Defining the Precision with Which a Protein Structure Is Determined by NMR. Application to Motilin†

John Shriver* and Stephen Edmondson

Department of Medical Biochemistry, School of Medicine, Southern Illinois University, Carbondale, Illinois 62901

Received August 11, 1992; Revised Manuscript Received November 24, 1992

ABSTRACT: A simple procedure is introduced for accurately defining the precision with which the Cartesian coordinates of any macromolecular structure are determined by nuclear Overhauser data. The method utilizes an ensemble of structures obtained from an array of independent simulated data sets derived from a final structure. Using the noise-free, back-calculated NOE spectrum as the "true" NOE spectrum, simulated Monte Carlo data sets are created by superimposing onto the "true" spectrum Gaussian distributed noise with a standard deviation equal to that of the residuals. Full relaxation matrix refinements of the simulated data sets provide probability distributions of the Cartesian coordinates for each atom in the model. Molecular dynamics simulations are included to estimate the effect of sparse information on the precision. The procedure is applied here to the 22-residue peptide hormone motilin, and the results are compared to those obtained using the conventional method of analyzing multiple refinements using a single distance constraint set. The average root mean square deviation for α-carbon atoms in the central portion (Arg12–Arg18) of the single helix of motilin was determined to be 0.72 Å by the Monte Carlo method, compared to 1.3 Å determined by an analysis of the 10 best DIANA structures using the same number of constraints between the same atoms. The origin of the bias of the conventional method is discussed.

Rapid progress has been made in the last few years in developing techniques for obtaining the solution structures of proteins and nucleic acids from NMR[1] data with a precision that appears to rival that obtained by X-ray crystallography (Braun, 1987; Gippert et al., 1990; Havel, 1991; Wüthrich, 1986, 1990). As with any model derived from a fitting of quantitative experimental data, if meaningful conclusions are to be drawn from the results, it is necessary not only to obtain the model parameters from the fitting process but also to accurately characterize the precision with which the parameters are defined by the data. In the case of structure determinations, the precision of the atomic coordinates in the form of a three-dimensional confidence probability distribution is required. This is especially true if the structural information is to be used with confidence to discuss topics such as ligand binding, catalysis, packing densities, solvent accessibility, electrostatic interactions, and stability.

Precision should be distinguished from accuracy. Each refers to different aspects of an experimental measurement or data analysis, although they are often used interchangeably, sometimes by the same author. We adopt the definitions given by Bevington (1969), i.e., the *accuracy* of an experiment or parameter estimation is a measure of how close the result agrees with the true value, and the *precision* is a measure of how exactly the result is determined. The question of the accuracy of a structure determined by NMR has been addressed by numerous workers (Braun & Go, 1985; Clore et al., 1986; Duben & Hutton, 1990; Fejzo et al., 1991; Havel, 1991; Havel & Wüthrich, 1985; Kaluarachchi et al., 1991; Landy & Rao, 1989; Lane, 1988; Pardi et al., 1988; Post et al., 1990) and can only be studied by using model structures and simulated spectra. The accuracy of a structure is not the focus of this paper.

Present methods for defining the precision of a structure derived from NMR data rely on executing repeated, independent fits of the same NOE data set using distance geometry, restrained molecular dynamics, or simulated annealing (Bax, 1989; Braun, 1987; Gippert et al., 1990; Wüthrich, 1989, 1990). Each fitting process is typically begun with a randomly generated structure in an effort to sufficiently sample conformational space. Identical constraints are applied to each starting structure; the constraints being either distance constraints derived from the NOE intensities or the NOE intensities themselves. Additional constraints from other types of information can also be incorporated, e.g., torsion angles from coupling constants, hydrogen bonds from proton exchange rates, and coordinating atom–metal distance constraints (Summers et al., 1992). The widths, or RMSD's, of the distributions of the atomic coordinates of each atom in the molecule calculated from an ensemble of final structures is used as a representation of the quality, or precision, of the structure.

Distance geometry cannot give a unique structure, but only a collection of structures which satisfy the constraints (Havel, 1991; Havel et al., 1979), and therefore it is not surprising that an overlay of independently determined structures was used to define the quality of the final model in the first paper describing a protein structure based on NMR constraints (Braun et al., 1983). The philosophy behind obtaining structures using distance geometry differs from that underlying a least-squares fitting of data (Havel & Wüthrich, 1984). In distance geometry, the data are viewed as being "imprecisely known but completely correct" (Havel, 1991) (i.e., the conversion of the NOE intensities to distance constraints is inaccurate, but the data are noise free); whereas in a least-squares fitting, the data are viewed as being "precisely known but subject to random measurement errors" (Havel, 1991)

† This work was supported by a research grant from the Biotechnology Research Development Corporation and by Southern Illinois University School of Medicine. J.W.S. is a recipient of a National Institutes of Health Research Career Development Award.

* Author to whom correspondence should be addressed.

[1] Abbreviations: DG, distance geometry; FIRM, iterative full relaxation matrix refinement program; MD, molecular dynamics; MM, molecular mechanics; NMR, nuclear magnetic resonance; NOE, nuclear Overhauser enhancement; NOESY, nuclear Overhauser enhancement spectroscopy.

0006-2960/93/0432-1610$04.00/0

(i.e., the NOE data and derived distances are subject to error due to noise). Because of the inaccuracies associated with converting NOE's to distances (due to spin diffusion, anisotropic rotations, flexibility, inaccurate spectral density functions, etc.), precise distance constraints are commonly not used in distance geometry or restrained molecule dynamics refinements of NMR structures. A distance is calculated by assuming that the system is composed of pairs of isolated, two-spin interactions (i.e., neglecting spin diffusion) using an internal calibration distance such as the tyrosine H5–H6 ring protons. The resulting distances are typically grouped into user-defined classes of distance bounds: for example, strong (1.8–2.7 Å), medium (1.8–3.3 Å), and weak (1.8–5.0 Å) classes (Nilges et al., 1988; Williamson et al., 1985). Protons within each class are permitted to move freely during the refinement within these ranges with little or no penalty. The imprecision is built into the beginning of the analysis, rather than allowing the quality of the fit define it. *Such treatments clearly remove any contribution that noise in the data might make to the imprecision, and the only way to improve the quality of the fit is to increase the number of constraints.*

In order to achieve a more accurate measure of the precision with which a structure is determined, it is advantageous to preserve the integrity of the data and perform a quantitative fitting of the NOE data directly without first converting it into broad distance constraints or bounds. Even if the NOE intensities could be converted into discrete distances, the transformation results in a conversion of the normally distributed noise in the data into a non-Gaussian distribution. Such a transformation violates one of the central requirements for reliably performing a fit driven by an attempt to minimize the sum of the squares of the residuals (Johnson & Faunt, 1992).

Another requirement for obtaining an accurate measure of the precision is to base the calculation on the magnitude of the NOE residuals, i.e., the differences between the experimental NOE intensities and those back-calculated using the final structure. A calculation of the residuals and a χ^2 are central to the fitting of data in other disciplines where the magnitude of the residuals serves as a quantitative measure of the overall goodness of fit; they are used to drive the adjustments of the parameters in the proper direction (so as to minimize the residuals), and they determine when convergence has been achieved.

Obtaining the precision of a structure by comparing a collection of independent fits of the same data set can be expected to be biased for the reasons shown in Figure 1. (This example is intentionally exaggerated to clearly portray the point.) The data are depicted by the asterisks (*), and 10 possible fits that might have resulted from an iterative, nonanalytical fitting procedure starting with 10 random initial choices for the parameters are indicated with the solid lines. In this example, the fits have all converged well away from the original data. If the precision were calculated by comparison of the ensemble of fits, it might be incorrectly concluded that the quality of the fit was good since the magnitude of the residuals is not considered.

Another difficulty with the presently accepted method of defining the quality of structures stems from the dependence of the apparent precision of the structure on the relative and absolute magnitudes of the weights placed on the constraints, e.g., the weighting factors in the distance geometry objective function (Havel, 1991; Kuntz et al., 1989) or target function (Güntert et al., 1991) or the force constants used in the simulated annealing or molecular dynamics (Nilges et al., 1988; Scheek et al., 1989). (The weighting factors in some distance geometry programs are not written explicitly and are therefore equal to 1). Using increasingly larger values for the weights or force constants results in decreasing the RMSD's of the atomic coordinates of the fitted structures. The choice of the magnitude of these parameters (and therefore the precision) is subjective, and while some authors have argued that particular values of the weights are optimal, it could just as easily be argued that a fit of the data should be accomplished with the highest emphasis on the data, and essentially no bias from the procedure should be present in the results of the fit (e.g., the force field in the dynamics should have little influence on the result).

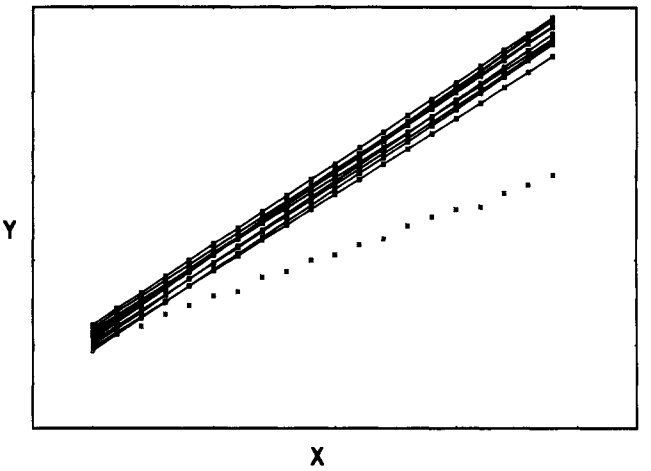

FIGURE 1: Graphical representation of the origin of the bias associated with defining the precision of a model based on the width of the distribution of the parameters obtained by repeated fitting of the same data set. The experimental data are indicated with asterisks, and the results of 10 (nonanalytical) fits are shown with the open squares connected by straight lines. The widths of the distributions of parameters defining the fitted curves cannot be used to define the precision of the parameters since they will not reflect the magnitude of the residuals.

We define the precision with which the atomic coordinates are determined by NMR by using the Monte Carlo algorithm for measuring confidence interval probability distributions as described by Press et al. (1989) [see also Bard (1974) and Straume and Johnson (1992)]. The structure obtained from the data via a full relaxation matrix refinement is accepted as the best representation of the "true" structure. In the algorithm used here, the precision is derived from the noise of the fit characterized by the standard deviation of the NOE residuals. An ensemble of NOE data sets is simulated by superimposing noise of the same magnitude as the standard deviation of the fit on the NOE spectrum back-calculated from the "true" structure. Each of the artificial data sets contains the same number of NOE's between the same proton pairs. Fitting each of these data sets gives a set of results which differ due to the noise of the fit and no other factor. The distribution of each of the atomic positions resulting from fitting the ensemble of simulated data sets represents the confidence limits or precision of the coordinates of each atom.

MATERIALS AND METHODS

Structure Determination. Two-dimensional NOESY spectra of motilin (22 residues, MW 2700) in 30% hexafluoroisopropanol have been presented elsewhere (Khan et al., 1990). Initial structures were obtained from the NOE constraints by distance geometry [DIANA (Güntert et al., 1991)] and restrained molecular dynamics [AMBER (Weiner & Kollman, 1981; Weiner et al., 1984)] using distance

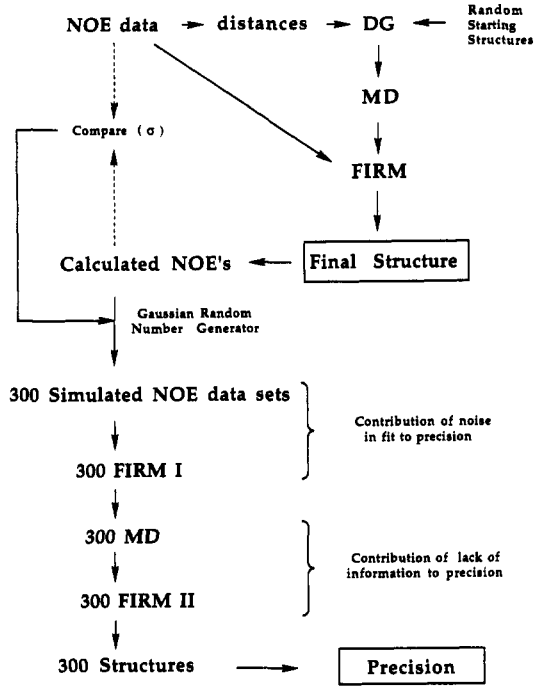

FIGURE 2: Protocol used here to obtain the precision of a solution structure of a macromolecule from NOE intensities. The final structure is obtained by using a combination of distance geometry (DG), restrained molecular dynamics (MD), and a full relaxation matrix refinement (FIRM). The back-calculated NOE spectrum is compared with the experimental data to obtain a standard deviation of the residuals. An ensemble of simulated data sets is calculated using a Gaussian random number generator. Each of these are fit using a full relaxation matrix refinement to obtain the imprecision due to noise and the inadequacy of the fit. Further refinement using molecular dynamics at elevated temperature followed by a final full relaxation matrix refinement allows an estimation of the contribution to the imprecision from a lack of information.

constraints obtained by using the two-spin approximation, as detailed in an earlier publication (Edmondson et al., 1991). One of the DG/MD structures was further refined (Edmondson et al., 1991) by using an iterative full relaxation matrix analysis (Boelens et al., 1989; Borgias et al., 1990; Borgias & James, 1990; Post et al., 1990) using the program FIRM (Edmondson, 1992) by iteratively substituting experimental NOE's into the calculated volume matrix, back-calculating distances from the "hybrid" relaxation matrix, and energy-minimizing the structure with these distance constraints now corrected for spin diffusion. This final FIRM-refined structure had an NOE R factor of 0.33 and serves as the final "true" structure for which we would like to know the precision. (The noise-free NOE spectrum back-calculated from the "true" structure can be referred to as the "true" NOE spectrum and should not be confused with the experimental data). Note that we reported an R factor of 0.27 in Edmondson et al. (1991), where we were using a slightly larger set of NOE's and an earlier version of FIRM which treated methyl group rotation in a different manner from that used here.

Precision Determination. The procedure used here for defining the precision of the fit is derived from that described by Press et al. (1989) and also Bard (1974), Straume and Johnson (1992), Kamath and Shriver (1989), and Motulsky and Ransnas (1987). It is summarized graphically in Figure 2. In this report, the final "true" structure obtained from a combined DG, MD, and full relaxation matrix refinement was used to back-calculate a set of NOE's using the program FIRM (Edmondson, 1992). Isotropic rotation of the molecule was assumed, with a rotational correlation time of 2 ns

Table I: Summary of Constraint Violations, Energies, and R Factors following Refinements of 300 Monte Carlo Data Sets[a]

	sum of violations (Å)	mean violation (Å)	E_{tot}[e] (kcal/mol)	E_{con}[f] (kcal/mol)	R_{NOE}	R_d
FIRM I[b]	1.40 (0.30)	0.03 (0.006)	−59 (54)	78 (23)	0.27 (0.04)	0.12 (0.03)
MD[c]	0.61 (0.14)	0.01 (0.003)	−200 (39)	48 (21)	0.38 (0.05)	0.13 (0.02)
FIRM II[d]	2.09 (0.71)	0.04 (0.01)	−123 (71)	61 (27)	0.31 (0.05)	0.12 (0.03)

[a] Results are the averages obtained from the first 20 structures from the set of 300, with the standard deviations in parentheses. [b] Results from the first FIRM refinement of the Monte Carlo data sets. [c] Results from restrained molecular dynamics refinements of the first FIRM refinement structures. [d] Results from the final FIRM analysis of the MD structures. [e] E_{tot} is the total potential energy at the structures following minimization with AMBER including the contribution made by the constraints. [f] E_{con} is the constraint contribution to the total energy of the structures following minimization.

throughout. A 12-site jumping model was used to simulate rotating methyl groups (Tropp, 1980), and a correlation time of 0.1 ns was assumed for the methyl rotors. The flipping motions of aromatic rings were simulated by using an r^{-6} averaging as described by Koning et al. (1990). Since only experimental NOE's involving backbone protons are considered in this report, errors arising from an approximate treatment of methyl group and aromatic ring rotations enter indirectly and are negligible.

Simulated NOE data sets containing normally distributed random noise superimposed on the noise-free back-calculated "true" NOE spectrum were constructed by using a Gaussian random number generator algorithm described by Miller (1987) which requires as input the "true" NOE values (which will become the mean of the generated distribution) and the standard deviation desired for the distribution. The standard deviation was derived from the residuals between the experimental NOE's and those back-calculated from the FIRM-refined model (viz., 0.011). NOE's that became negative after adding the noise were set to zero intensity, since only distances were allowed to fluctuate during the refinement, and the rotational correlation time was held constant. A total of 300 simulated NOE data sets were constructed in this fashion and iteratively substituted into the starting model's NOE matrix using the program FIRM. This FIRM refinement procedure was repeated until the NOE R factor was less than 0.05. The distances in the "hybrid" relaxation matrix were then translated into an actual structure by energy minimization of the "true" motilin model, which was subjected to 200 cycles of restrained molecular mechanics (MM) using a flat-well potential with an NOE force constant (K_{NOE}) of 100 kcal/Å^2. The average constraint violation was less than 0.03 Å for the first 20 structures from the set of 300. The average total potential energy for these structures was −137 kcal/mol if the constraint contribution was removed and −59 kcal/mol if the constraint contribution was included (Table I). The spread in the coordinates of the 300 structures reflects the contribution of noise in the NOE data and the inadequacy of the model to the imprecision of the structure.

In order to assess the contribution of sparse information to the imprecision of the structure determination, the 300 FIRM-refined structures were subjected to restrained molecular dynamics. Each was subjected to 5 ps (5000 steps) of restrained molecular dynamics (MD) without SHAKE at 300 K using K_{NOE} = 500 kcal/Å^2. This allowed increased conformational sampling by those atoms not constrained while

very tightly holding the constrained atoms in place. This was followed by 200 cycles of restrained molecular mechanics with K_{NOE} = 100 kcal/Å^2 to minimize the energy. The resulting set of 300 structures was further refined with another round of FIRM refinement and 200 cycles of restrained MM as described above. A summary of the energies and constraint violations of the structures at the end of MD and FIRM is given in Table I.

MM and MD calculations described above were performed with AMBER 4.0 (Pearlman et al., 1991) modified to incorporate a flat-well potential for NOE distance constraints (Edmondson et al., 1991). A pseudo/united atom force field with reduced charges on ionizable side chains was used as described previously (Edmondson et al., 1991). Since only NOE's involving backbone protons were included in this analysis, pseudoatom correction factors were required only for glycine residues. Except for these glycine residues, the sizes of the flat-wells were given by the fractional errors in the "hybrid" NOE's estimated as previously described [e_{ij} in Edmondson (1992)]. In general, the flat-wells were less than 0.1 Å wide and thus could be considered harmonic. However, distances back-calculated from simulated NOE's with near zero intensity had greater fractional error. To minimize the influence of NOE's with near zero intensity, the maximum values of the lower and upper distance bounds were set to 4.5 and 99 Å, respectively.

Two different R factors are utilized here for comparison of experimental and back-calculated NOE intensities (Baleja et al., 1990; Baleja & Sykes, 1991; Edmondson, 1992; Gonzalez et al., 1991; Nikonowicz et al., 1990). The R factors are defined similarly to the normalized mean deviation of structure factors used in X-ray diffraction, with R_{NOE} using the NOE intensities directly, and R_d using the inverse sixth power of the NOE intensities (so that it reflects more accurately differences in distances):

$$R_{NOE} = \frac{\sum|NOE_o - NOE_c|}{\sum|NOE_o|}$$

$$R_d = \frac{\sum||NOE_o|^{-1/6} - |NOE_c|^{-1/6}|}{\sum|NOE_o|^{-1/6}}$$

where NOE_o is the experimentally observed NOE intensity and NOE_c is the back-calculated NOE intensity using the model structure.

Structures were visualized on a Silicon Graphics 4D25G graphics workstation using MIDASPlus from the Computer Graphics Laboratory, University of California, San Francisco (Ferrin et al., 1988).

The Monte Carlo procedure presented here for determining confidence probability distributions for atomic coordinates should not be confused with the Monte Carlo search procedure of Bassolino et al. (1988) and Levy et al. (1989), which is used to effectively scan conformational space during a distance geometry refinement.

RESULTS AND DISCUSSION

A theoretical analysis of the precision of a least-squares fit is reliable only for linear models since an analytical solution is not possible (Bard, 1974; Bevington, 1969; Johnson & Faunt, 1992; Motulsky & Ransnas, 1987). An analysis based on the parameter distributions obtained from the results of numerous (e.g., 100) experiments is the most straightforward method for defining precision, but this is not practical for NMR structure determinations. The Monte Carlo procedure for determining the confidence probability distributions for parameters obtained from a nonlinear least-squares fitting of experimental data is a brute force method which is conceptually simple and reliable (Bard, 1974; Press et al., 1989; Straume & Johnson, 1992). It requires only an accurate estimate of the noise of the experimental data and an accurate model for fitting the data (Straume & Johnson, 1992). It provides the most complete description of the probability distributions of all of the parameters of the fit, with a resolution limited by the number of simulated, artificial data sets analyzed, and it intrinsically takes full account of all correlations between parameters.

We have previously determined the solution structure of the 22-residue gastrointestinal peptide hormone motilin (MW 2700) by a combination of distance geometry (DIANA), restrained molecular dynamics (using AMBER), and a full relaxation matrix refinement (using FIRM) (Edmondson et al., 1991). We address the question here of how precisely this structure has been determined. The full relaxation matrix refined structure is accepted as the "true" structure, i.e., the best representation of the real structure of motilin given the data and the assumptions of the model (e.g., a single rotational correlation time for the overall motion of the molecule). Note that it is not necessary that the "true" structure be the actual, real structure in order to evaluate the precision of the fit (Press et al., 1989).

Using the methods described in Edmondson et al. (1991), we calculated a noise-free NOE spectrum based on the "true" structure of motilin. The deviations (residuals) between the calculated, noise-free "true" spectrum and the experimental NOE data represent the noise of the fit. A comparison of the calculated and experimental NOE intensities has been presented previously [Figure 7 of Edmondson et al. (1991)], and a complete listing of the calculated NOE's, experimental NOE's, and residuals is provided as supplementary material. At first glance, it appears that the differences between the experimental and calculated NOE's are a function of the value of the NOE. Indeed this might be expected since small position differences that occur during the energy minimization after the FIRM will lead to larger errors for the larger NOE's due to the $1/r^6$ dependence of the NOE on distance. The largest deviations involve NOE's between backbone protons in the Thr6–Gly8 region (Figure 3), where considerable flexibility is expected in solution (Edmondson et al., 1991). We therefore treat the error as independent of the NOE magnitude and calculate a standard deviation of the fit using equal weights for all of the NOE's. Future work with a larger protein containing a larger set of NOE's may require weighting of the individual deviations in the standard deviation calculation.

The overall goodness of fit can be described by the standard deviation of the residuals. (The distribution of the residuals or deviations is shown in Figure 1 of the supplementary material.) A fit of the residual distribution to a Gaussian gives a standard deviation of 0.0065, whereas the calculated σ is 0.011. Given the noise in the distribution, we have chosen to use the calculated rather than the fitted value of σ as the value which characterizes the noise in the fit. The standard deviation characterizing the noise in the data (obtained by integration of many peak-free regions of the data) was 0.00169, or a factor of 3.8–6.5 smaller than that characterized by the residuals of the fit. The origin of this discrepancy is not clear, but given its size it is most likely real and indicates an inadequacy of the model used to fit the data. This is not surprising considering that we are assuming a single confor-

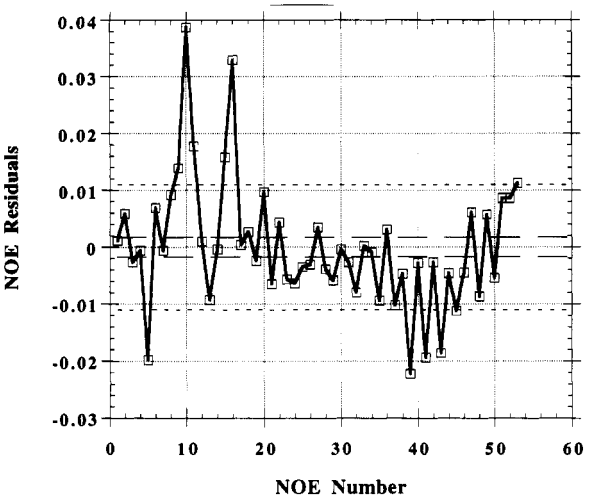

FIGURE 3: NOE residuals (i.e., calculated NOE − experimental NOE) derived from the final, "true" structure of motilin. The first 20 NOE's involve interactions between backbone protons in the 10 amino-terminal residues (Phe1–Leu10), whereas the others arise from backbone interactions in the 13 carboxy-terminal residues (Glu9–Gln22). (A complete list of the calculated and experimental NOE's is provided as supplementary material.) The standard deviation of the residuals (viz., 0.011) is indicated by the short dashed line, and the standard deviation of the NOE spectral noise (viz., 0.00169) is indicated by the long dashed line.

mation and a single rotational correlation time for a linear peptide of 22 residues. In this work we have chosen to use the larger value for the noise in order to obtain the most conservative estimate of the precision. In the future, it will be interesting to determine if the standard deviation of the fit is more similar to that of the experimental noise for a larger, globular protein whose structure is defined by an interconnecting network of NOE's.

Artificial data sets were constructed which were distributed normally about the "true" NOE spectrum with a noise level equal to the standard deviation of the fit. Using a Gaussian random number generator, we generated 300 new data sets, each containing NOE intensities between the same proton pairs as in the original experimental data. (As an example, the distribution of the 300 NOE intensities generated for the NH to $C_\alpha H$ interaction of Thr6 is shown in Figure 2 of the supplementary material. A smooth curve through the distribution represents a Gaussian fit with a mean of 0.033 and a σ of 0.0126, reasonably close to the input "true" NOE of 0.0345 and a σ of 0.011.) Any of these data sets could be expected to represent a legitimate data set that might be obtained if additional NOESY experiments were performed and the final model were a true representation of the solution structure.

The fitting of the 300 Monte Carlo simulations was performed using a full relaxation matrix analysis with the "true" structure as the starting structure, i.e., the single "true" structure was refined with 300 different "data" sets. The distribution of structures obtained from the fits represents the precision with which the atomic coordinates can be specified due to noise in the fit. An overlay of the first 10 structures from the set of 300 is shown in Figure 4. It can be seen that the structure remains well defined even after introduction of considerable noise. RMSD's for the α-carbons (calculated using all 300 structures) range from 0.16 to 0.56 Å (Table II). In the central part of the α-helical region (Arg12–Arg18) the average RMSD for the C_α carbons is 0.20 Å. A closer look at a specific locus is provided by a three-dimensional plot of all of the 300 vectors between each of the C_α carbon atoms of Glu15 and the mean of the distribution (Figure 5). The distribution can be seen to be quite narrow with an average spread of about 0.1–0.2 Å, and some asymmetry is clear. It is not necessary that the distributions of the parameters be a normal distribution, and resolution of the shape of the distribution could be increased if necessary by analyzing a larger number of Monte Carlo data sets. In three dimensions, the distribution is not spherical, and ellipsoids could be used to describe the confidence probability distribution of each parameter. The degree of asymmetry does not appear to be significant enough to preclude the use of a single, overall RMSD to characterize the distribution width for our purposes here. If necessary, cross-sections through the confidence probability distribution can be used to more clearly describe the shape and asymmetry of the distribution (Figure 6).

One important contribution lacking in the above analysis is the influence of insufficient data on the precision of the result. In a full relaxation matrix refinement, those regions of the molecule which are ill defined (i.e., for which there are few NOE constraints) will change negligibly during the refinement. This leads to an overestimation of the precision (smaller RMSD's) in regions of the molecule lacking constraint information than is warranted by the data. Most of the apparent movement of these atoms is the result of the alignment necessary to calculate RMSD's if the alignment region includes atoms that have experienced extensive position adjustments during the refinement.

Incorporation of the effect of a lack of information on the precision of the result is accomplished here by performing a restrained molecular dynamics refinement on each of the 300 FIRM structures to allow increased sampling of conformational space, followed by a second FIRM refinement to optimize the fits (Figure 2). The confidence probability distributions obtained from the resulting structures are essentially spherical, and the RMSD's are reported in Table II. There is a sizable increase in distribution widths of atoms in the amino terminus compared to those between residues 12 and 18, consistent with there being few constraints in the amino terminus. An increase is also observed for the distribution widths in the helical region from Glu10 to Lys20, indicating some contribution from a lack of information here also. The average R factor for the final ensemble of 300 structures (0.31) is essentially the same as for the "true" structure (0.33), indicating that the array of structures are satisfactory "fits" of the data.

The RMSD's for the α-carbon atom positions at the end of the Monte Carlo analysis reported in column four of Table II represent the most accurate definition available of the precision with which the solution structure of motilin is defined by the NMR data. A full accounting is taken of the noise and incompleteness of the data and the inadequacy of the model. Over the central portion of the α-helix (Arg12–Arg18) the precision ranges from 0.57 to 0.93 Å with a mean of 0.72 Å. For comparison, the precision of a motilin structure determination was also calculated using the conventional method, i.e., comparing the family of the best structures obtained from repeated refinements with a single set of constraints. Three hundred structures were calculated with DIANA using constraints between the same backbone proton pairs as used in the Monte Carlo analysis with the lower limits set to the sum of the van der Waals radii, the upper limits set to the isolated two-spin approximation distance plus 0.5 Å (Table I, Supplementary Material), and the minimization strategy was that determined by the DIANA program. RMSD's using the 10 best DIANA structures (i.e., the 10 with the 10 lowest

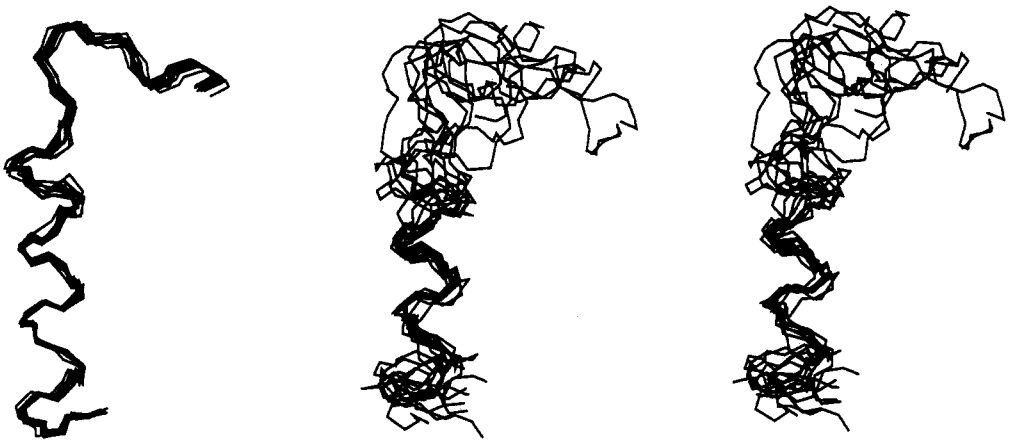

FIGURE 4: Overlay of 10 structures of motilin at three successive points in the precision analysis: after FIRM refinement of the 300 Monte Carlo data sets (left), after restrained molecular dynamics of the resulting 300 structures (center), and after a final FIRM refinement of the 300 MD structures (right). The N-terminus is at the top of each overlay. Only the first 10 structures of each of the sets of 300 are shown for clarity.

Table II: Distribution Widths (Expressed as Root Mean Square Deviations in Å) for the C_α Carbon Atom Positions of Motilin after Each of the Three Stages of the Monte Carlo Precision Analysis[a]

	RMSD's from Monte Carlo analysis		
C_α position	FIRM I[b]	MD	FIRM II
Phe1	0.54	7.11	7.07
Val2	0.53	6.93	6.90
Pro3	0.48	6.38	6.34
Ile4	0.56	6.52	6.47
Phe5	0.53	6.35	6.29
Thr6	0.41	5.70	5.63
Tyr7	0.37	4.68	4.61
Gly8	0.45	3.90	3.86
Glu9	0.40	3.29	3.12
Leu10	0.39	2.69	2.67
Gln11	0.44	1.74	1.69
Arg12	0.25	0.94	0.93
Met13	0.21	0.75	0.72
Gln14	0.27	0.62	0.62
Glu15	0.18	0.68	0.68
Lys16	0.18	0.71	0.69
Glu17	0.16	0.60	0.57
Arg18	0.16	0.86	0.82
Asn19	0.26	1.77	1.71
Lys20	0.30	2.21	2.15
Gly21	0.33	2.63	2.61
Gln22	0.23	2.51	2.47

[a] Three hundred structures within each group were aligned by minimizing the deviations of the C_α carbon atom positions from Arg12 through Arg18. [b] The RMSD's at the end of each of the three stages of the Monte Carlo precision analysis summarized in Table I are calculated using all 300 structures.

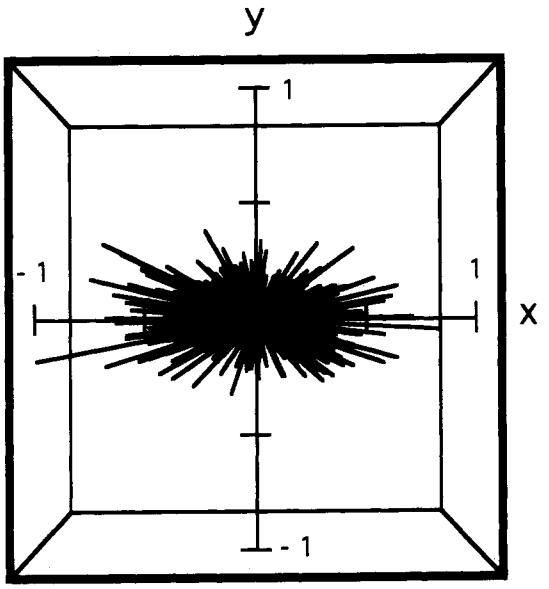

FIGURE 5: Three-dimensional display of the positions of the C_α proton of Glu15 relative to the mean position (center of the cube) for the 300 structures obtained after a full relaxation matrix refinement using the 300 simulated data sets. (Each side of the cube is 2 Å in length).

target functions) ranged from 0.94 to 1.9 Å (mean = 1.3 Å) over the same region (Arg12–Arg18). [Essentially identical results (within 0.1 Å) were obtained if the constraints were grouped into three constraint classes of 2.0–2.5, 2.0–3.5, and 2.0–4.5 Å.] The size and magnitude of the bias of the conventional estimate of the precision can be attributed to the various factors discussed in the introduction. Since the magnitude of the bias cannot be predicted, the Monte Carlo method should be used to obtain an accurate measure of the precision.

A precision level of 0.72 Å is surprisingly good considering the size of motilin and the small number of constraints. The single extended helix in motilin is defined by NOE interactions that extend largely along the axis of the helix. A network of interlocking NOE's in the interior of a larger, globular protein may reduce the contribution of the lack of information to a level significantly less than that presented here so that it becomes negligible compared to that caused by the noise in the fit. If this were the case, the true precision of an NMR structure could easily approach 0.2 Å, the level obtained for a refined X-ray structure and considerably better than the 0.5-Å precision derived from independent X-ray structures of homologous proteins (Janin, 1990).

A number of workers have argued that the quality of the NOE data is not of major importance in determining the accuracy or the precision of the structure obtained by NMR [see Kuntz et al. (1989)]. However, Kaluarachchi et al. (1991) have demonstrated that the precision with which NOESY volumes are measured significantly affects the precision of the final structure when using a hybrid full relaxation matrix/restrained molecular dynamics procedure. The results presented here indicate that if the assessment of the precision of the structure is based on the noise of the fit, then it is important to carefully integrate each NOE cross peak to obtain the most precise NOE constraints.

The calculation of the above distributions requires the alignment of the final ensemble of refined structures due to rotation and translation of the molecule as a whole during

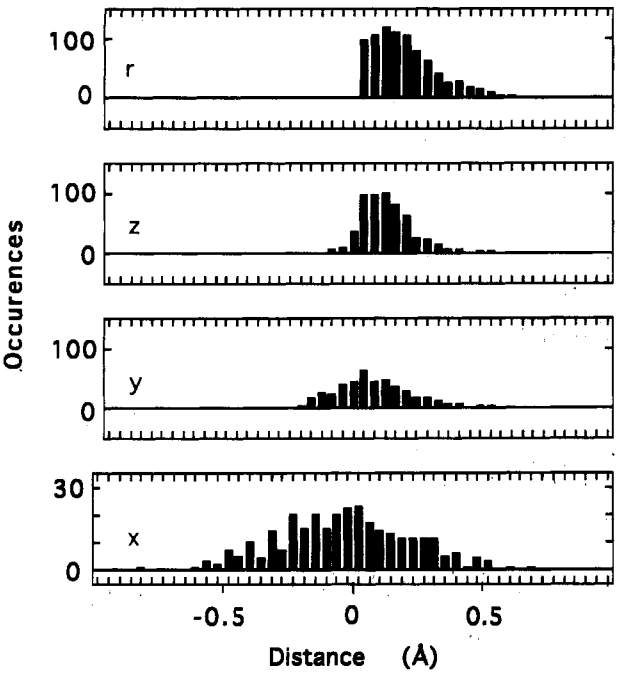

FIGURE 6: Orthogonal slices through the three-dimensional confidence probability distribution for the Cartesian coordinates of the C_α proton of Glu15. The probability distribution of the distance from the mean is shown in the top panel.

Table III: Distances between the i and i+3 C_α Carbon Atom Positions of Motilin after Each of the Three Stages of the Monte Carlo Precision Analysis[a]

C_α positions (i,i+3)	distances (Å)		
	FIRM I[a]	MD	FIRM II[a]
1,4	8.57 (0.51)[b]	7.37 (1.00)	7.34 (1.00)
2,5	8.35 (0.55)	7.04 (1.23)	7.04 (1.22)
3,6	7.57 (0.48)	7.05 (1.22)	7.05 (1.19)
4,7	5.56 (0.43)	6.33 (0.90)	6.35 (0.89)
5,8	8.53 (0.53)	8.21 (0.95)	8.18 (0.96)
6,9	7.18 (0.44)	7.00 (1.23)	6.99 (1.23)
7,10	6.09 (0.41)	6.34 (1.46)	6.31 (1.44)
8,11	4.86 (0.38)	5.26 (0.72)	5.31 (0.73)
9,12	5.29 (0.46)	5.74 (1.00)	5.50 (0.95)
10,13	4.78 (0.41)	5.89 (1.05)	5.83 (1.04)
11,14	5.14 (0.47)	4.99 (0.66)	5.06 (0.67)
12,15	4.85 (0.39)	4.96 (0.81)	5.01 (0.80)
13,16	4.75 (0.36)	4.76 (0.65)	4.73 (0.66)
14,17	5.32 (0.42)	5.35 (0.56)	5.01 (0.57)
15,18	5.17 (0.33)	5.08 (0.76)	5.12 (0.75)
16,19	5.59 (0.40)	5.45 (0.81)	5.39 (0.78)
17,20	5.96 (0.42)	5.52 (0.85)	5.50 (0.80)
18,21	5.45 (0.35)	4.93 (0.94)	4.91 (0.98)
19,22	6.65 (0.39)	6.38 (0.79)	6.34 (0.79)

[a] FIRM I, MD, and FIRM II refer to the three stages of the Monte Carlo precision analysis summarized in Table I. The distances reported are the averages of 300 structures, and the standard deviations are given in parentheses.

refinement. The choice of the alignment region can contribute to the magnitude of the root mean square deviations of the atomic positions, especially in a molecule the size of motilin. In order to circumvent this problem, we also investigated the distribution of selected distances rather than positions within the structure. In fact, this is most likely the kind of information which would be required in a typical application of structural data. Table III reports the mean and standard deviations of distances between the C_α carbons at positions i and i+3. The average distribution width for the i,i+3 distances in the central part of the α-helical region of motilin (i.e., i,i+3 interactions 12–15, 13–16, 14–17, and 15–18) is 0.70 Å, somewhat smaller than the value predicted using a standard deviation for the atom positions [i.e., $\sqrt{(0.72)^2 + (0.72)^2} = 1.02$ Å].

As in any fitting of experimental data, the equations and assumptions in the model chosen for the fitting will affect the accuracy and the precision of the fit. Inaccuracies due to assuming a single solution conformation and a single rotational correlation time are certainly present. Systematic errors will result from incorrect NMR peak assignments and T_1 noise. In addition, inaccuracies occur due to assumptions made on how to model methyl group rotations (i.e., the methyl group correlation time) when attempting to back-calculate an NOE spectrum (Edmondson, unpublished results; Koning et al., 1990; Tropp, 1980). The effect of each of these can be reliably assessed by investigating the effect of their modification on the precision of the fit.

The small peptide hormone motilin has been used to demonstrate the Monte Carlo method to allow for decreased computation time. The total time required for the combined MD/FIRM analysis was approximately 500 h of CPU time on a Silicon Graphics 4D25G Personal IRIS. For a protein of 60 amino acids, it is estimated that the Monte Carlo analysis on 300 simulated data sets as outlined here would require approximately 150 days of CPU time on a Personal IRIS. In some preliminary calculations, we have found that although the distributions become quite noisy, as few as 25 data sets can accurately represent a Gaussian distribution, as determined by the close agreement between the standard deviation and mean input into the Gaussian generator with those obtained by fitting the resulting distribution. For larger molecules where computational time may be limited, we estimate that 50–100 simulated data sets would probably be satisfactory for a reliable estimate of the precision. While this is greater than the number of structures commonly refined using a single data set [for a summary of recent structures, see Hendrickson and Wüthrich (1992)], it is comparable to the number generated in some of the more thorough recent studies [60 in Gronenborn et al. (1991), 48 in Kay et al. (1991), and 120 in Weber et al. (1991)].

ACKNOWLEDGMENT

We acknowledge many stimulating discussions with Dr. Wayne Bolen and Dr. Astrid Gräslund at various stages of this work.

SUPPLEMENTARY MATERIAL AVAILABLE

One table listing the backbone NOE's used for a precision analysis of the final structure of motilin and two figures showing the distribution of the residuals between the experimental and calculated NOE intensities derived from the final "true" structure of motilin and the distribution of the 300 NOE intensities generated for Thr6 NH to the C_αH proton (6 pages). Ordering information is given on any current masthead page.

REFERENCES

Baleja, J., & Sykes, B. (1991) *J. Magn. Reson. 91*, 624–629.
Baleja, J., Moult, J., & Sykes, B. (1990) *J. Magn. Reson. 87*, 375–384.
Bard, Y. (1974) *Nonlinear Parameter Estimation*, Academic Press, New York.
Bassolino, D., Hirata, F., Kitchen, D., Kominos, D., Pardi, A., & Levy, R. (1988) *Int. J. Supercomput. Appl. 2*, 41–61.
Bax, A. (1989) *Annu. Rev. Biochem. 58*, 223–256.
Bevington, P. (1969) *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, Inc., New York.

Boelens, R., Koning, T., van der Marel, G., van Boom, J., & Kaptein, R. (1989) *J. Magn. Reson. 82*, 290–308.
Borgias, B., & James, T. (1990) *J. Magn. Reson. 87*, 475–487.
Borgias, B., Gochin, M., Kerwood, D., & James, T. (1990) *Prog. Nucl. Magn. Reson. Spectrosc. 22*, 83–100.
Braun, W. (1987) *Q. Rev. Biophys. 19*, 115–157.
Braun, W., Wider, G., Lee, K., & Wüthrich, K. (1983) *J. Mol. Biol. 169*, 921–948.
Edmondson, S. (1992) *J. Magn. Reson. 98*, 283–298.
Edmondson, S., Khan, N., Shriver, J., Zdunek, J., & Gräslund, A. (1991) *Biochemistry 30*, 11271–11279.
Ferrin, T., Huang, C., Jarvis, L., & Langridge, R. (1988) *J. Mol. Graphics 6*, 13–27.
Gippert, G., Yip, P., Wright, P., & Case, D. (1990) *Biochem. Pharmacol. 40*, 15–22.
Gonzalez, D., Rullmann, J., Bonvin, A., Boelens, R., & Kaptein, R. (1991) *J. Magn. Reson. 91*, 659.
Gronenborn, A., Filpula, D., Essig, N., Achari, A., Whitlow, M., Wingfield, P., & Clore, G. (1991) *Science 253*, 657–661.
Güntert, P., Braun, W., & Wüthrich, K. (1991) *J. Mol. Biol. 217*, 517–530.
Havel, T. (1991) *Prog. Biophys. Mol. Biol. 56*, 43–78.
Havel, T., & Wüthrich, K. (1984) *Bull. Math. Biol. 46*, 673–698.
Havel, T., Crippen, G., & Kuntz, I. (1979) *Biopolymers 18*, 73–81.
Hendrickson, W., & Wüthrich, K. (1992) *Macromolecular Structures 1992*, Current Biology, Ltd., London.
Janin, J. (1990) *Biochimie 72*, 705–709.
Johnson, M., & Faunt, L. (1992) *Methods Enzymol. 210*, 1–37.
Kaluarachchi, K., Meadows, R., & Gorenstein, D. (1991) *Biochemistry 30*, 8785–8797.
Kamath, U., & Shriver, J. (1989) *J. Biol. Chem. 264*, 5586–5592.
Kay, L., Forman-Kay, J., McCubbin, W., & Kay, C. (1991) *Biochemistry 30*, 4323–4333.
Khan, N., Gräslund, A., Ehrenberg, A., & Shriver, J. (1990) *Biochemistry 29*, 5743–5751.
Koning, T., Boelens, R., & Kaptein, R. (1990) *J. Magn. Reson. 90*, 111–123.
Kuntz, I., Thomason, J., & Oshiro, C. (1989) *Methods Enzymol. 177*, 159–204.
Levy, R., Bassolino, D., Kitchen, D., & Pardi, A. (1989) *Biochemistry 28*, 9361–9372.
Miller, A. (1987) *Turbo BASIC Programs for Scientists and Engineers*, SYBEX, San Francisco.
Motulsky, H., & Ransnas, L. (1987) *FASEB J. 1*, 365–374.
Nikonowicz, E., Meadows, R., & Gorenstein, D. (1990) *Biochemistry 29*, 4193–4204.
Nilges, M., Clore, G., & Gronenborn, A. (1988) *FEBS Lett. 229*, 317–324.
Pearlman, D., Case, D., Caldwell, J., Seibel, G., Chandra Singh, U., Weiner, P., & Kollman, P. (1991) *AMBER 4.0*, University of California, San Francisco.
Post, C., Meadows, R., & Gorenstein, D. (1990) *J. Am. Chem. Soc. 112*, 6796–6803.
Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1989) *Numerical Recipes: The Art of Scientific Computing (Fortran Version)*, Cambridge University Press, Cambridge.
Scheek, R., van Gunsteren, W., & Kaptein, R. (1989) *Methods Enzymol. 177*, 204–218.
Straume, M., & Johnson, M. (1992) *Methods Enzymol. 210*, 117–129.
Summers, M., Henderson, L., Chance, M., Bess, J., South, T., Blake, P., Sagi, I., Perez-Alvarado, G., Sowder, R., Hare, D., & Arthur, L. (1992) *Protein Sci. 1*, 563–574.
Tropp, J. (1980) *J. Chem. Phys. 72*, 6035–6043.
Weber, C., Wider, G., von Freyberg, G., Traber, R., Braun, W., Widmer, H., & Wüthrich, K. (1991) *Biochemistry 30*, 6563–6574.
Weiner, P., & Kollman, P. (1981) *J. Comput. Chem. 2*, 287–303.
Weiner, S., Kollman, P., Case, D., Singh, U., Ghio, C., Alagona, G., Profeta, S., & Weiner, P. (1984) *J. Am. Chem. Soc. 106*, 765–784.
Williamson, M., Havel, T., & Wüthrich, K. (1985) *J. Mol. Biol. 182*, 295–315.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley and Sons, New York.
Wüthrich, K. (1989) *Science 243*, 45–50.
Wüthrich, K. (1990) *J. Biol. Chem. 265*, 22059–22062.