Journal of Biomolecular NMR, 5 (1995) 357–361
ESCOM

J-Bio NMR 253

Trajectory analysis of NMR structure calculations

Daisuke Kohda* and Fuyuhiko Inagaki

Department of Molecular Physiology, the Tokyo Metropolitan Institute of Medical Science, 18–22, Honkomagome 3-chome, Bunkyo-ku, Tokyo 113, Japan

Received 11 July 1994
Accepted 5 November 1994

Keywords: Trajectory analysis; Multidimensional scaling; Sampling problem; Optimization of protocols

Summary

NMR as well as X-ray crystallography are used to determine the three-dimensional structures of macromolecules at atomic resolution. Structure calculation generates coordinates that are compatible with NMR data from randomly generated initial structures. We analyzed the trajectory taken by structures during NMR structure calculation in conformational space, assuming that the distance between two structures in conformational space is the root-mean-square deviation between the two structures. The coordinates of a structure in conformational space were obtained by applying the metric multidimensional scaling method. As an example, we used a 22-residue peptide, μ-Conotoxin GIIIA, and a simulated annealing protocol of XPLOR. We found that the three-dimensional solution of the multidimensional scaling analysis is sufficient to describe the overall configuration of the trajectories in conformational space. By comparing the trajectories of the entire calculation with those of the converged calculation, random sampling of conformational space is readily discernible. Trajectory analysis can also be used for optimization of protocols of NMR structure calculation, by examining individual trajectories.

Introduction

The three-dimensional structure of a macromolecule determined by NMR is represented as a group of structures that are compatible with NMR data (Wüthrich, 1986). These structures can be obtained by using the metric matrix approach known as distance geometry, and by repeated minimization of a target function with the method known as restrained molecular dynamics (MD) from randomly generated initial structures. The latter method is based on Newton's equations of motion and thus generates time-dependent coordinates, i.e., a trajectory. The former method directly generates coordinates by diagonalizing a metric matrix, but refinement with restrained MD must be performed subsequently, which generates a trajectory. In this paper, we propose a 'Trajectory Analysis', which makes it possible to follow the trajectory taken by structures in conformational space during the restrained MD calculation. We assume that the distance between two structures in conformational space is the root-mean-square deviation (rmsd) between the two structures. This assumption has been used previously (Bruccoleri and Karplus, 1990). The coordinates of a structure in conformational space are obtained by application of a statistical method, multidimensional scaling (MDS; Kruskal and Wish, 1978), to a matrix comprising rmsd values for all pairs of the structures that were saved periodically during the restrained MD calculation. Metric MDS is based on a theorem of Young and Householder (Young and Householder, 1938; Torgerson, 1958) and it is identical to the embedding process of the metric matrix algorithm (Crippen and Havel, 1978) used in programs such as DISGEO and DSPACE. It should be noted that a similar method has been used to examine the conformers of a protein during MD simulation (Levitt, 1983a,b) and the converged structures generated with the program EMBOSS (Nakai et al., 1993), but we extended the analysis to cover the entire NMR structure calculation.

We chose a 22-residue peptide, μ-Conotoxin GIIIA, as a model macromolecule (Lancelin et al., 1991; Wakamatsu et al., 1992), and a simulated annealing protocol, YASAP, as a restrained MD algorithm. Simulated an-

*To whom correspondence should be addressed.

Abbreviations: MD, molecular dynamics; MDS, multidimensional scaling; rmsd, root-mean-square deviation; armsd, angular rmsd; R, multiple correlation coefficient; YASAP, yet another simulated annealing protocol; PCA, principal component analysis.

nealing is a kind of restrained MD, where the temperature is raised and then lowered. YASAP stands for 'yet another simulated annealing protocol' and is described in the XPLOR v. 2.1 manual (Brünger, 1990). Since only a part of the NMR structure calculation (say, 5 to 50% of the total calculation) produces final structures with acceptably low values of the target function, the initial set of structures that produce converged structures is a subset of the generated initial structures. The conformational randomness of the initial structures that will converge must be checked to demonstrate unbiased sampling of conformational space. Usually, this randomness is proven by comparison between the average of the rmsd values for all pairs of the initial structures that *will converge*, and that of the rmsd values for all pairs of random-coil polypeptides consisting of the same number of residues (see for example: Braun and Gō, 1985; Montelione et al., 1987; Wagner et al., 1987). Rmsd values for the intermediate structures that will converge are also used to show random sampling of conformational space (Nilges et al., 1988). The Trajectory Analysis described in this paper provides a more systematic view for a demonstration of random sampling of conformational space than the simple rmsd comparisons of initial or intermediate structures. We also show that Trajectory Analysis is useful to optimize parameters in protocols used in NMR structure calculations.

Methods

Distance and dihedral angle constraints for wild-type μ-Conotoxin GIIIA were obtained as described previously (Wakamatsu et al., 1992). μ-Conotoxin GIIIA (also named Geographutoxin I) is a 22-residue peptide with three disulfide bridges, having a selective inhibitory effect on muscle-type sodium channels (Ohizumi et al., 1986). The coordinates of the mean structure and of 10 individual structures have been deposited in the Protein Data Bank. The identification codes are 1TCG and 1TCJ. The NMR structure calculation was done with YASAP (Brünger, 1990) in the program XPLOR (v. 2.1, Molecular Simulations, Waltham, MA) on a Silicon Graphics Personal Iris 35 workstation. The target function used in XPLOR can be expressed as (Driscoll et al., 1989):

$$F_{total} = F_{bond} + F_{angle} + F_{impr} + F_{repel} + F_{NOE} + F_{tor}$$

The last two terms represent geometric constraints from NMR data. We use

$$F_3 = F_{repel} + F_{NOE} + F_{tor}$$

for the evaluation of convergence of the calculation. The YASAP protocol comprises three restrained MD steps, flanked by initial and final minimization steps (see Fig. 3). The first restrained MD calculation consists of high-temperature dynamics at 1000 K, with a very small weight for the repel term in order to create rough folding of a target molecule. The second restrained MD calculation also consists of high-temperature dynamics, with a gradually increasing weight of the repel term and a tilting asymptote of the NOE potential function for refinement of the structure. The system is cooled to 300 K during the third restrained MD calculation.

The metric MDS calculation was performed with *S-plus* (Statistical Science, Inc., Seattle, WA) on an Apollo DN3500 workstation. Briefly, a transformation known as double centering converts an rmsd matrix, $[d_{ij}]$ (where $i,j = 1 \sim N$), to an inner product matrix. The coordinates of individual structures are obtained by spectral decomposition of the inner product matrix, in which eigenvalues ($\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq \ldots \geq \lambda_N$) and their associated normalized eigenvectors ($x_{p1}, x_{p2}, \ldots, x_{pN}$) ($p = 1 \sim N$) are found. The pth coordinate of the qth structure is obtained as $\lambda_p^{1/2} \times x_{pq}$.

In an NMR distance geometry calculation, the dimensionality of the coordinates of atoms is three. By contrast, we do not know the dimensionality of the conformational space in advance. Stress is a measure of how well the MDS solution fits the input data and this can be used to determine the appropriate dimensionality of the conformational space. Stress is defined as:

$$\text{Stress}(k) = 1 - \sum_{i=1}^{k} \lambda^2 i \Big/ \sum_{i=1}^{N} \lambda^2 i$$

where λ_i is the ith largest eigenvalue, N is the total number of points (structures), and k is the number of dimensions of the MDS solution.

Results

Construction and dimensionality of the trajectory map

Figure 1 shows an example of Trajectory Analysis. We performed 200 simulated annealing calculations for the structure determination of μ-Conotoxin GIIIA, using YASAP. For each of the 189 calculations that did not abort, the coordinates of 10 structures (one initial, eight intermediate and one final) were saved. The sampling schedule is shown in Fig. 3a. Rmsd values for all pairs ($_{1890}C_2 = 1\,785\,105$) of the 1890 structures were computed using a program written in C according to the algorithm of Kabsch (1978). The resulting rmsd matrix was subjected to metric MDS analysis. About 70% (1300/1890) of the eigenvalues were positive; coordinates up to the 1300th dimension of the conformational space can be defined, showing that the contradiction among rmsd values in the rmsd matrix was almost negligible. The points corresponding to structures on the same trajectory were connected by lines. To determine the minimum dimension of the trajectory space, stress values for solutions with one to five dimensions were plotted (Fig. 1b). Because of the 'elbow' (i.e. a point of inflection) in the graph at Dimension 2, two dimensions were sufficient to

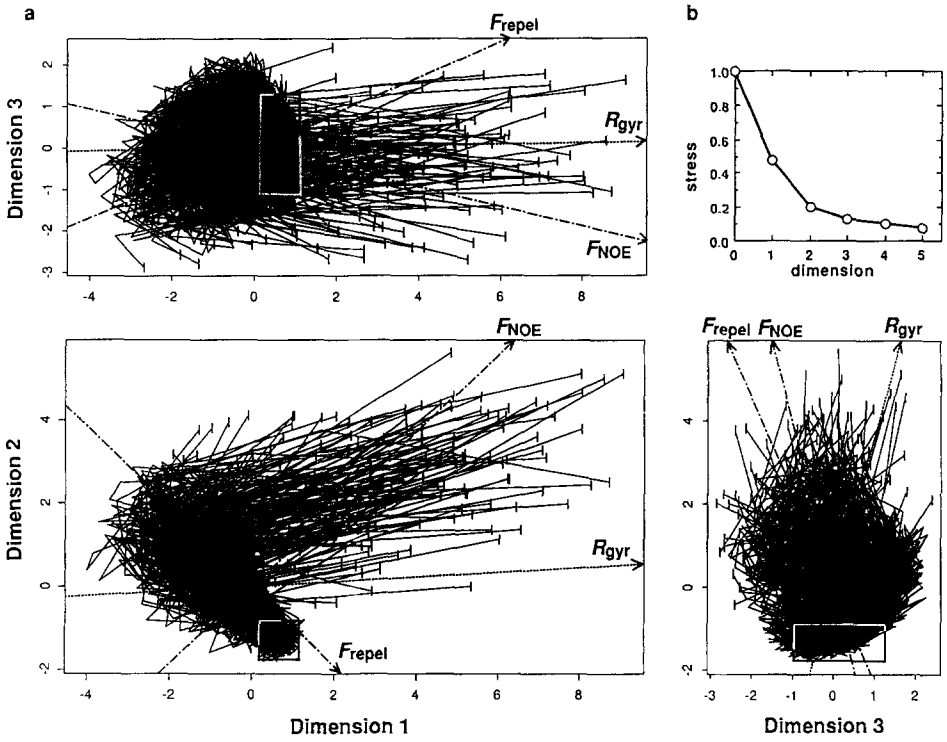

Fig. 1. (a) Three-dimensional trajectory map containing 189 simulated annealing calculations of μ-Conotoxin GIIIA. All dimensions are expressed in Å. Short vertical bars indicate the positions of the initial structures. The box encloses a region containing all the final structures except three, which are located near the point (0,1.5,1.5). F_{NOE} = energy for distance constraints; F_{repel} = repulsion energy for van der Waals contact; R_{gyr} = radius of gyration of a molecule. (b) Stress for 1–5 dimensions of the MDS solution.

describe the overall configuration of the trajectories, but we added the third dimension for further analysis. The stress value at Dimension 3 indicates that the three dimensions accounted for about 85% of the total variance, i.e. useful information, of the original rmsd matrix. It is surprising that only two or three dimensions are sufficient to describe the relationships among 1890 structures. This low dimensionality of the trajectory space can be understood better by plotting a scatter diagram (Fig. 2). In this diagram, the horizontal axis displays the rmsd values and the vertical axis displays distances in three-dimensional conformational space. The majority of the points are located nearly along the diagonal line, indicating that the three-dimensional solution is sufficient to describe the trajectories. Deviations from the diagonal are mainly seen when the rmsd is relatively small, and the direction is downward, i.e., the distance in three-dimensional conformational space is always smaller than the corresponding rmsd value. These observations enable an interpretation where low dimensionality of the trajectory space is achieved by overlooking local short-range relationships, while preserving long-range relationships.

Dimensional interpretation of the trajectory map

The ensemble of the trajectories in three-dimensional conformational space displays an overall 'L-shape' (Fig. 1a). Note that the map is independent of translation and rotation. The L-shape of the trajectories is a general feature, because similar trajectory maps were obtained from different numbers of simulated annealing calculations, different proteins, different measurements of distance in conformational space and a different statistical method. The following trajectory analyses gave a similar trajectory map (data not shown): subsets (10 and 50) of the 189

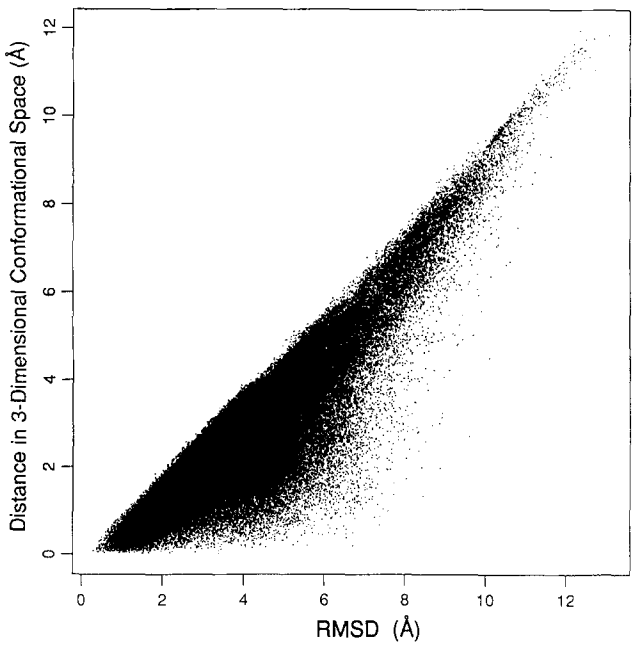

Fig. 2. Scatter diagram for the three-dimensional MDS solution shown in Fig. 1.

calculations of μ-Conotoxin GIIIA; a 62-residue snake neurotoxin, erabutoxin b; an analysis using armsd (angular rmsd) instead of rmsd. Armsd is defined as the root-mean-square difference between the ϕ and ψ torsion angles in two structures; principal component analysis (PCA) is another method to generate a trajectory map from the coordinates of the structures. A structure containing N atoms can be regarded as a point in 3N-dimensional space. PCA finds the first principal axis for the projection of the points to have the maximum possible variation, and then it finds the second principal axis in a similar manner, with the restriction that the second axis must be orthogonal to the first axis, and so on. PCA yielded a similar configuration of trajectories. However, PCA consists of simple rotation and projection of points in a multidimensional space, but MDS reconstructs the coordinates of points from the distance matrix. Thus, MDS gives a more appropriate solution of low dimensionality than does PCA.

In fact, an L-shaped solution is closely related to the algorithm used in YASAP. We interpreted the configuration of the trajectories using multiple regression analysis. The total energy and each energy term of the target function, the radius of gyration of the structures, and other geometric values were computed for every structure, and regressed to the three-dimensional solution of the MDS analysis (Fig. 1a). The multiple correlation coefficients (R) of F_{total}, F_{bond} and F_{NOE} were greater than 0.90 (significant at $P<0.001$) and that of F_{tor} was 0.79 ($P<0.001$). The directions of these four axes in conformational space were almost identical. As an example, the F_{NOE} axis is drawn in Fig. 1a. In contrast, the F_{repel} axis is perpendicular (89.8°) to the F_{NOE} axis with $R=0.87$ ($P<0.001$). The other values had a less significant correlation, except for the radius of gyration ($R=0.95$, $P<0.001$). Along the trajectory from an initial to a final structure, the value of F_{NOE} (F_{total}, F_{bond}, F_{tor}) decreased rapidly, while the F_{repel}

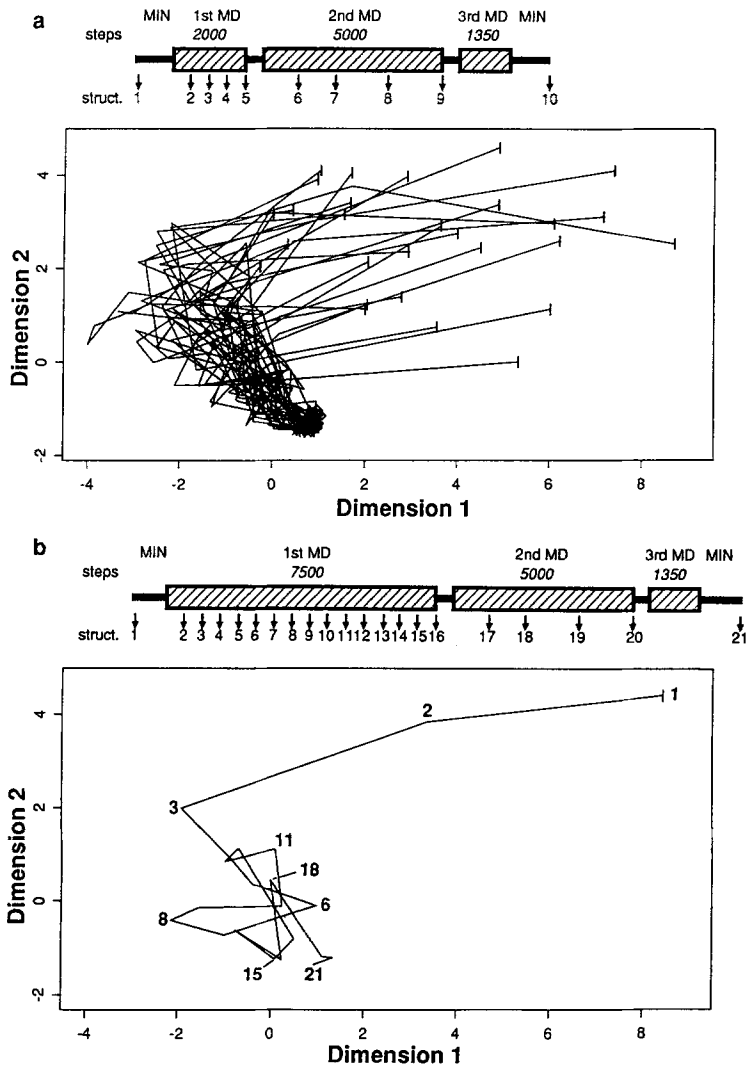

Fig. 3. (a) Dimensions 1 and 2 for 30 converged trajectories selected out of the 189 trajectories of μ-Conotoxin GIIIA. (b) Typical trajectory of an XPLOR calculation performed with the original version of YASAP. The first dynamics calculation of the original protocol comprised 7500 steps. The insets show time schedules of saving structures. MIN = restrained minimization; MD = restrained molecular dynamics. One step of MD corresponds to 2 fs. All dimensions are expressed in Å.

value first decreased and then increased. The change in the radius of gyration indicated that the volume of the protein molecule decreased temporarily during the calculation. This is a reflection of a small force constant of F_{repel} during the first restrained MD stage in the YASAP protocol (Brünger, 1990); $k_{repel} = 0.002$ kcal mol^{-1} Å^{-4} is used, which corresponds to 0.05% of the value for the final structures. Thus, the linear combination of Dimensions 1 and 2 represents the energies and size of the molecule. The interpretation of Dimension 3 is not straightforward, but seems to correlate to asymmetry of the molecule.

Random sampling of conformational space

A total of 30 trajectories, whose final structures had the smallest F_3 values, were selected as computationally converged trajectories, and their Dimensions 1 and 2 are shown in Fig. 3a. Visual comparison between Figs. 1a and 3a reveals that random sampling of the conformational space is readily discernible over the entire trajectories. Clearly, this visual comparison is not a substitute for more rigorous tests of random sampling using a simulated data set of constraints (Metzler et al., 1978; Nakai et al., 1993). However, it provides an intuitive criterion for the random sampling problem. For example, one can readily recognize that the converged trajectories do not pass through any particular common intermediate states. It is difficult to obtain such information from a simple rmsd comparison of initial and intermediate structures.

Optimization of protocols for NMR structure calculation

One interesting application of trajectory analysis is optimization of protocols used in NMR structure calculations. At first, we had calculated the structure of μ-Conotoxin GIIIA using the original YASAP protocol in which the first MD calculation comprised 7500 steps. Figure 3b shows one typical trajectory extracted from the trajectory map containing the 30 preliminary calculations. The trajectories frequently tangled in the conformational space during the last two-thirds of the first 7500-step dynamics, indicating that this first MD calculation is sufficiently long for peptides of about 20 residues to reach an equilibrium state, i.e., it appears to be redundant. We expected that a reduction of the first dynamics step would greatly shorten the computational time, without affecting the convergence of the calculation. Hence, the first calculation was reduced from 7500 to 2000 steps. The CPU time was reduced by half, but the yield of final good structures was the same. In fact, the trajectory maps in Figs. 1a and 3a were made using the tuned YASAP.

The length of the first dynamics calculation is dependent primarily on the size of the peptide/protein of interest. The optimization of YASAP should thus be conducted considering the molecular size. We found that the optimized number of steps for the first dynamics calculation is 2000 for about 20-residue peptides, 5000 for about 50-residue peptides and 10 000 or more for about 100-residue proteins.

Conclusions

Trajectory Analysis as proposed in this paper enables visualization of how structures converge to final structures during NMR structure calculations. This method is based on the assumption that the distance between two structures in conformational space is the rmsd between the two structures, and thus independent of the algorithms used in various NMR structure calculation programs. The trajectory map is useful to prove random sampling of conformational space intuitively, and to improve protocols of various algorithms used in NMR structure calculations.

Acknowledgements

We thank the Ministry of Science, Education and Culture of Japan and the Human Frontier Science Program Organization for supporting this work. We thank our colleague Dr. Hideki Hatanaka for invaluable discussions.

References

Bruccoleri, R.E. and Karplus, M. (1990) *Biopolymers*, **29**, 1847–1862.
Brünger, A.T. (1990) XPLOR software manual, version 2.1, Yale University, New Haven, CT.
Braun, W. and Gō, N. (1985) *J. Mol. Biol.*, **186**, 611–626.
Crippen, G.M. and Havel, T.F. (1978) *Acta Crystallogr.*, **A34**, 282–284.
Driscoll, P.C., Gronenborn, A.M., Beress, L. and Clore, G.M. (1989) *Biochemistry*, **28**, 2188–2198.
Kabsch, W. (1978) *Acta Crystallogr.*, **A34**, 827–828.
Kruskal, J.B. and Wish, M. (1978) *Multidimensional Scaling*, Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07–011, Sage Publications, Beverly Hills, CA.
Lancelin, J.-M., Kohda, D., Tate, S., Yanagawa, Y., Abe, T., Satake, M. and Inagaki, F. (1991) *Biochemistry*, **30**, 6908–6916.
Levitt, M. (1983a) *J. Mol. Biol.*, **168**, 621–657.
Levitt, M. (1983b) *J. Mol. Biol.*, **170**, 723–764.
Metzler, W.J., Hare, D.R. and Pardi, A. (1989) *Biochemistry*, **28**, 7045–7052.
Montelione, G.T., Wüthrich, K., Nice, E.C., Burgess, A.W. and Scheraga, H.A. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 5226–5330.
Nakai, T., Kidera, A. and Nakamura, H. (1993) *J. Biomol. NMR*, **3**, 19–40.
Nilges, M., Gronenborn, A.M., Brünger, A.T. and Clore, G.M. (1988) *Protein Eng.*, **2**, 27–38.
Ohizumi, Y., Nakamura, H., Kobayashi, J. and Catterall, W.A. (1986) *J. Biol. Chem.*, **261**, 6149–6152.
Torgerson, W.S. (1958) *Theory and Methods of Scaling*, Wiley, New York, NY.
Wagner, G., Braun, W., Havel, T.F., Schaumann, T., Gō, N. and Wüthrich, K. (1987) *J. Mol. Biol.*, **196**, 611–639.
Wakamatsu, K., Kohda, D., Hatanaka, H., Lancelin, J.-M., Ishida, Y., Oya, M., Nakamura, H., Inagaki, F. and Sato, K. (1992) *Biochemistry*, **31**, 12577–12584.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY, pp. 186–199.
Young, G. and Householder, A.S. (1938) *Psychometrika*, **3**, 19–22.