JOURNAL OF MAGNETIC RESONANCE **135,** 454–465 (1998)
ARTICLE NO. MN981564

Analysis of Error Propagation from NMR-Derived Internuclear Distances into Molecular Structure of *Cyclo*-Pro-Gly

Željko Džakula,*,[1] Nenad Juranić,† Michele L. DeRider,*,‡ William M. Westler,*
Slobodan Macura,† and John L. Markley*,[2]

National Magnetic Resonance Facility at Madison, Department of Biochemistry, University of Wisconsin-Madison, 420 Henry Mall, Madison, Wisconsin 53706; †Department of Biochemistry and Molecular Biology, Mayo Graduate School, Mayo Clinic/Foundation, Rochester, Minnesota 55905; and ‡Department of Chemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706

Received February 19, 1998; revised July 6, 1998

Analytical expressions have been derived that translate uncertainties in distance constraints (obtained from NMR investigations) into uncertainties in atom positions in the maximum likelihood (ML) structure consistent with these inputs. As a test of this approach, a comparison was made between test structures reconstructed by the new ML approach, which yields a single structure and a covariance matrix for coordinates, and those reconstructed by metric matrix distance-geometry (MMDG), which yields a family of structures that sample uncertainty space. The test structures used were 560 polyhedra, with edges of arbitrary length containing up to 50 vertices, and one polyhedron, with 100 vertices; randomized distance constraints generated from these structures were used in reconstructing the polyhedra. The uncertainties derived from the two methods showed excellent agreement, and the correlation improved, as expected, with increasingly larger numbers of MMDG structures. This agreement supports the validity of the rapid analytical ML approach, which requires the calculation of only a single structure. As a second test of the ML method, the approach was applied to the determination of uncertainties in the structure of a cyclic dipeptide, *cyclo*(DL-Pro-Gly) (cPG), derived from NMR cross-relaxation data. The input data were interproton distances calculated from NOEs measured for a solution of the peptide in 2:1 DMSO:H_2O at −40°C (so as to yield large negative NOEs). In order to evaluate effects of the quality of the input spectral parameters on the precision of the resulting NMR structure, information from the covalent geometry of cPG was not used in the structure calculations. Results obtained from the analytical ML approach compared favorably with those from the much slower random-walk variant of the Monte Carlo method applied to the same input data. As a third test, the ML approach was used with synthetic structural constraints for a small protein; the results indicate that it will be feasible to use this rapid method to translate uncertainties associated with a given set of distance restraints into uncertainties in atom positions in larger molecules.

Key Words: **error propagation; NMR structure determination; distance geometry, maximum likelihood.**

INTRODUCTION

Structures of biological macromolecules derived from NMR data are based primarily on interproton distances deduced from multidimensional cross-relaxation (NOESY) spectra (*1, 2*). A starting model for the structure can be determined from distance-geometry calculations from crude distance estimates, and then the measured NOE intensities can be converted into refined interproton distances by application of the full relaxation matrix (FRM) approach (*3–5*). These refined interproton distances finally can be used as experimental constraints that coerce the macromolecular structure into a more or less well defined conformational subspace (*6, 7*).

NOE intensities and interproton distances are determined experimentally only to within fairly broad error limits, and the uncertainties arise from a variety of sources (*8*). The problem that we address in this paper is how these errors translate into uncertainties in atomic coordinates, *i.e.*, how to find the family of structures that satisfy the given experimental data set. This problem has been addressed by several different approaches, all of which are deficient in one way or another. The conventional approach, consisting of random generation of families of structures consistent with the input data, is time-consuming and potentially unreliable, because the methods for structure determination may have limited capabilities for sampling conformational space. Methods that compare back-calculated NOESY spectra with an original experimental spectrum solve only one-half of the problem: even if the derived structure back-calculates into the original spectrum perfectly, it is not known whether other structures, possibly quite different, will back-calculate into the same spectrum. In addition, one can encounter the paradox that the poorer the original spectrum, the better the comparison between the original and back-calculated data. Similarly, refinement methods based on calculating gradients of peak volumes with respect to interproton distances (*9, 10*) fail to find the size of the conformational subspace of proton pairs with small gradients when the minimum of the penalty function is broad.

[1] Current address: Molecular Simulations, Inc., 9605 Scranton Rd., San Diego, CA 92121.

[2] To whom correspondence should be addressed.

1090-7807/98 $25.00

Both the heuristic approach based on random sampling of distances with the use of doubly iterated Kalman filters (*11*) and the Monte Carlo approach (*8, 12*) are computationally costly. Few studies have taken into account the great variability in the signal-to-noise of the original cross-relaxation data in assessing errors in individual distance constraints. In this paper, we present a direct method, based on the maximum likelihood approach (*13*), for translating the experimental NMR observations and their associated errors into uncertainties in atomic coordinates.

THEORY

The experimental observations used to reconstruct the molecular structure, collected in the column vector **E**, include interproton distances derived from NOEs and dihedral angles obtained from spin–spin couplings. The vector **E** also contains the bond lengths, valence angles, chiralities, and planarity constraints that are imposed by the covalent structure of the molecule. The variances and covariances of the experimental measurements along with uncertainties in the covalent constraints, are contained in the variance matrix $\mathbf{S_E}$ whose dimensions are $N \times N$, where N represents the length of the vector **E**. The $3M$ nuclear coordinates of the M atoms in the molecule are gathered in the vector **X** and arranged so that $X_{3k-2} = x_k$, $X_{3k-1} = y_k$, and $X_{3k} = z_k$ ($k = 1, \ldots, M$). Only $3M - 6$ coordinates are independent. Six constraint functions, $C_1(\mathbf{X}), \ldots, C_6(\mathbf{X})$, assembled in the vector **C,** are needed to account for three orientational and three translational degrees of freedom of the molecule as a whole. Constraints are satisfied when all the constraint functions become equal to zero. The goal is to find the set of values of the coordinates, $\mathbf{X_0}$, that satisfies simultaneously the experimental measurements $\mathbf{E_0}$ and the constraints **C**, and to estimate the matrix $\mathbf{S_X}$ of variances and covariances of $\mathbf{X_0}$ knowing $\mathbf{E_0}$ and the matrix $\mathbf{S_E}$. The determination of the most likely structure is achieved on employing the matrices (Appendix; also see (*13*))

$$\mathbf{Q_0} = \left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)^T_{\mathbf{X^{(0)}}} \mathbf{S_E^{-1}} \left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)_{\mathbf{X^{(0)}}},$$

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q_0} & (\partial \mathbf{C}/\partial \mathbf{X})^T_{\mathbf{X^{(0)}}} \\ (\partial \mathbf{C}/\partial \mathbf{X})\mathbf{X^{(0)}} & \mathbf{0} \end{bmatrix} \qquad [1]$$

and vectors

$$\mathbf{R_0} = \left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)^T_{\mathbf{X^{(0)}}} \mathbf{S_E^{-1}}(\mathbf{E} - \mathbf{E_0}), \quad \mathbf{R} = \begin{bmatrix} (\mathbf{R_0}) \\ -\mathbf{C} \end{bmatrix}_{\mathbf{X^{(0)}}}, \qquad [2]$$

where $\mathbf{0}$ is a zero matrix of dimensions 6×6 (*cf.* Appendix). Initially, Eqs. [1]–[2] are evaluated using a guess $\mathbf{X^{(0)}}$ for the coordinates. The initial guess is improved by employing the expression

$$[(\mathbf{X}^{(i+1)})^T \ \lambda_1 \cdots \lambda_6]^T = [(\mathbf{X}^{(i)})^T \ 0 \cdots 0]^T - \mathbf{Q}^{-1}\mathbf{R}, \qquad [3]$$

where λ_i, $i = 1, \ldots, 6$ are intermediate results of no interest. Equation [3] is used iteratively, and the iterated solution $\mathbf{X}^{(i)}$ converges toward the maximum likelihood structure $\mathbf{X_0}$ within a few steps provided that **E** is an approximately linear function of **X**. The desired variance matrix $\mathbf{S_X}$ is found by truncating the $(3M + 6) \times (3M + 6)$ matrix $\mathbf{Q}^{-1}$ to the order $3M \times 3M$. Finally, the area of space that is statistically accessible to each atom "k" ($k = 1, \ldots, M$) is visualized approximately by drawing an ellipsoid centered at the most likely atom position (x_{k0}, y_{k0}, z_{k0}) and determined by the 3×3 submatrix $\mathbf{S_k}$ from the diagonal of the matrix $\mathbf{S_X}$. The submatrices $\mathbf{S_k}$ are composed of the variances and covariances of coordinates of the atom k (*11*). The axes of the ellipsoid coincide with the principal axes of the matrix $\mathbf{S_k}$, and the lengths of the axes are square roots of the eigenvalues of $\mathbf{S_k}$.

The ML procedure described above computes uncertainties in Cartesian coordinates, and so its results depend on the choice of the frame of reference. This unwanted dependence on the arbitrary choice of the frame of reference is eliminated by extending the analysis to the chemically relevant internal descriptors of the molecular geometry, such as dihedral angles. These internal coordinates are invariant to the choice of reference frame. The variances and covariances in dihedral angles or other geometric descriptors (vector $\mathbf{\Phi}$) of the molecular structure can be found easily (*14*) as

$$\mathbf{S_\Phi} = (\partial \mathbf{\Phi}/\partial \mathbf{X})^T \mathbf{S_X} (\partial \mathbf{\Phi}/\partial \mathbf{X}). \qquad [4]$$

The assumption that second- and higher-order terms in the Taylor expansion of $\mathbf{E}(\mathbf{X})$ are negligible is essential for successful evaluation of both the maximum likelihood structure and $\mathbf{S_X}$. Cross-relaxation rates are strongly nonlinear functions of the relevant coordinates within the interval of ± 0.5 Å, whereas distances D_{ij} often vary almost linearly with coordinates within much wider ranges. The expressions that are available for the first derivatives of NOE intensities with respect to coordinates (*9, 10*) can be used only when the distance errors δD_{ij} are relatively small. Finite distance errors require computation of the second derivatives, which would render the method impractical. That is the reason why the present method calculates error propagation from internuclear distances D_{ij}, which are "secondary" experimental observations, rather than from the cross-relaxation rates, which are the "primary" experimental data (*14*). Therefore, the $N \times 3M$ matrix $(\partial \mathbf{E}/\partial \mathbf{X})_{\mathbf{X}_0}$, used above, contains the elements

$$[(\partial D_{ij}/\partial x_k) \ (\partial D_{ij}/\partial y_k) \ (\partial D_{ij}/\partial z_k)]_{\mathbf{X_0}} = [x_i - x_j \ \ y_i - y_j \ \ z_i - z_j]_{\mathbf{X_0}} (\delta_{ik} - \delta_{jk})/D_{ij}, \qquad [5]$$

where $\delta_{ij} = 1$ if $i = j$, 0 otherwise. We have chosen to estimate the uncertainties in the distances (*i.e.*, the diagonal elements of the matrix $\mathbf{S_E}$) from the lower and upper bounds,

L_{ij} and U_{ij}, respectively, assuming that distances are uniformly distributed within their limits:

$$\langle \delta D_{ij}^2 \rangle = (U_{ij} - L_{ij})^2/12. \quad [6]$$

The matrix $\mathbf{Q_0}$ (Eq. [1]) is singular because of the interdependence of the coordinates. The inversion of $\mathbf{Q_0}$ (Eq. [3]) is made possible by augmenting it with the blocks containing derivatives of the constraint functions $C_i(\mathbf{X})$ (Eq. [1]). A convenient choice of constraints includes the condition that the origin is located at the center of mass

$$C_1(\mathbf{X}) \equiv \sum_{k=1}^{M} x_k, \qquad C_2(\mathbf{X}) \equiv \sum_{k=1}^{M} y_k, \qquad C_3(\mathbf{X}) \equiv \sum_{k=1}^{M} z_k \quad [7]$$

and that the axes of the reference frame coincide with the principal axes of the inertia tensor

$$C_4(\mathbf{X}) \equiv \sum_{k=1}^{M} x_k y_k, \qquad C_5(\mathbf{X}) \equiv \sum_{k=1}^{M} x_k z_k,$$

$$C_6(\mathbf{X}) \equiv \sum_{k=1}^{M} y_k z_k. \quad [8]$$

When the constraints are given by Eqs. [7]–[8], the $6 \times 3M$ matrix $(\partial \mathbf{C_k}/\partial \mathbf{X})_{\mathbf{X_0}}$ becomes

$$\left(\frac{\partial \mathbf{C_k}}{\partial \mathbf{X}}\right)_{\mathbf{X_0}} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & \cdots & 0 & 0 & 1 \\ y_1 & x_1 & 0 & y_2 & x_2 & 0 & \cdots & y_M & x_M & 0 \\ z_1 & 0 & x_1 & z_2 & 0 & x_2 & \cdots & z_M & 0 & x_M \\ 0 & z_1 & y_1 & 0 & z_2 & y_2 & \cdots & 0 & z_M & y_M \end{bmatrix}_{\mathbf{X_0}}. \quad [9]$$

EXAMPLES

Arbitrarily Generated Polyhedra

As an initial test of the method, comparisons were made between structures and their uncertainties calculated from the same set of input distances by the new maximum likelihood (ML) approach and by the conventional metric matrix distance-geometry (MMDG) method. In order to keep the calculations small, but still test possible effects of differing geometry on the outcome, both methods were applied to a series of 563 arbitrarily generated polyhedra. The sizes of polyhedra (their numbers of vertices) ranged from 4 to 50; in addition, one polyhedron with 100 vertices was included. For each polyhedron size, between 5 and 15 polyhedra were constructed with their vertices positioned randomly within a cube, 20 Å/edge. The errors in distances were confined to ± 0.6 Å. The ML approach was applied both to reconstruct the known target vertex positions (from the exact distances) and to calculate their uncertainties (from the uncertainties in these distances). The MMDG ensembles were generated in the usual manner in which protein structures are calculated from NMR data: bounds were imposed on distances in such a way that Eq. [6] was satisfied; triangular inequalities were used to smooth both upper and lower bounds; trial values for distances were generated randomly within the smoothed bounds and used as inputs for MMDG to reconstruct the initial polyhedra. No metrization was employed. The embedded structures were projected into three dimensions and regularized to eliminate violations of distance bounds. Successfully regularized structures were accepted, while those with residual violations or with incorrect chiralities of substructures were discarded, and the procedure was repeated until 100–500 structures were accumulated for each polyhedron. As an example, an MMDG ensemble created for a sextahedron is shown on Fig. 1A, together with the ML results.

In all the polyhedra of all different sizes, the variances and covariances of coordinates calculated with the ML and MMDG approaches were virtually the same when the ensembles were sufficiently large (Figs. 1B, 1D). However, as the complexity of the system under study increased, sampling of conformational space by MMDG structure generation required progressively more time, encountered more difficulties, and became less reliable than the ML method. As an illustration, we randomly assembled subsets of structures from each ensemble and calculated r^2 between the ML covariances and the MMDG covariances computed from the subsets. The variations of these r^2 values as functions of the sub-ensemble sizes are displayed in Fig. 1C. In all cases, r^2 increased monotonically with the sub-ensemble size, suggesting that eventually MMDG results must converge toward the ML results and thus validating ML. These examples reinforce two conclusions: (a) ML gives the same results as MMDG in the limit of infinitely large ensembles, and (b) ML does so in a much shorter time.

Cyclic Dipeptide cyclo(DL-Pro-Gly)

As a second test of the method, error propagation from NOE measurements into the positions of hydrogen atoms was examined in the cyclic dipeptide *cyclo*(DL-Pro-Gly) (cPG). The dipeptide consists of 21 atoms, including 10 protons. The geometry of the dipeptide is completely determined by its covalent structure.[3] This

[3] The optimal independent set of distances in *cPGall* was obtained by analyzing the relevant pentangles (method *A* in the text). There are $3N - 6 = 57$ independent distances and they connect the following pairs of atoms: N_{Pro}–C^{α}_{Pro}, N_{Pro}–H^{α}_{Pro}, N_{Pro}–C^{β}_{Pro}, N_{Pro}–C^{γ}_{Pro}, N_{Pro}–C^{δ}_{Pro}, N_{Pro}–$H^{\delta R}_{Pro}$, N_{Pro}–$H^{\delta S}_{Pro}$, N_{Pro}–C'_{Pro}, N_{Pro}–C^{α}_{Gly}, N_{Pro}–C'_{Gly}, N_{Pro}–O_{Gly}, C^{α}_{Pro}–H^{α}_{Pro}, C^{α}_{Pro}–C^{β}_{Pro}, C^{α}_{Pro}–$H^{\beta S}_{Pro}$, C^{α}_{Pro}–$H^{\beta R}_{Pro}$, C^{α}_{Pro}–C^{γ}_{Pro}, C^{α}_{Pro}–C^{δ}_{Pro}, C^{α}_{Pro}–C'_{Pro}, C^{α}_{Pro}–O_{Pro}, C^{α}_{Pro}–N_{Gly}, C^{α}_{Pro}–C'_{Gly}, C^{α}_{Pro}–O_{Gly}, H^{α}_{Pro}–C^{β}_{Pro}, H^{α}_{Pro}–C'_{Pro}, C^{β}_{Pro}–$H^{\beta S}_{Pro}$, C^{β}_{Pro}–$H^{\beta R}_{Pro}$, C^{β}_{Pro}–C^{γ}_{Pro}, C^{β}_{Pro}–$H^{\gamma R}_{Pro}$, C^{β}_{Pro}–$H^{\gamma S}_{Pro}$, C^{β}_{Pro}–C^{δ}_{Pro}, $H^{\beta S}_{Pro}$–$H^{\beta R}_{Pro}$, $H^{\beta S}_{Pro}$–C^{γ}_{Pro}, C^{γ}_{Pro}–$H^{\gamma R}_{Pro}$, C^{γ}_{Pro}–$H^{\gamma S}_{Pro}$, C^{γ}_{Pro}–$H^{\delta R}_{Pro}$, C^{γ}_{Pro}–$H^{\delta S}_{Pro}$, $H^{\gamma R}_{Pro}$–$H^{\gamma S}_{Pro}$, $H^{\gamma R}_{Pro}$–C^{δ}_{Pro}, C^{δ}_{Pro}–$H^{\delta R}_{Pro}$, C^{δ}_{Pro}–$H^{\delta S}_{Pro}$, C^{δ}_{Pro}–C'_{Gly}, C'_{Pro}–O_{Pro}, C'_{Pro}–N_{Gly}, C'_{Pro}–H^{N}_{Gly},

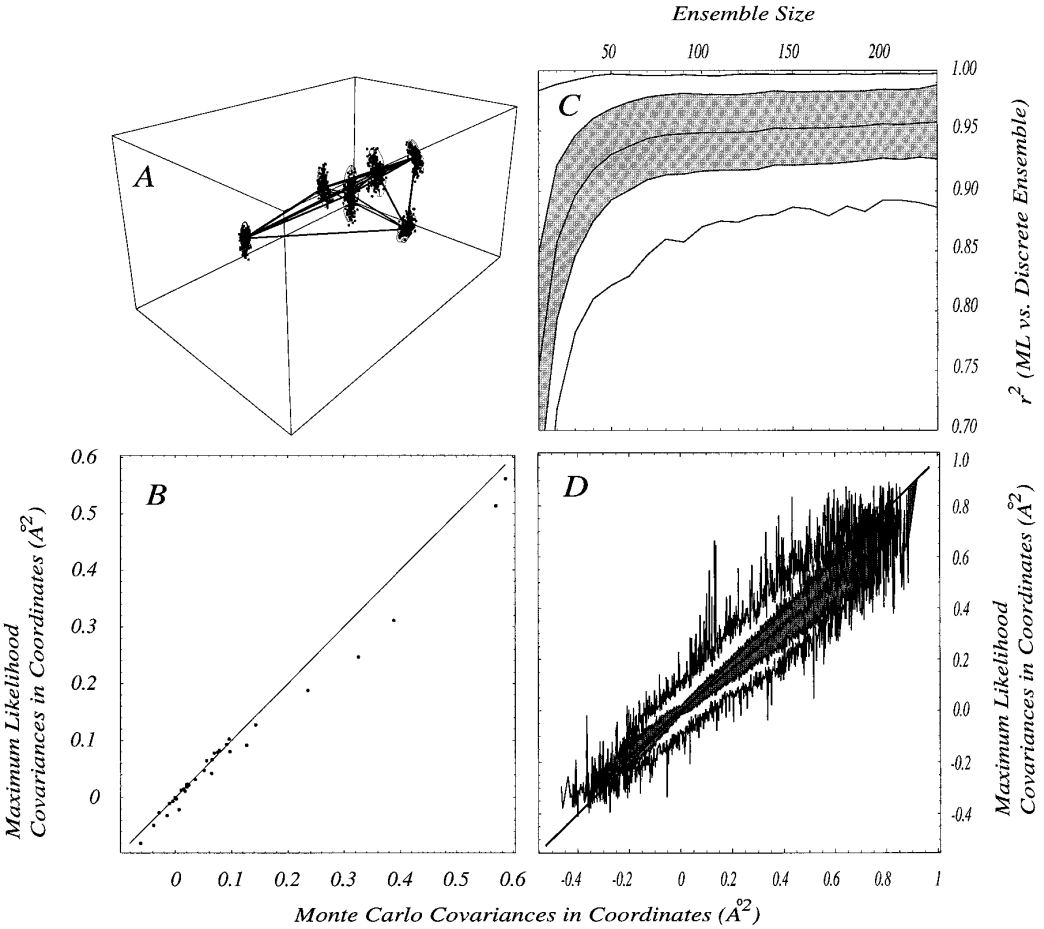

FIG. 1. (A) Areas of uncertainties in vertex positions in an arbitrarily constructed sextahedron, as calculated by the ML approach (ellipsoids) and as derived from the MMDG ensemble of structures that satisfy the constraints imposed on internode distances (points). (B) Correlation between covariances of coordinates of the sextahedron obtained from a discrete ensemble (abscissa) and with the ML approach (ordinate) ($r^2 = 0.985$). (C) The variation of correlation coefficient values r^2 between ML and MMDG covariances of coordinates in 563 polyhedra as a function of the size of MMDG ensembles. The gray area represents two standard deviations around the average r^2 values. Maximal and minimal r^2 values are also shown as functions of the ensemble sizes. (D) Correlation between covariances of coordinates of 563 polyhedra obtained from discrete ensembles (abscissa) and with the ML approach (ordinate). The gray area represents one standard deviation around the average (continuous line in the center). The ideal straight line with a slope of 1 is shown for comparison, as well as the range of variation. More than 10^5 data points were used after being divided with normalization factors of the form $\max[\delta(xyz)] - \min[\delta(xyz)]$, which were different for each polyhedron and which served to scale the coordinate covariances for various polyhedra so they can fit the same display window.

was demonstrated by using X-PLOR (*15*) to calculate the positions of all atoms from tight bounds imposed on all covalent distances (bond lengths, distances between geminal neighbors, and peptide bond planarities) and no bounds at all on non-covalent distances. The resulting all-atom structure, labeled *cPGall*, is well-defined and rigid, in spite of the fact that it was calculated in absence of any experimental NMR data. Thus, when the covalent information is used, experimental errors become completely masked by the high precision of the covalent bond lengths, valence angles, and planarity constraints.

In order to establish relationships between input experimental NMR data and the quality of the resulting structure, we eliminated the geometric information contained in the covalent structure by considering only the substructure consisting exclusively of protons (*cPGH*). The uncertainties in the structure of *cPGH* were calculated, using as input only the interproton distances derived from a full relaxation matrix analysis of NOESY spectra collected at a mixing time of 80 ms (Table 1). The interproton distances were classified into six categories depending on whether the values calculated from the full relaxation matrix fell within one of the following ranges: <2.0 Å, 2.0–2.5 Å, 2.5–3.2 Å, 3.2–4.0 Å, 4.0–5.0 Å, and >5.0 Å. The following lower and upper bounds pairs were imposed on the distances belonging to the first five classes: (1.6, 1.9) Å, (2.0, 3.0) Å, (2.2, 3.7) Å, (2.8, 4.8) Å, and (2.3, 6.6) Å, respectively. The upper bounds for the distances in the sixth class were obtained by adding 1.5 Å to the value obtained from

C'_{Pro}–C^{α}_{Gly}, O_{Pro}–N_{Gly}, N_{Gly}–H^{N}_{Gly}, N_{Gly}–C^{α}_{Gly}, N_{Gly}–$H^{\alpha R}_{Gly}$, N_{Gly}–$H^{\alpha S}_{Gly}$, N_{Gly}–C'_{Gly}, H^{N}_{Gly}–C^{α}_{Gly}, C^{α}_{Gly}–$H^{\alpha R}_{Gly}$, C^{α}_{Gly}–$H^{\alpha S}_{Gly}$, C^{α}_{Gly}–O_{Gly}, $H^{\alpha R}_{Gly}$–$H^{\alpha S}_{Gly}$, and $H^{\alpha R}_{Gly}$–C'_{Gly}. These independent distances are all known from the covalent structure and their uncertainties are negligibly small.

TABLE 1
Interproton Distances in cPG[a]

Proton 1	Proton 2	Lower bounds[a] (Å)	Upper bounds[a] (Å)	Distances[b] (Å)	Standard deviations[c] (Å)
H^{α}_{Pro}	$H^{\beta S}_{Pro}$	2.0	2.9	2.4	0.3
H^{α}_{Pro}	$H^{\beta R}_{Pro}$	2.2	3.7	3.0	0.3
H^{α}_{Pro}	$H^{\gamma R}_{Pro}$	2.2	3.7	2.9	0.3
H^{α}_{Pro}	$H^{\gamma S}_{Pro}$	2.8	4.8	3.9	0.4
H^{α}_{Pro}	$H^{\delta R}_{Pro}$	2.8	4.8	3.4	0.4
H^{α}_{Pro}	$H^{\delta S}_{Pro}$	2.8	4.8	3.9	0.4
H^{α}_{Pro}	H^{N}_{Gly}	2.8	4.8	3.9	0.4
H^{α}_{Pro}	$H^{\alpha R}_{Gly}$	2.2	3.7	3.1	0.3
H^{α}_{Pro}	$H^{\alpha S}_{Gly}$	3.2	5.6	4.3	0.4
$H^{\beta S}_{Pro}$	$H^{\beta R}_{Pro}$	1.6	1.9	1.8	0.1
$H^{\beta S}_{Pro}$	$H^{\gamma R}_{Pro}$	2.0	2.9	2.4	0.3
$H^{\beta S}_{Pro}$	$H^{\gamma S}_{Pro}$	2.2	3.7	2.7	0.3
$H^{\beta S}_{Pro}$	$H^{\delta R}_{Pro}$	3.2	5.8	4.1	0.4
$H^{\beta S}_{Pro}$	$H^{\delta S}_{Pro}$	3.2	5.6	4.0	0.4
$H^{\beta S}_{Pro}$	H^{N}_{Gly}	3.2	6.7	5.0	0.4
$H^{\beta S}_{Pro}$	$H^{\alpha R}_{Gly}$	3.2	6.6	5.3	0.4
$H^{\beta S}_{Pro}$	$H^{\alpha S}_{Gly}$	3.2	8.2	5.9	0.4
$H^{\beta S}_{Pro}$	$H^{\gamma R}_{Pro}$	2.2	3.7	3.0	0.3
$H^{\beta R}_{Pro}$	$H^{\gamma S}_{Pro}$	2.0	2.9	2.4	0.3
$H^{\beta R}_{Pro}$	$H^{\delta R}_{Pro}$	2.8	4.8	3.9	0.4
$H^{\beta R}_{Pro}$	$H^{\delta S}_{Pro}$	2.2	3.7	3.1	0.3
$H^{\beta R}_{Pro}$	H^{N}_{Gly}	3.2	6.0	4.3	0.5
$H^{\beta R}_{Pro}$	$H^{\alpha R}_{Gly}$	3.2	6.8	5.0	0.4
$H^{\beta R}_{Pro}$	$H^{\alpha S}_{Gly}$	3.2	7.0	5.2	0.4
$H^{\gamma R}_{Pro}$	$H^{\gamma S}_{Pro}$	1.6	1.9	1.8	0.1
$H^{\gamma R}_{Pro}$	$H^{\delta R}_{Pro}$	2.0	2.9	2.4	0.3
$H^{\gamma R}_{Pro}$	$H^{\delta S}_{Pro}$	2.2	3.7	3.0	0.3
$H^{\gamma R}_{Pro}$	H^{N}_{Gly}	3.2	8.5	6.2	0.5
$H^{\gamma R}_{Pro}$	$H^{\alpha R}_{Gly}$	3.2	7.4	5.5	0.4
$H^{\gamma R}_{Pro}$	$H^{\alpha S}_{Gly}$	3.2	8.6	6.2	0.4
$H^{\gamma S}_{Pro}$	$H^{\delta R}_{Pro}$	2.2	3.7	2.8	0.3
$H^{\gamma S}_{Pro}$	$H^{\delta S}_{Pro}$	2.0	2.9	2.4	0.3
$H^{\gamma S}_{Pro}$	H^{N}_{Gly}	3.2	8.7	6.2	0.5
$H^{\gamma S}_{Pro}$	$H^{\alpha R}_{Gly}$	3.2	8.5	6.1	0.4
$H^{\gamma S}_{Pro}$	$H^{\alpha S}_{Gly}$	3.2	8.9	6.4	0.4
$H^{\delta R}_{Pro}$	$H^{\delta S}_{Pro}$	1.6	1.9	1.8	0.1
$H^{\delta R}_{Pro}$	H^{N}_{Gly}	3.2	8.0	5.8	0.5
$H^{\delta R}_{Pro}$	$H^{\alpha R}_{Gly}$	3.2	6.3	4.6	0.5
$H^{\delta R}_{Pro}$	$H^{\alpha S}_{Gly}$	3.2	6.7	5.0	0.4
$H^{\delta S}_{Pro}$	H^{N}_{Gly}	3.2	7.2	5.3	0.5
$H^{\delta S}_{Pro}$	$H^{\alpha R}_{Gly}$	3.2	6.6	4.8	0.5
$H^{\delta S}_{Pro}$	$H^{\alpha S}_{Gly}$	3.2	6.4	4.7	0.5
H^{N}_{Gly}	$H^{\alpha R}_{Gly}$	2.2	3.7	2.7	0.3
H^{N}_{Gly}	$H^{\alpha S}_{Gly}$	2.0	2.9	2.3	0.3
$H^{\alpha R}_{Gly}$	$H^{\alpha S}_{Gly}$	1.6	1.9	1.8	0.1

[a] The bounds on interproton distances were calculated from NOESY data acquired at 500 MHz (Bruker AMX) for a mixing time of 80 ms at 233 K. The sample was 50-m*M cyclo*(DL-Pro-Gly) (cPG, Sigma) dissolved in 2:1 volume mixture of DMSO-d_6:H_2O. At this temperature, the dipeptide behaves like a small protein at room temperature in water (τ_c = 4 ns, as estimated from the cross-relaxation rate of 7 s^{-1} between geminal protons). Errors in all NOE intensities were estimated from NOESY cross-peak volumes at zero mixing time to be $\pm$0.015. Triangular smoothing was applied both to upper and lower bounds.

[b] Interproton distance values were derived from *cPGall* calculated by X-PLOR (*15*).

[c] Standard deviation for interproton distances were derived from the statistical ensemble cPGrrw.

the full-relaxation matrix calculations, and their lower bounds were fixed at 3.2 Å. Thus defined, the bounds ranges obtained from the experiment are identical to those inferred by *cPGall* or overlap with them significantly. None of the interproton distances in *cPGall* violates the experimental lower and upper bounds. Prior to use, the distance bounds were smoothed using the triangular inequality (*7*).

An attempt was made to create an MMDG ensemble in the same fashion as for the arbitrarily generated polyhedra. However, the gaps between lower and upper bounds on interproton distances were large enough to allow many different combinations of chiralities for four-proton subsets within *cPGH*. MMDG ensembles created in the usual way contained, therefore, a huge number of *cPGH* structures with wrong local chiralities and only a negligible fraction of structures with correct chiralities. This situation was not improved when a penalty term for chiralities was incorporated into the regularization procedure.

Therefore, to enable comparison between ML and a statistical ensemble, a procedure was devised which produced a high yield of structures that belong to the same class as *cPGH*. This procedure, which we call "restrained random-walk Monte Carlo" (*rrwMC*) is somewhat similar to the *CONCOORD* method of de Groot *et al.* (*17*); the difference is that *rrwMC* makes a random search outward from an initial structure to probe conformational space until violations are found, whereas CONCOORD starts from a conformational excursion from a RANDOM initial structure that leads to violations and searches randomly until the violations vanish. The starting structures used with *rrwMC* were those found in *cPGall*. Small steps are made in randomly chosen directions, with the step size ($\leq$0.01 Å) small when compared with the uncertainties in the interproton distances. If a step results in a violation of a distance bound, the procedure bounces back, ensuring that the distance bounds are never violated and that the whole trajectory remains confined within the same class of conformers as the initial, correct conformation (*cPGall*). To prevent the trajectory from remaining too close to the initial conformation, the procedure retracts only the increments to coordinates of the proton pair whose distance bound was violated. Other protons, not involved in distance bounds violations, retain their newly acquired positions. Although, in principle, this could create new violations of distance bounds between protons whose positions are updated and those whose coordinates were retracted, it rarely happened in practice. Also, it turned out that correct chiralities were maintained without the need for any special manipulations. All the relevant chiralities remained correct throughout the restrained random walk, which enabled fast creation of a large *rrwMC* ensemble (710 structures). For each member of the ensemble, the Brownian-like motion freezes after a randomly selected number of steps. The length of the trajectory ranged from 1,000 to 10,000 steps. Termination of the conformation space search left the structure somewhere within the "allowed" area of the space that needed to be sampled. The *rrwMC* ensemble was large enough to fill this

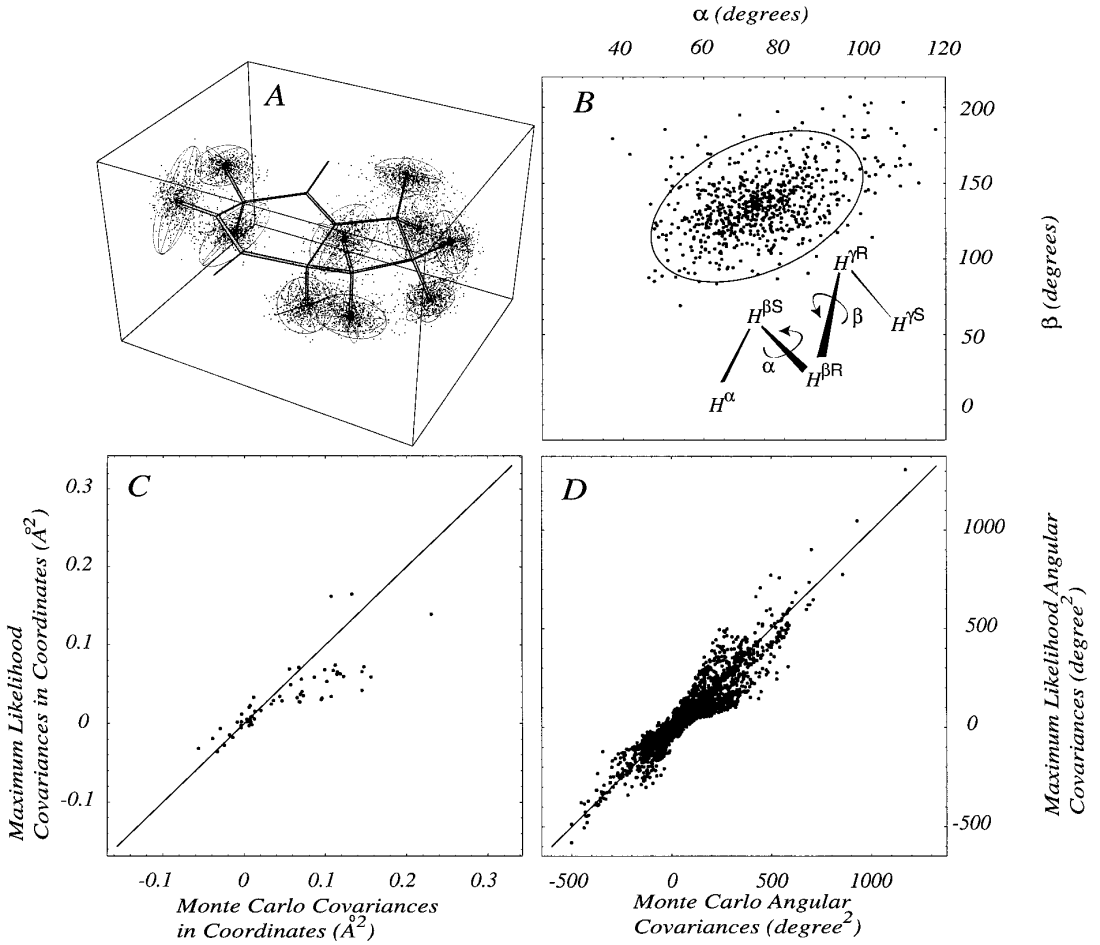

FIG. 2. (A) Conformational space of the proton subsystem of *cyclo*(DL-Pro-Gly) (*cPGH*) consistent with the experimental NOE data, estimated using the ML approach (ellipsoids) and *rrwMC* discrete ensemble (dots). The discrete ensemble consists of structures whose interproton distances are consistent with the NOESY spectrum recorded at 80 ms mixing time (Table 1). The structure of the molecule as a whole (*cPGall*) is superimposed on the proton substructure. (B) Correlation between covariances and variances for coordinates obtained from the discrete ensemble of structures (abscissa) and those obtained analytically using the ML approach (ordinate) ($r^2 = 0.862$). (C) Correlation diagram of two consecutive dihedral angles $H^{\alpha}_{\mathrm{Pro}}$–$H^{\beta S}_{\mathrm{Pro}}$–$H^{\beta R}_{\mathrm{Pro}}$–$H^{\gamma R}_{\mathrm{Pro}}$ and $H^{\beta S}_{\mathrm{Pro}}$–$H^{\beta R}_{\mathrm{Pro}}$–$H^{\gamma R}_{\mathrm{Pro}}$ –$H^{\gamma S}_{\mathrm{Pro}}$. Dots are values of the dihedral angles obtained from the discrete ensemble of structures. The ellipse was calculated from the ML matrix of covariances of the torsion angles (Eq. [4]). (D) Covariances of 3192 pairs of consecutive torsion angles within the proton subsystem (*cPGH*) obtained from the discrete ensemble of structures (abscissa) and with the ML approach (ordinate) ($r^2 = 0.917$).

whole area (dots in Fig. 2A), enabling a meaningful comparison with the ML results (ellipsoids in Fig. 2A).

In most cases, the distribution of distances within the discrete *rrwMC* ensemble was nonuniform and narrow around the most likely value of a distance. To account for this, we estimated the input uncertainties in distances from the discrete ensemble, rather than using Eq. [6]. The ML covariances for the proton Cartesian coordinates in *cPGH* were derived from the dispersions in interproton distances using Eqs. [1], [5], and [9]. The center of mass and products of inertia in *cPGH* were fixed in these calculations. When Eq. [3] was used with experimental interproton distances, the ML proton positions were close to those found in *cPGall.* ML proton coordinates calculated with interproton distances extracted from *cPGall* are identical to those in *cPGall* (Fig. 2A). The ellipsoids in Fig. 2A visualize the ML uncertainties in the proton coordinates of *cPGH.* The structure of the whole molecule (*cPGall*) is superimposed on these for comparison. The correlation found between the ML results and the covariances from the *rrwMC* ensemble improved with the size of the *rrwMC* ensemble, reaching the value $r^2 = 0.862$ when the complete ensemble (710 structures) was used for the comparison (Fig. 2B). This trend suggests that in the limit of an infinitely large *rrwMC* ensemble, the results from two methods will converge. The advantage of the ML approach is obvious.

Figure 2C shows correlations between two consecutive dihedral angles in the *cPGH* proton subsystem. The ellipse is calculated from covariances of the two torsion angles (Eq. [4]), whereas the points represent the values of the dihedral angles from the discrete ensemble of structures. Figure 2D shows covariances of 3192 consecutive pairs of dihedral angles within the proton subsystem of cPG calculated from the *rrwMC* ensemble (abscissa) and with the ML approach (ordinate). The correlation found between the covariances obtained by the two

methods ($r^2 = 0.917$ for all protons; $r^2 = 0.968$ for protons within the proline ring) shows that the ML approach can be applied successfully to determine uncertainties in the descriptors of the molecular geometry that are invariant to the choice of the frame of reference. The dependence of the linear correlation coefficient r^2 on the size of the MC ensemble is similar to that observed for Cartesian coordinates.

In terms of computational cost, the limiting step of the ML procedure is inversion of the matrix Q (Eq. [1]). To test whether the method will be useful for evaluating protein structures determined from NMR data, we applied ML to simulated NOE constraints plus the usual covalent constraints for a small protein (turkey ovomucoid third domain, OMTKY3). OMTKY3 has ~800 atoms and the size of the inverted matrix is $\approx 2400 \times 2400$. Both the MATHEMATICA (*16*) and the FORTRAN implementations of the ML method required only about 10 min of CPU time to invert the matrix on SGI Iris or on IBM RISC R/6000 workstations. The precision of the inverted matrix was excellent, as judged from the unit matrix obtained as the product of the original and the inverted matrices Q, even with the single precision FORTRAN implementation (data not shown). The fact that the matrix Q is composed of sparse matrices implies that it should be possible to use sparse matrix algorithms to achieve a significant reduction in the time needed to invert Q. When Eq. [3] was used to improve a distorted OMTKY3 structure which contained constraint violations, the resulting ML structure of OMTKY3 had no constraint violations. This suggests that ML could be useful as a refinement step.

All the calculations (ML, bounds smoothing, MMDG, regularization, *rrwMC*, full relaxation matrix calculations, selection of independent distances, and molecular graphics) were performed with our own software written in MATHEMATICA or FORTRAN (*18*).

DISCUSSION

The somewhat arbitrary choice of lower and upper distance bounds in *cPGH* was made to imitate the procedure in wide use for proteins (*19*). In the present form of the approach, one can account for internal molecular mobility by increasing the error ranges on distances, as is done with pseudoatoms. When distance constraints imposed on atoms are insufficient to fix their positions, *i.e.,* when they allow free rotation of groups of atoms, the areas of maximum likelihood for such atoms are rings. This violates the starting assumption that the probability distribution is unimodal and can be approximated by a Gaussian (*cf.* Appendix). The rows and columns of the matrix $\mathbf{Q}$ (Eq. [1]) which correspond to the coordinates of these atom groups are inversely proportional to the relatively large error ranges in the distances connecting them to the rest of the molecule. The elements of these rows and columns are therefore much smaller than the remaining matrix elements, and the condition number of the matrix is large (compared to the inverse of the machine precision). The calculated uncertainties in the underdetermined atom positions are grossly overestimated. However, in spite of dealing with an ill-conditioned matrix, the matrix inversion yielded covariances in coordinates of the remaining atoms that were largely unaffected by the lack of information localized at the underdetermined groups (data not shown).

The procedure described here is valid within the immediate vicinity of the maximum of the probability distribution for the coordinates. It employs derivatives and therefore only accounts for the continuous contributions to the overall uncertainties in coordinates. When the procedure is applied to only one, among several, distinct local probability maxima (discrete classes of conformers), the uncertainties contributed by the other discrete classes of conformers remain unaccounted for. One way to address this problem would be to identify the optimal set of $3N - 6$ independent distances, use them to reconstruct the whole set of discrete classes of conformers, and apply the maximum likelihood method to each probability maximum separately. A set of distances is independent if the rank of its Jacobian with respect to coordinates is $3N - 6$. It is optimal if the sum of dispersions of its elements is minimal compared to all other independent sets. When the structure contains several dozens of points or less, as in cPG, the optimal independent distances can be selected from the distances with minimal error ranges by scanning pentangles that contain the "candidate" distance and by using the fact that a pentangle contains one dependent and nine independent distances (method *A*). The inclusion of a dependent distance into the independent set is thus prevented. We used method *A* to identify the optimal set of independent distances in *cPGall* (see Footnote 3). Independent distances in larger molecules can be identified quickly by the "anchoring" algorithm used by Altman and Jardetzky (*11*) in their stepwise inclusion of structural motifs in the course of heuristic structure building (method *B*). Once identified, the independent distances can be used to build the structure, either by constructing pentangles (method *A*) or by successive anchoring of new points (method *B*). In cases where a pentangle within the structure can adopt two different configurations, the ambiguity can be removed on the basis of chirality constraints (if available for that pentangle), by tightening the error ranges on the distances within the pentangle by means of tetrahedral bounds smoothing, or by analyzing inconsistent distances from the pentangle to points that have already been "anchored." A combinatorial explosion of conformations thus can be avoided. If the input data set is such that ambiguities cannot be eliminated, the procedure results in multiple structures that are representative of discrete classes of conformers. Although these are probably not the most likely structures, they can be used as starting points in the iterative search for the maximum likelihood structures within their respective discrete classes.

Our approach analyzes the precision of NMR-derived molecular structures, as opposed to their accuracy (*20, 21*). The heuristic approach to structure building, which makes use of Kalman filters (*11*), yields uncertainties in coordinates as a byproduct. That approach, in analogy to our method, uses matrices of covariances to determine ellipsoids of uncertainties in atomic positions. The essential difference between the two

approaches is that the heuristic method bases its estimates of covariances on iterative sampling of distances, whereas the ML procedure uses analytical formulae to determine the covariances in a single computational cycle. Similarly, the advantage of the ML approach over Monte Carlo (MC) simulations (*8, 12*) again is that ML needs only a *single* cycle of computation to evaluate the size of the conformational space compatible with the experimental data. A prohibitively large number of lengthy computational cycles must be performed with MC to obtain comparable results.

The often-used assumption (*22*) that distances of equal lengths have the same error ranges is not always justified (*23, 24*). The true error limits on a distance depend on its value, the surrounding cross-relaxing network, and details of the experiment, such as the mixing time. Simulations of NOESY data corresponding to different mixing times indicate, as expected, that the optimal mixing time varies for different proton pairs (*23, 24*) and that the overall uncertainty in a given distance increases when the mixing time is either too short or too long. Uncertainties in distance constraints can be decreased by buildup curve analysis of data acquired at different mixing times or by using specialized pulse sequences, such as BD-NOESY (*25*) or CBD-NOESY (*26*), with a single mixing time. The effects these approaches have on the precision of structures can now be evaluated by the ML approach in a rapid, quantitative manner.

Although the discussion above has focused on errors propagated from distances derived from NOE data, errors in measured couplings can be incorporated straightforwardly into the procedure. The method is applicable to structures obtained with any method currently available (for instance MMDG or simulated annealing).

CONCLUSIONS

In summary, this new method of quantifying error propagation in NMR structure calculations provides a means to assess the reliability of NMR-derived structures. The approach should be applicable to the validation of conclusions concerning bioactive conformations and mechanisms of ligand binding or enzyme catalysis. Moreover, the method should make it possible to design improved strategies for NMR structure determination or refinement in which the most relevant structural data are collected in the most efficient manner.

APPENDIX: MAXIMUM LIKELIHOOD WITH CONSTRAINTS

A molecule with M atoms has $3M$ nuclear coordinates, but only $3M - 6$ coordinates are independent. Six constraints must be considered to account for the mutual dependence of coordinates. The constraints have been incorporated into the maximum likelihood equations according to the procedure described by Taupin (*13*). The number N of experimental observations and covalent constraints, gathered in the vector $\mathbf{E}$, must satisfy the condition $N > 3M - 6$. The column vector $\mathbf{E_0}$ represents the value of $\mathbf{E}$ obtained in the actual measurement. The diagonal elements of the $N \times N$ variance matrix $\mathbf{S_E}$ are squared experimental errors. The $3M$ coordinates are grouped in the vector $\mathbf{X} = [x_1\ y_1\ z_1 \cdots x_M\ y_M\ z_M]$. The maximum likelihood approach estimates the most likely values of the coordinates, $\mathbf{X_0}$, and their variance-matrix, $\mathbf{S_X}$, as the position and the width, respectively, of the peak of the *a posteriori* conditional probability $p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})$. The function $p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})$ represents the probability density for $\mathbf{X}$ when the measured values $\mathbf{E_0}$ of observables $\mathbf{E}$ are known with experimental errors $\mathbf{S_E}$. The conditional probability $p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})$ is obtained from the assumed Gaussian conditional probability distribution $p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})$ that the experimental observables (internuclear distances) will take values $\mathbf{E}$ knowing that the theoretical parameters (nuclear coordinates) have the values $\mathbf{X_0}$

$$p(\mathbf{E}|\mathbf{X} = \mathbf{X_0}) = [\sqrt{2\pi}\ \det(\mathbf{S_E})]^{-1} \times \exp[-\tfrac{1}{2}(\mathbf{E} - \mathbf{E_0})^T \mathbf{S_E^{-1}} (\mathbf{E} - \mathbf{E_0})], \quad [A1]$$

where $\mathbf{E} = \mathbf{E}(\mathbf{X})$ and $\mathbf{E_0} = \mathbf{E}(\mathbf{X_0})$.

To determine $p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})$, the *a priori* probability distribution for the coordinates, $p_{ap}(\mathbf{X})$, is also needed. There is no prior knowledge about the atomic coordinates, and $p_{ap}(\mathbf{X})$ is therefore uniform over the whole space allowed. However, the six constraints on coordinates reflecting the translational and rotational degrees of freedom of the molecule as a whole must be incorporated into the *a priori* distribution $p_{ap}(\mathbf{X})$. That is achieved by multiplying the uniform distribution with Dirac delta-functions that enforce the constraints on the coordinates (*13*):

$$p_{ap}(\mathbf{X}) = \text{const}_1 \prod_{i=1}^{6} \delta[C_i(\mathbf{X})] = \frac{\text{const}_1}{(2\pi)^3} \prod_{i=1}^{6} \lim_{s_i \to 0} \frac{1}{s_i} \exp[-C_i(\mathbf{X})^2/(2s_i^2)]. \quad [A2]$$

The delta-functions in Eq. [A2] are represented as the limiting cases of Gaussian distributions with vanishing widths. The terms $C_i(\mathbf{X})$ stand for the constraint functions (*cf.* Eqs. [7]–[8], main text); a set of coordinates $\mathbf{X}$ satisfies the constraints when $C_i(\mathbf{X}) = 0,\ \forall i$.

The determination of $p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})$ also requires the knowledge of the *a priori* probability $p_{ap}(\mathbf{E_0})$ that the observables will take the values $\mathbf{E_0}$. The distribution $p_{ap}(\mathbf{E_0})$ is the projection of the *joint* probability $p(\mathbf{E_0}|\mathbf{X})p_{ap}(\mathbf{X})$ of simultaneous occurrence of $\mathbf{E_0}$ and $\mathbf{X}$

$$p_{ap}(\mathbf{E_0}) = \int_{\mathrm{x}} p(\mathbf{E_0}|\mathbf{X})\, p_{ap}(\mathbf{X})\, d\mathbf{X}. \quad [A3]$$

Equations [A1]–[A3], combined with the Bayes formula

$$p(\mathbf{X}|\mathbf{E}=\mathbf{E_0})\,p_{ap}(\mathbf{E_0}) = p(\mathbf{E_0}|\mathbf{X})\,p_{ap}(\mathbf{X}) \qquad [A4]$$

lead to the following expression for the desired *a posteriori* conditional probability $p(\mathbf{X}|\mathbf{E}=\mathbf{E_0})$

$$\begin{aligned} p(\mathbf{X}|\mathbf{E}=\mathbf{E_0}) &= \frac{p(\mathbf{E_0}|\mathbf{X})\,p_{ap}(\mathbf{X})}{\int_{\mathbf{X}} p(\mathbf{E_0}|\mathbf{X})\,p_{ap}(\mathbf{X})\,d\mathbf{X}} \\ &= \text{const}_2 \exp[-\tfrac{1}{2}(\mathbf{E}-\mathbf{E_0})^T\mathbf{S_E^{-1}}(\mathbf{E}-\mathbf{E_0})] \\ &\quad \times \prod_{i=1}^{6} \lim_{s_i\to 0} \exp[-C_i(\mathbf{X})^2/(2s_i^2)] \\ &= \text{const}_2 \exp\left[-\tfrac{1}{2}(\mathbf{X}-\mathbf{X_0})^T\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)^T \right. \\ &\quad \left. \times \mathbf{S_E^{-1}}\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)(\mathbf{X}-\mathbf{X_0})\right] \\ &\quad \times \prod_{i=1}^{6} \lim_{s_i\to 0} \exp[-C_i(\mathbf{X})^2/(2s_i^2)], \end{aligned} \qquad [A5]$$

where the functional dependence of $\mathbf{E}$ upon $\mathbf{X}$

$$\mathbf{E} = \mathbf{E}(\mathbf{X}) \equiv \begin{bmatrix} E_1(x_1, y_1, z_1, \ldots, x_M, y_M, z_M) \\ \vdots \\ E_N(x_1, y_1, z_1, \ldots, x_M, y_M, z_M) \end{bmatrix} \qquad [A6]$$

was approximated by the truncated Taylor series

$$\begin{aligned} \mathbf{E} &= \mathbf{E}(\mathbf{X_0}) + \left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)_{\mathbf{X_0}} (\mathbf{X}-\mathbf{X_0}) \\ &\quad + \tfrac{1}{2}(\mathbf{X}-\mathbf{X_0})^T\left(\frac{\partial^2 \mathbf{E}}{\partial \mathbf{X}^2}\right)_{\mathbf{X_0}} (\mathbf{X}-\mathbf{X_0})\cdots \equiv \mathbf{E}(\mathbf{X_0}) \\ &\quad + \begin{bmatrix} (\partial E_1/\partial x_1) & \cdots & (\partial E_1/\partial z_M) \\ \vdots & \ddots & \vdots \\ (\partial E_N/\partial x_1) & \cdots & (\partial E_N/\partial z_M) \end{bmatrix}_{\mathbf{X_0}} (\mathbf{X}-\mathbf{X_0}) \\ &\quad + \tfrac{1}{2}(\mathbf{X}-\mathbf{X_0})^T \left[\begin{bmatrix} \left(\frac{\partial^2 E_1}{\partial x_1^2}\right) & \cdots & \left(\frac{\partial^2 E_1}{\partial x_1 \partial z_M}\right) \\ \vdots & \ddots & \vdots \\ \left(\frac{\partial^2 E_1}{\partial x_1 \partial z_M}\right) & \cdots & \left(\frac{\partial^2 E_1}{\partial z_M^2}\right) \end{bmatrix} \right. \\ &\quad \times \begin{bmatrix} \left(\frac{\partial^2 E_2}{\partial x_1^2}\right) & \cdots & \left(\frac{\partial^2 E_2}{\partial x_1 \partial z_M}\right) \\ \vdots & \ddots & \vdots \\ \left(\frac{\partial^2 E_2}{\partial x_1 \partial z_M}\right) & \cdots & \left(\frac{\partial^2 E_2}{\partial z_M^2}\right) \end{bmatrix} \cdots \\ &\quad \left. \times \begin{bmatrix} \left(\frac{\partial^2 E_N}{\partial x_1^2}\right) & \cdots & \left(\frac{\partial^2 E_N}{\partial x_1 \partial z_M}\right) \\ \vdots & \ddots & \vdots \\ \left(\frac{\partial^2 E_N}{\partial x_1 \partial z_M}\right) & \cdots & \left(\frac{\partial^2 E_N}{\partial z_M^2}\right) \end{bmatrix}^T \right]_{\mathbf{X_0}} (\mathbf{X}-\mathbf{X_0}) + \ldots \end{aligned} \qquad [A7]$$

assuming that quadratic and higher-order terms are negligible within the (still unknown) region of uncertainties of the coordinates.

Approximating, on the other hand, the *a posteriori* conditional probability distribution $p(\mathbf{X}|\mathbf{E}=\mathbf{E_0})$ with a Gaussian function

$$\begin{aligned} p(\mathbf{X}|\mathbf{E}=\mathbf{E_0}) &= [\sqrt{2\pi}\,\det(\mathbf{S_X})]^{-1} \\ &\quad \times \exp[-\tfrac{1}{2}(\mathbf{X}-\mathbf{X_0})^T\mathbf{S_X^{-1}}(\mathbf{X}-\mathbf{X_0})] \end{aligned} \qquad [A8]$$

and assuming that the variation of the matrix $\mathbf{S_X^{-1}}$ with $\mathbf{X}$ is negligible, one can find the uncertainties in the coordinates, $\mathbf{S_X}$, from the width of the peak of $p(\mathbf{X}|\mathbf{E}=\mathbf{E_0})$ at $\mathbf{X_0}$

$$\begin{aligned} \mathbf{S_X^{-1}} &= -\left[\frac{\partial^2 \ln p(\mathbf{X}|\mathbf{E}=\mathbf{E_0})}{\partial \mathbf{X}^2}\right]_{\mathbf{X_0}} \\ &= -\begin{bmatrix} \left(\frac{\partial^2 \ln p}{\partial X_1^2}\right) & & \\ \left(\frac{\partial^2 \ln p}{\partial X_1}\right)\left(\frac{\partial^2 \ln p}{\partial X_2}\right) & & \\ \vdots & & \\ \left(\frac{\partial^2 \ln p}{\partial X_1}\right)\left(\frac{\partial^2 \ln p}{\partial X_{3M}}\right) & & \end{bmatrix} \\ &\quad \times \begin{bmatrix} \left(\frac{\partial^2 \ln p}{\partial X_1}\right)\left(\frac{\partial^2 \ln p}{\partial X_2}\right) & \cdots & \left(\frac{\partial^2 \ln p}{\partial X_1}\right)\left(\frac{\partial^2 \ln p}{\partial X_{3M}}\right) \\ \left(\frac{\partial^2 \ln p}{\partial X_1^2}\right) & \cdots & \left(\frac{\partial^2 \ln p}{\partial X_2}\right)\left(\frac{\partial^2 \ln p}{\partial X_{3M}}\right) \\ \vdots & \ddots & \vdots \\ \left(\frac{\partial^2 \ln p}{\partial X_2}\right)\left(\frac{\partial^2 \ln p}{\partial X_{3M}}\right) & \cdots & \left(\frac{\partial^2 \ln p}{\partial X_{3M}^2}\right) \end{bmatrix}_{\mathbf{X_0}} \\ &= \left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)^T_{\mathbf{X_0}} \mathbf{S_E^{-1}} \left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)_{\mathbf{X_0}} + \sum_{i=1}^{K} \lim_{s_i\to 0} \frac{1}{s_i^2} \begin{bmatrix} \left(\frac{\partial C_i}{\partial X_1}\right)^2 & & \\ \left(\frac{\partial C_i}{\partial X_1}\right)\left(\frac{\partial C_i}{\partial X_2}\right) & & \\ \vdots & & \\ \left(\frac{\partial C_i}{\partial X_1}\right)\left(\frac{\partial C_i}{\partial X_{3M}}\right) & & \end{bmatrix} \end{aligned}$$

$$\times \begin{bmatrix} \left(\frac{\partial C_i}{\partial X_1}\right)\left(\frac{\partial C_i}{\partial X_2}\right) & \cdots & \left(\frac{\partial C_i}{\partial X_1}\right)\left(\frac{\partial C_i}{\partial X_{3M}}\right) \\ \left(\frac{\partial C_i}{\partial X_2}\right)^2 & \cdots & \left(\frac{\partial C_i}{\partial X_2}\right)\left(\frac{\partial C_i}{\partial X_{3M}}\right) \\ \vdots & \ddots & \vdots \\ \left(\frac{\partial C_i}{\partial X_2}\right)\left(\frac{\partial C_i}{\partial X_{3M}}\right) & \cdots & \left(\frac{\partial C_i}{\partial X_{3M}}\right)^2 \end{bmatrix}_{\mathbf{X_0}} . \quad [A9]$$

Note that the second derivatives of the constraint functions do not appear in Eq. [A9] because they multiply the constraint functions, which vanish at $\mathbf{X_0}$. The inversion of the matrix on the right side of Eq. [A9] is equivalent to the solution to the following set of linear equations involving "dummy" unknowns λ_k and arbitrary constants d_i

$$\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{s=1}^{3M}\left(\frac{\partial E_j}{\partial X_i}\right)_{\mathbf{X_0}}(S_E^{-1})_{jk}\left(\frac{\partial E_k}{\partial X_s}\right)_{X_0} X_s + \sum_{k=1}^{6}\left(\frac{\partial C_k}{\partial X_i}\right)_{\mathbf{X_0}}\lambda_k = d_i, \qquad i = 1, \ldots, 3M \quad [A10]$$

$$\sum_{j=1}^{3M}\left(\frac{\partial C_k}{\partial X_j}\right)_{X_0} X_j = s_k^2\lambda_k, \qquad k = 1, \ldots, 6. \quad [A11]$$

The desired variance matrix $\mathbf{S_X}$ is found from the solution to Eqs. [A10]–[A11], in the limit when s_k tends to zero, by truncating the following $(3M + 6) \times (3M + 6)$ matrix to the order $3M \times 3M$

$$\mathbf{Q}^{-1} = \begin{bmatrix} \left[\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)^T \mathbf{S}_{\mathbf{E}}^{-1}\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)_{\mathbf{X_0}}\right] & \left[\begin{matrix} \left(\frac{\partial C_1}{\partial X_1}\right) & \left(\frac{\partial C_2}{\partial X_1}\right) & \cdots & \left(\frac{\partial C_6}{\partial X_1}\right) \\ \left(\frac{\partial C_1}{\partial X_2}\right) & \left(\frac{\partial C_2}{\partial X_2}\right) & \cdots & \left(\frac{\partial C_6}{\partial X_2}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{\partial C_1}{\partial X_{3M}}\right) & \left(\frac{\partial C_2}{\partial X_{3M}}\right) & \cdots & \left(\frac{\partial C_6}{\partial X_{3M}}\right) \end{matrix}\right]^{-1} \\ \left[\begin{matrix} \left(\frac{\partial C_1}{\partial X_1}\right) & \left(\frac{\partial C_1}{\partial X_2}\right) & \cdots & \left(\frac{\partial C_1}{\partial X_{3M}}\right) \\ \left(\frac{\partial C_2}{\partial X_1}\right) & \left(\frac{\partial C_2}{\partial X_2}\right) & \cdots & \left(\frac{\partial C_2}{\partial X_{3M}}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{\partial C_6}{\partial X_1}\right) & \left(\frac{\partial C_6}{\partial X_2}\right) & \cdots & \left(\frac{\partial C_6}{\partial X_{3M}}\right) \end{matrix}\right] & \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \end{bmatrix} . \quad [A12]$$

The matrix $(\partial \mathbf{E}/\partial \mathbf{X})^T_{\mathbf{X_0}}\mathbf{S}_{\mathbf{E}}^{-1}(\partial \mathbf{E}/\partial \mathbf{X})_{\mathbf{X_0}}$ is singular because of interdependence of coordinates. The inversion is not possible without augmenting that matrix with blocks containing the derivatives of the constraint functions.

The most likely structure, $\mathbf{X_0}$, is found as the position of the maximum of the function $p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})$, with the additional requirement that the constraints are satisfied at $\mathbf{X_0}$

$$\partial p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})/\partial \mathbf{X}|_{\mathbf{X_0}} = \mathbf{0}, \qquad \mathbf{C}(\mathbf{X_0}) = \mathbf{0}. \quad [A13]$$

The first condition is equivalent to

$$\partial \ln p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})/\partial \mathbf{X}|_{\mathbf{X_0}} = \mathbf{0} \quad [A14]$$

because $p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})$ is nonnegative and $\ln(p)$ increases monotonically with p. The logarithm is used to separate the terms that make up $p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})$ (*cf.* Eq. [A5]). The search for $\mathbf{X_0}$ starts by expanding the conditions (Eqs. [A13], [A14]) into a Taylor series around an initial guess $\mathbf{X^{(1)}}$:

$$\left.\frac{\partial \ln p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})}{\partial \mathbf{X}}\right|_{\mathbf{X_0}} = \mathbf{0} = \left.\frac{\partial \ln p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})}{\partial \mathbf{X}}\right|_{\mathbf{X^{(1)}}} + \left.\frac{\partial^2 \ln p(\mathbf{X}|\mathbf{E} = \mathbf{E_0})}{\partial \mathbf{X}^2}\right|_{\mathbf{X^{(1)}}} (\mathbf{X_0} - \mathbf{X^{(1)}}) \quad [A15]$$

$$\mathbf{C}(\mathbf{X_0}) = \mathbf{0} = \mathbf{C}(\mathbf{X^{(1)}}) + \left.\frac{\partial \mathbf{C}}{\partial \mathbf{X}}\right|_{\mathbf{X^{(1)}}} (\mathbf{X_0} - \mathbf{X^{(1)}}). \quad [A16]$$

The derivatives in Eq. [A15] are found from the expression (*cf.* Eq. [A5])

$$\ln p(\mathbf{X}|\mathbf{E}=\mathbf{E_0}) = \text{const} - \tfrac{1}{2}[\mathbf{E}(\mathbf{X})-\mathbf{E_0}]^T\mathbf{S_E^{-1}}[\mathbf{E}(\mathbf{X})-\mathbf{E_0}] - \sum_{k=1}^{6}\lim_{s_k\to 0}[C_k(\mathbf{X})^2/(2s_k^2)] \quad [A17]$$

and their explicit forms are

$$\frac{\partial \ln p(\mathbf{X}|\mathbf{E}=\mathbf{E_0})}{\partial \mathbf{X}} = -\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)^T\mathbf{S_E^{-1}}[\mathbf{E}(\mathbf{X})-\mathbf{E_0}] - \sum_{k=1}^{6}\lim_{s_k\to 0}\left[\frac{C_k(\mathbf{X})}{s_k^2}\left(\frac{\partial C_k(\mathbf{X})}{\partial \mathbf{X}}\right)\right] \quad [A18]$$

$$\frac{\partial^2 \ln p(\mathbf{X}|\mathbf{E}=\mathbf{E_0})}{\partial \mathbf{X}^2} = -\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)^T\mathbf{S_E^{-1}}\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right) - \sum_{k=1}^{6}\lim_{s_k\to 0} \times\left[\frac{1}{s_k^2}\left(\frac{\partial C_k(\mathbf{X})}{\partial \mathbf{X}}\right)\otimes\left(\frac{\partial C_k(\mathbf{X})}{\partial \mathbf{X}}\right)^T\right]. \quad [A19]$$

Equation [A15] can therefore be rewritten as

$$\left\{\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)^T_{\mathbf{X^{(1)}}}\mathbf{S_E^{-1}}\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)_{\mathbf{X^{(1)}}} + \sum_{k=1}^{6}\lim_{s_k\to 0} \times\left[\frac{1}{s_k^2}\left(\frac{\partial C_k(\mathbf{X})}{\partial \mathbf{X}}\right)_{\mathbf{X^{(1)}}}\otimes\left(\frac{\partial C_k(\mathbf{X})}{\partial \mathbf{X}}\right)^T_{\mathbf{X^{(1)}}}\right]\right\}(\mathbf{X_0}-\mathbf{X^{(1)}}) = \left.\frac{\partial \ln p(\mathbf{X}|\mathbf{E}=\mathbf{E_0})}{\partial \mathbf{X}}\right|_{\mathbf{X^{(1)}}}. \quad [A20]$$

Equation [A20] is further transformed as

$$\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)^T_{\mathbf{X^{(1)}}}\mathbf{S_E^{-1}}\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)_{\mathbf{X^{(1)}}}(\mathbf{X_0}-\mathbf{X^{(1)}}) + \sum_{k=1}^{6}\lambda_k\left(\frac{\partial C_k(\mathbf{X})}{\partial \mathbf{X}}\right)_{\mathbf{X^{(1)}}} = \left.\frac{\partial \ln p(\mathbf{X}|\mathbf{E}=\mathbf{E_0})}{\partial \mathbf{X}}\right|_{\mathbf{X^{(1)}}}, \quad [A21]$$

where λ_k has the meaning

$$\lambda_k = \lim_{\substack{s_k\to 0\\ \mathbf{X}\to\mathbf{X_0}}}\left(\frac{\partial C_k}{\partial \mathbf{X}}\right)_{\mathbf{X^{(1)}}}\left(\frac{\mathbf{X_0}-\mathbf{X}}{s_k}\right) = \lim_{\substack{s_k\to 0\\ \mathbf{X}\to\mathbf{X_0}}}\left(\frac{C_k(\mathbf{X})}{s_k}\right). \quad [A22]$$

When Eqs. [A16] and [A22] are combined into one matrix equation

$$\begin{bmatrix}\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)^T_{\mathbf{X^{(1)}}}\mathbf{S_E^{-1}}\left(\frac{\partial \mathbf{E}}{\partial \mathbf{X}}\right)_{\mathbf{X^{(1)}}} & \left(\frac{\partial C_k(\mathbf{X})}{\partial \mathbf{X}}\right)^T_{\mathbf{X^{(1)}}}\\ \left(\frac{\partial C_k(\mathbf{X})}{\partial \mathbf{X}}\right)_{\mathbf{X^{(1)}}} & \mathbf{0}\end{bmatrix} \times\begin{bmatrix}(\mathbf{X_0}-\mathbf{X^{(1)}})\\ \boldsymbol{\Lambda}\end{bmatrix} = \begin{bmatrix}\left(\frac{\partial \ln p(\mathbf{X}|\mathbf{E}=\mathbf{E_0})}{\partial \mathbf{X}}\right)_{\mathbf{X^{(1)}}}\\ -\mathbf{C}(\mathbf{X^{(1)}})\end{bmatrix} \quad [A23]$$

then lead to Eq. [3] (main text).

ACKNOWLEDGMENTS

This work was supported by NIH Grants RR 02301 and GM 35976. The authors thank Dr. Charles G. Hoogstraten for stimulating discussions and Ahmad Abualsamid for assistance in calculations on OMTKY3, which were performed on IBM SP2 RISC computers of the DoIT Engineering Computer Center, University of Wisconsin-Madison.

REFERENCES

1. A. Kumar, R. R. Ernst, and K. Wüthrich, *Biochem. Biophys. Res. Commun.* **95,** 1 (1980).
2. S. Macura and R. R. Ernst, *Mol. Phys.* **41,** 95 (1980).
3. J. Keepers and T. James, *J. Magn. Reson.* **57,** 404 (1984).
4. B. Borgias and T. James, *J. Magn. Reson.* **79,** 493 (1988).
5. S. Edmondson, N. Khan, J. Shriver, J. Zdunek, and A. Graslund, *Biochemistry* **30,** 11271 (1991).
6. G. M. Crippen, "Distance Geometry and Conformational Calculations," Wiley, New York (1981).
7. G. M. Crippen and T. F. Havel, "Distance Geometry and Molecular Conformation," Wiley, New York (1988).
8. J. W. Shriver and S. Edmondson, *Methods Enzymol.* **240,** 415 (1994).
9. P. Yip and D. A. Case, *J. Magn. Reson.* **83,** 643 (1989).
10. M. J. Dellwo and J. Wand, *J. Biomol. NMR* **3,** 205 (1993).
11. R. B. Altman and O. Jardetzky, *Methods Enzymol.* **177,** 218 (1989).
12. J. W. Shriver and S. Edmondson, *Biochemistry* **32,** 1610 (1993).
13. D. Taupin, "Probabilities, Data Reduction, and Error Analysis in the Physical Sciences," Les Editions de Physique, Les Ulis Cedex (1988).
14. A. A. Clifford, "Multivariate Error Analysis," Wiley, New York (1973).
15. A. T. Brünger, "X-PLOR, Version 3.1: A System for X-ray Crystallography and NMR," Yale Univ. Press, New Haven (1987).
16. S. Wolfram, "The MATHEMATICA Book, Version 3," Cambridge Univ. Press, New York (1996).
17. B. L. de Groot, D. M. F. van Aalten, R. M. Scheek, A. Amadei, G.

Vriend, and H. J. C. Berendsen, *Proteins Struct. Funct. Genet.* **29,** 240 (1997).

18. ML software package, available from the authors; information may be obtained from the World Wide Web (http://www.nmrfam.wisc.edu).

19. K. Wüthrich, "NMR of Proteins and Nucleic Acids," Wiley, New York (1986).

20. O. Jardetzky, M. G. Clore, D. Hare, and A. Torda, *in* "Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy" (J. C. Hoch, F. M. Poulsen, and C. Redfield, Eds.), pp. 375–389, Plenum, New York (1991).

21. P. R. Bevington, "Data Reduction and Error Analysis for the Physical Sciences," McGraw–Hill, New York (1969).

22. D. Zhao and O. Jardetzky, *J. Mol. Biol.* **239,** 601 (1994).

23. S. Macura, *J. Magn. Reson. B* **104,** 168 (1994).

24. S. Macura, *J. Magn. Reson. A* **112,** 152 (1995).

25. S. Macura, J. Fejzo, C. G. Hoogstraten, W. M. Westler, and J. L. Markley, *Israel J. Chem.* **32,** 245 (1992).

26. C. G. Hoogstraten, W. M. Westler, S. Macura, and J. L. Markley, *J. Am. Chem. Soc.* **117,** 5610 (1995).