COMPARING TWO NMR STRUCTURE REFINEMENT METHODS

By

YISHA YAO

A thesis submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Biochemistry

Written under the direction of

Gaetano T. Montelione

And approved by

____

____

____

____

New Brunswick, New Jersey

October, 2017

ABSTRACT OF THE THESIS

Comparing Two NMR Structure Refinement Methods

by YISHA YAO

Thesis Director:

Gaetano T. Montelione

High throughput and automatic procedures for NMR structure determination are under intensive study in the current era of structural genomics. The major steps include data collection, data processing, resonance assignment with validation, derivation of structural restraints, generation of 20 conformers satisfying the structural restraints (the ensemble), and refinement of the conformers. The final step structure refinement typically refers to further energy minimization based on certain force fields. Refinement could improve the structure quality to a large extent due to the sparseness of NMR experimental measurements. A scientific and robust refinement methodology is desired as a vital part of the standard protocols of automatic NMR structure determination. In this study, we compare the performances of two refinement methods, CNS refinement and AMBER refinement. The core algorithm of CNS refinement is simulated annealing with gradient descent while AMBER uses molecular dynamics simulated annealing.

Eight protein targets are chosen randomly from the NESG depository and the two refinement methods tested on these targets. All the targets have chemical shifts and NOESY peak lists available, and 4 of them also have RDC data. Using the available NMR experiment data, initial coarse structures are generated by ASDP-CYANA. These coarse structures further go through CNS refinement and AMBER refinement. Then the CNS-refined and AMBER-refined structures are evaluated in terms of RMSD (reference to X-ray PDB structure) and DP score. We find that AMBER refinement achieves better results than CNS on 7 out of 8 targets—AMBER refined structures have smaller average RMSDs and higher ensemble-average DP scores. The differentiated performance of the two refinement methods could stem from the different algorithms and force fields implemented.

Acknowledgement

Thanks to Prof. Gaetano T. Montelione who kindly supported me for the completion of my master study and gave many constructive suggestions.

Thanks to Prof. David Case who devoted a lot of energy and time on this project.

Thanks to Janet Y. Huang who helps me overcome many technical difficulties.

Table of Contents

List of Tables

List of Illustrations

1 Introduction

1.1 General principles of NMR-based structure determination

Spin is an intrinsic property of the elementary particles. The particle behaves as though it were spinning. However, spin is a quantum mechanical phenomenon without any analogue in classical mechanics. Not all the nuclei have spin. When both the number of protons and the number of neutrons are even, the nucleus has no spin; If one of them is odd, the spin is half-integer; If both of them are odd, the spin is positive integer valued. Nuclei with positive spins are spin-active, have a magnetic moment, and thus can be studied by magnetic resonance techniques (Günther 2013). In the presence of an external magnetic field, the spin-active nuclei will splits into two energy levels. The nuclear state corresponding to the spin quantum number -1/2 has higher energy than the one corresponding to +1/2. The nuclei are populated between the two energy levels according to the Boltzmann distribution. The energy gap between the two levels corresponds to a transition frequency, namely the Larmor frequency in NMR. In Modern NMR techniques, a pulse sequence will first be applied to the protein sample in solution, and then the facilities monitor the re-equilibrium of the system. The signal is digitized at regularly spaced time points followed by Fourier transform. The resulting spectrum expresses the signal as a histogram of frequencies and their intensities. By recognizing the spectrum patterns, NMR is able to explore the structures, interactions, and motions of macromolecules.

Chemical shift is a critical measurement that can help determine the secondary structure (Sebastiani, Goward et al. 2002). Different structural geometries give rise to different local molecular environment and hence varied magnetic fields. H atoms in these specific geometric configurations would their respective signals. Following such logic, chemical shifts are informative of the local structural characteristics. Generally, the chemical shifts obtained in NMR experiments are with respect to the reference frequency, *i.e.* a quotient of the difference between the observed frequency and the reference frequency over the reference frequency. This value is independent of the external magnetic field. A commonly used reference compound is 2,2-methyl-2-silapentane-5-sulfonate (DSS) whose 1H NMR frequency is defined as 0 ppm. The electrons around the nuclei, including electron pairs and π electron current, cause the variation in the local magnetic field. These electrons form a weak magnetic field opposite to the main magnetic field and thus shield the nuclei from the main magnetic field to some extent. Besides, the spin states of adjacent nuclei will be affected by each other through intervening bonds. This phenomenon is named as spin-spin coupling, scalar coupling, or J-coupling. Due to the through-bond interactions, the signal of a set of equivalent nuclei is split into a multiplet whose pattern contains rich information about molecular configuration and conformation. In summary, chemical shift values are determined by electron densities, bond orbitals, spin-spin interactions, and some other physical-chemical factors, and therefore are characteristic of the local structures.

Correlation Spectroscopy (COSY) and Total Correlation Spectroscopy (TOCSY) are the center of homonuclear nuclear magnetic resonance. Both COSY and TOCSY produce a two-dimensional spectrum, where both the axes measure the hydrogen chemical shifts. Their working principle is that the magnetization transferred through chemical bonds generates a cross-peak. Yet there is a major difference. COSY is only able to transfer magnetization between protons on adjacent atoms while in TOCSY magnetization can be transferred through all the protons connected by chemical bonds. Hence TOCSY usually contains richer information than COSY. For unlabeled proteins, these two types of experiments are used to build the spin systems. A common drawback in COSY and TOCSY is peak overlap, especially in larger proteins. Therefore, homonuclear NMR is usually limited to small proteins (Kessler, Gehrke et al. 1988).

Nuclear Overhauser Effect (NOE) are the fundamentals of some advanced NMR techniques for resolving molecular tertiary structures, including Nuclear Overhauser Effect Spectroscopy (NOESY), Heteronulcear Overhauser Effect Spectroscopy (HOESY), transferred Nuclear Overhauser Effect (TRNOE), *etc*. The general procedure is as following: apply the radiation to the sample at the transition frequency of one type of protons; this saturates these particular protons; the saturated nuclei will affect proximate nuclei through dipole-dipole interaction and change the spin population of proximate nuclei from their equilibrium distribution; the intensities of the NMR peaks of nearby nuclei will change. There is something common between spin-spin coupling and NOE—they are both due to the interaction of two nuclei. But NOE is through space rather than through chemical bonds. The strength of an NOE cross-peak is approximately proportional to r^{-6}, so only close neighbors can give observable peaks. Since NOE is extremely sensitive to space distance, it can be used to determine the neighbors of a target nucleus and their quantitative distances to the nucleus. If an NOE cross-peak is observed between two protons that are far away in amino acid sequence, then the peptide must have folded in such a way that these two protons are close in space. Therefore, NOE can provide distance constraints for determining the macromolecule conformation. There have been multi-dimensional NOE spectroscopies, where each frequency domain corresponds to a type of nuclei that are being correlated and has a time delay in the pulse sequence. A cross peak indicates that the nuclear pair with the corresponding frequencies is interacting with each other. Multi-dimensional NMR spectroscopy spreads the one-dimensional spectrum and makes it simpler to exploit the wealthy spectrum information (Günther 2013).

Residual Dipolar Coupling (RDC) proves complementary information for exploiting the global folding of a biomolecule. It tells the orientation of the dipole-dipole interaction vectors (nucleus pairs) with respect to the common reference frame. When the molecules in solution are partially aligned, which results in an incomplete averaging of anisotropic magnetic interaction, the phenomenon of residual dipolar coupling can be observed. Partially oriented media was first introduced and explained in 1960s (Saupe and Englert 1963). After that, large progress has been made in the development of alignment methods, including biocelles made of dimyristoylphosphatidylcholine (DMPC), dihexanoylphosphatidylcholine (DHPC), filamentous phages, stretched or compressed polyacrylamide gel, and poly (ethylene glycol) / hexanol mixture (Chen and Tjandra 2012). The dipolar coupling between two nuclei depends on the distance between them, and the angle of bond relative to the external magnetic field, which can further tell the relative orientation information of parts of the molecule that are far apart in space.

From a systematic point of view, NMR structure determination will combine all the information derived from the experiments listed above. Generally, the first step is sequential assignment of the resonances, including linking each spin system to its corresponding amino acid residue (spin system identification) and assigning each spin system to the amino acid sequence (sequence specific assignment). In the stage of spin system identification, some amino acids have very unique COSY patterns and are easy to identify while other amino acids might be confusing and would need other NMR experiments (*e.g.* NOESY, TOCSY). The second stage—assigning each spin system to a particular amino acid residue in the sequence, relies on the through-space connectivity observed in NOESY spectra. For an isotope-labeled (usually15N-labeled) protein, the first experiment performed is 2D heteronuclear single quantum correlation spectrum (HSQC). It is very sensitive and regarded as the footprint of a protein. There are also some 3D extension of HSQC, such as TOCSY-HSQC and NOESY-HSQC, which resolve the overlapped peaks in a 2D 1H-1H spectrum. Triple resonance NMR spectroscopy emerged in 1990. The triple resonance assignment strategy is different from the sequential assignment method described above. In this experiment, proteins are labeled with 13C and 15N. Magnetization can be transferred over the peptide bonds and thus different spin systems are connected. Six types of spectroscopy are commonly used—HNCO, HN(CA)CO, HNCA, HN(CO)CA, HNCACB, CBCA(CO)NH. Taking HNCO as an example, it consists the 1H-15N plane expanded by the 13C dimension. Each 1H-15N plan contains the peaks originated from the proceeding carbonyl carbon. In this case, sequential assignment is by matching the chemical shifts of each residue and its processor’s carbons. The second step of NMR structure determination is to derive several categories of restraints from all the information collected, including torsion angle restraints, distance restraints, and orientation restraints. Torsion angle restraints are calculated from the chemical shifts and coupling constants. Each crosspeak in NOE spectrum stands for the spatial proximity of two nuclei. The distance is not precise and a distance range is used. Orientation restraints are derived from the RDC data, which measures the relative orientations of the bond vectors to the reference frame (Günther 2013). These restraints are input into a structure-calculation computer program and it generates structure models satisfying as much restraints as possible. See Table 1 for a summary of the above.

Experiment	**Labeling**	**Dimensionality**	**Utility**
COSY; TOCSY	None	2D	Identification of the spin systems
NOESY	None	2D	Sequential assignment
1H-15N HSQC	15N	2D	15N assignment; NH assignment
TOCSY-HSQC; NOESY-HSQC	15N	3D	Reducing the overlap in TOCSY and NOESY

HNCO;
HN(CA)CO;
HNCA;
HN(CO)CA;
HNCACB;
CBCA(CO)NH
	13C, 15N	3D	Resonance assignment

Table 1. Elements of NMR spectroscopy

With the progress in the development of NMR techniques, the size of accessible molecules keeps increasing. Isotope labeling and multidimensional spectra allow correlation and assignment of thousands of nuclei. Although NMR solved structures may not be as accurate as X-ray solved ones (especially for large proteins), but they would approximate the native conditions of protein and provide complementary information to X-ray solved ones.

1.2 Automatic NMR structure determination

Traditionally NMR structures determination is manually solved by an NMR expert, which is often laboratory-specific, or expert-specific, and cannot be reproduced from one laboratory to another. Moreover, the manual analysis requires tremendous expertise in NMR principles. Scientists are kept from NMR technique due to the large expertise barrier. Yet most part of spectrum analysis is relatively sample-independent. Actually, the whole process of NMR structure determination is possibly turned automated except a few adjustable parameters. For the past few decades, the community is devoted to realizing a fully automated procedure for spectrum interpretation and data analysis, whose success would eventually evolve the field of NMR structure determination. We are expecting that computers solve the structures with higher efficiency and robust.

To achieve the above goal, certain experiments that ease automation are to be created or modified. The use of multidimensional triple resonance NMR has become a routine in many biomolecular structural laboratories (Ikura, Kay et al. 1990). With the power of reduced-dimensionality (RD) (Szyperski, Wider et al. 1994) and G-matrix Fourier Transformation (GFT) (Kim and Szyperski 2004), triple resonance NMR gains the advantage of rapid data collection and generates spectra that are more amenable to automation. There are also hybrid approaches, which combine experimental and computational methods, such as nonlinear sampling with maximum entropy reconstruction (Schmieder, Stern et al. 1994), Hadamard techniques for selective multichannel excitation and selection (Kupce and Freeman 2003), spectral reconstruction from tilted planes (Kupce and Freeman 2004), *etc*. Even with the standardization of some data interpretation and structure validation rules, the process of NMR structure determination still retains some subjective aspects, which limits its scientific utility. To break this bottleneck, several algorithms or computer programs have been created to automate the NMR structure determination. These algorithms will be discussed later. Parallel computation hardware architecture also speeds up the calculation.

The general procedure for automated NMR structure determination is summarized in Figure 1, five steps—data collection, pre-processing, peak editing, resonance assignment, and structure calculation. A big challenge in building a reliable automation platform is how to generate complete and self-consistent data in each step. Validations have been added to the last three steps to improve the data quality (Zolnai, Lee et al. 2003). Self-inconsistency rises in data collection when mixing data from different NMR spectrometers or using different samples of the protein. It is known that each implementation of a sophisticated NMR experiment has a unique set of procedures and parameters. And different types of proteins require respective collection strategies. Therefore, efforts should be made to standardize the sets of NMR experiments (with certain parameters or conditions adjustable) for each type of protein samples. NMR data archiving remains to be a problematic issue in the application of modern NMR techniques. The conventional way of archiving is to store the data in the form of time domain free induction decay (FID), which is inefficient and error-prone. It is desired that appropriate database structure be used, which is simple to track and recover (Baran, Huang et al. 2004). With the emergence of world-widely public databases, such as The Protein Data Bank (PDB) and Biological Magnetic Resonance Bank (BMRB) (Seavey, Farr et al. 1991), it demands a standard set of deposit formats for the multiple data lists. Standardizing formats would allow public sharing of the data, make the NMR structural laboratories more productive, and make it possible to test novel computational algorithms. Currently a Self-defining Text Archival and Retrieval (NMR-STAR) format is used for depositing NMR solved structures. Coordinating with the worldwide PDB, the NMR community is developing a common format for encoding the various data lists generated by the available experimental techniques, the

NMR Exchange Format (NEF) (Gutmanas, Adams et al. 2015). This common format would promisingly make the output and input of multiple programs compatible. As the high-throughput omics projects started, NMR data grows almost exponentially. It brings the problem how to archive, organize, and process the data efficiently. Sesame (Zolnai, Lee et al. 2003) and SPINS (Baran, Moseley et al. 2002) systems are computing infrastructures designed for managing the high-volume NMR projects. Besides data storage, they can link all the steps in the process of NMR structure determination, from data collection to database depository.

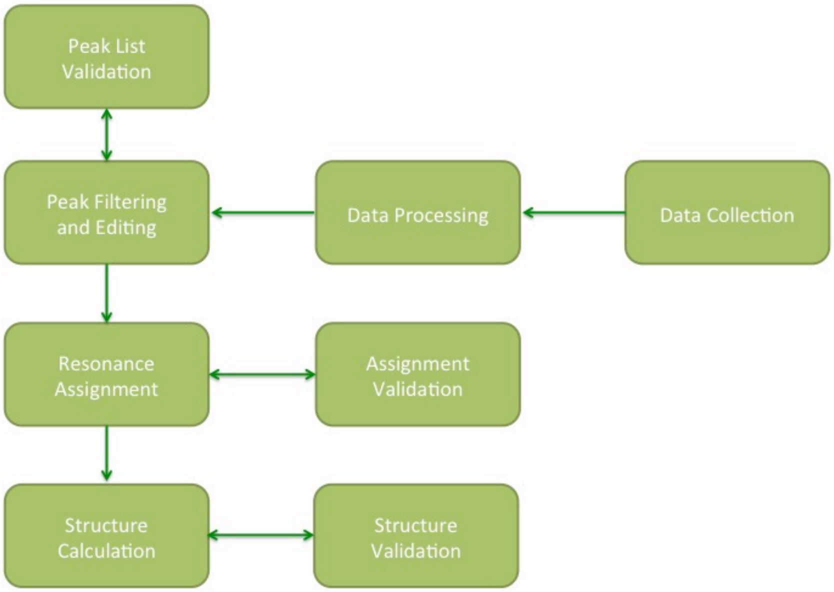


Figure 1. Flow chat of the procedure for NMR structure determination

1.3 Algorithms for automated NMR structure determination

To fully get rid of subjective aspects in the process of NMR structure determination, it is reasonable to employ computational algorithms for resonance assignment and structure calculation. Two categories of programs have been developed—for sequence-specific resonance assignment and for NOE assignment. The first category (automated sequential assignment) includes neural network based algorithm (Hare and Prestegard 1994), connectivity tracing assignment tools (CONTRAST) (Olson and Markley 1994), AUTOASSIGN (Baran, Huang et al. 2004), ALPS (Assignment for Labeled Protein Spectra) (Morelle, Brutscher et al. 1995), GARANT (General Algorithm for Resonance Assignment) (Bartels, Guntert et al. 1997), SAGA (Sequential assignment of GSs Algorithm) (Crippen, Rousaki et al. 2010), *etc*. Relatively fewer attempts have been made to automate NOE assignment (the second category). Indeed, many difficulties exist in NOE data interpretation, including noisy bands, artificial peaks, peak missing due to fast relaxation, and peak overlap. Only those protons that are spatially close due to covalent bonds or secondary structure can be assigned without ambiguity. With these inevitable issues, interactive programs are the mainstream tools. In early 90's, several computer programs were created to automate the NOE assignment problem. Most of these algorithms iteratively determine structures since a large portion of cross peaks cannot be assigned with certainty at the very beginning. The small portion of unambiguous NOEs that can be assigned initially are used to calculate the preliminary structure and other NOEs can be assigned based on the preliminary structure. The updated NOE assignment will be the input for the next iteration. Such cycles continue until no further update can be made on the NOE assignment. ASNO (Assign NOEs) implemented in XEASY program uses the explicitly assigned NOE cross peaks/inter-nuclear distances to calculate a preliminary structure. Subsequently, the preliminary structure is taken as a reference to eliminate the possible pairs of protons that violate the reference to a large extent (Guntert, Berndt et al. 1993). ARIA (Ambiguous Restraints for Iterative Assignment) incorporated spin diffusion correction in the iterative assignment of NOEs. It calculates NOE intensities based on the intermediate structure of current iteration, takes the theoretic NOE intensities as a correction factor, and calculates distance restraints for the next iteration (Linge, Habeck et al. 2003). It also integrated refinement module, refinement in explicit solvent using PARALLHDG 5.3 force field. However, such iterative approach can fail since it heavily depends on the correctness of the structure generated in previous cycle. If two protons are far apart in the preliminary structure, it will not be assigned even they are the true "answer" of the cross peak. Therefore, the iterative approach depends on how well the preliminary structure samples the conformation space. While ANSRS (Assignment of NOESY Spectra in Real Space) (Kraulis 1994) and another publication (Oshiro and Kuntz 1993) use a novel procedure—an inversion of the traditional strategy. First they get a three-dimensional real-space geometry model/conformation based on the NOE data, and then assign the sequential spectral by matching the measured frequency to the theoretical ones calculated from the three-dimensional real-space geometry model. Assignment ambiguity caused by chemical shift degeneracy and cross-peak overlap is not considered seriously in the above methods. The first method that addressed this issue works as following: ambiguity of a NOE cross peak is represented as the distances of all the proton pairs that may explain this cross peak. A new ambiguous distance restraint that allows all the possible assignments, is added to the energy minimization based on simulated annealing (Nilges 1995). NOAH treats ambiguous assignments as separate distance restraints, and iteratively calculates an ensemble of structures by distance geometry from unambiguous assignments and selected ambiguous assignments (Mumenthaler, Guntert et al. 1997). The wrong assignments are eliminated in subsequent cycles according to the principle of "self-consistency". Combined automated NOE assignment and structure determination module (CANDID) takes the similar iterative approach. Meanwhile it incorporates two new elements, network-anchoring and constraint-combination (Herrmann, Guntert et al. 2002). It is known that any possible set of assignments that may explain the collected NOE spectra forms a self-consistent set. Therefore the weights given to the multiple possible assignments of a cross peak are adjusted by the extent to which they can be embedded into the network formed by all other cross-peak assignments. One way to reduce the error induced by the artifact NOE upper distance constraints is to combine the assignments for two or several peaks into a single upper limit distance constraint. Network-anchoring and constraint-combination make the algorithm robust to high ambiguously NOE data. KNOWNOE uses a knowledge-driven Bayesian algorithm for dealing with the ambiguity in NOE data (Gronwald, Moussa et al. 2002). PASD (probabilistic assignment algorithm for automated structure determination) (Kuszewski, Schwieters et al. 2004) has three features that initial errors will not propagate through successive cycles—a linear item representing NOE restraint in energy function; treating all the possible assignments of a cross peak as independent; a probabilistic model to allow the elimination and re-entering of a certain NOE restraint during simulated annealing.

In this project, we use ASDP, an updated version of the AutoStructure program for NOE assignment (Huang, Tejero et al. 2006). Its core algorithm is a bottom-up topology-constrained network anchoring approach. Given a set of sequence-specific resonance assignments (each atom of the protein has been assigned a resonance frequency) R and the NOE peak lists, an ambiguous NOE network can be constructed by linking each NOE peak to one or more proton pairs. The distance between a pair of protons is decided based on the relationship that intensity of a peak is proportional to the inverse of the sixth power of inter-proton distance. To take into account of the experimental error, the algorithm allows certain tolerances for matching chemical shifts in the resonance assignment set R with those in the NOE peak lists. The true solution network is a subgraph of this ambiguous network. There are two major parts consisted in the algorithm, initial fold analysis and iteratively generating structures. First part is mostly preparation for iterative structure generation. First, the input data sets (chemical shifts, NOE peak lists, scalar coupling constant—optional, dihedral constraints—optional, RDC—optional) are preprocessed. Second, the algorithm will construct an initial ambiguous distance network derived from the sequence-specific resonance assignment set R and NOE peak lists, followed by validation of the input data sets using M score. Then it will build heuristic distance network starting from close proton pairs (within four covalent bonds). Based on the heuristic network, the set R of chemical shifts is refined. And reversely using the refined set R and NOE peak lists to prune the initial distance network results in a new ambiguous distance network. The heuristic distance network will be replenished by adding secondary-structure-specific proton pair contacts and well-assigned proton pair contacts gradually. With all these preparation, the initial structure models are generated based on the distance constraints, dihedral angle constraints, H-bond distance constraints, *etc*. The information extracted from the initial structure models are again used to refine the heuristic distance network. The second part consists of several iterative cycles. In each cycle the program generates structure models based on the heuristic distance network, and then refines the heuristic distance network using the intermediate structure models and topology constraints. The workflow of ASDP algorithm is shown in Figure 2.

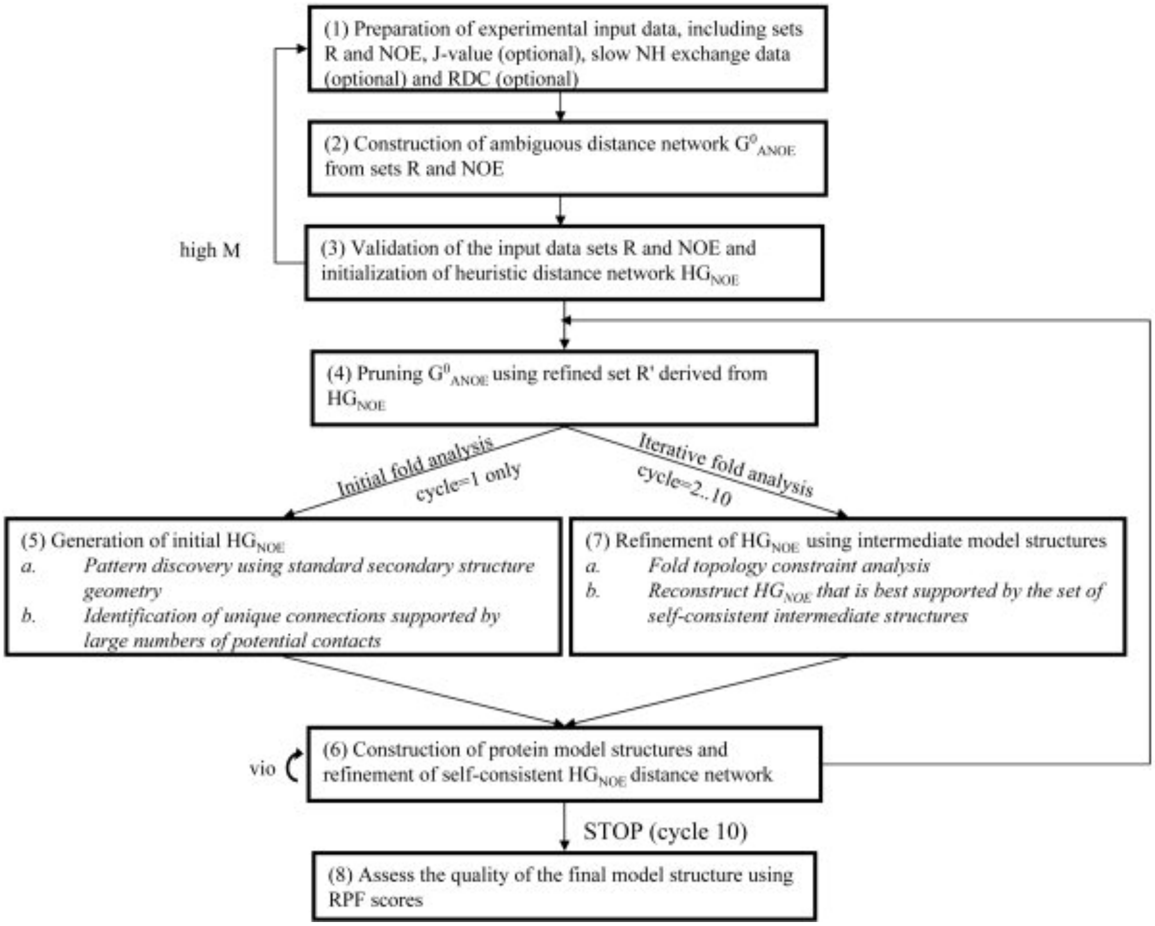

(1) Preparation of experimental input data, including sets R and NOE, J-value (optional), slow NH exchange data (optional) and RDC (optional)
(2) Construction of ambiguous distance network G^0_{ANOE} from sets R and NOE
high M
(3) Validation of the input data sets R and NOE and initialization of heuristic distance network HG_{NOE}
(4) Pruning G^0_{ANOE} using refined set R' derived from HG_{NOE}
Initial fold analysis
cycle=1 only
Iterative fold analysis
cycle=2..10
(5) Generation of initial HG_{NOE}
a. Pattern discovery using standard secondary structure geometry
b. Identification of unique connections supported by large numbers of potential contacts
(7) Refinement of HG_{NOE} using intermediate model structures
a. Fold topology constraint analysis
b. Reconstruct HG_{NOE} that is best supported by the set of self-consistent intermediate structures
vio
(6) Construction of protein model structures and refinement of self-consistent HG_{NOE} distance network
STOP (cycle 10)
(8) Assess the quality of the final model structure using RPF scores

Figure 2. Workflow of the algorithm implemented in AutoStructure

(Huang *et al*. 2006)

The first cycle (initial fold analysis) includes steps 1-6, and cycles 2-6 (the iterative fold analysis) include steps 4, 7, and 6. The initial ambiguous network in step 2 is reanalyzed in each iterative cycle.

1.4 General principles of NMR structure refinement

NMR experimental data gives sparse spatial constraints. They are not sufficient to completely determine the tertiary structure of a macromolecule. Additional information (e.g. force fields) is needed to generate a reasonably accurate model. Irrespective of the algorithm used, NMR refinement attempts to minimize the energy function that has many local minima (see Figure 3 for an example) (Bertini, Case et al. 2011). Usually the force field energy function includes three parts, covalent geometry (bonds, angles, planarity, and chirality), non-bonded interactions, and terms representing constraints derived from NMR experimental data. The covalent geometry term cannot introduce much variability since the values of bond lengths, bond angles, planes, and chirality are all accurately known. While there are many ways to represent the non-bonded interaction term, like a simple van de Waals repulsion, or Lennard-Jones potential, a considerable amount of variability would be introduced. However, the major determinant of structure accuracy resides in the number and quality of the constraints derived from experimental data, the last energy term discussed above (Clore and Gronenborn 1998). Common algorithms for NMR structure refinement include simulated annealing in Cartesian or torsion angle space, metric matrix distance geometry, and minimization of a defined energy function (Clore and Gronenborn 1998).

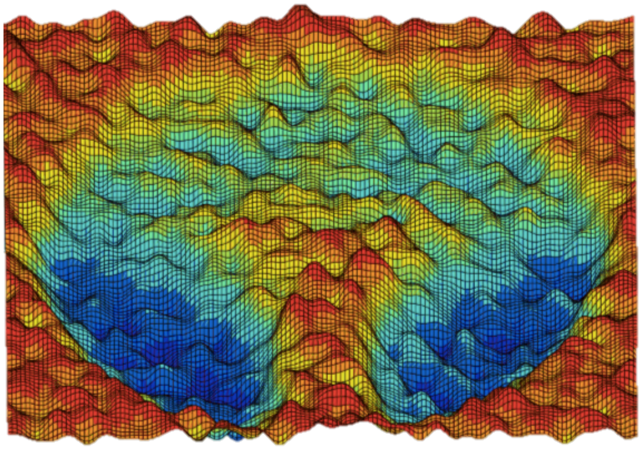


Figure 3. An example of a rough energy landscape

(Ren *et al.* 2015)

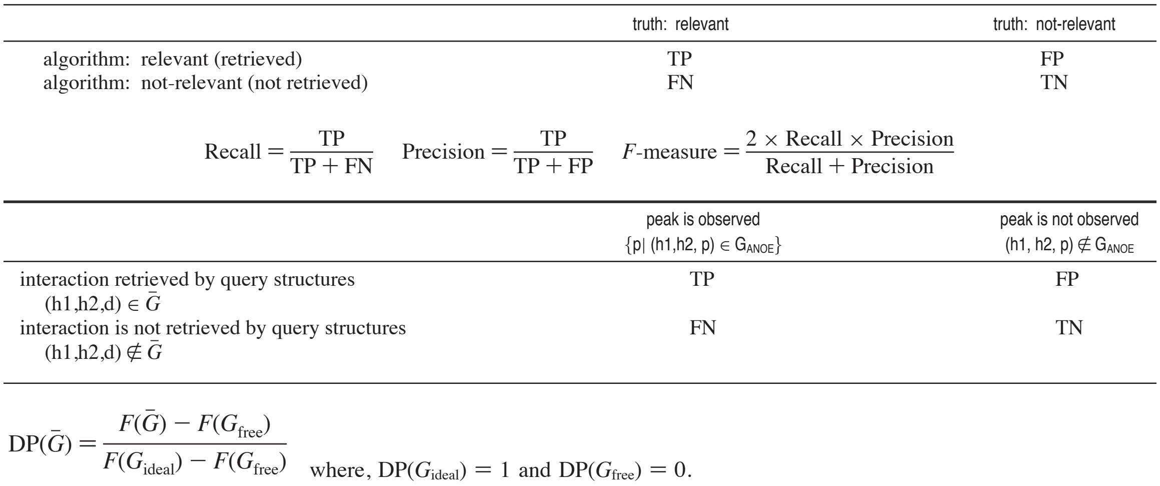
CNS (Crystallography and NMR system) is a software package designed for structure calculation, refinement, and modeling molecular dynamics. Scientists also use it as a computational framework to explore how to integrate the available information at multiple stages of structure determination (Brunger 2007). It has been a routine step in NMR structure determination procedure. In this project, we use CNS refinement protocol for energy minimization in water. The core algorithm in CNS water refinement is simulated annealing and gradient descent (Brunger, Adams et al. 1998). The detailed energy function is described as below: distance restraints are represented as harmonic functions, quadratic square-well functions, or quadratic asymptotic functions; torsion angle restraints are taken as harmonic or quadratic square-well functions; for RDC restraints, the axis system can be magnetic susceptibility or molecular alignment tensor, and the orientation of the axis system is allowed to float during simulated annealing; Lennard-Jones potential is adopted for the non-bonded interaction. One feature of CNS is direct refinement against NOE intensities via full relaxation matrix or quasi-relaxation matrix. The CNS water refinement protocol works as following: first, creating a topology file according to the target sequence and disulfide bonds; second, generating an extended starting model, which has good local stereo properties but no folding patterns; third, folding the extended model using distance geometry and generating five accepted structures which are consistent to the experimental derived constraints and also self-consistent; finally, the structures are further refined through simulated annealing.

Molecular Dynamics (MD) studies the biochemical systems at the atomistic levels and on timescale of milliseconds, complementing conventional experimental techniques. AMBER is widely used for the purpose of Molecular Dynamics simulation (Salomon-Ferrer, Case et al. 2013). Yet AMBER refers to more than just a molecular dynamics package. Besides a collection of modules that setup, conduct MD simulation and analyze the results, it also includes a series of classical molecular mechanics force fields (Ponder and Case 2003). The multiple modules possess their stand-alone utilities (*e.g. sqm* for semiempirical and DFTB quantum chemistry, *pbsa* for Poisson-Boltzmann modeling, *3D-RISM* for solvation integral equation modeling, *cpptraj/pytraj* for trajectory analysis) and can also be integrated for comprehensive analysis. The molecular dynamics simulation engine in AMBER consists of three parts, referred to as *sander*, *pmemd*, and *pmemd.cuda*. Among the three forks, *sander* is the most important for computation and originally development in the history of AMBER, whereas *pmend* and *pmemd.cuda* are designed to optimize the efficiency of MD simulation.

1.5 Assessment of the structure quality

With the rapid progress in the automation of NMR structure determination, it is critical to find a fast and sensitive way to evaluate the structure quality. A good structure quality measurement would help guide the automated structure determination process, and prevent deviation from the native structure. The community has been seeking for validation methods for NMR determined structure. Since NOESY peak lists are the prominent clues for structure determination, it is reasonable to evaluate the structure quality against NOE data. Conventional validation is to compare the back-calculated dihedral angles, or inter-proton distances with the corresponding constraints interpreted from experimental data (Doreleijers, Raves et al. 1999). Yet it has been realized that such comparison is biased since the constraints are manually derived (Nabuurs, Spronk et al. 2003). Analogous to the R-factor used in X-ray crystallography, the goodness-of-fit of the structure to the NOE data can be used as a quality assessment. Several programs take the strategy that compares the back-calculated NOESY peak list from the structure with the experimental NOESY peak list (Gronwald, Kirchhofer et al.

2000). However, cross-peak overlaps, spin diffusion, molecular tumbling, and other factors make it difficult to estimate NOESY peak intensities from three-dimensional structures, even with explicit relaxation matrix. Here we use DP score to measure the structure quality. DP score is derived from the RPF (Recall, Precision, F-measure) score, which relies on information retrieval (Huang, Powers et al. 2005). The key definitions are described as following: An ambiguous NOE network G_{Amb} is built based on the experimental NOESY peak lists and resonance assignment, where the vertices represent all the protons in this protein, and the edges linking vertices stand for all potential proton pairs that may explain the NOESY peak lists within a pre-defined match tolerance. Each cross peak in NOESY peak list corresponds to one, two, or multiple edges which may explain this cross peak. The correct solution network $G_{solution}$ corresponding to the true three-dimensional structure is a subgraph of G_{Amb}, assuming no artifacts in the NOESY peak lists. For a given query ensemble of structures, an ensemble-average distance network $G_{ensemble}$ is built. Then the differences between G_{Amb} and $G_{ensemble}$ is a measure of the goodness-of-fit between the structure and NOESY data. Assuming no artifacts in NOESY peak lists, the edges in both G_{Amb} and $G_{ensemble}$ are true positives (TPs); the edges in $G_{ensemble}$ but not in G_{Amb} are false positives (FPs); the edges not in G_{Amb} and not in $G_{ensemble}$ are true negatives (TNs); a peak will be assigned false negative (FN) if none of the potential proton pairs are in $G_{ensemble}$. Recall, Precision, and F-measure statistics are defined in Table 2.



	truth: relevant	truth: not-relevant
algorithm: relevant (retrieved)	TP	FP
algorithm: not-relevant (not retrieved)	FN	TN

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad F\text{-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

	peak is observed $\{p \mid (\text{h1,h2}, p) \in G_{\text{ANOE}}\}$	peak is not observed $(\text{h1, h2}, p) \notin G_{\text{ANOE}}$
interaction retrieved by query structures $(\text{h1,h2,d}) \in \bar{G}$	TP	FP
interaction is not retrieved by query structures $(\text{h1,h2,d}) \notin \bar{G}$	FN	TN

$$\text{DP}(\bar{G}) = \frac{F(\bar{G}) - F(G_{\text{free}})}{F(G_{\text{ideal}}) - F(G_{\text{free}})} \quad \text{where, } \text{DP}(G_{\text{ideal}}) = 1 \text{ and } \text{DP}(G_{\text{free}}) = 0.$$

Table 2. Definitions of Recall, Precision, F-measure statistics

(Huang, Powers et al. 2005)

We also calculated the average root mean squared deviation (RMSD) between the X-ray crystal structure and the NMR ensemble to measure the accuracy of NMR conformers. The wwPDB NMR-VTF has recommended that the “representative conformer of the ensemble” is the conformer that is most similar to all the other conformers, *i.e.* the medoid of the conformer distribution. The ill-defined regions that do not converge across the conformers are excluded from RMSD calculation. The FindCore2 algorithm of PDBStat is used to detect these ill-defined regions (Snyder, Grullon et al. 2014). In summary, we superimpose the X-ray crystal structure with the medoid, based on the superimposing calculate the backbone RMSD between the X-ray crystal structure and each conformer of the ensemble, and take the average RMSD.

1.6 Northeast Structural Genomics (NESG) consortium

Northeast Structure Genomics consortium is one of the four large-scale centers that conduct Protein Structure Initiative (PSI) project funded by NIH. The long-term goal of PSI is accurate prediction of the three-dimensional atomic-level structure of most proteins from their DNS sequences (Montelione, Zheng et al. 2000) To achieve this goal, the Structural Genomics centers are making efforts to build a paradigm for high-throughput structure determination. The accumulation of resolved structures will eventually increase our knowledge on the natural law of peptide chain folding with regard to its sequence. With the abundant data generated by Structural Genomics Consortia, we are able to conduct experiments on the data sets, *e.g.* testing algorithms, comparing methods. In this project, we tested two refinement methods, CNS refinement and AMBER refinement, on 8 NESG targets (Everett, Tejero et al. 2016). Each of these targets has NMR solved structure deposited in the BMRB database and X-Ray solved structure deposited in the PDB database. Automated calculated structures are going through CNS refinement and AMBER refinement respectively. The CNS-refined and AMBER-refined ensembles are evaluated based on their RMSD to the "golden standard"—X-Ray solved structure, and their DP scores that measure the goodness of the fit between the structure and the NOESY data (Bhattacharya, Tejero et al. 2007).

2 Material and Methods

2.1 Constraints preparation

We choose 8 targets (Table 3) from the NESG X-ray/NMR pair repository, ranging from 78 residues to 175 residues, and various folding patterns (α helix, β sheet, α helix + β sheet). For each target, the sequence, chemical shifts and NOE peak lists are extracted from the NMR-STAR v3.1 file deposited in BMRB, and transformed into CYANA format. Four of the eight targets also have RDC data available. The dihedral constraints are obtained by using the TALOSN program with the chemical shift data as input. TALOSN is a program that implements artificial neural network (ANN) to predict protein backbone and sidechain X1 torsion angles and secondary structure based on the chemical shifts of HN, Hα, Cα, Cβ, CO, N. There are two major methods used by TALOSN—search a high-resolution structural database for the 25 best matches to the secondary chemical shifts of a given residue triplet and averaging these best matches; cut the Ramachandran map into 324 pixels and use ANN to classify a given residue into one of the 324 pixels; combine the database mining result with the classification result (Shen and Bax 2013). Moreover, the TALOSN program outputs a measure of uncertainty in the prediction (the quality score) and the RCI-S^2 value (Berjanskii and Wishart 2005) for each residue. The quality score indicates how many out of the 25 best matches fall into a "consistent" region of the Ramachandran map. If the quality score is 25, then it is a "strong" prediction. The distance constraints are calculated by automated NOE assignment using the ASDP program. For the RDC data, we use the CYANA program to calculate the magnitude and rhombicity. Before structure calculation, the dihedral constraints and RDC constraints are filtered based on their S^2 score and quality score—the ones with S^2 < 0.75 or quality score < 25 are removed. The S^2 is a measure of order. The larger S^2 score is, the higher plausibility that this region has rigid structure and hence more confidence in the constraints.

NESG-ID	**Residue number**	**Availability of RDC data**	**Reso. of X-ray structure**
GmR137	78	RDC 1 medium	1.9 Å
CcR55	115	No RDC	1.8 Å
HR4694F	94	No RDC	1.99 Å
HR41	175	No RDC	2.54 Å
DhR29B	90	RDC 2 media	1.9 Å
HR4435B	83	No RDC	1.2 Å
PfR193A	127	RDC 1 medium	1.7 Å
CtR107	158	RDC 2 media	1.81 Å

Table 3. Targets summary

2.2 Structure calculation by ASDP-CYANA

The chemical shift lists, NOE peak lists, dihedral constraints, and RDC lists (when available) are input to the ASDP program. The workflow of ASDP is analogous to the routine analysis by an NMR expert. First, it will generate an ambiguous distance network by matching NOE peaks to the chemical shifts with certain tolerance and make ambiguous assignments. Then this ambiguous distance network will be used to evaluate the quality of the input data. After validating the input data sets, a heuristic distance network is gradually built from the sequential contacts, to the secondary structure specific contacts, then to the uniquely assigned contacts, and finally to the ambiguous assignments. The initial structure ensemble will be generated based on the distance constraints, dihedral constraints, H-bond distance constraints, *etc*. In the following iterative cycles, the program alternates between updating the heuristic distance network using the intermediate structure models and generating new structure models based on the refined distance constraints. The structure calculation takes 6 cycles. The DP scores of the final structure models will be reported for quality assessment. These structure models generated by ASDP-CYANA are further going through CNS and AMBER refinement.

2.3 CNS refinement

The conformers generated by ASDP-CYANA together with other information (the peptide sequence, chemical shift list, dihedral angle constraints, distance constraints, RDC constraints) are transformed into CNS format using the PDBStat program (Tejero, Snyder et al. 2013) and input into CNS water refinement. The commend used for CNS water refinement is WaterRefinement_cns/WaterRefCNS -na CNS -que pbs -par PARAM19 -tsc 0.001, where -na denotes the name of the target, -que indicates which que system is used, -par specifies the force field used for non-bonded interactions, and -tsc denotes the cooling time steps.

2.4 AMBER refinement

In this project, we use the generalized Born explicit solvent protocol. The conformers obtained from ASDP-CYANA and the restraints (the peptide sequence, chemical shift list, dihedral angle constraints, distance constraints, RDC constraints) are taken as input. It first carries out local minimization (small changes to the atom coordinates) to get rid of the bad contacts, and then conducts molecular dynamics simulated annealing to reach or approximate global energy minimum. At the start of the molecular dynamics simulated annealing, the system will be heated up (rise the temperature dramatically in a short time) so as to explore a wide range of the conformational space. Such regime would reduce the risk of getting trapped in local minima. Then it will cool down gradually and the molecules descend gently to the global minima (but not guaranteed of the global minimum).

2.5 Evaluation of the structure models

The quality of a structure is evaluated by two criteria—the RMSD between the obtained NMR structure and the corresponding X-ray PDB structure; the DP score of the obtained NMR structure, which measures the fit between the structure and the NOE data. For each target, the RMSDs as well as the DP scores of CNS-refined conformers and the corresponding AMBER-refined conformers are collected. Then the distributions of the RMSDs and DP scores are plotted. To rigorously differentiate the performance of CNS refinement and AMBER refinement, paired one-tail t-tests are applied to the RMSDs and DPs. Note, two types of DP scores are used in this work. The ensemble-average DP score is the DP score of the average structure of the ensemble. The average DP score of an ensemble is the average of the DP scores of the individual structures in the ensemble. The DP score of the X-ray PDB structure would measure any discrepancy between the solution and crystal conformation. We expect that the structures obtained by the automatic procedure of structure determination would achieve comparable or slightly less accuracy than the ones solved manually by NMR experts. We also expect that refinement would improve the structures output from ASDP-CYANA. Which refinement method works better, CNS or AMBER, is the main question we are going to address. Another question to explore is whether the DP score always highly correlates to the RMSD. Assuming there is no spurious or missing data in the NOE peak lists, the DP scores and corresponding RMSDs should be highly correlated.

3 Results

3.1 The structures obtained by the procedure of automatic NMR structure determination are reasonably accurate.

The average RMSDs of ASDP-CYANA model, CNS-refined model, and AMBER-refined model are around 2-3 Å for all the targets except CtR107 (Table 4-11 (a)). The ensemble-average and average DP scores of the ASDP-CYANA models, CNS-refined models, and AMBER-refined models are around 0.7-0.8 for all the targets except CtR107 (Table 4-11 (a)). These structures obtained by the fully automatic procedure are comparable to the NMR structures deposited in PDB, which are manually solved by NMR spectroscopy experts (Table 4-11 (a)). Figures 4-11 show that most regions in the 7 targets (except CtR107) converge and align well with the X-ray structure. It is reasonable that some coil and loop regions are divergent from the X-ray structure as they are flexible in solution.

3.2 AMBER refinement achieves better results than CNS.

For all the targets except PfR193A, AMBER refinement achieves smaller average RMSDs and higher average DP scores. Tables 4-11 (a) show that the average RMSDs of CNS-refined structures are larger than those of AMBER-refined structures. While the ensemble-average and average DP scores of CNS-refined structures are smaller than those of AMBER-refined structures. The paired one-tail t-tests confirm that

AMBER refinement achieves better results in 5 targets (Tables 4-11 (b)). Figure 12 (a) and (b) visualize the distributions of RMSDs and DP scores. Similar phenomena can be found: the RMSD distributions of AMBER-refined structures shift downward while the DP distributions of AMBER-refined structures shift upward compared to those of CNS-refined structures.

(a)

Residue number	78		
X-ray PDB	**DP**	0.784	
NMR PDB	**RMSD**	1.66 Å	
	DP	**Ensemble average**	0.881
		Average of the ensemble	0.762
ASDP-CYANA	**RMSD**	1.72 Å	
	DP	**Ensemble average**	0.873
		Average of the ensemble±SD	0.801 ± 0.0100
CNS-refined	**RMSD**	1.62 Å	
	DP	**Ensemble average**	0.835

		Average of the ensemble±SD	0.688 ± 0.0254
AMBER-refined	**RMSD**	1.16 Å	
	DP	**Ensemble average**	0.869
		Average of the ensemble±SD	0.775 ± 0.0170

(b)

	RMSD			DP		
	ASDP-CYANA	CNS	AMBER	ASDP-CYANA	CNS	AMBER
Model1	1.89	1.46	1.10	0.81	0.709	0.775
Model2	1.74	1.77	0.87	0.81	0.643	0.776
Model3	1.81	1.77	1.06	0.81	0.688	0.776
Model4	1.64	1.82	1.41	0.81	0.679	0.73
Model5	1.46	1.71	1.21	0.80	0.699	0.789
Model6	1.77	1.37	1.21	0.80	0.674	0.787
Model7	1.88	1.39	1.59	0.81	0.689	0.781
Model8	1.88	1.44	0.90	0.82	0.709	0.785
Model9	1.72	1.76	1.08	0.80	0.692	0.751
Model10	1.64	1.77	0.99	0.80	0.708	0.771
Model11	1.58	1.71	1.11	0.79	0.662	0.758

Model12	1.54	1.58	1.21	0.79	0.674	0.743
Model13	1.69	1.99	1.16	0.79	0.621	0.768
Model14	1.77	1.82	1.06	0.79	0.636	0.784
Model15	1.67	1.46	1.11	0.79	0.713	0.781
Model16	1.95	1.48	1.19	0.79	0.669	0.776
Model17	1.67	1.74	1.28	0.78	0.667	0.764
Model18	1.76	1.32	1.22	0.79	0.692	0.78
Model19	1.72	1.47	1.38	0.79	0.699	0.763
Model20	1.71	1.60	1.11	0.80	0.669	0.743
Paired one-tail t-test 3.13E-07				Paired one-tail t-test 7.76E-12		

Table 4. Structure evaluation for target GmR137

(a) Summary of the ensemble-average and average RMSDs and DP scores

(b) Paired one-tail t-test to compare the CNS-refined and AMBER-refined conformers

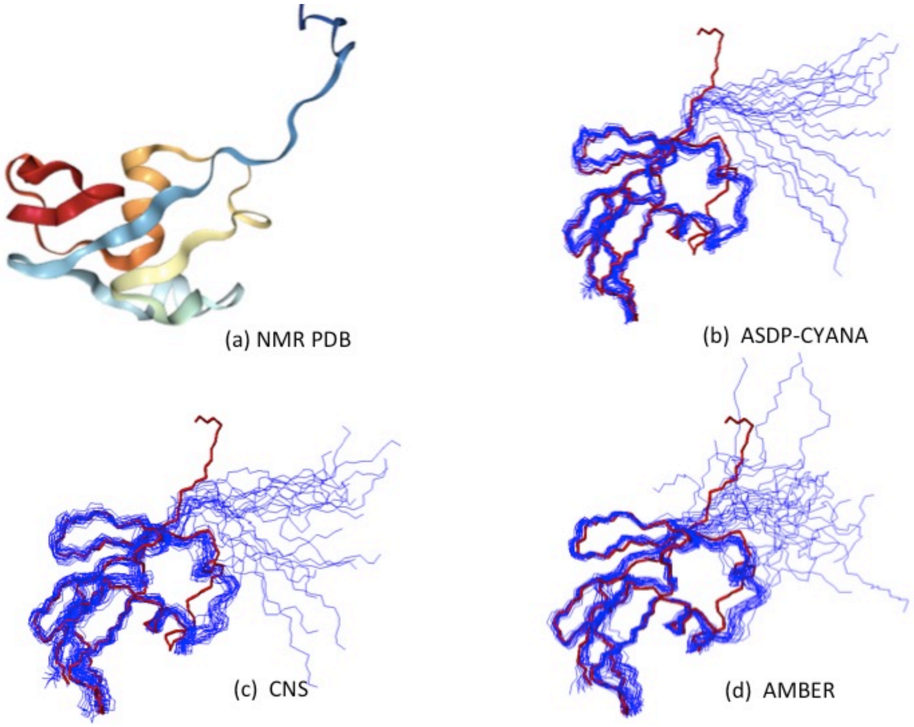


Figure 4. Visualization of the results for target GmR137

(a) The NMR structure deposited in PDB

(b) The NMR structure ensemble output from ASDP-CYANA (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(c) The CNS-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(d) The AMBER-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(a)

Residue number	115		
X-ray PDB	**DP**	0.699	
NMR PDB	**RMSD**	1.45 Å	
	DP	**Ensemble average**	0.791
		Average of the ensemble	0.641
ASDP-CYANA	**RMSD**	2.23 Å	
	DP	**Ensemble average**	0.770
		Average of the ensemble±SD	0.670 ± 0.0105
CNS-refined	**RMSD**	1.97 Å	
	DP	**Ensemble average**	0.779
		Average of the ensemble±SD	0.609 ± 0.0181
AMBER-refined	**RMSD**	1.51 Å	
	DP	**Ensemble average**	0.786
		Average of the ensemble±SD	0.667 ± 0.0132

(b)

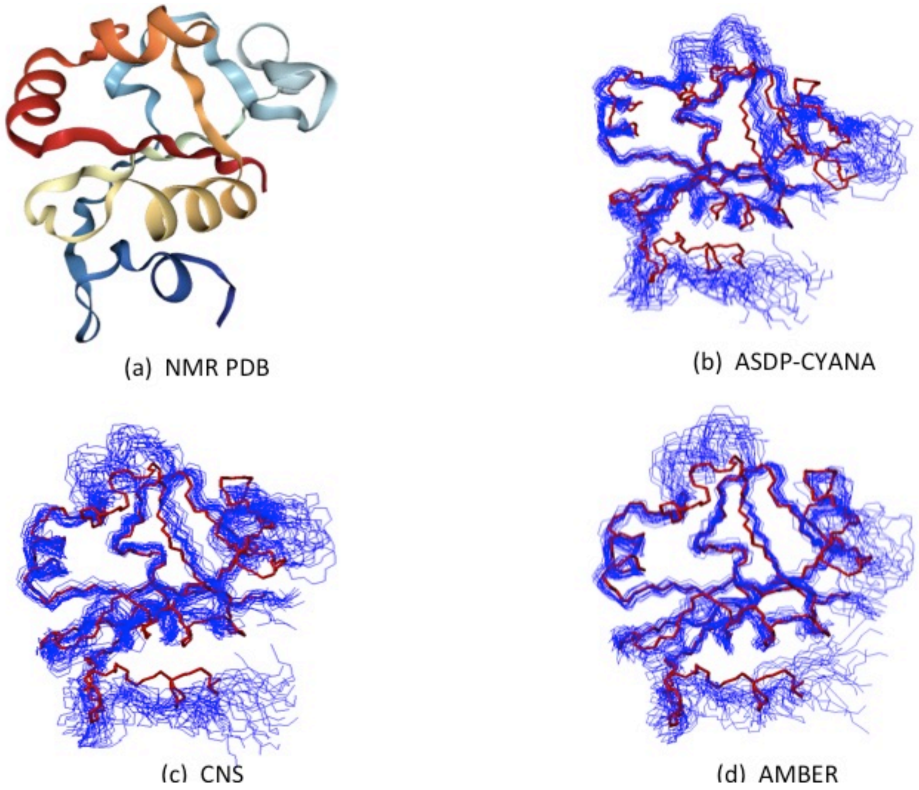
	RMSD			DP		
	ASDP-CYANA	CNS	AMBER	ASDP-CYANA	CNS	AMBER
Model1	1.95	1.66	1.705	0.6770	0.6	0.632
Model2	2.54	2.045	1.696	0.681	0.634	0.672
Model3	1.92	1.663	1.759	0.665	0.613	0.675
Model4	2.11	2.069	1.299	0.668	0.585	0.677
Model5	2.16	2.013	1.599	0.671	0.613	0.658
Model6	2.10	1.826	1.643	0.671	0.594	0.65
Model7	2.34	2.029	1.469	0.658	0.609	0.66
Model8	2.55	2.276	1.725	0.66	0.568	0.643
Model9	2.66	2.336	1.565	0.658	0.609	0.64
Model10	2.25	2.003	1.333	0.662	0.607	0.675
Model11	2.16	1.884	1.444	0.675	0.579	0.671
Model12	2.11	1.881	1.163	0.654	0.579	0.668
Model13	2.04	1.904	1.303	0.668	0.594	0.664
Model14	2.20	1.918	1.479	0.684	0.591	0.662
Model15	2.39	2.358	1.611	0.66	0.643	0.668
Model16	2.28	2.07	1.242	0.654	0.583	0.667
Model17	2.34	1.819	1.68	0.647	0.594	0.65

Model18	2.18	1.849	1.468	0.664	0.6	0.675
Model19	2.31	2.065	1.598	0.655	0.604	0.645
Model20	2.07	1.693	1.501	0.647	0.604	0.66
Paired one-tail t-test 2.20E-07				Paired one-tail t-test 3.43E-11		

Table 5. Structure evaluation for target CcR55

(a) Summary of the ensemble-average and average RMSDs and DP scores

(b) Paired one-tail t-test to compare the CNS-refined and AMBER-refined conformers


(a) NMR PDB

(b) ASDP-CYANA

(c) CNS

(d) AMBER

Figure 5. Visualization of the results for target CcR55

(a) The NMR structure deposited in PDB

(b) The NMR structure ensemble output from ASDP-CYANA (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(c) The CNS-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(d) The AMBER-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(a)

Residue number	96		
X-ray PDB	**DP**	0.565	
NMR PDB	**RMSD**	1.02 Å	
	DP	**Ensemble average**	0.806
		Average of the ensemble	0.743
ASDP-CYANA	**RMSD**	2.93 Å	
	DP	**Ensemble average**	0.797

		Average of the ensemble±SD	0.701 ± 0.0210
CNS-refined	**RMSD**	2.76 Å	
	DP	**Ensemble average**	0.788
		Average of the ensemble±SD	0.658 ± 0.0212
AMBER-refined	**RMSD**	1.78 Å	
	DP	**Ensemble average**	0.788
		Average of the ensemble±SD	0.705 ± 0.0180

(b)

	RMSD			DP		
	ASDP-CYANA	CNS	AMBER	ASDP-CYANA	CNS	AMBER
Model1	2.63	2.50	1.63	0.73	0.68	0.71
Model2	2.69	2.54	2.48	0.73	0.68	0.69
Model3	2.56	2.03	2.16	0.70	0.66	0.70
Model4	2.40	2.20	1.31	0.70	0.65	0.71
Model5	2.35	2.24	1.58	0.72	0.68	0.69
Model6	2.98	2.82	1.75	0.69	0.65	0.67
Model7	4.21	3.77	1.39	0.70	0.68	0.74

Model8	3.28	2.79	1.64	0.70	0.60	0.70
Model9	2.19	2.32	1.88	0.70	0.64	0.69
Model10	2.78	2.90	1.61	0.70	0.65	0.74
Model11	2.99	2.64	1.98	0.69	0.65	0.68
Model12	2.48	2.17	1.51	0.68	0.64	0.70
Model13	2.68	2.52	1.58	0.70	0.63	0.71
Model14	3.32	3.12	3.65	0.69	0.65	0.68
Model15	3.65	3.30	1.74	0.68	0.64	0.68
Model16	3.26	3.11	1.43	0.69	0.65	0.71
Model17	2.17	2.25	1.55	0.69	0.67	0.72
Model18	2.45	2.42	2.03	0.69	0.65	0.69
Model19	3.73	3.54	1.34	0.67	0.67	0.70
Model20	3.83	4.04	1.37	0.68	0.64	0.71
Paired one-tail t-test 1.73E-05				Paired one-tail t-test 3.40E-08		

Table 6. Structure evaluation for target HR4694F

(a) Summary of the ensemble-average and average RMSDs and DP scores

(b) Paired one-tail t-test to compare the CNS-refined and AMBER-refined conformers

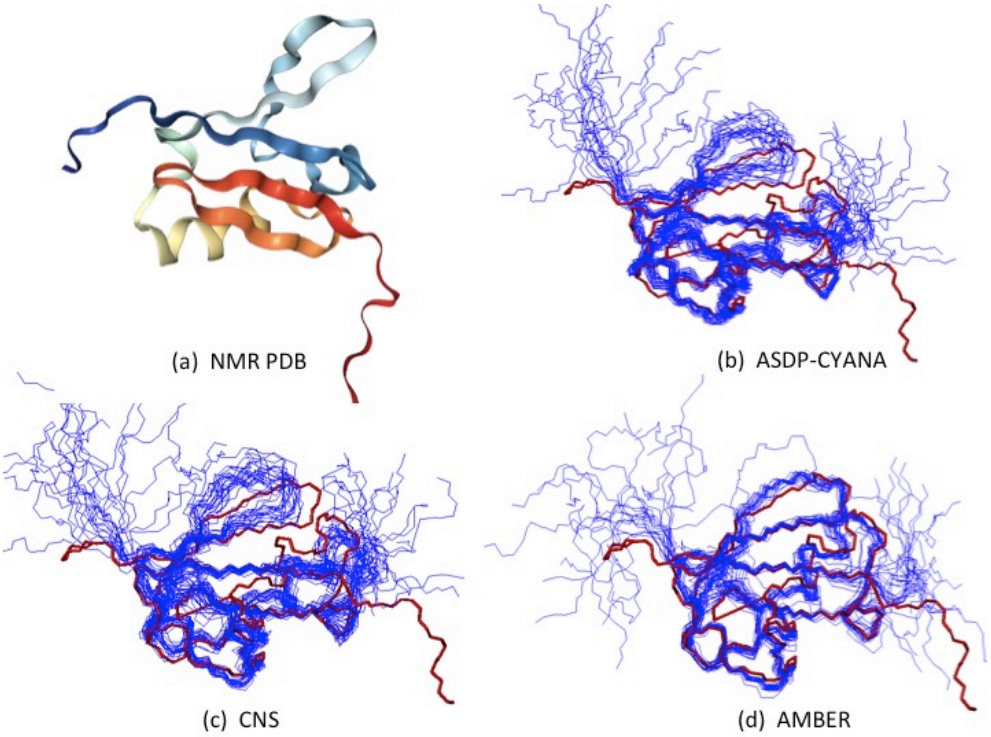


Figure 6. Visualization of the results for target HR4694F

(a) The NMR structure deposited in PDB

(b) The NMR structure ensemble output from ASDP-CYANA (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(c) The CNS-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(d) The AMBER-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(a)

Residue number	175		
X-ray PDB	**DP**	0.763	
NMR PDB	**RMSD**	1.52 Å	
	DP	**Ensemble average**	0.852
		Average of the ensemble	0.768
ASDP-CYANA	**RMSD**	2.46 Å	
	DP	**Ensemble average**	0.818
		Average of the ensemble±SD	0.741 ± 0.0042
CNS-refined	**RMSD**	2.22 Å	
	DP	**Ensemble average**	0.754
		Average of the ensemble±SD	0.651 ± 0.0114
AMBER-refined	**RMSD**	1.79 Å	
	DP	**Ensemble average**	0.83
		Average of the ensemble±SD	0.761 ± 0.00685

(b)

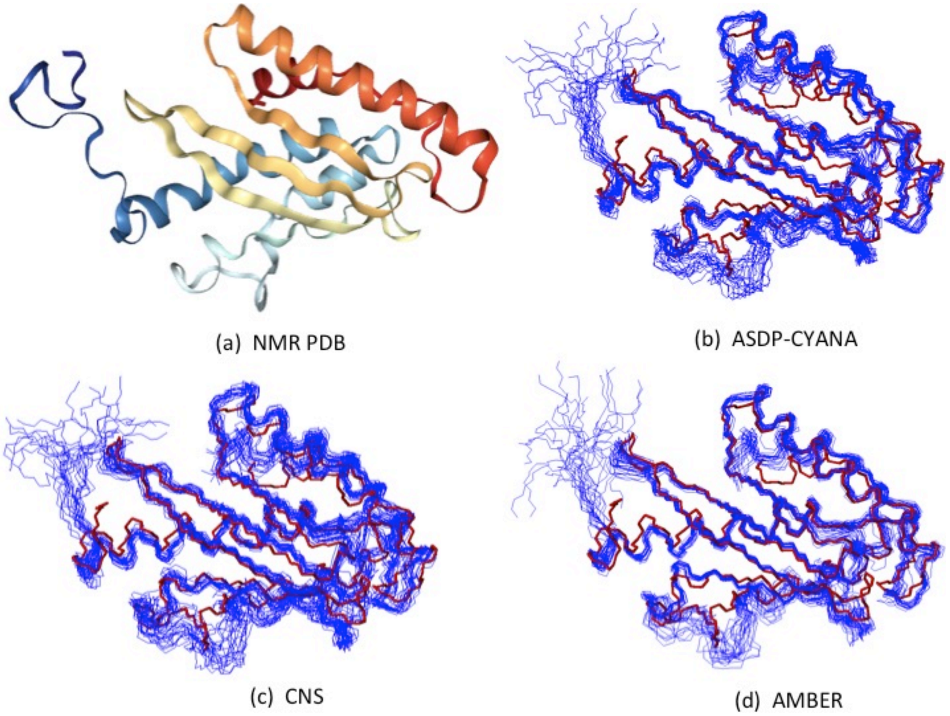
	RMSD			DP		
	ASDP-CYANA	CNS	AMBER	ASDP-CYANA	CNS	AMBER
Model1	2.07	2.36	1.76	0.74	0.66	0.76
Model2	2.10	2.33	1.52	0.74	0.65	0.76
Model3	2.09	2.16	1.95	0.74	0.64	0.76
Model4	2.14	2.49	1.71	0.74	0.65	0.77
Model5	2.18	2.29	2.08	0.74	0.66	0.76
Model6	2.14	2.29	1.49	0.74	0.65	0.77
Model7	1.95	2.28	1.56	0.74	0.63	0.76
Model8	2.15	1.88	1.84	0.74	0.63	0.76
Model9	2.07	2.21	2.12	0.74	0.65	0.75
Model10	2.21	2.12	1.97	0.74	0.67	0.76
Model11	2.10	2.23	1.79	0.74	0.64	0.75
Model12	2.08	2.28	1.89	0.74	0.64	0.76
Model13	2.11	2.11	1.62	0.74	0.65	0.74
Model14	2.06	2.13	1.98	0.73	0.63	0.76
Model15	2.09	2.12	1.59	0.74	0.66	0.77
Model16	2.14	2.31	1.70	0.73	0.63	0.76

Model17	2.15	2.33	1.93	0.73	0.64	0.76
Model18	2.10	2.17	1.92	0.74	0.65	0.76
Model19	2.15	2.08	1.72	0.73	0.65	0.77
Model20	2.13	2.25	1.57	0.73	0.65	0.77
Paired one-tail t-test 1.27E-07				Paired one-tail t-test 2.21E-20		

Table 7. Structure evaluation for target HR41

(a) Summary of the ensemble-average and average RMSDs and DP scores

(b) Paired one-tail t-test to compare the CNS-refined and AMBER-refined conformers


(a) NMR PDB

(b) ASDP-CYANA

(c) CNS

(d) AMBER

Figure 7. Visualization of the results for target HR41

(a) The NMR structure deposited in PDB

(b) The NMR structure ensemble output from ASDP-CYANA (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(c) The CNS-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(d) The AMBER-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(a)

Residue number	90		
X-ray PDB	**DP**	0.676	
NMR PDB	**RMSD**	1.54 Å	
	DP	**Ensemble average**	0.798
		Average of the ensemble	0.692
ASDP-CYANA	**RMSD**	3.57 Å	

	DP	**Ensemble average**	0.759
		Average of the ensemble±SD	0.657 ± 0.0161
CNS-refined	**RMSD**	3.27 Å	
	DP	**Ensemble average**	0.778
		Average of the ensemble±SD	0.602 ± 0.0302
AMBER-refined	**RMSD**	2.49 Å	
	DP	**Ensemble average**	0.800
		Average of the ensemble±SD	0.689 ± 0.02

(b)

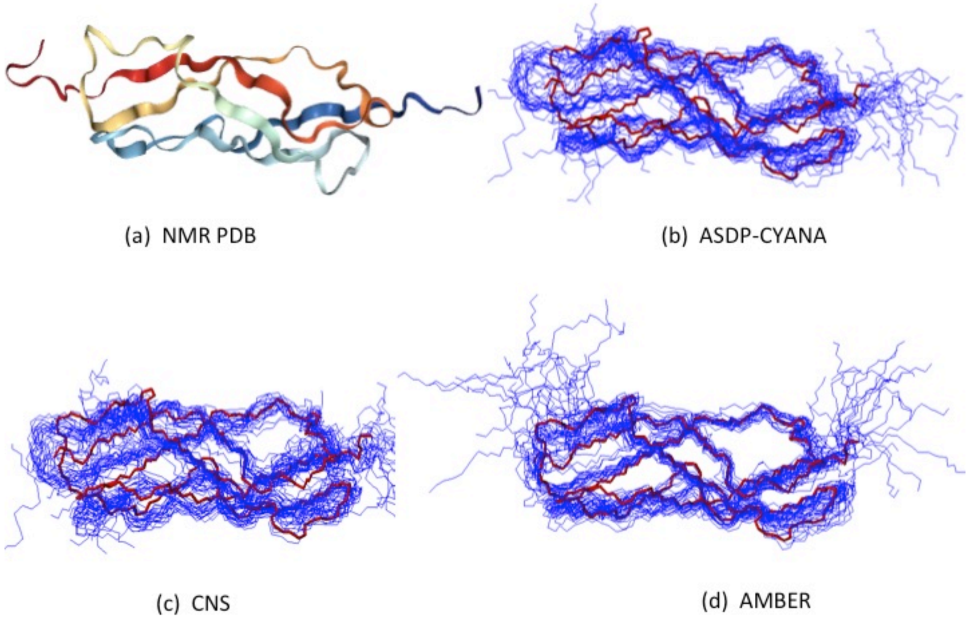
	RMSD			DP		
	ASDP-CYANA	CNS	AMBER	ASDP-CYANA	CNS	AMBER
Model1	2.67	2.79	3.43	0.69	0.57	0.68
Model2	2.82	2.69	1.83	0.67	0.60	0.72
Model3	3.40	3.12	1.86	0.67	0.60	0.69
Model4	4.02	3.27	1.52	0.66	0.58	0.70
Model5	3.00	2.65	3.27	0.66	0.61	0.68

Model6	3.38	3.28	3.62	0.66	0.63	0.65
Model7	3.90	3.77	1.99	0.66	0.60	0.71
Model8	2.87	2.97	1.91	0.66	0.65	0.71
Model9	4.00	3.59	3.65	0.66	0.62	0.66
Model10	4.66	4.38	2.67	0.65	0.60	0.67
Model11	3.32	3.17	2.28	0.65	0.57	0.66
Model12	3.10	2.77	3.32	0.64	0.57	0.70
Model13	3.31	2.85	3.57	0.64	0.58	0.67
Model14	3.69	3.24	2.18	0.64	0.56	0.66
Model15	4.70	4.18	2.31	0.64	0.63	0.69
Model16	4.67	4.31	1.48	0.63	0.52	0.69
Model17	3.00	2.74	3.13	0.63	0.57	0.68
Model18	3.05	2.85	1.54	0.63	0.59	0.69
Model19	3.75	3.06	1.86	0.64	0.60	0.68
Model20	4.09	3.71	2.38	0.63	0.63	0.68
Paired one-tail t-test 1.72E-03				Paired one-tail t-test 5.82E-10		

Table 8. Structure evaluation for target DhR29B

(a) Summary of the ensemble-average and average RMSDs and DP scores

(b) Paired one-tail t-test to compare the CNS-refined and AMBER-refined conformers



(a) NMR PDB

(b) ASDP-CYANA

(c) CNS

(d) AMBER

Figure 8. Visualization of the results for target DhR29B

(a) The NMR structure deposited in PDB

(b) The NMR structure ensemble output from ASDP-CYANA (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(c) The CNS-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(d) The AMBER-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(a)

Residue number	83		
X-ray PDB	**DP**	0.654	
NMR PDB	**RMSD**	2.11 Å	
	DP	**Ensemble average**	0.773
		Average of the ensemble	0.660
ASDP-CYANA	**RMSD**	3.58 Å	
	DP	**Ensemble average**	0.788
		Average of the ensemble±SD	0.724 ± 0.0069
CNS-refined	**RMSD**	3.47 Å	
	DP	**Ensemble average**	0.781
		Average of the ensemble±SD	0.661 ± 0.0222
AMBER-refined	**RMSD**	3.45 Å	
	DP	**Ensemble average**	0.793
		Average of the ensemble±SD	0.687 ± 0.013

(b)

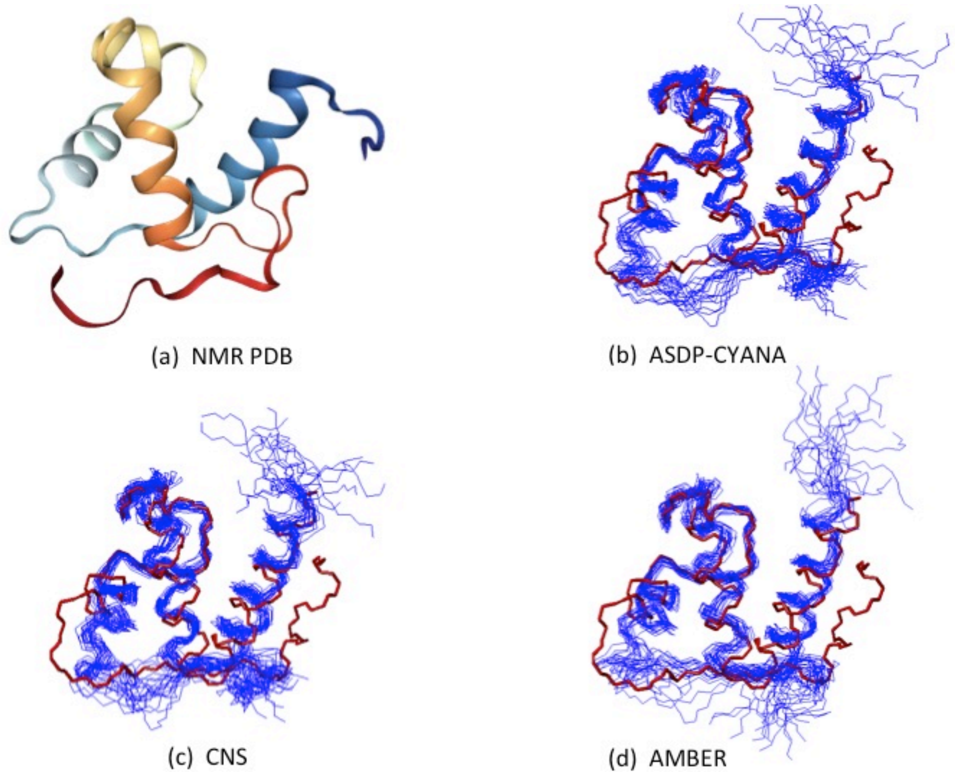
	RMSD			DP		
	ASDP-CYANA	CNS	AMBER	ASDP-CYANA	CNS	AMBER
Model1	3.19	3.36	3.24	0.73	0.700	0.71
Model2	3.62	3.69	3.16	0.73	0.66	0.69
Model3	3.56	3.57	3.92	0.72	0.65	0.68
Model4	3.52	3.43	3.28	0.74	0.63	0.68
Model5	3.58	3.44	3.78	0.73	0.66	0.69
Model6	3.43	3.42	3.04	0.72	0.63	0.66
Model7	3.37	3.35	3.04	0.73	0.68	0.66
Model8	3.70	3.21	3.27	0.72	0.68	0.68
Model9	3.71	3.53	3.22	0.72	0.61	0.68
Model10	3.66	3.49	3.65	0.71	0.66	0.69
Model11	3.47	3.38	4.24	0.72	0.63	0.66
Model12	3.61	3.57	2.98	0.72	0.62	0.70
Model13	3.67	3.43	3.79	0.72	0.66	0.68
Model14	3.73	3.69	3.4	0.72	0.64	0.66
Model15	3.68	3.69	3.65	0.71	0.64	0.67
Model16	3.64	3.57	4.01	0.72	0.66	0.68
Model17	3.68	3.58	3.36	0.71	0.68	0.70

Model18	3.70	3.50	3.85	0.73	0.67	0.69
Model19	3.54	3.31	4.20	0.71	0.66	0.68
Model20	3.51	3.26	3.54	0.71	0.65	0.67
Paired one-tail t-test 0.27				Paired one-tail t-test 8.73E-06		

Table 9. Structure evaluation for target HR4435B

(a) Summary of the ensemble-average and average RMSDs and DP scores

(b) Paired one-tail t-test to compare the CNS-refined and AMBER-refined conformers



(a) NMR PDB

(b) ASDP-CYANA

(c) CNS

(d) AMBER

Figure 9. Visualization of the results for target HR4435B

(a) The NMR structure deposited in PDB

(b) The NMR structure ensemble output from ASDP-CYANA (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(c) The CNS-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(d) The AMBER-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(a)

Residue number	127		
X-ray PDB	**DP**	0.808	
NMR PDB	**RMSD**	0.85 Å	
	DP	**Ensemble average**	0.875
		Average of the ensemble	0.854
ASDP-CYANA	**RMSD**	1.74 Å	
	DP	**Ensemble average**	0.854

		Average of the ensemble±SD	0.787 ± 0.0133
CNS-refined	**RMSD**	1.36 Å	
	DP	**Ensemble average**	0.868
		Average of the ensemble±SD	0.789 ± 0.0153
AMBER-refined	**RMSD**	1.42 Å	
	DP	**Ensemble average**	0.871
		Average of the ensemble±SD	0.827 ± 0.005

(b)

	RMSD			DP		
	ASDP-CYANA	CNS	AMBER	ASDP-CYANA	CNS	AMBER
Model1	2.05	1.22	1.58	0.79	0.77	0.83
Model2	1.71	1.34	1.45	0.79	0.77	0.83
Model3	2.04	1.73	1.42	0.79	0.78	0.83
Model4	1.65	1.24	1.47	0.78	0.78	0.82
Model5	1.39	1.05	1.63	0.79	0.81	0.83
Model6	2.32	1.61	1.55	0.76	0.79	0.83
Model7	1.93	1.52	1.58	0.77	0.76	0.82

Model8	1.08	1.21	1.45	0.80	0.79	0.83
Model9	1.14	0.96	1.22	0.81	0.82	0.83
Model10	2.00	1.37	1.38	0.77	0.80	0.82
Model11	1.88	1.42	1.52	0.78	0.77	0.82
Model12	2.11	1.61	1.33	0.77	0.78	0.82
Model13	1.25	1.12	1.39	0.81	0.79	0.82
Model14	1.93	1.46	1.24	0.78	0.77	0.82
Model15	1.56	1.28	1.67	0.77	0.76	0.83
Model16	1.30	1.02	1.48	0.78	0.80	0.83
Model17	1.83	1.61	1.07	0.80	0.80	0.83
Model18	1.67	1.54	1.10	0.79	0.79	0.82
Model19	2.33	1.54	1.65	0.77	0.77	0.82
Model20	1.72	1.42	1.25	0.78	0.77	0.83
Paired one-tail t-test 0.20				Paired one-tail t-test 1.31E-10		

Table 10. Structure evaluation for target PfR193A

(a) Summary of the ensemble-average and average RMSDs and DP scores

(b) Paired one-tail t-test to compare the CNS-refined and AMBER-refined conformers

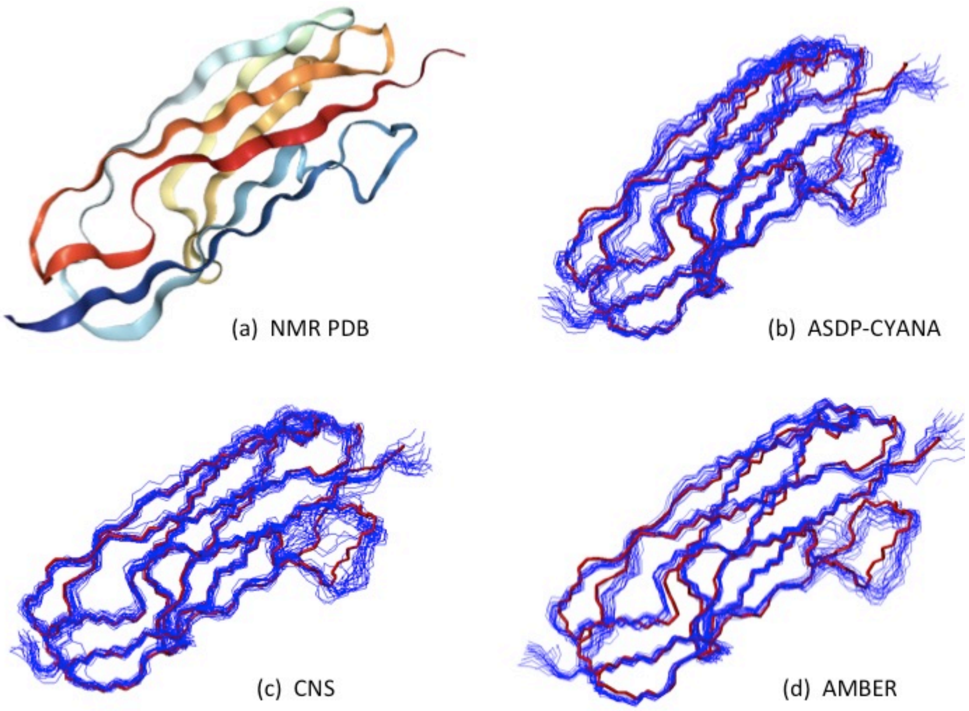


Figure 10. Visualization of the results for target PfR193A

(a) The NMR structure deposited in PDB

(b) The NMR structure ensemble output from ASDP-CYANA (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(c) The CNS-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(d) The AMBER-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(a)

Residue number	158		
X-ray PDB	**DP**	0.481	
NMR PDB	**RMSD**	3.09 Å	
	DP	**Ensemble average**	0.734
		Average of the ensemble	0.410
ASDP-CYANA	**RMSD**	4.69 Å	
	DP	**Ensemble average**	0.793
		Average of the ensemble±SD	0.531 ± 0.0194
CNS-refined	**RMSD**	5.41 Å	
	DP	**Ensemble average**	0.805
		Average of the ensemble±SD	0.400 ± 0.0277
AMBER-refined	**RMSD**	5.14 Å	
	DP	**Ensemble average**	0.811
		Average of the ensemble±SD	0.511 ± 0.0431

(b)

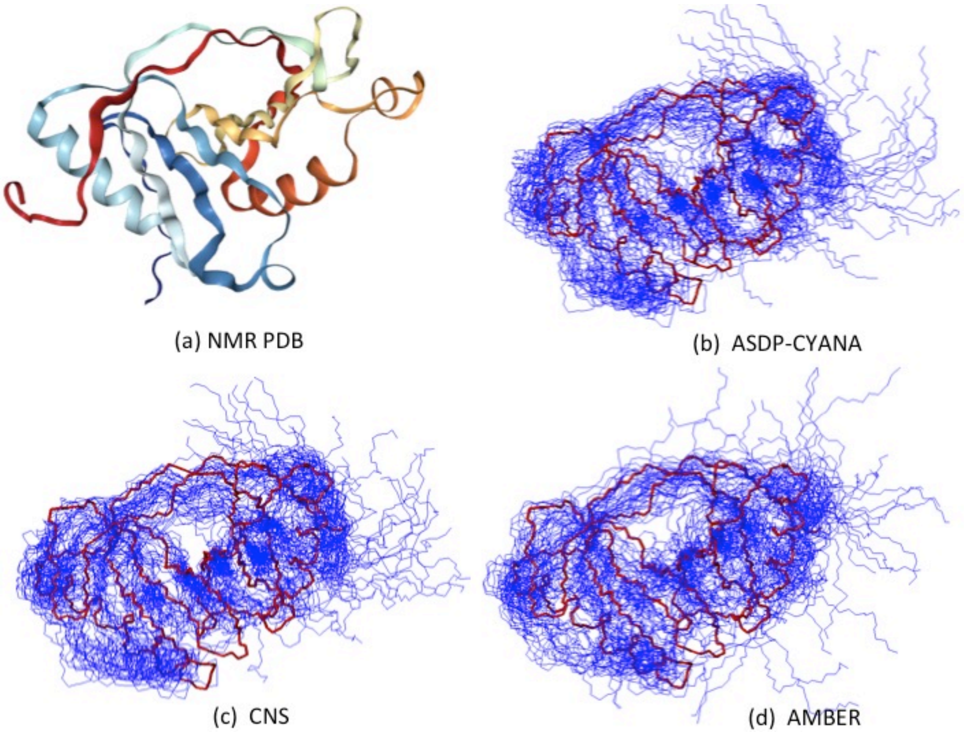
	RMSD			DP		
	ASDP-CYANA	CNS	AMBER	ASDP-CYANA	CNS	AMBER
Model1	4.72	4.87	4.12	0.56	0.40	0.57
Model2	4.25	5.38	5.76	0.56	0.36	0.52
Model3	4.76	6.23	3.78	0.54	0.40	0.50
Model4	3.49	5.55	4.82	0.54	0.36	0.50
Model5	4.80	5.16	6.67	0.52	0.39	0.43
Model6	5.44	6.17	5.59	0.52	0.36	0.43
Model7	5.33	3.86	5.33	0.52	0.39	0.47
Model8	3.74	3.79	5.07	0.51	0.38	0.57
Model9	3.81	6.10	5.38	0.52	0.43	0.54
Model10	4.72	7.35	5.38	0.51	0.32	0.48
Model11	5.55	5.97	4.42	0.52	0.39	0.48
Model12	4.77	5.67	4.68	0.51	0.36	0.53
Model13	3.96	4.51	5.43	0.51	0.41	0.50
Model14	5.08	5.94	5.58	0.50	0.38	0.49
Model15	4.85	4.35	6.78	0.50	0.36	0.46
Model16	5.68	4.46	4.76	0.50	0.39	0.43

Model17	4.30	4.54	4.84	0.50	0.36	0.49
Model18	5.55	5.90	4.76	0.50	0.35	0.56
Model19	4.55	6.15	4.25	0.50	0.44	0.49
Model20	4.38	6.28	5.46	0.50	0.38	0.53
Paired one-tail t-test 0.18				Paired one-tail t-test 4.95E-10		

Table 11. Structure evaluation for target CtR107

(a) Summary of the ensemble-average and average RMSDs and DP scores

(b) Paired one-tail t-test to compare the CNS-refined and AMBER-refined conformers



(a) NMR PDB

(b) ASDP-CYANA

(c) CNS

(d) AMBER

Figure 11. Visualization of the results for target CtR107

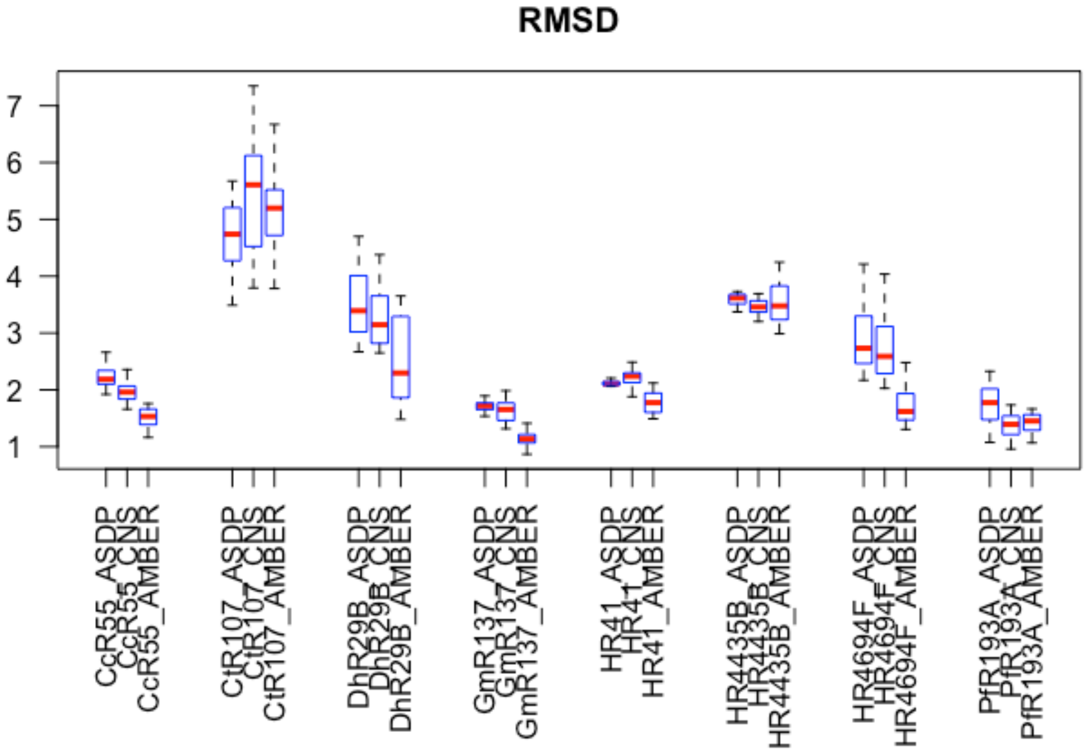
(a) The NMR structure deposited in PDB

(b) The NMR structure ensemble output from ASDP-CYANA (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(c) The CNS-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(d) The AMBER-refined NMR structure ensemble (blue lines) superimposed to the X-ray structure deposited in PDB (red stick)

(a)



(b)

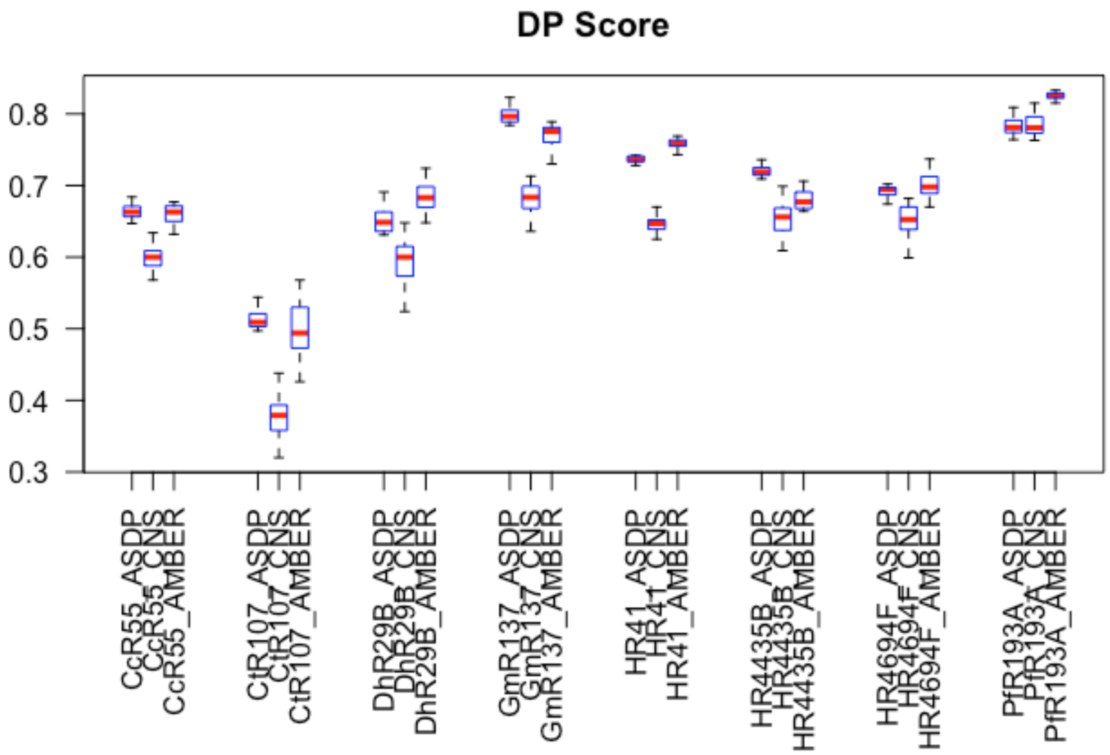


Figure 12. The distributions of RMSDs and DP scores

(a) The RMSD distributions of the ASDP-CYANA, CNS-refined, and AMBER-refined ensembles of each target.

(b) The DP distributions of the ASDP-CYANA, CNS-refined, and AMBER-refined ensembles of each target.

3.3 When the structure models are beyond certain accuracy, the linear correlation between DP scores and RMSDs becomes weaker.

The DP scores and RMSDs of each structure model in all the ensembles generated by

ASDP-CYANA, CNS, and AMBER are collected. Then DP scores are plot against the corresponding RMSDs (Figure 13). From a global point of view, there is a certain trend between DP score and RMSD, an approximately linear relationship negative correlated. The larger the RMSD, the lower the DP score. However, when we zoom in the trend, there is some discontinuity. The data points with RMSD $\geq$ 2 Å have a linear correlation equal to -0.73 while the data points with RMSD $<$ 2 Å have a correlation equal to -0.33. This implies that when the structures reach certain accuracy, DP score becomes less sensitive to the goodness of the structure in terms of RMSD.

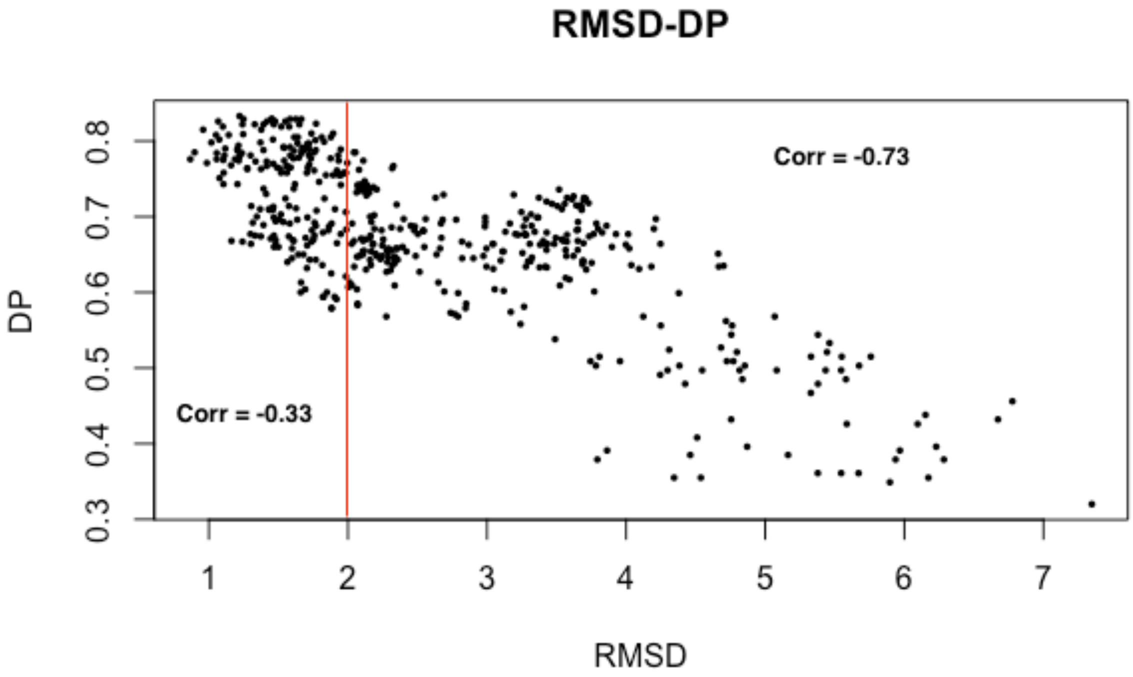


Figure 13. The correlation between the RMSDs and DP scores

4 Discussion

4.1 Fully automata structure determination is promising.

The results listed in 3.1 are encouraging. For proteins of moderate size (under 200 residues), the automatic procedure generates reasonably accurate structures compared to the manually solved PDB structures. Further efforts may result in more accurate and efficient automatic procedure, which even beats the NMR experts. Then high throughput determination of protein structures or structural genomics would come true. Yet there is one exception, target CtR107. The RMSDs are around 4-5 Å. One possible reason is that its conformation in solution is different from its conformation in crystal, supported by the large RMSD between the PDB X-ray structure and NMR structure as well as the low DP score of the PDB X-ray structure. Inconsistency between solution conformation and crystal conformation also explains why the RMSDs get larger after CNS or AMBER refinement—refinement makes the NMR structure closer to the true conformation in solution and thus more deviant from the X-ray structure. In Figure 11, the ensemble of CtR107 is loose, implying the structure calculation does not converge. This might be due to the fact that this protein becomes flexible in solution since there is a large amount of loops in it (Figure 11 a).

4.2 AMBER refinement achieves better results than CNS.

Judging from the average RMSDs and average DP scores, AMBER refinement results in better structures than CNS for all 8 targets except PfR193A. PfR193A has regular folding and is relatively rigid in solution. The automatic procedure generates sufficiently accurate structures (with RMSD a little more than 1 Å). At this level of accuracy, little space left for further improvement, and hence AMBER refinement does not beat CNS in this case. As for CtR107, the structure calculation does not even converge. The conformers output from ASDP-CYANA are not worthy further refinement. Excluding these two special targets (PfR193A and CtR107), paired one-tail t-tests provide statistical significance that AMBER refinement performs better than CNS refinement in 5 out of 6 targets.

One interesting phenomena in Figure 12 is that the DP score is sometimes lowed by further refinement. This could be explained by the nature of the algorithm implemented in ASDP. In the step of assigning the NOE peaks, structures with higher DP scores are selected. On the other hand, refinement is generally based on some potential energy function.

4.3 DP score becomes less sensitive for the structures beyond certain accuracy

Generally, there is a negative correlation between DPs and RMSDs. However, as the RMSD becomes smaller, the correlation tends to be weaker. Such phenomenon indicates that DP score would be less discriminant for evaluating structures beyond certain accuracy, say with RMSD $<$ 2 Å. There are two potential explanations. First, there is inevitably some measuring error in the NOE data, including spurious or missing peaks, which makes the NOE data deviant from the native solution conformation. Second, even assuming the NOE data is of perfect quality or there is no measuring error, the references based on which DP scores and RMSDs are computed are different. DP score measures the agreement between the NMR structure and the NOE data, while RMSD measures the similarity between the NMR structure and the X-ray structure. NOE data is a time-average measurement of solution conformers varying fast along time. The solution conformers might be close to the crystal structure, but not exactly the same. The NOE data does not fit X-ray structure perfectly, *i.e.* the DP score of the X-ray structure is not very close to 1. That is why DP score is sensitive when the structures are deviant far from the X-ray structure, and becomes less sensitive when the structures are close to the X-ray structure.

5 Conclusions

In this project, the procedure of automatic NMR structure determination is use to generate conformers for 8 targets randomly selected from the NESG depository. These conformers further go through CNS and AMBER refinements. The results show that structures obtained by the automatic procedure have comparable qualities to the ones solved by NMR experts. This promises the fully automata of NMR structure determination which is more efficient and economic than the manual approach. The results also imply that AMBER refinement has better performance that CNS refinement on normal cases. The differentiated performance might be due to the different algorithms implemented in these two programs. The detailed mechanism needs to be further studied. When the structure accuracy has reached certain threshold, RMSD does not correlated strongly with DP score.

Bibliography

Baran, M. C., Y. J. Huang, H. N. Moseley and G. T. Montelione (2004). "Automated analysis of protein NMR assignments and structures." Chem Rev **104**(8): 3541-3556.

Baran, M. C., H. N. Moseley, G. Sahota and G. T. Montelione (2002). "SPINS: standardized protein NMR storage. A data dictionary and object-oriented relational database for archiving protein NMR spectra." J Biomol NMR **24**(2): 113-121.

Bartels, C., P. Guntert, M. Billeter and K. Wuthrich (1997). "GARANT - A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra." Journal of Computational Chemistry **18**(1): 139-149.

Berjanskii, M. V. and D. S. Wishart (2005). "A simple method to predict protein flexibility using secondary chemical shifts." Journal of the American Chemical Society **127**(43): 14970-14971.

Bertini, I., D. A. Case, L. Ferella, A. Giachetti and A. Rosato (2011). "A Grid-enabled web portal for NMR structure refinement with AMBER." Bioinformatics **27**(17): 2384-2390.

Bhattacharya, A., R. Tejero and G. T. Montelione (2007). "Evaluating protein structures determined by structural genomics consortia." Proteins-Structure Function and Bioinformatics **66**(4): 778-795.

Brunger, A. T. (2007). "Version 1.2 of the Crystallography and NMR system." Nat Protoc **2**(11): 2728-2733.

Brunger, A. T., P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson and G. L. Warren (1998). "Crystallography & NMR system: A new software suite for macromolecular structure determination." Acta Crystallographica Section D-Biological Crystallography **54**: 905-921.

Chen, K. and N. Tjandra (2012). "The Use of Residual Dipolar Coupling in Studying Proteins by NMR." Nmr of Proteins and Small Biomolecules **326**: 47-67.

Clore, G. M. and A. M. Gronenborn (1998). "New methods of structure refinement for macromolecular structure determination by NMR." Proceedings of the National Academy of Sciences of the United States of America **95**(11): 5891-5898.

Crippen, G. M., A. Rousaki, M. Revington, Y. Zhang and E. R. Zuiderweg (2010). "SAGA: rapid automatic mainchain NMR assignment for large proteins." J Biomol NMR **46**(4): 281-298.

Doreleijers, J. F., M. L. Raves, T. Rullmann and R. Kaptein (1999). "Completeness of NOEs in protein structure: a statistical analysis of NMR." J Biomol NMR **14**(2): 123-132.

Everett, J. K., R. Tejero, S. B. K. Murthy, T. B. Acton, J. M. Aramini, M. C. Baran, J.

Benach, J. R. Cort, A. Eletsky, F. Forouhar, R. J. Guan, A. P. Kuzin, H. W. Lee, G. H. Liu, R. Mani, B. C. Mao, J. L. Mills, A. F. Montelione, K. Pederson, R. Powers, T. Ramelot, P. Rossi, J. Seetharaman, D. Snyder, G. V. T. Swapna, S. M. Vorobiev, Y. B. Wu, R. Xiao, Y. H. Yang, C. H. Arrowsmith, J. F. Hunt, M. A. Kennedy, J. H. Prestegard, T. Szyperski, L. Tong and G. T. Montelione (2016). "A community resource of experimental data for NMR/X-ray crystal structure pairs." Protein Science **25**(1): 30-45.

Gronwald, W., R. Kirchhofer, A. Gorler, W. Kremer, B. Ganslmeier, K. P. Neidig and H. R. Kalbitzer (2000). "RFAC, a program for automated NMR R-factor estimation." J Biomol NMR **17**(2): 137-151.

Gronwald, W., S. Moussa, R. Elsner, A. Jung, B. Ganslmeier, J. Trenner, W. Kremer, K. P. Neidig and H. R. Kalbitzer (2002). "Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE)." Journal of Biomolecular Nmr **23**(4): 271-287.

Guntert, P., K. D. Berndt and K. Wuthrich (1993). "The Program Asno for Computer-Supported Collection of Noe Upper Distance Constraints as Input for Protein-Structure Determination." Journal of Biomolecular Nmr **3**(5): 601-606.

Günther, H. a. (2013). NMR spectroscopy : basic principles, concepts, and applications in chemistry, Third completely revised and updated edition. Weinheim : Wiley-VCH, [2013]. ©2013.

Gutmanas, A., P. D. Adams, B. Bardiaux, H. M. Berman, D. A. Case, R. H. Fogh, P. Guntert, P. M. Hendrickx, T. Herrmann, G. J. Kleywegt, N. Kobayashi, O. F. Lange, J. L. Markley, G. T. Montelione, M. Nilges, T. J. Ragan, C. D. Schwieters, R. Tejero, E. L. Ulrich, S. Velankar, W. F. Vranken, J. R. Wedell, J. Westbrook, D. S. Wishart and G. W. Vuister (2015). "NMR Exchange Format: a unified and open standard for representation of NMR restraint data." Nat Struct Mol Biol **22**(6): 433-434.

Hare, B. J. and J. H. Prestegard (1994). "Application of neural networks to automated assignment of NMR spectra of proteins." J Biomol NMR **4**(1): 35-46.

Herrmann, T., P. Guntert and K. Wuthrich (2002). "Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA." Journal of Molecular Biology **319**(1): 209-227.

Huang, Y. J., R. Powers and G. T. Montelione (2005). "Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics." J Am Chem Soc **127**(6): 1665-1674.

Huang, Y. J., R. Tejero, R. Powers and G. T. Montelione (2006). "A topology-constrained distance network algorithm for protein structure determination from NOESY data." Proteins-Structure Function and Bioinformatics **62**(3): 587-603.

Ikura, M., L. E. Kay and A. Bax (1990). "A novel approach for sequential assignment of proton, carbon-13, and nitrogen-15 spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin."

Biochemistry **29**(19): 4659-4667.

Kessler, H., M. Gehrke and C. Griesinger (1988). "Two-Dimensional NMR Spectroscopy: Background and Overview of the Experiments [New Analytical Methods (36)]." Angewandte Chemie International Edition in English **27**(4): 490-536.

Kim, S. and T. Szyperski (2004). "GFT NMR experiments for polypeptide backbone and 13Cbeta chemical shift assignment." J Biomol NMR **28**(2): 117-130.

Kraulis, P. J. (1994). "Protein three-dimensional structure determination and sequence-specific assignment of 13C and 15N-separated NOE data. A novel real-space ab initio approach." J Mol Biol **243**(4): 696-718.

Kupce, E. and R. Freeman (2003). "Fast multi-dimensional NMR of proteins." J Biomol NMR **25**(4): 349-354.

Kupce, E. and R. Freeman (2004). "Fast reconstruction of four-dimensional NMR spectra from plane projections." J Biomol NMR **28**(4): 391-395.

Kuszewski, J., C. D. Schwieters, D. S. Garrett, R. A. Byrd, N. Tjandra and G. M. Clore (2004). "Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments." J Am Chem Soc **126**(20): 6258-6273.

Linge, J. P., M. Habeck, W. Rieping and M. Nilges (2003). "ARIA: automated NOE assignment and NMR structure calculation." Bioinformatics **19**(2): 315-316.

Montelione, G. T., D. Y. Zheng, Y. P. J. Huang, K. C. Gunsalus and T. Szyperski (2000). "Protein NMR spectroscopy in structural genomics." Nature Structural Biology **7**: 982-985.

Morelle, N., B. Brutscher, J.-P. Simorre and D. Marion (1995). "Computer assignment of the backbone resonances of labelled proteins using two-dimensional correlation experiments." Journal of Biomolecular NMR **5**(2): 154-160.

Mumenthaler, C., P. Guntert, W. Braun and K. Wuthrich (1997). "Automated combined assignment of NOESY spectra and three-dimensional protein structure determination." J Biomol NMR **10**(4): 351-362.

Nabuurs, S. B., C. A. E. M. Spronk, E. Krieger, H. Maassen, G. Vriend and G. W. Vuister (2003). "Quantitative evaluation of experimental NMR restraints." Journal of the American Chemical Society **125**(39): 12026-12034.

Nilges, M. (1995). "Calculation of Protein Structures with Ambiguous Distance Restraints. Automated Assignment of Ambiguous NOE Crosspeaks and Disulphide Connectivities." Journal of Molecular Biology **245**(5): 645-660.

Olson, J. B., Jr. and J. L. Markley (1994). "Evaluation of an algorithm for the automated sequential assignment of protein backbone resonances: a demonstration of the connectivity tracing assignment tools (CONTRAST) software package." J Biomol NMR **4**(3): 385-410.

Oshiro, C. M. and I. D. Kuntz (1993). "Application of distance geometry to the proton assignment problem." Biopolymers **33**(1): 107-115.

Ponder, J. W. and D. A. Case (2003). "Force fields for protein simulations." Adv Protein Chem **66**: 27-85.

Salomon-Ferrer, R., D. A. Case and R. C. Walker (2013). "An overview of the Amber biomolecular simulation package." Wiley Interdisciplinary Reviews-Computational Molecular Science **3**(2): 198-210.

Saupe, A. and G. Englert (1963). "High-Resolution Nuclear Magnetic Resonance Spectra of Orientated Molecules." Physical Review Letters **11**(10): 462-464.

Schmieder, P., A. S. Stern, G. Wagner and J. C. Hoch (1994). "Improved resolution in triple-resonance spectra by nonlinear sampling in the constant-time domain." J Biomol NMR **4**(4): 483-490.

Seavey, B. R., E. A. Farr, W. M. Westler and J. L. Markley (1991). "A relational database for sequence-specific protein NMR data." J Biomol NMR **1**(3): 217-236.

Sebastiani, D., G. Goward, I. Schnell and M. Parrinello (2002). "NMR chemical shifts in periodic systems from first principles." Computer Physics Communications **147**(1): 707-710.

Shen, Y. and A. Bax (2013). "Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks." Journal of Biomolecular Nmr **56**(3): 227-241.

Snyder, D. A., J. Grullon, Y. J. Huang, R. Tejero and G. T. Montelione (2014). "The expanded FindCore method for identification of a core atom set for assessment of protein structure prediction." Proteins-Structure Function and Bioinformatics **82**: 219-230.

Szyperski, T., G. Wider, J. H. Bushweller and K. Wuthrich (1994). "Reduced Dimensionality in Triple-Resonance Nmr Experiments (Vol 115, Pg 9307, 1993)." Journal of the American Chemical Society **116**(4): 1601-1601.

Tejero, R., D. Snyder, B. Mao, J. M. Aramini and G. T. Montelione (2013). "PDBStat: a universal restraint converter and restraint analysis software package for protein NMR." J Biomol NMR **56**(4): 337-351.

Zolnai, Z., P. T. Lee, J. Li, M. R. Chapman, C. S. Newman, G. N. Phillips, Jr., I. Rayment, E. L. Ulrich, B. F. Volkman and J. L. Markley (2003). "Project management system for structural and functional proteomics: Sesame." J Struct Funct Genomics **4**(1): 11-23.