Estimating Protein Fold from Incomplete and Approximate NMR Data

Peter J. Connolly, Alan S. Stern, and Jeffrey C. Hoch*

Rowland Institute for Science
100 Edwin H. Land Boulevard
Cambridge, Massachusetts 02142

Received September 14, 1993

Advances in multidimensional NMR methods for the determination of protein structure in solution have led to continual improvements in the size of the molecules that can be studied and the precision of the structures that can be obtained. These improvements have been a result of the increasing number of distance restraints that can be obtained from NOE data and from the use of dihedral angle restraints derived from vicinal coupling constants. Since a structure determination of even moderate precision requires approximately 10 distance restraints per residue, the assignment and quantitation of resonances in the NOE spectrum remains a time-consuming and tedious task. An approximate description of the structure would be useful in resolving ambiguous assignments and to focus the search for assignments that will be particularly useful for defining the structure of specific parts of the protein. We demonstrate that a method we recently developed for determining protein fold from NMR distance restraints is capable of discerning the overall fold using highly incomplete and approximate distance restraints, as low as 0.5 long-range restraint per residue. In contrast to previously published methods for determining overall protein fold, our method does not require elements of secondary structure to be predefined.[1,2] In addition to aiding the assignment process, knowledge of the overall fold can provide insights into biological activity, establish relationships to other protein folding motifs, and serve as a starting point for subsequent structure refinement.

Our method is based on a two-particle-per-residue representation of protein structure: one particle to represent the trace of the backbone and the other to represent the direction of the side chain relative to the backbone. We define a simple force field that includes pseudobond, pseudoangle, and distance restraint terms and an electrostatic term that serves as a generalized repulsion to prevent structures from collapsing. NMR distance restraints are converted to the two-particle representation by addition of appropriate correction factors. The overall fold of the protein can then be determined using simulated annealing or distance geometry followed by minimization of the two-particle energy. The method is readily implemented using commercial molecular modeling software; details are given elsewhere.[3]

We illustrate the method using preliminary data for LSIII, a long neurotoxin from the venom of *Laticauda semifasciata*. Sequential assignments for this 66-residue protein have been determined, and 54 long-range interresidue nuclear Overhauser effects have been assigned, though not quantified, giving 39 qualitative two-particle distance restraints (as some of the restraints become redundant when reduced to the two-particle representation). Using upper distance bounds of 5.0 Å for these qualitative NOE restraints and adding restraints for the five known disulfide bonds, the method yielded the family of folds depicted in Figure 1. The root mean square difference from the mean for this family of structures is 4.98 Å for all particles and 4.68 Å for the backbone trace. This family of structures contains all the significant features of the fold, namely the three strands of antiparallel β-sheet from residues 19–25, 36–42, and 52–58. In addition, the three major loops and globular head (the structural motif common to long neurotoxins) are evident and consistent with the X-ray and NMR structures for the homologous protein α-cobratoxin.[4,5]

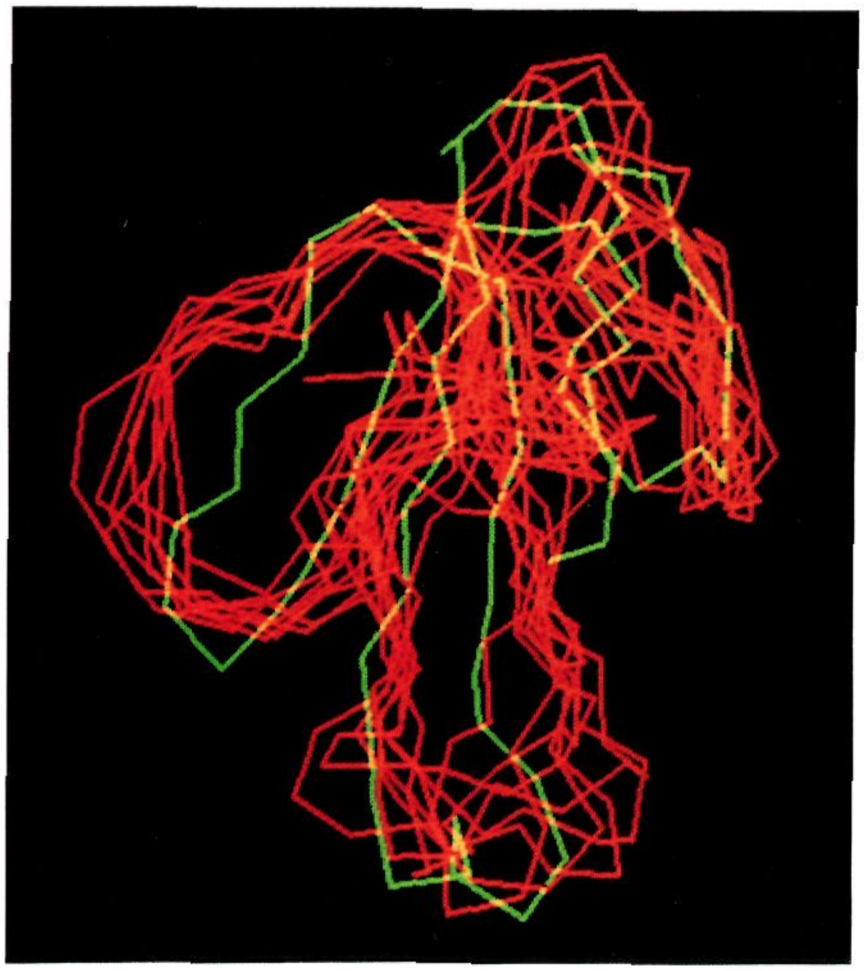

Figure 1. Ten two-particle simulated annealing structures of LSIII are displayed in red. Each structure required 3.8 min to calculate on an IBM RS/6000 Model 320H. The two-particle structure of α-cobratoxin derived from the X-ray crystal coordinates (Brookhaven Data Bank entry 1CTX)[4] is shown in green for comparison.

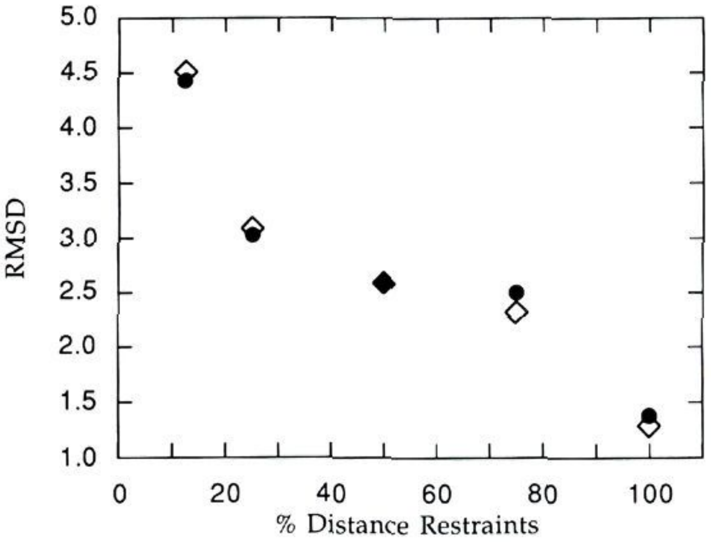

Figure 2. Results of two-particle simulated annealing calculations for BDS-I using randomly chosen subsets of NOE restraints. The results obtained using quantitative restraints are shown by ◇ and those using qualitative restraints by ●.

To assess the accuracy and precision of the two-particle method, control calculations were performed on the 43-residue protein BDS-I, whose solution structure has been determined by Driscoll et al.[6] (Brookhaven Data Bank entry 1BDS). Randomly chosen subsets of the complete set of NOEs were used. Quantitative distance restraints were taken from Driscoll et al.,[6] and qualitative restraints were derived by setting the upper and lower bounds to 5.0 and 0.0 Å, respectively. Figure 2 shows the results of our

[1] Altman, R. B.; Jardetzkey, O. *Methods Enzymol.* **1989**, *177*, 218–245.

[2] Smith-Brown, M. J.; Kominos, D.; Levy, R. M. *Protein Eng.* **1993**, *6*, 605–614.

[3] Hoch, J. C.; Stern, A. S. *J. Biomol. NMR* **1992**, *2*, 535–543.

[4] Walkinshaw, M. D.; Saenger, W.; Maelicke, A. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 2400–2404.

[5] Le Goas, R.; Laplante, S. R.; Mikou, A.; Delsuc, M.-A.; Guittet, E.; Robin, M.; Charpentier, I.; Lallemand, J.-Y. *Biochemistry* **1992**, *31*, 4867–4875.

[6] Driscoll, P. C.; Gronenborn, A. M.; Beress, L.; Clore, J. M. *Biochemistry* **1989**, *28*, 2188–2198.

0002-7863/94/1516-2675$04.50/0

calculations. The structures obtained using only 0.5 approximate restraint per residue faithfully represent the overall fold despite the relatively high root mean square deviation. We note that the precision of the structures is independent of the quantitation of the restraints, a result consistent with the observations of Havel[7] and Clore.[8] Calculations on other proteins (not shown) show that regions of helical secondary structure can also be identified.

In our control calculations, the randomly chosen subsets of NOEs were evenly distributed over the molecule. However, in practice, it is frequently the case that during the early stages of sequential assignment, NOEs are not so evenly distributed. Under such circumstances, different portions of the molecule may be well defined but their relative orientation may not be, and the folding topology for poorly defined regions will necessarily be indeterminate.

This leads to a more general question: when are experimental restraints sufficient to define a unique fold? We know of no way to answer this question a priori. As with all methods, short of exhaustive search, it is impossible to guarantee that any fold is unique or optimal. In practice, we observe that the two-particle force field tends to eliminate nonphysical folds. In cases when there are many feasible folds consistent with the experimental restraints, we observe wide variation in the ensembles of structures. We have also seen cases where the structures cluster around a single fold but include some outliers, which generally have a higher two-particle energy.[3]

In summary, the simplicity of the two-particle representation, its low computational cost, and its robustness provide a powerful method for determining overall protein fold, capable of being used with incomplete and approximate distance restraints. The approach serves as a useful adjunct to conventional methods that require larger amounts of more accurate data.

Acknowledgment. We are grateful to Molecular Simulations Inc. and to Tripos Associates Inc. for use of their software. This work was supported by the Rowland Institute for Science.

[7] Havel, T. F.; Wütrich, K. *J. Mol. Biol.* **1985**, *182*, 281–294.

[8] Clore, G. M.; Robien, M. A.; Gronenborn, A. M. *J. Mol. Biol.* **1993**, *231*, 82–102.