Biochemical Pharmacology, Vol. 40, No. 1, pp. 15–22, 1990.
Printed in Great Britain.

0006-2952/90 $3.00 + 0.00

COMPUTATIONAL METHODS FOR DETERMINING PROTEIN STRUCTURES FROM NMR DATA

GARRY P. GIPPERT, PING F. YIP, PETER E. WRIGHT and DAVID A. CASE*
Department of Molecular Biology, Research Institute of Scripps Clinic, La Jolla, CA 92037, U.S.A.

Abstract—The general procedures by which solution structures of proteins may be deduced from distance and angular constraints derived from NMR are reviewed, with an emphasis on practical aspects of the calculations. In addition, novel methods based on chemical shift calculations and on quantitative fits to nuclear Overhauser effect intensities are presented; these should provide improved understanding of the limits of our ability to simulate complex spectra, and may permit higher precision structures to be determined.

The development over the past decade of high-field NMR spectrometers and novel two-dimensional techniques has made it possible in favorable cases to determine the three-dimensional (3D) structures of small proteins in solution. The general procedure is by now fairly well known [1–4]: distance constraints based on nuclear Overhauser effect (NOE) intensities and angular constraints based on coupling constants are used to first deduce and then refine a three-dimensional structure consistent both with general stereochemical constraints derived from the covalent structure and with the particular experimental data for the protein in question. The most popular method for deducing a three-dimensional structure that (approximately) satisfies the experimental constraints is based on a metric matrix or "distance geometry" approach [5]; other methods have been tried, and the entire field has been reviewed recently [6]. It is typical for these initial structures to be refined in 3D space by some combination of energy minimization and stimulated annealing, in an attempt to simultaneously minimize energy terms related to strain in the covalent structure and penalty terms based on experimental information [6, 7]. Finally, systematic comparisons of the spectra predicted by these structures with the experiment can be used to further refine the constraints and to gain insight into the motional and spin diffusion effects that influence spectral intensities [8–11]. Figure 1 gives a general overview of the flow of information in this process.

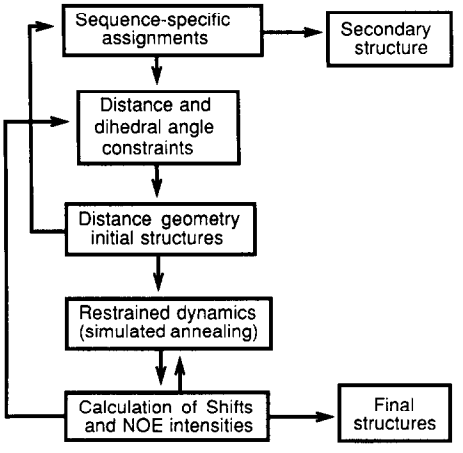

Fig. 1. Flow of information in a general refinement.

It should be emphasized that no generally accepted method for structure refinement exists, and that many algorithms are being tried in various laboratories. In this paper, we review the foundations of various approaches and give some details of the approach we have used for several proteins. Prospects for new approaches, particularly involving quantitative assessment of NOE intensities and analyses of chemical shifts, will be assessed.

Distance geometry methods

Basic formulas. If we consider describing a macromolecule in terms of the distances between atoms, it is clear that there are many constraints that these distances must satisfy, since for N atoms there are $N(N+1)/2$ distances but only $3N$ coordinates. General considerations for the conditions required to "embed" a set of interatomic distances into a realizable three-dimensional object form the subject of distance geometry [5, 12], and applications to macromolecular conformation pre-date modern NMR techniques for deriving large numbers of distance constraints [13]. In this section we outline, in a non-rigorous fashion, the fundamental mathematics of the procedure and describe some of our experiences in using it on proteins of various sizes.

The basic idea of distance geometry arises from consideration of the *metric matrix* that contains the scalar products of the vectors x_i that give the positions of the atoms:

$$g_{ij} \equiv x_i \cdot x_j \tag{1}$$

These matrix elements can be expressed in terms of the distances d_{ij}, d_{i0}, and d_{j0}:

$$g_{ij} = \tfrac{1}{2}(d_{i0}^2 + d_{j0}^2 - d_{ij}^2) \tag{2}$$

* Author to whom correspondence should be addressed.

If the origin ("0") is chosen at the centroid of the atoms, then it can be shown that distances from this point can be computed from the interatomic distances alone [5]:

$$d_{i0}^2 = \frac{1}{N}\sum_{j=1}^{N} d_{ij}^2 - \frac{1}{N^2}\sum_{k>j=1}^{N} d_{jk}^2 \qquad (3)$$

A fundamental theorem of distance geometry states that a set of distances can correspond to a three-dimensional object only if the metric matrix g is rank three, i.e. if it has three positive and N-3 zero eigenvalues. This is not a trivial theorem, but it may be made plausible by thinking of the eigenanalysis as a singular value decomposition: all of the distance properties of the molecule should be describable in terms of three "components," which would be the x, y and z coordinates. If we denote the eigenvector matrix as w and the eigenvalues λ_k, then the metric matrix can be written in two ways:

$$g_{ij} = \sum_{k=1}^{3} x_{ik}x_{jk} = \sum_{k=1}^{3} w_{ik}w_{jk}\lambda_k \qquad (4)$$

The first equality follows from the definition of the metric tensor, Eq. 1; the upper limit of three in the second summation reflects the fact that a rank three matrix has only three non-zero eigenvalues. Equating these two provides an expression for the coordinates x_{ik} (k = 1–3) in terms of the eigenvalues and eigenvectors of the metric matrix:

$$x_{ik} = \lambda_k^{1/2} w_{ik} \qquad (5)$$

This, then, is a prescription for determining coordinates from distances. If the distances are not exact, then in general the metric matrix will have more than three non-zero eigenvalues, but an approximate scheme can be made by using Eq. 5 with the three largest eigenvalues. Since information is lost by discarding the remaining eigenvectors, the resulting distances will not agree with the input distances, but will approximate them in a certain optimal fashion [5]. A further "refinement" of these structures in three-dimensional space can then be used to improve agreement with the input distances.

In practice, even approximate distances are not known for most atom pairs; rather, one can set upper and lower bounds on acceptable distances, based on the covalent structure of the protein and on the observed NOE crosspeaks. Then particular instances can be generated by choosing (often randomly) distances between the upper and lower bounds, and embedding the resulting metric matrix. Different choices of specific distances will lead to new structures, so that it is typical to generate a family of approximate structures in this way.

Practical implementation. Although the actual mathematics of embedding and refining structures is straightforward, it is in practice a complex operation to encode all of the covalent and NOE information into distance bounds, to check it for errors and redundancies, and to analyze the output. The most commonly used programs to accomplish this are DISGEO [14] and DSPACE [15]. Our experience is with the former program, and here we give an overview of its implementation.

The first step in distance geometry involves the creation of a bounds matrix, which stores upper and lower distance bounds for each pair of atoms. For covalently-linked atoms, the upper and lower bounds are the same, and set equal to the known covalent bond length. Many other distances are known exactly; these include: 1–3 distances where fixed bond angles are assumed, distances between aromatic ring atoms and 1–4 distances across the peptide linkage. For each rotatable bonds, the 1–4 distance bounds are set to *cis* and *trans* limits. All other lower bounds are taken to be the sum of the corresponding van der Waals' radii, except for potential hydrogen bonds, where the minimum distance is somewhat shorter. Upper bounds also come from experimental NOE constraints, from limitations imposed by the covalent structure, or simply from estimates of the maximum extended length of a polypeptide chain. Dihedral angle constraints can be expressed in terms of distances between the 1–4 atoms involved, and are also imposed as constraints during subsequent refinements.

Where stereospecific assignments are not available, pseudoatoms can be placed at appropriate positions, and distance constraints measured from these points [16]. We have found it useful to use a mixed pseudoatom/all atom representation in many cases, where both the pseudoatoms and the "real" atoms are retained in the calculation, and distance relations among all of these are inferred from experimental data and from the chemical structure. The initial implementations of DISGEO did not include all of the bounds it is possible to infer from the chemical structure (such as the complete set of intra-ring distances in aromatic side chains), and the embedding proceeds more smoothly if these are included.

These initial bounds are then improved ("smoothed") by application of the triangle inequality, which effectively extends information about proton–proton distances to the nearby heavy atoms [5, 13]. Additional techniques can be used to improve bounds in cases where distances are fixed by the covalent structure [15]. At this point, the smoothed bounds matrix represents all that is known geometrically about the protein structure, though it by no means represents a unique molecular conformation. A family of conformations is generated by repeatedly choosing particular sets of distances between the upper and lower bounds. An eigenanalysis of the resulting metric matrix yields an embedded structure, as described above.

This eigenanalysis step is the single most time-consuming step of the calculation for large systems, and sometimes only a subset of atoms is embedded in order to save time. It is not known what the "best" subset of atoms is, but we find it useful to add many sidechain atoms to the set originally proposed [14], so that the substructure size is almost half that of the final calculation. Our current subset includes the heavy atoms N, C, Cβ, Cγ, Sγ, Cδ (Pro), Nδ (Asn), Cδ2 and Cε3 (Trp), and Cζ (Tyr, Phe, Arg), plus all protons attached to Cα and Cβ. This is an empirical list, and the exact specification of the substructure is probably not crucial; adding Cα atoms to the above list generally leads to markedly poorer results, however. After the substructure atoms have been embedded and refined, the resulting distances are inserted

into the full distance bounds matrix, and the entire problem is solved, now with the (presumably) improved distances arising from the substructure part of the calculation. It is also appropriate to discard some of the poorer structures at this point, to save computer time.

In distance geometry calculations on a number of proteins we have observed that fewer problems are encountered with incorrect local structures in regions of β sheets than in helical segments. Right-handed α helices are virtually indistinguishable from left-handed ones in their distance behavior unless connectivities can be observed between the Hα protons of residue i and the Hβ protons of residue $i+3$. These distances fall in the range 2.5 to 4.4 Å for right-handed helices and in the range 7.0 to 8.5 Å for left-handed ones. In our experience, unless a significant number of such connectivities are observed, distance geometry structures will often have distorted helices, sometimes beginning with one handedness and ending with the other. Although these poor initial structures can be corrected in the molecular dynamics refinement stage (discussed below), it is clearly advisable to search carefully for the diagnostic $d_{\alpha\beta}(i, i+3)$ NOEs in regions of the protein where helices are suspected.

Once the distances have been entered, smoothed and checked for internal consistency, a three-dimensional structure is determined, as outlined above. Much information is lost in the process of discarding all but the first three eigenvectors, so the resulting structures violate many of the distance constraints, and may also have incorrect chirality about the Cα atoms (and Cβ of Thr and Ile), non-planar conformations about sp^2 carbons and nitrogens, or incorrect *cis* peptide bonds. These problems can be ameliorated by a real-space (3D) refinement in which violations of distances, chirality or planarity are penalized by positive functions that approach zero when the constraints are satisfied. These refinements can be carried out by conventional nonlinear optimization techniques or by more powerful simulated annealing procedures. In fact, this 3D refinement stage, although a part of the DISGEO or DSPACE programs, conceptually belongs with the molecular dynamics refinement procedure outlined below, and various implementations differ primarily in the details of the functions to be optimized. What is unique to the "distance geometry" approach is the embedding step, which in our experience is remarkably robust in determining the correct *global fold* of a polypeptide chain.

An alternative approach, implemented in the computer code DISMAN [17], works in three-dimensional space from the outset, using an ingenious variable target approach to avoid (as much as possible) being trapped in local minima. There is not space to describe this approach here, but it appears to be comparable to distance geometry approaches in determining protein structures [18]. The correspondence of results is somewhat remarkable because the two procedures work in opposite directions: distance geometry first obtains a global structure whose details are refined at a later step, whereas DISMAN begins by building up local structures (using local constraints) and gradually takes into account longer range constraints. This and other evidence provide strong support for the notion that the results of NMR refinements of proteins are not, in general, dependent upon the details of the refinement procedures.

Molecular dynamics refinements

Refinements in three-dimensional space start from some initial conformation and attempt both to preserve the covalent geometry of the protein and to satisfy the constraints derived from the NMR measurements. The relative weighting of these two components is an important feature of the function to be optimized, which is almost always a sum of terms representing the energetics of the molecule *per se* and the experimental constraints. The molecular terms may be represented by simplified functions that maintain bond distances and angles and prevent non-bonded overlaps [15, 19], or by use of molecular mechanics expressions such as that from the AMBER potential [20]:

$$E_{MM} = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} + \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] + \sum_{H\text{-bonds}}\left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}\right] \quad (6)$$

This expression does not include explicit solvent molecules, but the expense of such calculations, plus difficulties in heating solvated systems to high temperatures, currently make solvated simulations impractical. We do reduce the net charge on Asp, Glu, His, Lys and Arg sidechains to ± 0.2, in order to mimic partially dielectric effects and to reduce the tendency to form salt bridges. We have also found it useful to increase the torsional force constant about the peptide bond to reduce the tendency otherwise found for these angles to deviate significantly from 0 or 180°.

The penalty functions arising from NMR observations can be of several forms, but the most common are "half-parabolas" related to upper bounds d_0:

$$E_{\text{NOE}} = \tfrac{1}{2}K_d(d - d_0)^2 \text{ for } d > d_0 = 0 \text{ otherwise} \quad (7)$$

In the absence of stereospecific assignments, pseudoatoms can be used, as described above. As an alternative, one can avoid pseudoatoms by using some average of the distances to the protons involved. The theoretical basis for such a procedure depends upon the nature of the internal motions, in particular whether they are fast or slow compared to the overall molecular tumbling [21]. As long as one is involved with deliberately conservative distance bounds, such distinctions make little difference, and an r^{-6} averaging procedure is typically used. For angular constraints we typically use a trigonometric function $K_\phi[1 - \cos(\phi - \phi_0)]$, where ϕ_0 is the nearest endpoint of an "allowed" range, e.g. backbone ϕ angles with large NH-CαH coupling constants

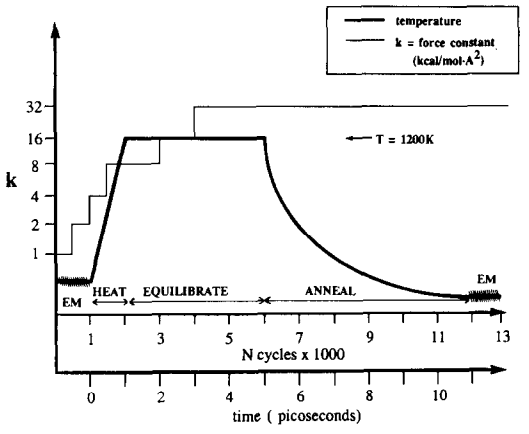

Fig. 2. Annealing scheme in use in our laboratory, giving temperature and force constant for the NOE penalty terms as a function of time.

($J > 8$ Hz) are typically restrained to the range $-160°$ to $-80°$. As with distance constraints, this potential is applied only when ϕ is in violation of the constraint.

The first applications of these procedures were reported about four years ago [22, 23]. Since that time it has become a fairly standard method in the field, although there are sometimes significant differences in implementation, and thermal procedures such as Monte Carlo can also be used [24]. The discussion here necessarily reflects our experience; as with other aspects of refinement, the optimal strategies are not yet clear.

As stated above, the relative weight given to molecular mechanics and NMR constraints is an important parameter in the optimization procedure, and one for which there is no general agreement. General practice can be divided into two camps, differentiating those who use relatively strong constraints (with K_d in the range 30–50 kcal/mol · Å^2) from those who use much weaker NOE penalties. For well-constrained structures, where the final constraint energies are low (on the order of 25 kcal/mol) the larger values are probably acceptable and introduce little strain into the overall protein structure. For less well-defined proteins the situation is less clear, and care should be taken in all cases to look for large residual constraint forces, particularly where conformational averaging may be present.

Although standard procedures (such as conjugate gradients) may be used to minimize the sum of E_{MM} and E_{NOE}, these typically get trapped in nearby local minima, and it has become common practice to use molecular dynamics or Monte Carlo schemes to obtain more robust optimization. These techniques add kinetic energy to the system to allow it to pass over some barriers in an attempt to find lower (local) minima. The scheme we have used is outlined in Fig. 2, and contains features common to most implementations. There is an equilibration phase in which the system is allowed to explore conformational space at a high temperature. Systems with large numbers of constraints are quite stable even at such temperatures, at least in vacuum calculations. The higher the temperature, the larger the barrier that can be overcome in a fixed length of simulation (on the average), but very high temperatures can lead to instabilities in the dynamics algorithm, and the length of time required to anneal the structure increase as the temperature is raised. Our temperature of 1200 °K for the equilibration phase represents a rough compromise between these competing factors.

The cooling or annealing phase is crucial to the success of the procedure and should be carried out as slowly as possible. We use a modified dynamics algorithm that removes kinetic energy in an exponential fashion [25], controlled by a time constant τ. (Because of partitioning between kinetic and potential energies, the actual time constant for the temperature decay is approximately 2τ.) Most of our calculations to date have used $\tau = 2$ psec for the cooling step, although recent experiments suggest that a larger value may be worth the added expense. Again, a tradeoff exists between the length of the computation and the care with which the cooling takes place, and other algorithms may prove superior to the one described here. The entire heating and cooling cycle can be repeated several times; in our experience, no further improvement is found after 2–5 cycles (depending on the protein involved).

Refinement by NOE intensities

Two-dimensional ^{1}H NOE spectra contain valuable information about molecular structures and dynamics, and a number of efforts are underway to use this information in a more quantitative manner. Between protons i and j, the 2D NOE crosspeak intensity I_{ij} for a mixing time τ_m is given by [26, 27]:

$$I_{ij} = \exp(-R\tau_m)_{ij} \tag{8}$$

where R is the dipolar longitudinal relaxation matrix, which depends on the set of interproton distances and the motional characteristic of the protein [26, 28]. In particular, $R_{ij} \propto r_{ij}^{-6}$, with r_{ij} the distance between the i-th and the j-th protons. Thus, if the crosspeak intensities can be measured accurately for a sufficiently small mixing time τ_m, we have as a linear approximation to Eq. 1,

$$I_{ij} = -R_{ij}\tau_m \tag{9a}$$

or,

$$r_{ij} \propto (I_{ij}/\tau_m)^{-1/6} \tag{9b}$$

The short mixing time crosspeak intensities thus directly yield interproton distances. It is possible to "calibrate" the intensities against known distances such as that between two methylene protons, or between two ring protons in phenylalanine or tyrosine. Alternatively, to eliminate potential zero-quantum contributions, the calibration can be based on inter-residue distances that are approximately fixed in regions of regular secondary structure. A reasonably complete set of distance can then serve to determine the molecular conformation.

However, practical considerations can severely compromise the simplicity of the above method. The intensities at small mixing times may be strongly influenced by noise, especially for weak peaks corresponding to large (>3.5 Å) distances. This can make it difficult to determine the range of validity

of the linear approximation, Eq. 9a. Peaks whose primary intensities arise from indirect (multiple-spin) effects may have very short linear regions (<20 msec) that may be missed entirely. A number of authors have considered the likely effect of such errors on estimates of interproton distances [8, 29–31]. A general conclusion is that longer distances may be significantly underestimated, often by more than 1 Å. While such uncertainties can be offset by assigning distance bounds deliberately higher than estimates of the true distances, this can entail a significant loss of information.

In a qualitative way, "back calculation" of the NOESY spectrum using the initial rate approximation is an important part of current refinement techniques. The computed spectra can be used to aid in the assignment of additional NOESY peaks, and the spectrum can be carefully checked for peaks predicted by the structure but absent from the spectrum. To exploit the full content of the 2D NOE intensities in a more quantitative way, one would like to adopt a refinement procedure that can treat spectra at longer mixing times. If this were merely a matter of accounting for spin diffusion effects, the calculation would be relatively straightforward, as outlined below. However, a number of other factors inhibit quantitation of NOE intensities; these include limited digital resolution in the spectrum, t_1 noise from the solvent or from sharp or intense protein resonances, "leakage" due to alternative relaxation mechanisms [21]. Even if it were possible to obtain near-"perfect" NOE spectra, where only dipolar relaxation contributed, the resulting intensities would depend upon molecular motions as well as interproton distances, and detailed motional characterizations of biopolymers are not known. Nevertheless, it appears that significant gains could be made by at least accounting for spin diffusion (multiple-spin) effects, and several groups are pursuing research in this direction [8–11]. The following paragraphs give an outline of these efforts.

The general procedure for such refinement is as follows (see Fig. 1): (a) start with an initial trial structure, say, from a distance geometry/molecular dynamics calculation as described above; (b) account for spin diffusion effects by simulating NOE spectra, using (at present) relatively simple models for the motions of the protons; (c) use the difference between the simulated and the experimental intensities to improve the trial structures; and (d) iterate until a desirable agreement is reached between the simulated and observed intensities.

In the simplest, but potentially very effective approach, the distance bounds are modified by the spectral comparisons: if the computed NOE intensity is too large, the lower bound is raised by some empirically determined amount; if it is too small, the corresponding upper bound is lowered. One then returns to distance geometry or real space refinements to incorporate this new information into the structures, and iterates the procedure. There is no guarantee that this procedure will converge, and in some cases the direction of the error may be different for different mixing times, making it hard to decide what correction to make. Nevertheless, the work by South *et al.* [32] indicates that this may be a promising approach.

A second method is being pursued in our laboratory [11]. We use the difference between calculated and experimental intensities to form a penalty function P,

$$P = \sum_{\text{peaks},\tau_m} (I_{ij}^{\text{exp}} - I_{ij}^{\text{calc}})^2 \tag{10}$$

The gradient of P, ∇P, with respect to proton coordinates provides a contribution to the force in a molecular dynamics annealing scheme that is used to improve the structures. The key advantages are (a) the refinement scheme works directly to maximize agreement between computed and experimental spectra, (b) results from several mixing times can be refined simultaneously, and (c) it is easy to incorporate other penalty functions (such as those arising from coupling constants or chemical shifts) in the same calculation.

Generating an expression for ∇P is not trivial, although the final result resembles the well-known two-spin expression. The rate matrix can be diagonalized as $R = L\lambda L^T$, where L is the unitary eigenvalue matrix, and λ the diagonal eigenvalue matrix. The gradient of P is determined by the gradient of I_{ij}, which is given by [11]:

$$\nabla_\mu I_{ij} = \sum_{pqlm} L_{ip} L_{pq} \nabla_\mu R_{ql} L_{lm}^T L_{mj}^T \times \left[\frac{\exp(\lambda_p \tau_m) - \exp(-\lambda_m \tau_m)}{\lambda_p - \lambda_m}\right] \tag{11}$$

where λ_i is the i-th element of the eigenvalue λ matrix and μ is any proton coordinate. $\nabla_\mu R_{ql}$ can easily be computed since the elements of R are simple functions of distances.

The penalty function P and its gradient ∇P are incorporated into the code AMBER [33], which has the force field to handle other (non-NOE-based) constraints, and an annealing routine for minimization. Test studies for a small peptide with simulated "dry" lab intensities are promising [11]. Further studies on crambin and zinc-fingers are currently in progress. In the current implementation it can be quite time-consuming to carry out these calculations for systems with more than a hundred spins, and approximation schemes are being studied [11, 34]. Nevertheless, having ∇P available should help many refinement schemes, since it indicates the *direction* that a conformational update should take in order to improve agreement with the experiment, a feature that is lacking in current alternatives. Furthermore, since the elements of R depend upon correlation times as well as distances, this same general method allows calculation of the gradient with respect to motional parameters, and thus could be used to refine effective correlation times as well as interproton distances. All of these potential advantages remain to be demonstrated in a practical computational scheme, however.

A third approach uses an ingenious combination of experimental and computed spectra to estimate spin diffusion effects. It is based on the observation that the rate matrix R could in principle be determined directly from the (complete) intensity matrix

I. Diagonalizing I, we have $I = MDM^T$, where M and D are the eigenvector and eigenvalue matrices of I. Then it is easy to show that:

$$R_{ij} = -\frac{1}{\tau_m} \sum_l M_{il} \ln(D_l) M_{lj}^T \qquad (12)$$

Once the rate matrix R is known, the set of inter-proton distances can be determined. In practice, however, one can only obtain a small fraction of the set of intensities. For example, the diagonal peaks are often not resolved well enough to be measured with accuracy. This problem of incomplete intensities is attacked by merging the experimental data with intensities computed from a trial structure. Inversion via Eq. 12 then yields a new rate matrix and set of distances, which can be used to update distance bounds as described above. This general procedure is also very new and goes by the acronyms IRMA [10, 35] or MARDIGRAS [9].

Although each of these methods is in its infancy, it seems clear that considerable progress has been made in incorporating spin diffusion effects into the refinement process, and that this should allow an increase in the confidence one has in the correctness of the structure and (one hopes) in the precision of structural results themselves.

Refinements based on chemical shifts

In principle, chemical shifts depend upon conformation in ways that might provide useful information, particularly if approximate structures are already known by the methods outlined above. As with refinements based on NOE intensities, the principal question to be decided is whether the expressions used to compute the shifts for a particular structure are sufficiently accurate and reliable to be used as input in a refinement procedure. For chemical shift calculations, the situation is complicated by the fact that many physical contributions to the chemical shift are not known, and only empirical relations are available for ring-current effects, which we consider here. Hence, this method has not yet been successfully applied to structure determination; we give here a preliminary analysis of prospects in this direction.

The conformation-dependent contribution to chemical shifts that is best understood is the ring-current contribution associated with conjugated groups [36]. In the Haigh–Mallion model, for example, the magnitude of the secondary field at a proton position is proportional to the geometric factor:

$$K(r) = \sum_{\text{ring}} s_{ij} \left\{ \frac{1}{r_i^3} + \frac{1}{r_j^3} \right\} \qquad (13)$$

where s_{ij} is the area of the projection into the plane of the aromatic ring of the triangle defined by the proton and the bonded ring atoms i and j, and r_i and r_j are the distances from the proton to the i-th and j-th ring atoms. In the context of a protein calculation, each aromatic sidechain contributes to the local magnetic field at each proton, and the contributions can be significant (1–4 ppm) for nearby protons.

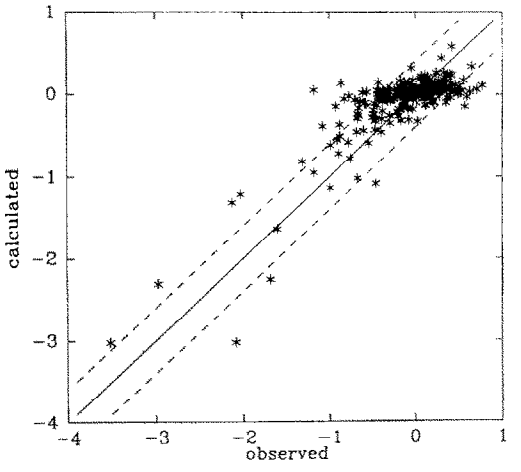

Fig. 3. Comparison of calculated and observed secondary shifts for sidechain protons in hen egg-white lysozyme. See text.

As an example, in Fig. 3 we plot the observed chemical shifts for protons bonded to sidechain carbons in lysozyme [37] with the predictions from a crystal structure of the same protein [38]. The values on the horizontal axis are the "secondary" or conformational shifts obtained by subtracting reference values obtained from short peptides [39] from the observed shifts. The values on the vertical axis are the computed ring-current contributions [36]. The general pattern is seen in many proteins: there are a small number of significant secondary shifts (>1 ppm), and there is a reasonable but flawed linear correlation between the observed and calculated values. Most points lie near the origin of this plot (i.e. have small secondary shifts), and here the spread of observed values is greater than that calculated, leading to a dense "ellipse" in this region. The simplest and most likely explanation for this behaviour is that other contributions to the chemical shift are present, with magnitudes on the order of 0.5 ppm. (Alternatively, some of this error may arise from deficiencies in the empirical ring-current expressions and/or from differences between solution and crystal conformations; it is not possible at present to sort out all of these potential errors.) Overall, the linear correlation between observed and calculated values is $r = 0.81$ for 276 side-chain protons in lysozyme, and the r.m.s. error in the predicted values is 0.30 ppm. For protons near aromatic rings, the dependence of computed shift on position is very strong, so that small changes in conformation (such as using a different crystal form of lysozyme for the calculation) can lead to significant changes in computed shifts. This is somewhat like the situation with regard to NOE intensities, and suggests that refinements may be meaningful even with imperfect theories, since sizable errors in the computed shifts translate into only small errors in computed positions.

Heme proteins form a special class, where large secondary shifts arise primarily from proximity to the porphyrin ring. Models for heme ring-current

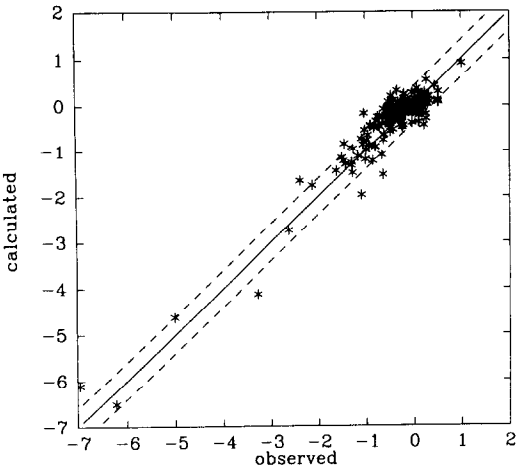

Fig. 4. Comparison of calculated and observed secondary shifts for side chain protons in sperm whale carbonmonoxy myoglobin. See text.

effects (reviewed in Ref. 40) treat the heme group as five or eight coplanar loops, and work quite well. Figure 4 shows results for the carbon monoxide complex of myoglobin, based on assignments from this laboratory [41, *] and the 1.5 Å X-ray structure of the same protein [42]. Here $r = 0.95$ for 213 side-chain protons, with an r.m.s. error in the predictions of 0.29 ppm.

Calculations of ring-current shifts are already being used in certain cases to aid in peak assignments [43, 44]. How useful such data will be for refinement purposes is not yet clear. Similar comparisons of calculated versus observed shifts for NMR-determined structures of plastocyanin show significantly poorer correlations (with values of r around 0.4), suggesting that improvements should be possible, and that predictions of chemical shifts can at least provide one parameter (among others) of the "quality" of the resultant structures. As more protein assignments become available, it may be possible to calibrate additional contributions to chemical shifts (e.g. from hydrogen bonds or from the peptide group [45, 46]) and to extend consideration to amide or Cα protons.

Acknowledgements—We thank Jane Dyson, Jonathan Moore and Walter Chazin for helpful discussions. This work was supported in part by NIH Grants GM 38794 and GM 36643.

REFERENCES

1. Wüthrich K, Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science* **243**: 45–50, 1989.
2. Kaptein R, Boelens R, Scheek RM and van Gunsteren WF, Protein structures by NMR. *Biochemistry* **27**: 5389–5395, 1988.
3. Clore GM and Gronenborn AM, Determination of three-dimensional structures of proteins in solution by nuclear magnetic resonance spectroscopy. *Protein Engineering* **1**: 275–288, 1987.
4. Wright PE, What can two-dimensional NMR tell us about proteins? *Trends Biochem Sci* **14**: 255–260, 1989.
5. Crippen GM and Havel TF, *Distance Geometry and Molecular Conformation*. Research Studies Press, Taunton, U.K., 1988.
6. Braun W, Distance geometry and related methods for protein structure determination from NMR data. *Q Rev Biophys* **19**: 115–157, 1987.
7. Clore GM, Brünger AT, Karplus M and Gronenborn AM, Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination. A model study of crambin. *J Mol Biol* **191**: 523–551, 1986.
8. Borgias BA and James TL, COMATOSE, a method for constrained refinement of macromolecular structure based on two-dimensional nuclear Overhauser effect spectra. *J Magn Reson* **79**: 493–512, 1988.
9. James TL and Borgias BA, Determination of DNA and protein structures in solution via complete relaxation matrix analysis of 2D NOE spectra. In: *Frontiers of NMR in Molecular Biology, UCLA Symposium on Molecular and Cellular Biology* (Eds. Live D, Armitage I and Patel D) Vol. 109. Alan R. Liss, New York, 1989.
10. Boelens R, Koning TMG, van der Marel GA, van Boom JH and Kaptein R, Iterative procedure of structure determination from proton–proton NOEs using a full relaxation matrix approach. Application to a DNA octamer. *J Magn Reson* **82**: 290–308, 1989.
11. Yip P and Case DA, A new method for refinement of macromolecular structures based on nuclear Overhauser effect spectra. *J Magn Reson* **83**: 643–648, 1989.
12. Blumenthal L, *Theory and Applications of Distance Geometry*. Cambridge University Press, Cambridge, U.K., 1953.
13. Havel TF, Kuntz ID and Crippen GM, The theory and practice of distance geometry. *Bull Math Biol* **45**: 665–720, 1983.
14. Havel T and Wüthrich K, A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of ^1H-^1H proximities in solution. *Bull Math Biol* **46**: 673–698, 1984.
15. Nerdal W, Hare DR and Reid BR, Three-dimensional structure of the *wild-type lac* Pribnow promoter DNA in solution. Two-dimensional nuclear magnetic resonance studies and distance geometry calculations. *J Mol Biol* **201**: 717–739, 1988.
16. Wüthrich K, Billeter M and Braun W, 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 Mol Biol* **169**: 949–961, 1983.
17. Braun W and Go N, Calculation of protein conformations by proton–proton distance constraints. A new efficient algorithm. *J Mol Biol* **186**: 611–626, 1985.
18. Wagner G, Braun W, Havel TF, Schaumann T, Go N and Wüthrich K, 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. *J Mol Biol* **196**: 611–639, 1987.
19. Nilges M, Clore GM and Gronenborn AM, Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Lett* **239**: 317–324, 1988.
20. Weiner SJ, Kollman PA, Nguyen DT and Case DA,

* Additional unpublished assignments have been made by T. Pochapsky.

An all-atom force field for simulations of proteins and nucleic acids. *J Computat Chem* **7**: 230–252, 1986.
21. Neuhaus D and Williamson M, *The Nuclear Overhauser Effect in Structural and Conformational Analysis*. VCH Publishers, New York, 1989.
22. Kaptein R, Zuiderweg ERP, Scheek RM, Boelens R and van Gunsteren WF, A protein structure from nuclear magnetic resonance data. Lac repressor headpiece. *J Mol Biol* **182**: 179–182, 1985.
23. Clore GM, Gronenborn AM, Brünger AT and Karplus M, Solution of 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 Mol Biol* **186**: 435–455, 1985.
24. Bassolino DA, Hirata F, Kitchen DB, Kominos D, Pardi A and Levy RM, Determination of protein structures in solution using NMR data and IMPACT. *Int J Supercomputer Appl* **2**: 41–61, 1988.
25. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A and Haak JR, Molecular dynamics with coupling to an external bath. *J Chem Phys* **81**: 3684–3690, 1984.
26. Macura S and Ernst RR, Elucidation of cross relaxation in liquids by two-dimensional N.M.R. spectroscopy. *Mol Phys* **41**: 95–117, 1980.
27. Keepers JW and James TL, A theoretical study of distance determinations from NMR. Two-dimensional nuclear Overhauser effect spectra. *J Magn Reson* **57**: 404–426, 1984.
28. Lane AN, The influence of spin diffusion and internal motions on NOE intensities in protein. *J Magn Reson* **78**: 425–439, 1988.
29. Clore GM and Gronenborn AM, Assessment of errors involved in determination of interproton distance ratios and distances by means of one- and two-dimensional NOE measurements. *J Magn Reson* **61**: 158–164, 1985.
30. Madrid M and Jardetzky O, Comparison of experimentally determined protein structures by solution of Bloch equations. *Biochim Biophys Acta* **953**: 61–69, 1988.
31. Olejniczak ET, Gampe RT Jr and Fesik SW, Accounting for spin diffusion in the analysis of 2D NOE data. *J Magn Reson* **67**: 28–41, 1986.
32. South TL, Kim B, Hare DR and Summers MF, Zinc fingers and molecular recognition. Structure and nucleic acid binding studies of an HIV zinc finger-like domain. *Biochem Pharmacology* **40**: 123–129, 1990.
33. Singh UC, Caldwell P, Weiner P and Kollman PA, *AMBER* 3.0. University of California, San Francisco, CA.
34. Yip PF, Calculating NOESY intensities by perturbation expansion. *Chem Phys Lett* **161**: 50–54, 1989.
35. Boelens R, Koning TMG and Kaptein R, Determination of biomolecular structures from proton–proton NOE's using a relaxation matrix approach. *J Mol Struct* **173**: 299–311, 1988.
36. Haigh CW and Mallion RB, Ring current theories in nuclear magnetic resonance. *Prog NMR Spectr* **13**: 303–344, 1980.
37. Redfield C and Dobson CM, Sequential ^{1}H NMR assignments and secondary structure of hen egg white lysozyme in solution. *Biochemistry* **27**: 122–136, 1988.
38. Kurachi K, Sieker LC and Jensen LH, Structures of triclinic mono- and di-*N*-acetylglucosamine. Lysozyme complexes—A crystallographic study. *J Mol Biol* **101**: 11–24, 1976.
39. Bundi A and Wüthrich K, ^{1}H NMR parameters of the common amino acid residues measured in aqueous solutions of the linear tripeptides H-Gly-Gly-X-L-Ala-OH. *Biopolymers* **18**: 285–297, 1979.
40. Cross KJ and Wright PE, Calibration of ring-current models for the heme ring. *J Magn Reson* **64**: 220–231, 1985.
41. Dalvit C and Wright PE, Assignment of resonances in the ^{1}H nuclear magnetic resonance spectrum of the carbon monoxide complex of sperm whale myoglobin by phase-sensitive two-dimensional techniques. *J Mol Biol* **194**: 313–327, 1987.
42. Kuriyan J, Wilz S, Karplus M and Petsko GA, X-ray structure and refinement of carbonmonoxy Fe(II)-myoglobin at 1.5 Å resolution. *J Mol Biol* **192**: 133–154, 1986.
43. Weiss MA and Hoch JC, Interpretation of ring-current shifts in proteins: Application to phage λ repressor. *J Magn Reson* **72**: 324–333, 1987.
44. Reid DG and Saunders MR, A proton nuclear magnetic resonance and nuclear Overhauser effect (NOE) study of human plasma prealbumin, including the development and application to spectral assignment of a combined ring current shift and NOE prediction program. *J Biol Chem* **264**: 2003–2012, 1989.
45. Pardi A, Wagner G and Wüthrich K, Protein conformation and proton nuclear-magnetic-resonance chemical shifts. *Eur J Biochem* **137**: 445–454, 1983.
46. Wagner G, Pardi A and Wüthrich K, Hydrogen bond length and ^{1}H NMR chemical shifts in proteins. *J Am Chem Soc* **105**: 5948–5949, 1983.