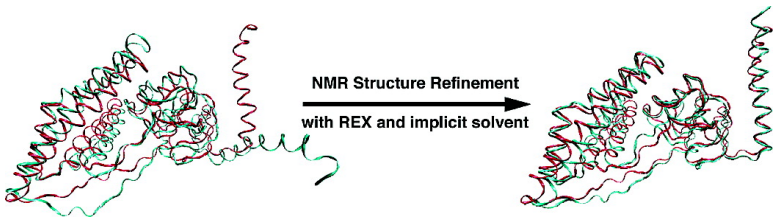
Article

Refinement of NMR Structures Using Implicit Solvent and Advanced Sampling Techniques

Jianhan Chen, Wonpil Im, and Charles L. Brooks

J. Am. Chem. Soc., **2004**, 126 (49), 16038-16047• DOI: 10.1021/ja047624f • Publication Date (Web): 18 November 2004



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 1 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

J|A|C|S
ARTICLES
Published on Web 11/18/2004

Refinement of NMR Structures Using Implicit Solvent and Advanced Sampling Techniques

Jianhan Chen, Wonpil Im, and Charles L. Brooks, III*

Contribution from the Department of Molecular Biology, Center for Theoretical Biological Physics, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

Received April 24, 2004; E-mail: brooks@scripps.edu

Abstract: NMR biomolecular structure calculations exploit simulated annealing methods for conformational sampling and require a relatively high level of redundancy in the experimental restraints to determine quality three-dimensional structures. Recent advances in generalized Born (GB) implicit solvent models should make it possible to combine information from both experimental measurements and accurate empirical force fields to improve the quality of NMR-derived structures. In this paper, we study the influence of implicit solvent on the refinement of protein NMR structures and identify an optimal protocol of utilizing these improved force fields. To do so, we carry out structure refinement experiments for model proteins with published NMR structures using full NMR restraints and subsets of them. We also investigate the application of advanced sampling techniques to NMR structure refinement. Similar to the observations of Xia et al. (*J. Biomol. NMR* **2002**, *22*, 317−331), we find that the impact of implicit solvent is rather small when there is a sufficient number of experimental restraints (such as in the final stage of NMR structure determination), whether implicit solvent is used throughout the calculation or only in the final refinement step. The application of advanced sampling techniques also seems to have minimal impact in this case. However, when the experimental data are limited, we demonstrate that refinement with implicit solvent can substantially improve the quality of the structures. In particular, when combined with an advanced sampling technique, the replica exchange (REX) method, near-native structures can be rapidly moved toward the native basin. The REX method provides both enhanced sampling and automatic selection of the most native-like (lowest energy) structures. An optimal protocol based on our studies first generates an ensemble of initial structures that maximally satisfy the available experimental data with conventional NMR software using a simplified force field and then refines these structures with implicit solvent using the REX method. We systematically examine the reliability and efficacy of this protocol using four proteins of various sizes ranging from the 56-residue B1 domain of Streptococcal protein G to the 370-residue Maltose-binding protein. Significant improvement in the structures was observed in all cases when refinement was based on low-redundancy restraint data. The proposed protocol is anticipated to be particularly useful in early stages of NMR structure determination where a reliable estimate of the native fold from limited data can significantly expedite the overall process. This refinement procedure is also expected to be useful when redundant experimental data are not readily available, such as for large multidomain biomolecules and in solid-state NMR structure determination.

Introduction

Even though solvation plays an essential role in defining the native conformation of proteins,[1] it has generally been ignored in traditional NMR structure calculations.[2−4] Additional force field simplifications are also exploited in NMR structure refinement. While the covalent nature of the bonding geometry is explicitly used, the nonbonded interactions are typically simplified, for example, by using a repulsive soft-sphere interaction and ignoring the electrostatic interactions.[5] The main purpose for these expedients is to reduce the roughness of the energy landscape and thereby improve the computational efficiency of conformational sampling. As such, conventional NMR structure calculations rely almost completely on the experimental restraints to determine three-dimensional (3D) structures. Aside from the desire to determine structure based solely on experimental data, this approach may, in principle, be justified by the fact that molecular mechanics force fields without explicit consideration of solvent effects did not provide an accurate description of the nonbonded interactions. Recent developments of efficient generalized Born (GB) based implicit solvent models[6−10] now make it feasible to incorporate realistic

(1) Feig, M.; Brooks, C. L., III. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217−224.
(2) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R.; Jiang, J.-S.; Kuszewski, J.; Nilges, N.; Pannu, N.; Read, R.; Rice, L.; Simonson, T.; Warren, G. *Acta Crystallogr.* **1998**, *D54*, 905−921.
(3) Güntert, P.; Mumenthaler, C.; Wüthrich, K. *J. Mol. Biol.* **1997**, *273*, 283−298.
(4) Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, G. M. *J. Magn. Reson.* **2003**, *160*, 66−74.
(5) Nilges, M.; Gronenborn, A. M.; Brunger, A. T.; Clore, G. M. *Protein Eng.* **1988**, *2*, 27−38.

10.1021/ja047624f CCC: $27.50

solvent effects into NMR structure determination with minimal additional computational cost. With these implicit solvent models important structural and equilibrium dynamical properties of proteins can be reproduced reasonably well compared to expensive explicit solvent simulations.[1] Incorporation of these improved force fields into NMR structure calculations is expected to improve the quality of the final structures and/or decrease the amount of experimental data required for convergence with an affordable increase in the computational cost. Furthermore, as NMR experiments are typically carried out on solution samples, including solvent effect in the structure calculation can potentially lead to structures with improved agreement with experimental measurements.

Recently, Xia et al. showed that simulated annealing refinement in a GB implicit solvent could lead to noticeable improvement in the final protein NMR structures in terms of the backbone dihedral angle distribution and hydrogen bond pattern, which was comparable to what could be obtained by refinement with explicit water.[11] However, the structures obtained with and without implicit or explicit solvent were very similar, and the overall impact of the solvent was rather small. This may arise from several origins. First, the amount of experimental data was sufficient and thus the structures were already extremely well-defined. Second, when all of the experimental restraints were applied at full strength during refinement, the structures were deeply trapped in local energy minima. Finally, restrained simulated annealing of limited time scales did not sufficiently explore the conformation space. Therefore, it is not clear that the improvement observed in the study of Xia et al. represents the limit of implicit solvent on improving protein NMR structures. Neither is it clear that simulated annealing refinement with implicit solvent during the final stage of NMR structure determination is the best way of utilizing the improved force fields.

To address these questions, we have furthered the study of the influence of an accurate force field on protein NMR structures by using a GB implicit solvent throughout the structure calculations. To identify the best way of utilizing these force fields, we also carried out numerical experiments to mimic the situation of insufficient experimental data such as in the early stages of NMR structure determination. Three model proteins with various sizes and topologies have been used. Subsets of the full NMR restraints are chosen such that the structures cannot be sufficiently well-determined solely by the experimental data, leaving room for possible improvement by refinement with implicit solvent. Carrying out such numerical experiments also makes it possible to identify an optimal sampling technique for refining NMR structures with implicit solvent. The prevailing sampling technique used in NMR structure calculation and refinement is based on simulated annealing molecular dynamics (MD) in both Cartesian and torsion space.[2,3,12] However, advanced sampling techniques such as the replica exchange method (REX)[13] exist and have been shown to offer generally better conformational sampling in applications such as protein folding and unfolding studies.[14,15] In this study, we investigate the application of the replica exchange method to NMR structure refinement, particularly, in the context of exploiting an accurate empirical force field with implicit solvent. Finally, we test the proposed refinement procedure using a 370-residue maltose-binding protein with full published NMR restraints, where the initial structures are poorly converged due to insufficient experimental data.[16]

Table 1. Model Protein Systems

protein	PDB ID[a]	topology	residues	NOE (long-range)	DIHE
GB1	3gb1	α/β	56	735 (302)	0
GAIP	1cmz	all α	128	1427[b] (311)	146
EIN	2eza	α/β	259	2820 (608)	546
MBP	1ezo	α/β	370	1991[c] (826)	555

[a] 3gb1;[18] 1cmz;[19] 2eza;[20] 1ezo.[16] [b] Including 70 hydrogen bonding restraints. [c] Including 48 hydrogen bonding restraints.

Methods

Model Systems and NMR Restraints. Four proteins of various sizes and topologies were used in this study: the B1 domain of Streptococcal protein G (GB1); the human Gα interacting protein (GAIP); the N-terminal domain of enzyme I (EIN) of *Escherichia coli* (*E. coli*); and the 370-residue Maltose-binding protein (MBP). All NMR experimental restraints were obtained from the Protein Data Bank.[17] Table 1 summarizes the model systems and experimental NMR restraints used in this study. Note that the NMR structures of MBP are not converged even with additional residual dipolar coupling restraints.[16] Therefore the refinement of MBP offers a realistic, challenging test for the protocol proposed.

MD Simulations. All MD simulations were carried out using the CHARMM program[21] with the PARAM22 all-hydrogen parameter set.[22] The average solvent effect was characterized by a generalized Born implicit solvent model with a simple switching function (GBSW).[10] GBSW closely reproduces the electrostatic solvation energy given by the Poisson−Boltzmann (PB) equation but is computationally much more efficient. The nonpolar solvation energy is estimated from the solvent-exposed surface area (SA) using a phenomenological surface tension coefficient. For the simulations used in this study, an approximation to the molecular surface was used;[23] otherwise, default GBSW parameters were used. The surface tension coefficient was set to be 40 cal/(mol·Å^2) for GB1 and GAIP and 10 cal/(mol·Å^2) for EIN and MBP. A harmonic flat-bottom NOE restraint potential with a soft asymptote[5,24] was used. Distance summation ($r = (\sum_j r_j^{-6})^{-1/6}$) was used for NOE restraints that involve more than two protons. Dihedral restraints were applied using a simple harmonic potential function with a flat bottom.

(6) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(7) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1−20.
(8) Bashfold, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129−152.
(9) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Chem. Phys.* **2002**, *116*, 10606−10614.
(10) Im, W.; Lee, M. S.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1691−1702.
(11) Xia, B.; Tsui, V.; Case, D. A.; Dyson, H. J.; Wright, P. E. *J. Biomol. NMR* **2002**, *22*, 317−331.
(12) Clore, G. M.; Gronenborn, A. M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5891−5898.
(13) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(14) Hansmann, U.; Okamoto, Y. *Curr. Opin. Struct. Biol.* **1999**, *9*, 177−183.
(15) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96−123.
(16) Mueller, G. A.; Choy, W. Y.; Yang, D.; Forman-Kay, J. D.; Venters, R. A.; Kay, L. E. *J. Mol. Biol.* **2000**, *300*, 197−212.
(17) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. *Nucl. Acids Res.* **2000**, *28*, 235−242.
(18) Kuszewski, K.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 2337−2338.
(19) De Alba, E.; De Vries, L.; Farquhar, M. G.; Tjandra, N. *J. Mol. Biol.* **1999**, *291*, 927−939.
(20) Tjandra, N.; Garrett, D. S.; Gronenborn, A. M.; Bax, A.; Clore, G. M. *Nature Struct. Biol.* **1997**, *4*, 443−449.
(21) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.
(22) MacKerell, A. D., Jr.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(23) Feig, M.; Onufriev, A.; Lee, M.; Im, W.; Case, D.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 265−284.
(24) Stein, E. G.; Rice, L. M.; Brünger, A. T. *J. Magn. Reson.* **1997**, *124*, 154−164.

To examine the influence of implicit solvent when the experimental data are limited, poorly converged structures were generated using subsets of the full NMR restraints for GB1, GAIP, and EIN by the CNS program (version 1.1[2]). These initial structures were then refined by the replica exchange method with the GBSW implicit solvent. How the subsets were chosen will be discussed in detail in Results and Discussion. The simulated annealing protocol implemented in the CNS script anneal.inp[5,24] was used in all CNS calculations. This protocol includes high-temperature dynamics in torsion space, followed by two slow-cooling stages implemented in torsion and Cartesian space, respectively. Longer dynamics was used for larger proteins to ensure reasonable acceptance rates. Note that while more sophisticated simulated annealing protocols exist and might be more commonly used in current practice, the CNS:anneal.inp protocol has been able to successfully produce initial structures that satisfy all experimental restraints sufficiently well for all the systems used in this study. In the following numerical experiments, the initial CNS structures are not well-converged primarily due to a lack of experimental data rather than the simulated annealing protocol. More sophisticated protocols have been shown to make little difference here and therefore are not used. The adequacy of the CNS:anneal.inp protocol will be further addressed in Results and Discussion.

Replica Exchange Refinement Protocol. The replica exchange method, also known as *parallel tempering*, is a generalized-ensemble method. Multiple copies (replicas) of the system are simulated at different temperatures independently and simultaneously by conventional Monte Carlo (MC) or MD methods.[13] The temperatures are usually distributed exponentially within a specified range, and there is always one single replica simulated at each temperature. Pairs of replicas at neighboring temperatures attempt to exchange simulation temperatures according to a Metropolis type algorithm after a number of steps of MC or MD simulation. Replicas with lower potential energy tend to occupy the lower temperature conditions, while exchanging to higher temperature is highly probable even for replicas with lower energies compared with their higher temperature neighbors. In the course of an REX simulation, replicas can travel up and down the temperature space automatically in a self-regulated fashion, which, in turn, induces a nontrivial walk in temperature space. REX can greatly reduce the probability of being trapped in states of local energy minima and sample a larger conformation space.[14,15] In addition, the occupancy of the structures at a given temperature is determined by the relative average potential energy, providing a simple way of ranking structures.

The replica exchange simulations were carried out using the Multiscale Modeling Tools in Structural Biology (MMTSB) tool set (available from http://mmtsb.scripps.edu)[25,26] and CHARMM. Each replica started from a different conformation of the same initial ensemble generated by CNS. A replica exchange was attempted every 0.5 ps of restrained molecular dynamics, where the force constants were set to be 10 kcal/(mol·Å^2) for NOE restraints and 50 kcal/(mol·radian^2) for dihedral angle restraints. These constants were chosen empirically to allow some balance between flexibility and stability of the structures. To maintain a reasonable exchange acceptance ratio (i.e., the ratio between the number of actual exchanges and exchange attempts), the number of replicas required for a given temperature range increases as the system size N increases (according to $\sqrt{N}$).[27] As such, we used 16 replicas between 300 and 550 K for GB1 and GAIP, 32 replicas between 300 and 700 K for EIN, and 48 replicas between 300 K to 800 K for MBP. 400 REX steps for GB1 and GAIP, 200 REX steps for EIN and 1000 REX steps for MBP were carried out. The overall exchange ratio ranged between 0.2 and 0.3. The structures of all replicas were saved at the end of each REX step for analysis. Finally, the last 20−25% of the structures from the lowest temperature ensemble were energetically minimized to provide an ensemble of refined structures. Force constants of 75 kcal/(mol·Å^2) for NOE restraints and 200 kcal/(mol·radian^2) for dihedral angle restraints were used during minimization. Further clustering can also be applied to extract a few representative structures. All these calculations were enabled by the MMTSB tool set together with CHARMM.

Results and Discussion

GB1. GB1 is a small 56-residue protein with a stable but nontrivial α/β fold topology. Due to the small size and availability of both NMR and X-ray structures,[18,28] we have primarily used this model protein to study the optimal conditions of NMR structure calculation/refinement with implicit solvent models.

Impact of Molecular Force Field with Full NMR Restraints. Both simulated annealing and replica exchange simulations in Cartesian space have been performed with CHARMM to study the influence of using implicit solvent throughout the structure calculation. As shown in Table 1, there is a sufficiently large number of NMR restraints to determine the structure of GB1 to within 0.6 Å backbone root mean square deviation (RMSD) from the X-ray structure. It was found that the force field had little influence on the calculated final structures in terms of backbone RMSD from the X-ray structure, NOE violation statistics, RMS fluctuation of the structures around the mean, and distribution of backbone torsion angles on the Ramanchandran plots (data not shown). Instead, incorporating full nonbonded interactions (i.e., with full van der Waals and electrostatic interactions and implicit solvent) in the initial stage of structure calculations (from extended structures) significantly increases the roughness of the energy landscape and reduced the computational efficiency. Similar observations were made when REX was used, even though REX is believed to offer better sampling (more discussion in the following sections). Instead, the popular simulated annealing protocols with a simplified representation of nonbonded interactions, and particularly with torsion angle molecular dynamics (TAMD) from very high temperature, is indeed an extremely efficient way of finding an ensemble of structures that satisfy the experimental restraints. Therefore, in the following sections, we will focus on utilizing the implicit solvent model in the context of refining the structures generated by simulated annealing calculations using conventional programs.

Refinement Experiments with Reduced NMR Restraints. Since current force fields and simulation techniques are not mature enough to predict the native structure of a protein from its sequence in general, it is necessary to rely on the experimental restraints in order to unambiguously determine the 3D structures. However, one might be able to benefit significantly from a more accurate empirical force field when the experimental data are limited. In practice, one often needs to deal with limited experimental data. For example, only a limited number of NOE restraints can be assigned in early stages of NMR structure determination, which is particularly true for large proteins with multiple domains largely due to increased resonance overlap and degraded spectral quality. The complete evaluation and assignment of NOEs typically rely on recursive algorithms,[29−31] where structures calculated from preliminary

(25) Feig, M.; Karanicolas, J.; Brooks, C. L., III. *MMTSB Tool Set*, MMTSB NIH Research Resource; The Scripps Research Institute: La Jolla, CA, 2001.
(26) Feig, M.; Karanicolas, J.; Brooks, C. L., III. *J. Mol. Graph. Model.* **2004**, *22*, 3777−3795.
(27) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604−1608.
(28) Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. *Science* **1991**, *253*, 657−661.

Table 2. Results of REX Refinement Experiments on GB1[a]

		initial statistics		final statistics	
subset	LR NOE[b]	RMSD[c]	NOE[d]	RMSD[c]	NOE[d]
GB1:1	44 (15%)	1.45 ± 2.00	0.0/1.38/0.018	0.77 ± 0.54	0.0/0.06/0.008
GB1:2	36 (12%)	1.62 ± 2.40	0.0/1.69/0.020	1.09 ± 0.97	0.0/0.91/0.017
GB1:3	31 (10%)	1.88 ± 2.68	0.0/1.76/0.023	1.24 ± 1.21	0.0/0.92/0.013
GB1:4	17 (5.6%)	3.93 ± 3.77	0.0/0.65/0.014	2.32 ± 1.86	0.0/1.18/0.014
GB1:5	13 (4.3%)	3.82 ± 4.35	0.0/0.60/0.011	1.52 ± 0.56	0.0/0.66/0.011

[a] The initial ensembles of structures were generated by CNS using randomly selected subsets of the full NOE restraints. [b] Number and percentage of long-range NOE restraints randomly selected and used. [c] Backbone RMSD of the mean structure from X-ray structure ± backbone RMS fluctuation around the mean structure (Å). The X-ray structure (PDB: 1pgb) was relaxed by energy minimization in CHARMM with GBSW implicit solvent. [d] Average number of NOE restraints violated by more than 0.5 Å/average number of NOE restraints violated by more than 0.2 Å/RMSD of NOE restraints for all structures in the ensemble (Å).

assignments are used to verify and adjust existing assignments and to predict new assignments in each cycle. Another case is solid-state NMR structure determination, where only a limited amount of experimental data is typically available. To further explore the influence of implicit solvent in these cases, numerical experiments were carried out using subsets of the full experimental restraints. Only a portion of the available long-range NOE restraints (i.e., those involving atoms separated by more than four residues sequentially) were randomly selected and used with other restraints in the structure calculations. The RMS fluctuation of the structures around the mean was used to identify the appropriate percentage of long-range NOE restraints to be used. For GB1, only about 15−40 randomly selected long-range NOE restraints were retained to produce the initial structures. The backbone RMS fluctuations around the mean for the resulting ensembles were 2−4 Å, and the RMSD values of the average structures from the X-ray structure ranged from 2 to 4 Å. Including more long-range NOE restraints led to well-converged structures (with the RMS fluctuation around the mean on the order of 1 Å or less) and leaves little room for refinement with implicit solvent. Note that while in practice the NOEs that can be readily assigned might not be randomly distributed, random selection of subsets of long-range NOEs still reflects the real world situation to a certain extent. The most difficult NOEs, therefore the last NOEs assigned, are often those involving long-range contacts. The reason is that the fold is not known and little help is available when there is peak overlapping and/or chemical shift degeneracy. On the contrary, a significant portion of intraresidue, sequential, and medium range NOEs can be readily assigned even in the very early stages of structural determination, because information about the primary sequence and secondary structure is available and can be utilized to resolve ambiguous NOEs.

Table 2 summarizes the results of REX refinement of five sets of initial structures computed by CNS using five randomly selected subsets of the full NOE restraints. Significant improvement in the RMSD values from the X-ray structure was observed in all experiments. The NOE violation statistics are also slightly improved after the REX refinement in some cases. Such an improvement might reflect that with the implicit solvent less violation in NOE restraints is needed to sustain the correct fold. However, the change in NMR statistics is minimal compared to the structural improvement in most cases. Ensembles of REX refined structures show smaller RMS fluctuation around the mean. However, here a smaller RMS fluctuation around the mean does not necessarily reflect a tighter convergence as in conventional NMR structure calculations. This is partially due to the fact that typically only a few replicas contribute to the lowest temperature ensemble. This is an advantage of the REX refinement, because, as will be discussed further later in this section, only most native-like (lowest energy) conformations among the potentially diverse initial structures contribute to the lowest temperature ensemble, providing an effective way to select a few best structures from an ensemble of structures that satisfy all the experimental restraints similarly well.

It should also be stressed here that the initial structures produced by CNS:anneal.inp protocol already satisfy all the restraints extremely well and a different annealing protocol is not expected to introduce substantial improvement in the initial structures. For example, there is no NOE violation greater than 0.5 Å in all cases and the average numbers of NOEs violated by over 0.2 Å is less than 2.0 and sometimes less than 1.0. As we already noted in the previous section, additional refinement steps are often carried out after CNS:anneal.inp and these may improve both the structures and NMR statistics in some cases. To verify that the initial structures have already reached the limit of the data and will not be substantially improved without more data or a better force field with implicit solvent, additional conventional refinement calculations were performed for two initial structure ensembles, GB1:2 and GB1:3, which had the worst initial NOE statistics. The Xplor-NIH[4] script refine_gentle.inp was used. This script was designed to further relax the structures through room-temperature MD with full Lennard-Jones nonbonded functions, electrostatic interactions with a distance-dependent dielectric constant (RDIE), and dihedral terms. An NOE scale of 150 was used in all calculations. As expected, after the refinement, there is some slight improvement in the NOE statistics (to 0/1.00/0.022 and 0/1.25/0.022 for GB1:2 and GB1:3, respectively), but the structures stay virtually the same in terms of the backbone RMSD of the mean structure from X-ray and backbone RMS fluctuation around the mean structure (e.g., the final RMSD values are 1.55 ± 2.44 and 1.94 ± 2.70 Å, respectively).

Figure 1 shows several representative structures before and after the REX refinement. All of the representative structures were chosen from replicas that contributed significantly to the lowest temperature ensemble, which are presumably the most-native-like structures as predicted by REX with implicit solvent. The occupancies of these representative structures at the lowest temperature during the sampling period (last 100 REX steps in this case) are listed in the figure caption. Figure 1 illustrates that the REX refinement with implicit solvent can quickly bring the initial conformers closer to the native conformation. In particular, when only 13 long-range NOE restraints are used, as in GB1:5, some starting structures were completely different from the native with RMSD values as large as 10 Å while others deviated less with RMSD values of about 4 Å. The former are typically too distorted for refinement to bring any meaningful improvement in the given simulation length, but the latter can be improved significantly to an RMSD of less than 1 Å from the X-ray structure after the REX refinement with implicit

(29) Nilges, M.; Macias, M. J.; O'Donoghue, S. I.; Oschkinat, H. *J. Mol. Biol.* **1997**, *269*, 408−422.
(30) Duggan, B. M.; Legge, G. B.; Dyson, H. J.; Wright, P. E. *J. Biomol. NMR* **2001**, *19*, 321−329.
(31) Hermann, T.; Güntert, P.; Wüthrich, K. *J. Mol. Biol.* **2002**, *319*, 209−227.

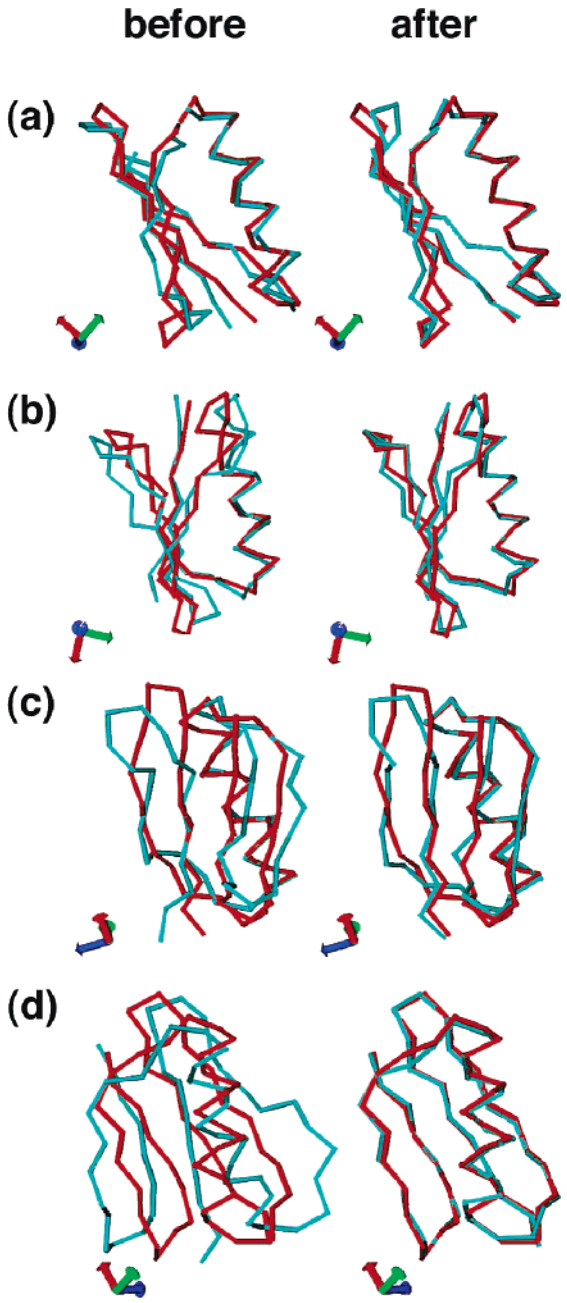


Figure 1. Representative structures before and after the REX refinement (blue) of the replicas that contributed to the lowest temperature ensemble. Panels (a)−(d) are from GB1:1−3 and 5, respectively (see Table 2). The occupancies of these replicas at the lowest temperature during the last 100 REX steps are 77, 28, 24, and 4%, respectively. The RMSD values from the X-ray structure (red) before and after refinement are as follows: (a) 1.7 Å/0.94 Å; (b) 2.3 Å/0.94 Å; (c) 2.3 Å/1.6 Å; (d) 4.0 Å/0.8 Å.

solvent, as illustrated in Figure 1d. Furthermore, these refined native-like conformations have lower energy in the CHARMM force field with the GB implicit solvent and occupy the lowest temperatures. For example, Figure 2 shows the energy, temperature, and RMSD profiles of the same replicas shown in Figure 1b,d. It appears that the REX refinement, on one hand, significantly improves the near-native structures and, on the other hand, is able to identify the most native-like conformers. The latter can be particularly useful as the structures computed with limited experimental data generally have a wide structural distribution. Structures that deviate significantly from the native

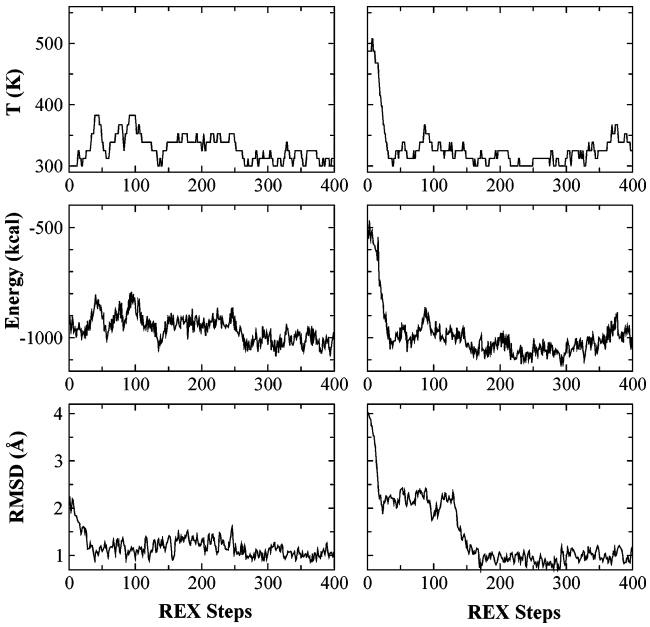


Figure 2. Representative profiles of temperature, energy, and backbone RMSD from the X-ray structure for two replicas that contribute to the lowest temperature ensembles. Replicas shown in the left and right columns are from GB1:2 and GB1:5 of Table 2, respectively. Corresponding structures before and after refinement are shown in Figure 1b,d.

conformation should be excluded, such as in structure-assisted recursive assignment and evaluation of NOE. To examine the ability to predict NOE restraints from the structures, we compare the histograms of the average number of long-range NOE violations per structure for the initial and REX-refined structures. All 302 long-range NOE restraints were included in the violation analysis, even though the structures were computed and refined with only small subsets of the restraints. As demonstrated in Figure 3, the REX-refined structures show much better agreement with the full long-range NOE restraints, reflecting a greater ability to predict new NOE restraints. It should be stressed here that it is important to use a high-quality implicit solvent model, because the ability to improve the structures and the reliability of selecting native-like conformations are both ultimately connected to the quality of the force field. Control calculations without the implicit solvent (i.e., with a constant dielectric constant of 1.0 or with a distance-dependent dielectric constant) all failed to significantly improve the structures or correctly select the most native-like conformations (data not shown).

The ability to correctly "rank" the refined structures completely relies on the ability of the employed force field to separate decoys from native-like conformations.[32] To further verify this ranking ability of the CHARMM force field with the GBSW implicit solvent, in Figure 4 we examine the correlation between RMSD from the X-ray structure and the dynamic average energy and the energy of the minimized structures. The dynamic average energy was computed as the average of the potential energy of snapshots during a 200 ps restrained MD simulation. Weak harmonic restraints were also applied to the backbone to prevent significant deviation from the initial conformation during the dynamics. It is clear that the RMSD from the X-ray structure shows a reasonable correlation with the dynamic average energy, in agreement with

(32) Feig, M.; Brooks, C. L., III. *Proteins* **2002**, *49*, 232−245.

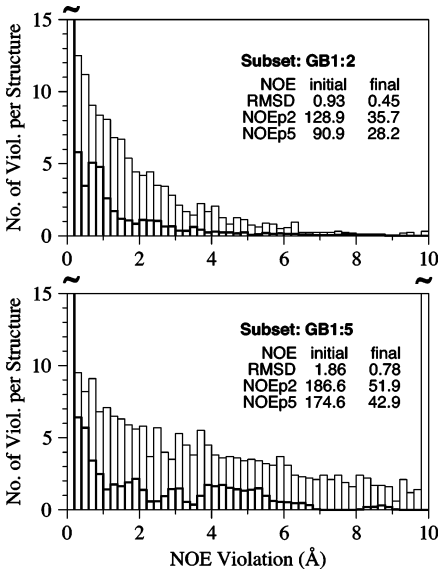


Figure 3. Histograms of the average number of long-range NOE violations per structure for the initial and REX-refined structures. While the structures were computed and refined with only subsets of the full experimental restraints, all long-range NOE restraints were included in the violation analysis. The bins at 10 Å include all NOE restraints violated by 10 Å or more. Note that the bins at 0.2 Å in both panels and 10 Å in the lower panel are truncated for better plotting. The height of the bin at 10 Å in the lower panel is 21.6. Notations: NOEp2 is the number of NOE restraints violated by 0.2 Å or more; NOEp5 is the number of NOE restraints violated by 0.5 Å or more. The light and heavy lines represent the initial and final histograms, respectively.

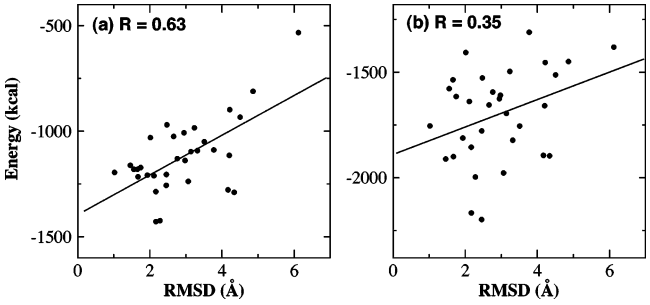


Figure 4. Correlation of (a) dynamic average energy and (b) minimized energy with the RMSD from X-ray structure. A total of 32 initial structures from sets GB1:1 and GB1:3 were used. In (a) average RMSD values of the trajectories were used. *R* is the correlation coefficient of linear fitting. The solid lines indicate the best fits.

what we observed in the REX refinement calculations. However, note that the energy of minimized structures does not correlate strongly with the RMSD from the X-ray structure, and thus is not a good scoring benchmark for selecting native-like conformers.[33]

Sampling Efficiency of Replica Exchange. The enhanced sampling of the REX method arises partially from the "self-regulated" automatic heating and cooling, which leads to efficient, adaptive sampling in the temperature space as well as in the conformational space. This is illustrated by the examples given in Figure 2 (more examples are shown in Figures 7 and 10). To demonstrate that such self-regulated annealing can enhance the sampling and accelerate convergence to native-like conformations, restrained simulated annealing simulations were carried out to refine the same initial conformations as those shown in Figure 2. The simulated annealing was implemented as a simple linear temperature ramp, with a typical starting temperature of 1000 K and an ending temperature of 300 K. Constant-temperature MD simulations were also carried out at 300 and 550 K for comparison. The same NOE force constants as those used in the REX refinement protocol were applied. The simulation length was 200 ps (matching the total length of the REX calculations), and snapshots were saved every 0.5 ps. As demonstrated by the two examples shown in Figure 5, the REX refinement appears to sample more efficiently and thus moves the structures toward the native conformation more rapidly in most cases.

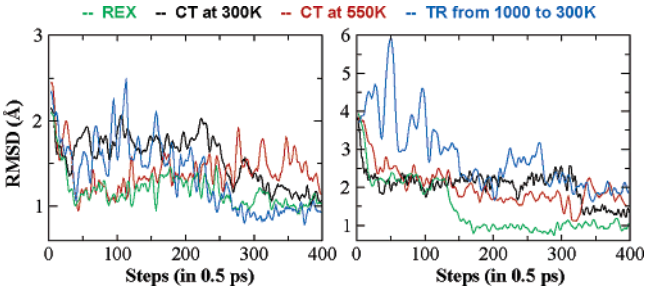


Figure 5. Backbone RMSD values from the X-ray structure as a function of the number of time steps during the simulated annealing simulations that started from the same initial conformations as the two replicas shown in Figure 2. Running averages over 5 points were used in plotting for clarity. Notations: CT corresponds to constant temperature; TR corresponds to a linear temperature ramp.

Larger Proteins. The numerical experiments on GB1 show that REX refinement with an implicit solvent can significantly improve the structures when the amount of experimental data alone is not sufficient to unambiguously define the 3D conformation. Since GB1 is a small protein with a very robust fold, the efficacy of the REX refinement protocol is further assessed by additional numerical experiments on three larger model proteins with varied folds.

GAIP and EIN. GAIP is a 217-residue protein with a 128-residue all α core and EIN is a 259-residue α/β protein. Both protein structures have been determined by NMR[19,20,34] (see Table 1). Due to the multidomain nature of both proteins, either residual dipolar couplings or rotational diffusion anisotropy has been used to improve the long-range order during structure refinement. In the following numerical experiments, we have used only the NOE (including hydrogen bonding) and dihedral angle restraints, and thus mainly rely on the force field to correctly define the relative orientation of the subdomains.

Table 3 summarizes the results of these numerical experiments. Subsets of the full NOE restraints were randomly selected and used in the CNS calculations to obtain poorly converged initial structures. Similar to the case of GB1, significant improvement in the structures, as reflected in the RMSD values, was generally achieved. The structures before and after the REX refinement show comparable, sometimes better, NOE and dihedral angle restraint statistics. Figures 6 and 7 show examples of the structures before and after refinement and corresponding temperature and RMSD profiles for replicas that contribute to the lowest temperature ensembles. Improvement in both local structure and long-range order is evident. Comparison of the histograms of the average number of long-range NOE violations

(33) Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. *J. Mol. Biol.* **2001**, *313*, 417−430.

(34) Garrett, D. S.; Seok, Y. J.; Liao, D. I.; Peterkofsky, A.; Gronenborn, A. M.; Clore, G. M. *Biochemistry* **1997**, *36*, 2517−2530.

Table 3. Results of the REX Refinement Experiments on GAIP and EIN[a]

		initial statistics			final statistics		
subset	LR NOE[b]	RMSD[c]	NOE[b]	DIHE[d]	RMSD[c]	NOE[b]	DIHE[d]
GAIP:1	46 (15%)	3.05 ± 5.96	0/0.69/0.020	0.034	1.55 ± 0.64	0/0.37/0.018	0.87
GAIP:2	34 (11%)	3.60 ± 3.87	0/2.9/0.023	0.064	1.68 ± 1.05	0/3.2/0.021	0.56
EIN:1	54 (8.9%)	4.82 ± 8.17	0/0.88/0.0093	0.15	2.54 ± 1.48	0/0.25/0.012	1.85
EIN:2	57 (9.4%)	4.14 ± 7.75	0/0.13/0.0068	0.081	2.42 ± 0.91	0/0.18/0.012	1.72

[a] The initial ensembles were generated by CNS using randomly selected subsets of the full NOE restraints. [b] See Table 2. [c] Backbone RMSD of the mean from the minimized average PDB NMR structure ± backbone RMS fluctuation around the mean structure (Å). The original PDB structures were relaxed by energy minimization in CHARMM with the GBSW implicit solvent. Only residues 2−230 were included in the RMSD calculations for EIN. [d] RMSD of dihedral angle restraints for all structures in the ensemble (deg).

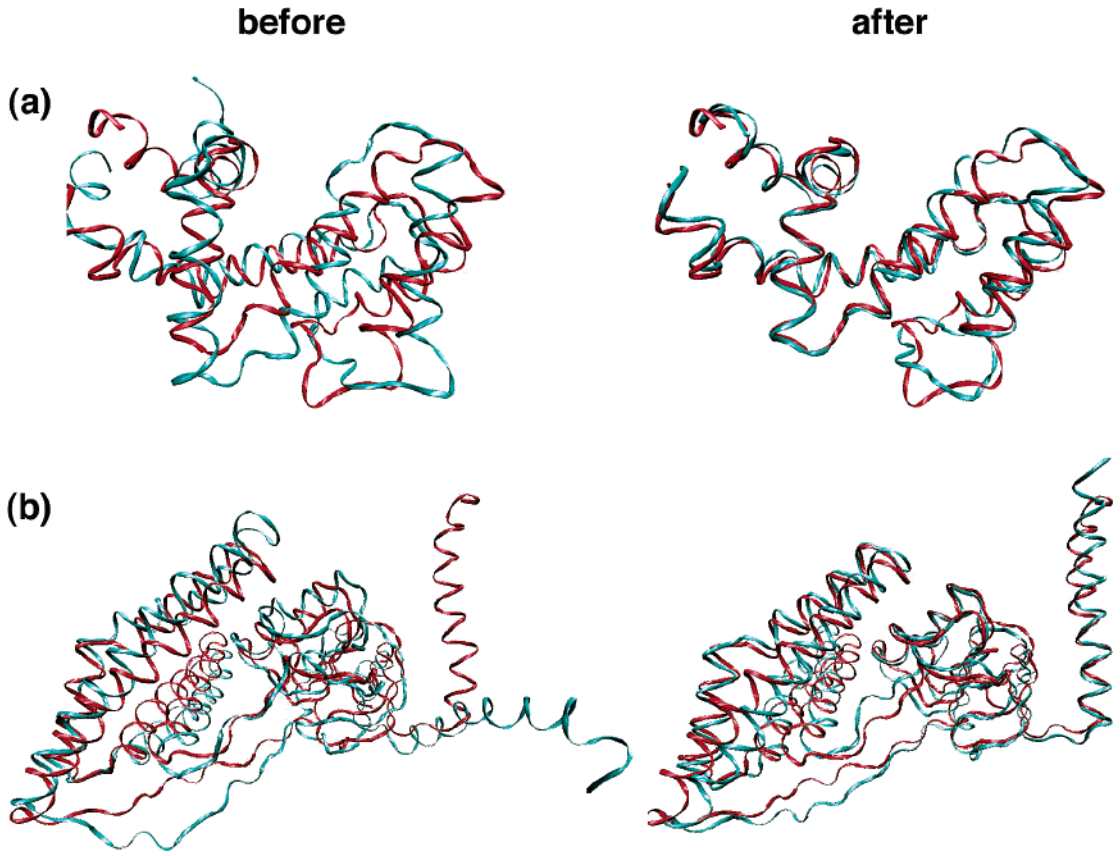


Figure 6. Examples of structures before and after the REX refinement (blue) for replicas that contribute to the lowest temperature ensembles, with corresponding temperature and RMSD profiles shown in Figure 7. Panels (a) and (b) are from experiments GAIP:2 and EIN:2, respectively, of Table 3 with occupancies of 11 and 16% at the lowest temperature during the last 25% of the REX calculations. The backbone RMSD values from the minimized average PDB structures (red) before/after refinement are (a) 6.3 Å/1.6 Å and (b) 4.5 Å/2.6 Å (residues 2−230).

per structure (using the full long-range NOE restraint sets) before and after the REX refinement, shown in Figure 8, also demonstrates that the refined structures can provide a much more accurate prediction of unassigned NOE restraints.

Typically, the GAIP replicas that contribute to the lowest temperature ensemble have smallest RMSD from PDB:1cmz. However, for EIN with longer and more flexible interdomain linkers, examination of the RMSD profiles of the two subdomains of EIN reveals some difficulty in predicting the relative orientation of α and α/β domains. For example, for the replica shown in Figure 7b, while the RMSD of each domain quickly converges to below 1.5 Å, the global RMSD fluctuates between 2 and 3 Å. This fluctuation might also be due to the intrinsic flexibility of the protein. Furthermore, we observed several cases where the lowest temperature ensemble was occupied by structures that did not have the lowest RMSD. Close examination of these structures revealed a limitation in the GB/SA model currently employed. Namely, the nonpolar solvation free energy, estimated from the surface area with a single phenomenological surface tension coefficient is not accurate enough.[35] The delicate balance between intramolecular and solute−solvent dispersion interactions, which is important in defining the interdomain orientation, is not well-maintained. For example, the C-terminal helix (residues 233−250) can bend and interact with α/β domain, resulting in more favorable nonnative intramolecular van der Waals interactions. These structures can have lower energy than more native-like structures, which lack these interactions. We found that using a smaller surface tension coefficient slightly improved the identification of native-like interdomain orientation. However, more accurate models of the nonpolar solvation free energy, such as a recent GB/NP model,[36]

(35) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *125*, 9523−9530.
(36) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479−499.

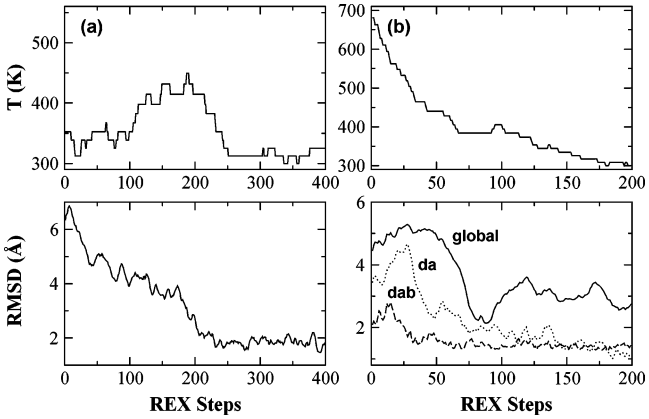


Figure 7. Temperature and backbone RMSD profiles for the two replicas shown in Figure 6. Notations: global = residues 2−230; da = α domain (residues 33−142); dab = α/β domain (residues 2−20 and 148−230).

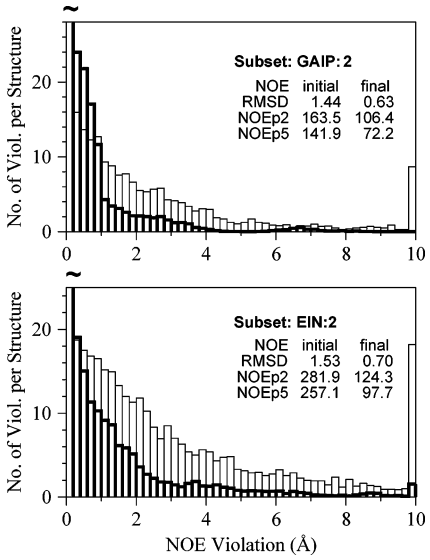


Figure 8. Histograms of the average number of long-range NOE violations per structure for the initial and REX refined structures. Notations as in Figure 3.

may be necessary for a more reliable and general prediction of the packing of flexible subdomains. In practice, when the GBSW/SA implicit solvent model is used, we recommend using a small surface tension coefficient (e.g., ∼10 cal/(mol·Å^2)) for multidomain proteins, while for single-domain proteins any reasonable value (e.g., 5−80 cal/(mol·Å^2)) can be safely used.

Realistic Test: Refinement of the 370-Residue Maltose-Binding Protein. The 370-residue MBP contains two domains and has a molecular weight of 42 kDa. The X-ray structure has been determined at 1.8 Å resolution[37] (PDB:1dmb). However, due to the difficulty in obtaining high-quality NMR spectra for larger proteins, the NMR structure of MBP has not been well-determined even with a combination of NOE, hydrogen bonding, dihedral angle, and residual dipolar coupling measurements.[16] It was shown that without the residual dipolar coupling restraints, the average pairwise backbone RMSD of the ensemble was 5.5 Å with an average backbone RMSD of 5.1 Å from PDB:1dmb. With residual dipolar coupling restraints, the average pairwise RMSD was reduced to 2.2 Å with an average RMSD of 3.3 Å from PDB:1dmb.[16] Therefore, MBP provides a very realistic and challenging test for the present REX refinement protocol.

Table 4. Results of the REX Refinement of MBP[a]

	initial	final
RMSD to PDB:1dmb[b] (Å)		
global	4.3 ± 4.1	2.3 ± 2.6
N-domain	2.5 ± 2.1	2.2 ± 1.4
C-domain	3.0 ± 3.2	2.0 ± 1.9
ϕ/ψ space: residues[c] (%)		
most favored region	72.2	84.3
additionally allowed region	22.8	13.3
generously allowed region	3.8	1.6
disallowed region	1.2	0.8
violation statistics		
RMSD of NOEs (Å)	0.047	0.014
NOE violations > 0.2 Å	2.85	4.42
RMSD of DIHE (deg)	0.53	6.25

[a] All available NOE and dihedral angle restraints were used in the structure calculation and refinement. [b] Backbone RMSD of the mean structure from the X-ray structure ± the RMS fluctuation around the mean. Global, residues 6−235 and 241−370; N-domain, residues 6−109 and 264−309; C-domain, residues 114−235, 241−258 and 316−370. [c] Calculated with PROCHECK_NMR.[38]

All 1943 NOE, 555 dihedral angle and 48 hydrogen bonding restraints were used in CNS to generate an initial set of 256 structures. The standard simulated annealing protocol (CNS input file: anneal.inp) was used with 120 ps high-temperature TAMD at 50 000 K followed by 45 ps TAMD and 25 ps Cartesian MD for slow cooling stages. A set of 48 structures with the lowest overall CNS energies was selected and then refined by REX with GBSW implicit solvent. The average RMSD of these initial structures with respect to PDB:1dmb is 5.7 Å for global residues (residues 6−235 and 241−370 as defined in ref 16). Forty-eight replicas were used at temperatures ranging from 300 to 800 K. The calculation was carried out until the energy of the lowest temperature ensemble converged, which occurred within 1000 REX cycles. The last 200 structures from the lowest temperature ensemble were minimized with 200 steps of mixed steepest descent and adopted basis Newton−Raphson (ABNR) minimization and then analyzed. The results are summarized in Table 4. Figure 9 shows the initial and final structures for two representative replicas that contribute to the lowest temperature ensemble between REX step 800 to 1000. The corresponding temperature and RMSD profiles are shown in Figure 10. After the REX refinement, significant improvement in both global and subdomain structures was observed, to some extent, at the cost of a moderate increase in dihedral angle restraint violations. Significant improvement was also seen in the percentage of backbone torsion angles in the most favored region of the ϕ/ψ space. Even though there is a slight increase in the average number of NOE restraints violated by over 0.2 Å, the RMSD of the NOE restraints is reduced from 0.047 to 0.014 Å after refinement, indicating that the final structures satisfy the overall NOE restraints better. Histograms of the global backbone RMSD from PDB:1dmb for the initial and final ensembles are shown in Figure 11. The average RMSD from PDB:1dmb of the final ensemble is improved to 3.4 Å, which is comparable to the reported value of 3.3 Å for the structures computed with additional residual dipolar coupling restraints.[16]

(37) Sharff, A. J.; Rodseth, L. E.; Quiocho, F. A. *Biochemistry* **1993**, *32*, 10553−10559.

(38) Laskowski, R. A.; Rullmann, J. A. C.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M. *J. Biomol. NMR* **1996**, *8*, 477−486.

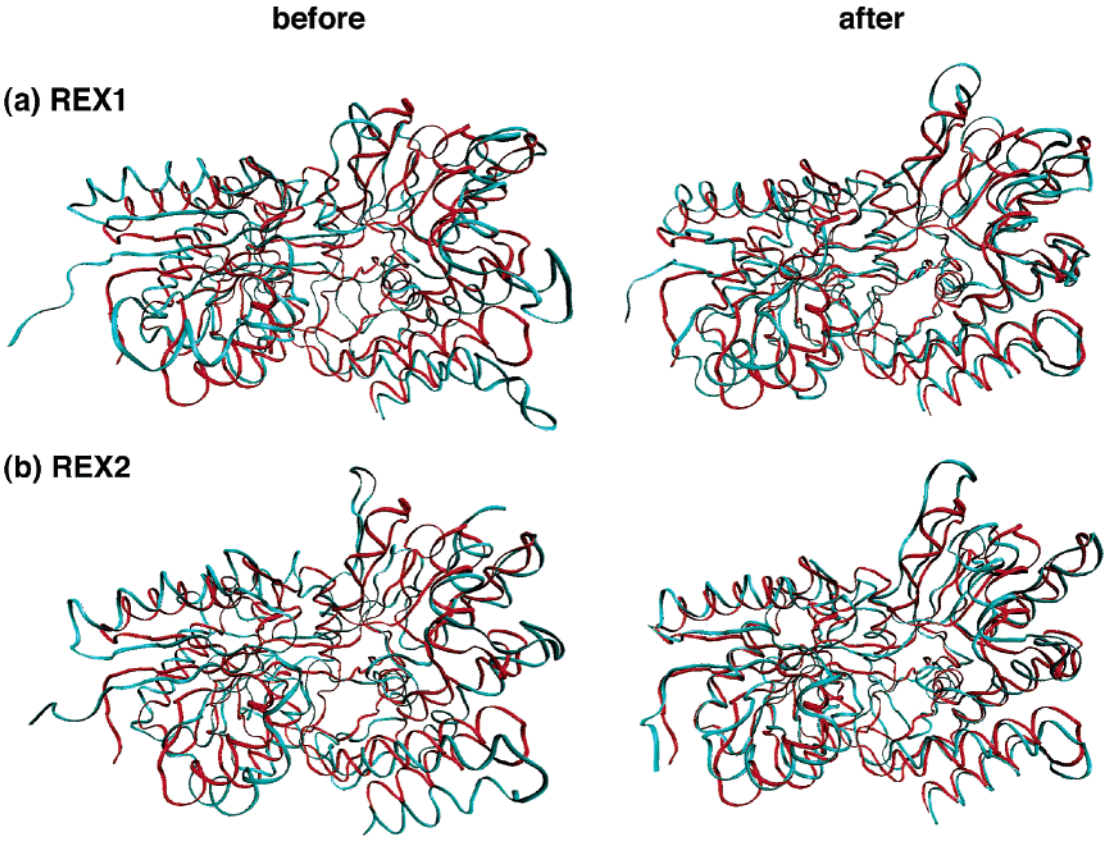


Figure 9. Examples of structures before and after the REX refinement (blue) for replicas that contribute to the lowest temperature ensembles, with corresponding temperature and RMSD profiles shown in Figure 10. The occupancies of REX1 and REX2 at the lowest temperature during the last 200 REX steps are 52 and 28%, respectively. Global backbone RMSD values with respect to PDB:1dmb (red) before/after refinement are (a) 5.8 Å/2.9 Å and (b) 5.7 Å/3.5 Å.

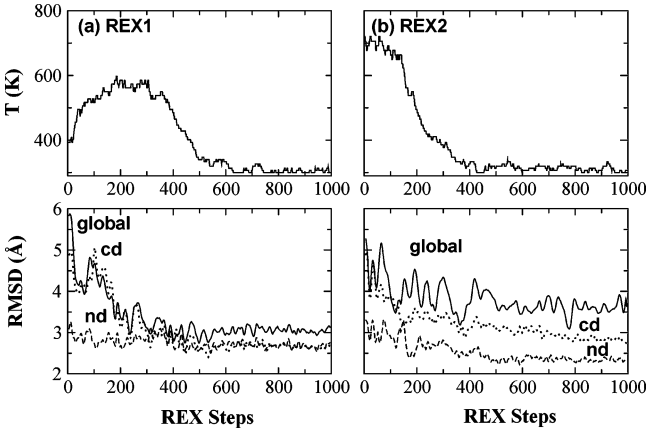


Figure 10. Temperature and backbone RMSD profiles of the replicas shown in Figure 9. Notations: nd = N-domain; cd = C-domain. See caption of Table 4 for the residue ID ranges. The RMSD was plot with running averages over 10 points for clarity.

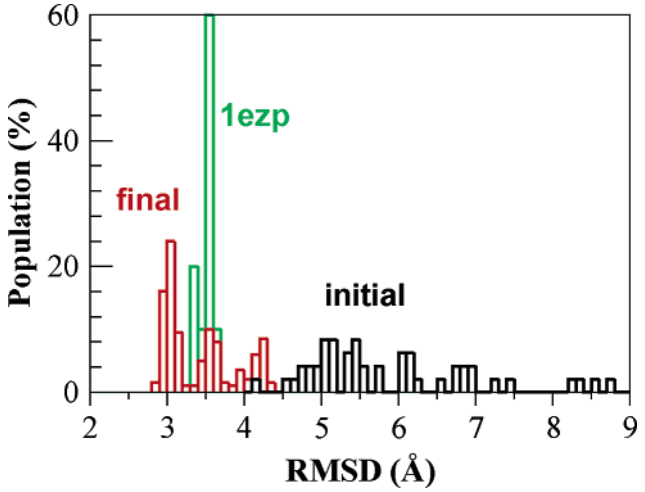


Figure 11. Histograms of the global backbone RMSD from PDB:1dmb of the initial and final structure ensembles, in comparison with that of PDB: 1ezp (10 structures, obtained with additional residual dipolar coupling restraints).

Conclusions

We have carried out numerical experiments to study the influence of molecular mechanics force fields on the refinement of NMR structures by NMR inferred restraints using several model protein systems of various sizes and topologies. We have also investigated the application of advanced sampling techniques such as the REX method to the NMR structure refinement process. It was found that when there were sufficient experimental restraints, the force field had very little influence on the final structures, even though some improvement in the backbone ϕ, ψ distribution and hydrogen bond pattern was previously observed.[11] In this case, empirical conformational database potentials derived from high-resolution X-ray structures[39,40] might be more effective in further refinement, or at least, in bringing the final NMR structures closer to the X-ray structures. However, an accurate force field with implicit solvent can have a significant impact when the experimental restraints alone are

(39) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *Protein Sci.* **1996**, *5*, 1067−1080.
(40) Grishaev, A.; Bax, A. *J. Am. Chem. Soc.* **2004**, *126*, 7281−7292.

not sufficient to unambiguously determine the structure. Initial structures generated by conventional NMR software such as CNS and DYANA can be quickly moved toward the native basin when refined with implicit solvent using the REX method. Slight improvement in the NMR restraint violation statistics was also observed in some cases. The REX refinement method has advantages over the conventional simulated annealing methods in terms of enhanced conformation sampling. Furthermore, an ensemble of most native-like conformations can be automatically selected through the REX refinement process. Application of the proposed refinement protocol to the 370-residue MBP was able to achieve an improvement in the structures that was comparable to what was obtained by refinement with additional residual dipolar coupling restraints,[16] demonstrating the efficacy of the proposed protocol. We expect REX refinement with an implicit solvent to be very useful in the early stages of NMR structure determination where only limited data are available. The improved preliminary structures can be effectively used to evaluate the NOE restraints and resolve ambiguous NOEs. We also expect the proposed protocol, in combination with a recently developed implicit membrane GB model,[41] to be useful when it is practically difficult to obtain redundant experimental restraints such as in solid-state NMR structure determination.

Acknowledgment. C.J.H. thanks the La Jolla Interfaces in Science interdisciplinary training program for fellowship support. The authors are grateful to Carol Post for helpful discussion. We also thank Peter E. Wright, H. Jane Dyson, and Hyung-Sik Won for useful comments. This work was supported by grants from the National Institutes of Health (Grants GM48807 and RR12255) and National Science Foundation (Grants PHY0216576 and PHY0225630).

JA047624F

(41) Im, W.; Feig, M.; Brooks, C. L., III. *Biophys. J.* **2003**, *85*, 2900−2918.