-Bio NMR 090

A method for determining overall protein fold from NMR distance restraints

Jeffrey C. Hoch and Alan S. Stern

The Rowland Institute for Science, 100 Cambridge Parkway, Cambridge, MA 02142, U.S.A.

Received 1 May 1992
Accepted 14 September 1992

Keywords: Protein fold; Distance restraint; Molecular dynamics; Distance geometry; Reduced representation; Structure determination

SUMMARY

We describe a simple method for determining the overall fold of a polypeptide chain from NOE-derived distance restraints. The method uses a reduced representation consisting of two particles per residue, and a force field containing pseudo-bond and pseudo-angle terms, an 'electrostatic' term, but no van der Waals or hard shell repulsive terms. The method is fast and robust, requiring relatively few distance restraints to approximate the correct fold, and the correct mirror image is readily determined. The method is easily implemented using commercially available molecular modeling software.

INTRODUCTION

The determination of small protein structures from NOE-derived distance restraints (possibly augmented by torsion angle restraints) is now almost routine. A number of powerful computational methods have been developed to solve this problem (for a review, see James and Basus, 1991). Indeed, the rate-determining step in structural studies of proteins in solution is usually the assignment of spectral resonances to specific atoms, not the determination of the three-dimensional (3D) structure from distance restraints. There remains, however, a powerful imperative for the development of improved methods for deriving 3D structures from distance restraints. In particular, error analysis requires repeated, independent realizations of the structure, with the spread of the resulting ensemble of structures providing an indication of precision (Hoch, 1991). Faster methods will permit more independent realizations, increasing confidence in the structure and in the precision estimate.

Abbreviations: NOE, Nuclear Overhauser Effect; r.m.s.d., root-mean-square deviation; R_g, radius of gyration; RMD, restrained molecular dynamics; DG, distance geometry.

The robustness of an algorithm is just as important as the speed with which it can be executed. An algorithm that converges to a valid solution every time is just as good as an algorithm that is twice as fast, but converges only half the time. Another characteristic of a robust algorithm is that it degrades gracefully as the problem becomes increasingly underdetermined, for example, when fewer distance restraints are available.

Reduced representations of protein structure have been used extensively to lessen the computational burden of modeling proteins. An early example is the work of Levitt and Warshel, in which they simulated aspects of protein folding (Levitt and Warshel, 1975). Most existing methods for computing 3D structures from experimental distance restraints use a hierarchical approach, in which a reduced representation is used to obtain a coarse structure that is subjected to further refinement using a detailed representation (Havel and Wüthrich, 1984; Clore et al., 1987; Nilges et al., 1988). The advantage of using a reduced representation is that the requisite computations need only be applied to a small number of particles, thereby simplifying and speeding up the calculations, and yielding a method that is more robust.

We describe here a drastically simplified representation of protein structure, using only two particles per residue, and a corresponding force field. The backbone fold can be determined in this representation by restrained molecular dynamics or by distance geometry techniques. We present results of the method as applied to four proteins.

METHODS

The basis of the method is a two particle per residue representation of protein structure. One particle stands for the main chain, and can be identified with the C^{α} atom; the other stands for the orientation of the side chain with respect to the main chain and can be identified with the C^{β} atom. (The second particle is present even in glycine residues, which have no C^{β} atom.) The potential energy is given by:

$$E = \sum_{\text{bonds}} k_b(r-r_0)^2 + \sum_{\text{angles}} k_\theta(\theta-\theta_0)^2 + \sum_{i \neq j} \frac{k_e q_i q_j}{4\pi\varepsilon r_{ij}}$$

$$+ \sum_{\substack{\text{restraints} \\ r_{ij}}} \begin{cases} k_r(r_{ij}-r_{ij}^{\text{lower}})^2, & r_{ij} \leq r_{ij}^{\text{lower}} \\ 0, & r_{ij}^{\text{lower}} \leqslant r_{ij} \leqslant r_{ij}^{\text{upper}} \\ k_r(r_{ij}-r_{ij}^{\text{upper}})^2, & r_{ij} r_{ij}^{\text{upper}}. \end{cases} \qquad (1)$$

The force field parameters are listed in Table 1. The C^{α}–C^{α} equilibrium distance is correct for trans peptides; no special provision has been made for cis prolines. The C^{α}–C^{α}–C^{α} angle term serves to keep the chain extended in the absence of other forces. The equilibrium value of 109° is an average; polypeptides display a range of values for this angle, depending on the torsion angles of the middle residue. The results are not sensitive to the exact value used. The electrostatic energy uses a 1/r dielectric and serves as a generalized repulsion to keep the structure from collapsing. The use of relatively small charges permits the particles to pass close to one another during high-temperature dynamics without becoming trapped. Levitt and Warshel (1975) previously pointed

TABLE I
TWO-PARTICLE FORCE FIELD PARAMETERS

Bonds	C^{α}–C^{α}	r_O = 3.7 Å	k_b = 105 kcal/mol/Å^2
	C^{α}–C^{β}	r_O = 1.54 Å	k_b = 105 kcal/mol/Å^2
Angles	C^{α}–C^{α}–C^{α}	θ_O = 109°	k_θ = 2.5 kcal/mol/rad^2
	C^{α}–C^{α}–C^{β}	θ_O = 109°	k_θ = 50 kcal/mol/rad^2
Electrostatic		Dielectric $\varepsilon = r_{ij}$ $q(C^{\alpha}) = q(C^{\beta})$ = 0.18 electron charge	k_e = 4173 Å-kcal/mol/e^2
NOE restraint			k_r = 50 kcal/mol/Å^2

out this advantage of soft repulsions. The value of the charge (which has no physical meaning) affects the radius of gyration, but does not affect the overall fold so much.

Distance restraints for the two-particle representation were derived from experimental distance restraints or known distances in the crystal structures. Restraints on the H^N and H^{α} protons were mapped to the C^{α} particle; all others were mapped to the C^{β} particle. Correction factors were added to the upper bound of the restraints, depending on the remoteness of the restrained hydrogen, according to Table 2. The correction factors were about half of the maximum possible distance from the proton in question to the corresponding carbon atom. Lower bounds for NOE-derived distance restraints were set to zero. In addition, where known disulfide bonds were present, they were included as distance restraints between the C^{β} particles of the appropriate residues, with a lower bound of 3.5 Å and an upper bound of 4.8 Å.

One method used for determining the fold in the reduced representation was restrained molecular dynamics optimization, in which the particles were initially distributed randomly within a volume several times larger than that of the expected final structure. The structure was then subjected to the following protocol:

(1) 50 steps of conjugate gradient minimization.
(2) 2500 steps of restrained molecular dynamics at a temperature of 5000 K, with a step size of 1 fs and temperature equilibration every 20 steps.
(3) 2500 steps of restrained molecular dynamics as before, with an initial temperature of 5000 K and a final temperature of 300 K.
(4) Conjugate gradient minimization until the average force was less than 1.5 kcal/mol/Å, typically 50 to 60 steps.

TABLE 2
CORRECTION FACTORS (IN Å) ADDED TO THE UPPER BOUND OF THE TWO-PARTICLE DISTANCE RESTRAINTS[a]

H^N	H^{α}	H^{β}	H^{γ}	H^{δ}	H^{ϵ}	H^{ζ}	H^{η}
1.10	0.55	0.55	1.15	1.75	2.50	2.95	2.95

[a] A correction was added for each of the two restrained protons.

TABLE 3
HOLONOMIC CONSTRAINTS (IN Å) FOR TWO-PARTICLE DISTANCE GEOMETRY

$1.35 \leqslant C_i^\alpha - C_i^\beta \leqslant 1.55$	$4.20 \leqslant C_i^\alpha - C_{i+2}^\alpha \leqslant 7.30$
$3.50 \leqslant C_i^\alpha - C_{i+1}^\alpha \leqslant 3.85$	$3.70 \leqslant C_i^\alpha - C_j^\alpha, \quad \lvert i-j \rvert > 2$
$3.50 \leqslant C_i^\alpha - C_{i+1}^\beta \leqslant 5.00$	$3.50 \leqslant C_i^\alpha - C_j^\beta, \quad \lvert i-j \rvert > 1$
$3.25 \leqslant C_i^\beta - C_{i+1}^\alpha \leqslant 5.00$	$3.20 \leqslant C_i^\beta - C_j^\beta, \quad i \neq j$

A different protocol, based on distance geometry, was also used. In this approach, distance bounds were computed from the NOE restraints, the bond constraints, and several holonomic constraints (see Table 3). Bound smoothing via triangle inequalities was performed. Trial distances were choosen with a cubic bias toward the upper bound, given by the formula

$$d_{trial} = d_{upper} - t^3 (d_{upper} - d_{lower}), \tag{2}$$

where t is a uniform random deviate between zero and one. Structures computed using embedding were subjected to energy minimization using the two-particle force field as in step 4 above.

The molecular dynamics and mechanics calculations were carried out using the Biograf/NMRgraf program from Molecular Simulations, Inc. However, the simple nature of the force field means that it can be easily implemented using most molecular modeling software packages. Distance geometry calculations were performed using our own program, which is based on the algorithms given in Crippen and Havel (1988).

RESULTS AND DISCUSSION

We present results for four proteins: hirudin (Folkers et al., 1989), BDS-I (Driscoll et al., 1989), crambin (Teeter, 1984), and iodine-inactivated lysozyme from hen egg-white (Beddell et al., 1975), ranging in size from 43 residues to 129 residues. The structures are in the Brookhaven Protein Data Bank (Bernstein et al., 1977; Abola et al., 1987): entries 5HIR, 1BDS, 1CRN, and 8LYZ, respectively. Experimentally derived distance restraints were used for hirudin and BDS-I. For crambin and lysozyme, 12.5% of the interproton distances less than 5 Å computed from the crystal structures were selected at random. Experimental distance restraints for BDS-I and hirudin were generously provided by G.M. Clore and A. Gronenborn, and are now available from the Brookhaven Protein Data Bank (entries 2BDS and 2HIR). Distance restraints derived from assigned hydrogen bonds and dihedral angle restraints were not used.

Table 4 summarizes the results using the protocols described above. Figures 1 and 2 illustrate C^α traces for representative ensembles consisting of ten structures obtained using restrained molecular dynamics and distance geometry protocols, respectively. Using the RMD protocol, the average r.m.s.d. from the published structure was less than 2.5 Å in each case, and for many individual structures the r.m.s.d. was less than 1.5 Å. Results obtained using the DG protocol were slightly worse. Considering the simplicity of the approach, the agreement was suprisingly good for both methods.

Although nearly all of the calculated structures lie fairly close to one another, in the lysozyme ensemble one of the structures is distinctly different (shown in blue in Fig. 1D). This outlier has

TABLE 4
RESULTS OF TWO-PARTICLE STRUCTURE CALCULATIONS

Protein	Hirudin	BDS-I	Crambin	Lysozyme
No. of residues	49	43	46	129
No. NOE restraints	700	482	328	593
No. two-particle restraints	197 + 3 S–S	118 + 3 S–S	143 + 3 S–S	391 + 4 S–S
Target R_g (Å)	9.4	9.3	10.0	13.8
Average R_g (Å)				
RMD protocol	9.6	9.2	9.4	12.8
DG protocol	9.4	9.1	8.9	12.5
Average r.m.s.d. from target (Å)				
RMD protocol	2.2	2.4	1.9	2.4 (2.1)[a]
DG protocol	3.0	2.9	3.6	4.1
Ensemble r.m.s.d. (Å)				
RMD protocol	2.0	2.2	1.7	2.6 (1.5)[a]
DG protocol	3.8	3.1	4.6	4.3
Average two-particle energy (kcal/mol)				
RMD protocol	278	213	256	875 (859)[a]
DG protocol	334	260	359	1181
CPU time (min)[b]				
RMD protocol	3.1	3.6	3.9	17
DG protocol	0.4	0.4	0.4	1.1

[a] Values in parentheses refer to the ensemble excluding one outlier.
[b] Using Biograf on an IBM RS/6000 Model 320H.

a different fold for part of the chain, but still satisfies all of the distance restraints. The two-particle energy, dominated by the electrostatic component, was significantly larger for this structure than for the other members of the ensemble. The values given in parentheses in Table 4 reflect the ensemble excluding this outlier.

We studied the robustness of the two-particle method by randomly eliminating distance restraints. Results for BDS-I (using the RMD protocol) are given in Table 5 and Fig. 3. While the average r.m.s.d. from the published structure degraded rapidly, it can be seen from Fig. 3D that the gross characteristics of the overall fold were still correctly represented when only 21 restraints were used.

The embedding step of the DG protocol yielded structures with numerous restraint violations. Conjugate gradient minimization sufficed to satisfy the restraints in most instances. Metrization alone (Havel, 1990) generated structures that were far too compact, probably due to the absence of excluded volume from side chain and other atoms not present in the two-particle representation. The biased distribution given by Eq. 2 generated structures that were more extended.

TABLE 5
RESULTS OF THE TWO-PARTICLE CALCULATIONS FOR BDS-I USING PARTIAL DISTANCE RESTRAINT SETS

Number of restraints	Average R_g (Å)	Average r.m.s.d. from target (Å)	Ensemble r.m.s.d. (Å)
100	9.6	2.7	1.9
76	9.8	3.6	1.9
40	10.4	4.4	2.8
21	10.8	6.1	3.4

The published structures do not correspond to local minima of the two-particle energy potential. However, they are close: 1000 steps of conjugate gradient minimization applied to the published structures resulted in r.m.s. shifts of 1.3 Å, 1.7 Å, 1.3 Å and 1.4 Å for the C^{α} positions of hirudin, BDS-I, crambin and lysozyme, respectively – differences that are smaller than the distributions of our two-particle structures. This fact suggests that the method does contain a systematic bias. This suggestion is supported by the results in Table 4 showing that in many cases the ensemble r.m.s.d. was smaller than the average r.m.s.d. from the target structure. In addition, inspection of Fig. 1 reveals several locations where the C^{α} traces of the two-particle ensembles deviate systematically from the target traces.

The differences between the published structures and the minima of the two-particle potential are not suprising, given the approximations involved. This is the reason the minimizations were

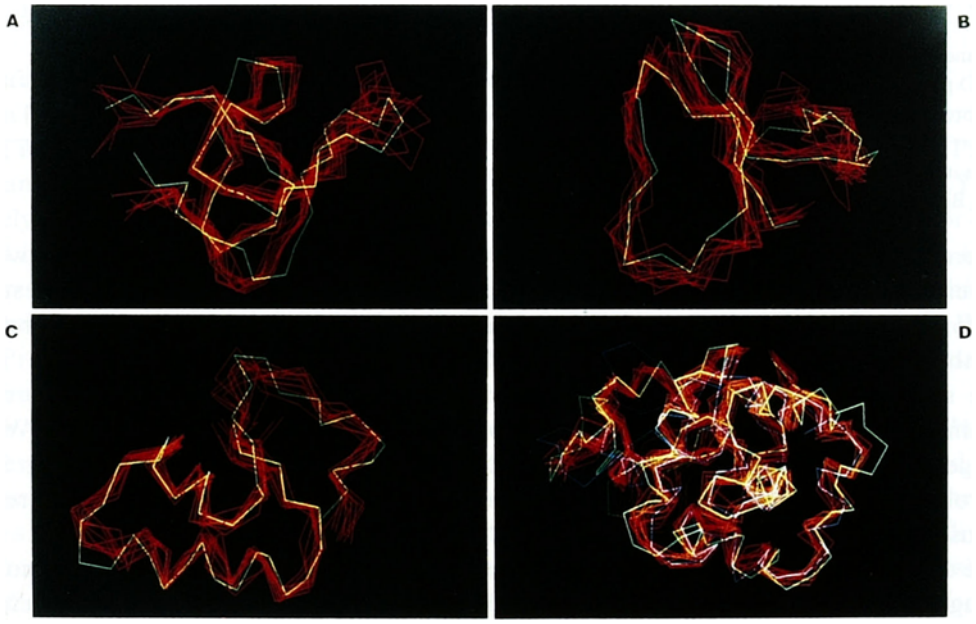

Fig. 1. C^{α} traces for representative families of two-particle structures computed using the restrained molecular dynamics protocol. (A) Hirudin. (B) BDS-I. (C) Crambin. (D) Lysozyme. The target structure is shown in green. In (D), a high-energy structure that satisfies the distance restraints is shown in blue.

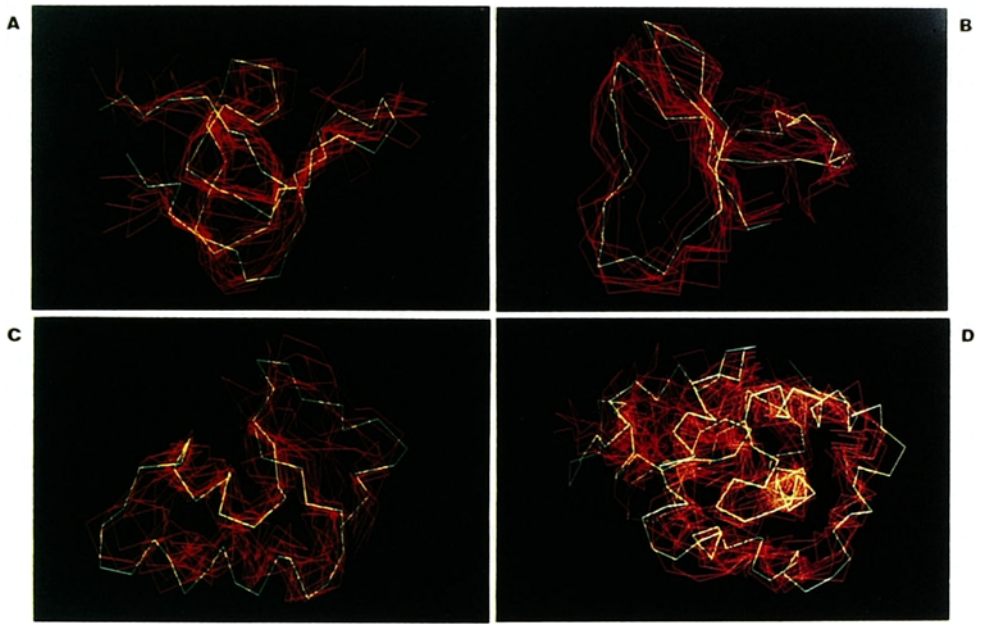

Fig. 2. C^{α} traces for families of structures computed using the distance geometry protocol. (A) Hirudin. (B) BDS-I. (C) Crambin. (D) Lysozyme. The target structure is shown in green.

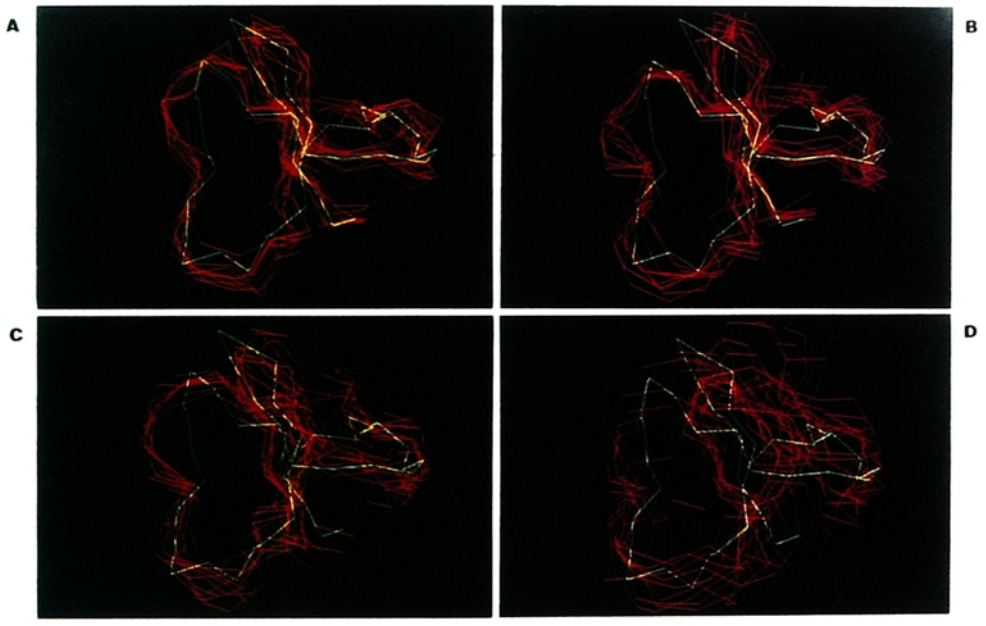

Fig. 3. C^{α} traces for BDS-I obtained using the restrained molecular dynamics protocol with partial restraint sets. (A) 75% of the available experimental restraints. (B) 50%. (C) 25%. (D) 12.5%.

stopped when the average force reached a cut-off level of 1.5 kcal/mol/Å. Regardless of whether one is optimizing a two-particle or an all-atom structure, there is no point in optimizing beyond the level of accuracy of the force field. To do so is over-fitting and will force the structure to mirror the inaccuracies in the force field. It also results in a narrower distribution of structures, with no corresponding improvement in accuracy. The value we chose for the minimization cut-off was just sufficient to insure that the distance restraints were satisfied.

The close-packed nature of the interiors of globular proteins means that accurate modeling of their 3D structures demands an all-atom representation. Transition from the two-particle representation to an all-atom representation is straightforward. For each residue, a structure is taken from an amino acid fragment library. The fragment is placed with its C^{α} atom coinciding with the C^{α} particle, is oriented so that its C^{β} atom lies in the same direction as the C^{β} particle, and is rotated so that its N and C atoms point as directly as possible toward the C^{α} particles of the preceding and following residues. This procedure is fast and provides a simple means of determining the correct mirror image of the overall fold. Indeed, for every instance in which we calculated an all-atom structure from both a two-particle structure and its mirror image, the correct form was always the one that yielded the lower r.m.s. violation of distance restraints on H^{α} protons.

Quantitative statements about the extent to which the two protocols – restrained molecular dynamics and distance geometry – sample the feasible conformational space would require knowledge of the total hypervolume of the feasible region. Without this knowledge, we are limited to qualitative comparative statements. For two methods that generate feasible structures, the one that gives rise to the broadest distribution of structures provides better sampling. A difficulty with applying even this conservative statement is that the definition of what constitutes a feasible structure is elusive. One possible characterization is that feasible structures must satisfy the distance restraints and have reasonable energies. The difficulty lies in deciding what constitutes a reasonable energy. This difficulty arises irrespective of whether one is using a reduced representation force field or an all-atom force field. Although the ensembles of structures we obtained using DG are more broadly distributed than those obtained using RMD (Table 4), they also have significantly higher two-particle energies, and consequently do not necessarily represent a better sample of feasible structures.

Given the approximations inherent to the two-particle force field, the distributions obtained (using either method) should not be construed as being indicative of the range of structures consistent with the experimental uncertainty. Short of a systematic search, we know of no method capable of proving that there is only one possible fold. However, we have demonstrated that the two-particle method reliably finds protein folds that are consistent with experimental restraints.

Possible future improvements in the force field include explicit provision for differences in the C^{α}–C^{α} pseudo-bond length for cis X–Pro bonds, more precise derivation of two-particle distance restraints from explicit atom restraints, use of torsion angle restraints, and variation of the pseudo-charge placed on the C^{β} particles to reflect the sizes of the corresponding side chains.

CONCLUSIONS

The two-particle representation of protein structure provides a simple method for determining the overall fold. The calculations require relatively little computer time and are highly reliable, converging to a correct fold even when few distance restraints are available. The method is easy

to implement with commercial molecular modeling software. It may prove useful during early stages of the assignment process, where a knowledge of the overall fold could facilitate subsequent assignments. In addition to its intended purpose, the method may also prove useful in other structure–prediction contexts, such as homology building.

ACKNOWLEDGMENTS

We are grateful to Polygen/Molecular Simulations Inc. for granting a software license for Biograf/NMRgraf. We thank Barry Olafson for useful discussions, and Marius Clore and Angela Gronenborn for providing the structures and experimental restraints for hirudin and BDS-I. Tim Havel provided valuable suggestions on distance geometry methodology. Finally, we would like to thank Jay Scarpetti and MaryAnn Nilsson for technical assistance with the manuscript.

REFERENCES

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. and Weng, J. (1987) In *Crystallographic Databases – Information Content, Software Systems, Scientific Applications* (Eds, Allen, F.H., Bergerhoff, G. and Sievers, R.), Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, pp. 107–132.

Beddell, C.R., Blake, C.C.F. and Oatley, S.J. (1975) *J. Mol. Biol.*, **97**, 643–654.

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

Clore, G.M., Sukumaran, D.K., Nilges, M. and Gronenborn, A.M. (1987) *Biochemistry*, **26**, 1732–1745.

Crippen, G.M. and Havel, T.F. (1988) *Distance Geometry and Molecular Conformation*, John Wiley and Sons; New York.

Driscoll, P.C., Gronenborn, A.M., Beress, L. and Clore, G.M. (1989) *Biochemistry*, **28**, 2188–2198.

Folkers, P.J.M., Clore, G.M., Driscoll, P.C., Dodt, J., Köhler, S. and Gronenborn, A.M. (1989) *Biochemistry*, **28**, 2601–2617.

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

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

Hoch, J.C. (1991) In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy* (Eds, Hoch, J.C., Poulsen, F.M. and Redfield, C.), Plenum Press, New York, pp. 253–267.

James, T.L. and Basus, V.J. (1991) *Ann. Rev. Phys. Chem.*, **42**, 501–542.

Levitt, M. and Warshel, A. (1975) *Nature*, **253**, 694–698.

Nilges, M., Gronenborn, A.M. and Clore, G.M. (1988) *FEBS Lett.*, **229**, 317–324.

Teeter, M.M. (1984) *Proc. Nat. Acad. Sci. USA*, **81**, 6014–6018.