

Accurate Automated Protein NMR Structure Determination Using Unassigned NOESY Data

Srivatsan Raman,^{†,||} Yuanpeng J. Huang,[‡] Binchen Mao,[‡] Paolo Rossi,[‡] James M. Aramini,[‡] Gaohua Liu,[‡] Gaetano T. Montelione,[‡] and David Baker^{*,†,§}

Department of Biochemistry, University of Washington, Seattle, Washington 98195, Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, and Northeast Structural Genomics Consortium, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, and Howard Hughes Medical Institute, Seattle, Washington 98195

Received July 16, 2009; E-mail: dabaker@u.washington.edu

Abstract: Conventional NMR structure determination requires nearly complete assignment of the cross peaks of a refined NOESY peak list. Depending on the size of the protein and quality of the spectral data, this can be a time-consuming manual process requiring several rounds of peak list refinement and structure determination. Programs such as Aria, CYANA, and AutoStructure can generate models using unassigned NOESY data but are very sensitive to the quality of the input peak lists and can converge to inaccurate structures if the signal-to-noise of the peak lists is low. Here, we show that models with high accuracy and reliability can be produced by combining the strengths of the high-resolution structure prediction program Rosetta with global measures of the agreement between structure models and experimental data. A first round of models generated using CS-Rosetta (Rosetta supplemented with backbone chemical shift information) are filtered on the basis of their goodness-of-fit with unassigned NOESY peak lists using the DP-score, and the best fitting models are subjected to high resolution refinement with the Rosetta rebuild-and-refine protocol. This hybrid approach uses both local backbone chemical shift and the unassigned NOESY data to direct Rosetta trajectories toward the native structure and produces more accurate models than AutoStructure/CYANA or CS-Rosetta alone, particularly when using raw unedited NOESY peak lists. We also show that when accurate manually refined NOESY peak lists are available, Rosetta refinement can consistently increase the accuracy of models generated using CYANA and AutoStructure.

Introduction

NMR is a powerful method for protein structure determination. Conventional structure determination by NMR requires complete assignment of the chemical shifts (backbone and side chain) and complete assignment of the NOESY peak list. In general, the structure determination process goes through several iterations of compiling a NOESY peak list, assignment of NOESY cross peaks to sequence-specific interactions, structure generation and assessment, refinement of NOESY peak lists (i.e., distinguishing the real peaks from noise and artifacts), reassignment of the cross peaks, etc. The process evolves into an iterative effort to refine the NOESY peak list while simultaneously refining the 3D protein structure. While automated structure determination programs such as Aria,¹ CYANA,² or AutoStructure³ can successfully assign a large fraction of the NOESY peaks for small proteins when provided with high-quality NOESY peak list data, resulting in accurate structures,

challenges arise when the NOESY peak lists contain artifacts or when key long-range NOESY data are weak and/or not well distinguished from noise. In cases where the initial structures of the trajectory are not well-defined by the available unambiguous data, inaccurate initial structures may cause mis-assignment of NOESY cross peaks, which are then propagated in the process of assigning additional NOESY cross peaks in subsequent steps. Accordingly, the programs are less robust for intermediate-sized and larger proteins (e.g., >150 residues) and do not perform well with poorer quality NOESY data. In general, for these systems a substantial part of the effort of structure refinement involves manual NOESY peak list refinement.

Rosetta can consistently generate high-accuracy models for small proteins starting from backbone chemical shift information alone.⁴ However, the CS-Rosetta method does not generally converge for complex protein folds or for proteins of >110 residues. Here, we demonstrate that for these more challenging proteins, the lack of convergence resulting from the increase in the size of the conformational space that must be sampled can be overcome in part by using the unassigned NOESY peak list as a filter to select out the best models, followed by intensive sampling around these models. In cases where the NOESY data

[†] University of Washington.

[‡] Rutgers, The State University of New Jersey.

[§] Howard Hughes Medical Institute.

^{||} Current address: Department of Genetics, Harvard Medical School, Boston, MA 02115.

(1) Rieping, W.; Habeck, M.; Bardiaux, B.; Bernard, A.; Malliavin, T. E.; Nilges, M. *Bioinformatics* **2007**, *23*, 381–382.

(2) Guntert, P. *Eur. Biophys. J.* **2009**, *38*, 129–43.

(3) Huang, Y. J.; Tejero, R.; Powers, R.; Montelione, G. T. *Proteins* **2006**, *62*, 587–603.

(4) Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G.; Eletsky, A.; Wu, Y.; Singarapu, K. K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C. H.; Szyperski, T.; Montelione, G. T.; Baker, D.; Bax, A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4685–4690.

are sparse or incomplete, the resulting energy-optimized structures can be more accurate than those generated from such data with conventional semiautomated NOESY assignments methods. In particular, we demonstrate that the need for manual intervention for NOESY peak list refinement, a significant bottleneck for many automated analysis methods, can be reduced or eliminated by exploiting the Rosetta force field and high-resolution sampling methodology to resolve the ambiguities inherent in unassigned NOESY NMR spectra. Finally, we show that high-resolution Rosetta refinement can improve the accuracy of close to native models generated automatically by AutoStructure and CYANA from refined peak lists, reducing the efforts required in the final stages of protein NMR structure refinement.

Methods

We describe two methods to combine the Rosetta methodology with unassigned NOESY peak lists to determine protein structures at atomic-level accuracy. Both approaches require NOESY peak list data and essentially complete chemical shift assignments (backbone and side chain). The first method, called CS-DP-Rosetta, uses minimally edited raw NOESY peak lists prepared by automatic peak picking of the NOESY spectra using 2D HSQC root spectra. The second method, called AssignNOE-Rosetta, uses more refined NOESY peak lists generated by expert human manual refinement of the raw peak list (see Supporting Information for a complete description). The models generated by the second approach, which rely on iteratively refined high-quality NOESY peaks lists, are generally more accurate. However, the manual intervention required for NOESY peak list refinement is time-consuming and dependent on user expertise. The approach is demonstrated on a set of proteins produced by the Northeast Structural Genomics Consortium (NESG). The following proteins were used (Swiss-Prot entries): Q9AAR9_CAUCR, Q8ZRJ2_SALTY, YPPE_BACSU, UFC1_HUMAN, P95883_SULSO, Q67Z52_ARATH(11-97), ARI3A_HUMAN(218-351), and A6B4U8_VIBPA (hereafter referred to by their respective NESG IDs: CcR55, StR65, SR213, HR41, SsR10, AR3436A, HR4394C, and VpR247). A description of the protein production and purification and the systematic method for obtaining raw and refined peaklists are given in Supporting Information together with the data acquisition and processing scheme. The statistics for the peaklists used in the study are reported in Table S1 in Supporting Information.

Model Generation with Raw Peak Lists (CS-DP-Rosetta Protocol). The first step in this protocol is the generation of 50,000 models using CS-Rosetta. The lowest-energy ~ 1000 CS-Rosetta⁴ models are then filtered on the basis of their fit to the unassigned NOESY data. Briefly, given a model, essentially complete backbone and side chain resonance assignments, and unassigned NOESY peaks, the RPF NMR software⁵ assesses the global agreement between the experimental NOESY peak list and a NOESY peak list simulated from the structure. The program reports a discriminating power (DP) score that is normalized on the basis of an estimate of the completeness of the NOESY peak list data and the goodness-of-fit to a random coil structure; models with DP-score of 1 are excellent fits to the NOESY peak list data, whereas a model with DP-score of 0 fits the data no better than a random coil. The DP-score is correlated with the accuracy of the model⁵ and so can be used to identify CS-Rosetta models that have more native-like global structures.

The best 20 models based on a linear combination of CS-Rosetta all-atom energy + 1000(1 - DP-score) are chosen for a second stage of refinement. In the second stage, the Rosetta rebuild-and-

refine⁶ protocol is carried out to focus sampling on regions that have not adequately converged to the lowest energy conformation in the first round. The regions to be rebuilt are identified by choosing residues with the largest C- α deviations in the lowest energy 20 models from the first stage. In the rebuild-and-refine protocol, these selected regions are stochastically rebuilt by fragment insertion and CCD loop closure⁷ followed by all-atom refinement of the entire structure using the physically realistic Rosetta forcefield.⁸ After the second step, the best 10 models by Rosetta all-atom energy and DP-score are chosen as the final models.

Model Generation with Refined Peak Lists (AssignNOE-Rosetta Protocol). With refined peak lists, programs such as AutoStructure or CYANA are capable of generating nearly correct models with unassigned NOESY data. However, these models can still show significant backbone and side chain differences compared with the native structure, providing ample scope for further refinement. In the AssignNOE-Rosetta protocol, models from the ensemble generated by CYANA or AutoStructure are used as starting points for the rebuild-and-refine protocol described above. The residues with maximum C- α deviation in the CYANA/AutoStructure ensemble are chosen for rebuilding; these regions are usually loops, edges of regular secondary structure elements, or chain termini.

Detailed descriptions of how to run the CS-DP-Rosetta and AssignNOE-Rosetta protocols complete with Rosetta and AutoStructure command line arguments are given in the Supporting Information.

Results

There are two sources of information available for determining protein structures. First, any available experimental data greatly constrains the space of possible structures. Programs such as Aria, AutoStructure, and CYANA use elegant algorithms to generate structures consistent with input NOESY data. Second, native structures, to be highly populated, must be the lowest free energy accessible conformations for their amino acid sequences, and this in principle is sufficient to completely determine protein structures. In practice, finding the global free energy minimum is a formidable search problem, and experimental data can be extremely valuable in constraining the search.

We have explored two methods for combining the CYANA/AutoStructure capabilities of generating models based on unassigned NOESY peak lists with the global energy optimization algorithms in Rosetta. We begin by illustrating the two approaches for the *Bacillus subtilis* protein SR213 in Figure 1. Using a refined NOESY peak list produced with expert curation of the raw peak list, CYANA and AutoStructure generate topologically correct models (Figure 1D). In this case, we have found it quite effective to start Rosetta high resolution refinement searches from these starting points, which can further increase the accuracy of the models (compare Figure 1D to 1E) by minimizing the energy (Figure 1A, from purple to light blue). We refer to this approach as AssignNOE-Rosetta. This energy minimization with Rosetta of the automatically generated NMR structure produced with CYANA or AutoStructure builds on previous work refining PDB deposited NMR structures for use in molecular replacement.^{6,9}

(6) Qian, B.; Raman, S.; Das, R.; Bradley, P.; McCoy, A. J.; Read, R. J.; Baker, D. *Nature* **2007**, *450*, 259–264.

(7) Canutescu, A. A.; Dunbrack, R. L., Jr. *Protein Sci.* **2003**, *12*, 963–972.

(8) Bradley, P.; Misura, K. M.; Baker, D. *Science* **2005**, *309*, 1868–1871.

(9) Ramelot, T. A.; Raman, S.; Kuzin, A. P.; Xiao, R.; Ma, L. C.; Acton, T. B.; Hunt, J. F.; Montelione, G. T.; Baker, D.; Kennedy, M. A. *Proteins* **2009**, *75*, 147–167.

(5) Huang, Y. J.; Powers, R.; Montelione, G. T. *J. Am. Chem. Soc.* **2005**, *127*, 1665–1674.

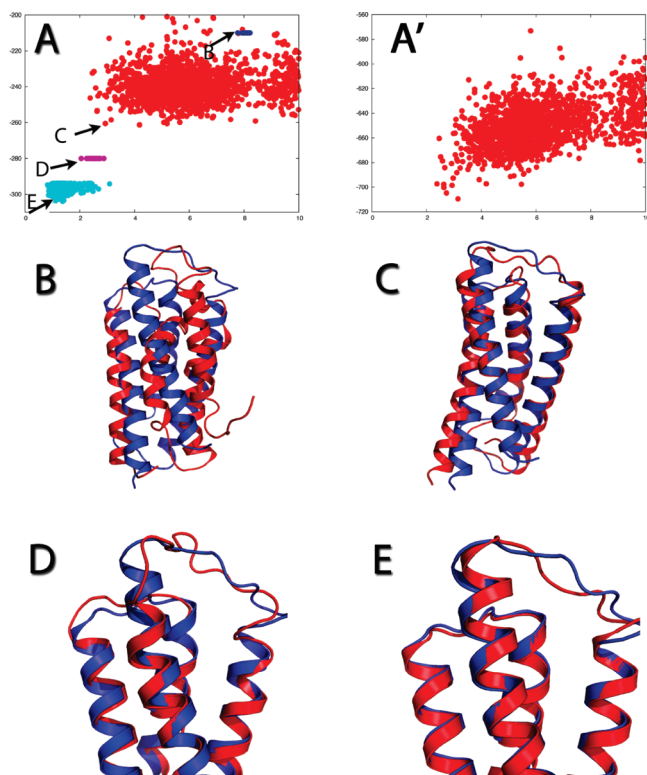


Figure 1. Model generation from raw and refined peak lists with CYANA/AutoStructure and Rosetta for protein SR213. (A) Rosetta all-atom energy vs rmsd to the X-ray structure. Dark blue points are CYANA/AutoStructure models from raw peak lists with energy set to arbitrary value. Red points are Rosetta models after the CS-DP-Rosetta protocol using raw peak lists. Purple points are CYANA/AutoStructure models from refined peak lists with energy set to arbitrary value. Light blue points are Rosetta models generated by AssignNOE-Rosetta refinement protocol starting from the purple points. (A') Rosetta all-atom energy + DP-score vs rmsd to X-ray structure for Rosetta models after the CS-DP-Rosetta protocol from raw peak lists (red points in panel A). It should be noted that the Rosetta energy function correctly assigns very low energies to the models less than 2 Å from the native structure in light blue in panel A; adding the DP-score improves discrimination of models somewhat further from the native structure (2–3 Å). (B–E) Superposition of the X-ray structure (dark blue) with the best CYANA/AutoStructure model from raw peak lists (B), best Rosetta model after the CS-DP-Rosetta protocol using raw peak lists (C), best CYANA/AutoStructure model from refined peak lists (D), and the best Rosetta model after the AssignNOE-Rosetta model generation protocol using refined peak lists. The arrows in panel A indicate the models chosen for superposition in panels B–E.

If on the other hand the NOESY peak lists are not refined and contain extensive spurious noise peaks, automated NOESY analysis methods such as CYANA and AutoStructure may produce models that are much less accurate and even topologically incorrect (Figure 1B). In this case, we have found it most effective to generate models using Rosetta with chemical shift information to guide fragment selection (CS-Rosetta) and to then select from the lowest energy models generated those for which the unassigned NOESY peak list data, back calculated with RPF, agrees best with the unrefined NOESY peak list data (the DP-score, Figure 1A'). The DP-score accounts for all possible assignments of each NOESY cross peak, given the list of resonance assignments and an estimate of the uncertainty in matching NOESY peaks to chemical shift values.⁵ This is a less deterministic use of noisy NOESY peak list data than in traditional NMR structure determination protocols, and can avoid inaccurate interpretation of spurious noise peaks. The selected models are then

subjected to the previously described Rosetta rebuild-and-refine protocol with sampling focused on the regions that differ in the selected models. We refer to this approach as CS-DP-Rosetta. This approach can produce quite good models (Figure 1C) that are generally somewhat higher in energy and rmsd (Figure 1A, colored red) than those produced by the first method because the starting point is further from the native structure. This approach has the important feature of being able to generate high quality structures without the need for manual iterative refinement of the NOESY peak list data.

The results with the two new methods on a series of test cases are described in the following sections. Since AutoStructure and CYANA consistently produce good models only when refined peak lists are available, we focused our testing of the AssignNOE-Rosetta protocol on cases with refined peak lists and tested the CS-DP-Rosetta protocol on cases with raw unedited NOESY peak lists. The native structure and all homologous structures were excluded from the database used in the initial fragment selection to mimic the new fold structure determination scenario.

Test Cases with CS-DP-Rosetta Protocol. The CS-DP-Rosetta protocol was initially tested on four proteins (CcR55, SR213, StR65, and HR41) ranging in size from 100 to 160 residues for which raw unedited NOESY peak list data were provided by the NESG (www.nesg.org). For comparison, we used both CYANA or AutoStructure and CS-Rosetta alone. The models generated by the new protocol were consistently better than those generated by either CYANA/AutoStructure or CS-Rosetta alone (Table 1A; Tables S3 and S4 in Supporting Information for DP-score, recall, and precision measures using raw and refined peak lists, respectively; Table S6 in Supporting Information for inter-ensemble rmsd). For all cases, except HR41, the low energy models were very close to the native structure. The combined Rosetta all-atom energy and DP-score identified the near-native models better than the Rosetta all-atom energy alone (see Figure 1A'). The 20 models with the best combined score converged to the same fold with an average inter-ensemble rmsd of 0.96 Å over the core residues. The regions with large coordinate deviations were largely loops or edges of secondary structure elements. Resampling these regions in the second refinement phase resulted in much better converged models. HR41 is a relatively large protein (160 residues), and the new protocol is unsuccessful (data not shown) because CS-Rosetta does not generate models close enough to the native structure for the Rosetta all-atom energy and DP-score to favorably discriminate.

Blind Test Cases. After benchmarking the protocol with proteins with known structure, we tested the CS-DP-Rosetta protocol on three blind test cases (VpR247, AR3436A, and HR4394C). Two of the three proteins VpR247 and AR3436A, were targets in the E-NMR blind structure determination experiment.¹⁰ Following the public release of the native

- (10) Rosato, A.; Bagaria, A.; Baker, D.; Bardiaux, B.; Cavalli, A.; Dorelejers, J. F.; Giachetti, A.; Guerry, P.; Guntert, P.; Herrmann, T.; Huang, Y. J.; Jonker, H. R.; Mao, B.; Malliavin, T. E.; Montelione, G. T.; Nilges, M.; Raman, S.; van der Schot, G.; Vranken, W. F.; Vuister, G. W.; Bonvin, A. M. *Nat. Methods* **2009**, *6*, 625–626.
- (11) Snyder, D. A.; Montelione, G. T. *Proteins* **2005**, *59*, 673–686.
- (12) Bradley, P.; Baker, D. *Proteins* **2006**, *65*, 922–929.
- (13) Bertone, P.; Kluger, Y.; Lan, N.; Zheng, D.; Christendat, D.; Yee, A.; Edwards, A. M.; Arrowsmith, C. H.; Montelione, G. T.; Gerstein, M. *Nucleic Acids Res.* **2001**, *29*, 2884–2898.
- (14) Goh, C. S.; Lan, N.; Echols, N.; Douglas, S. M.; Milburn, D.; Bertone, P.; Xiao, R.; Ma, L. C.; Zheng, D.; Wunderlich, Z.; Acton, T.; Montelione, G. T.; Gerstein, M. *Nucleic Acids Res.* **2003**, *31*, 2833–2838.

Table 1. Improvement in Model Accuracy Using Unassigned NOESY Peak Lists^a

(A) CS-DP-Rosetta (raw NOESY peak lists)			
protein name (length)	CS-DP-Rosetta model	CYANA/AutoStructure model	CS-Rosetta model
CcR55 (116 aa)	2.42 (1.86)	1.71 (1.68)	7.40 (5.68)
SR213 (123 aa)	2.93 (2.37)	8.03 (7.76)	6.15 (3.65)
StR65 (100 aa)	1.40 (1.10)	2.84 (1.45)	7.44 (5.91)
(B) AssignNOE-Rosetta (refined NOESY peak lists)			
protein name (length)	AssignNOE-Rosetta model	CYANA/AutoStructure model	PDB-deposited NMR ensemble
CcR55 (116 aa)	1.40 (1.15)	2.36 (2.04)	1.39 (1.21)
SR213 (123 aa)	0.99 (0.92)	2.54 (2.05)	2.30 (2.00)
StR65 (100 aa)	1.26 (1.02)	1.27 (1.13)	1.21 (1.10)
HR41 (160 aa)	1.41 (1.08)	1.68 (1.58)	1.44 (1.23)
SsR10 (129 aa)	1.19 (1.08)	1.93 (1.59)	1.25 (1.02)

^a Column 2 in sections A and B are the median rmsd to native of the 10 lowest energy models. Column 3 in sections A and B are the median rmsd to native in the CYANA/AutoStructure ensemble using the raw and refined peak lists, respectively. Column 4 in section A denotes the median rmsd of the 10 lowest energy models generated using CS-Rosetta (without DP-score filtering) and in section B denotes the median rmsd to the X-ray structure of all the conformers in the PDB-deposited NMR ensemble. The numbers in parentheses denote the lowest rmsd model in the ensemble. All rmsd's are computed with reference to the X-ray over the core residues as identified by FindCore.¹¹ The number of core residues are the following: CcR55, 85 aa; SR213, 103 aa; StR65, 77 aa; HR41, 125 aa; and SsR10, 107 aa. The rmsd's over the full length are shown in Table S2 in Supporting Information. The protein names are NESG target id's; detailed protein sequence data for these targets are available from the SPINE database.^{13,14}

structures, we found that our model ensembles agreed well with the native structures, as shown in Figure 2.

For VpR247 and AR3436A, the CS-DP-Rosetta models were generated using refined peak lists for DP-score calculations, while raw peak lists were used for HR4394C.

For VpR247, the CS-DP-Rosetta protocol converged on an ensemble of low energy models in good agreement with the final refined NOESY peak list (DP-score = 0.62). The average rmsd of the low energy models to the first structure in the NMR ensemble was 2.4 Å over the full length and 1.8 Å over the core residues. As shown in Figure 2A, most regions of the model ensemble are nearly as well converged as the NMR ensemble including the relatively long loop spanning residues 13–20. However, for loop residues 46–52, our model ensemble shows greater variation than the NMR ensemble.

In the case of AR3436A (Figure 2B), the CS-DP-Rosetta model ensemble had a well-packed hydrophobic core and showed excellent convergence with an inter-ensemble rmsd of 0.26 Å over the core residues, but the rmsd to the independently determined NMR structure was surprisingly high (~4 Å). More detailed comparison of the CS-DP-Rosetta models to the manually refined NMR models showed that the former had a well-packed hydrophobic core, whereas the latter were much less well-packed. The overall arrangement of secondary structure elements is more similar to other members of the fold family in the Rosetta models than the NMR models, and given the well-packed core, it seems plausible that the Rosetta model is more accurate. We are currently investigating the possibility that the differences in the Rosetta structure and the manually refined NMR structure are due to the lack of NOEs between core side chains that could result from protein dynamics; this would disfavor close approach of core side chains in the manually refined models but have less impact on Rosetta's ability to determine the native structure once guided to the correct region of conformational space by the rest of the NOESY data.

As the largest protein in this study, the HR4394C blind prediction (Figure 2C) is particularly noteworthy. At the end of the first-stage sampling, CS-DP-Rosetta protocol clearly converged on the "correct" core of the protein, whereas CS-Rosetta models diverged significantly. Although the core had

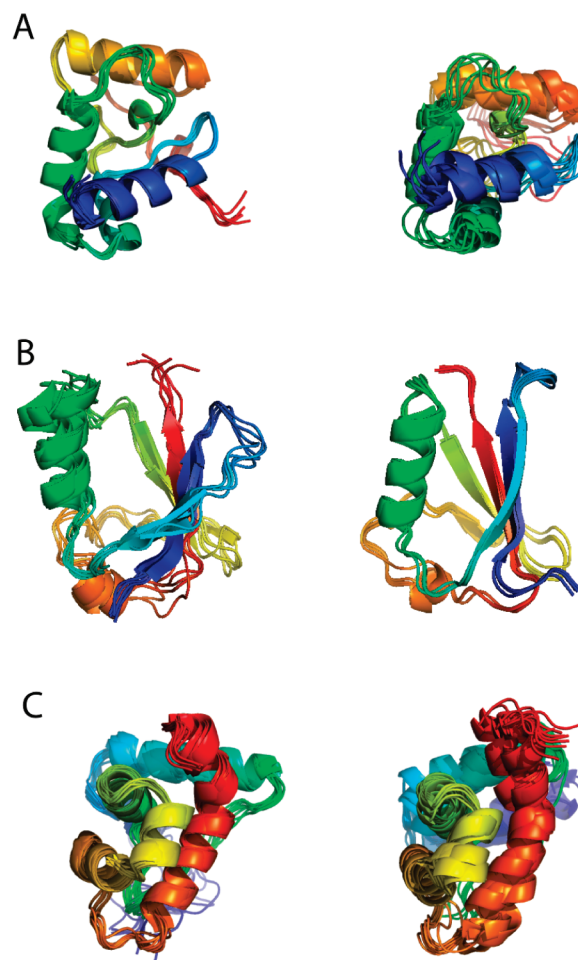


Figure 2. Blind structure determinations with CS-DP-Rosetta protocol: (A) VpR247, (B) AR3436A, (C) HR4394C. (Left) Experimentally solved NMR ensemble. (Right) Ensemble of lowest energy structures by the CS-DP-Rosetta protocol. Refined peak lists were used for VpR247 and AR3436A; raw peak lists were employed for HR4394C.

converged, the per-residue deviation analysis showed significant variations in the terminal helices at either end. Preferential sampling of the termini of models identified using the DP-score

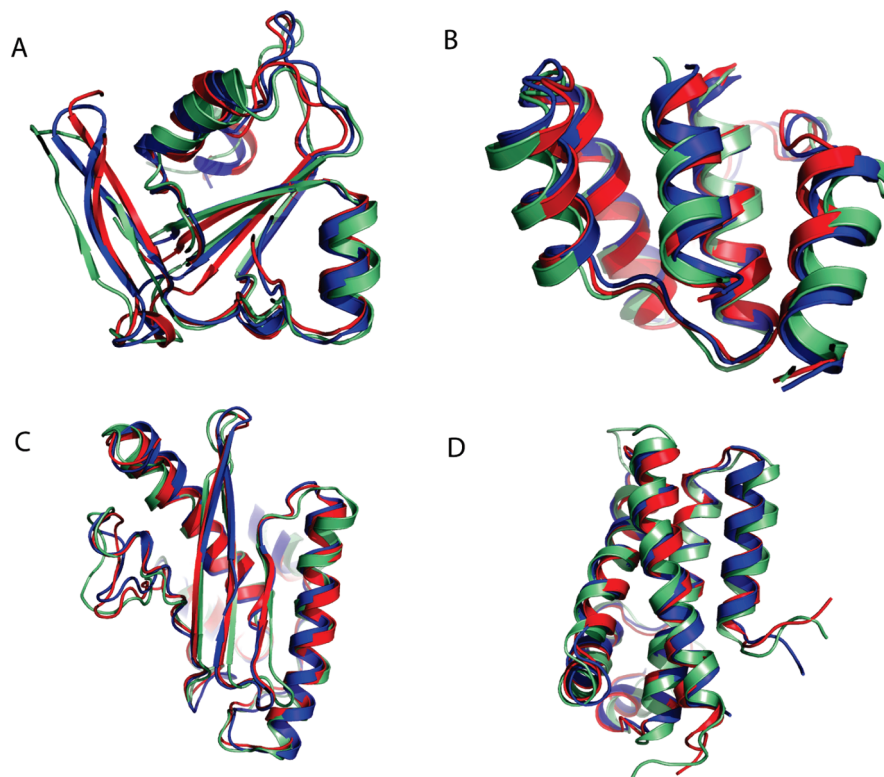


Figure 3. Superposition of the AssignNOE-Rosetta model (red) with the starting model generated by CYANA/AutoStructure using refined peak lists (light green) and the X-ray structure (dark blue): (A) CcR55, (B) StR65 (flexible loop residues 14–22 not shown), (C) HR41, (D) SsR10.

in the second stage generated a tighter ensemble with lower Rosetta all-atom energy, better DP-score, and in good agreement with the native structure (with an average rmsd of 2.3 Å to the first structure of the native NMR ensemble).

Test Cases with AssignNOE-Rosetta Protocol. We tested the AssignNOE-Rosetta protocol on five proteins ranging in size from 100 to 160 residues for which a high-resolution X-ray structure was available. For these structures, models generated by fully automated analysis of the refined NOESY peak list data with CYANA or AutoStructure were generally 2–3 Å rmsd from the native structure (determined following careful manual refinement of the NOESY peak list data). Although these structures can be refined even further by expert interactive analysis of the NOESY peak list data, this is a time-consuming and expertise-dependent process.

Starting from these refined NOESY peak list data, the AssignNOE-Rosetta protocol generated models with close to native side chain packing and ~ 1 Å backbone rmsd from the X-ray structure (see Figure 3). As shown in columns 2 and 3 of Table 1, section B, the Rosetta-refined models have lower rmsd to the X-ray structure over the full length and the core residues (as identified by FindCore¹¹) compared to the starting CYANA/AutoStructure model (Table S5 in Supporting Information for DP-score, Recall and Precision measures using refined peak lists, Table S6 in Supporting Information for inter-ensemble rmsd). Interestingly, the Rosetta-refined model was closer to the X-ray structure than the PDB-deposited manually refined NMR structure in all five cases, which is consistent with our previous findings⁶ (see Table 1, section B, columns 1 and 3). This suggests that refinement to the global energy minimum can consistently improve the accuracy of close to native structures generated by fully automated NOESY assignment programs, avoiding the need for tedious final stage manual refinement. We also note that surface loop regions, which could be inherently more dynamic in solution, have tighter convergence

in the AssignNOE-Rosetta structures compared to the published NMR structures. However, the rmsd of a disordered region in an ensemble of structures depends on multiple factors including the fraction of the total number of conformers computed used to represent the ensemble.¹¹ Hence, without independent solution data (i.e., NMR relaxation data), it is difficult to meaningfully compare the rmsd of dynamic regions in protein structures obtained by NMR, X-ray, and CS-DP-Rosetta or AssignNOE-Rosetta structures.

Discussion

The two methods presented in this paper offer exciting alternatives to determining NMR structures that do not require manual or semimanual assignment of the NOESY spectrum. While Rosetta refinement of CYANA/AutoStructure structures using refined NOESY peak lists models provides much higher accuracy models than the CS-DP-Rosetta protocol, significant human effort goes into refining the NOESY peak lists to distinguish between noise and real peaks. The CS-DP-Rosetta protocol, in contrast, is fully automated and robust and does not require expertise in analysis of NOESY spectra, making it especially useful for a first pass determination of the structure and data prior to investing more effort in manual peak list refinement. In cases where refined peak lists are available, the Rosetta refinement of CYANA/AutoStructure models is particularly advantageous because the refinement is carried out using the accurate Rosetta all-atom force field without the bias of experimental restraints.

The DP filter is a powerful global fold score that can sometimes overcome the lack of convergence for larger proteins using CS-Rosetta alone.⁴ We expect the CS-DP-Rosetta protocol with raw peak lists to find wide applicability in the NMR community. Since the raw peak lists used in this study are automatically generated from the FIDs, minimal human intervention is required with this method. As the unassigned NOESY

data is used to filter models and not to drive conformation space sampling, it is relatively less insensitive to potential mis-assignments of NOESY cross peaks. This avoids the “garden path” problem, in which incorrectly assigned NOESY cross peaks subsequently rule-in other mis-assignments and drive the trajectory to an incorrect structure. The DP-score provides a global filter to eliminate non-native-like topologies, resulting in enrichment of native-like structures. This leads to enhanced sampling of conformation space close to the native structure in the subsequent rebuild-and-refine step. However, this method is constrained by the sampling that can be achieved by CS-Rosetta in the first step, as evidenced in the HR41 test case where the best CS-Rosetta models had ~ 5 Å rmsd, which is outside the radius-of-convergence of the Rosetta all-atom energy and the DP-score. In the case of HR41, the key N-terminal helix that is not accurately positioned by CS-DP-Rosetta protocol is connected to the rest of the protein by a long flexible loop. Although this N-terminal helix was poorly packed, the core of HR41 was predicted relatively well. Hence, this “failure” stems from both the size and complexity of HR41 fold. For larger proteins with complex nonlocal β sheet topologies, it may be possible to overcome this sampling limitation using the Rosetta broken chain folding protocol.¹²

The need for complete chemical shift assignment (backbone and side chain) to calculate the DP-score limits the applicability of this protocol to proteins under 150 amino acids, where side chain chemical shift assignment is relatively less time-consum-

ing. Accordingly, efforts are in progress to explore the use of CS-DP-Rosetta in cases where only backbone and limited side chain (e.g., methyl) resonance assignments are obtained. This approach could allow extension of the CS-DP-Rosetta protocol to larger proteins, including membrane proteins, which require perdeuteration in order to provide sufficient signal-to-noise.

Acknowledgment. We thank Rosetta@home participants and the DOE INCITE program for access to the Blue Gene/P super-computer at the Argonne National Laboratory. We thank Drs. O. Lange, Y. Wu, R. Mani, GVT Swapna, and Y. Tang for providing NMR data and for helpful discussions and comments on the manuscript. This work was supported in part by the National Institutes of Health grant GM76222 (to D.B) and National Institutes of General Medical Science Protein Structure Initiative Grant U54 GM074958 (to G.T.M).

Supporting Information Available: Description of sample preparation, data collection, and methods employed to generate the raw and refined NOESY peak lists for all the proteins used in this study. The NOESY data and corresponding lists are subdivided into specific spectral region and editing nuclei as ¹³C-aliphatic and aromatic and ¹⁵N edited spectra. The statistics for all peaklists are listed in Table S1. This material is available free of charge via the Internet at <http://pubs.acs.org>.

JA905934C