IMPROVING THE QUALITY OF PROTEIN NMR STRUCTURES BY ROSETTA REFINEMENT AND ITS APPLICATION IN MOLECULAR REPLACEMENT

By

BINCHEN MAO

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

And

The Graduate School of Biomedical Sciences

University of Medicine and Dentistry of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Biochemistry

written under the direction of

Gaetano T. Montelione

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October , 2012

ABSTRACT OF THE DISSERTATION

IMPROVING THE QUALITY OF PROTEIN NMR STRUCTURES BY ROSETTA REFINEMENT AND ITS APPLICATION IN MOLECULAR REPLACEMENT

By BINCHEN MAO

Dissertation Director:

Professor Gaetano T. Montelione

This dissertation demonstrates restrained Rosetta refinement can improve the quality of protein NMR structures and describes a protocol to improve their phasing power. Recent studies manifest unrestrained Rosetta refinement can improve the stereochemical quality and geometry of protein NMR structures, to move NMR structures closer to their X-ray counterparts and consequently to improve their phasing power in a few cases. In this study, we intend to explore whether those observations stand corrected in general and the impact of incorporating NMR experimental restraints into Rosetta refinement.

We developed a newer version of PdbStat software to convert Cyana/Xplor formatted restraints into Rosetta formatted restraints. Based on a dataset of 41 NESG NMR/X-ray structure pairs, we have done unrestrained and restrained Rosetta refinement for all the NMR structures. The knowledge based structural quality Z-scores are significantly improved by Rosetta refinement with or without restraints. Compared with unrestrained Rosetta refined structures, restrained Rosetta refined structures fit the

experimental data better, are in better agreement with their X-ray counterparts and are generally of better phasing power, while unrestrained Rosetta refinement often drives the NMR structures further from their X-ray counterparts especially when the structural similarity between NMR structures and X-ray structures is high. To summarize, a majority of the experimental NMR restraints still apply for X-ray crystal structures determined at crystalline environment, and they can be utilized to guide Rosetta refinement to improve the quality of NMR structures.

Molecular replacement (MR) is widely used for addressing the phase problem in X-ray crystallography. Historically, crystallographers have had limited success using NMR structures as MR search models. Here, we report a comprehensive investigation of the utility of protein NMR structures as MR search models, using a dataset of 25 NESG NMR/X-ray structure pairs. Starting from NMR ensembles prepared by an improved protocol, FindCore, correct MR solutions were obtained for 22 targets. Rosetta refinement of NMR structures provided MR solutions for another two proteins. We also demonstrate that such properly prepared NMR ensembles and X-ray crystal structures have similar performance when used as MR search models for homologous structures, particularly for targets with sequence identity >40%.

ACKNOWLEDGEMENT

First and foremost I would like to express my deepest gratitude to my advisor, Dr. Gaetano T. Montelione, for his expertise, leadership, caring, and providing me with an excellent atmosphere for doing research. The work presented in this dissertation would not have been possible without his guidance, constant encouragement and support.

I would also like to give my special thanks to all other members of my thesis committee - Dr. Helen Berman, Dr. Steve Anderson and Dr. Vikas Nanda, for their constructive feedback, support and patience.

The members of the Montelione group have contributed immensely to my personal and professional time at Rutgers. The group has been a source of friendships as well as good advice and collaboration. Especially, I would like to thank the following people for their contribution to my thesis: Dr. Janet Yuanpeng Huang has trained me when I first came to the lab and instructed me on a variety of structural bioinformatics projects I have been involved in, such as disMeta, CS-DP-Rosetta, and CASD-NMR. Dr. Rongjin Guan has collaborated with me on NMR MR project and Dr. Roberto Tejero has collaborated with me on Rosetta-MR project. I would also like to acknowledge all the people in Dr. David Baker's lab, for developing Rosetta and giving me precious instructions on Rosetta applications. Moreover, my special thanks goes out to all the members in Montelione lab and in Northest Structural Genomics Consortium, without their tremendous efforts on protein production and structure determination, this dissertation would be like a fish out of water.

Lastly, I am grateful to acknowledge my debt to my family, especially for my parents and my wife, for their love and consistent support along the journey.

TABLE OF CONTENTS

Improved Technologies Now Routinely Provide Protein NMR Structures Useful for Molecular Replacement

LIST OF TABLES

LIST OF ILLUSTRATIONS

INTRODUCTION

Nuclear magnetic resonance (NMR) is a powerful tool to determine protein structures in solution and in the solid state. It typically requires both the resonances assignment and multidimensional nuclear Overhauser effect spectroscopy (NOESY) spectra analysis, which could be quite time consuming and error-prone if done manually . In the past few years, a variety of programs have been developed to fully automate the NOESY assignment and the structure calculation steps, which has the potential to boost the efficiency, reproducibility and reliability of NMR structure determination. To evaluate their respective strength and deficiency, critical assessment of automated structure determination of proteins by NMR (CASD-NMR)[1] project was launched at 2009.

The Northeast Structural Genomics (NESG) consortium is one of the large-scale structure production centers of the Protein Structure Initiative (PSI). The NESG has contributed more than 450 NMR structures to the PDB over the past ten years, representing a large fraction of the NMR structures deposited into the PDB by the PSI. One of the most important objectives of NESG is to develop new techniques for NMR structure determination. In this regard, since the onset of CASD-NMR project, NESG has played a key role as both the data provider and one of the competitors. Several groups across the Europe and Unite States have participated in this project, and an analysis of 10 blind targets has revealed that routine application of NMR structure calculation methods integrating NOE crosspeak assignment and structure generation is both feasible and reliable, under the condition that the NOESY peak lists are carefully curated by the NMR spectroscopists[2].

Conventional structure determination by NMR requires complete assignment of the chemical shifts (backbone and side chain) and complete assignment of the NOESY peak list, which could be quite labor-intensive. While automated structure determination programs can successfully assign a large fraction of the NOESY peaks for small proteins when provided with high quality NOESY peak list data, challenges arise when the size of protein is considerably large of the quality of NOESY data is poor. Recently, the CS-Rosetta method has been demonstrated to be able to consistently generate high-accuracy models for small proteins starting from backbone chemical shift information alone[3]. However, the CS-Rosetta method does not generally converge for proteins of complex folds or more than 110 residues due to the enormous conformational search space, this could be ameliorated by the guidance of the NOESY peak lists data for Rosetta fold trajectory search. Calculated from RPF software suite[4], the DP-score is utilized to evaluate the goodness-of-fit of protein NMR structures to experimental data, therefore it can be integrated into CS-Rosetta calculation to filter the decoys when the NOESY data are available. In this regard, CS-DP-Rosetta protocol was proposed, which 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[5].

NMR spectroscopy has contributed a substantial fraction of structures in Protein Data Bank (PDB), and it is currently the only technique to determine the structure of macromolecule in solution state. In structural biology community, it is commonly accepted there is a quality gap between NMR structures and X-ray structures and the value of NMR structures being used as molecular replacement (MR) starting models is limited. Although there were a few cases being reported of successfully using NMR structures in MR studies in the past, a systematic investigation of using NMR structures

as MR templates had still been lacking until our work. As of December 2009 the NESG consortium had solved 27 pairs of protein structures for identical construct sequences using both X-ray crystallography and NMR methods. These 3D structures of proteins with identical sequences,together with the raw NMR and crystallography data available in the BioMagRes and PDB, are an ideal starting point for our NMR for MR study.

Model preparation is a cornerstone of MR success. A number of protocols to prepare the MR search model had been proposed previously. These are generally designed to exclude structurally disordered residues or trimming the long side chains to their common bases. One of the major deficiencies of those protocols is that the structural precision information is only considered at the level of amino acid at best. Therefore, based on the interatomic variance matrix calculation, we use FindCore program[6] to calculate the atomic pseudo B-factor of protein NMR ensembles, which is a good estimation of structural precision at an atomic level. MR starting models are prepared based on those pseudo B-factors in our study, which are generally of better phasing ability than the models prepared otherwise. We are able to get correct MR solutions for 22 out of 25 targets. Rosetta refinement of NMR structures can provide MR solutions for another two proteins. We have also demonstrate that such properly prepared NMR structures and X-ray crystal structures have similar performance when used as MR search models for homologous structures, particularly for targets with sequence identity >40%[6].

The NMR structure quality indicators generally fall into two categories: one is related to experimental data, such as restraint violations, NOE completeness and goodness-of-fit with NMR NOESY peak list data; the other is the knowledge-based normality scores relative to high-resolution X-ray crystal structure database, such as bond length, bond angle, backbone or side chain dihedral angle, and packing statistics. CASD-NMR study has shown that the algorithm and force field utilized in NMR structure

determination and structure refinement have a big impact on NMR structure quality, for example, NMR structures refined in Rosetta force field are generally of excellent stereochemical and geometric quality scores[2].

Recent studies have also demonstrated that unrestrained Rosetta refinement can move NMR structures closer to their X-ray counterparts and consequently to improve their phasing power in a few cases. NMR restraints can be incorporated into Rosetta refinement nowadays. We intend to explore whether those findings stand corrected in general and to investigate the impact of incorporating NMR experimental restraints into Rosetta refinement. A newer version of PdbStat software has been developed to convert Cyana/Xplor formatted restraints into Rosetta formatted restraints. In this work, we have done both unrestrained Rosetta refinement and restrained Rosetta refinement for all the NMR structures of 41 NESG NMR/X-ray structure pairs. The quality of PDB NMR structures, unrestrained Rosetta refined structures and restrained Rosetta refined structures have been evaluated by PSVS web server(http://psvs.nesg.org), including restraint violations analysis, ensemble RMSD calculation, knowledge based normality analysis and RPF analysis. We have also calculated their structural similarity with the corresponding X-ray structures and how well they could be utilized as molecular replacement templates.

The knowledge based structural quality Z-scores are significantly improved by Rosetta refinement with or without restraints. Compared with unrestrained Rosetta refined structures, restrained Rosetta refined structures have significantly less restraint violations, fit the NOESY peak list better, are in better agreement with their X-ray counterparts and are generally of improved phasing power, while unrestrained Rosetta refinement often drives the NMR structures away from their X-ray counterparts especially when initially the structural similarity between NMR structures and X-ray structures is high. For small size protein NMR structures of poor structural similarity with

their corresponding X-ray structures, CS-Rosetta calculation with the experimental restraints is proven to be a better choice than restrained Rosetta refinement. To summarize, a majority of the experimental NMR restraints still apply for X-ray crystal structures determined at crystalline environment, and they can be utilized to guide Rosetta refinement to improve the quality of NMR structures.

CHAPTER 1

CASD-NMR : Critical Assessment of Automated Structure Determination by NMR

Introduction

The typical protocol for protein structure determination by NMR spectroscopy involves a number of sequential steps[8]. First, the chemical shifts (CS) observed in multidimensional NMR spectra are assigned sequence-specifically to their corresponding protein atoms (the resonance assignment step). Second, thousands of through-space dipolar coupling effects, known as nuclear Overhauser effects (NOEs), are identified in multidimensional NOESY spectra (peak picking), assigned and converted into inter-atomic distance restraints (NOESY assignment step). Additional conformational restraints can result from e.g. measurements of residual dipolar couplings (RDCs), scalar couplings, and CS data. Third, software programs are used to generate a set of protein conformations (called a bundle of conformers) that should satisfy these experimental restraints (structure generation step). The bundle of conformers is often energetically refined through restrained molecular dynamics simulations (structure refinement step).

The *NOESY assignment* and *structure generation* steps are performed in an integrated manner over several iterations in order to maximize the number of conformational restraints obtained while guaranteeing the self-consistency of all distance restraints (measured *a* posteriori from the absence of significant distance restraint violations). Many of the tasks in the *NOESY assignment* step are repetitive, although

non-trivial, yet typically they must be performed by a skilled researcher. A considerable bookkeeping effort is also needed in order to converge to a self-consistent set of conformational restraints from which the final bundle of low pseudo-energy conformers is calculated. For these reasons, and to enhance reproducibility, automation of the aforementioned steps has been actively pursued[9,10,11]. Protocols aimed at the integration of all steps of the protocol for protein structure determination by NMR have also appeared[12].

In 2009 the community-wide initiative called "Critical assessment of Automated Structure Determination of proteins by NMR (CASD-NMR)" was launched (http://www.wenmr.eu/wenmr/casd-nmr)[1], with the aim to assess whether automated methods addressing the NOESY assignment (if needed), structure generation and structure refinement steps can, in a fully automated manner, produce protein structures that closely match the structures manually determined by experts using the same experimental data ("reference structures"). To this end, we have released regularly over one year NMR data sets consisting of assigned chemical shift lists and unassigned NOESY peak lists, while the reference structures determined from the same data were kept "on hold" by the Protein Data Bank (PDB)[13], and were thus unavailable to the participants. Each of these data sets is referred to as a masked data set. The protocols used to determine the reference structures are summarized in the methods section, and typically involved manual refinement (such as fixing assignments or removing artifacts) of initial, partly automated NOESY assignments performed with various tools. The final, iteratively obtained lists of resonance assignments and NOESY peak positions were subsequently provided to the CASD initiative.

Here we report the results obtained in the first round of CASD-NMR (CASD-NMR2010) for a total of ten masked data sets, provided by the NIH Protein Structure Initiative, for monomeric proteins of 60 to 150 amino acids. All the input data as well as

the structures generated in the present CASD-NMR2010 study can be freely downloaded from http://www.wenmr.eu/wenmr/casd-nmr. CASD-NMR2010 did not address automation methods for determining resonance assignments and for NOESY peak picking. We chose to postpone the assessment of these parts of the process until the NOE assignment and structure calculation steps will have been demonstrated to be truly robust.

The results of the CASD-NMR2010 study presented here demonstrate that routine application of NMR structure calculation methods integrating NOE cross peak assignment and structure generation is both feasible and reliable. Furthermore, the recently developed approaches based on the use of NMR chemical shift data to generate structural models were found to benefit significantly when supplemented with information from unassigned NOESY peak lists**.**

Methods

Determination of reference structures

The reference structures were manually solved by the NESG co-authors (see the corresponding PDB entries for details).

The cloning, expression, and purification of ^{13}C and ^{15}N isotopically-enriched samples of the following proteins for solution NMR structure determination were conducted using standard protocols of the NESG. Automated backbone ^{1}H, ^{13}C, and ^{15}N resonance assignments were made using either AutoAssign[14], PINE[15], or FAWN[16], followed by manual side chain assignment. The assignments of AtT13 were obtained by an ABACUS approach[17]. In general, iterative structure calculations were done using constraints derived from automated NOESY assignments determined with CYANA[18,19],

manual curation of NOESY spectral peak lists, and structure generation using either CYANA or Xplor-NIH[20]. These NOE-based distance constraints were supplemented with other constraints, such as manually-assigned NOESY-based distance constraints and/or hydrogen bond constraints derived from N-H exchange data, backbone dihedral angle constraints computed by TALOS or TALOS+[21,22] for residues in well-defined secondary structure elements, and in some cases residual dipolar coupling data, as described below. The near final structures were carefully inspected, and in some cases manually-defined dihedral angle constraints were used in the final stages of structure refinement to constrain side chains to low energy rotamer states. In all cases, the final ensemble of structures was refined by restrained molecular dynamics in explicit water using CNS[23,24] or using an Xplor+ refinement protocol[25].

Structure calculations on VpR247[26], HR5537A[27], NeR103A[28], CgR26A[29], and CtR69A[30] were performed using CYANA 2.1 (CgR26A and CtR69A) or CYANA 3.0 (VpR247, HR5537A, and NeR103A), and the 20 conformers out of 100 with the lowest target functions were refined in explicit water using CNS 1.2 supplied with NOE-derived distance constraints and backbone dihedral angle constraints; hydrogen bond constraints were also applied in the case of VpR247. For AR3436A[31], structures were computed using AutoStructure 2.2.1 interfaced with CYANA 2.1 and the 20 conformers out of 140 with the lowest target functions were refined in explicit water using CNS 1.2 and NOE-derived distance constraints, backbone dihedral angle constraints, and hydrogen bond constraints. Structure calculations on AtT13[32] and the reduced and oxidized forms of ET109A were performed using FMCGUI interfaced with CYANA 2.1, and the best 20 out of 100 structures from the final cycle were refined in explicit water using CNS 1.2 supplied with NOE-derived distance constraints and backbone dihedral angle constraints; residual dipolar couplings ($^1D_{NH}$, $^1D_{CC'}$, and $^1D_{NC')}$ were applied in the

final refinement stage of the ET109A structure determinations[33]. In the case of Pgr122A[34], structures were calculated using Xplor-NIH (version 2.25) with a simulated annealing protocol, followed by refinement of the 20 structures out of 150 with the lowest energies using Xplor+ augmented with NOE-based distance constraints, backbone dihedral angle constraints, and hydrogen bond constraints. For each final structural ensemble, structural statistics and global structure quality factors were computed using the PSVS software package[35], and the global goodness-of-fit of the coordinates with the NOESY peak list data was assessed using the RPF analysis program[4].

Data distribution

Masked data sets for CASD-NMR 2010 (whose amino acidic sequences are given in Supplemental Table S1) comprised chemical shift assignments in BMRB format and unassigned NOESY peak lists in SPARKY and/or XEASY/CARA format. The data were made available both via the CASD-NMR website (www.wenmr.eu/wenmr/casd-nmr) and a dedicated page at the Protein Structure Initiative (PSI Knowledge Base (http://kb.psi-structuralgenomics.org/). For two targets raw NOESY spectra were also made available. At the time of release, all participants were notified of the availability of a new data set as well as of the date of release of the corresponding structure from the PDB (about eight weeks later). The automatically calculated structures and all restraints were deposited directly by the participants into a password-protected database again via the CASD-NMR website.

Residual dipolar coupling data and hydrogen bond restraints were not used in the CASD-NMR 2010 project.

Calculation Protocols

Each method developer team carried out calculations with their own program, as detailed below.

CYANA

Structure calculations by the CYANA method[10] used as input data from the blind data sets the protein sequence, the list of assigned chemical shifts, and the unassigned NOESY peak lists. Torsion angle restraints were generated on the basis of the chemical shift values with the program TALOS+[22] for the backbone torsion angles and of non-proline residues with a prediction classified as "Good" by TALOS+. The torsion angle restraints were centered at the predicted average value and their full width was set to four times the predicted standard deviation or 20°, whichever was larger. The program CYANA was used for seven cycles of combined automated NOE assignment[19] and structure calculation by torsion angle dynamics[18]. The tolerance for the matching of chemical shifts and NOESY peak positions was set to 0.03 ppm for ^{1}H and 0.5 ppm for ^{13}C and ^{15}N. Peak intensities were converted into upper distance bounds according to a $1/r^6$-relationship. Each structure calculation was started from 100 conformers with random torsion angle values, the standard CYANA simulated annealing schedule was applied with 15000 torsion angle dynamics steps, and the 20 conformers with lowest CYANA target function values were analyzed. NOE distance restraints involving ^{1}H atoms with degenerate chemical shifts, e.g., methyl groups, were treated as ambiguous distance restraints using $1/r^6$-summation over the distances to the individual ^{1}H atoms. Non-stereospecifically assigned methyls and methylene protons were treated by automatic swapping of restraints between diastereotopic partners during the seven cycles of automated NOE assignment and by pseudoatom correction and

symmetrization for the final structure calculation. The 20 conformers with the lowest final CYANA target function values were embedded in an 8 Å shell of explicit water molecules and subjected to restrained energy refinement against the AMBER force field[36] using the program OPALp[37,38]. A maximum of 3000 steps of restrained conjugate gradient minimization were applied, using the standard AMBER force field and pseudo-potentials proportional to the sixth power of the NOE upper distance bound violations and the square of the torsion angle restraint violations, respectively. The entire procedure was driven by the program CYANA, which was also used for parallelization of all time-consuming steps on 10–100 processors of a Linux cluster system with Intel quad-core 2.4 GHz processors.

UNIO

For all blind data sets NOE assignment were performed using the modules ATNOS/CANDID and/or the CANDID module alone incorporated into the software UNIO[39.40], depending if NOE peak lists or NOESY spectra were provided for a given CASD-target. The standard UNIO protocol with seven cycles of peak picking with ATNOS, if NOESY spectra were provided, and NOE assignment with CANDID was used. During the first six UNIO-ATNOS/CANDID cycles, ambiguous distance restraints were used. At the outset of the spectral analysis, UNIO-ATNOS/CANDID used highly permissive criteria to identify and assign a comprehensive set of peaks in the NOESY spectra or the unassigned peak lists provided. Only the knowledge of the covalent polypeptide structure and the chemical shifts were exploited to guide NOE cross peak identification and NOE assignment. In the second and subsequent cycles, the intermediate protein three-dimensional structures were used as an additional guide for the interpretation of the NOESY spectra or unassigned peak lists. The output in each ATNOS/CANDID cycle consisted of assigned NOE peak lists for each input spectrum

and a final set of meaningful upper limit distance restraints which constituted the input for the torsion angle dynamics algorithm of CYANA for 3D structure calculation. In addition, torsion angle restraints for the backbone dihedral angles ϕ and ψ derived from C^{α} chemical shifts were automatically generated in UNIO and added to the input for each cycle of structure calculation[41,42]. For the final structure calculation in cycle 7, only distance restraints were retained from UNIO that could be unambiguously assigned based on the protein three-dimensional structure from cycle 6.

The 20 conformers with the lowest residual CYANA target function values obtained from cycle 7 were energy-refined in a water shell with the program OPALp[37,38] using the AMBER force field[36].

ASDP

^{13}C chemical shift was first referenced based on the LACS method[43]. AutoStructure's topology-constrained distance network algorithm[44] was used to assign NOE peaks, using the list of resonance assignments, and the unassigned NOESY peak lists. The tolerance to match chemical shifts with NOE peak positions was set to 0.05 ppm for ^{1}H and 0.5 ppm for ^{13}C and ^{15}N. Distance constraints were generated based on these NOE assignments. Dihedral angle constraints were generated using TALOS+[22], using only sites with TALOS+ scores = 10 and constraining the dihedrals to the defined range ± 20° or twice the standard deviation, whichever was larger. One hundred structures were generated using CYANA[18] standard structural calculation module and DP-scores[4] were calculated for all 100 structures. We then computed a new score: (target function/100)-DP for each model, and the 20 models with highest scores were selected for additional iterative 5 cycles of NOE analysis with AutoStructure and structure generation with CYANA[18]. After six cycles of ASDP analysis, the resulting

structures were energy-refined using CNS[45] with explicit water. If any TALOS+ dihedral angle constraints were observed to be violated in all 20 models, they were removed and the ASDP / CNS refinement process was repeated.

ARIA

Two protocols were used: one (ARIA-Soft) based on the standard soft-square distance restraint potential, the other (ARIA-BayW) based on a log-harmonic potential shape[46] and iterative determination of the optimal data weight[47,48]. ARIA 2.2[49] was used with the ARIA-Soft protocol, and ARIA 2.3 with the more recent ARIA-BayW protocol. ARIA-Soft was applied to targets VpR247, HR5537A, ET109A, AtT13, PgR122A, whereas ARIA-BayW was applied to targets NeR103A, CgR26A and CtR69A.

Dihedral angle restraints were generated on the basis of the chemical shift values with the program TALOS+[22] for the backbone torsion angles ϕ and ψ. The predictions classified as “good” by TALOS were converted into dihedral angle restraints with the script talos2xplor.tcl. For analyzing NOESY crosspeaks, the tolerance for matching chemical shifts and peak positions was set to 0.04 and 0.02 ppm for indirect and direct ^{1}H dimensions and to 0.5 ppm for ^{13}C and ^{15}N.

For each calculation, we ran eight ARIA iterations in a simplified, geometric force field, and one refinement iteration in water with full electrostatics[24]. Structures were calculated with CNS[45], recompiled with specific ARIA subroutines. The standard four-phase ARIA simulated annealing protocol was used, with 2200 TAD steps at 20 000 K, 2200 TAD steps cooling from 20 000K down to 0K, 10000 Cartesian cooling steps from 2000K to 1000K, and 8000 cooling steps from 1000K to 50K. Molecular dynamics was followed by 200 steps of conjugate gradient minimization. For the water refinement, we used heating from 100 to 500 K in steps of 100 K with 750 dynamics steps at each

temperature, during which positional restraints on the heavy atom positions were progressively relaxed; 2000 steps of refinement at 500K; cooling to 25K in steps of 25K, with 1000 integration steps at each temperature, followed by 200 steps conjugate gradient minimization. The log-harmonic potential and the Bayesian weight determination were only used in the final cooling phase, minimization and water refinement. 50 conformers were generated in each iteration. The 15 conformers with lowest (extended) hybrid energy were analyzed to refine the restraint list. After the eighth iteration, the 10 conformers with the lowest energy were further refined in water.

CHESHIRE

In the structure calculations two protocols were used, CHESHIRE and CHESHIRE-YAPP. CHESHIRE uses only chemical shifts, while CHESHIRE-YAPP uses a combination of chemical shifts and unassigned NOESY peak lists.

CHESHIRE consists of a three-phase computational procedure[50]. In the first phase, the chemical shifts and the intrinsic secondary structure propensities of amino acid triplets are used to predict the secondary structure of the protein. In the second phase, the secondary structure predictions and the chemical shifts are used to predict backbone torsion angles. These angles are screened against a database to create a library of trial conformations of three and nine residue fragments spanning the sequence of the protein. In the third phase, a molecular fragment replacement strategy is used to assemble low-resolution structural models. The information provided by chemical shifts is used in this phase to guide the assembly of the fragments. The resulting structures are refined with a hybrid molecular dynamics and Monte Carlo conformational search using a scoring function defined by: (1) the agreement between experimental and calculated chemical shifts, and (2) the energy of a molecular mechanics force field. This

scoring function ensures that a structure is associated with a low CHESHIRE score only if it has a low value of the molecular mechanics energy and is highly consistent with experimental chemical shifts. Typically 50,000 structures were generated for each target and the best scoring one was submitted. This protocol was used for five targets (VpR247, AR3436A, HR5537A, PGR122A and CtR69A).

The CHESHIRE-YAPP protocol uses the best scoring 500-1000 high-resolution structures generated by CHESHIRE to select compatible NOEs from the unassigned NOESY peak lists. NOEs are selected using an iterative protocol. In the first step, atoms are assigned to each spectral dimension using a chemical shift tolerance of 0.03 ppm for ^{1}H and 0.3 ppm for ^{13}C and ^{15}N. Then, chemical shift-based assignments that are violated by more than 2Å in 50 or more of the best 500 CHESHIRE structures are removed. The remaining restraints are used to refine the best scoring 100 CHESHIRE structures. The last two steps are repeated 4 times with a threshold for violations of 1.5, 1.0, 0.5 and 0.2Å. This protocol was used for three targets (ET109A, NeR103A and CGR103A).

CS-DP-ROSETTA

Fragments were picked using the original CS-Rosetta fragment picker[3] . Decoys were generated on Rosetta@home using 50,000 boinc work units (ca. 200,000 CPU hours). This resulted in 10^5-10^6 decoys, depending on the target. Decoys were generated with the standard CS-Rosetta protocol[3] and relaxed in full-atom resolution, as described by Raman et al.[5] . The best 1000 decoys were selected by score and their DP-score was calculated with AutoStructure (version 2.2.1)[44]. To finally rank the models, we computed the final score S = R + 1000(1-DP), with R for the Rosetta full-atom score

and DP for the DP-score, and selected the 10-20 best models for submission to the CASD website.

CS-ROSETTA (Web Server)

The CS-Rosetta webserver developed under the eNMR project[51] was used. Firstly, the supplied NMR chemical shift data were pre-checked on chemical shift referencing and possible errors, using the standard pre-check option of the TALOS+ program[22]. TALOS+ was then used to identify flexible residues at the termini of the protein (those classified as either “Dynamic” or “Not classified” by TALOS+). These and any histidine tags were removed. The resulting cleaned TALOS+ file was submitted to the server. For each target 50000 models were generated on the Grid following the standard CS-ROSETTA protocol[3] using the original CS-Rosetta fragment picker and Rosetta version 2.3.0. The 1000 best ROSETTA score models were rescored using chemical shift rescoring as in the CS-ROSETTA protocol. After rescoring, if convergence was observed in the top five models (backbone RMSD below 2Å), these were submitted as prediction for CASD-NMR, otherwise only the top scoring model was submitted.

For the last two targets, we implemented a novel smoothing procedure on the Rosetta raw score: for each model, a smoothed score was calculated as a Gaussian-weighted average score calculated over all structural neighbors within a 4.5Å C^{α}-RMSD cutoff. The smoothing was performed on the top 5000 models. The top 1000 models after smoothing were then rescored using the regular CS-scoring in CS-ROSETTA. This smoothing procedure removes some of the noise in the raw score and strengthens any weak correlation that might be present in the data set.

Results

Accuracy and convergence of structure calculations

CASD-NMR2010 involved three groups of automated methods (Table 1.1): those using NOESY data to obtain distance restraints for structure calculations (CYANA, UNIO, ASDP and ARIA), those using chemical shift data augmented by NOESY data (CS-DP-Rosetta, which uses NOESY information to re-rank its CS-based results, and Cheshire-YAPP, which uses CS-generated structures to perform NOESY assignments and extract distance restraints), and those relying exclusively on CS data as experimental information (Cheshire and CS-Rosetta). The NOESY-based methods include a structure refinement step after structure generation with the aforementioned programs. Both steps exploit all automatically assigned restraints. A variety of programs has been used for the refinement (also in the case of the reference structures).

For each data set, we used the deviation of the backbone coordinates (RMSD) to quantify the degree of convergence (i.e. the similarity) among the automatically generated structures as well as their closeness to the reference structure determined under manual supervision. Assuming that the reference structure is correct, the RMSD to it becomes a measure of accuracy. We computed the RMSD to the reference for the structures generated by all the methods (Tables 1.1 and 1.2, Figure 1.1A). As the RMSD calculations require the *a priori* definition of residue ranges to be superimposed, a consensus RMSD range comprising the well-ordered residues in the reference structure was chosen for each dataset. In order to avoid a possible bias from this selection when evaluating the similarity to the reference structure, we computed also the Global Distance Test Total Score (GDT_TS, Figure 1.1B), which does not require residue ranges to be predefined and is independent of protein size. The GDT_TS score has

been developed in the frame of the Local-Global alignment method[52]. It is defined by GDT_TS = (P_1 + P_2 + P_4 + P_8)/4, where P_d is the percentage of residues that can be superimposed under a distance cutoff of *d* Å. This definition reduces the dependence on the choice of the distance cutoff by averaging over four different distance cutoff values.

The backbone RMSD values to the reference for the structures generated by NOESY restraint-based methods were in the range 0.6-2.7 Å whereas the range for GDT_TS scores were 61-94% (Table 1.2 and Table 1.3). Setting thresholds for an acceptable structural accuracy (here assumed to be quantified by similarity to the reference structure) at an RMSD from the reference structure $\leq$ 2 Å and GDT_TS $\geq$ 80%, three of the four NOESY-based programs (CYANA, UNIO and ASDP) automatically and consistently generated acceptable structures, based on one (90-100% of the instances) or simultaneously both (80-90% of the instances) parameters (Table 1.2). The RMSD was always $\leq$ 2.2 Å, whereas the lowest GDT_TS was 61% (78% upon exclusion of target AR3436). The fourth program, ARIA, performed acceptably for nearly 80% of the targets, with the best results obtained with a recently developed logharmonic potential combined with a Bayesian determination of restraint weights (protocol ARIA-BayW)[53], which produced structures with excellent GDT_TS and RMSD values for the three most recent targets.

Regarding CS-based methods augmented with NOESY data, Cheshire-YAPP, which was developed during CASD-NMR2010 and run on three randomly selected targets, featured a similarity to the corresponding reference structures in-line with NOESY restraint-driven methods. Cheshire-YAPP uses initial (pure CS) Cheshire models to assign NOESY distance restraints used to refine the models. For CS-DP-Rosetta, which uses NOESY information only to re-rank the CS-based models, the deviation from the manual reference structures was close to that of the NOESY restraint

methods, with a range of RMSD and GDT_TS values of, respectively, 0.3-3.3 Å and 55-90% and 70% of targets falling within the thresholds described above. Finally, pure CS-based methods had the poorest performance in terms of closeness to the reference structures, as it is apparent from Table 1.2 and Figure 1.1. Note that the poorer appearance of the CS-Rosetta server, which was run via the web server developed in the e-NMR project, is partly due to inclusion of non-converged solutions in the comparison. It can be concluded that NOESY-based methods delivered more consistent and robust performances than CS-based methods (resulting in smaller boxes in Fig. 1.1A-B), yielding structures on average closer to the reference. NOESY-filtering as in CS-DP-Rosetta could recover some but not all of the consistency and reliability of the restraint-driven methods (see also below). Notably, the CS-methods (regardless of whether augmented with NOESY information) are computationally much more demanding (several orders of magnitude) than NOESY-based methods.

Regarding individual targets, the one with the lowest performance across all methods was AR3436A (Table 1.2), a 97-amino acid protein. Our target selection included three proteins with more than 100 residues (HR5536A, AtT13 and CgR26A), for all of which NOESY-based methods were able to automatically generate accurate structures. Instead, purely CS-based methods failed for all of them, whereas CS-based methods augmented with NOESY data were successful in nearly all cases.

All the results examined in the preceding paragraphs address the degree of similarity to the manually solved reference structure. Additional insight can be obtained by the evaluation of the degree of convergence among the different programs. This has been measured as the mean RMSD among the average conformers obtained with the automatically generated methods (Table 1.4). For the NOESY-based algorithms, the mean RMSD for each target was in the range 0.9-3.0 Å, with four targets featuring a

mean RMSD lower than 1.0 Å and eight targets being within 2.0 Å. If CS-based methods augmented with NOE cross peak information are also included, the mean RMSD range widens slightly up to 3.3 Å, still with eight targets having a mean RMSD lower than the 2.0 Å threshold. Instead, inclusion of all methods yielded values as large as 6.2 Å (Table 1.4). Note that the present evaluation of convergence is the much more stringent than the standard re-calculation with different random number seeds, because in each calculation the NOE assignments have been determined in an independent manner, with different methods.

Finally, an independent measure of accuracy would be the comparison with a completely independent structure determination. This is at present possible for only two targets (VpR247 and PgR122A), for which the PDB contains X-ray structures of relatively close homologues (40-50% sequence identity). These allowed us to build reliable structural models that can be used as the structural reference for comparisons (Table 1.5). For PgR122A, the relevant structure is 3HVZ. The homology model of PgR122A built on this structure shows a backbone RMSD of 0.77 Å to the average coordinates of the reference structure. All methods yielded structures within 1.5 Å from the homology model, with the majority being actually within 1 Å. For VpR247 there are several related crystal structures of the *S. pombe* homologue, in the free or ligand-bound form. The model built on the DNA-complexed protein (3GX4) is closer to the reference VpR247 structure than the model built on the free protein (3GVA), with backbone RMSD values of 1.4 Å and 2.1 Å, respectively. Similarly, nearly all the automatically generated structures are more similar to the former than the latter model. With the exception of the ARIA and CS-Rosetta server structures (Table 1.5), all structures are within 2.0 Å from the 3GX4-based model, whereas they are in the range 1.7-2.2 Å from the 3GVA model. These results may suggest that the free VpR247 protein in solution populates a different

conformational state than its *S. pombe* homologue in the crystal structure. This state would be relatively similar to the DNA-bound conformation.

Geometric and stereochemical quality

The geometric and stereochemical quality is another important property of a structure that must be checked prior to deposition in the PDB. We evaluated this aspect using the PSVS[35] (http://psvs-1_4-dev.nesg.org/) and CING (http://nmr.cmbi.ru.nl/cing/) validation suites (Table 1.3), which assess several quality measures. The Verify3D[54] and ProsaII[55] scores, which evaluate the global fold likelihood, were not significantly different for the CASD-NMR or the reference structures and featured relatively wide ranges for all the various algorithms. Instead, the Procheck-all[56] score, which assesses the distribution of all the protein dihedral angles, and the MolProbity clashscore[57], which assesses the occurrence of high-energy interatomic contacts, differed among the CASD-NMR structures, even though their ranges over all targets overlapped with the reference structures (Fig. 1.2). The ranges of Procheck-all values for the structures generated by the Rosetta-based algorithms are narrow and on average significantly better than for the other structures (Figure 1.2B). Also the MolProbity clashscore tends to be better for the Rosetta-based structures (Figure 1.2A). Given the fact that the latter structures tend to be the most dissimilar from the reference, it appears that the geometric and stereochemical quality of the structures is not a good indicator of their accuracy, as defined above (Fig. 1.2 and Table 1.2). The geometric and stereochemical quality of the structures is largely determined by the algorithm and the force field used in the structure refinement step. This can be appreciated also by comparing the scores of the various NOESY-based results, which can vary appreciably even when for structures closely similar to the reference. The importance of force fields is due partly to the fact that NMR data cannot define parameters such as bond lengths or bond angles, which however are

often restrained also during X-ray structure determinations. Studies affording a deeper understanding of the effects of structure refinement as a function of the quantity and quality of the NMR data available would be quite useful. Nonetheless, it can be stated that accurate structures should satisfy both stereochemical requirements and the available experimental information.

Goodness-of-fit with the experimental data

A different kind of structure validation assesses the completeness of experimental data and its agreement with the structure. Because it is difficult to compare structures directly to the raw experimental NMR data, these analyses were performed with respect to partially interpreted experimental data, e.g. after peak picking and CS assignment. The DP-score[4] (Figure 1.2C) is a measure of the goodness-of-fit of the unassigned NOESY peak lists to a structure, ranging from 0 to 1. This data-based quality measure featured a significant correlation to structure accuracy (Figure 1.2D). A DP-score cutoff of ≥0.7 allowed the identification of acceptable CASD-NMR structures with a reliability of 94%, based on the available refined peak lists. On the other hand, all structures with an RMSD to the reference larger than 3.0 Å or a GDT_TS score lower than 60% had DP-scores lower than 0.6, except for a single CS-DP-Rosetta structure. For comparison, the DP-score values for the reference structures were in the 0.64-0.90 range. It is important to note that the 0.7 DP-score threshold value was determined using refined peak lists, which might facilitate the discrimination, e.g. by reducing the number of artifact peaks that cannot be accounted for. If automatically peak-picked NOESY lists, which potentially contain a significant amount of artifacts that however cannot be excluded at the outset of a NMR structure determination, were used, presumably the DP-score threshold would be shifted toward lower values. It is interesting to observe that for the AR3436A target, which was previously mentioned as the one for which we

observed the poorest overall performance, the average DP-score was as low as 0.60; for the other targets the range of average DP-scores was 0.72-0.81.

Discussions

On average, the automatically generated and the reference structures are of comparable geometric and stereochemical quality. These quality measures do not correlate with the similarity to the reference structure, as measured by either the backbone RMSD or the GDT_TS score. Indeed, the present data demonstrate that even structures with a significantly wrong fold can feature excellent geometric and stereochemical quality measures. Our findings thus reinforce previous indications that the structure refinement protocol is a major determinant of these parameters[58]. The use of an indicator, the DP-score, quantifying the agreement between the structures and the unassigned NOESY data was useful to discriminate good or problematic structures. The DP-score featured a good correlation with both the backbone RMSD and the GDT_TS score; with the present refined peak lists, a DP-score threshold of 0.7 could be applied to identify accurate structures with a 94% precision. Conversely, all structures further than 3.0 Å from the reference had a DP-score lower than 0.6. For the AR3436A target the automated methods obtained the lowest accuracy (Table 1.2) and the poorest convergence (Table 1.4). AR3436A is also the target with the lowest DP-score for the reference structure as well as on average over all CASD-NMR2010 structures. It is possible that the available data did not permit capturing some features of the protein, e.g. related to its dynamics.

For a given target, the various automated NOESY-based methods could yield varying levels of NOESY assignments and, consequently, quite different numbers of structural restraints. Interestingly, this factor did not correlate appreciably with the DP-

score (which refers to the unassigned lists) of the calculated structure nor with its geometric and stereochemical quality, as mentioned above. Overall, we can thus conclude that indicators of agreement with non-interpreted experimental data are useful to validate NMR structures. Geometric and stereochemical parameters are not sufficient to guarantee accuracy; nevertheless they should be taken into account as necessary features of high-quality protein structures; i.e. good structures should have both good agreement with non-interpreted experimental data (e.g. DP-score) and good geometric and stereochemical parameters.

The automated structure calculations addressed in this contribution are unsupervised, with the exclusively NOESY-based methods being typically fast (with calculation times on a single CPU of the order of hours, including refinement) and routine and CS-based methods being relatively CPU-intensive (with estimated calculation times on a single CPU of the order of 10^3-10^4 hours, making it mandatory to employ large clusters or distributed computing for these calculations) and less dependable. A fair criticism to the setup of CASD-NMR2010 is that the NOESY peak lists provided had been refined against initial structural models during the determination of the reference structure and were therefore almost devoid of artifacts. This simplifies the task for NOESY-based approaches and for CS-methods augmented by NOESY data. However, considering their highly satisfactory performance observed here, the peak list refinement may not be necessary if the quality of the NOESY spectra and the completeness of the chemical shift assignments are high. To investigate this, we have initiated a second round of CASD-NMR using new masked NOESY data sets that have been generated using exclusively automated peak-picking procedures. This second round will further consolidate the methodological improvements fostered by the 2010 round.

Conclusions

In summary, the CASD-NMR 2010 initiative has successfully proven, without the possible bias inherent in test calculations of targets with previously known structure, that, given almost complete CS assignments, the automated calculation of NMR structures of small proteins from “clean”, unassigned NOESY peak lists is routinely feasible. NOESY-based methods yield structures that are typically within 2.0 Å of the corresponding manually solved structures and within 2.5 Å in all but one of the 49 cases reported here. This conclusion is also supported by the good convergence of these algorithms, which is within 3.0 Å for all targets and within 2.0 Å for eight targets out of ten. Comparison with the crystal structures of homologous proteins, limited to the Pgr122A and VpR247 targets, provided similar conclusions.

Another notable result of the present investigation is that whereas the performance of methods for NMR structure determination based only on CS data is not yet fully reliable, augmenting these methods with different schemes to exploit unassigned (refined) NOESY peak lists recovers to a significant extent the robustness of the NOESY-based methods, as judged both by similarity to the manually solved structures and by looking at the convergence of the various methods. For the size range addressed by our target selection (up to 150 amino acids), the protein size does not impact significantly on the success rate of the approaches that include NOESY data.

Table 1.1. Features of the programs used in CASD-NMR2010. Y indicates this type of information is directly used in structure calculations, **s** indicates that it is used as a support to derive additional restraints for refinement and/or to improve scoring. Details are given in the Methods section.

Software	NOEs	Chemical shifts*	Comments
CYANA	Y	s	Includes torsion angle restraints generated on the basis of the chemical shift values
UNIO	Y	s	Includes torsion angle restraints generated on the basis of the chemical shift values
ARIA	Y	s	Includes torsion angle restraints generated on the basis of the chemical shift values
ASDP	Y	s	Includes torsion angle restraints generated on the basis of the chemical shift values; uses the DP-score[4] measure to re-rank the structural models
Cheshire-Yapp	s	Y	Uses structural models initially generated using only CS data to assign NOEs, derive distance restraints and refine the best-scoring initial 100 models
CS-DP-Rosetta	s	Y	Uses the unassigned NOESY peak lists and the DP-score[4] measure to re-rank the structural models
Cheshire		Y	
CS-Rosetta		Y	

* Used as direct structural restraints, rather than to derive secondary structure information or torsion angle restraints

Table 1.2. Targets for CASD-NMR and overview of the accuracy of the various approaches.

Target	PDB Code	Sequence length	Average pairwise RMSD within the reference (Å)*	Backbone RMSD (Å) *,+ / GDT_TS * score (%) to the reference structure							
				CYANA	UNIO	ARIA	ASDP	Cheshire-Yapp	CS-DP-Rosetta	Cheshire	CS-Rosetta
VpR247	2KIF	106	0.7	0.8 / 91	0.9 / 92	2.7 / 71§	1.8 / 81	n.a.	1.4 / 78	1.7 / 78	14.6 / 43
AR3436A	2KJ6	97	1.4	2.0 / 65	2.2 / 61	n.a.	1.4 / 66	n.a.	3.3 / 55	4.5 / 56	3.3 / 47
HR5537A	2KK1	135	1.0	1.3 / 89	1.6 / 83	2.4 / 76§	1.7 / 84	n.a.	1.6 / 86	2.1 / 77	2.2 / 76
ET109A((reduced)	2KKX	102	0.6	1.2 / 90	1.7 / 85	1.5 / 87§	1.4 / 90	1.5 / 86	2.0 / 82	n.a.	4.2 / 58
ET109A (oxidized)	2KKY	102	0.6	0.9/ 92	1.1 / 90	1.2 / 89§	1.0 / 91	n.a.	1.6 / 84	n.a.	14.3 / 30
AtT13	2KNR	121	0.6	1.9 / 85	1.7 / 91	2.5 / 84§	2.1 / 84	n.a.	6.8 / 65	n.a.	11.2 / 32
PgR122A	2KMM	73	0.7	1.1 / 85	1.0 / 87	1.6 / 74§	1.0 / 86	n.a.	0.9 / 88	1.1 / 87	1.3 / 83
NeR103A	2KPM	105	1.7	1.0 / 86	0.9 / 89	1.0 / 86#	1.6 / 80	1.5 / 78	1.4 / 81	n.a.	2.8 / 62
CgR26A	2KPT	148	1.6	0.8 / 94	0.8 / 94	0.5 / 87#	1.0 / 93	0.8 / 97	2.6 / 78	n.a.	4.0 / 62
CtR69A	2KRU	63	0.4	0.6 / 92	0.9 / 86	0.6 / 90#	0.7 / 89	n.a.	0.6 / 90	1.2 / 79	1.0 / 83
Number of submitted targets				10	10	9	10	3	10	5	10
Number of successful targets				10	9	7	10	3	7	3	2

* For the backbone atoms of ordered residues, as defined by PSVS using dihedral angle order parameters

+ Backbone RMSD between the average conformer of each structure and the average conformer of the reference structure

§ Determined with the ARIA-Soft protocol

Determined with the ARIA-BayW protocol

Table 1.3. Performance measures and quality scores for all CASD-NMR and reference structures. The column "Successful" reports YES when the condition RMSD ≤ 2.0Å or GDT_TS ≥ 80 is met (see also Table 1), NO otherwise; reference structures are labeled as "Manual". The DP-score ranges from 0.0 to 1.0; it has not been calculated for some CHESHIRE submissions consisting of a single conformer. All other scores are given as Z-scores. The targets are ordered by the time of release, from the oldest (VpR247) to the most recent (CtR69A).

Target	Method	RMSD (Å)	GDT-TS (%)	Success-full	DP-score	Verify3D	ProsaII	Procheck (phi-psi)	Procheck (all)	MolProbity Clashscore	ROG-score ()	Distance/angle restraints	Information (bits/atom)
VpR247	ARIA **(ARIA-Soft)**	2.7	71	NO	0.564	-2.09	-1.2	-7.4	-10.17	-23.34	68/16/17	1630/0	0.23
VpR247	ASDP	1.8	81	YES	0.837	0.48	0.66	0	-0.35	-2.6	12/16/72	2644/133	0.10
VpR247	CHESHIRE	1.7	78	YES	N/A	0.8	0.45	-0.67	-1.6	-0.03	18/31/52	N/A	N/A
VpR247	CS-DP-ROSETTA	1.4	78	YES	0.622	-0.32	1.16	0.87	1.83	0.9	4/4/92	N/A	N/A
VpR247	CS-ROSETTA	14.6	43	NO	0.588	-1.12	0.37	1.34	2.42	0.34	1/1/98	N/A	N/A
VpR247	CYANA	0.8	91	YES	0.849	0.48	0.79	-0.75	-1.66	1.1	7/29/64	2582/0	0.54
VpR247	Manual	N/A	N/A	Manual	0.841	0.32	1.08	-0.2	-0.12	-1.72	8/16/76	2583/125	0.31
VpR247	UNIO	0.9	92	YES	0.837	0	0.87	-1.06	-2.78	0.3	22/29/49	2159/269	0.51
AR3436A	ASDP	1.4	66	YES	0.688	-4.33	-1.86	-0.94	-2.54	-2.36	19/23/59	1045/105	0.076
AR3436A	CHESHIRE	4.5	56	NO	N/A	-0.32	-0.54	-2.99	-2.9	-0.28	36/34/30	N/A	N/A
AR3436A	CS-DP-ROSETTA	3.3	55	NO	0.405	-0.96	0.33	-0.75	0.24	0.8	10/16/74	N/A	N/A
AR3436A	CS-ROSETTA	3.3	47	NO	0.562	-1.12	-0.29	0.08	1.01	0.53	3/6/91	N/A	N/A
AR3436A	CYANA	2.0	65	YES	0.662	-4.01	-1.53	-1.3	-3.31	1.2	14/29/57	793/142	0.27
AR3436A	Manual	N/A	N/A	Manual	0.641	-4.49	-1.61	-1.69	-1.89	-1.13	14/22/64	917/94	0.061
AR3436A	UNIO	2.2	61	NO	0.664	-4.17	-2.11	-2.44	-4.44	0.84	45/31/24	657/258	0.22

HR5537A	ARIA **(ARIA-Soft)**	2.37	76	NO	0.757	-2.25	-0.04	-2.08	-4.85	-26.38	39/44/17	2340/168	0.32
HR5537A	ASDP	1.75	84	YES	0.815	-3.21	-0.41	1.97	0.41	-1.24	15/21/64	2734/182	0.64
HR5537A	CHESHIRE	2.12	77	NO	0.692	-1.61	0.91	1.93	1.42	0.58	4/15/81	N/A	N/A
HR5537A	CS-DP-ROSETTA	1.65	86	YES	0.748	-1.44	0.91	2.79	3.55	1.02	2/2/96	N/A	N/A
HR5537A	CS-ROSETTA	2.17	76	NO	0.677	-0.8	0.91	2.83	3.19	0.7	4/8/88	N/A	N/A
HR5537A	CYANA	1.29	89	YES	0.811	-2.41	-0.5	0.87	-1.66	-0.23	24/27/48	4082/170	0.56
HR5537A	Manual	N/A	N/A	Manual	0.807	-2.41	-0.37	1.46	0.65	-1.62	16/17/67	4133/172	0.36
HR5537A	UNIO	1.59	83	YES	0.819	-2.73	-0.79	-0.87	-3.67	-0.81	36/38/27	3151/349	0.26
ET109Ared	ARIA **(ARIA-Soft)**	1.47	87	YES	0.757	-0.64	-0.12	-2.12	-4.85	-37.46	39/41/20	2521/152	0.78
ET109Ared	ASDP	1.4	90	YES	0.82	-0.64	0.37	-0.63	-1.36	-2.42	9/23/69	2177/170	0.87
ET109Ared	CHESHIRE-YAPP	1.54	86	YES	0.809	-0.8	0.17	-4.17	-6.03	-1.49	39/39/22	N/A	N/A
ET109Ared	CS-DP-ROSETTA	2.02	82	YES	0.749	-0.64	0.66	-0.12	1.06	1.12	3/8/89	N/A	N/A
ET109Ared	CS-ROSETTA	4.2	58	NO	0.583	-2.41	-0.25	0.28	1.48	0.32	8/18/75	N/A	N/A
ET109Ared	CYANA	1.19	90	YES	0.809	-0.8	0.37	-1.22	-2.96	0.23	24/37/39	3168/174	0.86
ET109Ared	Manual	N/A	N/A	Manual	0.797	-0.48	0.29	-0.94	-1.36	-1.44	15/27/58	3024/164	0.31
ET109Ared	UNIO	1.71	85	YES	0.816	-1.28	0.08	-3.66	-6.21	-3.94	41/30/28	2574/268	0.58
ET109Aoxi	ARIA **(ARIA-Soft)**	1.2	89	YES	0.795	0.96	-0.29	-2.71	-5.14	-40.17	43/39/20	2670/152	0.74
ET109Aoxi	ASDP	1.01	91	YES	0.823	-0.16	0.66	-0.71	-1.48	-2.99	11/25/64	2332/173	0.93
ET109Aoxi	CS-DP-ROSETTA	1.6	84	YES	0.755	-0.8	0.7	0	1.24	1.13	2/7/91	N/A	N/A
ET109Aoxi	CS-ROSETTA	14.29	30	NO	0.445	-2.09	-0.87	0.12	1.6	-0.22	11/26/63	N/A	N/A
ET109Aoxi	CYANA	0.87	92	YES	0.816	-0.48	0.45	-1.14	-3.08	-0.87	26/38/35	3318/174	0.74

ET109Aoxi	Manual	N/A	N/A	Manual	0.809	-1.44	0.74	-1.1	-1.6	-2.15	16/30/54	3147/160	0.35
ET109Aoxi	UNIO	1.13	90	YES	0.802	-0.96	0.37	-3.7	-6.33	-6.31	45/29/25	2800/0	0.73
AtT13	ARIA **(ARIA-Soft)**	2.48	84	YES	0.819	-1.28	-1.65	-2.52	-3.73	-12.65	31/31/38	5024/158	0.48
AtT13	ASDP	2.11	84	YES	0.816	-1.12	-1.28	-0.9	-2.01	-2.92	14/25/61	2827/204	0.63
AtT13	CS-DP-ROSETTA	6.77	65	NO	0.681	-2.41	-1.74	-0.43	0.53	0.98	4/10/86	N/A	N/A
AtT13	CS-ROSETTA	11.21	32	NO	0.531	-4.33	-2.23	0.79	1.54	0.79	3/5/92	N/A	N/A
AtT13	CYANA	1.9	85	YES	0.816	-0.96	-1.08	-1.73	-4.14	-0.81	30/25/45	4149/208	0.69
AtT13	Manual	N/A	N/A	Manual	0.825	-1.12	-0.99	-1.57	-2.9	-1.44	21/23/56	4062/138	0.72
AtT13	UNIO	1.75	91	YES	0.844	-1.28	-1.12	-2.4	-4.49	-0.44	34/34/32	3262/319	0.52
PgR122A	ARIA **(ARIA-Soft)**	1.65	74	YES	0.747	-2.73	-1.16	-1.85	-3.61	-19.69	42/36/22	1515/112	0.72
PgR122A	ASDP	1.05	86	YES	0.801	-3.37	-1.41	-0.51	-1.3	-1.83	5/12/82	1483/118	1.02
PgR122A	CHESHIRE	1.1	87	YES	0.744	-3.05	-1.32	-2.08	-2.19	0.56	10/18/64	N/A	N/A
PgR122A	CS-DP-ROSETTA	0.9	88	YES	0.785	-2.09	-0.87	-0.12	1.3	0.91	0/5/95	N/A	N/A
PgR122A	CS-ROSETTA	1.32	83	YES	0.689	-2.41	-0.29	-0.12	1.3	0.79	0/4/96	N/A	N/A
PgR122A	CYANA	1.09	85	YES	0.814	-3.37	-1.53	-0.87	-2.25	0.07	12/27/60	1950/108	0.81
PgR122A	Manual	N/A	N/A	Manual	0.798	-3.21	-1.99	0.12	0.41	-1.52	7/8/85	1730/78	0.50
PgR122A	UNIO	0.98	87	YES	0.807	-2.73	-1.32	-2.24	-3.67	0.15	32/29/40	1556/194	0.75
NeR103A	ARIA **(ARIA-BayW)**	1	86	YES	0.771	0.32	-0.91	-0.08	-1.3	-1.07	17/21/62	1563/128	0.47
NeR103A	ASDP	1.6	78	YES	0.783	-2.41	-1.24	-0.08	-1.42	-1.61	16/25/59	1750/146	0.45
NeR103A	CHESHIRE-YAPP	1.52	78	YES	0.745	-2.09	-0.74	-3.11	-6.15	0.29	53/21/26	N/A	N/A
NeR103A	CS-DP-ROSETTA	1.37	81	YES	0.776	-1.44	-0.45	0.2	1.24	0.98	0/5/95	N/A	N/A

NeR103A	CS-ROSETTA	2.84	62	NO	0.705	-1.44	-1.03	0.28	1.12	0.34	7/7/85	N/A	N/A
NeR103A	CYANA	0.96	86	YES	0.789	-1.77	-0.79	-0.24	-2.6	0.97	20/29/51	2176/138	0.65
NeR103A	Manual	N/A	N/A	Manual	0.788	-1.93	-1.08	-0.31	0.06	-0.18	8/18/74	2093/258	0.40
NeR103A	UNIO	0.93	89	YES	0.788	-2.25	-0.95	-2.64	-5.62	-2.48	55/19/26	1622/0	0.45
CgR26A	ARIA **(ARIA-BayW)**	0.48	97	YES	0.87	0.32	1.49	1.06	0.95	-1.36	8/10/82	2393/195	0.44
CgR26A	ASDP	1	93	YES	0.871	-1.77	0.04	0.43	-0.3	-1.36	11/13/76	2484/234	0.36
CgR26A	CHESHIRE-YAPP	0.82	95	YES	0.865	-0.48	0.41	-0.87	-2.72	0.83	30/31/39	N/A	N/A
CgR26A	CS-DP-ROSETTA	2.56	75	NO	0.683	-0.64	0.83	0.63	1.36	1.02	2/2/96	N/A	N/A
CgR26A	CS-ROSETTA	4.02	62	NO	0.603	-1.12	0.87	1.3	2.25	0.95	2/4/94	N/A	N/A
CgR26A	CYANA	0.77	94	YES	0.875	-1.77	-0.21	0.2	-1.06	1.31	15/20/65	2954/208	0.53
CgR26A	Manual	N/A	N/A	Manual	0.903	-1.44	0	0.55	0.3	-1.34	11/14/74	2819/146	0.57
CgR26A	UNIO	0.77	94	YES	0.892	-1.44	-0.08	-1.65	-3.61	-3.16	32/34/34	2448/0	0.41
CtR69A	ARIA **(ARIA-BayW)**	0.61	90	YES	0.793	-1.77	-0.37	1.53	1.83	-0.48	6/16/78	1300/86	0.66
CtR69A	ASDP	0.73	89	YES	0.817	-2.73	-0.79	1.49	0.77	-0.71	3/17/79	1162/100	0.12
CtR69A	CHESHIRE	1.2	79	YES	0.697	-2.09	0.25	1.22	0.53	0.61	12/12/76	N/A	N/A
CtR69A	CS-DP-ROSETTA	0.6	90	YES	0.78	-1.44	0.37	1.97	3.19	1.34	5/4/91	N/A	N/A
CtR69A	CS-ROSETTA	0.96	83	YES	0.743	-1.28	0.62	2.05	3.02	1.11	2/8/90	N/A	N/A
CtR69A	CYANA	0.58	92	YES	0.809	-2.41	-0.83	1.22	0.12	1.22	11/17/71	829/98	0.56
CtR69A	Manual	N/A	N/A	Manual	0.809	-2.57	-0.45	1.61	1.48	-1.32	6/22/71	1013/70	0.71
CtR69A	UNIO	0.92	86	YES	0.796	-2.89	-1.36	-0.16	-1.95	1.24	24/29/48	674/0	0.65

Table 1.4. Convergence of the automated structure calculation methods. The convergence of the structures has been calculated as the average pairwise RMSD among the mean conformers of the bundles generated with the selected methods.

Target name	NOESY-based methods (Å)	NOESY-based methods + CS-based methods using NOESY data (Å)	All methods (Å)
VpR247	1.81	1.81	3.92
AR3436A	2.97	3.13	3.62
HR5537A	1.77	1.81	1.90
ET109A (reduced)	1.07	1.32	1.37
ET109A (oxidized)	0.95	1.24	1.57
AtT13	2.64	3.31	6.17
PgR122A	0.96	0.98	1.12
NeR103A	1.24	1.36	1.70
CgR26A	0.93	1.26	1.85
CtR69A	0.92	0.91	1.12

Table 1.5. RMSD (Å) of automatically generated structures to homology models of the PgR122 and VPr247 targets

PDB	CYANA	UNIO	ARIA	ASDP	CS-DP-Rosetta	Cheshire	CS-Rosetta
PgR122							
0.77	0.89	0.61	1.36	0.99	0.86	0.82	1.19
VpR247 (Based on 3GVA, chain A)							
2.07	1.74	2.02	3.61	1.78	2.19	2.03	14.9
VpR247 (Based on 3GX4, chain X)							
1.43	1.37	1.40	3.21	1.98	1.94	1.66	15.1

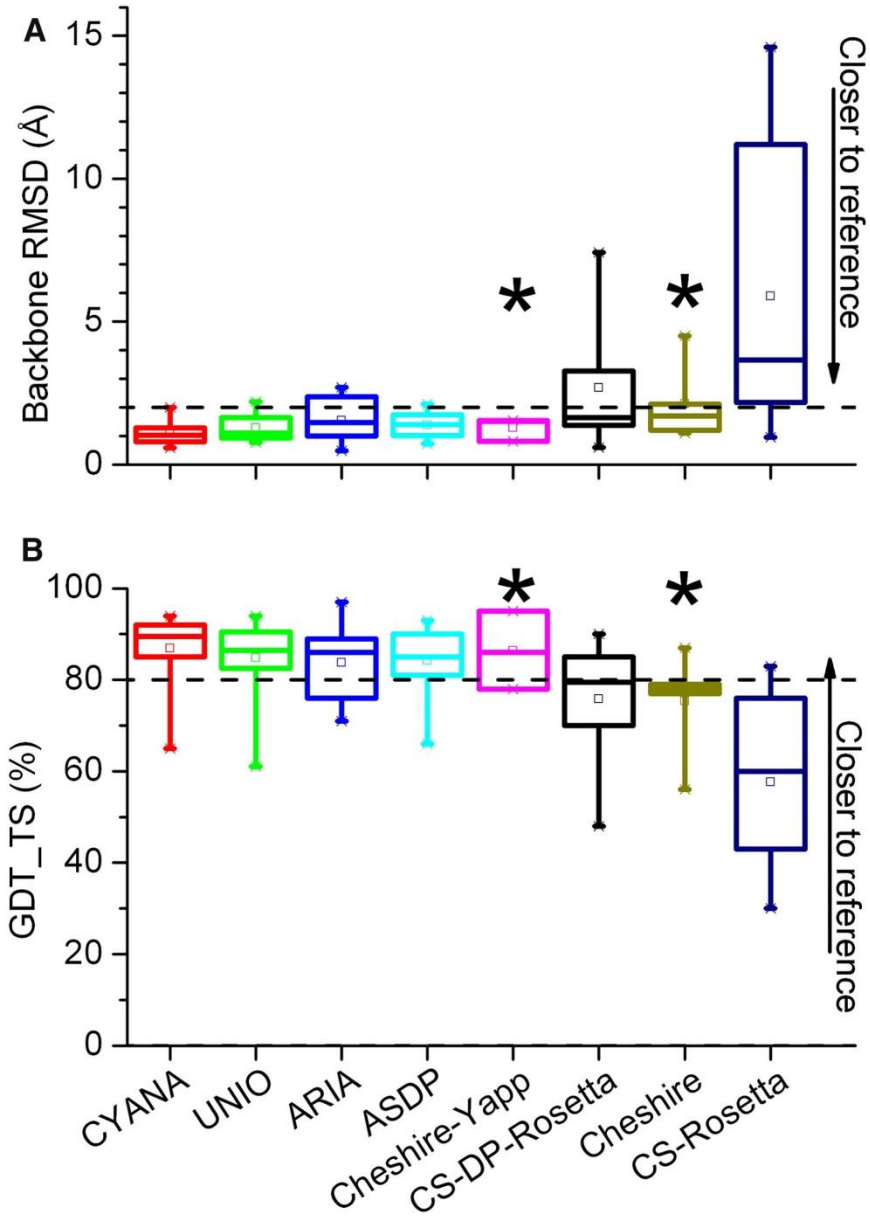

Figure 1.1. Structural similarity between reference and CASD-NMR2010 structures. RMSD (A) and GDT_TS score (B) deviation of the backbone coordinates (for ordered residues only) with respect to the reference structure for the various algorithms. The box parameters are as follows: the box range goes from the first to the third quartile; box whiskers identify the minimum and maximum values; the square within the box identifies the mean; the thick line in the box identifies the median.The starred boxes correspond to algorithms for which less than 60% of the targets were submitted.

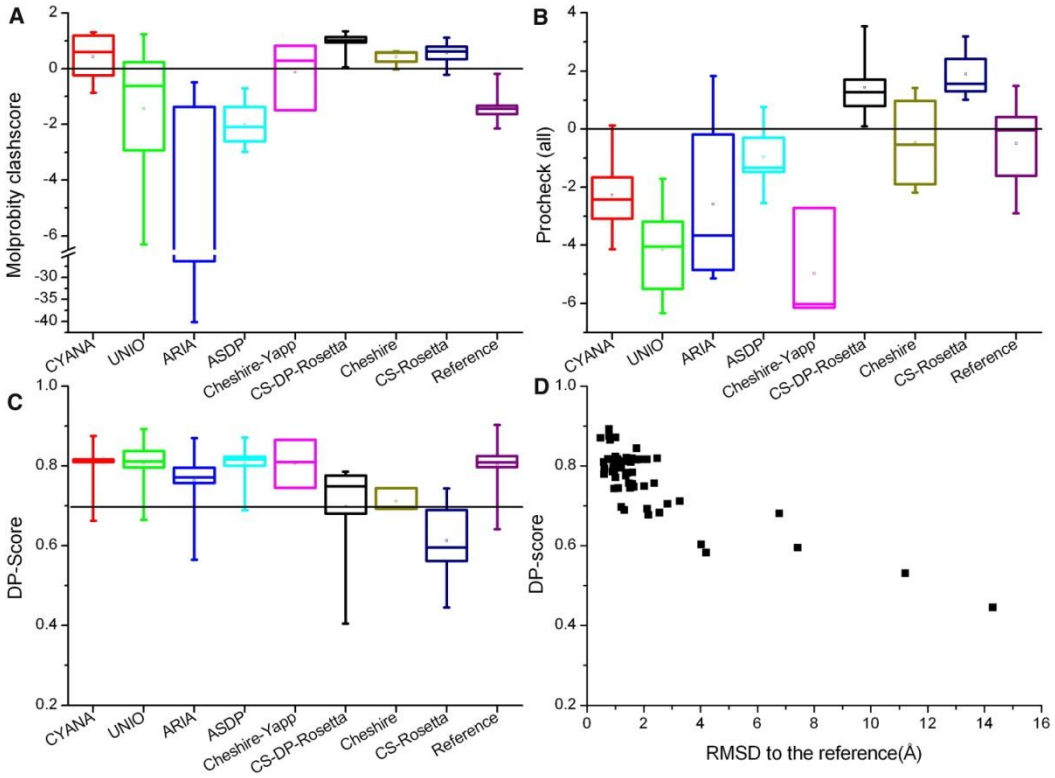

Figure 1.2. Quality of CASD-NMR2010 structures.

Molprobity (A) and Procheck-all (B) Z-score values describe the distribution of, respectively, all protein dihedral angles and high energy interatomic contacts for the automatically generated and the reference structures. The Z-score is the deviation of the value calculated for a given structure from the average calculated for a set of 150 high-resolution X-ray structures, expressed in units of the standard deviation. A positive Z-score indicates that the corresponding structure quality score is better than the average, whereas a negative value indicates that the structure analyzed is worse than the average. DP-Scores (C) describe the agreement between the structures and the *unassigned* NOESY peak lists, and range from 0 (worst) to 1 (best). The dashed line corresponds to the 0.7 threshold described in the main text. The box parameters are as in Figure 1. Panel (D) reports DP-scores as a function of the backbone RMSD to the reference structure, for all CASD-NMR2010 structures.

Chapter 2

Accurate Automated Protein NMR Structure Determination Using Unassigned NOESY Data

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[49], CYANA[59], or AutoStructure[44] 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[3]. 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 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 Auto-Structure 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). The statistics for the peaklists used in the study are reported in Table 2.1.

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-Rosetta 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 program[4] 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 DPscore is correlated with the accuracy of the model 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 refine[60] 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-R 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 closure[61] followed by all-atom refinement of the entire structure using the physically realistic Rosetta forcefield[62]. 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 (AssignNOERosetta 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.

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 2.1. Using a refined NOESY peak list produced with expert curation of the raw peak list, CYANA and AutoStructure generate topologically correct models (Figure 2.1D). 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 2.1D to 2.1E) by minimizing the energy (Figure 2.1A, 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[60,63].

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 2.1B). 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 2.1A′). 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. 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 2.1C) 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 Assign---NOE-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 2.2A). 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 2.1A′). 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[1]. Following the public release of the native structures, we found that our model ensembles agreed well with the native structures, as shown in Figure 2.2. 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 2.2A, 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 2.2B), 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 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 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 Assign-NOE-Rosetta protocol generated models with close to native side chain packing and ~ 1 Å backbone rmsd from the X-ray structure (see Figure 2.3). As shown in columns 2 and 3 of Table 2.2, 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 FindCore[6]) compared to the starting

CYANA/AutoStructure model. 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 findings[60,63](see Table 2.2, 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. 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.

Discussions

The DP filter is a powerful global fold score that can sometimes overcome the lack of convergence for larger proteins using CS-Rosetta alone. 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[64].

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-consuming. 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.

Conclusions

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.

Acknowledgment

We thank Rosetta@home participants and the DOE INCITE program for access to the Blue Gene/P supercomputer 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).

Table 2.1. Details of peak lists used in this study

ID of NMR Deposition			Number of peaks in NOESY lists					
			^{13}C-aliphatic		^{13}C-aromatic		^{15}N	
NESG	PDB	BMRB	*raw*	*refined*	*raw*	*refined*	*raw*	*refined*
Test Datasets								
HR41	2k07	6546	7701	6806	918	1028	2836	2713
SR213	2hfi	16113	7312	3945	154	440	1451	1595
StR65	2jn8	15089	7432	2250	293	254	1840	899
CcR55	2jqn	15281	5801 (4965)[b]	2398 (133)[b]	455	223	2995	1096
SsR10	2q00	15265		3342		202		1588
Blind Datasets								
VpR247	2kif	16272		3900		419		1437
AR3436A	2kj6	16313		1402*		97		577
HR4394C	2kk0	16348	10172		246		2081	

*peaks folded in the 13C dimension are present (+/- 24ppm sweep width). The peaks in the 13C dimension of the aliphatic NOESY are match-tested by AutoStructure NOESY assign routine.

[b]In parenthesis peaks from aliphatic NOESY spectrum after exchange into 100% D2O solvent.

Table 2.2. 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[6].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 protein names are NESG target id's; detailed protein sequence data for these targets are available from the SPINE database[65,66]

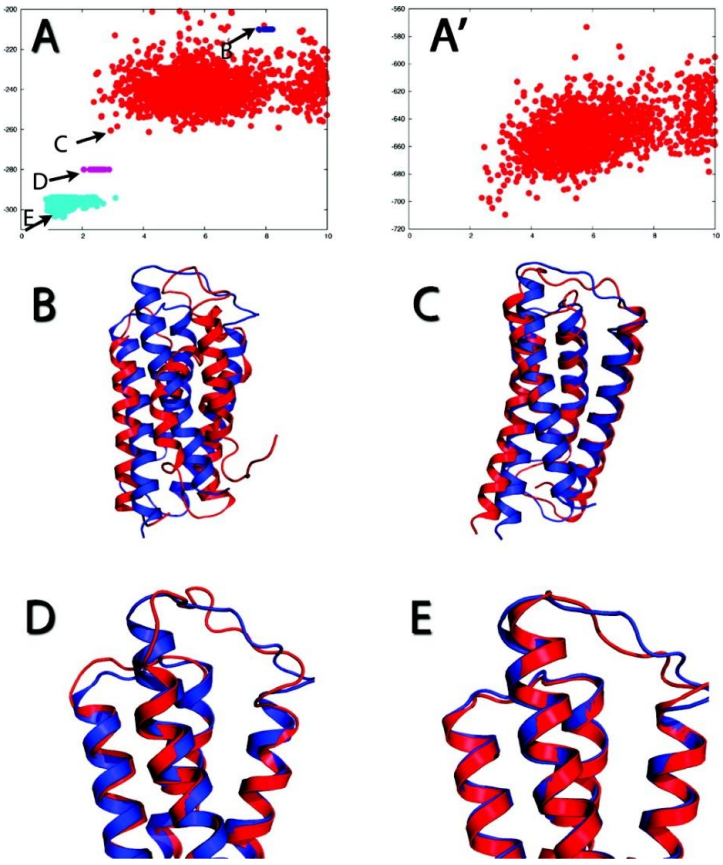

Figure 2.1. Model generation from raw and refined peak lists with CYANA/AutoStructure and Rosetta for protein SR213. (A) Rosetta all-atom energy vs rmsd to the X-ray structure. Dark blue points are CYANA/AutoStructure models from raw peak lists with energy set to arbitrary value. Red points are Rosetta 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.

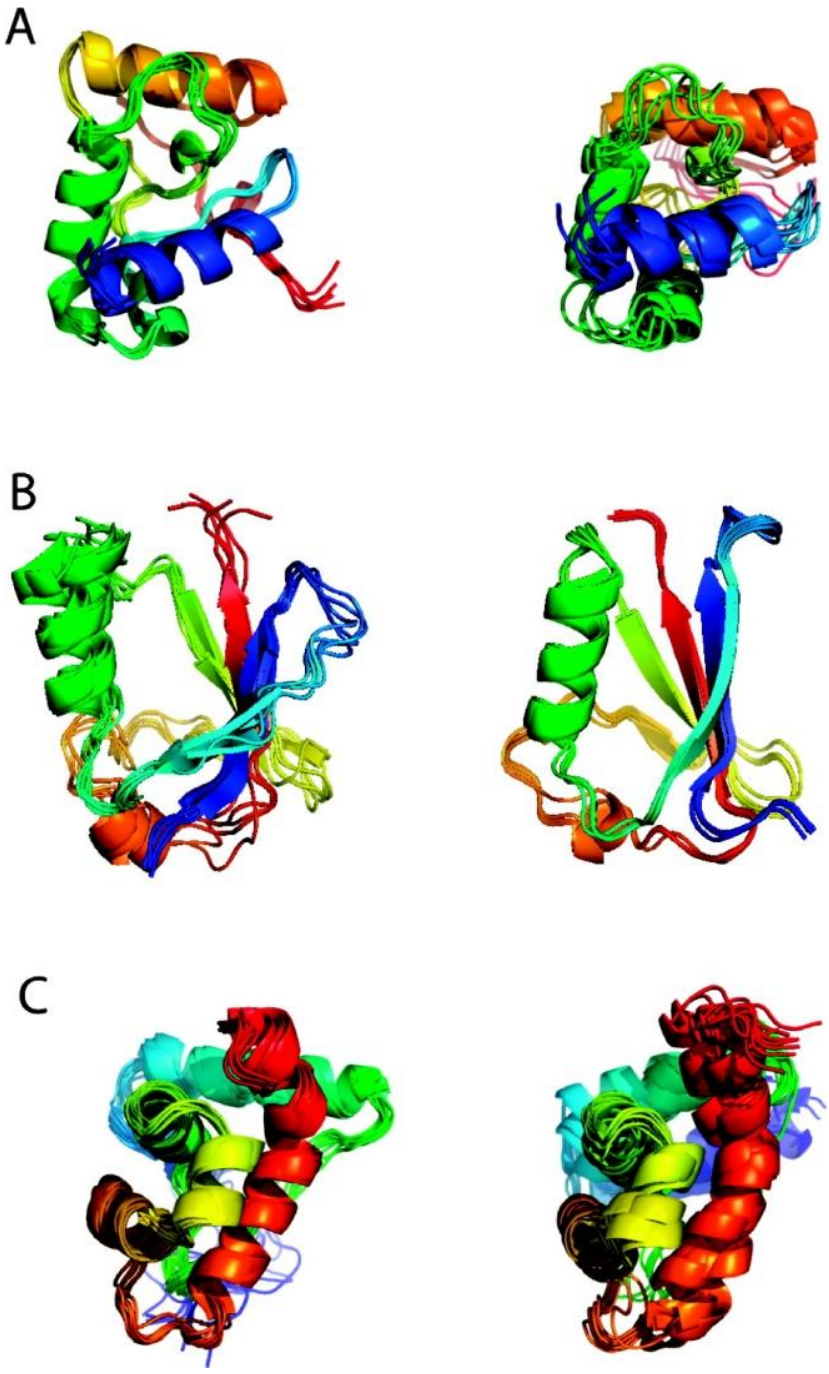

Figure 2.2. Blind structure determinations with CS-DP-Rosetta protocol (A) VpR247, (B) AR3436A, (C) HR4394C. (Left) Experimentally solved NMR ensemble. (Right) Ensemble of lowest energy structures by the CS-DP-Rosetta protocol. Refined peak lists were used for VpR247 and AR3436A; raw peak lists were employed for HR4394C.

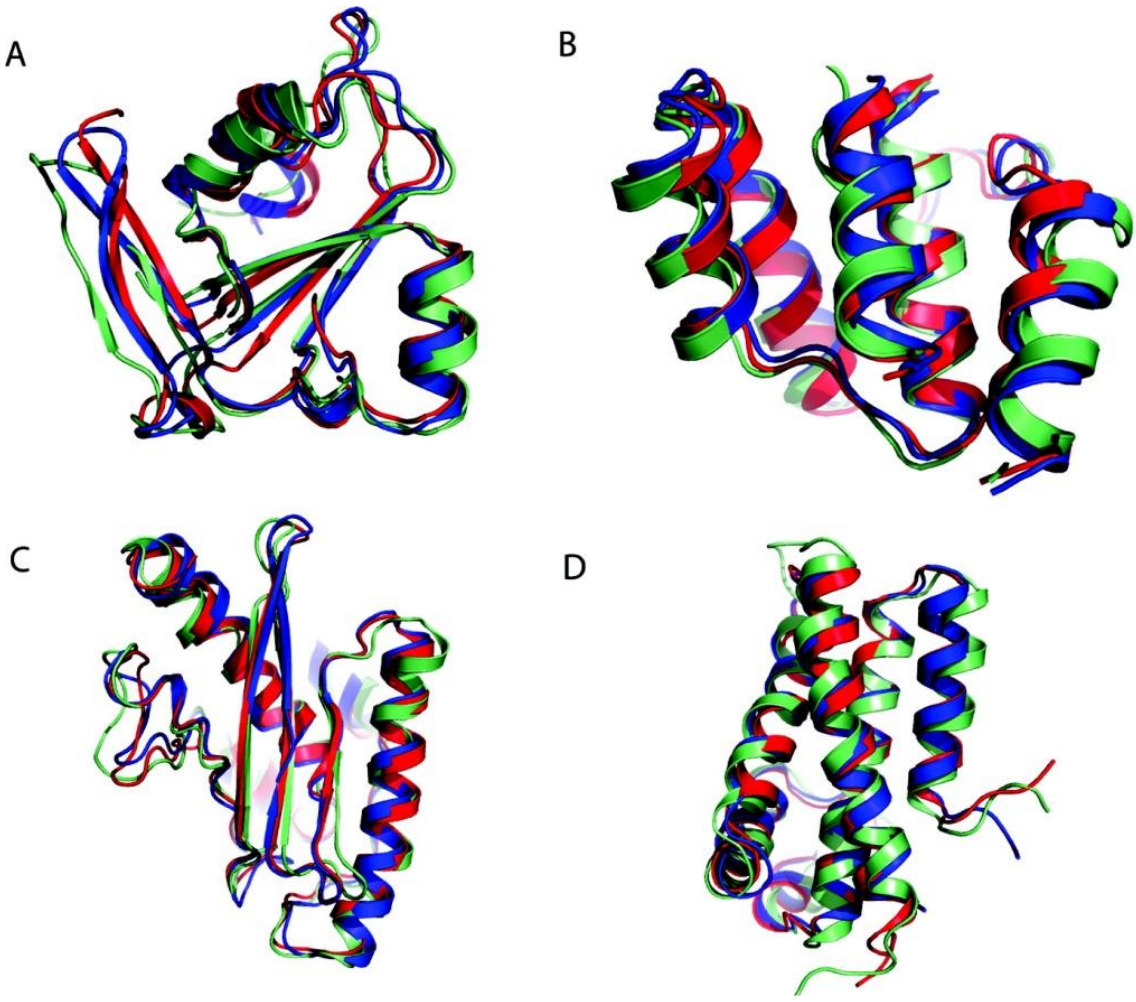

Figure 2.3. Superposition of AssignNOE-Rosetta models to the X-ray structures

Superposition of 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.

Chapter 3

Improved Technologies Now Routinely Provide Protein NMR Structures Useful for Molecular Replacement

Introduction

One of the most critical stages in the process of determining the crystal structure of a protein involves estimating the phases of X-ray diffraction data. There are several ways to address this phase problem, including direct methods[67], multi-wavelength or single-wavelength anomalous diffraction (MAD or SAD)[68,69], multiple or single isomorphous replacement (MIR or SIR)[70,71,72], molecular replacement (MR)[73,74], and/or a combination of these methods. Molecular replacement, first described by Rossmann and Blow[75], involves estimating the initial phases of diffraction data based on a known similar structure. In comparison to the experimental phase determination techniques, molecular replacement has the advantage of not requiring preparation of heavy atom derivatives, hence can be cost and time effective. In recent years, around 70 percent of deposited macromolecular structures have been solved by molecular replacement[76]. Additionally, both the number of structures deposited in PDB and the coverage of structure space are increasing rapidly[77,78,79]. These data, in combination with advances in homology modeling[80,81,82,83] and MR programs, make molecular replacement an increasingly important approach to the phase problem in protein X-ray crystallography.

In principle, given an accurate search model for a target protein structure, MR is quite straightforward. However, it can sometimes be very difficult to get a correct MR solution due to the enormous search space. Therefore, for successful MR phasing, it is

critical to effectively prepare the initial search model so as to maximize its signal/noise ratio, and to enhance the signal detection capabilities of MR algorithms by finding an optimal target function and effective search strategy that can identify correct solutions. Significant efforts have been made to develop and improve both of these aspects in the last two decades. A number of protocols to prepare MR search model have been proposed. These are generally designed to exclude structurally-disordered regions (e.g. by truncating long flexible side chains) or to incorporate structural flexibility information into search models by using a composite search model[84,85,86] or pseudo B-factors[87,88,89]. Armed with more accurate target functions, more advanced mathematical models and more effective search strategies, a number of software packages have been developed which have greatly improved the effectiveness of the MR approach, such as COMO[90], XPLOR/CNS[23], AMoRe[91], MOLREP[92],EPMR[93], Queen of Spades[94], SoMoRe[95], MrBUMP[96], Phaser[97],and others.

Nuclear magnetic resonance (NMR) is a powerful tool to determine protein structures in solution and in the solid state. Solution NMR methods have contributed a substantial fraction of the structures deposited in the Protein Data Bank (PDB)[13]. In 1987, Brunger et al. showed that solution NMR structures could be employed as search models for MR[98]. Since this early work, quite a few successful cases using NMR structures for MR have been published (for a useful review of this progress, see Chen et al., 2000[99]). However, a common notion in the structural biology community is that the quality of NMR structure is often not good enough for MR, even when the sequence of the search model is identical to the target X-ray structure. There are various explanations for this observation. Some NMR structures, or parts of the structure, may be under-constrained due to insufficient data; in other cases, there may be genuine differences between structures in solution and in the crystal. Chen et. al have demonstrated, based on a few individual successful cases reported in previous literature,

that success rate of using NMR structures in MR can be significantly improved by carefully preparing the initial search[99,100,101]. However, in most studies only the successful examples are reported and, to date, there have been no systematic studies to evaluate the general utility of NMR structures as initial search models for MR.
Over the last 10 years, there have been significant improvements in both phasing algorithms and the NMR structure determination process, particularly in structural genomics projects where state-of-the-art refinement and quality assessment tools are employed. These advances beg the question: given modern technologies for NMR structure determination and refinement, can NMR structures be used routinely as initial search models for molecular replacement? If that is the case, can we define an optimal protocol to prepare NMR structure ensembles as MR search models in order to maximize their phasing power in MR?

The Northeast Structural Genomics Consortium (NESG; www.nesg.org) is one of the large-scale structure production centers of the Protein Structure Initiative (PSI). NESG has contributed more than 400 NMR structures, as well as some 600 X-ray crystal structures, to the Protein Data bank (PDB) over the past ten years, representing a large fraction of the NMR structures deposited into the PDB by the PSI. The NESG Consortium, involving several NMR groups, has focused efforts on improving the efficiency and accuracy of its NMR structure determination pipeline, and has implemented strict quality control measures to ensure the production of high quality structures[4,35,102]. Although most NESG structures have been solved by either NMR or X-ray crystallography, as of December 2009 the NESG consortium had solved 27 pairs of protein structures for identical construct sequences using both X-ray crystallography and NMR methods. These 3D structures of proteins with identical sequences, together with the raw NMR and crystallography data available in the BioMagRes[103] and PDB, are an extremely valuable composite dataset for understanding structural variations between

solution and crystal states, providing insights into protein dynamics and the effects of lattice packing in selecting conformations from solution, and for new methods development.

Model preparation is a cornerstone of many successful molecular replacement trials, given the fact that every atom in the search model contributes in MR analysis. In particular, it is critical to estimate structural variability in order to decide which portion of structure should be kept in the search model. There are alternative ways to assess the precision of a NMR structural ensemble, including RMSD (the root-mean-square-deviations from the average model), dihedral angle circular variance or order parameters[104], and inter-atomic variance matrices[6,105]. RMSD statistics depend on details of how the structural ensemble is superimposed. Dihedral angle order parameters are good estimators of local structural uncertainty, but generally do not provide a good measure of global consistency. Methods based on the inter-atomic variance matrix can identify one or more sets of “core atoms” whose positions are well defined with respect to one another. The FindCore algorithm[6] uses the inter-atomic variance matrix to define an “order parameter” for each atom, then identifies sets of “core atoms” using hierarchical clustering methods with an empirically-motivated stopping rule based on Chauvenet’s criterion for outlier detection. In some cases it partitions the protein structure into “multiple cores”, each of which is well-defined internally but exhibit structural variation between “cores”. The FindCore algorithm thus allows identification of the well-defined regions (i.e. groups of atoms) of the protein structure from the ensemble of NMR structures without the assumptions involved in generating a molecular superimposition.

We have used 25 NESG NMR/X-ray crystal structure pairs in a systematic investigation of the utility of NMR structures as initial search models for molecular replacement. Staring from NMR ensembles prepared by an improved protocol, FindCore,

we obtained correct MR solutions for 22 of 25 targets. The NMR ensembles for two (2) additional proteins could also be used successfully for MR following Rosetta refinement. Based on these solutions, automatic model rebuilding could also be successfully done with high sequence completeness and model accuracy. We also demonstrate that these NMR structure ensembles can be used successfully as MR search models for homologous target X-ray structures, given sequence coverage and sequence identity of NMR structures to X-ray structures no less than 70% and 40% respectively. These studies indicate the high quality of the NMR structures that are being generated by structural genomics projects using routine modern NMR methods, and demonstrate that the FindCore protocol generally provides high success rates using NMR ensembles for phasing by MR.

Methods

Data acquisition and preprocessing

The coordinates files of NMR structures and the structure factor files of X-ray structures were downloaded from PDB directly. The structure factor files, downloaded in mmCIF format, were converted to mtz format using the *CCP4* program *CIF*2*MTZ*((Collaborative Computational Project, Number 4, 1994). Another *CCP4* program *uniquefy* was used to standardize the mtz files and select reflections for free R calculation.

MR Search model preparation

For each NMR ensemble, eight different search models were prepared with various levels of simplification as detailed below. These methods are also summarized in

Table 3.3. For all those models, hydrogen atoms were deleted from NMR coordinates files.

1. nh model: A composite model including all the individual models in NMR ensemble and the coordinates of all the non-hydrogen atoms are kept.
2. bsm model: Single NMR model which has the highest structural similarity with X-ray structure.
3. aveB model: Average structure of NMR ensemble with distance based pseudo B-factor[89] ; coordinates of ‘not-well-defined’ residues calculated by the *PSVS* program based on dihedral order parameter values are deleted.
4. AG model: Composite model including all the individual models in NMR ensemble residues with side chains longer than Ala are truncated to Ala. This model is based on the protocol as defined in the script *multiprobe* (ftp: //X-ray.bmc.uu.se/pub/gerard/omac/ multi_probe) .
5. SAG model: Composite model including all the individual models in NMR ensemble and residues with side chains longer than Ser are truncated to Ser. This model is based on the protocol as defined in the script *multiprobe* (ftp: //X-ray.bmc.uu.se/pub/gerard/omac/ multi_probe) .
6. nd model: Composite model including all the individual models in NMR ensemble for which coordinates of ‘not-well-defined’ residues calculated by *PSVS* program based on dihedral order parameter values are deleted.
7. ndSAG model: Composite model including all the individual models in NMR ensemble. Coordinates of ‘not-well-defined’ residues calculated by *PSVS* program based on dihedral order parameter values are removed, and surface residues with side chains longer than Ser are truncated to Ser .

8. fc model: Composite model with NMR ensemble trimmed by results of *FindCore* analysis. The atomic precision of the NMR structure ensemble was assessed by a pseudo B-factor, which was calculated from a variance distance matrix using the *FindCore* program. Each residue was treated as a tree data structure with backbone atoms (N, C_α, C, O) being defined as the root , and side chain heavy atoms were defined as child nodes and their precedence were determined by their relative distance to C_α; e.g., C_β is the child node of C_α, and C_γ is the child node of C_β. Any nodes together with their child nodes were removed from search model if their pseudo B-factors calculated by *FindCore*, were equal or greater than 60.

MR trials and automatic model building and refinement

The program *Phaser* (version 2.1)[97] was used for molecular replacement. *MR_AUTO* mode was adopted with RMS being set to 1.5. Program *ARP/wARP* version 7.0[106] was used for automatic model building starting from the *Phaser* MR solution. The *ARP/wARP* expert system mode was employed for automatic model building, and *Refmac5*[107] was used in refinement, staring from the positioned search model and a maximum of 10 building cycles were allowed. *Phenix.autobuild*[108] was employed for automatic model rebuilding if *ARP/wARP* failed to generate good quality models. No manual model building was applied to any case, to allow a fair comparison of each MR trials.

We developed a pipeline using Perl script language to run *Phaser* and *ARP/wARP* jobs on a cluster of 128 CPUs in a highly automated manner. TFZ and LLG values were extracted from *Phaser* solutions to assess the quality of MR solutions. The quality of models automatically built by *ARP/wARP* was judged by *R*, *R*-free, and the

completeness of auto-tracing. In addition, structural similarity between *ARP/wARP* models and corresponding X-ray structures were evaluated by GDT-TS score[52].

Coot [109] was used to check the models and electron density maps, after molecular replacement, and after model building in *ARP/wARP*. The *TM-score* program[110] was used to perform structural alignment and GDT-TS calculation.

Rosetta **loop rebuilding and all atom refinement**

The *Robetta* fragment server (http://robetta.bakerlab.org/fragmentsubmit.jsp)[80,111] was used to generate fragment library, based on sequence and chemical shift data of each target protein. Then loop rebuilding and all atom refinement[62,112] was done by *Rosetta* cyclic coordinate descent (CCD) and kinematic closure (KIC) loop modeling application (Version 3.0), '*fastrelax*' mode was used to allow the whole structure to relax in *Rosetta* all-atom force field, and could be 5-10 times faster than normal relaxation mode. For each target protein, loop regions were defined by the consensus of secondary structure, "not-well-defined" residues were identified by the *PSVS* program based on dihedral order parameter values, and non-core residues defined by *FindCore* program. 1000 decoys were generated from each individual model of NMR structure ensemble, and the overall top 20 decoys with the lowest *Rosetta* energy were selected and combined as a composite model to be used in molecular replacement the same way as their NMR counterpart.

Results

22 of 25 NESG NMR structures successfully provide MR solutions

The NESG project uses the *Protein Structure Validation Suite* (*PSVS*)[35] to monitor the quality of structures. Based on a set of 252 high resolution X-ray structures, *PSVS* provides Z scores for a variety of widely adopted structural quality measures, such as *Procheck* G-factor[56]， *Molprobity* clashscore[57], and other structure quality assessment metrics. The analysis aims to provide a multi-criteria estimate of protein structure quality. A time course study of the evolution of various *PSVS* Z scores for NESG NMR structures indicates that the quality of NESG structures has steadily improved over time. For example, significant improvements of knowledge-based stereochemical, geometric, and interatomic packing properties of protein NMR structures over the past few years are illustrated in Figure 3.1. Most of the NMR structures used in this study were solved since 2006 (Table 3.2).

Of 27 NESG NMR / X-ray crystal structure pairs available at the time this study was initiated, two were excluded from this investigation due to the following facts: One target (GR4) was reported as only a single structure, rather than as an ensemble. The NMR structure of target ER382A (PDB id: 2jn0) was solved as a monomer without a ligand, while its crystal structure counterpart (PDB id:3fif) has eight subunits in the asymmetric unit and was solved in complex with a heptapeptide ligand and appears to have a distinct structure; i.e. the C_α-rmsd between the NMR structure and chain A of crystal structure is 2.44 Å.

For each of the remaining 25 structures, MR search models were prepared from the NMR structure ensemble, using eight different methods to define the search models. We obtained definite MR solutions with *Phaser,* which have positive log likelihood gain

(LLG) scores and translation function Z-score (TFZ) scores greater than 8, for 20 of 25 targets. For two additional targets, HR3646E and StR65, although their TFZ scores were relatively low (3.6 and 5.8 respectively), using the MR solutions with the highest TFZ scores, more than half of the residues could be accurately traced by *ARP/wARP* program; this indicates that the MR solutions were actually correct even though the TFZ scores were lower than 8 (see more details below). All together, useful phase information for 22 of 25 X-ray structures could be determined by *Phaser* based on their corresponding NMR structure ensembles (Figure 3.2A, Table 3.2). In addition, for most targets with correct MR solutions and resolution better than 2.5 Å, highly accurate *ARP/wARP* models could be built with great sequence completeness. However, for five targets with definite *Phaser* solutions (TFZ >8), *ARP/wARP* either failed to build any legitimate model (HR41, StR70, PsR293) or eventually generated models with free R value worse than 0.4 (BeR31, SR213). To address these cases, we used *phenix.autobuild* for automatic model rebuilding, which was less sensitive to low resolution X-ray diffraction data. For all five of the targets that failed model building using *ARP/wARP*, we could build models using *phenix.autobuild* with free R factors better than 0.45. The free R factors of some models (HR41, PsR293) were even comparable with the free R factors of the corresponding crystal structures deposited in PDB (Table 3.4). These results are particularly impressive since no manual intervention was used in these analyses. From this study, we conclude that good quality NMR structures, like those solved by the NESG consortium using standard modern NMR methods, are generally of sufficient accuracy to be routinely used as search models in MR.

Structure similarity limit of search models to X-ray structures

A rule of thumb in MR is that a correct MR solution requires a C_α-RMSD between search model and target structure no greater than 1.5 Å over a large fraction of the

molecule. In 2005, Giorgetti *et. Al*[113] demonstrated that the *Global Distance Test* (*GDT*) algorithm provides an even more robust measure to assess the usefulness of protein search model for MR than C_α-RMSD. They concluded that a GDT-TS higher than 0.84 is generally sufficient to guarantee the success of MR procedure, while a GDT-TS lower than 0.80 is essentially never successful in MR trials; GDT-TS values between 0.80 and 0.84 are in the “twilight zone” of mixed success rates. Our analysis confirms the first part of this conclusion. However, for two cases (NESG targets CtR107 and HR3646E), we obtained correct MR solutions using initial search structures with GDT-TS values lower than 0.8. In addition, we had almost perfect success rate of MR trials for targets in the “twilight zone” (Table 3.2).

We are in a better position today to push the limits of the application of MR than five years ago. In particular, recent advances in MR programs such as *Phaser* offer more powerful signal detection and more effective search strategies. In addition, improvements in NMR data analysis and structure refinement methods provide more accurate NMR models, and model uncertainty is better described by the reported NMR structure ensembles.

The *FindCore* protocol provides better search models for MR

The basic problem of preparing NMR search models for MR can be reduced to determining which subset of atoms have highest probability to contribute to signal instead of noise, and assigning appropriate weight to each atom proportional to its S/N ratio. Since it is impossible to know the X-ray structure beforehand without phase information, there is no direct criteria to assess the S/N level of each atom; i.e., the consistency of its relative position between solution and crystal states. However, structurally-ordered regions of the protein, such as atoms buried in the hydrophobic cores, generally have better "phasing power" than disordered residues, such as atoms in

large surface side chains. This conclusion is supported by the work of Chen et al. which demonstrated that phasing power of NMR structure ensemble can be significantly improved by removing structurally-disordered regions and by truncating long side chains to their common bases (C_β or C_γ)[99,100,101]. Ensemble-derived pseudo-B factors or composite models can also improve the phasing power of NMR ensembles as search models[89].

The “dihedral angle order parameter” (S), a measure of dihedral angle circular variance, is one of the most commonly used measures to calculate the ordered region of a protein[104].In our study, the *PSVS* server[35] was used to identify ordered residues with S(phi) + S(psi) ≥ 1.8. Then, the *areaimol* program[114,115] in the *CCP4* software package was used to identify surface exposed residues. As described in methods section and in Supplementary Table 3.3, eight search models were prepared for each target in order to compare their relative performance in MR experiments based on both *Phaser* solutions and *ARP/wARP* model building results. Most of these methods utilize the ensemble of NMR structures, trimmed in various ways, as the search model. We plotted TFZ scores against model preparation protocols for all the targets (Figure 3.2B). TFZ scores of *Phaser* solutions derived using the whole ensemble model (nh) or single (best) NMR conformer (bsm) as the search model were among the lowest. Better TFZ scores could be attained by removing disordered residues (nd, aveB) or by truncating long side chain residues to common base (AG, SAG), but the level of improvement was case specific, and these protocols failed to find optimal MR solutions for some targets. A combination of removing disordered residues and truncating long surface side chains (ndSAG) showed no further significant improvement. TFZ scores of *Phaser* solutions using NMR ensembles trimmed to “core atom sets”, defined by the *FindCore* program (fc) which allows a robust estimate of model uncertainty at an atomic level, were always the

highest or among the highest. Starting from these ‘fc’ MR solutions, more than half of the residues could be accurately built (C_α - rmsd < 1 Å) using *ARP/wARP* for 18 of 19 targets (i.e. except for StR65) which had both correct MR solutions and X-ray diffraction data resolution better than 2.5 Å (Table 3.5). For target StR65, we only obtained a relatively weak solution using the ‘fc’ search model ensemble (TFZ = 5.8), and the quality of *ARP/wARP* model for this target was less satisfying (R-free=0.39 and GDT-TS=0.71). For targets BeR31 and SR213, although their *ARP/wARP* models were close to target X-ray structures, the free R values were relatively poor (> 0.4). In addition, for targets HR41, StR70 and PsR293 with resolution of X-ray diffraction data > 2.50 Å, no legitimate *ARP/wARP* models could be built from the ‘fc’ MR solutions (Table 1).

To validate the correctness of ‘fc’ MR solutions for targets that could not be modeled automatically with *ARP/wARP*, *phenix.autobuild* was used as an alternative automatic model rebuilding method. Models built by *phenix.autobuild* were generally of high quality (except for target StR70), with free R factors < 0.4, map correlation coefficient better than 0.75. and GDT-TS score to target X-ray structures > 0.85. For target StR70, although the quality of *phenix.autobuild* model was relatively poor with free R factor of 0.44 and map correlation coefficient of 0.62, it was still acceptable given the resolution of X-ray diffraction data is 2.80 Å (Table 3.4); the R and R_{free} values of the PDB deposited X-ray structure are 0.29 and 0.34 respectively. In conclusion, correct MR solutions were obtained and automatic model building of the crystal structure was done successfully for 22 of 25 of these NESG NMR / X-ray pairs, using the ‘fc”-trimmed NMR ensemble coordinates deposited in the PDB, *Phaser*, and either *ARP/wARP* or *Phenix*.

NMR structures can also be used as partial search models in solving complexes by MR

X-ray structure of NESG target OR8C, the “effector domain” of the influenza A

virus non-structural protein 1 (NS1A), was determined as a tetrameric complex bound to the F2F3 Zn-finger fragment of human cellular polyadenylation and specificity factor 30 (CPSF30)[116]. In this complex, the asymmetric unit has four chains, two for OR8C and two for F2F3. The solution NMR structure of target OR8C is a monomer[117]. NMR search model ensembles trimmed using “core atom sets” determined by *FindCore* provide an unambiguous *Phaser* solution for the two OR8C chains, with final TFZ=19.5 and LLG=352. Starting from this MR solution from *Phaser* and using the 1.95 Å resolution X-ray data, *ARP/wARP* could build the structure of the entire complex automatically with high accuracy and almost complete sequence coverage. More specifically, for the *ARP/wARP* model, the *R* factor is 0.22, *R*-free is 0.27, and 344 of 361 residues were traced successfully. The C_α-rmsd between X-ray structure of the complex and the automated *ARP/wARP* model is less than 0.3 Å (Figure 3.5A, Figure 3.5B). These results demonstrate that NMR structures can also be used as partial search models for MR experiments, and can be used to solve the structures of protein-protein complexes when there are minimal structural rearrangements upon complex formation.

NMR structures that fail to provide good MR models can be improved by *Rosetta* refinement

Three NMR structures in our MR experiments failed to generate correct MR solutions with the methods described above. For NESG target DrR147D, the GDT-TS between NMR structure (PDB id: 2kcz) and X-ray structure (PDB ID: 3ggn) is quite low (0.48), as a large portion of the NMR structure [46 residues (i.e. residues 24 – 69) out of 155 residues] is not well defined. The X-ray crystal structure of target SR478 is a dimer of three-helix bundle domains, and the orientation of two N-terminal helices is somewhat different between NMR and X-ray structure, which accounts for about 40 percent of the X-ray structure. For ZR18, the overall agreement between secondary structure elements

of the X-ray structure and the NMR structure are acceptable, however, the relative orientation between helix α1 (residues 40-47) and helix α2 (residues 71-81) is different in the NMR and X-ray structures; *viz*, the angles between those two helices in X-ray structure and NMR structure ensemble are 155.7 degree and 160.5-166.6 degree respectively. In addition, there are only 10 models in the reported NMR ensemble, which may not be large enough to properly sample the conformation space, providing an inaccurate estimate of precision that precludes proper elimination of inaccurately-defined regions in the initial model.

It has been pointed out previously that the phasing power of NMR structures that fail to provide good MR solutions can be significantly improved by *Rosetta* refinement[60,63]. Therefore we carried out *Rosetta* loop rebuilding and all-atom refinement for NMR structure ensembles of NESG targets SR478 and ZR18, respectively. Improved agreement was observed between the X-ray structure and *Rosetta*-refined NMR structure compared to the NMR structure deposited in the PDB. For example, the angles between helix α1 and helix α2 of some *Rosetta* decoys for target ZR18 were within one degree variance from their corresponding X-ray structure. Both average GDT-TS and best GDT-TS between *Rosetta* models and X-ray structures were much higher than their PDB-deposited counterparts for those two targets (Table 3.6). Using these *Rosetta*-refined NMR models, search models were prepared the same way as was done for the NMR structure ensembles. In both cases, we were able to obtain definite *Phaser* solutions starting from fc models with TFZ > 8 (Figure 3.2A). Specifically, we obtained a solution with TFZ=9.9 for target ZR18 (identified by ZR18_R) and a solution with TFZ=11.3 for target SR478 (identified by SR178_R), which are significantly higher than the values of TFZ=4.5 for target ZR18 and TFZ=4.8 for target SR478, respectively, before *Rosetta* loop rebuilding and all-atom refinement. These results confirm the high

value of the *Rosetta* loop-rebuilding and refinement protocol when using NMR structures for MR.

NMR structures can be successfully used as MR search models for homologous X-ray structures

As indicated by previous results, NESG NMR structures which have 100% sequence identity with target X-ray structures generally can be utilized successfully as MR search models. To further explore the value of NMR structures as MR search models, we identified homologous proteins in the PDB for nine (9) of the NESG NMR/X-ray structure pairs. These homologous X-ray structures were selected using the following criteria: (i) sequence identity with template sequence ≥ 20%, (ii) sequence coverage of the target by the template ≥ 70%, (iii) better than 3-Å diffraction data, and (iv) no more than 4 copies of the molecule in the asymmetric unit. These data sets for 9 homologous proteins are summarized in Table 3.7.

For each target, we aligned the sequence of homologous protein with the sequence of our NMR / X-ray structure pair using the *align2D* function of *Modeller* software[81] . Unaligned residues were deleted from template NMR/X-ray structures, and unmatched sidechains were stripped back to the CG/OG coordinates. Based on these pre-processed NMR structure ensembles or X-ray structure coordinates, search models were prepared using each of the eight protocols summarized in Table 3.3. *Phaser* was used to find MR solutions, and *ARP/wARP* was used for automatic model rebuilding.

The results of this study can be divided into two subsets, distinguished by the sequence identity between the NMR / X-ray structure pair and the corresponding homologous X-ray crystal structures. For all five homologues with sequence identity > 40%, (i.e. for templates CsR4, HR41, MrR110B, OR8C and SoR77) correct MR solutions

were found by *Phaser*, and a majority of residues could be successfully traced using *ARP/wARP*, with free R factors lower than 0.45 (Figure 3.3B, Table 3.8). On the other hand, for the four cases where the sequence identity between target X-ray sequence and template NMR/X-ray sequence is ≤ 30%, valid MR solutions were identified for only one case, SR213, with sequence identity of 24% and *Phaser* TFZ value of Z = 4.4. Subsequent model rebuilding demonstrates that this is indeed a correct solution, because the free R factor of the *ARP/wARP* model is only 0.24, and the GDT-TS value between the *ARP/wARP* model and target PDB structure is 0.94.

The same MR study was done using the corresponding NESG X-ray crystal structures, instead of the NMR structure ensembles, as MR templates. For all five targets with sequence identity greater than 40%, correct MR solutions could also be found using X-ray crystal structures as search models. Judged by TFZ scores of *Phaser* solutions and free R values of *ARP/wARP* models, for targets CsR4, OR8C and SoR77, the quality of MR solutions originating from either the NMR or X-ray search models was equally good. For target HR41, a better MR solution could be found using X-ray structure as a search model, while for target MrR110B a better MR solution was found using the ‘fc” trimmed NMR ensemble as the search model (Figure 3.3, Table 3.8). These results lead us to conclude that modern NMR structures can be as effective as X-ray crystal structures for MR of homologous protein structures, when the NMR coordinate ensemble is properly prepared.

Discussions

In this paper, we have shown that NESG NMR structures usually serve as excellent search models to estimate the phase information of their corresponding X-ray counterparts. Compared with X-ray crystallography, protein NMR structure determination is a relatively new field. The process of NMR structure determination is not as mature as the process of X-ray structure determination, and is still subject to intensive development. It is generally recognized that there is a gap between the quality of typical solution NMR structures and the best X-ray crystal structures[35]. However, over the last decade protein NMR analysis of small (< 160-residue) proteins has become more routine, and the quality of protein NMR structures has improved significantly. NMR structures of such proteins generally have accuracies comparable to medium-resolution (2.0 – 2.5 Å) X-ray crystal structures[35]. Moreover, as demonstrated in Figure 1, the quality of NMR structures solved by structural genomics consortia, such as the NESG, has consistently improved over the past several years, as improved methods of data analysis and structure validation tools have been incorporated into the protein structure refinement process.

In this study, we failed to obtain MR solution for target DrR147D by all of the methods tested. Further investigation revealed that there are *bona fide* structural differences between these NMR and X-ray structures due to the fact they were solved at different pH values. Specifically, the solution NMR structure is a monomer solved at pH 4.5, while the crystal structure is a dimer solved at pH 6.0; most residues on the dimer interface observed in this crystal structure are disordered in the corresponding monomeric NMR structure (Figure 3.4), and this disorder to order transition is pH dependent (unpublished results).

In our 22 successful MR experiments, one case, NESG target HR3646E, is

particularly interesting. Using the NMR ensemble to generate a 'fc'-trimmed search model ensemble, we obtained one solution with TFZ=3.6 and LLG = 26, which was also the single solution reported by *Phaser*. Although we tried various model preparation methods and different *Phaser* parameters, this solution with low TFZ score was the best we could obtain; this was not unexpected since the best GDT-TS score between any individual NMR model and X-ray structure was only 0.77. None the less, a highly accurate model (GDT-TS relative to X-ray structure equals to 0.97) could be built by *ARP/wARP* using the initial MR solution, with 93 of 98 residues automatically-traced (Figure 3.5C). Although the resolution of the X-ray data is high (1.45 Å), *ARP/wARP* worked so well as to indicate that starting MR model produced by *Phaser* must be correct, even with a relatively low TFZ score of 3.6.

Recent developments in structural bioinformatics have further expanded the application of NMR data in molecular replacement. For example, for small proteins with less than 130 residues, CS-*Rosetta* models generated using only chemical shift data and energy calculations can be quite accurate[3], and have been used successfully as MR search models[118] . In addition, as shown in Figure 3.2A for NESG targets SR478 and ZR18, by focusing sampling on the most structurally variable regions, and then relaxing the whole NMR structure in the *Rosetta* all-atom energy field, *Rosetta* loop rebuilding protocol can be used to improve their agreement with X-ray structures to provide better phasing power[60,63]. In this study, two NMR structures which did not initially provide MR solutions could be improved, both in phasing power and similarity with the crystal structure, by unconstrained *Rosetta* refinement. The generality of these results in using NMR structure ensembles as phasing models will be explored in future studies.

Conclusions

Stating from 25 pairs of X-ray and NMR structures solved by NESG, this work has demonstrated that by preparing MR ensembles using an interatomic variance matrix based protocol, *FindCore*, correct MR solutions can be found for the majority of the cases. Based on these solutions, automatic model rebuilding could be done successfully by either *ARP/wARP* or *Phenix*. *Rosetta* refinement has the potential of improving the phasing power of NMR structures, when the agreement of NMR structures with their corresponding X-ray structures is low. Our new MR model preparing protocol 'FindCore' outperforms other protocols, due to the fact it can make a good estimation of NMR ensemble precision at an atomic level. We also demonstrate that such properly prepared NMR ensembles and X-ray crystal structures have similar performance when used as MR search models for homologous structures, particularly for targets with sequence identity > 40%.

Acknowledgment

We thank Drs. L. Tong and F. Faroud for helpful comments on this manuscript. This work was supported by the National Institutes of General Medical Science Protein Structure Initiative program, grants U54 GM074958 and U54 GM094597. PDB and BMRB codes for the NMR NOESY peak list and chemical shift data, as well as crystallographic structure factor, data for the 25 proteins used in this study are summarized in Table 3.1, and available online at http://psvs-1_4-dev.nesg.org/MR/dataset.html and from the BioMagResDB (http://www.bmrb.wisc.edu/published/improve_tech/improve.html).

Table 3.1. Data for protein NMR / X-Ray structure pairs used in MR studies

NESG_ID	X-RAY					NMR							
	PDB_ID	Res(Å)	Space group	Coordinates	Structure factor	PDB_ID	Molecule	Coordinates	Constraints	BMRB ID	Chemical Shift	Peaks List[b]	FID[b]
BeR31	3CPK	2.5	$P4_32_12$	3CPK.pdb	3CPK-sf.cif	2K2E	monomer	2K2E.pdb	2K2E.mr	15702	15702.bmrb	NA	NA
CcR55	2O0Q	1.8	C222	2O0Q.pdb	2O0Q-sf.cif	2JQN	monomer	2JQN.pdb	2JQN.mr	15281	15281.bmrb	NA	15281.fid
CsR4	2OTA	2.2	P212121	2OTA.pdb	2OTA-sf.cif	2JR2	dimer	2JR2.pdb	2JR2.mr	15317	15317.bmrb	15317.peaks	15317.fid
CtR107	3E0H	1.81	P212121	3E0H.pdb	3E0H-sf.cif	2KCU	monomer	2KCU.pdb	2KCU.mr	16097	16097.bmrb	submitted	submitted
CtR148A	3IBW	1.93	P43212	3IBW.pdb	3IBW-sf.cif	2KO1	dimer	2KO1.pdb	2KO1.mr	16486	16486.bmrb	16486.peaks	16486.fid
DrR147D[a]	3GGN	2	P1211	3GGN.pdb	3GGN-sf.cif	2KCZ	monomer	2KCZ.pdb	2KCZ.mr	16100	16100.bmrb	submitted	submitted
GmR137	3CWI	1.9	P43212	3CWI.pdb	3CWI-sf.cif	2K5P	monomer	2K5P.pdb	2K5P.mr	15844	15844.bmrb	15844.peaks	15844.fid
HR1958	1TVG	1.6	C121	1TVG.pdb	1TVG-sf.cif	1XPW	monomer	1XPW.pdb	1XPW.mr	6344	6344.bmrb	6344.peaks	6344.fid
HR3646E	3FIA	1.45	C121	3FIA.pdb	3FIA-sf.cif	2KHN	monomer	2KHN.pdb	2KHN.mr	16250	16250.bmrb	submitted	submitted
HR41	3EVX	2.54	P1	3EVX.pdb	3EVX-sf.cif	2K07	monomer	2K07.pdb	2K07.mr	6546	6546.bmrb	NA	6546.fid
MbR242E	3GW2	2.1	P6422	3GW2.pdb	3GW2-sf.cif	2KKO	dimer	2KKO.pdb	2KKO.mr	16368	16368.bmrb	16368.peaks	16368.fid
MrR110B	3E0E	1.6	P212121	3E0E.pdb	3E0E-sf.cif	2K5V	monomer	2K5V.pdb	2K5V.mr	15849	15849.bmrb	15849.peaks	15849.fid
OR8C	2RHK	1.95	P41	2RHK.pdb	2RHK-sf.cif	2KKZ	monomer	2KKZ.pdb	2KKZ.mr	16376	16376.bmrb	16376.peaks	NA
PfR193A	3IDU	1.7	P1211	3IDU.pdb	3IDU-sf.cif	2KL6	monomer	2KL6.pdb	2KL6.mr	16385	16385.bmrb	16385.peaks	NA
PsR293	3H9X	2.51	P1	3H9X.pdb	3H9X-sf.cif	2KFP	monomer	2KFP.pdb	2KFP.mr	16186	16186.bmrb	16186.peaks	16186.fid
SR213	2IM8	2	P212121	2IM8.pdb	2IM8-sf.cif	2HFI	monomer	2HFI.pdb	2HFI.mr	16113	16113.bmrb	NA	16113.fid
SR384	3BHP	2.01	C121	3BHP.pdb	3BHP-sf.cif	2JVD	monomer	2JVD.pdb	2JVD.mr	15476	15476.bmrb	15476.peaks	15476.fid
SR478	2GSV	1.9	P121	2GSV.pdb	2GSV-sf.cif	2JS1	dimer	2JS1.pdb	2JS1.mr	15350	15350.bmrb	15350.peaks	submitted
SgR42	3C4S	1.7	P32	3C4S.pdb	3C4S-sf.cif	2JZ2	monomer	2JZ2.pdb	2JZ2.mr	15604	15604.bmrb	15604.peaks	NA
SoR77	2QTI	2.3	P43212	2QTI.pdb	2QTI-sf.cif	2JUW	dimer	2JUW.pdb	2JUW.mr	15456	15456.bmrb	15456.peaks	15456.fid
SsR10	2Q00	2.4	I4122	2Q00.pdb	2Q00-sf.cif	2JPU	monomer	2JPU.pdb	2JPU.mr	15265	15265.bmrb	15265.peaks	NA
StR65	2ES9	2	I213	2ES9.pdb	2ES9-sf.cif	2JN8	monomer	2JN8.pdb	2JN8.mr	15089	15089.bmrb	NA	NA
StR70	2ES7	2.8	P1211	2ES7.pdb	2ES7-sf.cif	2JZT	monomer	2JZT.pdb	2JZT.mr	7178	7178.bmrb	NA	NA
XcR50	1TTZ	2.11	P65	1TTZ.pdb	1TTZ-sf.cif	1XPV	monomer	1XPV.pdb	1XPV.mr	6363	6363.bmrb	NA	NA
ZR18	2FFM	2.51	P41212	2FFM.pdb	2FFM-sf.cif	1PQX	monomer	1PQX.pdb	1PQX.mr	5844	5844.bmrb	NA	5844.fid

a: Part of the NMR structure is not well defined (residue 24-69 out of 155 residues).
b: 'Submitted' means data has been sumbitted to BMRB but has not been updated by far.

Table 3.2. Summary of MR results

Target	X-ray structure				NMR Structure		GDT-TS[2]		Phaser Solution[3]		ARP/wARP or Phenix model[7]				
	PDB_id	Resolution	Space_Group	Length[1]	PDB_id	Year	Mean	Max	LLG	TFZ	R	R-Free	Docked	Matched[4]	GDT-TS
BeR31	3cpk	2.50	P43212	150	2k2e	2008	0.85	0.88	111	13.6	0.27(0.26)	0.43(0.34)	115	89 (118)	0.87(0.95)
CcR55	2o0q	1.80	C222	115	2jqn	2007	0.79	0.84	154	13.1	0.18	0.23	112	110 (114)	0.98
CsR4	2ota	2.20	P212121	76 (2)	2jr2	2007	0.95	0.97	388	27.7	0.23	0.30	123	116 (128)	0.96
CtR107	3e0h	1.81	P212121	158	2kcu	2009	0.72	0.77	54	8.7	0.23	0.29	136	120 (153)	0.88
CtR148A	3ibw	1.93	P43212	88 (2)	2ko1	2009	0.94	0.96	219	15.7	0.20	0.24	154	149 (156)	0.99
GmR137	3cwi	1.90	P43212	78	2k5p	2008	0.79	0.84	64	8.8	0.23	0.26	67	65 (73)	0.97
HR1958	1tvg	1.60	C121	153	1xpw	2004	0.78	0.81	150	9.5	0.22	0.26	134	102 (136)	0.87
HR3646E	3fia	1.45	C121	121	2khn	2009	0.75	0.78	26	3.6	0.20	0.26	93	90 (98)	0.97
MbR242E	3gw2	2.10	P6422	108	2kko	2009	0.88	0.93	178	18.1	0.23	0.26	89	84 (93)	0.95
MrR110B	3e0e	1.60	P212121	97	2k5v	2008	0.93	0.96	136	12.7	0.20	0.25	94	91 (95)	0.98
OR8C	2rhk	1.95	P41	140(2),72(2)	2kkz	2009	0.92	0.94	352	19.5	0.22	0.27	344	327 (361)	0.98
PfR193A	3idu	1.70	P1211	127 (2)	2kl6	2009	0.87	0.88	262	17.1	0.23	0.27	209	188 (226)	0.9
SgR42	3c4s	1.70	P32	66 (2)	2jz2	2008	0.94	0.96	210	21	0.16	0.20	107	102 (112)	0.95
SoR77	2qti	2.30	P43212	80	2juw	2007	0.93	0.97	173	16	0.23	0.30	64	61 (67)	0.96
SR213	2im8	2.00	P212121	131 (2)	2hfi	2006	0.82	0.86	234	14.1	0.25(0.29)	0.47(0.39)	201	183 (242)	0.92(0.89)
SR384	3bhp	2.01	C121	60 (3)	2jvd	2007	0.8	0.83	188	16.1	0.19	0.31	135	124 (157)	0.96
SsR10	2q00	2.40	I4122	129 (2)	2jpu	2007	0.84	0.88	454	24.6	0.27	0.33	218	155 (242)	0.84
StR65[5]	2es9	2.00	I213	115	2jn8	2007	0.82	0.86	38	5.8	0.24(0.30)	0.39(0.35)	77	51 (100)	0.71(0.86)
XcR50	1ttz	2.11	P65	87	1xpv	2004	0.9	0.94	81	10.9	0.19	0.24	72	69 (75)	0.96
HR41	3evx	2.54	P1	175 (4)	2k07	2008	0.82	0.85	445	16.7	0.29(0.24)	0.62(0.30)	46	NA	NA(0.96)
PsR293	3h9x	2.51	P1	125 (4)	2kfp	2009	0.81	0.85	227	12	0.28(0.18)	0.57(0.23)	10	NA	NA(0.99)
StR70	2es7	2.80	P1211	142 (4)	2jzt	2008	0.76	0.82	927	28.6	0.40(0.37)	0.58(0.44)	0	NA	NA(0.82)
DrR147D	3ggn	2.00	P1211	155 (2)	2kcz	2009	0.48	0.52	48	4.7	0.53	0.57	0	NA	NA
SR478	2gsv	1.90	P121	80 (2)	2js1	2007	0.74	0.78	51	4.8	0.47	0.54	0	NA	NA
ZR18	2ffm	2.51	P41212	91	1pqx[6]	2004	0.78	0.8	23	4.5	0.37	0.66	0	NA	NA

1-The number of subunits in the asymmetric unit is indicated in parentheses.
2-TM-score program is used to calculate GDT-TS between X-ray structure and NMR models.
3-TFZ and LLG values are extracted from MR solution with the highest TFZ score given LLG>0
4-The number of residues with C^{α}-rmsd < 1Å between *ARP/wARP* model and X-ray structure
5-rms=1.8 is used in MR_AUTO mode of *Phaser*.
6-All NMR structures contain 20 models except for ZR18 (10 models).
7-R,R-free and GDT-TS values of *Phenix* models are in the parentheses

Table 3.3. Methods of preparing MR templates for NMR structure ensemble

Method	Definition
nh	Composite ensemble model, hydrogen atoms are deleted from NMR ensemble
bsm	Single NMR model which has the highest structural similarity[a] with X-ray structure
aveB	Average model of NMR structure ensemble with distance based pseudo B-factor, "not-well-defined" residues[b] are deleted
AG	Composite ensemble model, residues with side chain longer than Ala are truncated to Ala
SAG	Composite ensemble model, residues with side chain longer than Ser are truncated to Ser
nd	Composite ensemble model , "not-well-defined" residues[b] are deleted from NMR structure ensemble
ndSAG	Composite ensemble model , "Not-well-defined" residues[b] are deleted from NMR structure ensemble, and long side chains of surface residues are truncated to Ser
fc	Composite ensemble model, all of the residues are trimmed using the *FindCore* pseudo B-factor[c]

a. *TMAlign* is used for structural alignment, and GDT-TS score is used to assess structural similarity.
b. "Not-well-defined" residues are calculated by *PSVS*, based on dihedral angle order parameters S(phi) + S(psi) < 1.8.
c. Each residue is treated as a hierarchical tree, nodes with pseudo B-factor larger than 60 are deleted along with all its children and sibling nodes.

Table 3.4. Models built by phenix.autobuild for cases *ARP/wARP* failed to build high quality models

	Target X-ray structure				aser Solution[a]		Model built[b]			
NESG_ID	PDB_ID	Res(Å)	R	Rfree	TFZ	LLG	R	Rfree	CC[c]	GDT-TS[d]
BeR31	3cpk	2.50	0.21	0.25	13.6	111	0.26	0.34	0.8	0.95
SR213	2im8	2.00	0.24	0.26	14.1	234	0.29	0.39	0.8	0.89
StR65	2es9	2.00	0.24	0.26	5.8	38	0.3	0.35	0.77	0.86
HR41	3evx	2.54	0.23	0.28	16.7	445	0.24	0.3	0.78	0.96
PsR293	3h9x	2.51	0.21	0.25	12	227	0.18	0.23	0.82	1.00
StR70	2es7	2.80	0.29	0.34	28.6	927	0.37	0.44	0.62	0.82

a. *Phaser* solutions from 'fc' starting models

b. Models were built by phenix.autobuild

c. Correlation coefficient between model and density map

d. GDT-TS score between models and X-ray structures deposited in PDB

Table 3.5. Comparison of performance for different model preparation protocols

Model preparing protocol	fc	nh	bsm	aveB	AG	SAG	nd	ndSAG
Correct Phaser solutions[1]	22	11	11	17	17	18	18	17
Accurate models built[2]	18	10	11	15	12	12	13	13

1-Number of *Phaser* solution with TFZ > 8 and LLG > 0 or being able to successfully guide subsequent automatic *ARP/wARP* model building

2-Number of *ARP/wARP* models with more than half of the residues in X-ray structures being accurately built (C_α-RMSD < 1Å)

Table 3.6. Comparison of *Rosetta*-refined NMR structures with X-ray structures

Target[a]	**Type[b]**	**Model #**	**Angle[c]**				**GDT-TS**	
			Mean	**SD**	**Closest**	**Min_dev[d]**	**Average**	**Best**
SR478	X-ray	1	132.4	0	132.4	0	1	1
	NMR	20	142	2.9	135.4	3	0.78	0.8
	Rosetta	20	133	5.6	131.9	0.5	0.85	0.93
ZR18	X-ray	1	155.7	0	155.7	0	1	1
	NMR	10	163.6	1.9	160.6	4.9	0.74	0.78
	Rosetta	20	163.6	4.7	155.3	0.4	0.81	0.87

a. NESG protein target name

b. Structure determination method. *Rosetta* refers to *Rosetta*-refinement of the NMR coordinates.

c. Angle between helices α1 and α2 of target ZR18, and angle between two N-terminal helices of target SR478

d. The minimum angle deviation of any individual model to X-ray structure

Table 3.7. Dataset of homologous proteins used in MR study

Target	Homologous target X-ray structure							Template NMR structure		
NESG_ID	PDB_ID	Res(Å)	R	R-free	n_chain	MW	Length	PDB_ID	Coverage	Seq_id
CsR4	2qti	2.3	0.22	0.23	1	9040.1	80	2jr2	0.83	0.41
HR1958	3f2z	1.3	0.16	0.18	1	17904.8	159	1xpw	0.87	0.21
HR41	3kpa	2.2	0.22	0.27	3	19809.9	168	2k07	0.94	0.57
MrR110B	3dm3	2.4	0.23	0.27	3	11742.4	105	2k5v	0.9	0.42
OR8C	2gx9	2.1	0.21	0.22	2	14737.1	129	2kkz	0.96	0.82
PsR293	2a1v	2.15	0.16	0.23	1	16303.6	144	2kfp	0.77	0.26
SR213	2huj	1.74	0.18	0.21	1	17219.8	140	2hfi	0.83	0.24
SoR77	2ota	2.2	0.24	0.26	2	8698.9	76	2juw	0.88	0.43
XcR50	1h75	1.7	0.2	0.21	1	9152.5	81	1xpv	0.96	0.3

Table 3.8. Results of MR of homologous proteins

	Target		Alignment		Template NMR Structure					Template X-ray Structure				
NESG_ID	PDB_ID	Res(Å)	Cov[a]	seq_id[b]	PDB_ID	TFZ	LLG	R	R-free	PDB_ID	TFZ	LLG	R	R-free
OR8C	2gx9	2.1	0.96	0.82	2kkz	22	420	0.25	0.30	2rhk	24	390	0.25	0.29
HR41	3kpa	2.2	0.94	0.57	2k07	11	450	0.25	0.43	3evx	18	436	0.29	0.37
SoR77	2ota	2.2	0.88	0.43	2juw	25	334	0.23	0.31	2qti	25	213	0.23	0.32
MrR110B	3dm3	2.4	0.90	0.42	2k5v	7.6	188	0.26	0.31	3e0e	5.1	167	0.27	0.44
CsR4	2qiI	2.3	0.83	0.41	2jr2	13	98	0.24	0.33	2ota	12	72	0.24	0.34
XcR50	1h75	1.7	0.96	0.30	1xpw	3.9	20	0.41	0.60	1ttz	3.7	17	0.52	0.61
PsR293	2a1v	2.15	0.77	0.26	2kfp	3.6	18	0.33	0.66	3h9x	7.1	36	0.20	0.27
SR213	2huj	1.74	0.83	0.24	2hfi	4.4	14	0.20	0.24	2im8	3.8	14	0.51	0.59
HR1958	3f2z	1.3	0.87	0.21	1xpw	4.6	17	0.49	0.53	1tvg	4.6	22	0.50	0.52

a. Sequence coverage of template structure to target X-ray structure
b. Above the dash line, sequence identity > 40%; below the dash line, sequence identity <= 30%

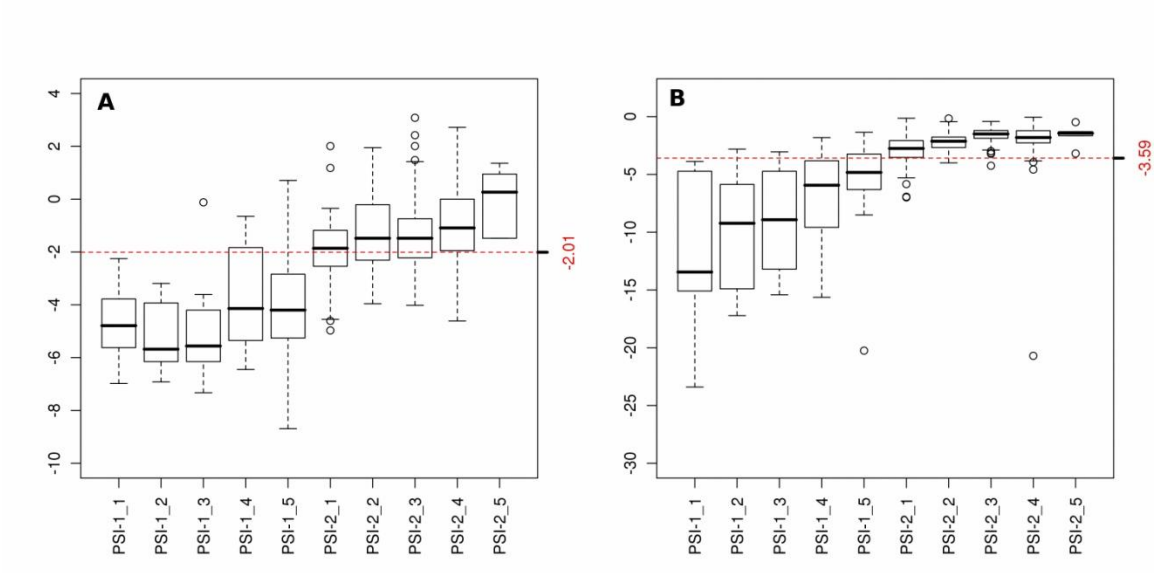

Figure 3.1. Structure quality Z-scores of NESG NMR structures~Fiscal Year

Knowledge-Based Structure Quality Scores for NESG NMR Structures Have Consistently Improved as NMR Methods Have Matured over the Past Several Years (A) and (B) show box plots of the distribution of Z scores (y axis) of PROCHECK ‘‘all-dihedral-angle’’ G factor and MolProbity clash scores, respectively, for all NMR structures solved by the NESG consortium in each PSI fiscal year (x axis). The red dashed lines represent the average Z scores. One PSI fiscal year is a 12 month time period generally spanning July 1st through June 30th of the following year. The PROCHECK all-dihedral-angle G factor is determined by the stereochemical quality of both backbone and side-chain dihedral angles of proteins, and MolProbity clash score is a measure to reflect the number of high-energy contacts in a structure calculated by the program probe. PSVS Z scores are calculated based on a calibrated data set of 252 high-quality X-ray crystal structures from the PDB with resolution %1.80 A°, R factor %0.25, and Rfree %0.28.

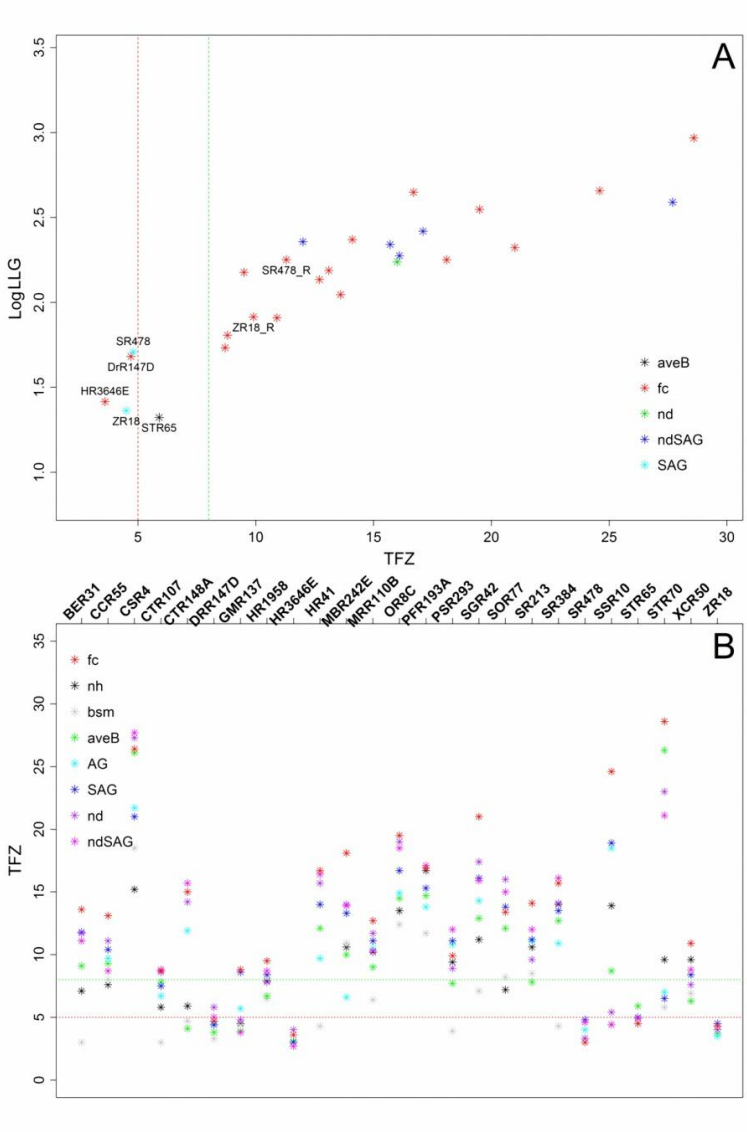

Figure 3.2. **TFZ plot for each target against model preparation protocols**

Using the fc Method, Phaser Phasing Scores Obtained Using NMR Structure Ensembles as Templates Are Generally Sufficient to Provide Good MR Solutions (A) LLG-TFZ scatter plot. LLG and TFZs are calculated by Phaser, and $\log_{10}$(LLG) and TFZs are plotted on y axis and x axis, respectively. The red vertical-dashed line delimits (TFZ = 5) the typical cutoff of an invalid Phaser solution, whereas the green vertical-dashed line (TFZ = 8) delimits the typical cutoff of a definite Phaser solution, according to the Phaser manual. For each individual target only the model with the highest TFZ solution is plotted. Colors are coded by different model preparation methods. SR478_R and ZR18_R denote the two models following Rosetta refinement. (B) Comparisons of TFZs from different MR models prepared by the eight model preparation methods. Models are color coded by their respective preparation method. TFZs calculated by Phaser are plotted on y axis, whereas each NESG target is plotted on x axis in alphabetical order. The red horizontal-dashed line (at TFZ = 5) delimits the typical cutoff of an invalid Phaser solution, whereas the green horizontal-dashed line (at TFZ = 8) delimits the typical cutoff of a definite Phaser solution, according to the Phaser manual.

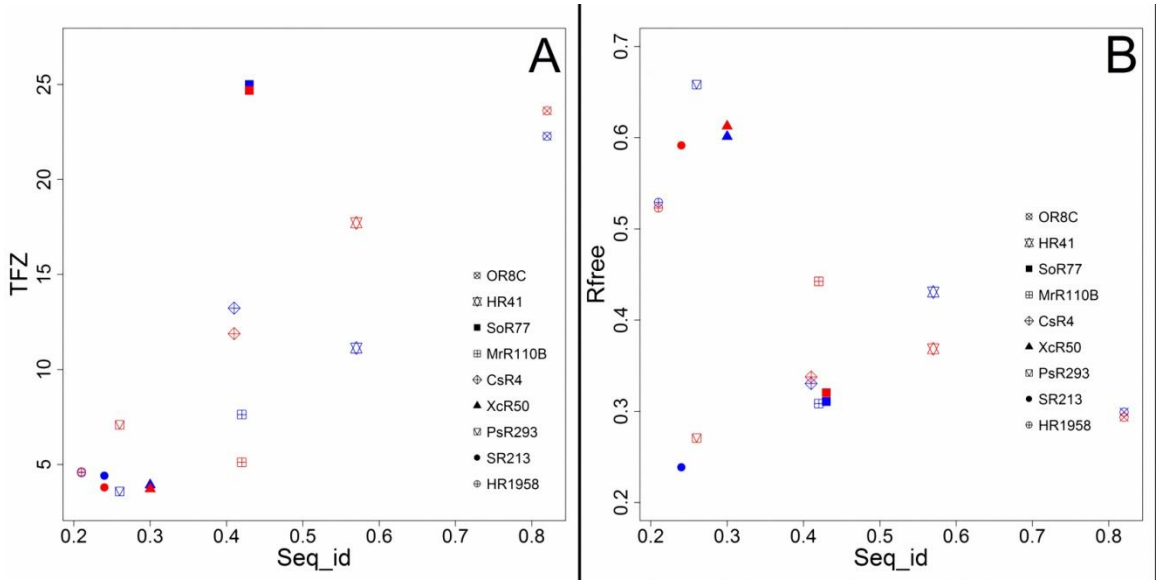

Figure 3.3. TFZ plot of homologous study

NMR and X-ray Structures Are About Equally Useful as Templates for Obtaining MR Solutions for Homologous Protein Structures (A) Plot of TFZs of Phaser solutions versus sequence identity (Seq_ID) between search model and target X-ray crystal structure. Solutions derived from X-ray crystal structure search models are colored red, and solutions derived from ‘‘fc’’-trimmed NMR structure ensemble search models are colored blue. (B) Plot of free R factor values of final ARP/wARP models versus sequence identity between search models and target X-ray structures. Solutions derived from X-ray crystal structure search models are colored red, and solutions derived from ‘‘fc’’-trimmed NMR structure ensemble search models are colored blue.

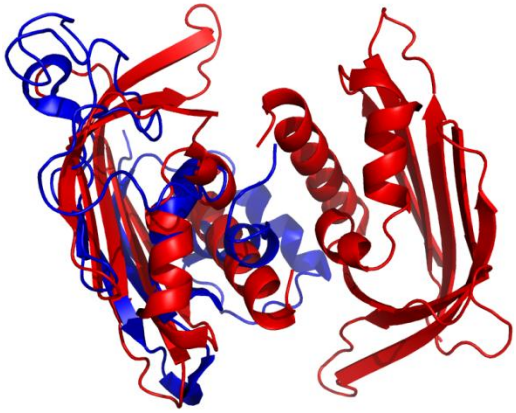

Figure 3.4. Structure superimposition of NMR and X-ray structures of DrR147D
Superimposition of X-ray structure (PDB id: 3ggn) chain A and NMR structure (PDB id: 2kcz) for target DrR147D. The X-ray structure is colored red and NMR structure is colored blue. X-ray structure is a dimer solved at pH 6.0, while NMR structure is a monomer solved at pH 4.5. The dimer interface of the crystal structure is largely disordered in the monomeric NMR structure. The figure was prepared using Pymol[119] .

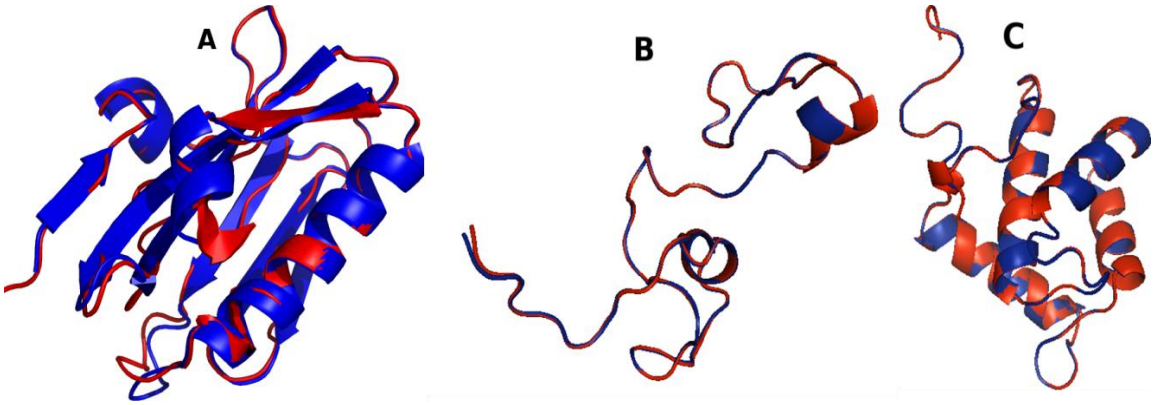

Figure 3.5. Structure Superimposition of *ARP/wARP* models and X-ray structures for OR8C-F2F3 complex and HR3646E

Structure Superimposition of *ARP/wARP* model (blue) and X-ray crystal structure deposited in PDB (red) for OR8C-F2F3 complex and HR3646E. (A) Chain A of OR8C-F2F3 complex, the effector domain of influenza A virus NS1A protein. The C_α rmsd between the *ARP/wARP* model and the deposited PDB structure is 0.24 Å. (B) Chain C of OR8C-F2F3 complex, the F2F3 fragment of human CPSF30. The C_α rmsd between the *ARP/wARP* model and the deposited PDB structure is 0.11 Å. (C) Target HR3646E-EH1 domain from human intersectin-1 protein. *ARP/wARP* models were automatically built by *ARP/wARP* expert system and refined by Refmac5 with no manual intervention, starting from *Phaser* MR solution. The C_α rmsd between the *ARP/wARP* model and the deposited PDB structure is 0.06 Å. Structures are aligned by C_α atoms and cartoon representation are displayed. All the figures were prepared using Pymol[119].

Chapter 4

Improving the quality of protein NMR structure by restrained Rosetta refinement

Introduction

Protein 3D structure is a cornerstone of investigating its functionality. The majority of the protein structures deposited in the Protein Data Bank are determined either by X-ray crystallography or Nuclear magnetic resonance (NMR). While X-ray crystal structures are derived from electron density data and are generally of higher resolution, solution protein NMR structure determination can reflect molecular dynamics and has the advantage of requiring no crystallization.

NMR structure determination is mainly based on three classes of experimental restraints: distance restraints, dihedral angle restraints and orientational restraints. In combination with those constraints, different algorithms and force fields are implemented to determine NMR structure by a variety of programs. Currently two groups of simulated annealing based programs are commonly used by the NMR community: XPLOR/CNS[20,23] and DYANA/CYANA[18,120]. Aside from the accuracy and completeness of experimental data, the quality of NMR structures also depends on the programs utilized in structure calculation and structure refinement. As demonstrated by many studies, the quality of NMR structures can be improved by structure refinement in state-of-the-art force field with explicit/implicit solvent[24,121,122,123,124,125]. By using the more advanced refinement protocols, a few re-refinement efforts have been done for the sake

of improving the quality of NMR structures especially for those determined at early days[126,127,128,129].

The NMR structure quality indicators generally fall into two categories: one category is related to experimental data, such as restraint violations[35], NOE completeness[130] and goodness-of-fit with NMR NOESY peak list data[4]; the other one is the knowledge-based normality scores relative to high-resolution X-ray crystal structure database, such as bond length, bond angle, backbone or side chain dihedral angle, and packing statistics[35,56,57,131]. CASD-NMR study has shown that the algorithm and force field utilized in NMR structure refinement has a heavy impact on those normality scores, for example, NMR structures refined by Rosetta are generally of excellent stereochemical and geometric quality scores[2].

Although the Rosetta molecular modeling program was first developed for de-novo protein structure prediction[132,133], homology modeling[112] and protein design[134], recently it has shown great potential in the fields of both molecular replacement[60,135,136] and NMR structure determination and refinement[5,137]. Theresa et al. have shown that unrestrained Rosetta refinement can improve the phasing power of NMR structure by moving it closer to its X-ray counterpart[63], which is corroborated by the results of Mao et al.'s study of a systematic investigation of using NMR structures in molecular replacement[7]. However, only one or two cases are reported in those two papers; to prove its generality, it is necessary to perform a systematic study of refining NMR structures by Rosetta on the basis of a much larger dataset. Another intriguing observation is that there are numerous restraint violations emerged after unrestrained Rosetta refinement, which begs the question: Do those violated restraints actually reflect de-facto intrinsic structural disagreement between NMR structures and X-ray crystal structures thus driving NMR structures away from their X-ray counterparts? If that is the

true, would incorporating NMR experimental restraints into Rosetta refinement have the reverse effect?

The Northeast Structural Genomics (NESG; http://www.nesg.org) consortium is one of the large-scale structure production centers of the Protein Structure Initiative (PSI). The NESG has contributed more than 450 NMR structures, as well as some 600 X-ray crystal structures, to the PDB over the past ten years, representing a large fraction of the NMR structures deposited into the PDB by the PSI. Although most NESG structures have been solved by either NMR or X-ray crystallography, as of December 2011 the NESG consortium had solved 41 pairs of protein structures for identical construct sequences using both X-ray crystallography and NMR methods. These 3D structures of proteins with identical sequences, together with the raw NMR and crystallography data available in the BioMagResBank(BMRB)[103] and Protein Data Bank(PDB)[13], are an extremely valuable composite dataset for understanding structural variations between solution and crystal states, and for new methods development. This dataset would be an ideal starting point for our investigation of using Rosetta to refine protein NMR structures.

In this study, we have done both unrestrained Rosetta refinement and restrained Rosetta refinement for all the NMR structures of 41 NESG NMR/X-ray structure pairs. The quality of PDB NMR structures, unrestrained Rosetta refined structures and restrained Rosetta refined structures has been evaluated by PSVS web server(http://psvs.nesg.org)[35], including restraint violations analysis, ensemble RMSD calculation, knowledge based normality analysis and RPF analysis. Then we have assessed the structural similarity with their corresponding X-ray structures and how well they could be utilized as molecular replacement templates.

Methods

Data preparation

In this study, we selected 41 NESG targets solved by both solution NMR and X-ray crystallography by December, 2011. The coordinates files of both NMR structures and X-ray crystal structures were downloaded from the PDB[13] database, and the NMR restraints files and X-ray structure factor files were also retrieved from the PDB. The structure factor files, downloaded in mmCIF format, were converted to mtz format using the CCP4 program CIF2MTZ (Collaborative Computational Project, Number 4, 1994). Another CCP4 program uniquefy was used to standardize the mtz files and select reflections for free R calculation. The NMR restraints files are either in CYANA format or in Xplor/CNS format, PDBStat v5.4 has been utilized to convert them to Rosetta format. The chemical shift and peak list data are retrieved from BMRB[103] database.

Rosetta Refinement

The Robetta fragment server(http://robetta.bakerlab.org/fragmentsubmit.jsp)[80,111] was used to generate fragment library, based on sequence and chemical shift data of each target protein. Loop rebuilding and all-atom refinement were done with Rosetta version 3.3 loopmodeling applications based on cyclic coordinate descent (CCD) and kinematic closure (KIC) [62,112]. The'fastrelax' mode was used to allow the whole structure to relax in Rosetta all-atom force field. For each target protein, loop regions were defined by the consensus of secondary structure, ''not-well-defined'' residues identified by the PSVS server based on dihedral angle order parameter values[35,104], and noncore residues identified by FindCore[6]. For each individual conformer of the NMR structure ensemble, we generated 500 decoys and picked the one with the lowest Rosetta energy as the final Rosetta refined model for this specific conformer, then those Rosetta refined

models derived from each conformer of the NMR ensemble were combined into an ensemble. If the NMR structure is an oligomer, a symmetry definition file would be generated by Rosetta and used to guide Rosetta refinement.

For restrained Rosetta refinement, the Rosetta formatted distance restraints and dihedral angle restraints were merged into a single restraints file, which would be used by Rosetta refinement with a weight of 1.0. The other steps were exactly the same as unrestrained Rosetta refinement.

Structure quality assessment

The quality of NMR structures, unrestrained Rosetta refined structures, and restrained Rosetta refined structures was evaluated by *PSVS* web server (http://psvs.nesg.org)[35], we calculated ensemble RMSD, restraint violations, RPF statistics[4], and various structural quality Z-scores such as Procheck[56] all dihedral angle Z-scores and Molprobity Clashscore[57] Z-scores. The results can be accessed by the following table: http://psvs-1_4-dev.nesg.org/results/rosetta_MR/rosettaMR_PSVS_summary.html

To evaluate the structural similarity between NMR structures and their X-ray counterparts, we utilized *PDBStat* v5.4 to calculate RMSD of backbone atoms or all heavy atoms for both well defined residues and all residues. We also used *TM-score*[110] program to calculate the GDT.TS[52] and TM-score[110]. To further determine RMSD for specific subset of atoms, such as side chain atoms of α-helix residues, we used *Pymol*[119] to superimpose NMR structures with reference X-ray crystal structures, and calculated RMSD based on the structural superimposition. The same procedures were performed to evaluate the structural similarity between Rosetta refined structures and X-ray crystal structures of the same targets.

The ordered regions are defined by dihedral angle order parameters with S(phi)+S(psi)>=1.8, and the core atoms are calculated by *FindCore* program based on interatomic distance variance matrix. The *DSSP*[138,139] program was utilized for secondary structure elements assignment, and solvent accessible areas of atoms were calculated by *areaimol* program in *CCP4* package[114,115].

Molecular Replacement

The program *Phaser*[97] (version 2.1) was used for molecular replacement. MR_AUTO mode was adopted with rms being set to 1.5. The program *ARP/wARP*[106] version 7.0 was used for automatic model building based on the Phaser MR solution. The *ARP/wARP* expert system mode was employed for automatic model building, and Refmac5[107] was used in refinement, staring from the positioned search model, and a maximum of ten building cycles were allowed. *Phenix.autobuild*[108] was also employed for automatic model rebuilding.

Results

Restrained Rosetta refinement significantly reduces the number of restraint violations

In this study, we calculate distance restraint violations and dihedral angle restraint violations for NMR structures, unrestrained Rosetta refined structures and restrained Rosetta refined structures. We divided distance restraint violations into three categories based on the level of severity, that is, the number of distance restraint violations between 0.1 Å and 0.2 Å, between 0.2 Å and 0.5 Å, and higher than 0.5 Å. Similarly, dihedral angle restraint violations were divided into two categories, one is between 1 degree and 10 degrees, and the other is higher than 10 degrees. The mean

and standard deviation of the number of restraint violations in each category were listed in Table 4.1. Clearly, for protein NMR structures, unrestrained Rosetta refinement would end up with a large number of restraint violations especially in the most severe violation category, while the number of restraint violations for restrained Rosetta refinement structures was in par with or slightly higher than the number of restraint violations of the NMR structures(Table 4.1, Figure 4.1). Therefore, incorporating NMR restraints into Rosetta refinement is an effective endeavor which would make Rosetta refined NMR structures meet the experimental restraints reasonably well.

Unrestrained Rosetta refinement decreases the Ensemble RMSD

The resolution of electron density map and atomic B-factor can reflect the precision of X-ray crystal structure, however, there are no equivalent experimental observables to define the precision of solution NMR structure. Usually the ensemble RMSD of NMR structure is considered to be a useful measure of its overall precision, although it could be problematic when intra-molecular dynamics is present. Here we calculated the ensemble RMSD of PDB NMR structures, unrestrained Rosetta refined structures and restrained Rosetta refined structures for 40 NESG targets, except for target GR4 which has only one single conformer in its NMR structure. Four categories of RMSD were calculated, which are backbone RMSD of well-defined residues defined by dihedral angle order parameters, backbone RMSD of all residues, heavy atom RMSD of well-defined residues defined by dihedral angle order parameters, and heavy atom RMSD of all residues. The mean and standard deviation of RMSD are listed in Table 1. The average RMSD of unrestrained Rosetta refined structures are higher than PDB NMR structures in all four categories, which indicates ignoring experimental restraints in Rosetta refinement would generally increase structural uncertainty for all the heavy atoms of all residues. For restrained Rosetta refined structures, in well defined region,

the average RMSD of backbone atoms is comparable with PDB NMR structures, and the average RMSD of all heavy atoms is about 10% lower than PDB NMR structures, which indicates that restrained Rosetta refinement has the potential of improving the precision of side chain heavy atoms. The same conclusion could be reached by inspecting Panel C of Figure 4.2. The RMSD of PDB NMR structures are plotted on X-axis, while the RMSD of Rosetta refined structures are plotted on Y-axis, with the unrestrained and restrained Rosetta refined structures represented by red solid triangle symbols and blue solid rectangle symbols respectively. It is evident that a majority of blue solid rectangle symbols are under the black dash line y=x, which demonstrates that for most targets restrained Rosetta refined structures are of smaller heavy atom RMSD than that of PDB NMR structures in well defined regions. On the other hand, for all residues, the average RMSD of both unrestrained and restrained Rosetta refined structures are higher than PDB NMR structures, which is not unexpected due to the loop rebuilding process implemented in our Rosetta refinement protocol.

Unrestrained Rosetta refinement fits the experimental data less well then restrained Rosetta refinement

RPF[4] is a tool to evaluate how well Protein NMR structure fits the experimental NOESY peak list and resonance assignment data. The program calculates recall, precision and DP score: Recall is defined as the percentage of peaks in the NOESY peak list that are consistent with the inter-proton distances of the 3D structures, Precision is defined as the percentage of close distance proton pairs in the query structures whose back calculated NOE cross peaks are also actually detected in NMR experiments, and DP score is a normalized F-score calculated from the recall and precision to measure the overall fit between the query structure and the experimental

data, with the freely-rotating chain model and the ideal model being used as the lower bound and upper bound respectively.

Since peak lists data are not available in BMRB database for 7 targets, we performed RPF analysis for the remaining 34 targets of the NMR/X-ray structure pairs dataset. The mean and standard deviation of recall, precision and DP-score are listed Table 1, and boxplots of recall, precision and DP-score are shown in Figure 4.3. For unrestrained Rosetta refined structures, they are of similar precision with PDB NMR structures but are of lower recall and DP-score than PDB NMR structures. On the other hand, for restrained Rosetta refined structures, they have almost identical average recall, precision and DP-score with PDB NMR structures (Table 4.1; Figure 4.3,Panel A,B,C). We draw the 2D DP-score scatterplot with the DP-score of PDB NMR structures being plotted on the X-axis and DP-score of Rosetta refined structures being plotted on the Y-axis, using red solid triangle or blue solid rectangle to represent unrestrained or restrained Rosetta refined structures respectively. The black dashed line represents y=x. Clearly, for a majority of the targets, structures generated by unrestrained Rosetta refinement are of lower DP-score than the PDB NMR structures; while structures generated by restrained Rosetta refinement are of similar DP-score with the PDB NMR structures (Fig 4.3, Panel D). Therefore, because no distance restraints are enforced during the unrestrained Rosetta refinement process, the refined structures will not satisfy the distance restraints as well as the PDB NMR structures, therefore they would fit the NOESY peak lists data less well because the distance restraints are directly derived from the NOESY peak lists. On the other hand, if the distance restraints are incorporated into Rosetta refinement, the refined structures can fit the NESOY peak lists data equally well as the PDB NMR structures.

Rosetta refinement consistently improves stereochemical quality and geometry of NMR structures

The NESG project uses the PSVS (http://psvs.nesg.org/) [35] to monitor the quality of structures. Based on a set of 252 high-resolution X-ray structures, PSVS provides Z-scores for a variety of widely adopted structural quality measures, such as PROCHECK G factor[56], MolProbity clashscore[57] and some other structure quality assessment metrics. The PROCHECK all dihedral angle G factor is determined by the stereochemical quality of both backbone and side-chain dihedral angles of proteins, and MolProbity clashscore calculated by the program probe is a measure to reflect the number of high-energy contacts in a structure. We ran PSVS to calculate a variety of knowledge-based structural quality Z-scores, including Verify3D, Prosa, Procheck backbone G factor (Procheck_bb), Procheck all dihedral angle G factor (Procheck_all) and Molprobity clashscore. The mean and standard derivation of those Z-scores are listed in Table 1. Both unrestrained and restrained Rosetta refined structures achieve better Z-scores for all the five metrics, especially for Procheck all dihedral angle G factor and Molprobity Clashscore Z-scores, which is consistent with the results of CASD-NMR study[2]. We draw boxplots of Procheck_bb, Procheck_all, Molprobity Clashcore Z-scores for PDB structures, unrestrained Rosetta refined structures and restrained Rosetta refined structures, which also shows Rosetta refined structures are of much improved Procheck_all and Molprobity clashscore Z-scores (Figure 4.4, Panel C,E). Therefore, the stereochemical quality and geometry of PDB NMR structures can be significantly improved after Rosetta refinement, no matter the experimental restraints being used or not.

We also made 2D Procheck_bb, Procheck_all and Molprobity Clashscore Z-scores scatterplots for unrestrained Rosetta refined structures and restrained Rosetta

refined structures to investigate the effect of the incorporation of experimental restraints into Rosetta refinement on protein structure quality. The Z-scores of unrestrained Rosetta refined structures (R3) and restrained Rosetta refined structures (R3cst) are plotted on X-axis and Y-axis respectively. The black dashed line represents y=x. The Procheck_bb Z-scores of restrained Rosetta refined structures are almost consistently better than unrestrained Rosetta refined structures, which indicates that the experimental dihedral angle restraints is helpful to guide Rosetta to generate decoys of better backbone stereochemcial quality (Figure 4.4, Panel B). On the contrary, the Molprobity Clashscore Z-scores of unrestrained Rosetta refined structures are generally better than restrained Rosetta refined structures, which can be explained by the fact that some experimental restraints may be responsible for the close contacts existed in the structure (Figure 4.4, Panel F).

Restrained Rosetta refinement mostly moves NMR structures closer to their X-ray counterparts

Theoretically, solution NMR structures need not necessarily be the same as X-ray crystal structures determined from a crystalline environment due to molecular motion and crystal packing. However, since X-ray structures are highly hydrated, one might expect such effects are not significant in most cases and X-ray and NMR structures should be similar. Therefore it is worthwhile to evaluate if NMR structures can be moved closer to their X-ray counterparts by Rosetta refinement with or without experimental restraints.

We calculated GDT.TS between PDB NMR structures, unrestrained Rosetta refined structures and restrained Rosetta refined structures with their corresponding X-ray structures, DrR147D (NESG id) was left out for this analysis because its solution

NMR structure is a monomer solved at PH 4.5 while its X-ray structure is a dimer solved at PH 6.0. Based on the results of previous studies[7,63], it is expected that unrestrained Rosetta refinement generally could move NMR structures closer to their X-ray counterparts; however, this was proved to be not the case. After unrestrained Rosetta refinement, only 18 targets achieve better GDT.TS values, 6 targets remain the same and 16 targets get worse GDT.TS values, with the average GDT.TS improved by merely 0.5 percent. On the other hand, after restrained Rosetta refinement, 33 targets achieve better GDT.TS values, 4 targets remain the same and only 3 targets get worse GDT.TS values, with the average GDT.TS improved by 2.4 percents. We drew a 2D GDT.TS scatterplot with the GDT.TS of NMR structures on the Y-axis and GDT.TS of Rosetta refined structures on the Y-axis, the black dashed line represents y=x. It is observed that if the similarity between NMR structures and X-ray structures is moderate (0.7 <= GDT.TS <= 0.85), more often than not, Rosetta refinement can move NMR structures closer to their X-ray counterparts especially when the experimental restraints are incorporated. However, if the similarity between NMR structures and X-ray structures is high (GDT.TS > 0.85), more often than not, unrestrained Rosetta refinement would move NMR structures away from their X-ray counterparts, but this effect can be reversed if the experimental restraints are utilized in the refinement process(Figure 4.5). Therefore, the majority of restraints derived from NMR experiments are not the source of structural differences between NMR structures and X-ray structures as implied by the previous studies, but are consistent between NMR structures and X-ray structures.

Furthermore, we were interested with how restrained Rosetta refinement can improve the similarity between NMR structures and X-ray structures. The atoms of NMR structures are grouped into different subsets based on their locations, then the RMSD between NMR structures, Rosetta refined structures with X-ray structures were

calculated for each subset. After restrained Rosetta refinement, the agreement between NMR structures and X-ray structures is consistently improved for both backbone and side chain atoms of ordered residues defined by dihedral angle order parameters, core residues calculated by *FindCore* program, buried residues and secondary structure elements (Figure 4.6). More often than not, the agreement between NMR structures and X-ray structures is improved over the disordered regions, non-core residues, surface residues and loop regions, but this improvement is much less consistent across different targets in our dataset. On the other hand, unrestrained Rosetta refinement would more or less randomly move NMR structures closer or away to their corresponding X-ray structures for any of those subsets of atoms(Figure 4.7).

Rosetta refinement could improve the phasing power of poor NMR MR templates

Molecular replacement (MR) is widely used for addressing the phase problem in X-ray crystallography. Historically, the common notion in structural biology community is that the quality of NMR structure is often not good enough for MR, even when the sequence of the search model is identical to the target X-ray structure. However, as demonstrated by a recent study[7], protein NMR structures prepared by an interatomic variance matrix based protocol are quite successful to be utilized as MR templates. Additionally, the phasing power of NMR structures that failed to provide good MR solutions can be improved by unrestrained Rosetta refinement in two cases. In this paper, we proposed to testify whether or not this assumption would stand correct in general, and to investigate the impact of incorporating experimental restraints into Rosetta refinement on the phasing power.

We prepared the MR starting models for PDB NMR structures, unrestrained Rosetta refined structures and restrained Rosetta refined structures by 'FindCore' protocol, *Phaser* was used to search for MR solutions. Three targets (DrR147D, ER382A and GR4) were excluded in this study due to the following facts: NMR structure of GR4 (PDB ID: 1rzw) consists of only a single model, therefore it can not be prepared by 'FindCore' protocol. The NMR structure of target ER382A (PDB ID: 2jn0) was solved as a monomer without a ligand, whereas its crystal structure counterpart (PDB ID: 3fif) has eight subunits in the asymmetric unit and was solved in complex with a heptapeptide ligand and appears to have a distinct structure, i.e., the Ca rmsd between the NMR structure and chain A of crystal structure is 2.44 Å. The NMR structure of target DrR147D (PDB ID: 2kcz) is a monomer solved at PH 4.5 while its crystal structure counterpart (PDB ID: 3ggn) is a dimer solved at PH 6.0.

For the initial Rosetta refinement protocol, the decoys are picked solely by Rosetta energy, that is, we picked the top 20 decoys with the lowest Rosetta energy from the entire pool of decoys generated from all the conformers in the NMR structure ensemble. It is observed that frequently those 20 decoys are originated from only one conformer or two and are highly similar with each other, thus the structural variance information within the NMR ensemble is lost during this kind of decoy picking process. In order to preserve all the conformers' information within the NMR ensemble, we proposed another protocol to pick the Rosetta decoys based on both conformer and energy, that is, to pick one decoy with the lowest Rosetta energy from the decoys generated from each conformer of the NMR ensemble, then to merge those conformer based decoys into an ensemble. The resulting Rosetta ensembles are much better MR templates and fit the NOESY peak lists data better than the Rosetta structures generated by the initial protocol, as manifested by the significantly improved TFZ scores and DP-scores for the

majority of the targets (Fig 4.8, Panel A, B). Therefore, to reflect structural uncertainty related to either insufficient experimental data or molecular dynamics, it is necessary to use the whole ensemble rather than a single model to represent NMR structure.

Starting from Phaser MR solutions, we utilized *Phenix* and *Arp/Warp* for automatic model rebuilding and refinement, and models with the lower R.free values were chosen as the final structures solved by MR. The detailed results of MR are presented in Table 4.2. For each target, we plotted the R.free values of the final MR structures against the sources of their MR templates, more specifically, which are PDB NMR structures, unrestrained Rosetta refined structures, and restrained Rosetta refined structures represented by black dots, red dots and green dots respectively (Figure 4.9). The green dashed line indicates R.free = 0.3 and the red dashed line indicates R.free = 0.45, any data points above the red dashed line (R.free > 0.45) are considered as failed MR solutions. Starting from their NMR structures as MR templates, seven targets (ZR18,SgR145, RpR324,StR65, SpR104, SR478, HR4435B) failed to provide valid MR solutions, four targets (RpR324,StR65, SR478, HR4435B) can provide good MR solutions after Rosetta refinement with or without experimental restraints, one target (ZR18) can provide a good MR solution and another target (SpR104) can provide a borderline acceptable MR solution (GDT.TS between MR structure and X-ray structure is 0.875) only after restrained Rosetta refinement. One target (SgR145) failed to provide good MR solutions even after restrained Rosetta refinement, which is a sparse restraints NMR structure and the C_{α}-RMSD to its corresponding X-ray structure is relatively large (3.07 Å). Surprisingly, two targets (HR41, SrR115C) which originally can provide valid MR solutions by using their NMR structures as MR templates failed to provide valid MR solutions after unrestrained Rosetta refinement, which was not the case if the experimental restraints were utilized in Rosetta refinement. The same conclusion can

be reached by inspecting the 2D GDT.TS plot (Figure 10), which demonstrates that when the NMR structures are poor MR templates to start with, that is, the GDT.TS values between the final MR structures derived from NMR structures and their corresponding X-ray structures are less than 0.8, mostly their phasing power can be significantly improved by Rosetta refinement especially with the experimental restraints being utilized. On the other hand, if the NMR structures are good MR templates initially, their phasing power can potentially be deteriorated by unrestrained Rosetta refinement, i.e., for targets CtR107, StR70, SrR115C, and HR41. Therefore, ignoring the experimental restraints is not recommended for the practice of Rosetta refinement.

Discussions

The quality of solution NMR structures is mainly determined by two factors: the accuracy and completeness of experimental data and the program used in structure calculation and refinement. In the past few years, several papers have demonstrated that unrestrained Rosetta refinement can improve the stereochemical quality of NMR structures and move NMR structures closer to X-ray crystal structures, which might be explained by two hypothesis: one is that all atom relaxation in Rosetta energy field can produce more energy favorable structures, the other is that some NMR experimental restraints are in conflict with X-ray structures solved at crystalline environment. In this study, we are interested to test whether the aforementioned observations stand correct for a large-scale investigation, and do the experimental restraints actually matter in Rosetta refinement. With that in mind, our final objective is to design an optimal protocol of using Rosetta to improve the quality of protein NMR structures.

The restrained Rosetta refinement produces structures with much less number of restraint violations than structures generated by unrestrained Rosetta refinement, which

proves that our restrained Rosetta refinement protocol is effective in meeting the experimental restraints. The weights of both distance restraints and dihedral angle restraints are set to 1 by default, we found that if those weights are too high, the final Rosetta refined models would be over restrained and often end up with poor Rosetta energy. On the other hand, if the weights of restraints are too low, the final Rosetta refined models would end up with a large number of restraint violations thus the restraints information is not properly utilized. Judged by ensemble RMSD, unrestrained Rosetta refinement will decrease the precision of NMR structures, while restrained Rosetta refinement can increase the precision of side chain heavy atoms of well defined residues. Additionally, the restrained Rosetta refined structures fit the NOESY peak list data better than unrestrained Rosetta refined structures. Rosetta refinement can generally improve the stereochemical quality and geometry of NMR structures, more specifically, the addition of dihedral angle restraints can guide Rosetta to generate models with even better backbone rotamers than otherwise. In most cases, restrained Rosetta refinement will move protein NMR structures closer to their X-ray counterparts, while unrestrained Rosetta refinement often fails to do so especially when the structural similarity between NMR structures and X-ray structures is considerably high (GDT.TS>0.85). For NMR structures with poor phasing power, mostly Rosetta refinement can be used to generate MR templates which are able to guide phasing software such as *Phaser* to identify correct MR solutions especially when the experimental restraints are utilized in Rosetta refinement. However, one must be aware of unrestrained Rosetta refinement can sometimes make NMR structures less useful MR templates, if they are good MR templates to start with; while this kind of pitfall does not come along with restrained Rosetta refinement. Therefore, we can safely declare that the majority of NMR experimental restraints still apply for their corresponding X-ray

structures, and it is the more sophisticated algorithm and the more advanced force field of Rosetta that helps to improve the quality of NMR structures.

Although Rosetta refinement could modify the input structure to some extent, it is expected that the refined structure won't deviate significantly from the input structure because Rosetta refinement would only sample conformations close to it. Therefore if the NMR structures are in poor agreement with their X-ray counterparts from the beginning, that kind of structural differences cannot be fixed by Rosetta refinement only. It is of great interest for us to find out if we utilize exclusively experimental restraints information for Rosetta calculation without the input NMR structure, whether or not the Rosetta models would be in better agreement with X-ray structures than restrained Rosetta refined structures. For nine NMR structures with GDT.TS to their X-ray counterparts less than 0.8, we have run CS-Rosetta calculation with the experimental restraints. If the length of the target is below 100, the CS-Rosetta structure of such target is more closer to X-ray structure than its corresponding restrained Rosetta refined structure, especially for target ER382A and ZR18; on the other hand, if the length of the target is above 120, the CS-Rosetta structure of such target are more distant to X-ray structure than its corresponding restrained Rosetta refined structure, which might be partially explained by the under sampling in our calculation (Table 4.3).

Conclusions

Starting from a dataset of 41 NESG NMR/X-ray structure pairs, we have done unrestrained Rosetta refinement and restrained Rosetta refinement for all the NMR structures. The knowledge based structural quality Z-scores are significantly improved by Rosetta refinement with or without restraints, especially for Procheck all dihedral angle G-factor and Molprobity clashscore. Incorporating NMR restraints into Rosetta refinement can significantly reduce the number of restraint violations. In addition, restrained Rosetta refined structures fit the NOESY peak lists data better, are in better agreement with their corresponding X-ray structures and are generally of better phasing power; while sometimes unrestrained Rosetta refinement could drive the NMR structures away from their X-ray counterparts especially when initially they are of high structural similarity. For small size protein NMR structures of poor structural similarity with their corresponding X-ray structures, CS-Rosetta calculation with the experimental restraints is proved to be a better choice than restrained Rosetta refinement. To summarize, a majority of the experimental NMR restraints still apply for X-ray crystal structures determined at crystalline environment, and they can be utilized to guide Rosetta to improve the quality of NMR structures, therefore the restraints should always be utilized in Rosetta refinement if available.

Acknowledgment

We thank Drs. Oliver Lange for his helpful advices on restrained Rosetta refinement. This work was supported by the National Institutes of General Medical Science Protein Structure Initiative program, grants U54 GM074958 and U54 GM094597.

Tabel 4.1. Summary of PSVS statistics. The mean and standard deviation of each measure listed in this table are formatted as mean±sd. The detailed PSVS statistics for each target can be access by the link below: http://psvs-1_4-dev.nesg.org/results/rosetta_MR/rosettaMR_PSVS_summary.html

		PDB	R3	R3cst
NOE violiations(Å)	[0.1,0.2)	5.4±7.2	15.1±7.7	7.1±5.2
	[0.2,0.5)	2.6±5.0	30.9±17.5	5.6±4.7
	>0.5	2.2±7.7	74.8±51.6	3.8±4.4
ACO violations(◦)	[1,10)	5.4±6.9	8.0±7.0	1.3±1.6
	>10	0.2±0.5	6.1±6.6	1.0±1.5
Ensemble RMSD(Å)	bb_ord	0.79±0.69	1.05±0.83	0.80±0.73
	hvy_ord	1.19±0.64	1.43±0.79	1.07±0.70
	bb_all	2.92±1.85	3.32±1.80	3.14±1.64
	hvy_all	3.46±1.85	3.85±1.83	3.61±1.68
RPF statisics	Recall	0.94±0.07	0.92±0.07	0.94±0.06
	Precision	0.90±0.06	0.91±0.06	0.90±0.06
	DP-score	0.79±0.08	0.76±0.07	0.79±0.08
PSVS Z-scores	Verify3D	-2.26±1.18	-1.22±1.05	-1.29±1.04
	Prosa	-0.61±1.05	-0.20±0.98	-0.29±1.01
	Procheck_bb	-0.37±1.67	0.11±1.44	0.59±1.55
	Procheck_all	-1.02±1.90	1.21±1.42	1.26±1.54
	Molprobity Clashscore	-2.15±1.23	0.84±0.37	0.41±0.64

Table 4.2. Summary of MR results

Target	Source	Phaser		Model(GDT.TS)		Arp/wARP				Phenix				
		TFZ	LLG	Mean	Best	R	R.free	Docked	GDT.TS	R	Rfree	Docked	GDT.TS	Map_CC
BeR31	PDB	13.6	112	0.85	0.88	0.26	0.41	85	0.77	0.26	0.35	115	0.95	0.79
BeR31	R3	19.6	202	0.86	0.90	0.24	0.35	112	0.95	0.27	0.36	111	0.92	0.8
BeR31	R3cst	19.7	180	0.87	0.93	0.27	0.47	88	0.69	0.25	0.36	113	0.93	0.8
CcR55	PDB	13.2	155	0.79	0.84	0.21	0.27	107	0.93	0.25	0.28	103	0.89	0.83
CcR55	R3	19.5	325	0.85	0.90	0.18	0.23	111	0.97	0.27	0.31	93	0.81	0.82
CcR55	R3cst	18.7	294	0.86	0.92	0.18	0.23	110	0.97	0.23	0.27	99	0.87	0.85
CsR4	PDB	27.8	405	0.95	0.97	0.22	0.30	126	0.99	0.24	0.29	130	0.97	0.86
CsR4	R3	24.9	592	0.9	0.96	0.24	0.31	122	0.93	0.24	0.31	126	0.94	0.85
CsR4	R3cst	26.5	471	0.97	0.99	0.23	0.32	127	0.99	0.24	0.33	130	0.97	0.86
CtR107	PDB	9	51	0.71	0.76	0.31	0.60	0	NA	0.26	0.3	122	0.81	0.81
CtR107	R3	7.4	-21	0.69	0.77	0.34	0.59	0	NA	0.29	0.34	108	0.72	0.82
CtR107	R3cst	10.2	74	0.72	0.80	0.19	0.25	147	0.96	0.28	0.31	118	0.78	0.81
CtR148A	PDB	15	230	0.94	0.96	0.20	0.26	153	1.00	0.26	0.29	166	1.00	0.79
CtR148A	R3	18.8	319	0.93	0.97	0.20	0.25	150	0.98	0.26	0.29	160	1.00	0.79
CtR148A	R3cst	16.6	246	0.96	0.98	0.20	0.24	153	1.00	0.25	0.29	166	1.00	0.79
DhR29B	PDB	9	74	0.84	0.89	0.27	0.35	83	0.90	NA	NA	0	NA	NA
DhR29B	R3	12.8	137	0.87	0.92	0.23	0.31	86	1.00	NA	NA	0	NA	NA
DhR29B	R3cst	11.1	126	0.91	0.94	0.25	0.33	83	0.96	0.23	0.28	87	1.00	0.81
DhR8C	PDB	13.5	240	0.82	0.84	0.26	0.32	122	0.90	0.28	0.32	120	0.86	0.8
DhR8C	R3	17.5	251	0.82	0.84	0.26	0.32	127	0.95	0.29	0.35	114	0.80	0.78
DhR8C	R3cst	15.2	242	0.83	0.87	0.26	0.32	120	0.90	0.3	0.35	116	0.81	0.79
ER382A	PDB	4.7	159	0.77	0.80	0.35	0.67	0	NA	0.43	0.54	368	0.69	0.53
ER382A	R3	4.4	139	0.8	0.89	0.35	0.68	0	NA	0.44	0.55	368	0.71	0.56

ER382A	R3cst	4.4	241	0.8	0.86	0.35	0.68	7	NA	0.38	0.49	368	0.80	0.65
GmR137	PDB	8.4	65	0.78	0.82	0.24	0.28	65	0.96	0.27	0.3	64	0.93	0.79
GmR137	R3	11.2	103	0.79	0.88	0.26	0.29	67	0.96	0.27	0.31	62	0.90	0.79
GmR137	R3cst	9.3	81	0.8	0.85	0.23	0.28	67	0.97	0.25	0.28	64	0.94	0.8
HR1958	PDB	9.6	150	0.78	0.81	0.21	0.26	134	0.58	0.23	0.26	139	0.88	0.81
HR1958	R3	12	224	0.83	0.85	0.22	0.27	130	0.86	0.24	0.26	139	0.87	0.8
HR1958	R3cst	11.5	232	0.83	0.85	0.23	0.27	130	0.86	0.25	0.27	139	0.88	0.81
HR3102A	PDB	16	159	0.92	0.96	0.21	0.30	74	0.99	0.21	0.27	74	0.97	0.86
HR3102A	R3	14.9	171	0.9	0.94	0.22	0.29	73	0.97	0.22	0.28	74	0.97	0.85
HR3102A	R3cst	13.2	134	0.93	0.96	0.22	0.30	72	0.96	0.22	0.29	74	0.97	0.86
HR3646E	PDB	4.5	26	0.75	0.79	0.20	0.28	93	0.97	0.29	0.31	92	0.87	0.74
HR3646E	R3	8.6	96	0.79	0.86	0.20	0.25	97	0.98	0.3	0.3	89	0.85	0.74
HR3646E	R3cst	7.5	61	0.82	0.85	0.35	0.56	0	NA	0.26	0.29	92	0.90	0.76
HR41	PDB	16	457	0.82	0.85	0.30	0.66	30	NA	0.25	0.3	616	0.95	0.78
HR41	R3	16.2	532	0.78	0.84	0.32	0.68	0	NA	NA	NA	0	NA	NA
HR41	R3cst	20.7	679	0.84	0.90	0.29	0.63	96	NA	0.25	0.31	604	0.94	0.76
HR4435B	PDB	3.9	18	0.71	0.80	NA	NA	0	NA	0.49	0.54	64	0.69	0.48
HR4435B	R3	7.2	15	0.77	0.85	NA	NA	0	NA	0.29	0.33	58	0.79	0.66
HR4435B	R3cst	4.1	31	0.79	0.86	NA	NA	0	NA	0.27	0.29	58	0.80	0.67
HR4527E	PDB	12.1	182	0.86	0.89	0.23	0.30	132	0.97	0.24	0.27	138	0.90	0.85
HR4527E	R3	14.4	204	0.85	0.88	0.24	0.30	132	0.96	0.24	0.26	132	0.89	0.86
HR4527E	R3cst	13.3	182	0.88	0.90	0.22	0.27	132	0.99	0.24	0.27	134	0.90	0.85
HR4694F	PDB	20.5	804	0.87	0.90	NA	NA	0	NA	0.28	0.32	308	0.93	0.8
HR4694F	R3	22.4	907	0.89	0.91	NA	NA	0	NA	0.28	0.32	308	0.92	0.8
HR4694F	R3cst	23.8	813	0.9	0.91	NA	NA	0	NA	0.28	0.33	312	0.93	0.8
HR5546A	PDB	10.3	128	0.79	0.82	0.26	0.48	124	NA	0.23	0.28	206	0.96	0.83
HR5546A	R3	12.5	183	0.79	0.83	0.31	0.67	0	NA	0.25	0.31	198	0.90	0.81

HR5546A	R3cst	11.4	166	0.79	0.82	0.31	0.62	31	NA	0.26	0.33	200	0.92	0.81
LkR112	PDB	22.3	394	0.94	0.97	0.19	0.23	257	1.01	0.21	0.23	264	0.99	0.82
LkR112	R3	20.1	481	0.92	0.96	0.21	0.25	246	0.98	0.21	0.23	261	0.99	0.82
LkR112	R3cst	21.2	397	0.96	0.98	0.19	0.23	255	1.00	0.21	0.23	264	0.99	0.81
MbR242E	PDB	18.3	178	0.88	0.93	0.23	0.27	88	0.94	0.25	0.3	97	0.97	0.82
MbR242E	R3	14.2	210	0.88	0.94	0.22	0.27	91	0.95	0.27	0.29	85	0.90	0.82
MbR242E	R3cst	18.4	235	0.91	0.96	0.23	0.28	88	0.94	0.24	0.29	95	0.98	0.83
MrR110B	PDB	12.7	136	0.93	0.95	0.20	0.26	94	0.98	0.22	0.26	91	0.97	0.86
MrR110B	R3	9.8	138	0.93	0.96	0.22	0.29	92	0.95	0.22	0.27	89	0.96	0.87
MrR110B	R3cst	11.5	126	0.95	0.97	0.20	0.26	92	0.96	0.22	0.26	92	0.98	0.87
OR8C	PDB	19.5	352	0.92	0.94	0.30	0.35	243	0.97	0.34	0.37	238	0.98	0.7
OR8C	R3	21.6	459	0.91	0.94	0.31	0.37	254	0.96	0.35	0.38	234	0.98	0.7
OR8C	R3cst	17.6	434	0.92	0.93	0.31	0.36	231	0.97	0.35	0.38	236	0.98	0.7
PfR193A	PDB	16.9	266	0.87	0.88	0.24	0.28	206	0.90	0.24	0.26	214	0.90	0.82
PfR193A	R3	13.4	225	0.86	0.89	0.24	0.29	204	0.89	0.24	0.26	214	0.90	0.82
PfR193A	R3cst	16.3	242	0.88	0.89	0.23	0.27	209	0.90	0.25	0.28	214	0.90	0.82
PsR293	PDB	12.5	245	0.81	0.84	0.29	0.59	0	NA	0.18	0.22	472	1.00	0.82
PsR293	R3	17.2	503	0.82	0.87	0.29	0.47	270	0.55	0.23	0.29	432	0.92	0.8
PsR293	R3cst	15	368	0.83	0.88	0.28	0.58	20	NA	0.2	0.24	464	0.98	0.81
RpR324	PDB	4.2	5	0.8	0.83	0.39	0.55	0	NA	0.44	0.51	93	0.69	0.55
RpR324	R3	9.6	59	0.8	0.83	0.21	0.26	92	0.98	0.26	0.29	94	0.95	0.79
RpR324	R3cst	11	86	0.82	0.85	0.21	0.27	89	0.95	0.24	0.29	94	0.96	0.8
SgR145	PDB	4.2	57	0.64	0.66	0.35	0.59	0	NA	0.43	0.51	316	0.63	0.6
SgR145	R3	4.4	-92	0.62	0.66	0.36	0.57	0	NA	0.41	0.47	276	0.50	0.62
SgR145	R3cst	3.8	21	0.63	0.65	0.36	0.64	0	NA	0.4	0.47	306	0.53	0.6
SgR209C	PDB	17.5	491	0.81	0.86	0.26	0.53	197	NA	0.22	0.27	512	0.95	0.87
SgR209C	R3	16.8	382	0.8	0.88	0.20	0.35	461	0.86	0.23	0.27	504	0.93	0.86

SgR209C	R3cst	25	811	0.85	0.92	0.21	0.30	495	0.98	0.22	0.27	512	0.94	0.87
SgR42	PDB	17.9	226	0.94	0.96	0.17	0.21	108	0.97	0.2	0.23	112	0.98	0.86
SgR42	R3	20.3	230	0.96	1.00	0.16	0.21	107	0.95	0.2	0.21	112	0.98	0.86
SgR42	R3cst	17.5	182	0.97	0.99	0.16	0.21	108	0.95	0.2	0.22	112	0.98	0.87
SoR77	PDB	17.2	163	0.93	0.97	0.23	0.29	65	0.97	0.22	0.28	67	0.98	0.87
SoR77	R3	14.4	145	0.89	0.95	0.24	0.32	64	0.96	0.24	0.33	63	0.94	0.86
SoR77	R3cst	15.1	162	0.94	0.99	0.24	0.31	65	0.97	0.22	0.28	68	0.99	0.87
SpR104	PDB	4	46	0.82	0.87	0.29	0.68	41	NA	0.42	0.51	134	0.71	0.4
SpR104	R3	3.6	67	0.81	0.94	0.28	0.62	42	NA	0.49	0.51	108	0.41	0.35
SpR104	R3cst	4.3	64	0.86	0.90	0.27	0.66	26	NA	0.37	0.46	126	0.88	0.49
SR213	PDB	14.1	235	0.81	0.86	0.24	0.45	196	0.92	0.3	0.38	218	0.88	0.81
SR213	R3	16.9	351	0.83	0.88	0.26	0.36	215	0.96	0.31	0.38	210	0.85	0.8
SR213	R3cst	15	265	0.84	0.88	0.25	0.48	187	0.92	0.3	0.37	214	0.87	0.8
SR384	PDB	15.8	203	0.78	0.80	0.19	0.33	134	0.94	0.26	0.29	120	0.78	0.78
SR384	R3	14.3	191	0.79	0.82	0.20	0.32	131	0.94	0.26	0.29	117	0.76	0.79
SR384	R3cst	10.4	152	0.78	0.83	0.20	0.33	135	0.95	0.26	0.28	117	0.76	0.79
SR478	PDB	NA	NA	0.73	0.77	NA	NA	0	NA	NA	NA	0	NA	NA
SR478	R3	12.1	188	0.81	0.90	0.24	0.29	129	1.01	0.28	0.31	122	0.91	0.73
SR478	R3cst	4.7	-64	0.73	0.80	0.34	0.54	0	NA	0.31	0.35	130	0.89	0.71
SrR115C	PDB	21	367	0.93	0.97	0.37	0.57	0	NA	0.29	0.32	276	1.00	0.77
SrR115C	R3	4.4	169	0.93	0.97	0.39	0.69	0	NA	NA	NA	0	NA	NA
SrR115C	R3cst	4.5	337	0.92	0.98	0.37	0.63	0	NA	0.34	0.42	368	1.00	0.72
SsR10	PDB	24.4	454	0.84	0.88	0.26	0.35	213	0.98	0.26	0.3	232	0.94	0.76
SsR10	R3	26.5	546	0.87	0.90	0.25	0.34	218	0.94	0.26	0.3	226	0.92	0.76
SsR10	R3cst	24.6	498	0.89	0.91	0.25	0.33	237	0.97	0.26	0.3	232	0.93	0.76
StR65	PDB	4.6	14	0.81	0.85	0.52	0.58	0	NA	0.45	0.56	88	0.65	0.57
StR65	R3	7.6	69	0.82	0.90	0.31	0.50	49	NA	0.32	0.38	84	0.82	0.77

StR65	R3cst	11.4	101	0.83	0.89	0.27	0.45	66	0.71	0.32	0.38	89	0.86	0.75
StR70	PDB	30.8	936	0.76	0.81	0.32	0.60	0	NA	0.38	0.47	420	0.81	0.66
StR70	R3	26.8	963	0.75	0.81	0.32	0.59	11	NA	0.37	0.44	384	0.69	0.64
StR70	R3cst	28.4	1053	0.79	0.83	0.28	0.50	60	NA	0.36	0.42	396	0.81	0.7
UuR17A	PDB	8.3	64	0.72	0.75	0.35	0.61	25	NA	0.3	0.34	109	0.86	0.72
UuR17A	R3	12	103	0.73	0.79	0.25	0.35	98	0.82	0.31	0.37	101	0.81	0.72
UuR17A	R3cst	11.5	107	0.75	0.80	0.30	0.43	82	0.66	0.29	0.34	108	0.85	0.75
XcR50	PDB	10.9	81	0.9	0.94	0.23	0.30	72	0.95	0.2	0.22	72	0.96	0.89
XcR50	R3	11.6	120	0.86	0.92	0.27	0.54	22	NA	0.2	0.21	74	0.98	0.87
XcR50	R3cst	11.1	89	0.88	0.92	0.20	0.27	73	0.97	0.18	0.2	75	0.99	0.89
ZR18	PDB	4.2	24	0.77	0.79	0.34	0.71	0	NA	0.47	0.57	79	0.62	0.55
ZR18	R3	9.7	93	0.76	0.85	0.34	0.66	0	NA	0.44	0.55	70	0.57	0.53
ZR18	R3cst	10	79	0.79	0.85	0.32	0.63	20	NA	0.3	0.38	77	0.90	0.7

Table 4.3. GDT.TS to corresponding X-ray structures

Target	Length	PDB[a]	R3[b]	R3cst[c]	CSRcst[d]
ER382A	61	0.77	0.80	0.80	0.91
GmR137	78	0.78	0.79	0.80	0.84
HR4435B	83	0.71	0.77	0.79	0.80
ZR18	91	0.77	0.76	0.79	0.90
UuU17A	121	0.72	0.73	0.75	0.72
HR3646E	121	0.75	0.79	0.82	0.81
HR5546A	122	0.79	0.79	0.79	0.74
GR4	123	0.51	0.53	0.55	0.54
SgR145	202	0.64	0.62	0.63	0.61

a. NMR structures deposited in PDB
b. Unrestrained Rosetta refined structures
c. Restrained Rosetta refined structures
d. CS-Rosetta structures with experimental restraints

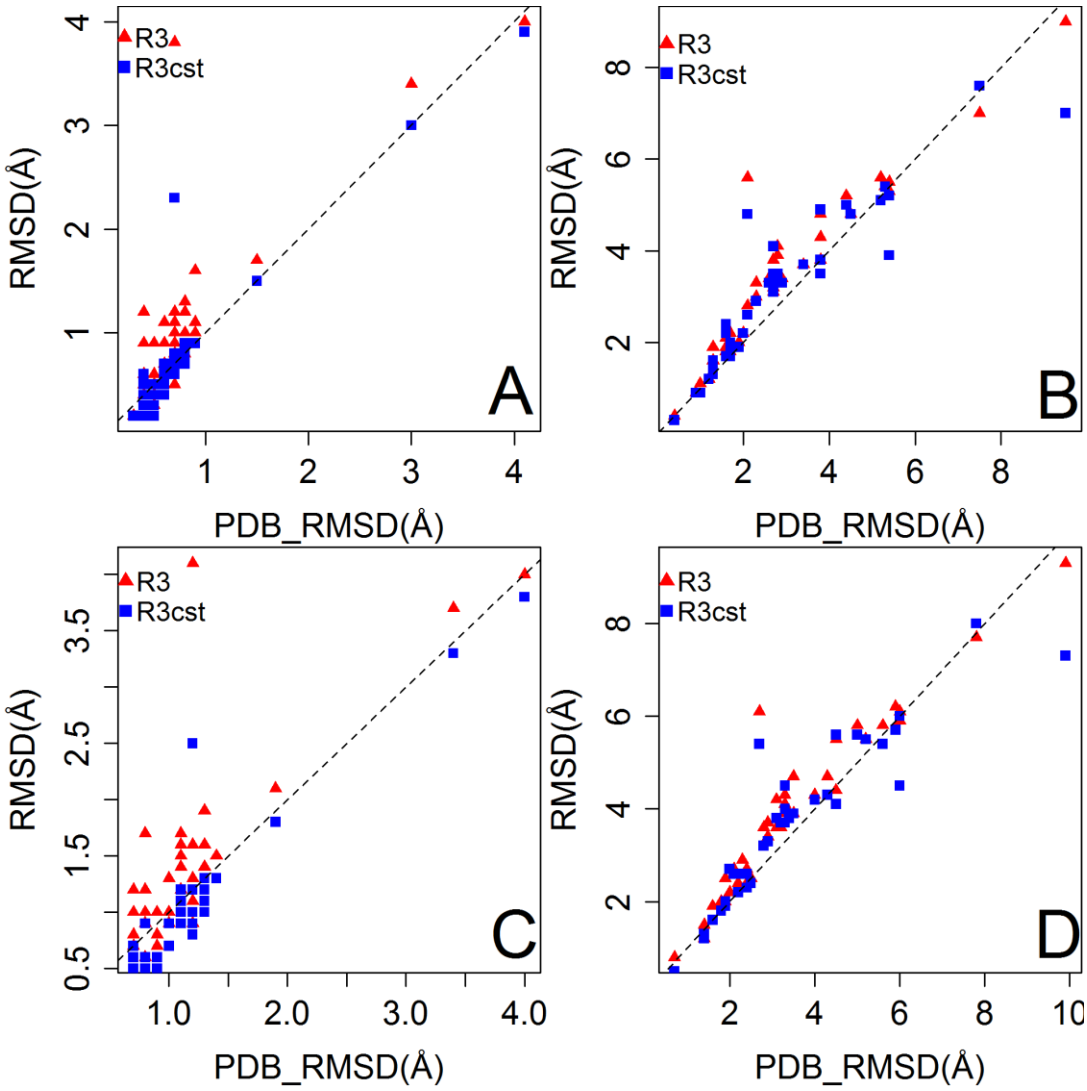

Figure 4.1. 2D ensemble RMSD scatterplots. The X-axis is the ensemble RMSD of the PDB NMR structures, and the Y-axis is the ensemble RMSD of the unrestrained Rosetta refined structures represented by red solid triangle symbols(R3) and restrained Rosetta refined structures represented by blue solid rectangle symbols(R3cst). The black dashed line represents y=x. (A): 2D ensemble RMSD scatterplot of backbone atoms for well defined residues defined by S(phi)+S(psi)>=1.8. (B): 2D ensemble RMSD scatterplot of backbone atoms for all residues. (C): 2D Ensemble RMSD scatterplot of all heavy atoms for well defined residues defined by S(phi)+S(psi)>=1.8. (C): 2D Ensemble RMSD scatterplot of all heavy atoms for all residues.

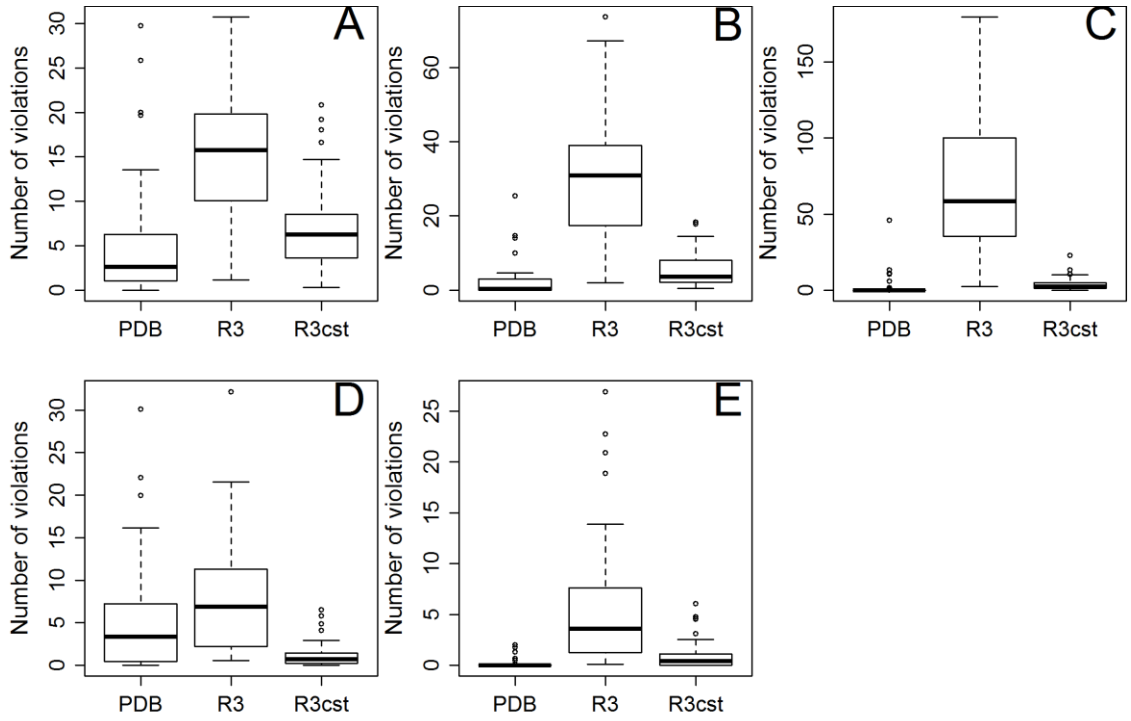

Figure 4.2. Boxplot of the number of restraint violations against structure sources

The statistics is assessed across the complete set of 41 NESG targets. PDB: NMR structures deposited in PDB; R3: Structures generated by unrestrained Rosetta refinement; R3cst: Structures generated by restrained Rosetta refinement. (A): Boxplot of the number of distance restraint violations between 0.1Å and 0.2Å. (B): Boxplot of the number of distance restraint violations between 0.2Å and 0.5Å. (C): Boxplot of the number of distance restraint violations larger than 0.5Å. (D): Boxplot of the number of dihedral angle restraint violations between 1˚and 10˚. (E): Boxplot of the number of dihedral angle restraint violations larger than 10˚.

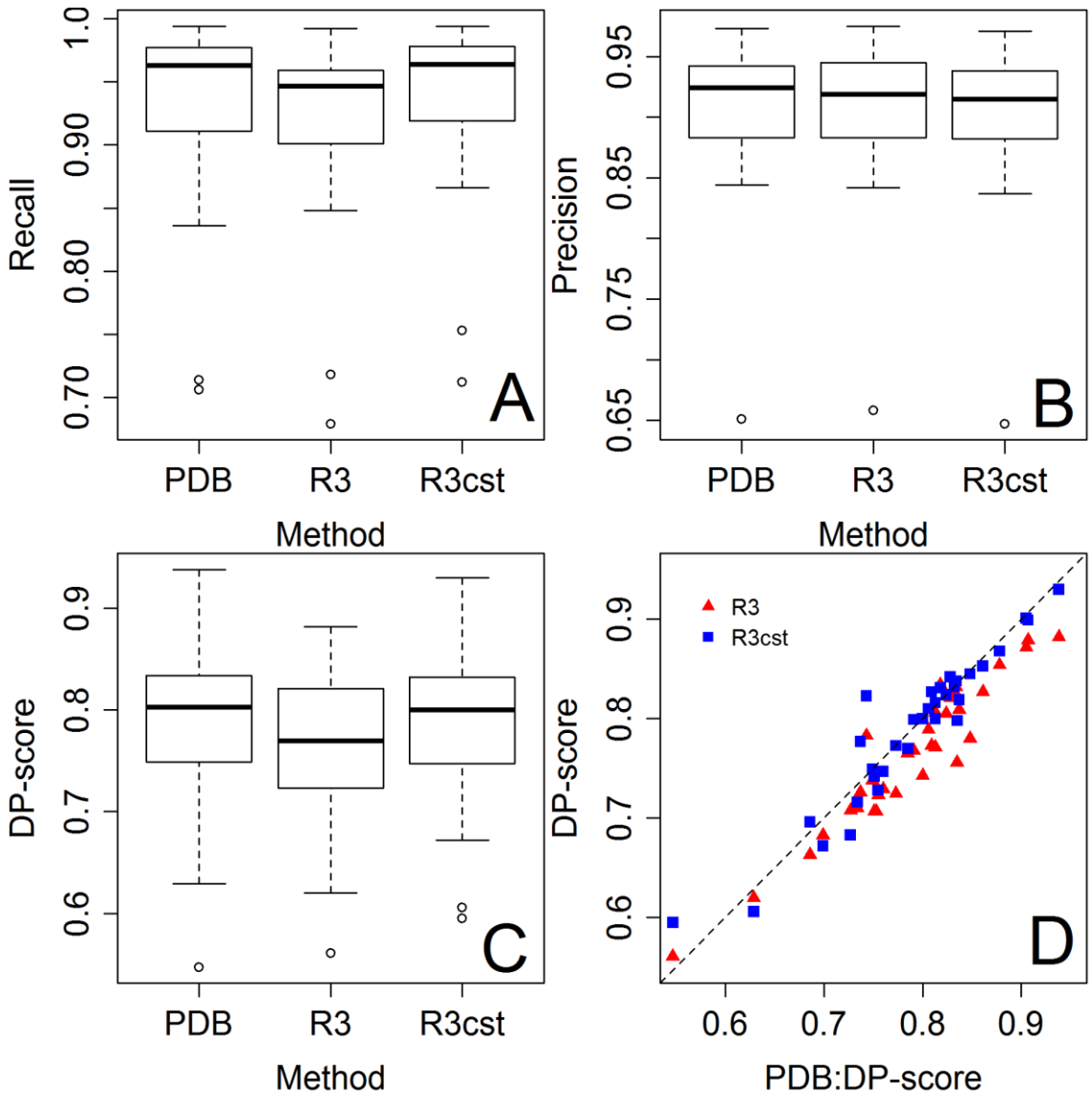

Figure 4.3. RPF analysis statistics.

PDB: NMR structures deposited in PDB; R3: Structures generated by unrestrained Rosetta refinement; R3cst: Structures generated by restrained Rosetta refinement. (A): Boxplots of Recall against different structure sources. (B): Boxplots of Precision against different structure sources. (C): Boxplots of DP-score against different structure sources. (D): 2-D DP-score scatterplot. DP-scores of the PDB NMR structures are plotted on the X-axis, while the DP-scores of both the unrestrained Rosetta refined structures represented by red solid triangle symbols(R3) and restrained Rosetta refined structures represented by blue solid rectangle symbols(R3cst) are plotted on the Y-axis. The black dashed line represents y=x.

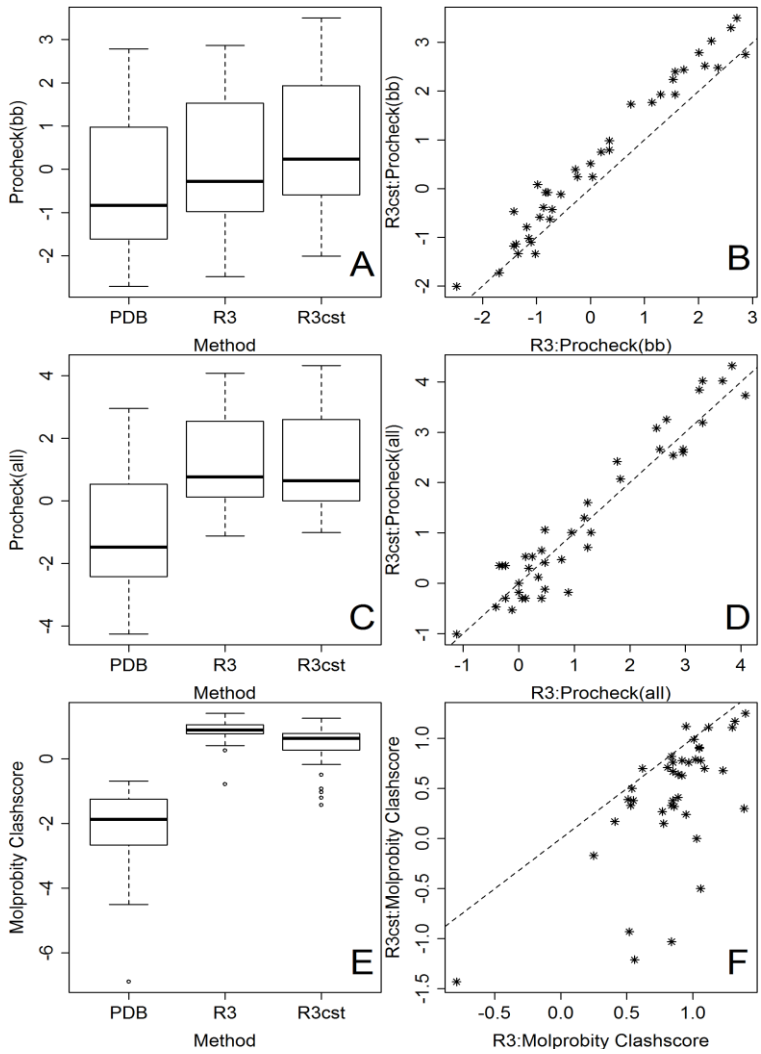

Figure 4.4. Boxplot of PSVS Z-scores and 2D satterplot of PSVS Z-scores

PDB: NMR structures deposited in PDB; R3: Structures generated by unrestrained Rosetta refinement; R3cst: Structures generated by restrained Rosetta refinement. (A): Boxplot of Procheck backbone dihedral angle G-factor Z-scores against different structure sources. (B): 2D scatterplot of Procheck backbone dihedral angle G-factor Z-scores. (C): Boxplot of Procheck all dihedral angle G-factor Z-scores against different structure sources. (D): 2D scatterplot of Procheck all dihedral angle G-factor Z-scores. (E): Boxplot of Molprobity clashscore Z-scores against different structure sources. (F): 2D scatterplot of Molprobity clashscore Z-scores.

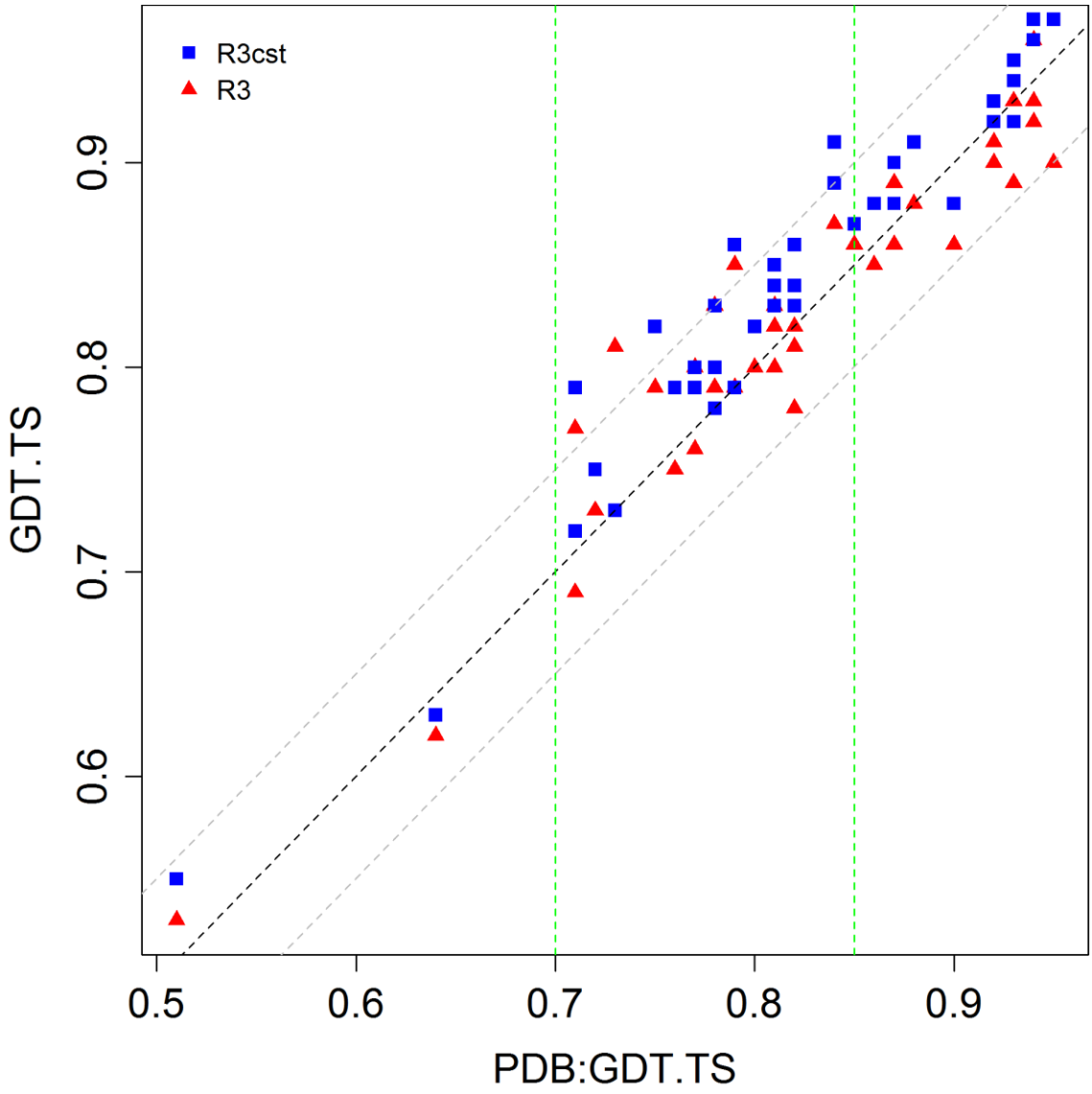

Figure 4.5. 2-D GDT.TS scores scatterplot.

GDT.TS values of PDB NMR structure to corresponding X-ray structure are plotted on the X-axis, while GDT.TS values of both unrestrained Rosetta refined structures(R3, represented by red solid triangle symbols) and restrained Rosetta refined structures(R3cst, represented by blue solid rectangles symbols) to their corresponding X-ray structures are plotted on the Y-axis. The two green dash lines indicate GDT.TS of PDB NMR structures equal to 0.7 and 0.85 respectively. The black dash line represents y=x, and the two gray dash lines represent y=x+0.05 and y=x-0.05 respectively.

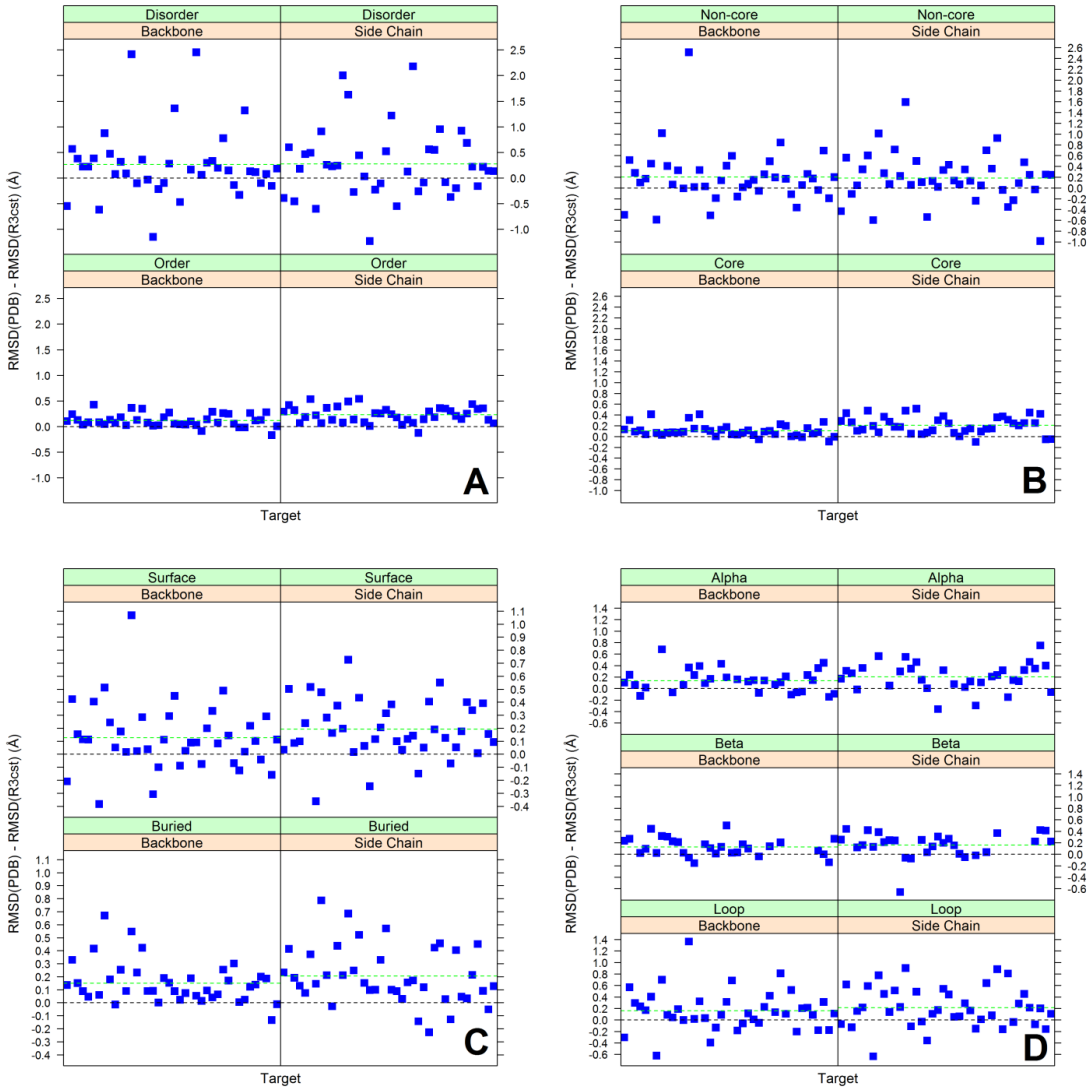

Figure 4.6. Plot of differences of RMSD to X-ray structures before and after restrained Rosetta refinement.

The target index is plotted on the X-axis, and the differences between the RMSD of PDB NMR structures to their corresponding X-ray structures and the RMSD of restrained Rosetta refined structures to their corresponding X-ray structures are plotted on the Y-axis. The four panels represent the RMSD differences for different subset of residues, which are ordered and disordered residues defined by S(phi)+S(psi)>=1.8 (A), core and non-core residues calculated by *FindCore* (B), buried and surface residues calculated by *areaimol* of CCP4 (C), alpha-helix, beta-sheet and loop residues calculated by DSSP (D). Each subset is further divided by atomic position- Backbone atoms and side chain atoms.

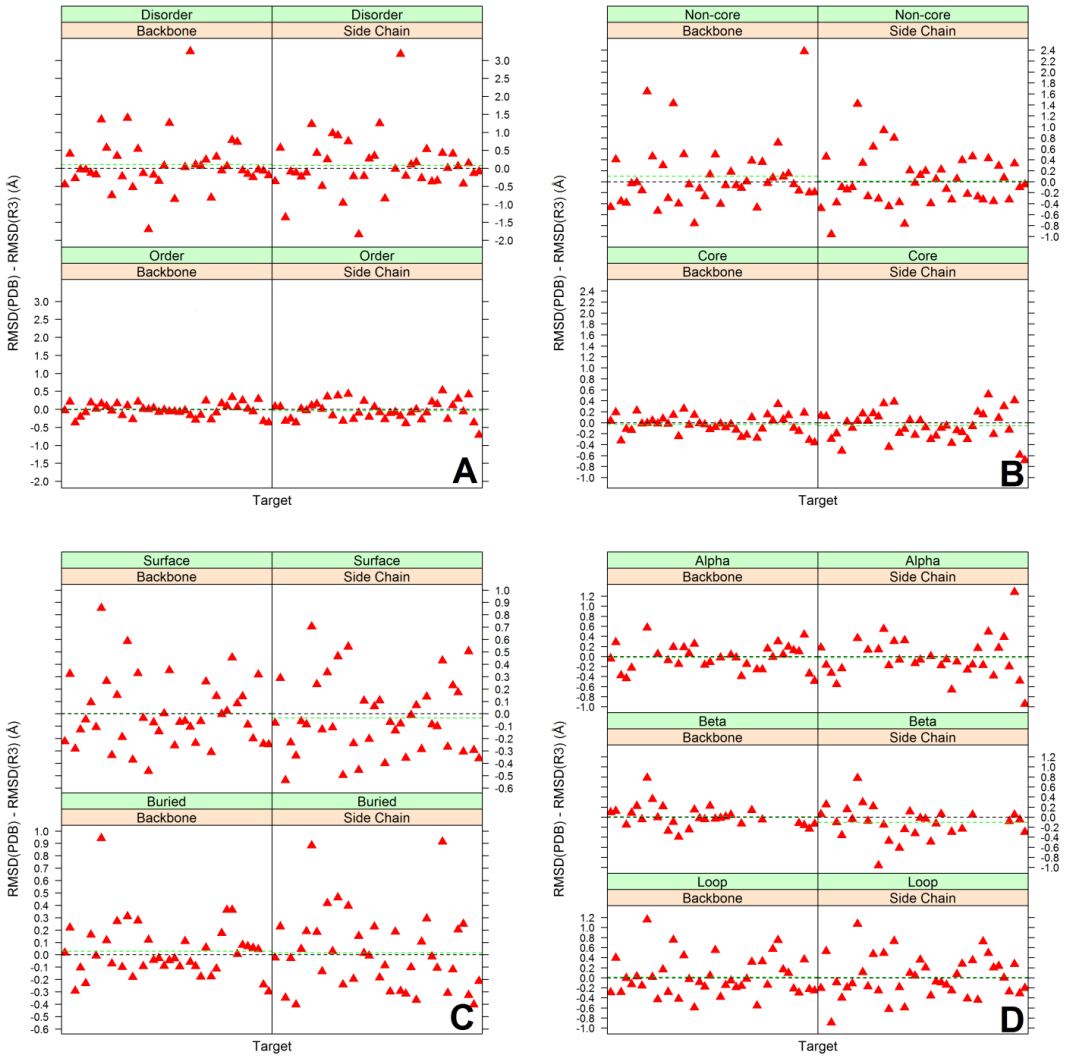

Figure 4.7. Plot of differences of RMSD to X-ray structures before and after unrestrained Rosetta refinement.

The target index is plotted on the X-axis, and the differences between the RMSD of PDB NMR structures to their corresponding X-ray structures and the RMSD of unrestrained Rosetta refined structures to their corresponding X-ray structures are plotted on the Y-axis. The four panels represent the RMSD differences for different subset of residues, which are ordered and disordered residues defined by S(phi)+S(psi)>=1.8 (A), core and non-core residues calculated by *FindCore* (B), buried and surface residues calculated by *areaimol* of CCP4 (C), alpha-helix, beta-sheet and loop residues calculated by DSSP (D). Each subset is further divided by atomic position: Backbone atoms and side chain atoms.

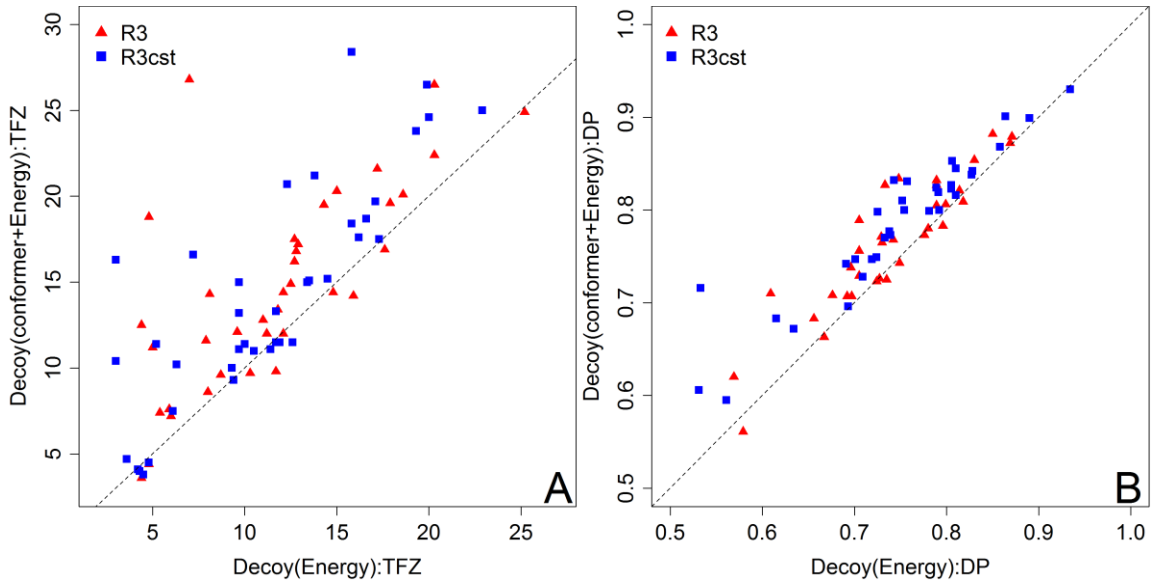

Figure 4.8. 2D scatterplot of TFZ scores (A) and DP-scores (B) for different model picking protocols.

Decoy(Energy): The final Rosetta refined structures are picked by Rosetta energy only. Decoy(Conformer+Energy): The final Rosetta refined structure is composed of the lowest Rosetta energy decoy generated from each NMR conformer. The scores of structures picked by Decoy(Energy) protocol are plotted on the X-axis, and the scores of structures picked by Decoy(Conformer+Energy) protocol are plotted on the Y-axis. Unrestrained Rosetta refined structures are represented by red solid triangle symbols and restrained Rosetta refined structures are represented by blue solid rectangle symbols.

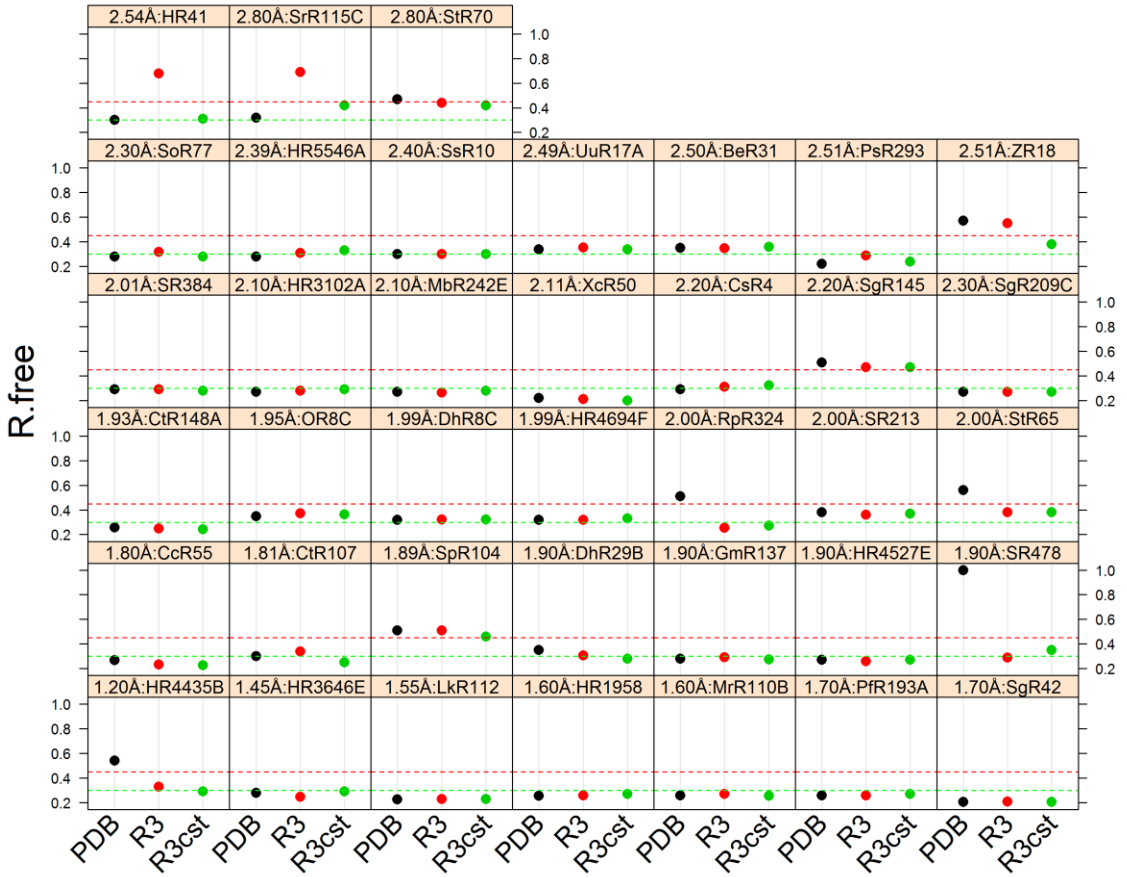

Fig. 4.9. Dotplot of R.free values of MR structures against the source of their MR templates for 38 NESG targets.

The MR structures are solved either by Phenix or Arp/WARP. PDB: NMR structures deposited in PDB; R3: Structures generated by unrestrained Rosetta refinement; R3cst: Structures generated by restrained Rosetta refinement. The R.free values are plotted on the Y-axis, each subpanel represents one NESG target, and the subpanels are organized in ascending order of the resolution of its X-ray crystal structure from bottom left corner to top right corner.

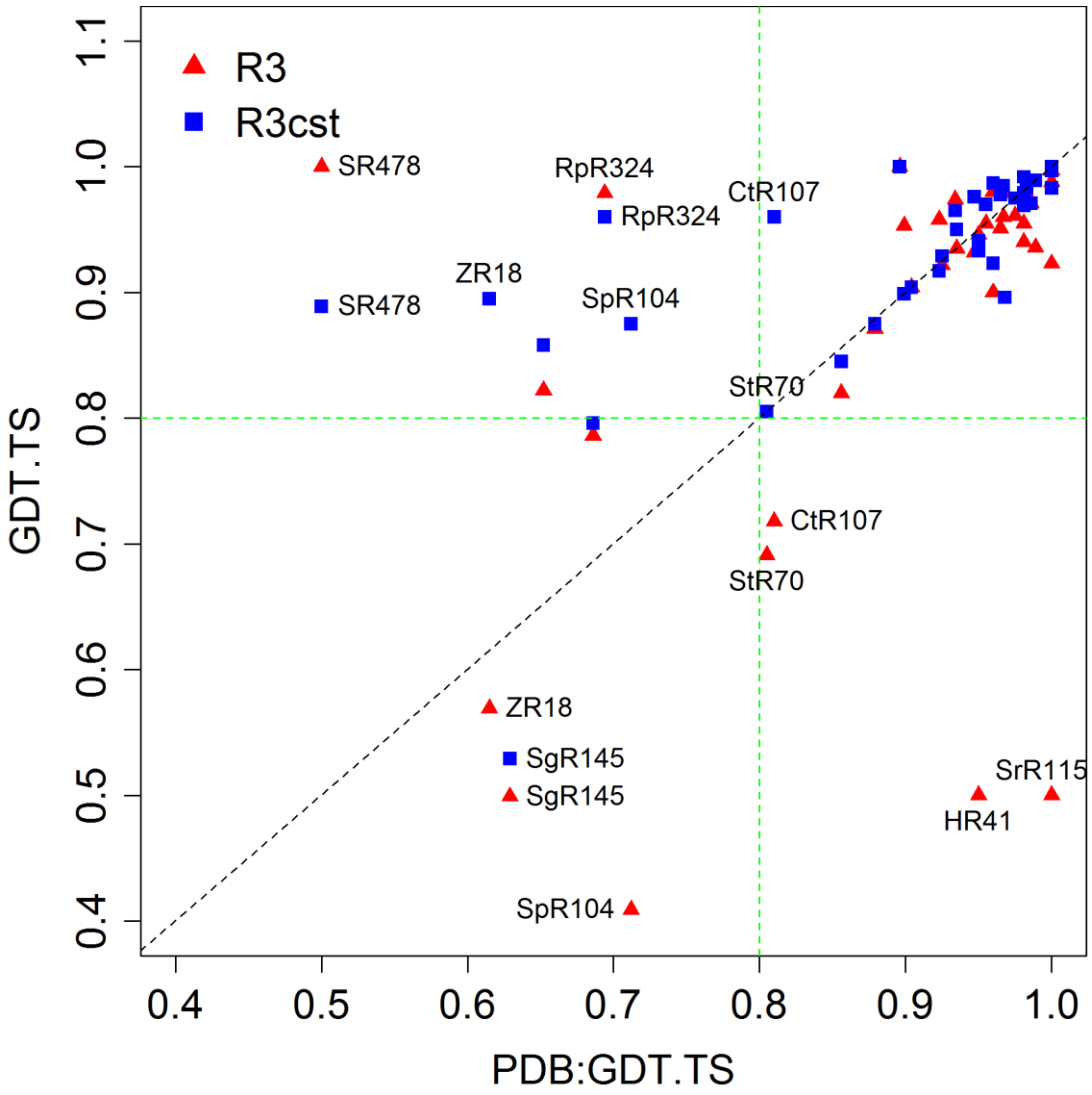

Figure 4.10. 2D GDT.TS scatterplot of MR structures to their corresponding X-ray structure

The GDT.TS values of MR structures solved by PDB NMR templates are plotted on the X-axis, and the GDT.TS values of MR structures solved by both unrestrained Rosetta refined structures (R3, represented by red solid triangle symbols) and restrained Rosetta refined structures (R3cst, represented by blue solid rectangle symbols) are plotted on the Y-axis. The black dashed line indicates y=x. We use a cutoff of GDT.TS=0.8 (represented by the two green dash lines) to classify the quality of MR structures.

REFERENCE

1. Rosato, A., Bagaria, A., Baker, D., Bardiaux, B., Cavalli, A., Doreleijers, J.F., Giachetti, A., Guerry, P., Güntert, P., Herrmann, T.*, et al.* (2009). CASD-NMR: critical assessment of automated structure determination by NMR. Nat Methods *6*, 625-626.

2. Rosato, A., Aramini, J.M., Arrowsmith, C., Bagaria, A., Baker, D., Cavalli, A., Doreleijers, J.F., Eletsky, A., Giachetti, A., Guerry, P.*, et al.* (2012). Blind testing of routine, fully automated determination of protein structures from NMR data. Structure *20*, 227-236.

3. Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A.*, et al.* (2008). Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci U S A *105*, 4685-4690.

4. Huang, Y.J., Powers, R., and Montelione, G.T. (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*, 1665-1674.

5. Raman, S., Huang, Y.J., Mao, B., Rossi, P., Aramini, J.M., Liu, G., Montelione, G.T., and Baker, D. (2010). Accurate automated protein NMR structure determination using unassigned NOESY data. J Am Chem Soc 132, 202-207.

6. Snyder, D.A., and Montelione, G.T. (2005). Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. Proteins 59, 673-686.

7. Mao, B., Guan, R., and Montelione, G.T. (2011). Improved technologies now routinely provide protein NMR structures useful for molecular replacement. Structure 19, 757-766.

8. Wüthrich, K. (1986). NMR of Proteins and Nucleic Acids (New York: Wiley).

9. Markwick, P.R., Malliavin, T., and Nilges, M. (2008). Structural biology by NMR: structure, dynamics, and interactions. PLoS Comput. Biol. 4, e1000168.

10. Güntert, P. (2009). Automated structure determination from NMR spectra. Eur. Biophys. J. 38, 129–143.

11. Guerry, P., and Herrmann, T. (2011). Advances in automated NMR protein structure determination. Q. Rev. Biophys. 44, 257–309.

12. Lopez-Mendez, B., and Güntert, P. (2006). Automated protein structure determination from NMR spectra. J Am Chem Soc *128*, 13112-13122.

13. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., and Bourne,P.E. (2000). The Protein Data Bank. Nucleic Acids Res. 28, 235-242.

14. Moseley,H.N., Monleon,D. & Montelione,G.T. (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data Methods Enzymol. 339, 91-108.

15. Bahrami,A., Assadi,A.H., Markley,J.L. & Eghbalnia,H.R. (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy PLoS. Comput. Biol. 5, e1000307.

16. Lemak,A. et al. (2011) A novel strategy for NMR resonance assignment and protein structure determination J. Biomol. NMR 49, 27-38.

17. Lemak,A., Steren,C.A., Arrowsmith,C.H. & Llinas,M. (2008) Sequence specific resonance assignment via Multicanonical Monte Carlo search using an ABACUS approach J. Biomol. NMR 41, 29-41.

18. Guntert,P., Mumenthaler,C. & Wüthrich,K. (1997) Torsion Angle Dynamics for NMR Structure Calculation with the new program DYANA. J. Mol. Biol. 273, 283-298.

19. Herrmann,T., Güntert,P. & Wüthrich,K. (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J. Mol. Biol. 319, 209-227.

20. Schwieters,C.D., Kuszewski,J., Tjandra,N. & Clore,G.M. (2003) The Xplor-NIH NMR molecular structure determination package J. Magn. Reson. 160, 65-73.

21. Cornilescu,G., Delaglio,F. & Bax,A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology J. Biomol. NMR 13, 289-302.

22. Shen,Y., Delaglio,F., Cornilescu,G. & Bax,A. (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts J. Biomol. NMR 44, 213-223.

23. Brunger,A.T. et al. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination Acta Crystallogr D Biol Crystallogr 54, 905-921.

24. Linge,J.P., Williams,M.A., Spronk,C.A.E.M., Bonvin,A.M.J.J. & Nilges,M. (2003) Refinement of protein structures in explicit solvent. Proteins 50, 496-506.

25. Cai,M. et al. (2007) Solution NMR structure of the barrier-to-autointegration factor-Emerin complex J. Biol. Chem. 282, 14525-14535.

26. Aramini,J.M. et al. (2010) Structural basis of O6-alkylguanine recognition by a bacterial alkyltransferase-like DNA repair protein J. Biol. Chem. 285, 13736-13741.

27. Liu,G., Huang,Y.J., Xiao,R., Wang,D., Acton,T.B. & Montelione,G.T. (2010) NMR structure of F-actin-binding domain of Arg/Abl2 from Homo sapiens Proteins 78, 1326-1330.

28.Rossi, P., Xiao, R., Acton, T. B., Rost, B., and Montelione, G. T. Solution NMR Structure of uncharacterized protein from gene locus NE0665 of Nitrosomonas europaea. Northeast Structural Genomics Target NeR103A. 2009.

29.Eletsky, A., Sathyamoorthy, B., Sukumaran, D. K., Wang, D., Buchwald, W. A., Ciccosanti, C., Janjua, H., Nair, R., Rost, B., Acton, T. B., Xiao, R., Everett, J. K., Montelione, G. T., and Szyperski, T. Solution NMR structure of the N-terminal domain of cg2496 protein from Corynebacterium glutamicum. Northeast Structural Genomics Consortium Target CgR26A. 2009.

30. He, Y., Eletsky, A., Lee, D., Ciccosanti, C., Janjua, H., Acton, T. B., Xiao, R., Everett, J. K., Montelione, G. T., and Szyperski, T. Solution NMR structure of the PCP_red domain of light-independent protochlorophyllide reductase subunit B from Chlorobium tepidum. Northeast Structural Genomics Consortium Target CtR69A. 2009.

31. Mani, R., Swapna, G. V., Shastry, R., Foote, E., Ciccosanti, C., Jiang, M., Xiao, R., Nair, R., Everett, J., Huang, Y. J., Acton, T. B., Rost, B., and Montelione, G. T. NMR Solution Structure of a Tubulin folding cofactor B obtained from Arabidopsis thaliana: Northeast Structural Genomics Consortium target AR3436A. 2009.

32. Gumanas, A., Lemak, A., Yee, A., Semesi, A., Fares, C., Montelione, G. T., and Arrowsmith, C. H. Solution structure of protein Atu0922 from A. tumefaciens. Northeast Structural Genomics Consortium target AtT13. Ontario Center for Structural Proteomics target ATC0905. 2009.

33. Wu,B. et al. (2010) NleG Type 3 effectors from enterohaemorrhagic Escherichia coli are U-Box E3 ubiquitin ligases PLoS. Pathog. 6, e1000960.

34. Yang, Y., Ramelot, T. A., Cort, J. R., Lee, D., Ciccosanti, C., Hamilton, K., Nair, R., Rost, B., Acton, T. B., Xiao, R., Swapna, G. V., Everett, J. K., Montelione, G. T., and Kennedy, M. A. Solution NMR structure of the TGS domain of PG1808 from Porphyromonas gingivalis. Northeast Structural Genomics Consortium Target PgR122A (418-481). 2009.

35. Bhattacharya,A., Tejero,R. & Montelione,G.T. (2007) Evaluating protein structures determined by structural genomics consortia Proteins 66, 778-795.

36. Ponder,J.W. and Case,D.A. (2003). Force fields for protein simulations. Adv. Prot. Chem. 66, 27-85.

37. Koradi,R., Billeter,M., and Güntert,P. (2000). Point-centered domain decomposition for parallel molecular dynamics simulation. Computer Physics Communications *124*, 139-147.

38. Luginbühl,P., Güntert,P., Billeter,M., and Wüthrich,K. (1996). The new program OPAL for molecular dynamics simulations and energy refinements of biological macromolecules. J. Biomol. NMR *8*, 136-146.

39. Herrmann,T., Güntert,P., and Wüthrich,K. (2002a). Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J. Mol. Biol. 319, 209-227.

40. Herrmann,T., Güntert,P., and Wüthrich,K. (2002b). Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. J. Biomol. NMR 24, 171-189.

41. Spera,S. and Bax,A. (1991). Empirical Correlation between Protein Backbone Conformation and C_α and C_β ^{13}C Nuclear Magnetic Resonance Chemical Shifts. J. Am. Chem. Soc. 113, 5490-5492.

42. Luginbühl,P., Szyperski,T., and Wüthrich,K. (1995). Statistical Basis for the Use of $^{13}C_\alpha$ Chemical-Shifts in Protein-Structure Determination. J. Magn. Reson. Ser. B *109*, 229-233.

43. Wang,L., Eghbalnia,H.R., Bahrami,A., and Markley,J.L. (2005). Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. J. Biomol. NMR 32, 13-22.

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

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

46. Rieping,W., Habeck,M., and Nilges,M. (2005). Modeling errors in NOE data with a log-normal distribution improves the quality of NMR structures. J. Am. Chem. Soc. *127*, 16026-16027.

47. Habeck,M., Rieping,W., and Nilges,M. (2006). Weighting of experimental evidence in macromolecular structure determination. Proc. Natl. Acad. Sci. U. S. A *103*, 1756-1761.

48. Nilges,M., Bernard,A., Bardiaux,B., Malliavin,T., Habeck,M., and Rieping,W. (2008). Accurate NMR structures through minimization of an extended hybrid energy. Structure. *16*, 1305-1312.

49. Rieping,W., Habeck,M., Bardiaux,B., Bernard,A., Malliavin,T.E., and Nilges,M. (2007). ARIA2: Automated NOE assignment and data integration in NMR structure calculation. Bioinformatics 23, 381-382.

50. Cavalli,A., Salvatella,X., Dobson,C.M., and Vendruscolo,M. (2007). Protein structure determination from NMR chemical shifts. Proc. Natl. Acad. Sci. USA *104*, 9615-9620.

51. Bonvin,A.M.J.J., Rosato,A., and Wassenaar,T. (2010). The eNMR platform for structural biology. J. Struct. Funct. Genomics 11, 1-8.

52. Zemla,A. (2003). LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. *31*, 3370-3374.

53. Bernard,A., Vranken,W.F., Bardiaux,B., Nilges,M., and Malliavin,T.E. (2011). Bayesian estimation of NMR restraint potential and weight: A validation on a representative set of protein structures. Proteins 79, 1525-1537.

54. Eisenberg,D., Luthy,R., and Bowie,J.U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. *277*, 396-404.

55. Sippl,M.J. (1993). Recognition of Errors in the Three-Dimensional Structures. Proteins Struct. Funct. Genet. 17, 355-362.

56. Laskowski,R.A., Rullmann,J.A.C., MacArthur,M.W., Kaptein,R., and Thornton,J.M. (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. J. Biomol. NMR 8, 477-486.

57. Davis,I.W., Leaver-Fay,A., Chen,V.B., Block,J.N., Kapral,G.J., Wang,X., Murray,L.W., Arendall,W.B., III, Snoeyink,J., Richardson,J.S., and Richardson,D.C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 35, W375-W383.

58. Saccenti,E. and Rosato,A. (2008). The war of tools: how can NMR spectroscopists detect errors in their structures? J. Biomol. NMR 40, 251-261.

59. Guntert, P. (2009). Automated structure determination from NMR spectra. Eur Biophys J 38, 129-143.

60. Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A.J., Read, R.J., and Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. Nature 450, 259-264.

61. Canutescu, A.A., and Dunbrack, R.L., Jr. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci *12*, 963-972.

62. Bradley, P., Malmstrom, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D.E., Meiler, J., Misura, K.M., and Baker, D. (2005). Free modeling with Rosetta in CASP6. Proteins *61 Suppl 7*, 128-134.

63. Ramelot, T.A., Raman, S., Kuzin, A.P., Xiao, R., Ma, L.C., Acton, T.B., Hunt, J.F., Montelione, G.T., Baker, D., and Kennedy, M.A. (2009). Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study. Proteins 75, 147-167.

64. Bradley, P., and Baker, D. (2006). Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. Proteins 65, 922-929.

65. Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T., and Gerstein, M. (2001). SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. Nucleic Acids Res 29, 2884-2898.

66. Goh, C.S., Lan, N., Echols, N., Douglas, S.M., Milburn, D., Bertone, P., Xiao, R., Ma, L.C., Zheng, D., Wunderlich, Z., et al. (2003). SPINE 2: a system for collaborative structural proteomics within a federated database framework. Nucleic Acids Res 31, 2833-2838.

67. Woolfson, M.M. (1971). Direct methods in crystallography. Rep Prog Phys, 369-434.

68. Pahler, A., Smith, J.L., and Hendrickson, W.A. (1990). A probability representation for phase information from multiwavelength anomalous dispersion. Acta Crystallogr A 46 (Pt 7), 537-540.

69. Hendrickson, W.A. (1991). Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science 254, 51-58.

70. Green, D.W., Jngram, V.M., and Perutz, M. F. (1954). The structure of haemoglobin IV. Sign determination by the isomorphous replacement method. Proc Roy Soc A225 (1954), 287-307.

71. Perutz, M.F. (1956). Isomorphous replacement and phase determination in noncentrosymmetric space groups. Acta Cryst 9 (1956), 867-873.

72. Blow, D.M., and Rossmann, M.G. (1961). the single isomorphous replacement method. Acta Cryst 14 (1961), 1195-1202.

73. Rossmann, M.G. (1972). The Molecular Replacement Method. Godon & Breach, New York.

74. Rossmann, M.G., and Arnold, E. (1993). Patterson and molecular-replacement techniques. International Tables for Crystallography Vol. B, 230-263.

75. Rossmann, M.G., and Blow, D.M. (1962). The detection of sub-units within the crystallographic asymmetric unit. Acta Cryst 15, 24-31.

76. Evans, P., and McCoy, A. (2008). An introduction to molecular replacement. Acta Crystallogr D Biol Crystallogr 64, 1-10.

77. Liu, J., Montelione, G.T., and Rost, B. (2007). Novel leverage of structural genomics. Nat Biotechnol 25, 849-851.

78. Burley, et al., 2008) S.K.; Joachimiak, A.; Montelione, G.T.; Wilson, I.A. Contributions to the NIH Protein Structure Initiative from the four large-scale production centers. Structure 2008, 16: 5 - 11.

79. Nair, R., Liu, J., Soong, T.T., Acton, T.B., Everett, J.K., Kouranov, A., Fiser, A., Godzik, A., Jaroszewski, L., Orengo, C., et al. (2009). Structural genomics is the largest contributor of novel structural leverage. J Struct Funct Genomics 10, 181-191.

80. Chivian, D., Kim, D.E., Malmstrom, L., Bradley, P., Robertson, T., Murphy, P., Strauss, C.E., Bonneau, R., Rohl, C.A., and Baker, D. (2003). Automated prediction of CASP-5 structures using the Robetta server. Proteins 53 Suppl 6, 524-533.

81. Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U., and Sali, A. (2006). Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics Chapter 5, Unit 5 6.

82. Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 69 Suppl 8, 108-117.

83. Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H.M., Jones, D., Brenner, S.E., Burley, S.K., Das, R., Dokholyan, N.V., et al. (2009). Outcome of a workshop on applications of protein models in biomedical research. Structure 17, 151-159.

84. Kleywegt, G.J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K., and Jones, T.A. (1994). Crystal structures of cellular retinoic acid binding proteins I and II in complex with all-trans-retinoic acid and a synthetic retinoid. Structure 2, 1241-1258.

85. Leahy, D.J., Axel, R., and Hendrickson, W.A. (1992). Crystal structure of a soluble form of the human T cell coreceptor CD8 at 2.6 A resolution. Cell 68, 1145-1162.

86. Muller, T., Oehlenschlager, F., and Buehner, M. (1995). Human interleukin-4 and variant R88Q: phasing X-ray diffraction data by molecular replacement using X-ray and nuclear magnetic resonance models. J Mol Biol 247, 360-372.

87. Anderson, D.H., Weiss, M.S., and Eisenberg, D. (1996). A challenging case for protein crystal structure determination: the mating pheromone Er-1 from Euplotes raikovi. Acta Crystallogr D Biol Crystallogr 52, 469-480.

88. Baldwin, E.T., Weber, I.T., St Charles, R., Xuan, J.C., Appella, E., Yamada, M., Matsushima, K., Edwards, B.F., Clore, G.M., Gronenborn, A.M., et al. (1991). Crystal structure of interleukin 8: symbiosis of NMR and crystallography. Proc Natl Acad Sci U S A 88, 502-506.

89. Wilmanns, M., and Nilges, M. (1996). Molecular replacement with NMR models using distance-derived pseudo B factors. Acta Crystallogr D Biol Crystallogr 52, 973-982.

90. Jogl, G., Tao, X., Xu, Y., and Tong, L. (2001). COMO: a program for combined molecular replacement. Acta Crystallogr D Biol Crystallogr 57, 1127-1134.

91. Navaza, J. (2001). Implementation of molecular replacement in AMoRe. Acta Crystallogr D Biol Crystallogr 57, 1367-1372.

92. Vagin, A., and Teplyakov, A. (2000). An approach to multi-copy search in molecular replacement. Acta Crystallogr D Biol Crystallogr 56, 1622-1624.

93. Kissinger, C.R., Gehlhaar, D.K., and Fogel, D.B. (1999). Rapid automated molecular replacement by evolutionary search. Acta Crystallogr D Biol Crystallogr 55, 484-491.

94. Glykos, N.M., and Kokkinidis, M. (2000). A stochastic approach to molecular replacement. Acta Crystallogr D Biol Crystallogr 56, 169-174.

95. Jamrog, D.C., Zhang, Y., and Phillips, G.N., Jr. (2003). SOMoRe: a multi-dimensional search and optimization approach to molecular replacement. Acta Crystallogr D Biol Crystallogr 59, 304-314.

96. Keegan, R.M., and Winn, M.D. (2008). MrBUMP: an automated pipeline for molecular replacement. Acta Crystallogr D Biol Crystallogr 64, 119-124.

97. Mccoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J Appl Crystallogr 40, 658-674.

98. Brunger, A.T., Campbell, R.L., Clore, G.M., Gronenborn, A.M., Karplus, M., Petsko, G.A., and Teeter, M.M. (1987). Solution of a Protein Crystal-Structure with a Model Obtained from Nmr Interproton Distance Restraints. Science 235, 1049-1053.

99. Chen, Y.W., Dodson, E.J., and Kleywegt, G.J. (2000). Does NMR mean "not for molecular replacement"? Using NMR-based search models to solve protein crystal structures. Structure *8*, R213-220.

100. Chen, Y.W., and Clore, G.M. (2000). A systematic case study on using NMR models for molecular replacement: p53 tetramerization domain revisited. Acta Crystallogr D Biol Crystallogr 56, 1535-1540. Standard journal name

101. Chen, Y.W. (2001). Solution solution: using NMR models for molecular replacement. Acta Crystallogr D Biol Crystallogr 57, 1457-1461.

102. Kim, S., and Szyperski, T. (2003). GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. J Am Chem Soc 125, 1385-1393.

103. Ulrich, E.L., Akutsu, H., Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., et al. (2008). BioMagResBank. Nucleic Acids Res 36, D402-408.

104. Hyberts, S.G., Goldberg, M.S., Havel, T.F., and Wagner, G. (1992). The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. Protein Sci 1, 736-751.

105. Kelley, L.A., Gardner, S.P., and Sutcliffe, M.J. (1997). An automated approach for defining core atoms and domains in an ensemble of NMR-derived protein structures. Protein Eng 10, 737-741.

106. Perrakis, A., Harkiolaki, M., Wilson, K.S., and Lamzin, V.S. (2001). ARP/wARP and molecular replacement. Acta Crystallogr D Biol Crystallogr 57, 1445-1450.

107. Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53, 240-255.

108. Terwilliger, T.C., Grosse-Kunstleve, R.W., Afonine, P.V., Moriarty, N.W., Zwart, P.H., Hung, L.W., Read, R.J., and Adams, P.D. (2008). Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Crystallogr D Biol Crystallogr 64, 61-69.

109. Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 60, 2126-2132.

110. Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins 57, 702-710.

111. Kim, D.E., Chivian, D., and Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 32, W526-531.

112. Misura, K.M., and Baker, D. (2005). Progress and challenges in high-resolution refinement of protein structure models. Proteins 59, 15-29.

113. Giorgetti, A., Raimondo, D., Miele, A.E., and Tramontano, A. (2005). Evaluating the usefulness of protein structure models for molecular replacement. Bioinformatics 21 Suppl 2, ii72-76.

114. Lee, B., and Richards, F.M. (1971). JMolBiol 55, 379-400.

115. Saff, E.B., and Kuijlaars, A.B.J. (1997). The Mathematical Intelligencer 19, 5-11.

116. Das, K., Ma, L.C., Xiao, R., Radvansky, B., Aramini, J., Zhao, L., Marklund, J., Kuo, R.L., Twu, K.Y., Arnold, E., et al. (2008). Structural basis for suppression of a host antiviral response by influenza A virus. Proc Natl Acad Sci U S A 105, 13093-13098.

117. Aramini, J.M., Ma, L., Lee, H., Zhao, L., Cunningham, K., Ciccosanti, C., Janjua, H., Fang, Y., Xiao, R., Krug, R.M., Montelione, G.T. (2009). Solution NMR structure of the monomeric W187R mutant of A/Udorn NS1 effector domain. Northeast Structural Genomics target OR8C[W187R].

118. Szymczyna, B.R., Taurog, R.E., Young, M.J., Snyder, J.C., Johnson, J.E., and Williamson, J.R. (2009). Synergy of NMR, computation, and X-ray crystallography for structural biology. Structure 17, 499-507.

119. DeLano, W.L. (2002). The PyMOL Molecular Graphics System (San Carlos, CA, USA., DeLano Scientific).

120. Guntert, P. (2004). Automated NMR structure calculation with CYANA. Methods Mol Biol 278, 353-378.

121. Bashford, D., and Case, D.A. (2000). Generalized born models of macromolecular solvation effects. Annu Rev Phys Chem 51, 129-152.

122. Xia, B., Tsui, V., Case, D.A., Dyson, H.J., and Wright, P.E. (2002). Comparison of protein solution structures refined by molecular dynamics simulation in vacuum, with a generalized Born model, and with explicit water. J Biomol NMR 22, 317-331.

123. Feig, M., and Brooks, C.L., 3rd (2004). Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr Opin Struct Biol 14, 217-224.

124. Chen, J., Im, W., and Brooks, C.L., 3rd (2004). Refinement of NMR structures using implicit solvent and advanced sampling techniques. J Am Chem Soc 126, 16038-16047.

125. Chen, J., Brooks, C.L., 3rd, and Khandogin, J. (2008). Recent advances in implicit solvent-based methods for biomolecular simulations. Curr Opin Struct Biol 18, 140-148.

126. Nabuurs, S.B., Nederveen, A.J., Vranken, W., Doreleijers, J.F., Bonvin, A.M., Vuister, G.W., Vriend, G., and Spronk, C.A. (2004). DRESS: a database of REfined solution NMR structures. Proteins 55, 483-486.

127. Nederveen, A.J., Doreleijers, J.F., Vranken, W., Miller, Z., Spronk, C.A., Nabuurs, S.B., Guntert, P., Livny, M., Markley, J.L., Nilges, M., et al. (2005). RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. Proteins 59, 662-672.

128. Lee, S.Y., Zhang, Y., and Skolnick, J. (2006). TASSER-based refinement of NMR structures. Proteins 63, 451-456.

129. Yang, J.S., Kim, J.H., Oh, S., Han, G., Lee, S., and Lee, J. (2012). STAP Refinement of the NMR database: a database of 2405 refined solution NMR structures. Nucleic Acids Res 40, D525-530.

130. Nabuurs, S.B., Spronk, C.A.E.M., Vriend, G. and Vuister, G.W. (2004). Concepts and tools for NMR restraint analysis and validation. Concepts in Magnetic Resonance, 90-105.

131. Morris, A.L., MacArthur, M.W., Hutchinson, E.G., and Thornton, J.M. (1992). Stereochemical quality of protein structure coordinates. Proteins 12, 345-364.

132. Simons, K.T., Bonneau, R., Ruczinski, I., and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl 3, 171-176.

133. Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268, 209-225.

134. Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364-1368.

135. DiMaio, F., Tyka, M.D., Baker, M.L., Chiu, W., and Baker, D. (2009). Refinement of protein structures into low-resolution density maps using rosetta. J Mol Biol 392, 181-190.

136. Terwilliger, T.C., Dimaio, F., Read, R.J., Baker, D., Bunkoczi, G., Adams, P.D., Grosse-Kunstleve, R.W., Afonine, P.V., and Echols, N. (2012). phenix.mr_rosetta: molecular replacement and model rebuilding with Phenix and Rosetta. J Struct Funct Genomics 13, 81-90.

137. Lange, O.F., Rossi, P., Sgourakis, N.G., Song, Y., Lee, H.W., Aramini, J.M., Ertekin, A., Xiao, R., Acton, T.B., Montelione, G.T., et al. (2012). Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc Natl Acad Sci U S A 109, 10873-10878.

138. Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577-2637.

139. Joosten, R.P., te Beek, T.A., Krieger, E., Hekkelman, M.L., Hooft, R.W., Schneider, R., Sander, C., and Vriend, G. (2011). A series of PDB related databases for everyday needs. Nucleic Acids Res 39, D411-419.

CURRICULUM VITA

Binchen Mao

2006-2012 Ph.D. joint program of Molecular Biosciences at Rutgers University and University of Medicine and Dentistry of New Jersey (UMDNJ)

2003-2006 M.S. in Biochemistry, Nanjing University

1999-2003 B.S. in Biochemistry, Nanjing University

Publications

1. Mao, B.; Guan, R.; Montelione, G.T. Improved technologies now routinely provide protein NMR structures useful for molecular replacement. Structure 2011, 19: 757 - 766.

2. Raman, S.; Huang, Y.J.; Mao, B.; Rossi, P.; Aramini, J.M.; Liu, G.; Montelione, G.T.; Baker, D. Accurate automated protein NMR structure determination using unassigned NOESY data. J. Am. Chem. Soc. 2010, 132: 202 - 207.

3. Rosato, A.; Bagaria, A.; Baker, D.; Bardiaux, B.; Cavalli, A.; Doreleijers, J.F.; Giachetti, A.; Guerry, P.; Güntert, P.; Herrmann, T.; Huang, Y.J.; Jonker, H.R.A.; 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.J.J. CASD-NMR: critical assessment of automated structure determination by NMR. Nature Methods 2009, 6: 625 - 626.

4. Rosato, A.; Aramini, J.M.; Arrowsmith, C.; Bagaria, A.; Baker, D.; Cavalli, A.; Doreleijers, J.F.; Eletsky, A.; Giachetti, A.; Guerry, P.; Gutmanas, A.; Güntert, P.; He, Y.; Herrmann, T.; Huang, Y.J.; Jaravine, V.; Jonker, H.R.A.; Kennedy, M.A.; Lange, O.F.; Liu G.; Malliavan, T.E.; Mani, R.; Mao, B.; Montelione, G.T.; Nilges, M.; Possi, P.; van der Schot, G.; Schwalbe, H.; Szyperski, T.A.; Vendruscolo, M.; Vernon, R.; Vranken, W.F.; de Vries, S.; Vuister, G.W.; Wu, B.; Yang, Y.; Bonvin, A.M.J.J. Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 2012, 20: 227 – 236.

5. Kobayashi, H., Swapna, G.V., Wu, K.P., Afinogenova, Y., Conover, K., Mao, B., Montelione, G.T., and Inouye, M. (2012). Segmental isotope labeling of proteins for NMR structural study using a protein S tag for higher expression and solubility. J Biomol NMR *52*, 303-313.

6. PDBStat: Constraint Analysis Software for Protein NMR. R Tejero , D Snyder, B Mao, G.T. Montelione. J Biomol NMR (In Preparation)

7. Protein Structure Validation Software. GT Montelione, J. Block, B Mao, Y.P. Huang, L Ferella, A Rosato. J Biomol NMR (In preparation)

8. Improve the quality of protein NMR structure by restrained Rosetta refinement. B Mao , R Tejero, G.T. Montelione. J. Am. Chem. Soc. (In prepration)