

Technische Universität München

Lehrstuhl für biomolekulare NMR-Spektroskopie

Department Chemie

Rapid and automatic structure determination using sparse NMR data and Rosetta

Zaiyong Zhang

Vollständiger Abdruck der von der Fakultät für Chemie der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.- Prof. Dr. Ville R. I. Kaila

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Michael Sattler

2. TUM Junior Fellow Dr. Tobias Madl

Die Dissertation wurde am 24.06.2014 bei der Technischen Universität München eingereicht und durch die Fakultät für Chemie am 15.09.2014 angenommen.

Brüder, ich denke von mir selbst nicht, ergriffen zu haben; eines aber: Ich vergesse, was dahinten, strecke mich aber aus nach dem, was vorn ist, und jage auf das Ziel zu, hin zu dem Kampfpreis der Berufung Gottes nach oben in Christus Jesus.

Philipper 3:13-14

Content

DECLARATION	6
Abstract	7
Zusammenfassung	8
Chapter 1 Introduction	9
1.1 Protein structure determination methods	9
1.2 Nuclear magnetic resonance (NMR).....	11
1.2.1 NMR spectroscopy for protein determination.....	11
1.2.2 NMR chemical shift assignment.....	13
1.2.3 Nuclear Overhauser effect (NOE) assignment	15
1.3 Macromolecular modeling program Rosetta	16
1.3.1 Fragment picking based on protein sequence and chemical shifts	17
1.3.2 Rosetta score	20
1.3.3 Fragment assembly by Monte Carlo method	22
1.3.4 RASREC protocol	23
1.4 Aims of the present thesis.....	25
1.4.1 Structure prediction by Rosetta from chemical shift data.....	26
1.4.2 New algorithm for automatic NOESY assignment and structure determination with Rosetta.....	26
1.4.3 Performance of automatic NOESY assignment and structure determination algorithms with scramble chemical shift data.....	27
1.5 References.....	28
Chapter 2 Improving 3D structure prediction from chemical shift data	35
2.1 Introduction	35
2.2 Materials and Methods.....	36
2.2.1 Target Selection and Fragment Picking.....	37
2.2.2 Structure Generation.....	38

2.2.3 Calculation of converged regions.....	39
2.2.4 RMSD Calculations.....	39
2.2.5 Criteria used for annotations.....	40
2.2.6 Classification of 3D structure predictions.....	41
2.2.7 Weak/strong-classification with CS-Rosetta toolbox.....	42
2.3 Results.....	43
2.3.1 Performance of new fragment picker (R3FP).....	43
2.3.2 RASREC with chemical shift rescoring.....	45
2.3.3 Restriction to converged regions.....	45
2.3.4 Reliability measure: Annotation of weak/strong predictions.....	51
2.3.5 The WeNMR CS-ROSETTA web server.....	58
2.4 Discussion.....	59
2.5 References.....	61
2.6 My contribution to this project.....	63
Chapter 3 Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta.....	64
3.1 Significance Statement.....	64
3.2 Introduction.....	64
3.3 Results.....	67
3.4 Discussion.....	78
3.5 Methods.....	79
3.5.1 Benchmark.....	79
3.5.2 AutoNOE-Rosetta.....	80
3.5.3 Cyana structure calculations.....	81
3.6 References.....	82
3.7 My contribution to this project.....	85
Chapter 4 Effect of incorrect chemical shift assignments on automated NOE assignments and NMR structure calculation.....	86

4.1 Introduction	86
4.2 Materials and Methods.....	88
4.2.1 Preparation of benchmark datasets	88
4.2.2 Datasets with incomplete chemical shift assignments	89
4.2.3 Datasets with swapped chemical shift assignments	89
4.2.4 Datasets with combined chemical shift assignments.....	90
4.2.5 Structure generation with CYANA.....	90
4.2.6 Structure generation with AutoNOE-Rosetta	90
4.2.7 Structure generation with ASDP	91
4.3 Results	91
4.3.1 Effect of missing chemical shift assignments.....	94
4.3.2 Effect of swapping chemical shift assignments.....	96
4.3.3 Effect of combining chemical shift assignments	99
4.3.4 Effect of missing resonances with low-fidelity assignments.....	100
4.3.5 Indicating problematic runs of AutoNOE-Rosetta	102
4.4 Conclusion	106
4.5 References.....	108
Chapter 5 Conclusion and Discussion	111
Appendix.....	113
A.1 Supplementary Figures.....	113
A.2 Supplementary Tables.....	126
A.3 Supplementary Methods.....	161
A.3.1 AutoNOE-Rosetta calculations	161
A.3.2 Checking for unsuccessful calculations	166
A.3.3 Robustness of AutoNOE-Rosetta in repeated runs	168
A.4 References	168
Acknowledgement	170

DECLARATION

I hereby declare that parts of this Thesis are already published in scientific journal:

[1] G. Schot, **Z. Zhang**, R. Vernon, Y. Shen, W. F. Vranken, D. Baker, A. M. J. J. Bonvin, and O. F. Lange, “Improving 3D structure prediction from chemical shift data,” J Biomol NMR, 2013.

[2] **Z. Zhang**, J. Porter, K. Tripsianes and O. F. Lange, “Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta,” J Biomol NMR, 2014.

Abstract

During my PhD study, I improved the performance of CS-Rosetta on protein structure determination by following routes: 1) I reported advances in the calculation of protein structures from chemical shift NMR data alone. I demonstrated that combination of a new and improved fragment picker and the iterative sampling algorithm RASREC yield significant improvements in convergence and accuracy. Moreover, I introduced improved criteria for assessing the accuracy of the models produced by the method. 2) I benchmarked the performance of AutoNOE-Rosetta, a novel and robust approach for automatic and unsupervised simultaneous nuclear overhauser effect (NOE) assignment and structure determination within the CS-Rosetta framework on 50 protein targets ranging from 50 to 200 residues in size. The approach proved to be able to tolerate incomplete and raw NOE peak lists as well as incomplete or partially incorrect chemical shift assignments. 3) I studied the effect of incomplete and erroneous chemical shifts on automatic NOE assignments and protein structure determinations. With 3 automatic NOE assignment and protein de novo programs CYANA, ASDP and AutoNOE-Rosetta, the test was carried out on a benchmark of three proteins and 10 typical kinds of problems in chemical shift assignments.

Zusammenfassung

Während meiner Doktorarbeit verbesserte ich die Leistung von CS-Rosetta bei der Proteinstrukturvorhersage auf folgende Wege: 1) Ich verbesserte die Berechnung von Proteinstrukturen mit Hilfe von chemischen Verschiebungen der NMR-Spektroskopie. Ich zeigte, dass die Kombination eines neuen und verbesserten Fragment Pickers und dem iterativen Sampling-Algorithmus RASREC zu signifikanten Verbesserungen von Konvergenz und Präzision führen. Des Weiteren habe ich ein verbessertes Kriterium eingeführt um die Genauigkeit der von der Methode generierten Strukturmodelle zu beurteilen. 2) Ich habe die Effizienz von AutoNOE-Rosetta, einem neuen und robusten Ansatz zur automatischen und nicht überwachten Zuweisung des Nuclear-Overhauser-Effekts (NOE) und gleichzeitigen Strukturbestimmung innerhalb des CS-Rosetta Frameworks, mit 50 verschiedenen Proteinen (mit Proteinlängen von 50 bis 200 Aminosäuren) bewertet. Es konnte gezeigt werden, dass sowohl unvollständige und unbearbeitete NOE Peak Listen als auch unvollständige und teilweise fehlerhafte Zuweisungen der chemischen Verschiebung von der Methode toleriert werden. 3) Ich untersuchte den Einfluss von unvollständigen und fehlerhaften chemischen Verschiebungen auf die automatische Zuweisung des Nuclear-Overhauser-Effekts und auf Proteinstrukturvorhersagen. Hierfür wurden die drei Programme zur automatischen NOE-Zuweisung und de novo Proteinstrukturvorhersage CYANA, ASDP und AutoNOE-Rosetta auf einem Benchmark Datensatz von drei Proteinen und 10 typischen Problemen in der Zuweisung von chemischen Verschiebungen getestet.

Chapter 1 Introduction

1.1 Protein structure determination methods

Proteins are large molecules made up of one or more chains of amino acids. As main components, proteins play lots of important roles in organisms: enzymes catalyzing chemical reactions in metabolic processes; receptor proteins on cell membranes getting extracellular signals and transmitting into cells; keratin protein making up nail plates of animals, etc. To recognize the molecular-scale functions(e.g. ligand binding) of proteins, their 3D structures should be determined and studied. Nowadays three methods, X-ray crystallography, nuclear magnetic resonance(NMR) spectroscopy and Cryo-electron microscopy(Cryo-EM) are usually employed to determine protein structures, and the determined structures are deposited in Protein Data Bank (PDB)(www.rcsb.org) which is a repository contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. Table 1.1 shows the statistics of current depositions in PDB(up to April, 2014).

Exp. Method	Proteins	Nucleic Acids	Protein/NA	Other	Total
X-RAY	81610	1516	4249	4	87379
NMR	9076	1076	204	7	10363
EM	514	51	170	0	735
HYBRID	59	3	2	1	65
other	155	4	6	13	178
Total	91414	2650	4631	25	98720

Table 1.1: Numbers of deposited structures in Protein Data Bank by different methods(www.rcsb.org).

As shown in Table 1.1, nearly 90% of proteins are determined by X-ray crystallography since it's the most powerful method to obtain macromolecular structures(www.rcsb.org). The steps to determine high resolution structures of proteins by x-ray are demonstrated in Figure 1.1. The proteins are purified and crystallized firstly, and normally this is the slowest step in the experiment(Pechkova and Nicolini 2003), particularly for membrane proteins. Once we have the crystal, we position it in X-ray beam and the crystal produces a diffraction pattern which is recorded and analyzed to determine the electron density distribution map of the molecule in the crystal. Finally, the electron density

CHAPTER 1 INTRODUCTION

is interpreted to determine atom coordinates. To improve the accuracy of structures, an iterative refinement is carried out by fitting the calculated diffraction pattern to the experimental patterns. Because there is no size limitation of molecules for X-ray crystallography determination, it can be applied to very large macromolecular complexes (>~100 kDa) and provide detailed atom locations. However, X-ray crystallography works only if the molecule could be crystallized but this step sometimes limit its applications to particular proteins, e.g. flexible proteins. In addition, we cannot study the dynamics of proteins in solutions by X-ray because the proteins are crystallized and in solid phase.

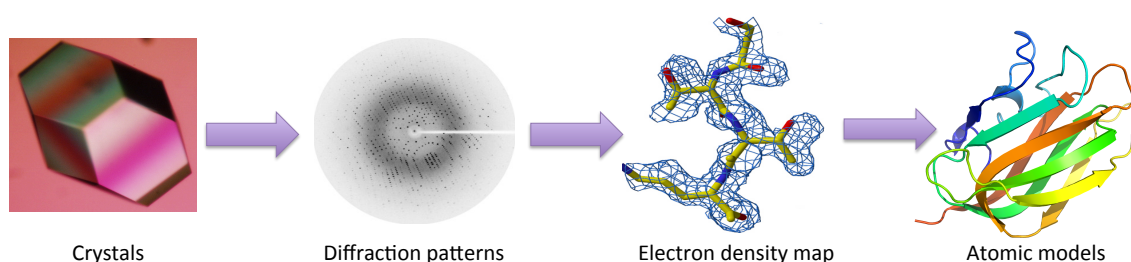


Figure 1.1: Diagram of structure determination by X-ray crystallography

The second popular method to determine protein structures is nuclear magnetic resonance (NMR). Different from X-ray method, proteins should be purified and dissolved in solutions and then placed in a strong magnetic field. Conventionally two kinds of data are recorded in NMR experiment, the first is the nuclear magnetic resonances of protons and labeled carbons and nitrogens in the protein, the second is the nuclear overhauser effects (NOE) (Noggle 1971) which consist the intra-protein distance information. Traditionally these data are picked from NMR spectrum and assigned to atoms manually but lots of automatic peak-picking algorithms and peaks assignment methods are developed and implemented in recent years. With assigned resonances and NOEs, computational programs, for example CYANA, are employed to calculate the protein structures. Since the NOE distance restraints are not always super-precise, more than one models may match the NMR observation, so NMR structures are normally deposited as ensembles. Compare to X-ray method, the major disadvantage of NMR is that it's normally not applicable to large proteins (weights > ~50 kDa) (Yu 1999) because of slower tumbling of the molecule in solution; as a consequence, the efficiency of magnetization transfer through bonds employed in NMR experiments decreases (Clore and Gronenborn 1998). What's more, large proteins will introduce more complexity, e.g. overlap peaks to the NMR spectrum and make it hard to analyze. The major advantage of NMR method is that it determine structures in solution, and thus it's the premier method for studying the atomic structures of flexible

proteins(Goodsell). Besides structures, NMR can also measure information on the dynamics of various parts of the proteins over wide time durations.

Start from the end of last century, Cryo-Electron microscopy is also employed to determine protein structures, especially for large protein complexes. The first protein resolved in atomic resolution is bacteriorhodopsin which is determined by Richard Henderson(Henderson et al. 1990) in 1990. On one hand, the advantage of Cryo-EM is that it is able to tackle very large or heterogeneous assemblies(Saibil 2000) and provides 3D image of molecules directly. On the other hand, several drawbacks of Cryo-EM limit its application in structural biology field. The main problem of Cryo-EM is that its resolution is still not as high as X-ray or NMR. And because Cryo-EM cannot provide the atom positions so it usually combine information from X-ray crystallography or NMR spectroscopy to sort out atomic details(Goodsell).

1.2 Nuclear magnetic resonance (NMR)

Nuclear magnetic resonance, abbreviated as NMR is a physical phenomenon that nuclei of atoms absorb and release electromagnetic radiation in a static magnetic field. The resonance frequency depends on the intensity of the magnetic field and the spin quantum number. In detail, when nuclear spins are placed in an external magnetic field, different spin states have different magnetic potential energies. according to the theory of Quantum mechanics, the states and energies are not continuous distribution, but have several energy levels. The number of nuclear spins in different states are approximately equal at thermal equilibrium. In the presence of the static magnetic field, a radio frequency signal of the proper frequency can induce a transition of nuclear spins from their lower to higher energy state. If the radio frequency signal is then switched off, the nucleus spins return to the thermodynamic state and produce radio frequency signals. Then the nuclear spins can be induced again and repeat the above process.

1.2.1 NMR spectroscopy for protein determination

Nuclear magnetic resonance spectroscopy (Figure 1.2) is a powerful and theoretically technique that commonly used to determine the structures and properties of organic compounds. In a molecule, due to different chemical environments and shielding effect by neighbor atoms, the resonant frequencies of different nucleuses are various and occur as peaks in NMR spectroscopy. However, the difference between two nearby peaks are quite small so that the frequencies are measured as relative values to a standard which is usually Tetramethylsilane(TMS). This relative frequency called chemical shifts(δ) and

CHAPTER 1 INTRODUCTION

calculated by Eq. 1.1 where ν_{sample} is the frequency of sample and ν_{TMS} is the frequency of TMS.

$$\delta = \frac{\nu_{sample} - \nu_{TMS}}{\nu_{TMS}} \times 10^6 \quad (1.1)$$

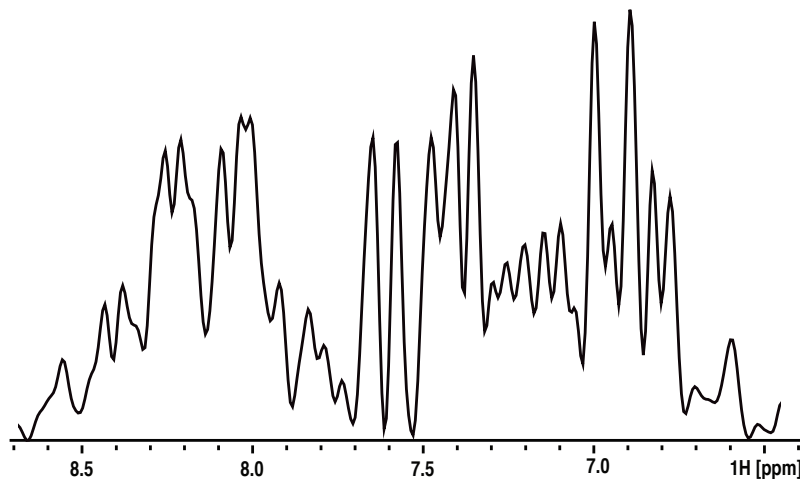


Figure 1.2: Example of 1D NMR spectroscopy of ^1H provided by Diana C. Rodriguez Camargo. The experiment was recorded using a 500-MHz Bruker spectrometer equipped with a cryo-triple resonance probe. The proton chemical shift was referenced with respect to the water resonance frequency (4.78 ppm at 4°C).

If the sample is large organic molecule, most of the signals will be overlap heavily in 1D NMR spectrum. To resolve this problem, people introduce additional spectral dimensions as high-dimensional spectra (Figure 1.3) which shows the correlation of atoms and provides more information about a molecule than 1D NMR spectra and are especially useful in determining the structure of a molecule. In the high-dimensional NMR spectroscopy, each cross peak shows a correlation between nucleus spins and the coordinates are the chemical shifts of spins. The common high-dimensional spectroscopies for protein structure determination include Heteronuclear single-quantum correlation spectroscopy (HSQC), correlation spectroscopy (COSY) (Noda and Ozaki 2005), total correlation spectroscopy (TOCSY) and nuclear overhauser effect spectroscopy (NOESY). Heteronuclear single-quantum correlation spectroscopy (HSQC) discovers the correlations between ^1H and another type of nucleus spin (normally is ^{13}C or ^{15}N) which are directly coupled by one bond, for example ^1H - ^{15}N HSQC. Correlation spectroscopy (COSY) is the simplest and most popular NMR experiment used for determining spin-spin couplings. Similar to HSQC, The correlation signals (cross peaks) appear when spins are directly coupled, and if there is no coupling, no correlation is expected to appear. However, this correlation is between the same type of spins. The total correlation spectroscopy (TOCSY) experiment is similar to the

COSY experiment, on which signals (cross peaks) of coupled spins are observed. However, correlation signals are seen between distant spins coupled through a chain of spins. Based on this feature, TOCSY can be used to identify the large spin network of molecules. In the application of protein structure determination by NMR, the high-dimensional HSQC, COSY and TOCSY, e.g. HCCH-TOCSY(Olejniczak et al. 1992), HC(CC)(CO)NH(Montelione et al. 1992) are generally employed to assign chemical shifts. Different from the above through-bond spectroscopy, the through space nuclear overhauser effect spectroscopy (NOESY) shows the effect of dipolar interaction from one nuclear spin to another by cross-relaxation through space. Since NOE dipolar coupling interacts are throughout space, it provides structural information of molecules, then it becomes a very useful tool to study the conformation of proteins.

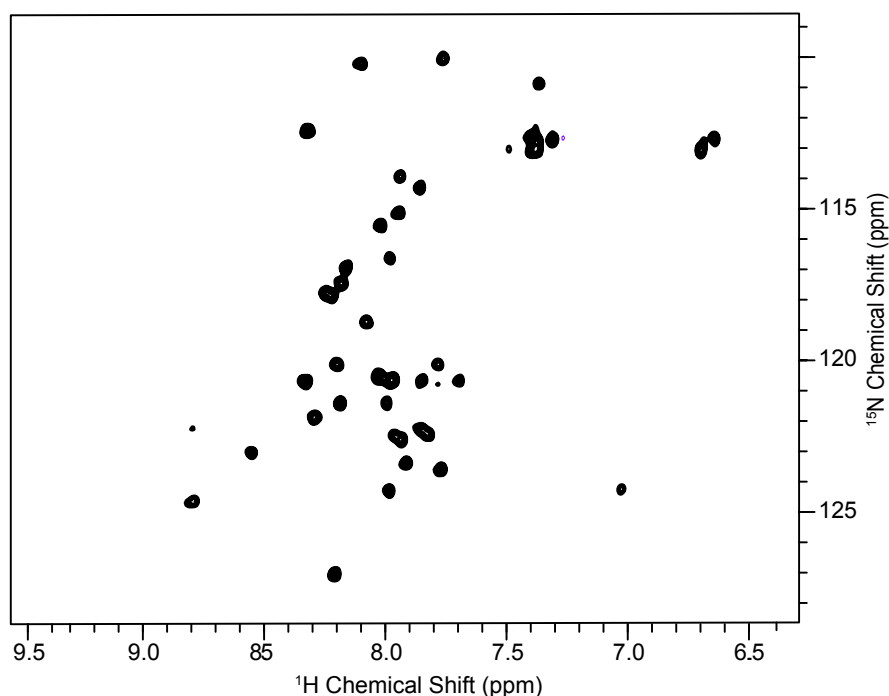


Figure 1.3: Example of 2D NMR spectroscopy provided by Diana C. Rodriguez Camargo.

1.2.2 NMR chemical shift assignment

The chemical shifts picked from NMR spectrum are raw data and not related to proteins. The assignment is to find out which chemical shift corresponds to which atom and this work can be carried out manually or automatically with the help of computers. Separately, the automatic backbone chemical shift assignment have been well-developed and generally easy to be done, but the automatic assignment of sidechain is relatively less explored and still a bottleneck in NMR structure determination(Zeng et al. 2011).

a) Manual NMR chemical shift assignment

In recent years, although lots of automatic chemical shift assignment methods have been developed and applied (Guerry and Herrmann 2011), most of the assignment work is still done manually. For backbone chemical shift assignment, HSQC and triple NMR spectra CBCANNH (GRZESIEK and Bax 1992a) and CBCA(CO)NNH (GRZESIEK and Bax 1992b) (Figure 1.4) are necessarily adopted. As shown in Figure 1.4, in CBCANNH there is strong correlation among $C\alpha_i$, $C\beta_i$, N_i and HN_i (as detection) as well as weak correlation among $C\alpha_{i-1}$, $C\beta_{i-1}$, N_i and HN_i . In CBCA(CO)NNH the correlation happens only among $C\alpha_{i-1}$, $C\beta_{i-1}$, N_i and HN_i . From HSQC, we select a clear peak p_i which is not overlapped as the original point. Start with it, we are guided to its corresponding peak rp_i of the same residue in CBCANNH spectra according to the chemical shift value. In parallel, we also locate the relative peak lp_i in CBCA(CO)NNH spectra. To confirm the found peaks are consistent on the same residue, we compare rp_i and lp_i then we can assign the chemical shifts of $C\alpha_i$ and $C\beta_i$. From this new assigned $C\alpha_i$ and $C\beta_i$, in CBCA(CO)NNH locate new N-HN pairs and repeat the above steps.

There are various manual methods and NMR spectra e.g. HBHA(CO)ONH (GRZESIEK and Bax 1993), H(CCCO)NNH (Montelione et al. 1992; GRZESIEK et al. 1992), CC(CO)NNH (GRZESIEK et al. 1992), HCCH-TOCSY (Bax et al. 1989) available for sidechain chemical shift assignment. However, it is still much more challenging than the backbone resonance assignment because of its complexity and serious overlap.

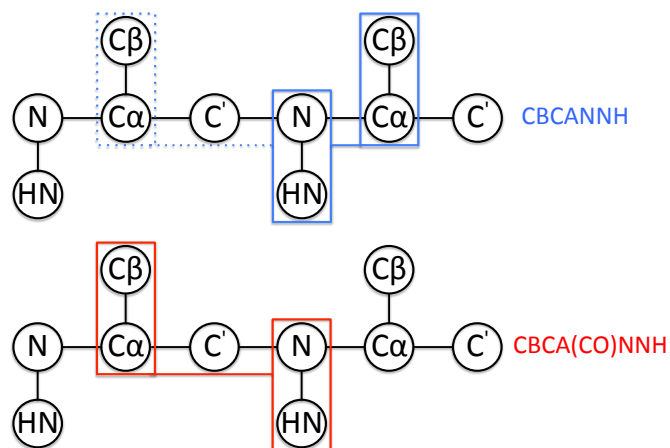


Figure 1.4: CBCANNH and CBCA(CO)NNH spectra

b) Automatic NMR chemical shift assignment

In the past two decades, not less than 44 publications for automated backbone and/or sidechain resonance assignment algorithms are published (Guerry and Herrmann

CHAPTER 1 INTRODUCTION

2011; Schmidt and Güntert 2012). Among them, only 19 programs can work purely based on NMR peaks, and all the others need additional input data like 3D structures(Jung and Zweckstetter 2004), residual dipolar couplings(Wang et al. 2011), assigned backbone chemical shifts, etc. In addition, many programs(Lukin et al. 1997; Buchler et al. 1997; Zimmerman et al. 1997; Leutner et al. 1998; Jung and Zweckstetter 2004) only focus on assigning the backbone and C β chemical shifts. Moreover, only 3 exclusively NMR peak based programs have been used to determine protein structures deposited in the Protein Data Bank(PDB). One is AutoAssign(Zimmerman et al. 1997) for automated backbone assignment and the other two, PINE(Bahrami et al. 2009) and GARANT(Bartels et al. 1997) are appropriate for full resonance assignment. In 2012, Schmidt E, Güntert P presented a new algorithm FLYA for reliable and general NMR resonance for both sidechain and backbone(Schmidt and Güntert 2012), which works much better than GARANT and PINE. The input of this new algorithm is only the sequence of proteins and any combination of peak lists from high-dimensional through-bond or through-space NMR spectra experiments(Schmidt and Güntert 2012). For each kind of NMR spectra, FLYA defines through-bond or through-space magnetization transfer network in the database, based on which FLYA generates expected peaks where each atom has chemical shift range from the BMRB statistics. Then FLYA maps the expected peaks to the measures peaks from experiment spectra by chemical shift match. With optimization, the accurate of FLYA is as high as 96–99% for the backbone and 90–91% for all resonances.

1.2.3 Nuclear Overhauser effect (NOE) assignment

In order to extract distance information from the NOESY spectrum, the cross-peaks have to be assigned, in other words, the pairs of hydrogen atoms of the peak as well as labelled Carbon or Nitrogen need to be identified(Herrmann et al. 2002a). Because the precision of the chemical shift values of NOE peaks and assigned atoms are limited and there are usually about 0.03 ppm tolerance for Hydrogen and 0.3 ppm for carbon and nitrogen, so it's difficult to assign a NOE peak to a single hydrogen pair. The assignments will be either ambiguous or with serious errors. Nowadays, several programs for automatic NOE assignment and protein structure determination, such as ARIA(Linge et al. 2003; Rieping et al. 2007), CYANA(Güntert et al. 1997; Herrmann et al. 2002b), Auto-Structure(Huang et al. 2005; Huang et al. 2006), UNIO(Serrano et al. 2012) and PASD(Kuszewski et al. 2004) are presented and applied. Among them, CYANA is most popular and widely used. The input of the NOE peak assignment algorithm CANDID(Herrmann et al. 2002b) for CYANA are protein sequence, assignment chemical shifts and raw NOE peak list without assignment. If there are conformational information from other sources available, CANDID also uses them(Kuszewski et al. 2004). Based on the

CHAPTER 1 INTRODUCTION

chemical shift fitting within a pre-defined tolerance, a list of hydrogen pairs as well as their bonded Nitrogen or Carbon are assigned to each NOE peak initially. Then for each NOE peak, all assignments are sorted by some criteria and low ranking assignments are eliminated. The criteria for assignment ranking includes: 1) the closeness of the chemical shifts of assigned hydrogen pairs fitting NOE peak. 2) Normally there are more than one NOE spectra are implemented for structure determination, and then there would be symmetric peaks between two relative peaks. 3) the distance of hydrogen pairs if prior structural knowledge, for example fragments for Rosetta exists. 4) In the NOE peaks, if the hydrogen pairs of one assignment are indirectly connected through a third atom, this assignment has high probability to be correct. After the assignment elimination, restraints are generated for each cross peak with at least 1 assignment. For the restraints, the upper-distance bound is defined based on the intensity of the peak. Since there are noises and artifacts that are picked from NOE spectra as peaks, several filters are applied to eliminate spurious cross-peaks that should not be considered for restraint generation. For one peak, if all its assignments don't quite conform to the above criteria, it should be removed. In addition, the peak whose assignments exceed a maximum threshold is also eliminated. Third, peaks are also eliminated if the network anchoring score of their assignments remains low (Herrmann et al. 2002a). During the sampling, peaks are also eliminated if they are violated by decoy structures (Zhang et al. 2014).

1.3 Macromolecular modeling program Rosetta

The macromolecular modeling software Rosetta is initially presented in 1999 for protein structure prediction based on only the primary structures without experimental data (Simons et al. 1999). The predictions for protein domains with fewer than 125 amino acids regularly have a backbone root-mean-square deviation of better than 5.0 Å (Kaufmann et al. 2010). More impressively, there are several cases in which Rosetta has been used to predict structures with atomic level accuracy better than 2.5 Å (Kaufmann et al. 2010). In addition to de novo structure prediction, more functions has been developed and implemented in Rosetta in the past quarter century. Rosetta also has methods for protein-protein docking (Zhang and Lange 2013), protein-ligand docking (Davis et al. 2009), homology modeling (Thompson et al. 2012), determining protein structures from experimental NMR (Raman et al. 2010; Lange et al. 2012; Schot et al. 2013), and protein design (Liu and Kuhlman 2006).

Based on the sequence match, Rosetta picks hundreds of fragments for each segment (3-9 residues) of the proteins, and then it samples the conformations by switching fragments for every segment based on Monte Carlo algorithm (Rohl et al. 2004) to minimize

the knowledge-guided Rosetta energy functions. After a pre-defined trajectory of conformational sampling, thousands or even more structures are generated. Rosetta selected 10 structures with lowest energy as the final predicted models.

The correlations between isotropic chemical shifts and structural information is largely based on empirical statistics gained from the mining of protein chemical shifts deposited in the BMRB as well as its corresponding 3D structures in the PDB. Building on top of Rosetta, Chemical-Shift-Rosetta (CS-Rosetta)(Schot et al. 2013) is a framework for structure calculation of biological macromolecules with the input of backbone NMR chemical shifts (^{13}CA , $^{13}\text{C}\beta$, $^{13}\text{C}'$, ^{15}N , ^1HA and ^1HN) which are easily measured and assigned in NMR experiments as described in above. As shown in Figure 1.5, the main functions of chemical shifts in CS-Rosetta are to improve the accuracy of fragment picking and rescore the sampled structures. CS-Rosetta selects fragments from the PDB with both sequence and chemical shift match between fragments and target proteins. Then a regular Rosetta Monte Carlo assembly and relaxation procedure is carried out. The all-atom models produced by Rosetta sampling are rescored by value match between experimental chemical shifts and computational chemical shifts of Rosetta models. The same to Rosetta, 10 structures with lowest energy are selected finally. In the procedure of CS-Rosetta, the simulated chemical shifts of fragments and sampled models are both computed by SPARTA+ (Shen and Bax 2010).

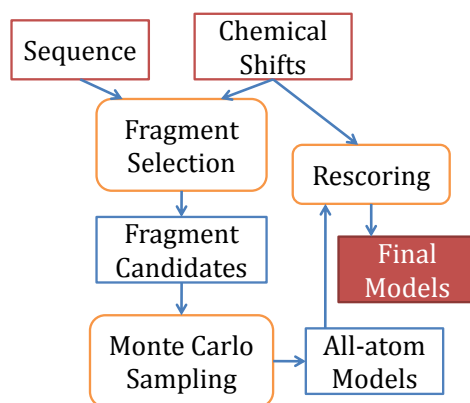


Figure 1.5: Procedure of CS-Rosetta

1.3.1 Fragment picking based on protein sequence and chemical shifts

In Rosetta, a fragment represents a small continuous segment (typically comprising 3–15 residues) of protein with known 3D backbone structures, which is defined by ϕ , ψ and ω torsion angles(Figure 1.6). Based on the fragment library covering every residue position,

CHAPTER 1 INTRODUCTION

the fragment assembly algorithm can efficiently generate a wide variety of compactly folded structural models(Vernon et al. 2013). Among the fragment library, only a small percentage of accurate fragments are usually enough for de novo programs to generate a few models close to the native protein structure.

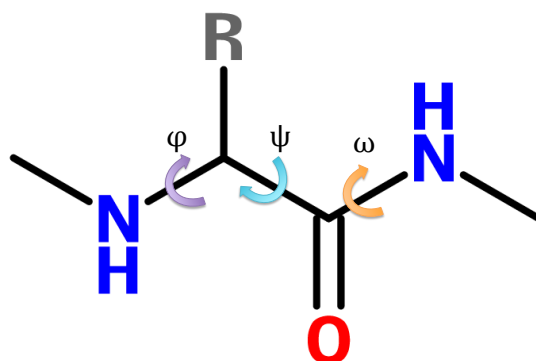


Figure 1.6: torsion angles of protein backbone

The original fragment picking algorithm for CS-Rosetta was the multiple fragment replacement (MFR) method of the NMRPipe software package(Delaglio et al. 1995). From a large pool of crystal structures, MFR method scores all fragment candidates by 1) chemical shift similarity between the target's values measured in NMR experiments and the shift values predicted by SPARTA+ for fragment candidates, 2) sequence match between target and the fragment candidates, 3) the torsion angle probabilities of fragment candidates. Then it selects low-score fragments to compose the fragment library for following structure predictions. The drawback of this method is that its chemical shift match score will decrease if the chemical shifts are incomplete in certain regions, and then it bias selects fragments whose secondary structure is alpha helix(Vernon et al. 2013). This problem was resolved by an upgrade of CS-Rosetta fragment picker(R2FP) (Simons et al. 1997; Rohl et al. 2004) which involves sequence based secondary structure prediction(Jones 1999; Meiler et al. 2001; Karplus et al. 2003). In 2013, R. Vernon et al. presented a new fragment picking algorithm Rosetta3 Fragment Picker(R3FP) combining the advantages of both MFR and R32FP and introduce new concepts for scoring fragment candidates(Vernon et al. 2013).

In the new protocol, 200 fragments in both 3-residue and 9-residue size for each residue position are selected from a database of about 2.3 million fragment candidates generated from ~9,000 proteins. To improve the accuracy of fragments, following information of each residue in the database is analyzed or predicted: 1) sequence profiles from PSI-BLAST(Altschul et al. 1997), 2) chemical shifts predicted by SPAETA+(Shen and Bax 2010), 3) secondary structure assignment by DSSP(Kabsch and Sander 1983). To select the best match fragments, 5 independent scores are calculated for each residue in the database.

CHAPTER 1 INTRODUCTION

CS-Score S_δ : this score only depends on the chemical shifts and use a sigmoid potential for the error between predicted and experimental chemical shifts. Then it can filter out badly matching data points. where N_T and N_{DB} are the number of available shifts that can be compared in the database or target, respectively. In Eq. 1.2, δ_T , δ_{CD} and $\Delta\delta_{CD}$ are the secondary shift of target, predicted secondary shift and prediction error, respectively.

$$S_\delta = \frac{N_T}{N_{CD}} \sum_{shifts}^{N_{CD}} \frac{1}{1+e^{-2\frac{\delta_T-\delta_{CD}}{\Delta\delta_{CD}}+4}} \quad (1.2)$$

Profile-Score S_p : as described in Eq. 1.3, this score is to compare the sequence profile of target residues and candidates by Manhattan distance. P_T and P_{CD} are the sequence profiles of target residue and candidate residue, respectively.

$$S_p = \sum_{aa} |P_T - P_{CD}| \quad (1.3)$$

Rama-Score S_r : this score is to compute the probabilities of the candidate residues' backbone torsion angles in Ramachandran plot. In Eq. 1.4, $R(\phi, \psi, k, aa)$ is the Ramachandran density for residue type aa and secondary structure k , weighted by the TALOS+(Shen et al. 2009) predicted secondary structure propensities $P_{TSS}(k)$.

$$S_r = \frac{1}{1+e^{-\sum_{k \in \{h,e,l\}} \log[R(\phi, \psi, k, aa) P_{TSS}(k)]}} \quad (1.4)$$

SS-similarity-Score S_{SS} : here, the secondary structure score from R2FP:NNMAKE is implemented, but the chemical shift based secondary structure propensities $P_{TSS}(k_{DSSP})$ is used to instead secondary structure profile predicted based on sequence. Similar to CS-Score, sigmoid potential is also used in this score function. C_{TSS} represents the TALOS+ prediction confidence.

$$S_{SS} = \sqrt{C_{TSS}} \frac{1}{1+e^{-7(P_{TSS}(k_{DSSP}))^{+4}}} \quad (1.5)$$

Phi/Psi-Squarewell-Score $S_{\phi\psi}$: the last score compares the backbone torsion angles between candidates and targets whose backbone structure is predicted by TALOS+ based on chemical shift.

$$S_{\phi\psi} = \sqrt{\frac{1}{1+e^{-\frac{\max(0, d(\phi_T, \phi_{CD}) - \Delta\phi_T)}{2 \cdot \Delta\phi_T} + 5}}} + \frac{1}{1+e^{-\frac{\max(0, d(\psi_T, \psi_{CD}) - \Delta\psi_T)}{2 \cdot \Delta\psi_T} + 5}}} \quad (1.6)$$

In Eq. 1.6, ϕ_T , ψ_T are predicted backbone torsion angles for targets and ϕ_{CD} , ψ_{CD} are candidates' torsion angles. $\Delta\phi_T$, $\Delta\psi_T$ are the tolerances.

Overall, the final score function of a fragment at position k with N residues is:

$$S_{all} = N^{-1} \sum_{i=k}^{N+k-1} S_\delta^{(i)} + 1.5 S_p^{(i)} + S_r^{(i)} + 0.25 S_{SS}^{(i)} w(r_{SS}^{(i)}) + 5.0 S_{\phi\psi}^{(i)} w(r_{\phi\psi}^{(i)}) \quad (1.7)$$

1.3.2 Rosetta score

Rosetta scoring function is a model generated using various contributions that describe the protein-likelihood of a predicted structure, independent or dependent of the sequence (Simons et al. 1997; Simons et al. 1999). Two levels of conformations, centroid mode and full-atom mode (Figure 1.7), are utilized to represent protein side chains in Rosetta. In centroid mode, each side chain is represented by a centroid located at the side-chain center of mass (Rohl et al. 2004). Consequently, the items comprising Rosetta score function are also divided into simple centroid scores and more sophisticated full atom scores.

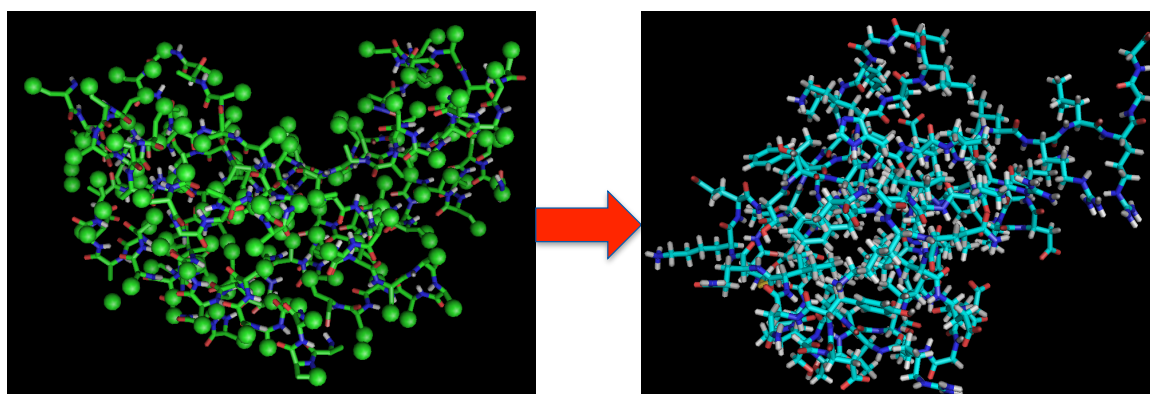


Figure 1.7: centroid mode and full-atom mode representations of proteins in Rosetta.

Following are typical items comprising the low-resolution Rosetta score function (Rohl et al. 2004):

vdw score: generally, van der Waals' force is the sum of the attraction and repulsion between molecules. In Rosetta, only repulsive force is considered. D is the distance between two atoms and r_{vdw} is the pre-determined van der Waals radius (Rohl et al. 2004).

$$vdw = \sum_{D^2 < r_{vdw}^2} \frac{(r_{vdw}^2 - D^2)^2}{r_{vdw}^2}; \quad (1.8)$$

env score: this score represents the probability of given residue in given environment. m is the residue index, a_m is the amino acid type of residue m , nn_m is the number of residues surrounding residue m (Rohl et al. 2004).

$$env = -\ln[\prod_m P(a_m | nn_m)] \quad (1.9)$$

pair score: the interactions of each atom pair. m and n are the residue indices, a_m, a_n are the amino acid type of residue m and n , sd_{mn} is the sequence distance between residue m, n and cd_{mn} is the centroid-centroid space distance (Rohl et al. 2004).

$$pair = -\ln \left\{ \prod_m \prod_{n>m} \left[\frac{P(a_m a_n | sd_{mn} cd_{mn})}{P(a_m | sd_{mn} cd_{mn}) P(a_n | sd_{mn} cd_{mn})} \right] \right\} \quad (1.10)$$

CHAPTER 1 INTRODUCTION

sheet score: this score represents the stand arrangement into sheets. n_{sh} is the number of sheets and n_{st} is the number of strands(Rohl et al. 2004).

$$sheet = -\ln[P(n_{sh}|n_{st})] \quad (1.11)$$

cbeta score: this score stands for the neighbor frequency. m and n are the residue indices, sh is the shell radius(6, 12 Å)(Rohl et al. 2004), $P_{ensemble}$ is the probability in ensembles from fragments and P_{rand} is the probability randomly from fragments. nn is the number of neighboring residues within in the shell.

$$cbeta = -\ln\left[\prod_m \prod_{sh} \left[\frac{P_{ensemble}(nn_{m,sh})}{P_{rand}(nn_{m,sh})}\right]\right] \quad (1.12)$$

rg score: The radius of gyration shows the root-mean-square distance between all atoms in a molecule and the centroid. N is the number of residues, m, n are the residue indices, r_m is the position of residue m .

$$rg = \left\{\sum_{n=1}^N [r_n - (\sum_{m=1}^N r_m / N)]^2 / N\right\}^{1/2} \quad (1.13)$$

In some situations, e.g. intro-protein pathway detection, high-resolution with full described sidechain is necessary. New score items are also should be introduced into Rosetta score function.

rama score: the Ramachandran torsion angle probabilities. ϕ_m, ψ_m are the backbone torsion angles, a_m is the amino acid type, and ss_m is the secondary structure type.

$$rama = -\ln[\prod_m P(\phi_m, \psi_m | a_m, ss_m)] \quad (1.14)$$

hb score: the hydrogen bonding energies. m is the donor residue index, n is the acceptor residue index, d_{mn} is the acceptor-proton interatomic distance, h_n is the hybridization, ss_{mn} is the secondary structure type

$$hb = -\ln\{\prod_m \prod_n [P(d_{mn} | h_n, ss_{mn})P(\cos \phi_{mn} | d_{mn}, h_n, ss_{mn})P(\cos \psi_{mn} | d_{mn}, h_n, ss_{mn})]\} \quad (1.15)$$

dun score: the Rotamer self-energy. m is the residue index, dr_m is the Dunbrack backbone-dependent rotamer(Shapovalov and Dunbrack 2011), ϕ_m, ψ_m are backbone torsion angles. a_m is amino acid type.

$$pair = -\ln\left\{\prod_m \left[\frac{P(dr_m | \phi_m, \psi_m)P(a_m | \phi_m, \psi_m)}{P(a_m)}\right]\right\} \quad (1.16)$$

Besides the described score above, there are still more score items comprising Rosetta score function in both centroid and full-atom modes(Rohl et al. 2004).

1.3.3 Fragment assembly by Monte Carlo method

Rosetta assemble fragments into a protein-like structure by Monte Carlo simulation method starting from an fully extended conformation(Rohl et al. 2004). As shown in Figure 1.8, the assembly process started with a random position in the sequence, A 9-residue fragment insertion segment is randomly located and from the fragment library a fragment for this window is randomly selected. In fact, Rosetta not really uses the selected fragment to replace the 9-residue segment in the protein chain, but just passes the backbone torsion angles. As a result, the protein structure and its Rosetta energy are changed. Rosetta retains or abandons the change according to Metropolis criterion(described below, Figure 1.9). After decision, Rosetta randomly switches to another segment and repeats above steps. The 9-residue fragment insertions have 4 steps. In step 1, only Van der Waals score is evaluated, and this stage continues until all extended protein chain is replaced by fragments. In step 2, residue interaction scores and secondary structure scores, e.g., *pair*, *env*, *sheet*, *ss_pair*, *hs_pair*, are included and this step normally has 2000 fragment insertion attempts. In step 3, the secondary structure scores are increased to full weight to extensively search for secondary structure interactions and packing. Step 3 usually has 20,000 fragment insertion attempts. In the last step of 9-residue fragment insertion, the full centroid score function with all score items are evaluated. After the structure assembly from 9-residue fragments, short segment refinement by 3-residue fragment insertion is carried out to slightly compact the structures. To avoid the local minimum problem, many samplings start in parallel from different random positions to generate ensembles with both favorable local interaction and protein-like global properties(Rohl et al. 2004).

Metropolis criterion:

- 1) propose an unbiased random structure change $\vec{m}_* = \vec{m}_{t-1} + \vec{\epsilon}$
- 2) calculate the energy change: $\Delta E = E(\vec{m}_*) - E(\vec{m}_{t-1})$
- 3) always accept the change if $\Delta E < 0$
- 4) if step 3 doesn't pass, accept the change with probability $p = \exp(-\Delta E/k_B T)$
- 5) if accepted $\vec{m}_t \equiv \vec{m}_*$, otherwise $\vec{m}_t \equiv \vec{m}_{t-1}$.

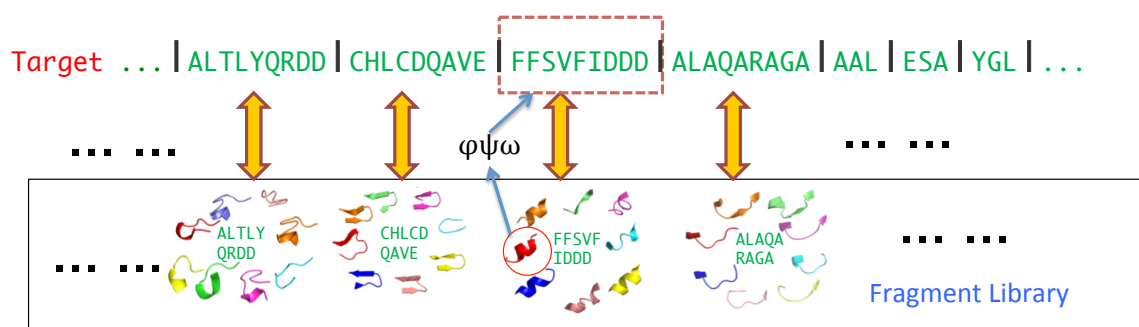


Figure 1.8: Process of Rosetta fragment assembly

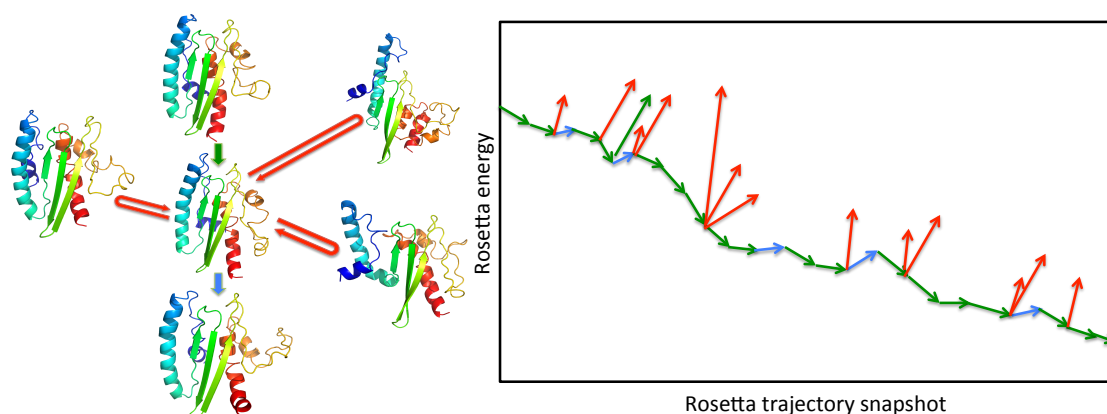


Figure 1.9: Metropolis criterion is used to determine the acceptance of structure change. Green means Rosetta energy decreases and acceptance; red means Rosetta energy increases and rejection; blue means Rosetta energy increases but acceptance.

1.3.4 RASREC protocol

As mentioned in chapter 1.1, NMR structure determination is a big challenge if the size of proteins are larger than 15 kDa. For larger proteins, there were two major problems from experiment aspect: first, the NMR spectra will be overlapped, and second, the resonance line widths are related to the size of the protein.(Bax 1994). As the result, large protein determinations will involve more ambiguous NOESY-derived distance restraints(Lange and Baker 2011). From computation aspect, large size and increased complexity are also challenges for protein structure ab initio programs(Bonneau et al. 2002; Kryshtafovych et al. 2005). For Rosetta, its normal de-novo structure calculation usually works for proteins within 100 amino acids. With additional sparse NMR data—chemical shifts, RDCs, and backbone HN-HN contacts, the size of Rosetta sampling limitation increases slightly, to 120–130 amino acids(Shen et al. 2008b; Raman et al. 2010), and hence the original CS-Rosetta protocol abrelax does not have a robust success rate for proteins over 15 kDa(Raman et al. 2010; Lange and Baker 2011). To overcome the experimental problems, deuteration is introduced(Nietlispach et al. 1996). To resolve the de novo sampling limitations, O. Lange and D. Baker presented an iterative sampling protocol that recombines structural features found in intermediate structures, named as resolution-adapted structural recombination (RASREC)(Lange and Baker 2011).

As shown in Figure 1.10, the new RASREC protocol has 6 sampling stages, on initial exploration stage(not shown in Figure 1.10) and 5 resampling stages, of which, the first 4 stages use low-resolution(centroid) Rosetta score function and stage 5, 6 use high-resolution(full-atom) Rosetta score function. With RASREC protocol, Rosetta firstly get the secondary structure from chemical shift and determine the beta-sheet region in stage

CHAPTER 1 INTRODUCTION

1(Figure 1.10-A) because strand is the difficult region in protein de novo prediction with following reasons(Bradley and Baker 2006): firstly, the random fragment insertion attempts hardly achieve the precise relative geometry of these long-range beta-sheet pairings; secondly, the beta-sheet pairings play core roles in the protein 3D structures, if they are changed, the Rosetta score will change significantly, thus they will be effectively fixed once they are formed; thirdly, the factual nonlocal beta-pairings are replaced by competing local beta-pairings, just because the latter are easier to sample; finally, the large number of possible nonlocal beta-sheet topologies expands the searching space of conformations. In stage 1, chain break is introduced to avoid the cyclic fold-tree(Karplus et al. 2003; Bradley and Baker 2006; Leaver-Fay et al. 2011). Then Rosetta generates the possible beta-sheet topologies including the orders and directions, based on the determined strand pairings in stage 2(Figure 1.10-B). In stage 3-6, fragment resampling is carried out in parallel together with other operations (Figure 1.10-C). The fragment resampling here is the same as original Rosetta abrelax protocol. Since at the beginning of sampling, it's nearly impossible to predict which trajectory reaches the lowest-energies region, Rosetta launches several initial trajectories and samples different folds in stage 4(Figure 1.10-D1). Then Rosetta starts more trajectories from the earlier snapshot of trajectory which is lowest in energy since the assembly cannot be corrected once it's compacted(Figure 1.10-D2,D3). In the last two stages, the compacted protein is relaxed in high-resolution and removes chain-break by idealization or rebuilding.

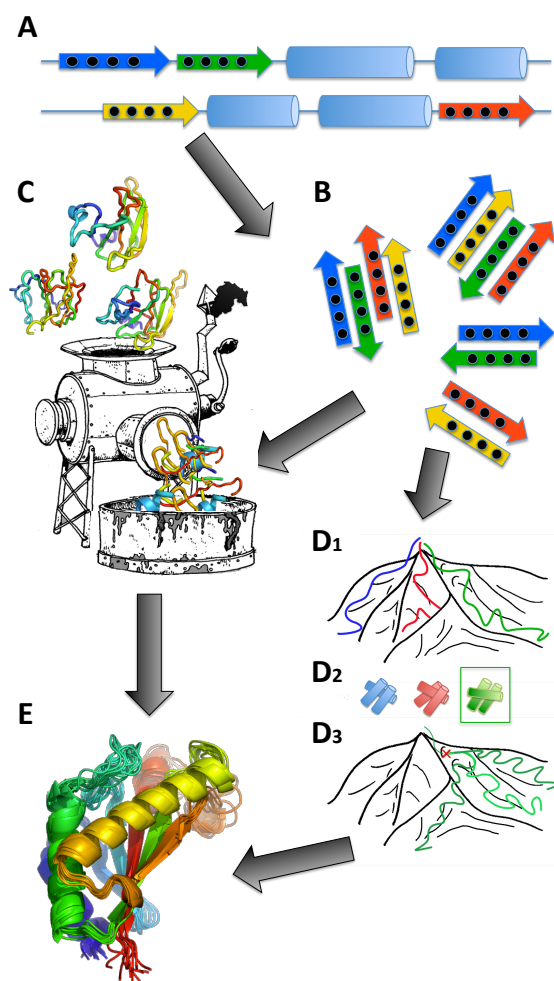


Figure 1.10: Illustration of RASREC protocol. (A) strand sampling with random pairs protocol; (B) Beta-sheets topology sampling and topology resampling; (C) fragment resampling; (D1-D3) proto-fold resampling; (E) loop-rebuilding and all-atom refinement. (Lange and Baker 2011)

1.4 Aims of the present thesis

In this thesis, the main tasks are to contribute to the understanding of 3D protein structure determination by Rosetta and NMR data by three different processes (1) improvement of 3D structure prediction from chemical shift data (chapter 2), (2) a new algorithm for automatic NOESY assignment and structure determination with Rosetta (chapter 3), and (3) testing the performance of automatic NOESY assignment and structure determination algorithms with scramble chemical shift data (chapter 4). Particular questions, corresponding reviews and additional background are detailed at the beginning of each study.

1.4.1 Structure prediction by Rosetta from chemical shift data

Fragment picking for ROSETTA was originally carried out using the MFR (molecular fragment replacement) method from the NMRPipe software package which combined chemical shift information with peptide sequence matching to score fragment candidates. However, the drawback is that ROSETTA would outperform MFR for regions where no experimental data was present. In this project, a new kind of fragment picker named R3FP (ROSETTA3 Fragment picking) combines salient features of MFR and chemical shifts is incorporated into CS-ROSETTA.

Compared to standard CS-ROSETTA algorithm, RASREC (resolution-adapted structural recombination) is an iterative sampling algorithm that has been shown to significantly increase the sampling efficiency for larger proteins (10-40kDa) when additional restraint data such as RDCs and NOEs are used. It is crucial for the performance of RASREC that pseudo-energies (e.g., from RDC and NOE restraint data) must be available to assist ROSETTA in predominantly selecting the structures with native features for this pool. In this project, I extend the RASREC algorithm to allow chemical shift rescoring of intermediate structures (CS-RASREC) and test the performance of this extended method.

Since the chemical shifts are dominated by local backbone conformations, CS-ROSETTA structures based alone on chemical shifts tend to be locally correct while globally still unconverged. Hence a post-analysis procedure that identifies locally converged regions of the structure is introduced to CS-ROSETTA and converged CS-Rosetta structures have been shown to be generally accurate. However, with the decreasing convergence there is also an increasing probability that the conformation of the converged regions becomes inaccurate. We address this issue here by testing a number of criteria including the quality of chemical shift data, the number of converged residues and the significance of the ROSETTA score gap to detect inaccurate predictions. These criteria are aggregated to annotate each CS-ROSETTA prediction as weak or strong, thereby providing users with a reliable metric to assess the results.

1.4.2 New algorithm for automatic NOESY assignment and structure determination with Rosetta

The structure calculations of solution-NMR proteins are carried out based on following input: distance constraints from NOE spectroscopy(Kumar et al. 1980; Wüthrich 1989), dihedral angle constraints (Güntert and Wüthrich 1991), residual dipolar couplings(Rohl and Baker 2002; Prestegard et al. 2005), chemical shifts, etc. Therein, one of the key information is NOE distance constraints(Herrmann et al. 2002b). To get this data,

two-dimensional or higher-dimensional heteronuclear-resolved H-H NOESY peaks are manually assigned to individual atoms based on assigned chemical shifts, which is a time-consuming work and usually take several months to finish for only one protein. Nowadays automation of this process is a major goal for NMR structure calculation. Combined with structure de novo programs, a list of automated NOE peak assignment algorithms are recently developed (Mumenthaler et al. 1997; Herrmann et al. 2002b; Rieping et al. 2007; Zhang et al. 2014) and proved to produce comparable structures as those solved manually.

However, there are several limitations for current programs. Firstly, These program must generate a sufficiently accurate model from the initial assignments, which usually limits their application to only small proteins with high quality NMR spectra, complete and accurate chemical shift assignments and well-refined cross-peaks. Secondly, manual hydrogen bond restraints are also usually needed by these programs. In this project, I focus on developing a NOE assignment and structure determination algorithm that can produce results that are both reliable and accurate within only chemical shift assignments and unassigned NOE peak-lists.

1.4.3 Performance of automatic NOESY assignment and structure determination algorithms with scramble chemical shift data

The success of automated NOE assignment and structure calculation strongly relies on the completeness and precision of chemical shift assignments (Jee and Güntert 2003). Although lots of automatic semi-automatic methods for NOESY assignment have been developed in the past two years, only 3 exclusively NMR peak based programs have been used to determine protein structures deposited in the Protein Data Bank (PDB). Therefore, although the computational methods of automated chemical shift assignment have been developed significantly, most assignments are still done manually (Shen et al. 2008a) in practice and it has high probability to involve mistakes. In addition, the accuracy of side chain assignments are always much lower than those of backbone assignments (Moseley and Montelione 1999; Schmidt and Güntert 2012).

Until now, there are only a few systematic studies with respect to the influence of chemical shift assignments on NMR protein structure determination. In 2003, Jee and Güntert presented a study about the influence of resonance assignment on automated NOE assignment and NMR structure calculations (Jee and Güntert 2003). The limitation of this study is that it tested only one de novo program CYANA, one automated NOE assignment algorithm CANDID, and one kind of problem omission. In 2008, Shen et al. stated a work about the protein structure determination from incomplete chemical shift assignments (Shen

et al. 2008a). Nevertheless, this work is not about the effect on NOE assignments but the effect on fragment picking and final model selection of CS-Rosetta.

In this project, I forward the research with two popular de novo programs CYANA(Güntert et al. 1997; Herrmann et al. 2002a) and ASDP(Huang et al. 2005; Huang et al. 2006) and contrast them with the new program AutoNOE-Rosetta(Zhang et al. 2014), as well as more types of missing or erroneous assignments.

1.5 References

Altschul SF, Madden TL, Schaffer AA, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.

Bahrami A, Assadi AH, Markley JL, Eghbalnia HR (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. doi: 10.1371/journal.pcbi.1000307

Bartels C, Güntert P, Billeter M, Wüthrich K (1997) GARANT - A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J Comput Chem* 18:139–149.

Bax A (1994) Multidimensional nuclear magnetic resonance methods for protein studies. *Curr Opin Struct Biol* 4:738–744. doi: 10.1016/S0959-440X(94)90173-2

Bax A, Clore G, Gronenborn AM (1989) 1H-1H correlation via isotropic mixing of ¹³C magnetization, a new three-dimensional approach for assigning 1H and ¹³C spectra of ¹³C-enriched proteins. *Journal of Magnetic Resonance* 88:425–431. doi: 10.1016/0022-2364(90)90202-K

Bonneau R, Ruczinski I, Tsai J, Baker D (2002) Contact order and ab initio protein structure prediction. *Protein Sci* 11:1937–1944. doi: 10.1110/ps.3790102

Bradley P, Baker D (2006) Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. *Proteins* 65:922–929. doi: 10.1002/prot.21133

Buchler N, Zuiderweg E, Wang H, Goldstein RA (1997) Protein heteronuclear NMR assignments using mean-field simulated annealing. *J Magn Reson* 125:34–42.

Clore GM, Gronenborn AM (1998) NMR structure determination of proteins and protein complexes larger than 20 kDa. *Current Opinion in Chemical Biology*

CHAPTER 1 INTRODUCTION

Davis IW, Raha K, Head MS, Baker D (2009) Blind docking of pharmaceutically relevant compounds using RosettaLigand. *Protein Sci* 18:1998–2002. doi: 10.1002/pro.192

Delaglio F, Grzesiek S, Vuister G, et al. (1995) NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6:277–293. doi: 10.1007/BF00197809

GRZESIEK S, Anglister J, Bax A (1992) Correlation of Backbone Amide and Aliphatic Side-Chain Resonances in $^{13}\text{C}/^{15}\text{N}$ -Enriched Proteins by Isotropic Mixing of ^{13}C Magnetization. *J Magn Reson B* 101:114–119. doi: 10.1006/jmrb.1993.1019

GRZESIEK S, Bax A (1992a) An Efficient Experiment for Sequential Backbone Assignment of Medium-Sized Isotopically Enriched Proteins. *Journal of Magnetic Resonance* 99:201–207.

GRZESIEK S, Bax A (1992b) Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *J. Am. Chem. Soc.*

GRZESIEK S, Bax A (1993) Amino acid type determination in the sequential assignment procedure of uniformly $^{13}\text{C}/^{15}\text{N}$ -enriched proteins. *J Biomol NMR* 3:185–204.

Guerry P, Herrmann T (2011) Advances in automated NMR protein structure determination. *Quart Rev Biophys* 44:257–309. doi: 10.1017/S0033583510000326

Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *Journal of Molecular Biology* 273:283–298. doi: 10.1006/jmbi.1997.1284

Güntert P, Wüthrich K (1991) Improved efficiency of protein structure calculations from NMR data using the program DIANA with redundant dihedral angle constraints. *J Biomol NMR* 1:447–456.

Henderson R, Baldwin JM, Ceska TA, et al. (1990) Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *Journal of Molecular Biology* 213:899–929. doi: 10.1016/S0022-2836(05)80271-2

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

Herrmann T, Güntert P, Wüthrich K (2002b) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of Molecular Biology* 319:209–227. doi: 10.1016/S0022-2836(02)00241-3

CHAPTER 1 INTRODUCTION

Huang YJ, Powers R, Montelione GT (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. doi: 10.1021/ja047109h

Huang YJ, Tejero R, Powers R (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data - Huang - 2005 - *Proteins: Structure, Function, and Bioinformatics* - Wiley Online Library. *Proteins: Structure*

Jee J, Güntert P (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. *J Struct Funct Genomics* 4:179–189.

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* 292:195–202. doi: 10.1006/jmbi.1999.3091

Jung YS, Zweckstetter M (2004) Backbone assignment of proteins with known structure using residual dipolar couplings. *J Biomol NMR* 30:25–35.

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637. doi: 10.1002/bip.360221211

Karplus K, Karchin R, Draper J, et al. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* 53:491–496. doi: 10.1002/prot.10540

Kaufmann KW, Lemmon GH, DeLuca SL, et al. (2010) Practically Useful: What the ROSETTA Protein Modeling Suite Can Do for You. *Biochemistry* 49:2987–2998. doi: 10.1021/bi902153g

Kryshtafovych A, Venclovas C, Fidelis K, Moult J (2005) Progress over the first decade of CASP experiments. *Proteins* 61 Suppl 7:225–236. doi: 10.1002/prot.20740

Kumar A, Ernst RR, Wüthrich K (1980) A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. *Biochemical and Biophysical Research Communications* 95:1–6.

Kuszewski J, Schwieters CD, Garrett DS, et al. (2004) Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. *J Am Chem Soc* 126:6258–6273.

CHAPTER 1 INTRODUCTION

Lange OF, Baker D (2011) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins* 80:884–895.

Lange OF, Rossi P, Sgourakis NG, 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 USA* 109:10873–10878.

Leaver-Fay A, Tyka M, Lewis SM, et al. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Meth Enzymol* 487:545–574.

Leutner M, Gschwind RM, Liermann J, et al. (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *J Biomol NMR* 11:31–43.

Linge JP, Habeck M, Rieping W, Nilges M (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* 19:315–316. doi: 10.1093/bioinformatics/19.2.315

Liu Y, Kuhlman B (2006) RosettaDesign server for protein design. *Nucleic Acids Res* 34:W235–8.

Lukin JA, Gove AP, Talukdar SN, Ho C (1997) Automated probabilistic method for assigning backbone resonances of (¹³C,¹⁵N)-labeled proteins. *J Biomol NMR* 9:151–166.

Meiler J, Muller M, Zeidler A, Schmaschke F (2001) Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Journal of Molecular Modeling* 7:360–369.

Montelione GT, Lyons BA, Emerson SD (1992) An efficient triple resonance experiment using carbon-13 isotropic mixing for determining sequence-specific resonance assignments of isotopically-enriched proteins. *J. Am. Chem. Soc.*

Moseley HN, Montelione GT (1999) Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struct Biol* 9:635–642.

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

Nietlispach D, Clowes RT, Broadhurst RW, et al. (1996) An Approach to the Structure Determination of Larger Proteins Using Triple Resonance NMR Experiments in Conjunction with Random Fractional Deuteration. *J Am Chem Soc* 118:407–415.

Noda I, Ozaki Y (2005) *Two-Dimensional Correlation Spectroscopy: Applications in Vibrational and ...* - Isao Noda, Yukihiko Ozaki - Google Books. John Wiley & Sons

Noggle J (1971) *The Nuclear Overhauser Effect*. Elsevier

CHAPTER 1 INTRODUCTION

Olejniczak ET, Xu RX, Fesik SW (1992) A 4D HCCH-TOCSY experiment for assigning the side chain ^1H and ^{13}C resonances of proteins. *J Biomol NMR* 2:655–659.

Pechkova E, Nicolini C (2003) *Proteomics and Nanocrystallography*. Springer

Prestegard JH, Mayer KL, Valafar H, Benison GC (2005) Determination of protein backbone structures from residual dipolar couplings. *Meth Enzymol* 394:175–209. doi: 10.1016/S0076-6879(05)94007-X

Raman S, Lange OF, Rossi P, et al. (2010) NMR structure determination for larger proteins using backbone-only data. *Science* 327:1014–1018. doi: 10.1126/science.1183649

Rieping W, Habeck M, Bardiaux B, et al. (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics* 23:381–382. doi: 10.1093/bioinformatics/btl589

Rohl CA, Baker D (2002) De Novo Determination of Protein Backbone Structure from Residual Dipolar Couplings Using Rosetta. *J Am Chem Soc* 124:2723–2729. doi: 10.1021/ja016880e

Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein Structure Prediction Using Rosetta. In: *Methods in enzymology*. Elsevier, pp 66–93

Saibil HR (2000) Macromolecular structure determination by cryo-electron microscopy. *Acta Crystallographica Section D: Biological ...*

Schmidt E, Güntert P (2012) A New Algorithm for Reliable and General NMR Resonance Assignment. *J Am Chem Soc* 134:12817–12829. doi: 10.1021/ja305091n

Schot G, Zhang Z, Vernon R, et al. (2013) Improving 3D structure prediction from chemical shift data. *J Biomol NMR*. doi: 10.1007/s10858-013-9762-6

Serrano P, Pedrini B, Mohanty B, et al. (2012) The J-UNIO protocol for automated protein structure determination by NMR in solution. *J Biomol NMR* 53:341–354. doi: 10.1007/s10858-012-9645-2

Shapovalov MV, Dunbrack RL Jr. (2011) A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* 19:844–858. doi: 10.1016/j.str.2011.03.019

Shen Y, Bax A (2010) SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network - Springer. *J Biomol NMR*

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. doi: 10.1007/s10858-009-9333-z

CHAPTER 1 INTRODUCTION

Shen Y, Vernon R, Baker D, Bax A (2008a) De novo protein structure generation from incomplete chemical shift assignments. *J Biomol NMR* 43:63–78. doi: 10.1007/s10858-008-9288-5

Shen Y, Zhang Z, Delaglio F, et al. (2008b) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105:4685–4690. doi: 10.1073/pnas.0800256105

Simons KT, Bonneau R, Ruczinski I (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA - Simons - 1999 - *Proteins: Structure, Function, and Bioinformatics* - Wiley Online Library. *Proteins: Structure*

Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*

Thompson JM, Sgourakis NG, Liu G, et al. (2012) Accurate protein structure modeling using sparse NMR data and homologous structure information. *Proc Natl Acad Sci USA* 109:9875–9880. doi: 10.1073/pnas.1202485109

Vernon R, Shen Y, Baker D, Lange OF (2013) Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker. *J Biomol NMR* 57:117–127. doi: 10.1007/s10858-013-9772-4

Wang X, Tash B, Flanagan JM, Tian F (2011) RDC derived protein backbone resonance assignment using fragment assembly. *J Biomol NMR* 49:85–98. doi: 10.1007/s10858-010-9467-z

Wüthrich K (1989) Protein-Structure Determination in Solution by Nuclear Magnetic-Resonance Spectroscopy. *Science* 243:45–50.

Yu H (1999) Extending the size limit of protein nuclear magnetic resonance. *Proc Natl Acad Sci USA* 96:332–334.

Zeng J, Zhou P, Donald BR (2011) Protein side-chain resonance assignment and NOE assignment using RDC-defined backbones without TOCSY data. *J Biomol NMR* 50:371–395. doi: 10.1007/s10858-011-9522-4

Zhang Z, Lange OF (2013) Replica Exchange Improves Sampling in Low-Resolution Docking Stage of RosettaDock. *PLoS ONE*. doi: 10.1371/journal.pone.0072096

Zhang Z, Porter J, Lange OF (2014) Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta. *J Biomol NMR*

CHAPTER 1 INTRODUCTION

Zimmerman DE, Kulikowski CA, Huang Y, et al. (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology* 269:592–610.

Chapter 2 Improving 3D structure prediction from chemical shift data

2.1 Introduction

Knowledge of the three-dimensional (3D) structure of proteins at atomic accuracy is important to understand protein function, protein-ligand interactions and for rational drug design. Over the last two decades nuclear magnetic resonance spectroscopy (NMR) has become an established complement to X-ray crystallography for the determination of 3D structures. The most challenging bottleneck in determining NMR structures, the assignment of side-chain chemical shifts and of NOE cross-peaks, can be avoided with methods for computing structures from backbone-only NMR experiments. Backbone chemical shift values reflect a wide array of structural information including backbone and side-chain conformations, secondary structure, hydrogen bond strength, and the position of aromatic rings. This information can be exploited to predict the 3D structure of proteins using software packages such as CS-ROSETTA, CHESHIRE and CS23D (Cavalli et al. 2007; Shen et al. 2008; Wishart et al. 2008).

The convergence and reliability of CS-ROSETTA calculations have been shown to improve by utilizing additional NMR data, such as residual dipolar couplings (RDC) (Raman et al. 2010a), NOE-derived distance restraints (Lange et al. 2012) and pseudo-contact shifts (PCS) (Schmitz et al. 2012). In the context of available RDC and NOE data an iterative sampling scheme, RASREC (Raman et al. 2010a; Lange and Baker 2011; Lange et al. 2012), was shown to greatly extend the applicability towards larger protein structures. Here we introduce a number of algorithmic advances whose cumulative effect significantly improves reliability, convergence and accuracy of final structures for chemical shift-only calculations. Moreover, we describe the WeNMR (Wassenaar et al. 2012) webserver that accesses the European Grid Initiative (EGI, www.egi.eu) computational resources, allowing efficient CS-ROSETTA computations via the simplicity of a web interface to academic users.

The CS-ROSETTA methodology consists of three stages: 1) fragment picking, 2) sampling, and 3) model selection. Originally, backbone chemical shift information was only used in stages 1) and 3). Fragment picking for CS-ROSETTA was originally carried out using the MFR method of the NMRPipe software package (Delaglio et al. 1995; Lange et al. 2012) which combined chemical shift information with peptide sequence matching to score fragment candidates. However, for regions where no experimental data was present the

ROSETTA2 method (Rohl et al. 2004; Schmitz et al. 2012) outperformed MFR (Shen et al. 2009b; Lange and Baker 2011). In the present work the chemical shift based fragment picking is incorporated directly into a new ROSETTA3 fragment picker. This new fragment picker (denoted R3FP in the following) combines salient features of both original algorithms (MFR and ROSETTA2)(RV, YS, DB and OFL; in preparation). The performance of the new method is benchmarked on a set of target proteins that have not been used for development or optimization of the R3FP protocol.

RASREC is an iterative sampling algorithm that has been shown to significantly increase sampling efficiency for larger proteins (10-40kDa), if additional restraint data such as RDCs and NOEs are used (Lange and Baker 2011; Lange et al. 2012). Instead of running 10,000 or more independent structure calculations with increased cycle number as in CS-ABRELAX (the standard CS-ROSETTA algorithm (Shen et al. 2008)), RASREC performs iterative batches of short simulations. Similar to a genetic algorithm, a pool of best performing structures is maintained throughout the iterative procedure and sampling is focused around previously identified conformations. It is crucial for the performance of RASREC that pseudo-energies (e.g., from RDC and NOE restraint data) be available to assist ROSETTA in predominantly selecting structures with native features for this pool. In this study we extend the RASREC algorithm to allow chemical shift rescoring of intermediate structures (CS-RASREC) and test the performance of this extended method.

Chemical shifts are dominated by local backbone conformations, and thus CS-ROSETTA structures based solely on chemical shifts tend to be locally correct but globally unconverged. Here we introduce a post-analysis procedure that identifies locally converged regions of the structure, which have been shown to be generally accurate (Rohl et al. 2004; Shen et al. 2008; Raman et al. 2010a). However, with decreasing convergence there is an increasing probability that also the conformation of the converged part is inaccurate. We address this issue here by testing a number of criteria, including the quality of chemical shift data, the number of converged residues and the significance of the ROSETTA energy gap, to detect inaccurate predictions. These criteria are aggregated to annotate each CS-ROSETTA prediction as *weak* or *strong*, thereby providing users with a reliability metric to assess the results.

2.2 Materials and Methods

We benchmarked the performance of the new fragment picker (R3FP) and CS-RASREC on a set of 39 proteins in the size range of 50-100 residues that have neither been used for training of the R3FP, nor in CS-ROSETTA, SPARTA+ or TALOS+ development

(Table 2.2+2.3). All input files (fragments, reference coordinate and chemical shift files) are available for download at www.csrosetta.org/benchmarks.

2.2.1 Target Selection and Fragment Picking

The benchmark set was selected from a larger set of 206 proteins for which recently released chemical shift information from the BMRB was linked to coordinate information from the PDB in the CCPN framework (Vranken and Rieping 2009) and re-referenced using the VASCO protocol (Rieping and Vranken 2010). For NMR resolved structures, only proteins of sequence length 50-100 with at least 40% secondary structure were retained from this set. Homologous proteins using an e-value cutoff of 0.05 (sequence identity > 20 %) were excluded from MFR and R3FP fragment picking. The resulting set of 39 proteins covers a wide range of secondary structure content, as determined by DSSP from the PDB deposited structures. Since TALOS+ is used to pre-filter CS-ROSETTA submissions, and because the TALOS+ predicted secondary structure content is similar to what DSSP determines from the coordinates (Figure 2.1), this set of 39 is expected to be representative of typical CS-ROSETTA input.

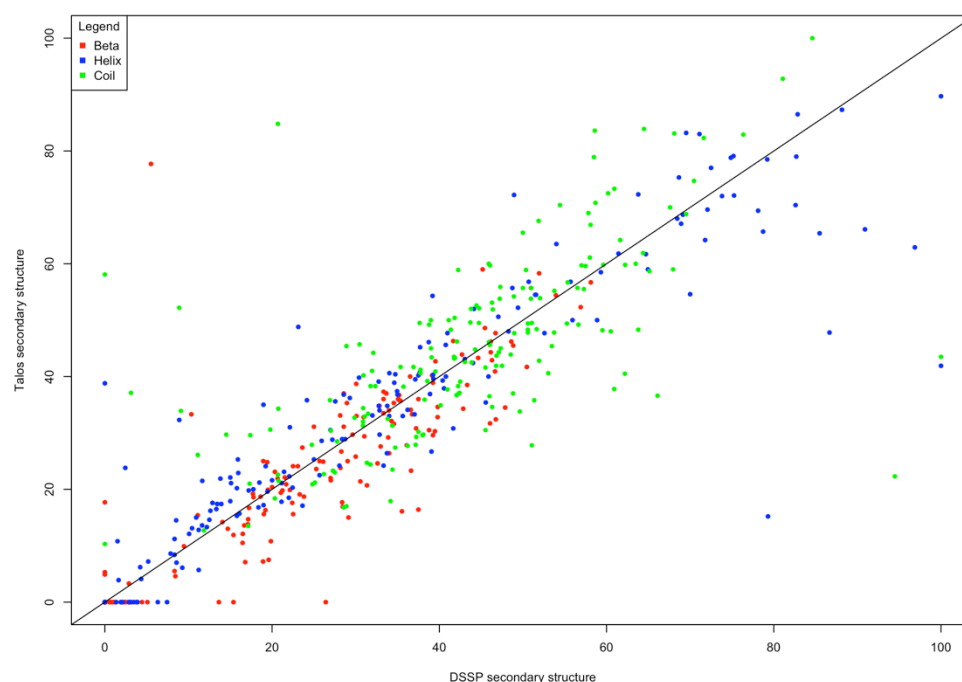


Figure 2.1: The secondary structure content as predicted by TALOS+ from the chemical shifts versus the secondary structure content as determined by DSSP from the PDB-deposited structures for 181 proteins for which sufficient heteronuclear chemical shift data was available to run TALOS+.

2.2.2 Structure Generation

CS-ROSETTA (Server)

The latest version of the CS-ROSETTA webserver runs ROSETTA 3.3 including the new fragment selection method R3FP. For each target in the benchmark, 50,000 models were automatically generated by the CS-ROSETTA web server, using the standard CS-ABRELAX protocol with the ABRELAX cycle factor (command-line flag *-increase_cycles*) set to 10 as in Ref (Shen et al. 2008). The jobs are automatically distributed to available computational resources part of the worldwide WeNMR grid under the European Grid Initiative (EGI). As input, only a backbone NMR chemical shift list is required, which can be supplied in any of the common NMRPipe (TALOS), NMR-Star 2.1, or NMR-Star 3.1 (BMRB) formats.

The webserver uses SPARTA+ (Shen and Bax 2010) for final model selection in analogy to the original procedure based on SPARTA (Shen et al. 2008). Additionally, the server can combine the chemical shifts score with the DP score (Huang et al. 2005) based on unassigned NOE data for model selection, which has been shown to improve model selection from CS-ROSETTA calculations (Raman et al. 2010b; Rosato et al. 2012).

An overview of the CS-ROSETTA web portal workflow can be found in Figure 2.2.

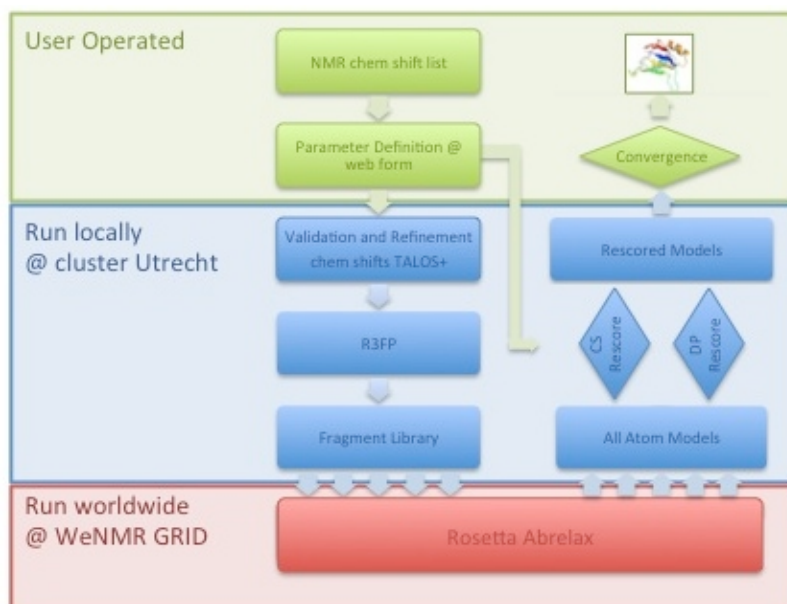


Figure 2.2: Workflow of the grid-enabled CS-ROSETTA web portal (Wassenaar et al. 2012). Green indicates user-operated steps. Blue indicates calculations run locally at the cluster in Utrecht, and red indicates calculations run worldwide on the WeNMR grid (www.wenmr.eu). Step 1: The user submits a NMR chemical shift file and defines several input parameters in the web form; 2) Using these parameters, molecular fragments are selected for the query

CHAPTER 2 IMPROVING 3D STRUCTURE PREDICTION FROM CHEMICAL SHIFT DATA

protein; 3) the fragments are assembled using the ROSETTA ABRELAX protocol; 4) Optionally (as indicated in input form) the generated models are rescored using several methods; 5) The user evaluates the top ten selected models. If prediction is reliable, the models are selected.

RASREC

RASREC structure calculations (Lange and Baker 2011) with a pool-size of 500 conformers (command-line flag *-iterative:pool_size 500*) were started from the same fragment libraries as CS-ABRELAX calculations. As in the standard protocol (Lange and Baker 2011), *Recombination-Stages* were terminated when the acceptance ratio into the pool dropped below 10% (*-iterative:accept_ratio 0.1*) and the cycle factor was set to 2.0 (*-increase_cycles 2*). The protocol was modified to add chemical shift pseudo-energies with a weight of 5.0 to the ROSETTA energy to bias the RASREC pool of low-energy structures towards conformations in agreement with the experimental chemical shifts. Chemical shifts were computed from conformations using SPARTA+ (Shen and Bax 2010) and compared to the experimental chemical shifts to yield a pseudo-energy as described previously (Shen et al. 2008). To improve the prediction of chemical shifts from intermediate low-resolution structures a shortened refinement procedure was applied that uses only 1 of the usual 5 relax cycles (Raman et al. 2010a). SPARTA+ was implemented as a module of ROSETTA to allow computation of chemical shift pseudo-energies during RASREC iterations.

2.2.3 Calculation of converged regions

To determine the converged region of a protein structure predicted with CS-ROSETTA an adaption of the Gaussian-weighted RMSD method (Damm and Carlson 2006) was implemented in ROSETTA. The 30 lowest energy structures were superimposed using a scaling factor of 2\AA^2 (Damm and Carlson 2006). This procedure iteratively determines a set of rigid residues on which the structures can be superimposed; residues with a root-mean square fluctuation (RMSF) below 2\AA are considered to be converged. Gaps of less than 3 residues between regions of low RMSF ($<2\text{\AA}$) are ignored.

2.2.4 RMSD Calculations

All reported RMSDs are C_{α} -RMSD to the PDB deposited reference structure or its first model. If the reference structure stems from an NMR solution ensemble only residues that superimpose within 1\AA in the deposited ensemble were used. Where indicated, C_{α} -RMSD computations are further *restricted* to regions converged in the ROSETTA calculations (see Methods).

2.2.5 Criteria used for annotations

The criteria of *strong/weak* prediction annotation are slightly different between CS-RASREC and CS-ABRELAX. *cs-consensus*, *convergence* and *energy gap* are used to annotate the prediction from CS-RASREC, and for CS-ABRELAX the criteria are *cs-class*, *convergence* and *energy gap*. *cs-consensus* is the fraction of residues for which TALOS+ finds more than 7 consensus matches in the database. *cs-class* is the fraction of residues annotated by TALOS+ with 'GOOD'. *convergence* is the fraction of residues which are *converged* with an RMSF cutoff of 2Å (see Methods). The *energy gap* is the difference in ROSETTA all-atom energy between the lowest energy decoys and the lowest energies obtained for decoys far away (>4Å) from the lowest energy decoys. Specifically, it is calculated as follows: the median energy of the 10 lowest energy structures is subtracted from the median energy of the 10 lowest energy structures within the subset of structures that are more than >4Å (converged region; see Methods) from the lowest energy structure. In CS-RASREC annotations, the raw energy gap is divided by the number of residues and mapped to an interval 0...1 using a sigmoidal function (Figure 2.3) with its inflection point at 0.05 Rosetta Energy Units (REU) per residue (see *Classification of 3D structure predictions*). Differently, the raw energy gap is directly mapped to [0,1] using sigmoidal function in CS-ABRELAX annotations.

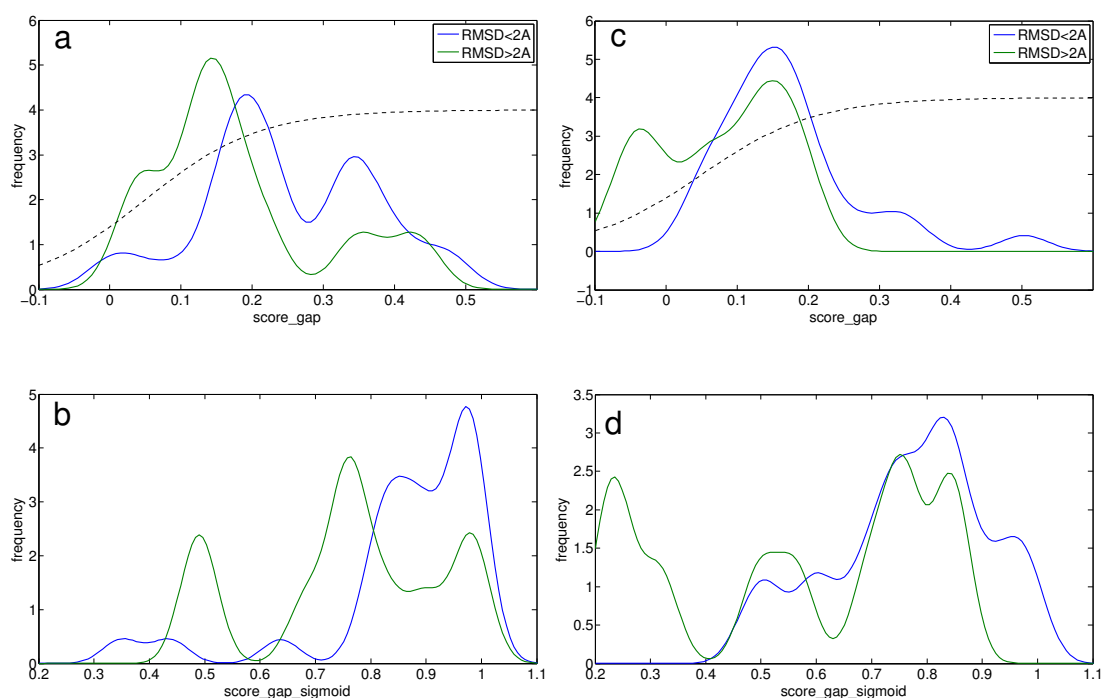


Figure 2.3: Choice of sigmoid parameters to map energy gap onto interval [0,1].

Panels a+c) The inflection point and slope of the sigmoid (dashed line), was chosen such that the steepest slope of the sigmoid (and thus the region of highest sensitivity) coincides

with the transition zone between the distributions of energy gaps for good (RMSD <2Å, blue) and bad (RMSD >2Å, green) final models of CS-RASREC (a) and CS-ABRELAX (c).

Panels b+d) Distribution of the sigmoid-enhanced output for energy gaps of good and bad final models. Clearly, the sigmoid has served to enhance the separation of the two distributions for CS-RASREC (b) and CS-ABRELAX (d) models.

2.2.6 Classification of 3D structure predictions

For both CS-ABRELAX and CS-RASREC, we separately developed a classification of 3D structure predictions into *strong* and *weak* predictions based on the criteria *cs-consensus*, *convergence* and *energy-gap*(CS-RASREC) or *cs-class*, *convergence* and *energy-gap*(CS-ABRELAX), as described in Methods of the main-text. The three criteria yield values in the range 0...1 and

$$P_{\text{sum}} = \sum_{i=1}^3 w_i c_i, \quad (2.1)$$

is used as predictor model. The classification *strong* is reached if $P_{\text{sum}} > T$. The threshold T and weights w_i were determined using a cross-validated fit-procedure that optimized the receiver operator characteristics (ROC). To this end, the 39 structure prediction results for CS-ABRELAX and CS-RASREC were classified manually into *strong* and *weak*. Targets 2jov, and 2k5c were excluded from this analysis, due to questionable packing quality of the respective reference structures (Figure 2.8). The remaining classifications were randomly separated into a *Training* set with 75% and a *Test* set with 25% of data points. A grid-search for the weights w_i was performed and the set of weights with the largest area between diagonal and train-ROC curve were selected. The optimal weight set was evaluated on the Test set to yield the test-ROC curve. This procedure was repeated 100 times with different random separations into Training and Test sets. The resulting test-ROC curves were averaged and are shown as Figure 2.7.

Different statistics for c_i were evaluated as model 1-4. In Models 1-3 the energy gap was mapped to the interval 0...1 using a sigmoidal function, whereas in Model 4 the energy gap was taken directly. In Model 1 the energy gap is divided by the number of residues before the sigmoid is applied. In Model 3 we swapped *cs_consensus* (fraction of residues with more than 7 consensus matches in the database) against *cs_class* (fraction of residues annotated by TALOS+ with 'GOOD'). The cross-validated ROCs allowed to select Model 3 for CS-ABRELAX and Model 1 for CS-RASREC as the procedure that yielded best prediction characteristics (Figure 2.7). Selecting thresholds by fixing the false positive rate to 6% and 3% thresholds of 0.69 ± 0.05 and 0.82 ± 0.06 are obtained for CS-ABRELAX and CS-

RASREC, respectively. A scatter plot of convergence and energy gap is shown in Figure 2.6.

Additionally, we evaluated a fourth criterion, which counted the residues for which the RCI predicted S^2 order parameter⁶ was above 0.7. However, this criterion was discarded, since its addition did not improve cross-validated ROC curves.

The parameters for the sigmoidal function were not fitted but were fixed to 5 Rosetta energy units (REU) for the mid-point and the response parameter such that 10 REU above the mid-point yield a response of 90%. For energies normalized by number of residues the parameters were set such that energy gaps of 0.05 REU/residue and 0.2 REU/residue yield 50% and 90% response, respectively. These values were determined by visual inspection of a scatter plot of energy gap and CA RMSD (Figure 2.3).

2.2.7 Weak/strong-classification with CS-Rosetta toolbox

To obtain the classification *weak/strong* for a finished CS-Rosetta calculation, run *annotate_target* from the CS-Rosetta Toolbox Version 2.x or higher (www.csrosetta.org).

```
annotate_target -type abrelax -pred 3cwi/pred.tab -run_folder 3cwi/abrelax/run/
```

with the following inputs:

1. *type*: algorithm type for structure calculation, it should be *abrelax* or *rasrec*.
2. *pred*: the *pred.tab* file generated by talos+, it contains the cs-consensus of the target.
3. *run_folder*: directory of the job.

The command firstly outputs the values of cs-consensus(or cs-class), convergence and energy gap of the structure and then the annotation result.

Example output of annotation of CS-RASREC calculation:

#	<i>cs-consensus</i>	<i>convergence</i>	<i>energy gap</i>	<i>classification</i>
	0.937	0.810	0.978	STRONG

Example output of annotation of CS-ABRELAX calculation:

#	<i>cs-class</i>	<i>convergence</i>	<i>energy gap</i>	<i>classification</i>
	0.461	0.670	0.524	WEAK

2.3 Results

2.3.1 Performance of new fragment picker (R3FP)

Figure 2.4a shows the mean C_{α} -RMSD of the best 10 generated models with respect to the reference structure for MFR and R3FP. As can be seen, more targets appear above the diagonal, i.e., ABRELAX samples closer to the native structure, if R3FP fragments are used. Necessary for the success of a CS-ROSETTA structure calculation is that sufficient conformations below a C_{α} -RMSD of 2.0Å to the reference structure are generated (Shen et al. 2008). This is the case for significantly more targets, if R3FP fragments are used (Figure 2.4a, Table 2.1).

We also compared the performance in sampling near-native conformations of ABRELAX between software versions Rosetta 2.6 (used here (Shen et al. 2008; Wassenaar et al. 2012)) and Rosetta 3.x (used here (Raman et al. 2010b; Schmitz et al. 2012)). As shown in Table 2.1, a performance gain is observed for Rosetta 3.x.

Version ^a	Fragments ^b	native sampling rate ^c	RMSD (Å) ^d
Rosetta2-ABRELAX ^e	MFR	62%	1.12 ± 0.42
Rosetta3-ABRELAX ^e	MFR	72%	1.23 ± 0.48
Rosetta3-ABRELAX ^e	R3FP	77%	1.18 ± 0.39
Rosetta3-RASREC	R3FP	64% ^f	1.31 ± 0.41 ^f
Rosetta3-CS-RASREC	R3FP	74% ^f	1.27 ± 0.43 ^f

Table 2.1: Success of structure generation for MFR and R3FP fragment picker

Footnote:

^a Major version number of Rosetta

^b Fragment picking protocols

^c Success rate of the structural sampling step defined as the percentage of targets for which the mean C_{α} -RMSD of the 10 lowest RMSD structures is lower than 2.0 Å; this reflects if the method samples the native structure, not how well it predicts it. C_{α} -RMSDs are calculated over all residues that are converged within 1Å in the reference NMR structural ensemble (Table 2.2).

^d Average and standard deviation of the distribution of mean C_{α} -RMSDs, when restricted to those targets where the mean C_{α} -RMSD of 10 lowest RMSD structures is lower than 2.0Å.

^e In CS-ABRELAX the chemical shifts are only used for final model selection. The native sampling rate is independent of final model selection, and thus CS-ABRELAX

and ABRELAX are equivalent in this analysis. Note, however, that chemical shifts are used for fragment picking for all protocols analysed in this table.

^f For RASREC protocols, the native sampling rate is systematically lower than for ABRELAX, since instead of 50,000 full-atom models in ABRELAX, only ca. 1,500 full-atom models are generated in RASREC.

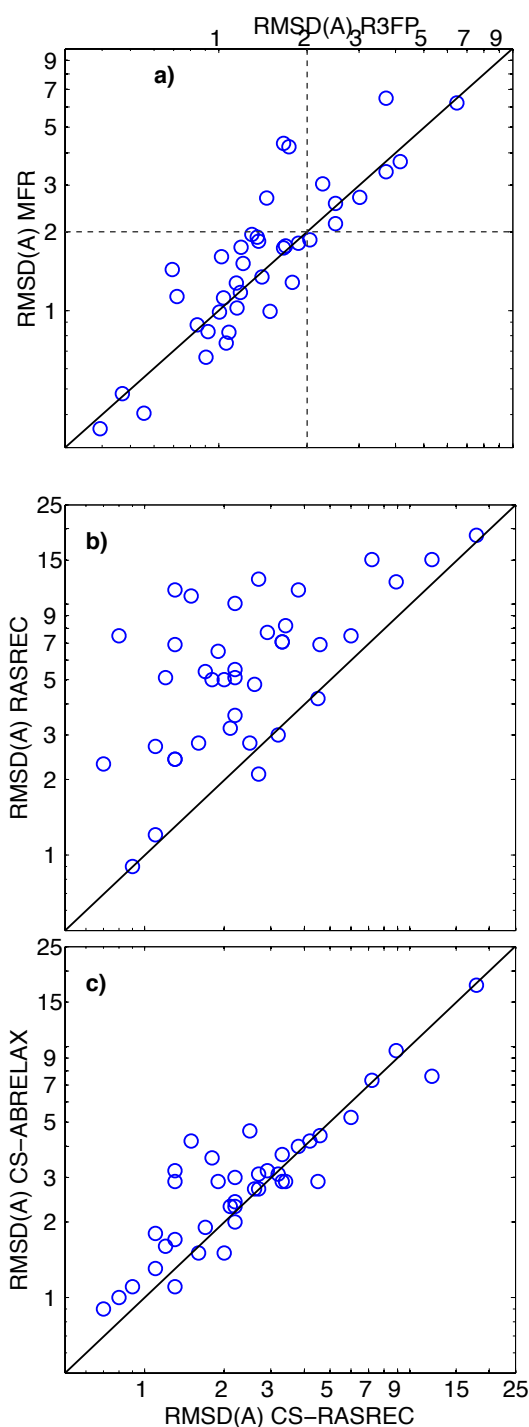


Figure 2.4: Performance comparison. **a)** Comparison of MFR and R3FP fragment picking methods using the ABRELAX sampling protocol. Shown are the mean C_{α} -RMSD of the

lowest 10-RMSD structures (Table 2.1). Dashed lines indicate the 2Å RMSD threshold, which is often predictive whether CS-Rosetta yields converged ensembles after energy-based selection. **b)** Comparison of CS-RASREC(x-axis) and RASREC(y-axis). Shown are the median RMSDs of the ten lowest energy models selected by Rosetta energy and chemical shift score. **c)** Comparison of CS-RASREC (x-axis) with CS-ABRELAX(y-axis). Shown are RMSDs of 10 lowest energy models selected by Rosetta energy and chemical shift score as in b).

2.3.2 RASREC with chemical shift rescoring

As shown previously (Rohl et al. 2004; Shen et al. 2008), chemical shift rescoring improves precision and accuracy of final target selection for the CS-ABRELAX method. We have now implemented SPARTA+ rescoring directly into ROSETTA which allows us to apply the chemical shift score as a filter between iterations of the RASREC method (Shen et al. 2009b; Lange and Baker 2011). However, chemical shift rescoring is usually applied to fully refined structures, whereas intermediate structures in RASREC are without atomic detail (i.e., they use only a single *centroid* to represent the side-chain). To allow chemical shift rescoring nevertheless, a short refinement to atomic detail models that requires only ca. 20% of the usual computer time is applied (see Methods).

We investigated whether chemical shift rescoring of intermediate structures improves the performance of the new CS-RASREC protocol on the benchmark set of 39 proteins. Indeed, a significant improvement in the RMSDs of the final energy selected models (Figure 2.4b) is seen for CS-RASREC (points left of diagonal). Thus, CS-RASREC (but not RASREC) can further improve the accuracy of final models in comparison to CS-ABRELAX with R3FP fragments (Figure 2.4c).

2.3.3 Restriction to converged regions

Figure 2.5a shows the C_{α} -RMSD to the reference structure of the lowest 10 scoring models from CS-ABRELAX calculations. As can be seen, only a small fraction of targets (~25%) yields accurate (<2Å) solutions. The reason for this apparent bad accuracy of CS-ROSETTA predictions is that RMSDs were computed on regions that are not converged in the CS-ROSETTA ensemble. To address this issue we added an auxiliary application called *ensemble_analysis* to the CS-ROSETTA toolbox (www.csrosetta.org), which detects residues whose RMSD fluctuations are less than 2Å (see Methods). Restriction of the structural prediction to these converged residues drastically changes the appearance of the results and shows that the converged regions are actually quite accurate for the majority of targets, with only five targets where the accuracy is worse than 2.5Å (Figure 2.5b).

However, Figure 2.5b also reveals that for many targets significant portions of the structures remain unconverged in the CS-ABRELAX calculations. As can be seen in Figure 2.5c, the convergence is significantly improved in CS-RASREC calculations.

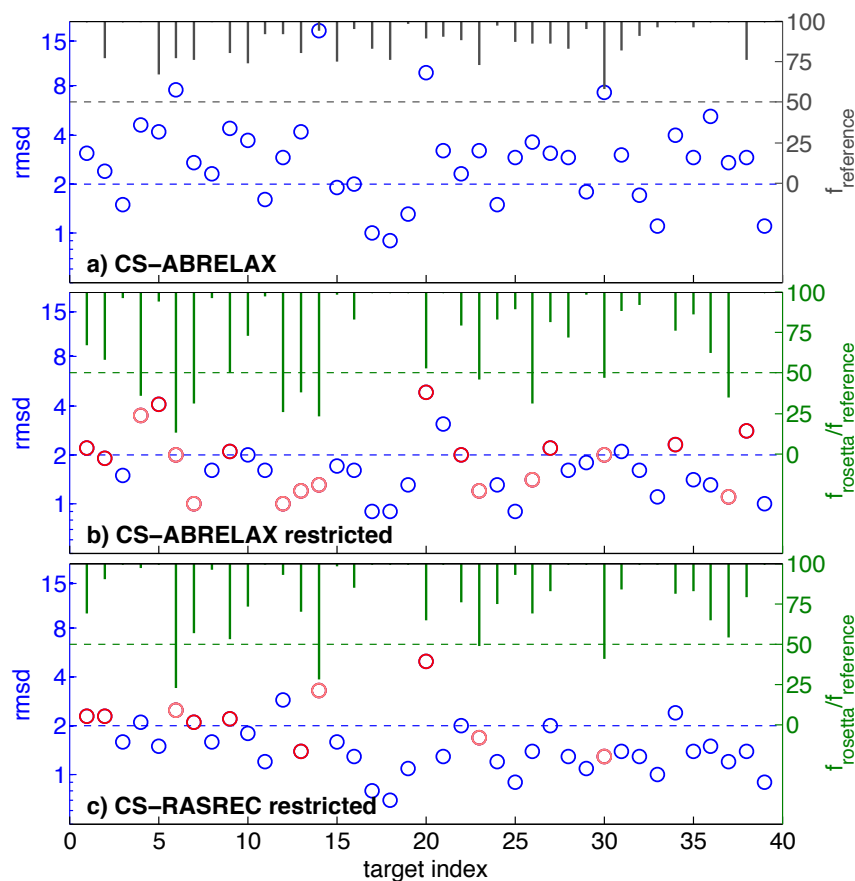


Figure 2.5: Overview of accuracy of 10 lowest scoring structures from the 39 protein benchmark. C_{α} -RMSD to the reference structure (circles) are calculated over a subset of residues (bars). Predictions annotated as *weak* are shown in red (convergence is more than 50%) or pink (convergence is less than 50%) (Table 2.4+2.5). **a)** The RMSDs are calculated over all residues that are converged within 1Å in the reference NMR structural ensemble (Table 2.2). The number of residues used for RMSD calculation are shown as fraction of total length $f_{reference}$ (gray). **b)** The RMSD calculation is restricted to residues converged within 2Å in the CS-ROSETTA structural ensemble (and within 1Å in the references) (Table 2.2). The additional restriction in RMSD calculation is given as ratio $f_{rosetta} / f_{reference}$ (green). **c)** RMSD restriction as in b) but using the CS-RASREC method.

PDB	Exp. method	Residue NO.				RMSD ^e
		length ^a	trimmed ^b	converged in NMR ^c	converged in CS-ABRELAX ^d	

CHAPTER 2 IMPROVING 3D STRUCTURE PREDICTION FROM CHEMICAL SHIFT DATA

1	1ig5	XRAY	75	1-75	1-75	23-72 ^g	2.2(2.0-2.6)
2	1q02	NMR	52	1-52	10-49	27-49	1.9(1.8-2.1)
3	2ckx	XRAY	83	1-83	1-83	4-83	1.5(1.1-2.0)
4	2dm2	NMR	110	8-104	8-103	17-19,54-59,65-84,94-99	3.5(3.3-3.7)
5	2htj	NMR	81	2-77	2-46,54-59 ^f	2-45,56-59	4.1(4.0-4.5)
6	2hx6	NMR	153	18-115	18-79,103-115	28-37	2.0(1.9-2.9)
7	2ike	NMR	54	1-54	1-28,36-43,49-54	16-28	1.0(0.6-1.3)
8	2jmb	NMR	79	1-79	2-79	5-79	1.6(1.4-1.8)
9	2jml	NMR	81	1-80	6-35,44-77	46-77	2.1(2.0-2.2)
10	2jmp	NMR	100	2-100	3-22,31-83	31-83	2.0(1.7-2.5)
11	2joq	NMR	91	6-89	13-89	15-89	1.6(1.5-2.5)
12	2jov	NMR	85	2-79	3-74	27-45	1.0(0.7-1.3)
13	2jpn	NMR	79	4-79	12-72	49-71	1.2(1.0-1.5)
14	2jq3	NMR	79	1-79	6-79	48-64	1.3(1.0-1.7)
15	2jrm	NMR	60	1-60	5-49	6-49	1.7(1.4-2.0)
16	2jso	NMR	88	1-88	1-84	3-72	1.6(1.3-2.6)
17	2jsx	NMR	95	3-88	5-75	5-75	0.9(0.7-1.1)
18	2jt1	NMR	71	2-71	5-18,24-57,66-70	5-18,24-57,66-70	0.9(0.7-1.0)
19	2jtv	NMR	65	1-64	2-64	2-64	1.3(1.1-1.6)
20	2jub	NMR	76	1-76	2-11,18-75	30-65	4.8(4.5-5.2)
21	2jvf	NMR	94	1-94	4-22,29-94	4-22,29-94	3.1(1.5-3.4)
22	2jvr	NMR	80	1-80	3-38,45-78	4-24,45-78	2.0(1.7-2.2)
23	2jvw	NMR	82	2-82	15-73	47-73	1.2(0.9-2.3)
24	2jxt	NMR	86	3-81	4-80	5-68	1.3(1.1-1.5)

CHAPTER 2 IMPROVING 3D STRUCTURE PREDICTION FROM CHEMICAL SHIFT DATA

25	2jz5	NMR	91	1-86	6-81	12-79	0.9(0.9-1.0)
26	2k0m	NMR	104	2-98	6-70,76-93	6-31	1.5(0.9-1.7)
27	2k14	NMR	84	1-84	13-84	24-81	2.2(1.9-2.5)
28	2k19	NMR	98	1-98	12-92	35-92	1.6(1.3-2.5)
29	2k2p	NMR	64	1-64	2-63	2-62	1.8(1.1-2.2)
30	2k37	NMR	59	1-59	4-37	10-25	2.0(0.5-3.4)
31	2k3d	NMR	87	2-83	2-27,35-69,75-80	3-27,36-69	2.1(0.8-2.3)
32	2k4y	NMR	86	2-83	4-78	4-47,54-78	1.6(1.3-2.7)
33	2k52	NMR	74	1-74	1-71	2-71	1.1(1.0-1.5)
34	2k5c	NMR	88	1-88	1-88	2-68	2.3(2.1-3.0)
35	2k5n	NMR	74	2-69	2-66	2-50,60-66	1.4(1.2-1.9)
36	2k5s	NMR	73	1-69	1-69	7-49	1.3(1.1-1.9)
37	2osq	NMR	74	2-73	2-73	30-54	1.1(0.9-2.8)
38	2ot2	NMR	90	2-90	4-71	4-70	2.8(2.6-3.5)
39	2qmt	XRAY	56	1-56	1-56	1-56	1.0(0.8-1.5)

Table 2.2: Protein structures used to evaluate the performance of the CS-ABRELAX method.

Footnote:

^a The number of residues in the deposited structure

^b Residue range used for structure calculation in ROSETTA

^c Residues that superimpose within 1Å in solution NMR structures, if the experimental method is Xray, we consider it 100% converged.

^d The residues superimpose within 2Å in the 30 lowest-energy structures predicted by CS-ABRELAX. Predictions classified as *weak* (Table 2.4) are shown in red.

^e Median C_{α} -RMSD of the 10 lowest-energy structures calculated on the residues converged in CS-ABRELAX. The lowest and highest C_{α} -RMSD within the 10 lowest-energy structures is given in parenthesis. Predictions classified as *weak* (Table 2.4) are shown in red.

CHAPTER 2 IMPROVING 3D STRUCTURE PREDICTION FROM CHEMICAL SHIFT DATA

^f Target 2htj is an NMR structure, but only 1 model is deposited. Residues with S^2 order parameter predicted from TALOS+ smaller than 0.7 are considered as flexible.

^g Results in red and italics indicate structure calculations that are annotated as *weak* predictions.

	PDB	Exp. method	Residue NO.				RMSD ^e
			length ^a	trimmed ^b	converged in NMR ^c	converged in CS-RASREC ^d	
1	1ig5	XRAY	75	1-75	1-75	<i>13-14,21-70^g</i>	<i>2.3(1.8-2.8)</i>
2	1q02	NMR	52	1-52	10-49	<i>14-49</i>	<i>2.3(2.2-2.3)</i>
3	2ckx	XRAY	83	1-83	1-83	1-83	1.6(1.4-1.8)
4	2dm2	NMR	110	8-104	8-103	11-103	2.1(2.1-2.6)
5	2htj	NMR	81	2-77	2-46,54-59 ^f	2-47,55-62	1.7(1.3-2.0)
6	2hx6	NMR	153	18-115	18-79,103-115	<i>58-74</i>	<i>2.5(1.7-3.6)</i>
7	2ike	NMR	54	1-54	1-28,36-43,49-54	<i>12-22,36-43,50-54</i>	<i>2.1(1.7-2.5)</i>
8	2jmb	NMR	79	1-79	2-79	5-79	1.6(1.4-1.9)
9	2jml	NMR	81	1-80	6-35,44-77	<i>34-35,44-77</i>	<i>2.2(2.2-2.3)</i>
10	2jmp	NMR	100	2-100	3-22,31-83	31-83	1.8(1.7-3.2)
11	2joq	NMR	91	6-89	13-89	14-89	1.2(0.9-2.0)
12	2jov	NMR	85	2-79	3-74	4-70	2.9(2.4-3.6)
13	2jpn	NMR	79	4-79	12-72	<i>29-71</i>	<i>1.4(0.9-2.1)</i>
14	2jq3	NMR	79	1-79	6-79	<i>47-77</i>	<i>3.3(2.7-4.0)</i>
15	2jrm	NMR	60	1-60	5-49	6-49	1.6(1.6-1.7)
16	2jso	NMR	88	1-88	1-84	1-71	1.3(1.2-1.8)
17	2jsx	NMR	95	3-88	5-75	5-75	0.8(0.8-1.4)
18	2jt1	NMR	71	2-71	5-18,24-57,66-70	5-18,24-57,66-70	0.7(0.6-0.7)

CHAPTER 2 IMPROVING 3D STRUCTURE PREDICTION FROM CHEMICAL SHIFT DATA

19	2jtv	NMR	65	1-64	2-64	2-64	1.1(1.1-1.2)
20	2jub	NMR	76	1-76	2-11,18-75	29-69	5.0(4.7-5.7)
21	2jvf	NMR	94	1-94	4-22,29-94	4-22,29-94	1.3(1.2-1.8)
22	2jvr	NMR	80	1-80	3-38,45-78	3-21,45-78	2.0(1.8-2.6)
23	2jvw	NMR	82	2-82	15-73	15-43	1.7(1.2-2.0)
24	2jxt	NMR	86	3-81	4-80	5-62	1.2(1.0-1.6)
25	2jz5	NMR	91	1-86	6-81	11-81	0.9(0.8-1.0)
26	2k0m	NMR	104	2-98	6-70,76-93	20-51,59-70,76-91	1.4(1.1-1.9)
27	2k14	NMR	184	1-84	13-84	23-82	2.0(1.8-2.3)
28	2k19	NMR	98	1-98	12-92	12-92	1.3(1.3-1.7)
29	2k2p	NMR	64	1-64	2-63	2-63	1.1(0.9-1.9)
30	2k37	NMR	59	1-59	4-37	12-24	1.3(0.5-2.6)
31	2k3d	NMR	87	2-83	2-27,35-69,75-80	2-27,37-66	1.4(0.9-1.9)
32	2k4y	NMR	86	2-83	4-78	4-78	1.3(1.1-1.5)
33	2k52	NMR	74	1-74	1-71	2-71	1.0(0.9-1.5)
34	2k5c	NMR	88	1-88	1-88	2-72	2.4(2.2-2.9)
35	2k5n	NMR	74	2-69	2-66	2-49,60-66	1.4(1.2-1.8)
36	2k5s	NMR	73	1-69	1-69	6-53	1.5(1.3-1.9)
37	2osq	NMR	74	2-73	2-73	24-35,42-72	1.2(1.0-2.5)
38	2ot2	NMR	90	2-90	4-71	4-31,43-68	1.4(1.2-2.0)
39	2qmt	XRAY	56	1-56	1-56	1-56	0.9(0.7-1.3)

Table 2.3: Protein structures used to evaluate the performance of the CS-RASREC method.^a The number of residues in the deposited structure^b Residue range used for structure calculation in ROSETTA^c Residues that superimpose within 1Å in solution NMR structures, if the experimental method is Xray, we consider it 100% converged.

^d The residues superimpose within 2Å in the 30 lowest-energy structures predicted by CS-RASREC. Predictions classified as *weak* (Table 2.5) are shown in red.

^e Median C_{α} -RMSD of the 10 lowest-energy structures calculated on the residues converged in CS-RASREC. The lowest and highest C_{α} -RMSD within the 10 lowest-energy structures is given in parenthesis. Predictions classified as *weak* (Table 2.5) are shown in red.

^f Target 2htj is an NMR structure, but only 1 model is deposited. Residues with S^2 order parameter predicted from TALOS+ smaller than 0.7 are considered as flexible.

^g Results shown in red and italics indicate structure calculations that are annotated as *weak* predictions.

2.3.4 Reliability measure: Annotation of weak/strong predictions

Originally, CS-Rosetta calculations were discarded if they did not converge on all residues (Shen et al. 2008; Schmitz et al. 2012). However, as shown above, some of the calculations that contain converged segments yield quite acceptable models. Thus, we looked for additional criteria to detect accurate predictions. We speculated that, in addition to a) the overall convergence of the calculation, also the significance of b) the chemical shift consensus or class (Shen et al. 2009a) and c) the ROSETTA energy gap (Raman et al. 2010a; Fleishman and Baker 2012) should be informative on the likelihood of obtaining accurate structures.

To this purpose we define a predictor model that yields the signal *strong* if the weighted sum of the criteria c_i

$$P_{\text{sum}} = \sum_{i=1}^3 w_i c_i \quad (2.2)$$

exceeds a threshold of 0.82 or 0.69 for CS-RASREC and CS-ABRELAX calculations, respectively. Optimizing the predictor model against manual classifications of the benchmark results, we obtained for CS-RASREC the weights 0.58, 0.29 and 0.13 for the criteria *cs-consensus*, *convergence*, and *energy-gap*, respectively. For CS-ABRELAX the weights 0.08, 0.54 and 0.38 for criteria *cs-class*, *convergence*, and *energy-gap*. The criteria are defined in Material and Methods section. In 100 rounds of cross-validated training using a different random selection of 25% of the data as test-set for each round, for CS-RASREC the cutoff was selected by fixing the false positive rate (FPR) to 3% and 6% for CS-RASREC and CS-ABRELAX, respectively. The resulting thresholds of 0.82 ± 0.06 and 0.69 ± 0.05 yielded true positive rates (TPR) of $(89 \pm 20)\%$ and $(80 \pm 33)\%$, respectively. The standard deviation of the weights trained on the 100 different selections of training data during cross-validation were 0.08, 0.10 and 0.08, for CS-RASREC and 0.25, 0.13, 0.14 for CS-ABRELAX. The

higher variation of weights for CS-ABRELAX reflects the less pronounced energy gap and the lower rate of convergence observed in CS-ABRELAX simulations (Figure 2.6). Alternative predictor models were discarded based on inferior receiver operating characteristic (ROC) in the cross-validation and the compound predictor model outperforms the individual criteria (Figure 2.7). The final set of weights was obtained by optimizing the most successful predictor model against all data points.

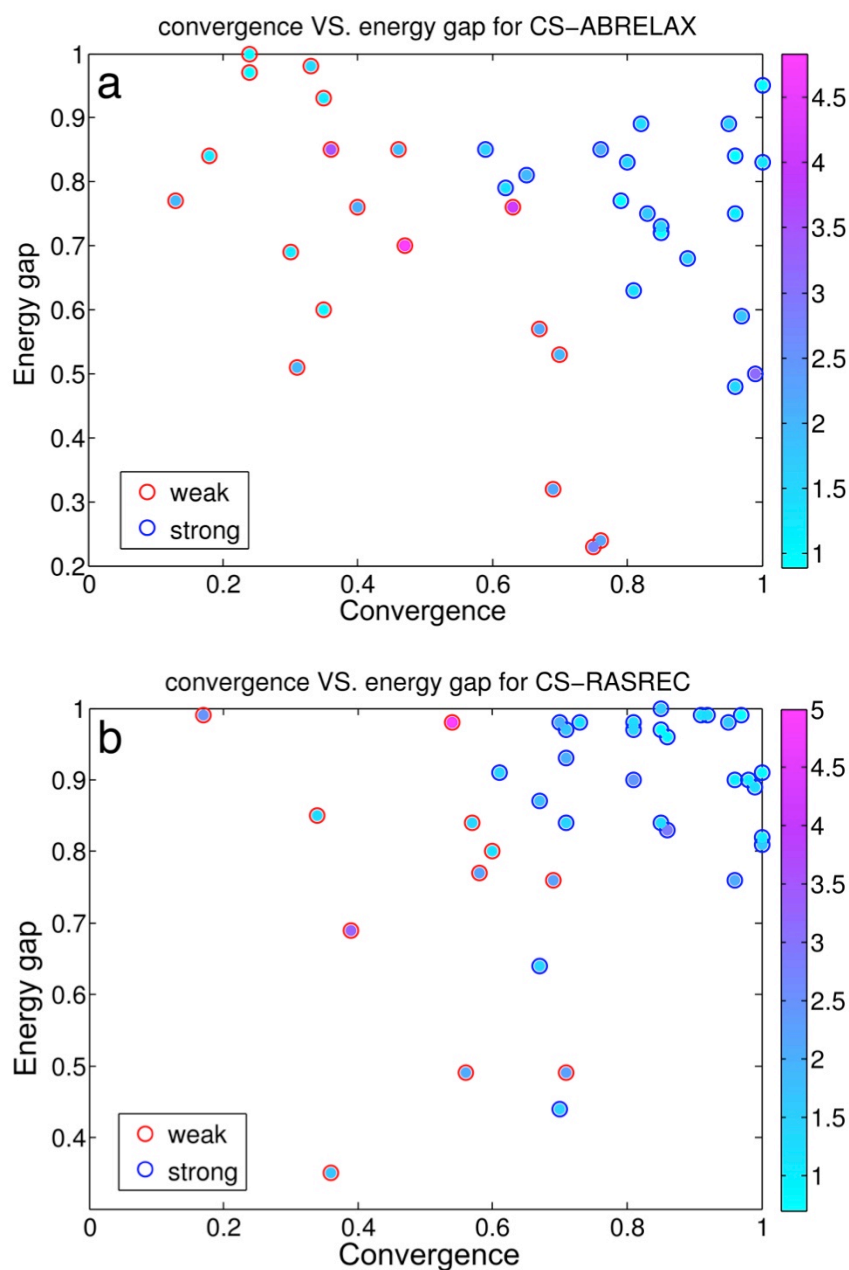


Figure 2.6: Energy gap and convergence of final models for a) CS-ABRELAX and b) CS-RASREC. The convergence (fraction of converged residues; x-axis) and the energy gap per residue after application of the sigmoid function (y-axis). The accuracy of the final models (C_{α} -RMSD of converged residues against reference structure) is depicted by the color inside

the circle (colorbar). The outline of the circle is shown in red or blue, and depicts the given annotation for this run (legend).

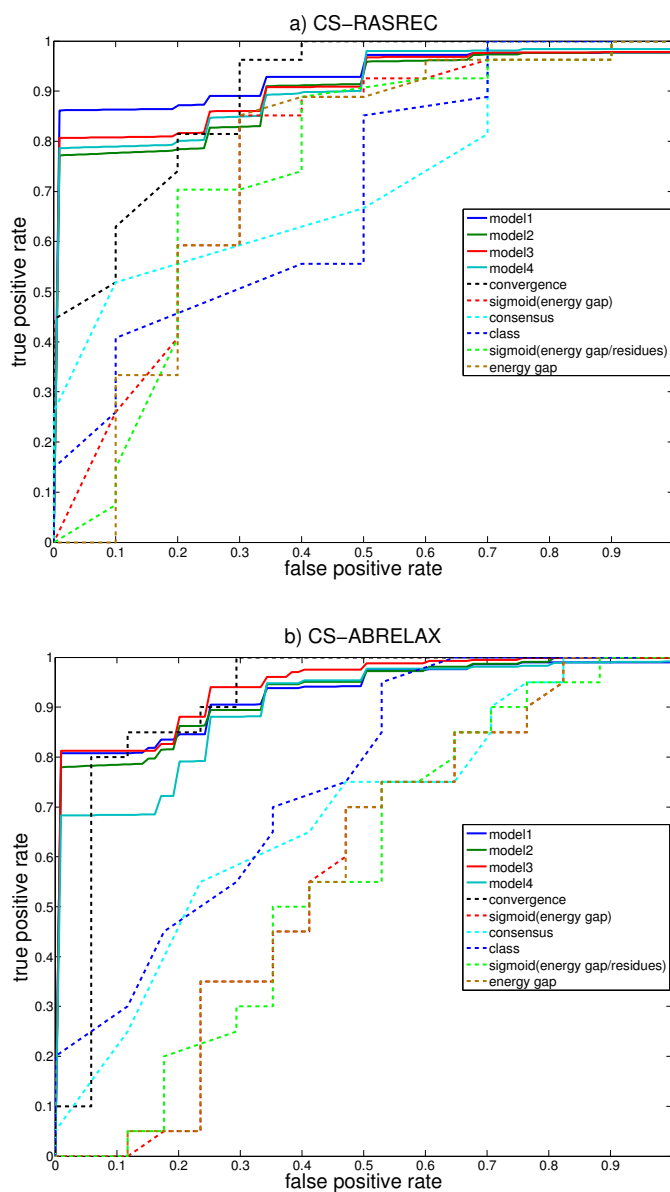


Figure 2.7: Receiver Operator Characteristics (ROC) for *strong/weak* classification of 3D structure predictions of a) CS-RASREC and b) CS-ABRELAX calculations. The ROCs of model 1-4 (solid-lines, legend) were obtained via 100 rounds of cross-validation with random separation into training and test data (75%/25%). The ROC of individual input criteria used for the predictor models 1-4 are shown as dashed lines. Clearly, the linear combination of the input criteria in the predictor models 1-4 does improve the ROC. Model 1 and Model 3 are selected as the predictor for CS-RASREC and CS-ABRELAX, respectively, based on their highest area under the curve (AUC).

CHAPTER 2 IMPROVING 3D STRUCTURE PREDICTION FROM CHEMICAL SHIFT DATA

Indeed, the classification scheme successfully annotates those predictions as *weak* that yield bad accuracy (red in Figure 2.5b+c). 20 of 39 targets (51%) listed in Table 2.4 computed with CS-ABRELAX are considered as *strong* structure calculations. For 18 of these the accuracies range from 0.9Å to 2.0Å, and for the remaining two, accuracies are 3.1Å and 2.1Å for targets #21(2jvf) and #31(2k3d), respectively. From the targets computed with CS-RASREC, 29 of 39 (74%) results are considered *strong*. For 26 of the *strong* targets, accuracies range from 0.7Å to 2.0Å and for the remaining three, targets #4 (2dm2), #12 (2jov) and #34 (2k5c), accuracies are 2.1Å, 2.9Å and 2.4Å, respectively (Table 2.5).

	PDB	Annotation parameters			final class ^d
		convergence ^a	energy gap ^b	cs-class ^c	
1	1ig5	0.67	0.52	0.46	<i>weak</i>
2	1q02	0.46	0.72	0.72	<i>weak</i>
3	2ckx	0.96	0.43	0.86	<i>strong</i>
4	2dm2	0.36	0.93	0.78	<i>weak</i>
5	2htj	0.63	0.76	0.67	<i>weak</i>
6	2hx6	0.13	0.87	0	<i>weak</i>
7	2ike	0.24	0.93	0.83	<i>weak</i>
8	2jmb	0.95	0.92	0.82	<i>strong</i>
9	2jml	0.40	0.78	0.61	<i>weak</i>
10	2jmp	0.65	0.91	0.82	<i>strong</i>
11	2joq	0.89	0.71	0.81	<i>strong</i>
12	2jov	0.24	1.00	0.75	<i>weak</i>
13	2jpn	0.30	0.67	0.78	<i>weak</i>
14	2jq3	0.18	0.86	0.8	<i>weak</i>
15	2jrm	0.83	0.65	0.72	<i>strong</i>
16	2jso	0.80	0.89	0.78	<i>strong</i>
17	2jsx	0.85	0.76	0.8	<i>strong</i>
18	2jt1	0.96	0.83	0.83	<i>strong</i>
19	2jtv	1.00	0.78	0.81	<i>strong</i>
20	2jub	0.47	0.68	0.81	<i>weak</i>
21	2jvf	0.99	0.49	0.83	<i>strong</i>
22	2jvr	0.70	0.49	0.73	<i>weak</i>

CHAPTER 2 IMPROVING 3D STRUCTURE PREDICTION FROM CHEMICAL SHIFT DATA

23	2jvw	0.35	0.96	0.72	<i>weak</i>
24	2jxt	0.81	0.61	0.9	<i>strong</i>
25	2jz5	0.79	0.82	0.84	<i>strong</i>
26	2k0m	0.33	1.00	0.82	<i>weak</i>
27	2k14	0.69	0.24	0.69	<i>weak</i>
28	2k19	0.59	0.94	0.79	<i>strong</i>
29	2k2p	0.97	0.51	0.75	<i>strong</i>
30	2k37	0.31	0.41	0.81	<i>weak</i>
31	2k3d	0.76	0.90	0.83	<i>strong</i>
32	2k4y	0.85	0.75	0.82	<i>strong</i>
33	2k52	0.96	0.74	0.8	<i>strong</i>
34	2k5c	0.76	0.14	0.81	<i>weak</i>
35	2k5n	0.82	0.88	0.77	<i>strong</i>
36	2k5s	0.62	0.77	0.85	<i>strong</i>
37	2osq	0.35	0.55	0.77	<i>weak</i>
38	2ot2	0.75	0.14	0.7	<i>weak</i>
39	2qmt	1.00	0.91	0.77	<i>strong</i>

Table 2.4: Automatic annotation result of CS-ABRELAX predictions

^a The fraction of residues which are converged in 30 lowest-energy structures predicted by CS-ABRELAX with an RMSF cutoff of 2 Å

^b The difference in ROSETTA all-atom energy between the lowest energy decoys and the lowest energies obtained for decoys far away (>4 Å) from the lowest energy decoys

^c The fraction of residues annotated by TALOS+ with 'GOOD'

^d Annotation of prediction reliability

PDB	Annotation parameters			final class ^d	
	convergence ^a	energy gap ^b	cs-consensus ^c		
1	1ig5	0.69	0.76	0.55	<i>weak</i>
2	1q02	0.71	0.49	0.94	<i>weak</i>
3	2ckx	1.00	0.81	0.98	<i>strong</i>
4	2dm2	0.96	0.76	0.97	<i>strong</i>
5	2htj	0.71	0.97	0.90	<i>strong</i>

CHAPTER 2 IMPROVING 3D STRUCTURE PREDICTION FROM CHEMICAL SHIFT DATA

6	2hx6	0.17	0.99	0.00	<i>weak</i>
7	2ike	0.56	0.49	0.94	<i>weak</i>
8	2jmb	0.95	0.98	0.94	<i>strong</i>
9	2jml	0.58	0.77	0.78	<i>weak</i>
10	2jmp	0.67	0.87	0.95	<i>strong</i>
11	2joq	0.91	0.99	0.95	<i>strong</i>
12	2jov	0.86	0.83	0.96	<i>strong</i>
13	2jpn	0.57	0.84	0.88	<i>weak</i>
14	2jq3	0.39	0.69	0.91	<i>weak</i>
15	2jrm	0.85	1.00	0.90	<i>strong</i>
16	2jso	0.81	0.98	0.94	<i>strong</i>
17	2jsx	0.85	0.97	0.91	<i>strong</i>
18	2jt1	0.97	0.99	0.96	<i>strong</i>
19	2jtv	1.00	0.82	0.95	<i>strong</i>
20	2jub	0.54	0.98	0.92	<i>weak</i>
21	2jvf	0.99	0.89	0.95	<i>strong</i>
22	2jvr	0.70	0.98	0.86	<i>strong</i>
23	2jvw	0.36	0.35	0.93	<i>weak</i>
24	2jxt	0.73	0.98	0.98	<i>strong</i>
25	2jz5	0.86	0.96	0.95	<i>strong</i>
26	2k0m	0.67	0.64	0.96	<i>strong</i>
27	2k14	0.71	0.93	0.92	<i>strong</i>
28	2k19	0.85	0.84	0.93	<i>strong</i>
29	2k2p	0.98	0.90	0.89	<i>strong</i>
30	2k37	0.34	0.85	0.95	<i>weak</i>
31	2k3d	0.71	0.84	0.96	<i>strong</i>
32	2k4y	0.92	0.99	0.95	<i>strong</i>
33	2k52	0.96	0.90	0.93	<i>strong</i>
34	2k5c	0.81	0.90	0.93	<i>strong</i>
35	2k5n	0.81	0.97	0.91	<i>strong</i>
36	2k5s	0.70	0.44	0.97	<i>strong</i>
37	2osq	0.60	0.80	0.94	<i>strong</i>

38	2ot2	0.61	0.91	0.92	<i>strong</i>
39	2qmt	1.00	0.91	0.95	<i>strong</i>

Table 2.5: Automatic annotation result of CS-RASREC predictions

^a The fraction of residues which are converged in 30 lowest-energy structures predicted by CS-RASREC with an RMSF cutoff of 2 Å

^b The difference in ROSETTA all-atom energy between the lowest energy decoys and the lowest energies obtained for decoys far away (>4 Å) from the lowest energy decoys

^c The fraction of residues for which TALOS+ finds more than 7 consensus matches in the database

^d Annotation of prediction reliability

For these three targets(#4, #12, and #34), CS-RASREC predicted structures have the same fold as the reference structure, but show better packing with less and smaller solvent inaccessible cavities in the protein core (Figure 2.8) (Sheffler and Baker 2008). Given the clear packing deficiencies in the deposited NMR ensembles, we believe that the 2.1-2.9Å RMSDs do not actually reflect the accuracy of the CS-RASREC structures, and that these targets can be ignored for the overall assessment of CS-RASREC accuracy of *strong* predictions. Representative examples of the remaining *strong* predictions are shown in Figure 2.9.

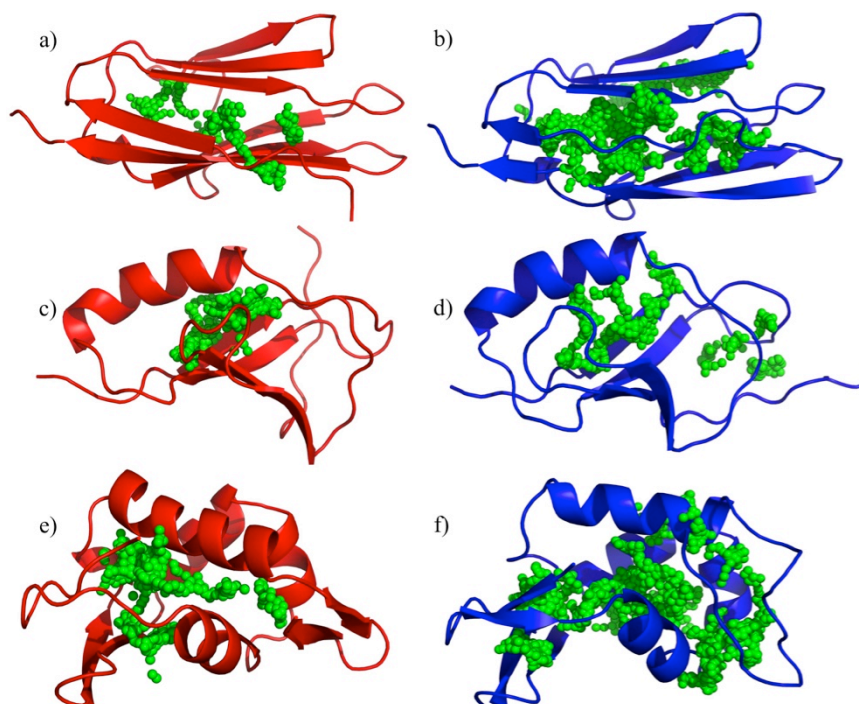


Figure 2.8: Packing analysis of final RASREC (red) models and deposited NMR reference structures (blue) for targets #4(2dm2, Figure a+b), #12(2jov, Figure c+d) and #34(2k5c, Figure e+f), respectively. Cavities detected by RosettaHoles are illustrated as green spheres.

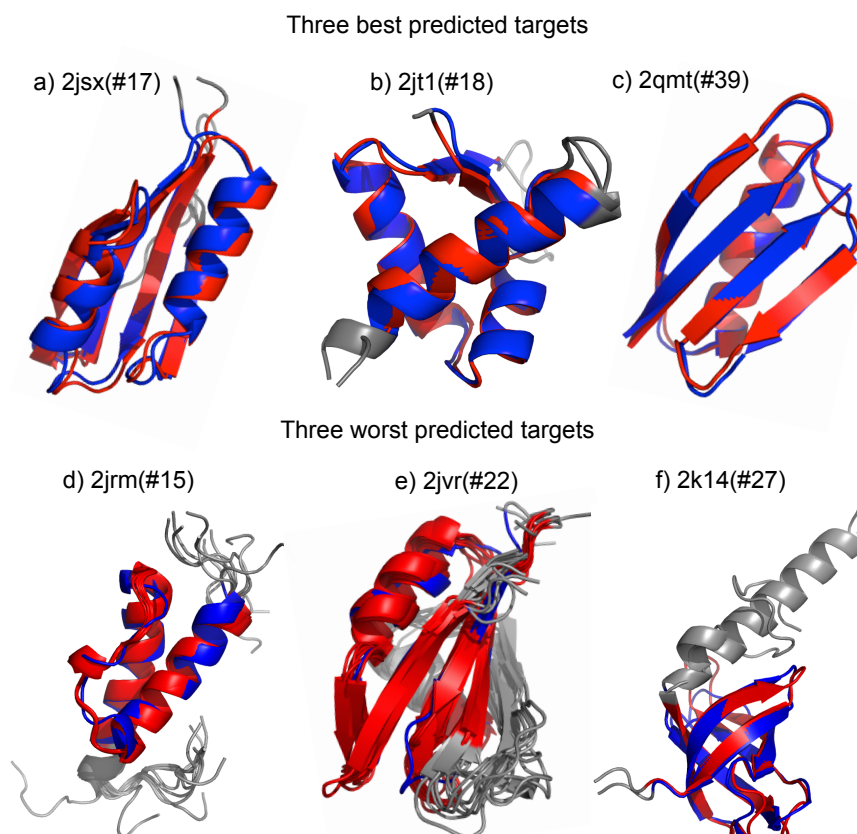


Figure 2.9: Overview of structures obtained with RASREC structure calculations that passed the filter (i.e., annotated as *strong prediction*). Shown are the three best, 2jsx(0.8Å), 2jt1(0.7Å) and 2qmt(0.9Å), respectively, and the three worst, 2jrm (1.6Å), 2jvr(2.0Å) and 2k14(2.0Å), respectively. For each target, the reference structure is in blue and the predicted structures are in red with unconverged regions (see Methods) shown in gray.

2.3.5 The WeNMR CS-ROSETTA web server

The most time consuming part of a typical CS-ROSETTA run consists of a large number (500 to 2500) of independent Monte Carlo calculations to calculate in the order of 10000 to 50000 structures. The WeNMR (www.wenmr.eu) CS-Rosetta web server (Wassenaar et al. 2012) conveniently distributes those calculations over the grid resources made available through the European Grid Infrastructure (EGI, www.egi.eu). The original server has now been extended to allow DP scoring and include the reliability measure

described above. Table 2.6 shows the results of the DP rescoring option (using the CS-ABRELAX setup), using a different benchmark of 6 CASD-NMR targets (Rosato et al. 2009; Rosato et al. 2012). Consistent with previous observations (Raman et al. 2010b; Rosato et al. 2012) the combination of DP rescoring (Huang et al. 2005; Raman et al. 2010b) and CS rescoring outperforms the other rescoring option, including CS rescoring, both in successful predictions and reliability (100%).

Selection ^a	Converged ^b		Not	Reliability ^d
	TP	FP		
raw	2	2	2	50%
cs	3	1	2	75%
dp	5	1	0	83%
dpcs	5	0	1	100%

Table 2.6: Reliability of different structure selection methods

^a Final structure selection methods, raw: rosetta score; cs: cs-rescoring; dp: dp-rescoring; dpcs: cs-rescoring+dp-rescoring.

^b Number of targets for which the average RMSD of selected models is below the threshold of 2.0 Å and are counted as true/false positive.

^c Number of targets for which the average RMSD of selected models is above the threshold of 2.0 Å.

^d Reliability of different structure selection methods

2.4 Discussion

We have considerably improved the scope, convergence and reliability of CS-ROSETTA calculations from chemical shifts only. On a representative benchmark of 39 small proteins in the size range of 50-100 residue size range, we demonstrated that CS-ROSETTA calculations yield successful and accurate 3D structure predictions in 74% of the cases when using the new CS-RASREC method. CS-ABRELAX is still successful in 51% of the cases but generally yields less converged residues per target. Most importantly, we introduced a classification scheme that can be used to detect whether a successful prediction has been made, which increases the reliability to >89% and >80% for CS-RASREC and CS-ABRELAX calculations, respectively. Reliable predictions have accuracies of 2Å and better on the converged residues. This renders the presented CS-ROSETTA structure calculation protocols a reliable tool for rapid and accurate structure determination at atomic resolution.

CHAPTER 2 IMPROVING 3D STRUCTURE PREDICTION FROM CHEMICAL SHIFT DATA

CS-ROSETTA calculations entail a considerably computational effort; a reliable structure prediction requires 10000 or more models to be generated with an overall cost of several thousand CPU-hours. We implemented a webserver that utilizes the WeNMR grid infrastructure to farm out the time-consuming model generation part of CS-ROSETTA calculations. The service is available for the whole scientific community and is free of charge to academic users. It only requires a backbone chemical shift list as input and offers several options to re-evaluate the generated models, including NOE based rescoring with the DP-score (Huang et al. 2005; Raman et al. 2010b).

Currently, the WeNMR grid cannot support CS-RASREC calculations due to the requirement of communication between RASREC processes that is not supported by the grid-infrastructure. However, RASREC calculations are considerably more time-efficient than CS-ABRELAX; for targets in the size range addressed here, they require on the order of 200-1000 CPU hours, which is available on medium sized in-house clusters or at adjunct computer centers of universities. We made considerable advances to simplify running these calculations by providing a Python-based toolbox for pre- and post-processing of CS-ROSETTA related data files and fragment picking. This allows easy setup of CS-ABRELAX and RASREC CS-ROSETTA structure generation runs including integrated support for queuing systems such as SLURM and MOAB. The computational infrastructure has to support jobs that utilize the common Message Passing Interface (MPI) protocol (e.g., openMPI, LAM, MPICH, MPICH2) for inter-process communication. Additionally, a website providing documentation and tutorials (www.csrosetta.org) has been launched in support of the growing user community.

The main advantage of CS-RASREC calculations over the CS-ABRELAX is that a larger fraction of residues converges and that the energy gap becomes more pronounced. This in turn generates a higher chance of a *strong* prediction. On the 39 benchmark cases, the average fraction of converged residues (as shown in Figure 2.5b/c (green bars) is 72% for CS-ABRELAX and 80% for CS-RASREC. From the 9 targets that are classified *strong* in CS-RASREC but *weak* in CS-ABRELAX, 4 have improved classification due to a drastic increase in convergence (from ~30% to >70%), whereas the remaining 5 have similar convergence but improved energy-gaps (Table 2.4+2.5). Finally, the mean accuracy (RMSD) for *strong* predictions is 1.76Å for CS-ABRELAX and 1.44Å for CS-RASREC. Thus, if local computer resources can be obtained it is advisable to run CS-RASREC rather than CS-ABRELAX, if such resources cannot be secured, running just the webservice-based CS-ABRELAX remains a reasonable and valuable alternative. Adaption of the RASREC protocol to a grid or cloud computing platform is in principle possible as only very low-bandwidth

communication is required, but technically involved as the entire communication layer of the protocol has to be adapted.

A program to apply the reported classification scheme into *strong* and *weak* 3D structure predictions is provided with the CS-ROSETTA toolbox versions 1.5 and higher at www.csrosetta.org and is implemented in the CS-ROSETTA web server.

2.5 References

Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. *Proceedings of the National Academy of Sciences* 104:9615–9620

Damm KL, Carlson HA (2006) Gaussian-Weighted RMSD Superposition of Proteins: A Structural Comparison for Flexible Proteins and Predicted Protein Structures. *Biophysical Journal* 90:4558–4573

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6:277–293

Fleishman SJ, Baker D (2012) Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution. *Cell* 149:262–273

Huang YJ, Powers R, Montelione GT (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

Lange OF, Baker D (2011) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins* 80:884–895

Lange OF, Rossi P, Sgourakis NG, Song Y, Lee H-W, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, 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 USA* 109:10873–10878

Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, et al. (2010a) NMR Structure Determination for Larger Proteins Using Backbone-Only Data. *Science* 327:1014–1018

CHAPTER 2 IMPROVING 3D STRUCTURE PREDICTION FROM CHEMICAL SHIFT DATA

Raman S, Huang YJ, Mao B, Rossi P, Aramini JM, Liu G, Montelione GT, Baker D (2010b) Accurate Automated Protein NMR Structure Determination Using Unassigned NOESY Data. *J Am Chem Soc* 132:202–207

Rieping W, Vranken WF (2010) Validation of archived chemical shifts through atomic coordinates. *Proteins* 78:2482–2489

Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein Structure Prediction Using Rosetta. *Methods in enzymology* 383:66–93

Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, Doreleijers JF, 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

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

Schmitz C, Vernon R, Otting G, Baker D, Huber T (2012) Protein Structure Determination from Pseudocontact Shifts Using ROSETTA. *Journal of Molecular Biology* 1–10

Sheffler W, Baker D (2008) RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design and validation. *Protein Science* NA–NA

Shen Y, Bax A (2010) SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR* 48:13–22

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

Shen Y, Vernon R, Baker D, Bax A (2009b) De novo protein structure generation from incomplete chemical shift assignments. *J Biomol NMR* 43:63–78

Shen Y, Zhang Z, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, et al. (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105:4685–4690

Vranken WF, Rieping W (2009) Relationship between chemical shift value and accessible surface area for all amino acid atoms. *BMC Structural Biology* 9:20

Wassenaar TA, van Dijk M, Loureiro-Ferreira N (2012) WeNMR: structural biology on the grid. *Journal of Grid ...*

Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucleic Acids Res* 36:W496–W502

2.6 My contribution to this project

As one of the two first authors in the paper-improving 3D structure prediction from chemical shift data, my main contribution is calculating protein structures with RASREC-Rosetta. I also carried out all the analysis of the final structures, e.g. RMSDs calculation, determination of converged region and packing analysis by RosettaHole. I developed the annotation method to classify the structure calculations and implemented the method into CS-Rosetta toolbox. I prepared nearly all the figures and tables in this paper except Figure 2.1, 2.2 and Table 2.1. I also participated in the paper writing.

Chapter 3 Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta

3.1 Significance Statement

NMR structure determination is next to X-ray crystallography the only available high-resolution method for protein structure determination. NMR spectroscopy is conducted in aqueous solution and thus might be the only route towards high-resolution 3D structures for proteins that cannot be crystallized. However, the analysis of NMR data is very time-consuming and can generally only be conducted by highly trained NMR experts, which require from weeks to several months to obtain accurate and precise structures. Here we provide a novel method to automate the analysis process and show that accurate structures can be obtained. Remarkably, the automatically generated structures are in a majority of the cases more accurate than the structures laboriously generated by NMR experts. The method promises to significantly increase the attractiveness and viability of NMR structure determination.

3.2 Introduction

Structure determination by nuclear magnetic resonance (NMR) spectroscopy is largely driven by distance information gathered through Nuclear Overhauser Effect Spectroscopy (NOESY). To use such data as distance restraints, the NOESY crosspeaks in multidimensional spectra have to be assigned to individual atoms of the biomolecular system. NOESY cross-peak assignment and structure generation steps are usually performed in an integrated, iterative manner. This maximizes the number of conformational restraints, while guaranteeing self-consistency amongst distance restraints(Wüthrich 1986).

Many of the repetitive tasks in NMR structure determination have been successfully automated(Moseley and Montelione 1999; Baran et al. 2004; Güntert 2008; Guerry and Herrmann 2011). Two such crucial tasks in the chain of the data analysis are the assignment of NOE cross-peaks and the determination of accurate structural models. Popular programs that perform these two tasks are ARIA(Linge et al. 2003a), CYANA(Güntert et al. 1997;

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

Güntert 2008), AutoStructure(Huang et al. 2006) and UNIO(Serrano et al. 2012) and have recently been tested with good results in a blind-testing challenge(Rosato et al. 2012). However, a limitation of these programs is that they have to be able to generate a sufficiently accurate model from the initial set of assignments. This usually limits the methods to small proteins with high quality spectra, complete and accurate chemical shift assignments, and well-refined peak lists. When conditions are suboptimal, a calculation either does not converge, or worse, converges to a precise but inaccurate fold(Guerry and Herrmann 2011). Accordingly, these programs are not usually used unsupervised, and must instead be applied in combination with manual assignment and possibly peak list refinement by a skilled NMR expert. Indeed, in our own work on larger proteins, a few manual assignments were required to bootstrap the automated analysis with CYANA(Lange et al. 2012).

Here, we aim to develop a NOE assignment and structure determination algorithm that can – unsupervised – produce results that are both reliable and accurate. This algorithm should take chemical shift assignments and unassigned NOE peak lists as input and produce, without further user interaction, refined models of protein structures in atomic resolution.

To achieve this goal, we combine Rosetta structure prediction with automatic NOE assignment. It has been demonstrated that Rosetta, which searches for the lowest energy conformation of the polypeptide chain using physically realistic force fields, requires only very sparse NMR data to guide its search to accurate structures(Raman et al. 2010; Lange et al. 2012). The question we ask here is whether the very noisy automatically assigned NOE restraints might be able to provide sufficient guidance for Rosetta to yield accurate initial models. These models would then allow iterative refinement of NOE assignments until accurate high-quality structures and self-consistent assignments can be generated. Iteration of automatic NOE assignment with structural modeling is, however, also the basis of established algorithms. Thus, the crucial question to be explored in this study is not whether iteration between modeling and assignment is a successful strategy, but rather if a significant benefit is gained by using the improved, but computationally more demanding, ROSETTA structural modeling, and if we can solve the engineering challenge to render the ROSETTA structure calculation sufficiently robust against the very noisy automatically assigned NOE restraints of the initial assignment stage. In cases where established programs cannot find converged initial models, and thus fail, the new approach might converge and thus applicability is broadened to include more challenging cases. Additionally, the more accurate modeling provided by the ROSETTA energy function might render the method more robust against erroneous input data and yield more accurate final 3D models.

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

To couple NOE assignment with ROSETTA, we build on the previously developed iterative structural modeling algorithm, RASREC, and extend it to become an algorithm for automatic NOE assignment. This entails the implementation of a new ROSETTTA module for automatic NOE assignment as well as the development of a robust protocol to couple the iterative search for the near-native protein structures in RASREC with iterative NOE assignment. The assignment module employs among other techniques, network anchoring(Herrmann et al. 2002), ambiguous restraints(Nilges et al. 1997), covalent structure compliance(Herrmann et al. 2002; Huang et al. 2005), structure dependent and independent peak calibration, and restraint combination(Herrmann et al. 2002). In our final protocol, the calculation consists of multiple iterations of structural sampling guided by automatically assigned NOE restraints. In early iterations, cross-peak assignments compatible with preliminary models are reinforced, but incompatible assignments are not removed. In later iterations, incompatible cross-peak assignments are removed from the restraint list. Throughout the whole process, however, a pool of best fitting structures is maintained that is ranked by the initial NOE assignments. This is a major difference to existing programs and helps us to prevent convergence on inaccurate but self-consistent solutions. Implementation details of the new method will be described elsewhere(Lange).

To investigate the performance of the new methodology, we carried out a benchmark on 50 NOE data sets obtained from 41 protein samples of 63-370 residues length. To test the impact of *difficult* inputs on the performance of AutoNOE-Rosetta, we have included unrefined and automatically picked peak lists, as well as sparse data sets obtained from perdeuterated ILV-methyl labelled protein samples. To avoid unwittingly cherry-picking targets that work especially well for our method, we chose three pre-existing benchmark sets and used *all* monomeric proteins from each(Mao et al. 2011; Rosato et al. 2012; Lange et al. 2012).

In the following we report on the results of the benchmark. First, we will contrast the performance of AutoNOE-Rosetta with CYANA. Subsequently, we compare the accuracy of the unsupervised method with the state of the art of expert guided NMR structure determination as reflected in PDB-deposited NMR models. This is followed by an analysis of structure validation metrics and NOE completeness scores. Finally, we stress test the method with non-ideal input data, such as raw or unrefined peak lists or incomplete and erroneous chemical shift assignments.

3.3 Results

We have defined a single set of parameters that is used to run all targets, including data preparation (e.g., automatic trimming of flexible tails), structure calculation and final model selection. Thus, results in similar quality as reported here should be achievable from application of the method to as yet unknown targets. We also provide a suite of scripts that allow the user to run the software in this unsupervised fashion. The entire benchmark set and the final models can be obtained from our website (www.csrosetta.org/benchmarks) and our results can be scrutinized by interested readers using our software and accompanying tool-chain.

The benchmark comprises 50 NOE data sets derived from 40 different proteins ranging in size from 5.5 kDa to 40 kDa. Input data are the sequence, chemical shift assignments and NOE peak lists (Methods, *Appendix Table S1*). In 20 cases, RDC data of the N-H bond vectors in one or more alignment medium was also included (*A Table S1*).

Multiple calculations are carried out with different weighting of the NOE data against the Rosetta Energy. One is selected from these based on a combination of final Rosetta Energy and the intrinsic precision of the resulting models (Methods). Finally, to be accepted as a successful solution, the structures must fulfill two criteria: convergence and intrinsic NOE consistency (*Appendix Methods Section A.3.2.2*). AutoNOE-Rosetta was run successfully on 42 of 50 data sets, comprising 35 different proteins. Final models are shown in *Appendix Figure S1* for all targets, and their accuracy is reported in *Appendix Table S2* as C_{α} -RMSD with respect to the reference structure. A number of targets have only been used after the AutoNOE-algorithm was finalized, including all parameters, and the run selection protocol. These targets are DrR147D, MrR110B, OR8C, PfR13A, PsR293, SR384, SgR42, VpR247, and HmR11 and display similar performance as the other targets (*Appendix Table S2*).

To provide a reference for the performance of AutoNOE-Rosetta, we chose to run the popular program CYANA 3.0, which obtained the most accurate models in a recent community-wide blind structure determination challenge (CASD)(Rosato et al. 2012). In analogy to AutoNOE-Rosetta we have defined an acceptance rule for CYANA. Based on suggestions of CYANA's creator, Peter Güntert, we use a combination of convergence and CYANA's target function (*Appendix Methods Section A.3.2.1, Figure S2*).

CYANA was *successful* for 31 of 50 data sets according to its acceptance rule (Methods). Thus, a significant improvement in both accuracy and radius of convergence for

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

AutoNOE-Rosetta is observed with respect to CYANA (Figure 3.1). All structures that failed the automatic acceptance criteria in AutoNOE-Rosetta also failed in CYANA, but eleven of the failing targets in CYANA were acceptable according to the criteria in AutoNOE-Rosetta, and yielded accurate structures below 2.5Å RMSD (Figure 3.1). Furthermore, 10 of 17 inaccurate CYANA-structures (RMSDs >2.5Å) were determined accurately by AutoNOE-Rosetta (RMSDs <2.5Å). Numerical values of the C_{α} -RMSD against the reference structures for CYANA and AutoNOE-Rosetta can be found in (*Appendix Table S2*).

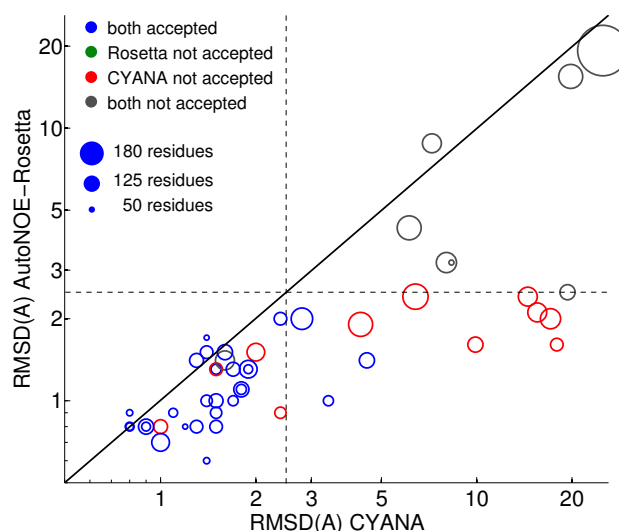


Figure 3.1: Comparison of AutoNOE-Rosetta with CYANA. Shown are the median C_{α} -RMSDs of final models with respect to their reference structure on logarithmic scale. The diagonal line indicates points of equal performance, points above the line correspond to targets for which CYANA yields lower RMSDs, and points below the line correspond to targets for which AutoNOE-Rosetta yields lower RMSDs. The dashed lines mark 2.5Å RMSD. **The size** of the proteins is proportional to the area of the symbol as indicated by the legend. **The color** indicates whether for CYANA, AutoNOE-Rosetta or for both programs the final models are considered as success based on convergence and NOE consistency (*Appendix Method Section A.3.2*). RMSDs are capped at a maximum of 25 Å. Assignment statistics, convergence and accuracy of final models can be found in *Appendix Table S7* and *Appendix Table S8* for AutoNOE-Rosetta and CYANA models, respectively. Comparing heavy-atom RMSDs instead of C_{α} -RMSDs yields a similar picture (*Appendix Figure S6*).

The state-of-the-art in high-resolution NMR structure determination typically involves not just a single CYANA run, but performing several rounds of CYANA-based NOE assignment and refinement of the input peak lists (or even manual assignments, going through peak-by-peak), followed by simulated annealing in XPLOR or CNS (considered to

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

have a better force field than CYANA), and finally a high-resolution refinement in explicit water(Linge et al. 2003b), where RDCs are used if present. To directly compare AutoNOE-Rosetta to this more complex structure determination protocol, we included 20 protein targets in our benchmark for which both a conventionally determined solution NMR structure and an X-ray crystal structure are available. We further assume that the state-of-the-art in NMR structure calculation is well reflected in these 20 PDB-deposited NMR solution structures. Indeed, all these structures were deposited in the last decade, the program CNS is listed in all PDB headers (except 1xpv), and whenever the respective remark section is provided in the PDB header (12 of 20 cases), water refinement is mentioned explicitly.

In this study, we assume that the X-ray structure is an accurate representation of the dominant solution structure; accordingly, the RMSD of atomic coordinates between NMR and X-ray structure provides a measure for the accuracy of the NMR structure. This view is supported by the NMR data (*Appendix Table S3*). Based on this criterion, AutoNOE-Rosetta significantly outperforms conventional supervised NMR structure determination (Figure 3.2a and *Appendix Table S4*). For 10 of 21 targets, accuracy is significantly improved, and only for 2 of 21 it is decreased (CcR55, partially converged; ER690 unconverged). Moreover, if we restrict the analysis to the 19 converged targets, accuracy never deteriorates more than 33%, whereas it improves for 7 targets significantly beyond 33%. This is in stark contrast to the performance of established automatic assignment programs. Only 3 of the smallest targets of the benchmark set (<80 residues) yield sufficiently accurate results in CYANA to compete with PDB deposited NMR structures. For the other 18 of 21 targets, the structures obtained unsupervised with CYANA are >25% worse in accuracy than PDB-deposited NMR structures (Figure 3.2b). Of these 18 with deteriorated accuracy, 13 yield a tight structural bundle and 10 are acceptable according to the success criteria introduced above (*Appendix Tables S2+S8*).

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

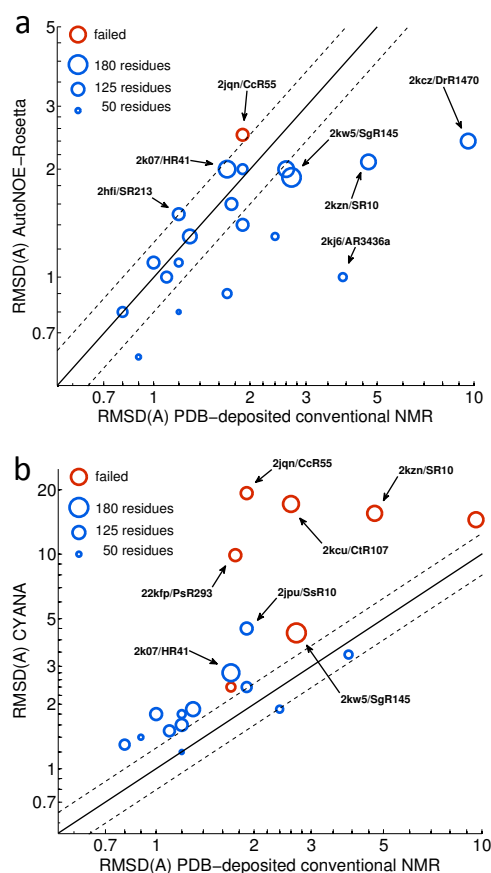


Figure 3.2: Comparison of unsupervised automatic NOE models with expert-analyzed NMR solutions structures. The C_{α} -RMSDs of PDB deposited NMR models is plotted against final models obtained with **(a)** AutoNOE-Rosetta and **(b)** CYANA. For AR3436a no X-ray structure is available as reference, but a new manually refined NMR solution structure, which supersedes 2kj6 (Figure 3.4 and Results). **The solid diagonal line** indicates points of equal performance, points above the line correspond to targets where PDB-deposited NMR structures have higher accuracy, and points below the line correspond to targets with higher accuracy of the AutoNOE-Rosetta models. Dashed lines mark +/- 25% accuracy. **The size** of the proteins is proportional to the area of the symbol as indicated by the legend. AutoNOE-Rosetta or CYANA runs that are not converged (<90% of residues converged) are shown in red. Comparing heavy-atom RMSDs instead of C_{α} -RMSDs yields a similar picture (*Appendix Figure S7*).

In addition to a high accuracy, we generally would like to obtain 3D models of proteins with a high structural quality. This quality is generally assessed by structural validation packages through various metrics, such as packing quality, ramachandran consistency, and Janin-plots. NMR solution structures based on NOE distance restraints are

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

prone to show deficits(Doreleijers et al. 2012b), whereas un-restrained CS-Rosetta models were previously reported to show high structural quality but significantly lower accuracy than NOE-driven structure calculations(Rosato et al. 2012). We were curious to see whether AutoNOE-Rosetta preserves the high structural quality, despite being subjected to a large number of automatically assigned NOE restraints and yields more accurate structures than CS-Rosetta. To assess the structural quality of AutoNOE-Rosetta models we used the online validation server iCING(Doreleijers et al. 2012a), which performs WhatIF(Vriend 1990), PROCHECK(Laskowski et al. 1996) and its own structural analysis.

The iCING-ROG score summarizes and integrates different validation measures into a single score and annotates individual residues as *green*, *orange* and *red* to convey an increasing level of alertness for unphysical local structure(Doreleijers et al. 2012a). AutoNOE-Rosetta models produce generally less red and orange residues than PDB NMR-models or CYANA models(Figure 3.3a-c). WhatIF compares local structure of the protein against common structural knowledge derived from high-resolution X-ray structures(Vriend 1990). Figure 3.3d-f shows the WhatIF structure Z-scores on *Ramachandran plot appearance*, *backbone quality*, *1st generation packing quality*, and *chi-1/chi-2 rotamer quality*. AutoNOE-Rosetta models generally are of higher quality than PDB-NMR models or CYANA models.

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

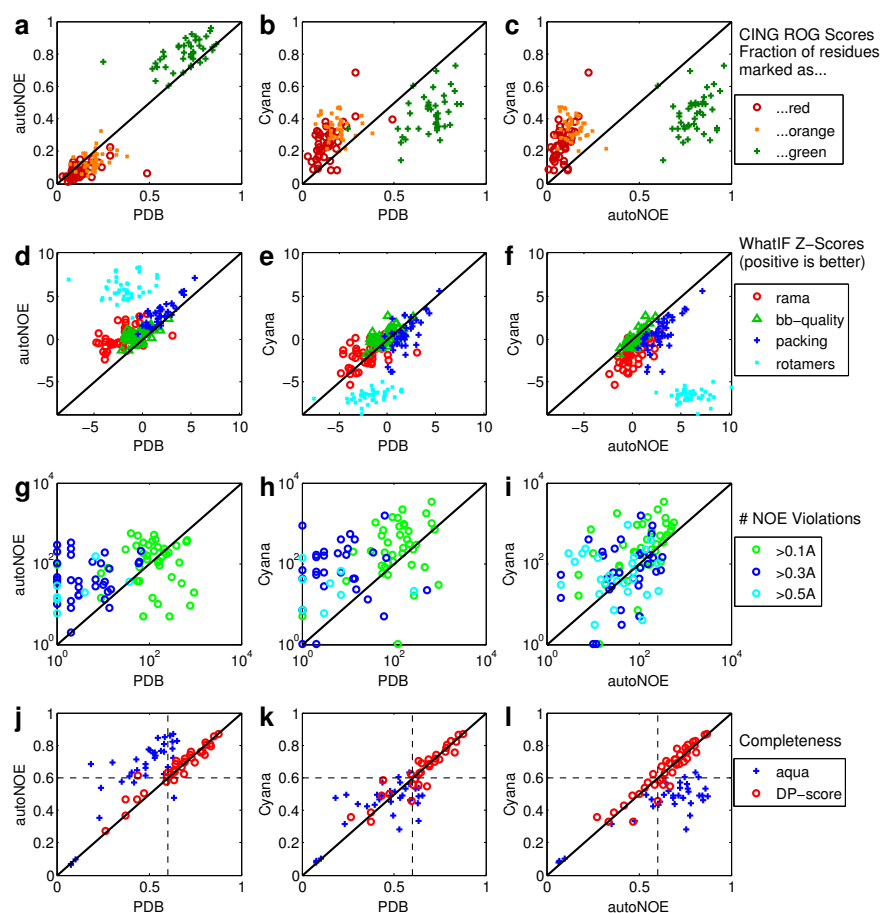


Figure 3.3: Validation metrics for AutoNOE-Rosetta, CYANA and PDB-deposited NMR models. Metrics computed for AutoNOE-Rosetta and CYANA-models are compared to metrics computed on PDB-models, in panel-columns 1 and 2, respectively. Metrics between AutoNOE and CYANA are compared directly in panel-column 3. **(a-c)** Fraction of residues annotated as *red*, *orange* and *green* by the iCING server's ROG score(legend). Less *red* and *orange* and more *green* residues is better. **(d-f)** WhatIF Z-scores for Ramachandran plot appearance, backbone-quality, packing and chi-1/chi-2 rotamer normality (legend). Higher Z-scores are better. **(g-i)** The number of NOE restraints violated by structural models. Structural models of CYANA and AutoNOE are analyzed together with the restraints produced by the respective algorithms. PDB-deposited models are analyzed with respect to the NOE restraints uploaded with the structures. **(j-l)** Completeness scores computed with AQUA(Doreleijers et al. 1999) and AutoStruct-DP(Huang et al. 2005). Higher numbers are better.

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

Another popular criterion for judging NMR structure quality is a low count of restraint violations by the final models. Figure 3.3g-i shows how often the final models violate the NOE-derived restraints by $>0.1 \text{ \AA}$, $>0.3 \text{ \AA}$ and $>0.5 \text{ \AA}$. Generally, NMR restraint-sets deposited with their corresponding PDB structures have less violations above $>0.3 \text{ \AA}$ or $>0.5 \text{ \AA}$ than those obtained with CYANA or AutoNOE-Rosetta, but CYANA and AutoNOE-Rosetta yield similar results. We found that, for AutoNOE-Rosetta ensembles, many of the violations occurred at side-chains that adopted multiple conformations. In these cases, each conformation would actually be consistent with a subset of the violated NOE restraints involving this side-chain, and it would be plausible that dynamic averaging causes the assigned NOE cross-peaks. Since it is well possible that dynamic averaging might be the reason for some of the observed violations, as well as the fact that programs could trivially remove any violated restraint from the restraint-list, it is questionable whether the count of restraint violations is actually a valuable criterion for NMR structure validation. Indeed, we see no particular correlation between this measure and accuracy of the final models (C_{α} -RMSD) regardless whether they were downloaded from the PDB or generated with CYANA or AutoNOE-Rosetta (*Appendix Figure S3*).

Since the AutoNOE-Rosetta structures fit more accurately to X-ray structural models, a possible concern might be that Rosetta modeling is biased towards X-ray crystallographic artifacts rather than solution state structure. To verify that this is not the case we show that AutoNOE-Rosetta models yield a better or equivalent interpretation of the NMR data in comparison to conventional NMR solution structures, as quantified by the AQUA completeness (Doreleijers et al. 1999) and the AutoStruct DP score (Huang et al. 2005). AQUA reads the models and restraint list and checks how many of the proton-proton contacts in the model are actually observed as assigned NOEs. The more modern DP score uses chemical shift assignments and unassigned peak lists as input, and is thus independent of the specific restraint list. AQUA's completeness score is systematically better for AutoNOE-Rosetta than for PDB-NMR or CYANA models (Figure 3.3j-l, black circles). For most targets the DP scores are comparable between the different methods (Figure 3.3j-l, blue crosses). However, for some PDB-NMR structures with low DP-scores (<0.6) AutoNOE-Rosetta was able to yield significant improvements. Overall these quality measures show that AutoNOE-Rosetta models yield as good or better an interpretation of the NMR data as the PDB-deposited NMR models.

Next, we were interested how AutoNOE-Rosetta behaves when provided with problematic data. Accordingly, we tested both automatic (raw) and refined peak lists for 8 targets from round II of the blind, community-wide NMR structure determination challenge

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

(CASD)(Rosato et al. 2009). In addition to the 8 *raw* data sets, we use 7 *unrefined* data sets from previous work(Lange et al. 2012) and one from CASD round I. For these *unrefined* data sets, peaks have been picked manually and chemical shift assignments have been validated, but the peak lists and chemical shift assignments have not yet undergone iterative refinement using structural models. Of the *unrefined* data sets, 6 stem from ILV-methyl labeled perdeuterated protein samples. Restraints obtained from such ILV-samples are inherently sparse, rendering structure calculation more challenging due to a lower restraint density. Moreover, the sparser NOE networks render the automatic validation of NOE cross-peak assignments via network anchoring less effective.

The availability of 9 targets with both *raw/unrefined* and *refined* data allows us to investigate the robustness of AutoNOE-Rosetta. AutoNOE-Rosetta turns out to be remarkably robust; for 7 of the 9 *raw/unrefined* peak lists differences in accuracy are insignificant($<0.3\text{\AA}$). In only two cases, StT322 and HR5460, was the accuracy significantly decreased. The automatic acceptance criteria successfully identified both these raw data sets as having produced untrustworthy results. Interestingly, AutoNOE-Rosetta tends to select a lower weight for the NOE-based pseudo-energy contribution for *raw* peak lists compared to *refined* peak lists (Table 3.1), which is consistent with the presumed lower quality of the data.

Target	Reference	size	Residue ranges				weight ratio ¹	C _α -RMSD (Å) to reference structure			
			of NMR sample	used in Rosetta	RMSD analysis	Raw peak list		Refined peak list			
						CYANA		AutoNOE-Rosetta	CYANA	AutoNOE-Rosetta	
StT322	2loj	38	1-63	26-63	26-63	0.04	8.3	3.2	1.4	1.7	
HR6470	2l9r	48	1-69	11-58	11-58	1.00	0.8	0.9	0.8	0.8	
OR135	2ln3	69	1-79	5-73	5-73	0.50	0.9	0.8	1.1	0.9	
AR3436a	tbd ³	80	1-97	14-93	14-93	0.40	3.4	1.0	1.7	1.0	
HR6430	2la6	89	1-99	11-99	11-99	1.00	1.4	1.0	1.5	0.9	
HR2876	2ltm	95	1-107	13-107	13-107	0.04	not converged	1.6	1.4	1.5	
YR313	2ltl	102	1-119	18-119	18-40, 46-115	0.10	1.6	1.4	1.7	1.3	
OR36	2lci	128	1-134	1-128	1-128	0.50	2.0	1.5	n/a ²	1.2	
HR5460	2lah	150	1-160	11-160	19-160	0.02	not converged	3.2	n/a ²	1.8	

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

Table 3.1: Impact of raw peak lists.

Footnotes:

- 1) Ratio of NOE-pseudo energy weights selected automatically by AutoNOE-Rosetta for *raw* vs. *refined* data sets.
- 2) a segmentation fault in CYANA 3.0 prohibited us from finishing the structure calculation
- 3) The reference structure for AR3436a is a new NMR structure (PDB accession code: TBD) that results from a (manual) re-evaluation of the original NMR spectra.

Another type of challenging input is given by the 8 *unrefined* ILV data sets. AutoNOE-Rosetta succeeded on four of these data sets and yielded a partially converged structure for another (HmR11). CYANA, however, did not succeed on any of these 8 data sets. Is the deciding factor, which makes these data sets so challenging, the sparseness of the ILV data, the quality of the data sets (*unrefined* vs. *refined*), or the increased molecular weight (ILV data sets have a molecular weight between 15-21 kDa)? One can mostly exclude the increased molecular weight, as the driving factor for these failures, since both AutoNOE and CYANA were significantly more successful on the *refined* data sets in the same size range. Furthermore, we showed above that the influence of data quality (*raw* vs. *refined*) on AutoNOE-Rosetta is low for small, double-labeled data sets. Thus, the lower success rate is likely a result of the sparseness of the ILV data.

To run AutoNOE-Rosetta or CYANA unsupervised, it is important to have clear criteria to flag problematic runs. This filter mechanism has to catch most, if not all, problematic results. In other words, the filter should produce little or no false positives. Some false negatives, on the other hand, are not as worrisome, as human experts can inspect a few such calculations. Here, we have introduced clear definitions for such a filter rule based on convergence of structures and NOE self-consistency for CYANA and AutoNOE-Rosetta (*Appendix Methods A.3.2*). Of the eight declined calculations performed with AutoNOE-Rosetta, four failed both criteria, and four (two each) failed only one of the criteria. The data sets that only failed the consistency criterion, are YR313(*raw*) and StT322(*raw*). While YR313 yielded accurate structures (1.4Å) in AutoNOE-Rosetta, StT322 did not (C_{α} -RMSD 3.2Å; *Appendix Figure S4*). For CcR55, HR5460(*raw*), and HmR11(*unrefined*) only 88%, 79% and 75% of residues converged, respectively, failing the criterion of 90% convergence by only a small margin. In these cases, the converged part of the structure is reasonably accurate (C_{α} -RMSDs of 1.3Å, 2.0Å, and 3.1Å, respectively) and would provide an advanced starting point for further iterative and structure based refinement of the data set (*Appendix*

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

Figure S4). Hence, as intended, the filter has been successful in producing no false positives and only very few false negatives.

The data presented here shows that AutoNOE-Rosetta yields accurate results even when the peak lists are not well refined. In the following we discuss a fortuitous discovery that demonstrates that AutoNOE-Rosetta is not only robust against problematic peak lists, but also shows remarkable accuracy in the face of incomplete or erroneous side-chain chemical shift assignments. During our work on the here-presented benchmark we were initially puzzled by one outlier. For this outlier, AR3436a, AutoNOE-Rosetta yielded structures that were 3.8Å away from the PDB-deposited NMR solution structure (2kj6). The AR3436a data set stems from the CASD set, and was originally posed as a blind challenge to the community. The results of this competition seemed fairly standard except the CS-Rosetta models were identified as an outlier (Rosato et al. 2012): all NOE driven programs produced structures close to the PDB-deposited structure (1.4-2.2 Å) and with acceptable, albeit slightly borderline, validation scores. However, a closer inspection of the NMR solution models (2kj6) reveals that the main helix is at an angle causing the hydrophobic core of the protein to be exposed (Figure 3.4a+c). In the AutoNOE-Rosetta models, in contrast, the helix is well packed against the core (Figure 3.4b+d), which is more consistent with our understanding of the physical chemistry of hydrophobic protein cores. Moreover, the CS-Rosetta based submissions to the blind structure determination challenge also packed the helix against the core (with RMSD >4Å to the reference NMR structure), but did not converge to a high-precision structural bundle.

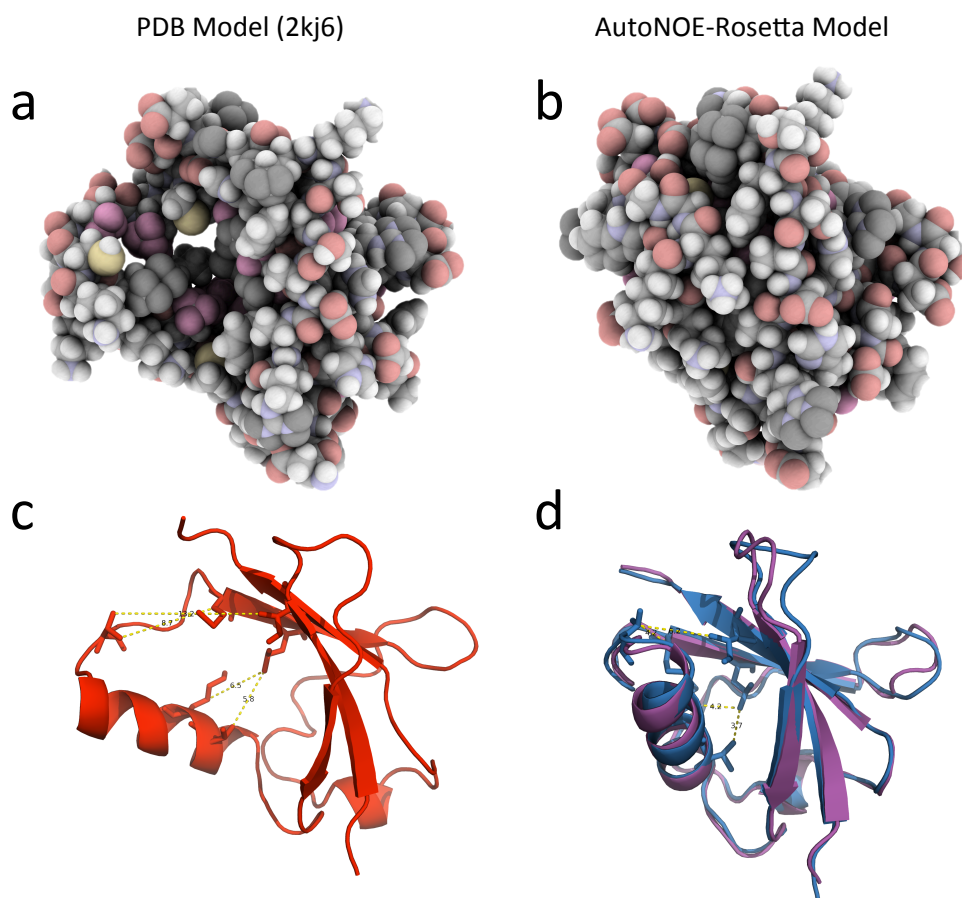


Figure 3.4: Structure determination of AR3436A from incomplete and erroneous input data. Shown are two models of AR3436A in space-fill (**a,b**) and cartoon visualization (**c,d**) to highlight the differences in packing of the hydrophobic core between the PDB-deposited NMR solution structure (**a,c**) and the structure obtained with AutoNOE-Rosetta from the same input data (**b,d**). Due to the incomplete and erroneous chemical shift assignments AutoNOE-Rosetta can only assign a few NOE-crosspeaks (yellow lines) that support the packing of the helix, nevertheless, these are sufficient to yield well packed structures. (**d**). The PDB-deposited models violate these NOE crosspeaks, demonstrating that the respective assignments were discarded because they didn't fit initial models.

These observations prompted us to investigate whether the better-packed structure obtained with AutoNOE-Rosetta might actually be better supported by the raw NMR data as well. Indeed, a careful analysis of the raw input data conducted together with members of the laboratory that authored the original data set revealed a number of problems. Although the backbone assignment was nearly complete and correct, the side-chain chemical shifts were incomplete and had miss-assignments. Additionally, the NOESY data were under-picked as indicated by the unbalanced Recall-Precision scores of the PSVS analysis, such

that many potentially well resolved peaks were not contained in the original peak list. These issues hindered the structure calculations of NOE-driven programs, but had no influence on the CS-Rosetta calculations. After correcting these issues with the input data, the structures obtained with conventional methods matched with the AutoNOE-Rosetta models obtained with either the original data (1.0Å) or the new data (1.0Å). This shows that AutoNOE-Rosetta is not only reliable with unrefined (raw) peak lists but also with raw (i.e., incomplete and erroneous) sidechain chemical shift assignments. We are now in the process of systematically investigating the influence of such raw chemical shift assignments on automatic NOE assignment methods and our preliminary results support the anecdotal case reported here. The advantage of this robustness of AutoNOE-Rosetta for the full NMR pipeline is obvious. Assignment of side-chain chemical shifts is often a major bottleneck to progress in an NMR structure determination project. Automatic methods, such as FLYA(Schmidt and Güntert 2012), might take the burden of manual assignment, but cannot be relied on to always yield the highest quality of resonance assignments. However, paired with AutoNOE-Rosetta, which is more fault-tolerant than other methods, an accurate structure might still be generated either as final result, or as a starting point for further refinement of the chemical shift assignments.

3.4 Discussion

We developed a new method for automatic NOE assignment and NMR structure determination, which we tested on a benchmark of 50 data sets including 20 for which X-ray crystallographic reference structures were available. A final convergence and NOE consistency filter accurately discriminates between successful and failed runs, and all 42 runs that pass this filter yield an accuracy better than 2.5Å C_{α} -RMSD. Thus, we successfully combined the most important traits of CS-Rosetta with those of NOE-driven structure determination. The new algorithm is robust against missing or erroneous data as CS-Rosetta, but in the end exploits the full NOESY data to achieve the optimal precision and accuracy in final structures. In particular the lack of precision is problematic for CS-Rosetta, even if NOE-based filtering is applied (CS-DP-Rosetta(Raman et al. 2010)), as shown by the community wide assessment of structure determination (CASD)(Rosato et al. 2012).

The usefulness of an automatic NOESY assignment algorithms hinges on its ability to handle a wide variety of data. In fact, the quality of NOESY peak lists can vary dramatically as a function of the quality of the raw data, the method of picking peaks, and the level of peak list refinement. With 50 data sets from 41 different proteins, we are confident that our benchmark covers a realistic range of NMR data quality. To enhance the

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

variety in the benchmark, we also included data sets at different stages of refinement (termed *raw*, *unrefined*, and *refined*). And in spite of this wide variety of input data quality AutoNOE-Rosetta yields accurate results with striking consistency, which demonstrates a remarkable robustness of the method against challenging input data. Thus, AutoNOE-Rosetta is a significant advance in *fully automatic* analysis of NMR data.

We were able to compare AutoNOE-Rosetta ensembles with PDB-deposited NMR ensembles which reflect the state-of-the-art in NMR structure determination including final refinement in explicit water. Remarkably, the AutoNOE-Rosetta results are either very close in accuracy (within 25%) or significantly better (Figure 3.2) than the PDB-deposited models. The most significant improvements were from 9.6Å to 2.3Å for the double-labelled sample, DrR1470, and from 4.7Å to 2.1Å for the triple-labelled, ILV-protonated sample, SR10 for which our calculations started from an *unrefined* data set.

AutoNOE-Rosetta ensembles' high accuracy—both relative and absolute—is especially remarkable considering that we are comparing an automated, unsupervised method with expert driven iterative and structure based refinement, as it is reflected in PDB deposited structures. For experts in NMR data analysis the method will provide better starting points for refining challenging data sets. For non-experts it will allow a safe and straightforward application of NMR structure determination to routine cases. Thus, we are confident that our method provides a significant progress towards unsupervised automatic NMR structure determination, which is likely to broaden the applicability of NMR for structure determination in academic and non-academic labs.

3.5 Methods

3.5.1 Benchmark

The 50 data sets comprising target sequence, assigned chemical shifts, and unassigned peak lists were obtained from three published sources (*Appendix* Table S1): 1) all data sets available by December 2012 at the community wide assessment of NMR structure determination (CASD) (Rosato et al. 2009; Rosato et al. 2012) (currently hosted at <http://www.wenmr.eu/wenmr/casd-nmr-data-sets>), 2) all monomer data sets from a recent molecular replacement (MR) benchmark (Mao et al. 2011) (http://psvs-1_4-dev.nesg.org/MR/dataset.html) (*Appendix* Figure S5), 3) all targets from our previous work (Lange et al. 2012).

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

Peak lists from the first prediction period of CASD (CASDI) are refined. For targets from the second prediction period of CASD (CASDII), both, *refined* and *raw* (automatically picked) peak lists are available. For MR targets, the status of the peak lists is unknown but assumed *refined*, and for ILV-targets the peak lists and chemical shift files are *unrefined*, that is chemical shift assignments have been verified and peaks have been picked by a human expert(Lange et al. 2012), but the data sets have not undergone iterative refinement using structural models.

To analyze the accuracy of final structures, we computed the C_{α} -RMSD on all residues that are structured in the reference. Tails that were not well defined (flexible) in the reference structure are excluded from RMSD computation as specified in Table S1. For 11 reference structures, also internal loop-regions were not well defined and had to be excluded from RMSD calculations. Detailed justifications for these exclusions are given in Table S9. For a given method, AutoNOE-Rosetta or CYANA, the ten final models are superimposed with the reference structures to compute C_{α} -RMSDs and heavy-atom RMSDs.

3.5.2 AutoNOE-Rosetta

AutoNOE-Rosetta structure calculations were run with parameters as detailed here(Lange). Fragments were picked by the Rosetta3 fragment picker(Vernon et al. 2013) using the provided chemical shift data. Homologous proteins using an e-value cutoff of 0.05 (sequence identity > 20 %) were excluded from fragment picking. Tolerances for NOESY cross-peak assignment were set for all targets to 0.3, 0.3, 0.03 and 0.04 for ^{13}C , ^{15}N , direct ^1H , and indirect ^1H dimension, respectively. Residual Dipolar Coupling data were used where available (*Appendix* Table S1).

For data sets with *unrefined* or *refined* peak lists, NOE-restraint strengths of 5, 10, 25 and 50, respectively are chosen, and for targets with *raw* peak lists restraint strengths of 1, 2, 5, 10, 25, and 50. For each restraint weight 3 independent runs were carried out with different random seeds. The 10 lowest energy structures yield the final ensemble of a given run.

To identify the optimal run the resulting ensembles were ranked as follows: The converged residues are identified as those with a C_{α} -RMS fluctuation of less than 2 Å, as reported previously(Lange et al. 2012). The average pairwise RMSD is computed on converged regions (*Appendix* Methods), and an effective precision (EP) is computed from pairwise RMSD and fraction of converged residues. For each run with constraint weight w_{cst} a cumulative score $S = E - 12 \log w_{\text{cst}} + 5\text{EP}$ is computed, where E denotes the median

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

Rosetta all-atom energy of the ensemble. If in any of the runs more than 2000 peaks with initial assignments are removed, because final models violate them, only E is considered for selection of runs, otherwise the final run is selected using S.

The models of the top-ranking ensemble are further relaxed against the automatically assigned NOEs including intra-residue and sequential NOEs using a 10-fold increased NOE-restraint weight. If this procedure reduces the number of NOE violations to less than 40% of the violations counted in the ensemble of un-relaxed models, the relaxed models are accepted as final models, otherwise the un-relaxed models are kept as final models. This was the case for data sets HR2876(raw), YR313(raw), and CtR107. For all other data sets the relaxed models are kept as final models. This refinement step generally reduces NOE-violations without significantly affecting backbone RMSD to the reference structure.

We established two criteria for *successful* calculations: 1) reasonable NOE consistency (target-function<500) and 2) convergence (*Appendix Methods A.3.2.2*). For the convergence criterion the number of well defined residues has to reach 90% or more of the total number of residues with random coil index (RCI) derived S² order parameter (Berjanskii and Wishart 2005) larger than 0.7 (Methods and *Appendix Table S2d*).

A few NOE data sets were recorded with reduced sweep width leading to *peak folding*. AutoNOE-Rosetta unfolds such frequencies on the fly, if the sweep-window is noted in the header of the respective peak list. For CYANA calculations we manually *unfolded* by replicating peaks with integer multiples of the sweep width subtracted or added to the respective frequencies. This applies to 4 peak lists of 2 proteins of our benchmark and the corresponding sweep-width parameters are given in (*Appendix Table S6*).

AutoNOE-Rosetta is parallelized for the MPI framework and runs were either carried out on our in-house cluster or on JUROPA at the Juelich Supercomputer center using 184 or 192 parallel processes, respectively.

Instructions to run AutoNOE-Rosetta including command-lines can be found in the Manual or Tutorial sections of our website (www.csrosetta.org) and in *Appendix Methods*.

3.5.3 Cyana structure calculations

Cyana 3.0 calculations were carried out to provide readers with a familiar reference for each target. TALOS+ restraints were generated from the chemical shift data, and 100 initial, and 20 final models were generated using 20,000 steps of torsion angle dynamics. RMSDs were computed from the 20 final models using the same residues and reference structure as for AutoNOE-Rosetta models (*Appendix Methods* for example script). All

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

TALOS+ predicted phi and psi angles with prediction class 'Good' are used. Two schemes to derive torsion restraints from TALOS+ predictions were tested. ACO_TIGHT restraints were generated by computing the lower- and upper bound as $\phi \pm \Delta\phi$, where ϕ denotes the TALOS+ predicted torsion angle in degree, and $\Delta\phi$ the TALOS+ estimated standard deviation. For ACO_LOOSE, we obtained bounds as $\phi \pm 2\max(\min(\Delta\phi, 35), 10)$. ACO_TIGHT is the recommended protocol at the NMR facility of the Center for Advanced Biotechnology and Medicine (CABM) as described here (http://www.nmr2.buffalo.edu/enter/NMRWiki/images/2/2e/Talos2dyana_taloserrors.txt).

ACO_LOOSE is the protocol that derives from applying the talos2dyana.com executable packaged with the TALOS+ software. A comparison of both protocols shows that ACO_TIGHT yields better accuracy over all targets (*Appendix Figure S2a*). Thus, ACO_TIGHT is used in all further CYANA calculations.

Where RDC data was available, CYANA runs were carried out both, with and without RDC data. A weight of 0.02 was used for the RDC restraint, and 0.2 as cutoff for RDC violation output. For each alignment medium 5 additional pseudo-residues of type LL5 and 1 of type ORI are attached at the end of the protein sequence. Alignment tensor parameters, Dzz and R, are estimated using the macro FindTensor.cya which employs the histogram method (Clare et al. 1998). This protocol was obtained from <http://www.nmr2.buffalo.edu/neg.wiki/CYANA>. RMSDs of CYANA calculations with RDCs were generally higher than CYANA calculations without RDCs (*Appendix Figure S2b*), whereas RDC data leads to improved results for AutoNOE-Rosetta (*Appendix Figure S2c*). Thus, CYANA calculations *without* RDCs are compared to AutoNOE-Rosetta *with* RDCs throughout the study.

3.6 References

Baran MC, Huang YJ, Moseley HNB, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. *Chem Rev* 104:3541–3556. doi: 10.1021/cr030408p

Berjanskii MV, Wishart DS (2005) A simple method to predict protein flexibility using secondary chemical shifts. *J Am Chem Soc* 127:14970–14971. doi: 10.1021/ja054842f

Clare GM, Gronenborn AM, Bax A (1998) A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. *J Magn Reson* 133:216–221. doi: 10.1006/jmre.1998.1419

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

Doreleijers JF, Raves ML, Rullmann T, Kaptein R (1999) Completeness of NOEs in protein structures: A statistical analysis of NMR data. *J Biomol NMR* 14:123–132. doi: 10.1023/A:1008335423527

Doreleijers JF, Sousa da Silva AW, Krieger E, et al. (2012a) CING: an integrated residue-based structure validation program suite. *J Biomol NMR* 54:267–283. doi: 10.1007/s10858-012-9669-7

Doreleijers JF, Vranken WF, Schulte C, et al. (2012b) NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. *Nucleic Acids Res* 40:D519–24. doi: 10.1093/nar/gkr1134

Guerry P, Herrmann T (2011) Advances in automated NMR protein structure determination. *Quart Rev Biophys* 44:257–309. doi: 10.1017/S0033583510000326

Güntert P (2008) Automated structure determination from NMR spectra. *Eur Biophys J* 38:129–143. doi: 10.1007/s00249-008-0367-z

Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *Journal of Molecular Biology* 273:283–298. doi: 10.1006/jmbi.1997.1284

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

Huang YJ, Powers R, Montelione GT (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. doi: 10.1021/ja047109h

Huang YJ, Tejero R, Powers R, Montelione GT (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* 62:587–603. doi: 10.1002/prot.20820

Lange OF Implementation of automatic NOE assignment in Rosetta. *J Biomol NMR*

Lange OF, Rossi P, Sgourakis NG, 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 USA* 109:10873–10878. doi: 10.1073/pnas.1203013109

Laskowski RA, Rullmannn JA, MacArthur MW, et al. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* 8:477–486.

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

Linge JP, Habeck M, Rieping W, Nilges M (2003a) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* 19:315–316. doi: 10.1093/bioinformatics/19.2.315

Linge JP, Williams MA, Spronk CAEM, et al. (2003b) Refinement of protein structures in explicit solvent. *Proteins* 50:496–506. doi: 10.1002/prot.10299

Mao B, Guan R, Montelione GT (2011) Improved Technologies Now Routinely Provide Protein NMR Structures Useful for Molecular Replacement. *Structure* 19:757–766. doi: 10.1016/j.str.2011.04.005

Moseley HN, Montelione GT (1999) Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struct Biol* 9:635–642. doi: 10.1016/S0959-440X(99)00019-6

Nilges M, Macias MJ, O'Donoghue SI, Oschkinat H (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β -spectrin. *Journal of Molecular Biology* 269:408–422. doi: 10.1006/jmbi.1997.1044

Raman S, Huang YJ, Mao B, et al. (2010) Accurate Automated Protein NMR Structure Determination Using Unassigned NOESY Data. *J Am Chem Soc* 132:202–207. doi: 10.1021/ja905934c

Rosato A, Aramini JM, Arrowsmith C, et al. (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure* 20:227–236. doi: 10.1016/j.str.2012.01.002

Rosato A, Bagaria A, Baker D, et al. (2009) CASD-NMR: critical assessment of automated structure determination by NMR. *Nat Meth* 6:625–626. doi: 10.1038/nmeth0909-625

Schmidt E, Güntert P (2012) A New Algorithm for Reliable and General NMR Resonance Assignment. *J Am Chem Soc* 134:12817–12829. doi: 10.1021/ja305091n

Serrano P, Pedrini B, Mohanty B, et al. (2012) The J-UNIO protocol for automated protein structure determination by NMR in solution. *J Biomol NMR* 53:341–354. doi: 10.1007/s10858-012-9645-2

Vernon R, Shen Y, Baker D, Lange OF (2013) Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker. *J Biomol NMR* 57:117–127. doi: 10.1007/s10858-013-9772-4

CHAPTER 3 ROBUST AND HIGHLY ACCURATE AUTOMATIC NOESY ASSIGNMENT AND STRUCTURE DETERMINATION WITH ROSETTA

Vriend G (1990) WHAT IF: A molecular modeling and drug design program. *Journal of Molecular Graphics* 8:52–56. doi: 10.1016/0263-7855(90)80070-V

Wüthrich K (1986) *NMR of proteins and nucleic acids*. Wiley-Interscience

3.7 My contribution to this project

In the paper- Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta, my main contribution is preparing of part of the test proteins and carrying out part of AutoNOE-Rosetta calculations. I also validated the structures using DP-score.

Chapter 4 Effect of incorrect chemical shift assignments on automated NOE assignments and NMR structure calculation

4.1 Introduction

Structure determination by nuclear magnetic resonance (NMR) spectroscopy is largely driven by distance information gathered through Nuclear Overhauser Effect Spectroscopy (NOESY). To use such data as distance restraints, the NOESY cross peaks in multidimensional spectra have to be assigned to interactions between individual atoms of the biomolecular system. Obtaining an almost complete list of the chemical shifts of each N, C, and H atom in the system usually precedes the assignment of NOE cross peaks. Obviously, wrong or missing chemical shifts have a negative impact on the subsequent NOE assignment and the resulting set of distance restraints.

Yet, such mistakes commonly occur. It has been estimated that for about 1% of all structures in the PDB at least 1 chemical shift assignment is wrong (Zhang et al. 2003). Moreover, the accuracy of side chain assignments is generally much lower than that of backbone assignments (Moseley and Montelione 1999; Schmidt and Güntert 2012). Computational methods of automated chemical shift assignment have been developed with remarkable advances in recent years (Schmidt and Güntert 2012), and yet do not routinely reach the completeness and accuracy of manual assignments in practice (Shen et al. 2008a). Multiple iterations between NOE and chemical shift assignment could improve accuracy further, but are time-consuming, and hinder full automation of NMR structure determination.

We have recently introduced a new program for automatic assignment, AutoNOE-Rosetta(Lange), which showed remarkable accuracy on a benchmark of 50 proteins(Zhang et al. 2014). Most notably, we found that AutoNOE-Rosetta also displayed a remarkable robustness against problems of incomplete or wrong chemical shift assignments. Indeed, one of the data sets, AR3436A, contained considerable errors in the chemical shift assignments. In a recent blind structure determination challenge (CASD), where this data set was introduced originally, these errors and missing resonances in the chemical shift

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

assignments caused all participating programs that use automatic NOE assignment to drive structure determination (i.e., CYANA, AutoStructure, UNIO, and ARIA), to generate models with its helix sticking out at an angle instead of packing against the core. These inaccuracies remained largely unnoticed in the original publication of the CASD results (Rosato et al. 2012), because the reference structure suffered from the same packing deficiencies presumably caused by the same problems in the input data.

Later, however, when we tested the new program, AutoNOE-Rosetta, on this data set, we obtained a well-packed structure about 4 Å away from the hitherto known reference structure. This finding prompted us to revisit the raw NMR data, and we obtained a corrected set of chemical shift assignments and a new NOE peak list, with which CYANA and AutoStructure, two of the previously failing programs, generated the same well-packed structure as AutoNOE-Rosetta with the original erroneous data set (1Å C_α-RMSD).

The strong effect that erroneous and missing chemical shift assignments had on the original structure calculations, and the obvious difficulty to detect these problems, as well as the remarkable robustness against these errors displayed by AutoNOE-Rosetta, prompted us to systematically study the effect of missing and erroneous chemical shifts on automatic NOE assignment with various programs. In 2003, Jee and Güntert studied the effect of missing resonance assignments on the automatic NOE assignment with the CANDID algorithm, and concluded that CANDID can tolerate about 10% missing chemical shifts if heteronuclear-resolved three-dimensional NOESY spectra are used (Jee and Güntert 2003). Here, we systematically studied the effect of actual errors in the assignments rather than just omissions. Moreover, we studied different patterns of missing or erroneous shifts, i.e., we ask whether it makes a difference if all shifts on a given side-chain are missing, or if the same amount of shifts is distributed uniformly across the protein.

As discussed above, AutoNOE-Rosetta displayed remarkable robustness compared to established programs on the erroneous data set of protein AR3436A. In this study we explore the generality of the robustness of AutoNOE-Rosetta to incorrect sidechain resonance assignments. Using three proteins of known structure and diverse fold-classes covering i) HR5537A (alpha-helical), ii) OR135 (alpha-beta), and iii) PfR193A (beta), we generate resonance assignment lists with missing or scrambled resonance assignments and systematically compare the performance of AutoNOE-Rosetta (Zhang et al. 2014) with the performance of two other well-established programs for automatic NOE assignment and structure determination: CYANA (Güntert et al. 1997; Herrmann et al. 2002) and AutoStructure-DP (ASDP) (Huang et al. 2005; Huang et al. 2006). The original experimental chemical shifts for these three proteins are highly complete and sufficiently correct such that

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

all programs yield highly accurate structures ($<1.7\text{\AA}$ C_{α} -RMSD to reference). Since backbone chemical shift assignment is generally highly reliable (Moseley and Montelione 1999; Baran et al. 2004; Jung and Zweckstetter 2004; Schmidt and Güntert 2012), we focus here on problems with side-chain resonances. Thus, we kept the resonances H_N , N, C_{α} , H_{α} , and CO fixed and artificially modified the sidechain resonance assignments with various levels of severity and *scramble types* to simulate incompleteness and errors. Based on these comparisons we conclude that the AutoNOE-Rosetta program is less sensitive than either CYANA or AutoStructure to errors in resonance assignments or missing resonance assignments. The improved robustness to errors in assignments results from the power of the Rosetta force field to correctly model protein structures even when some restraints are incorrectly interpreted from the experimental data.

4.2 Materials and Methods

4.2.1 Preparation of benchmark datasets

We selected three proteins PfR193A (Tejero et al. 2013), HR5537A (Liu et al. 2009) and OR135 (Koga et al. 2012) from our previous benchmark set of 50 proteins (Zhang et al. 2014). Chemical shift assignments are sufficiently complete and accurate to yield high-quality 3D models with all tested programs for all selected test proteins. The protein data sets were selected such that major fold-classes are covered (Table 4.1): a purely alpha helical protein, a beta-protein and an alpha-beta-protein.

name	PDB	residues	residues for RMSD	NOE peaks	chemical shifts	RDC	LR NH-NH ¹	LR NH-Methyl ²	LR Methyl-Methyl ³
PfR193A	2KL6	114	1-108	6191	1252	NaN	38	118	51
HR5537A	2KK1	135	39-104, 118-134	13995	1122	NaN	0	125	74
OR135	2LN3	83	5-73	6359	933	101	13	94	54

Table 4.1: The proteins selected for the benchmark and some statistics about the available NMR data for these test cases. The NMR data was originally published by NESG and is publicly available on the website http://psvs-1_4-dev.nesg.org/MR/dataset.html (Mao et al. 2011; Mao et al. 2014).

¹ Long Range(>4 residues) NH-NH restraints.

² Long Range(>4 residues) NH-Methyl restraints.

³ Long Range(>4 residues) Methyl-Methyl restraints.

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

Flexible terminal residues were predicted using TALOS+ (Shen et al. 2009) and all tail residues with predicted S_{RCI}^2 order parameter lower than 0.7 were removed (Schot et al. 2013). Internal residues of HR5537A with a predicted S_{RCI}^2 order parameter smaller than 0.7 were excluded from RMSD calculations (Table 4.1).

4.2.2 Datasets with incomplete chemical shift assignments

Incomplete chemical shift data sets were generated in sub-categories METHYL, SIDECHAINS and PROTONS as detailed below. For these sub-categories the severity levels were determined by the percentage of resonances within each sub-category that were removed from the data sets. We tested 0%(CONTROL), 10%, 30%, 50%, 70%, and 90% omission rates.

For sub-category METHYL, a given percentage of all methyl-bearing residues (ALA, LEU, ILE, VAL, MET and THR) were selected and the chemical shifts of all their methyl protons were removed. For sub-category SIDECHAINS, a given percentage of all residues were selected and the chemical shift of all sidechain atoms were removed. For sub-category PROTONS, a certain percentage of all protons on sidechains were selected and their chemical shifts were removed.

4.2.3 Datasets with swapped chemical shift assignments

In this error category, pairs of atoms were formed randomly and their respective chemical shifts swapped. Swapped chemical shift data sets were generated in the sub-categories METHYL, C-H, STEREO, CARBON, and SIDECHAIN.

To generate the sub-category METHYL, we randomly paired all methyl groups in the protein and selected 3, 6, 9, 12, and 15, pairs respectively to swap their resonances. The sub-category C-H is designed to test whether a swap of resonances of a carbon and its proton together has a more severe effect than independent swaps of protons and carbon atoms. Thus, we randomly paired side-chain carbon atoms and selected 6, 12, 18, 24 and 30 of these pairs, respectively, to swap their resonances as well as the resonances of one of their respective protons. In sub-category STEREO we tested the effect of erroneous stereospecific assignments, and swapped the resonances of 10%, 30%, 50%, 70% and 90% of all diastereotropic protons in the data sets. For sub-category CARBON, we randomly paired all sidechain carbon atoms that have the same atom name (i.e., CB's are paired with other CB's). Subsequently, we picked 10%, 30%, 50%, 70% or 90% of these pairs and swapped their chemical shifts. Finally, for sub-category SIDECHAIN, we randomly paired

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

residues with the same amino-acid type. Subsequently, we selected 1, 2, 3, 4 or 5 of these pairs and swapped the resonances of all side-chain atoms.

4.2.4 Datasets with combined chemical shift assignments

In this error category for specific pairs of atoms or atom groups one member of the pair is replaced with the chemical shifts of the corresponding atom(s) of the other member of the pair. For sub-category METHYL, we selected 10%, 30%, 50%, 70%, or 90% of all LEU, ILE and VAL residues. For all selected residues we combined the proton and carbon resonances of their two methyl groups. For sub-category STEREO, we selected 10%, 30%, 50%, 70%, or 90% of diastereo specifically assigned protons and combined their resonances.

4.2.5 Structure generation with CYANA

Seven cycles of 20,000 steps of torsion angle dynamics were run in CYANA 3.0 generating 100 initial, and 20 final models. All TALOS+ predicted phi and psi angles with prediction class 'Good' are used for torsion restraints by computing the lower- and upper bound as $\phi \pm \Delta\phi$, where ϕ denotes the TALOS+ predicted torsion angle in degree, and $\Delta\phi$ the TALOS+ estimated standard deviation. All CYANA calculations were distributed on 48 processes.

4.2.6 Structure generation with AutoNOE-Rosetta

AutoNOE-Rosetta structure calculations were run as described in Ref(Zhang et al. 2014). Chemical shift based fragments were picked using the Rosetta3 fragment picker(Schot et al. 2013; Vernon et al. 2013). The standard protocol for AutoNOE-Rosetta prescribes to run 4-6 calculations for each data set using NOE-restraint weights ranging from 2 to 100. For this study we have generated a total of 720 chemical shift data sets, such that running 4-6 calculations per data set would be prohibitively expensive. Thus, we abstain from scanning NOE-restraint weights and instead fix this weight to 10 in all runs, accepting thus somewhat diminished performance for AutoNOE-Rosetta with respect to a real application. All AutoNOE-Rosetta calculations were distributed on 178 processes.

We recorded 3 parameters to facilitate detection of failed AutoNOE-Rosetta calculations. These were 1) the number of initially assigned NOE peaks, 2) the number of finally assigned NOE peaks, and 3) the number of converged residues in the final models (Rohl et al. 2004; Shen et al. 2008b; Raman et al. 2010; Schot et al. 2013).

4.2.7 Structure generation with ASDP

ASDP (Huang et al. 2005; Huang et al. 2006) utilizes (1) a topology-based algorithm to build secondary structures including anti-parallel and parallel beta-sheets from the unassigned NOEs and resonance assignments in the first cycle, and (2) a bottom-up iterative strategy to assign NOE peaks and generate distance restraints from the list of resonance assignments and the unassigned NOEs. In the current version of ASDP, the DP score (Huang et al. 2005; Huang et al. 2006) is used to rank and filter intermediate structures which are used direct the trajectory of NOESY assignment process.

Input files for ASDP included resonance assignments, dihedral angle restraints, 3D N15-NOESY and C13-NOESY peak lists, and/or RDC. RDC data were only used here to calculate the OR135 structure. The dihedral angle restraints were generated from the chemical shifts using TALOS+(Shen et al. 2009). Only the dihedral angles from the regions predicted to be α -helices or β -strands and also classified to be 'good' by TALOS+ were used as restraints. The ranges of these dihedral angles were set by TALOS+ default. 100 structures were calculated by CYANA using the distance, dihedral angle, and hydrogen bond restraints provided by the ASDP, together with RDC data when available. Among these 100 structures, 20 structures with the best combined scores of DP and CYANA's target function [e.g. target function/weight)-DP, where weight = min(target function of 100 models)*100], were selected and iteratively calculated for additional five cycles of NOE analysis (Huang et al. 2005; Huang et al. 2006). Subsequently, the generated structures were refined with the WaterRefCNS protocol(Brünger et al. 1998) with slow cooling steps (tsc) = 0.001. RDC weight (wrdc1) = 0.2 was used in the WaterRefCNS refinement for OR135.

For CYANA, ASDP and AutoNOE-rosetta, 20, 20 and 10 structures with best scores were selected to compose the final ensembles, respectively. For comparison within the three programs, we recorded the median C_{α} -RMSD of the final structures with respect to the reference structure evaluated on the residue ranges as given in Table 4.1.

4.3 Results

The goal of this study is to test robustness of automatic NOE assignment methods against problems in chemical shift input files. Thus, we keep the peak-lists as in the original data sets and introduce errors into the chemical shift inputs in a systematic manner. The original chemical shift files are used as CONTROL.

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

To generate the test cases we started from the respective original data set and introduced three distinct categories of errors, which are inspired from our experiences with existing errors in real-world chemical shift assignments. To study the effect of incompleteness, we remove chemical shifts entirely from the data set. To study the effect of miss-assigned chemical shifts we swap chemical shifts between two distinct atoms of similar type. Furthermore, as a combination of the previous error categories we generate data sets, where two entities with distinct resonances are combined and are both assigned the same resonance, whereas the other resonance is omitted from the data set. The details, on how these error categories are generated, are given in Methods. Furthermore, within each of the three error categories outlined above, we define sub-categories based on which groups of atoms are affected. Finally, to allow a systematic study of the effect of each error sub-category, we define a severity level, which controls the amount of errors of the respective type. Accordingly, we generated data sets for a range of severity levels ending either at a value, where each program fails to yield 3D models of any reasonable quality, or at 100% severity. Within each error sub-category and for a given severity level the respective atoms or groups of atoms to affect are chosen randomly. To obtain sufficient statistics we generated 6 independent scrambled data sets at each severity level of each error sub-category. As a particular choice of scrambled resonances can have a drastically different impact on the accuracy than other choices, we do not regenerate scrambled data sets for each program, but instead use each scrambled data set with all three programs.

Figure 4.1 shows a performance overview across all different error categories. Clearly, AutoNOE-Rosetta shows a remarkable improvement in final accuracy for structure calculations from erroneous chemical shift data. In the following, we present the results for each individual error category. See also *Appendix* Table S9-S17 for the complete set of C_{α} -RMSDs of final structures generated for scrambled data sets by CYANA, ASDP and AutoNOE-Rosetta, respectively.

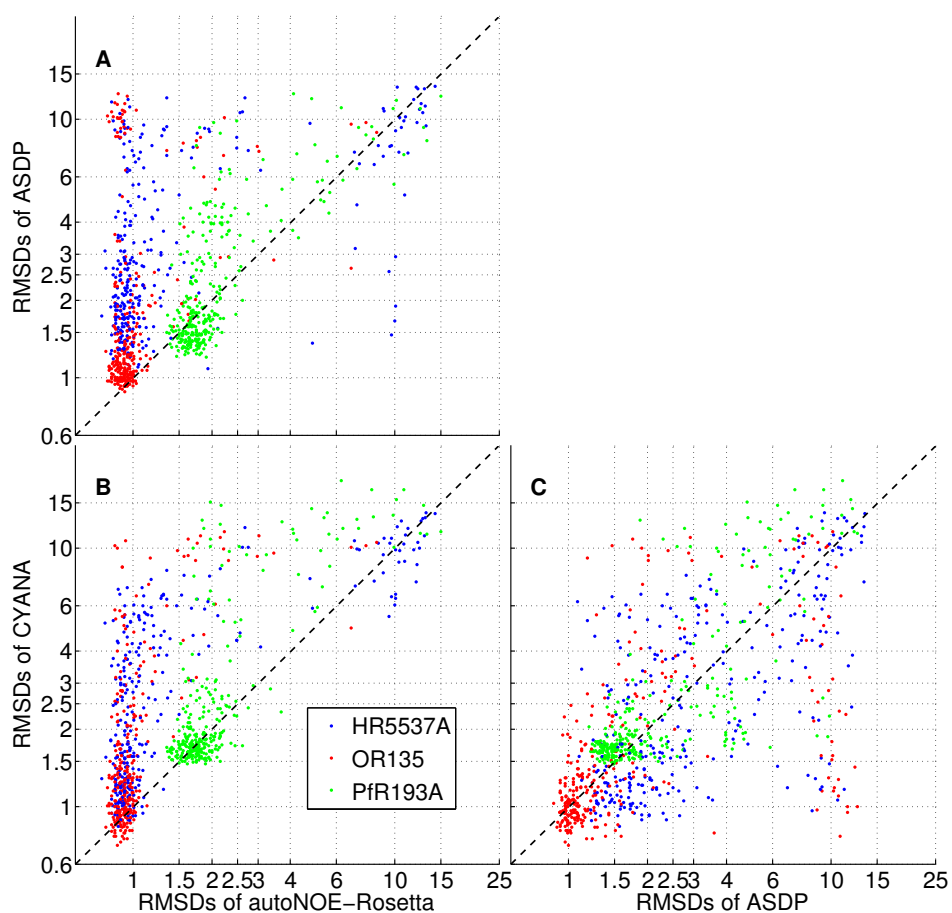


Figure 4.1: C_{α} -RMSD comparison of AutoNOE-Rosetta, ASDP and CYANA using incomplete and/or erroneous chemical shift lists. Figure 4.1A shows the comparison of ASDP and AutoNOE-Rosetta, Figure 4.1B shows the comparison of CYANA and AutoNOE-Rosetta, and Figure 4.1C shows the comparison of ASDP and CYANA.

4.3.1 Effect of missing chemical shift assignments

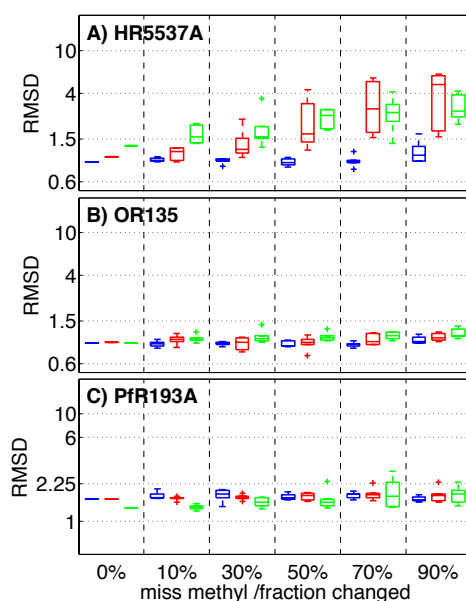


Figure 4.2: α -RMSDs statistics of final structures generated with CYANA (red), ASDP (green) and AutoNOE-Rosetta (blue) from missing methyls (A-C). X-axis shows the percentage of swapped fraction. For each severity level, 6 independent runs for AutoNOE-Rosetta, CYANA and ASDP were performed.

Figure 4.2 shows the effect of missing a fraction of the methyl chemical shifts on the programs CYANA, ASDP and AutoNOE-Rosetta, respectively. For HR5537A and OR135, we defined that structures with RMSDs $< 1.5\text{\AA}$ is accurate, RMSDs $< 4\text{\AA}$ but $> 1.5\text{\AA}$ is inaccurate and RMSDs $> 4\text{\AA}$ is failed. For PfR193A, the cutoffs are 2.25\AA and 6\AA because its RMSD with original chemical shifts is a little higher than HR5537A and OR135. This description regulation is appropriate for all analysis in this paper. for the three proteins in our benchmark set. Generally, missing the chemical shifts of methyls has no severe effect on the structure calculations with AutoNOE-Rosetta. For two of the three proteins, these missing data also had little impact on the performance of CYANA and ASDP. However, for protein target HR5537A, which is an alpha-helical protein, the performance of both CYANA and ASDP deteriorates drastically, when methyl chemical shifts are missing. For only 10% and 30% missing methyls the deterioration is still minor, but at and beyond 50% of missing methyls, both programs frequently fail to generate accurate structures. This suggests that for alpha-helical proteins methyl resonances are more important than for proteins with substantial fraction of beta content. This reliance of ASDP and CYANA on methyl contacts for alpha-helical proteins seems greatly reduced in AutoNOE-Rosetta, at least for proteins in

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

the <15kDa size range. To support this general conclusion we tested a further alpha-helical protein SR213 with CYANA and found very similar results (*Appendix Figure S10*).

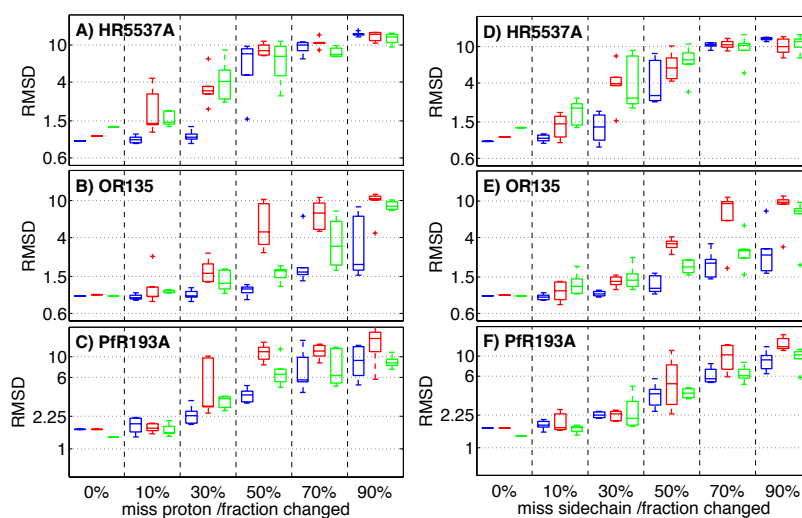


Figure 4.3: α -RMSDs statistics of final structures generated with CYANA (red), ASDP (green) and AutoNOE-Rosetta (blue) from missing side-chain proton resonance assignments (A-C) and complete omission of sidechain resonance assignments (D-F). X-axis shows the percentage of swapped fraction. For each severity level, 6 independent runs for AutoNOE-Rosetta, CYANA and ASDP were performed.

Next, we removed individual side-chain proton resonance assignments (Figure 4.3, left panels) or removed all assignments from entire sidechains at once (Figure 4.3, right panels). Naturally, the severity of this change is systematically higher than that of missing methyls, since in these new error categories at 100% severity all protons are removed such that no distance restraints at all can be derived from the NOE data.

All three programs are quite sensitive to such missing sidechain data. AutoNOE-Rosetta was the least sensitive to these classes of errors, followed by ASDP and CYANA. For the alpha-helical protein HR5537A data set analyzed with CYANA or ASDP, significant deterioration of structural accuracy is observed with as little as 10% of missing sidechain protons. At 30% missing assignments, most of the data sets result in inaccurate structure calculations, and with more than 50% missing all calculations fail. For the other two proteins the situation is somewhat better. For these proteins ASDP and CYANA remain mostly robust at 10% missing protons, with the exception of a single data set for which CYANA produced an inaccurate structure. At 30% missing, ASDP and CYANA produce inaccurate structures some times for OR135 and all the time for PfR193A. For PfR193A, CYANA calculations frequently fail already at 30% missing protons. Across all protein, AutoNOE-Rosetta can

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

handle 20-30% more missing protons than the other two programs before deteriorating. It remains accurate up to 30%, 50% and 30% for HR5537A, OR135 and PfR193A. However, for PfR193A at 30% missing protons a significant deterioration of the accuracy is already noticeable.

Generally, the deterioration in performance of all three programs due to removal of entire sidechains concerted and removal of protons independently, parallel one another. However, in some cases it seems that removing entire sidechains is less severe than removing protons independently. This behavior is observed for PfR193A at 30% removed protons for CYANA and for OR135 at >70% missing assignments for ASDP and AutoNOE-Rosetta.

4.3.2 Effect of swapping chemical shift assignments

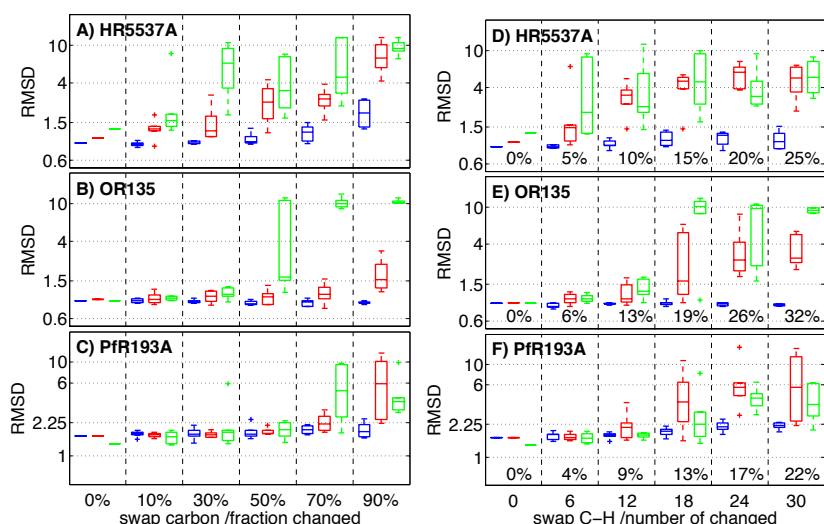


Figure 4.4: C_{α} -RMSDs statistics of final structures generated with CYANA (red), ASDP (green) and AutoNOE-Rosetta (blue) from swapped chemical shifts in SWAP-CARBON(A-C), SWAP-C-H(D-F). X-axis shows the severity level, as percentage of swapped fraction(A-C) or number of swapped pairs(D-F). For each severity level, 6 independent runs for AutoNOE-Rosetta, CYANA and ASDP were performed.

Above, we have studied the effect of incomplete chemical shifts and found that programs are generally robust against a small fraction of missing shifts. Conceivably, it's much more dangerous if chemical shifts are miss-assigned rather than just incomplete. To simulate such errors, one could randomize some chemical shifts. However, such an approach would yield mostly chemical shifts that are not reflected at all in any of the NOE cross-peaks, and thus would not be assigned. Thus, such a randomization of chemical shifts

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

would be more similar to missing chemical shifts than to miss-assigned chemical shifts. Instead, we chose to simulate miss-assignments by swapping chemical shifts. To be close to reality we further swapped shifts of atoms that have similar characteristics, i.e., same amino-acid type and same position in the side-chain, as such swaps probably reflect realistic assignment errors more closely. The different swapping types are CARBON, C-H, METHYL and SIDECHAIN as detailed in Methods.

The results of automatic NOE assignment based on chemical shift data sets with swapped resonances are shown in Figure 4.4. Overall, swapping carbons does not appear to be a highly severe error category. For OR135 and PfR193A, CYANA and AutoNOE-Rosetta remain stable up to about 50% and 70% of swapped sidechain carbons, respectively. For HR5537A, the programs performance is more different, and CYANA and ASDP start to incur inaccuracies above 10%, whereas AutoNOE-Rosetta remains stable up to 70% of swapped sidechain carbons. Surprisingly, ASDP is affected significantly more strongly by this error class than CYANA. ASDP calculations frequently fail at 50% and 70% for OR135 and PfR193A, respectively. In both cases, CYANA remains accurate at most cases. Also for HR5537A, ASDP shows a stronger deterioration of its results than CYANA. As ASDP was more similar to CYANA in other error categories, this might point to a possible avenue of improvement for ASDP.

Swapping protons together with their bound carbons has a much more detrimental effect on the structure calculations than swapping isolated carbons and thus we evaluated the effect on a scale of individual swapping events rather than a percentage of swapped resonances. For HR5537A, at six pairs of swapped C-H (equivalent to 5% of swapped entities), some scrambled data sets cause failure of CYANA and ASDP, whereas other data sets with six swapped C-H pairs cause no deterioration at all (Figure 4.4D). Similar to the previous error category, ASDP seems somewhat more sensitive to this error. At 12 pairs and more, that is at >10%, both CYANA and ASDP start to have systematic problems in producing accurate structures. AutoNOE-Rosetta, in contrast, remains nearly unperturbed, and only at severities of >50% did we see deterioration of accuracy for AutoNOE-Rosetta (Figure 4.4D and *Appendix* Figure S12). For OR135 and PfR193A, robustness of CYANA and ASDP is improved compared to HR5537A (Figure 4.4E-F), and only at higher variation (>= 18 swaps) significant deterioration of the results is observed. For OR135, AutoNOE-Rosetta is unperturbed even at high severity levels of >97% (Figure 4.4E, *Appendix* Figure S13), and for PfR193A it shows a small deterioration effect >24 pairs (17%), and significant deterioration at >40 pairs (29%) (*Appendix* Figure S14).

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

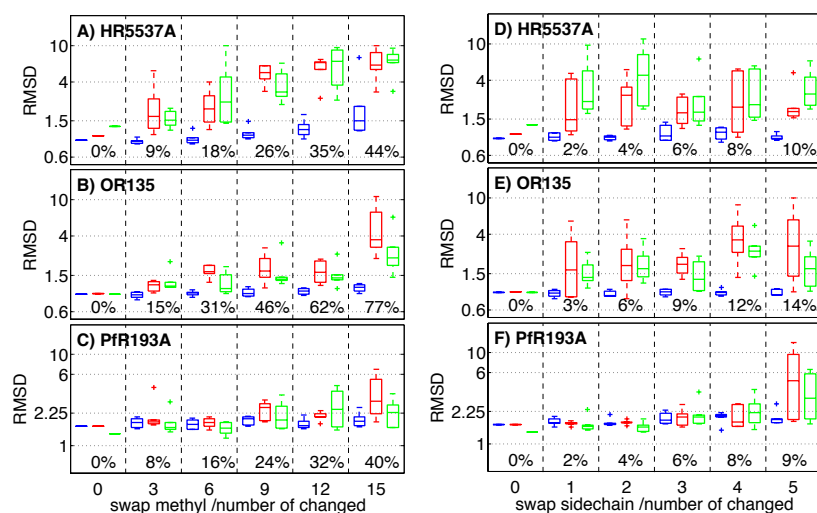


Figure 4.5: C_{α} -RMSDs statistics of final structures generated with CYANA (red), ASDP (green) and AutoNOE-Rosetta (blue) from swapped chemical shifts in SWAP-METHYL(A-C) and SWAP-SIDECHAIN(D-F). X-axis shows the number of swapped pairs. For each severity level, 6 independent runs for AutoNOE-Rosetta, CYANA and ASDP were performed.

Figure 4.5A-C shows the effect of swapping pairs of methyls on CYANA, ASDP and AutoNOE-Rosetta. Similar to other error types, CYANA and ASDP are more sensitive to errors when applied to the alpha-helical protein HR5537A than to the other two test proteins. For HR5537A, inaccuracies already appear at 10% of swapped methyls and failures at >25%. For OR135 and Pfr193A individual scrambled data sets already cause inaccuracies at ~10% of swapped methyls, but real failures only occur above 15 or more swapped pairs. For HR5537A, we even found individual swaps of methyls can cause deterioration in the accuracy of CYANA calculations (*Appendix Figure S11*).

Among all swapping error categories, the swapping of entire sidechains has the most severe effect (Figure 4.5D-F). If a whole sidechain is switched, network-anchoring filters, which are supposed to filter out spurious cross peak assignments, are less effective. When all sidechain resonances are swapped consistently, a putative NOE networks stays intact. Indeed, CYANA and ASDP calculations might fail already with a single pair of swapped sidechains for HR5537A. For both proteins, HR5537A and OR135, CYANA and ASDP are likely to generate inaccurate structures if only 1 pair of sidechains is swapped. For Pfr193A, the beta-protein of the test set, sidechain swapping is less severe, and CYANA and ASDP remain fairly robust until 4-5 swapped sidechains. AutoNOE-Rosetta, in contrast, yields accurate results for both HR5537A and OR135 with up to 5 swapped sidechains and shows only minor performance deterioration with 5 and more swapped pairs for Pfr193A. In fact,

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

AutoNOE-Rosetta remains stable until 15, 15, 5 pairs of swapped sidechains for the proteins HR5537A, OR135 and PfR193A, respectively (Figure 4.5F, *Appendix* Figure S15-16).

In addition to the four types of swapping errors discussed above, we also tested SWAP-STEREO, which swaps diastereotropic protons in the dataset (*Appendix* Figure S9). Except a single ASDP run at 30% severity, all calculations were robust against these errors. This reflects, that the programs are usually interpreting the input data of diastereotropic protons as ambiguous.

4.3.3 Effect of combining chemical shift assignments

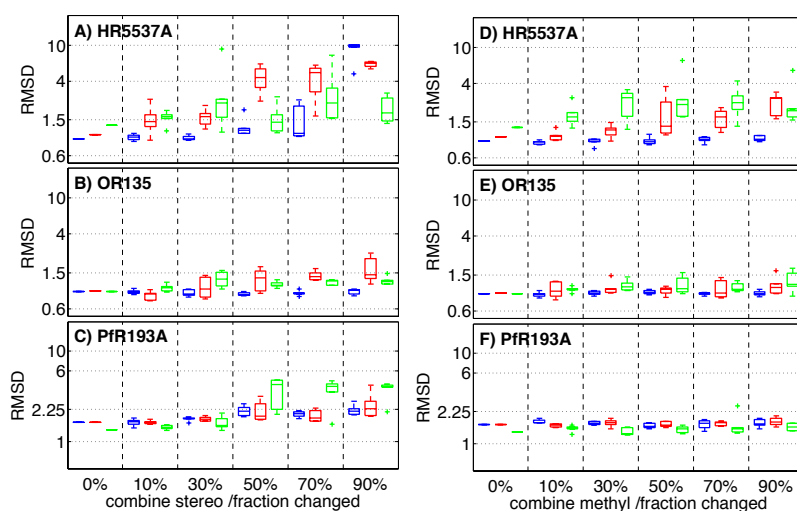


Figure 4.6: α -RMSDs statistics of final structures generated with CYANA(red), ASDP(green) and AutoNOE-Rosetta(blue) from combined chemical shifts in STEREO(A-C) and METHYL(D-F). X-axis shows the percentage of combined fraction. For each severity level, 6 independent runs for AutoNOE-Rosetta, CYANA and ASDP were performed.

In addition to missing and miss-assigned resonances, one also commonly observes errors in which one NMR resonance is assigned to two (or more) atoms, when in fact the two (or more) atoms have different resonance frequencies. This problem occurs most often within groups of similar atoms, e.g., for diastereotropic protons, or for the isopropyl methyl groups of Leucine, Isoleucine or Valine sidechains. We simulated this problem by combining chemical shifts in sub-category STEREO and METHYL (Methods).

Also for sub-category COMBINE-STEREO (Figure 4.6A-C), the programs are more affected by the errors in the purely alpha helical protein HR5537A. CYANA and ASDP start to deteriorate significantly as early as a severity of 10%, with individual failed calculations above 30% severity. Surprisingly, ASDP recovers at 90% severity. For the other proteins,

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

effects of the scrambling are weak but noticeable at >30% severity for OR135 and > 50% for PfR193A. For PfR193A at >50% severity ASDP calculations are worst.

In line with the other error categories, AutoNOE-Rosetta is more robust. For COMBINE-STEREO it shows nearly no effect for proteins OR135 and PfR193A at all severity levels, and only weak effect for HR5537A at 70%. At 90% of HR5537A, however, AutoNOE-Rosetta calculations also fail.

At any given severity level the expected effect of COMBINE-METHYL is generally less severe than that of COMBINE-STEREO, since there are fewer protons in methyl-groups than there are diastereotropic protons. Indeed, a loss of performance with COMBINE-METHYL is only observed for HR5537A at levels above 50% or 10% with CYANA or ASDP, respectively.

4.3.4 Effect of missing resonances with low-fidelity assignments

Above we have analyzed the effects of missing or erroneous assignments on protein structure determinations. A practical situation that might arise in the process of NMR structure determination is that a certain number of assignments have been made, but they are far from complete. However, the accuracy of the assignments is quite high. At this point in the process the NMR expert has three choices: a) accept the achieved structural quality, b) obtain more experimental data to improve assignment coverage, or c) run computational methods to obtain missing assignments. Whereas choice b) is costly, choice c) might incur a substantially higher error rate in the new sub-set of assignments compared to the already available set of assignments. To understand, whether choosing c) improves performance despite the substantial error rate in the automatic assignments, we re-analyzed our data in this way. Indeed, as seen in Figure 4.7, when starting from incomplete assignments, adding more resonance assignments (even with a substantial assignment error rate) might actually still lead to an overall improvement of final structural accuracy when interpreted with AutoNOE-Rosetta. To find the critical proportion where improved accuracy can still be expected, we simulate cases where a substantial fraction (50%/70%) of all chemical shifts are correctly assigned but the rest of the assignments are provided by a method with a higher error rate (10%/20%).

As expected more data is better at any base error rate, and with a higher quality of the assignments, calculations are more accurate with less complete data sets (solid lines, Figure 4.7). If we now combine data sets of different qualities to simulate the scenario motivated above, we see that under most circumstances an improvement in final accuracy

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

can be gained, if incomplete assignments of high quality are supplemented by additional low-quality assignments (dashed lines, Figure 4.7).

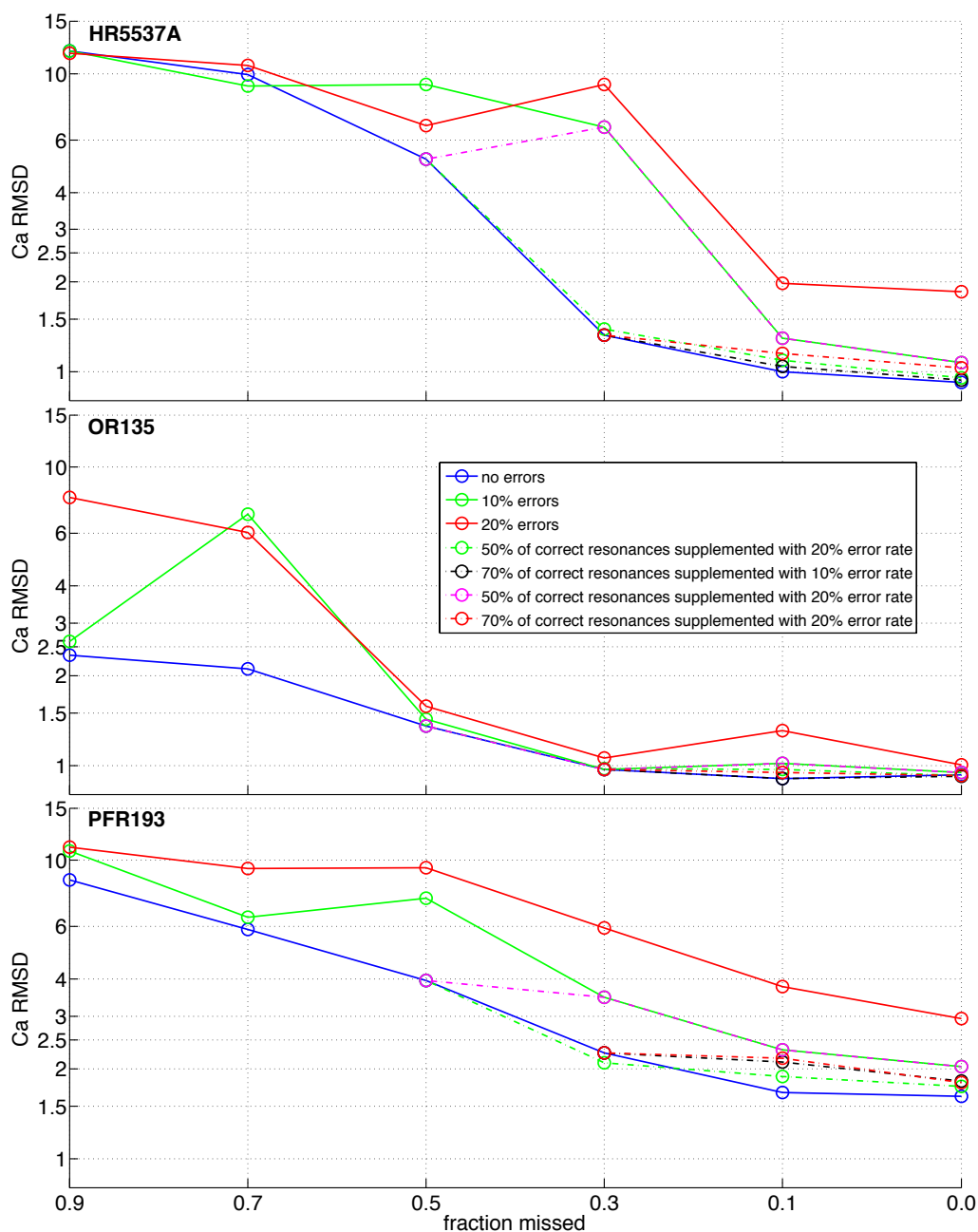


Figure 4.7: Effect of assigning missing resonances with a low-fidelity assignment method. Here we show the effect of combining incomplete but high-fidelity chemical shift assignments with further assignments that are less correct. The solid curves show expected the accuracy given a base error rate in the assignments. Dashed curves that start from the solid curves at 50% and 30% missing resonances, respectively, show the expected accuracy, if the remaining resonance assignments are provided by a method with a higher error rate (Legend). One can see, for instance, that final accuracy is significantly improved if

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

a fully-correct (e.g., manual) half-assigned data set is complemented with an erroneous assignment method (e.g., computations) that is expected to produce ca. 20% of errors.

4.3.5 Indicating problematic runs of AutoNOE-Rosetta

As shown above, the calculation of protein structures will be various if the precision of chemical shifts are different, so it's quite necessary to have clear criteria to indicate the problematic calculations. In a previous paper(Zhang et al. 2014), two automatically computed criteria: convergence and intrinsic NOE consistency are presented to flag inaccurate calculations for AutoNOE-Rosetta. Although highly inaccurate structures may exhibit good convergence(Huang et al. 2005), but particularly for Rosetta, adopting convergence to identify the structure accuracy has been proved to be generally accurate(Rohl et al. 2004; Shen et al. 2008b; Raman et al. 2010; Schot et al. 2013). Intrinsic NOE consistency is an empirical criterion which works quite well in previous research but is proved to be inappropriate to current cases. In 2005, Huang et al. presented RPF(Recall, Precision, and F-measure) score for structure quality assessment(Huang et al. 2005). Here, a new criterion similar to the Recall score, the percentage of NOE peaks matched to final structures is introduced to filter out problematic runs. For classification, it's significant that a reliable classification method tries to output few false positives and false negatives. Among the two, false positives are prior to be avoided. Therefore, we manually determine a linear separatrix in this research, which can filter out all high RMSD calculations(0 false positive) with only a few false negatives(Figure 4.8-10).

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

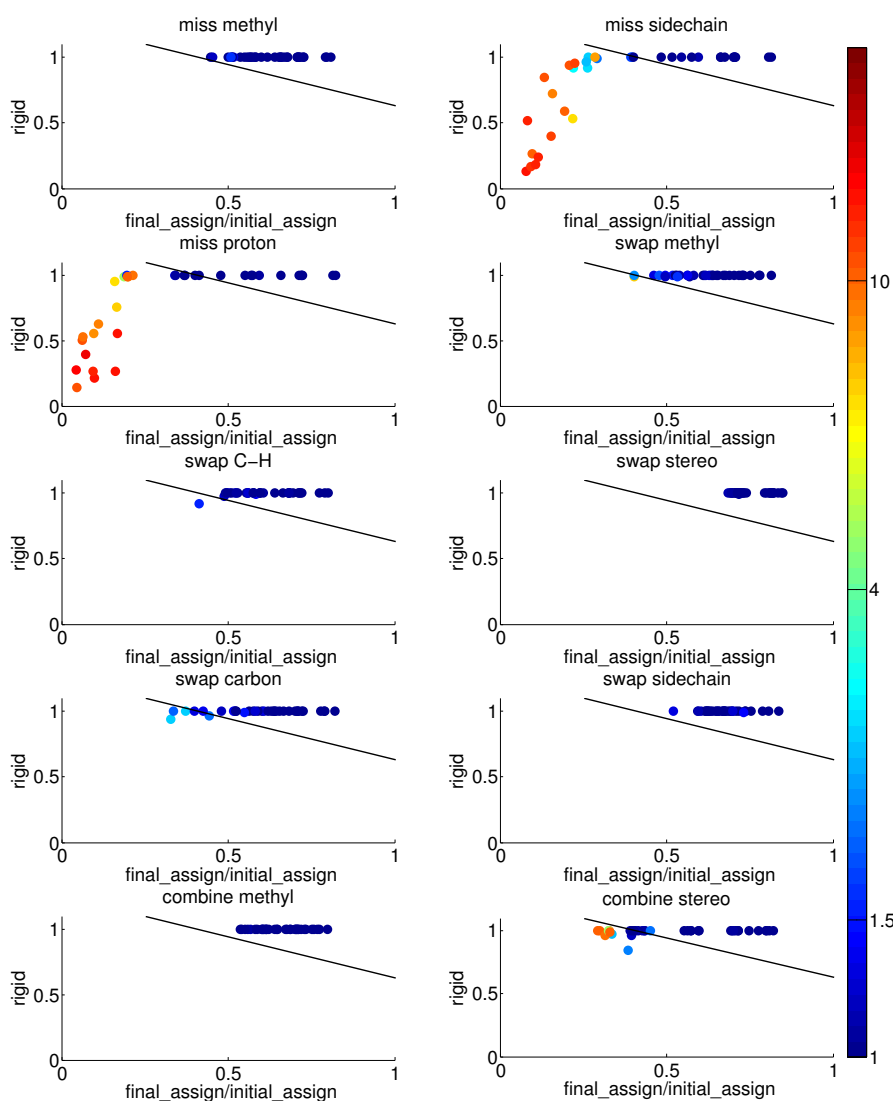


Figure 4.8: Distinguish of questionable calculations of AutoNOE-Rosetta on target HR5537A. Y-axis is the percentage of converged residues and x-axis value is described by follow: number of NOE cross peaks explained by final structures divided by number of initial assigned peaks. Each dot represents one calculation of AutoNOE-Rosetta and color shows the value $C\alpha$ RMSDs. The black solid line described by $y = -0.63x + 1.26$ classifies all calculations into success (above the lines) and failure (below the lines).

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

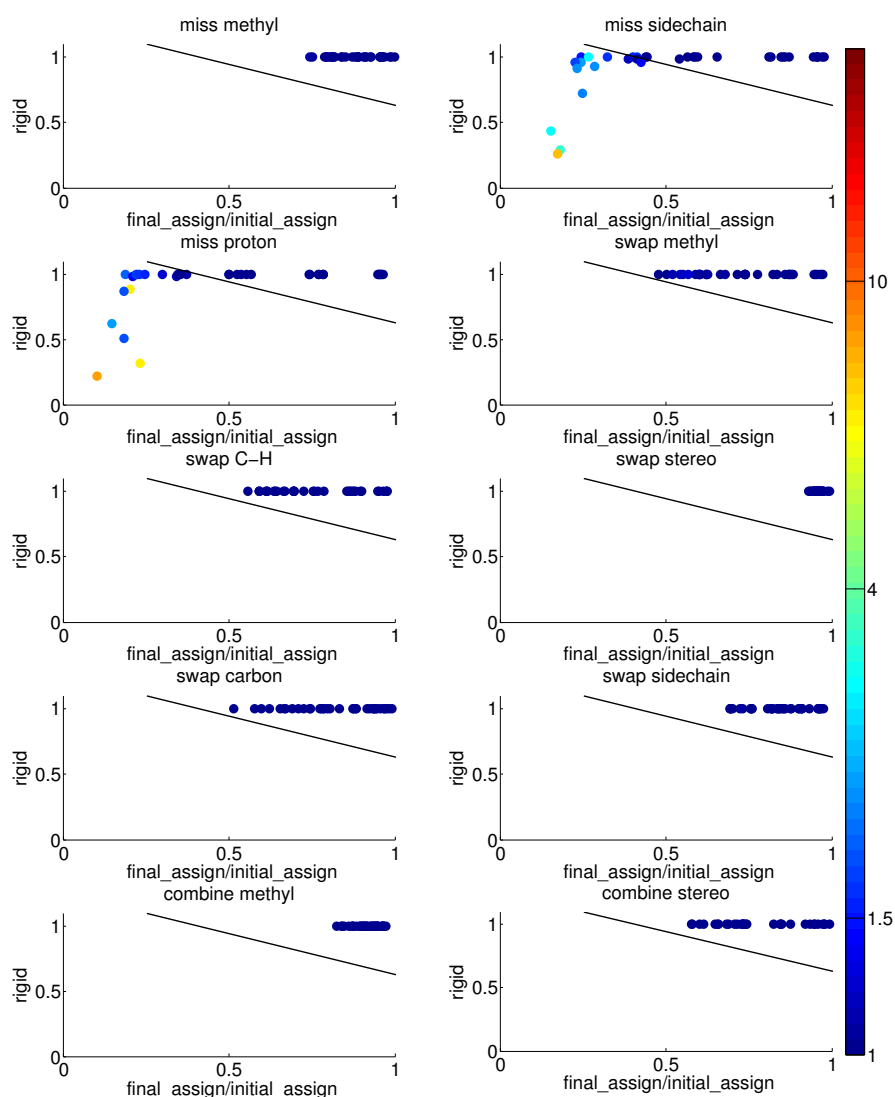


Figure 4.9: Distinguish of questionable calculations of AutoNOE-Rosetta on target OR135. Y-axis is the percentage of converged residues and x-axis value is described by follow: number of NOE cross peaks explained by final structures divided by number of initial assigned peaks. Each dot represents one calculation of AutoNOE-Rosetta and color shows the value $C\alpha$ RMSDs. The black solid line described by $y = -0.63x + 1.26$ classifies all calculations into success (above the lines) and failure (below the lines). In the sub-figure of miss proton, there is one case with poor convergence but low RMSD which is infrequent. The reason for this situation is that 9 of 10 finally selected structures are accurate and similar to native structure (Appendix Figure S17). However, the rest 1 is quite far away from the native structure and its RMSD is higher than 8\AA , then the convergence of this case is low but its mean RMSD is still good.

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

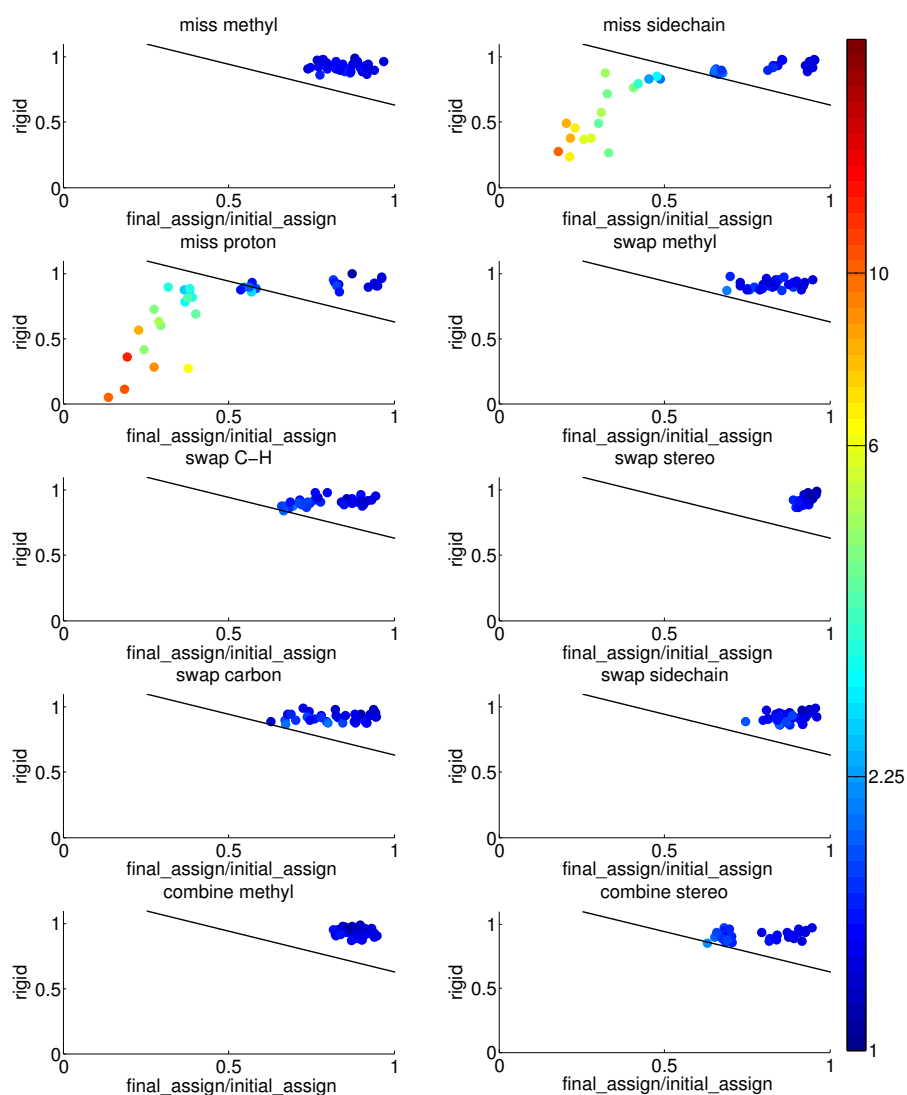


Figure 4.10: Distinguish of questionable calculations of AutoNOE-Rosetta on target PFR193A. Y-axis is the percentage of converged residues and x-axis value is described by follow: number of NOE cross peaks explained by final structures divided by number of initial assigned peaks. Each dot represents one calculation of AutoNOE-Rosetta and color shows the value $C\alpha$ RMSDs. The black solid line described by $y = -0.63x + 1.26$ classifies all calculations into success(above the lines) and failure(below the lines).

4.4 Conclusion

We studied the effect of incomplete and erroneous chemical shifts on automatic NOE assignments and protein structure determinations. With 3 automatic NOE assignment and protein de novo programs CYANA, ASDP and AutoNOE-Rosetta, the test was carried out on a benchmark of three proteins whose structures are known and their original experimental chemical shifts are highly complete and correct. We started from the original data set and introduced three distinct categories of errors missing, swapping and combining to the correct chemical shifts

The results in this paper confirm our general intuition that miss-assignments are worse than missed assignments. About 9-10% of swapped sidechains have already caused serious deterioration to structural accuracy obtained with CYANA and ASDP but at 10% missing sidechains the programs still yield accurate structures with only minor deteriorations. For methyls, at 90% missing methyls CYANA and ASDP still produce accurate structures for the proteins with beta- or alpha-beta fold class, and AutoNOE-Rosetta for all fold-classes, whereas already at ca. 10% of swapped methyls, CYANA and ASDP can be affected unfavorably. In particular, for alpha-helical proteins errors in the methyl-groups have a high impact. Indeed, even AutoNOE-Rosetta fails to produce accurate structures at more than 40% of swapped methyls, whereas it is unaffected by more than 90% missing methyls even for the alpha-helical proteins tested here.

Among the three proteins, the purely alpha helical protein HR5537A is more likely to be influenced by scrambled assignments of sidechains. We can rationalize this result by its purely alpha-helical nature, such that the tertiary structure of the protein is mostly determined by long-range restraints involving sidechain protons. In contrast, for beta-sheet containing proteins their tertiary structure is determined to a large extent by NOEs involving backbone protons.

Comparing the performances of the three programs CYANA, ASDP and AutoNOE-Rosetta, AutoNOE-Rosetta generally outperforms the others when there is an extensive resonance assignment incompleteness and/or error rate. For the cases of missing and swapping methyls, AutoNOE-Rosetta, yielded accurate structures at all runs. For the other "scramble-type" errors, the scramble level where AutoNOE-Rosetta starts to fail is always higher than that of CYANA and ASDP. For the same protein with the same scrambled resonance list, structures of CYANA and ASDP are generally less accurate than AutoNOE-Rosetta structures. CYANA and ASDP also proved to be quite sensitive to some key chemical shift assignments. In some cases, a minor number of missing or swapped chemical

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

shifts can deteriorate these calculations, whereas the corresponding AutoNOE-Rosetta calculation would still produce accurate results. The structure calculation algorithm in CYANA is based on the torsion angle dynamics(Güntert et al. 1997), ASDP uses XPLOR/CNS(Huang et al. 2006) for 3D structure determination and the sampling protocol for AutoNOE-Rosetta is RASREC-Rosetta(Zhang et al. 2014). within these three protocols, only RASREC-Rosetta could predict accurate structures without any experimental restraints because of its fragment-based sampling method. Since the methods of initial NOE assignment and constraint distance calibration of CYANA, ASDP and AutoNOE-Rosetta are similar, their different sampling algorithms are probably the main reason of different performances on incorrect data.

For reliable NOE assignment and structure calculation, CYANA and ASDP generally require 90% correct sidechain resonance assignments, whereas only 70% correct assignments are enough for AutoNOE-Rosetta. With erroneous chemical shift list, CYANA and ASDP cannot guarantee accurate results because even minor miss-assign errors would deteriorate their calculations. AutoNOE-Rosetta, on the other hand, can yield correct structures beyond 10% severity for nearly all types of miss-assignment errors.

In the course of this work, several features of the ASDP program were identified which make it sensitive to resonance assignment errors. Code modifications to correct these aspects of the algorithm were observed to significantly improve the performance of ASDP with incomplete or scramble resonance assignment lists. These improvements in ASDP benchmarked on the scrambled data sets introduced in this current study will be described elsewhere.

Besides the statistic research of performances of CYANA, ASDP and AutoNOE-Rosetta on incomplete/incorrect chemical shifts, this paper also introduces a new scheme to classify AutoNOE-Rosetta calculations into success or failure based on percentages of converged residues and number of assigned NOE peaks, which is proved to be reliable and can filter out all calculations with high RMSD structures.

A program to artificially scramble chemical shifts with kinds of problems and severity levels is available within the CS-ROSETTA toolbox versions 2.x and higher at www.csrosetta.org.

4.5 References

Baran MC, Huang YJ, Moseley HNB, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. *Chem Rev* 104:3541–3556. doi: 10.1021/cr030408p

Brünger ATA, Adams PDP, Clore GMG, et al. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* 54:905–921. doi: 10.1107/S0907444998003254

Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *Journal of Molecular Biology* 273:283–298. doi: 10.1006/jmbi.1997.1284

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

Huang YJ, Powers R, Montelione GT (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. doi: 10.1021/ja047109h

Huang YJ, Tejero R, Powers R, Montelione GT (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* 62:587–603. doi: 10.1002/prot.20820

Jee J, Güntert P (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. *J Struct Funct Genomics* 4:179–189.

Jung YS, Zweckstetter M (2004) Backbone assignment of proteins with known structure using residual dipolar couplings. *J Biomol NMR* 30:25–35.

Koga N, Tatsumi-Koga R, Liu G, et al. (2012) Principles for designing ideal protein structures. *Nature* 491:222–227. doi: 10.1038/nature11600

Lange OF Implementation of automatic NOE assignment in Rosetta. *J Biomol NMR*

Liu G, Huang YJ, Xiao R, et al. (2009) NMR structure of F-actin-binding domain of Arg/Abi2 from *Homo sapiens*. *Proteins* 78:1326–1330. doi: 10.1002/prot.22656

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

Mao B, Guan R, Montelione GT (2011) Improved Technologies Now Routinely Provide Protein NMR Structures Useful for Molecular Replacement. *Structure* 19:757–766. doi: 10.1016/j.str.2011.04.005

Mao B, Tejero R, Baker D, Montelione GT (2014) Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. *J Am Chem Soc* 136:1893–1906. doi: 10.1021/ja409845w

Moseley HN, Montelione GT (1999) Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struct Biol* 9:635–642. doi: 10.1016/S0959-440X(99)00019-6

Raman S, Lange OF, Rossi P, et al. (2010) NMR structure determination for larger proteins using backbone-only data. *Science* 327:1014–1018. doi: 10.1126/science.1183649

Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein Structure Prediction Using Rosetta. In: *Methods in enzymology*. Elsevier, pp 66–93

Rosato A, Aramini JM, Arrowsmith C, et al. (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure* 20:227–236. doi: 10.1016/j.str.2012.01.002

Schmidt E, Güntert P (2012) A New Algorithm for Reliable and General NMR Resonance Assignment. *J Am Chem Soc* 134:12817–12829. doi: 10.1021/ja305091n

Schot G, Zhang Z, Vernon R, et al. (2013) Improving 3D structure prediction from chemical shift data. *J Biomol NMR*. doi: 10.1007/s10858-013-9762-6

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. doi: 10.1007/s10858-009-9333-z

Shen Y, Vernon R, Baker D, Bax A (2008a) De novo protein structure generation from incomplete chemical shift assignments. *J Biomol NMR* 43:63–78. doi: 10.1007/s10858-008-9288-5

Shen Y, Zhang Z, Delaglio F, et al. (2008b) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105:4685–4690. doi: 10.1073/pnas.0800256105

Tejero R, Snyder D, Mao B, et al. (2013) PDBStat: a universal restraint converter and restraint analysis software package for protein NMR. *J Biomol NMR* 56:337–351. doi: 10.1007/s10858-013-9753-7

CHAPTER 4 EFFECT OF INCORRECT CHEMICAL SHIFT ASSIGNMENTS ON AUTOMATED NOE ASSIGNMENTS AND NMR STRUCTURE CALCULATION

Vernon R, Shen Y, Baker D, Lange OF (2013) Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker. *J Biomol NMR* 57:117–127. doi: 10.1007/s10858-013-9772-4

Zhang HY, Neal S, Wishart DS (2003) RefDB: A database of uniformly referenced protein chemical shifts. *J Biomol NMR* 25:173–195.

Zhang Z, Porter J, Lange OF (2014) Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta. *J Biomol NMR*

Chapter 5 Conclusion and Discussion

The scope of this work is the investigation of automotive protein structure calculation by NMR data and CS-Rosetta. In this dissertation, I improved the performance of structure calculation by CS-Rosetta and extend its functions by the following routes. 1. I improved CS-Rosetta for computing structures from backbone-only chemical shifts by introducing CS-Score to RASREC-rosetta and a new calculation annotation method. 2. I presented a new NOE assignment and structure determination algorithm that can—unsupervised—produce results that are both reliable and accurate. 3. I tested the robustness and reliability of de novo programs against low quality chemical shift assignments.

Firstly, I report advances in the calculation of protein structures from chemical shift nuclear magnetic resonance data alone. Our previously developed method, CS-Rosetta, assembles structures from a library of short protein fragments picked from a large library of protein structures using chemical shifts and sequence information. Here I demonstrate that combination of a new and improved fragment picker and the iterative sampling algorithm RASREC yield significant improvements in convergence and accuracy. Moreover, I introduce improved criteria for assessing the accuracy of the models produced by the method. The method was tested on 39 proteins in the 50–100 residue size range and yields reliable structures in 70 % of the cases. All structures that passed the reliability filter were accurate ($<2\text{\AA}$ RMSD from the reference).

Secondly, I have developed a novel and robust approach for automatic and unsupervised simultaneous nuclear Overhauser effect (NOE) assignment and structure determination within the CS-Rosetta framework. Starting from unassigned peak lists and chemical shift assignments, auto-NOE-Rosetta determines NOE cross-peak assignments and generates structural models. The approach tolerates incomplete and raw NOE peak lists as well as incomplete or partially incorrect chemical shift assignments, and its performance has been tested on 50 protein targets ranging from 50 to 200 residues in size. We find a significantly improved performance compared to established programs, particularly for larger proteins and for NOE data obtained on perdeuterated protein samples. X-ray crystallographic structures allowed comparison of Rosetta and conventional, PDB-deposited, NMR models in 20 of 50 test cases. The unsupervised AutoNOE-Rosetta models were often of significantly higher accuracy than the corresponding expert-supervised NMR models deposited in the PDB. We also tested the method with unrefined peak lists and found that performance was nearly as good as for refined peak lists. Finally, demonstrating our

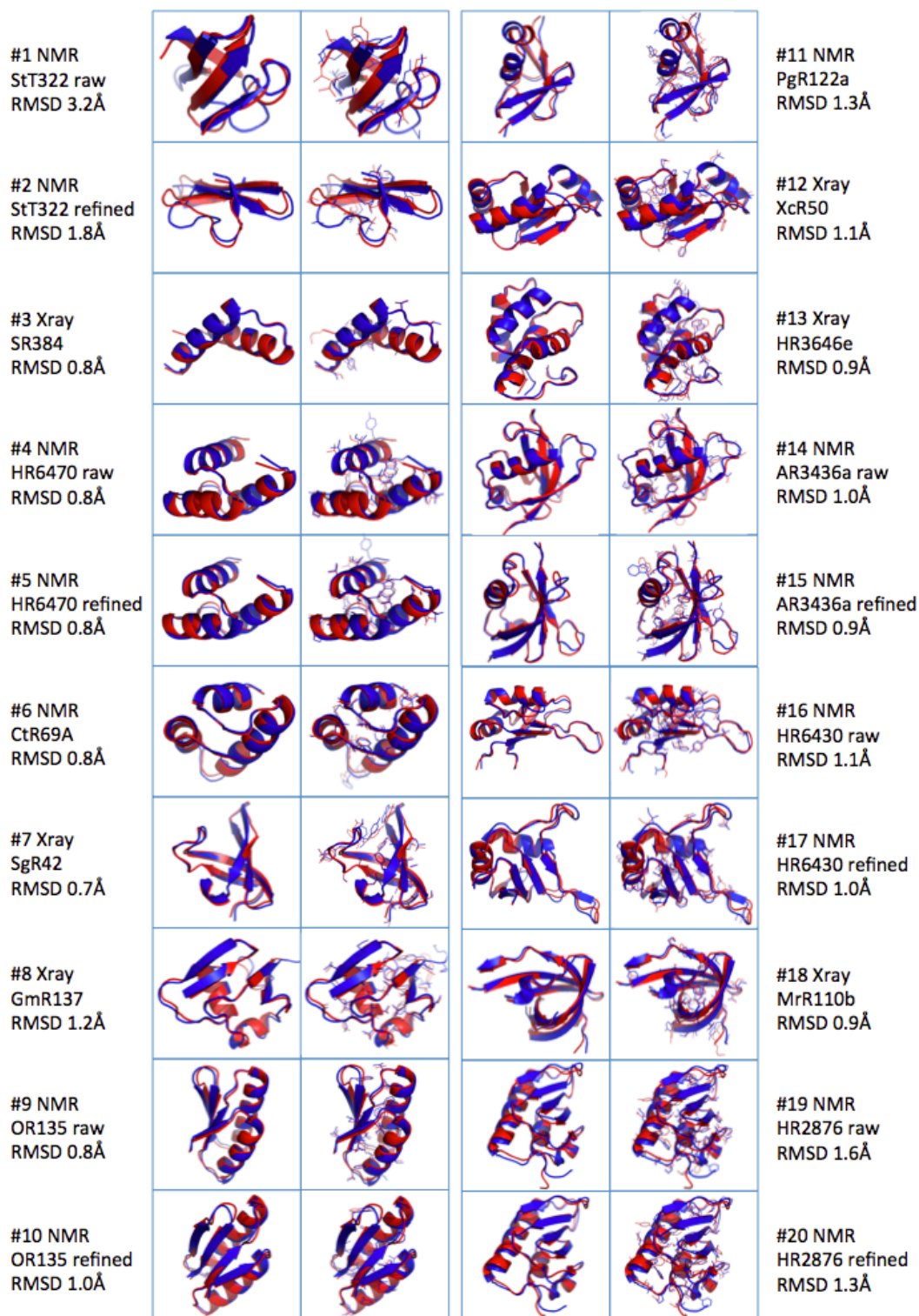
CHAPTER 5 CONCLUSION AND DISCUSSION

method's remarkable robustness against problematic input data we provided correct models for an incorrect PDB-deposited NMR solution structure.


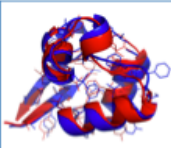

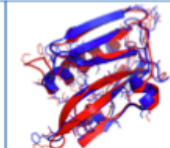
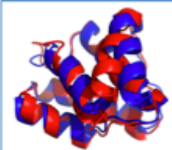
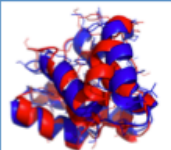
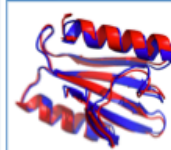
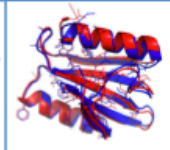
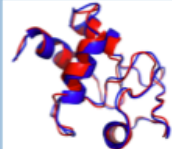

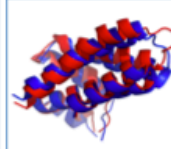
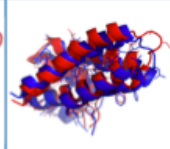
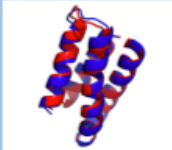
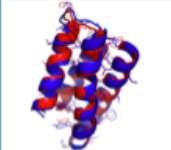

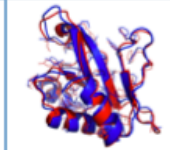
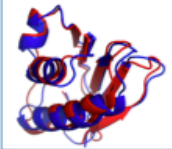
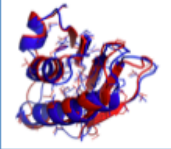
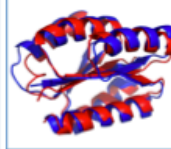
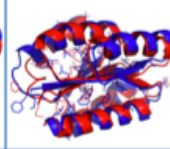
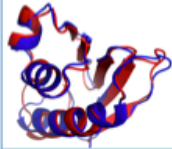
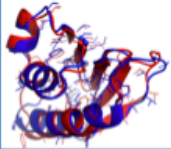
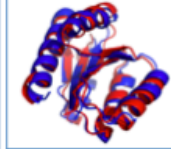
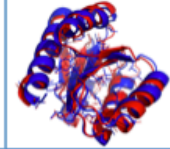


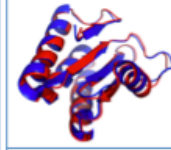
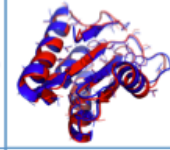

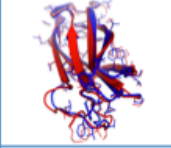

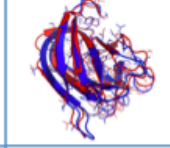

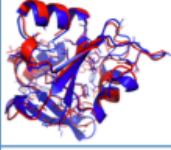
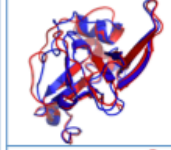
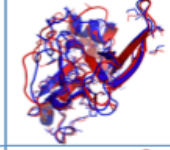
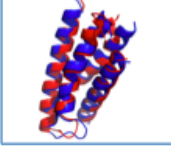
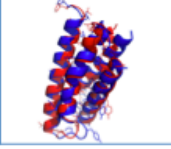
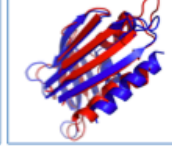
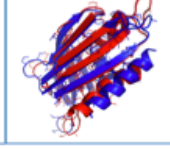
Thirdly, I investigate the influence of incomplete and wrong chemical shifts on the reliability of NMR structures obtained with automated NOE assignment. Three programs: CYANA, ASDP and AutoNOE-Rosetta were used for automatic NOE assignment and structure determination based on chemical shifts with various levels of severity and scramble types to simulate incompleteness and errors. The result proves that AutoNOE-Rosetta generally outperforms CYANA and ASDP with incompleteness and/or erroneous resonance assignments. Among the three types of proteins, the purely alpha helical protein is more likely to be influenced by incomplete or wrong assignments of sidechains. In addition, a new discriminating protocol to flag inaccurate calculations for AutoNOE-Rosetta is also presented.

Appendix

A.1 Supplementary Figures



APPENDIX

#21 NMR NeR103a RMSD 1.3Å					#31 Xray OR8C RMSD 1.2Å
#22 Xray StR65 refined RMSD 2.1Å					#32 NMR AtT13 RMSD 0.9Å
#23 Xray VpR247 RMSD 0.9Å					#33 Xray SsR10 RMSD 1.5Å
#24 NMR HR5573 RMSD 0.8Å					#34 Xray PsR293 RMSD 1.6Å
#25 NMR YR313 raw RMSD 1.4Å					#35 NMR OR36 raw RMSD 1.5Å
#26 NMR YR313 refined RMSD 1.3Å					#36 NMR OR36 refined RMSD 1.2Å
#27 NMR ET109 RMSD 1.3Å					#37 NMR CgR26a RMSD 0.7Å
#28 Xray Pfr193 RMSD 1.1Å					#38 Xray HR1958 RMSD 1.3Å
#29 Xray CcR55 unconverged RMSD 2.5Å					#39 XRAY SR10 RMSD 2.1Å
#30 Xray SR213 RMSD 1.5Å					#40 Xray DrR1470 RMSD 2.3Å

APPENDIX

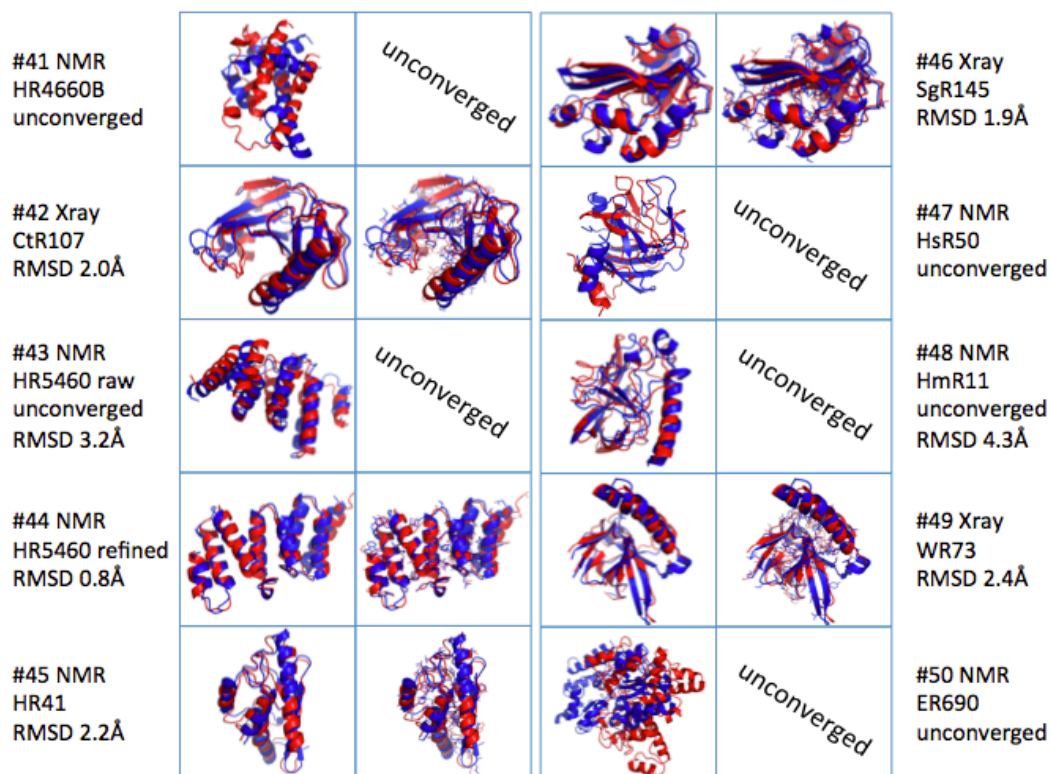


Figure S1: Final structures (red) are shown superimposed with the reference structure (blue). The same structures are shown with hydrophobic and aromatic sidechains as sticks in the right-hand panel. If the AutoNOE-Rosetta calculation is not converged, the panel for sidechain details is omitted.

APPENDIX

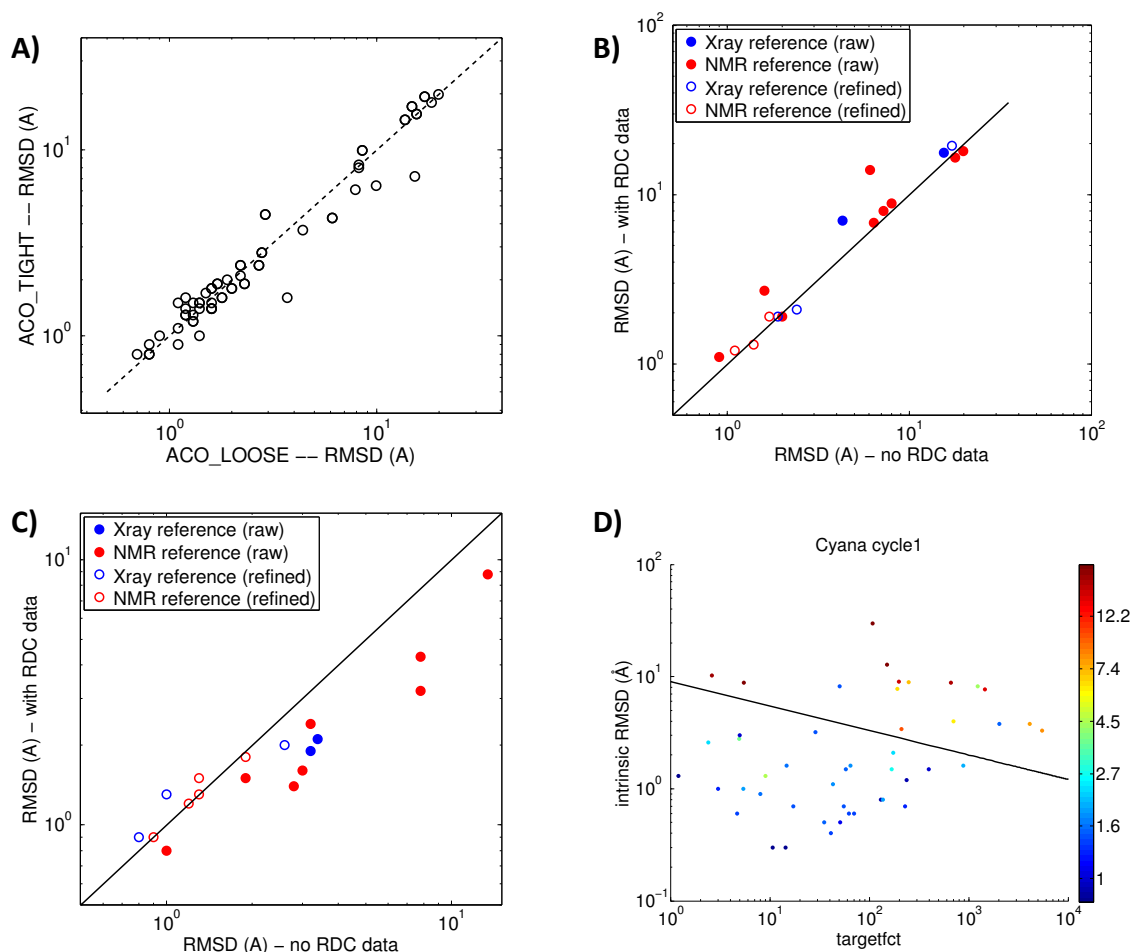


Figure S2: Supporting data for choices in the CYANA structure determination protocol. **A-C)** C_α -RMSD against a PDB deposited reference structure. **A)** CYANA calculations carried out with dihedral restraints derived from TALOS+ using either the ACO_LOOSE or the ACO_TIGHT protocol (Methods main text). **B)** CYANA was run with and without the RDC data for the 17 targets, where RDC data was available. Targets for which an Xray structure was available as reference are shown in blue, all other targets in red. Targets with data type *raw* or *unrefined* are shown with closed faces, whereas targets with *refined* data are shown as open faced circles. As shown here, CYANA generally performs worse with RDC data, and hence AutoNOE-Rosetta (with RDC) is compared to CYANA (without RDC) throughout the study. **C)** AutoNOE-Rosetta with and without the RDC data (colors and symbols as in panel B). **D)** Failure of CYANA runs can best be seen after the first cycle (P. Güntert, private communication). Shown are the target function (x-axis, log) and intrinsic bb-RMSD (y-axis, log) after cycle 1. The color of the points is given by the C_α -RMSD against the reference structure of the final CYANA structures after cycle 7 (colorbar). By manual inspection we determined the linear decision boundary described by $y = 10^{-0.22\log_{10}(x)+0.95}$, that yields the optimal classification into failed and successful runs, such that runs above the line are

APPENDIX

classified as failures. This classification has been used to classify CYANA runs. From this data it follows that for a successful run $10^{-0.22\log_{10}(x)+0.95} - y > 0$, where x denotes the target-function after cycle 1, and y the backbone RMSD after cycle 1. These values can be found using the command *cyanatable*.

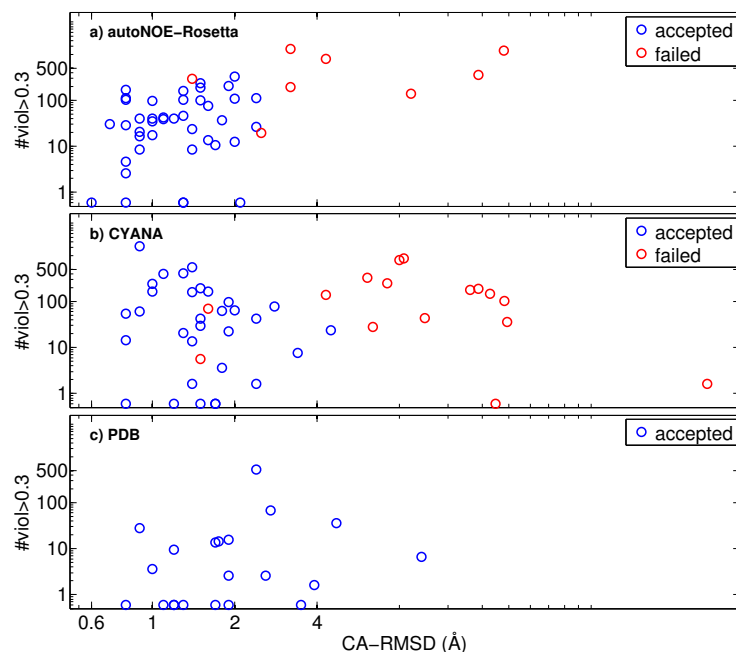


Figure S3: Scatter plots of number of violations above 0.3Å vs. the C_{α} -RMSD against the reference structure computed for models obtained from AutoNOE-Rosetta, CYANA, and the PDB. The numbers of violations are computed for the NOE restraint sets generated by the respective methods (AutoNOE-Rosetta, CYANA) or for the NOE restraint set downloaded from the PDB. For CYANA and AutoNOE calculations that passed the automatic acceptance criteria of the respective method are shown in blue, the others in red (*Appendix Method A.3.2*). For PDB only targets with an X-ray reference structure are shown. As obvious from the plots, no correlation between RMSD and number of violations can be detected, whereas the automatic acceptance criteria succeed quite well in discriminating failed calculations.

APPENDIX

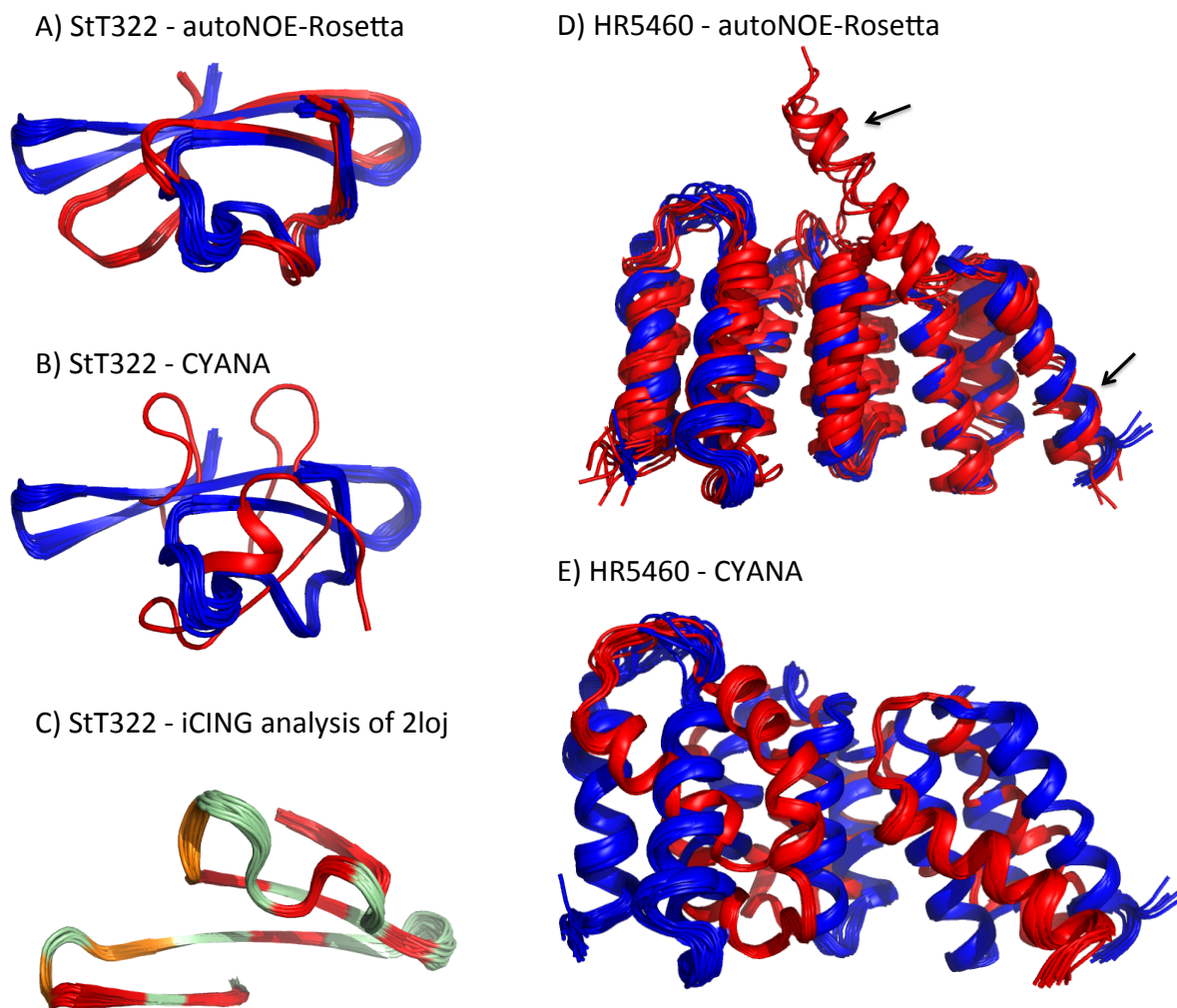


Figure S4: Results for the raw data sets of targets StT322 (**A-C**) and HR5460 (**D-E**). Final ensembles (red) are compared to the respective reference ensemble (blue). Ensemble generated with AutoNOE-Rosetta (**A+D**), ensembles generated with CYANA (**B+E**). **C**) NRG-CING report(Doreleijers et al. 2012) of reference structure shows many residues with warnings (red, orange). **D**) Black arrows mark the positions of the C-terminal helix in the AutoNOE-Rosetta models. In 4 models the helix is not packed against the remainder of the structure (up), in the remaining 6 models, the helix is packed against helix 4+5 in the same location as in the reference structure (down).

APPENDIX

Dataset for MR study

http://psvs-1_4-dev.nesg.org/MR/dataset.html

NESG_ID	XRay_PDB_ID	Sequence	XRAY				NMR										
			Length	Resolution(A)	Space group	Coordinates	Structure factor	NMR_PDB_ID	Sequence	Oligomer	Coordinates	Constraints	BMRB ID	Chemical Shift	Peaks List ^b	FID ^b	RDC
IbR31	3CPK	3CPK.seq	150	2.5	P43212	3CPK.pdb	3CPK-sf.cif	2K2E	2K2E.seq	monomer	2K2E.pdb	2K2E.mr	15702	15702.hmrh	NA	NA	Phage
CcR55	200Q	200Q.seq	115	1.8	C222	200Q.pdb	200Q-sf.cif	2JQN	2JQN.seq	monomer	2JQN.pdb	2JQN.mr	15281	15281.hmrh	15281.peaks	15281.fid	NA
CsR4	20TA	20TA.seq	76 (2)	2.2	P212121	20TA.pdb	20TA-sf.cif	2JR2	2JR2.seq	dimer	2JR2.pdb	2JR2.mr	15317	15317.hmrh	15317.peaks	15317.fid	PEG PEG+CTAB Phage
CtR107	3E0H	3E0H.seq	158	1.8	P212121	3E0H.pdb	3E0H-sf.cif	2KCU	2KCU.seq	monomer	2KCU.pdb	2KCU.mr	16097	16097.hmrh	16097.peaks	16097.fid	PEG Phage
CtR148A	3BW	3BW.seq	88 (2)	1.9	P43212	3BW.pdb	3BW-sf.cif	2KO1	2KO1.seq	dimer	2KO1.pdb	2KO1.mr	16486	16486.hmrh	16486.peaks	16486.fid	PAG PEG
DrR147D ^a	3GGN	3GGN.seq	155 (2)	2.0	P1211	3GGN.pdb	3GGN-sf.cif	2KCZ	2KCZ.seq	monomer	2KCZ.pdb	2KCZ.mr	16100	16100.hmrh	16100.peaks	16100.fid	NA
GmR137	3CW1	3CW1.seq	78	1.9	P43212	3CW1.pdb	3CW1-sf.cif	2KSP	2KSP.seq	monomer	2KSP.pdb	2KSP.mr	15844	15844.hmrh	15844.peaks	15844.fid	PEG Phage
HR1958	1TVG	1TVG.seq	153	1.6	C121	1TVG.pdb	1TVG-sf.cif	1XPW	1XPW.seq	monomer	1XPW.pdb	1XPW.mr	6344	6344.hmrh	6344.peaks	6344.fid	NA
HR3646E	3FA	3FA.seq	121	1.5	C121	3FA.pdb	3FA-sf.cif	2KHN	2KHN.seq	monomer	2KHN.pdb	2KHN.mr	16250	16250.hmrh	16250.peaks	16250.fid	PAG PEG
HR41	3EVX	3EVX.seq	175 (4)	2.5	P1	3EVX.pdb	3EVX-sf.cif	2K07	2K07.seq	monomer	2K07.pdb	2K07.mr	6546	6546.hmrh	6546.peaks	6546.fid	NA
MbR242E	3GW2	3GW2.seq	108	2.1	P6422	3GW2.pdb	3GW2-sf.cif	2KKO	2KKO.seq	dimer	2KKO.pdb	2KKO.mr	16368	16368.hmrh	16368.peaks	16368.fid	PAG PEG
MrR110B	3E0E	3E0E.seq	97	1.6	P212121	3E0E.pdb	3E0E-sf.cif	2K5V	2K5V.seq	monomer	2K5V.pdb	2K5V.mr	15849	15849.hmrh	15849.peaks	15849.fid	NA
OR8C	2RHK	2RHK.seq	140 (2), 72 (2)	2.0	P41	2RHK.pdb	2RHK-sf.cif	2KKZ	2KKZ.seq	monomer	2KKZ.pdb	2KKZ.mr	16376	16376.hmrh	16376.peaks	16376.fid	NA
PR193A	3DU	3DU.seq	127 (2)	1.7	P1211	3DU.pdb	3DU-sf.cif	2KL6	2KL6.seq	monomer	2KL6.pdb	2KL6.mr	16385	16385.hmrh	16385.peaks	16385.fid	Phage
PsR293	3H9X	3H9X.seq	125 (4)	2.5	P1	3H9X.pdb	3H9X-sf.cif	2KFP	2KFP.seq	monomer	2KFP.pdb	2KFP.mr	16186	16186.hmrh	16186.peaks	16186.fid	NA
SR213	2M8	2M8.seq	131 (2)	2.0	P212121	2M8.pdb	2M8-sf.cif	2HFI	2HFI.seq	monomer	2HFI.pdb	2HFI.mr	16113	16113.hmrh	16113.peaks	16113.fid	NA
SR384	3BHP	3BHP.seq	60 (3)	2.0	C121	3BHP.pdb	3BHP-sf.cif	2IVD	2IVD.seq	monomer	2IVD.pdb	2IVD.mr	15476	15476.hmrh	15476.peaks	15476.fid	NA
SR478	2GSV	2GSV.seq	80 (2)	1.9	P121	2GSV.pdb	2GSV-sf.cif	2IS1	2IS1.seq	dimer	2IS1.pdb	2IS1.mr	15350	15350.hmrh	15350.peaks	15350.fid	NA
SgR42	3C4S	3C4S.seq	66 (2)	1.7	P32	3C4S.pdb	3C4S-sf.cif	2JZ2	2JZ2.seq	monomer	2JZ2.pdb	2JZ2.mr	15604	15604.hmrh	15604.peaks	15604.fid	PEG
SoR77	2QTI	2QTI.seq	80	2.3	P43212	2QTI.pdb	2QTI-sf.cif	2IUW	2IUW.seq	dimer	2IUW.pdb	2IUW.mr	15456	15456.hmrh	15456.peaks	15456.fid	PAG
SsR10	2Q00	2Q00.seq	129 (2)	2.4	I4122	2Q00.pdb	2Q00-sf.cif	2JPU	2JPU.seq	monomer	2JPU.pdb	2JPU.mr	15265	15265.hmrh	15265.peaks	15265.fid	NA
StR65	2ES9	2ES9.seq	115	2.0	I213	2ES9.pdb	2ES9-sf.cif	2JN8	2JN8.seq	monomer	2JN8.pdb	2JN8.mr	15089	15089.hmrh	15089.peaks	15089.fid	NA
StR70	2ES7	2ES7.seq	142 (4)	2.8	P1211	2ES7.pdb	2ES7-sf.cif	2JZT	2JZT.seq	monomer	2JZT.pdb	2JZT.mr	7178	7178.hmrh	NA	NA	NA
XcR50	1TZ	1TZ.seq	87	2.1	P65	1TZ.pdb	1TZ-sf.cif	1XPV	1XPV.seq	monomer	1XPV.pdb	1XPV.mr	6363	6363.hmrh	6363.peaks	NA	NA
ZR18	2FFM	2FFM.seq	91	2.5	P41212	2FFM.pdb	2FFM-sf.cif	1POX	1POX.seq	monomer	1POX.pdb	1POX.mr	5844	5844.hmrh	NA	5844.fid	NA

^aif the NMR structure is not well defined/residue 24-69 out of 155 residues.
^bnr means data has been submitted to BMRB but has not been updated by far.

Figure S5: Targets marked MR in Table S1 have been taken from the website http://psvs-1_4-dev.nesg.org/MR/dataset.html. As more targets or missing data to already featured targets might be added later, we provide here a screenshot of the state of the website on January the 31st 2013.

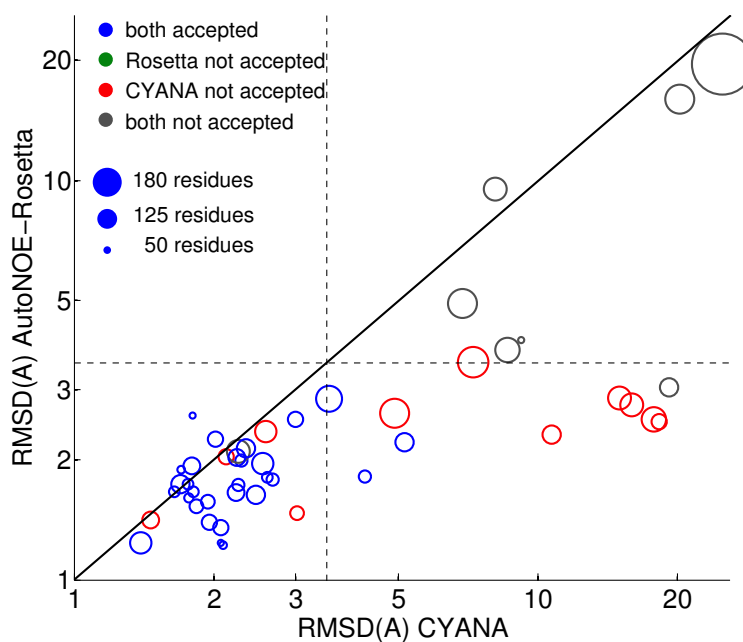


Figure S6: Median heavy-atom RMSDs (log-scale) of final models with respect to the reference structure. The diagonal line indicates points of equal performance, points above the line correspond to targets for which CYANA yields lower RMSDs, and points below the line correspond to targets for which AutoNOE-Rosetta yields lower RMSDs. The dashed

APPENDIX

lines mark 3.5Å RMSD. **The size** of the proteins is proportional to the area of the symbol as indicated by the legend. **The color** indicates whether for CYANA, AutoNOE-Rosetta or for both programs the final models are considered as success (*Appendix Methods A.3.2*). RMSDs are capped at 25 Å.

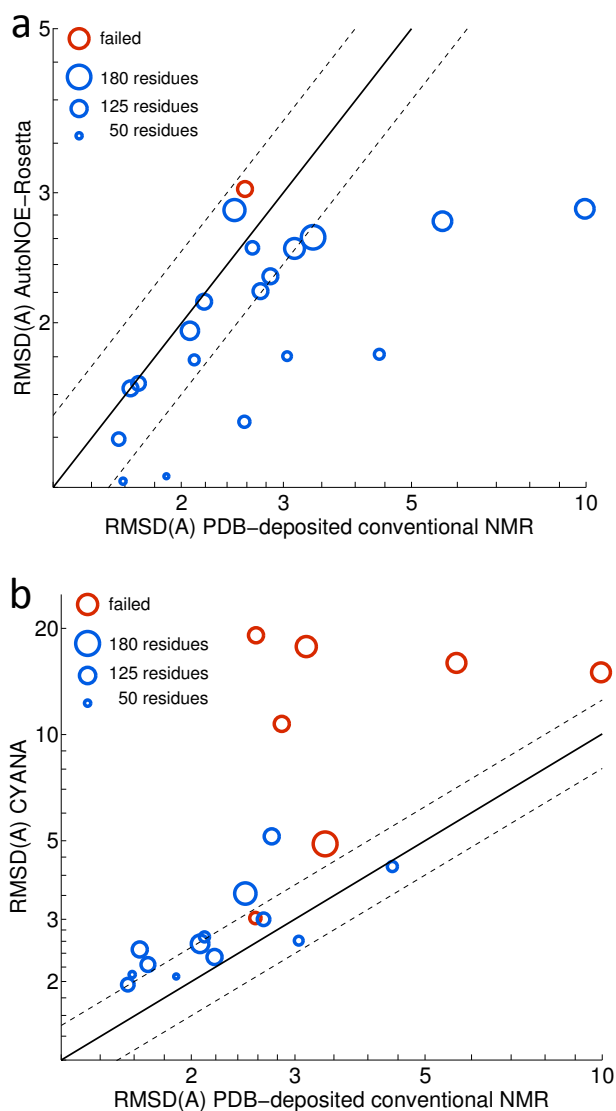


Figure S7: Comparison of AutoNOE-Rosetta with manually solved best-effort PDB-deposited NMR structures. Shown are the median heavy RMSDs of final models vs. the median C_{α} -RMSDs of all PDB deposited models computed against the reference structure. For all targets but AR3436a an Xray structure is used as reference; 2kj6/AR3436a is compared to a new manually refined NMR solution structure, which supersedes 2kj6. **The diagonal line** indicates points of equal performance, points above the line correspond to targets with higher accuracy of the PDB-deposited models, points below the line to targets with higher accuracy of the AutoNOE-Rosetta models. Dashed lines mark +/- 25% accuracy.

APPENDIX

The size of the proteins is proportional to the area of the symbol as indicated by the legend. **The color** indicates whether final AutoNOE-Rosetta models are considered as success based on intrinsic convergence criteria (>90% of residues converged).

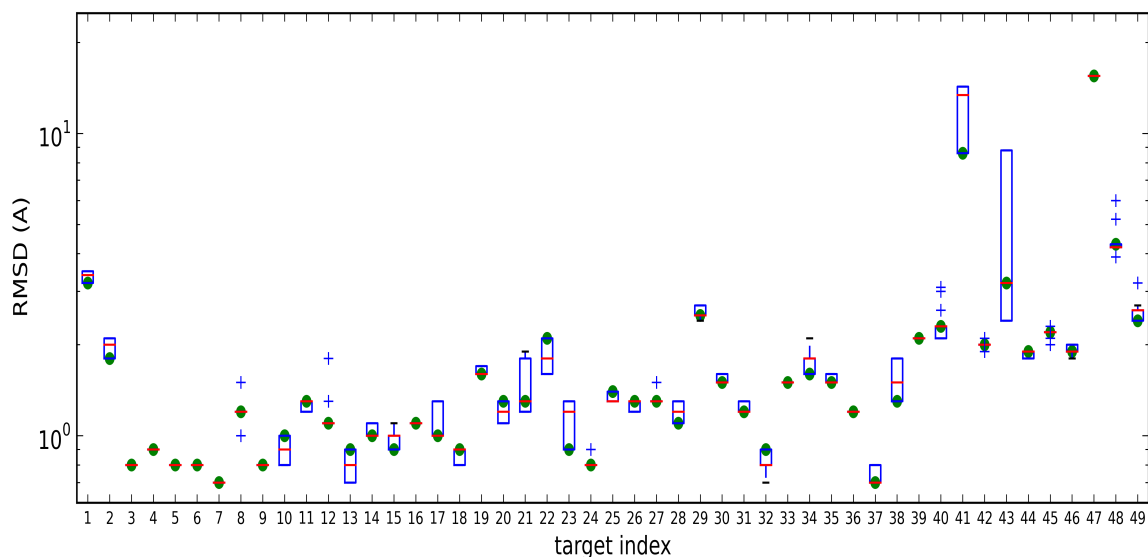


Figure S8: Error analysis of AutoNOE-Rosetta. We have carried out three runs of AutoNOE-Rosetta at each NOE restraint weight (1, 2, 5, 10, 25, 50). The green dots represent the reported RMSD values (main text) obtained by selecting from all generated runs together according to the described procedure. The boxplot shows the expected statistics of C_{α} -RMSD if only a single run per NOE restraint weight were carried out. To obtain these statistics, we randomly selected one of the 3 runs with identical parameter settings before applying the selection criterion. This procedure was repeated 1000 times with new random selections and the boxplot illustrates the obtained statistics. The box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend from the box to show the range of the data but do not extend more than half the box size. Any data point beyond whisker range is considered an outlier and shown as flier point. The post-selection relaxation protocol to reduce NOE violations (Suppl. Methods S1.11) is not applied here, and numerical values of the green-dots thus can differ from reported values in *Appendix Table S2*

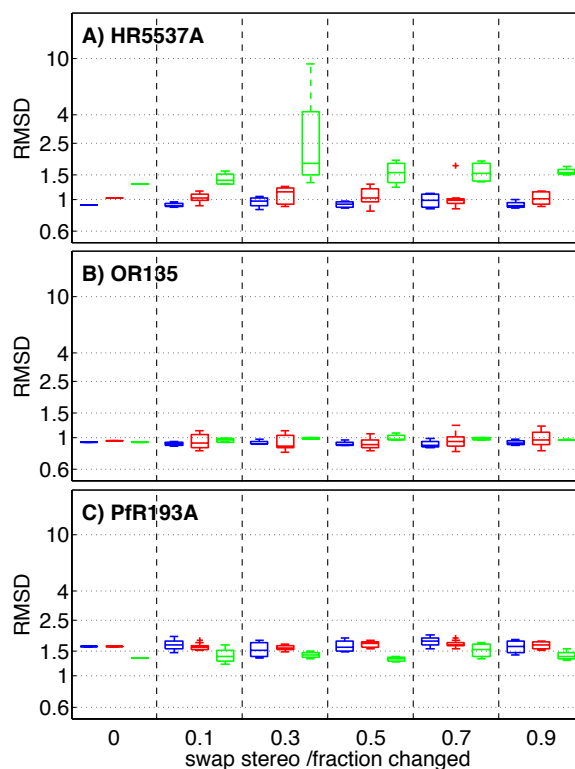


Figure S9: C α -RMSDs statistics of final structures generated with CYANA(red), ASDP(green) AutoNOE-Rosetta(blue) from swapping stereos(A-C). X-axis shows the percentage of swapped fraction. For each severity level, 6 independent runs for AutoNOE-Rosetta, CYANA and ASDP were performed. In addition to the four types of swapping errors discussed above, we also tested SWAP-STEREO, which swaps diastereotropic protons in the dataset. Except a single ASDP run at 30% severity, all calculations were robust against these errors. This reflects, that the programs are usually interpreting the input data of diastereotropic protons as ambiguous.

APPENDIX

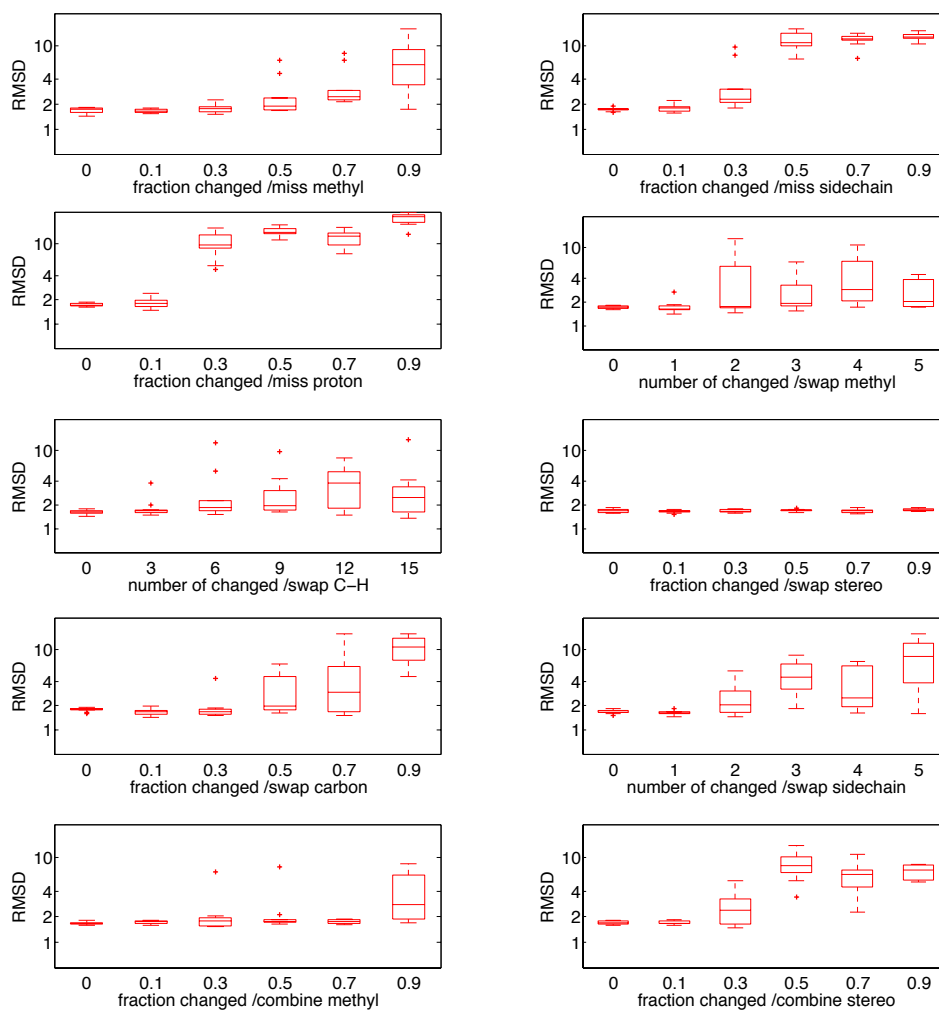


Figure S10: α -RMSDs statistics of final structures generated with CYANA for protein SR213.

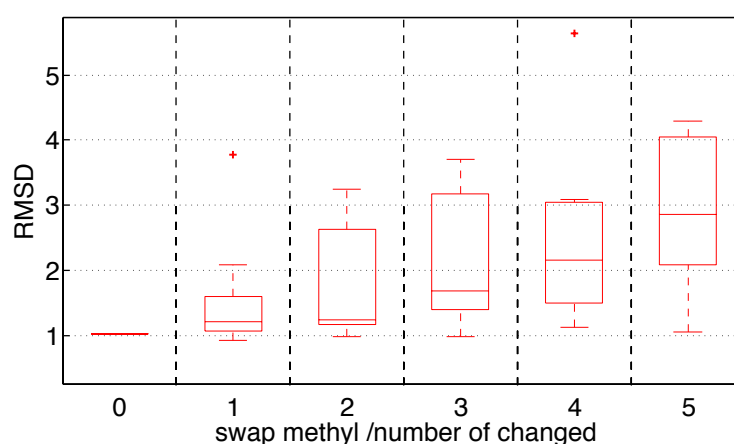


Figure S11: α -RMSDs statistics of final structures generated with CYANA for HR5537a from low severity levels of swapping methyls.

APPENDIX

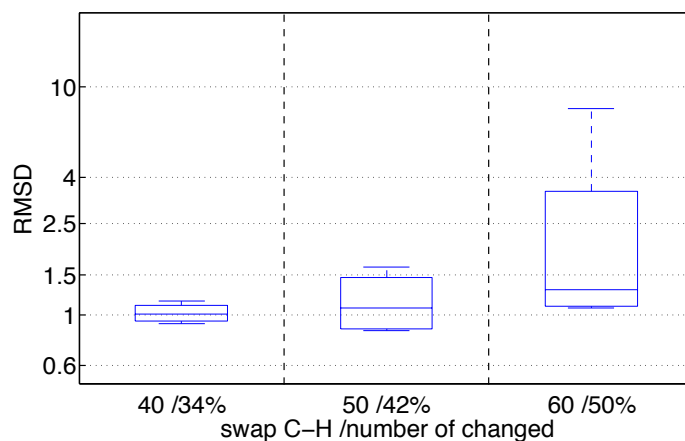


Figure S12: C α -RMSDs statistics of final structures generated with AutoNOE-Rosetta for HR5537a from high severity levels of swapping C-H.

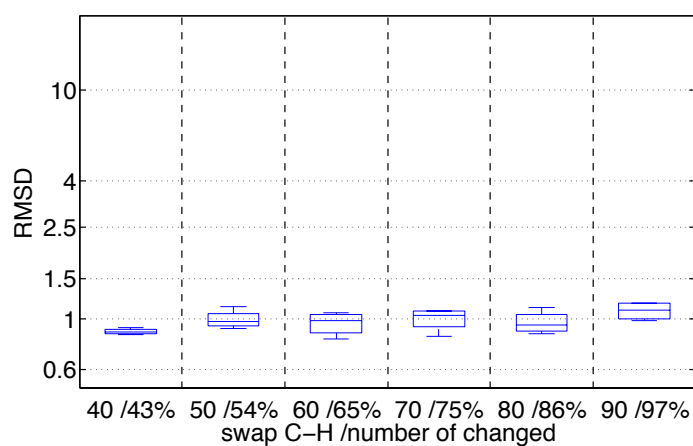


Figure S13: C α -RMSDs statistics of final structures generated with AutoNOE-Rosetta for OR135 from high severity levels of swapping C-H.

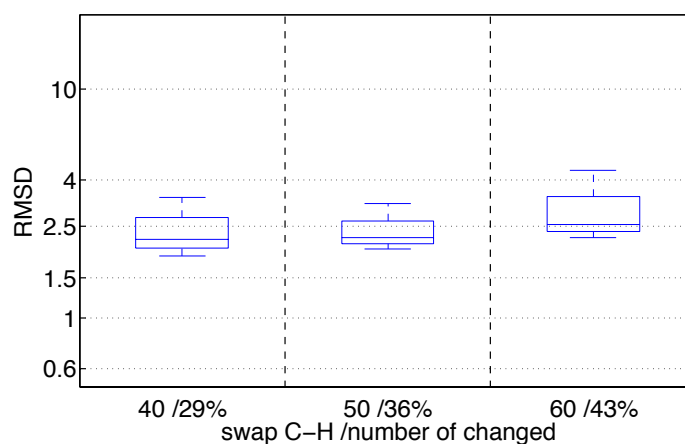


Figure S14: C α -RMSDs statistics of final structures generated with AutoNOE-Rosetta for PfR193A from high severity levels of swapping C-H.

APPENDIX

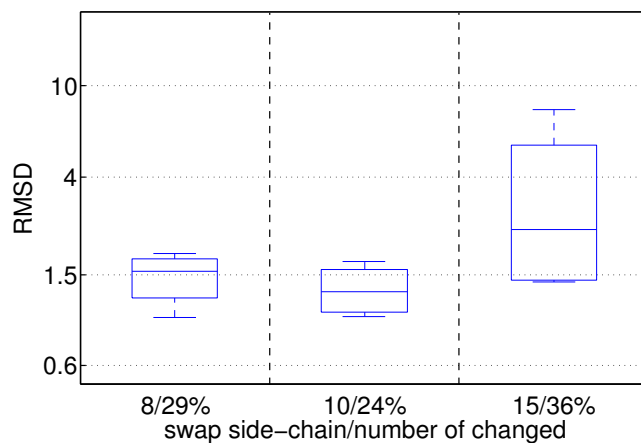


Figure S15: C α -RMSDs statistics of final structures generated with AutoNOE-Rosetta for HR5537A from high severity levels of swapping side-chain.

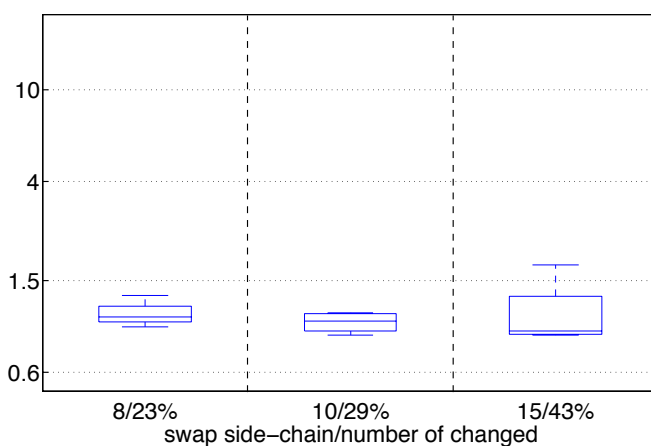


Figure S16: C α -RMSDs statistics of final structures generated with AutoNOE-Rosetta for OR135 from high severity levels of swapping side-chain.

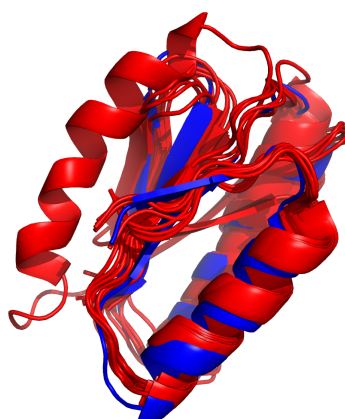


Figure S17: Final structures of OR135 with 90% missing protons at run 5. The blue structure is native OR135 and Red shows the calculated structures.

A.2 Supplementary Tables

No:	NESG-Code	NMR-PDB	Source ¹	mol. weight (kDa)	reference structure	peak-quality ²	RDC Data ³	chemical shift completeness ⁴
1	StT322	2loj	CASDII	7.1	NMR	raw	-	97.4%
2	StT322	2loj	CASDII	7.1	NMR	refined	-	97.4%
3	SR384	2jvd	MR	5.5	XRAY	-	-	98.7%
4	HR6470	2l9r	CASDII	8.4	NMR	raw	-	98.6%
5	HR6470	2l9r	CASDII	8.4	NMR	refined	-	98.6%
6	CtR69a	2kru	CASDI	7.4	NMR	-	-	97.1%
7	SgR42	2jz2	MR	7.7	XRAY	-	-	99.7%
8	GmR137	2k5p	MR	8.6	XRAY	-	PEG, Phage	98.6%
9	OR135	2ln3	CASDII	9.9	NMR	raw	PEG, Phage	99.8%
10	OR135	2ln3	CASDII	9.9	NMR	refined	PEG, Phage	99.8%
11	PgR122a	2kmm	CASDI	8.2	NMR	-	-	96.2%
12	XcR50	1xpv	MR	8.8	XRAY	-	-	97.3%
13	HR3646e	2khn	MR	13.7	XRAY	-	PAG,PEG	93.1%
14	AR3436a	2kj6	CASDI	10.9	NMR	raw ⁵	-	93.9%
15	AR3436a	2kj6	CASDI	10.9	NMR	refined ⁵	-	98.9%
16	HR6430	2la6	CASDII	11.1	NMR	raw	-	99.2%
17	HR6430	2la6	CASDII	11.1	NMR	refined	-	99.2%
18	MrR110B	2k5v	MR	10.9	XRAY	-	-	99.8%
19	HR2876	2ltm	CASDII	12.3	NMR	raw	PEG,Phage	99.4%
20	HR2876	2ltm	CASDII	12.3	NMR	refined	PEG,Phage	99.4%
21	NeR103a	2kpm	CASDI	12.0	NMR	-	-	92.5%
22	StR65	2jn8	MR	12.2	XRAY	-	-	98.1%
23	VpR247	2kif	CASDI	11.6	NMR	-	-	99.2%
24	HR5537	2kk1	CASDI	14.7	NMR	-	-	98.2%
25	YR313	2ltl	CASDII	13.7	NMR	raw	PEG,Phage	98.7%
26	YR313	2ltl	CASDII	13.7	NMR	refined	PEG,Phage	98.7%
27	ET109	2kky	CASDI	11.4	NMR	-	-	98.6%
28	PfR193	2kl6	MR	12.1	XRAY	-	-	99.3%
29	CcR55	2jqn	MR	12.8	XRAY	-	-	98.4%
30	SR213	2hfi	MR	14.6	XRAY	-	-	93.1%
31	OR8C	2kkz	MR	15.2	XRAY	-	-	99.1%
32	AtT13	2knr	CASDI	13.4	NMR	-	-	98.8%
33	SsR10	2jpu	MR	15.2	XRAY	-	-	98.8%

APPENDIX

34	PsR293	2kfp	MR	14.9	XRAY	-	-	95.9%
35	OR36	2lci	CASDII	16.1	NMR	raw	PEG,Phage	97.0%
36	OR36	2lci	CASDII	16.1	NMR	refined	PEG,Phage	97.0%
37	CgR26a	2kpt	CASDI	16.0	NMR	-	-	99.7%
38	HR1958	1xpw	MR	16.3	XRAY	-	-	97.9%
39	SR10	2kzn	ILV	17.3	XRAY	unrefined	PAG,PEG, Phage	93.9%
40	DrR1470	2kcz	MR	17.5	XRAY	-	-	84.8%
41	HR4660 B	2lmd	ILV	20.9	NMR	unrefined	PAG	91.8%
42	CtR107	2kcu	MR	18.3	XRAY	-	PEG,Phage	93.1%
43	HR5460	2lah	CASDII	19.3	NMR	-	Phage	98.3%
44	HR5460	2lah	CASDII	19.3	NMR	-	Phage	98.3%
45	HR41	2k07	MR	20.7	XRAY	-	-	96.1%
46	SgR145	2kw5	ILV	22.6	XRAY	unrefined	PEG, Phage	90.7%
47	HsR50	2lok	ILV	20.9	NMR	unrefined	PEG, Phage	81.7%
48	HmR11	2lnu	ILV	21.6	NMR	unrefined	PEG, Phage	95.3%
49	WR73	2loy	ILV	21.0	NMR	unrefined	Phage	97.2%
50	ER690	1ezp	ILV	41.1	XRAY	unrefined	Phage	91.7%
targets not used ⁶				reason to ignore target				
	BeR31		MR	no peak lists available				
	CsR4		MR	dimer				
	CtR148A		MR	dimer				
	MbR242E		MR	dimer				
	SR478		MR	dimer				
	SoR77		MR	dimer				
	StR70		MR	no peak lists available				
ZR18		MR	no peak lists available					

Table S1: List of data sets used to benchmark AutoNOE-Rosetta.

Footnotes:

- 1) The data was taken from published lists (1+2) and our own work (3). 1) CASDI, CASDII: <http://www.wenmr.eu/wenmr/casd-nmr-data-sets> (Rosato et al. 2012), 2) MR: http://psvs-1_4-dev.nesg.org/MR/dataset.html (Mao et al. 2011), 3) ILV(Lange et al. 2012). All data sets available by January 31st 2013 at these sources were considered.

APPENDIX

- 2) For CASDII targets *raw* and *refined* data sets are available: *raw* data sets comprise automatically picked peak-lists. ILV data sets are *unrefined*: peaks have been picked manually and chemical shift assignments have been validated, but the peak-lists and chemical shift assignments have not yet undergone iterative refinement using structural models. In all other cases the quality of the peak-lists is unspecified and is denoted as '-' in the table.
- 3) RDC data were in the calculations. Some of the RDC sets linked on the MR website were not downloadable or not assigned; some RDC sets could be retrieved directly from the responsible researchers, or from the BMRB.
- 4) Chemical Shift completeness is computed by CYANA for non-ILV targets. For ILV-targets we expect shifts for backbone atoms C, CA, CB, N, H and both δ (γ) methyls for Leucine (Valine), and the δ 1-methyl for Isoleucine.
- 5) This data set was manually re-analyzed by us and then by the original authors of the structure. Both independent re-analyses led to consistent corrections in the sidechain chemical shift assignments and a new manually picked peak-list. The original data set is denoted as *raw*, the new data set as *refined*.
- 6) Data sets that are dimers were ignored for the current study. For three of the MR data sets no peak-list was available.

APPENDIX

No:	Target	data type ²	Reference PDB-Id	Size	Residue ranges			C _α -RMSD (Å) to reference structure ¹		
					of reference	AutoNO E-Rosetta	RMSD ³	CYAN A	PDB	AutoNO E-Rosetta
1	StT322	raw	2loj	38	1-63	26-63	26-63	8.3	-	3.2
2	StT322	-	2loj	38	1-63	26-63	26-63	1.4	-	1.7
3	SR384	-	3bhp	39	1-48	1-39	1-39	1.2	1.2	0.8
4	HR6470	raw	2l9r	48	1-69	11-58	11-58	0.8	-	0.9
5	HR6470	-	2l9r	48	1-69	11-58	11-58	0.8	-	0.8
6	CtR69a	-	2kru	57	1-63	1-57	4-52	0.8	-	0.8
7	SgR42	-	3c4s	54	1-66	1-54	1-54	1.4	0.9	0.6
8	GmR137	-	3cwi	66	1-78	1-66	1-66	1.9	2.4	1.3
9	OR135	raw	2ln3	69	1-79	5-73	5-73	0.9	-	0.8
10	OR135	-	2ln3	69	1-79	5-73	5-73	1.1	-	0.9
11	PgR122a	-	2kmm	73	1-73	1-73	1-64	1.5	-	1.3
12	XcR50	-	1ttz	73	1-78	2-74	2-74	1.8	1.2	1.1
13	HR3646e	-	3fia	89	1-121	24-112	24-99	2.4	1.7	0.9
14	AR3436a	raw	2kj6	80	1-97	14-93	14-93	3.4	3.9	1.0
15	AR3436a	-	2kj6	80	1-97	14-93	14-93	1.7	-	1.0
16	HR6430	raw	2la6	89	1-99	11-99	11-99	1.4	-	1.0
17	HR6430	-	2la6	89	1-99	11-99	11-99	1.5	-	0.9
18	MrR110B	-	3e0e	92	1-98	1-92	1-29, 36-78, 83-90	1.3	0.8	0.8
19	HR2876	raw	2ltm	95	1-107	13-107	13-107	(+)	-	1.6
20	HR2876	-	2ltm	95	1-107	13-107	13-107	1.4	-	1.5
21	NeR103a	-	2kpm	99	1-105	1-99	23-82,90-96	1.5	-	1.3
22	StR65	-	2es9	99	1-109	10-108	10-108	2.4	1.9	2.0
23	VpR247	-	2kif	99	1-102	1-99	1-99	1.5	-	0.8
24	HR5537	-	2kk1	101	1-135	35-135	39-104,118-134	1.0	-	0.8
25	YR313	raw	2ltl	102	1-119	18-119	18-40, 46-115	1.6	-	1.4
26	YR313	-	2ltl	102	1-119	18-119	18-40, 46-115	2.1	-	1.3
27	ET109	-	2kky	102	1-102	1-102	2-101	1.3	-	1.4
28	PfR193	-	3idu	106	1-116	11-118	11-118	1.5	1.1	1.0
29	CcR55	-	2o0q	112	1-116	2-113	2-113	(+)	1.9	2.5

APPENDIX

30	SR213	-	2im8	113	1-123	7-119	7-119	1.6	1.2	1.5
31	OR8C	-	2rhk	116	1-134	5-120	5-120	1.8	1	1.1
32	AtT13	-	2knr	118	1-118	1-118	2-53,70-114	0.9	-	0.8
33	SsR10	-	2q00	118	1-129	6-123	6-55,58-123	4.5	1.9	1.4
34	PsR293	-	3h9x	117	1-117	1-117	1-117	9.9	1.7	1.6
35	OR36	raw	2lci	128	1-134	1-128	1-128	2.0	-	1.5
36	OR36	-	2lci	128	1-134	1-128	1-128	(*)	-	1.2
37	CgR26a	-	2kpt	131	1-148	1-131	15-130	1.0	-	0.7
38	HR1958	-	1tvq	133	1-143	5-137	5-137	1.9	1.3	1.3
39	SR10	unr	3e0o	141	1-147	1-141	1-141	(+)	4.7	2.1
40	DrR1470	-	3ggn	142	1-155	3-144	3-110,127-144	14.5	9.6	2.4
41	HR4660B	unr	2lmd	144	1-174	31-174	39-106, 114-136, 142-154	(+)	-	(+)
42	CtR107	-	3e0h	147	1-166	9-155	9-155	(+)	2.6	2.0
43	HR5460	-	2lah	150	1-160	11-160	19-160	8.0	-	3.2
44	HR5460	-	2lah	150	1-160	11-160	19-160	(*)	-	1.8
45	HR41	-	3evx	158	1-175	5-162	5-162	2.8	1.7	2.0
46	SgR145	unr	3mer	177	1-202	21-197	21-173,184-197	4.3	2.7	1.9
47	HsR50	unr	2lok	177	1-191	11-188	11-23,55-157,167-181	(+)	-	(+)
48	HmR11	unr	2lnu	181	1-185	1-181	4-180	6.1	-	4.3
49	WR73	unr	2loy	183	1-183	1-183	1-37,66-180	6.4	-	2.4
50	ER690	unr	1dmb	366	1-370	5-370	5-370	(+)	3.5	(+)

Table S2: Complete list of C_{α} -RMSDs against reference structure for CYANA, best-effort PDB-deposited NMR structure and AutoNOE-Rosetta. This table is reproduced in parts by main-text Table 3.1.

Footnotes:

- 1) In these columns, the symbols '(+)' and '(*)' mark unconverged (<60% residues converged) and crashed structure calculations, respectively. We assume that the two crashed CYANA calculations would yield accurate models, if succeeded. Calculations that failed the final acceptance criteria (*Appendix A.3.2*) are shown in red.
- 2) Data quality (*raw*, *unrefined* or *refined*) is abbreviated as *raw*, *unr* and -, respectively.
- 3) Residues used for RMSD calculation. Tails or loops that are not well defined in the reference structures were removed. For removed loops explicit justifications are given in Table S9.

APPENDIX

Target	PDB accession codes		DP-Scores		
	NMR	X-ray	NMR	Xray	AutoNOE
SR384	2jvd	3bhp	0.72	0.74	0.75
SgR42	2jz2	3c4s	0.59	0.64	0.62
GmR137	2k5p	3cwi	0.62	0.67	0.64
XcR50	1xpv	1ttz	0.67	0.72	0.66
HR3646e	2khn	3fia	0.43	0.52	0.46
MrR110B	2k5v	3e0e	0.71	n/a ¹	0.73
StR65	2jn8	2es9	0.64	0.56	0.69
PfR193	2kl6	3idu	0.82	0.81	0.80
PsR293	2kfp	3h9x	0.62	0.61	0.58
CcR55	2jqn	2o0q	0.76	0.79	0.72
SR213	2hfi	2im8	0.58	0.62	0.57
OR8C	2kkz	2rhk	0.61	0.63	0.62
SsR10	2jpu	2q00	0.46	n/a ¹	0.49
HR1958	1xpw	1tvq	0.63	0.68	0.63
SR10	2kzn	3e0o	0.59	0.55	0.59
DrR1470	2kcz	3ggn	0.36	n/a ¹	0.37
CtR107	2kcu	3e0h	0.26	0.36	0.26
HR41	2k07	3evx	0.68	0.71	0.61
SgR145	2kw5	3mer	0.36	n/a ¹	0.36
ER690	1ezp	1dmb	0.29	0.36	0.22

Table S3: The AutoStruct DP-Score(Huang et al. 2005) has been computed on PDB-deposited and AutoNOE-Rosetta structures. All models were trimmed to match the sequences, used in AutoNOE-Rosetta (*Appendix Table S1*).

Footnotes:

- 1) missing density in X-ray structure. DP-score not calculated.

APPENDIX

Target	molecular weight (kDa)	residue ranges		PDB accession codes		C_{α} -RMSD to Xray structure (Å)	
		modelled ¹	RMSD analysis	conventional NMR structure	Xray reference structure	AutoNOE	conventional NMR structure
SR384	5.5	1-39	1-39	2jvd	3bhp	0.8	1.2
SgR42	7.7	1-54	1-54	2jz2	3c4s	0.6	0.9
GmR137	8.6	1-66	1-66	2k5p	3cwi	1.3	2.4
XcR50	8.8	2-74	2-74	1xpv	1ttz	1.1	1.2
HR3646E	13.7	24-99	24-99	2khn	3fia	0.9	1.7
AR3436A	10.9	14-93	14-93	2kj6	n/a ⁷	1.0	3.9
MrR110B	10.9	1-92	1-29, 36-78, 83-90 ²	2k5v	3e0e	0.8	0.8
StR65	12.2	10-108	10-108	2jn8	2es9	2.0	1.9
PfR193	12.1	11-118	11-116 ³	2kl6	3idu	1.1	1.1
PsR293	14.9	1-117	1-117	3h9x	2kfp	1.6	1.75
CcR55	12.8	2-113	2-113	2jqn	2o0p	2.5 ⁸	1.9
SR213	14.6	7-119	7-119	2hfi	2im8	1.5	1.2
OR8C	15.2	5-120	5-120	2kkz	2rhk	1.1	1
SSR10	15.2	6-123	6-55, 58-123 ⁴	2jpu	2q00	1.4	1.9
HR1958	16.3	5-137	5-137	1xpw	1tvq	1.3	1.3
SR10	17.3	1-141	1-141	2kzn	3e0o	2.1	4.7 ⁹
DrR1470	17.5	3-144	3-110, 127-144 ⁵	2kcz	3ggn	2.4	9.6 ¹⁰
CtR107	18.3	9-155	9-155	2kcu	3e0h	2.0	2.6
HR41	20.7	5-162	5-162	2k07	3evx	2.0	1.7
SgR145	22.6	21-197	21-173, 184-197 ⁶	2kw5	3mer	1.9	2.7
ER690	41.1	5-370	5-370	1ezp ¹¹	1dmb	19.2 ¹¹	3.5 ¹¹

Table S4: Accuracy of AutoNOE structures vs. PDB-deposited NMR structures

Footnotes:

- 1) target sequences are trimmed automatically to remove flexible tails based on TALOS+ RCI S² prediction. All numbering is relative to the respective NMR sample.
- 2) residues 31-34 and 80-81, have missing density in Xray reference structure
- 3) residues 117-118 have missing density in Xray reference structure
- 4) residues 55-58 are unconverged in conventional NMR structure. This is consistent with high flexibility predicted by TALOS+

APPENDIX

- 5) residues 115-129 have missing density in Xray reference structure
- 6) residues 174-183 have missing density in Xray reference structure
- 7) An independent expert analysis of the raw NMR data yielded a more accurate NMR reference structure for this target, which supersedes 2kj6.
- 8) The AutoNOE-Rosetta calculation did not fully converge. The accuracy of the converged residues (88% of total length) is 1.3Å.
- 9) The pdb-deposited NMR ensemble is of low precision. Nevertheless, in at least two regions the Xray reference is not within the structural bundle of the ensemble. These regions have high (>0.7) TALOS+ calculated RCI S^2 values, suggesting that they are well-structured in solution.
- 10) Major parts of the pdb-deposited NMR models are not well converged, giving rise to the high RMSD.
- 11) The AutoNOE-Rosetta structure calculation did not converge. The pdb-deposited NMR ensemble has been obtained from different NMR data (Mueller et al. 2000). No NMR solution structure was deposited for the ER690 data set.

APPENDIX

No:	Target	modeled residues ¹	residues used for RMSD	Reason for excluding internal regions from RMSD calculations
17	MrR110B	1-92	1-29, 36-78, 83-90	residues 31-34 and 80-81, have missing density in Xray reference structure
20	NeR103a	1-99	23-82,90-96	residues 83-89 have RCI-S2 of less than 0.7 according to TALOS+. ²
23	HR5537	35-135	39-104,118-134	residues 105-116 have RCI-S2 of less than 0.7 according to TALOS+. ³
24	YR313	18-119	18-40, 46-115	residues 41-45 are heterogeneous in NMR reference ensemble
25	YR313	18-119	18-40, 46-115	residues 41-45 are heterogeneous in NMR reference ensemble
31	AtT13	1-118	2-53,70-114	residues 54-69 are heterogeneous in NMR reference ensemble
32	SsR10	6-123	6-55,58-123	residues 56-57, missing density in the Xray reference structure
39	DrR1470	3-144	3-110,127-144	residues 111-126, missing density in the Xray reference structure
40	HR4660B	31-174	39-106, 114-136, 142-154	residues 107-113 and 137-141 are heterogeneous in NMR reference ensemble
45	SgR145	21-197	21-173,184-197	residues 175-183, missing density in the Xray reference structure
46	HsR50	11-188	11-23,55-157,167-181	residues 24-53, and 158-166 are heterogeneous in NMR reference ensemble
48	WR73	1-183	1-37,66-180	residues 38-65 are heterogeneous in NMR reference ensemble

Table S5: Justification for omitting internal residues from RMSD calculation: In some cases internal regions had to be excluded from RMSD calculations due to missing density in Xray structures or flexible regions in NMR structures. Justifications for individual cases are given in this table.

Footnotes:

- 1) Residues modeled in the CYANA and AutoNOE-Rosetta calculations. The numbering is relative to the corresponding NMR PDB structure (column 6 in Table S2).
- 2) Heterogeneity in NMR ensemble is moderate. Including the loop in RMSD calculation, increases the C_{α} -RMSD by 0.2 Å for both, AutoNOE and CYANA final models.

APPENDIX

- 3) Heterogeneity in NMR ensemble is moderate. Including the loop in RMSD calculation, increases the C_{α} -RMSD by 0.6 Å for both, AutoNOE and CYANA.

APPENDIX

Target	peak-list	dimension	sweep window (ppm)	
			start	end
AR3436 (raw)	CHH-ali	^{13}C	30.0	54.0
HR1958	CCHH	$^{13}\text{C}-1$	13.6	34.5
HR1958	CCHH	$^{13}\text{C}-2$	13.6	55.4
HR1958	CHH-ali	^{13}C	23.1	47.0

Table S6: A small subset of peak-lists contained aliased frequencies. The parameters to set the *folding window* are given in columns *start* and *end*. To set the folding window in AutoNOE-Rosetta the line

```
#FOLD 1 30.0 54.0
```

is added to the respective peak-file, where 1 would be the index of the dimension that is folded.

APPENDIX

No:	Target	peaks available	diagonal and zero intensity	un-assigned	long-range	violated	prec. (Å) ¹	conv'd ²	target ³	rmsd (Å)
1	StT322	12437	266	9683	376	2261	0.6	1.00	1599.7	3.2
2	StT322	2727	215	953	194	155	0.7	1.00	9	1.8
3	SR384	2626	225	1075	48	82	0.4	1.00	9	0.8
4	HR6470	4262	429	802	121	240	0.5	1.00	8.9	0.9
5	HR6470	4262	490	700	147	214	0.6	1.00	19.1	0.8
6	CtR69a	1975	0	210	159	126	1.3	1.00	4.4	0.8
7	SgR42	1658	0	327	152	106	0.7	1.00	18.9	0.7
8	GmR137	2604	218	394	147	167	0.7	1.00	3.9	1.2
9	OR135	7749	357	3508	401	629	0.5	1.00	76.5	0.8
10	OR135	6359	578	1283	467	317	0.4	1.00	12.1	1.0
11	PgR122a	3515	232	742	286	229	2.3	1.00	10.9	1.3
12	XcR50	4156	499	636	277	228	0.8	1.00	27.7	1.1
13	HR3646e	6372	2126	2067	183	232	0.8	1.00	46.6	0.9
14	AR3436a	2076	182	223	109	147	0.6	1.00	36.5	1.0
15	AR3436a	2453	8	305	243	171	0.7	1.00	17.4	0.9
16	HR6430	6825	627	1023	758	327	0.9	1.00	13.7	1.1
17	HR6430	6643	621	967	775	271	0.8	1.00	12.6	1.0
18	MrR110B	4270	478	722	299	186	0.7	1.00	8.3	0.9
19	HR2876	14102	341	13530	6	179	1.4	1.00	46.5	1.6
20	HR2876	7054	678	1265	708	299	0.6	1.00	11.3	1.3
21	NeR103a	4658	346	809	422	368	2.4	1.00	53.5	1.3
22	StR65	3428	114	1764	61	108	1.3	1.00	3.8	2.1
23	VpR247	5756	580	925	436	321	0.8	1.00	19	0.9
24	HR5537	13995	1462	4755	773	855	1.5	1.00	32.7	0.8
25	YR313	12303	330	8518	394	2082	1.4	1.00	1658.7	1.4

APPENDIX

26	YR313	6592	776	1508	393	273	0.9	1.00	8.8	1.3
27	ET109	6752	79	1103	761	439	0.7	1.00	21.6	1.3
28	PfR193	6191	603	837	726	364	0.9	1.00	29.9	1.1
29	CcR55	3756	267	1703	124	156	2.2	0.88	9.6	2.5
30	SR213	5980	1065	1439	264	310	1.6	0.99	8.9	1.5
31	OR8C	6092	652	1444	325	325	1.0	1.00	22.2	1.2
32	AtT13	8036	670	1643	837	548	2.0	1.00	27.4	0.9
33	SsR10	5131	715	1498	121	253	1.4	1.00	9.3	1.5
34	PsR293	6870	518	2694	469	486	1.4	1.00	39.9	1.6
35	OR36	13794	338	9517	479	1718	1.1	1.00	258.9	1.5
36	OR36	9459	1036	3980	367	435	1.0	1.00	9.8	1.2
37	CgR26a	5131	3	896	539	301	1.6	1.00	11.3	0.7
38	HR1958	8712	503	3514	836	660	1.2	0.97	123.7	1.3
39	SR10	2100	330	315	382	119	1.7	0.98	13.6	2.1
40	DrR1470	8579	4888	2053	169	313	2.5	0.92	167.8	2.3
41	HR4660B	9200	573	5184	170	2524	17.2	0.14	2000	8.6
42	CtR107	7180	3475	931	243	338	1.8	1.00	180.8	2.0
43	HR5460	17250	248	14205	329	2568	7.7	0.79	3092	3.2
44	HR5460	12015	1073	4689	685	881	1.1	1.00	21.6	1.9
45	HR41	10879	1507	3456	669	696	1.8	0.95	36.6	2.2
46	SgR145	4302	453	974	847	415	2	0.98	345.7	2.0
47	HsR50	5058	546	3253	160	1184	3.9	0.51	496.5	15.5
48	HmR11	10348	708	3433	1016	855	3.3	0.65	601.2	4.3
49	WR73	2740	379	349	411	188	2.9	1.00	45.1	2.4
50	ER690	7153	780	3420	346	2262	9	0.19	1466.8	19.2

Table S7: Assignment Statistics for AutoNOE-Rosetta.

Footnotes:

- 1) precision calculated with ROSETTA as described in Methods. Runs that failed the convergence criterion of 90% are marked in red.

APPENDIX

- 2) fraction of converged residues: The converged residues are determined as described in *Appendix* Methods A.3.1.8. The number of converged residues is divided by the total number of non-flexible residues. A residue is considered as flexibility if TALOS+ predicted RCI $S_2 < 0.7$. If the fraction of drops below 90% the calculation is considered a failure. Accuracy of the converged part is not guaranteed anymore.
- 3) target function: The target function value is computed as described in *Appendix* Methods A.3.2.2. Values above 500 are considered dangerous and the computational results might not be accurate.

APPENDIX

No:	Target	peaks available	un-assigned	long-range	target-function,		preciscion (Å)			fraction converged	rms d (Å)
					cycle 1	cycle 7	cyana		our ¹		
							cycle 1	cycle 7			
1	StT322	12437	10056	356	5483.4	110.9	3.3	0.01	0	1.00	8.3
2	StT322	2727	1010	355	62.1	3.8	0.6	0.09	0.1	1.00	1.4
3	SR384	2626	854	188	0.1	0.0	0.6	0.33	0.5	1.00	1.2
4	HR6470	4262	665	293	14.4	0.8	0.3	0.23	0.3	1.00	0.8
5	HR6470	4262	494	379	10.7	1.1	0.3	0.16	0.2	1.00	0.8
6	CtR69a	1975	80	253	1.2	0.3	1.3	1.21	1.9	1.00	0.8
7	SgR42	1658	245	271	4.7	0.1	0.6	0.35	0.5	1.00	1.4
8	GmR137	2604	264	290	5.4	0.2	1	0.39	0.6	1.00	1.9
9	OR135	7749	3148	690	130.9	3.1	0.8	0.07	0.1	1.00	0.9
10	OR135	6359	527	994	51.1	10.2	0.5	0.06	0.1	1.00	1.1
11	PgR122a	3515	250	634	28.7	0.8	3.2	4.24	6.5	1.00	1.5
12	XcR50	4156	507	419	43	11.2	1.1	0.37	0.6	1.00	1.8
13	HR3646e	12744	9037	261	173.7	4.5	2.1	0.62	0.9	1.00	2.4
14	AR3436a	6282	4425	137	4.9	0.0	2.8	1.2	1.8	0.96	3.7
15	AR3436a	2453	195	416	8	0.1	0.9	0.45	0.7	1.00	1.7
16	HR6430	6825	302	1341	70.6	14.1	0.6	0.11	0.1	1.00	1.4
17	HR6430	6643	184	1314	35.4	14.6	0.5	0.15	0.2	1.00	1.6
18	MrR110B	4270	340	610	3	0.4	1	0.42	0.6	1.00	1.3
19	HR2876	14102	13473	3	2.6	0.8	10.2	12.14	29.9	0.22	17.9
20	HR2876	7054	319	1247	41.2	26.0	0.4	0.1	0.1	1.00	1.4
21	NeR103a	4658	561	697	50.3	1.6	8.2	6.23	9.9	1.00	1.5
22	StR65	3428	1658	168	2.4	0.3	2.6	1.08	1.6	1.00	2.4
23	VpR247	5756	281	804	17.3	2.1	0.7	0.31	0.5	1.00	1.5
24	HR5537	13995	2256	940	398.1	10.9	1.5	0.07	0.1	1.00	1.0

APPENDIX

25	YR313	12303	7977	701	2027.6	6.3	3.8	0.36	0.5	1.00	1.6
26	YR313	6592	586	731	14.8	10.3	1.6	0.49	0.8	1.00	1.7
27	ET109	6752	650	1418	227.8	45.9	0.7	0.36	0.5	1.00	1.3
28	PfR193	6191	289	1249	55.2	10.5	0.7	0.15	0.2	1.00	1.5
29	CcR55	3756	1681	90	5.5	0.6	8.8	9.54	25.3	0.18	19.3
30	SR213	5980	571	520	58.2	3.7	1.5	0.74	1.2	1.00	1.6
31	OR8C	6092	735	697	64.4	1.1	1.6	0.4	0.6	1.00	1.8
32	AtT13	8036	933	1472	236.8	66.3	1.2	0.28	0.4	1.00	0.9
33	SsR10	5131	326	381	9	1.5	1.3	0.79	1.2	1.00	4.5
34	PsR293	6870	2415	529	211.8	5.9	3.4	0.62	0.9	1.00	9.9
35	OR36	13794	7830	1198	886.7	7.9	1.6	0.44	0.7	1.00	2.0
36	OR36 ²	9459									
37	CgR26a	5131	88	1103	5	1.2	3	2.55	4.2	1.00	1.0
38	HR1958	8712	15527	989	137.3	8.0	0.8	0.33	0.5	1.00	1.9
39	SR10	2100	590	236	197.7	11.4	9	8.46	21.2	0.27	15.5
40	DrR1470	17158	13190	109	1451.4	12.4	7.7	2.31	3.7	0.61	14.5
41	HR4660B	9200	7681	242	249.4	7.1	8.9	5.6	10.1	0.45	7.2
42	CtR107	7180	1864	143	659.8	6.5	8.8	8.35	20.9	0.47	17.1
43	HR5460	17250	12675	693	4090.7	80.2	3.8	0.33	0.5	1.00	8.0
44	HR5460 ²	12015									
45	HR41	10879	1164	1129	168.1	10.2	1.5	0.87	1.3	0.99	2.8
46	SgR145	4302	1637	430	1229.7	12.9	8.2	1.76	2.8	0.93	4.3
47	HsR50	5058	3781	119	150.8	3.3	12.8	12.74	33.7	0.14	19.8
48	HmR11	10348	2397	1344	704.1	12.7	4	0.54	0.8	1.00	6.1
49	WR73	2740	536	335	191.2	5.3	7.8	3.52	5.7	0.87	6.4
50	ER690	7153	6377	2	108	4.8	29.9	51.01	0	0.00	106

Table S8: Assignment Statistics for CYANA. The assignment statistics are computed with the command *cyanatable* in the CYANA run directory.

Footnotes:

APPENDIX

- 1) precision calculated with ROSETTA as described in Methods, comparable with column 7 in *Appendix Table S7*
- 2) CYANA calculations crashed on this data set.

APPENDIX

		RMSD (A)					
Severity Level		0	1	2	3	4	5
		Repetition ID					
miss_methyl	1	0.92	0.97	0.84	1.01	1.14	0.93
	2	0.92	0.92	0.93	0.86	0.94	1.27
	3	0.92	1.02	0.97	0.82	0.94	0.93
	4	0.92	0.99	0.93	0.88	0.93	1.18
	5	0.92	0.93	0.95	0.98	0.78	0.96
	6	0.92	0.93	0.97	0.93	0.91	1.67
miss_sidechain	1	0.92	1.11	1.78	2.57	10.47	11.22
	2	0.92	0.88	1.03	8.35	10.19	12.13
	3	0.92	0.93	1.70	7.05	9.09	12.65
	4	0.92	1.08	0.79	2.75	10.09	11.84
	5	0.92	1.04	0.95	3.07	10.95	12.07
	6	0.92	0.96	1.96	2.49	10.88	12.40
miss_proton	1	0.92	1.00	1.03	4.75	10.81	13.09
	2	0.92	0.86	0.95	9.71	8.63	12.26
	3	0.92	1.09	0.86	1.57	10.37	12.78
	4	0.92	0.87	0.98	9.03	10.56	13.21
	5	0.92	0.95	1.30	9.05	9.57	12.94
	6	0.92	0.94	1.08	7.16	7.03	14.28
swap_methyl	1	0.92	1.00	0.84	1.47	1.05	7.31
	2	0.92	0.91	0.89	0.97	0.95	1.17
	3	0.92	0.84	1.24	1.09	1.36	1.35
	4	0.92	0.90	0.86	1.11	1.31	2.15
	5	0.92	0.85	0.97	0.95	1.76	1.18
	6	0.92	0.90	0.95	0.99	1.10	1.67
swap_coupled	1	0.92	0.93	0.94	1.05	0.98	1.53
	2	0.92	0.89	1.15	0.92	1.25	0.88
	3	0.92	0.96	0.83	1.38	1.17	1.28
	4	0.92	0.90	0.95	0.95	1.31	1.01
	5	0.92	0.89	0.95	1.32	1.28	0.88
	6	0.92	0.96	1.05	1.11	0.84	1.07
swap_stereo	1	0.92	0.91	1.05	0.90	0.92	0.88
	2	0.92	0.96	1.00	0.95	1.12	1.00
	3	0.92	0.92	0.96	0.97	0.86	0.92
	4	0.92	0.89	0.85	0.88	1.08	0.88
	5	0.92	0.88	1.01	0.90	1.01	0.93
	6	0.92	0.87	0.86	0.91	0.98	0.99
swap_carbon	1	0.92	0.97	0.89	1.31	1.35	2.61
	2	0.92	0.89	0.90	1.07	1.21	1.94
	3	0.92	0.88	0.90	0.97	1.19	1.35
	4	0.92	0.86	0.96	0.92	1.50	2.67
	5	0.92	0.92	0.93	0.88	0.96	1.29
	6	0.92	0.82	0.96	0.92	0.90	1.88
swap_sidechain	1	0.92	0.97	0.86	1.02	1.20	0.91
	2	0.92	0.87	0.98	1.28	1.06	0.90

APPENDIX

	3	0.92	1.05	0.99	0.88	1.09	0.98
	4	0.92	0.86	0.97	0.93	0.91	0.87
	5	0.92	1.05	0.89	0.87	1.22	1.10
	6	0.92	0.92	0.92	1.39	0.83	0.93
	1	0.92	0.91	0.76	0.92	0.93	1.06
	2	0.92	0.85	0.96	0.84	0.84	1.00
	3	0.92	0.91	0.92	1.10	0.99	1.06
combine_methyl	4	0.92	0.83	0.92	0.86	1.02	0.91
	5	0.92	0.94	0.91	0.90	0.93	0.93
	6	0.92	0.89	0.96	0.95	1.01	0.94
	1	0.92	0.97	0.98	1.10	1.01	10.07
	2	0.92	0.86	0.90	1.20	1.12	9.99
	3	0.92	0.90	0.89	1.20	2.49	9.48
combine_stereo	4	0.92	1.01	0.90	1.05	0.97	9.71
	5	0.92	0.956	0.95	1.05	0.992	4.838
	6	0.92	1.058	1.047	1.931	2.11	10.04

Table S9: C α -RMSDs of HR5537A structures calculated by AutoNOE-Rosetta on different scramble types and severity levels.

APPENDIX

		RMSD (A)					
Severity Level		0	1	2	3	4	5
		Repetition ID					
miss_methyl	1	0.93	0.94	0.90	1.00	0.97	1.13
	2	0.93	0.90	0.95	0.88	0.92	1.06
	3	0.93	0.83	0.90	0.87	0.92	0.96
	4	0.93	1.01	0.86	0.86	0.84	0.93
	5	0.93	0.93	0.93	0.98	0.91	0.97
	6	0.93	0.87	0.96	0.88	0.87	0.93
miss_sidechain	1	0.93	0.85	0.90	1.53	1.43	2.99
	2	0.93	0.91	1.07	1.16	2.07	1.63
	3	0.93	0.83	0.98	1.64	1.48	1.74
	4	0.93	1.01	0.93	1.10	3.45	3.03
	5	0.93	0.89	1.04	0.97	2.32	7.79
	6	0.93	0.95	0.96	1.03	2.15	2.21
miss_proton	1	0.93	0.83	1.14	1.11	1.57	2.24
	2	0.93	0.94	1.04	1.16	1.87	6.80
	3	0.93	0.86	0.81	1.09	1.35	1.56
	4	0.93	0.90	0.90	1.22	6.81	8.55
	5	0.93	1.01	0.95	0.85	1.73	1.76
	6	0.93	0.87	0.91	1.01	1.63	1.81
swap_methyl	1	0.93	0.81	0.95	0.89	1.03	1.01
	2	0.93	0.90	0.96	0.96	1.08	0.94
	3	0.93	0.86	0.97	1.12	0.93	1.18
	4	0.93	0.94	0.86	1.05	1.07	1.15
	5	0.93	0.99	0.91	0.87	0.90	1.21
	6	0.93	0.98	1.03	0.94	0.98	1.07
swap_coupled	1	0.93	0.86	0.92	0.93	0.88	0.90
	2	0.93	0.95	0.93	0.90	0.95	0.87
	3	0.93	0.81	0.94	0.95	0.86	0.92
	4	0.93	0.87	0.89	0.95	0.92	0.91
	5	0.93	0.85	0.92	1.03	0.95	0.88
	6	0.93	0.93	0.91	0.86	0.91	0.91
swap_stereo	1	0.93	0.90	0.90	0.89	0.89	0.89
	2	0.93	0.87	0.98	0.89	0.85	0.91
	3	0.93	0.93	0.91	0.97	0.89	0.92
	4	0.93	0.94	0.91	0.90	0.99	0.98
	5	0.93	0.95	0.91	1.00	0.94	1.02
	6	0.93	0.97	0.92	0.92	0.80	0.95
swap_carbon	1	0.93	0.97	0.87	0.92	0.81	0.87
	2	0.93	0.87	0.98	0.95	0.80	0.87
	3	0.93	0.88	0.91	0.85	0.98	0.93
	4	0.93	0.93	0.89	0.85	0.91	0.84
	5	0.93	0.98	0.93	0.86	0.93	0.89
	6	0.93	0.92	0.88	0.83	0.88	0.90
swap_sidechain	1	0.93	0.97	0.96	1.00	0.94	1.02
	2	0.93	0.90	0.85	0.83	0.85	0.89

APPENDIX

	3	0.93	0.79	1.01	0.91	1.05	0.87
	4	0.93	0.84	0.87	0.87	0.94	0.87
	5	0.93	1.01	0.85	1.01	0.91	0.87
	6	0.93	0.93	0.85	0.98	0.88	1.00
	1	0.93	1.00	0.94	0.95	0.96	0.93
	2	0.93	0.93	0.96	0.93	0.95	0.92
	3	0.93	0.83	0.87	1.03	0.92	1.03
combine_methyl	4	0.93	0.88	1.01	0.90	0.95	0.90
	5	0.93	0.89	0.99	0.96	0.95	0.86
	6	0.93	0.93	0.92	1.00	0.88	0.98
	1	0.93	0.95	0.86	0.83	0.99	0.98
	2	0.93	1.01	0.90	0.87	0.87	0.96
	3	0.93	0.90	0.81	0.93	0.89	0.99
combine_stereo	4	0.93	0.92	0.99	0.85	0.89	0.96
	5	0.93	0.853	0.849	0.849	0.903	0.871
	6	0.93	0.891	0.968	0.903	0.819	0.836

Table S10: C α -RMSDs of OR135 structures calculated by AutoNOE-Rosetta on different scramble types and severity levels.

APPENDIX

		RMSD (A)					
Severity Level		0	1	2	3	4	5
		Repetition ID					
miss_methyl	1	1.62	1.68	1.83	1.66	1.81	1.77
	2	1.62	1.79	1.97	1.78	1.91	1.56
	3	1.62	1.66	1.96	1.65	1.80	1.52
	4	1.62	2.00	1.38	1.89	1.80	1.56
	5	1.62	1.66	1.78	1.59	1.66	1.69
	6	1.62	1.63	1.66	1.61	1.58	1.59
miss_sidechain	1	1.62	1.74	2.26	2.46	5.23	9.85
	2	1.62	1.48	2.44	5.59	7.18	6.35
	3	1.62	1.68	2.22	4.06	6.01	10.13
	4	1.62	1.76	2.11	3.68	5.08	8.17
	5	1.62	2.02	2.47	2.88	5.28	7.26
	6	1.62	1.94	2.07	4.29	8.26	12.52
miss_proton	1	1.62	1.53	2.29	3.13	5.63	6.24
	2	1.62	2.19	2.55	3.70	4.11	12.86
	3	1.62	1.36	3.35	3.22	9.91	4.97
	4	1.62	1.85	1.90	4.06	14.99	7.76
	5	1.62	2.15	1.84	4.83	5.54	10.80
	6	1.62	1.89	2.34	4.22	5.36	13.29
swap_methyl	1	1.62	1.52	1.50	1.63	1.51	2.06
	2	1.62	1.95	1.73	2.04	1.61	1.59
	3	1.62	2.05	1.50	2.05	1.56	1.73
	4	1.62	1.86	1.91	1.89	1.85	1.97
	5	1.62	1.58	1.65	2.07	2.15	2.59
	6	1.62	1.68	2.00	1.66	1.66	1.64
swap_coupled	1	1.62	1.72	1.74	1.86	1.76	2.33
	2	1.62	1.48	1.77	1.98	2.34	2.19
	3	1.62	1.54	1.48	1.94	1.98	2.09
	4	1.62	1.94	1.68	1.56	2.03	2.35
	5	1.62	1.76	1.86	2.15	2.57	1.87
	6	1.62	1.76	1.70	1.74	2.21	2.17
swap_stereo	1	1.62	1.65	1.80	1.87	1.79	1.82
	2	1.62	1.66	1.35	1.49	1.76	1.54
	3	1.62	1.91	1.42	1.66	1.56	1.73
	4	1.62	1.47	1.64	1.54	1.96	1.41
	5	1.62	1.87	1.53	1.75	1.47	1.58
	6	1.62	1.64	1.74	1.41	1.80	1.43
swap_carbon	1	1.62	1.68	1.75	1.66	1.74	2.16
	2	1.62	1.50	2.14	1.54	2.08	2.49
	3	1.62	1.79	1.87	1.75	1.68	1.61
	4	1.62	1.87	1.38	1.86	2.04	1.70
	5	1.62	1.78	1.63	1.62	2.16	1.56
	6	1.62	1.70	1.67	2.46	1.80	1.97
swap_sidechain	1	1.62	1.67	1.65	2.15	2.09	1.70
	2	1.62	1.98	1.60	1.66	2.18	1.84

APPENDIX

	3	1.62	1.73	1.66	1.74	1.97	1.88
	4	1.62	1.88	2.08	1.65	1.95	2.71
	5	1.62	1.67	1.65	1.88	1.41	1.71
	6	1.62	1.52	1.63	2.31	2.04	1.85
	1	1.62	1.70	1.75	1.68	1.75	1.44
	2	1.62	1.91	1.62	1.58	1.83	1.64
	3	1.62	1.82	1.74	1.46	1.50	1.83
combine_methyl	4	1.62	1.71	1.59	1.50	1.60	1.60
	5	1.62	1.70	1.61	1.71	1.79	1.66
	6	1.62	1.69	1.74	1.52	1.37	1.89
	1	1.62	1.65	1.82	2.18	2.11	1.99
	2	1.62	1.41	1.60	2.14	1.99	2.30
	3	1.62	1.69	1.77	2.61	2.18	1.95
combine_stereo	4	1.62	1.55	1.86	1.92	1.91	2.18
	5	1.62	1.62	1.78	1.867	1.792	2.76
	6	1.62	1.815	1.781	2.367	2.062	2.146

Table S11: C α -RMSDs of PfR193A structures calculated by AutoNOE-Rosetta on different scramble types and severity levels.

APPENDIX

Severity Level	Repetition ID	RMSD (A)					
		0	1	2	3	4	5
miss_methyl	1	1.03	0.94	1.01	1.56	5.53	5.99
	2	1.03	1.17	1.53	1.41	2.11	4.13
	3	1.03	0.92	1.30	1.80	3.85	5.53
	4	1.03	1.24	1.12	4.32	1.54	1.56
	5	1.03	1.13	2.28	1.18	5.14	1.78
	6	1.03	1.24	1.11	3.19	1.73	5.85
miss_sidechain	1	1.03	1.62	3.89	4.70	9.82	11.99
	2	1.03	1.03	3.97	7.45	12.17	8.66
	3	1.03	1.89	4.61	7.13	10.97	12.78
	4	1.03	1.73	1.55	10.07	9.90	9.88
	5	1.03	1.27	3.71	4.15	11.26	7.43
	6	1.03	0.89	7.83	4.41	8.90	10.02
miss_proton	1	1.03	1.40	2.93	7.65	10.42	13.14
	2	1.03	1.13	2.04	10.94	10.47	10.55
	3	1.03	1.38	3.11	7.84	8.83	13.01
	4	1.03	4.36	3.58	9.47	12.74	13.73
	5	1.03	1.43	3.31	8.07	10.50	11.18
	6	1.03	2.95	7.05	9.98	10.69	13.64
swap_methyl	1	1.03	1.22	2.47	1.06	1.18	4.06
	2	1.03	1.00	2.08	2.58	1.21	1.39
	3	1.03	2.10	1.19	5.27	1.13	2.20
	4	1.03	1.00	2.23	1.23	1.09	3.72
	5	1.03	0.95	1.08	1.72	1.44	3.60
	6	1.03	1.35	3.05	1.61	2.12	2.84
swap_C-H	1	1.03	1.66	1.54	1.32	1.42	3.56
	2	1.03	1.16	0.96	5.23	2.63	3.22
	3	1.03	1.42	1.43	2.26	3.51	7.65
	4	1.03	1.09	1.05	6.31	3.71	5.42
	5	1.03	1.00	6.65	1.76	4.99	2.16
	6	1.03	0.99	1.56	1.03	3.04	1.23
swap_stereo	1	1.03	1.05	1.21	1.04	1.02	0.98
	2	1.03	1.09	0.92	1.30	1.74	0.92
	3	1.03	1.02	1.18	0.88	1.03	0.98
	4	1.03	1.13	1.17	1.21	1.00	1.04
	5	1.03	0.93	1.11	1.20	0.99	1.14
	6	1.03	0.99	1.24	1.01	0.97	1.15
swap_carbon	1	1.03	1.32	1.06	1.17	3.87	10.10
	2	1.03	1.38	1.06	4.27	1.61	4.19
	3	1.03	0.84	2.96	1.66	2.99	8.23
	4	1.03	1.27	1.19	2.97	2.27	12.08
	5	1.03	1.22	1.28	2.06	2.43	6.57
	6	1.03	1.81	1.75	3.46	3.00	5.81
swap_sidechain	1	1.03	1.01	1.26	1.18	5.04	1.61
	2	1.03	1.11	5.17	2.28	5.28	1.67

APPENDIX

	3	1.03	1.39	2.32	1.36	0.94	1.96
	4	1.03	1.54	1.17	1.34	1.57	1.55
	5	1.03	4.12	3.17	2.60	2.60	4.84
	6	1.03	4.74	3.35	2.80	1.07	1.97
	1	1.03	1.30	1.28	3.67	1.15	1.74
	2	1.03	0.93	0.92	2.47	1.29	2.77
	3	1.03	1.03	1.18	1.08	1.56	3.20
combine_methyl	4	1.03	0.94	1.47	1.21	1.83	2.65
	5	1.03	1.04	1.08	1.12	1.96	1.63
	6	1.03	1.05	1.24	1.48	2.16	2.76
	1	1.03	1.70	1.68	2.41	6.03	6.63
	2	1.03	2.51	2.16	4.12	5.57	6.06
	3	1.03	0.89	1.33	4.58	4.42	5.44
combine_stereo	4	1.03	1.61	1.18	5.42	1.65	6.49
	5	1.03	1.26	1.75	3.41	3.04	5.85
	6	1.03	1.26	1.56	6.23	5.48	6.42

Table S12: C α -RMSDs of HR5537A structures calculated by CYANA on different scramble types and severity levels.

APPENDIX

Severity Level	Repetition ID	RMSD (A)					
		0	1	2	3	4	5
miss_methyl	1	0.95	1.15	0.92	1.11	0.98	1.01
	2	0.95	0.85	0.97	0.99	1.16	1.15
	3	0.95	1.02	0.81	0.71	0.94	0.96
	4	0.95	0.96	0.77	1.01	0.91	1.00
	5	0.95	1.06	1.04	0.90	1.15	1.19
	6	0.95	1.00	1.06	0.90	0.90	1.07
miss_sidechain	1	0.95	1.32	1.48	2.63	1.86	9.28
	2	0.95	1.20	1.30	4.08	6.10	9.40
	3	0.95	0.85	1.23	3.10	9.81	3.16
	4	0.95	1.34	1.37	3.63	9.60	11.22
	5	0.95	0.93	1.57	3.56	8.98	10.23
	6	0.95	0.75	1.08	3.28	11.01	9.94
miss_proton	1	0.95	0.80	1.78	4.34	9.34	11.61
	2	0.95	0.91	1.30	9.20	4.62	10.11
	3	0.95	1.16	1.32	2.73	5.98	4.42
	4	0.95	0.92	2.68	3.34	4.90	10.56
	5	0.95	2.47	2.05	10.25	9.01	10.47
	6	0.95	0.92	1.49	4.75	10.84	11.21
swap_methyl	1	0.95	0.97	2.39	1.01	1.22	1.29
	2	0.95	0.89	1.43	1.31	1.28	1.29
	3	0.95	1.10	1.27	1.19	0.79	1.51
	4	0.95	1.05	0.90	1.28	2.42	0.87
	5	0.95	0.94	1.11	1.18	1.11	2.26
	6	0.95	1.31	0.96	0.96	1.35	1.14
swap_C-H	1	0.95	0.94	0.94	1.58	0.98	0.99
	2	0.95	0.75	1.24	1.39	1.47	0.78
	3	0.95	0.97	1.03	1.15	0.96	0.86
	4	0.95	1.14	0.86	0.90	1.75	0.81
	5	0.95	1.43	1.16	1.23	1.10	0.86
	6	0.95	1.40	1.03	0.89	0.89	4.09
swap_stereo	1	0.95	0.84	0.89	0.88	1.23	0.90
	2	0.95	0.87	0.86	0.96	0.81	0.98
	3	0.95	0.85	1.04	0.81	0.93	1.10
	4	0.95	0.81	0.88	0.84	1.07	0.97
	5	0.95	1.13	1.09	1.01	0.99	0.95
	6	0.95	1.08	0.79	0.93	1.01	0.90
swap_carbon	1	0.95	1.06	1.19	0.84	1.29	2.27
	2	0.95	0.88	0.91	1.36	1.18	1.58
	3	0.95	0.95	0.83	1.00	1.57	1.27
	4	0.95	1.22	0.95	1.05	0.77	1.15
	5	0.95	0.95	1.11	0.85	0.97	1.52
	6	0.95	0.85	1.18	1.10	1.00	3.12
swap_sidechain	1	0.95	1.20	1.28	1.29	4.89	2.24
	2	0.95	3.34	1.25	2.23	2.55	5.73

APPENDIX

	3	0.95	0.83	2.63	2.79	3.18	4.00
	4	0.95	0.83	2.73	1.74	3.76	9.99
	5	0.95	2.21	5.76	1.51	8.44	1.40
	6	0.95	5.53	0.79	2.05	1.35	0.94
	1	0.95	1.02	1.04	0.94	1.05	1.03
	2	0.95	1.26	0.94	0.85	0.86	0.93
	3	0.95	1.28	0.96	1.05	0.86	1.15
combine_methyl	4	0.95	0.95	1.05	1.07	1.39	1.67
	5	0.95	0.88	1.47	1.02	0.83	0.95
	6	0.95	0.79	1.03	1.13	1.29	1.19
	1	0.95	0.97	1.05	0.94	1.26	2.45
	2	0.95	0.89	1.40	1.73	1.33	1.56
	3	0.95	0.73	0.94	1.18	1.24	1.13
combine_stereo	4	0.95	0.75	0.81	1.44	1.67	1.31
	5	0.95	0.87	0.77	0.89	1.40	2.13
	6	0.95	0.87	1.33	1.57	1.50	1.30

Table S13: C α -RMSDs of OR135 structures calculated by CYANA on different scramble types and severity levels.

APPENDIX

		RMSD (A)					
Severity Level	Repetition ID	0	1	2	3	4	5
miss_methyl	1	1.62	1.71	1.54	1.74	1.66	1.73
	2	1.62	1.52	1.84	1.59	1.75	1.75
	3	1.62	1.65	1.64	1.74	1.78	1.52
	4	1.62	1.65	1.64	1.54	1.83	1.79
	5	1.62	1.65	1.71	1.87	2.28	1.56
	6	1.62	1.63	1.66	1.83	1.55	2.33
miss_sidechain	1	1.62	1.66	2.47	5.16	9.77	13.46
	2	1.62	1.54	2.25	11.39	12.92	11.91
	3	1.62	1.64	1.97	4.81	10.87	15.41
	4	1.62	1.57	2.39	2.32	5.92	16.90
	5	1.62	2.61	2.56	2.94	7.08	11.92
	6	1.62	2.34	1.92	7.91	13.26	11.47
miss_proton	1	1.62	1.57	2.45	8.21	8.52	18.34
	2	1.62	1.73	10.16	12.66	11.42	14.70
	3	1.62	1.63	2.86	9.44	10.11	5.70
	4	1.62	1.91	2.88	11.89	11.59	25.21
	5	1.62	1.87	2.91	14.30	13.71	16.91
	6	1.62	1.46	9.72	10.81	13.15	11.37
swap_methyl	1	1.62	1.59	1.70	4.29	1.72	1.73
	2	1.62	1.76	1.55	1.72	1.67	1.53
	3	1.62	1.60	1.68	1.86	1.61	1.87
	4	1.62	1.60	1.60	1.68	1.66	1.95
	5	1.62	1.47	1.75	1.91	1.60	1.62
	6	1.62	1.63	1.59	1.75	1.59	1.62
swap_C-H	1	1.62	1.97	1.90	2.35	1.51	14.81
	2	1.62	1.60	1.59	1.60	1.61	1.85
	3	1.62	1.71	1.50	1.50	2.17	1.73
	4	1.62	1.62	1.54	1.91	3.88	1.88
	5	1.62	1.55	1.63	1.79	2.33	1.57
	6	1.62	2.93	1.74	1.49	1.98	3.24
swap_stereo	1	1.62	1.61	1.55	1.76	1.66	1.78
	2	1.62	1.73	1.69	1.79	1.68	1.75
	3	1.62	1.60	1.52	1.72	1.80	1.55
	4	1.62	1.64	1.69	1.61	1.65	1.52
	5	1.62	1.53	1.55	1.60	1.64	1.74
	6	1.62	1.55	1.49	1.69	1.64	1.67
swap_carbon	1	1.62	1.53	1.75	1.70	2.57	12.41
	2	1.62	1.68	1.69	1.74	3.10	2.24
	3	1.62	1.78	1.91	1.87	1.80	2.45
	4	1.62	1.61	1.67	1.74	1.86	10.04
	5	1.62	1.66	1.58	2.12	1.88	3.41
	6	1.62	1.73	1.59	1.72	2.64	10.17
swap_sidechain	1	1.62	1.77	1.56	1.92	1.55	3.08
	2	1.62	1.52	1.71	2.01	2.67	12.82

APPENDIX

	3	1.62	1.69	1.71	1.60	2.71	1.75
	4	1.62	1.70	1.72	2.11	1.55	9.44
	5	1.62	1.65	1.69	1.53	1.66	7.80
	6	1.62	1.67	1.86	2.70	1.81	1.83
	1	1.62	1.49	1.65	1.74	1.70	1.80
	2	1.62	1.51	1.62	1.50	1.56	2.01
	3	1.62	1.63	1.91	1.56	1.73	1.62
combine_methyl	4	1.62	1.65	1.72	1.75	1.78	1.53
	5	1.62	1.58	1.46	1.65	1.71	1.63
	6	1.62	1.65	1.77	1.55	1.55	1.91
	1	1.62	1.58	1.67	2.61	1.66	1.89
	2	1.62	1.61	1.84	1.84	1.79	4.17
	3	1.62	1.74	1.75	1.72	2.31	2.78
combine_stereo	4	1.62	1.65	1.93	3.13	1.82	1.94
	5	1.62	1.55	1.73	1.93	2.19	2.41
	6	1.62	1.64	1.66	1.76	1.67	2.15

Table S14: C α -RMSDs of PfR193A structures calculated by CYANA on different scramble types and severity levels.

APPENDIX

		RMSD (A)					
Severity Level		0	1	2	3	4	5
Repetition ID							
miss_methyl	1	1.29	1.78	1.59	2.81	3.12	3.09
	2	1.29	2.09	3.57	1.82	4.09	2.04
	3	1.29	2.03	1.96	1.88	2.18	3.87
	4	1.29	1.37	1.26	2.83	2.22	4.18
	5	1.29	1.39	1.52	2.17	3.10	2.40
	6	1.29	1.39	1.53	2.83	1.37	2.44
miss_sidechain	1	1.29	2.23	2.13	10.62	10.19	7.43
	2	1.29	1.30	2.77	6.71	13.37	9.74
	3	1.29	1.39	7.78	3.17	10.81	11.51
	4	1.29	2.72	2.66	7.56	9.03	11.94
	5	1.29	2.04	2.35	6.38	10.23	13.34
	6	1.29	2.36	8.97	8.50	5.09	10.91
miss_proton	1	1.29	1.94	2.41	9.64	9.89	11.28
	2	1.29	1.94	2.61	7.94	7.93	13.13
	3	1.29	1.30	4.19	2.80	7.49	10.59
	4	1.29	1.54	3.89	11.01	9.28	9.48
	5	1.29	1.39	5.32	7.18	7.86	13.13
	6	1.29	1.40	8.79	4.66	7.49	13.47
swap_methyl	1	1.29	1.35	9.90	3.33	7.04	6.72
	2	1.29	1.62	1.41	6.34	6.45	3.12
	3	1.29	1.18	4.48	2.27	2.50	9.36
	4	1.29	1.41	1.49	4.92	3.68	6.45
	5	1.29	2.06	3.90	2.76	8.84	7.91
	6	1.29	1.89	1.45	2.91	9.52	7.09
swap_C-H	1	1.29	9.22	2.79	1.69	4.63	7.56
	2	1.29	8.46	1.41	9.91	2.62	3.49
	3	1.29	1.25	11.69	2.66	3.43	3.04
	4	1.29	1.28	2.14	6.28	9.21	7.28
	5	1.29	2.37	2.23	9.13	2.98	3.57
	6	1.29	1.97	5.61	3.36	2.51	8.54
swap_stereo	1	1.29	1.45	1.91	1.43	1.72	1.72
	2	1.29	1.59	1.32	1.71	1.33	1.54
	3	1.29	1.30	1.70	1.22	1.88	1.57
	4	1.29	1.29	9.14	1.90	1.38	1.50
	5	1.29	2.03	1.44	1.53	1.63	1.38
	6	1.29	1.82	1.70	1.95	1.21	1.72
swap_carbon	1	1.29	8.19	10.58	2.16	12.11	10.71
	2	1.29	1.37	3.53	7.49	2.27	9.15
	3	1.29	1.56	1.81	2.48	4.13	7.24
	4	1.29	1.62	9.28	1.68	5.11	12.10
	5	1.29	1.85	8.76	4.42	11.98	8.66
	6	1.29	1.24	4.82	8.03	3.09	9.27
swap_sidechain	1	1.29	1.73	2.08	6.83	5.72	2.12
	2	1.29	2.23	4.03	1.89	2.00	2.75

APPENDIX

	3	1.29	1.94	7.61	1.28	2.29	2.86
	4	1.29	2.45	11.23	1.70	1.58	1.92
	5	1.29	5.02	5.10	2.64	5.29	6.52
	6	1.29	9.62	1.92	1.42	1.46	4.33
	1	1.29	1.51	1.73	7.15	2.64	1.58
	2	1.29	1.90	2.65	2.65	2.05	5.56
combine_methyl	3	1.29	1.76	3.08	1.69	1.34	2.06
	4	1.29	1.27	2.93	1.72	4.25	1.69
	5	1.29	1.62	3.42	2.39	2.88	1.89
	6	1.29	2.75	1.23	2.28	2.28	2.10
	1	1.29	1.54	2.10	2.69	3.34	1.89
	2	1.29	1.67	9.00	1.57	1.56	1.66
combine_stereo	3	1.29	1.90	2.48	1.26	7.73	2.58
	4	1.29	1.70	2.51	1.71	2.77	1.46
	5	1.29	1.12	1.60	1.12	1.88	1.36
	6	1.29	1.57	1.10	1.08	1.55	2.94

Table S15: C α -RMSDs of HR5537A structures calculated by ASDP on different scramble types and severity levels.

APPENDIX

		RMSD (A)					
Severity Level		0	1	2	3	4	5
Repetition ID							
miss_methyl	1	0.93	0.93	1.09	1.04	1.19	1.07
	2	0.93	0.99	1.04	1.25	1.18	1.26
	3	0.93	1.00	1.01	1.00	1.03	1.02
	4	0.93	1.18	1.38	0.98	1.15	1.09
	5	0.93	1.04	0.98	1.05	1.04	1.11
	6	0.93	0.99	0.94	1.09	0.98	1.35
miss_sidechain	1	0.93	1.95	1.52	1.93	1.58	7.85
	2	0.93	0.97	1.18	1.89	5.37	2.02
	3	0.93	0.99	2.42	1.66	2.39	8.30
	4	0.93	1.13	1.25	2.27	2.86	7.55
	5	0.93	1.42	1.62	1.55	2.95	9.72
	6	0.93	1.23	1.08	2.27	2.92	7.23
miss_proton	1	0.93	1.01	1.11	1.70	3.82	10.17
	2	0.93	1.07	1.77	1.18	5.98	9.61
	3	0.93	1.07	1.11	1.79	7.60	8.08
	4	0.93	1.09	1.82	1.95	2.65	8.92
	5	0.93	1.02	1.44	1.47	2.01	7.83
	6	0.93	1.00	0.99	1.80	1.76	8.50
swap_methyl	1	0.93	1.10	1.52	1.42	2.54	1.42
	2	0.93	1.10	0.99	1.31	1.07	6.38
	3	0.93	1.16	1.86	1.22	1.33	2.08
	4	0.93	1.13	1.12	1.34	1.47	1.91
	5	0.93	2.10	1.04	3.38	1.52	2.56
	6	0.93	1.25	0.95	1.35	1.35	3.02
swap_C-H	1	0.93	1.21	1.15	NaN	NaN	NaN
	2	0.93	0.93	1.76	NaN	1.61	NaN
	3	0.93	0.97	0.95	NaN	2.35	NaN
	4	0.93	1.12	1.70	1.01	NaN	NaN
	5	0.93	1.01	1.24	NaN	NaN	NaN
	6	0.93	1.08	1.27	NaN	NaN	NaN
swap_stereo	1	0.93	0.93	0.98	1.08	0.97	0.97
	2	0.93	0.99	1.00	0.98	0.98	0.97
	3	0.93	0.99	0.99	0.99	0.99	0.98
	4	0.93	0.93	0.99	0.97	1.00	0.97
	5	0.93	0.93	0.98	0.97	0.98	0.97
	6	0.93	0.94	0.99	1.01	0.97	0.97
swap_carbon	1	0.93	1.03	1.30	1.13	NaN	NaN
	2	0.93	0.95	1.03	1.77	NaN	NaN
	3	0.93	0.94	0.90	NaN	NaN	NaN
	4	0.93	1.04	1.07	1.56	NaN	NaN
	5	0.93	1.04	1.27	1.54	NaN	NaN
	6	0.93	0.93	1.11	NaN	NaN	NaN
swap_sidechain	1	0.93	1.43	1.64	0.98	2.77	1.48
	2	0.93	2.56	2.31	0.96	2.26	3.37

APPENDIX

	3	0.93	1.26	1.37	2.33	2.47	2.26
	4	0.93	1.03	1.18	1.16	2.95	1.90
	5	0.93	1.87	1.72	1.45	5.01	1.08
	6	0.93	1.28	3.59	2.00	1.39	0.96
	1	0.93	1.05	1.06	0.97	1.01	0.88
	2	0.93	0.93	1.42	1.38	1.01	1.14
	3	0.93	1.04	1.16	1.06	1.06	1.57
combine_methyl	4	0.93	1.05	1.21	1.04	1.20	1.12
	5	0.93	1.14	1.02	1.59	1.29	1.76
	6	0.93	1.02	1.01	0.93	0.97	1.22
	1	0.93	1.06	0.99	1.08	1.25	1.23
	2	0.93	1.06	1.49	1.15	1.20	1.23
	3	0.93	0.97	1.07	1.07	1.23	1.17
combine_stereo	4	0.93	1.17	1.51	1.26	1.22	1.12
	5	0.93	0.94	1.59	1.11	1.09	1.47
	6	0.93	0.91	1.08	1.00	1.08	1.13

Table S16: C α -RMSDs of OR135 structures calculated by ASDP on different scramble types and severity levels.

APPENDIX

		RMSD (A)					
Severity Level		0	1	2	3	4	5
Repetition ID							
miss_methyl	1	1.34	1.39	1.46	2.36	1.36	1.40
	2	1.34	1.35	1.66	1.41	1.52	1.89
	3	1.34	1.31	1.40	1.60	1.95	1.71
	4	1.34	1.47	1.59	1.34	2.91	1.95
	5	1.34	1.34	1.32	1.40	2.32	2.31
	6	1.34	1.25	1.68	1.61	1.37	1.51
miss_sidechain	1	1.34	1.76	2.43	4.54	4.88	11.01
	2	1.34	1.59	3.18	4.35	6.46	5.80
	3	1.34	1.50	1.70	3.71	7.18	11.89
	4	1.34	1.37	1.75	4.07	5.75	9.40
	5	1.34	1.68	1.76	3.38	5.81	9.32
	6	1.34	1.70	4.64	3.49	8.42	11.00
miss_proton	1	1.34	1.49	3.58	7.00	5.22	11.07
	2	1.34	1.50	2.59	5.38	12.60	9.34
	3	1.34	1.75	3.70	4.63	5.62	7.97
	4	1.34	1.38	3.38	5.85	12.30	8.71
	5	1.34	2.04	2.81	12.07	6.87	7.29
	6	1.34	1.48	3.71	7.36	4.82	8.27
swap_methyl	1	1.34	3.00	1.37	3.54	3.94	2.77
	2	1.34	1.50	1.41	1.54	1.59	1.97
	3	1.34	1.77	1.79	1.75	1.47	1.57
	4	1.34	1.40	1.20	1.53	3.56	3.68
	5	1.34	1.56	1.71	2.69	1.70	2.70
	6	1.34	1.58	1.81	2.04	4.52	1.57
swap_C-H	1	1.34	1.79	1.53	1.40	2.86	2.93
	2	1.34	1.42	1.72	2.30	6.39	4.58
	3	1.34	1.45	1.82	3.03	4.80	1.96
	4	1.34	1.75	1.72	1.67	3.94	6.16
	5	1.34	1.39	1.65	2.19	3.55	2.76
	6	1.34	1.90	1.82	7.91	4.70	6.27
swap_stereo	1	1.34	1.41	1.33	1.37	1.64	1.36
	2	1.34	1.34	1.50	1.27	1.45	1.30
	3	1.34	1.66	1.41	1.35	1.32	1.39
	4	1.34	1.21	1.43	1.30	1.72	1.57
	5	1.34	1.29	1.39	1.78	1.32	1.23
	6	1.34	1.42	1.50	1.53	1.23	1.51
swap_carbon	1	1.34	1.78	1.46	1.77	2.62	3.98
	2	1.34	1.47	5.88	1.40	1.76	9.86
	3	1.34	1.30	1.74	2.36	2.78	3.08
	4	1.34	1.87	1.35	1.62	NaN	3.66
	5	1.34	1.36	1.88	2.02	NaN	4.15
	6	1.34	1.73	1.81	2.27	NaN	2.91
swap_sidechain	1	1.34	1.57	1.86	2.02	2.14	2.47
	2	1.34	1.59	1.48	1.66	3.92	1.87

APPENDIX

	3	1.34	2.38	1.60	2.06	1.43	1.65
	4	1.34	1.52	1.33	1.64	1.69	5.85
	5	1.34	1.46	1.36	1.88	2.73	6.51
	6	1.34	1.41	1.58	3.70	2.25	4.00
	1	1.34	1.46	1.24	1.60	1.33	1.68
	2	1.34	1.51	1.29	1.31	2.60	1.38
	3	1.34	1.25	1.51	1.52	1.45	1.47
combine_methyl	4	1.34	1.59	1.26	1.53	1.30	1.38
	5	1.34	1.47	1.49	1.27	1.47	1.59
	6	1.34	1.48	1.28	1.37	1.49	1.68
	1	1.34	1.38	1.47	4.17	1.54	3.96
	2	1.34	1.31	2.06	2.26	3.94	4.00
	3	1.34	1.38	1.31	2.00	3.50	4.19
combine_stereo	4	1.34	1.52	1.51	4.76	4.67	3.97
	5	1.34	1.49	1.46	4.73	4.30	4.30
	6	1.34	1.46	1.78	4.26	4.21	2.12

Table S17: C α -RMSDs of PfR193A structures calculated by ASDP on different scramble types and severity levels.

A.3 Supplementary Methods

A.3.1 AutoNOE-Rosetta calculations

The following method section is written in the style of a tutorial, and reflects exactly what has been done for each target. The benchmark set can be downloaded from our website (www.csrosetta.org/benchmarks/). To run targets in the benchmark set, proceed directly with Section 1.5. Note, that for a given target multiple different *setups* can be maintained by using the flag `-label`. This has been used in the benchmark set to differentiate between *raw* and *refined* peak-lists, if omitted the *label* defaults to *standard*.

Rosetta applications are denoted in the following with the extension `<.ext>`, which should be replaced with the system and compiler dependent extension. For instance, for gcc compiled Rosetta on a linux system use `.default.linuxgccrelease`. Commands shown here without extension are python programs from the csrosetta toolbox and do not require an extension for their execution (`pick_fragments`, `setup_target`, ...).

A.3.1.1 Pre-requisites

The CS-Rosetta toolbox version 2.0 or higher has to be downloaded and installed from (www.csrosetta.org/downloads/). The Rosetta software package version 3.6 or higher has to be obtained from (www.rosettacommons.org).

A.3.1.2 Fragment Selection

We have run the fragment picker(Vernon et al. 2013) for all targets as follows

```
pick_fragments -cs t000.tab -trim -nohom
```

The flag `-nohom`, leads to exclusion of fragments from homologous proteins. This flag should be omitted when AutoNOE-Rosetta is not used for benchmarking. The flag `-trim` leads to automatic removal of flexible tails based on TALOS+ computed RCI S2 values. The `pick_fragments` application will correct chemical shifts if TALOS+ complains about referencing problems. For instance TALOS+ output like this

```
[...]
```

```
Estimated Referencing Offset for CA/CB: 1.138 +/- 0.162ppm Size: 78
```

will lead to subtraction of 1.138 from all CA and CB shifts. The fragment picker will write the file `t000.corrected.tab`. The reference-corrected chemical shifts are used for fragment picking and chemical shift rescoring, but not for automatic NOE assignment.

APPENDIX

Automatic correction of referencing issues can be turned off, using flag `-nocheck`. Referencing can be manually corrected using the command `correctCSoffset`.

A.3.1.3 Preparing Shift Assignment Files

The shift-assignment files (typical extension `.prot`), are obtained from BMRB files using `bmr2prot` and trimmed to the sequence determined by the auto-trim method of the fragment picker using `renumber_prot`.

A.3.1.4 Preparing Peak Files

The peak-files have to be prepared to be read-able with AutoNOE-Rosetta. This can be done with the provided tool `clean_peak_files`, which reads column based peak data and converts it into a file-format similar to XEASY and CYANA peak-list formats, including an appropriate header. Please refer to the application reference (<http://www.csrosetta.org/reference>) for details.

A.3.1.5 Starting the runs

Using the automated setup tools (www.csrosetta.org) we combine all input data files for one target (e.g., `t000`) as follows

```
setup_target --method autoNOE --target t000 \<\  
-fasta t000.fasta \<\  
-frags t000.frag3.nohom.dat.gz t000.frag9.nohom.dat.gz \<\  
-peaks ali.peaks aro.peaks n.peaks \<\  
-shifts t000.prot \<\  
-cs t000.tab \<\  
[-label raw]  
[-rdc t000_med1.rdc t000_med2.rdc] \<\  
[-native native.pdb [-native_restrict reference.rigid]]
```

and start a run from this setup with the command (BASH-syntax)

```
for cst in 5 10 25 50; do  
  setup_run --method autoNOE --target t000 [-label raw]\<\  
  -dir . \<\  
  -run_label run_cst$cst \<\  
  -noesy_cst_strength $cst \<\  
  -job slurm  
done
```

This will start runs at constraint weights 5, 10, 25 and 50, respectively. For *raw* peak-files we changed these to 1,2,5,10, 25, and 50, respectively. The flag `-job` is used to indicate which queuing system to use (here SLURM). Template files for different queuing systems are

APPENDIX

provided in the csrosetta-toolbox (see www.csrosetta.org for more information on how to adapt these templates to your own queuing system).

A.3.1.6 Final model selection per run

RASREC calculations are finished when the directory `fullatom_pool_stage8` is created. The final decoys of the structural pool are found in `fullatom_pool/decoys.out`. Using the command

```
extract_decoys -formula 'score-atom_pair_constraint-rdc' -N 10 -verbose 0 > low_10.out
```

we extract the 10 best models with Rosetta-energy only. The `atom_pair_constraint` and `rdc` pseudo-energies are contained in `score` of final decoys and thus have to be subtracted. Alternatively, one could pick final structures by the weighted sum of Rosetta Energy and restraint energy, that is contained in column `score`. This is done with the command

```
extract_decoys -score 10 -verbose 0 > low_10.out.
```

On our benchmark set both selection methods yield comparable results.

A.3.1.7 Median energy of a run

The command

```
silent_data low_10.out score atom_pair_constraint rdc|awk '{print $1-$3-$2}' | median
```

yields the median Rosetta energy of the lowest 10 models in the first column of the output (for instance):

```
median   Q1    Q3    hi    lo
-221.831 -223.333 -220.377 -219.375 -223.906
```

This example yields the median energy ME=-221.831.

A.3.1.8 Convergence of a run

In this step we figure out which residues remain unconverged and what the precision (measured as pairwise RMSD) is on the remaining (converged) residues. To this end, use

```
ensemble_analysis.<ext> -residues:patch_selectors replonly -in:file:silent low_10.out -wRMSD 2 -calc:rmsd -out:levels
all:error main:info
```

which yields (for example)

```
main: (0) make rigid with cutoff 2 and calculate RMSD
main: (0) computer RMSD on RIGID 2 6 0 0 0
main: (0) RIGID 16 76 0 0 0
main: (0) RIGID 82 113 0 0 0
main: (0)
main: (0) number of atoms from 116 for mean RMSD: 93
main: (0) mean RMSD to average structure: 1.45
main: (0) mean pairwise RMSD: 2.15
```

APPENDIX

main: (0) mean pairwise RMSD * superposed_fraction_of_atoms^-1: 2.68

The converged residues are given by the lines starting with RIGID, and are residues 16-76 and 82-113 in the example output. For N models with M residues of which C have converged, the mean pairwise RMSD is computed as

$$P = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N \sqrt{C^{-1} \sum_{k=1}^M w_k \|x_{k,i} - x_{k,j}\|^2},$$

where N denotes the number of conformers, $x_{k,j}$ denotes the k th C_α -atom of conformation j , and the w_k are 1 for converged residues and 0 for unconverged residues. An effective precision (EP) is computed from P by multiplying with the inverse of the fraction of converged residues

$$EP = P \left(\frac{C}{M} \right)^{-1}$$

A.3.1.9 Analysis of final NOE assignment of a run

To use the NOE assignments and resulting restraints in further work or for deposition with the models, the NOE assignment module of ROSETTA has to be applied to the final models. To this end, a script is generated when the run directory is created (*setup_run*). Moreover, the assignments can also be used to provide statistics about the peaks and their assignments with command *noeout2txt*

To use the final models for NOE assignment run the two provided scripts

```
./final_assignments.sh fullatom_pool/low_10.out  
noeout2txt -peaks final_assignment/NOE_final.out -split_level 0
```

Which will provide output as follows:

```
Peaks:  
picked.....| 3705  
zero intensity.....| 0  
assigned.....| 2766  
with more than 5 assignments.....| 194  
unassigned.....| 939  
without assignment possibility.....| 139  
eliminated due to Network.....| 6  
eliminated due to MinPeakVol.....| 64  
eliminated due to DistViol.....| 341  
eliminated due to MaxAssign.....| 389  
Assignments:  
with unique assignment.....| 59
```


APPENDIX

with short-range assignment $ i-j \leq 1$	2273
with medium-range assignment $1 < i-j < 5$	206
with long-range assignment $ i-j \geq 5$	247
violated by conformers.....	341
between 0.0 and 0.5 Å.....	23
between 0.5 and 2.0 Å.....	109
between 2.0 and 5.0 Å.....	93
above 5.0 Å.....	116

This output provides some statistics about the assignment. We have found that these statistics do not allow to systematically select the most accurate runs. The only result used for run-selection is whether *violated by conformers* (VIOL) is below or above 2000. Here, this number is far below the threshold (shown in red in the example output).

A.3.1.10 Selection of best run

For each target we perform simulations with different constraint weights. $CST=1,2,5,10,25,50$. (where 1 and 2 is only used for *raw* data sets). It is possible to add more runs with other constraint weights, and one can also perform multiple runs at the same constraint weight. However, we recommend to stay below 50, as higher restraint weights seem to yield slightly worse accuracy on our benchmark. The selection procedure presented here is responsible for selecting a single run from all performed runs. Alternatively, one can select several top-ranking runs. To rank, we use from Section 2.2, 2.3 and 2.4 the values ME, EP and VIOL, that have been computed for each run. If VIOL is larger than 2000 we use only ME to rank runs, otherwise we rank by

$$S = ME - 12 \log(CST) + 5EP,$$

where the correction $-12 \log(CST) + 5EP$ allows to trade-off small differences in Rosetta energy for gaining higher precision and using stronger constraint weights. We usually find a general agreement between runs if the input data is easy. If the data is difficult, runs may differ in convergence.

A.3.1.11 Violation reduction in final models

During structure generation with AutoNOE-Rosetta restraint weights are kept relatively low, and final structure selection is mainly based on Rosetta energy. Moreover, NOE restraints that are intra-residue or between neighboring residues are usually ignored. This means that we will usually find some residual heterogeneity for rotamers that is not consistent with the automatic NOE restraints. Typically, the majority of the final models adopt a rotamer that is consistent with the NOE restraints, whereas the remainder adopt an alternative rotamer, which violates NOE restraints. To remove this unnecessary heterogeneity, we run a short relaxation of the 10 final models with all NOE restraints

APPENDIX

(including intra-residue) and with increase restraint weight. The backbone structure usually does not move significantly during this final refinement cycle. The median C_{α} -RMSD between models before and after this relaxation remained below 1 Å for all converged targets in the benchmark. Changes to the reported median C_{α} -RMSD against the reference structure were below <0.3 Å for all targets.

For refinement issue the command `post_relax` in the run folder of the selected run. This produces the output

```
[...]  
violations in      raw-decoys: 176  
violations in post-relaxed decoys: 63  
reduction of violations to 36%.
```

If the reduction of violations remains above 40% we recommend to not using the post-relaxed structures, but the un-relaxed models as final structures. This case occurred for 4 of 44 converged targets. 3 of these cases were *raw* data sets. The 4th was CtR107, where the reduction was only down to 63%.

A.3.2 Checking for unsuccessful calculations

For unsupervised methods it is important to raise flags if the final result is not solid. These flags should not be triggered too often, as then the method would be without value, but should be triggered as often as needed to not have inaccurate results slip through unnoticed.

A.3.2.1 Detecting failures in CYANA

Based on discussions with Peter Güntert we looked at the value of the target function and the level of convergence (measured by intrinsic RMSD) after cycle1 of a CYANA calculation to construct a criterion for failure. As shown in Figure S2-D, if either convergence is low (high RMSD) or the target function is high, the calculation is likely to produce inaccurate results (high RMSD to reference structure). A simple choice would be to determine an individual cutoff for both values and discard runs that fail either one. However, as seen in Figure S2-D, that would result in a poor discrimination. Hence, we introduced the linear separatrix described by $f(x, y) = 10^{-0.22\log_{10}(x)+0.95} - y$, where x denotes the target function, and y the backbone RMSD. Both values can be obtained by executing the command `cyanatable` in the directory of the CYANA calculation. Only if $f(x, y) > 0$ the run can be accepted.

APPENDIX

It is important to note, that the final value of the target function after cycle 7, is not informative towards the success of a calculation. Moreover, the final bundle of structures of failed calculations can be highly converged and yet inaccurate.

A.3.2.2 Detecting failures in AutoNOE-Rosetta

In contrast to CYANA we usually don't find highly converged structures that are inaccurate. Thus, we calculate the fraction of converged residues and divide by the fraction of residues that are expected to be rigid based on TALOS+ computed RCI S2 order parameter. If this ratio is above 0.9 the calculation is considered converged.

The only data set where this convergence criterion is fulfilled despite inaccurate structures is the raw peak list of StT322 from the CASDII set. In this case the topology of the structure is correct, but a 3.2Å C_{α} -RMSD to the reference is unacceptable. Generally, this data set seems particular in many ways. For most raw peak lists, we find about twice the number of peaks than for the refined peak list. However, for StT322 five times as many peaks were picked. Both methods, CYANA and AutoNOE, produce an unnaturally tight structural ensemble (intrinsic RMSD 0.0Å). Because all other CASD contributors reported problems with this data set, and also the reference structure has bad validation scores, the CASD community tentatively decided to remove this target from the CASD set.

Overall, it is not entirely clear which lessons should be drawn from the fact that AutoNOE-Rosetta converged on StT322(raw). As this is the single outlier with these peculiar characteristics (see above), it is difficult to give a definite answer what is wrong with this data set. However, fact remains, that this is a calculation, which requires further attention of an expert despite full convergence of the Rosetta calculation. Thus, we included a second criterion and calculated a target-function equivalent as follows: $MFS*STR/CST > 500$, where MFS is the lowest restraint score after StageIV, CST the restraint weight of the run (Section 1.10) and STR the strength of individual restraints. These values are obtained as follows: MFS) Relative to the run-folder of the selected run you'll find the file centroid_pool_stage4/decoys.out. Obtain the lowest score from the output of command silent_data decoys.out noesy_autoassign_cst. STR) in file initial_assignment/noe_auto_assign.cst this is the third parameter behind the keyword BOUNDED.

Using this criterion StT322 gets red-flagged with $MFS*STR/CST=1500$, whereas most other data sets yield values <100 . This criterion also raises the red-flag for YR313(raw), HR5460(raw), HmR11(unrefined) and ER690(unrefined). From these only YR313(raw) did actually not fail the convergence criterion. YR313(raw) is converged well and yielded a

formidable accuracy with 1.4Å C_{α} -RMSD to the reference structure. Similarly, CYANA runs were classified as failure, although an accurate structure was obtained (1.6Å). Thus, the second criterion serves to detect problematic (raw/unrefined) data sets, which deserve further attention. In the case of YR313 the final structures are accurate, and yet intervention of an expert is required to either improve the quality of the input data or to make the decision to ignore the red-flag.

A.3.3 Robustness of AutoNOE-Rosetta in repeated runs

AutoNOE-Rosetta uses non-deterministic sampling algorithms and thus might yield different results for the same input data. Moreover, small differences in the Rosetta energies of final models might lead to a switch between competing restraint weights, thus potentially further amplifying differences. To check that the reported results are robust we repeated all individual runs three times, and generated 1000 random sets that consist of a single AutoNOE-Rosetta calculation per restraint weight. The final run selection method is applied to these sets. This analysis shows that the results are robust for all successful runs and variations in RMSD are generally low between different sets of runs (*Appendix* Figure S8). Thus, we expect that a user would get results in accordance with the benchmark if a single calculation is performed at each recommended restraint weight and the final selection procedure is applied. The only striking variation between different randomly chosen sets of runs is seen for the data set, HR5460(raw). This target, however, is already considered as failure in the analysis and discussion of the benchmark results in the maintext. The final RMSD for this data set varies between 2Å and 8Å depending on the ability of AutoNOE-Rosetta to converge the C-terminal helix.

The final selection procedure can be applied to any number of runs and usually succeeds in selecting the runs with highest or near optimal accuracy. We recommend, however, keeping the NOE restraint strength parameter between 1 and 50, as smaller or larger parameters did not lead to improved results in our tests.

A.4 References

- Doreleijers JF, Vranken WF, Schulte C, et al. (2012) NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. *Nucleic Acids Res* 40:D519–24. doi: 10.1093/nar/gkr1134
- Huang YJ, Powers R, Montelione GT (2005) Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval

APPENDIX

statistics. *J Am Chem Soc* 127:1665–1674. doi: 10.1021/ja047109h

Lange OF, Rossi P, Sgourakis NG, 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 USA* 109:10873–10878. doi: 10.1073/pnas.1203013109

Mao B, Guan R, Montelione GT (2011) Improved Technologies Now Routinely Provide Protein NMR Structures Useful for Molecular Replacement. *Structure* 19:757–766. doi: 10.1016/j.str.2011.04.005

Mueller GA, Choy WY, Yang D, et al. (2000) Global folds of proteins with low densities of NOEs using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. *Journal of Molecular Biology* 300:197–212. doi: 10.1006/jmbi.2000.3842

Rosato A, Aramini JM, Arrowsmith C, et al. (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure* 20:227–236. doi: 10.1016/j.str.2012.01.002

Vernon R, Shen Y, Baker D, Lange OF (2013) Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker. *J Biomol NMR* 57:117–127. doi: 10.1007/s10858-013-9772-4

Acknowledgement

The work of this dissertation is helped and supported by a lot of people, to whom I would like to express the deepest appreciation.

First of all, I would like to thank my LORD Jesus Christ not only for leading my way to this research group, where I could study the knowledge of protein structures, mathematical algorithms, program languages and NMR spectroscopies, but also for blessing my 3 years' nice life in Munich.

My gratitude then goes to Dr. Oliver F. Lange for his strict and kind supervision to me. He accepted my application to be a Ph.D. student in his group in Chemistry department of TU Munich and provided me the project of protein structure determination by Rosetta and NMR experiments. At the beginning of my work, he taught me the basic knowledge of NMR experiments since this was a new field to me. When he was in the office, his door was always opened, and then I could ask him for help immediately when I met problems in my study. In the aspect of scientific writing and presentation, he also trained me patiently and improved my ability significantly.

I would also like to thank my thesis committee members, Prof. Dr. Michael Sattler and Dr. Tobias Madl for their precious suggestions to my work. Besides, Prof. Dr. Michael Sattler provided me many valuable opportunities to attend scientific seminars and conferences, which is quite helpful to my work.

I'm indebted to all members of computational structural biology group, Marcelino Arciniega Castro, Zhe Zhang, Tatjana Braun, Justin Porter, Shreyas Supekar, Atul Pawar and Alexej Grjasnow, for their help of technical problems, positive discussions and funny time.

I'm grateful to two cooperative groups, group of Prof. Dr. A.M.J.J. Bonvin in Utrecht University Netherlands, and group Prof. Dr. Gaetano T. Montelione in the State University of New Jersey USA, for pleasant cooperation with them. Together with Prof. Dr. A.M.J.J. Bonvin and his student Gijs van der Schot, we presented an improvement of protein structure prediction by CS-Rosetta in 2013. In 2014, we compared the performance of automatic NOE assignment and de novo programs along with Prof. Dr. Gaetano T. Montelione and his team member Fei Xu.

I would like to thank Technical University of Munich and German Research

Foundation for their financial support to this dissertation. They bring me chances to attend conferences in Germany or abroad too.

Lastly, I would like to thank my family for all their love and support. For my parents who raised me and encouraged me for many years. I cannot imagine, how my life would be like without my loving wife Jianing Gan. Thanks for all your faithful support during the final stages of this Ph. D. You have made my life a wonderful and positive experience.

Curriculum vitae

Zaiyong Zhang

Persönliche Daten

- Name: Zaiyong Zhang
- Adresse: Häberlstr. 11, 80337 München
- Geburtsdatum/-ort: 01.10.1988, Shandong China
- Nationalität: China

Ausbildung

Technische Universität München(DE) Doktorarbeit mit dem Titel: Rapid and automatic structure determination using sparse NMR data and Rosetta	09/2011-10/2014
Macau Universität(MO) Master of Science Elektromechanik	09/2009-08/2011
Nordwestlich Polytechnikum Universität (CN) Bachelor-Abschluss Maschinenbau	09/2009-08/2011

Veröffentlichung

- [1] G. Schot, **Z. Zhang**, R. Vernon, Y. Shen, W. F. Vranken, D. Baker, A. M. J. J. Bonvin, and O. F. Lange, "Improving 3D structure prediction from chemical shift data," *J Biomol NMR*, 2013.
- [2] **Z. Zhang**, J. Porter, and O. F. Lange, "Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta," *J Biomol NMR*, 2014.
- [3] M. Akimoto, **Z. Zhang**, S. Boulton, R. Selvaratnam, B. VanSchouwen, M. Gloyd, E. Accili, O. Lange and G. Melacini, "A Mechanism for the Auto-Inhibition of HCN Channel Opening and its Relief by cAMP," *Journal of Biological Chemistry*, 2014
- [4] W. Pak-kin, T. Lapmou, and **Z. Zhang**, "Ignition pattern analysis for automotive engine trouble diagnosis using wavelet packet transform and support vector machines," *Chinese Journal of Mechanical Engineering*, 2011.
- [5] P. K. Wong, C. M. Vong, **Z. Zhang**, and Q. Xu, "Fault Diagnosis of Automotive Engines Using Fuzzy Relevance Vector Machine," in *Communications in Computer and Information Science*, 2011.

