Methods for Automated Structure Determination by NMR Spectroscopy

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich
Biochemie, Chemie und Pharmazie
der Goethe-Universität
in Frankfurt am Main

von
Lena Buchner
aus Nisterau

Frankfurt am Main 2015
(D 30)

vom Fachbereich Biochemie, Chemie und Pharmazie der
Goethe-Universität Frankfurt am Main als Dissertation angenommen.

Dekan:	Prof. Dr. Michael Karas
1. Gutachter:	Prof. Dr. Peter Güntert
2. Gutachter:	Prof. Dr. Clemens Glaubitz
Datum der Disputation:	.....

dedicated to my mother

Contents

Summary

The knowledge of three-dimensional structures of biomolecules is fundamental for the understanding of their function. Nuclear magnetic resonance (NMR) spectroscopy represents besides X-ray crystallography one of the two most widely used techniques to study macromolecules at atomic resolution. Its application has long been a laborious task that could take months and required the expertise of an experienced scientist, however, owing to the tremendous effort that has been put into the development of respective computer algorithms, structure determination by NMR spectroscopy of small- to medium sized proteins is nowadays routinely performed. CYANA is one widely used software package, which combines the majority of individual steps towards a three-dimensional structure (Güntert, 2009; Güntert and Buchner, 2015). The most common application of the program, however, restricts to the combined automated NOE assignment and structure calculation based on NOESY peak lists and an existing chemical shift assignment. Completely automated structure determination starting from NMR spectra is to date technically possible with CYANA (López-Méndez and Güntert, 2006; Ikeya et al., 2009), however, not yet routinely applied. In order to achieve this long-term goal, the individual steps need to become more robust with regard to data imperfections such as peak overlap, spectral artifacts or a limited amount of NMR data.

The work presented in this thesis should be placed within the context of increasing the reliability and improving the accuracy of structures determined by CYANA on the basis of solution- as well as solid-state NMR data. The first project comprises an extensive study on the robustness of the combined automated NOE assignment and structure calculation algorithm based on experimental solution NMR data sets that were modified in several ways to mimic different kinds of data imperfections. Two additional projects represent methodological developments, i.e. the *Peakmatch* algorithm and a new protocol for combined automated NOE assignment and structure calculation, that aim to improve the input data quality and to increase the reliability of the structure calculation result, respectively. The last two projects are focused on structure determination by solid-state NMR and comprise a study on the impact of input data selection, including different labeling strategies, on the structure calculation result as well as a new method for spin

diffusion correction of experimental peak intensities, that aims to improve the quality of distance restraints obtained from NMR signals. The individual questions that have been addressed in this thesis and the respective results are summarized in the following.

The chapter "Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA" analyzes the performance of CYANA under a variety of conditions on the basis of the experimental NMR data sets of ten proteins. To evaluate the robustness of the algorithm, the original high-quality experimental data sets were modified in different ways to simulate the effect of data imperfections, i.e. incomplete or erroneous chemical shift assignments, missing NOESY cross peaks, inaccurate peak positions, inaccurate peak intensities, lower dimensionality NOESY spectra, and higher tolerances for the matching of chemical shifts and peak positions. The results show that the algorithm is remarkably robust with regard to imperfections of the NOESY peak lists and the chemical shift tolerances but susceptible to lacking or erroneous resonance assignments, in particular for nuclei that are involved in many NOESY cross peaks. The quality of a structure calculation result is evaluated as the average RMSD with respect to a known reference structure. Chemical shift omission rates of more than 15 % increase the average RMSD considerably above 3 Å indicating that structure calculations fail to converge to the correct global fold when using severely incomplete chemical shift data. The outcome in the range between 10 and 15 % chemical shift omission strongly depends on the protein and the quality of the respective NOESY data. In favorable cases, the correct structure can still be found with 20 % chemical shifts missing, whereas rather unfavorable cases may fail at 5 % missing chemical shifts. Compared to missing chemical shifts, deletion of NOESY peaks shows a less steep increase of the average RMSD. On average, the RMSD bias at 30 % deleted NOESY peaks is below 3 Å while the average RMSD rises slightly above 3 Å at 45 %. The much less pronounced increase can be explained by the fact that NOESY peaks contain rather redundant information through the dense NOE network, whereas one missing chemical shift leads to a whole set of NOESY peaks that remain unassigned in the more favorable case or get assigned incorrectly in the less favorable case.

Throughout this study, the structural accuracy was assessed as the RMSD with respect to the known reference structure. This is, however, not possible in case of *de novo* structure determinations without any knowledge about the true structure. For this reason, we have additionally investigated several criteria to assess the accuracy of a structure calculation result in a reference structure independent way. Two previously reported criteria comprise the convergence of the initial structure calculation cycle and the RMSD drift between the first and the last cycle. Our results suggest that the reliability of these measures can be significantly increased if they are combined in a weighted average which is thus recommended to be used as an indication for the quality of a structure calculation result.

In the chapter "*Peakmatch* – A simple and robust tool for peaklist matching" a method to achieve self-consistency of the chemical shift referencing among a set of peak lists is presented. The *Peakmatch* algorithm matches a set of peak lists to a specified reference peak list, neither of which have to be assigned. The chemical shift referencing offset between two peak lists is determined by optimizing an assignment-free match score function using either a complete grid search or downhill simplex optimization. The algorithm has been extensively tested on the basis of experimental NMR data sets of five different proteins. Each data set included typical backbone experiments for resonance assignment as well as through-space experiments for structure calculation. The peak lists of each data set were obtained from automatic peak picking and, in addition, manually refined peak lists were available for three of the five data sets. The results show that peak lists from many different types of spectra can be matched reliably as long as they contain at least two corresponding dimensions. Using a simulated peak list based on a given chemical shift list, the *Peakmatch* algorithm can also be used to obtain the optimal agreement between a chemical shift list and the corresponding experimental peak lists. Combining these features makes *Peakmatch* a useful tool that can be applied routinely before automated assignment or structure calculation in order to obtain an optimized input data set.

NMR structures are represented by bundles of conformers calculated from different randomized initial structures using identical experimental input data. The spread among these conformers indicates the precision of the atomic coordinates. However, there is as yet no reliable measure of structural accuracy, i.e. how close NMR conformers are to the "true" structure. Instead, the precision of structure bundles is widely (mis)interpreted as a measure of structural quality. Attempts to increase the precision thus often yield tight structure bundles where the precision overestimates the accuracy. To overcome this problem, the chapter "Increased reliability of NMR protein structures by consensus structure bundles" introduces a new protocol for NMR structure determination with the software package CYANA that produces, like the traditional method, bundles of conformers in agreement with a common set of conformational restraints, however with a realistic precision.

The algorithm performs 20 independent automated NOESY assignment and structure calculation runs using the same input data and different random number generation seeds, resulting in 20 individual structure bundles. The lowest energy structure of each of these 20 structure bundles is combined to obtain a new combined structure bundle. The precision of the combined structure bundle is a measure of the extent to which individual calculations differ from each other. Each of the 20 individual structure calculations leads to a different set of distance restraints as a result of the seven cycles of NOE assignment and structure calculation. These individual final sets of distance restraints are in optimal agreement

with the respective structure bundle, however, they do not represent the aforementioned combined structure bundle. The combination of the individual sets of distance restraints yields a consensus set of distance restraints that results in a structure bundle similar to the combined structure bundle when used as input for a further structure calculation. This final structure calculation is a simple standard CYANA structure calculation without automatic NOE assignment. It uses the consensus NOE distance restraints (and other conformational restraints, if available) as input and yields the consensus structure bundle as output.

The method has been extensively tested on the basis of the ten experimental data sets including all types of data modifications presented in the previous chapter "Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA" as well as eight experimental data sets provided as test data sets for the CASD-NMR project in 2011-2012 (Rosato et al., 2009; Rosato et al., 2012). The results show that the precision of the consensus structure bundle is throughout a variety of proteins and NMR data sets, a much better estimate of structural accuracy than the precision of conventional structure bundles.

Solid-state NMR is a powerful technique to study molecules which are not amenable to either solution NMR or X-ray crystallography. Two classes of these macromolecules, which are of special interest for medical questions, are membrane proteins in their native lipid environment and amyloid fibrils. Despite the reporting of individual atomic resolution structures of membrane proteins (Tang et al., 2013; Wang et al., 2013) and amyloid fibrils (Wasmer et al., 2008; Melckebeke et al., 2010; Lu et al., 2013; Schütz et al., 2014) based on solid-state NMR data, the application is far from routine. One major obstacle that hinders structure determination by solid-state NMR is the overall lower quality of the spectra, that can be attributed to the limited averaging of anisotropic interactions as well as the lower signal-to-noise ratio that prevents the recoring of 3D spectra in many cases. Further developments are therefore required on the spectroscopic side to improve spectral quality, however, it is also necessary to increase the robustness of the computer algorithms in order to improve the results when using lower quality solid-state NMR spectra.

CYANA can use common solid-state NMR experiments as input for combined automated NOE assignment and structure calculation. However, there is no systematic investigation about the structural quality that can be achieved using this method based on common solid-state NMR experiments . The chapter "Structure calculations of the model protein GB1 from solid-state NMR data" therefore presents structure calculations on the basis of a set of two-dimensional solid-state NMR experiments of the model protein GB1. In order to investigate the impact of different labeling strategies, NMR spectra have been recorded on three differently labeled GB1 samples based on uniformly ^{13}C-labeled glucose

(u-^{13}C/^{15}N GB1), 2-^{13}C-labeled glycerol (2-^{13}C/^{15}N GB1), and 1,3-^{13}C-labeled glycerol (1,3-^{13}C/^{15}N GB1). The final data set thus included a total of 10 two-dimensional ^{13}C-^{13}C correlation experiments recorded using different pulse sequences for magnetization exchange (e.g. DARR, PAR, CHHC) as well as different mixing times. Structure calculations were carried out for different combinations of input peak lists using the standard CYANA algorithm for combined automated NOE assignment and structure calculation. The first important finding is that it is in principle possible to reproducibly obtain 3D structures where the overall global fold is correct and inaccuracies occur mostly on the local scale, provided that spectra of GB1 samples with diluted labeling are included in the calculation. Diluted labeling strategies are beneficial to improve the spectral quality by reducing the overlap and thus increasing the number of visible long-range signals. Structure calculations from only u-^{13}C/^{15}N GB1 spectra do in the present case not yield the correct global fold. Attempts to further improve the structural quality have been made using a reference peak assignment, thus excluding potentially distorting effects arising from incorrect peak assignments. Structure calculations have then been performed as a basic CYANA structure calculation based on distance restraints that have been obtained from different methods for upper distance limit calibration. The most important result obtained from these test calculations is that, despite using a reference peak assignment lacking incorrect peak assignments, the structural accuracy is limited to ~1.5 Å, a quality range which is somewhat lower than expected based on typical results from solution NMR data of similar systems. The findings furthermore suggest that the limitation of structural accuracy can be attributed to inaccurate upper distance limits resulting from the limited correlation between peak intensities and distance, which is especially severe in spin diffusion-based solid-state NMR experiments.

The chapter “Full relaxation matrix-based correction of relayed polarization transfer for solid-state NMR structure calculation” therefore introduces a method which corrects experimental peak intensities for spin diffusion in order to improve the resulting upper distance limits. The method relies on a full relaxation matrix approach which predicts peak intensities based on a three-dimensional input structure, which can be the result of a conventional structure calculation, a homology model, or a structure determined by X-ray crystallography. These simulated peak intensities are subsequently used to calculate a correction factor which is applied to the corresponding experimental peak intensity. The corrected peak intensities are then recalibrated and used for an additional simple structure calculation based on distance restraints. In case of a *de novo* structure determination without prior knowledge about the true structure, the concept constitutes an iterative application of the correction procedure, initially starting from the result of a conventional structure calculation.

In order to investigate the potential improvement that can be achieved when applying the correction procedure, two-dimensional solution-NMR and solid-state NMR spectra of the protein ubiquitin have been simulated. Simulation of NMR spectra has the advantage that spectral properties can be exactly controlled, which allows a detailed investigation of the differences between solution NMR and solid-state NMR with emphasis on structure determination. Conventional structure calculations based on simulated two-dimensional solution and solid-state NMR spectra confirm the finding of the previous chapter that the limited correlation between peak intensity and distance is the major obstacle restricting the structural quality in the case of spin diffusion-based solid-state NMR data. The potential improvement obtained from the new correction procedure is investigated using different types of input structures for the relaxation matrix calculation. The results reveal a significant improvement if the known reference structure determined by X-ray crystallography is used as input. However, the result strongly depends on the quality and the type of input structure. The original concept of iterative application of the correction procedure did not turn out to be successful as the input structure obtained from a conventional structure calculation based on uncorrected input data can in fact improve the correlation between peak intensity and distance but the resulting 3D structure shows no improvement with respect to the input structure. This can most likely be attributed to the fact that structural distortions in the input structure bundle are homogeneous, thus affecting the correction procedure such that the corrected peak intensities still reflect the structural error.

Altogether, the results show that the presented correction method is ill-suited for routine application in *de novo* structure determinations due to the strong dependence on the input structure. In order to improve the quality of structures from solid-state NMR, it is therefore necessary to either develop NMR experiments that intrinsically possess more accurate distance information, or more robust methods for the correction of peak intensities need to be developed that are less dependent on preliminary structural information.

Zusammenfassung

Die Kenntnis der dreidimensionalen Struktur von Biomolekülen ist entscheidend für das Verständnis ihrer Funktion. Magnetische Kernspinresonanz (nuclear magnetic resonance; NMR) Spektroskopie stellt neben der Röntgenstrukturanalyse die wichtigste Methode zur Untersuchung von Makromolekülen mit atomarer Auflösung dar. Die Anwendung der NMR Spektroskopie zur Strukturaufklärung war lange Zeit eine Herausforderung, die einen Zeitraum von mehreren Monaten in Anspruch nehmen konnte und zudem die Expertise eines erfahrenen Spektroskopikers erforderte. Nicht zuletzt durch den großen Aufwand, der in die Entwicklung entsprechender Computeralgorithmen gesteckt wurde, ist es heutzutage möglich, Strukturen von kleinen bis mittelgroßen Proteinen routinemäßig in vergleichsweise kurzer Zeit zu berechnen. Das Softwarepaket CYANA (Güntert, 2009; Güntert and Buchner, 2015) ist dabei eines der meist verwendeten Programme, welches die Mehrzahl der notwendigen Schritte bis zur 3D Struktur automatisch ausführt. Die gebräuchlichste Verwendung des Programms beschränkt sich allerdings auf die Kombination aus automatischer nuklearer Overhauser-Effekt (NOE) Zuordnung und Strukturrechnung. Komplett automatisierte Strukturbestimmung basierend auf NMR Spektren ohne manuelle Intervention ist heute mit CYANA technisch möglich (López-Méndez and Güntert, 2006; Ikeya et al., 2009), wird jedoch bislang nicht routinemäßig durchgeführt. Um dieses langfristige Ziel zu erreichen, müssen die einzelnen Schritte robuster im Umgang mit nicht perfekten experimentellen Daten, so zum Beispiel Signalüberlappung, spektrale Artefakte oder limitierte Datenmenge, werden und es werden weiterhin verlässliche Kriterien zur Bewertung der Strukturrechnungsergebnisse benötigt.

Die Fragestellungen der vorliegenden Dissertation dienen alle dem Ziel, die Verlässlichkeit von NMR Strukturen zu erhöhen und deren Qualität zu verbessern. Das erste Projekt stellt eine umfassende Studie zur Anfälligkeit des in CYANA implementierten Strukturrechnungsalgorithmus in Bezug auf fehlerbehaftete Daten dar, die basierend auf experimentellen Lösungs-NMR Datensätzen auf verschiedene Arten modifiziert wurden, um gezielt diverse Fehlerquellen zu simulieren. Zwei weitere Projekte stellen methodische Entwicklungen dar, die zum einen die Qualität der experimentellen Daten und zum anderen die Verlässlichkeit der Strukturrechnungsergebnisse verbessern sollen. Der *Peak-*

match Algorithmus optimiert die relative Referenzierung verschiedener Signallisten um so die Zuordnung der Signale zu Atomen zu verbessern. Die Entwicklung der *consensus* Strukturbündel beinhaltet ein neues Protokoll für die kombinierte automatische NOE Zuordnung und Strukturrechnung mit CYANA, die aus 20 individuellen Strukturrechnungen sowohl *consensus* Zuordnungen als auch ein *consensus* Strukturbündel erstellt. Die letzten beiden Projekte dieser Arbeit beschäftigen sich mit der Strukturbestimmung basierend auf Festkörper NMR Daten. Das erste untersucht den Einfluss der Datenauswahl sowie die Bedeutung verschiedener Isotopenmarkierungsschemata auf die Strukturrechnungsergebnisse, während das zweite Projekt eine Methode zur Korrektur von Spindiffusion vorstellt. Die einzelnen Fragestellungen und deren Ergebnisse werden im Folgenden genauer erläutert.

In dem Kapitel "Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA" wird eine ausgedehnte Studie zur Anfälligkeit von CYANA in Bezug auf unterschiedliche Fehlerquellen vorgestellt. Die Strukturrechnungen wurden auf der Basis von zehn experimentellen Lösungs-NMR Datensätzen durchgeführt, die auf verschiedene Arten so modifiziert wurden, dass der Einfluss individueller Fehlerquellen untersucht werden konnte. Dazu gehörten unvollständige oder fehlerhafte Resonanzzuordnungen, fehlende NOESY Signale, ungenaue Signalpositionen, ungenaue Signalintensitäten, niedrigere Dimensionalität der NMR Spektren, sowie höhere Toleranzen für das Abgleichen der Signalposition und der chemischen Verschiebung der Atome. Die Ergebnisse zeigen, dass der Algorithmus bemerkenswert robust mit fehlerhaften oder unvollständigen NOESY Signallisten sowie hohen Zuordnungstoleranzen umgehen kann, jedoch anfällig für fehlende oder fehlerbehaftete Resonanzzuordnungen ist, insbesondere bei Atomen, die zu vielen Signalen beitragen. Die Qualität der Strukturrechnungsergebnisse wird durch den mittleren RMSD zur bekannten Referenzstruktur gemessen. Fehlende Resonanzzuordnungen von mehr als 15 % erhöhen den über alle zehn Proteine gemittelten RMSD deutlich über 3 Å. Da in diesem Qualitätsbereich die korrekte Faltung des Proteins nicht mehr gefunden wird, spricht das Ergebnis dafür, dass bereits 15 % fehlende Resonanzzuordnungen für den Algorithmus ein massives Problem darstellen. Das Ergebnis im Bereich zwischen 10 und 15 % hängt stark vom Protein und der Qualität des entsprechenden Datensatzes ab. In günstigen Fällen kann die korrekte Faltung noch mit 20 % fehlenden Resonanzzuordnungen gefunden werden, während in weniger günstigen Fällen bereits 5 % fehlende Resonanzzuordnungen zu einem fehlerhaften Endergebnis führen. Im Mittel bleibt der RMSD Wert bei 30 % fehlenden NOESY Signalen hingegen unter 3 Å, während er bei 45 % nur leicht auf über 3 Å ansteigt. Der wesentlich flachere Anstieg des RMSD Werts als Folge fehlender NOESY Signale lässt sich durch die Tatsache erklären, dass diese durch das dichte NOE Netzwerk eher redundante Information enthalten und die Informa-

tion eines fehlendes Signals in den meisten Fällen durch ein weiteres Signal ausgeglichen werden kann. Im Gegensatz dazu führt eine fehlende Resonanzzuordnung zu einer fehlenden oder falschen Zuordnung aller NOESY Signale, die von dem entsprechenden Atom stammen und diese fehlende Information kann nicht ausgeglichen werden.

Im Zuge der gesamten Studie wurde die Strukturqualität basierend auf dem RMSD zur bekannten Referenzstruktur bewertet. Dies ist jedoch im Fall einer *de novo* Strukturbestimmung ohne bekannte Referenzstruktur nicht möglich. Aus diesem Grund haben wir weiterhin die Verlässlichkeit zweier bereits früher vorgeschlagener Kriterien zur Referenzstrukturunabhängigen Bewertung von Strukturrechnungsergebnissen untersucht. Das erste Kriterium stellt die Konvergenz der Strukturrechnung im ersten Strukturrechnungszyklus dar, bei dem zweiten Kriterium handelt es sich um den RMSD *drift* vom ersten zum letzten Strukturrechnungszyklus. Die Ergebnisse zeigen deutlich, dass die Kombination beider Kriterien in ein gewichtetes Mittel die Verlässlichkeit der Qualitätsvorhersage deutlich verbessern kann. Daher wird die Verwendung dieses neuen Kriteriums zur Abschätzung der Strukturqualität generell empfohlen.

In dem Kapitel "*Peakmatch* – A simple and robust tool for peaklist matching" wird eine neue Methode vorgestellt, die dazu dient, die interne Referenzierung verschiedener Signallisten aufeinander abzustimmen. Der *Peakmatch* Algorithmus verschiebt eine Signalliste auf eine gegebene Referenzsignalliste, wobei keine Zuordnung der Signale vorausgesetzt wird. Der *Offset* dieser zwei Signallisten wird durch Optimierung einer zuordnungsunabhängigen Funktion erreicht, die entweder mithilfe einer kompletten Gittersuche durchgeführt werden kann oder mithilfe einer *downhill simplex* Optimierungsstrategie. Beide Optimierungsstrategieren wurden ausgiebig basierend auf experimentellen NMR Datensätzen von fünf unterschiedlichen Proteinen getestet. Jeder Datensatz umfasste dabei Signallisten typischer Experimente für Rückgrat-Resonanzzuordnung sowie solche für Strukturbestimmung. Signallisten wurden mithilfe von Programmen für automatische Signalidentifizierung erstellt, wobei zusätzlich im Fall von drei Datensätzen manuell verfeinerte Signallisten vorhanden waren. Die Ergebnisse zeigen, dass Signallisten von vielen verschiedenen Spektren verlässlich aufeinander abgestimmt werden können, solange zwei korrespondierende Dimensionen vorhanden sind. Mithilfe einer simulierten Signalliste basierend auf einer gegebenen Resonanzzuordnung kann außerdem eine optimale Übereinstimmung der experimentellen Signallisten mit der gegebenen Resonanzzuordnung hergestellt werden. Diese Eigenschaften machen den *Peakmatch* Algorithmus zu einem sinnvollen Werkzeug, das routinemäßig vor einer Strukturrechnung oder einer automatischen Resonanzzuordnung angewendet werden kann.

NMR Strukturen werden als Bündel von Konformeren dargestellt, die, ausgehend von verschiedenen zufälligen Startstrukturen, basierend auf identischen experimentellen Daten

berechnet werden, wobei die Abweichung dieser Konformere voneinander ein Maß für die Streuung der Atomkoordinaten darstellt. Allerdings gibt es bislang kein verlässliches Maß für die Korrektheit eines NMR Strukturbündels, das heißt, wie ähnlich die Strukturen des Bündels der "echten" Struktur sind. Stattdessen wird die Streuung des Strukturbündels weitläufig als Maß für die Qualität einer Struktur (miss)interpretiert. Bestrebungen, die Präzision zu erhöhen, führen daher häufig zu sehr engen Strukturbündeln, deren Präzision die Korrektheit häufig überschätzt. Um dieses Problem zu umgehen, wird in dem Kapitel "Increased reliability of NMR protein structures by consensus structure bundles" ein neues Protokoll für die automatische NOE Zuordnung und Strukturrechnung mit CYANA vorgestellt, das wie die ursprüngliche Methode ein Bündel von Konformeren erzeugt, das mit einem gemeinsamen Satz von Distanzeinschränkungen übereinstimmt, aber eine Streuung aufweist, die ein realistisches Maß für die Genauigkeit der Struktur ist.

Der Algorithmus führt 20 unabhängige automatische NOE Zuordnungen und Strukturrechnungen basierend auf den gleichen experimentellen Daten aber verschiedenen Startwerten für die Erzeugung von Zufallszahlen aus. Daraus ergeben sich 20 individuelle Strukturbündel, wobei die Strukturen mit der niedrigsten CYANA Zielfunktion aus jedem Bündel zu einem neuen kombinierten Bündel vereinigt werden. Die Streuung des kombinierten Bündels ist ein Maß für die Abweichung der einzelnen Strukturrechnungen. Jede der 20 individuellen Strukturrechnungen resultiert in einem unterschiedlichen Satz von Distanzeinschränkungen als Ergebnis der sieben Zyklen automatischer NOE Zuordnung und Strukturrechnung. Diese individuellen finalen Sätze von Distanzeinschränkungen sind in Übereinstimmung mit dem jeweiligen zugehörigen Strukturbündel, sie repräsentieren jedoch nicht das genannte kombinierte Strukturbündel. Aus diesem Grund werden die einzelnen Sätze von Distanzeinschränkungen zu einem *consensus* Satz vereinigt, der für eine weitere einfache CYANA Strukturrechnung ohne automatische NOE Zuordnung eingesetzt werden kann und so ein *consensus* Strukturbündel erzeugt, welches dem kombinierten Bündel ähnelt. Der Vorteil des *consensus* Bündels gegenüber dem kombinierten Bündel ist jedoch, dass es mit einem einzigen Satz von Distanzeinschränkungen übereinstimmt.

Das neue Protokoll wurde ausgiebig auf der Basis der zehn experimentellen Datensätze inklusive sämtlicher Arten von Datenmodifikationen aus dem Kapitel "Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA" getestet, sowie auf der Basis von acht weiteren Datensätzen, die als Test-Datensätze für das CASD-NMR Projekt in 2011-2012 (Rosato et al., 2009; Rosato et al., 2012) zur Verfügung gestellt wurden. Die Ergebnisse zeigen, dass die Genauigkeit des *consensus* Strukturbündels über eine große Anzahl von Proteinen und NMR Datensätzen ein deutlich besseres

Maß für die Richtigkeit darstellt als die Genauigkeit eines konventionell gerechneten Strukturbündels.

Festkörper NMR Spektroskopie ist eine leistungsfähige Methode, um Moleküle zu untersuchen, die weder der Lösungs-NMR Spektroskopie noch der Röntgenstrukturanalyse zugänglich sind. Zwei dieser Makromolekülklassen, die besondere Relevanz für medizinische Fragestellungen haben, sind Membranproteine in ihrer nativen Phospholipidumgebung sowie amyloide Fibrillen. Trotz der Veröffentlichung einzelner hochaufgelöster Strukturen von Membranproteinen (Tang et al., 2013; Wang et al., 2013) sowie amyloiden Fibrillen (Wasmer et al., 2008; Melckebeke et al., 2010; Lu et al., 2013; Schütz et al., 2014) basierend auf Festkörper NMR Daten, ist die Anwendung der Methode heute nach wie vor keine Routine. Eine der Hauptursachen ist die generell niedrigere Qualität der Festkörper NMR Spektren, die hauptsächlich auf die unvollständige Ausmittelung anisotroper Wechselwirkungen sowie das niedrigere Signal-zu-Rauschen Verhältnis, welches in vielen Fällen die Aufnahme von 3D Spektren verhindert, zurückzuführen ist. Weitere Entwicklungen werden daher sowohl auf Seite der NMR Experimente benötigt, um die Qualität der Spektren zu verbessern, als auch auf Seite der entsprechenden Computer Algorithmen, um deren Leistungsfähigkeit im Umgang mit niedrigerer spektraler Qualität zu verbessern.

CYANA bietet die Möglichkeit, Standard Festkörper NMR Experimente für die kombinierte automatische NOE Zuordnung und Strukturrechnung zu verwenden. Allerdings gibt es bisher keine genauen Untersuchungen zu den Ergebnissen, die mit dieser Methode erzielt werden können. Daher werden im dem Kapitel “Structure calculations of the model protein GB1 from solid-state NMR data” Strukturrechnungen mit der Standard CYANA Methode basierend auf einer Auswahl von experimentellen zwei-dimensionalen Festkörper NMR Spektren des Modellproteins GB1 vorgestellt. Um den Einfluss verschiedener Isotopenmarkierungsschemata zu untersuchen, wurden NMR Spektren an drei unterschiedlich markierten Proben basierend auf uniform ^{13}C markierter Glucose (u-^{13}C/^{15}N GB1), 2-^{13}C markiertem Glycerol (2-^{13}C/^{15}N GB1), sowie 1,3-^{13}C markiertem Glycerol (1,3-^{13}C/^{15}N GB1) aufgenommen. Der finale Datensatz umfasste 10 zwei-dimensionale ^{13}C-^{13}C Korrelationsexperimente, die mit verschiedenen Pulssequenzen für den Magnetisierungsaustausch (z.B. DARR, PAR, CHHC) sowie verschiedenen Mischzeiten aufgenommen wurden. Strukturrechnungen wurden für verschiedene Kombinationen der verfügbaren NMR Spektren mithilfe des kombinierten automatischen NOE Zuordnungs- und Strukturrechnungsalgorithmus in CYANA durchgeführt. Die erste wichtige Erkenntnis aus diesen Ergebnissen ist, dass es reproduzierbar möglich ist ohne manuelle Intervention basierend auf Festkörper NMR Daten eine 3D-Struktur zu berechnen, deren globale Faltung korrekt ist und Ungenauigkeiten hauptsächlich lokal auf Ebene der Seitenketten auftreten, vorausgesetzt, dass NMR Spektren von verdünnten Markierungsschemata ver-

wendet werden. Verdünnte Markierungsschemata dienen der Verbesserung der spektralen Qualität durch Reduktion des Überlapps, wodurch die Anzahl der sichtbaren langreichweitigen Signale erhöht wird. Strukturrechnungen ausschließlich basierend auf Spektren von uniform markiertem GB1 haben im untersuchten Fall nicht die korrekte Faltung ergeben. Um die Strukturqualität zu verbessern, wurde eine Referenzzuordnung der Signale aller vorhanden Signallisten erstellt, wodurch der potentiell negative Einfluss fehlerhafter Zuordnungen durch die automatische NOE Zuordnung verhindert werden kann. Strukturrechnungen wurden anschließend basierend auf den zugeordneten Signallisten mit verschiedenen Methoden zur Kalibrierung von oberen Distanzschranken durchgeführt. Die wichtigste Erkenntnis aus diesen Rechnungen ist, dass die erreichbare Qualität trotz korrekter Signalzuordnungen auf einen RMSD Wert von ~1,5 Å beschränkt ist. Diese Qualität ist niedriger, als es basierend auf typischen Strukturrechnungsergebnissen ausgehend von Lösungs-NMR Daten vergleichbarer Systeme zu erwarten wäre. Die Ergebnisse deuten weiterhin darauf hin, dass ungenaue obere Distanzschranken, die aus der geringen Korrelation zwischen Signalintensität und Distanz, insbesondere im Fall von Spindiffusions-basierten Experimenten, resultieren, die Ursache für die limitierte Strukturqualität darstellen.

Das Kapitel “Full relaxation matrix-based correction of relayed polarization transfer for solid-state NMR structure calculation” präsentiert daher eine Methode zur Spindiffusions-Korrektur der experimentellen Signalintensitäten, um die Qualität der kalibrierten oberen Distanzschranken zu verbessern. Die Methode beruht auf einem Relaxationsmatrix Ansatz, der ausgehend von einer 3D-Struktur Signalintensitäten simuliert. Die 3D-Struktur kann beispielsweise das Ergebnis einer konventionellen Strukturrechnung sein, ein Homologie-Modell, oder eine Röntgenstruktur. Die simulierten Signalintensitäten werden anschließend verwendet, um für jedes gemessene Signal einen Korrekturfaktor zu berechnen, der auf die experimentelle Signalintensität angewendet wird. Die korrigierten experimentellen Signalintensitäten können schließlich erneut kalibriert werden und für eine weitere Strukturrechnung eingesetzt werden. Im Falle einer *de novo* Strukturbestimmung ohne Kenntnis der richtigen 3D-Struktur, kann die Korrektur iterativ ausgehend von einer konventionell berechneten Struktur angewendet werden. Das Ergebnis der Rechnung mit korrigierten Daten kann anschließend erneut für die Korrektur eingesetzt werden.

Um die mögliche Verbesserung der Strukturrechnungsergebnisse durch die vorgestellte Methode zur Spindiffusionskorrektur zu untersuchen, wurden zwei-dimensionale Lösungs- sowie Festkörper NMR Spektren von Ubiquitin simuliert. Die Simulation von NMR Spektren hat dabei den Vorteil, dass sämtliche spektrale Eigenschaften genau kontrolliert werden können, sodass es möglich ist, den Unterschied zwischen Lösungs-NMR und Festkörper NMR in Bezug auf die Strukturrechnung genau zu untersuchen. Konventionelle Struktur-

rechnungen basierend auf den simulierten Lösungs- sowie Festkörper NMR Spektren von Ubiquitin bestätigen die Ergebnisse des vorherigen Kapitels, dass die geringe Korrelation zwischen Signalintensität und Distanz in Spindiffusions-basierten Experimenten die Hauptursache für die limitierte Strukturqualität im Fall von Festkörper NMR Daten ist. Die potentielle Verbesserung durch die neu eingeführte Korrekturmethode wurde basierend auf verschiedenen Eingabe-Strukturen für die Berechnung der Relaxationsmatrix getestet. Die Ergebnisse zeigen, dass die Strukturqualität signifikant verbessert werden kann, wenn die Ubiquitin Röntgenstruktur als Referenzstruktur für die Korrekturfaktorberechnung eingesetzt wird. Allerdings hängt das Ergebnis der Korrektur stark von der Qualität der Eingabe-Struktur ab. Der ursprüngliche Vorschlag, die Methode iterativ einzusetzen, hat sich als nicht erfolgversprechend herausgestellt, da das Ergebnis einer konventionellen Strukturrechnung sich nicht als Eingabe für die Korrekturmethode zu eignen scheint. Obwohl eine scheinbare Verbesserung der Korrelation zwischen Signalintensität und Distanz nach der Korrektur zu beobachten ist, zeigt das Ergebnis der anschließenden Strukturrechnung keine Verbesserung. Als potentielle Erklärung für dieses Verhalten wird angenommen, dass die Fehler einer solchen Eingabestruktur so homogen im Strukturbündel vorhanden sind, dass auch die korrigierten experimentellen Signalintensitäten diesen Fehler wiederspiegeln.

Alles in allem zeigen diese Ergebnisse, dass sich die Korrekturmethode in der aktuellen Form aufgrund der aufgeführten Probleme nicht für eine reguläre Anwendung eignet. Um die Qualität von Strukturen basierend auf Festkörper NMR Daten zu verbessern, ist es folglich nötig, entweder neue Experimenttypen zu entwickeln, die intrinsisch eine höhere Korrelation zwischen Signalintensität und Distanz aufweisen, oder die Entwicklung robusterer Korrekturmethoden voranzutreiben, die unabhängiger von gegebener Strukturinformation sind.

Part I

Introduction

Chapter 1

NMR spectroscopy

1.1 Overview

Nuclear magnetic resonance (NMR) spectroscopy is based on the fact that certain nuclei possess an intrinsic source of angular momentum called the nuclear spin angular momentum (Richard R. Ernst and Wokaun, 1987). The z-component of the nuclear spin, described by the operator $\hat{I}_z$, interacts with an applied magnetic field B_0. The energy of a nuclear spin experiencing an external magnetic field along the z-axis is described by the following Hamiltonian operator (Equation 1.1),

$$\hat{H}_{\text{Zeeman}} = -\gamma B_0 \hat{I}_z \tag{1.1}$$

where γ represents the gyromagnetic ratio of the nucleus and B_0 the external magnetic field strength. The Hamiltonian $\hat{H}$ is the energy operator and thus eigenfunctions as well as eigenvalues describing the energy states can be determined. The number of eigenfunctions and eigenvalues depends on the nuclear spin angular momentum quantum number I via $2I+1$, where I can take integer ($I = 0, 1, 2, ...$) or half-integer ($I = \frac{1}{2}, \frac{3}{2}, ...$) values. Every nucleus is characterized by a specific quantum number I, which determines the magnetic properties of the respective nucleus. NMR uses spin-half nuclei (i.e. $I = \frac{1}{2}$), where the Hamiltonian $\hat{H}_{\text{Zeeman}}$ has two eigenfunctions and eigenvalues describing two energy levels in the presence of an external magnetic field B_0. Nuclei with $I = 0$ have no magnetic moment which interacts with an external magnetic field and those with $I = 1$ (quadrupole nuclei) show very fast relaxation as well as broad signals, which makes them ill-suited for NMR measurements. Typical spin-half nuclei in biological molecules comprise ^{1}H, ^{13}C and ^{15}N. As ^{13}C and ^{15}N have a low natural abundance, it is necessary to perform isotopic labeling in order to measure these nuclei. Each of the two energy levels is characterized by a quantum number m, which again depends on I in the way that m takes values in

the range from $-I$ to I in integer steps (i.e. $m = -\frac{1}{2}$ and $m = +\frac{1}{2}$ for spin-half nuclei). The two eigenvalue equations for $\hat{I}_z$ of a spin-half nucleus are given in Equation 1.2, where $m = \pm\frac{1}{2}$ and Ψ_m is the wave function.

$$\hat{I}_z \Psi_m = m\hbar\Psi_m \tag{1.2}$$

The two eigenvalues for the operator $\hat{I}_z$ are $+\frac{1}{2}\hbar$ and $-\frac{1}{2}\hbar$. Combining Equation 1.1 and the two eigenvalues of $\hat{I}_z$ leads to the two eigenvalues of the Hamiltonian $\hat{H}_{\text{Zeeman}}$, which describe the energies of the two states that a spin-half nucleus can adopt in the presence of an external magnetic field (Equation 1.3).

$$E_\alpha = -\frac{1}{2}\hbar\gamma B_0 \qquad\qquad E_\beta = +\frac{1}{2}\hbar\gamma B_0 \tag{1.3}$$

The two states are commonly labeled α for the state corresponding to $m = +\frac{1}{2}$ and β for the state corresponding to $m = -\frac{1}{2}$. The energy difference between the two states ΔE corresponds to the frequency at which transitions between the two energy levels can occur (Equation 1.4).

$$\Delta E_{\alpha\rightarrow\beta} = \hbar\gamma B_0 \tag{1.4}$$

The energy difference between the two states expressed in frequency units ($rad\ s^{-1}$) is called *Larmor frequency* and is characteristic for every spin in an external magnetic field of a certain field strength B_0 (Equation 1.5).

$$\omega_0 = -\gamma B_0 \tag{1.5}$$

When a population of nuclear spins is placed in an external magnetic field, the individual spins interact with the magnetic field and it becomes energetically more favorable for a spin to align with the field (i.e. being in the α-state). However, the energy of thermal motion is greater than the interaction with the magnetic field and the alignment is consequently disrupted. Although the orientation of individual spins is rather random due to thermal motion, there is still a small preference to align with the field, giving rise to a bulk magnetization of the complete sample along the direction of the B_0 field (z-axis), which is not present along any other axis. After the magnetic field is applied, the bulk magnetization starts to build up until an equilibrium is reached (equilibrium bulk magnetization).

The process which brings the system back to equilibrium after any kind of disturbance is called *relaxation*. The bulk magnetization can be described by a vector model. During equilibrium, the vector describing the bulk magnetization has a defined size and direction along the z-axis, which is not changing with time. However, if the system is not at equilibrium (i.e. away from the z-axis), the bulk magnetization vector rotates around the z-axis at Larmor frequency, a process called *precession*.

The Larmor frequency is directly proportional to the magnetic field strength B_0 (Equation 1.5). However, nuclei within a molecule do not experience B_0, but a local field B_{loc}, which is influenced by the chemical environment of the nucleus. The electrons have a shielding effect, as the magnetic field B_0 induces a ring current and thus a magnetic moment, which operates in opposite direction of B_0. Consequently, the chemical environment can influence the local field experienced by a nucleus i and can thus change the Larmor frequency ω_i (Equation 1.6). The shielding constant σ describes the influence of the respective chemical environment.

$$\omega_i = -\gamma_i B_0(1 - \sigma_i) = \omega_0(1 - \sigma_i) \tag{1.6}$$

In order to avoid the dependence on the external magnetic field strength, the Larmor frequency of a certain nucleus is usually not measured in Hz, but with respect to a reference nucleus. The resulting *chemical shift* δ is measured in *ppm* (Equation 1.7).

$$\delta/ppm = \frac{\nu_i - \nu_{\text{ref}}}{\nu_{\text{ref}}} \cdot 10^6 \tag{1.7}$$

ν measures the Larmor frequency in Hz and relates to the Larmor frequency ω measured in $rad\ s^{-1}$ via $\nu = \omega/2\pi$. ν_i represents the Larmor frequency of the atom i and ν_{ref} of the reference compound.

Radio frequency-pulses (rf-pulses) can be used to align the bulk magnetization vector along a different axis. This is based on the fact that a small oscillating magnetic field at Larmor frequency (B_1 field) along the x- or y-axis induces the bulk magnetization vector to rotate in the yz- or xz-plane, respectively, although the field strength of B_0 is much larger. This can be attributed to the resonance condition between the Larmor precession of the bulk magnetization vector and the oscillating field. During the application of an rf-pulse, the bulk magnetization precesses around the axis of the oscillating field and the duration of the rf-pulse determines the direction to which the bulk magnetization vector points after the pulse. A 90°-pulse applied from the x-axis flips the bulk magnetization to the -y-axis, whereas a 180°-pulse flips the magnetization to the -z-axis.

The NMR signal is detected by a coil aligned in the xy-plane such that the bulk magnetization vector precessing in the xy-plane induces a current in the coil which can be measured. The signal reaches a maximum if the vector is aligned along the coil axis, a negative maximum is reached if it is aligned in the opposite direction and it vanishes while the vector is oriented perpendicular to the coil axis. When the signal is observed over time, it oscillates at Larmor frequency (*Free Induction Decay*, FID) showing a decay due to relaxation of the magnetization. Two different sources of relaxation are responsible for the return to equilibrium. Longitudinal relaxation (or spin-lattice relaxation) is triggered by small fluctuating local fields from nearby spins that have a similar effect on a spin as the oscillating B_1 field, which is applied during an rf-pulse, and thus bring the bulk magnetization back to the z-axis. Transverse relaxation (or spin-spin relaxation) is caused by the fact that spins precessing in the xy-plane have slightly different frequencies (depending on their individual chemical shift) and thus dephase after a certain amount of time. The dephasing in the xy-plane is the main source of the signal decay during the FID.

The FID measures the signal intensity over time (*time-domain* signal) representing an oscillation at Larmor frequency, which approaches zero after a certain amount of time due to relaxation. As the Larmor frequency is slightly different for every nucleus in a molecule, the FID is a superposition of several frequencies. In order to extract the frequencies included in the FID (i.e. determine the *frequency-domain* signal), Fourier transformation is applied, which gives rise to a spectrum showing the intensity against the frequency. As the processing is performed by a computer, the time-domain data is converted to a digital representation using data points. The raw data thus includes intensities measured at certain equally spaced time points. The space between two data points is called *dwell time* (dt) and relates to the spectral width (sweep width, sw), which specifies the measured frequency range in Hz, in the following way:

$$dt = \frac{1}{2sw} \tag{1.8}$$

The acquisition time in turn depends on the product of the dwell time and the number of data points. It determines the digital resolution (i.e. longer acquisition time yields higher digital resolution). The number of data points should thus be chosen such that the complete FID is recorded, however, acquisition times which exceed the FID increase the amount of noise and are not recommended.

The line width of the individual signals is determined by the length of the FID. As the FID decays exponentially due to relaxation, the frequencies cannot be determined exactly and each signal in the spectrum is represented by a *Lorentzian lineshape*. Truncation of the FID or application of non-exponential window functions during spectrum processing,

however, results in different lineshapes. Short transverse relaxation times are thereby responsible for broad NMR signals due to fast decay of the FID.

The NMR experiment is specified in the form of a *pulse sequence* defining the sequential order of one or several rf-pulses, optional indirect evolution periods, and the detection period where the FID is measured. The pulse sequence corresponding to the most simple 1D NMR experiment is composed of one 90°-pulse followed by the detection period. The resulting 1D NMR spectrum showing the intensity against the frequency can, however, be very crowded if the molecule in the sample is composed of many atoms that all give rise to a signal. This is especially true for biological macromolecules such as proteins. In such cases, individual signals cannot be resolved due to peak overlap resulting from the similarity of chemical shifts. In order to increase the resolution and therefore the information content of NMR spectra, more complex pulse sequences for multidimensional experiments have been developed. While the simple 1D experiment shows the intensity against the frequency, the intensity is plotted over several frequency axes in multidimensional NMR spectra.

Every signal in a multidimensional spectrum is located at a certain position on every frequency axis. Each frequency represents the chemical shift of a single atom and the respective atoms that contribute to a given signal are connected in an experiment-type-specific manner. The multidimensional NMR spectrum thus provides information about the connectivity of the atoms in a molecule. The kind of information that can be extracted from a spectrum depends on the type of magnetization exchange used in the respective pulse sequence.

A second frequency domain originates from an incremented evolution period (t1) during the pulse sequence. This results in an oscillation of the signal intensity of the respective 1D spectrum that can subsequently be processed via Fourier transformation (Fig. 1.1).

1.2 Nuclear spin interactions

Interactions of nuclear spins can be separated into external interactions with applied magnetic fields as well as internal interactions with magnetic fields originating from the surrounding nuclear spins.

Static external magnetic field B_0

The NMR spectrometer can apply different magnetic fields which interact with the nuclear spins in the sample. The static field B_0 is usually applied along the z-axis of the laboratory reference frame. The interaction between the nuclear spins and the static magnetic field

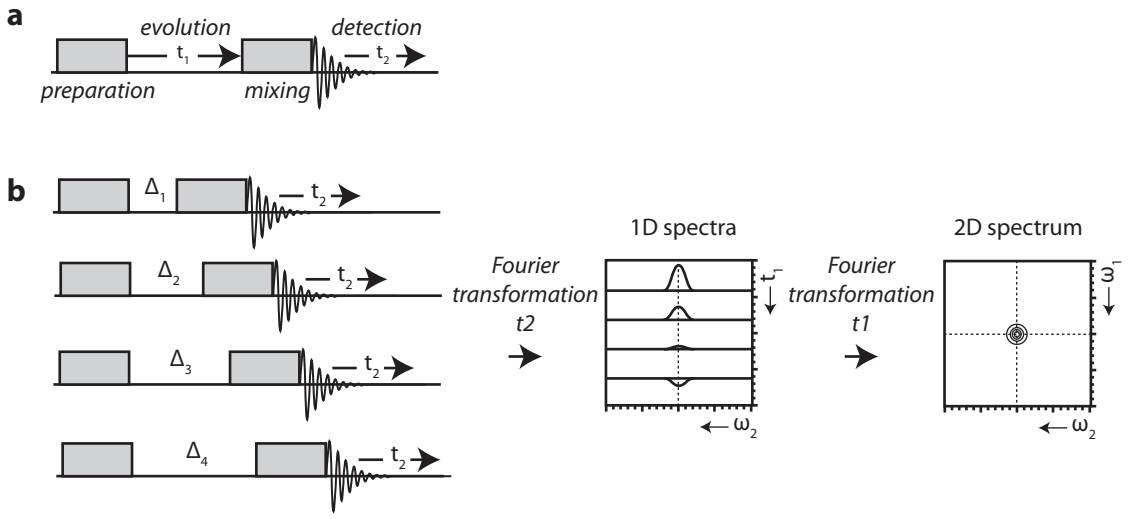


Figure 1.1: Two-dimensional NMR spectroscopy schematically. **a** General scheme of a two-dimensional pulse-sequence. Preparation and mixing periods can be varied depending on the desired spectrum type. **b** The FID of each increment is initially Fourier transformed in the direct detection period (t2) to gain a series of 1D spectra with the frequency domain ω_2. The oscillating peak intensity of the frequency domain (t1) represents an additional FID in the indirect dimension which is subsequently Fourier transformed to obtain a two-dimensional spectrum with two frequency domains ω_1 and ω_2.

is called *nuclear Zeeman interaction* and can be described by the following Hamiltonian $\hat{\mathcal{H}}_j^{\text{static}}$:

$$\hat{\mathcal{H}}_j^{\text{static}} = -\gamma_j B_0 \hat{I}_{jz} \tag{1.9}$$

The term $\gamma_j B_0$ represents the Larmor frequency.

Oscillating external magnetic field B_{RF}

In addition to the static field B_0, there is an rf-coil which generates an oscillating field B_{RF} in the xy-plane of the laboratory reference frame. This rf-field is characterized by a frequency, which represents the spectrometer reference frequency ω_{ref}, and an amplitude. The field is usually applied for a certain duration and called rf-pulse.

Chemical shift

The interaction with the static magnetic field B_0 is influenced by the surrounding electrons as they induce a magnetic field B_{induced}. The chemical environment thus determines the local field B_{loc} that a nucleus experiences, an effect called chemical shift. The induced field is not always oriented in the same direction as the static magnetic field and thus needs to be described by a three-dimensional chemical shift tensor δ, which is different for

every nucleus in a molecule. The component of the tensor in z-direction of the laboratory reference frame determines the actual local field that a nucleus experiences. The chemical shift consequently depends on the orientation of the chemical shift tensor with respect to the external magnetic field. In solution, the orientation of the molecule changes quickly due to molecular tumbling, resulting in an observed chemical shift which corresponds to the *isotropic chemical shift*. The isotropic chemical shift is the average of the *principal values* of the chemical shift tensor. The *chemical shift anisotropy* (CSA) measures the largest deviation in the chemical shift from the isotropic value. The Hamiltonian for the chemical shift is depicted in Equation 1.10,

$$\hat{\mathcal{H}}_j^{\mathrm{CS}} = -\gamma_j \delta_{\mathrm{zz}}^j(\theta) B_0 \hat{I}_{j\mathrm{z}} \tag{1.10}$$

where $\delta_{\mathrm{zz}}^j(\theta)$ is the z-component of the chemical shift tensor, which depends on the molecular orientation θ. This term can be replaced by the isotropic chemical shift value δ_{iso}^j in isotropic solutions.

J-coupling

J-coupling between two spins is mediated by the bonding electrons and is independent of the orientation of the bonding vector with respect to the external magnetic field. J-couplings are thus not averaged out by isotropic tumbling. Multidimensional NMR spectra that use J-coupling as magnetization exchange mechanism give information about the covalent structure of a molecule, i.e. every signal arises from covalently bonded atoms. The coupling can be observed for atoms connected via one bond (e.g. H-C), two bonds (e.g. H-C-H) or three bonds (e.g. H-C-C-H), however, the coupling constant decreases with increasing distance of the coupled spins. Three-bond J-couplings depend on the torsion angle θ between atoms i and $i+3$ and thus contain structural information which is described by the Karplus equation (Minch, 1994):

$$J(\theta) = A\ cos^2\theta + B\ cos\theta + C \tag{1.11}$$

A, B, and C are empirically derived constants whose values depend on the atom types.

Dipole-dipole coupling

Each nuclear spin generates a magnetic field which can be experienced by other nuclei. This mutual through-space interaction among two spins is called *dipole-dipole coupling*.

The Hamiltonian for the homonuclear dipole-dipole interaction between two nuclei i and j is shown in Equation 1.12.

$$\hat{\mathcal{H}}_{ij}^{\mathrm{DD}}(\theta_{ij}) = b_{ij}\frac{1}{2}(3cos^2\theta_{ij} - 1)(3\hat{I}_{iz}\hat{I}_{jz} - \hat{I}_i \times \hat{I}_j) \tag{1.12}$$

The interaction is described by the coupling constant b_{ij} which is defined in Equation 1.13,

$$b_{ij} = -\frac{\mu_0}{4\pi}\frac{\gamma_i\gamma_j\hbar}{r_{ij}^3} \tag{1.13}$$

where μ_0 represents the permeability of the vacuum, γ_1 and γ_2 the respective gyromagnetic ratios and r the internuclear distance.

The dipole-dipole coupling depends on the orientation of the internuclear vector with respect to the external magnetic field (i.e. the angle θ_{ij}). This leads to the fact that dipolar couplings vanish in isotropic liquids, but they can be measured in anisotropic liquids through molecular orientation as well as in solid-state samples.

1.3 Structural information from NMR

1.3.1 Dipolar relaxation and the Nuclear Overhauser Effect (NOE)

Dipolar relaxation is based on the fact that two nearby spins interact via dipolar coupling. As the behavior of spin 1 is consequently influenced by the state of spin 2, a phenomenon called *cross-relaxation*, magnetization can be exchanged between these two spins. This effect is described by the Solomon equations (Solomon, 1955) presented below:

$$\begin{aligned}
\frac{dI_{1z}}{dt} &= -R_z^{(1)}(I_{1z} - I_{1z}^0) - \sigma_{12}(I_{2z} - I_{2z}^0) \\
\frac{dI_{2z}}{dt} &= -R_z^{(2)}(I_{2z} - I_{2z}^0) - \sigma_{12}(I_{1z} - I_{1z}^0) \\
\frac{d2I_{1z}I_{2z}}{dt} &= -R_z^{(12)}(2I_{1z}I_{2z})
\end{aligned} \tag{1.14}$$

The rate constant $R_z^{(1)}$ measures the auto-relaxation of spin 1, analogously $R_z^{(2)}$ measures the auto-relaxation of spin 2. The rate constant for the cross-relaxation between the two spins is denoted σ_{12} and can be defined as

$$\sigma_{12} = b^2\frac{3}{10}j(2\omega_0) - b^2\frac{1}{10}j(0) \tag{1.15}$$

in the case of two spins of the same type. $j(2\omega_0)$ and $j(0)$ are the reduced spectral density functions, i.e. the probability of finding molecular motions at the respective angular frequencies $2\omega_0$ and 0, whereas b is defined in Equation 1.13.

The NOE

The NOE describes the exchange of magnetization between two spins via cross-relaxation. The Solomon equations (Equations 1.5) describe the influence of spin 2 on spin 1, whereas spin 2 is out of equilibrium (i.e. $I_{2z} - I^0_{2z} \neq 0$). The rate of magnetization transfer is determined by the rate constant σ_{12} (Equation 1.15), which is proportional to b^2 (Equation 1.13). The exchange rate thus depends on the internuclear distance via $1/r^6$, which makes the NOE observable only for atoms that are separated by a small distance (i.e. $r \leq 5$ Å).

Several experiments exist in order to detect the NOE. The *transient NOE experiment* inverts the z-magnetization of spin 2 by a selective 180°-pulse and thus brings spin 2 out of equilibrium. As a consequence, the z-magnetization of spin 1 increases via cross-relaxation from spin 2. The increase is proportional to the time τ as well as the relaxation rate constant σ_{12} (Equation 1.16).

$$I_{1z}(\tau) = 2\sigma_{12}\tau I^0_{2z} + I^0_{1z} \tag{1.16}$$

Equation 1.16 is only valid in the initial rate limit. This is attributed to the fact that for solving the differential equation (Equation 1.14) it was assumed that I_{1z} and I_{2z} have their initial values. Similarly, the change in z-magnetization can be calculated for spin 2 within the initial rate limit (Equation 1.17).

$$I_{2z}(\tau) = 2R^{(2)}_z\tau I^0_{2z} - I^0_{2z} \tag{1.17}$$

The initially inverted z-magnetization of spin 2 becomes less negative while approaching equilibrium. This spectrum is called the *irradiated spectrum*, whereas the pulse sequence which detects both spins at equilibrium magnetization (i.e. I^0_{1z} and I^0_{2z}) produces what is commonly called the *reference spectrum*. Subtracting the reference spectrum from the irradiated spectrum yields the *NOE difference spectrum* which shows the NOE enhancement for peak 1. If the cross-relaxation rate σ_{12} is zero, there will be no NOE enhancement for peak 1 and thus there is no peak intensity visible for spin 1 in the NOE difference spectrum. On the other hand, a signal in the NOE difference spectrum is a clear indication for cross-relaxation.

Due to auto-relaxation of spin 1, the NOE enhancement reaches a maximum and then decreases again. The time at which the maximum enhancement is reached as well as the size of the maximum enhancement depends on the auto-relaxation rate ($R_{1z}^{(1)}$) and the cross-relaxation rate (σ_{12}).

A second experiment, which can be used to measure the NOE, is the *steady-state NOE experiment*. While the transient NOE experiment flips the z-magnetization of spin 2 using a 180°-pulse, spin 2 is saturated by continuous weak irradiation in the steady-state NOE experiment. The z-magnetization of spin 2 is zero during irradiation which induces the cross-relaxation. In contrast to the transient NOE experiment, the NOE enhancement of the steady-state NOE experiment does not only depend on the cross-relaxation rate σ_{12} in the initial rate limit, but also on the auto-relaxation rate $R_{1z}^{(1)}$ of spin 1. This leads to the fact that the peak intensity of the NOE difference spectrum in the steady-state NOE experiment cannot be evaluated quantitatively with respect to the internuclear distance (Keeler, 2005).

Two-dimensional Nuclear Overhauser Enhancement Spectroscopy (NOESY)

A two-dimensional NOESY spectrum makes use of the NOE and cross-peaks arise from cross-relaxation between nearby spins. The pulse sequence is shown in Fig. 1.2.

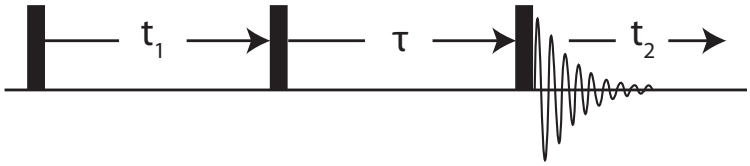


Figure 1.2: Pulse sequence of the 2D NOESY experiment. Black rectangles indicate 90°-pulses. t_1 refers to the indirect evolution time, τ represents the mixing time, and t_2 is the direct evolution time (detection).

The first two 90°-pulses and the t_1-time in between are responsible for the evolution of the indirect dimension in order to obtain a two-dimensional spectrum. The z-magnetization after the second pulse is able to undergo cross-relaxation. After the mixing time, a third 90°-pulse rotates the magnetization into the transverse plane for detection during t_2.

Solving the differential equation for a two-spin system using the initial rate approximation leads to the following equation, which describes the events during the mixing time τ.

$$\frac{I_{1z}(\tau)}{I_z^0} = -cos(\Omega_1 t_1)(1 - R_z\tau) + cos(\Omega_2 t_1)\sigma\tau + (R_z + \sigma)\tau \tag{1.18}$$

Equation 1.18 furthermore assumes that both spins are of the same type, i.e. only one cross-relaxation rate (σ) is used, the auto-relaxation rate R_z is equal for both spins as well as the equilibrium z-magnetization I_z^0. The change in z-magnetization of spin 1 $I_{1z}(\tau)$ is modulated as a sum of three terms in t_1. The first term describes the diagonal peak, as it is modulated at Ω_1 during t_1 and t_2. The intensity is negative as $R_z\tau << 1$ in the initial rate regime. The second term describes a cross peak which is modulated at Ω_2 during t_1 and at Ω_1 during t_2. This cross peak is positive or negative in intensity depending on the sign of σ (i.e. positive in the fast motion regime and negative in the slow motion regime). It indicates that cross-relaxation between the two spins took place and that they must consequently be located within small distance. The third term describes axial peaks, which arise during the mixing time and have thus no modulation during t_1. They appear at $\omega_1 = 0$ and $\omega_2 = \Omega_1$. As these peaks contain no useful information, they can be suppressed by a suitable phase cycle (Keeler (2005)).

1.3.2 Spin diffusion

Cross-relaxation does not only occur in two-spin systems, but also between individual pairs of spins in larger spin systems. The cross-relaxation between two spins, however, is affected by the presence of a third spin. If spin 1 transfers magnetization to spin 2, then spin 2 is out of equilibrium and can subsequently transfer magnetization to spin 3 which is close to spin 2 but not to spin 1. The transfer between spin 2 and spin 3 is called relayed NOE from spin 1. As the z-magnetization of spin 2 is above the equilibrium value after cross-relaxation from spin 1, the cross-relaxation with spin 3 leads to a negative NOE enhancement. Consequently, both peaks are negative in the slow motion regime and can thus not be distinguished from each other, whereas the intensity of direct and relayed NOE have opposite signs in the fast motion regime. Especially when looking at proteins, which are in the slow motion limit, direct cross-relaxation occurs relatively fast, favoring relayed transfer even along a chain of spins. The spread of magnetization through relayed cross-relaxation, which is especially prominent at longer mixing times, is called *spin diffusion* (Kalk and Berendsen, 1976).

The presence of spin diffusion complicates the interpretation of NOESY signals, since peaks can theoretically appear for spin pairs which are not actually close enough to ex-

change magnetization. Magnetization transfer via indirect pathways can have severe consequences for structure determination as the concept of distance restraints relies on the assumption that all peaks observed in the spectrum arise from atom pairs which are actually nearby. Furthermore, the signal intensities lose quantitative information if magnetization is passed on to a third spin (Keepers and James, 1984). Peak intensities are, however, usually calibrated into upper distance limits using the *isolated spin pair approximation* (ISPA) assuming that the peak intensity is correlated with the internuclear distance via r^{-6}.

1.3.3 Full relaxation matrix analysis

The effects of spin diffusion can be described by the *full relaxation matrix* approach. The Solomon equations (presented for a two-spin system in Equation 1.5) describe the build-up of NOE intensity and can be extended for a multi-spin system (Equation 1.19).

$$\frac{d}{dt}\boldsymbol{M}(t) = -\boldsymbol{R}(\boldsymbol{M}(t) - \boldsymbol{M_0}) \tag{1.19}$$

Solving this differential equation leads to Equation 1.20.

$$\boldsymbol{M}(\tau_\mathrm{m}) = e^{-\boldsymbol{R}\tau_\mathrm{m}}(\boldsymbol{M}(0) - \boldsymbol{M_0}) + \boldsymbol{M_0} \tag{1.20}$$

$\boldsymbol{R}$ is the relaxation matrix, $\boldsymbol{M_0}$ is the equilibrium magnetization, $\boldsymbol{M}(0)$ is the starting magnetization, and τ_m the mixing time. The relaxation matrix $\boldsymbol{R}$ is composed of auto-relaxation rates R_{ii} (Equation 1.21) as well as cross-relaxation rates R_{ij} (Equation 1.22).

$$R_{ii} = \rho_{ii} = \frac{1}{20}\,\frac{\mu_0^2\hbar^2\gamma_\mathrm{H}^4}{(4\pi)^2 r^6}\,(j(0) + 3j(\omega) + 6j(2\omega)) \tag{1.21}$$

$$R_{ij} = \sigma_{ij} = \frac{1}{20}\,\frac{\mu_0^2\hbar^2\gamma_\mathrm{H}^4}{(4\pi)^2 r^6}\,(6j(2\omega) - j(0)) \tag{1.22}$$

$j(\omega)$ is the reduced spectral density and defined by Equation 1.23.

$$j(\omega) = \frac{2\tau_\mathrm{c}}{1 + (\omega\tau_\mathrm{c})^2} \tag{1.23}$$

If all NOE cross peaks and diagonal peaks (i.e. the complete NOE intensity matrix) could be measured simultaneously, it would be possible to calculate the cross-relaxation rates σ_{ij}. These could then be used to estimate the interproton distances r_{ij}. This approach fails in the case of large molecules as the intensity matrix cannot be measured in a complete way due to spectral crowding. There are, however, several methods presented in literature that calculate theoretical peak intensities by using a full relaxation matrix approach, which is based on the distances in a 3D structure. Expected peak intensities are compared to the measured peak intensities and the contribution of spin diffusion to the peak intensity is estimated in this way.

Chapter 2

Structure determination by NMR spectroscopy

NMR spectroscopy provides structural information about biomolecules indirectly via structural restraints. These restraints are used as input for structure calculation algorithms which minimize a target function measuring the agreement between the set of restraints and a structural model. Starting from random structures, this target function minimization is performed several times. The subset of structural models with lowest target function values after a specified amount of minimization steps is combined into a structure bundle which represents the conformational space that is in agreement with the experimental input data. This approach is in fundamental contrast to the X-ray diffraction method, where the experimental input data are converted into an electron density map, which is a direct image of the structure.

Structure determination by NMR spectroscopy requires a sequence of steps that are introduced in more detail in the following section. Originally, all these steps had to be performed manually by experienced NMR spectroscopists. However, there is a strong ambition to fully automate the individual steps in order to (i) reduce the required time, (ii) make structure calculation more objective, and (iii) make the method more easily accessible to non-experts.

2.1 Sample preparation

In order to obtain high quality solution NMR data, the sample needs to be soluble at high concentrations (>0.05 mM), homogeneous, and stable at high temperatures. The spectral quality decreases with molecular weight due to reduced molecular tumbling rates, setting a molecular weight limit of ~25 kDa up to which structure determination is routinely performed. Multidimensional NMR spectroscopy, which is typically used to reduce signal

overlap, requires isotope labeling with ^{13}C and ^{15}N. Also, depending on the sample to be studied or the question to be addressed, deuteration may be necessary. Additional selective labeling strategies, such as amino acid type-selective labeling, stereo-selective labeling, segmental labeling, or SAIL methods (Kainosho et al., 2006), have been developed to further improve spectral quality (Takeda and Kainosho, 2011).

Expression is commonly carried out in *Escherichia coli* growing on isotope-labeled minimal medium (Kwan et al., 2011). An alternative cell-free expression approach possesses several advantages over conventional protein expression. Firstly, proteins can be produced that are either toxic to the host cell, that tend to aggregate in inclusion bodies, or that are degraded by host proteases. Secondly, labeled amino acids can be incorporated with a minimum of scrambling which improves the result especially in the case of selective isotope labeling (Kigawa et al., 1995; Torizawa et al., 2004).

2.2 Chemical shift assignment

Each nucleus is characterized by a chemical shift originating from its distinct chemical environment which in turn influences its Larmor frequency. The correlation of each chemical shift with the respective nucleus is called *chemical shift assignment*. It is an essential requirement for the majority of subsequent spectral analyses such as structure determination or dynamics. Chemical shift assignment is very time-consuming due to the large amount of required multi-dimensional NMR spectra as well as the manual interpretation of these spectra. A typical set of NMR spectra includes several spectra for the assignment of backbone atoms (HNCACB, CBCA(CO)NH, HNCA, HN(CO)A, HNCO, and HN(CA)CO) as well as additional experiments for the assignment of side-chain atoms ((H)CC(CO)NH, HCC(CO)NH, HNHA(CO)NH, HCCH-TOCSY, and ^{15}N-edited TOCSY) (Shin et al., 2008).

Despite the very time-consuming nature of the chemical shift assignment process, automation is currently only rarely applied. This originates from the fact that the available programs have not proven to be very robust especially when dealing with imperfect input data. A large number of algorithms for backbone and/or side-chain assignment have been introduced, however, the use of only a small subset of these has actually been reported in the Protein Data Bank (PDB): Autoassign (Zimmerman et al., 1997), PINE (Bahrami et al., 2009), and GARANT (Bartels et al., 1997) (Guerry and Herrmann, 2011). Recent

progress in the field was achieved by the development of the new FLYA program for chemical shift assignment (Schmidt and Güntert, 2012) whose robust performance and broad applicability has been demonstrated by a wide range of different applications such as purely NOESY-based assignment (Schmidt and Güntert, 2013), RNA assignment (Aeschbacher et al., 2013; Krähenbühl et al., 2014), or assignment of solid-state NMR spectra (Schmidt et al., 2013).

2.3 Automated NOE assignment and structure calculation

The classical NOE based structure determination procedure uses distance restraints from NOESY experiments, supplemented by dihedral angle restraints from chemical shift or residual dipolar coupling (RDC) analysis, as a major source of structural information. The assignment of NOESY signals to atom pairs is the crucial step to obtain distance restraints from NOESY spectra. Difficulties arising from spectral imperfections such as peak overlap, noise, and spectral artifacts usually prevent unambiguous assignment for the majority of signals. As a consequence, NOE assignment is typically performed in an iterative procedure using a preliminary structure calculated from a small set of unambiguous distance restraints in order to resolve further assignment ambiguities by incorporating structural information.

In order to accelerate this time-consuming procedure, several algorithms for automated NOE assignment have been developed: NOAH (Mumenthaler and Braun, 1995; Mumenthaler et al., 1997), ARIA (Rieping et al., 2007; Linge et al., 2003), AUTOSTRUCTURE (Huang et al., 2005; Huang et al., 2006), KNOWNOE (Gronwald et al., 2002), CANDID (Herrmann et al., 2002b), CYANA (Güntert, 2009; Güntert and Buchner, 2015), PASD (Kuszewski et al., 2004; Kuszewski et al., 2008). Automated NOE assignment is used much more frequently than automated chemical shift assignment which can be concluded from their number of reportings in the PDB.

The initial NOE assignment, which is typically performed in the absence of any structural information, is the most important and most difficult part during iterative NOE assignment and structure calculation. This results mostly from the large number of assignment possibilities. Several concepts have been introduced by different groups in order to improve the performance of the automated NOE assignment (network anchoring (Herrmann et al., 2002b)) or the results of the subsequent structure calculation (ambiguous distance restraints (Nilges, 1995) and constraint combination (Herrmann et al., 2002b)). These methods are now implemented in several of the aforementioned programs. The automated assignment algorithm of the software package CYANA (Combined assignment and dYnamics Algorithm for NMR Applications), which is a reimplementation of the CANDID algorithm, uses all of the aforementioned strategies and since the CYANA software was

used throughout this thesis, its assignment strategy is introduced in more detail. The following description is based on the comprehensive and detailed book chapter "Calculation of Structures from NMR Restraints" (Güntert, 2011).

CYANA commonly performs seven cycles of NOE assignment and structure calculation as well as one final structure calculation (Fig. 2.1). Each cycle uses as input the amino acid sequence, a set of unassigned peak lists from nD NOESY spectra containing peak positions and intensities, one or several chemical shifts lists as a result of the chemical shift assignment, and (except for the initial cycle) the 3D structure from the previous cycle. Irrespective of the assignment, each peak intensity is calibrated into an upper distance limit (upl) using the $1/r^6$ dependence introduced in Section 1.3.1. This requires either a calibration constant, or, if not available, the program determines the calibration constant based on a specified distance corresponding to the median intensity (in the program as well as throughout this thesis denoted as *dref*-value).

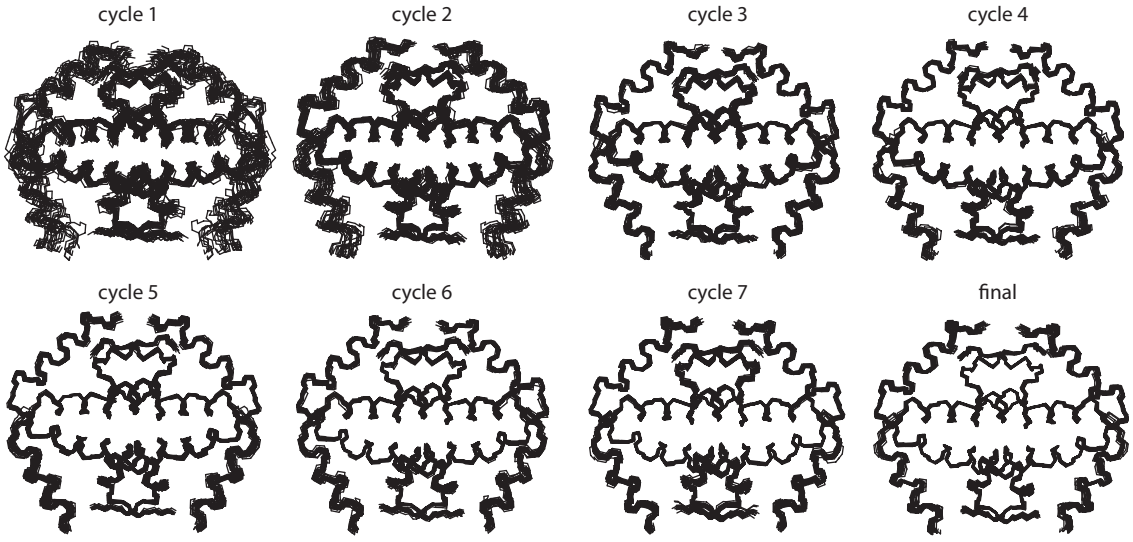


Figure 2.1: Seven cycles of NOE assignment and structure calculation as well as one final structure calculation using the program CYANA. The example calculation depicts the tetramerization domain of the heterotetrameric protein p63/p73 (Rocco and Ellisen, 2006; Coutandin et al., 2009). Input data included a total of ten different NOESY spectra recorded on two differently labeled samples (i.e. uniformly $^{13}C/^{15}N$ labeled p63 mixed with unlabeled p73, and uniformly $^{13}C/^{15}N$ labeled p73 mixed with unlabeled p63). The five NOESY experiments comprise NOESY-[^{13}C,^{1}H]-HSQC, NOESY-[^{15}N,^{1}H]-TROSY, ^{13}C,^{15}N filtered-NOESY-[^{13}C,^{1}H]-HSQC, ^{13}C,^{15}N filtered NOESY-[^{15}N,^{1}H]-TROSY, and NOESY-[^{13}C,^{1}H]-SOFAST-HMQC. As assignment tolerance 0.045 ppm for ^{1}H dimensions and 0.45 ppm for ^{13}C and ^{15}N dimensions were used. Simulated annealing was conducted for 200 random starting structures using 20,000 annealing steps. The 20 lowest energy structures were combined to form a structure bundle. Only ordered residues were selected for structure superposition and display.

The initial assignment is generated by selecting all atoms whose chemical shift values match the peak position in one of the dimensions within a given tolerance (Fig. 2.2a). Each of the assignment possibilities generated in this way is subsequently evaluated based on a probability P_{tot}, which is estimated based on the chemical shift agreement P_{CS}, the network anchoring score P_{Network}, and the agreement with the structure from the previous cycle $P_{\mathrm{Structure}}$ in all but the first cycle:

$$P_{\mathrm{tot}} = P_{\mathrm{CS}} \times P_{\mathrm{Network}} \times P_{\mathrm{Structure}} \tag{2.1}$$

The probability P_{CS} is calculated by a gaussian function. Network Anchoring is used to calculate a probability (P_{Network}) for a given assignment possibility between atoms A and B based on assignments of other peaks that indirectly connect atoms A and B via a third atom C through the assignments of other peaks to A and C and B and C, respectively (Fig. 2.2b). Network anchoring is especially helpful in the initial cycle where no structural information is available in order to exclude unlikely assignments. $P_{\mathrm{Structure}}$ is calculated in all but the first cycle and measures the percentage of structures from the structure bundle of the previous cycle that are in agreement with the upl-value of the given peak (Fig. 2.2c). The upl-value may be violated to a certain extent, however, the size of the tolerated violation is decreased from cycle to cycle. All assignment possibilities with a probability below a cycle-dependent threshold are discarded.

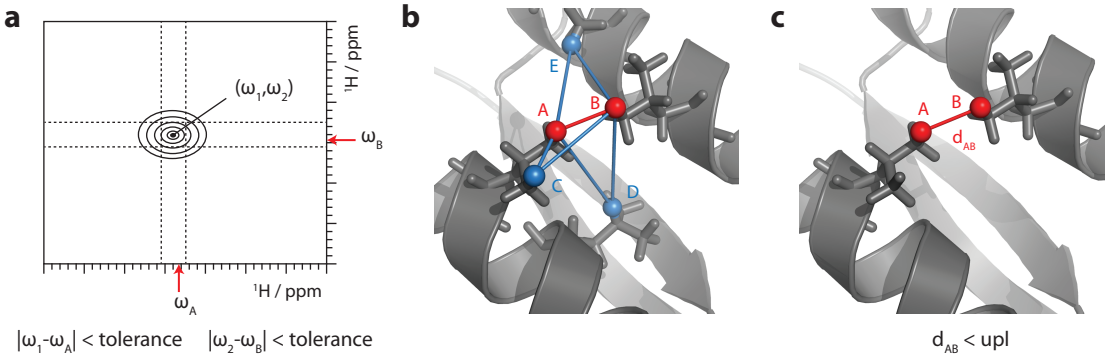


Figure 2.2: Criteria for assignment selection during automated NOE assignment with CYANA. **a** The initial set of assignment possibilities is generated based on chemical shift agreement such that all atoms whose chemical shift match the peak position within a given tolerance are selected for the respective dimension. **b** Network anchoring. Every assignment possibility is evaluated with respect to the network of other restraints that support the given assignment possibility (e.g. the given assignment between atoms A and B is supported by two restraints connecting atoms A and E and E and B, respectively). **c** In all but the first cycle, the structure from the previous cycle is used to evaluate any assignment possibility. The interatomic distance in the preliminary structure is required to be shorter than the corresponding upl-value plus a cycle-dependent violation cutoff.

By doing so, one distance restraint is created for each peak with at least one remaining assignment. The respective upper distance limit is obtained for each peak during calibration. As many peaks cannot be assigned unambiguously in early cycles of the calculation, a fundamental improvement was the concept of ambiguous distance restraints, first introduced by Nilges (Nilges, 1995), which allows the usage of distance restraints with multiple assignments. This method makes use of the fact that the "effective" distance d_{eff} (Equation 2.2), calculated as the r^{-6}-sum over all individual distances d_i, is always shorter than the shortest individual distance.

$$d_{\text{eff}} = (\sum_{i=1}^{n} d_i^{-6})^{-1/6} \tag{2.2}$$

Consequently, any ambiguous distance restraint which includes the correct assignment amongst others will be fulfilled by the correct structure and thus has no distorting influence during structure calculation (Fig. 2.3). In contrast, distance restraints containing only incorrect assignment possibilities will most likely cause distortions.

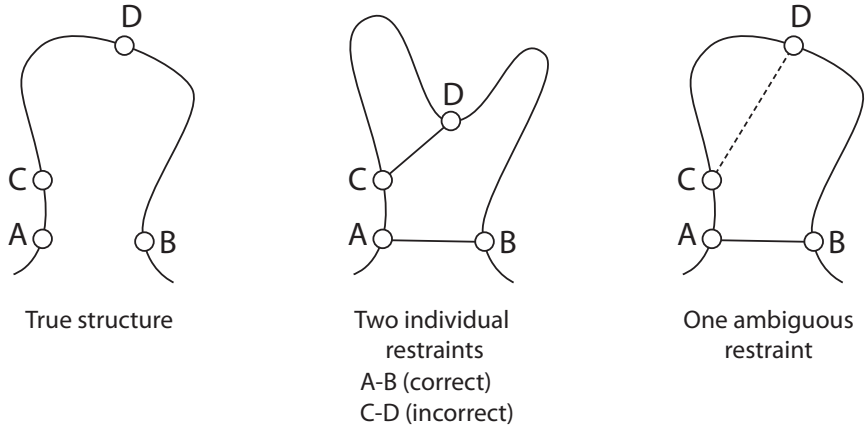


Figure 2.3: The concept of ambiguous distance restraints. If an incorrect distance restraint is treated individually, the structure will be distorted during structure calculation (*center*). If an incorrect distance restraint is treated as an ambiguous restraint with any number of additional restraints, the "effective" distance d_{eff} (Equation 2.2) is always shorter than the shortest individual distance. Consequently, this eliminates the harmful effect of incorrect restraints, provided the correct assignment is included (*right*).

In order to prevent their harmful effect, the concept of constraint combination was introduced first in the CANDID algorithm (Herrmann et al., 2002b). The idea is based on the same considerations as the ambiguous distances restraints (Fig. 2.3). If one distance restraint including only incorrect assignments is combined with a second independent restraint that contains at least one correct assignment, then the combined restraint is less harmful at the cost of a temporary loss of information. Constraint combination is typically

performed in the initial two cycles where erroneous assignments occur more frequently due to the lack of a structure bundle or the lower quality of the structure bundle in the first cycle.

CYANA uses a target function minimization strategy based on simulated annealing by molecular dynamics simulation in torsion angle space in order to calculate the 3D structure of a molecule. This algorithm for structure calculation was originally implemented in the DYANA software (Güntert et al., 1997). The target function, which is used as potential energy, measures the agreement of a conformation and the respective set of structural restraints (e.g. distance and/or angle restraints) as well as steric overlap. It is defined as zero if all restraints are fulfilled and no steric overlap occurs. Simulated annealing is characterized by the presence of kinetic energy which allows overcoming potential barriers and thus reduces the chance of becoming trapped in local minima. Performing the simulation in torsion angle space accelerates the calculation compared to using cartesian coordinates due to the reduced number of degrees of freedom, the simpler potential energy function and the use of longer time steps for the integration of the equations of motion. This can be rationalized by the conservation of the covalent structure during the simulation.

Alternative software packages for NMR structure calculation are CNS (Brünger et al., 1998) and XPLOR-NIH (Schwieters et al., 2003). All of these structure calculation programs use a variety of structural restraints in addition to distance restraints from NOESY experiments as input, among these are torsion angle restraints obtained from chemical shift analysis, RDCs, J-coupling constants, or H-bonds, to name only some of them. The following section gives a short introduction into some commonly used sources of structural information obtained by NMR spectroscopy.

2.4 Restraints for NMR structure calculation

Structural restraints can, in principle, be divided into three different classes: distance restraints, dihedral angle restraints, and orientational restraints. Different experimental methods can be used to obtain either of these.

2.4.1 Distance restraints

The most important source of distance restraints is the NOE, which is observed if two protons are separated by a small distance of <5-6 Å. NOESY spectra of medium sized proteins typically yield several thousand restraints. Especially the hydrophobic core of a globular protein is characterized by a dense network of protons that give rise to NOE signals (Fig. 2.4). Out of all observed NOESY peaks, long-range NOEs between distant residues in the amino-acid sequence yield most of the information content. Protein struc-

tures of medium sized proteins can be determined solely based on NOE-derived distance restraints.

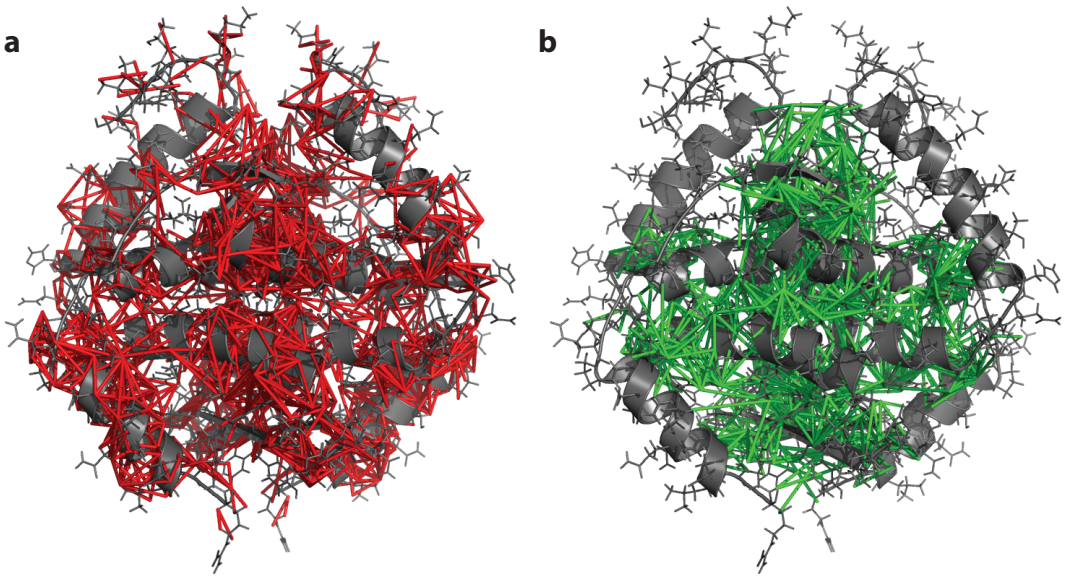


Figure 2.4: Final set of distance restraints originating from several 3D NOESY experiments. **a** 2502 short- and medium-range distance restraints (red), **b** 1421 long-range intra- and intermolecular distance restraints (green). The depicted heterotetrameric structure of p67/p73 is comprised of the tetramerization domains of each of two monomers p63 and p73 (Rocco and Ellisen, 2006; Coutandin et al., 2009). Distance restraints are the result of the combined automated NOE assignment and structure calculation with CYANA illustrated in Fig. 2.1.

Especially in the course of helical membrane protein structure determination, distance information from paramagnetic relaxation enhancement (PRE) (Gaponenko et al., 2000) and pseudo contact shifts (PCS) (Allegrozzi et al., 2000) has been proven to be very beneficial (Gottstein et al., 2012b; Crick et al., 2015). This results from the very limited observation of long-range NOEs. The physical background of the PRE is the distance-dependent influence of paramagnetic spin labels on NMR signals which causes the signals to either disappear completely, to be broadened, or to remain unaffected. Line broadening effects are observed for distances between the spin-label and the atom giving rise to the observed signal of 10-25 Å, shorter distances lead to the disappearance of the signal, whereas longer distances cause no observable effect. Spin labels, the most commonly used one being MTSL (methanethiosulfonate), are attached to cysteine residues which typically requires several single cysteine-mutants. Due to the time-consuming nature of sample preparations and NMR measurements, the use of PRE restraints is especially recommended in case of a very limited number of NOE-based distance restraints.

Another source of distance information that can be used for NMR structure calculation is the knowledge about hydrogen bonds. These are especially helpful for the definition of secondary structure elements. Several methods exist that allow the determination of hydrogen bonds, one of them being the amide proton-solvent exchange rate, which is very slow in hydrogen-bonded amide protons. Another option is the detection of trans-hydrogen bond scalar couplings, for instance through long-range HNCO-COSY spectra (Cordier et al., 2008). In contrast to the detection via exchange rates, J-couplings allow the determination of donor and acceptor atoms as well as an estimation of the geometry via the coupling strength.

2.4.2 Dihedral angle restraints

The most commonly used source of dihedral angle information is the atom chemical shift, which depends on the local conformation of the protein chain (i.e. secondary structure). However, the chemical shift additionally depends on other surrounding effects such as neighboring residues or solvent exposure. For this reason, it is common practice to use the secondary chemical shift, i.e. the difference between the observed chemical shift and the chemical shift of the same atom in a random coil, in order to predict the secondary structure type. Especially well established is the use of the TALOS-N software (Shen and Bax, 2013) which empirically determines Φ and Ψ torsion angles from backbone chemical shifts based on structure chemical shift relationships provided by the PDB (Berman et al., 2000) and the BioMagResBank (BMRB) (Doreleijers et al., 2005).

Structural information can as well be obtained from ^{3}J-coupling constants that are related to the dihedral angle via the Karplus relation shown in Equation 1.11 (Karplus, 1959; Karplus, 1963). Due to the ambiguous nature of this relation (i.e. several dihedral angles correspond to the same ^{3}J-coupling constant), it is not possible to determine a single dihedral angle value. Therefore, it is common practice to directly incorporate the ^{3}J-coupling constants into the structure calculation (Kim and Prestegard, 1990; Torda et al., 1993).

2.4.3 Orientational restraints – RDCs

The dipolar coupling of two nuclei depends on the length as well as the direction of the internuclear vector. Due to isotropic tumbling, the orientational component of the dipolar interaction is averaged to zero in solution and hence the dipolar Hamiltonian vanishes. If not averaged to zero, it provides valuable information about the relative orientation of the internuclear vector with respect to the external magnetic field. This can be achieved by restoring “residual“ dipolar couplings through weak alignment of the molecules using special alignment media. Several options exist for the choice of the optimal alignment

medium which needs to be optimized for every individual case (i.e. the biomolecule to be studied as well as the experimental condition such as pH, salt or temperature). Alignment medium options include amongst others bicelles, bacteriophages, or polyacrylamide gels (Prestegard et al., 2004).

Measurement of RDCs is based on the fact that the overall coupling between two spins is altered if the residual dipolar coupling is restored and this alteration can be quantified by the comparison of a spectrum recorded in an aligned medium and in isotropic solution. One-bonded interactions such as H_N-N or H-C vectors are especially well suited as the distance is fixed and known. The dipolar interaction thus mostly depends on the orientation of the vector. All experiments can be chosen which are typically used to record J-couplings, for larger molecules the IPAP experiment (Ottiger et al., 1998) has the great advantage that peak doublets arising from the splitting are separated into two spectra which greatly reduces the number of observed peaks in one spectrum (Vuister et al., 2011).

In order to use RDC restraints in structure determination, the following drawbacks have to be overcome. Firstly, the alignment tensor, described by two diagonal elements of the Saupe matrix (rhombicity and magnitude) as well as three Euler angles, of the molecule in the laboratory frame needs to be determined. Secondly, each residual dipolar coupling can arise from different vector orientations which leads to an ambiguity similar to that observed for J-coupling restraints based on the Karplus relation. For this reason, it is not straightforward to calculate a structure solely based on RDCs, but they represent a useful supplement for distance and dihedral angle restraints. The determination of the alignment tensor can be performed experimentally using the histogram method in the absence of structural information (Clore et al., 1998) or it can be determined by fitting calculated RDCs based on a known structure onto the experimental RDCs. During structure calculation, the RDCs are calculated based on the known tensor elements and the current structural model. The deviation between the calculated and measured RDCs is part of the target function.

RDCs are complementary to other sources of structural information due to their global nature. Each RDC gives structural information with respect to a common external frame of reference.

Chapter 3

What is different in the solid state?

The main difference between solid-state NMR and solution NMR is the presence of anisotropic interactions (i.e. homonuclear and heteronuclear dipolar coupling and chemical shift anisotropy (CSA)) which are not averaged out by molecular tumbling. These interactions are the main source of line broadening in solid-state NMR spectra, however, they are also a valuable source of information. In order to extract any information at all, it is necessary to improve the spectral quality. The most fundamental methodological developments are summarized in the following section.

3.1 Improving spectral quality

3.1.1 Technical advances

Magic angle spinning

The size of dipolar couplings as well as the chemical shift depend on the orientation of a molecular vector with respect to the external magnetic field. This orientation dependency can be described by the term $3\cos^2\theta$-1. As this term vanishes at the "magic angle" of $\theta = 57.74°$, a technique called *magic-angle spinning* (MAS) (Andrew et al., 1958; Lowe, 1959) was introduced in order to improve the quality of solid-state NMR spectra. When spinning the polycrystalline sample rapidly at the magic angle, its average CSA appears as an ellipsoid with its long axis aligned with the spinning axis, causing a complete disappearance of the CSA provided that the spinning frequency is high enough. Each atom then experiences its orientation-independent isotropic chemical shift. If the averaging is incomplete at lower frequencies, each peak shows spinning sidebands, i.e. additional signals of lower intensity at multiples of the spinning frequency. They can either be avoided

if the spinning frequency is greater than the width of the CSA or they can be placed in regions with no expected signals by choosing the spinning frequency accordingly (Fig. 3.1).

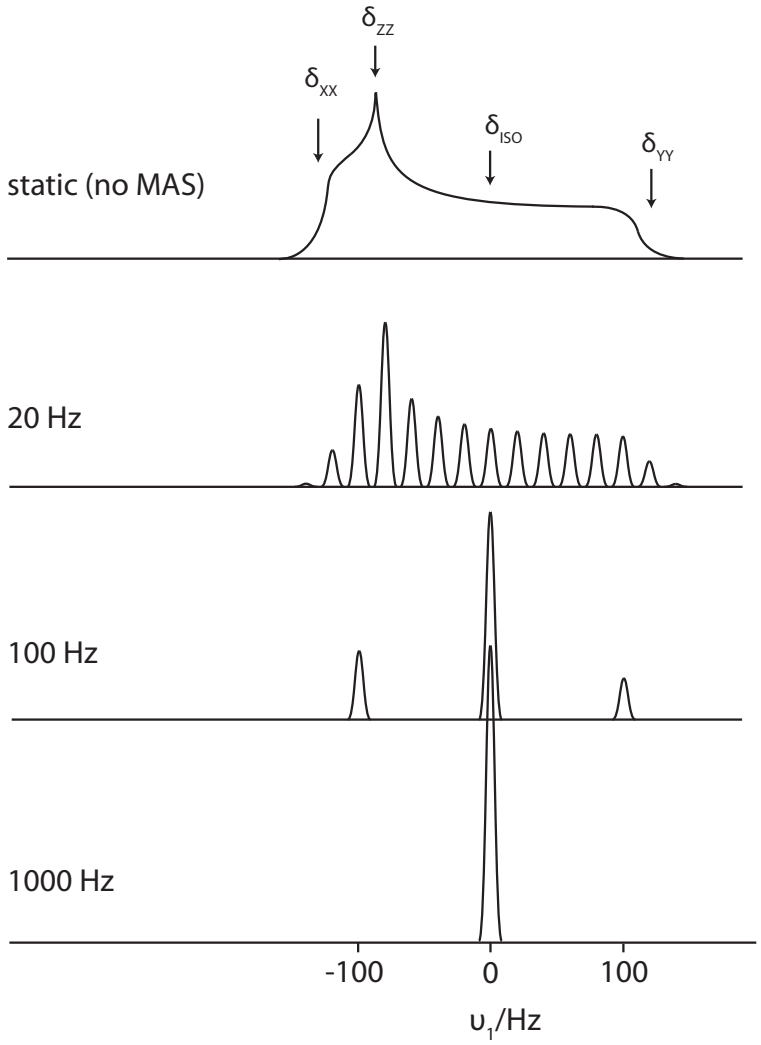


Figure 3.1: Effect of magic angle spinning (MAS) on solid-state NMR spectra. The static powder spectrum (*top*) shows the chemical shift anisotropy. The CSA tensor parameters δ_{XX}, δ_{YY}, and δ_{ZZ} can be deduced from the static lineshape. The isotropic chemical shift δ_{ISO} can then be calculated as $1/3 \times (\delta_{\mathrm{XX}} + \delta_{\mathrm{YY}} + \delta_{\mathrm{ZZ}})$. The distance between two spinning side band signals equals the MAS rate. A single peak at the isotropic chemical shift appears if the MAS frequency exceeds the CSA (i.e. the width of the static powder spectrum).

Whether interactions can be averaged out by magic-angle spinning depends on the strength of the respective interaction as well as the spinning frequency. A complete averaging is achieved if the spinning speed exceeds the size of the respective interaction in Hz. The size of a dipolar coupling depends on the internuclear distance and the gyromagnetic ratios of the spins involved (Equation 1.13). Therefore, dipolar couplings between low-γ

nuclei such as ^{13}C and ^{15}N are weak enough to be averaged out by MAS even at short distances. In contrast, couplings between and to ^{1}H spins are especially strong due to the large ^{1}H gyromagnetic ratio which prohibits the complete averaging of hetero- and homonuclear dipolar couplings involving protons by MAS alone. As a direct consequence, it is currently not feasible to obtain high-resolution ^{1}H solid-state NMR using fully protonated samples and moderate MAS frequencies, which makes the use of low-γ nuclei for detection the method of choice. The low sensitivity resulting from the low γ of these nuclei as well as the non-averaged heteronuclear coupling to the proton bath, however, make additional techniques inevitable to further improve spectral quality.

Cross polarization

A standard approach to increase sensitivity when using low-γ nuclei for detection is the transfer of ^{1}H polarization to the low-γ nucleus of interest, which is commonly achieved by a cross polarization (cp) step at the beginning of every pulse sequence. The Hartmann-Hahn method (Hartmann and Hahn, 1962) suggests the use of two simultaneous RF fields at the resonance frequencies of the respective atom types such that each spin rotates around the axis of the RF field at the same frequency. During this energy-conserving dipolar contact, magnetization flows from higher polarized nuclei to those with lower polarization. The nutation frequency is determined by the frequency and the amplitude of the applied RF-field. In this respect, the frequency is defined by the Larmor frequency of the respective nucleus which makes the amplitudes of the RF-fields the parameters used to adjust the resulting nutation frequencies (Laws et al., 2002).

Proton decoupling

As mentioned earlier, low-γ nuclei are coupled to the dense ^{1}H network, generating very broad signals if this interaction is not eliminated during periods of chemical shift evolution such as indirect evolution periods or the detection period. The effect of protons on low-γ nuclei can be suppressed by a number of proton decoupling sequences, the most simple of them being continuous-wave (cw) decoupling (Bloch, 1956; Bloch, 1958). The peak position of the low-γ nucleus is shifted in the presence of a coupled spin in opposite direction depending on the orientation of the latter (i.e. “spin-up” or “spin-down”). Consequently, if the orientation of the coupled spin changes continuously, the coupling effect is averaged out and thus no net effect on the peak position of the spin of interest is observed. This constant change in orientation is achieved by a continuous RF-field at the Larmor frequency of the spin to be decoupled, which is typically the proton spin.

More advanced techniques for heteronuclear decoupling include multi-pulse sequences such as the two-pulse phase modulation (TPPM) sequence (Bennett et al., 1995), which

relies on a special sequence of differently phased pulses, and a further development based on the TPPM sequence, the SPINAL64 sequence (Fung et al., 2000). These multi-pulse schemes improve the results especially in combination with MAS.

3.1.2 Isotope labeling

The ^{1}H spin is the only nucleus which provides sufficient natural abundance in biomolecules to perform NMR measurements. However, as mentioned in the previous section, the ^{1}H spin is not well suited for high resolution solid-state NMR. The low natural abundance of low-γ nuclei requires isotopic labeling techniques to increase the population of the NMR detectable spin 1/2 isotopes ^{13}C and ^{15}N. Heterologous expression in *E. coli* growing on ^{13}C-glucose and ^{15}N-NH_4-based minimal medium is the method of choice to obtain uniformly labeled protein samples. Uniform ^{13}C and ^{15}N labeling is the way to go for the typical multidimensional experiments for chemical shift assignment. However, depending on the size and the nature of the molecule of interest, it results in very crowded 2D spectra which are commonly recorded to obtain structural information. The detection of mostly weak but structurally important long-range peaks is hampered by the presence of strong short-range signals, which is discussed in more detail in the following sections about structural restraints from solid-state NMR. Attempts to suppress the disturbing influence of short-range signals promoted the development of patchwork labeling strategies with a reduced density of ^{13}C. This can be achieved via expression in *E. coli* growing on minimal medium based on either 1,3-^{13}C-glycerol, or 2-^{13}C-glycerol (LeMaster and Kushlan, 1996; Castellani et al., 2003; Franks et al., 2008). These labeling patterns (Fig. 3.2) eliminate the majority of directly bonded ^{13}C-^{13}C atom pairs and thus greatly reduce the number of signals visible in a 2D ^{13}C-^{13}C correlation spectrum. A different option for even further dilution of ^{13}C spins is the use of 1-^{13}C-glucose, or 2-^{13}C-glucose as carbon source for protein expression (Loquet et al., 2012). Very sparse isotope labeling is obtained when supplementing D-glucose-based deuterated minimal medium with 3-$^{13}CHD_2$-α-ketoisovalerate which results in a specific labeling of isoleucin, leucin, and valine methyl groups (Goto et al., 1999). This can be used to measure a small number of unambiguous long-range distance restraints which are especially helpful to define the global fold of a protein structure, because methyl groups are typically located in the hydrophobic core of a protein (Huber et al., 2011).

3.2 Structural restraints from solid-state NMR

As indicated previously, the interactions which cause the line broadening in solid-state NMR spectra are a valuable source of information. However, the averaging of these inter-

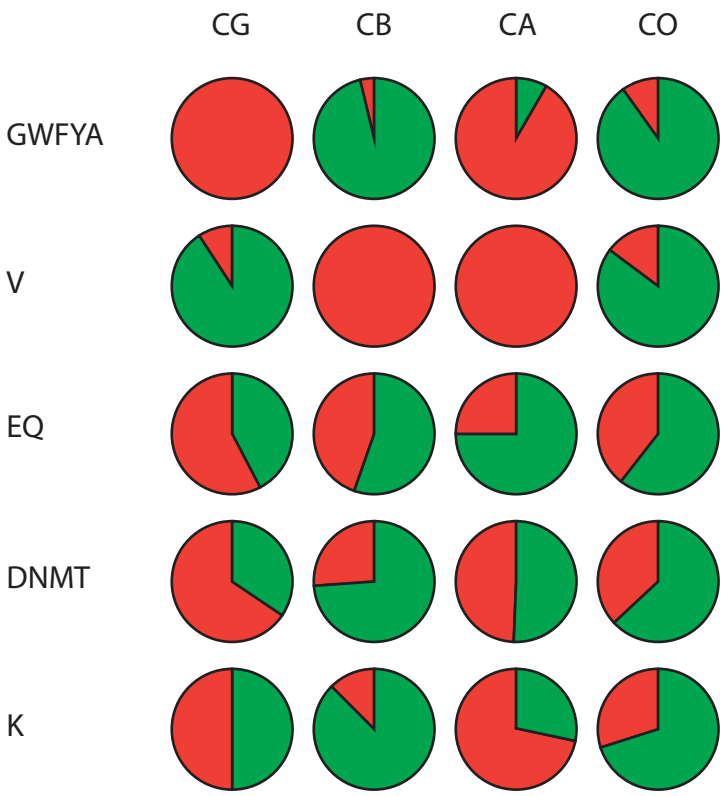


Figure 3.2: 1,3-^{13}C-glycerol (*green*) and 2-^{13}C-glycerol (*red*) labeling patterns. The percentage indicated by the circle filling defines the relative amount of labeling completeness for a given atom type in a given set of amino acids specified by the one-letter code on the left side of the figure (e.g. a complete red circle means that every atom of the corresponding type is labeled in the specified amino acid types in the 2-^{13}C-glycerol labeling. These percentages were estimated from TEDOR measurements of GB1 by the Rienstra group. The figure is based on Nieuwkoop et al., 2009.

actions in the scope of improving spectral quality requires the specific reintroduction of individual interactions during selected periods of the pulse sequence to obtain the desired information. Many experiments have been developed in the past in order to observe different types of interactions which can be used as restraints for structure calculation. An overview of the different experimental sources of structural information and their application to structure calculation of microcrystalline model proteins is given in the following sections. Fig. 3.3 and Table 3.1 summarize the methodological development in the field of structure determination by solid-state NMR. Fig. 3.3 includes only those structures that have been deposited in the PDB, whereas Table 3.1 includes all published structures. In principle, structural information can be divided into distance restraints, angle restraints, and orientational restraints. Angle restraints from chemical shift analysis are not mentioned specifically, as they can be obtained and applied in the same way as in solution NMR spectroscopy.

3.2.1 Distance restraints

Homonuclear dipolar recoupling

Homonuclear dipolar recoupling pulse sequences use rotor-synchronized radio-frequency pulses in order to interfere with the averaging effect of magic angle spinning and thus reintroduce dipolar couplings. This has been successfully used to determine internuclear distances in pair-labeled samples with very high accuracy. However, in order to determine structures of larger biomolecules such as proteins, it is necessary to determine a large number of distances which is not feasible when using pair-labeled samples. It is possible to apply dipolar recoupling sequences to uniformly labeled large molecules when combined with chemical shift correlation spectroscopy in order to resolve the individual labeled sites. This results in multi-dimensional spectra, where cross-peaks arise from dipolar-coupled spin pairs. The largest drawback is that structurally constraining carbon-carbon distances are usually in the order of several Å and the dipolar coupling of the corresponding atoms is consequently rather weak. It has been shown that polarization transfer between weakly coupled spins is especially prohibited in the presence of other strong couplings, an effect called *dipolar truncation*. This phenomenon has been extensively studied using simulations as well as experiments on three-spin systems and it was shown that alternating labeling schemes, which greatly reduce the amount of one-bond interactions, can only slightly improve the dipolar truncation effect (Bayro et al., 2009). This makes homonuclear dipolar recoupling ill-suited to obtain structural restraints from uniformly labeled proteins.

However, some effort has been put into the development of techniques to improve the situation, one example being the frequency-selective rotational resonance (R^2) experiment (Andrew et al., 1963; Creuzet et al., 1991) which was subsequently combined with chemical shift correlation spectroscopy in order to obtain multiple site-specific distances of a uniformly labeled dipeptide N-Ac-Val-Leu in the range between 2.5-5.3 Å (Ramachandran et al., 2003; Ramachandran et al., 2006). First applications have been presented on proteins in combination with uniform as well as diluted labeling schemes (Peng et al., 2008; van der Wel et al., 2009). In favor of obtaining several distances from a single experiment, it is necessary to reduce the frequency-selectivity for example through reduced decoupling power during the transfer step (Janik et al., 2007). Although this method is a first approach to improve the dipolar truncation effect, the number of distance restraints obtained in this way is not sufficient for structure determination.

Additional developments include a stochastic recoupling introduced by Tycko (Tycko, 2007), truncated dipolar recoupling by Levitt and co-workers (Marin-Montesinos et al., 2006) as well as oscillating-field techniques proposed by Nielsen and co-workers (Straasø et al., 2009; Straasø et al., 2014).

Heteronuclear dipolar recoupling

Aside from homonuclear recoupling sequences, there exists a number of pulse sequences which recouple dipolar interactions between nitrogen and carbon, such as REDOR (Gullion and Schaefer, 1989), TEDOR (Hing et al., 1992), and variants of these (Michal and Jelinski, 1997; Jaroniec et al., 1999; Jaroniec et al., 2001; Jaroniec et al., 2002). The magnetization transfer is carried out via rotor-synchronized π-pulses, which disturb the averaging of dipolar interactions by magic-angle spinning. The signal intensity as a function of the number of rotor cycles used for magnetization transfer yields an oscillation, which is modulated by the dipolar coupling strength. This oscillation can be fitted to obtain the internuclear distance between the interacting nuclei. Combination of the TEDOR oscillation measurement with chemical shift correlation spectroscopy (i.e. a pseudo 3D spectrum where the third dimension represents the TEDOR oscillation) yields multiple distances even for larger biomolecules such as proteins (Helmus et al., 2008). In order to obtain precise distances, the ^{13}C,^{15}N 2D spectrum must be completely resolved considering that peak overlap strongly biases the individual peak intensity. The resolution thereby strongly depends on the size and nature of the system to be studied. This makes the use of sparse labeling techniques especially interesting owing to the great reduction of the number of signals in the spectrum (Nieuwkoop et al., 2009). Along with the resolution in the 2D carbon-nitrogen spectrum, the long measuring time due to the requirement of a series of 2D spectra appears as another drawback of the method.

Nevertheless, it has been shown that it is possible to obtain a three-dimensional structure of a microcrystalline protein solely based on TEDOR distances of two glycerol-labeled samples with an above-average accuracy of 0.8 Å (Nieuwkoop et al., 2009; PDB 2KQ4). This can be explained by the fact that TEDOR oscillations can be fitted to obtain rather accurate distances, whereas spin diffusion-based experiments, introduced in the following section, suffer from relayed polarization transfer, which prohibits the determination of accurate distances. Using a mixture of u^{13}C- and u^{15}N-labeled protein, it was possible to specifically determine intermolecular contacts that allowed the calculation of a supramolecular structure including five individual monomers in the crystal (Nieuwkoop and Rienstra, 2010; PDB 2KWD).

Proton-mediated polarization transfer

Besides the dipolar recoupling sequences, there exists a variety of experiments which yield structural information by polarization transfer between heavy atoms with the help of the surrounding protons. The most widely used experiment of this type is the proton driven spin diffusion (PDSD) experiment owing to its simplicity and robustness. Despite the fundamental difference of the magnetization transfer mechanism during the mixing time,

the pulse sequence is very similar to that of the NOESY experiment. The transfer of magnetization is a relaxation process which is based on the dipolar coupling between two carbon spins as well as a frequency overlap of the two spins that exchange magnetization. The frequency overlap is accomplished by removing the proton-decoupling during the mixing time. This provokes significantly broadened carbon lineshapes and thus frequency overlap which allows magnetization transfer.

The transfer via proton-driven spin diffusion is governed by higher order Hamiltonians, since the zeroth-order Hamiltonian is averaged out by MAS. The first order average Hamiltonian includes terms that measure the carbon-carbon, carbon-proton, and proton-proton interactions. The averaging of the zeroth order Hamiltonian is the reason for reduced dipolar truncation in proton-driven spin diffusion. The carbon-proton and proton-proton interactions broaden the energy levels belonging to the zero-quantum transition which reduces the energy difference and hence allows the polarization transfer (Grommek et al., 2006).

The rate of polarization transfer under MAS was described by Kubo and McDowell (Kubo and McDowell, 1988) and is presented in Equation 3.1,

$$k_{ij} = \frac{1}{15}\omega_{ij,\mathrm{eff}}^2(G_{ij}(\nu_\mathrm{r}) + \frac{1}{2}G_{ij}(2\nu_\mathrm{r})) \tag{3.1}$$

where $\omega_{ij,\mathrm{eff}}^2$ is the squared effective dipolar coupling (Equation 3.2).

$$\omega_{ij,\mathrm{eff}} = \frac{\mu_0\gamma_\mathrm{C}^2\hbar}{4\pi r_{ij}^3} \tag{3.2}$$

The rate of polarization exchange thus depends on the squared effective dipolar coupling between the two spins as well as on a function G, which measures the zero-quantum lineshape at the MAS spinning frequency ν_r and twice the spinning frequency ($G(\nu_\mathrm{r})$ and $G(2\nu_\mathrm{r})$). The zero-quantum lineshape is the Fourier-transform of the FID of the two carbon spins i and j in the presence of the surrounding protons and magic angle spinning. This term measures the frequency overlap of the two carbon spins, which is equivalent to the energy difference of the zero-quantum transition in the energy level diagram. The highest transfer efficiency is observed if the energy levels are broadened, thus generating a small energy level difference. This corresponds to overlapping single-quantum lineshapes of the two carbon atoms (Dumez and Emsley, 2011).

While the PDSD pulse sequence does not use any rf-irradiation during the mixing time, there are several pulse sequences which are based on the same principle for polarization transfer but use rf-fields during the mixing time in order to increase the transfer rate.

Examples include dipolar assisted rotational resonance (DARR) (Takegoshi et al., 2001; Takegoshi et al., 2003) or proton-assisted recoupling (PAR) (Paëpe et al., 2008).

As opposed to pure dipolar recoupling sequences, it has been shown that dipolar truncation is not attenuating the polarization transfer between distant carbon atoms in proton-driven spin diffusion experiments (Grommek et al., 2006). This enables the measurement of meaningful long-range distance restraints for structure determination. Distance information can be obtained in rather large quantities, the individual distances are, however, of limited accuracy when using simple calibration methods as they are used to calibrate NOEs in solution NMR. This can be ascribed to the fact that the polarization transfer rate is not only dependent on the internuclear distance, but also on the zero-quantum lineshape. Furthermore, polarization is not always transferred directly between two spins, but relayed transfer via intermediate spins frequently occurs and biases the measured peak intensity. Nevertheless, distance restraints from PDSD-type experiments represent the most commonly used source of structural information, especially due to the simplicity and robustness of the experiment, resulting in a comparatively large number of long-range signals.

Another experiment type allows direct polarization exchange between protons combined with chemical shift evolution of heavy atoms, resulting in a spectrum where proton contacts are measured indirectly via heavy atom signals, i.e. CHHC and NHHC (Lange et al., 2002; Lange et al., 2003). The indirect detection of the signals via heavy atoms is required in order to obtain sufficient resolution. The experiment is based on two short cross-polarization steps prior to and after the ^{1}H-^{1}H mixing time which transfers the magnetization from a heavy atom to its directly bonded protons for magnetization exchange and back to the respective heavy atom for detection. The advantage of this type of experiment is that distances of structurally meaningful long-range contacts are smaller for protons than for the corresponding heavy atoms. At the same time, protons are never directly bonded, which makes the distance of short range interactions larger. Altogether, this decreases the distance difference between short-range and long-range contacts when compared to heavy atoms, leading to more equal peak intensities for the corresponding short-range and long-range signals. This improves one major problem of spectra relying on polarization transfer among carbon atoms, namely the disappearance of very weak long-range signals through the overlay of strong short-range signals. CHHC and NHHC spectra have been used as a sole source of structural information for the structure calculation of the Ktx-protein (Lange et al., 2005), however, they have as well been used in addition to previously mentioned experiment types (Loquet et al., 2008; Balayssac et al., 2008).

Paramagnetic effects

Pseudocontact shifts (PCS) in the presence of a paramagnetic center can be used to obtain distance information (Balayssac et al., 2008). This results from the fact that a paramagnetic compound influences the peak position of surrounding atoms in a distance- and orientation-dependent manner. The size of the disturbance can therefore be transformed into a distance restraint between the paramagnetic center and the respective atom. In order to use this information, it is necessary to distinguish intramolecular from intermolecular pseudo-contact shifts. This can be achieved by diluting a fully-labeled version of the paramagnetic species with an unlabeled diamagnetic version of the protein to be studied. A paramagnetic compound can either be naturally available, for example in case of metallo-proteins, or it can be added chemically via cysteine residues. The structure of the 16.8 kDa protein MMP-12 was calculated based on distance restraints from PCS in addition to a set of distance restraints from spin diffusion experiments (Balayssac et al., 2008; Bertini et al., 2010).

Paramagnetic relaxation enhancement (PRE) is a second effect caused by paramagnetic compounds and can be used to obtain structural information in a way similar to PCS. Its value for structure calculation has been demonstrated in solution NMR especially in cases of limited NOE data, however, PRE measurement is not limited to solution NMR. The possibility to measure PREs in the solid-state has been demonstrated using the microcrystalline protein GB1 (Nadaud et al., 2007). The diamagnetic nature of GB1 requires the addition of a paramagnetic compound such as Cu^{2+}-EDTA, which is achieved chemically via a cysteine residue. Distance information can be extracted when comparing NMR spectra in the presence and absence of the paramagnetic compound, whose effect is the attenuation or complete disappearance of NMR signals in a distance-dependent manner. Effects can be seen at distances up to 20 Å away from the spin label, yielding distance restraints of much more global nature when compared to experiments relying on dipolar couplings. In case of multimeric or microcrystalline samples, it is necessary to distinguish intramolecular from intermolecular effects in the same way as it is necessary for the detection of PCS. It is therefore required to use diluted samples in order to minimize intermolecular effects. The value of PRE restraints for structure calculation was demonstrated by their use as a sole source of structural information, yielding a structure bundle with an RMSD bias of 1.8 Å away from the x-ray reference structure (Sengupta et al., 2012; Sengupta et al., 2013).

3.2.2 Orientational restraints

Dipolar lineshape analysis

Dipolar lineshapes report on the relative orientation of two ^{1}H-^{15}N, ^{1}H-^{15}N (Reif et al., 2000) or ^{1}H-^{15}N, ^{1}H-^{13}C (Rienstra et al., 2002) vectors, thus providing information about torsion angles in the protein backbone with relatively high accuracy when compared to database-based approaches such as TALOS. The experiments combine 2D chemical shift correlation spectroscopy in order to resolve individual ^{15}N-^{15}N or ^{15}N-^{13}C atom pairs with the T-MREV sequence (Hohwy et al., 2000) during an indirect evolution time to recouple the dipolar interaction between ^{1}H-^{15}N and ^{1}H-^{15}N or ^{1}H-^{15}N and ^{1}H-^{13}C, depending on the nuclei chosen for the 2D spectrum. This results in a dipolar lineshape in the third dimension, which depends on the two dipolar tensors (^{1}H-^{15}N, ^{1}H-^{15}N or ^{1}H-^{15}N, ^{1}H-^{13}C), their relative orientation as well as the transfer between ^{15}N and ^{15}N or ^{15}N and ^{13}C. These dipolar lineshapes can subsequently be fitted in order to obtain the desired torsion angle information. Applications include the determination of the majority of torsion angles in the uniformly labeled protein GB1 based on a single 3D experiment (Franks et al., 2006) as well as their incorporation into the protein structure calculation of GB1 as vector angle (VEAN) restraints (Franks et al., 2008).

Chemical shift tensors (CST)

It has long been known that isotropic Cα chemical shift values are highly sensitive to the local structure of the protein backbone (i.e. alpha-helix or beta-sheet) (Spera and Bax, 1991). Similarly, the chemical shift tensors (CST) especially of ^{13}Cα and ^{15}N atoms provide information about the backbone torsion angles. Restoring the chemical shift anisotropy during magic-angle spinning can be performed using the Recoupling of the Chemical Shift Anisotropy (ROCSA) sequence (Chan and Tycko, 2003), which can be applied in combination with chemical shift correlation spectroscopy to obtain site-specific CSA information of proteins (Wylie et al., 2005). The ROCSA trajectories can be fitted to extract the CSA tensor parameters, which can then be transformed into structural information by comparison to *ab initio* chemical shielding surfaces (Pearson et al., 1997; Havlin et al., 2001; Sun et al., 2002) that are available for all 20 common amino acids. Cα CST have been included in structure refinement of the protein GB1 by minimizing the deviation between measured tensor magnitudes and calculated tensors based on the dihedral angles of a structural model (Wylie et al., 2009). A more advanced use of CST includes the determination of tensor magnitudes and additionally tensor orientations with respect to the respective dipole tensor of the ^{1}H,^{15}N (or ^{1}H,^{13}C) bond corresponding to the CST

using a series of several 3D experiments with tensor recoupling in the indirect dimension (Wylie et al., 2011).

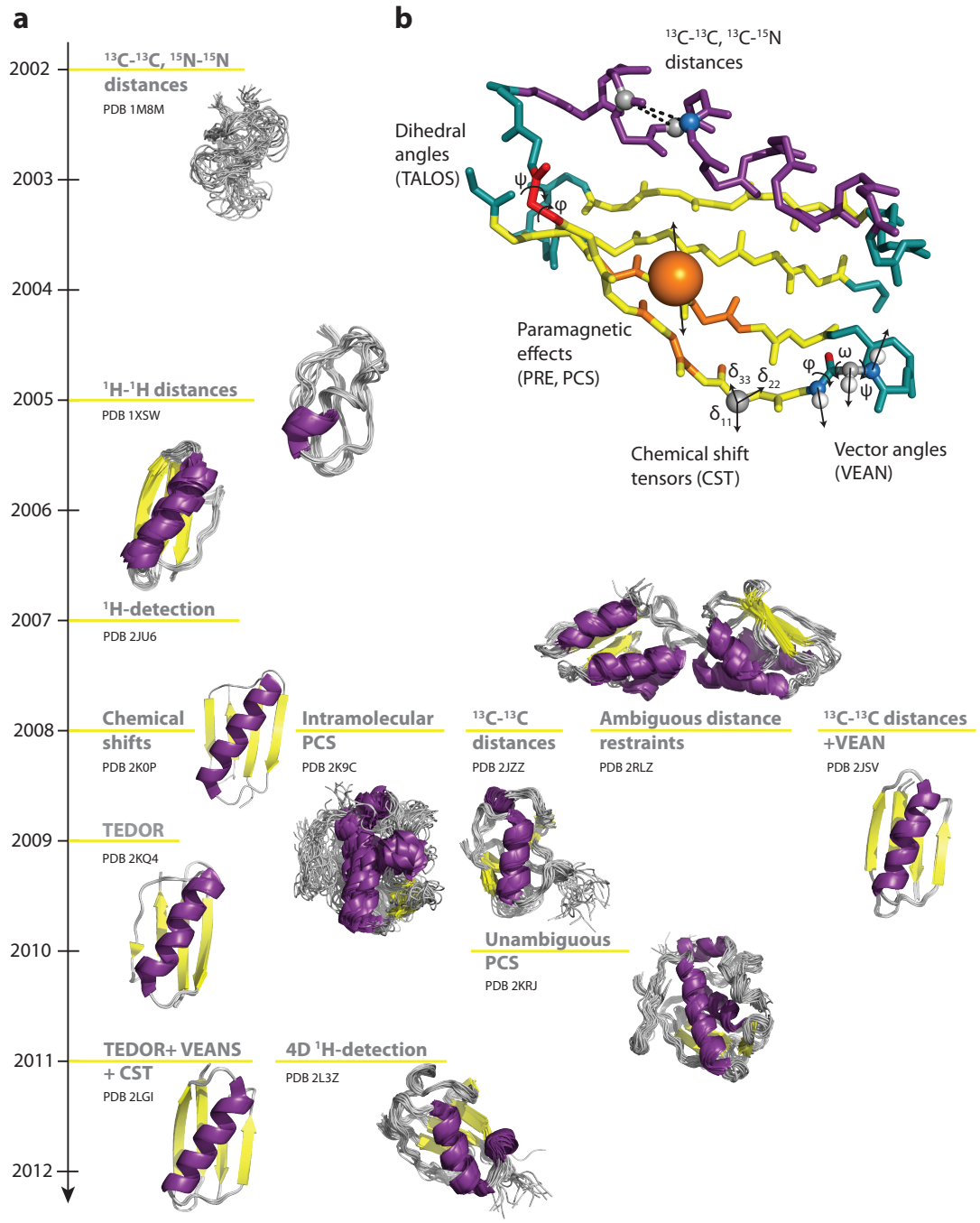


Figure 3.3: **a** Timeline for the methodological development of structure determination by solid-state NMR and its application to microcrystalline model proteins. Only those structures are depicted that have been deposited in the PDB. Secondary structure elements are highlighted in purple (α-helix) and yellow (β-sheet). A more complete overview is given in Table 3.1. **b** Overview of the different types of structural restraints available from solid-state NMR. The GB1 crystal structure (PDB 2QMT) was used for the graphical representation. This figure is based on Comellas and Rienstra, 2013.

Table 3.1: Protein structures determined by solid-state NMR for methodological development

Protein (MW, PDB code)	Reference	RMSD bias [Å]	RMSD radius [Å]	Software	Samples	Spectra	Assignment	Upl-values
SH3 (7.2 kDa, 1M8M)	Castellani et al. (2002)	2.6	1.6	CNS	u-^{13}C/^{15}N 1,3-^{13}C/^{15}N 2-^{13}C/^{15}N	DARR (500 ms)	Manual	distance classes
SH3 (7.2 kDa)	Castellani et al. (2003)	1.2	0.7	CNS	u-^{13}C/^{15}N 1,3-^{13}C/^{15}N 2-^{13}C/^{15}N	DARR (500 ms) NCACX (500 ms) NCOCX (500 ms)	Manual	Fixed (7.5 Å)
Ubiquitin (8.6 kDa)	Zech et al. (2005)	-	~1	CNS	u-^{13}C/^{15}N 2-^{13}C/^{15}N	DARR (10-500 ms)	Manual	distance classes
Ktx (4.2 kDa)	Lange et al. (2005)	1.9	0.8	CNS	u-^{13}C/^{15}N	CHHC (250-400 μs)	Manual	Calibration ($1/r^6$)
SH3 (7.2 kDa)	Fossi et al. (2005)	1.3	0.7	SOLARIA	u-^{13}C/^{15}N 1,3-^{13}C/^{15}N 2-^{13}C/^{15}N	DARR (500 ms) NCACX (500 ms) NCOCX (500 ms)	Automatic	Fixed (7.5 Å)
GB1 (6.2 kDa, 2JU6)	Zhou et al. (2007)	1.9	0.8	XPLOR-NIH	u-^{13}C/^{15}N/^{2}H (^{1}H backexchanged)	CON(H)H (2 ms RFDR) N(H)H (2 ms, 3 ms RFDR)	Manual	distance classes
GB1 (6.2 kDa, 2JSV)	Franks et al. (2008)	1.3	0.7	XPLOR-NIH	u-^{13}C/^{15}N 1,3-^{13}C/^{15}N 2-^{13}C/^{15}N	N(H)H N(HH)C DARR (50-500 ms) 3D NCC NN-PDSD DipolarLineShape (VEAN)	Manual	distance classes
Crh (9.4 kDa(x2), 2RLZ)	Loquet et al. (2008)	1.6	0.8	ARIA	u-^{13}C/^{15}N	CHHC (125 μs,200 μs) NHHC (100 μs) DARR (200 ms)	Manual and automatic	Fixed (5.0 Å H-H, 7.0 Å C-C)
Ubiquitin (8.6 kDa)	Manolikas et al. (2008)	1.6	0.7	ATNOS-CANDID	u-^{13}C/^{15}N	DARR (100 ms, 250 ms, 400 ms)	Manual, Semi-automatic Automatic	Calibration ($1/r^6$)

MMP-12 16.8 kDa, 2K9C)	Balayssac et al. (2008)	3.1	3.0	CYANA	u-^{13}C/^{15}N mixed u-^{13}C/^{15}N and unlabeled	PDSD CHHC PCS	Manual	Fixed (6.0 Å H-H, 9.0 Å C-C)
GB1 (6.2 kDa, 2KQ4)	Nieuwkoop et al. (2009)	0.8	0.25	XPLOR-NIH	1,3-^{13}C/^{15}N 2-^{13}C/^{15}N	TEDOR	Manual	TEDOR Oscillation fit
GB1 (6.2 kDa)	Wylie et al. (2009)	1.1	0.4	XPLOR-NIH	1,3-^{13}C/^{15}N 2-^{13}C/^{15}N	N(H)H N(HH)C DARR(50-500 ms) 3D NCC NN-PDSD CST Magnitudes	Manual	distance classes
GB1 (6.2 kDa, 2KWD)	Nieuwkoop and Rienstra (2010)	2.9	0.42	XPLOR-NIH	mixed u-^{13}C and u-^{15}N (50:50) u-^{15}N (50:50)	TEDOR	Manual	distance classes
MMP-12 (16.8 kDa, 2KRJ)	Bertini et al. (2010)	1.3	1.0	ARIA, CYANA	u-^{13}C/^{15}N	PDSD+DARR (50-800 ms) PAR (15,20 ms) PAIN (15 ms) PCS	Manual, Automatic	Fixed (6.5-9.0 Å)
GB1 (6.2 kDa, 2LGI)	Wylie et al. (2011)	0.51	0.16	XPLOR-NIH		TEDOR CST Magnitudes CST Orientations	Manual	
Ubiquitin (8.6 kDa, 2L3Z)	Huber et al. (2011)	1.57	0.72	CYANA	u-^{2}H/^{15}N (30 % backexchanged) VL-13CD2H	4D methyl-methl DREAM 3D HN-HN DREAM	Manual	
SH3 (7.2 kDa)	Bardiaux et al. (2012)	0.9	0.64	ARIA				
GB1 (6.2 kDa)	Sengupta et al. (2012), Sengupta et al. (2013)	1.8	-	XPLOR-NIH	6x u-^{13}C/^{15}N + Cu^{2+}-EDTA	PRE	Manual	Fit of relaxation trajectories

3.3 Application to membrane proteins and amyloid fibrils

Solid-state NMR is of special interest for the investigation of molecular systems that are not accessible to the standard structure determination methods, such as solution NMR and X-ray diffraction. These systems include amyloid fibrils as well as membrane proteins in their native phospholipid environment. It has been shown in recent years that 3D structures of complex molecules can in principle be determined by solid-state NMR at high resolution. The following section gives an overview of these structures with a main focus on amyloid fibrils and membrane proteins owing to their special relevance for biomedical questions.

In contrast to microcrystalline samples of rather rigid proteins, membrane proteins are characterized by limited order and dynamic properties causing an increase in the observed line width in solid-state NMR spectra. This is especially problematic if it occurs in combination with high molecular weight, resulting in a large number of overlapped signals. The general approach based on distance restraints is therefore not easily applicable for large membrane proteins. However, a fundamentally different solid-state NMR approach based on oriented samples instead of magic-angle spinning has been successfully applied to structure determination of several membrane proteins such as Gramicidin (Ketchem et al., 1993), AchR M2 channel (Opella et al., 1999), fd coat protein (Marassi and Opella, 2003), HIV Vpu channel (Park et al., 2003), the integral membrane domain of the mercury transporter Mer F (Angelis et al., 2006), or the Influenza M2 channel (Sharma et al., 2010). This technique requires the phospholipid bilayer including the protein to be oriented perpendicular to the external magnetic field, which can either be induced magnetically or on glass surfaces. Fast rotational diffusion around the bilayer-normal allows the measurement of accurate orientation restraints based on motionally averaged powder patterns for ^{1}H-^{13}Cα and ^{1}H-^{15}N amide heteronuclear dipolar couplings as well as ^{13}Cα,^{13}C', and ^{15}N chemical shift anisotropies. However, the dense network of ^{13}C atoms and their dipolar coupling in uniformly ^{13}C/^{15}N labeled proteins, which is not averaged out in the absence of magic-angle spinning, prohibits the direct detection of carbons and limits the use of triple-resonance experiments.

A new method introduced by Opella and coworkers, *rotationally aligned (RA) solid-state NMR* (Das et al., 2012), combines the advantages of oriented sample solid-state NMR (i.e. the measurement of accurate orientational restraints) with the ability to resolve and assign individual peaks in uniformly ^{13}C/^{15}N labeled proteins through the use of magic-angle spinning solid-state NMR. This was enabled by the development of experiments that allow the measurement of CSA and DC powder patterns during MAS (Chan and Tycko, 2003; Wylie et al., 2006). The method was successfully applied in the structure calculations of the integral membrane domain of the mercury transporter Mer F (Das et al., 2012), the full-length Mer F transporter (Lu et al., 2013), as well as the CXCR1

receptor (Park et al., 2012), representing the first structure of a receptor from the GPCR family determined without modifications of the protein sequence in a native phospholipid-environment. The structure calculation itself was performed using the chemical shift-based molecular fragment replacement approach implemented in CS-Rosetta (Shen et al., 2008).

Additional structures of membrane proteins have been determined using the classical approach based on distance restraints obtained from MAS solid-state NMR. The structure of the transmembrane domain of the *Yersinia enterocolitica* adhesin A (YadA) was calculated using distance restraints of 24 homo- and heteronuclear correlation spectra at different mixing times recorded on a single uniformly $^{13}C/^{15}N$ labeled microcrystalline sample (Shahid et al., 2012). The sample was obtained from crystallization trials that did not successfully yield well diffracting single-crystals suitable for X-ray diffraction. The structure of the trimeric YadA protein, calculated using the inferential structure determination method (Rieping et al., 2005), forms a β-barrel surrounding three N-terminal transmembrane helices that protrude into the extracellular space. A similar approach was used for the structure calculation of the disulfide bond generating membrane protein DsbB, which was determined in the lipid bilayer (Tang et al., 2013). Input data included X-ray reflections in addition to the ambiguous distance restraints obtained from 2D ^{13}C-^{13}C correlation spectra and dihedral angles from chemical shift analysis. The structure was subsequently used to generate a model in the lipid bilayer based on MD simulation. MAS solid-state NMR was used to calculate the structure of *Anabaena* sensory rhodopsin (ASR) reconstituted into a phospholipid bilayer (Wang et al., 2013). Structural restraints included a set of PRE restraints, distance restraints, as well as dihedral angles from chemical shift analysis obtained from NMR spectra recorded on three differently labeled samples based on 1,3-^{13}C glycerol, 2-^{13}C glycerol, and regenerated using doubly ^{13}C-labeled retinal. Structure calculation was performed through a combination of an initial manual assignment of cross-peaks with a subsequent automated assignment of additional cross-peaks. The topology of the trimeric ASR closely resembles that of the seven-transmembrane helix protein bacteriorhodopsin (BR). Details on the presented structure determinations can be found in Table 3.2 and a graphical representation of the structures is depicted in Fig. 3.4.

Amyloid fibrils typically form in the course of several known diseases, such as Alzheimer's disease, Parkinsons's disease, Huntington's disease, or prion diseases, however, they can also be of functional relevance, for example in the storage of peptide hormones. Amyloid fibrils are non-soluble and non-crystallizable which makes solid-state NMR uniquely positioned to study these systems at atomic resolution. The formation of amyloid fibrils is a nucleation-based process of normally soluble proteins, resulting in a fiber structure, which is usually dominated by β-sheet secondary structure combined with unstructured parts (Comellas and Rienstra, 2013).

For structure calculation based on distance restraints, it is helpful to have some information about the number and symmetry of the monomers within one fibril layer, which can for example be deduced from the mass per unit length (MPL) value measured by scanning transmission electron microscopy (STEM). Due to the tight packing of monomers within the fibril, NMR spectra contain intramolecular as well as intermolecular peaks, which need to be separated in order to guide the structure calculation in the correct direction. This is usually achieved by measuring NMR spectra of several samples. Uniformly $^{13}C/^{15}N$-labeled protein mixed with unlabeled protein yields mostly intramolecular signals, whereas the ^{13}C-labeled monomer mixed with ^{15}N-labeled monomer yields NMR spectra with purely intermolecular signals. Using these strategies, the structure of the fungus protein HET-s (Wasmer et al., 2008; Melckebeke et al., 2010) and two structures of the β-Amyloid fibril in Alzheimer's disease (Lu et al., 2013; Schütz et al., 2014) have been solved at atomic resolution (Table 3.2, Fig. 3.4).

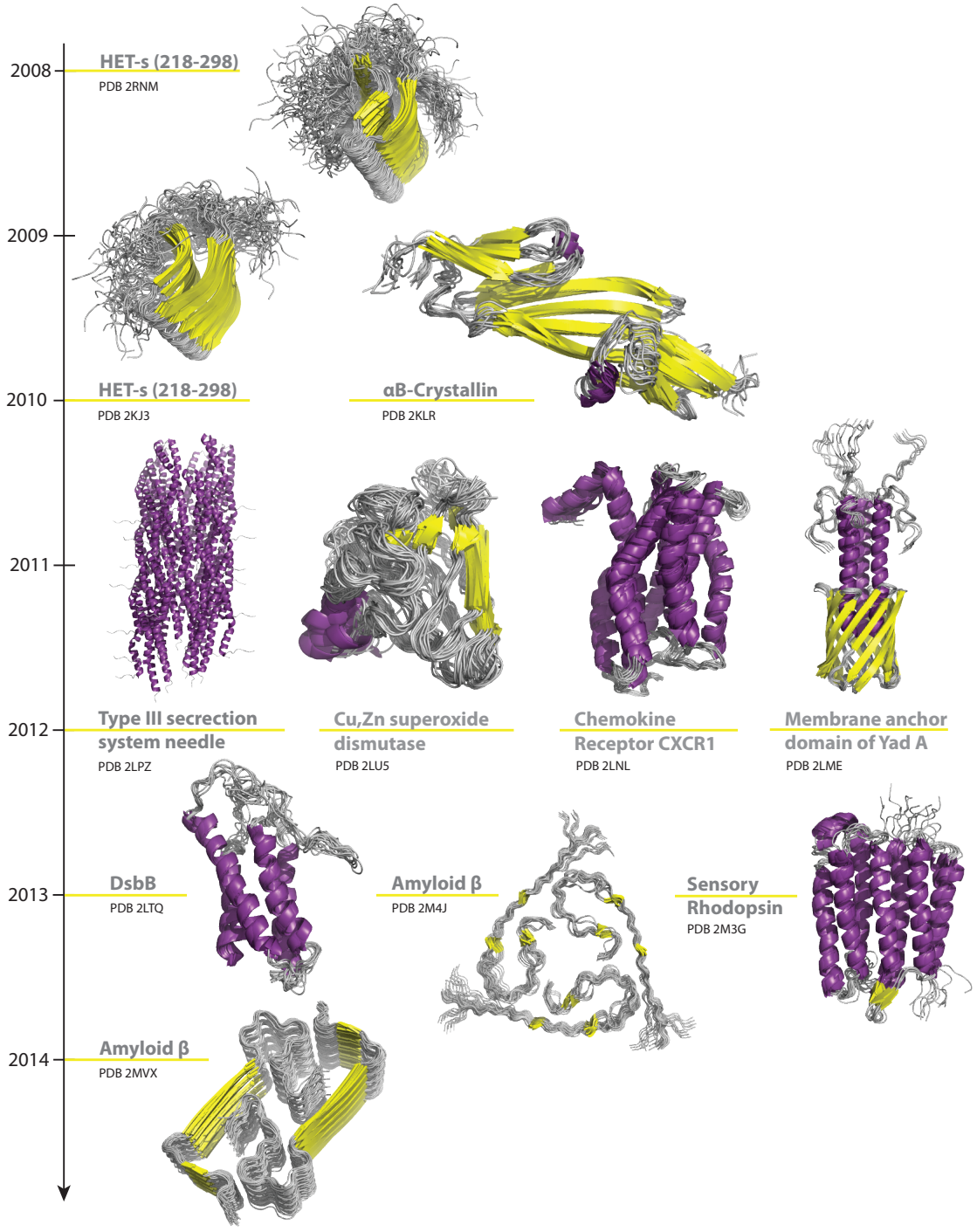


Figure 3.4: Timeline of membrane protein and amyloid fibril structures determined by solid-state NMR spectroscopy. Secondary structure elements are highlighted in purple (α-helix) and yellow (β-sheet). More details about each structure calculation are given in Table 3.2.

Table 3.2: De novo protein structures determined by solid-state NMR

Protein (MW[a], PDB code)	Reference	RMSD bias [Å]	RMSD radius [Å]	Software	Samples	Spectra	Assignment	Upl-values
Het-s (8.67 kDa, 2RNM)	Wasmer et al. (2008)	-	0.6	CYANA	u-^{13}C/^{15}N 2-^{13}C/^{15}N mixed u-^{13}C and u-^{15}N	CHHC NHHC PDSD	Manual	distance classes
Het-s (8.67 kDa, 2KJ3)	Melckebeke et al. (2010)	-	0.64	CYANA	u-^{13}C/^{15}N 2-^{13}C/^{15}N mixed u-^{13}C and u-^{15}N	PDSD CHHC NHHC PAIN PAR	Manual	Fixed (3.0-8.0 Å)
aB-Crystallin (20.2 kDa, 2KLR)	Jehle et al. (2010)		1.63	SOLARIA MDSA CNS ARIA	u-^{13}C/^{15}N 20 % u-^{13}C/^{15}N 50 % 2-^{13}C/^{14}N+ 50 % ^{12}C/^{15}N 50 % 1,3-^{13}C/^{14}N+ 50 % ^{12}C/^{15}N 2-^{13}C/^{15}N 1,3-^{13}C/^{15}N PDMW ^{13}C/^{15}N	PDSD CHHC NHHC TEDOR PAIN NCACX NCOCX	Semi-automatic	
Cu,Zn superoxide dismutase (15.84 kDa, 2LU5)	Knight et al. (2012)	-	1.66	UNIO	u-^{13}C/^{2}H/^{15}N (100 % backexchanged)	PRE (H)NHH-RFDR	Automatic	
Type-III Secretion System Needle (8.86 kDa, 2LPZ)	Loquet et al. (2012)	-	2.1	Rosetta	1-(Gluc) ^{13}C/^{15}N 2-(Gluc) ^{13}C/^{15}N	PDSD (850 ms, 1000ms) PAIN-CP	Manual	Fixed (5.0 Å)

[a] The molecular weight is specified for the monomeric subunit.

Chemokine Receptor CXCR1 (33.72 kDa, 2LNL)	Park et al. (2012)	-	1.7	Rosetta	u-^{13}C/^{15}N 2-^{13}C/^{15}N	13C detected SLF (1H-15N DC, 1H-13Ca DC)	Manual	
Membrane anchor domain of YadA (11.29 kDa, 2LME)	Shahid et al. (2012)	-	0.84	ARIA (ISD)	u-^{13}C/^{15}N	DARR (15-500 ms) PDSD (15-100 ms) CHHC NHHC (35-500 μs) PAR (2.25-15 ms) TEDOR (2.25-12 ms) ^{13}C-^{13}C methyl filtered DARR (300 ms)	Semi-automatic	Fixed (3.5-8.0 Å)
DsbB (70.39 kDa, 2LTQ)	Tang et al. (2013)	-	1.35	PASD (XPLOR-NIH)	1,3-^{13}C/^{15}N 2-^{13}C/^{15}N	DARR (300,500 ms)	Automatic	
Sensory rhodopsin (27.51 kDa, 2M3G)	Wang et al. (2013)	-	0.84	CNS, ARIA	1,3-^{13}C/^{15}N 2-^{13}C/^{15}N S26C u-^{13}C/^{15}N +Nitroxide label	PDSD (100,250,500 ms) CHHC (250 μs) 3D HBR (70 ms) PRE	Manual, Automatic	Fixed
Amyloid β fibril (4.33 kDa, 2M4J)	Lu et al. (2013)	-	-	XPLOR-NIH	u-^{13}C/^{15}N 6 selectively labeled samples	fpRFDR (11.77 ms) RAD (50,500 ms) PAR (13 ms) TEDOR (3.82,5.29 ms) ^{15}N- and ^{13}C-BARE fsREDOR PITHIRDS-CT	Manual	Fixed, Oscillation fit (fsREDOR, BARE, PITHIRDS-CT)
Amyloid β fibril (4.21 kDa, 2MVX)	Schütz et al. (2014)	-	0.19	CYANA	u-^{13}C/^{15}N 2-Glc-^{13}C 1:1 u-^{13}C,u-^{15}N 1:4 u-^{13}C/^{15}N, unlabeled	CHHC (500 μs) PAR (8 ms) PAIN (6 ms) DARR (400 ms) PDSD (4000 ms)	Manual, Automatic	Fixed

Part II

General method development

Chapter 4

Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA

This chapter is based on the following publication:

Buchner L. and Güntert P. Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA. *J Biomol NMR*, doi:10.1007/s10858-015-9921-z, 2015

4.1 Introduction

The structure determination of biological macromolecules by NMR in solution relies primarily on distance restraints derived from cross peaks in NOESY spectra. A large number of assigned NOESY cross peaks are necessary to compute an accurate three-dimensional (3D) structure because many of the NOEs are short-range with respect to the sequence and thus carry little information about the tertiary structure and because NOEs are generally interpreted as loose upper bounds in order to implicitly take into account internal motions and spin diffusion (although, in principle, accurate distance measurements are possible with NOEs (Vögeli et al., 2012; Vögeli et al., 2009)). Obtaining a comprehensive set of distance restraints from NOESY spectra is in practice not straightforward. The sheer amount of data, as well as resonance and peak overlap, spectral artifacts and noise, and the absence of expected signals because of fast relaxation turn interactive NOESY cross peak assignment into a laborious and error-prone task. Therefore, the development of computer algorithms for automating this often most time-consuming step of a protein structure determination by NMR has been pursued intensely and reviewed extensively (Altieri and Byrd, 2004; Baran et al., 2004; Billeter et al., 2008; Gronwald and Kalbitzer, 2004; Guerry and Herrmann, 2011; Güntert, 1998; Güntert, 2003; Güntert, 2009; Moseley and Montelione, 1999; Williamson and Craven, 2009). Besides semi-automatic approaches (Duggan et al., 2001; Güntert et al., 1993; Meadows et al., 1994), several algorithms have been developed for the automated analysis of NOESY spectra given the chemical shift assignments, namely NOAH (Mumenthaler and Braun, 1995; Mumenthaler et al., 1997), ARIA (Nilges et al., 1997; Rieping et al., 2007), ASDP (Huang et al., 2006), KNOWNOE (Gronwald et al., 2002), CANDID (Herrmann et al., 2002b), PASD (Kuszewski et al., 2004), AutoNOE-Rosetta (Zhang et al., 2014), and a Bayesian approach (Hung and Samudrala, 2006). Automated NOESY peak picking has been integrated into the method (Herrmann et al., 2002a). Automated NOESY assignment can be combined with automated sequence-specific resonance assignment with the Garant (Bartels et al., 1997) or FLYA (Schmidt and Güntert, 2012) algorithms in order to perform a complete NMR structure determination without manual interventions (López-Méndez and Güntert, 2006). In favorable cases, this can even be achieved using exclusively experimental data from NOESY spectra (Ikeya et al., 2011; Schmidt and Güntert, 2013).

The fundamental problem of NOESY assignment is the ambiguity of cross peak assignments. Assigning based solely on the match between cross peak positions and the chemical shift values of candidate resonances does in general not yield a sufficient number of unambiguously assigned distance restraints to obtain a structure (Mumenthaler et al., 1997). Ambiguous distance restraints make it possible to use also NOEs with multiple assignment possibilities in a structure calculation (Nilges, 1995). Nevertheless, additional

criteria have to be applied to resolve these ambiguities, such as using secondary structure information (Huang et al., 2006) or a preliminary structure that is refined iteratively in cycles of NOE assignment and structure calculation (Mumenthaler and Braun, 1995). The CANDID automated NOESY assignment method introduced the concepts of network anchoring to reduce the initial ambiguity of NOE assignments and constraint combination to reduce the impact of erroneous restraints (Herrmann et al., 2002b). In CYANA, the conditions applied by CANDID for valid NOE assignments have been reformulated in a probabilistic framework that is conceptually more consistent and better capable to handle situations of high chemical shift-based ambiguity of the NOE assignments (Güntert, 2009; Güntert and Buchner, 2015).

The aforementioned approaches can go wrong in two ways, especially with low-quality input data. Either the algorithm fails to ever assign enough NOE distance restraints to obtain a defined structure. This outcome, manifested by a divergent structure bundle with a high RMSD, is unfortunate but straightforward to detect. More problematic are failures of a second kind, where the algorithm, possibly gradually over several cycles, discards part of the NOE cross peaks (by letting them unassigned) and selects a self-consistent but incomplete subset of the data to compute a well-defined but erroneous structure, i.e. a tight bundle of conformers with low RMSD to its mean coordinates that, however, differs significantly from the (unknown) correct structure of the protein. If this outcome goes unnoticed, it may result in the publication or PDB deposition of erroneous structures that cannot be detected easily by coordinate-based validation tools (Nabuurs et al., 2006).

Given the widespread use of automated NOESY assignment algorithms (Guerry and Herrmann, 2011; Williamson and Craven, 2009) it is important to give criteria for their safe application (Herrmann et al., 2002b) and to assess their reliability. It is known that the CANDID algorithm generally requires a high degree of completeness of the backbone and side chain chemical shift assignments (Jee and Güntert, 2003). Recently, the CASD-NMR initiative (Rosato et al., 2009) has evaluated several NMR structure determination methods by blind testing. Using high-quality data sets of small proteins from a structural genomics project it was found that the NOESY-based methods included in the test yielded structures with an accuracy of 2 Å RMSD or better to the subsequently released reference structures (Rosato et al., 2012). However, the situation is less clear for more difficult cases, in which the resonance assignments may be incomplete, spectral crowding, overlap, and low signal-to-noise ratios prevent collecting a "complete" set of NOESY cross peaks, or the lack of isotope labeling may preclude the use of, intrinsically less ambiguous, 3D and 4D NOESY spectra. Further complications may arise with symmetric multimers or solid-state NMR data. We address these questions by an extensive, systematic analysis of the combined automated NOESY assignment and structure calculation algorithm in

CYANA under a variety of conditions mimicking data imperfections that may occur with challenging systems.

4.2 Methods

Experimental NMR data sets

The performance of CYANA was assessed on the basis of the NMR structure bundles of ten proteins to which we refer in the following by the four-letter acronyms given in Table 4.1: copz, the copper chaperone CopZ of *Enterococcus hirae* (Wimmer et al., 1999); cprp, the chicken prion protein fragment 128 – 242 (Calzolai et al., 2005); enth, the ENTH-VHS domain At3g16270 from *Arabidopsis thaliana* (López-Méndez et al., 2004; López-Méndez and Güntert, 2006); fsh2, the Src homology 2 domain from the human feline sarcoma oncogene Fes (Scott et al., 2004); fspo, the F-spondin TSR domain 4 (Pääkkönen et al., 2006); pbpa, the *Bombyx mori* pheromone binding protein (Horst et al., 2001); rhod, the rhodanese homology domain At4g01050 from *Arabidopsis thaliana* (Pantoja-Uceda et al., 2005; Pantoja-Uceda et al., 2004); wmkt, the *Williopsis mrakii* killer toxin (Antuch et al., 1996); scam, stereo-array isotope labeled (SAIL) calmodulin (Kainosho et al., 2006); ww2d, the second WW domain from mouse salvador homolog 1 protein (Ohnishi et al., 2007).

The proteins copz, cprp, enth, fsh2, pbpa, rhod and wmkt are proteins with a well-defined single-domain structure. The protein fspo has an unusual, less well-defined structure without regular secondary structure. The protein scam has two flexibly connected domains. The protein ww2d forms a symmetric dimer. For the original structure determinations the proteins were uniformly labeled with ^{13}C and ^{15}N, except for copz that was only ^{15}N labeled, wmkt that was unlabeled, and scam that was stereo-array isotope labeled (Kainosho et al., 2006). The completeness of the resonance assignments and the type and amount of NOESY data are summarized in Table 4.1. For most proteins the unassigned NOESY peak lists were the only source of conformational restraints. Exceptions are cprp and pbpa, whose data sets included 123 and 148 ϕ/ψ torsion angle restraints derived from Cα chemical shifts (Luginbühl et al., 1995), respectively, ww2d including 44 ϕ/ψ torsion angle restraints from TALOS (Cornilescu et al., 1999). In the data set of cprp the assignments of 18 NOESY cross peaks were kept fixed, as in the original structure determination (Calzolai et al., 2005). Disulfide bonds were restrained in cprp, fspo, pbpa, and wmkt. In scam the distances between the four calcium ions and their 16 ligands were restrained to the range 1.7 – 2.8 Å. No hydrogen bond restraints or other additional restraints were used. The original experimental data sets were used to determine a reference structure for each protein using the same computational schedule as for the subsequent

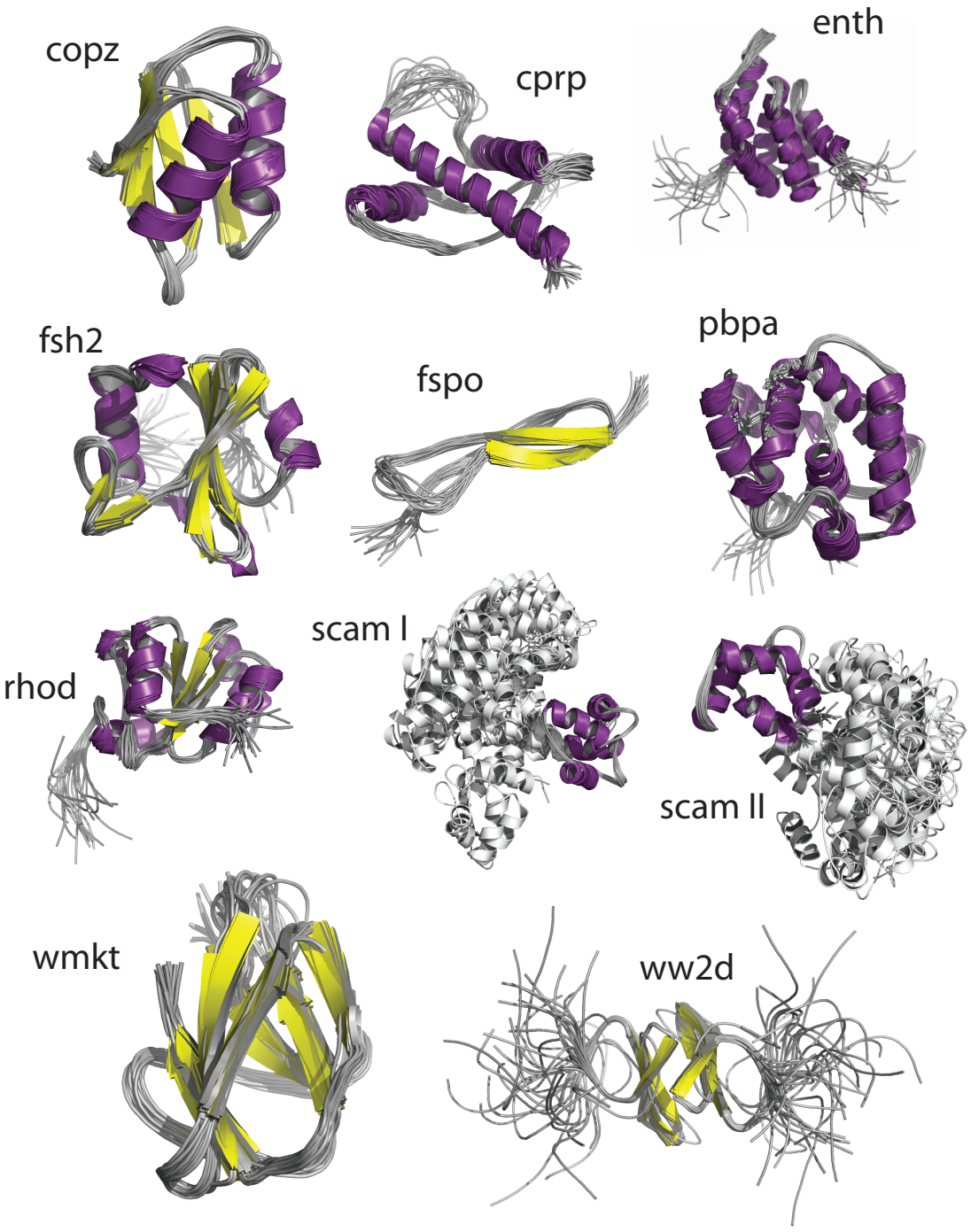


Figure 4.1: Bundle representations of the ten proteins included in the present study (see Table 4.1). Secondary structure elements are highlighted in purple (α-helix) and yellow (β-sheet). Atomic coordinates originate from the Protein Data Bank (PDB) entries 1CPZ (copz), 1U3M (cprp), 1VDY (enth), 1WQU (fsh2), 1VEX (fspo), 1GM0 (pbpa), 1VEE (rhod), 1X02 (scam), 1WKT (wmkt), and 2DWV (ww2d). Two separate superpositions are presented for the two-domain protein scam.

calculations with modified data. Seven cycles of combined automated NOESY assignment and structure calculation were performed, followed by a final structure calculation. In each cycle, structure calculations were started from 100 conformers with random values of the torsion angles, to which the standard CYANA simulated annealing schedule was applied with 10,000 torsion angle dynamics steps per conformer. The 20 conformers with the lowest final target function values were selected for analysis and are shown in Fig. 4.1.

Table 4.1: Summary of proteins and original NMR data sets

Acronym	PDB code	BMRB code	Amino acids	Chemical shifts	Ass. compl. [%]	NOESY spectra	Total Peaks
copz	1CPZ	4344	68	425	88.8	2D, ^{15}N	1175, 1063
cprp	1U3M	6269	117	1093	97.8	2D, ^{15}N, ^{13}C, aro	94, 1498, 2850, 166
enth	1VDY	5928	140	1471	96.0	^{15}N, ^{13}C	1730, 4169
fsh2	1WQU	6331	114	1101	97.2	^{15}N,^{13}C, aro	1313, 2979, 440
fspo	1VEX	10002	56	583	98.6	^{15}N, ^{13}C	494, 1398
pbpa	1GM0	4849	142	1409	99.3	^{15}N, ^{13}C, aro	1387, 3907, 320
rhod	1VEE	5929	134	1268	98.4	^{15}N,^{13}C, aro	1801, 3683, 354
wmkt	1WKT	5255	88	455	97.0	2D	1998
scam	1X02	6541	293	1214	100.0	^{15}N, ^{13}C, aro	1703, 3026, 85
ww2d	2DWV	10028	98	900	90.9	^{15}N, ^{13}C, inter	380, 1203, 62

The assignment completeness gives the percentage of the aliphatic and aromatic ^{1}H and backbone ^{1}HN resonances that are assigned. Codes for NOESY spectra types are: 2D, 2D [^{1}H,^{1}H]-NOESY; ^{15}N, 3D ^{15}N-resolved [^{1}H,^{1}H]-NOESY; ^{13}C, 3D ^{13}C-resolved [^{1}H,^{1}H]-NOESY; aro, 3D aromatic ^{13}C-resolved [^{1}H,^{1}H]-NOESY; inter, 3D ^{13}C-filtered ^{13}C-edited [^{1}H,^{1}H]-NOESY for the detection of intermolecular NOEs (Zwahlen et al., 1997). The numbers of NOESY peaks are given in the order of the spectra in the preceding column.

The experimental input data sets were modified in 14 different ways to mimic different kinds of data imperfections. All random data modifications were applied five times using different random numbers resulting in a total of 397 different data sets for each protein including the respective complete data set.

1. Missing chemical shift assignments

   A given percentage P between 0 and 40 % of randomly selected ^{1}H chemical shift assignments was deleted. Experimental NOESY peak lists were not changed.

   (a) Random deletion: The shifts to be deleted were chosen randomly among all assigned ^{1}H chemical shifts.

   (b) Deletion of side chain chemical shifts: The shifts to be deleted were chosen randomly among all side-chain ^{1}H chemical shift assignments.

   (c) Deletion of “important” chemical shift assignments: The shifts to be deleted were chosen among all assigned ^{1}H chemical shifts, but “important” shifts were

deleted with higher probability. Importance was defined according to the number of NOEs in the reference calculation that involve a given atom. Chemical shifts were divided into eleven classes occurring in 0 – 1, 2 – 3, 4 – 5,..., and ≥ 20 peaks, with class indices $i = 0, 2, 4, \ldots, 20$. Chemical shifts from class i were deleted with relative deletion probability $p_i = 1/(21 - i)$, resulting in higher deletion probabilities for more important chemical shifts.

(d) Deletion of "unimportant" chemical shift assignments: As in 1(c), but "unimportant" ^{1}H shifts were deleted preferably. Chemical shifts from class i were deleted with relative deletion probability $p_i = 1/(i + 1)$.

2. Erroneous chemical shift assignments

A given percentage P between 0 and 40 % of randomly selected assigned ^{1}H chemical shift values were modified. Experimental NOESY peak lists were not changed.

(a) Random new chemical shift values: The selected chemical shifts were set to randomly chosen values within fifteen times the assignment tolerance for a given atom.

(b) Chemical shift permutations: Each selected chemical shift values was replaced with the chemical shift value of another atom from the set of selected atoms. Only atoms with a chemical shift value within 2.5 times the standard deviation of the corresponding chemical shift distribution from the BMRB were used for replacement.

(c) Permuted locally with other chemical shifts: As in 2(b), but only atoms from the same or directly neighboring amino acid residues were used for replacement.

3. Missing NOESY Peaks

A given percentage P between 0 and 75 % of the NOESY peaks was deleted. Chemical shift lists were not changed.

(a) Random peak deletion: The peaks to be deleted were chosen randomly.

(b) Deletion of weak peaks: The weakest peaks were (non-randomly) deleted.

4. Erroneous NOESY peaks

The positions or volumes of all NOESY peaks were distorted. Chemical shift lists were not changed.

(a) Inaccurate peak positions: Peak positions were modified in all spectral dimensions by adding a random number from a normal distribution with mean 0 and

standard deviation equal to the corresponding assignment tolerance times a varying percentage P between 0 and 100 %.

(b) Inaccurate peak volumes: Peak volumes were multiplied by a normally distributed random number with mean 1 and standard deviation P between 0 and 150 %.

5. Projection to two dimensions

   NOESY peak lists of all data sets were reduced to the two proton dimensions.

6. Increased chemical shift tolerances

   Chemical shift tolerance for NOESY peak assignment was increased from the standard value of 0.03 ppm to 0.04, 0.05, 0.06, 0.08, and 0.1 ppm for ^{1}H, and proportionally from 0.5 ppm to 0.67, 0.83, 1.0, 1.33, and 1.67 ppm for ^{15}N and ^{13}C. Chemical shift lists and NOESY peak lists were not changed.

7. Increased number of random starting structures and annealing steps

   The calculations with randomly deleted chemical shifts of modification 1(a) were repeated with 200 instead of 100 random starting structures and 20,000 instead of 10,000 torsion angle dynamics steps during the simulated annealing protocol.

Structure calculations

Automated NOESY peak assignment was performed with a chemical shift tolerance of 0.03 ppm for ^{1}H and 0.5 ppm for heavy atoms (except for modification 6, see above). Twenty independent structure calculation runs starting from different random structures were performed for each data set of each protein. Each of these structure calculations (except for modification 7, see above) started from 100 random conformers to which the standard CYANA simulated annealing protocol with 10,000 torsion angle dynamics steps was applied, and the 20 conformers with lowest target function values were chosen for the final structure bundle.

Analysis of results

For each protein, the solution NMR structure calculated from the complete data set was used as the reference structure (Fig. 4.1). The accuracy of a structure was measured by the RMSD bias (Güntert, 1998), i.e. the backbone RMSD between the average structure of a given calculation and the average structure of the reference. The average structure of a structure bundle was obtained by optimally superimposing its individual conformers for minimal backbone RMSD of the ordered regions, and calculating the average coordinates. Ordered parts of each protein were determined by the program CYRANGE (Kirchner and

Güntert, 2011) applied to the reference structure. The average RMSD bias for each type of input data modification was averaged over all 10 proteins, 5 different random modifications and 20 independent structure calculation runs leading to averaging over 1000 structure calculations.

Important as well as unimportant chemical shifts were further analyzed by classification into six different 1H classes: Hα, HN, methyl protons, aromatic ring protons, lysine and arginine side chain protons beyond Hβ, and aliphatic protons. The number of NOE cross peaks involving a given atom was determined for each atom and the average was calculated for the different classes.

In de novo structure calculations there is usually no reference structure available. It is therefore necessary to have a measure independent from the RMSD bias to assess the quality of a structure calculation result. We analyzed two previously suggested criteria, i.e. the RMSD to the mean structure (RMSD radius) of cycle 1 (convergence) and the RMSD between the structure obtained in cycle 1 and in the final structure calculation (RMSD drift). The individual criteria were then combined into a weighted average calculated as $\sqrt{((1.5R)^2 + D^2)}$, where R denotes the RMSD radius in cycle 1 and D the RMSD drift.

4.3 Results and discussion

The effect of missing, erroneous, or inaccurate structure calculation input data was investigated by random deletion and modification of chemical shifts as well as NOESY peaks. Structure calculations were performed using original and modified experimental data sets of ten different proteins (Table 4.1 and Fig. 4.1) and the average RMSD bias was used as a measure of accuracy.

The consequence of random new chemical shifts in comparison to missing NOESY peaks is illustrated in Fig. 4.2 for the protein fsh2 as an example of the two principle kinds of structure calculation failures that were discussed in the Introduction. An incomplete set of NOESY peaks generally causes less well defined structure bundles indicative of a loss of long-range information. This is reflected in the RMSD radius which increases from 1.15 Å at 30 % deleted peaks (Fig. 4.2a) to 2.08 Å at 60 % deleted peaks (Fig. 4.2b) and 10.13 Å at 75 % deleted peaks (Fig. 4.2c). This example illustrates the first category of structure calculation failure, namely the inability to ever assign enough distance restraints to converge to a well-defined structure bundle. This type of error is straightforward to detect and therefore less problematic. The results for erroneous chemical shifts show a different effect. The bundle remains rather well defined with a low RMSD radius of 0.82 Å (10 % modified chemical shifts, Fig. 4.2c), 1.04 Å (30 %, Fig. 4.2d) and 1.8 Å (40 %, Fig. 4.2e) whereas the increasing RMSD bias of 2.07 Å (10 %), 7.64 Å (30 %) and 7.1 Å (40 %) shows that the structure calculation converges to an incorrect fold at a certain

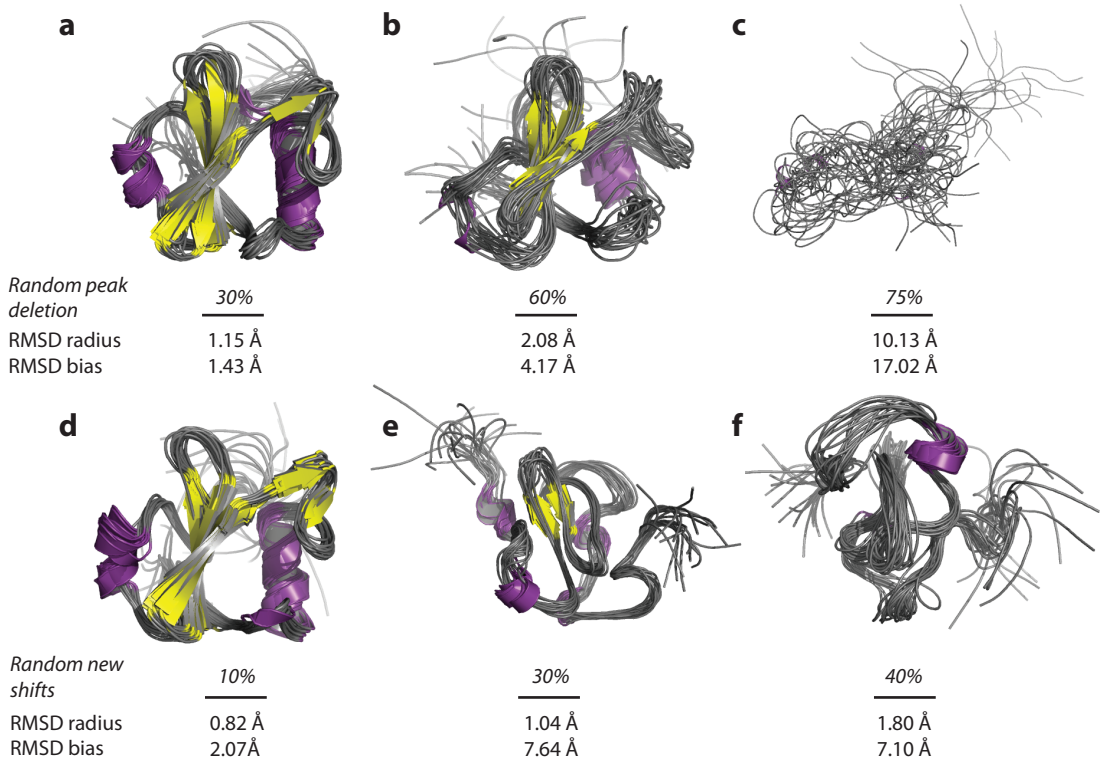


Figure 4.2: Effect of **a** 30 %, **b** 60 %, **c** 75 % missing NOESY peaks (modification 3(a) in Methods) and **d** 10 %, **e** 30 %, **f** 40 % erroneous chemical shift assignments (modification 2(a) in Methods) on the structure calculation result of the protein fsh2. Structures were calculated using the standard CYANA protocol for combined automated NOE assignment and structure calculation based on 100 random starting structures and 10,000 annealing steps. The final structure bundles comprise the 20 conformers with lowest target function values. The ordered residues 8 – 108 in the reference structure (PDB 1WQU) were used for superposition and RMSD calculation. The RMSD bias is calculated as the RMSD between the mean structure of the bundle and the mean reference structure and represents the accuracy. The RMSD radius is calculated as the average RMSD of each conformer to the mean structure of the bundle and represents the precision.

degree of erroneous shifts. This reflects the second kind of failure that can be attributed to the selection of a self-consistent, but incorrect subset of NOESY peak assignments. Due to the well-defined nature of the structure bundle, the error is more difficult to detect and hence potentially more dangerous.

For a systematic evaluation, the average RMSD bias was plotted against the percentage P of modified input data for the different types of modifications (Figs. 4.3 – 4.6). The dotted line indicates an RMSD value of 3 Å representing the threshold below which the global fold of the structure is still assumed to be correct. The results for each individual protein can be found in Fig. 4.4 and in Appendix A of the present work (Figs. A.1 – A.10).

The overall effect of chemical shift deletions is presented in Fig. 4.3a-d. Chemical shifts were deleted in four different ways: random deletion from the set of all shifts (Fig. 4.3a), random deletion only from side chain atoms (Fig. 4.3b), random deletion of “important”

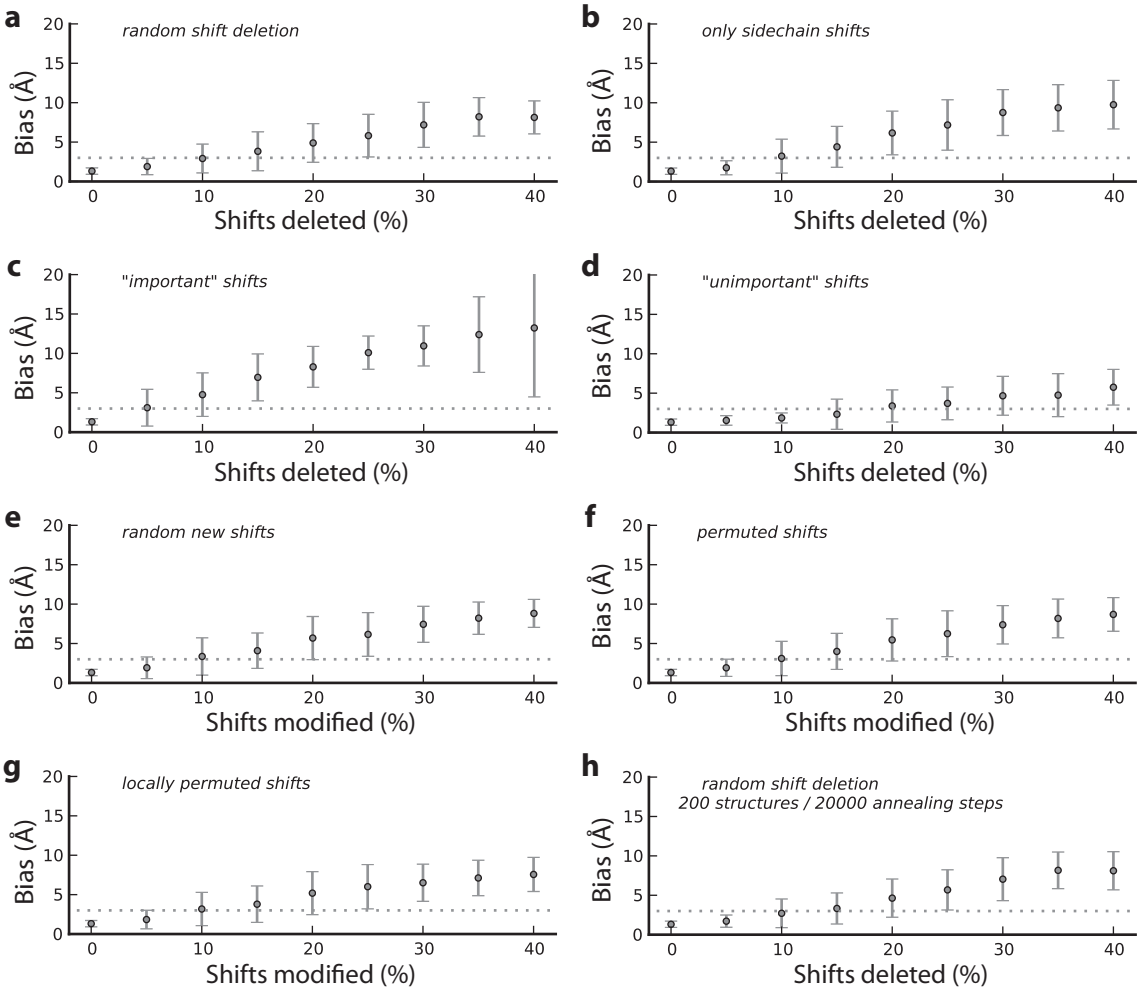


Figure 4.3: RMSD to the reference structure for different types of simulated chemical shift imperfections. For each data point, twenty independent automated NOESY assignment and structure calculation runs were performed for each of five randomly modified data sets of the ten proteins of Table 4.1. The average RMSD to the reference structure is plotted against the percentage P of modified chemical shifts. See Methods for details. The data point at 0 % shift modification denotes the RMSD for 20 runs with the complete, unmodified experimental data. **a** Random deletion of chemical shift assignments, **b** random deletion of side-chain chemical shift assignments, **c** random deletion of "important" chemical shift assignments, **d** random deletion of "unimportant" chemical shift assignments, **e** random new chemical shift values, **f** random permutation of chemical shift values, **g** local permutation of chemical shift values, **h** doubled number of random starting structures and annealing steps for randomly deleted chemical shift assignments.

shifts (Fig. 4.3c) and random deletion of "unimportant" shifts (Fig. 4.3d). Omission rates were varied between 0 and 40 % in steps of 5 %. In all four cases the average RMSD bias increases at increasing omission rates P. In most cases, random deletion of 5 % of the chemical shifts results in structures with an RMSD bias below 3 Å, whereas 10 and 15 % missing chemical shifts raise the average RMSD bias slightly above 3 Å (Fig. 4.3a). Omission rates of more than 15 % increase the average RMSD including the standard deviation considerably above 3 Å indicating that structure calculations reproducibly fail

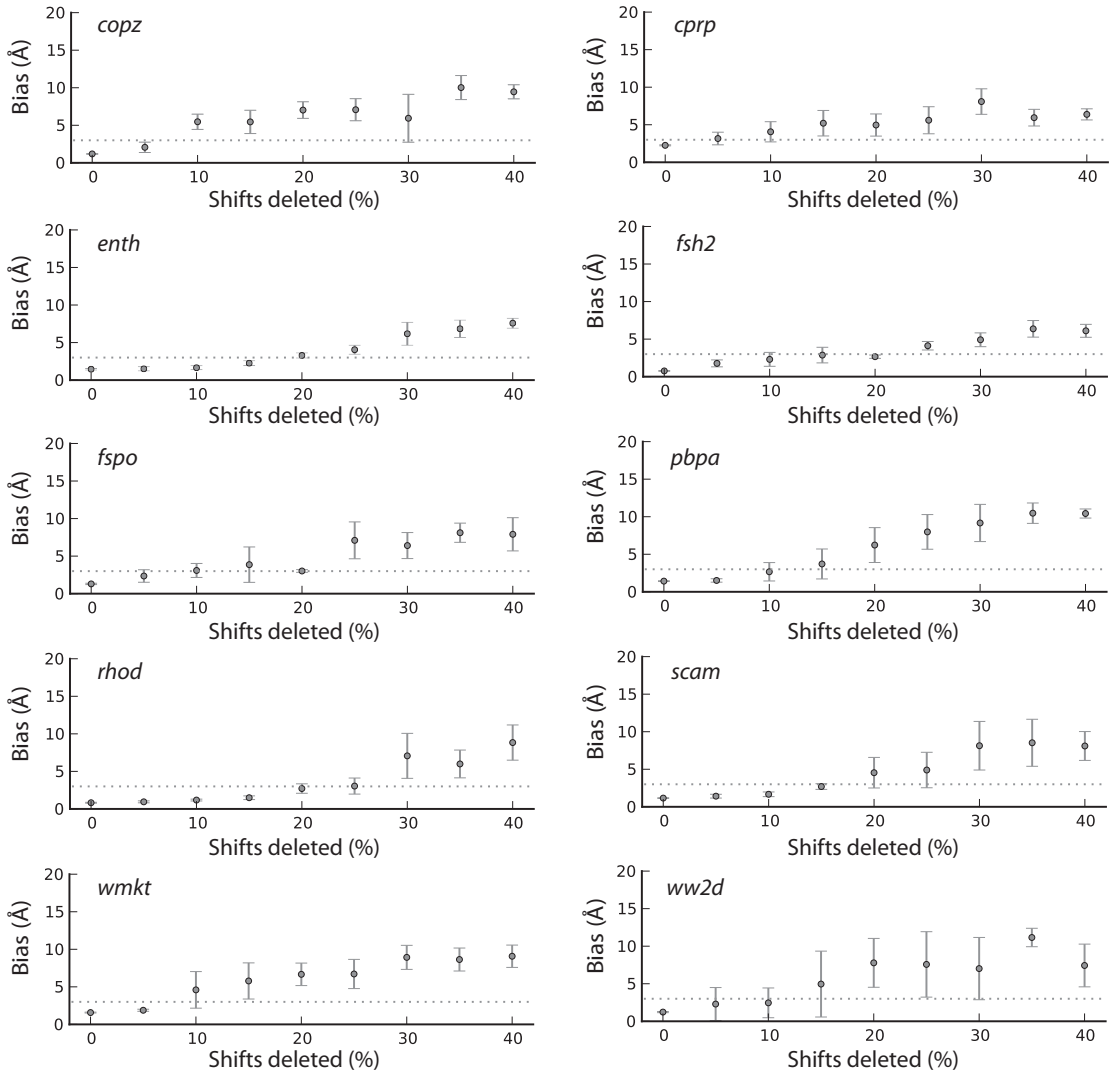


Figure 4.4: RMSD to the reference structure for different percentages of randomly deleted chemical shifts (modification 1(a) in Methods). Results are presented separately for each of the ten proteins of Table 4.1 and Fig. 4.1. For each data point, twenty independent automated NOESY assignment and structure calculation runs were performed for each of five randomly modified data sets. The RMSD bias from the reference structure is plotted against the percentage P of modified chemical shifts. The data point at 0 % shift deletion denotes the RMSD for 20 runs with the complete, unmodified experimental data.

to converge to the correct global fold when using severely incomplete chemical shift data. The outcome in the range between 10 and 15 % chemical shift omission strongly depends on the protein and the quality of the respective NOESY data, which becomes apparent when comparing the plots for the individual proteins presented in Fig. 4.4 and in the Supplementary Material. In favorable cases, the correct structure can still be found with 20 % chemical shifts missing, whereas rather unfavorable cases may fail at 5 % missing

chemical shifts. TALOS angle restraints can in some cases slightly improve the structure calculation result.

It does not make any significant difference whether random chemical shifts or only side-chain shifts are missing (Fig. 4.3a and b). Deletion of "important" shifts causes a steeper increase in the average RMSD bias compared to random deletion, whereas the slope is less steep in the case of "unimportant" shifts (Fig. 4.3c and d). This shows that it can make a difference for the structure calculation results which particular chemical shifts are missing. It is in practice more likely that "unimportant" shifts are missing, as they are typically more difficult to assign.

To further investigate the importance of individual types of protons, chemical shifts from all data sets were classified into six different classes: Hα, NH, methyl protons, aromatic protons, lysine and arginine side chain protons, and aliphatic protons. Importance is measured based on the amount of medium- and long-range NOESY peaks that involve the respective chemical shift (Fig. 4.5). Protons from methyl groups appear on average in 17.5 medium- and long-range NOE peaks, aromatic protons appear on average in 13.5 peaks, NH protons in 11.9 peaks, Hα protons in 10.3 peaks, aliphatic protons in 10.2 peaks and Lys/Arg sidechain protons in 9.0 peaks. Fig. 4.5 suggests that methyls and aromatic protons are very important, which can be attributed to their preferential occurrence in the hydrophobic, densely packed core of the protein enabling a large amount of NOE contacts.

Fig. 4.3e–g shows the effect of modified chemical shift values. Different simulated sources of errors such as random new chemical shift values (Fig. 4.3e), randomly permuted chemical shift values (Fig. 4.3f), and locally permuted chemical shift values (Fig. 4.3g) result in very similar average RMSD values as random missing chemical shifts. Even local permutations show the same result.

Compared to missing chemical shifts, deletion of NOESY peaks shows a less steep increase of average RMSD (Fig. 4.6a). On average, the RMSD bias at 30 % deleted NOESY peaks is below 3 Å while the average RMSD rises slightly above 3 Å at 45 %. The much less pronounced increase can be explained by the fact that NOESY peaks firstly contain a large amount of signals that contain no or very limited structural information due to their sequential nature and secondly contain rather redundant information through the dense NOE network. In contrast, one missing chemical shift leads to a whole set of NOESY peaks that remain unassigned in the more favorable case or get assigned incorrectly in the less favorable case. Fig. 4.6b shows the result for deletion of weak peaks. The RMSD bias at 30 % deletion is comparable to random deletion, whereas deletion of 45 % of the weakest peaks results in a significant increase of 7 Å compared to 3 Å at 45 % randomly deleted peaks. A higher average RMSD for deletion of weak peaks is expected as they contain important long-range information.

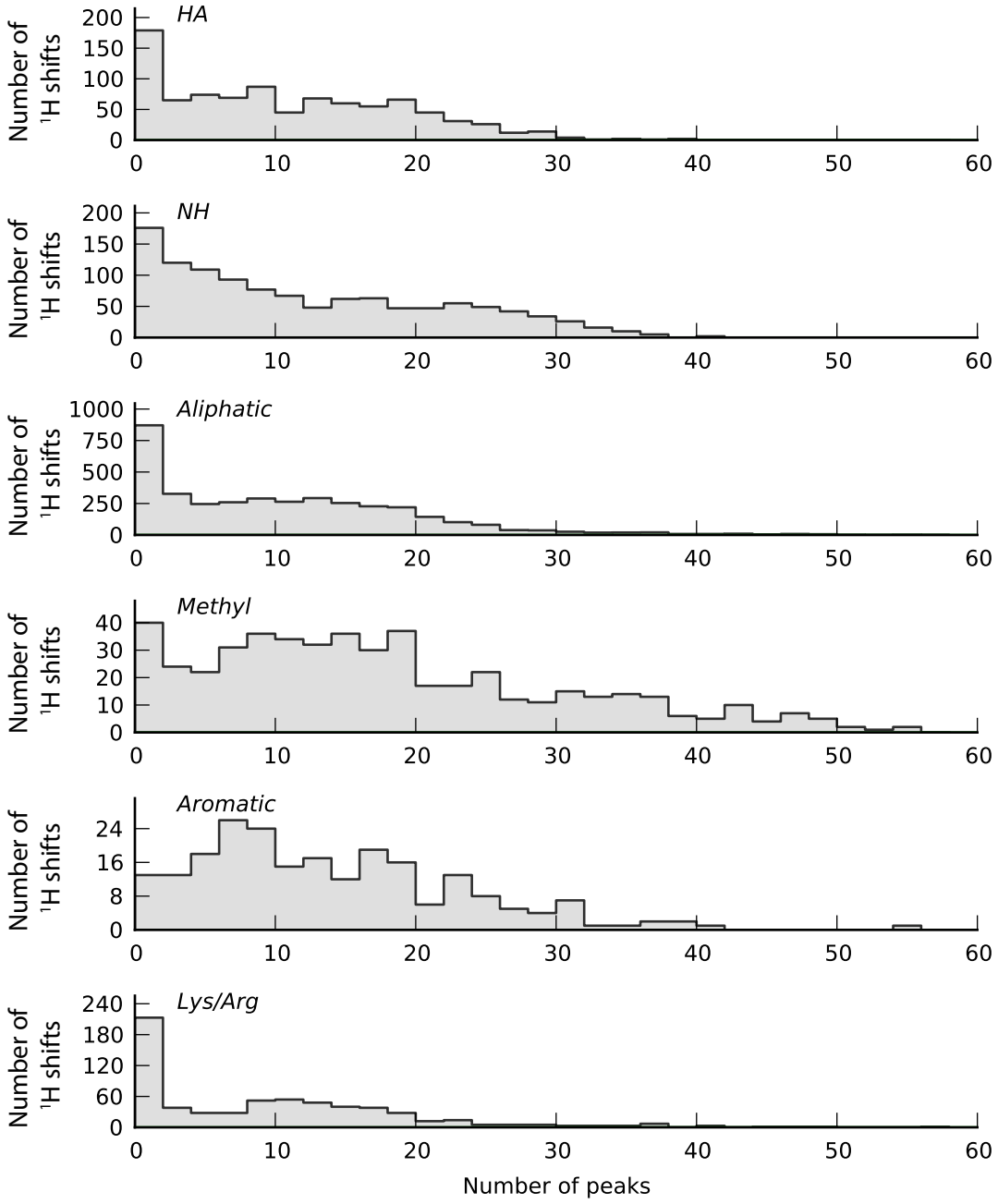


Figure 4.5: Number of ^{1}H resonances that are involved in a given number of medium- and long-range NOESY peaks. Proton chemical shifts were separated into six disjoint classes: Hα, NH, aliphatic, methyl, aromatic, and lysine and arginine sidechain atoms. Chemical shifts were taken from the data sets of ten different proteins (Table 4.1). Peaks were counted in the final assigned peak lists of the combined automatic NOE assignment and structure calculation run that yielded the reference structure.

Using the complete peak lists, but introducing errors in peak positions yields an average RMSD bias of 3 Å at 45 % error and of more than 5 Å at 60 % error (Fig. 4.6c). In contrast to errors in peak positions, errors in peak volumes have largely no effect on the average

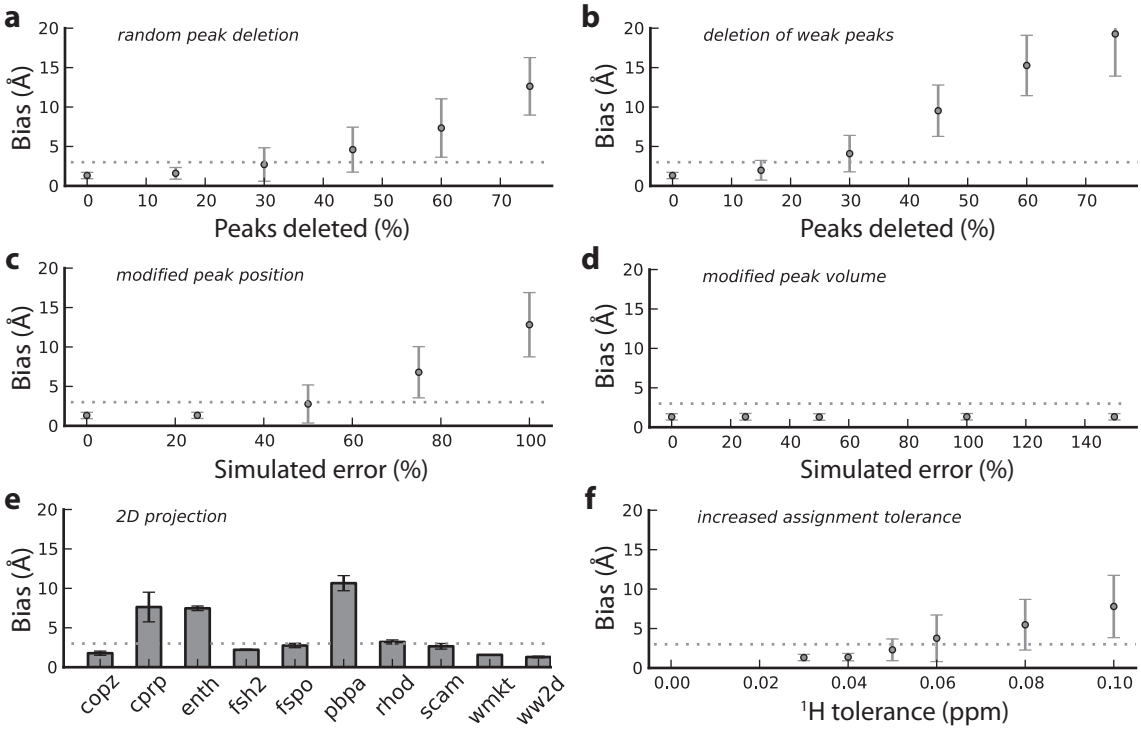


Figure 4.6: RMSD to the reference structure for different types of simulated peak list imperfections. For each data point, twenty independent automated NOESY assignment and structure calculation runs were performed for each of five randomly modified data sets of the ten proteins of Table 4.1. The average RMSD to the reference structure is plotted against the percentage P of modified data, where applicable. See Methods for details. The data point at 0 % peak modification denotes the RMSD for 20 runs with the complete, unmodified experimental data. **a** Random deletion of NOESY peaks, **b** random deletion of the weakest NOESY peaks, **c** erroneous peak positions, **d** erroneous peak volumes, **e** 2D projection of NOESY peaks, **f** increased assignment tolerances.

RMSD for the complete range tested up to 150 % error (Fig. 4.6d). A larger influence from erroneous peak positions can be explained by the fact that the number of incorrect assignments increases, creating potentially distorting restraints, whereas erroneous peak volumes only affect the upper distance limit value. This erroneous effect on the upper distance limit value is furthermore greatly reduced by the r-6-correlation between peak volume and calibrated distance. Using only two-dimensional peak lists has almost no effect on the structure calculation result in the case of three proteins (copz, ww2d and wmkt). This result can be explained by the fact that a significant part of the peaks of the original data set comes from 2D NOESY spectra. Reducing the remaining peaks to two dimensions has a less severe effect in these cases compared to other data sets, which contain mainly 3D data. For fsh2, fspo, rhod and scam the RMSD bias shows a slight increase but remains below 3 Å, and for cprp, enth and pbpa the RMSD bias increases above 5 Å (Fig. 4.6e).

Fig. 4.6f shows the effect of increased chemical shift tolerances, which simulates spectra with less resolution resulting in higher assignment ambiguities. Chemical shift tolerances for NOESY peak assignments were raised up to 3.33 times their original value, which corresponds to 0.1 ppm for 1H and 1.66 ppm for 15N and 13C. Up to 200 % increased tolerance, the average RMSD bias is still around 3 Å, whereas further increase results in RMSD bias values of around 5 Å. Increased chemical shift tolerances have very diverse consequences on the different data sets (Figs. A.1 – A.10). The effect is most severe in cases where the data sets contain a large amount of two-dimensional data (copz, ww2d and wmkt) as well as in the case of the data set of cprp. Two-dimensional data are especially sensible to reduced resolution as the amount of assignment possibilities is much higher. It should, however, be noted that these two simulations (reduction to two spectral dimensions and increased chemical shift tolerance) do not perfectly represent the situation of NMR spectra with poor resolution. In severely overlapped spectra, several peaks may be fused into one single peak with a biased peak position. In our simulation, all peaks are still considered individually at the correct peak position. This increases the probability of the correct assignment to be chosen despite the availability of a large number of additional assignment possibilities.

Finally, we tested whether the effect of missing data can be compensated by performing more annealing steps during structure calculation and using more random starting structures. For this purpose, we repeated all calculations with randomly deleted chemical shifts with 200 instead of 100 random starting structures and with 20,000 instead of 10,000 annealing steps. The calculation results show only marginal overall improvement (Fig. 4.3h), indicating that data imperfections can in general not be compensated by longer computation times. The only exception is the homodimeric protein ww2d, for which longer simulated annealing yielded significantly lower RMSD bias values for the data sets with 5 – 15 % deleted chemical shifts.

These results show that data imperfections of various natures can dramatically reduce the quality of NMR structures. In case of de novo structure determination with lack of a reference structure, it is important to be able to evaluate the structure calculation result based on a measure independent of the RMSD bias. Several criteria have been suggested previously. Two of these criteria are the convergence (RMSD to the mean structure) of the initial structure calculation cycle and the RMSD drift (RMSD between the first and the last cycle). If the initial cycle converges to an RMSD radius below 3 Å and the RMSD drift is simultaneously below 2 Å, the result is considered reliable (Herrmann et al., 2002a; Jee and Güntert, 2003). We have investigated these criteria using all aforementioned structure calculations and summarized the results in Fig. 4.7.

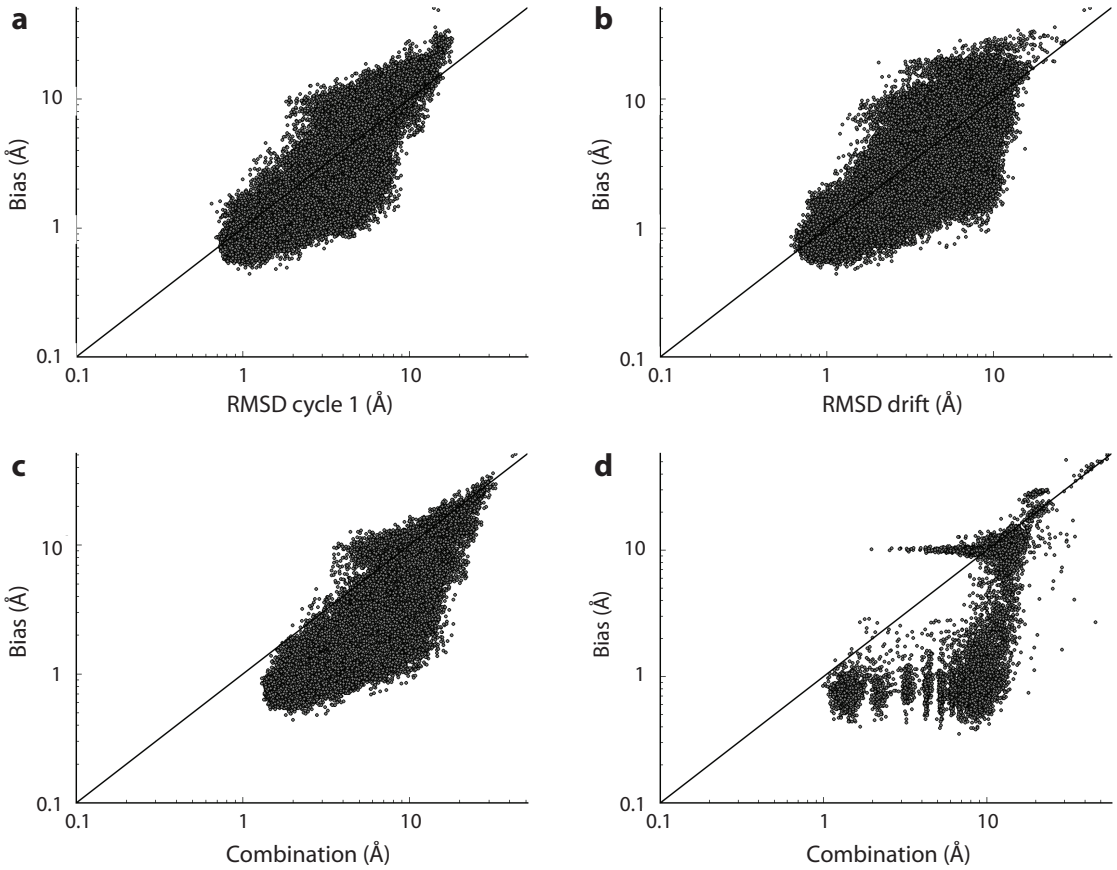


Figure 4.7: Structural accuracy plotted against commonly applied evaluation criteria for combined automated NOESY assignment and structure calculation runs. The accuracy is represented by the RMSD bias, i.e. the RMSD between the mean structure of the bundle and the mean reference structure. Every data point represents one combined automated NOESY assignment and structure calculation run. **a** Initial convergence measured by the RMSD to the mean structure of the structure bundle from the first structure calculation cycle, **b** RMSD drift measured by the RMSD between the final structure bundle and the structure bundle of the first cycle, **c** a combination of the two criteria calculated as $\sqrt{((1.5R)^2 + D^2)}$, where R denotes the RMSD radius in cycle 1 and D the RMSD drift, for all proteins except the homodimeric ww2d, and **d** same as in c for the structure calculations of the protein ww2d.

Fig. 4.7a and b show the accuracy plotted against the RMSD in cycle 1 and the RMSD drift. Especially dangerous are false positives, i.e. cases, where the evaluation parameters meet the required criteria (convergence < 3.0 Å, drift < 2.0 Å) but the structure is misfolded. Considering both criteria individually, the number of false positives is 2 % (convergence) and 0.4 % (drift), respectively. Calculation of a weighted average from both values (Fig. 4.7c) further reduces the number of false positives to 0.01 %. The correlation of the weighted average and the accuracy shows a significantly reduced number of data points above the diagonal (accuracy exceeding the criterion) which therefore allows it to be used as an upper limit on the accuracy. The distribution for the homodimeric protein ww2d is presented separately in Fig. 4.7d. In contrast to the monomeric proteins, it shows

multiple clusters that are presumably due to different ways of dimer formation. On the one hand, there are a large number of cases of structures with a high accuracy around 1 Å for which the combined criterion varies over a large range of 1 – 10 Å. On the other hand, there is a narrow cluster of structures with an RMSD bias of about 10 Å and values of 2 – 10 Å for the combined criterion.

In order to investigate the influence of artifacts such as water signals or baseline distortions on the structure calculation result, we have recalculated the structures of the three proteins enth, fsh2, and rhod based on peaks lists from automatic peak picking without subsequent refinement. Results are summarized in Table 4.2. Only slight differences between the refined and unrefined sets of peak lists can be observed in the case of enth with respect to the RMSD bias, the final CYANA target function, as well as the aforementioned evaluation parameters (convergence cycle1, RMSD drift and the combination thereof). This is in good agreement with the results obtained from the modified data sets, where enth is one of the rather stable structure calculations which yields an accurate structure bundle even at 15 % missing chemical shifts (Fig. 4.4). In the two other cases, the structural quality significantly drops when compared to the results obtained from refined peak lists (~three-fold), however, the RMSD bias is still <3.0 Å and the global fold is thus considered correct. In all three cases, the final CYANA target function increases and the RMSD radius decreases when using unrefined peak lists. This can be attributed to an increased number of potentially incorrect long- range restraints that result from artifact peaks. The combined criterion gives a good indication about the structural quality.

Table 4.2: Structure calculation results using refined and unrefined NOESY spectra.

	enth		fsh2		rhod	
	refined	unrefined	refined	unrefined	refined	unrefined
RMSD bias (Å)	0.90	1.00	0.67	2.11	0.40	1.09
RMSD radius (Å)	0.53	0.25	0.52	0.22	0.76	0.26
CYANA target function ($Å^2$)	1.18	3.95	1.84	15.99	0.95	6.17
Convergence cycle 1 (Å)	1.21	0.75	1.10	1.60	0.89	1.48
RMSD drift (Å)	1.01	0.96	1.10	2.46	0.82	2.26
Combined criterion	2.08	1.48	1.99	3.44	1.57	3.18

4.4 Conclusion

The results presented in this study clearly show that imperfections within the chemical shift assignment can cause severe problems during NOE assignment and structure calculation. In most of the data sets tested 10 % of missing or erroneous chemical shifts result

in inaccurate structures with RMSD bias values above 3 Å. In some cases of high quality data and large amounts of 3D peaks, higher percentages of missing or erroneous chemical shifts can be tolerated. Less severe problems arise from missing peaks, errors in peak positions and volumes as well as lower resolution simulated by using higher assignment tolerances. Furthermore, it was shown that data imperfections cannot be overcome by longer computation times. The convergence of the initial structure calculation cycle and the RMSD drift between the first and the last cycle can be combined in a weighted average and used as an indication for the reliability of a structure calculation result.

Chapter 5

Peakmatch - A simple and robust method for peaklist matching

This chapter is based on the following publication:

Buchner L., Schmidt E., and Güntert P. Peakmatch – A simple and robust method for peaklist matching. *J Biomol NMR* 55(3):267-277, 2013

5.1 Introduction

Protein structure determination by NMR spectroscopy has been accelerated by the development of programs that perform some or all of the necessary steps automatically (Baran et al., 2004; Guerry and Herrmann, 2011; Güntert, 2009; López-Méndez and Güntert, 2006; Williamson and Craven, 2009). The majority of these programs use the information from the NMR spectra in the form of peak lists rather than by accessing the spectra directly. For most applications a set of peak lists from different types of experiments is needed. It is important to have a consistently referenced data set for the resonance assignment, and automated NOE assignment and structure calculation require that the NOESY peak lists and the corresponding chemical shift list(s) are in optimal agreement. Several programs exist for correcting the referencing of chemical shifts or optimizing the agreement between chemical shift assignments and general chemical shift statistics (Aeschbacher et al., 2012; Ginzinger et al., 2007; Wang and Wishart, 2005; Wang et al., 2005). Methods are also available to adapt chemical shifts to NOESY spectra (Herrmann et al., 2002a). However, to the best of our knowledge there is no program available that optimizes automatically the mutual referencing of several unassigned, multidimensional peak lists to achieve a consistently referenced data set prior to automated assignment or structure calculation.

5.2 Methods

The new *Peakmatch* algorithm implemented in the CYANA software package (Güntert, 2009; Güntert and Buchner, 2015) calculates the optimal chemical shift referencing offsets between two peak lists by maximizing a match score using either a grid search or downhill simplex method.

One peak list is used as a reference and remains unchanged whereas each corresponding dimension in the second, target peak list is shifted by a constant offset. The offsets that yield the maximal match score represent the calculation result. An overview of the algorithm is given in Fig. 5.1.

5.2.1 Determination of corresponding dimensions

The user specifies the dimensions in the reference peak list. The algorithm can then determine the corresponding dimensions in the target peak list automatically based on the expected peak match. If more than one possibility is found, the one with the largest expected peak match is chosen.

To calculate the expected peak match score, the program generates expected peaks for the reference and target peak lists based on experiment type-specific connectivity patterns stored in the CYANA library and the covalent structure of the protein (Bartels et al., 1997;

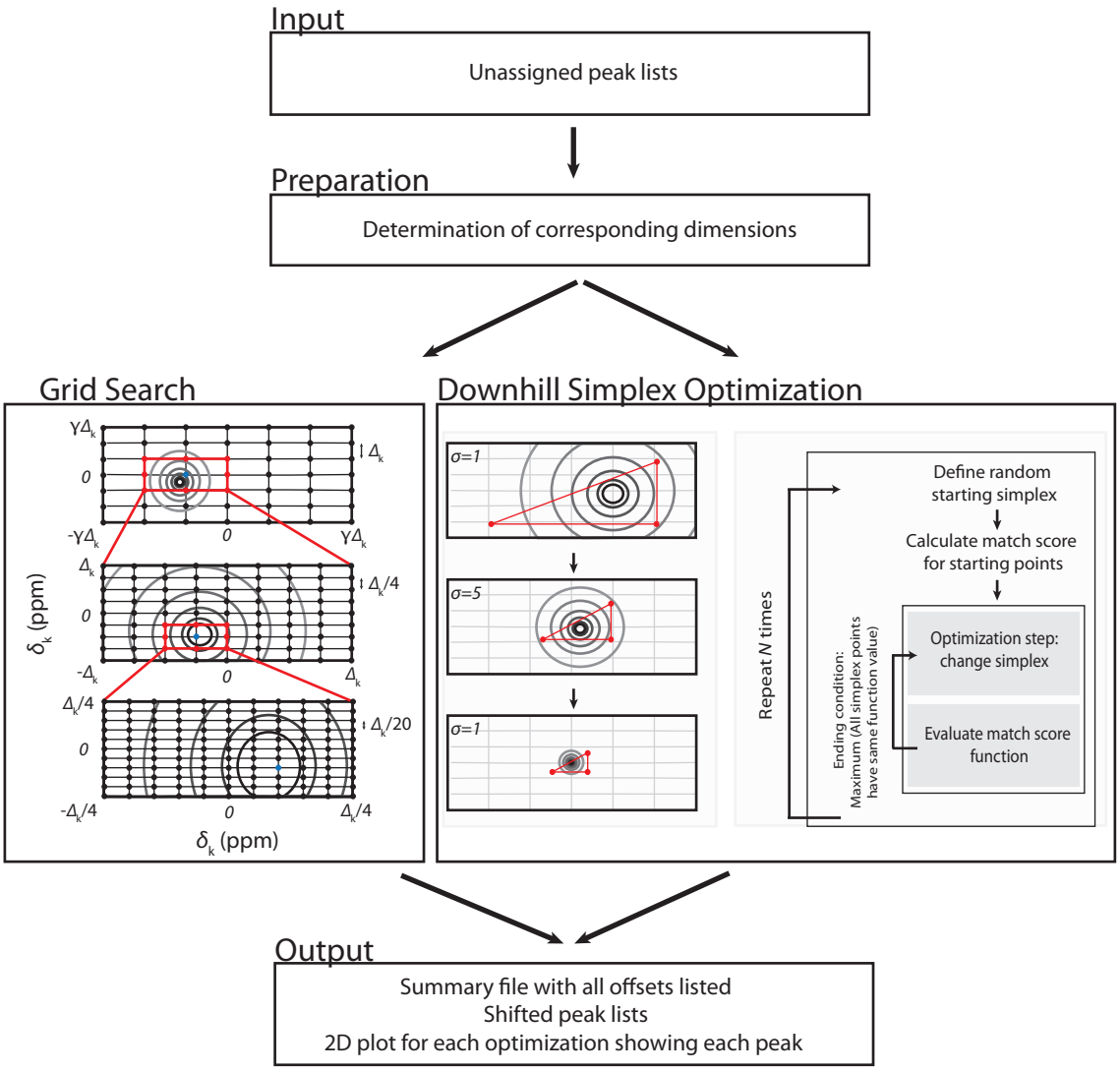


Figure 5.1: Flowchart of the *Peakmatch* algorithm.

Schmidt and Güntert, 2012; Schmucki et al., 2009). Through-space type experiments are approximated by the subset of short-range peaks, which is accurate enough for the present purpose. Details of the generation of expected peaks have been given elsewhere (Schmidt and Güntert, 2012). The expected peak match score is calculated using Equation 5.1,

$$E = \sum_{i=1}^{n_0} \sum_{j=1}^{m_0} \theta_{ij} \tag{5.1}$$

where n_0 and m_0 denote the number of expected peaks for the reference and target experiment, respectively, and θ_{ij} is one if the two expected peaks have the same assignment in all dimensions considered for match calculation, and zero otherwise. The expected peak

match is independent from the experimental peak lists and thus not influenced by the lack of peaks or the presence of artifacts.

5.2.2 Match score

For a reference peak list with $i = 1, ..., n$ peaks at positions $\omega_{ik}^{(1)}$ and a target peak list with $j = 1, ..., m$ peaks at positions $\omega_{jk}^{(2)}$ in the $k = 1, ..., d$ corresponding dimensions, we define the match score S as a function of the offsets $\delta_1, ..., \delta_d$:

$$S(\delta_1, ..., \delta_d) = \frac{1}{E} \times \sum_{i=1}^{n} \sum_{j=1}^{m} exp\left(- \prod_{k=1}^{d} \left(\frac{\omega_{ik}^{(1)} - \omega_{jk}^{(2)} + \delta_k}{\sigma \Delta_k} \right)^2 \right) \tag{5.2}$$

The contribution of an individual peak pair to the match score is given by a Gaussian function of the normalized distance between the two peaks. The chemical shift tolerances Δ_k represent the accuracy of the peak positions. They should be set by the user such that the positions of any two peaks assigned to the same atom differ by less than the chemical shift tolerance in the given dimension. The default values are 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N dimensions. The dimensionless scaling factor σ determines the significance of a deviation. By default, $\sigma = 1$. Deviations of peak positions smaller than $\sigma\Delta_k$ yield score contributions close to one, whereas those from deviations much larger than $\sigma\Delta_k$ are negligible. The overall match S between two peak lists is calculated as a sum over all n peaks in the reference peak list. For each reference peak i, the $q = E/n_0$ largest contributions from peaks in the target peak list are included in the match score calculation, as indicated by the prime in Equation 5.2. The parameter q represents the expected average number of peaks in the target peak list with the same assignment as a given reference peak in the corresponding dimensions. This results in larger q values when optimizing for instance ^{15}N-resolved [^{1}H,^{1}H]-NOESY against [^{15}N,^{1}H]-HSQC ($q \approx 13$) compared to HNCA against [^{15}N,^{1}H]-HSQC (q=2), or two peak lists from the same experiment type (q=1). Assignments for the input peak lists are not required.

The match score function of Equation 5.2 approximately counts the number of peaks in the two peak lists whose position matches within the ranges $\sigma\Delta_k$, and does so with minimal influence from other, non-matching peaks. The match score S is normalized by the expected peak match E of Equation 5.1. It thus has the value 1 for two ideally matched peak lists that contain exactly the expected peaks. In general, the match score shows one narrow optimum when using two or more corresponding dimensions (Fig. 5.2a) and gets broader as well as smoother with increasing σ (Figs. 5.2b and 5.2c).

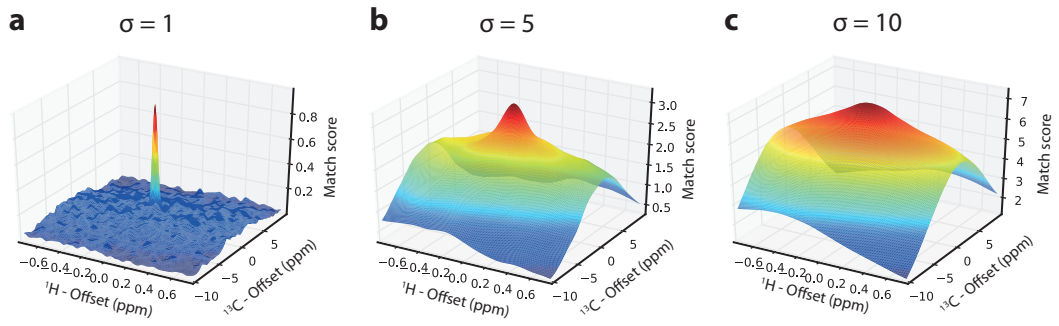


Figure 5.2: Match score function for two corresponding dimensions and different σ values (see Equation 5.2). The [^{15}N,^{1}H]-HSQC reference peak list and the CBCACONH target peak list are from the manually edited ENTH data set. **a** $\sigma = 1$, **b** $\sigma = 5$, **c** $\sigma = 10$

5.2.3 Optimization procedures

Grid search

The grid search evaluates the match score function at every point of a grid and takes as result the offset values $\delta_1, ..., \delta_d$ that yield the maximum score value. With an appropriate grid size and spacing, this procedure guarantees the identification of the global maximum, which is in general the correct offset. In order to save computation time, the grid search procedure performs several steps at different grid sizes and spacings (Fig. 5.1, left side). The first grid covers the largest offset range, which should be chosen larger than the expected offset to ensure that the region of the global maximum is found. To this end, the user specifies a dimensionless parameter γ to define a rectangular grid of size $[-\gamma\Delta_k, \gamma\Delta_k]$ and spacing Δ_k in the corresponding dimensions $k = 1, ..., d$. Two subsequent grid searches are performed using smaller grids of sizes $[-\Delta_k, \Delta_k]$ and $[-\Delta_k/4, \Delta_k/4]$ with smaller spacings of $\Delta_k/4$ and $\Delta_k/20$, respectively, centered at the optimum found in the preceding search. This procedure allows finding the correct offset at high precision without having to search a large grid with very small spacing between the grid points. Nevertheless, depending on the size of the initial grid, calculation times can be significant.

Downhill simplex optimization algorithm

To further reduce the computation time, a downhill simplex minimization algorithm (Nelder and Mead, 1965) can be used to find the optimal offsets $\delta_1, ..., \delta_d$ between two peak lists. This algorithm makes use of a simplex of $d + 1$ points in d dimensions, e.g. a triangle in 2 dimensions, or a tetrahedron in 3 dimensions, that should be initialized such that it encloses the optimum. For two corresponding dimensions the algorithm uses triangular start simplexes with the vertices $(c\Delta_1, c\Delta_2)$, $(-c\Delta_1, -c\Delta_2)$, and $(-c\Delta_1, c\Delta_2)$, where Δ_k represents the chemical shift tolerance for dimension k, and c is a random num-

ber from a normal distribution with zero mean and user-defined standard deviation $\sigma^{(s)}$. Analogous choices are made for start simplexes in more than two corresponding dimensions. The program performs a specified number of optimization runs with different start simplexes of randomly varying size. The same random number is used for all vertices, i.e. the start simplexes vary only in size but not in shape or position. Beginning with the start simplex, the algorithm then performs a number of optimization steps that move vertices of the simplex to a new position. The optimization ends as soon as one of the two conditions (i) all vertices have the same function value within a specified tolerance or (ii) a maximum of 10000 optimization steps was performed, occurs.

The downhill simplex optimization procedure requires a general slope towards the maximum of the function in order to reach the optimum. The match score function of Equation 5.2 has in general a very narrow maximum when optimizing two or more corresponding dimensions and choosing the default scaling factor $\sigma = 1$. To increase the probability for reaching the global maximum, the optimization is divided into three steps. The first step is performed with $\sigma = 10$ and $\sigma^{(s)} = 40$. The smoothed match score results in a high percentage of runs that reach the global optimum and large start simplexes increase the range of potential offsets which are covered. However, the optimum may be slightly shifted at high σ values. Therefore, two further local optimization runs with smaller σ values are added to determine the offsets with high precision. The second optimization with $\sigma = 5$ and $\sigma^{(s)} = 10$ is started from the optimum found in the first optimization. The final optimization is performed with $\sigma^{(s)} = 1$. The same number of runs with different random start simplexes is applied in the three optimization steps.

5.2.4 Algorithm input and output

The input to the *Peakmatch* algorithm consists of a reference peak list and one or more target peak lists in the format of the program XEASY (Bartels et al., 1995), the general CYANA library with the magnetization transfer pathway definitions for the corresponding NMR spectra (Schmidt and Güntert, 2012; Schmucki et al., 2009), and the protein sequence. Parameters that can be set by the user include the chemical shift tolerances Δ_k (default 0.03 ppm for H and 0.4 ppm for ^{13}C and ^{15}N dimensions), the scaling factor σ for peak matching (default $\sigma = 1$), the choice of optimization strategy (default: downhill simplex), the initial grid size parameter γ for the grid search (default $\gamma = 3.0$), the standard deviation $\sigma^{(s)}$ for generating start simplexes (default $\sigma^{(s)} = 40$), and the dimensions of the reference peak list for which corresponding dimensions in the target peak list(s) should be searched. In general, the default values can be used.

The algorithm outputs a summary table with the calculated optimal offsets, the initial match score and the match score after optimization for each pair of peak lists (Fig. 5.3a), plots overlaying the peaks from the reference and target peak list in the corresponding dimensions (Fig. 5.3b), and the shifted target peak lists.

a

Reference: N15H1.peaks dimensions: H+N

Peaklist	Dimensions	Offset	Match (ini)	Score (ini)	Match (opt)	Score (opt)
HNCA	N+HN	-4.995,-0.499	44.2	0.16	269.8	0.98
CBCANH	N+HN	-5.009,-0.500	72.0	0.14	440.2	0.83
C_CO_NH	N+HN	-4.997,-0.499	57.1	0.13	347.3	0.77
N15NOESY	N+HN	-4.997,-0.499	126.1	0.06	1524.9	0.79

b

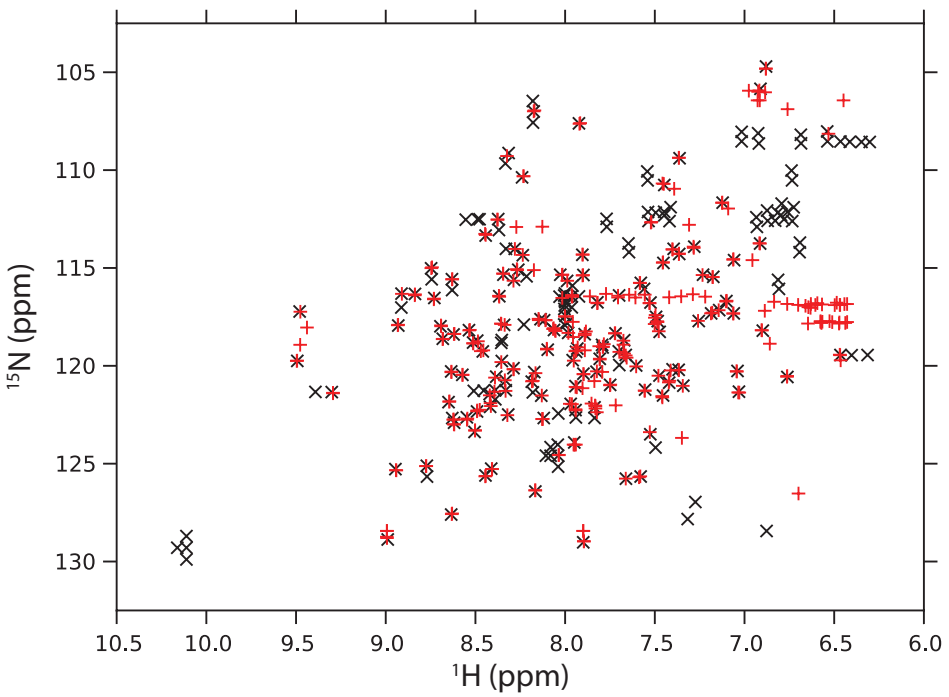


Figure 5.3: Example output from the *Peakmatch* algorithm. **a** Summary file of the application of the *Peakmatch* algorithm to automatically generated peak lists for the protein ENTH with artificially introduced offsets of 5 ppm for heavy atoms and 0.5 ppm for protons. The reference peak list and the specified dimensions are mentioned in the first line of the output file. The match result is listed for each target peak list in a separate line, which includes the corresponding dimensions, the offset for each dimension, and the absolute Match as well as the normalized Score prior to (ini) and after matching (opt). **b** Example plot of the optimized HNCA target peak list, projected on the HN dimensions (red, +), and the corresponding [^{15}N,^{1}H]-HSQC reference peak list (black, x).

5.2.5 Test data sets

The algorithm was evaluated with experimental data sets of five different proteins, i.e. the 140-residue ENTH-VHS domain At3g16270(9 – 135) from *Arabidopsis thaliana* (ENTH) (López-Méndez et al., 2004), the 134-residue rhodanese homology domain At4g01050(175 – 295) from *Arabidopsis thaliana* (RHO) (Pantoja-Uceda et al., 2004; Pantoja-Uceda et al., 2005), the 114-residue Src homology domain 2 from the human feline sarcoma oncogene Fes (SH2) (Scott et al., 2004; Scott et al., 2005), ubiquitin (Ikeya et al., 2009), and the DsbA. Stereo-array isotope labeling (SAIL) was used for ubiquitin and DsbA (Kainosho et al., 2006; Kainosho and Güntert, 2009). Each data set includes typical backbone experiments for resonance assignment as well as through-space experiments, i.e. [^{15}N,^{1}H]-HSQC, [^{13}C,^{1}H]-HSQC, HNCO, HN(CA)CO, CBCANH, CBCA(CO)NH, HCCH-COSY, HCCH-TOCSY (in the case of RHO only for the aromatic region), (H)CCH-TOCSY (only for DsbA and ENTH), H(CCCO)NH, ^{15}N-resolved NOESY, and ^{13}C-resolved NOESY spectra. The peak lists of all five data sets were generated automatically using automatic peak-picking algorithms of the programs NMRView (Johnson, 2004) and AZARA (http://www.ccpn.ac.uk/azara) without manual corrections (Ikeya et al., 2009; López-Méndez and Güntert, 2006). In addition, peak lists for ENTH, RHO, and SH2 were also available from manual, or manually curated peak picking.

5.3 Results and discussion

To evaluate the performance of the algorithm, we artificially introduced different offsets into the target peak lists and back-calculated the offset using the *Peakmatch* algorithm under various conditions.

The *Peakmatch* score can be calculated for any number of corresponding dimensions in two peak lists. Almost every pair of peak lists contains one corresponding dimension. Two corresponding dimensions occur mostly for HSQC planes, and three corresponding dimensions only for few peak list pairs. The standard application of the *Peakmatch* algorithm is to optimize the offsets for two corresponding dimensions, and this will be the focus of the presentation. Offset optimization for one and three corresponding dimensions will be discussed in separate sections.

5.3.1 Determination of corresponding dimensions

Corresponding dimensions among two peak lists are determined automatically prior to peak list matching. During this procedure expected peaks are generated for both experiment types and the expected peak match E (Equation 5.1) is used to evaluate different solutions. For many types of peak list pairs, such as typical backbone experiments being

matched to the respective HSQC spectrum, there is only one solution with an expected peak match larger than zero. However, for example NOESY spectra usually have more than one solution. As default, the solution with largest expected peak match is chosen, however, it is also possible to optimize all solutions in independent runs. Table 5.1 shows a summary of expected peak match values for all NOESY spectra. In all cases the expected peak match has a significantly higher value when using the HSQC-plane for matching compared to the indirect plane, which means that in general the HSQC-plane is the first choice for optimization.

Table 5.1: Expected peak match values for NOESY peak lists using the respective HSQC peak list as reference.

Peak list	Dimensions	Rho	Enth	SH2	Ubiquitin	Dsba
^{15}N resolved NOESY	N+HN[a]	1617	1941	1466	871	1495
	N+H[b]	211	270	195	150	181
^{13}C resolved NOESY	C+HC	4242	4666	3711	1293	10215
	C+H	1002	1085	860	251	2036

The expected peak match (Equation 5.1) for ^{15}N resolved NOESY and ^{13}C resolved NOESY was calculated with respect to the respective HSQC peak list. Both combinations of corresponding dimensions were compared for each pair of peak lists.
[a] H-N-plane including the directly bonded proton.
[b] H-N-plane including the distant proton.

5.3.2 Peak list matching for two corresponding dimensions

The performance of the *Peakmatch* algorithm was assessed using differently prepared peak lists. Manually generated peak lists are used as examples of high data quality and are thus expected to yield good results. Automatically picked peak lists, on the other hand, contain different levels of noise depending on the data set and the type of experiment. Finally, the robustness of the algorithm was evaluated systematically using a simulated SH2 data set by random deletion and addition of peaks. Downhill simplex optimization was used, unless noted otherwise.

Examples for automatically or manually prepared pairs of peak lists and the corresponding match score functions are shown in Fig. 5.4. The match score function for two corresponding dimensions shows a well-defined and narrow optimum at the optimal offset position even in the presence of many artifact peaks (Fig. 5.4c,d). Although this is true for every pair of peaklists in our test data set, there might be exceptions, and we want to mention one example briefly. In case of peak doublings due to very narrow lines in the spectrum, eg. for very small proteins at high magnetic field, two optima might occur in the match score function. For every peak being doubled, the match score function will show two optima of equal height and the algorithm will choose one of the two op-

tima randomly depending on the starting conditions of the optimization procedure. If one optimum is larger, a complete grid search will choose the larger optimum, whereas the downhill simplex optimization might get trapped at the smaller optimum.

The *Peakmatch* algorithm was applied to all automatically or manually prepared data sets using either [^{13}C,^{1}H]-HSQC or [^{15}N,^{1}H]-HSQC as reference peak lists. This resulted in a total of 91 pairs of peak lists. The quality of the automatically prepared peak lists depended strongly on the quality of the spectra. Especially some of the NOESY peak lists contained many noise peaks (see, for instance, Fig. 5.4a). Several offsets in the range between 0.1 ppm for heavy atoms and 0.01 ppm for protons and 10 ppm for heavy atoms and 1 ppm for protons were introduced into each target peak list and the *Peakmatch* algorithm was applied. Each optimization was performed using two different random starting simplexes and the offset with the highest match score was taken as the final result. An optimization result was considered correct if the difference between the introduced offsets and the calculation result was less than 0.33 times the chemical shift tolerance in all corresponding dimensions, i.e. if the offsets were correct within 0.01 ppm for ^{1}H and 0.13 ppm for ^{13}C and ^{15}N dimensions.

Using this criterion, the *Peakmatch* algorithm found the correct offsets for all automatically or manually prepared pairs of peak lists and all four offsets tested for each peak list pair. In all cases tested the optimal offsets determined by downhill simplex optimization coincided with the correct solution. Therefore, also a complete grid search would certainly find the correct result as long as the initial grid size is larger than the offset. This shows that the maximum of the match score function of Equation 5.2 approximately counts the number of peaks in the two peak lists whose position matches within the ranges $\sigma\Delta_k$, and does so with minimal influence from other, non-matching peaks. The mat describes correctly the optimal offsets also for peak lists that are far from perfect. The algorithm works reliably and with high precision over a large range of offsets for peak lists from a variety of spectra for backbone and side-chain assignment as well as ^{13}C- and ^{15}N-resolved NOESY experiments from five different proteins. The lower data quality from automatic peak picking did not have any significant effect on the offset determination by the *Peakmatch* algorithm.

The algorithm can match any combination of peak lists as long as they contain corresponding dimensions. Instead of using [^{15}N,^{1}H]-HSQC or [^{13},^{1}H]-HSQC reference peak lists, other suitable spectra can be chosen. For instance, we performed all offset determinations for the protein DsbA using the HSQC-planes of the CCH and HNCO spectra as reference peak lists. In all cases the correct offsets were found.

The robustness of the algorithm was also investigated systematically with respect to missing peaks, additional artifact peaks as well as unexpected shift changes among different

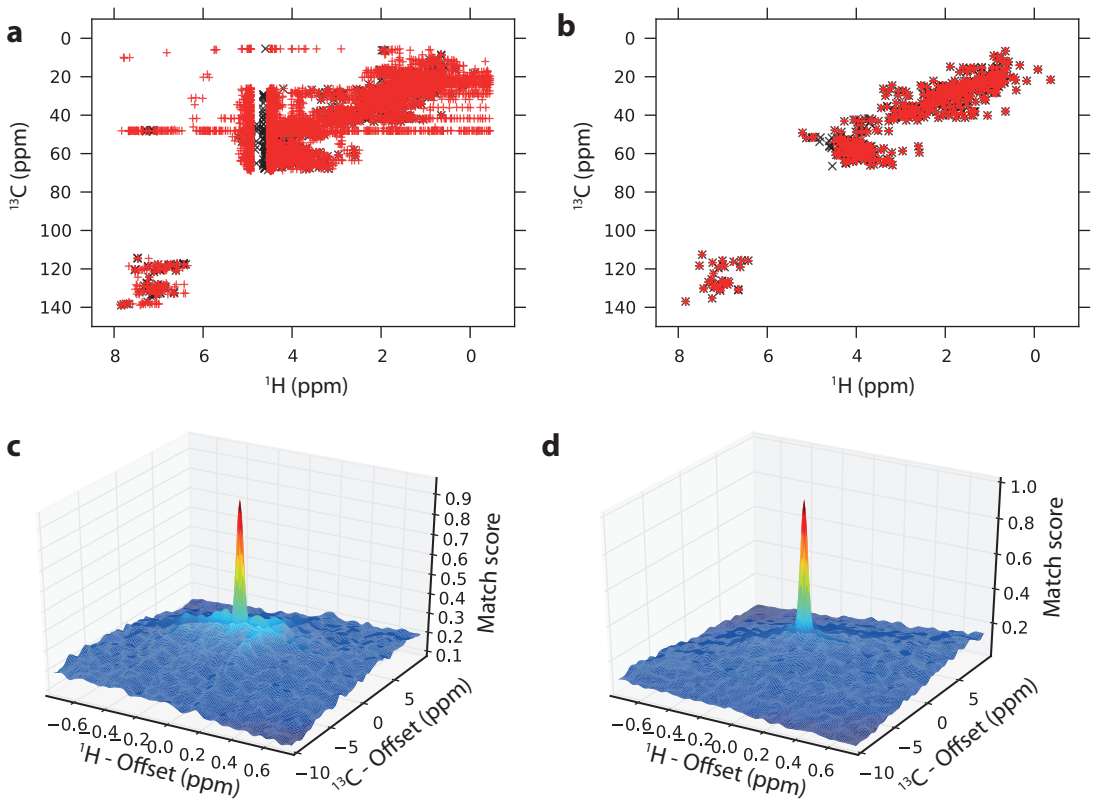


Figure 5.4: Graphical representation of manually and automatically generated peak lists and the corresponding match score functions for two corresponding dimensions. Peak lists are taken from the protein ENTH. The target peak list is ^{13}C-resolved NOESY (red in a and b) and the reference peak list [^{13}C,^{1}H]-HSQC (black in a and b). **a** Peak list from automatic peak picking. **b** Peak list from manual peak picking. **c** Match score function for automatic peak picking. **d** Match score function for manual peak picking.

peaklists, for example due to temperature changes between the experiments. Starting from a simulated data set for the protein SH2 consisting of all expected peaks (see Methods), randomly up to 90 % of the peaks in the reference and/or target peak lists were deleted. An offset was introduced in each target peak list and the *Peakmatch* algorithm was applied in the same way as with the experimental peak lists (Fig. 5.5). Deletions of up to 80 % of the peaks in either the reference or target peak list and deletion of up to 50 % of the peaks in both lists simultaneously had no effect on the offset determination. The amount of incorrect offsets increased up to 4 % for deletion of 90 % of the peaks in the target peak list (Fig. 5.5, triangles), up to 20 % for deletion in the reference peak list (Fig. 5.5, circles) and up to 76 % for deletion in both lists simultaneously (Fig. 5.5, stars). The effect of noise was evaluated by adding artifact peaks at random positions up to ten times the amount of peaks in the original peak list. The addition of randomly placed artifact peaks had no effect on the offset determination up to 500 %, independent of whether the peaks were added to the reference peak list, the target peak list, or to both peak lists simultaneously.

As for the deletion of peaks, the addition of more artifact peaks caused more incorrect offsets when adding peaks to both lists simultaneously (47 % incorrect offsets for 1100 % peaks). Addition of artifact peaks to the target peak list was effectless, whereas addition of artifact peaks to the reference peaklist increased the amount of incorrect offsets up to 16 % for 1100 % peaks. The effect of unexpected peak shifts among various peak lists for example due to sample heating was investigated starting from the simulated data set for the protein SH2 which was also used for investigation of the effect of missing peaks and artifact peaks. The chemical shift value of each atom was shifted by a gaussian distributed random number with a standard deviation of 1/2 the tolerance of the respective atom type and limited to a maximum of two times the respective tolerance value. Each peak list was then simulated using a different chemical shift file. A constant offset was introduced in each target peak list and the *Peakmatch* algorithm was applied. Random peak shifts up to two times the tolerance had no effect on the *Peakmatch* results. This again shows that the algorithm is very robust with respect to data imperfections.

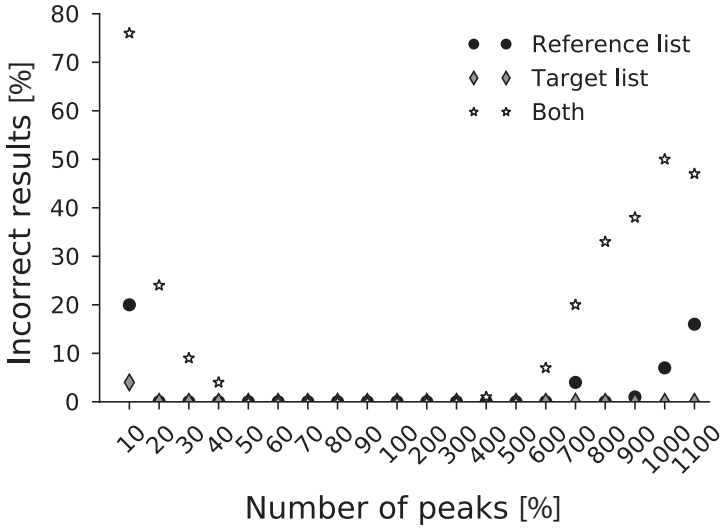


Figure 5.5: Robustness of the Peakmatch algorithm with respect to missing peaks as well as additional artifact peaks. Starting from a simulated data set for the protein SH2 consisting of all expected peaks (see Methods), randomly up to 90 % of the peaks in the reference and/or target peak lists were deleted and artifact peaks at random positions were added up to ten times the amount of peaks in the original peaklist. An offset was introduced in each target peak list and the *Peakmatch* algorithm was applied. The number of incorrect optimizations was measured in per cent. Every deletion or addition of artifact peaks was repeated five times using a different random number generator seed and results were averaged.

The number of independent downhill simplex optimization runs with different random start simplexes can be specified by the user. All optimizations mentioned were performed using two independent runs and the fact that no offset errors occurred indicates that in

general two runs are sufficient. To calculate the probability that the correct offsets are found when performing n independent runs, we performed 100 runs for each optimization and took the fraction of successful runs as the probability P_1 to find the correct offsets in a single run. Assuming that individual runs are mutually independent, the probability to find the correct offsets in n runs is $P_n = 1 - (1 - P_1)^n$. Using manually prepared peak lists, the percentage of correct optimizations was on average 99 % and in all cases above 95 %. This corresponds to an average probability of 99.99 % and a minimum probability of 99.75 % that two independent runs will yield the correct result. When using peak lists from automatic peak picking, the percentage of correct optimizations was on average 97 %, and the minimal percentage was 88 %. This leads to an average probability of 99.91 % and a minimal probability of 98.56 % that two independent runs will yield the correct result.

The match score S of Equation 5.2 is normalized by the expected match score E of Equation 5.1. For a perfect match of two ideal peak lists one thus obtains $S = 1$. The optimal match score for experimental peak lists, however, depends strongly on the quality of the peak lists. Missing peaks decrease and additional peaks potentially increase the score, which makes it difficult to judge the result of an offset determination simply by the match score value. The normalized match score values of all individual calculations performed with manually prepared peak lists were 0.19 – 1.08 (average 0.7) for the correct results and 0.04 – 0.14 (average 0.11) for the optimizations yielding incorrect results. The corresponding score values for the automatically prepared peak lists were 0.27-1.62 (average 0.89) for the correct results and 0.08 – 0.85 (average 0.21) for the optimizations yielding incorrect results. On average the correct results have thus much higher score values than the incorrect ones. Nevertheless, correct and incorrect results cannot be separated clearly by their individual match score values. In particular, the results for automatically prepared peak lists include correct results with match scores as low as 0.27 as well as incorrect results with match scores up to 0.85. Since it is not straightforward to distinguish correct from incorrect results by the match score value, the overlay of the peaks (projected onto the corresponding dimensions, if necessary) in the reference peak list and the optimally shifted target peak list is visualized as a graph (Fig. 5.3b). Based on this diagram the user can evaluate the result and decide whether to use the optimized peak lists or not.

The runtime of the algorithm depends on the number of peaks in the reference and target peak lists, and on the number of evaluations of the match score function of Equation5.2. The number of function evaluations differs for the different optimization procedures. We compared the runtime for downhill simplex optimization, a grid search with a limited grid size of 2 ppm for heavy atoms and 0.15 ppm for protons, and a grid search with a larger grid of 10 ppm for heavy atoms and 0.75 ppm for protons using an Intel E5-2690

2.9 GHz processor. The shortest average runtime of 3.9 s occurred for the grid search with limited grid size (281 function evaluations). The average calculation time using the downhill simplex optimization procedure was 5.2 s (on average 500 function evaluations), and the largest average calculation time of 33.5 s was required for the larger grid search (2731 function evaluations). Except in the case of small expected offsets, it is thus most efficient to use downhill simplex optimization.

5.3.3 Peak list matching for one corresponding dimension

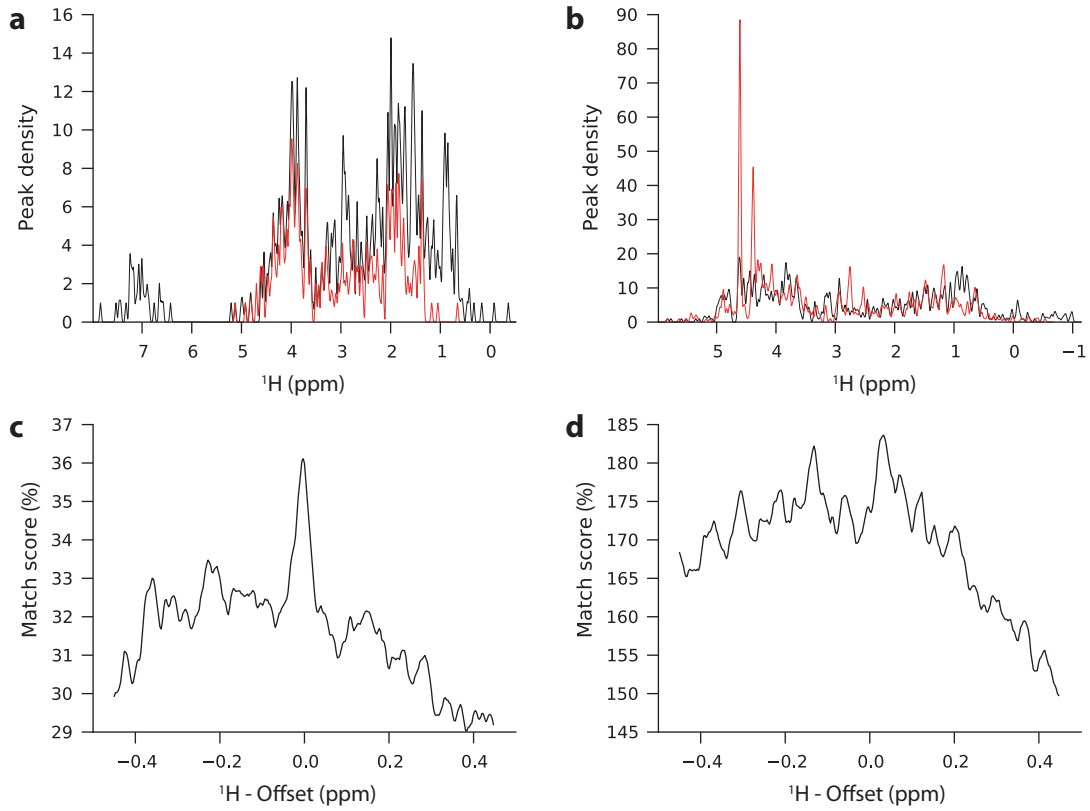


Figure 5.6: Graphical representation of manually and automatically generated peak lists and corresponding match score functions for one corresponding dimension. **a** Peak density for manually prepared peak lists from HBHACONH (black) and [^{13}C,^{1}H]-HSQC (red) spectra of the protein ENTH, obtained by plotting a Gaussian lineshape of unit height and standard deviation 0.03 ppm at the ^{1}H position of each peak. **b** Peak density for automatically picked peak lists from HC(CO)NH (black) and [^{13}C,^{1}H]-HSQC (red) spectra of the protein RHO. **c** Match score function for the manually prepared peak lists from a. **d** Match score function for the peak lists from automatic peak picking in b.

There are target peak lists that have only one corresponding dimension in common with the reference peak list. This makes the correct matching more difficult than with two or more corresponding dimensions. We tested the performance of the *Peakmatch* algorithm using only one corresponding dimension. The match score function for one correspond-

ing dimension does in general not show a single narrow maximum, but instead a larger number of local optima (Fig. 5.6c and d). Since the downhill simplex optimization might be trapped in local optima and the calculation time is not an issue for one-dimensional optimization, we limited the optimization procedure to a full grid search. We used again the aforementioned manually or automatically prepared peak lists and [^{13}C,^{1}H]-HSQC or [^{15}N,^{1}H]-HSQC as reference peak lists and performed a one-dimensional grid search to determine the optimal chemical shift offset, which was considered correct if it was within the chemical shift tolerance Δ_1 for the corresponding dimension, i.e. 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N. When using manually prepared peak lists, the match score function showed in all cases a global optimum at the correct solution (Fig. 5.6c). However, there are many local optima with match score values of similar magnitude, which makes it difficult to distinguish the correct from incorrect solutions based on the match score value. The superposition of the one-dimensional projections of the peak lists (Fig. 5.6a) shows that the peak lists overlay nicely, which explains the fact that all match score functions have their global maximum at the correct offset. In the case of automatically picked peak lists, however, the global optimum does in many cases not represent the correct solution. One example is shown in Fig. 5.6d. The graphical representation of the one-dimensional projections of the peak lists (Fig. 5.6b) show that it is difficult to see how the two peak lists should be overlayed correctly, reflecting the fact that the match score function has multiple maxima of similar size. These results indicate that the *Peakmatch* algorithm can also be used with only one corresponding dimension if good quality input data is available but results are much less reliable than with two or more corresponding dimensions and should be corroborated by visually checking the superposition of the one-dimensional projections of the peak lists.

5.3.4 Peak list matching for three corresponding dimensions

The *Peakmatch* algorithm can match any number of corresponding dimensions, even though in practice more than two corresponding dimensions occur rarely. One application is adapting the data from two three-dimensional spectra of the same type recorded under slightly different experimental conditions. As an example, we tested a pair of automatically picked peak lists for ENTH with three corresponding dimensions, CBCANH and CBCACONH, and performed a grid search as well as downhill simplex optimization. Both optimization strategies yielded the correct solution. The computation time for the grid search was 60 s due to a large number of function evaluations. In contrast, the computation time for the downhill simplex optimization increased only to 11.4 s. Compared to the two-dimensional case, the number of function evaluations by the downhill simplex

algorithm did not increase significantly and the increased runtime resulted mainly from the longer computation time for a single function evaluation.

5.3.5 Peak list matching against a chemical shift list

The *Peakmatch* algorithm can also be used to find the optimal offsets to match a peak list to a given chemical shifts list. For instance, NOESY peak lists can be matched to a chemical shift list obtained from through-bond spectra prior to automated NOE assignment and structure calculation. To this end, a reference peak list of the same spectrum type as the target peak list is simulated on the basis of the sequence and the chemical shift list, and then used as input to the *Peakmatch* algorithm treating all spectral dimensions as corresponding dimensions. Again, this approach has the advantage that it can be applied to unassigned peak lists.

5.3.6 Example for *Peakmatch* application

We have chosen automatic chemical shift assignment as one possible application of the *Peakmatch* algorithm. We have used three automatic data sets from our test data set (ENTH, SH2, and RHO) and performed automatic chemical shift assignment using the FLYA software (Schmidt and Güntert, 2012). To illustrate the consequences of chemical shift referencing inconsistencies among different peak lists of the same data set, we have introduced artificial random offsets in each peak list prior to automatic chemical shift assignment. Offsets were introduced either within the assignment tolerance (0.03 ppm for protons and 0.4 ppm for heavy atoms) or within 1.5 times the assignment tolerance and assignment results were compared to the optimized data set which was generated using the *Peakmatch* algorithm. A summary of results is presented in Table 5.2. In case of small offsets within the assignment tolerance all proteins show between 88 % and 89.8 % correct assignments when considering backbone as well as side chain atoms (first column in Table 5.2, third number). Results can be improved slightly to between 90.1 % and 90.8 % when optimizing the data sets using the *Peakmatch* program (third column in Table 5.2). In contrast, when using peak lists with larger chemical shift referencing offsets within 1.5 times the tolerance, the amount of correct assignments goes down to between 74.4 % and 85.2 % (second column in Table 5.2). This demonstrates nicely the possible improvement of results when applying the *Peakmatch* program prior to automatic chemical shift assignment.

Table 5.2: Expected peak match values for NOESY peak lists using the respective HSQC peak list as reference.

Protein	Small offsets[a]	Larger offsets[b]	Optimized data set
Rho	93.6 / 83.7 / 88.0	88.4 / 63.5 / 74.4	96.2 / 85.5 / 90.1
Enth	94.4 / 84.2 / 88.4	81.6 / 72.8 / 76.4	95.5 / 87.5 / 90.8
Sh2	97.7 / 84.4 / 89.8	96.9 / 77.3 / 85.2	98.4 / 85.6 / 90.8

Each offset is a random number within the specified range and each peak list is shifted independently by a different random offset.
The automatic assignment was performed using the FLYA program and automatically picked peak lists. All results are given as percentage of correctly assigned atoms with respect to the reference assignment and include backbone atoms (first number), side chain atoms (second number) and both (third number).
[a] Small offsets have a maximum value of 0.03 ppm for protons and 0.4 for heavy atoms.
[b] Larger offsets have a maximum value of 0.045 ppm for protons and 0.6 ppm for heavy atoms.

5.4 Conclusion

We have presented a new algorithm that determines the optimal offset between two multidimensional peak lists that contain corresponding dimensions. The algorithm identifies corresponding dimensions automatically based on the expected peaks for the given experiments and then optimizes a match score function for the experimental peak lists. Extensive tests showed that the algorithm works very reliably also with input peak lists that are far from ideal, e.g. those generated by automatic peak picking programs, provided that there are at least two corresponding dimensions. Principal advantages of the algorithm are that (i) it can be applied to unassigned peak lists, (ii) it is highly tolerant against the common imperfections of experimental peak lists, (iii) the criterion for optimal matching is mathematically simple and largely captures what an experienced spectroscopist would do manually, and (iv) its application is straightforward and quick. The optimization can be performed using a complete grid search or a downhill simplex optimization procedure. In all test cases, both procedures performed equally well when using two corresponding dimensions. When using only one corresponding dimension a complete grid search is recommended as the downhill simplex algorithm has a higher chance of getting trapped in a local optimum and computation time is no issue when using only one corresponding dimension. In case of more than two corresponding dimensions both methods can in general be used, however, the complete grid search can be very time consuming depending on the grid size, whereas computation time rises only slightly for the downhill simplex procedure when increasing the amount of corresponding dimensions.

5.5 Implementation in CYANA

The CYANA software package written in Fortran language includes numerous functionalities for the calculation and analysis of NMR structures. All these functionalities, accessible through a large number of CYANA commands, are used via the external scripting language INCLAN. Several CYANA commands can be stringed together in a macro which executes the specified commands one after another. Other macros can as well be called within a macro. The actual peak list optimization is implemented in Fortran and can be accessed via the two commands `peaks offset` and `peaks match`. All functions around the actual optimization, such as the determination of corresponding dimensions or the generation of output, are performed by the macro `peakmatch.cya`. The newly available commands and the peakmatch macro including available parameters are introduced in more detail in the following.

5.5.1 CYANA commands

peaks match

The command `peaks match` calculates one match value for two given peaks lists using the specified dimensions without optimization. The following parameters can or need to be provided. Either experimental peak lists or expected peaks of two types of spectra can be used. The experimental peak lists or the expected peaks, however, need to be available prior to executing this command. The output includes the experimental match value (available in INCLAN via `result('realmatch')`) or the expected match value (`result('artexpmatch')`).

- `peaks=string` (required if more than two peak lists are in memory)

  The peaks match command calculates the match value between two peak lists that need to be specified. If only two peak lists are in memory, it can be skipped, and the first peak list will be used as reference. The two peak list names are separated by "," (e.g. `peaks=1H15.peaks,N15NOESY.peaks`, or `peaks=1H15N,N15NOESY`).

- `dimensions=string` (required)

  Dimensions to be matched need to be specified for the reference peak list and the peak list to be matched. The labeling of each dimension has to match the library entry of the respective experiment in in cyana.lib. Several dimensions for one peak list are combined using "+" (e.g. "H+N" for [^{1}H,^{15}N] HSQC). The selected dimensions of the two peak lists are combined using "," (e.g. `dimensions=H+N,HN+N` for matching of the HSQC plane of a 3D [^{1}H,^{15}N] NOESY peaklist onto a 2D [^{1}H,^{15}N] HSQC peak list). No spaces are allowed in the specification.

- `sigma=real` (optional, default=1.0)

  The parameter `sigma` is used for match score calculation as a dimensionless scaling factor that determines the significance of a deviation. Details can be found in the description of the match score of the methods section.

- `maxpeaks=integer` (optional, default=10000)

  Not every peak of the peak list to be matched contributes to the individual match score of a given peak in the reference peak list, but only a specified number of the closest peaks (i.e. that have the highest match value). This number is specified by the parameter maxpeaks. When the default of 10000 is chosen, most likely all peaks do contribute.

- `expected` (optional)

  The option expected calculates the matchscore for two sets of expected peaks. These have to be generated, for example using the spectrum command, prior to calculating the expected match score.

peaks offset

The command peaks offset matches two peak lists using the specified dimensions and the chosen optimization procedure (i.e. downhill simplex optimization or complete grid search). The output includes the match score after optimization (available in INCLAN via `result('maxmatch')`) and the offset for each dimension (`result('offsets')`). The peak positions after optimization are stored in the `ppm`-array in CYANA, thus writing the peak lists using the `write peaks` command after the optimization can be used to obtain the shifted peak lists.

- `peaks=string` (required if more than two peak lists are in memory)

  See parameter `peaks` in `peaks match`.

- `dimensions=string` (required)

  See parameter `dimensions` in `peaks match`.

- `sigma=real` (optional, default=1.0)

  See parameter `sigma` in `peaks match`.

- `maxpeaks=integer` (otional, default=10000)

  See parameter `maxpeaks` in `peaks match`.

- `optimization=string` (optional, default=amoeba)

  Specifies the optimization procedure and can be either `amoeba` (which is the default) or `gridsearch`. If `gridsearch` is chosen, then the algorithm performs a fixed three-step optimization schedule which depends on the defined `gridsize` (see parameter `gridsize`) as well as the tolerance of each dimension. The grid is defined from `-gridsize` to `+gridsize` with the tolerance as space between two grid points in the first iteration. Around the optimum of the first iteration, the second grid is defined from *-tolerance* to *+tolerance* with 0.25 times the tolerance as spacing. The third grid is from -0.25 times the tolerance to +0.25 times the tolerance and 0.05 times the tolerance as spacing.

  If `amoeba` is chosen as optimization procedure, then a single downhill simplex minimization is performed unless the option `auto` (see parameter description `auto`) is selected in addition. The starting simplex for any minimization is defined based on the parameter `gridsize`, the tolerance and a gaussian distributed random number. The downhill simplex optimization can become trapped in local minima if the optimization landscape is not smooth, which can be influenced by the parameter `sigma`. If the option `auto` is selected, the algorithm performs a three-step optimization using different sigma-values and several runs for each step which greatly reduces the chance of getting trapped in a local minimum.

- `gridsize=real` (optional, default=3.0)

  The parameter `gridsize` influences the size of the initial grid during the gridsearch, as well as the size of the starting simplex during the amoeba-search. For the gridsearch, it should be chosen based on the expected offset. For the amoeba search, it is highly recommended to use the option `auto` where the algorithm performs a three-step optimization using 40.0, 10.0, and 1.0 as gridsize value in the respective iteration.

- `runs=integer` (optional, default=10)

  The parameter `runs` specifies the number of downhill simplex optimizations that is performed for each of the three steps during the optimization schedule when the option `auto` is selected. The result of the run with highest match score is used as starting position or the following optimization step.

- `auto`

  If the option `auto` is selected, the algorithm performs a three-step optimization schedule, where each step differs in the value for `gridsize` which influences the size of the starting simplex, as well as the value for `sigma` which influences the shape of

the match score landscape. In the first iteration, `sigma=10.0` and `gridsize=40.0`, in the second iteration `sigma=5.0` and `gridsize=10.0` and in the last iteration both values are set to 1.0.

5.5.2 Macro

peakmatch.cya

The macro `peakmatch.cya` uses the previously introduced CYANA commands in order to optimize a list of specified peak lists with respect to one specified reference peak list. The user can choose the dimensions of the reference peak list and corresponding dimensions in each peak list to be matched are determined automatically. The optimized peak lists are generated as output and additionally a summary file including the determined offset and the match score for each peak list that was optimized. The macro optionally generates a plot for each peak list showing peak positions of the reference peak list and the optimized peak list in case of one and two dimensions used for optimization.

- `reference=string` (required)

  The parameter `reference` specifies the reference peak list.

- `peaks=string` (required)

  The parameter `peaks` can be a list of peak lists, each separated by "," without spaces (e.g. `peaks=N15NOESY.peaks,HNCO.peaks,HNCA.peaks` or `peaks=N15NOESY,HNCO,HNCA`). Each peak list will be matched to the reference peak list.

- `dimensions=string` (required)

  See parameter `dimensions` in `peaks offset` or `peaks match`. It is, however, possible to just specify the dimensions of the reference peak list (e.g. `dimensions=H+N`) and the algorithm will determine corresponding dimensions in each of the peak lists to be matched based on the expected peak match of the two experiment types.

- `sigma=real` (optional, default=1.0)

  See parameter `sigma` in `peaks offset` or `peaks match`. The parameter is not used if the optimization is performed using the amoeba method and the option `simplexsize=auto` which calls the offset determination using the option `auto` in the `peaks offset` command.

- `maxpeaks=string` (optional, default=auto)

  See parameter `maxpeaks` in `peaks offset` or `peaks match`. The default value `maxpeaks=auto` determines an average value based on the expected peak match,

which is highly recommended to use. The resulting value measures the average number of peaks in the peak list to be matched that correspond to a certain peak in the reference peak list. Only the dimensions used for optimization are considered. This leads for example to a larger value in the case of matching N15NOESY to N15H1 when considering the HSQC-plane for matching as compared to matching HNCO to N15H1 considering the HSQC-plane for matching. When using this option, then the calculated match value of the experimental peak lists can be evaluated based on the expected match value, whereas the value itself has no meaning if all peaks of the peak list to be matched contribute to the match value of a given peak in the reference peak list. It should, however, be mentioned that this does not change the position of the optimum, but only the absolute match value.

- `optimization=string` (optional, default=amoeba)

  See parameter `optimization` in `peaks offset`.

- `gridsize=real` (optional, default=3.0)

  See parameter `gridsize` in `peaks offset` or `peaks match`. It should, however, only be used in combination with `optimization=gridsearch`. The respective parameter for the downhill simplex optimization is replaced by `simplexsize`.

- `simplexsize=string` (optional, default=auto)

  The parameter `simplexsize` combines the parameter `gridsize` and the option `auto` from the `peaks offset` command for the downhill simplex optimization and should be used only in combination with `optimization=amoeba`. The default value `auto` performs the optimization using the three-step optimization introduced in the description of the option `auto` for `peaks offset`. Any real number leads to one single downhill simplex optimization and the parameter `simplexsize` is then used in the same way as the parameter `gridsize` for the `peaks offset` command.

- `runs=integer` (optional, default=2)

  See parameter `runs` in `peaks offset`.

- `out=string` (optional, default=peakmatch.out)

  The parameter `out` specifies the name of the summary output file.

- `ending=string` (optional, default="_opt")

  The parameter `ending` specifies the ending of the optimized peak list file.

- `resolution=real` (optional, default=0.005)

The parameter `resolution` refers to the 1D plot which is created if only one dimension is used for matching.

- `noplot` (option)

  The option `noplot` skips the generation of a graphical representation (either 1D or 2D) of the optimized peak lists.

- `skipdiagonal` (option)

  The option `skipdiagonal` omits diagonal peaks during the generation of expected peaks. This is recommended, if diagonal peaks are not present in the experimental peak lists.

- `all` (option)

  The option `all` is used if all possible combinations of dimensions should be matched for a given pair of peak lists. If the option is not chosen, then only the one combination with the highest expected match value is used for optimization.

Chapter 6

Increased reliability of NMR protein structures by consensus structure bundles

This chapter is based on the following publication:

Buchner L. and Güntert P. Increased reliability of NMR proteins structures by consensus structure bundles. *Structure* 23(2):425-434, 2015

6.1 Introduction

NMR spectroscopy is besides X-ray crystallography the most widely used technique to determine three-dimensional (3D) structures of macromolecules. NMR structures are typically represented as bundles of conformers, each conformer being the result of a minimization procedure that optimizes the agreement between the 3D structure and the experimental data. The structural ensemble is characterized by its precision representing the positional uncertainty of the atomic coordinates as well as its accuracy, which is a measure of the closeness to the true structure (Spronk et al., 2004; Zhao and Jardetzky, 1994).

Structural precision is commonly quantified by the RMSD radius of the structure bundle, i.e. the average RMSD value between the individual conformers and the mean coordinates of the bundle. Structural accuracy can be quantified by the RMSD bias, i.e. the RMSD between the mean coordinates of the structure bundle and a reference structure (or the mean coordinates of a reference structure bundle) that is assumed to represent the "true structure" (Güntert, 1998). RMSD values are calculated for the atoms in the structured regions of the protein, which can be identified by visual inspection or algorithms such as CYRANGE (Kirchner and Güntert, 2011).

Experimental data are provided in the form of structural restraints, the most common ones being distance restraints from NOESY experiments as well as angular restraints for example from chemical shift analysis with the program TALOS+ (Shen et al., 2009). The conversion of NOESY peaks into distance restraints requires the assignment to atom pairs and the calibration of peak intensities into upper distance limits. This NOESY assignment is crucial for the outcome of a structure calculation and errors can have severe consequences for the quality of the resulting structure (Jee and Güntert, 2003). It should thus be performed as objectively as possible. Several software tools therefore combine automatic NOESY peak assignment and structure calculation in an iterative way (Herrmann et al., 2002a; Huang et al., 2006; Rieping et al., 2007). It has been shown that already a very small number of incorrect distance restraints can lead to a highly precise but completely misfolded protein structure that is not always recognized as such (Nabuurs et al., 2006). This can be attributed to the lack of an independent and reliable measure of NMR structure quality. Instead, some NMR spectroscopists tend to compare the precision of a structure ensemble to the X-ray resolution and use it as a measure of the structure quality. Consequently, there is a widespread ambition to improve the precision, i.e. to minimize the RMSD radius of structure bundles in the belief of increasing the quality of the structure. This misconception is the cause of a widely observed overestimation of NMR structure accuracy (Spronk et al., 2004; Spronk et al., 2003), which limits the reliability of NMR structures without further validation. Currently existing validation tools can reveal errors,

but they do not always guarantee a reliable result. They perform especially well when validating completely misfolded protein structures (Nabuurs et al., 2006). However, deviations in the range of 2 – 3 Å RMSD bias from the reference structure are not likely to be recognized although they occur much more frequently than severely erroneous structures (Saccenti and Rosato, 2008). There have also been attempts to combine various validation measures into an estimate of structural accuracy, e.g. the RMSD from the true structure can be estimated by a linear combination of (suitably normalized) validation parameters (Bagaria et al., 2012), or, similarly, an "equivalent resolution" can be obtained from multiple validation scores (Laskowski et al., 1996; Bagaria et al., 2013). However, while over a large number of different protein structures there is a visible correlation between these accuracy estimates and the true accuracy, the predictive power for a given, single NMR structure determination remains limited.

The precision of a structure bundle directly relates to the amount of meaningful long-range distance restraints, i.e. the information content of a restraint set (Nabuurs et al., 2003). The most severe possible problem of using distance restraints for structure calculation is the bias resulting from erroneous NOESY peak assignments. Potential error sources include the sequence-specific resonance assignments, the identification of true NMR signals, the assignment of NOESY peaks to atom pairs, and the calibration of upper distance limits. An incorrectly assigned NOE can distort a structure, whereas a too tight but correctly assigned NOE will have a much smaller impact. Although automation increases the reproducibility and reduces the bias originating from subjective user choices, the algorithms are still not perfect, especially in cases of limited data quality, e.g. signal overlap, low signal-to-noise ratios, or sparse data. Iterative combined automated NOESY peak assignment and structure calculation with CYANA can converge towards misfolded protein structures that are not immediately recognized as such due to their high bundle precision (Jee and Güntert, 2003). Similar problems can arise also with other algorithms for automated NOE assignment (Rosato et al., 2012). This occurs predominantly in cases where the protein fold is poorly defined in early cycles of the calculation and subsequent NOESY peak assignments in following cycles are based on incorrect assumptions about the protein fold. Errors are only rarely reflected in a lower precision of the final structure calculation result, but rather in a highly precise but inaccurate protein structure.

Since the outcome of a structure calculation in such cases depends partly on the random initial structures, structure calculations based on the same set of experimental data but different random starting structures converge potentially to different structure bundles. The degree of deviation strongly depends on the data quality. However, some extent of deviation is observed on a regular basis even when using input data of good quality, indicating that precision significantly exceeds accuracy. Despite this fact, in general only

one final structure calculation is performed and its results are reported, although a different solution, obtained with a different random number generator seed, would represent the NMR data equally well. Many structure bundles determined by NMR spectroscopy thus have a precision that overestimates the accuracy and errors remain unrecognized when using simply the bundle precision and the agreement between structure and experimental data as a measure of quality.

Several attempts to solve this problem have been conducted. Spronk et al. have developed a tool that maximizes the RMSD radius while maintaining the agreement with the experimental data and the geometric quality (Spronk et al., 2003). This approach improves the sampling of structures within a given set of distance restraints. However, structural distortions due to erroneous distance restraints are not addressed. Since the set of distance restraints remains unchanged, the method does not necessarily improve the accuracy of a protein structure. Inferential structure determination was introduced as a fundamentally different approach to structure determination by NMR spectroscopy (Rieping et al., 2005). The method uses a Bayesian inference to derive an objective probability distribution to evaluate the structural ensemble that is generated based on a Monte Carlo Simulation. It is independent of empirical parameter estimates and increases the completeness of the sampling of conformational space that is in agreement with the experimental data.

A new protocol for combined automatic NOESY peak assignment and structure calculation will be introduced that provides a solution to the problem of overestimated bundle precision. The new protocol aims at yielding protein structures for which the bundle precision is a reliable measure of the structural accuracy and where the structure bundle covers well the conformational space that is allowed by the experimental data.

6.2 Methods

6.2.1 Generation of consensus distance restraint set

The algorithm performs 20 independent automated NOESY assignment and structure calculation runs using the same input data and different random number generation seeds, resulting in 20 individual structure bundles (Fig. 6.1a). The lowest energy structure of each of these 20 structure bundles is combined to obtain a new combined structure bundle (Fig. 6.1b). The precision of the combined structure bundle is a measure of the extent to which individual calculations differ from each other.

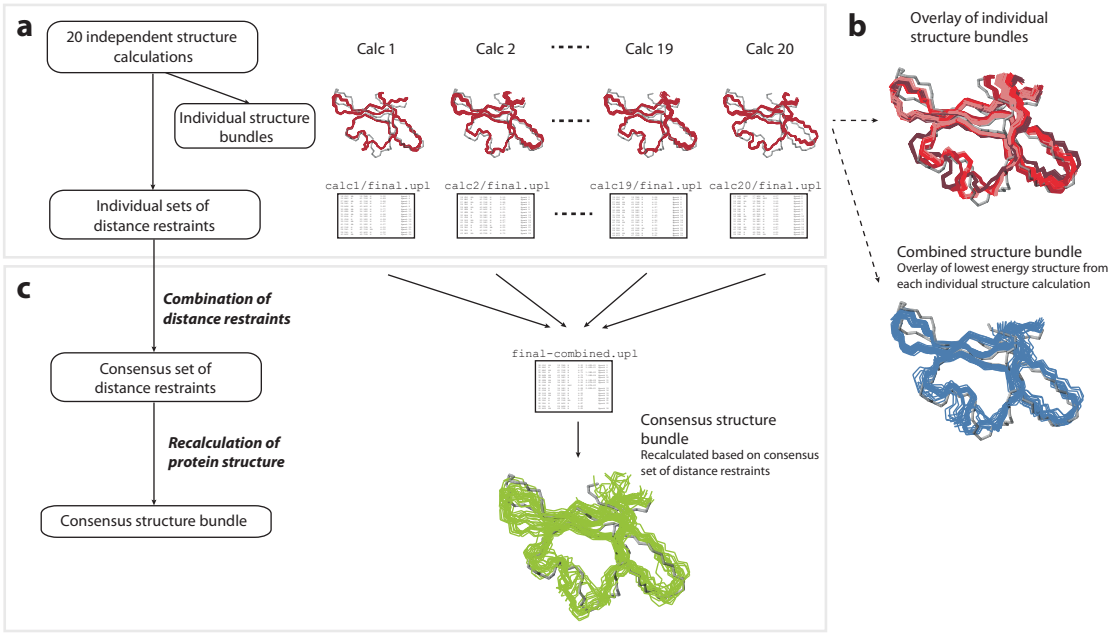


Figure 6.1: Schematic overview of the algorithm implemented in the CYANA software package. **a** Twenty independent runs of the standard CYANA automatic NOESY assignment and structure calculation procedure using the same input data, each starting from a different set of random structures. **b** Overlay of structure bundles from independent runs (top) and combined structure bundle consisting of the lowest target function structure of each run (bottom). **c** Combination of all individual sets of distance restraints and recalculation of the protein structure using the consensus set of distance restraints yields the consensus structure bundle.

Each of the 20 individual structure calculations leads to a different set of distance restraints as a result of the seven cycles of NOE assignment and structure calculation. These individual final sets of distance restraints are in optimal agreement with the respective structure bundle, however, they do not represent the aforementioned combined structure bundle. The combination of the individual sets of distance restraints yields a consensus set of distance restraints that represents the combined structure bundle and thus results in a structure bundle similar to the combined structure bundle when used as input for a further structure calculation (Fig. 6.1c). This final structure calculation is a simple, standard CYANA structure calculation without automatic NOE assignment. It uses the consensus NOE distance restraints (and other conformational restraints, if available) as input and yields the consensus structure bundle as output.

The combination makes use of the fact that each distance restraint is the result of a NOESY peak assignment. During the seven cycles of NOESY peak assignment and structure calculation peaks may have unambiguous assignments, ambiguous assignments, or remain unassigned. In the final structure calculation, all remaining ambiguities are resolved and non-stereospecifically assigned methyl- or methylene-protons are treated by symmetrization and pseudo-atom correction (Güntert et al., 1991). During the combina-

tion process, every distance restraint assignment originating from the same peak in all individual restraint sets is combined to obtain one ambiguous (or unambiguous) restraint.

Individual peaks may be assigned to different atom pairs in different structure calculation runs or they may remain unassigned in individual calculations. To form a consensus distance restraint data set that is suitable for recalculating the consensus structure bundle it is necessary to choose only those restraints that represent the combined structure bundle in a sufficient manner. Consequently, restraints are only chosen if the corresponding peak could be assigned (with any assignment(s)) in a specified minimal number of individual structure calculations, otherwise the complete peak will be discarded. If a restraint is chosen, then all atom pairs appearing in any of the respective peak assignments of the individual structure calculations are combined to obtain one ambiguous or unambiguous distance restraint. The threshold on the minimal number of individual structure calculations in which a peak must be assigned in order to be chosen for consensus restraint generation can be chosen by the user, however, after having tested the complete range of cutoff values, we recommend a threshold of 60 % of the individual structure calculations in which a peak needs to be assigned to any atom pair in order to be selected. Higher threshold values lead in a few cases to an unacceptably large loss of information by a very large number of discarded peaks, resulting in a severe underestimation of the achievable accuracy. Low threshold values, on the other hand, again increase the apparent precision due to a large number of restraints that are selected even though they represent only a small fraction of the conformers in the combined bundle.

Our choice of 60 % for the peak selection cutoff percentage can be rationalized from Fig. 6.3, Table 6.1, and Fig. 6.2. Overall, the results depend only weakly on the choice of the cutoff percentage. On the one hand, increasing the cutoff value slightly decreases the accuracy-to-precision ratio towards the ideal value of 1.0 (Fig. 6.3, left panels). On the other hand, lowering the cutoff increases the occurrence of structures with high accuracy (low RMSD to reference; Fig. 6.3, right panels). The average median accuracy-to-precision ratios and absolute accuracy values at different cutoff values are summarized in Table 6.1, which shows that a cutoff of 0.6 provides a good compromise between the two opposite trends. Fig. 6.2 shows the precision and accuracy as a function of the cutoff value for two examples from the CASD-NMR data set. In the first example, the results are almost independent from the cutoff, whereas in the second example there is a loss of accuracy with increasing cutoff.

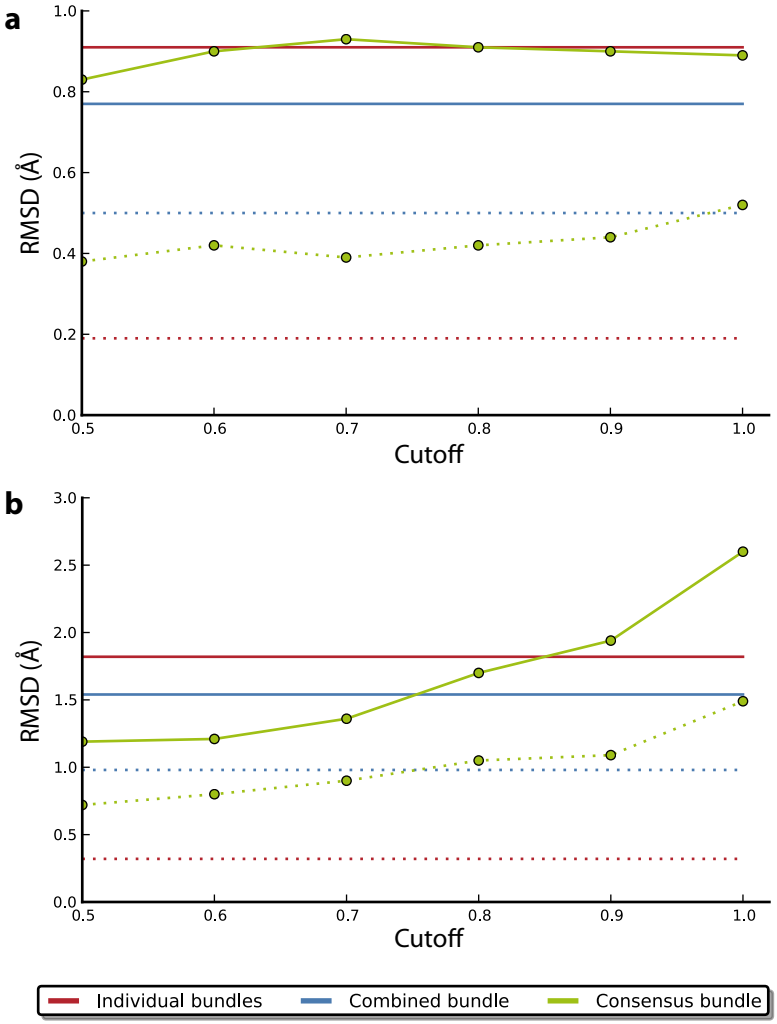


Figure 6.2: Precision and accuracy of individual bundles, combined bundles, and consensus bundles for different cutoff values. *Dotted lines*: Precision measured as the RMSD to the mean structure of the bundle. *Solid lines*: Accuracy measured as the RMSD bias with respect to the reference structure. Precision and accuracy of the individual bundles (red) and the combined bundles (blue) are not influenced by the cutoff-value. **a** OR135 raw data set, **b** HR5460A raw data set.

Table 6.1: Median accuracy-to-precision ratio and accuracy at different cutoff-values.

Cutoff	Accuracy-to-precision ratio		Accuracy	
	Median	Percent above 2.0	Median (Å)	Percent above 3.0 Å
0.5	1.47	22.3	3.61	55.3
0.6	1.40	15.9	3.71	56.8
0.7	1.38	11.3	3.97	58.5
0.8	1.35	7.8	4.39	62.0
0.9	1.32	6.0	5.00	65.6
1.0	1.31	5.5	6.18	74.0

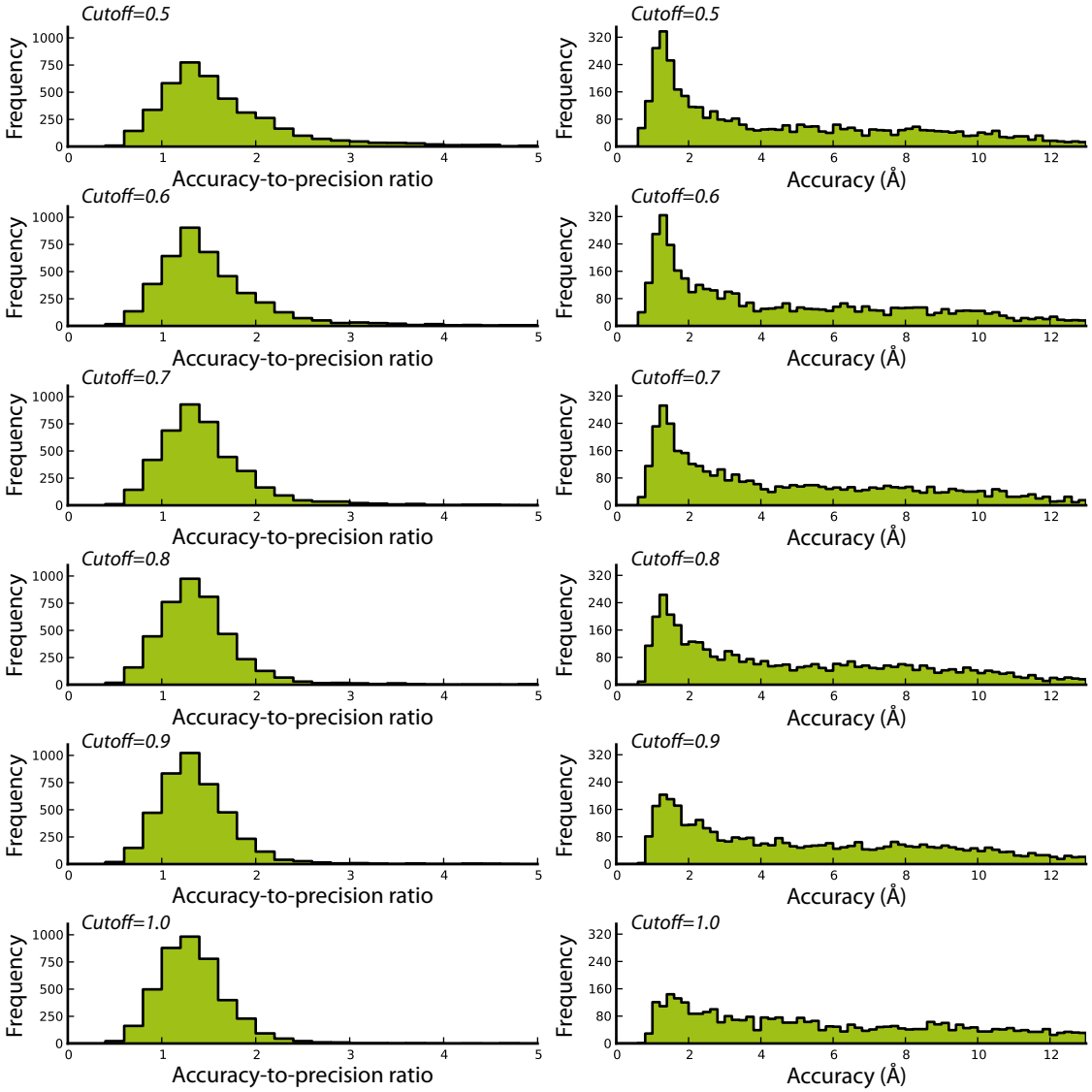


Figure 6.3: Histograms showing the accuracy-to-precision ratio and the consensus bundle accuracy for different cutoff values. The cutoff value represents the threshold on the minimal number of individual structure calculations in which a peak must be assigned in order to be chosen for consensus restraint generation (e.g. a cutoff of 1.0 means that a peak is only chosen, if the peak was assigned in every individual calculation). Results are presented for cutoff values in the range between 0.5 and 1.0. The test data is composed of 4050 solution NMR data sets from ten different proteins. It includes for each protein the original experimental data set and modifications that simulate a large variety of data imperfections (see Methods). *Left panels*: Accuracy-to-precision ratios, *Right panels*: Accuracy of the consensus bundle.

Combined structure bundles with low precision generally show large differences among the NOESY peak assignments from the individual runs. This in turn results in an increased ambiguity of the restraints as well as a larger number of discarded peaks in the combined data set. Altogether this reduces the information content of the combined restraint set, which in turn decreases the precision of the consensus structure bundle. Combined struc-

ture bundles with high precision on the other hand have very similar individual distance restraint sets which leads to the fact that most of the restraints are selected and have a low ambiguity in the consensus restraint set. The combined restraint list has thus more information content and, consequently, the consensus structure bundle will have a high precision.

The method generates a consensus set of distance restraints that essentially reproduces the combined structure bundle when used in a conventional structure calculation based on distance restraints. This is achieved since the precision of a structure bundle depends on the information content of the data set, which in turn is determined by the amount of meaningful long-range restraints as well as their ambiguity.

6.2.2 Individual structure calculations

Individual structure calculations are performed using the standard structure calculation procedure of the CYANA software (Güntert et al., 1997; Herrmann et al., 2002a; Güntert, 2009; Güntert and Buchner, 2015). The protein sequence as well as chemical shifts and unassigned peak lists from NOESY spectra are used as input for the structure calculations. Chemical shift assignments were taken from the BMRB whereas torsion angle restraints were taken from the Protein Data Bank. The chemical shift tolerance for NOESY peak assignments was set to 0.03 ppm for 1H and 0.3 ppm for 15N and 13C. The standard CYANA protocol was applied using 200 random starting structures and 15000 annealing steps during torsion angle dynamics. The 20 structures with lowest target function values were used as the final structure bundle. Details are described in (Schmidt and Güntert, 2013).

6.2.3 Data sets from CASD-NMR

In order to evaluate the advantages of consensus set distance restraints and consensus structure bundles for the reliability of protein structure calculations, we used input data of eight different proteins that were provided as test data sets for the CASD-NMR project in 2011 – 2012 (Rosato et al., 2012; Rosato et al., 2009). The same data set had already been used for the analysis of automatic chemical shift assignment based solely on NOESY spectra using the FLYA automated resonance assignment algorithm and subsequent structure calculations (Schmidt and Güntert, 2013). The present data set was of particular interest for our study since the structure calculation results presented in (Schmidt and Güntert, 2013) revealed a wide range of structural qualities. The most problematic cases are those that yield a structure with high precision but low accuracy, where problems generally remain hidden when performing just a single NOESY assignment and structure calculation run.

The eight proteins include: the human NFU1 iron-sulfur cluster scaffold homologue, Northeast Structural Genomics Consortium (NESG) target HR2876B (PDB accession code 2LTM, 107 amino acid residues, ordered residues 13 – 104); the human mitotic checkpoint serine/threonine-protein kinase BUB1 N-terminal domain, HR5460A (2LAH, 160 aa, 12 – 160); the RRM domain of RNA-inding protein FUS, HR6430A (2LA6, 99 aa, 12 – 99); the homeobox domain of the human homeobox protein Nkx-3.1, HR6470A (2L9R, 69 aa, 12 – 55); a *de novo* designed protein with IF3-like fold, OR135 (2LN3, 83 aa, 5 – 75) (Koga et al., 2012), a *de novo* designed protein with P-loop NRPase fold, OR36 (2LCI, 134 aa, 3 – 125); TSTM1273 from *Salmonella typhimurium* LT2, StT322 (2LOJ, 63 aa, 23 – 63); the NifI-like protein from *Saccharomyces cerevisiae* YR313A (2LTL, 119 aa, 16 – 116). The corresponding NMR structures deposited in PDB were used as reference structures in this work. In principle, also X-ray structures could be used as independently determined reference structures (but were not available for these proteins). The NMR data provided for this project were prepared according to standard NESG procedures (www.nesg.org). Two data sets were available for each protein, one containing "refined" NOESY peak lists that were used for the final structure calculations of the reference structures, and one containing the "raw" NOESY peak lists from an early stage of spectral analysis. Peak lists were generated from 15N-resolved NOESY spectra as well as 13C-resolved NOESY spectra. Chemical shift assignments were performed manually by experienced scientists and have been provided in addition to the NOESY peak lists.

6.2.4 Second test data set

The second test data set is composed of solution NMR data sets from ten different proteins. It includes for each protein the original experimental data set and modifications thereof that simulate a large variety of data imperfections. The ten proteins include: Copper chaperone of *Enterococcus hirae* (PDB accession code 1CPZ, BMRB accession code 4344, 68 amino acid residues) (Wimmer et al., 1999); Chicken prion protein fragment 128-242 (PDB 1U3M, BMRB 6269, 117 aa) (Calzolai et al., 2005); *Arabidopsis thaliana* ENTH-VHS domain At3g16270 (PDB 1VDY, BMRB 5928, 140 aa) (López-Méndez and Güntert, 2006; López-Méndez et al., 2004); Src homology 2 domain from the human feline sarcoma oncogene Fes (PDB 1WQU, BMRB 6331, 114 aa) (Scott et al., 2004; Scott et al., 2005); F-spondin TSR domain 4 (PDB 1VEX, BMRB 10002, 56 aa) (Pääkkönen et al., 2006); *Bombyx mori* pheromone binding protein (PDB 1GM0, BMRB 4849, 142 aa) (Horst et al., 2001); *Arabidopsis thaliana* rhodanese domain At4g01050 (PDB 1VEE, BMRB 5929, 134 aa) (Pantoja-Uceda et al., 2005; Pantoja-Uceda et al., 2004); *Williopsis mrakii* killer toxin (PDB 1WKT, BMRB 5255, 88 aa) (Antuch et al., 1996); stereo-array isotope labeled (SAIL) calmodulin (PDB 1X02, BMRB 6541, 293 aa) (Kainosho et al., 2006); Second WW

domain from mouse salvador homolog 1 protein (mWW45) (PDB 2DWV, BMRB 10028, 98 aa) (Ohnishi et al., 2007).

Each data set contains chemical shift lists, peak lists from 2D and/or 3D ^{15}N-resolved and ^{13}C-resolved NOESY spectra, as well as TALOS-generated angle restraints. Modifications include, amongst others, various percentages of randomly deleted chemical shifts, randomly permuted chemical shifts or randomly deleted NOESY peaks. A detailed description of the proteins as well as the 81 types of data set modifications is given in Chapter 4. Every type of modification was performed five times using a different seed for random number generation, resulting in a total of $10 \times 81 \times 5 = 4050$ data sets that were used to evaluate our new structure calculation protocol. Identical parameter values as for the CASD-data set were chosen for computing the consensus structure bundles.

6.2.5 Analysis of structure calculation results

Evaluation of structure calculation results was mainly based on RMSD values that were calculated with respect to the reference structure (accuracy) and to the mean coordinates of the bundle (precision). RMSD values were calculated for the twenty individual structure bundles, for the combined structure bundle consisting of the lowest target function structure from the 20 individual calculations as well as for the consensus structure bundle based on the consensus set of distance restraints. Only backbone atoms N, C_α, C' in the structured regions of each protein were considered for RMSD calculations.

6.2.6 Structure validation

One of the 20 individual structure calculation results of each of the CASD data sets was validated using the "validate" script of the CYANA software package that calls various validation software tools and summarizes their respective results into one file. Structure validation parameters were computed for one of the twenty individual structure calculation results and the following parameters were chosen: (1) zp-comb-score from the software ProSa2003 (Sippl, 1993). (2) The Verify3D score (Bowie et al., 1991; Lüthy et al., 1992). (3) The clashscore calculated by MolProbity, and the MolProbity score, which considers steric clashes, and Ramachandran plot and staggered rotamer outliers (Chen et al., 2010; Davis et al., 2007; Davis et al., 2004). (4) The packing, the Ramachandran plot appearance, the Chi1/Chi2 rotamer normality, and the backbone conformation quality scores calculated by the WHAT_CHECK program (Hooft et al., 1996). (5) The percentage of residues in the most favored region of the Ramachandran plot (Rama G-factor), and the Chi-1 rotamer normality (Chi-1 G-factor), as defined by the program PROCHECK-NMR (Laskowski et al., 1996; Morris et al., 1992).

6.3 Results and discussion

A schematic overview of the algorithm implemented in the CYANA software package (Güntert, 2009; Güntert and Buchner, 2015) is given in Fig. 6.1. The method first performs 20 independent runs of combined automated NOESY assignment and structure calculation using the standard CYANA automatic structure determination procedure with the same input data. Each run starts from a different set of random structures (Fig. 6.1a), comprises seven cycles and a final structure calculation, and yields a final structure bundle as well as the corresponding set of distance restraints. Because the NOESY peaks are assigned independently in each of the 20 runs, the sets of distance restraints from each run in general differ from each other. From each run, the conformer with the lowest CYANA target function value is collected to form a new bundle of 20 conformers, to which we refer in the following as the combined structure bundle (Fig. 6.1b). A further, crucial step is to combine the individual sets of distance restraints in order to obtain a consensus set of distance restraints including assignments from all individual runs, which is then used to recalculate the final protein structure bundle, which is again composed of 20 conformers (Fig. 6.1c) and to which we refer to as the consensus bundle. This final structure calculation is a simple, standard CYANA structure calculation without automatic NOE assignment. In contrast to the combined bundle, all conformers in the final consensus bundle are optimized against a single, "consensus" set of distance restraints. In the following, we will show that (i) the RMSD radius of the consensus bundle provides a good measure of structural accuracy, and (ii) the conformers of the consensus bundle fulfill the consensus set of distance restraints better than the conformers of the initial, individual runs fulfill their corresponding sets of distance restraints.

We evaluated the new method using NMR data sets of eight different proteins provided by the Northeast Structural Genomics Consortium (NESG) for the CASD-NMR project in 2011-2012 (Rosato et al., 2012; Rosato et al., 2009). Two sets of unassigned NOESY peak lists were available for each protein, one containing manually refined NOESY peaks lists and one containing raw NOESY peak lists from an early stage of spectral analysis. Structure calculations were performed using the standard combined automatic NOESY peak assignment and structure calculation procedure from the CYANA software package and the new method. The CASD-NMR data set was especially well suited for the present study because of the large variability of input data quality and subsequent considerable variety among the different structure calculation results. Despite large differences in structural accuracy, ranging from totally correct to severely erroneous structures, all structure bundles calculated by the standard CYANA approach exhibited a high precision. The test data set thus represents well the aforementioned problem of overestimated bundle accuracy.

In order to evaluate the reliability of a structure calculation result, we calculated the RMSD to the mean structure of the bundle (RMSD radius) representing the precision as well as the RMSD to the reference structure (RMSD bias) as a measure of the accuracy. A structure bundle is considered as reliable if precision and accuracy are in good agreement and the reference structure is thus included in the structure bundle. The structure quality can then be estimated solely based on the bundle precision, which is useful in cases where no reference structure is available. We compare the reliability of protein structures determined by the conventional structure calculation procedure to the results from the new method.

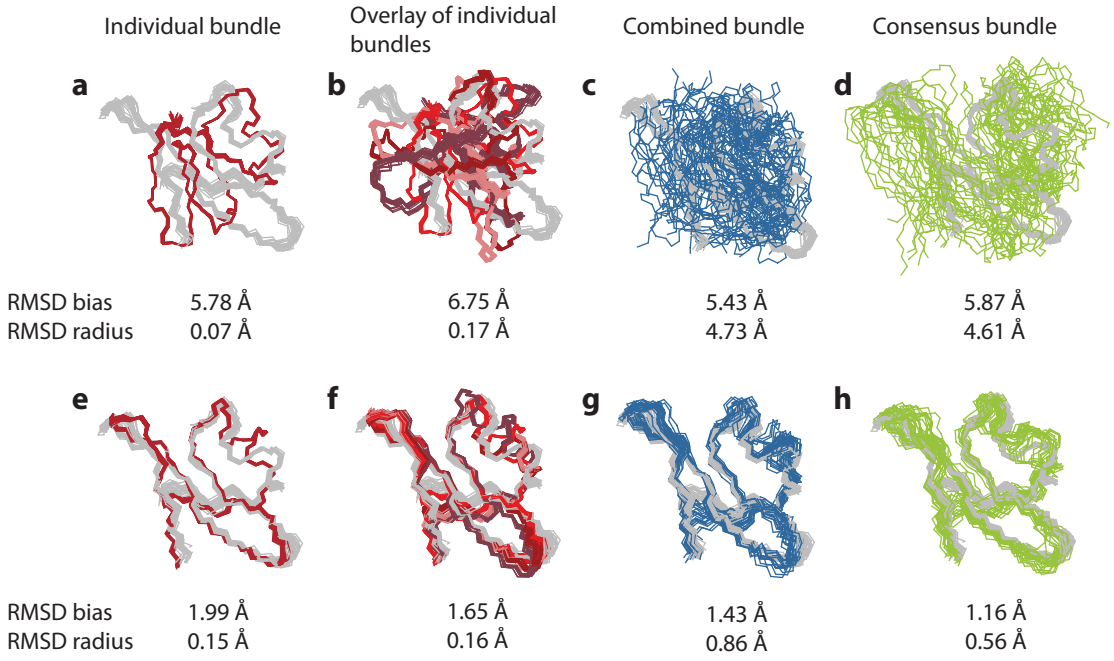


Figure 6.4: Structures of the protein StT322 calculated using the classic automatic structure calculation procedure and our new method. The structure bundle calculated using the classic CYANA automatic NOESY assignment and structure calculation protocol is shown in red (**a**,**e**). An overlay of four independent structure calculation results based on the same input data but different random starting structures using the classic structure calculation protocol is depicted in red (**b**,**f**). The combined structure bundle consisting of the lowest target function structure of each of the twenty individual structure bundles is shown in blue (**c**,**g**). The consensus structure bundle based on the consensus set of distance restraints is presented in green (**d**,**h**). The reference structure is always shown in grey for comparison. Structures were superimposed for optimal fit of the backbone atoms of the ordered residues 23 – 63. **a-d**: raw peak lists were used as input data. **e-h**: Refined peak lists were used as input data.

As an example, Fig. 6.4 shows structures of the protein StT322 (only the ordered parts in the reference structure) calculated from raw peak lists using the standard structure calculation procedure (Fig. 6.4a) as well as an overlay of four selected structure bundles from independent structure calculations using the same procedure (Fig. 6.4b). The reference

structure is presented in grey for comparison. Fig. 6.4a clearly shows the completely incorrect global fold of the protein structure when superimposed onto the reference structure. The error is also reflected by the high RMSD bias with respect to the reference structure of 5.78 Å. Nonetheless, the structure bundle is very tight and well defined (precision measured by an RMSD radius of 0.07 Å), illustrating clearly the misconception of precision being related to structural quality. The overlay of four selected structure bundles (Fig. 6.4b), each being the result of the same structure calculation using different random start structures for the minimization procedure, shows large deviations among the structures, indicating clearly that one individual structure calculation result is not fully representing the data set. The average accuracy is 6.75 Å whereas the average precision is 0.17 Å. The result of our new method is presented in Fig. 6.4c and d. Fig. 6.4c shows the combined structure bundle and Fig. 6.4d the consensus structure bundle based on the consensus distance restraints. The RMSD to the mean structure increases to 4.73 Å and 4.61 Å, respectively, indicating a complete lack of any common structural elements among the individual structures. This result furthermore illustrates well that the recalculated consensus structure bundle closely resembles the combined structure bundle. The RMSD bias remains in both cases similar to the average bias of the individual calculations (5.43 Å and 5.87 Å). Both structure bundles represent well the variety of individual structure calculations depicted in Fig. 6.4b and show that the structure calculation actually did not converge to a unique fold. This example illustrates well that our new method yields a structure bundle where the overestimation of accuracy is dramatically reduced and the precision is a faithful measure of the data quality. These results hold for both, the combined structure bundle and the consensus structure bundle but only the latter is calculated from a single set of conformational restraints such that all its conformers fulfill the same restraints.

The first example of Fig. 6.4a-d is based on experimental input data of very low quality and the incorrect global fold of a single structure calculation could be identified by the majority of validation tools (Fig. 6.5).

Figs. 6.4e-h shows the structure calculation results for the same protein using manually refined peak lists. The individual structure calculation result depicted in Fig. 6.4e shows an overall correct global fold. However, the reference structure is again not included in the structure bundle. The accuracy of the individual structure calculation result is 1.99 Å, indicating a correct global fold but not very accurate local structure. Again the precision of 0.15 Å overestimates the accuracy considerably. In contrast to the aforementioned example, the accuracy is in a range where currently available validation software does not produce reliable results as can be seen from the very limited correlation between the RMSD bias and the validation parameters of different validation software tools (Fig. 6.5).

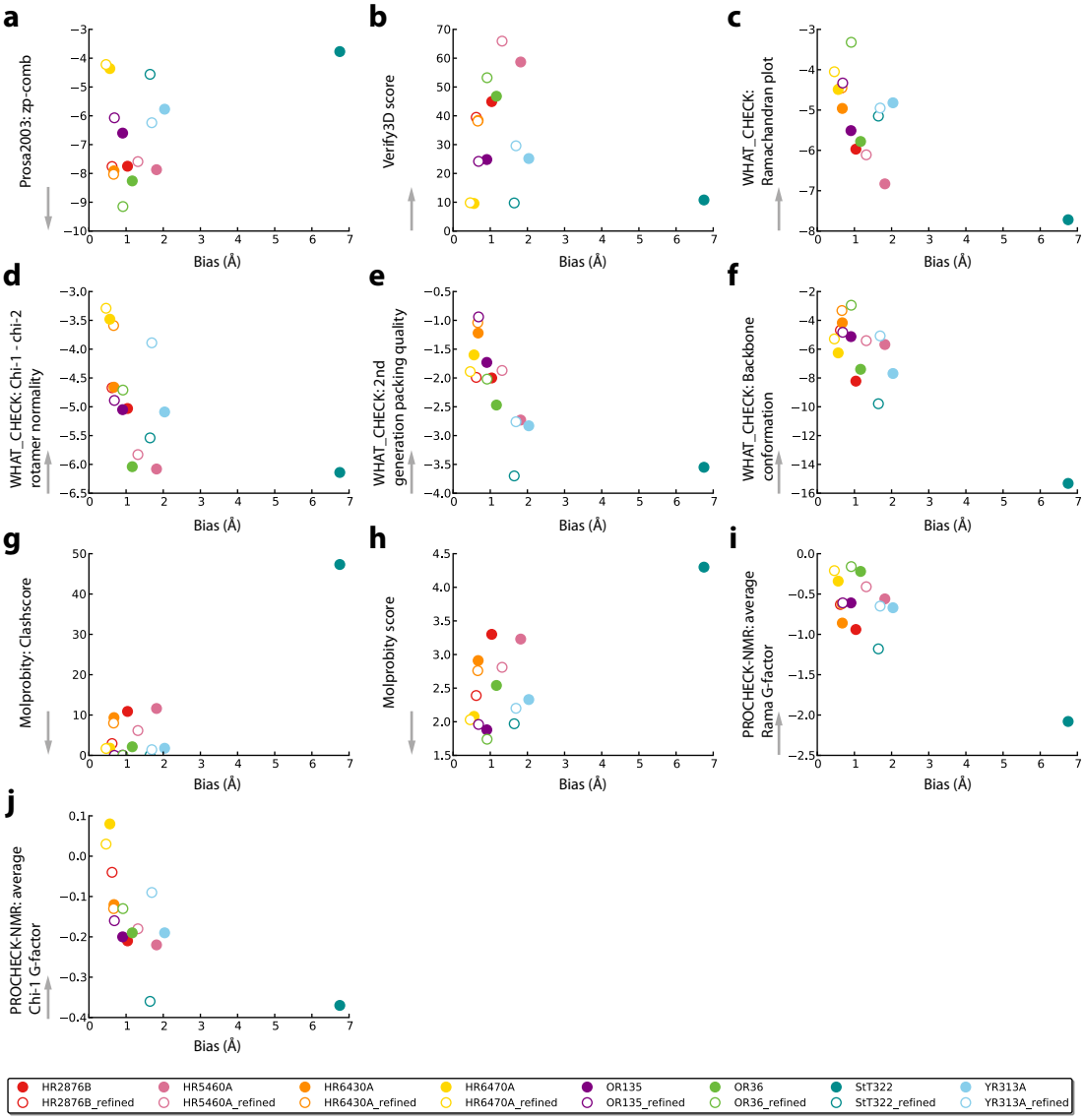


Figure 6.5: Structure validation scores for one representative individual structure bundle of each protein of the CASD data set plotted against the respective RMSD bias. Structures were calculated using the standard combined automated NOE assignment and structure calculation procedure of the CYANA software. The bias was estimated as the RMSD between the mean coordinates of the structure bundle with respect to the reference structure bundle. Only ordered residues were included in the RMSD calculation. The grey arrow at each subfigure indicates the direction to which the respective score value becomes more favorable. Filled dots originate from raw data sets and open circles from refined data sets. The validation scores include: **a** ProSa2003 zp-comb score, **b** Verify3D score, **c-f** WHAT_CHECK: Ramachandran plot appearance, Chi1/ Chi2 rotamer normality, packing, and backbone conformation quality, **g-h** Molprobity: Molprobity score and clashscore (number of serious clashes per 1,000 atoms), **i-j** PRO-CHECK-NMR: percentage of residues in the most favored region of the Ramachandran plot (Rama G-factor), and the Chi1 rotamer normality (Chi-1 G-factor). The value is averaged over all residues in a particular conformer.

The overlay of several structure bundles (Fig. 6.4f) shows significant deviations among the individual structure bundles. All these individual results represent equally well the experimental data when evaluating the final CYANA target function (data not shown). The consensus structure bundle from our new method is depicted in Fig. 6.4h. The accuracy of 1.16 Å is in the same range as the average accuracy of the twenty individual calculations. However, the RMSD radius increases almost four-fold from 0.15 Å to 0.56 Å. Visual inspection as well as the evaluation of RMSD values clearly shows the more complete representation of the experimental data by the new structural ensemble where the reference structure is included in the structure bundle. The consensus structure bundle again represents well the combined structure bundle presented in Fig. 6.4g. The example of Fig. 6.4 illustrates well that the new structure calculation method is beneficial in situations of good and bad input data quality alike since in both cases the reliability can be increased significantly.

In order to evaluate the reliability of all structure calculation results with the different proteins and experimental data sets from CASD-NMR, we calculated the ratio between accuracy (RMSD to the reference structure) and precision (RMSD to the mean structure). Ideally, this "accuracy-to-precision ratio" should be one in order to be able to estimate the structure quality solely based on the bundle precision, values above one indicate that the apparent precision overestimates the accuracy. All RMSD values that were used to calculate the ratios between accuracy and precision are given in Table 6.2. Fig. 6.6 shows the results for the individual structure calculations using the standard approach (red), the combined structure bundle based on twenty individual structure bundles (blue), and the recalculated structure bundle based on the consensus set of distance restraints (green). The results are presented for eight different proteins and two data sets for each protein.

For the conventional individual structure calculations the accuracy-to-precision ratio (averaged over the 20 individual runs with each data set) shows a very large variability in the range between 1.6 and 39.7 among the different proteins and data sets. One protein (HR6470A) has comparatively low ratios in the range between 1.6 and 1.9 and thus represents an exception among these test proteins. Rather low values were also observed for the proteins OR36 (2.2), YR313A (2.4), and HR2876B (2.5) when using refined peak lists. All other proteins show higher ratios indicating a considerable overestimation of accuracy by the bundle precision when using the conventional structure calculation procedure. For most proteins, the ratio decreases when optimizing the experimental input data. However, even when using highly correct input data in the form of manually refined peak lists, the ratios range between 1.6 and 10.3 (average 3.7).

The combined structure bundle contains the lowest energy structure from each individual structure calculation and thus represents a large part of the conformational space that

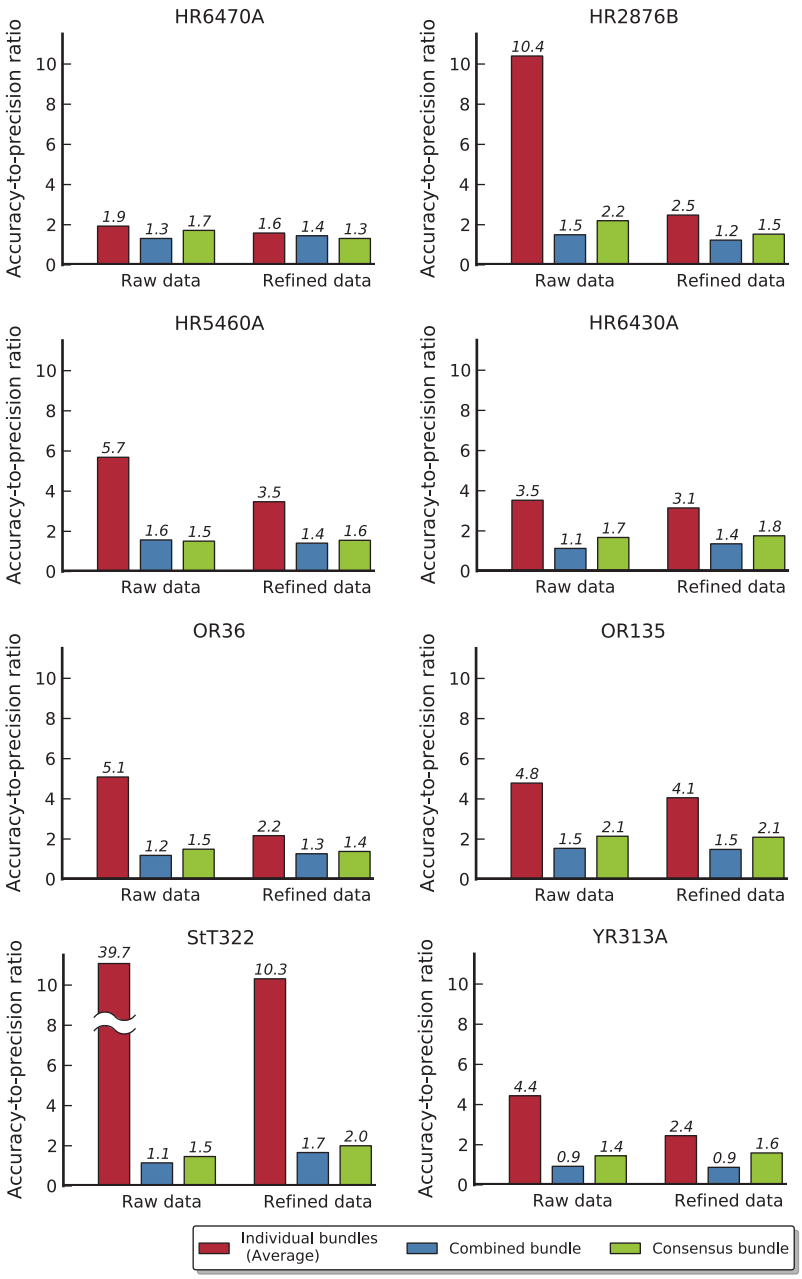


Figure 6.6: Accuracy overestimation by structure bundle precision. The degree of overestimation is quantified by the accuracy-to-precision ratio, where the structural accuracy is given by the RMSD between the mean coordinates of a structure bundle and the NMR reference structure from the PDB, which was determined and refined by experienced scientists, and the precision of a structure bundle is given by the average RMSD of the individual conformers of the structure bundle to its mean coordinates. Only ordered residues (see text) were used for RMSD calculation. The accuracy-to-precision ratio is presented for eight proteins from the CASD-NMR project and two different data sets for each protein (i.e. raw peak lists and refined peak lists). Results are given for the classic structure calculation process as the average from 20 independent runs (red), for the combined structure bundle consisting of the lowest target function conformer from each of the 20 independent runs (blue), and the consensus structure bundle based on the consensus set of distance restraints (green).

Table 6.2: Structure calculation results.

Protein	"Data type"	Individual bundles (Average)				Combined bundle		Consensus bundle		
		RMSD radius cycle1 (Å)	RMSD radius (Å)	RMSD bias (Å)	"Target function (Å²)"	RMSD radius (Å)	RMSD bias (Å)	RMSD radius (Å)	RMSD bias (Å)	"Target function (Å²)"
HR2876B	Raw	0.7	0.1	1.04	7.54	0.58	0.87	0.4	0.88	1.31
	Refined	0.6	0.25	0.62	1.24	0.43	0.53	0.49	0.75	0.52
HR5460A	Raw	3.16	0.32	1.82	13.9	0.98	1.54	0.8	1.21	1.94
	Refined	1	0.38	1.32	9.21	0.8	1.13	0.74	1.15	1.6
HR6430A	Raw	0.52	0.19	0.67	5.84	0.47	0.53	0.55	0.92	4.94
	Refined	0.48	0.21	0.66	5.37	0.42	0.57	0.45	0.79	4.77
HR6470A	Raw	0.48	0.29	0.56	0.32	0.38	0.5	0.39	0.67	0.28
	Refined	0.46	0.29	0.46	0.29	0.29	0.42	0.38	0.5	0.28
OR135	Raw	0.58	0.19	0.91	0.28	0.5	0.77	0.42	0.9	0
	Refined	0.48	0.17	0.69	0.48	0.39	0.58	0.32	0.67	0.01
OR36	Raw	0.98	0.23	1.17	4.59	0.76	0.9	0.63	0.94	0.06
	Refined	0.89	0.42	0.91	0.42	0.64	0.81	0.71	0.98	0.01
StT322	Raw	5.52	0.17	6.75	14.94	4.73	5.43	4.29	6.28	1.16
	Refined	0.8	0.16	1.65	0.52	0.86	1.43	0.6	1.2	0.09
YR313A	Raw	3.44	0.46	2.04	1.4	1.62	1.5	1.07	1.55	0.47
	Refined	1.44	0.69	1.69	0.47	1.52	1.33	0.83	1.32	0.35

The RMSD radius of a structure bundle is the average RMSD value between the individual conformers and the mean coordinates of the bundle. It characterizes the precision. For individual bundles, the first column reports the RMSD radius for the structure obtained in the first cycle of automated NOE assignment and structure calculation, the second column reports the RMSD radius for the final structure bundle. The RMSD bias is the RMSD between the mean coordinates of the structure bundle and a reference structure (or the mean coordinates of a reference structure bundle). RMSD radius and RMSD bias characterize the precision and accuracy of a structure bundle, respectively. RMSD values are calculated for the backbone atoms N, C_α, C' in the structured regions of the protein (see Methods).

can be explained by the given input data set. The ratio between accuracy and precision (Fig. 6.6, blue; values in the range between 0.9 and 1.7) decreases in all cases considerably when compared to the individual calculations. This clearly shows the beneficial effect of repeating the same structure calculation several times, making the calculation result more reliable and enabling the use of bundle precision as direct measure of the structural accuracy. One exception is again the protein HR6470A for which no significant difference is observed in the ratios between the individual structure calculations and the combined structure bundle. This is the only example where a single structure calculation already resulted in an accurate and reliable structure bundle, which could, however, not be recognized if the reference structure is unknown.

The essential part of the new method is the combination of the individual restraint data sets in order to obtain a single consensus set of distance restraints representing the entire conformation space allowed by the input peak lists. The ratios for the consensus structure bundles based on the combined set of distance restraints are shown in Fig. 6.6 (green). In general, the values are very similar to those of the combined structure bundle. This shows that the consensus structure bundle and the corresponding combined set of distance restraints are as well suited to represent the experimental data as the combined structure

bundle. Due to the significantly increased reliability, we recommend to use the consensus structure bundle and the corresponding combined restraints when presenting structure bundles determined by NMR spectroscopy. The precision of the consensus structure bundle can be used as a strong measure of the data quality and thus be compared to the resolution in X-ray crystallography.

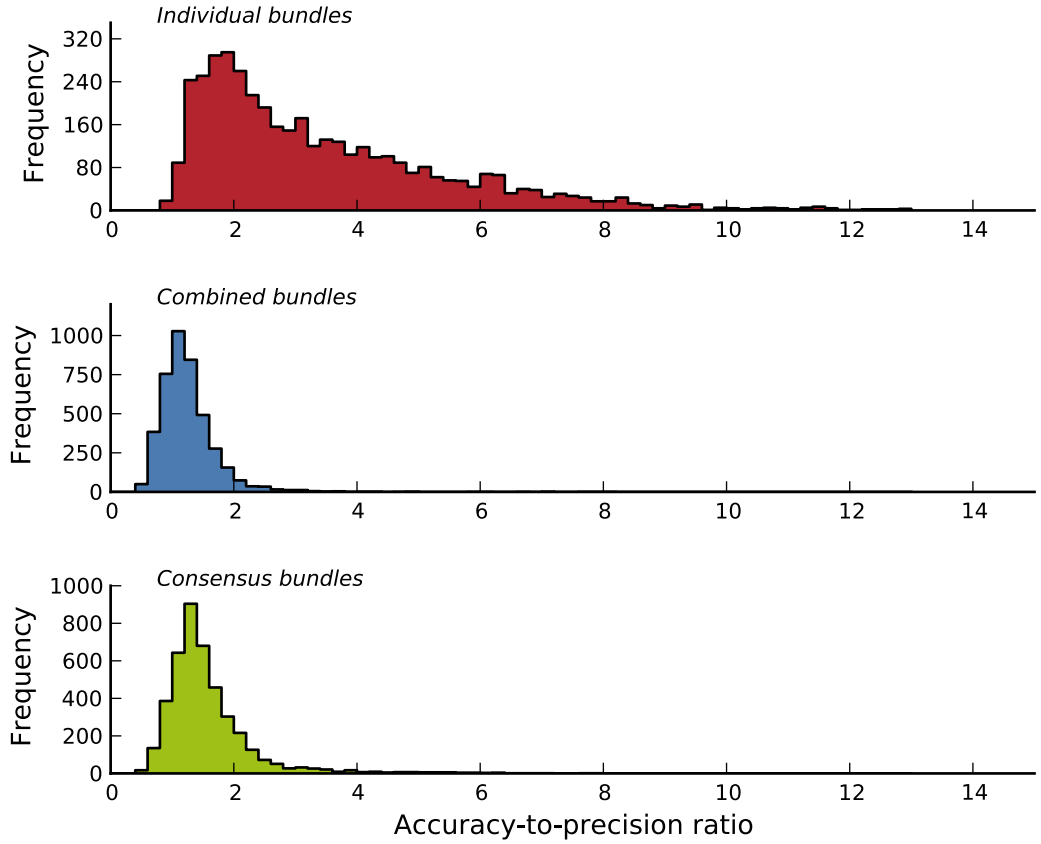


Figure 6.7: Accuracy-to-precision ratios for 4050 NMR data sets. Frequency distributions of the accuracy-to-precision ratios are given for the classic structure calculation process as the average from 20 independent runs (red), for the combined structure bundle consisting of the lowest target function conformer from each of the 20 independent runs (blue), and the consensus structure bundle based on the consensus set of distance restraints (green). The test data is composed of 4050 solution NMR data sets from ten different proteins. It includes for each protein the original experimental data set and modifications that simulate a large variety of data imperfections (see Methods).

In order to investigate the reproducibility, the method was applied to a second test data set comprised of ten proteins including various types of simulated data imperfections for each protein, i.e. randomly deleted chemical shifts, randomly modified chemical shifts, deleted NOESY peaks, etc. These modifications resulted in a total of 4050 restraint data sets covering a very large range of input data quality. A description of the data sets is given in the Experimental Procedures section and more details will be published

elsewhere. For every structure calculation, the overestimation ratio between the RMSD to the reference structure and the RMSD to the mean structure was analyzed and plotted as a histogram for the individual structure bundles (Fig. 6.7, top), the combined structure bundle (Fig. 6.7, center), and the consensus structure bundle (Fig. 6.7, bottom). The histograms show clearly that the overestimation ratios are in general significantly higher in the case of individual structure bundles (i.e. 71 % of the structure bundles have ratios above 2.0) with a median ratio of 2.9, whereas the median ratio for the combined structure bundles is 1.2 (5 % of ratios above 2.0) and for the consensus structure bundles 1.4 (15.8 % of ratios above 2.0). These results are in very good agreement with those from the CASD-NMR data set. Due to the large amount of structure calculations and the large variety of input data qualities resulting in large variations among the calculated structures, it can be concluded that the presented method works reproducibly and can thus be applied routinely.

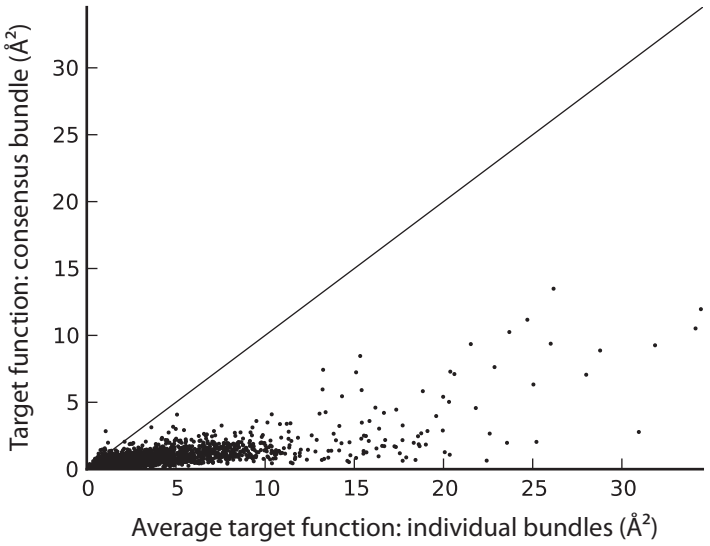


Figure 6.8: Target function values for 4050 conventional and consensus structure bundles. Each data point correlates the average target function value for a consensus structure bundle with the average target function value for the conformers of the corresponding conventional structure bundles. The CYANA target function measures the agreement between the structure bundle and the experimental and steric conformational restraints from which it was calculated. It is defined such that it is zero if all conformational restraints are fulfilled.

The CYANA target function measures the agreement between the structure bundle and the experimental restraints and is defined such that it is zero if all restraints are fulfilled. High target function values indicate problems during the structure calculation and need to be avoided by closer inspection of the experimental data. In order to show that our new method yields distance restraints that are still fulfilled by the recalculated

structure bundle, Fig. 6.8 compares the target function of the consensus structure bundles with the average final target function of the respective 20 individual structure calculations. The target function values of the consensus structure bundles are in almost all cases lower than those of the individual calculations. This shows that no inconsistencies or convergence problems result from using the consensus set of distance restraints.

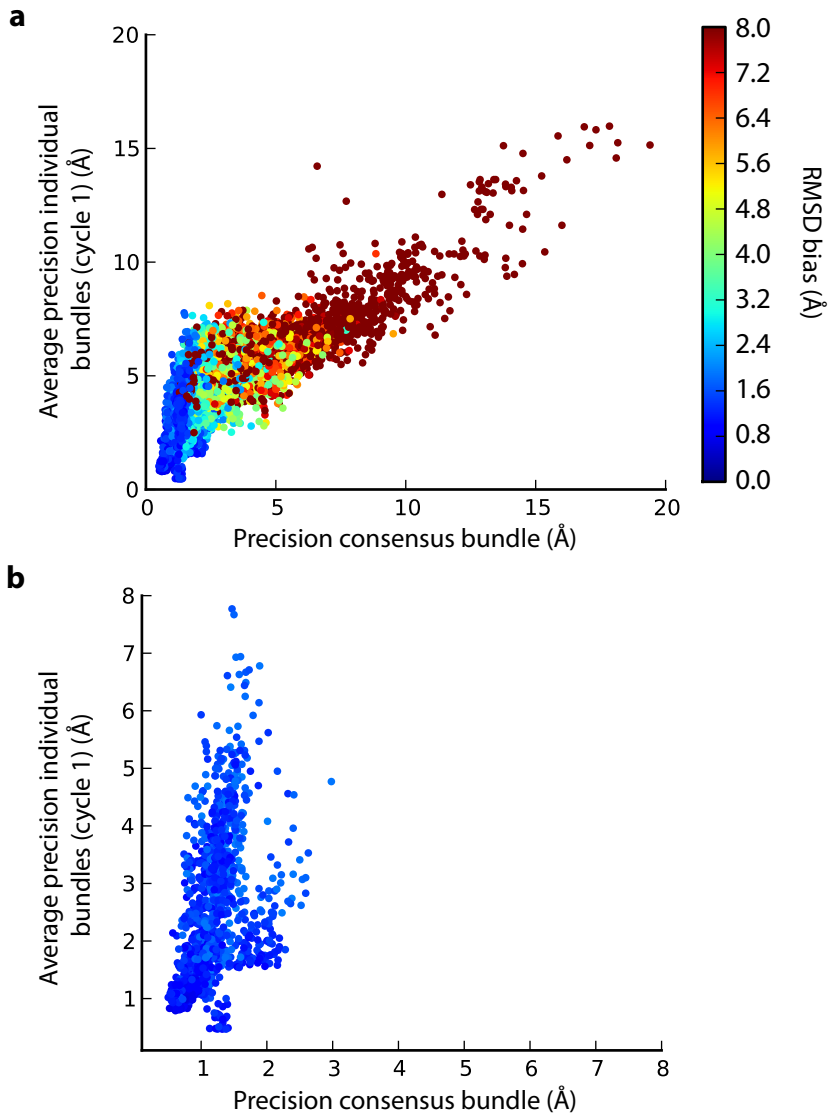


Figure 6.9: Correlation between the average precision of the individual bundles (first structure calculation cycle) and the precision of the consensus bundle. The accuracy of the consensus bundle is indicated by the color. The test data is composed of 4050 solution NMR data sets from ten different proteins. It includes for each protein the original experimental data set and modifications that simulate a large variety of data imperfections (see Methods of the original text). **a** All structure bundles, **b** Structure bundles with an RMSD bias of the consensus bundle below 2.0 Å.

The traditional criterion for evaluating the outcome of a CYANA structure calculation with automated NOE assignment is that an RMSD radius of less than 3 Å in the first cycle of automated NOE assignment and structure calculation is indicative of a final structure with low RMSD bias Herrmann et al., 2002a. Cycle 1 RMSD radii above 3 Å indicate that the resulting final structure may (but doesn't have to) be inaccurate. Therefore, the cycle 1 RMSD radius is not a direct measure of accuracy but rather provides a criterion to recognize potentially unreliable calculations. For comparison, Table 6.2 includes the cycle 1 RMSD radii, and Fig. 6.9 shows the correlation between the RMSD radii of the cycle 1 structure bundles and the consensus structure bundles.

6.4 Conclusion

We have presented a new method for combined automated NOE assignment and structure calculation implemented in the software package CYANA. The principal advantage of our method over simply repeating full calculations (referred to as combined structure bundles in the following) is that all conformers of the consensus structure bundle are calculated from the same restraint data, i.e. the consensus restraint list, in a single CYANA structure calculation. In the case of repeated full calculations, each calculation will lead to somewhat different NOESY peak assignments and restraints. Hence, the resulting structures will in general not fulfill a single set of restraints. This can be problematic if, for example, a combined structure bundle is submitted to the PDB along with the restraints from one of its individual NOE assignment/structure calculations because in this case a later evaluation of the agreement between the coordinates and the conformational restraints in the PDB will in general show additional restraint violations that have not been reported in the original publication. In contrast, we propose to deposit in the PDB the consensus structure bundle together with the consensus restraint list from which the consensus structure bundle was computed. NOESY peak lists containing the consensus peak assignments can also be produced by the program.

We have tested the new method using optimized and raw input peak lists of eight different proteins provided by the CASD-NMR project in 2011 – 2012 as well as a data set based on ten different proteins including various simulated data imperfections. We have measured the reliability of structure bundles as the ratio between accuracy (RMSD to the reference structure) and precision (RMSD to the mean structure) and compared the results from the classical structure calculation procedure to the results from our new method. The results clearly show that the new protocol for automatic structure calculation produces very reliable structure bundles where the precision can be used as a very good indication for the structure quality without having any prior information about the correct protein fold. It should be noted that the precision of the consensus structure bundles is

not strictly equal to the accuracy but proportional with a median proportionality factor of 1.4 (Fig. 6.4). For a conservative estimate, an upper bound on the accuracy, given by the RMSD bias, can be approximated as twice the precision, given by the RMSD radius, of the consensus structure bundle. The precision of the consensus bundle gives an estimate of the input data quality, however, additional criteria such as the assignment completeness of the assigned consensus peak lists as well as the average ambiguity of the latter can be used to assess experimental uncertainties (e.g. a large number of discarded peaks as well as a high ambiguity indicate inconsistencies within the input data).

The new method is helpful for input data optimization in the course of NMR structure determinations, and we recommend it especially for routine use in the final structure calculation, since the consensus bundle reflects the experimental data much better.

6.5 Implementation in CYANA

In order to calculate the consensus set of distance restraints as well as the combined and consensus structure bundle, two new CYANA commands, i.e. `peaks consolidate` and `distances consolidate` were implemented and a new macro `multnoeassign.cya` was written to automate the complete procedure.

6.5.1 CYANA commands

peaks consolidate

The command `peaks consolidate` uses all assigned peak lists from the individual structure calculations and generates consensus peak assignments. Thereby, all peak lists in the CYANA memory need to be of the same type in order to be used for combination and the corresponding chemical shift list needs to be in the memory as well. The consensus peak assignments are available in the CYANA memory after executing the `peaks consolidate` command at the position $n + 1$ whereas n refers to the number of individual calculations. After selecting the consensus peaks using the command `peaks select`, the command `write peaks` can be used to generate a file including the consensus peak assignments.

- mode=string (default=all)

  The value of the parameter `mode` can be chosen between “all” and “consolidated”, whereas “all” refers to the method described in the Methods section (i.e. all assignment possibilities that occur in any of the individual calculations are selected for the consensus peak assignment as long as the peak is assigned to any atom pair in a specified number of individual calculations). The option “consolidated”, in contrast,

chooses only assignment possibilities that occur in at least the specified number of individual calculations.

- cutoff=real (default=0.6)

  The parameter `cutoff` specifies the minimum relative amount of individual calculations in which a peak needs to be assigned in order to keep either all assignments or just those that occur in just as many individual calculations as the cutoff specifies depending on the parameter `mode`.

distances consolidate

The command `distances consolidate` generates a set of consensus distance restraints based on the individual final distance restraints from the individual structure calculations. The distance restraints need to be available to the program prior to executing the command. The resulting consensus distance restraints are stored within the CYANA memory and a file including these restraints can be generated using the CYANA command `write dco`. All distance restraints from the individual calculations are deselected after execution of the `distances consolidated` such that using the command `write dco` automatically generates a file containing only the consensus distance restraints.

- `nlist=integer` (default=20)

  `Nlist` specifies the number of individual structure calculations.

- `cutoff=real` (default=0.6)

  See parameter `cutoff` in `peaks consolidate`.

- `dcotype=string` (default=max)

  The parameter `dcotype` refers to the definition of the upper distance limit for the combined distance restraint. One of the two options “max” and “median” can be chosen, whereas the default value “max” selects the maximum upl-value of all individual values and the value “median” determines the median of all individual values.

6.5.2 Macro

multnoeassign.cya

The macro `multnoeassign.cya` uses the previously introduced CYANA commands for peak list and distance restraint combination in order to calculate the combined and consensus structure bundle. The first step includes a renumbering of peak lists, as it is required that each peak number occurs only once even in multiple peak lists. Unless the option `skipcalcs` is selected, the next step includes a specified number of individual

structure calculations based on the standard CYANA structure determination protocol using automatic NOE assignment via `noeassign.cya`. All parameters available for the macro `noeassign.cya` can as well be used for `multnoeassign`, therefore they will not be listed in the following. The assigned peak lists of the individual structure calculations are combined to obtain consensus peak lists via `peaks consolidate` and the final sets of distance restraints are combined to obtain the consensus distance restraints via `distances consolidate`. A specified number of structures from the final structure bundles of the individual structure calculations is combined to obtain the combined structure bundle. Based on the combined distance restraints as well as other types of restraints if available, a final structure calculation is performed to obtain the consensus structure bundle. The output thus includes (i) consensus peak lists, (ii) consensus distance restraints, (iii) combined structure bundle, (iv) consensus structure bundle, and (v) results of rmsd calculations with respect to a reference structure, if available (`rmsd-combined.txt`, `rmsd-consolidated.txt`.

- `numcalcs=integer` (default=20)

  `Numcalcs` defines the number of individual structure calculations that are performed using the standard CYANA structure determination protocol with automated NOE assignment.

- `numstrct=integer` (default=1)

  `Numstrct` specifies the number of structures from each individual calculation that are used for the combined structure bundle. It is recommended to choose the parameter such that the total number of structures in the combined bundle equals 20.

- `cutoff=real` (default=0.6)

  See parameter `cutoff` in `peaks consolidate` and `distances consolidate`.

- `file=string` (default=final)

  The parameter `file` specifies the name of the distance restraint file used as input for combination. The default value uses the `final.upl` file for combination.

- `skipcalcs`

  The option `skipcalcs` can be selected if the individual structure calculations have already been performed and just the combination of distance restraints as well as the subsequent structure calculation based on the consensus distance restraints is supposed to be repeated.

Part III

Solid-state NMR

Chapter 7

Structure calculations of the model protein GB1 from solid-state NMR data

7.1 Introduction

Solid-state NMR has proven to be a valuable tool to study molecules at atomic resolution which are not amenable to standard structure determination methods such as X-ray crystallography and solution NMR spectroscopy. Among these, amyloid fibrils and membrane proteins in their native phospholipid environment are of special interest for biomedical questions. Thus, great progress in the field was achieved when the first atomic resolution structures of amyloid fibrils or membrane proteins in the lipid bilayer have been released in the past few years (Wasmer et al., 2008; Park et al., 2012; Shahid et al., 2012; Wang et al., 2013; Lu et al., 2013; Schütz et al., 2014). However, in contrast to solution NMR, structure determination from solid-state NMR is still far from routine although many experiments have been developed in order to exploit the structural information which is available due to the anisotropic interactions such as dipolar couplings and CSA (Section 3.2).

The routine use of solution NMR spectroscopy for structure determination of small and medium sized soluble proteins has especially evolved due to the tremendous effort that has been put into the development of methods which automate each step of the structure determination procedure. Automation greatly accelerates the process and makes it more objective due to the independence from user choices. The recent development of more robust chemical shift assignment tools has especially flattened the path towards fully automated structure determination directly based on NMR spectra as input, thus reducing the amount of user interference to a minimum (Schmidt and Güntert, 2012; López-Méndez and Güntert, 2006; Serrano et al., 2012).

Automation of solid-state NMR structure calculation is particularly complicated by the large number of assignment ambiguities originating from peak overlap and low spectral resolution, which is especially severe when using two-dimensional NMR spectra. Higher-dimensional spectra, as they are commonly available from solution NMR spectroscopy, are limited to samples with high signal-to-noise ratios, which mainly depends on the amount of sample in the rotor. Low molecular weight microcrystalline proteins are therefore much more suited for higher-dimensional NMR experiments than large membrane proteins in the phospholipid bilayer. The use of 3D experiments for structure determination has consequently thus far only been reported for the two microcrystalline model proteins SH3 and GB1 (Castellani et al., 2003; Fossi et al., 2005).

Recent development in the field of proton-detection in combination with spin-dilution through perdeuteration and fast MAS may improve the sensitivity sufficiently in order to record multi-dimensional experiments for structure calculation routinely. First results on the structure calculation solely based on 3D and 4D proton-detected experiments of a sparsely-labeled ubiquitin sample have been presented (Huber et al., 2011).

Isotopic labeling schemes represent an additional approach to increase spectral reso-

lution, mainly by reducing the overall number of signals in the spectrum. An additional benefit of reduced isotopic labeling for structure calculation results from the increased number of visible long-range signals through the suppression of high-intensity short-range signals. Patchwork-labeling strategies have been introduced based on 1,3-^{13}C-glycerol and 2-^{13}C-glycerol (LeMaster and Kushlan, 1996; Castellani et al., 2003; Franks et al., 2008) or based on 1-^{13}C-glucose and 2-^{13}C-glucose (Loquet et al., 2012).

Another difference between solid-state NMR and solution NMR structure determination refers to the nature of magnetization exchange for the detection of through-space contacts. Spin diffusion-based experiments are especially popular in solid-state NMR due to their robustness, their comparatively high sensitivity as well as the reduced dipolar truncation effect (Section). The latter allows the measurement of long-range interactions in the presence of short-range contacts, which is not possible when applying for example pure dipolar recoupling sequences for magnetization transfer (Grommek et al., 2006).

Whereas the NOE in solution NMR follows a $1/r^6$-relation between peak intensity and distance, which can be used for the calibration of upper distance limits during structure calculation, there is no such relation which exactly describes the magnetization exchange in spin diffusion-based experiments in solid-state NMR. Consequently, there is no generally accepted procedure on how to determine upl-values from solid-state NMR spectra. Trade-offs include the solution NMR-based approach of distance calibration via $1/r^6$ (Lange et al., 2005; Manolikas et al., 2008), a peak intensity-based manual classification into several distance classes (Castellani et al., 2002; Zech et al., 2005; Zhou et al., 2007; Franks et al., 2008; Wasmer et al., 2008), or the most frequently applied method which uses spectrum type-dependent and mixing time-dependent constant upl-values (Castellani et al., 2003; Fossi et al., 2005; Loquet et al., 2008; Balayssac et al., 2008; Bertini et al., 2010; Shahid et al., 2012). A method to objectively determine the upl-value for a certain peak list was introduced by Melckebeke et al., 2010 on the basis of the structure calculation target function. A step-by-step decrease of the upl-value in distinct structure calculation runs thereby increases the resulting target function exponentially, which can be used as a guidance for the upl-value to be chosen for a certain peak list.

Automated assignment of 2D and 3D ^{13}C-detected spectra in combination with structure calculation has been performed using the solid-state NMR version of ARIA, SOLARIA, and the model protein SH3 (Fossi et al., 2005), as well as using ATNOS-CANDID and the model protein ubiquitin (Manolikas et al., 2008). An initial manual peak assignment resulting in a preliminary structure in combination with a subsequent automated assignment of additional peaks has been reported more frequently (Manolikas et al., 2008; Bertini et al., 2010; Jehle et al., 2010; Shahid et al., 2012; Wang et al., 2013; Tang et al., 2013; Schütz et al., 2014).

One of the most commonly used software tools for combined automated peak assignment and structure calculation in solution NMR is the CYANA software package (Güntert, 2009; Güntert and Buchner, 2015). We have thus used CYANA and a set of 11 two-dimensional solid-state NMR ^{13}C-^{13}C correlation spectra of the microcrystalline protein GB1 to address the following questions: (i) the influence of the input data selection, including the effect of intermolecular signals, on the results of automated peak assignment and structure calculation with the program CYANA, (ii) comparison of the results from automated peak assignment to the results using a manual reference peak assignment, and (iii) comparison of several methods for distance restraint calibration based on the reference peak assignment.

7.2 Experimental and computational methods

GB1 samples

In order to perform NMR measurements, recombinant GB1 was produced via overexpression in *E. coli* growing on M9 medium containing either u-^{13}C-labeled glucose (u-^{13}C/^{15}N GB1), 1,3-^{13}C-labeled glycerol (1,3-^{13}C/^{15}N GB1), or 2-^{13}C-labeled glycerol (2-^{13}C/^{15}N GB1) as sole carbon source, and ^{15}N labeled NH_4 as sole nitrogen source. Purification and crystallization were performed according to a modified version of the protocol introduced in Franks et al., 2005 (Appendix B). Expression from one liter of M9 medium typically yielded 60 mg of GB1 and 30 mg were generally used for crystallization and filled into a 3.2 mm rotor for MAS solid-state NMR experiments. The sample homogeneity was assessed via 1D ^{13}C-spectra. The final data set originated from various different crystallization trials as the quality of an initially homogeneous sample could only be maintained for a limited number of measurements. Sample heating as well as dehydration due to decoupling were assumed to represent at least partly the origin of the quality loss.

In the course of data collection, sample preparation issues of more serious nature prevented the final completion of the data set. Despite tremendous experimental effort, it was not possible to identify and solve the problem within one year and the experimental part of this project was subsequently stopped. Judging from the 1D spectra, the samples were inhomogeneous from the beginning on which led to the assumption that impurities in the crystallization solution caused the inhomogeneity. Attempts thus included, amongst others, the complete exchange of chemicals to ultrapure versions as well as additional purification steps of the protein, for example via ion exchange chromatography. The quality of the soluble protein was furthermore checked via [^{1}H-^{15}N]-HSQC solution NMR spectra which verified a well-folded monomeric protein sample.

NMR data

All GB1 NMR spectra used in the present work were recorded using a Bruker Avance 850WB spectrometer corresponding to a 20 T magnet with a 3.2 mm HCN triple resonance probe or a 3.2 mm HX double resonance probe (Bruker). The temperature was set to 250 K and SPINAL64 (71.6 kHz) was applied for ^{1}H-decoupling during evolution periods. Acquisition times were in the range between 35 and 40 ms in the direct dimension and, depending on the experiment type, between 5 and 20 ms in the indirect dimension. A complete list of experiments and experimental details is given in Table 7.1. All spectra were processed using the processing software NMRPipe (Delaglio et al., 1995) and spectra were subsequently analyzed with CcpNmr Analysis (Vranken et al., 2005).

Table 7.1: NMR Experiments of GB1

Pulse sequence	mixing time [ms]	MAS [kHz]	NS	AQ (F2) [ms]	SW (F2) [kHz]	TD (F2)	AQ (F1) [ms]	SW (F1) [kHz]	TD (F1)	Total time [h]
u-^{13}C/^{15}N GB1										
DARR	500	15	32	9.5	250.0	1024	39.9	299.8	5120	27.30
	800	15	32	14.4	250.0	1536	39.9	299.8	5120	40.96
CHHC	0.2	15	4x64	6.4	86.3	256	39.9	299.8	5120	54.60
	0.4	15	2.5x64	7.6	86.3	320	35.8	299.8	3838	42.70
PAR	3	19	16	6.3	177.7	480	35.8	299.8	3838	6.40
	6	19	32	6.3	177.7	480	35.8	299.8	3838	12.80
	9	19	64	6.3	177.7	480	35.8	299.8	3838	25.60
1,3-^{13}C/^{15}N GB1										
DARR	500	15	16	10.5	250.0	1127	39.9	299.8	5120	15.00
	800	15	32	14.4	250.0	1536	39.9	299.8	5120	40.96
PAR	15	19	32	22.4	80.0	768	34.9	299.8	4480	20.48
2-^{13}C/^{15}N GB1										
DARR	500	15	32	12.4	250.0	1330	39.9	299.8	5120	35.50

Generation of peak lists

Signals were manually picked into three different lists for each spectrum, in the following referred to as *peak list classes*. The first list contains broad and overlapped peaks, the second list well resolved peaks and the third list contains weak peaks below a chosen threshold. Overlapped and well resolved peaks were chosen subjectively. Peak maxima were determined by the signal identification function of CcpNmr. A Python script was written in order to automatically determine and remove signals arising from magic angle spinning sidebands, which can originate from diagonal peaks as well as from strong cross peaks. Sideband peaks were identified solely based on peak positions such that the difference between two peak positions equals an integer multiple of the MAS rate in at least one dimension within a given tolerance of 0.2 ppm. The MAS rate was converted to ppm for this purpose.

Table 7.2: Total number of peaks

Pulse sequence	mixing time [ms]	Number of peaks total	well resolved	overlapped	weak
u-^{13}C/^{15}N GB1					
DARR	500	2874	371	2142	361
DARR	800	3537	286	3071	180
CHHC	0.2	183	94	30	59
CHHC	0.4	538	170	213	155
PAR	3	333	132	174	27
PAR	6	459	161	245	53
PAR	9	747	222	457	68
13-^{13}C/^{15}N GB1					
DARR	500	2860	622	1752	486
DARR	800	3674	686	2605	383
PAR	15	407	161	189	57
2-^{13}C/^{15}N GB1					
DARR	500	2365	757	1279	329

Table 7.3: Number of intra-/intermolecular peaks

Pulse sequence	mixing time [ms]	Number of peaks total	well resolved	overlapped	weak
u-^{13}C/^{15}N GB1					
DARR	500	2267/607	298/73	1645/497	324/37
DARR	800	2860/677	249/37	2445/626	166/14
CHHC	0.2	147/36	74/20	22/8	51/8
CHHC	0.4	395/143	126/44	156/57	113/42
PAR	3	265/68	109/23	137/37	19/8
PAR	6	360/99	135/26	186/59	39/14
PAR	9	579/168	173/49	351/106	55/13
13-^{13}C/^{15}N GB1					
DARR	500	2359/501	553/69	1373/379	433/53
DARR	800	3042/632	600/86	2093/512	349/34
PAR	15	321/86	131/30	148/41	42/15
2-^{13}C/^{15}N GB1					
DARR	500	1859/506	622/135	969/310	268/61

In addition to MAS sidebands, microcrystalline samples such as GB1 usually possess intermolecular signals arising from atom pairs belonging to different molecules. These peaks potentially generate incorrect distance restraints if they are not recognized as such by the automated assignment program. In order to investigate their influence on structure calculation results, we have additionally created peak lists lacking intermolecular

peaks. This was achieved by simulation of all possible intermolecular signals based on a PDB structure of GB1 including five molecules in the correct microcrystalline symmetry. The structure was determined by Nieuwkoop and Rienstra, 2010 based on intermolecular restraints from TEDOR measurements (Nieuwkoop and Rienstra, 2010; PDB 2KWD). Peaks were simulated for every intermolecular atom pair with a distance of less than 8 Å using the GB1 chemical shift assignment, which allows the prediction of the peak position of a given atom pair. All peaks from the manually prepared peak lists with a corresponding peak in the set of simulated intermolecular peaks within a tolerance of 0.2 ppm were removed. An overview of all generated peak lists is given in Tables 7.2 and 7.3.

Structure calculations using automated peak assignment

All structure calculations were performed using the CYANA 3.96 software. Input peak lists for combined automated peak assignment and structure calculations were prepared as described in the previous section. The chemical shift assignment was used from the BMRB (BMRB entry 15380) and ^{13}C shifts were adapted to the sample and spectrometer based on a 5 ms 2D DARR spectrum of GB1. Individual chemical shift files were generated for peak lists from diluted 1,3-^{13}C/^{15}N GB1 and 2-^{13}C/^{15}N GB1 samples by removing all chemical shifts of those atoms that are to a very large extent unlabeled (Fig. 3.2). Angle restraints were generated using the TALOS+ software (Shen et al., 2009) based on backbone N, C_α, and CO chemical shifts. The assignment tolerance was set to 0.3 ppm for well resolved and weak peaks and to 0.5 ppm for overlapped peaks. Calibration of peak intensities into upl-values was performed using a $1/r^6$-condition where the upl-value corresponding to the median intensity (*dref*-value) was set to 8.5 Å for each peak list. According to the new protocol for combined automated peak assignment and structure calculation introduced in Chapter 6, twenty independent structure calculations were carried out for each combination of input data and the RMSD with respect to the reference structure (RMSD bias) was calculated for the combined structure bundle. The RMSD bias was furthermore averaged over five such calculations based on different seeds for random number generation. The reference structure was generated applying the regularization method introduced by Gottstein et al., 2012a to the GB1 crystal structure (PDB 2QMT).

Generation of a reference peak assignment

The reference peak assignment was generated by performing one cycle of automated peak assignment using the reference PDB as input structure from the previous cycle, thus excluding all assignment possibilities that are not in agreement with the correct structure. In order to exclude assignment possibilities based on the distance criterion, the distance in the structure of the preceding cycle is compared to the upl-value of the corresponding

peak and violations above a cycle-dependent cutoff lead to the exclusion of the respective assignment. However, as mentioned earlier, upl-values obtained from solid-state NMR spectra are rather inaccurate, which complicates the correct exclusion of assignments based on the distance criterion as it is applied normally. The difficulty arises from the fact that, depending on the mixing time, atom pairs with distances in the range between 2 Å and 12 Å can in theory generate signals in the spectrum. This would require a high upl-value in order to keep correct assignments for signals that arise from atom pairs separated by a distance of 12 Å. In the case of very large upl-values, however, many incorrect assignments are kept for peaks that actually arise from short-range atom pairs. In order to solve this problem, a CYANA function called *elasticity* was applied. This function is based on the assumption that upl-values are not always correct and a stepwise increase of the respective upl-value is performed, up to a specified maximum percentage of the original value, in case no proper assignment can be found for the original upl-value, but at least one assignment fulfills the new, increased upl-value. Based on the fact that transfer efficiency decreases with distance, the assignment with the shortest distance is most likely the correct one. Consequently, the upl-value was set to a rather small value of 5.0 Å for each peak and the elasticity was set to 3.0, allowing the upl-value to be increased stepwise up to a maximum of 15.0 Å (i.e. favoring assignments of atom pairs that are separated by a small distance, however, allowing assignments corresponding to atom pairs of larger distance in case no suitable assignment can be found otherwise). Network anchoring was switched off by setting the parameter *pathlength=0*. The assignment tolerance was set to 0.3 ppm for well resolved and weak peaks and 0.5 ppm for overlapped peaks.

Structure calculations based on the reference peak assignment

If input peak lists are assigned, CYANA calibrates distance restraints using all available assignments for a given peak and calculates the respective upl-value based on the peak intensity or uses a constant upl-value, if specified. For the calibration of peak intensities, the user defines the *dref*-value which represents the upl-value corresponding to the median intensity. Based on distance restraints generated in this way, structure calculations were performed using 100 random starting structures and 10,000 simulated annealing steps. The 20 lowest energy structures were combined into the final structure bundle. The structure quality was assessed based on the RMSD bias.

Calibration of distance restraints using L-shaped curves

The calibration method was originally introduced by Melckebeke et al., 2010 to determine the optimum upl-value for a given peak list. It is based on the fact that the final CYANA target function at the end of a structure calculation is a measure of how well the input

restraints are fulfilled by the final structure bundle. In order to find the optimum upl-value for a peak list, several upl-values in a range which depends on the type of experiment and the mixing time are applied and a structure calculation is performed. The final CYANA target function is plotted against the corresponding upl-value. The resulting curve typically shows an exponential increase for decreasing upl-values, which is the reason for the denomination *L-shaped curve*. The increase of the CYANA target function can be rationalized by the inability to converge to a structure bundle that fulfills all structural restraints if the upl-values are too small. According to the original method, the final upl-value for the respective peak list is chosen as the distance corresponding to the last increase of the CYANA target function which is less than 0.5 $Å^2$ (i.e. the kink of the L-shaped curve). If the data set is composed of several peak lists, then one L-shaped curve is generated for every peak list individually while the upl-value of every other peak list is set to a large value of 14.5 Å.

7.3 Results and discussion

7.3.1 NMR data

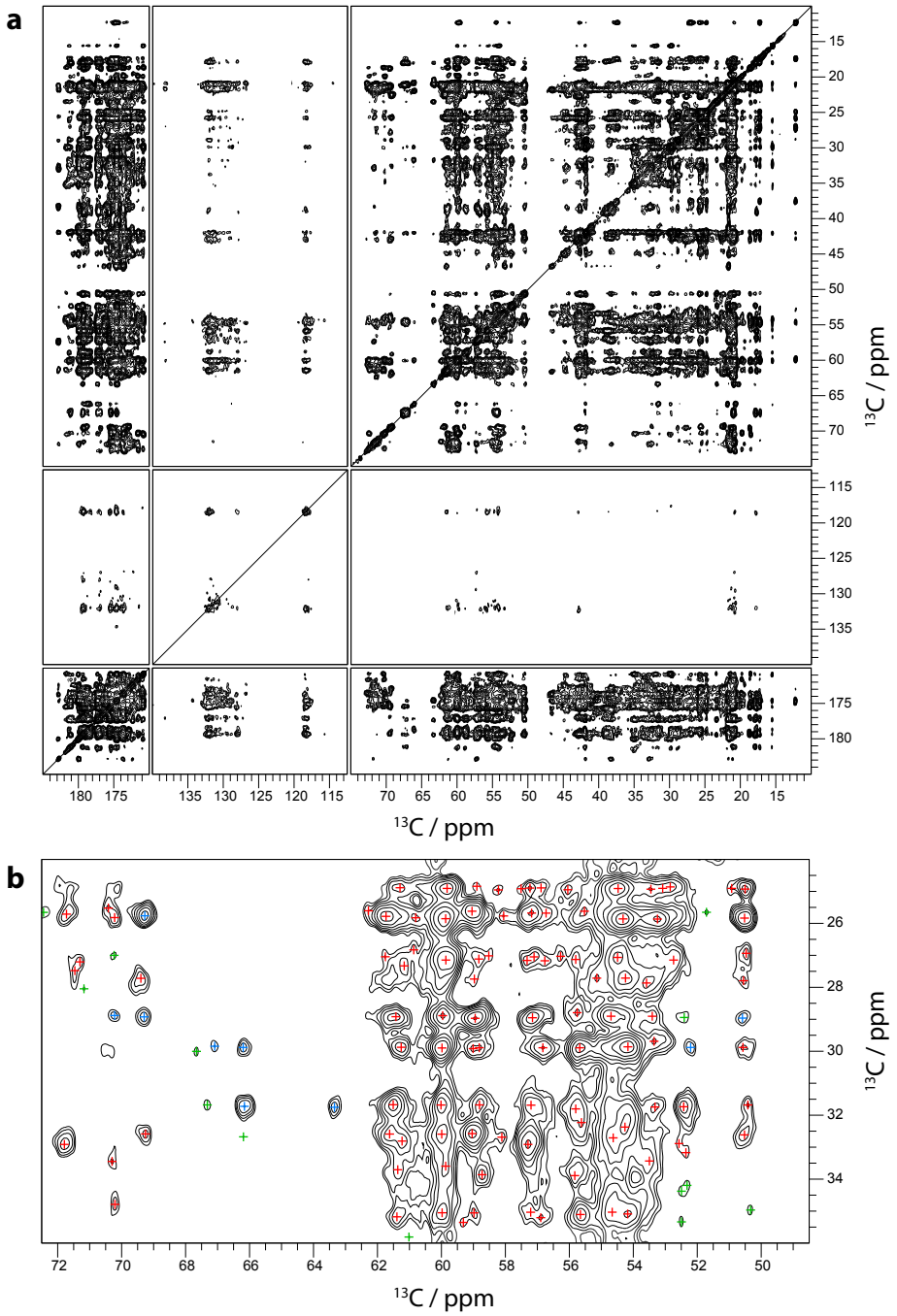


Figure 7.1: DARR spectrum of u-$^{13}C/^{15}N$ GB1 at 500 ms mixing time. **a** Graphical representation of the complete spectrum. **b** Enlarged section of the spectrum showing selected peaks which correspond to three different peak classes: Well resolved peaks (blue marks), weak peaks (green marks), and overlapped peaks (red marks).

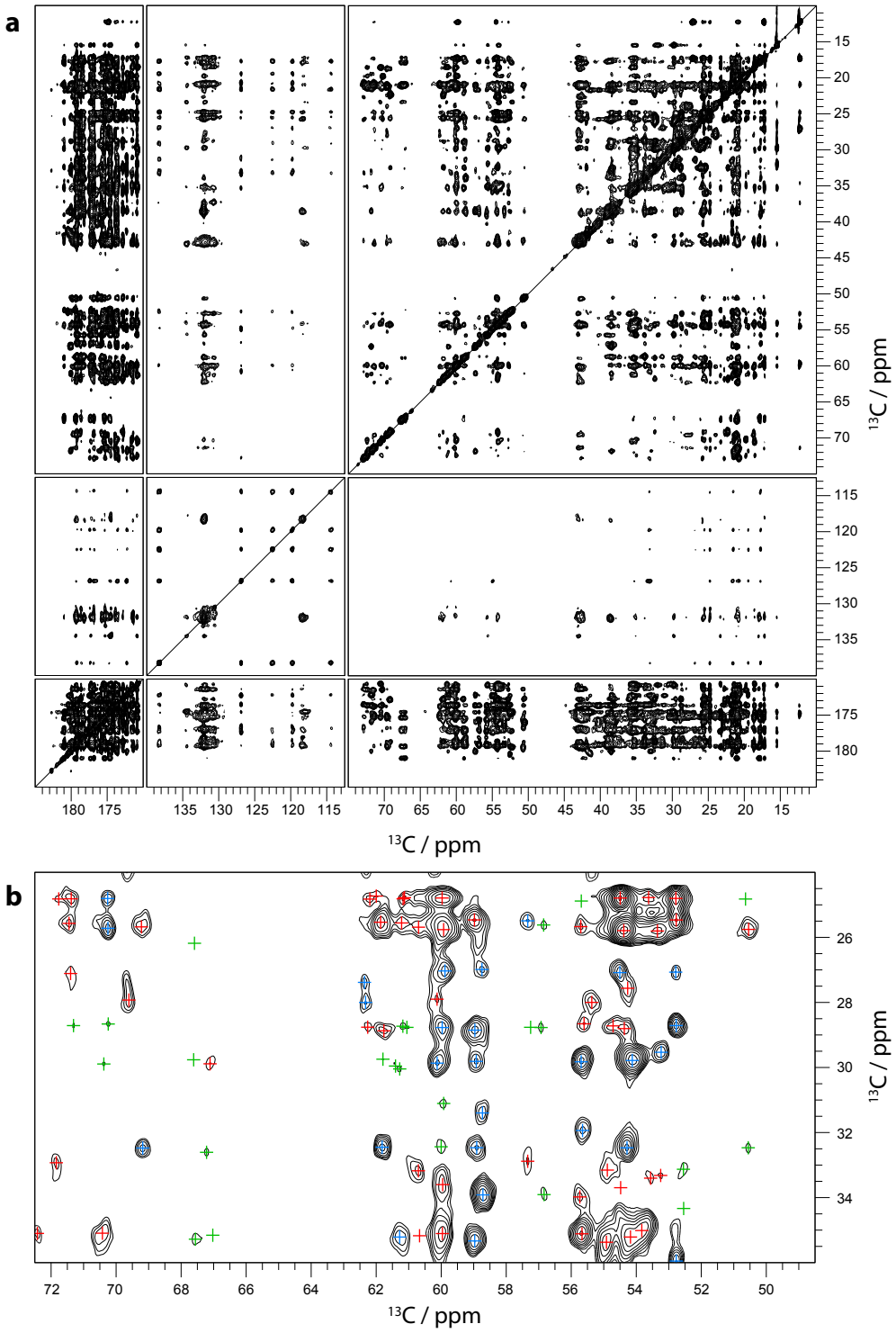


Figure 7.2: DARR spectrum of 1,3-^{13}C/^{15}N GB1 at 500 ms mixing time. **a** Graphical representation of the complete spectrum. **b** Enlarged section of the spectrum showing selected peaks which correspond to three different peak classes: Well resolved peaks (blue marks), weak peaks (green marks), and overlapped peaks (red marks).

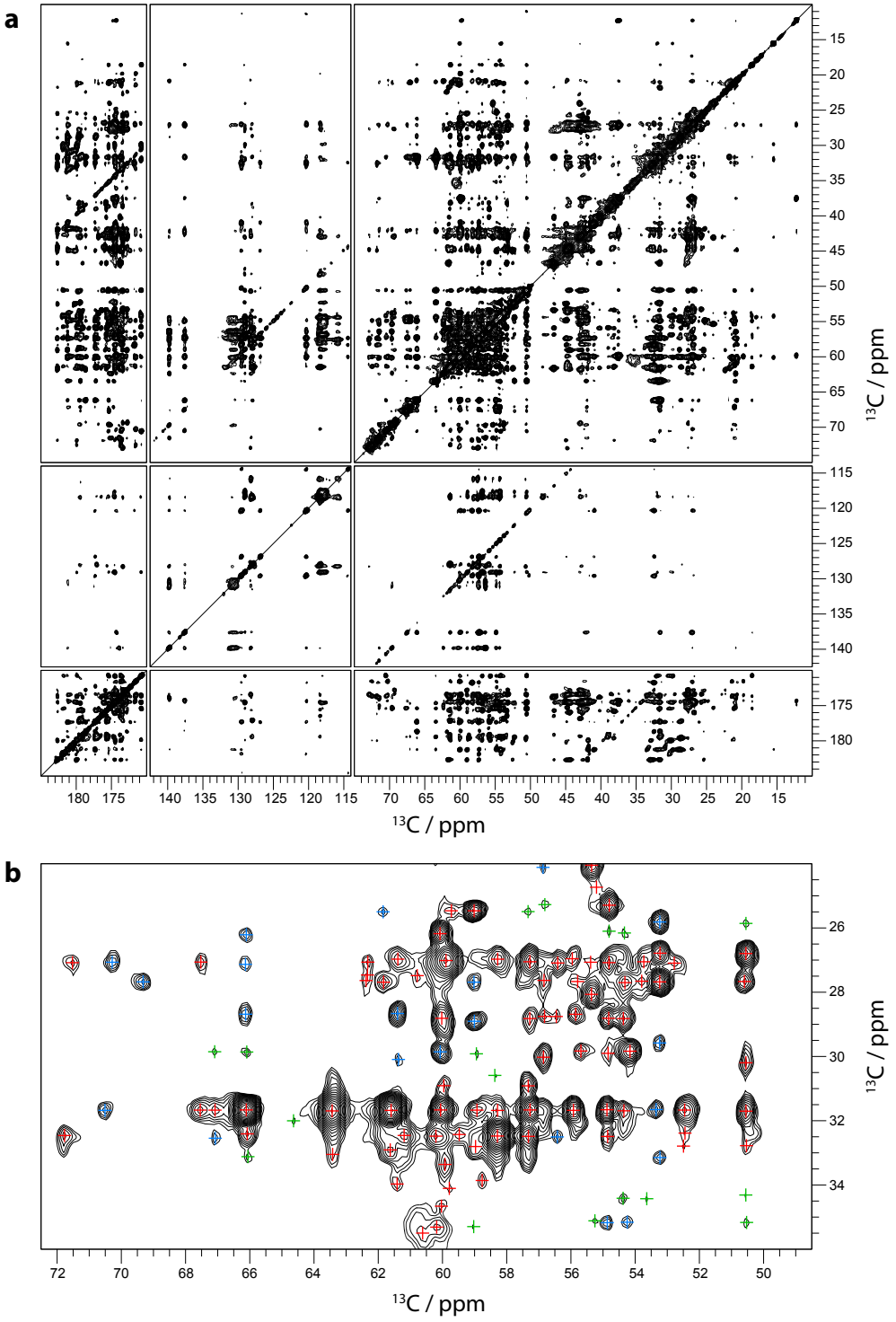


Figure 7.3: DARR spectrum of 2-^{13}C/^{15}N GB1 at 500 ms mixing time. **a** Graphical representation of the complete spectrum. **b** Enlarged section of the spectrum showing selected peaks which correspond to three different peak classes: Well resolved peaks (blue marks), weak peaks (green marks), and overlapped peaks (red marks).

The NMR data set of the protein GB1 used for structure calculations in the following sections includes several types of two-dimensional ^{13}C-^{13}C correlation spectra (i.e. DARR, CHHC, and PAR) at different mixing times recorded at 850 MHz ^{1}H Larmor frequency (Table 7.1). Three differently labeled samples based on ^{13}C-labeled glucose (u-^{13}C/^{15}N GB1), 2-^{13}C labeled glycerol (2-^{13}C/^{15}N GB1) as well as 1,3-^{13}C labeled glycerol (1,3-^{13}C/^{15}N GB1) were used for data acquisition.

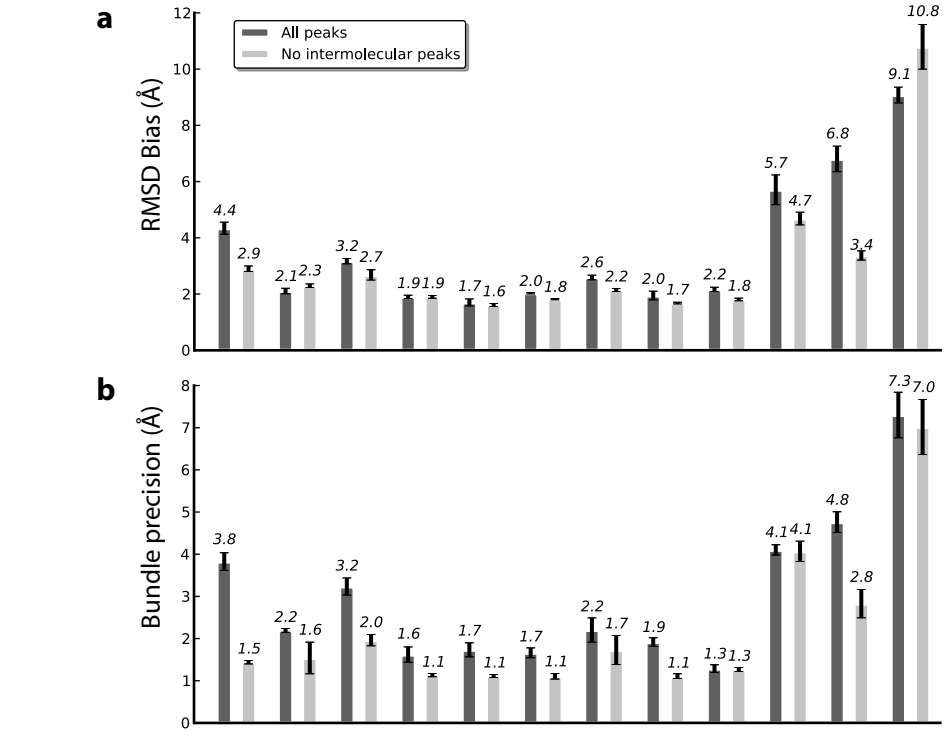
DARR spectra at 500 ms mixing time are presented for all three samples in Figs. 7.1a, 7.2a, and 7.3a, respectively, in order to demonstrate the impact of diluted labeling strategies on the spectral quality. Signals in the present set of GB1 spectra generally possess a rather narrow line width (~0.3 – 0.5 ppm) which is characteristic for microcrystalline samples with high molecular order and little dynamics. The narrow line width in combination with the small size of GB1 (56 amino acids) produces well resolved spectra. Glycerol-labeling furthermore increases the resolution as the number of labeled atoms is reduced.

NMR signals were identified automatically using the CcpNmr software package and manually separated into three different *peak list classes* for each spectrum containing (i) broad and overlapped peaks, (ii) well resolved peaks, and (iii) weak peaks with an intensity below a chosen threshold. Examples for peaks from different peak list classes are shown in Figs. 7.1b-7.3b.

7.3.2 Automated peak assignment and structure calculation for different selections of input peak lists

All structure calculations in this section were performed according to the new protocol for automated peak assignment and structure calculation introduced in Chapter 6 of the present work. According to this protocol, twenty independent structure calculations have been conducted for each combination of input peak lists and the combined structure bundle was used for subsequent analysis. Structural accuracy was evaluated using the RMSD to the reference structure (RMSD bias). Unless stated otherwise, a set of input peak lists from one sample and one peak list class includes all types of NMR spectra that were recorded on the respective sample.

Results are summarized in Fig. 7.4 (dark grey bars) in the form of RMSD bias (Fig. 7.4a) and bundle precision calculated as the average RMSD to the mean structure of the bundle (Fig. 7.4b). Several combinations of input peak lists exist where structure calculations yield the correct global fold, indicated by RMSD bias values below 2.5 Å. In nearly all cases, the results improve when overlapped peaks are excluded. Overlapped peaks in general possess an increased uncertainty with respect to peak positions and peak intensities. This potentially leads to erroneous peak assignments and upl-values when calibrating peak



u-^{13}C/^{15}N GB1													
well resolved							x	x	x	x	x	x	
weak							x	x		x	x		
overlapped							x			x			
13-^{13}C/^{15}N GB1													
well resolved	x	x	x	x	x	x	x	x	x				
weak	x	x		x	x		x	x					
overlapped	x			x			x						
2-^{13}C/^{15}N GB1													
well resolved				x	x	x	x	x	x				
weak				x	x		x	x					
overlapped				x			x						

Figure 7.4: Structure calculation results for different combinations of input peak lists using automated NOE assignment. "x" indicates that the corresponding peak list class has been used in the respective structure calculation. Four blocks comprising three consecutive columns can be separated: 1. NMR spectra from 1,3-^{13}C/^{15}N GB1 (column 1–3), 2. NMR spectra from 1,3-^{13}C/^{15}N GB1 and 2-^{13}C/^{15}N GB1 (column 4–6), 3. NMR spectra from all three samples (column 7-9), and 4. NMR spectra from u-^{13}C/^{15}N GB1 (column 10–12). Within each block, the first column represents calculations using all three peak list classes generated from each spectrum, the second column shows results lacking overlapped peaks and the third column shows calculations that were performed using only well resolved peaks. Automated peak assignment and structure calculation was performed according to the new protocol introduced in Chapter 6 of the present work. Five calculations were performed for each combination of input peak lists using a different random number generation seed. *Dark grey bars*: structure calculations based on the complete set of peaks, and *light grey bars*: results of the corresponding peak list combinations excluding intermolecular peaks. **a** Accuracy calculated as the RMSD to the reference structure, **b** Bundle precision calculated as the RMSD to the mean structure.

intensities. The influence of peak overlap on the intensity calibration is especially problematic since the peak intensities are overestimated, resulting in an underestimation of upl-values, which in turn can cause distortions during structure calculation. These facts explain the overall improvement when leaving out overlapped signals. However, the exclusion of overlapped peaks is only beneficial in case of a sufficient amount of long-range information among the remaining signals. Using spectra from only uniformly labeled GB1 constitutes one example where the amount of long-range information is not sufficient and the removal of overlapped peaks does consequently not improve the result. Similar findings are expected for the majority of solid-state NMR spectra of larger proteins or less ordered proteins that lack the required amount of resolution, especially when using two-dimensional spectra.

Leaving out weak peaks does in general result in a decrease of structural quality. Weak peaks have a higher risk of being artifacts or noise, however, they also have a high chance of containing important long-range information. Artifacts and random noise peaks can in many cases be recognized by the automated assignment algorithm, since the peak position has either no matching chemical shift or the low network anchoring score discards the respective peak. This is one potential explanation why the beneficial effect outweighs the risk of introducing errors. The result of a structure calculation is in general a trade-off between the overall amount of meaningful long-range information and the number of erroneous restraints coming from spurious peaks. Consequently, the selection of input peak lists is important for the outcome of a structure calculation. Calculations from uniformly labeled GB1 alone do not converge to the correct global fold (RMSD bias values above 2.5 Å) although the number of peaks exceeds that of the other two samples. The large number of spectra from uniformly labeled GB1 cannot outweigh the problems arising from the uniform labeling, the most severe of them being the large number of short-range signals that overlap with the relevant long-range signals.

Diluted labeling schemes (e.g. based on 1,3-^{13}C glycerol and 2-^{13}C glycerol) were proposed in the literature to circumvent this problem. The beneficial effect of using diluted samples can be confirmed according to the presented results. Using spectra from only 1,3-^{13}C/^{15}N GB1 instead of uniformly labeled GB1 reduces the RMSD bias from 5.7 Å to 4.4 Å when including all peak list classes and can be further improved to 2.1 Å when omitting overlapped peaks. Diluted labeling affects the spectral quality in the following ways: on the one hand, the reduced number of labeled atoms limits the overall number of observed signals which in turn improves the spectral resolution and consequently the accuracy of peak positions and peak intensities and on the other hand, the number of visible meaningful long-range signals increases due to the reduction of short-range signals. Adding peak lists from 2-^{13}C/^{15}N GB1 further improves the result to 1.9 Å when using

all peak list classes and to 1.7 Å when leaving out overlapped peaks. This improvement results from the fact that a different selection of atoms is labeled in these two samples and their spectra provide additional information, whereas different spectra from the same sample largely contain redundant information.

Using all peak list classes from all three samples slightly decreases the quality of the resulting structures to 2.6 Å (and to 2.0 Å when leaving out overlapped peaks). This indicates that spectra from uniformly labeled GB1 predominantly introduce errors instead of providing valuable additional information. The reason for the negative influence on the structural quality is most likely the increased number of errors in peak assignments and upl-values originating from a higher degree of inaccuracy concerning peak positions and peak intensities.

These results show that it is in principle possible to fully automatically calculate protein structures from solid-state NMR in a quality range where the global fold is correct. The accuracy is in good agreement with that reported in the literature from other model systems using similar approaches (Table 3.1). To the best of our knowledge, this is the first example of combined automated NOE assignment and structure calculation without manual interference which reproducibly yields decently accurate structures based on two-dimensional solid-state NMR data.

However, in comparison to the results typically obtained from solution NMR data, the outcome is rather unsatisfactory, considering the small system size and high spectral resolution. One major difference between solid-state NMR spectra of microcrystalline samples and solution NMR spectra is the presence of intermolecular signals. These signals are potentially harmful since their positions in the spectrum are not random as it is the case for other sources of artifact signals. Consequently, there will most likely be atoms whose chemical shifts match the peak positions and the assignments will have rather good network-anchoring scores. However, as the algorithm only generates intramolecular assignments, the resulting distance restraints will have a distorting effect during structure calculation. In order to investigate their influence on the structure calculation results, we have repeated all previously described structure calculations with peak lists lacking intermolecular signals (Fig. 7.4, light grey bars).

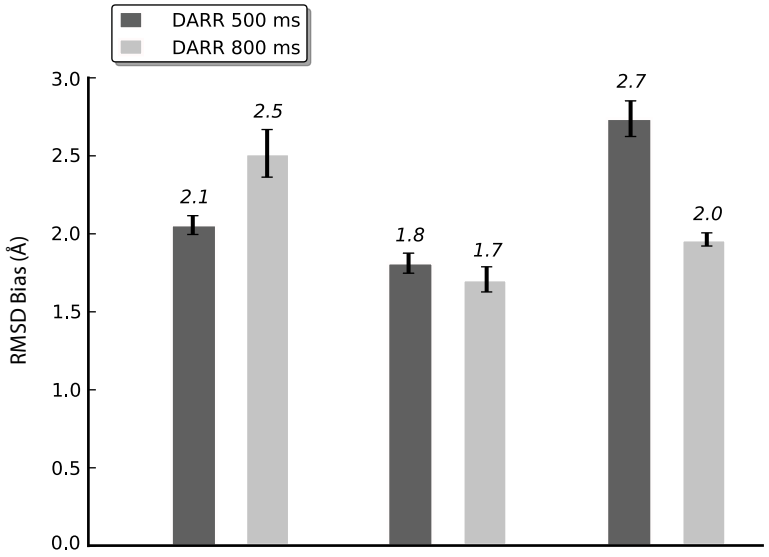
In most cases, the exclusion of intermolecular signals slightly improves the structure calculation result. This improvement is especially apparent in cases where the structure calculation using all signals completely fails to converge to the correct global fold, whereas those calculations with relatively high accuracy show less improvement. In none of the cases, it was possible to obtain very high quality structures with RMSD bias values below 1.0 Å. It can therefore be concluded that intermolecular signals are not the main reason for the limited accuracy of protein structures determined by solid-state NMR. However, since

their removal does not negatively influence the structural quality, either, the experimental identification of intermolecular peaks can be considered. It should, however, be noted that the detection of intermolecular peaks requires additional NMR experiments using diluted samples (i.e. fully labeled protein mixed with unlabeled protein), which generally requires longer measurement time due to the overall reduced number of spins. In this work, intermolecular signals could be identified without additional experimental effort since the 3D structure of the model protein GB1 is known.

The test data set used in the previous structure calculations included a total of 11 spectra recorded on three differently labeled samples. This required several weeks of measurement time. The previous structure calculations nicely demonstrated the highly beneficial effect of using spectra from differently labeled diluted samples. It was, however, not tested whether it is beneficial to use different types of carbon-carbon correlation experiments of the same sample. One of the most simple, sensitive and robust solid-state NMR experiments is the DARR experiment. The present data set includes DARR experiments recorded with two different mixing times (500 ms and 800 ms) for all three labeled samples, with the exception that the 800 ms spectrum is missing for 2-^{13}C/^{15}N GB1. The following structure calculations use the same structure calculation procedure as the previous calculations, but peak lists from only one DARR spectrum (either 500 ms or 800 ms) of each sample are used (Fig. 7.5, 500 ms dark grey bars, 800 ms light grey bars) instead of all peak lists that are available for each sample. For 2-^{13}C/^{15}N GB1, the 500 ms spectrum is used in all calculations.

Despite the significantly smaller total number of peaks compared to using peak lists from all available spectra (45 % in case of 500 ms and 53 % in case of 800 ms mixing time), the correct global fold is found in nearly all cases. Best results are obtained when leaving out overlapped peaks and including weak peaks (RMSD bias of 1.8 Å for 500 ms mixing time and 1.7 Å for 800 ms mixing time). This result is in good agreement with the previous calculations. When using high quality data of a model protein such as GB1, the information content of DARR spectra from differently labeled samples is consequently sufficient to define the global fold and additional spectra as they were used in the previous calculations do not improve the structural quality. It should, however, be noted that the available GB1 DARR spectra are characterized by very high resolution (due to the high order and small system size), allowing the identification of a large amount of signals. In other cases, it may be necessary to record additional types of spectra in order to increase the information content, although the majority of signals are redundant in the different types of ^{13}C-^{13}C correlation experiments. Using the shorter mixing time of 500 ms performs slightly better when using all peak list classes, whereas the spectra recorded with longer mixing time yield better results when using only well resolved peaks. In practice,

the mixing time strongly determines the signal-to-noise ratio and very long mixing times should only be considered in cases of very high sensitivity.



u-^{13}C/^{15}N GB1			
well resolved	x	x	x
weak	x	x	
overlapped	x		
13-^{13}C/^{15}N GB1			
well resolved	x	x	x
weak	x	x	
overlapped	x		
2-^{13}C/^{15}N GB1			
well resolved	x	x	x
weak	x	x	
overlapped	x		

Figure 7.5: Structure calculation results based on DARR spectra of three differently labeled samples using automated peak assignment and structure calculation. DARR spectra were recorded with two different mixing times (500 ms and 800 ms) for u-^{13}C/^{15}N GB1 and 1,3-^{13}C/^{15}N GB1 and with 500 ms for 2-^{13}C/^{15}N GB1. "x" indicates that the corresponding peak list class has been used in the respective structure calculation. The first column represents calculations using all three peak list classes generated from each spectrum, the second column shows results lacking overlapped peaks and the third column shows calculations that were performed using only well resolved peaks. Automated peak assignment and structure calculation was performed according to the new protocol introduced in Chapter 6 of the present work. Five calculations were performed for each combination of input peak lists using a different random number generation seed. The accuracy was calculated as the RMSD to the reference structure (RMSD bias). *Dark grey bars*: structure calculations based on peak lists from the 500 ms DARR spectra of all three samples, *light grey bars*: structure calculations based on peak lists from the 800 ms DARR spectra of u-^{13}C/^{15}N GB1 and 1,3-^{13}C/^{15}N GB1 and the 500 ms DARR spectrum of 2-^{13}C/^{15}N GB1).

7.3.3 Structure calculation results using the reference NOE assignment

In contrast to the available two-dimensional solid-state NMR data set, solution NMR data typically include three-dimensional NOESY spectra which simplifies the assignment of NOE peaks in the following ways: firstly, the increased resolution results in a higher accuracy of peak positions, and secondly, additional dimensions reduce the number of assignment possibilities. In order to investigate the general influence of assignment inaccuracies, following the lack of higher-dimensional solid-state NMR spectra, on the structural quality, we have generated a correct reference peak assignment for the available set of peak lists. This reference assignment was used for all subsequent structure calculations. These calculations were consequently not carried out using the combined automated peak assignment and structure calculation algorithm, but using a simple structure calculation based on distance restraints originating from the assigned peak lists, instead. Calibration of peak intensities into upl-values was performed based on the standard solution NMR approach using a $1/r^6$-relation between distance and peak intensity. Structure calculation input data included the same combinations of input peak lists as they were used for the automated peak assignment in the previous section. Results are summarized in Fig. 7.6.

In contrast to the automated peak assignment method, all combinations of input peak lists result in the correct overall global fold when using the reference peak assignment as input (Fig. 7.6a) and it appears that the result improves with increasing amount of input data, measured as the number of long-range restraints (Fig. 7.6b). Consequently, in contrast to the results obtained from the automated peak assignment, leaving out overlapped peaks has no beneficial effect. We attribute this to the fact that all assignment possibilities that violate the reference structure have been discarded during the generation of the reference assignment. This has eliminated the disturbing influence of overlapped peaks during the automated peak assignment and enabled a gain of information instead. The RMSD bias optimum of 1.4 Å was thus obtained if all available peak lists have been included in the calculation.

The structure calculation result could, however, not be significantly improved when compared to the automated peak assignment, although the calculation was carried out under nearly perfect circumstances (e.g. reference peak assignment, sufficient amount of input data). As assignment inaccuracies can be excluded through the use of a reference peak assignment, another potential explanation for the accuracy limitation is the calibration of distance restraints. Upl-values have been calibrated using the standard solution NMR approach (i.e. $1/r^6$-relation between distance and peak intensity) in all structure calculations presented so far. The calibration constant is typically estimated such that the median peak intensity corresponds to a distance referred to as *dref*-value. The default *dref*-value for solution NMR NOESY peak lists is 4.0 Å. Spin diffusion-based solid-state

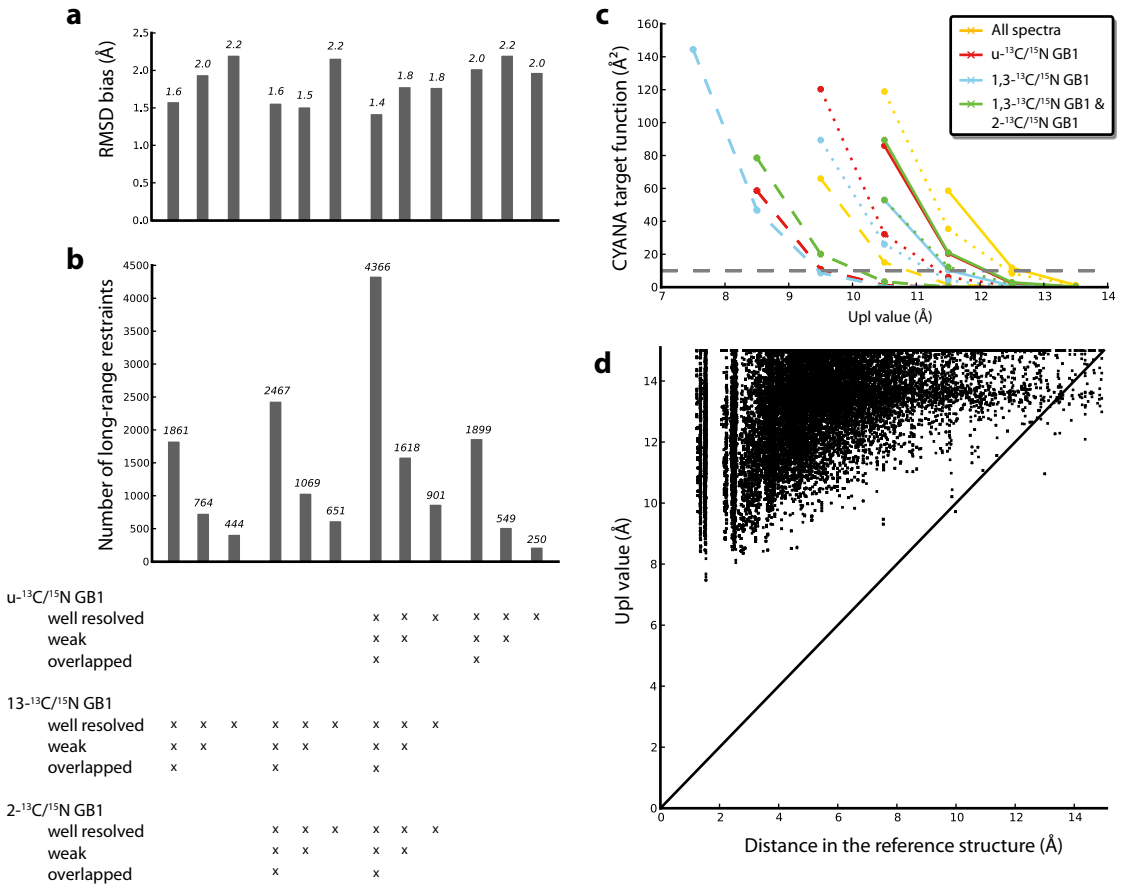


Figure 7.6: Structure calculation results for different combinations of input peak lists using the reference peak assignment. "x" indicates that the corresponding peak list class has been used in the respective structure calculation. Four blocks comprising three consecutive columns can be separated: 1. NMR spectra from 1,3-^{13}C/^{15}N GB1 (column 1–3), 2. NMR spectra from 1,3-^{13}C/^{15}N GB1 and 2-^{13}C/^{15}N GB1 (column 4–6), 3. NMR spectra from all three samples (column 7-9), and 4. NMR spectra from u-^{13}C/^{15}N GB1 (column 10–12). Within each block, the first column represents calculations using all three peak list classes generated from each spectrum, the second column shows results lacking overlapped peaks and the third column shows calculations that were performed using only well resolved peaks. Structure calculations were performed using a simple CYANA structure calculation based on distance restraints obtained from a $1/r^6$-calibration of the assigned peak lists. **a** RMSD with respect to the reference structure (RMSD bias). **b** Number of long range restraints. **c** L-shaped curves for the determination of the *dref*-value used for distance restraint calibration. *solid:* all peak list classes, *dashed:* no overlapped peaks, *dotted:* only well resolved peaks. **d** Correlation between upl-values and the respective distances in the reference structure for all peak lists (corresponding to column 7 in subfigures a and b).

NMR experiments at long mixing times allow magnetization transfer through much larger distances, which increases the required *dref*-value. It should be chosen such that the resulting upl-values cause as little violations of the reference structure as possible (i.e. high enough) while retaining the highest possible information content (i.e. as small as possible). The optimum *dref*-value for every combination of input data was determined empirically based on the analysis of L-shaped curves (Fig. 7.6c). The *dref*-value corresponding to the

first target function value below 10 Å^2 was selected. In contrast to the method originally proposed by Melckebeke et al., 2010, one *dref*-value for the complete set of peak lists was determined instead of one individual value for each peak list.

The correlation between the resulting upl-values and the corresponding distance in the reference structure is shown for the complete set of peak lists in Fig. 7.6d. Each dot represents one distance restraint. The information content of a distance restraint increases with decreasing upl-value, however, values smaller than the "true" distance (i.e. below the diagonal) cause distortions during the structure calculation and should thus be avoided. In order to minimize the amount of violations, the *dref*-value is chosen sufficiently high. In contrast to the NOESY experiment in solution NMR, whose spin dynamics can be well described by the Solomon-equations, there is no suitable theory which describes the spin dynamics during solid-state NMR spin diffusion experiments in a sufficiently accurate way. The experimental peak intensity and the distance thus have no perfect $1/r^6$-relation as it is assumed for calibration and the resulting upl-values thus do not correlate well with the distance in the true structure (Fig. 7.6d). This represents one potential explanation for the limited structural accuracy observed for the thus far presented structure calculations. The following section therefore presents a systematic evaluation of all available methods for distance restraint calibration that have been proposed in the literature. The general aim is to increase the information content while keeping the amount of violations as low as possible.

7.3.4 Evaluation of different distance restraint calibration methods

Constant upl-values

Since peak intensities and distances do not follow the $1/r^6$-relation as it was assumed for calibration in the previous sections, the following structure calculations investigate the effect of neglecting the relation between peak intensity and distance completely and using a single constant upl-value for each peak list instead. This procedure has been used in several structure calculations from solid-state NMR presented in the literature (Table 3.1). Upl-values are thereby usually chosen based on the experiment type and mixing time (i.e. longer mixing times allow magnetization transfer through larger distances, thus making larger upl-values necessary). Two different approaches are compared. The first approach uses the same upl-value for each peak list, independent of the experiment type and mixing time, in analogy to the method used in the previous section, where the same *dref*-value was used for calibration of each peak list. The second method determines an individual upl-value for each peak list.

For the first approach, a single upl-value of 14.5 Å was determined for the complete set of available peak lists through the analysis of the corresponding L-shaped curve. The

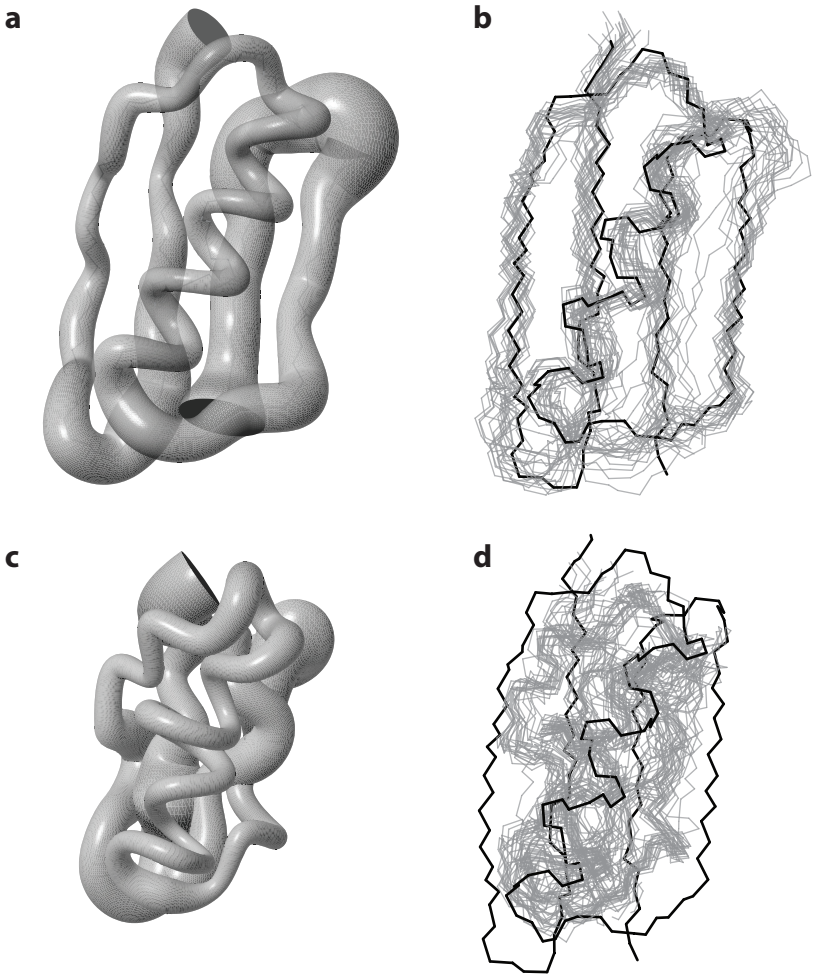


Figure 7.7: Structure calculation result displayed in sausage representation (a,c) and bundle representation (b,d) based on distance restraints obtained from the reference peak assignment. **a**,**b** Upl-values were set to 14.5 Å according to the analysis of the respective L-shaped curve. Backbone RMSD 1.5 Å and RMSD bias 1.6 Å. **c**,**d** Upl-values were set to 7.5 Å. Backbone RMSD 1.3 Å and RMSD bias 3.2 Å. The reference structure used for RMSD bias calculation is a regularized version of the GB1 X-ray structure (PDB 2QMT) and is displayed in black for comparison (b,d).

upl-value was selected as proposed in the original protocol by Melckebeke et al., 2010 as the last distance for which the increase of the CYANA target function was less than 0.5 Å^2. The RMSD bias of the resulting structure bundle (Fig. 7.7a,b) is 1.6 Å which is slightly higher but still in the same range as for the calibration of upl-values using the classic relation of $1/r^6$ (1.4 Å). In both cases, the bundle RMSD to the mean structure of 1.5 Å indicates a rather loose bundle and the precision is furthermore in the same range as the RMSD bias. This indicates that the structural inaccuracy is the result of a lack of information rather than the result of distorting distance restraints. If distorting distance restraints are the cause of structural inaccuracies, while the information content of the

data set is sufficiently high, it is expected that the precision is much smaller than the accuracy.

The information content is generally affected by the number of non-redundant long-range distance restraints as well as their individual capability of restraining the structure (i.e. a distance restraint with a higher upl-value carries less information as the same restraint with lower upl-value). Due to the large amount of distance restraints in the data set, the low information content does not result from the number of available distance restraints, but rather from their large upl-value of 14.5 Å that was chosen based on the L-shaped curve in order to minimize the number of violations.

The upl-value of 14.5 Å greatly overestimates the true distance for the majority of distance restraints and a more realistic value for most of the restraints would be ~7.5 Å. We have thus repeated the same structure calculation with a constant upl-value of 7.5 Å. The consequence is an overall distorted structure bundle (Fig. 7.7c,d) that violates a large number of structural restraints, indicated by a large final average CYANA target function value of 3085.5 $Å^2$. The RMSD bias of 3.2 Å is in good agreement with the distortions observed in the structure bundle. This indicates that a lower information content resulting from overestimated upl-values has less severe consequences on the structure calculation result as compared to violations obtained through underestimated upl-values.

The structure calculation using 14.5 Å as upl-value shows that it is possible to obtain the correct global fold of the protein GB1 without distortions if the upl-value is chosen sufficiently high in order to avoid violations. As a result of the low information content, the precision of the resulting structure bundle is, however, rather low. It is therefore necessary to compensate the loss of information as a consequence of the high upl-value by a sufficient number of distance restraints. Decreasing the upl-value to 7.5 Å increases the information content, but introduces a large number of violations, which greatly distorts the resulting structure bundle. In order to further improve the structure quality, it is thus necessary to increase the information content without introducing violations.

The second approach aims to increase the amount of information by taking the type of experiment and mixing time into consideration. This enables the decrease of the upl-values of individual peak lists. The method proposed by Melckebeke et al., 2010 was applied in order to determine the upl-value of every peak list based on the analysis of individual L-shaped curves (Fig. 7.8a,c,e) for every peak list. The corresponding upl-values are summarized in Fig. 7.8b,d,f. As expected, the resulting upl-values increase with increasing mixing time of the same experiment type (e.g. Fig. 7.8a, light and dark blue for the CHHC experiment at 200 and 400 μs). The upl-values of DARR-experiments are generally larger than those of CHHC and PAR experiments. This can be explained by the fact that the DARR experiment is purely spin diffusion-based, allowing magnetization to

spread over larger distances. As magnetization is exchanged via protons instead of carbon atoms in the CHHC experiment, the distance for magnetization transfer is further reduced as compared to the DARR experiment.

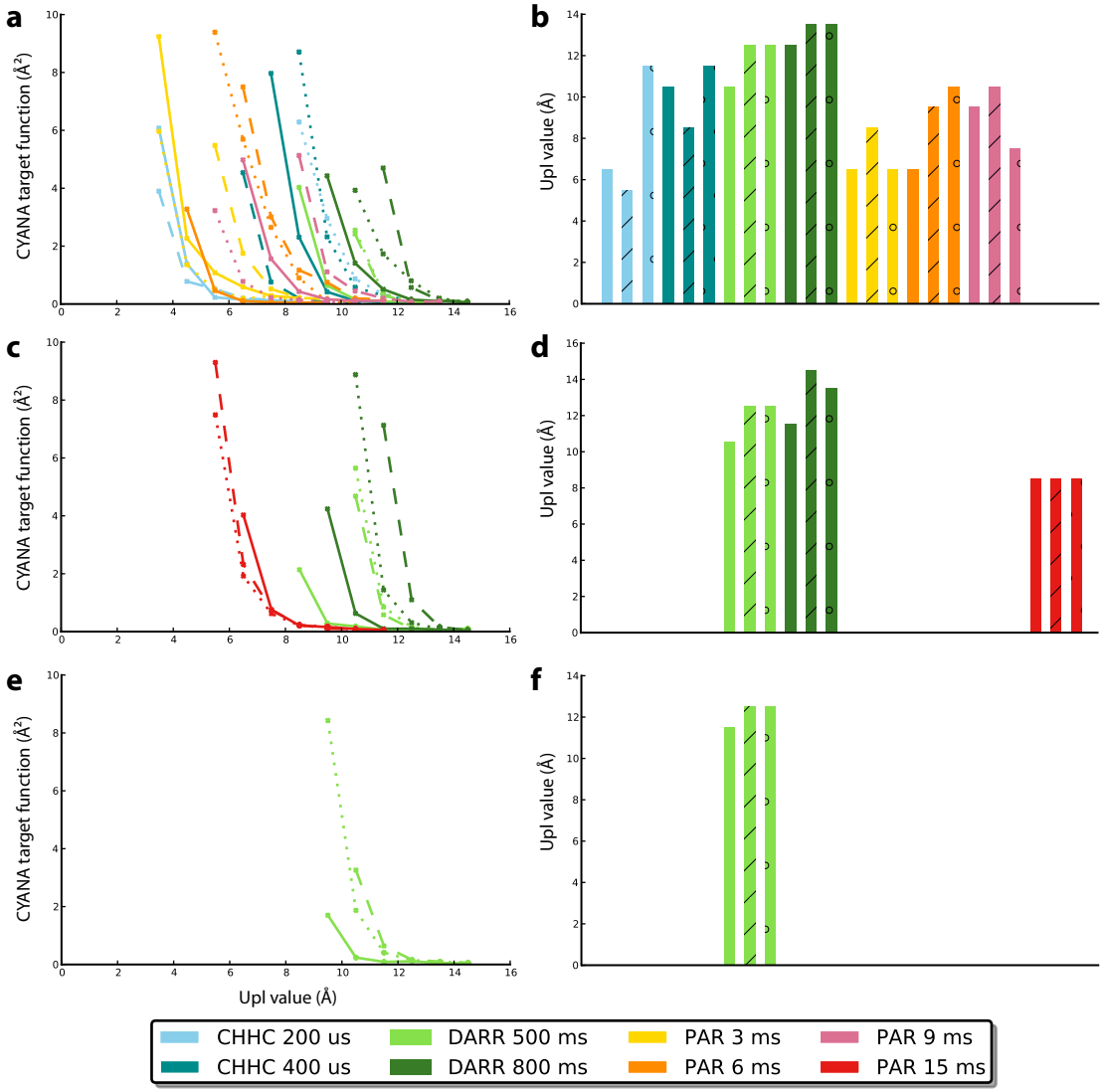


Figure 7.8: Determination of individual upl-values for every peak list of the GB1 data set. ***Left side:*** L-shaped curves for each peak list. *Solid:* all peak lists, *dashed:* no overlapped peaks, and *dotted:* only well resolved peaks. ***Right side:*** Upl-values for each peak list determined based on the L-shaped curves on the left side. The value corresponding to the last distance for which the increase in target function for decreasing upl-values is less than 0.5 Å^2, is selected as the final upl-value for the respective peak list. *Solid fill:* all peak lists, *striped fill:* no overlapped peaks, and *dotted fill:* only well resolved peaks. **a,b** u-^{13}C/^{15}N GB1 peak lists, **c,d** 1,3-^{13}C/^{15}N GB1 peak lists, **e,f** 2-^{13}C/^{15}N GB1 peak lists. Missing curves/bars arise from experiments that were not recorded for the respective sample.

Applying the respective upl-value of each peak list for an additional structure calculation, yields a structure bundle with an RMSD bias of 2.4 Å(Fig. 7.9c). This represents a

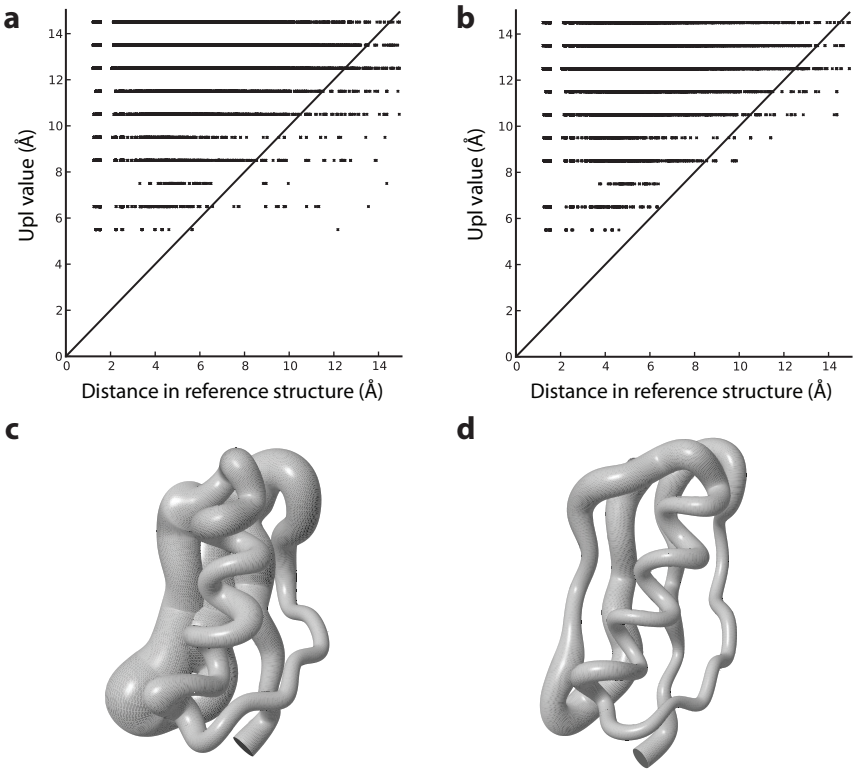


Figure 7.9: Structure calculations using the upl-values determined via analysis of the L-shaped curves presented in Fig. 7.8. The resulting structure bundle (subfigure c) was subsequently used to delete violated distance restraints and the structure calculation was then repeated without these restraints. **a** Correlation between upl-values and distance in the reference structure for all distance restraints. **b** Correlation between upl-values and distance in the reference structure without violated restraints. **c** Sausage representation of the structure bundle calculated based on the restraint set presented in subfigure a (RMSD to mean 1.8 Å and RMSD bias 2.4 Å). **d** Sausage representation of the structure bundle calculated based on the restraint set presented in b (RMSD to mean 1.0 Å and RMSD bias 1.3 Å). The reference structure used to calculate the RMSD bias and to extract the distances for subfigures a and b was generated via regularization (Gottstein et al., 2012a) of the GB1 X-ray structure (PDB 2QMT).

slight loss of quality when compared to the previous calculation results. The average final target function of this structure calculation is 137.8 Å^2, indicating that the algorithm did not completely converge to a structure bundle that fulfills all input distance restraints. Fig. 7.9a shows the correlation between upl-values and distance in the reference structure. All distance restraints originating from the same peak list are located on a horizontal line due to the equality of upl-values within the same peak list. All distance restraints below the diagonal violate the reference structure and are thus responsible for the high final target function and the loss of accuracy. This can be demonstrated when repeating the same structure calculation without all restraints below the diagonal, yielding a structure bundle with an RMSD bias of 0.9 Å and a final target function of 0.18 Å^2. The identification of these restraints is, however, not possible if the reference structure is not known.

Instead, all distance restraints that are violated by the calculated structure bundle presented in Fig. 7.9c can be identified and deleted. These distance restraints are responsible for the high target function value of 137.8 Å^2. Deletion of these restraints, which does not require the knowledge of the reference structure, results in the correlation between upl-values and distance in the reference structure displayed in Fig. 7.9b. The number of distance restraints below the diagonal is significantly reduced, especially those restraints are removed where the upl-value is several Å smaller than the true distance. Recalculation of the structure using the modified set of distance restraints yields the structure bundle presented in Fig. 7.9d with an RMSD bias of 1.3 Å. An iterative approach that uses the improved structure bundle to eliminate violated distance restraints of the original set of restraints does, however, not further improve the structure calculation result.

Altogether, these results show that the approach using constant individual upl-values for every peak list determined via the analysis of L-shaped curves yields a structure bundle which shows only a slight loss of structural quality when compared to the results of the previous structure calculations using a single constant upl-value for every peak list or calibration via $1/r^6$. In contrast to the other two methods, the result could, however, be further improved if the structure was recalculated excluding distance restraints that violate the original structure bundle. This approach to recalculate the structure bundle after deletion of violating distance restraints was additionally applied to the previous calibration methods, but no improvement could be observed, even when using smaller upl-values or *dref*-values.

All thus far presented methods yield structure bundles in the RMSD bias range around 1.5 Å where the global fold of the protein is correct but the local structure is rather inaccurate. The results of this section suggest that it is necessary to increase the information content while keeping the amount of violations as small as possible in order to further improve the structural quality.

Calibration using different exponents

Restraint calibration in one of the previous sections was performed using the $1/r^6$-relation between distance and peak intensity, a method which is commonly used in solution NMR for the calibration of NOESY spectra. This approach is based on the assumption of magnetization transfer between isolated spin pairs which can be used as an approximation if the extent of spin diffusion is small. The high degree of spin diffusion and relayed polarization transfer makes this method very inaccurate for the types of solid-state NMR experiments that were used in the present work. Relayed polarization transfer commonly increases the peak intensity of distant spin pairs if the magnetization can overcome part of the distance via chains of covalently bonded atoms such as amino acid side chains.

The subsequent underestimation of distances leads to the problem that the $1/r^6$-relation requires very large *dref*-values in order to keep the amount of violations to a minimum which, in turn, is responsible for the loss of information content discussed previously. A less steep relation between peak intensity and distance, such as $1/r^3$, results in a larger distance for the same intensity as compared to the steeper $1/r^6$-relation. As this reduces the degree of distance underestimation, it would allow to use a smaller *dref*-value without causing an increase in the number of violations, consequently leading to a gain of information content.

The following structure calculations aim to find out whether different relations between peak intensity and distance, out of which we have tried exponents in the range between $1/r^3$ and $1/r^7$, can improve the structure calculation result. The set of distance restraints was unchanged during all calculations.

Fig. 7.10 summarizes the results in the form of RMSD bias (Fig. 7.10a), information content (Fig. 7.10b) and violation of the reference structure (Fig. 7.10c). Violation of the reference structure is calculated as the CYANA target function of the distance restraint set with respect to the reference structure. The information content measures the potential of a given set of distance restraints to restrain the tertiary fold of a protein chain by taking into account the number of non-redundant long-range restraints as well as the individual upl-values. Improvements of the structure calculation result are expected if the information content can be increased while decreasing or keeping the amount of violations at the same level.

Fig. 7.10a shows that the best result was obtained based on the calibration using $1/r^6$, although the accuracy only slightly drops when applying other exponents. The calibration using different exponents affects the information content as it was intended (Fig. 7.10b). The gain of information content in combination with smaller exponents results from an overall decrease of upl-values, which is exemplary illustrated for the $1/r^3$ relation in Fig. 7.10d. The information content gives, however, no indication about the correctness of the restraints. This makes it necessary to additionally consider the CYANA target function calculated based on the reference structure as a measure of the amount of violations in order to obtain a meaningful estimation of the quality of the resulting distance restraints (Fig. 7.10c). The CYANA target function shows a clear increase when using smaller exponents which can be attributed to the increased number of violations. These additional violations are responsible for the loss of structural quality. The calibrated distance restraints are illustrated exemplary for the calibration using $1/r^3$ in Fig. 7.10d (calibration using $1/r^6$ is shown in Fig. 7.6d for comparison). Every dot below the diagonal represents one distance restraint that violates the reference structure. The presented structure calculation results do, however, not reveal whether the gain of information content

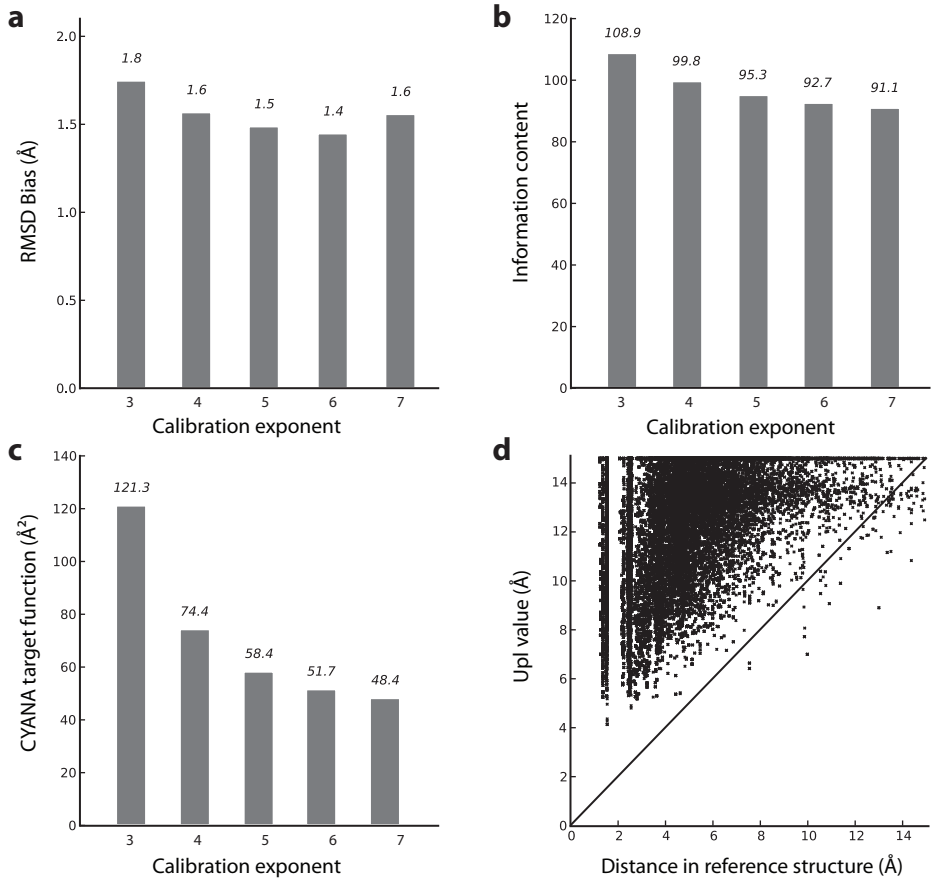


Figure 7.10: Calibration of distance restraints using different exponents for the relation between peak intensity and distance ($1/r^3 - 1/r^7$). The set of distance restraints is based on the manually generated reference assignment of all available peak lists. **a** RMSD bias calculated as the average RMSD to the reference structure, **b** information content of the respective distance restraint set calculated using the method introduced by Weber and Güntert, In preparation, **c** CYANA target function with respect to the reference structure, **d** correlation between upl-values and distance in the reference structure for the calibration using $1/r^3$. The reference structure was generated via regularization (Gottstein et al., 2012a) of the GB1 X-ray structure (PDB 2QMT).

as a consequence of the overall smaller upl-values would suffice to improve the structural quality if the number of violations could be decreased at the same time.

In order to investigate the net effect of the information content, we have repeated the structure calculation using the $1/r^3$ and the $1/r^6$-calibration and deleted all distance restraints below the diagonal. The RMSD bias slightly decreases to 1.35 Å for $1/r^3$ and 1.37 Å for $1/r^6$, respectively, but shows no significant improvement when compared to the $1/r^6$-calibration method including all restraints. This result clearly indicates that the amount of information content limits the structural quality in the absence of violations and the observed increase in information content is thus not sufficient. Altogether, all methods that enabled a gain of information content suffered from additional violations such that the overall structural quality could not be further improved.

7.4 Conclusion

Structure calculations were performed for the model protein GB1 using a set of several two-dimensional ^{13}C-^{13}C correlation spectra of three differently labeled samples (u-^{13}C/^{15}N GB1, 2-^{13}C/^{15}N GB1, 1,3-^{13}C/^{15}N GB1). Results from combined automated peak assignment and structure calculation show that it is in principle possible to fully automatically calculate protein structures from two-dimensional solid-state NMR data and obtain structures in a quality range where the global fold of the protein is correct and inaccuracies occur mostly on the local scale (RMSD bias to the reference structure ~1.5 Å). This is in good agreement with the results reported in the literature for similar test systems (Table 3.1).

Several conclusions can be drawn from the results using automated peak assignment: (i) spectra of samples with diluted labeling were necessary to obtain the correct global fold, (ii) exclusion of overlapped peaks could improve the result, (iii) deletion of intermolecular peaks only slightly improved the result, and (iv) one DARR spectrum of each sample was sufficient to obtain the correct global fold. As these effects were observed on only one test system, it is very likely that structure calculations especially on larger proteins with less resolved spectra behave differently.

In order to investigate which structural accuracy can be obtained with the present data set, we have eliminated errors originating from incorrect peak assignments through the generation of a reference peak assignment for all available peak lists. The best result was obtained when using all available peak lists, however, the structural quality only slightly improved (RMSD bias 1.4 Å) when compared to the automated peak assignment (1.7 Å). We attribute this to problems arising from the calibration of distance restraints and several methods for restraint calibration have thus been compared. Results obtained for different calibration methods did not show significant deviations and all methods were equally able to yield the correct global fold reproducibly. However, none of the methods could improve the accuracy significantly to a quality range that is commonly obtained in solution NMR for a comparable system. Due to the large extent of spin diffusion and relayed polarization transfer in the common ^{13}C-^{13}C correlation experiments, the peak intensity does not contain sufficiently accurate distance information. This leads to either a significant loss of information content if violations of the reference structure are avoided through rather high upl-values or an increased information content is at the cost of a large amount of violations that distort the resulting structure.

Altogether, it can be concluded that it is necessary to increase the information content of a data set without introducing violations of the true structure. Consequently, improvements of the structural quality based on solid-state NMR data need experiment types that allow better estimation of distances based on peak intensities. One such example is the

TEDOR experiment introduced by Hing et al., 1992 which allows the determination of exact distances through the analysis of TEDOR oscillations in a pseudo-three-dimensional spectrum. Using this technique, Nieuwkoop et al., 2009 were able to present a GB1 structure with high accuracy of 0.8 Å. Other developments focus on the use of protons instead of ^{13}C-nuclei for magnetization transfer which approaches the solution NMR situation.

The following section of the present work presents a correction module for relayed polarization transfer which aims to improve the correlation between peak intensity and distance in order to increase the information content without introducing large amounts of violations.

Chapter 8

Full relaxation matrix-based correction of relayed polarization transfer for solid-state NMR structure calculation

8.1 Introduction

Peak intensities carry in theory information about the distance of the corresponding atoms in space. This fact is used in structure calculation protocols to obtain upper distance limits (upls) from measured peak intensities. Especially in solution NMR, this is the most common way to obtain upls from NOESY spectra. In solid-state NMR, on the other hand, there is no consensus protocol for upl generation (Chapter 7) and methods range from intensity-independent fixed upl-values to the calibration as it is used in solution NMR. As shown in the previous section of the present work, the correlation between peak intensity and distance in the reference structure does not sufficiently follow the $1/r^6$ condition to obtain more accurate values when calibrating upls in comparison to simply using fixed values. It could furthermore be demonstrated in the previous chapter that the quality of the resulting NMR structures calculated from spin diffusion-based solid-state NMR experiments such as DARR is limited to a quality range of ~1.5 Å backbone RMSD bias even for a small model protein. This finding is in good agreement with results presented in the literature (Section 3.2). It was concluded that erroneous distance restraints originating from spin diffusion and relayed polarization transfer represent one major source of structural inaccuracies.

Spin diffusion is a problem for distance restraint calibration which does not only occur in solid-state NMR experiments. Solution NMR signals in NOESY spectra also suffer from inaccuracies that can be associated with spin diffusion, however, due to the much smaller degree of spin diffusion in solution NMR as compared to solid-state NMR, their impact on the resulting structural quality is much less pronounced.

Nevertheless, the group of Nilges developed a method for the correction of spin diffusion which can be applied during the iterative automated NOE assignment with ARIA (Linge et al., 2004) and a similar approach was proposed by the group of Riek in conjunction with the development of exact NOEs (Vögeli et al., 2012; Orts et al., 2012). If the complete set of peak intensities including diagonal peaks could be unambiguously measured, the full relaxation matrix approach would allow to back-calculate all individual NOE cross-relaxation rates from which all internuclear distances could be deduced. While this is feasible for small molecules, spectral crowding, peak overlap and missing signals make this approach non-applicable to larger biomolecules. Therefore, the methods presented both by Nilges and Riek rely on a preliminary input structure. This can for example be the result of a conventional structure calculation without the application of spin diffusion correction. The pairwise distances from the input structure are used to calculate theoretical peak intensities using the full relaxation matrix approach. Two sets of theoretical peak intensities are calculated in this way, the first one representing the isolated spin pair approximation (ISPA) without spin diffusion by setting the mixing time to

a very small value, and the second one representing the experimental condition including spin diffusion by setting mixing time to a value corresponding to that of the experimental input data. The two intensity values can be used to determine a correction factor which can subsequently be applied to correct the experimentally measured peak intensities for spin diffusion.

A correction procedure of this kind constitutes a promising approach to improve the structural accuracy of protein structures determined by solid-state NMR spectra. We have therefore implemented a similar idea into the structure determination software package CYANA (Güntert, 2009; Güntert and Buchner, 2015). We have applied the spin diffusion correction method to structure calculations from different simulated NMR spectra of the protein ubiquitin and investigated the performance under varying conditions such as the input structure. For comparison, structure calculations based on the same input data have been conducted according to the conventional CYANA protocol. The set of simulated NMR spectra comprised one solution NMR NOESY spectrum, as well as two solid-state NMR spectra of the CHHC- and the DARR-type. NMR spectrum simulation thereby offers the opportunity to be in complete control of the spectral quality in terms of line width and signal-to-noise ratio as well as to avoid spectral artifacts. This allows to compare the structure calculation results of solution NMR- and solid-state NMR spectra solely with respect to the spin dynamics of the respective experiment type.

8.2 Methods and theory

8.2.1 Full relaxation matrix approach

The intensity build-up of cross-peaks in a through-space NMR experiment (e.g. NOESY in solution NMR or DARR in solid-state NMR) depends on the cross-relaxation rates of the individual spin pairs as well as the initial magnetization of every spin. This process can be simulated for a multi-spin system by the full relaxation matrix approach:

$$\frac{d}{dt}\boldsymbol{M}(t) = -\boldsymbol{R}(\boldsymbol{M}(t) - \boldsymbol{M_0}) \tag{8.1}$$

Solving this differential equation leads to:

$$\boldsymbol{M}(\tau_\mathrm{m}) = e^{-\boldsymbol{R}\tau_\mathrm{m}}(\boldsymbol{M}(0) - \boldsymbol{M_0}) + \boldsymbol{M_0} \tag{8.2}$$

$\boldsymbol{R}$ is the relaxation matrix, $\boldsymbol{M_0}$ is the equilibrium magnetization, $\boldsymbol{M}(0)$ is the starting magnetization, and τ_m the mixing time. The equilibrium magnetization $\boldsymbol{M_0}$ will be ne-

glected in the following. $\mathbf{M}(\tau_{\mathrm{m}})$ contains the resulting peak intensities after the mixing time τ_{m}. The relaxation matrix $\mathbf{R}$ contains the auto-relaxation rates of the relevant nuclei as diagonal elements as well as the cross-relaxation rates between pairs of nuclei as off-diagonal elements. The cross-relaxation rate between the two nuclei i and j is denoted k_{ij} and corresponds to R_{ij} introduced in Equation 1.22 in Chapter 1.3.3 for the solution NMR NOESY experiment. Despite the overall difference between the cross-relaxation rate for NOESY experiments (which can be deduced from the Solomon equations) and the different solid-state NMR experiments (where no theoretical description exists), they all depend on the squared cubic interatomic distance r_{ij}. Equation 8.3 therefore represents a generally applicable description of the cross-relaxation rate which can be used for all experiment types considered in the following.

$$k_{ij} = \frac{c}{r_{ij}^6} \tag{8.3}$$

The constant c depends on the experiment type. For the NOE in solution NMR, c is a function of the rotational correlation time τ_{c} (Sections 1.3.3 and 8.2.2), whereas it is a function of the MAS spinning frequency and the zero-quantum lineshape of the individual spin pairs in the case of proton-driven spin diffusion (PDSD) in solid-state NMR (Section 7.1). In order to build the relaxation matrix $\mathbf{R}$, we have estimated the coefficient c theoretically for the NOESY experiment and determined experimentally for the two solid-state NMR experiments.

8.2.2 Theoretical estimation of the cross-relaxation rate constant for NOESY experiments

The cross-relaxation rate constant in NOESY experiments can be expressed as follows:

$$k_{ij} = \frac{b^2}{20}(6j(2\omega_0) - j(0)) \tag{8.4}$$

Assuming a rigid molecule with isotropic tumbling, the reduced spectral density function $j(\omega)$ is defined by Equation 8.5, and the dipolar coupling constant b by Equation 8.6. ω_0 depicts the Larmor frequency.

$$j(\omega) = \frac{2\tau_{\mathrm{c}}}{1 + (\omega\tau_{\mathrm{c}})^2} \tag{8.5}$$

$$b = \frac{\mu_0 \hbar \gamma_{\mathrm{H}}^2}{4\pi r^3} \tag{8.6}$$

τ_c is the rotational correlation time, r the internuclear distance, γ_{H} the proton gyromagnetic ratio, μ_0 the vacuum permeability, and $\hbar$ the reduced Planck constant. Combining Equations 8.3-8.6 allows the calculation of constant c in the following way:

$$c = \frac{1}{10} \frac{\mu_0^2 \gamma_{\mathrm{H}}^4 \hbar^2}{(4\pi)^2} \tau_c \left(\frac{6}{1 + (2\omega_0 \tau_c)^2} - 1\right) \tag{8.7}$$

The rotational correlation time of ubiquitin was estimated as 5.27 ns based on a molecular weight of 8.5 kDa and a temperature of 25°C using the tool provided at www.nick-anthis.com/tools/tau. This resulted in a value of $c = -296.4$ Å^6 s^{-1} for 850.3 MHz proton Larmor frequency ($\omega_0 = 5.34 \times 10^9$ $rad \times s^{-1}$).

8.2.3 Experimental estimation of the rate constant for solid-state NMR experiments

Experimental estimation of the ubiquitin constant c for the CHHC and DARR experiment at 850 MHz proton Larmor frequency was performed by Kathrin Székely in the group of Prof. Meier at ETH Zürich. This was achieved by peak intensity analysis from a series of NMR spectra of the respective type at three different mixing times. The resulting average constant c is -1200 Å^6 s^{-1} for the DARR experiment and -375,000 Å^6 s^{-1} for the CHHC experiment.

8.2.4 Calculation of peak intensities from an input structure

Peak intensities were calculated for a given mixing time τ_m and pairwise distances from a given 3D structure based on the master equation approach presented in Equation 8.2. A schematic is depicted in Fig. 8.1a.

The first step constitutes the buildup of the relaxation matrix **R**. Off-diagonal elements are calculated using Equation 8.3, where the internuclear distance r is taken from the input structure and the constant c is determined as described in the previous two sections for the respective type of experiment. Diagonal elements of the relaxation matrix are estimated as the negative sum of the off-diagonal elements in the corresponding row.

Only well-structured residues of the ubiquitin input structure were considered for the relaxation matrix. Consequently, all protons from residues 1 – 72 were chosen in the case of NOESY simulations, whereas all carbon atoms from residues 1 – 72 were selected in the case of DARR simulations. For CHHC experiments, all protons from residues 1 – 72

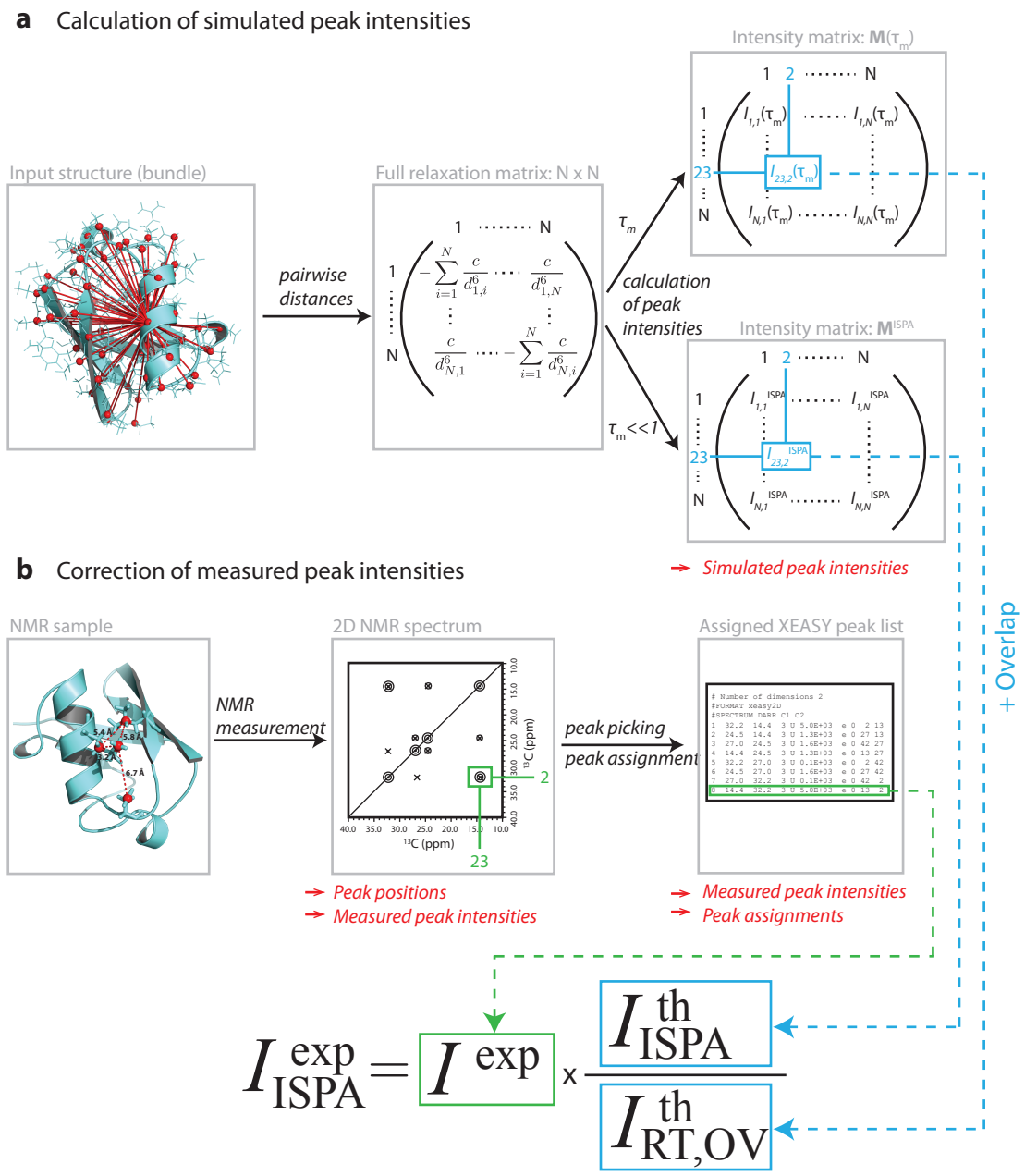


Figure 8.1: Overview of the presented method for peak intensity simulation (**a**) and correction of experimental peak intensities for relayed transfer (**b**).

were chosen excluding those from CH_3-groups. Due to the fast rotational averaging of CH_3-groups, these protons were treated as one pseudo-atom located at the position of the carbon atom.

The starting magnetization $\mathbf{M}(0)$ was set to an arbitrary value of 500.0 for every nucleus considered in the relaxation matrix for NOESY and DARR experiments. In the case of CHHC experiments, it was taken into consideration that the starting magnetization depends on the cross-polarization from carbons to protons prior to the mixing time.

The starting magnetization was therefore set to an arbitrary value of 250.0 for protons belonging to CH_2-groups and to 500.0 for protons belonging to CH- and CH_3-groups.

In order to calculate peak intensities from Equation 8.2, the relaxation matrix is diagonalized using the Jacobi algorithm, which returns the eigenvectors and eigenvalues of the relaxation matrix. Using these eigenvectors combined in the matrix $\mathbf{V}$ as well as the eigenvalues d_i as a diagonal matrix $\mathbf{D}$ containing the elements $e^{-d_i\tau_\mathrm{m}}$, the matrix $\mathbf{M}(\tau_\mathrm{m})$ comprising all peak intensities can be calculated in the following way:

$$\boldsymbol{M}(\tau_\mathrm{m}) = (\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1})\boldsymbol{M}(0) \tag{8.8}$$

8.2.5 Simulation of NMR spectra

Several types of NMR spectra of the protein ubiquitin were simulated using the SimTimeND tool from the NMRPipe System (Delaglio et al., 1995). Types of spectra include 2D [^{1}H-^{1}H]-NOESY, 2D [^{13}C-^{13}C]-DARR and 2D [^{13}C-^{13}C]-CHHC, the two latter being common solid-state NMR experiments for through-space magnetization transfer. The SimTimeND tool requires a list of expected signals, including peak positions in *points* and peak intensities, and calculates an FID which can subsequently be Fourier-transformed in order to obtain the NMR spectrum.

Additional required specifications for FID simulation include the number of dimensions and the types of nuclei, as well as the number of complex points, the sweep width in Hz, the observe frequency in MHz and the carrier position in ppm for every axis of the time domain data. Random noise was added using the AddNoise-tool which requires an RMS value as well as a random seed. RMS values were chosen such that the final signal-to-noise ratios resembled those of experimental spectra of the same type.

Fourier-transformation of the FID was performed using the NMRPipe command of the same software package. The lineshape was adjusted using the processing functions EM (Exponential Multiply Window), GM (Gauss-to-Lorentz Window), SP (Adjustable Sine Bell window) such that the peak shape was close to that of experimental peaks (Lorentz to Gauss ratio of 3:5). All experiments were simulated at 850.3 MHz proton Larmor frequency since comparable experimental solid-state NMR data were available from the Meier group at ETH Zürich that were recorded at the same magnetic field strength. All detailed values for FID simulation and processing are listed in Table 8.1.

The input list of NMR signals contains peak intensities from the intensity matrix $\mathbf{M}(\tau_\mathrm{m})$. An intensity cutoff was applied for the selection of cross peaks which was calculated as 0.01 times the maximum intensity. The intensity cutoff was estimated based on experimental ubiquitin spectra. Peak positions in points were calculated based on the

Table 8.1: NMR spectrum simulation and processing parameters.

	NOESY	CHHC	DARR
Overview			
Mixing time (ms)	50	0.6	400
Number of dimensions	2	2	2
Magnetic field (MHz)	850	850	850
Line width (Hz)	30	100	100
SimTimeND (FID simulation)			
Time domain size (points):			
xT	4096	4096	4096
yT	2048	2048	2048
Full spectral width (Hz):			
xSW	10,266	40,000	40,000
ySW	10,266	40,000	40,000
Observe frequency (MHz):			
xOBS	850.3	213.81	213.81
yOBS	850.3	213.81	213.81
Carrier position (ppm):			
xCAR	6.037	93.54	93.54
yCAR	6.037	93.54	93.54
NMRPipe tool (processing)			
Adjustable sine bell window (SP):			
-off	0.5	0.5	0.5
-end	0.98	0.98	0.98
-pow	2	2	2
-c	0.5	0.5	0.5
Exponential multiply window (EM):			
-lb	12	40	40
Lorentz-to-gauss window (GM):			
-g1	0	0	0
-g2	18	60	60
-g3	0	0	0
addNoise	0.000035	0.00014	0.00014

chemical shifts of ubiquitin as well as the total number of data points in the respective spectral dimension.

8.2.6 Signal identification in NMR spectra

The processed NMR spectra in UCSF-format were used to automatically identify signals using a peak picking routine implemented in the CYANA software package. Automatic peak picking was performed based on a manually determined intensity cutoff of 0.03 for DARR and NOESY spectra, and of 0.1 for the CHHC spectrum, such that the number of identified signals was similar to that of the respective experimental spectra.

8.2.7 Structure calculation

Combined automatic NOE assignment and structure calculation was carried out as described in Section 6 of the present work. Ten individual structure calculations based on the same input data but different random number generation seeds were conducted and the two lowest energy structures of each individual structure bundle were assembled to obtain a new combined structure bundle.

Individual calculations were performed based on the standard CYANA protocol. Chemical shift tolerances as well as calibration details for peak intensities into upper distance limits were determined empirically based on the RMSD bias values of the structure calculation result. An overview of the resulting values is given in Table 8.2. Each individual structure calculation was performed using 10,000 simulated annealing steps based on 100 random starting structures. The final structure bundle comprises the 20 lowest energy structures.

Table 8.2: Structure calculation parameters.

	NOESY	CHHC	DARR
Assignment tolerance (ppm)	0.02	0.2	0.4
Calibration parameter dref (Å)	3.5	5.0	6.0

8.2.8 Relay-correction of peak intensities

Each peak intensity measured from an NMR spectrum of a biomolecule (either experimental or simulated) is biased by contributions from relayed polarization transfer (RT) and overlap from surrounding peaks (OV). The correction procedure introduced in the following aims to determine this contribution in order to estimate a new corrected peak intensity value which represents the isolated spin pair approximation (ISPA). A complete overview of the method is given in Fig. 8.1.

As a first step, the required input structure is used to calculate theoretical peak intensity matrices $\mathbf{M}(\tau_\mathrm{m})$ for the experiment type and mixing time corresponding to the experimental input peak list and $\mathbf{M}^\mathrm{ISPA}$ for a very short mixing time of 10^{-9} s (ISPA simulation). Calculation of the intensity matrices is performed as described in Section 8.2.4 and illustrated in Fig. 8.1a.

The correction procedure is based on the assumption that the relation between the measured peak intensity (I^exp) and the respective peak intensity excluding the contributions from RT and OV ($I^\mathrm{exp}_\mathrm{ISPA}$) equals the relation of the corresponding calculated peak intensities from the peak intensity matrices $\mathbf{M}(\tau_\mathrm{m})$ and $\mathbf{M}^\mathrm{ISPA}$ ($I^\mathrm{th}_\mathrm{RT,OV}$ and $I^\mathrm{th}_\mathrm{ISPA}$). The calculated peak intensities therefore allow the calculation of a correction factor for every

measured peak which can subsequently be used to calculate the desired value $I^{\mathrm{exp}}_{\mathrm{ISPA}}$ from the measured value I^{exp} (Fig. 8.1b).

It is assumed that each measured peak is a superposition of individual peaks arising from the atom pairs representing the assignments of that peak (i.e. the peaks of the experimental input peak list are required to be assigned either ambiguously or unambiguously) as well as contributions from surrounding peaks. The respective calculated peak intensity $I^{\mathrm{th}}_{\mathrm{ISPA}}$ therefore also represents the sum of signal contributions of the n atom pairs corresponding to the assignments of the experimental peak at its position ω_1, ω_2:

$$I^{\mathrm{th}}_{\mathrm{ISPA}}(\omega_1, \omega_2) = \sum_{k=1}^{n} I^{\mathrm{ISPA}}_{k}(\omega_1, \omega_2) \tag{8.9}$$

The theoretical intensity $I^{\mathrm{th}}_{\mathrm{RT,OV}}$ furthermore includes the contributions of surrounding peaks. Equation 8.10 therefore takes into account all potential atom pairs independent of the peak assignments and calculates their intensity contribution at the experimental peak position ω_1, ω_2.

$$I^{\mathrm{th}}_{\mathrm{RT,OV}}(\omega_1, \omega_2) = \sum_{i=1}^{p} \sum_{j=1}^{p} I^{\mathrm{RT}}_{i,j}(\omega_1, \omega_2) \tag{8.10}$$

Every individual simulated signal for a given atom pair i and j ($I^{\mathrm{ISPA}}_{k}(\omega_1, \omega_2)$ or $I^{\mathrm{RT}}_{i,j}(\omega_1, \omega_2)$) is characterized by its peak position, intensity, line width and lineshape. The position of the signal corresponds to the chemical shifts of the two atoms δ_i and δ_j. The intensity is extracted from the respective calculated intensity matrix $\mathbf{M}$ and the line width is specified by the user in accordance with the experimental line width. The lineshape is assumed to be a weighted average of Gaussian and Lorentzian. The peak height $I_{i,j}$ for a given pair of atoms i and j is calculated at the position ω_1, ω_2 of the experimental peak:

$$I_{i,j}(\omega_1, \omega_2) = M_{i,j}(n_{\mathrm{l}} L_{i,j}(\omega_1, \omega_2) + n_{\mathrm{g}} G_{i,j}(\omega_1, \omega_2)) \tag{8.11}$$

The function describing the Lorentzian lineshape L is presented in Equation 8.12, where d represents the half width at half maximum (HWHM).

$$L(\omega_1, \omega_2) = \frac{1}{2\pi} \frac{d_1 d_2}{((\omega_1 - \delta_i)^2 + d_1^2)((\omega_2 - \delta_j)^2 + d_2^2)} \tag{8.12}$$

Equation 8.13 describes the Gaussian lineshape including the standard deviation σ which is connected to the HWHM as shown in Equation 8.14.

$$G(\omega_1, \omega_2) = \frac{1}{\sigma_1 \sigma_2 2\pi} e^{-\frac{(\omega_1-\delta_i)^2}{2\sigma_1^2}+\frac{(\omega_2-\delta_j)^2}{2\sigma_2^2}} \tag{8.13}$$

$$\sigma = \frac{d}{\sqrt{2}ln2} \tag{8.14}$$

The correction factor f is calculated as in Equation 8.15 and subsequently multiplied with the measured peak intensity I^{exp} in order to obtain the experimental peak intensity excluding relayed transfer and peak overlap $I^{\mathrm{exp}}_{\mathrm{ISPA}}$.

$$f = \frac{I^{\mathrm{th}}_{\mathrm{ISPA}}}{I^{\mathrm{th}}_{\mathrm{RT,OV}}} \tag{8.15}$$

8.3 Results and discussion

8.3.1 Conventional structure calculations from simulated NMR spectra

NMR spectra of three different experiment types were simulated for the protein ubiquitin as described in the methods section. Experiment types include solution NMR 2D [^{1}H,^{1}H]-NOESY, solid-state NMR 2D [^{13}C,^{13}C]-DARR and solid-state NMR 2D [^{13}C,^{13}C]-CHHC. All spectra were simulated at 850.3 MHz proton Larmor frequency and the full line width at half maximum (FWHM) was chosen in accordance with experimental NMR spectra of ubiquitin (NOESY 30 Hz, DARR and CHHC 100 Hz).

Automatically generated peak lists were used as input for combined automated peak assignment and structure calculation using the software package CYANA (Güntert, 2009; Güntert and Buchner, 2015). The quality of the resulting structure bundle was assessed via the RMSD with respect to the reference structure (RMSD bias). The reference structure was generated via regularization (Gottstein et al., 2012a) of the ubiquitin crystal structure (PDB 1UBQ). RMSD bias values are given in Table 8.3. Structure calculation results from the solution NMR NOESY spectrum as well as from the solid-state NMR CHHC spectrum are very similar (RMSD bias 0.73 Å vs. 0.78 Å). In contrast, the quality of the structure calculated from the solid-state NMR DARR spectrum is much lower (RMSD bias 2.11 Å) despite the fact that line width, peak density, and simulated inaccuracies of peak positions in the DARR spectrum equal that of the CHHC spectrum. This indicates that not the spectral quality *per se* is the reason for the observed limited structural quality when using

solid-state NMR spectra as input. The main difference between the two solid-state NMR spectra is the peak intensity resulting from either polarization transfer among protons (CHHC) or among carbon atoms (DARR), hence suggesting that calibration of upper distance limits from peak intensities introduces errors which are especially disturbing in the case of DARR spectra. This result is in good agreement with structure calculations from experimental solid-state NMR spectra (Chapter 7 of the present work).

Table 8.3: Structural statistics for conventional structure calculations based on simulated NMR spectra.

	NOESY	CHHC	DARR
Final structure calculation cycle:			
Average backbone RMSD to mean (Å)	0.53	0.71	1.98
Backbone RMSD to reference (Å)	0.73	0.78	2.11
First structure calculation cycle:			
Cross peaks:			
With long-range assignment |i-j|>=5	331	611	672
Distance restraints:			
Average assignments/constraint	4.9	5.3	7.2
Average target function with respect to the reference structure ($Å^2$)	169.2	53.1	417.2
Average information content	347.2	194.4	110.6
Average target function value ($Å^2$)	17.2	7.6	46.7
Average backbone RMSD to mean	1.3	1.2	4.7
Backbone RMSD to reference	1.3	1.2	3.4

The origin of the different structure calculation results is assessed in the following based on the mechanism of magnetization exchange, the resulting simulated peak intensities and their influence on the structure calculation. The calibration of distance restraints relies on the fact that the peak intensity scales with the squared cubic interatomic distance (isolated spin pair approximation, ISPA), however, in biological multi-spin systems interactions with surrounding atoms influence the peak intensity. In solution NMR NOESY experiments the peak intensity is usually weakened by spin diffusion whereas the so called relayed polarization transfer, which is especially prominent if polarization exchange is performed via carbon atoms in solid-state NMR experiments, causes an increase in peak intensity. This increase in peak intensity results in an underestimation of upper distance limits that violate the true distance in the reference structure and, consequently, distort the structure during structure calculation.

The simulated peak intensity in relation to the interatomic distance is presented in Fig. 8.2a-c for all three experiment types. Among the presented experiment types, the peak intensities from the NOESY experiment (Fig. 8.2a) reflect the isolated spin pair approximation (i.e. $1/r^6$-relation) better than those from the two solid-state NMR experi-

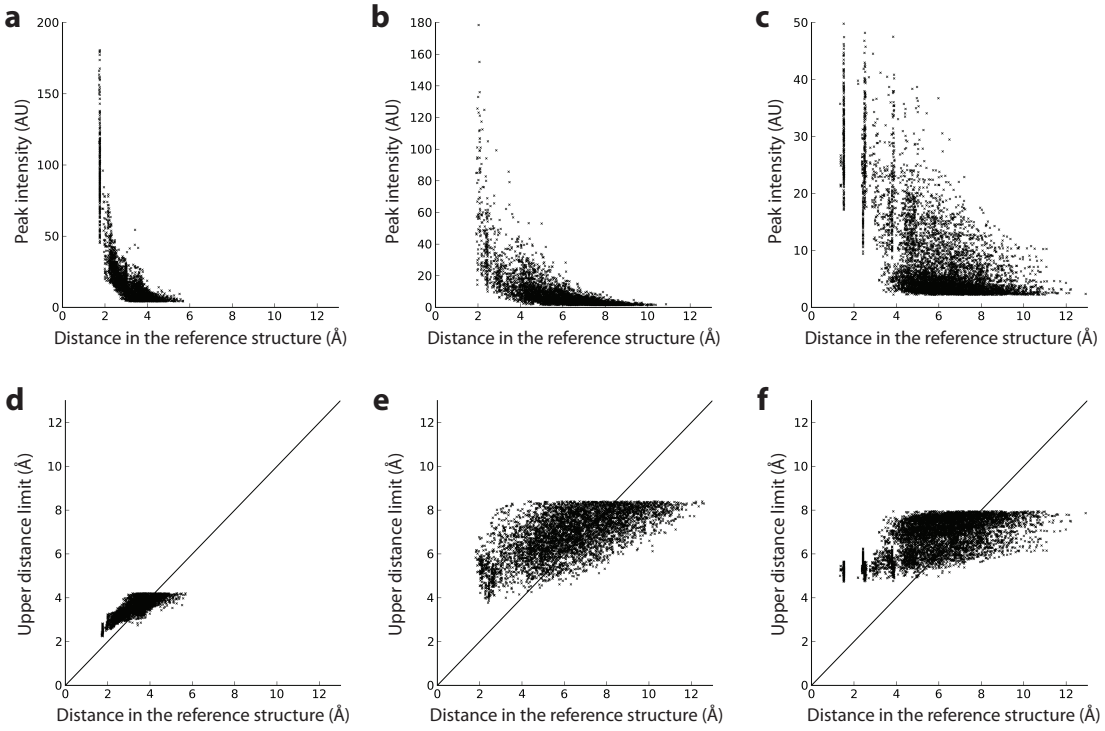


Figure 8.2: Simulated peak intensities (**a-c**) and calibrated upper distance limits using a $1/r^6$ relation (**d-f**) for the protein ubiquitin correlated with the internuclear distance in the reference structure. The reference structure was generated via regularization (Gottstein et al., 2012b) of the ubiquitin crystal structure (PDB 1ubq). Peak intensity simulation was performed using a full relaxation matrix approach, whereas off-diagonal elements of the relaxation matrix were determined based on the cross-relaxation rate constant and the inverse sixth-power of the internuclear distance. **a,d** solution NMR NOESY (100 ms mixing time, cross-relaxation rate constant -296.4 $Å^6 s^{-1}$), **b,e** solid-state NMR CHHC spectrum (0.6 ms mixing time, cross-relaxation rate constant -375x10^3 $Å^6 s^{-1}$), and **c,f** solid-state NMR DARR spectrum (300 ms mixing time, cross-relaxation rate constant -1200 $Å^6 s^{-1}$). All peaks with an intensity of at least 0.01 times the maximum intensity were selected.

ments (Figs. 8.2b and c). The difference between NOESY and CHHC, despite the transfer being mediated via protons in both cases, can be explained by the much stronger dipolar couplings in the solid-state, leading to a larger cross-relaxation rate and, consequently, increasing the distance through which two atoms can still exchange magnetization.

Relayed polarization transfer is most prominent in DARR experiments because magnetization is exchanged between carbon atoms instead of protons. The exchange rate of directly bonded carbon atoms is much higher due to their short distance as compared to the rate of polarization transfer between more distant nuclei. This leads to fast spin diffusion along amino acid sidechains followed by through-space transfer. Consequently, the measured peak intensity is then overestimated if the actual through-space transfer occurs between two less distant atoms as the two originally polarized atoms. This phenomenon is less severe in experiments exchanging magnetization via protons which can mainly be

rationalized by the fact that protons are never directly bonded (i.e. the shortest possible distance between two protons is larger than a covalent carbon-carbon bond) and the distance between structurally meaningful long-range proton pairs is smaller than the respective carbon distance. The discrepancy between the exchange rate of short-range atom pairs and long-range atom pairs is therefore much larger in carbon-mediated polarization transfer, largely favoring relayed transfer.

Figs 8.2d-f show the calibrated upper distance limits in correlation with the corresponding distance in the reference structure. Every dot below the diagonal represents one distance restraint which is distorting the correct protein fold during structure calculation due to an upl value which is smaller than the actual distance. This underestimation of upl-values, as a direct consequence of relayed polarization transfer, is expected to be the origin of the different structural qualities.

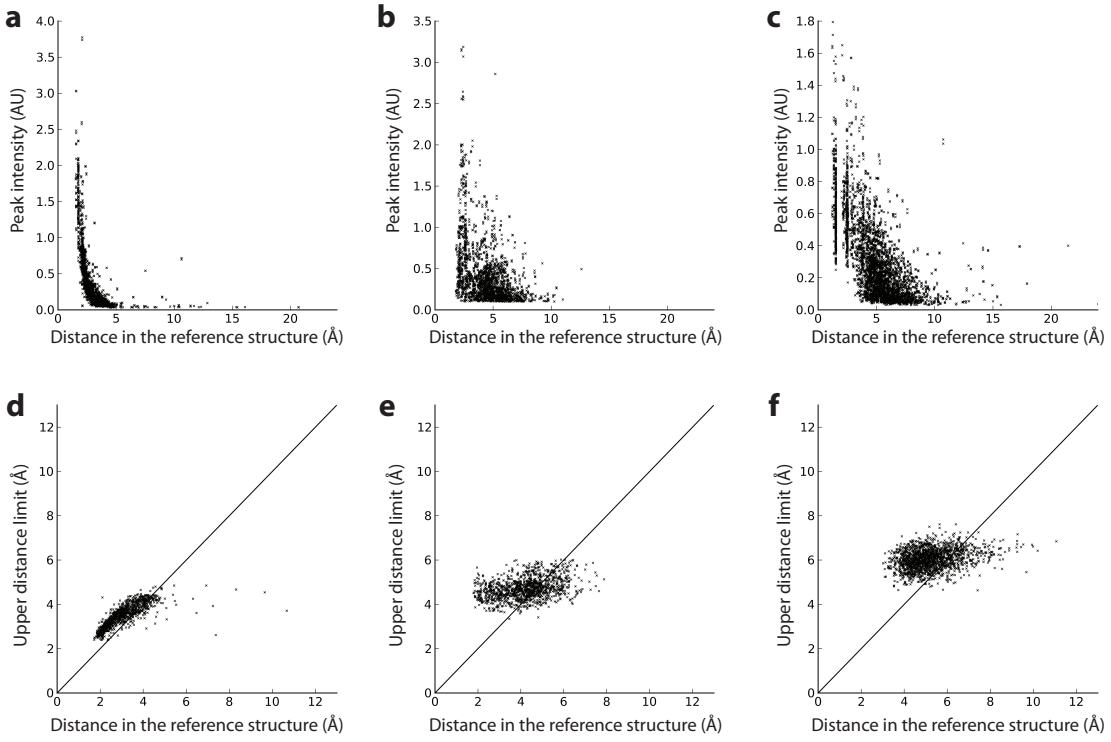


Figure 8.3: Measured peak intensities from simulated spectra (**a-c**) and calibrated upper distance limits using a $1/r^6$ relation (**d-f**) for the protein ubiquitin correlated with the internuclear distance in the reference structure. The reference structure was generated via regularization (Gottstein et al., 2012b) of the ubiquitin crystal structure (PDB 1ubq). Internuclear distances were determined based on peak assignments obtained from automatic peak assignment using CYANA and the r^6-summed distance was used in case of ambiguous assignments. **a,d** solution NMR NOESY (100 ms mixing time, line width 30 Hz), **b,e** solid-state NMR CHHC spectrum (0.6 ms mixing time, line width 100 Hz), and **c,f** solid-state NMR DARR spectrum (300 ms mixing time, line width 100 Hz).

The correlations in Fig. 8.2 show values directly from the full-relaxation matrix simulation instead of the measured peak intensities from the simulated NMR spectra that were used as input for the structure calculations. This allows a direct comparison of the spin dynamics without bias originating from incorrect peak assignments or peak overlap because the exact atom pair corresponding to every peak intensity is known. In order to discuss the structure calculation results, it is more obvious to consider peak intensities extracted from the simulated NMR spectra. As these peaks are unassigned and not completely resolved, the atom pairs corresponding to every peak were determined during iterative peak assignment and structure calculation with CYANA.

As shown in Fig. 8.3, the correlation between the measured peak intensity in the simulated spectrum and the distance in the reference structure is in general very similar to that shown for simulated peak intensities. Individual peak intensities, however, correspond to very large distances in the reference structure (e.g. the largest distance among the simulated NOESY peaks is 5.5 Å (Fig. 8.2a) whereas distances up to 20 Å occur among the peaks obtained from the simulated NOESY spectrum (Fig. 8.3a)). These discrepancies result from incorrect peak assignments. Some of these do not occur in the corresponding plot showing the calibrated upl-values in correlation with the true distance (Fig. 8.3d-f). This can be attributed to the application of constraint combination, a procedure which randomly combines pairs of distance restraints to generate one new restraint which includes all assignment possibilities of the two original distance restraints, thus, potentially decreasing the r^6-summed distance plotted on the x-axis of all subfigures in Fig. 8.3d-f.

The quality of a set of distance restraints can additionally be quantified using the following two measures: (i) CYANA target function with respect to the reference structure as a measure of the amount of violations, which is zero if none of the distance restraints has an upl-value that is smaller than the distance in the structure, and (ii) information content which is a measure of the amount and the restrictive nature of the distance restraints. Best results are obtained if the CYANA target function is low and the information content is simultaneously high.

Values determined for the distance restraints obtained from automatic assignment of the peaks used as input for structure calculation are summarized in Table 8.3. The target function is smallest for the set of distance restraints originating from the CHHC experiment (53.13 Å^2), in the medium range for the NOESY restraints (169.2 Å^2) and much larger for the DARR restraints (417.2 Å^2). The information content, in contrast, is highest for restraints from the NOESY spectrum (347.2), in the medium range for CHHC (194.4), and smallest for DARR restraints (110.6). These values in combination explain the results obtained from structure calculation. In the case of NOESY and CHHC, the violations originating from distance restraints with underestimated upl-values is compensated by a

rather high information content, which is not the case for distance restraints obtained from DARR spectra. Restraints from DARR experiments additionally suffer most from relayed polarization transfer, leading to a significantly larger amount of reference structure violations.

Altogether, it can be concluded from the presented results that the overall lower quality of solid-state NMR structures can to a large extent be attributed to relayed polarization transfer, which is most prominent in experiments relying on carbon-mediated polarization transfer. The present set of simulated NMR spectra is very well suited for this investigation as all parameters defining the spectral quality were equal for all three types of simulated spectra, thus, making the polarization transfer mechanism the only deviation. These results indicate that structural quality from solid-state NMR data could be significantly improved, especially when using DARR-like experiments, when taking relayed transfer into account.

8.3.2 Correction of relayed polarization transfer using a full-relaxation matrix approach

The previous section nicely demonstrated the disturbing influence of incorrect upl-values for structural accuracy. The benefit of the relay correction method for the quality of structure calculation results will therefore initially be investigated using the reference structure as input for the full relaxation matrix calculation. Since potential inaccuracies from errors in the input structure are excluded, this case represents the maximum possible improvement which can be obtained from solely increasing the accuracy of upl values while leaving all other parameters untouched. This is a proof of principle and not a realistic application, as the reference structure is typically unknown. The assigned peak lists from the previous section are used as input peak lists. The following structure calculations therefore differ from those of the previous section as no automatic peak assignment is performed. Instead, the assigned peak lists are simply calibrated into distance restraints after the relay correction procedure and subsequently used for a single target function minimization. Results are evaluated based on (i) the RMSD bias, (ii) the correlation between upl values and true distance in the reference structure, (iii) the CYANA target function with respect to the reference structure, and (iv) the information content (Table 8.4, top).

Fig. 8.4a-c shows the correlation between the corrected upl values and the distance in the reference structure for all three peak lists with the correction being conducted based on the reference structure as input. In all three cases, the upl-values correlate well with the true distance in the reference structure after the correction procedure. Structure calculation results are summarized in Table 8.4. In all three cases, the RMSD bias could be reduced (NOESY from 0.74 Å to 0.33 Å, CHHC from 0.78 Å to 0.49 Å, and DARR from

Table 8.4: Statistics for structure calculations using relay-corrected distance restraints.

	NOESY	CHHC	DARR
Reference structure as input:			
Backbone RMSD to mean (Å)	0.03	0.17	0.31
Backbone RMSD to reference (Å)	0.33	0.49	0.41
Target function with respect to the reference structure (Å^2)	6.87	2.4	2.13
Information content	837.2	560.2	358.8
Realistic structure as input:			
Backbone RMSD to mean (Å)	0.31	0.21	0.81
Backbone RMSD to reference (Å)	0.77	0.91	2.02
Target function with respect to the reference structure (Å^2)	11.0	317.6	1119.4
Information content	604.0	538.0	278.8

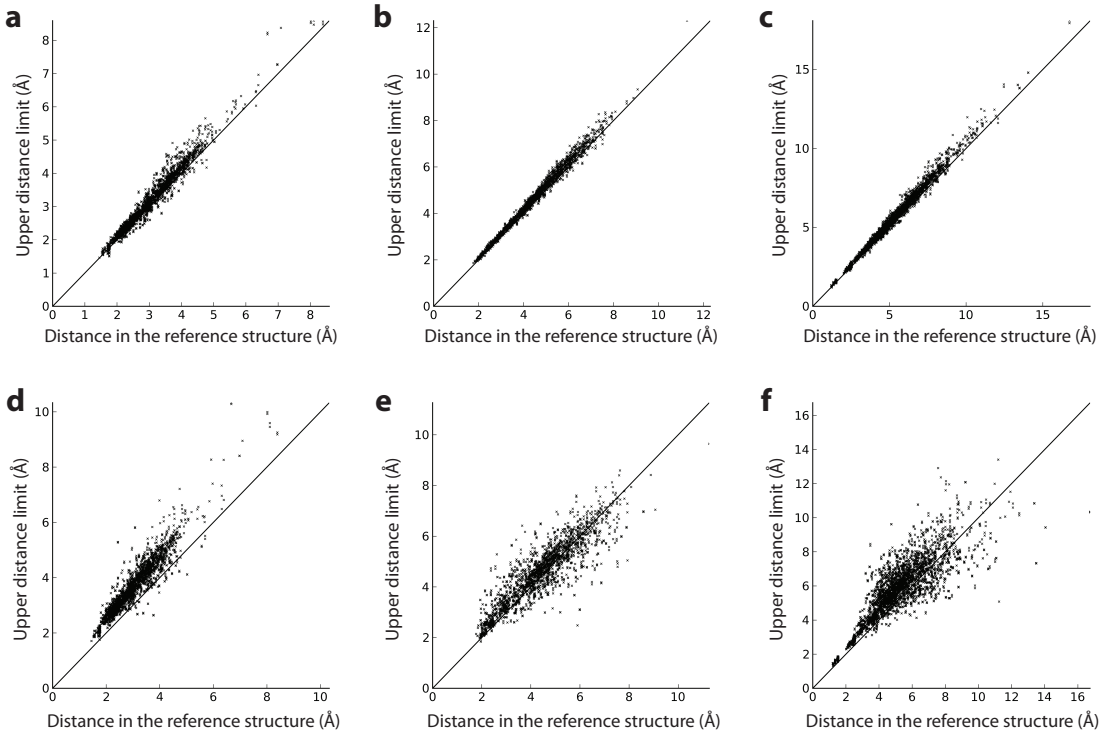


Figure 8.4: Correlation between calibrated upper distance limits using a $1/r^6$ relation and distance in the reference structure. Peak intensities for the calibration of distance restraints were obtained from simulated NMR spectra and subsequently corrected for relayed transfer using either the reference structure as input (**a-c**) or using the result of a conventional structure calculation with the software CYANA as input (**d-f**). The calibration constant was determined by setting the CYANA parameter *dref* such that the final target function of the subsequent structure calculation was below 10.0 Å^2. **a,d** solution NMR NOESY (100 ms mixing time, line width 30 Hz), **b,e** solid-state NMR CHHC spectrum (0.6 ms mixing time, line width 100 Hz), and **c,f** solid-state NMR DARR spectrum (300 ms mixing time, line width 100 Hz).

2.11 Å to 0.41 Å), indicating an overall improvement of the structural quality. This can be attributed to the very small number of violating upl values, which is illustrated in Fig. 8.4a-c and quantified by the low target function with respect to the reference structure (NOESY: 6.87 Å^2, CHHC: 2.4 Å^2, and DARR: 2.13 Å^2). The correction procedure furthermore increases the information content of the distance restraints (NOESY from 347.2 to 837.2, CHHC from 194.4 to 560.2, DARR from 110.6 to 358.8) which represents an additional benefit for the structure calculation result. The increase in information content results from the fact that upl-values have the lowest possible value without violating the true distance. An increase in information content is only beneficial if it is not in conjunction with an increased reference target function (i.e. upper distance limit values smaller than the true distance).

These results clearly show the large impact of the upl values on the structure calculation result. Improvements were obtained in all three cases although the number and type of peak assignments were identical to those from the previous section.

While these results present the theoretical benefit from correcting peak intensities for relayed transfer, this is not routinely applicable as the reference structure is usually unknown. We thus investigated the benefit of the method under more realistic circumstances using an input structure which was obtained as the result of a conventional structure calculation. We used the final combined structure bundle from the structure calculations in the previous section and the same assigned peak lists as for the correction using the reference structure. Results of the correction procedure are presented in Figs. 8.4d-f and the structure calculation results are summarized in the bottom part of Table 8.4. In all three cases, the correlation qualitatively improves when compared to the calibrated upl values without relay correction presented in Figs. 8.3d-f. It is, however, interesting to note that nevertheless none of the corresponding structure calculation results improved when compared to the structure used as input for the correction procedure. In the case of CHHC and DARR, this can be attributed to the significant increase in target function with respect to the reference structure in comparison to the target function without the correction procedure (NOESY: 169.2 Å^2 to 11.0 Å^2, CHHC: 53.1 Å^2 to 317.6 Å^2, DARR: 417.2 Å^2 to 1119.4 Å^2). On the other hand, the correlation resulting from the correction of the NOESY peak list on the basis of the realistic input structure is comparable to that obtained when using the reference structure as input but, against expectation, the structure calculation result shows no improvement when compared to the result without correction (0.77 Å vs. 0.73 Å).

The dependence of the correction result on the quality of the input structure was therefore investigated based on a set of structures with systematically decreasing quality. All of these input structures were obtained from a single CYANA target function minimiza-

tion based on varying numbers of simulated distance restraints with upl values matching the corresponding distances in the reference structure. Using a very small number of distance restraints thereby results in a broad structure bundle which is characterized by a high bundle RMSD as well as a high RMSD bias (e.g. input structure 3 in Table 8.5). Three structures of different qualities were generated and used as input for the correction procedure of the same input data as in the previous sections. Table 8.5 summarizes the results of the structure calculations using the corrected distance restraints as input. The target function with respect to the reference structure as well as the information content were used in order to characterize the corrected set of distance restraints. The correlation between upper distance limits and true distance in the reference structure is presented in Fig. 8.5.

Table 8.5: Statistics for structure calculations using relay-corrected distance restraints.

	1	2	3
Input Structure			
Backbone RMSD to mean (Å)	0.74	1.08	1.85
Backbone RMSD to reference (Å)	0.84	1.27	2.66
NOESY			
Backbone RMSD to mean (Å)	0.22	0.28	0.27
Backbone RMSD to reference (Å)	0.82	0.79	0.75
Target function with respect to the reference structure ($Å^2$)	7.9	9.4	14.6
Information content	585.7	576.5	521.7
CHHC			
Backbone RMSD to mean (Å)	0.54	0.38	0.68
Backbone RMSD to reference (Å)	0.75	0.87	1.03
Target function with respect to the reference structure ($Å^2$)	27.2	45.1	35.0
Information content	440.1	429.8	311.8
DARR			
Backbone RMSD to mean (Å)	0.64	0.57	1.01
Backbone RMSD to reference (Å)	0.91	0.92	1.47
Target function with respect to the reference structure ($Å^2$)	14.6	52.2	87.7
Information content	275.7	269.4	197.5

In direct comparison to the correlation displayed in Fig. 8.4d-f using a realistic input structure for correction, it can be noticed that none of the corrections summarized in Fig. 8.5 resulted in a significant number of distance restraints that violate the true distance in the reference structure, i.e. dots below the diagonal. This is in good agreement with the CYANA target function values with respect to the reference structure listed in Table 8.5 which is especially in the case of CHHC and DARR significantly lower as compared to the values obtained for the corrected restraints from the realistic input structure (Table 8.4).

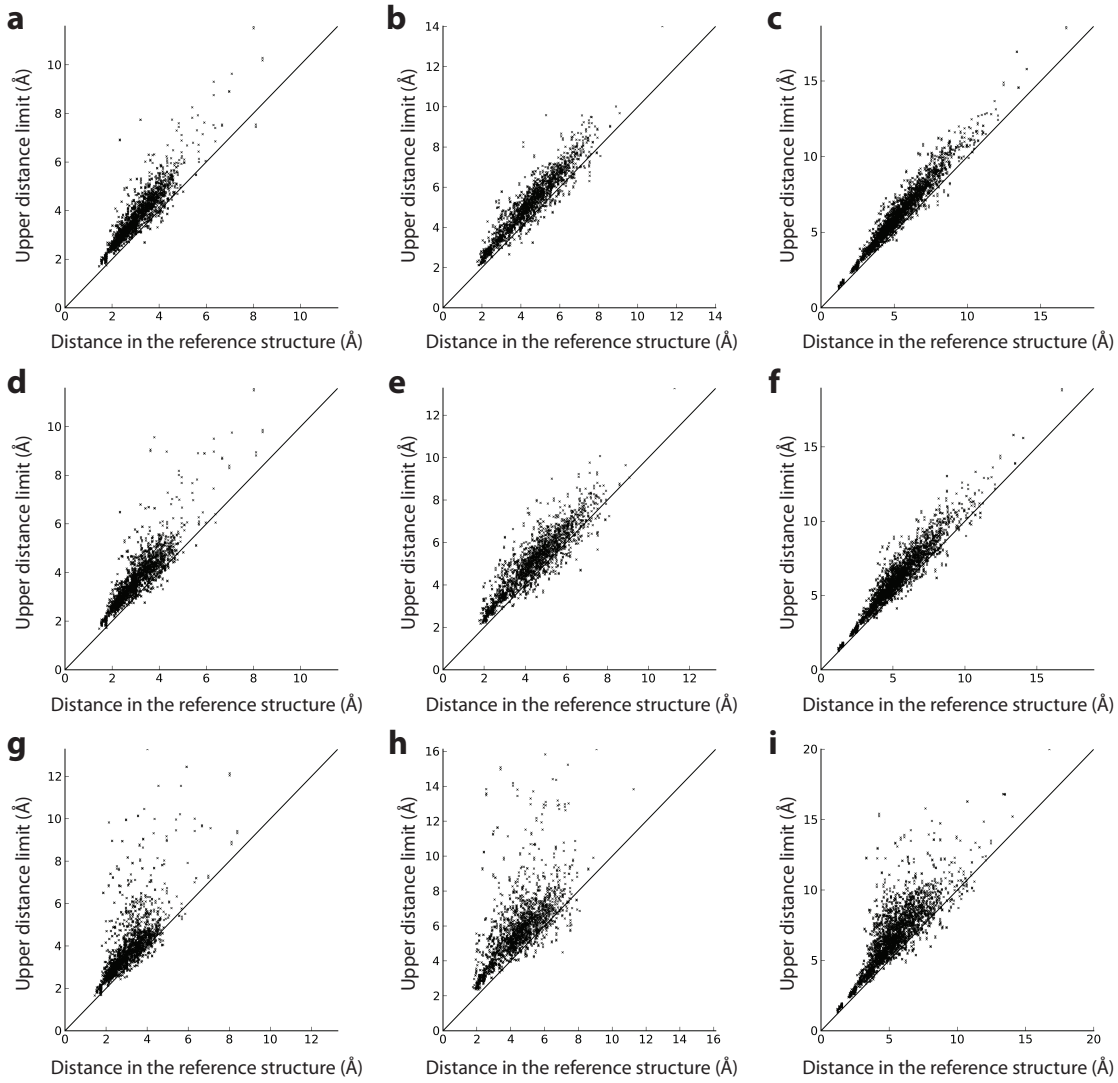


Figure 8.5: Correlation between calibrated upper distance limits using a $1/r^6$ relation and distance in the reference structure. Peak intensities for the calibration of distance restraints were obtained from simulated NMR spectra and subsequently corrected for relayed transfer using input structures of different quality. These structures were generated using a CYANA structure calculation based on different numbers of simulated distance restraints from the reference structure. **a-c** Input structure 1 (RMSD bias 0.84 Å, RMSD to mean 0.74 Å), **d-f** input structure 2 (RMSD bias 1.27 Å, RMSD to mean 1.08 Å), **g-i** input structure 3 (RMSD bias 2.66 Å, RMSD to mean 1.85 Å). The calibration constant was determined by setting the CYANA parameter *dref* such that the final target function of the subsequent structure calculation was below 10.0 Å^2. *left:* solution NMR NOESY (100 ms mixing time, line width 30 Hz), *center:* solid-state NMR CHHC spectrum (0.6 ms mixing time, line width 100 Hz), and *right:* solid-state NMR DARR spectrum (300 ms mixing time, line width 100 Hz).

The apparent "loss" of information content is the result of more accurate upl values in the case of otherwise violating ones and a general tendency of upl value overestimation rather than underestimation. Despite the overall improved correlation and the absence of additional violating restraints, the quality of the resulting structures shows no improvement

when compared to the structures obtained without the correction procedure in the case of NOESY and CHHC. In contrast, in the case of DARR, all structure calculation results significantly improved compared to the structure obtained without correction. Two of the three input structures tested even resulted in a better structure both in comparison to the structure calculated using the uncorrected restraints and in comparison to the input structure itself (Table 8.5 DARR, columns 2 and 3). This example nicely illustrates that it is hypothetically possible to apply the correction procedure using a preliminary structure as input and, as a matter of fact, improve the structure significantly.

However, it remains to be assessed why two structures of very similar quality in terms of RMSD bias (e.g. realistic DARR input structure (Table 8.4, 2.03 Å RMSD bias) and DARR structure 3 (Table 8.5, 2.66 Å RMSD bias) can have so different consequences on the correction result of the very same input peaks and the subsequent structure calculation. This gives an indication about the strong influence of the input structure, thereby not referring to the overall quality of the latter but rather to the specific nature of individual inaccuracies. A noticeable difference among the two structures relates to the distance restraints used as input for the respective structure generation. One structure was obtained based on a very small, but highly accurate set of distance restraints, whereas the other structure was calculated based on a large and rather inaccurate set of distance restraints including several highly violating ones. Depending on the type of violating restraints, this may have consequences on the orientation of individual side chains which does not necessarily reflect on the overall RMSD bias but may significantly influence the full-relaxation matrix calculation of peak intensities if new magnetization pathways are generated by these individual misoriented side chains which are not present in the true structure.

A closer inspection of the structural details illustrated in Fig. 8.6 indeed reveals one phenylalanine side chain, potentially one amongst others, which has a completely different orientation when comparing the reference structure (cyan in Fig. 8.6) with the realistic input structure (magenta in Fig. 8.6a). Fig. 8.6a furthermore nicely shows that this particular side chain is misoriented uniformly within the complete bundle which can be attributed to distance restraints in the input data set that restrain the side chain in this orientation. The second input structure calculated based on the sparse but correct set of restraints, on the other hand, possesses an almost random orientation of this particular side chain (magenta in Fig. 8.6b), most likely originating from a lack of restraints in this region. During the full relaxation matrix calculation, the conformation is averaged, thus avoiding inaccurate magnetization pathways which potentially influence the spin dynamics of the complete surrounding amino acids. Such incorrect but uniform misorientations in the input structure affect the peak intensity correction such that the corrected distance

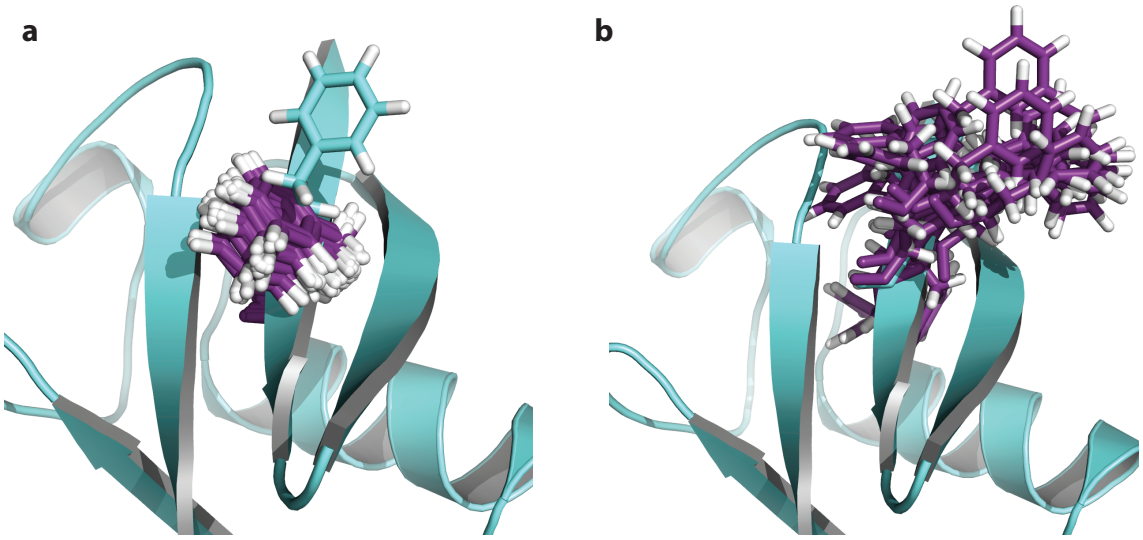


Figure 8.6: Detailed view of the phenylalanine side chain of residue 4 (displayed in purple) in two different structure bundles that were used as input for relay correction of solid-state NMR DARR spectra. The backbone atoms of the structure bundles were superimposed with the reference structure (shown in cyan), but only the phenylalanine 4 side chain was selected for display in the case of the two structure bundles. **a** Input structure obtained as the result of a conventional structure calculation (see also Fig. 8.4f), **b** input structure 3 (see also Fig. 8.5i).

restraints still restrain the structure into the incorrect position. This gives a potential explanation for the observed results, however, it is difficult to find definite proof.

Altogether, the presented results using different input structures for the correction of the same input peak list clearly show that the choice of input structure is very crucial and that especially those structures resulting from a preliminary structure calculation based on incorrect distance restraints are not well suited due to their potential mis-orientation of individual side-chains. On the other hand, structures obtained from an independent source of a limited number of distance restraints with accurate upper distance limits could rather be used to correct the peak intensities of additional spectra such as DARR for relayed transfer.

8.4 Conclusion

We have presented structure calculations of the model protein ubiquitin using input peak lists from simulated solution NMR NOESY as well as solid-state NMR CHHC and DARR spectra. Peak intensities for the simulation of NMR spectra with the software NMR-Pipe were obtained from a full relaxation matrix calculation. The simulation of these three experiment types with spectrum-specific atom selections and transfer rates revealed two main differences between solution NMR and solid-state NMR spectra: (i) signals in NOESY spectra generally arise from atom pairs with much shorter distances up to a max-

imum of ~5 Å compared to ~10 Å in solid-state NMR spectra and (ii) a significant smaller amount of relayed polarization transfer is observed in the NOESY spectrum.

Structure calculations using peak lists obtained from these simulated spectra show very similar high quality for the structures from NOESY and CHHC spectra and a noticeable lower quality in the case of the DARR spectrum. We attribute this to a significantly lower information content of DARR spectra as a result from distance restraints between carbon atoms, in combination with a much larger amount of violating distance restraints as a result of the high degree of relayed polarization transfer. These results indicate that the use of protons for magnetization exchange is very beneficial for the outcome of structure calculations in solid-state NMR.

However, especially due to their robust nature and wide applicability, proton-driven spin diffusion experiments such as DARR are very popular and therefore used in the majority of published structures determined by solid-state NMR. The knowledge about the inaccuracies of distance restraints obtained from DARR spectra made the correction of peak intensities for relayed transfer a promising approach to improve structural accuracy. The presented method using a full relaxation matrix approach relies on an input structure as well as one or several assigned peak lists. Thereby, peak assignments can be obtained from automatic peak assignment routines provided by structure calculation software tools such as CYANA. The original idea considering the input structure was an iterative approach using a preliminary structure resulting from a conventional structure calculation for the correction and subsequent recalculation of the structure using the corrected restraints.

The first important finding based on the results obtained when using the reference structure as input for the correction procedure shows that the signals in all of the simulated spectra have the potential to yield correct structures with very high accuracy even when using automatic peak assignment, leading to various degrees of ambiguity, as long as the upl values are sufficiently accurate (i.e. maximizing the information content while minimizing the amount of reference structure violations). The application of the original idea, i.e. using a preliminary structure of a conventional structure calculation as input, revealed that the method is not robust very with respect to inaccuracies in the input structure. Despite an apparent overall improvement of the correlation between the corrected upl-values and the true distance in the reference structure, the input structure inaccuracies influence the correction procedure to such a large extent that the corrected distance restraints do effectively not result in an improved structure calculation result.

A closer investigation of the input structure requirements led to the finding that an input structure determined from an independent source of highly accurate but potentially limited number of distance restraints can be used as input for the correction of DARR spectra and still improve the result both with respect to the result without the correction

procedure as well as with respect to the input used for the correction. This finding could be explained by the mis-orientation of individual side chains in structures obtained from inaccurate distance restraints. These have an especially disturbing effect on the correction procedure due to the potential generation of false magnetization pathways which negatively influence the spin dynamics of all atoms in the surrounding region.

The overall conclusion of the presented results is that spin diffusion-based experiments cause structural inaccuracies due to erroneous upl-values and that correction of these mainly relay transfer-based errors can significantly improve the structure calculation result. The presented method as a first suggestion of how to approach this problem did unfortunately not turn out to be robust enough to significantly improve the overall accuracy of structures determined by spin diffusion-based experiments. In this respect, it should furthermore be noted that the spectra were simulated using the exact same theory that was used as basis for the correction procedure, thus having no imperfections resulting from erroneous rate constants for example. Especially, when proceeding to the logical next step of using experimental NMR spectra for the correction procedure, it is thus necessary to develop a different, less input-structure dependent approach to tackle this problem.

8.5 Implementation in CYANA

Two new CYANA commands were implemented in order to simulate peak intensities based on a given input structure (`structure peakvolumes`) and to correct experimental peak intensities for relayed polarization transfer (`peaks spindiffusion`). The first command was used for the simulation of NMR spectra, whereas the second command was used for the relay-correction of different types of peak lists. A more detailed description of their use including available parameter settings is given in the following.

structure peakvolumes

The new CYANA command `structure peakvolumes` uses a given input structure and the corresponding chemical shifts in order to calculate peak intensities using the full relaxation matrix approach. In addition to the interatomic distances from the given structure, the specification of the mixing time as well as the rate constant for magnetization exchange is required for the calculation of the relaxation matrix. Prior to executing this command, it is furthermore necessary to select only those atoms which carry polarization at the beginning of the mixing time, which strongly depends on the experiment type to be simulated. The `structure peakvolumes` command can be used to generate a simulated peak list containing the peak positions, peak intensities and the peak assignment (`option=peaks`), however, it can also be used to generate the input file which is required for the simulation

of NMR spectra using the NMRPipe software package (`option=tab`). This NMRPipe input file requires the peak positions to be specified in points rather than ppm, hence, one needs to specify the spectral width in Hz (`swx`, `swy`), the number of data points in each dimension (`nx`, `ny`), as well as the observer frequency in MHz (`obsx`, `obsy`) in order to calculate the peak position in points from an atoms chemical shift in ppm. The routine is currently limited to two spectral dimensions.

- `tmix=real` (required)

  The mixing time `tmix` is specified in s.

- `constant=real` (required)

  The parameter `constant` refers to the rate constant of magnetization transfer and is used to calculate the off-diagonal elements of the relaxation matrix.

- `format=string` (required)

  The `format` refers to the CYANA format specifying the experiment type. It should match the syntax noted in the CYANA library `cyana.lib` for the respective experiment type.

- `swx=real` (required)

  Spectral width of the direct dimension specified in Hz.

- `swy=real` (required)

  Spectral width of the indirect dimension specified in Hz.

- `nx=integer` (required)

  Number of points in the direct dimension.

- `ny=integer` (required)

  Number of points in the indirect dimension.

- `obsx=real` (required)

  Carrier frequency in the direct dimension specified in MHz.

- `obsy=real` (required)

  Carrier frequency in the indirect dimension specified in MHz.

- `option=string` (default=tab)

  The parameter `option` specifies the output-format which can either be the Xeasy format for peak lists (`option=peaks`) or the input file format for spectrum simulation using NMRPipe (`option=tab`).

- `outfile=string` (default=peaks.tab)

  The parameter `outfile` specifies the name of the output file.

- `distance=string` (default=dave)

  The parameter `distance` refers to the determination of interatomic distances if a structure bundle is used as input. The default value `distance=dave` calculates the average distance, further options include the r^6-summed distance (`distance=dr6sum`), or the maximum distance (`distance=dmax`).

- `cutoff=real` (default=0.05)

  The parameter `cutoff` specifies the relative intensity cutoff with respect to the maximum intensity, i.e. peaks with an intensity below the threshold calculated as the product of `cutoff` and maximum intensity are discarded.

- `chhc` (option)

  The option `chhc` is supposed to be selected if the spectrum type to be simulated is CHHC. CHHC is a special case, as the chemical shifts are taken from the carbon atoms, but the covalently bonded protons are internally considered for the relaxation matrix. The starting magnetization of the protons is furthermore calculated based on the number of protons bonded to a specific carbon atom, such that a proton bonded to a methyl carbon obtains one third of the starting magnetization of a C_α atom.

peaks spindiffusion

The command `peaks spindiffusion` uses a given input structure, the corresponding chemical shifts and one or several assigned peak lists in order to calculate relay- and optionally overlap-corrected peak intensities using the full relaxation matrix approach. These corrected peak intensities can subsequently be calibrated to obtain distance restraints which can then be used as input for structure calculation. The result of the algorithm includes the corrected peak intensities which can be obtained by writing the peak list using `write peaks` after having executed the `peaks spindiffusion`-command. As the corrected peak intensities replace the original peak intensities in the storage, it is also possible to directly calibrate distance restraints without writing the peak list.

Prior to executing the command, it is necessary to have at least one structure, the chemical shifts and the assigned peaks in the storage, and to make an atom selection which corresponds to the experiment type of the peak lists. The routine is currently limited to 2D through space correlation experiments.

- `tmix=real` (required)

  See parameter `tmix` in `structure peakvolumes`.

- `constant=real` (required)

  See parameter `constant` in `structure peakvolumes`.

- `swx=real` (required)

  See parameter `swx` in `structure peakvolumes`.

- `swy=real` (required)

  See parameter `swy` in `structure peakvolumes`.

- `nx=integer` (required)

  See parameter `nx` in `structure peakvolumes`.

- `ny=integer` (required)

  See parameter `ny` in `structure peakvolumes`.

- `obsx=real` (required)

  See parameter `obsx` in `structure peakvolumes`.

- `obsy=real` (required)

  See parameter `obsy` in `structure peakvolumes`.

- `distance=string` (default=dave)

  See parameter `distance` in `structure peakvolumes`.

- `lwx=real` (default=25.0)

  Specifies the line width in the direct dimension of the experimental spectrum which was used as basis for the input peak list in Hz.

- `lwy=real` (default=25.0)

  Specifies the line width in the indirect dimension of the experimental spectrum which was used as basis for the input peak list in Hz.

- `cgauss=real` (default=0.5)

  Specifies the peak shape in the experimental spectrum as the relative amount of gaussian. The remaining part is considered to be lorentzian.

- `nooverlap` (option)

  If the option `nooverlap` is chosen, the experimental peak intensities are only corrected for relayed polarization transfer and the overlap of signals, which also contributes to the peak intensity inaccuracy, is ignored.

- `individual` (option)

  The option `individual` performs the correction of experimental peak intensities individually for each structure of the input structure bundle and averages the resulting corrected intensities. The standard approach without the option `individual` performs only one correction and uses the average distance (if `distance=dave`) of each atom pair for the relaxation matrix.

- `chhc` (option)

  See option `chhc` in `structure peakvolumes`.

Conclusion and outlook

The questions that have been addressed in this doctoral thesis all had the goal to increase the reliability and improve the accuracy of structures determined by solution- as well as solid-state NMR spectroscopy. The first project comprised an extensive study on the robustness of the combined automated NOE assignment and structure calculation algorithm based on ten experimental solution NMR data sets that were modified in several ways to mimic different kinds of data imperfections (Chapter 4). Two additional projects were concerned with methodological developments, i.e. the *Peakmatch* algorithm (Chapter 5) and a new protocol for combined automated NOE assignment and structure calculation (Chapter 6), that aim to improve the input data quality and to increase the reliability of the structure calculation result, respectively. The last two projects were focused on structure determination by solid-state NMR (Chapters 7 and 8).

The results from the large-scale study on the performance of the automated NOE assignment and structure calculation algorithm implemented in the CYANA software package with regard to input data imperfections clearly indicate that missing chemical shifts as well as any other type of error within the chemical shift assignment can cause severe errors in the structure calculation. Strongly depending on the protein and the quality of the input data, 10 % or more missing or erroneous chemical shifts result in structure bundles with an average RMSD to the reference structure above 3 Å. Missing or erroneous NOESY peaks, in contrast, cause less severe difficulties. This can be explained by the fact that NOESY peaks firstly contain a large amount of signals that contain no or very limited structural information due to their sequential nature and secondly contain rather redundant information through the dense NOE network. In contrast, one missing chemical shift leads to a whole set of NOESY peaks that remain unassigned in the more favorable case or get assigned incorrectly in the less favorable case.

A criterion for the reliable evaluation of a structure calculation result is especially important in the course of *de novo* structure determinations. We have therefore shown that combining the convergence of the initial structure calculation cycle and the RMSD drift between the first and the last cycle into a weighted average can be used as an indication for the accuracy of a structure calculation result.

The consequence of input data imperfections on the structure calculation result was thoroughly investigated. Especially when using a data set consisting of different input peak lists, it is important that the individual peak lists are consistent with each other and with the available chemical shifts. The *Peakmatch* algorithm was therefore introduced to determine the optimal offset between two multidimensional unassigned peak lists that contain corresponding dimensions. Principal advantages of the algorithm are that (i) it can be applied to unassigned peak lists, (ii) it is highly tolerant against the common imperfections of experimental peak lists, (iii) the criterion for optimal matching is mathematically simple and largely captures what an experienced spectroscopist would do manually, and (iv) its application is straightforward and quick. The two available optimization strategies, i.e. complete grid search and downhill simplex optimization, in general performed equally well. The complete grid search has the advantage that it cannot be trapped in a local minimum and its use is therefore generally recommended if the expected calculation time allows it, for example if the grid size or the number of dimensions to be matched are small.

It was demonstrated that combined automated NOE assignment and structure calculation can go wrong and yield a narrow structure bundle that is not in agreement with the reference structure, i.e. where the precision overestimates the accuracy. We have therefore presented a new method for combined automated NOE assignment and structure calculation implemented in the CYANA software package that performs 20 individual structure calculations that each yield a slightly different final NOE assignment and a different final structure bundle. Combining the individual NOESY peak assignments into one consensus assignment and repeating a simple CYANA structure calculation based on the resulting consensus distance restraints yields a new structure bundle, i.e. the consensus structure bundle. It could be demonstrated that the precision of the consensus structure bundle gives a reliable indication of the structural accuracy throughout a large number of test calculations. The precision of the consensus structure bundle is not strictly equal to the accuracy but proportional with a median proportionality factor of 1.4. The new method is helpful for input data optimization in the course of NMR structure determinations, and we recommend it especially for routine use in the final structure calculation, since the consensus bundle reflects the experimental data much better.

Solid-state NMR is a powerful tool to study molecules at atomic resolution that are not amenable to the classical structure determination methods, namely X-ray crystallography and solution NMR spectroscopy. Among these, membrane proteins in their native lipid environment as well as amyloid fibrils are of special relevance for medical questions. However, in order to routinely apply solid-state NMR for atomic resolution structure determination, reliable and robust protocols are still in need. We have extensively tested the CYANA automated NOE assignment and structure calculation algorithm when using a

set of various two-dimensional solid-state NMR spectra of differently labeled GB1 samples. The main conclusion of these test calculations is that it is in principle possible to fully automatically calculate atomic resolution structures from solid-state NMR spectra and reproducibly obtain the correct global fold with inaccuracies occurring mostly on the local scale, provided that NMR spectra from diluted labeling schemes are included. Another important finding is concerned with the generation of upper distance limits. Depending on the type of magnetization exchange in the NMR experiment, peak intensities have very limited correlation with the distance in the structure, which is especially severe in spin diffusion-based experiments that are very popular due to their simple and robust nature. Structure calculations based on assigned peak lists led to the conclusion that it is not possible to improve the accuracy of the structure calculation result to a level which is commonly obtained from solution NMR data when applying the common methods for upper distance limit calibration reported in the literature, i.e. calibration using a $1/r^6$ correlation between peak intensity and distance, constant upper distance limits, or analysis of L-shaped curves.

Due to this finding, we have implemented a correction module for spin diffusion in the CYANA software package, which aims to improve the quality of the resulting upper distance limits. This is achieved by a full relaxation matrix approach based on one or several assigned peak lists and a preliminary input structure, which can be the result of a conventional structure calculation, a homology model, or a structure determined by X-ray crystallography. The corrected peak intensities are then recalibrated into upper distance limits and applied for a final simple structure calculation based on distance restraints. The correction procedure was tested based on simulated solution NMR and solid-state NMR spectra of the protein ubiquitin. Structure calculation results from the conventional combined automated NOE assignment and structure calculation protocol yielded similar results as they were obtained from experimental data in the preceding chapter and as they are presented in the literature. Most importantly, it could be shown that the correction procedure can significantly improve the structural accuracy when the known reference structure determined by X-ray crystallography is used as input. This improvement is most prominent when using solid-state NMR spectra of the DARR type as input, which can be attributed to the large extent of spin diffusion in this type of experiment. Despite the proof of concept, the method did not turn out to be well suited for regular application, the most important reason being the strong dependency on the input structure. The result of a conventional structure calculation based on the uncorrected input data is ill-suited as input structure due to the distortions in the structure, which are most likely homogeneous in the complete structure bundle and therefore bias the correction procedure. It turned out that a structure calculated from a small number of highly accurate distance restraints can

be used as input for the correction as the structure bundle is rather broad and errors are not present homogeneously in the bundle. It is, however, not straightforward to determine the required amount of highly accurate distance restraints experimentally. The conclusion from the presented structure calculations based on simulated ubiquitin spectra is that the structural accuracy can only be improved if the distance information is more accurate. This could either be achieved by the development of NMR experiments that intrinsically deliver more accurate distance information or through the development of a more robust approach for the correction of spin diffusion that is less dependent on an input structure.

Whereas the questions addressed in the first part of this thesis all reached the intended goal and are therefore highly recommended to be applied routinely, structure determination by solid-state NMR still remains a very difficult task and no breakthrough could be achieved within this work. There is no general recommendation about the proceeding especially in more difficult cases such as large membrane proteins. Developments are definitely required on the spectroscopic side that focus on the increase in sensitivity and resolution, potentially using proton detection which allows the measurement of higher-dimensional solid-state NMR spectra that provide rather accurate long-range information. Considering more complex systems, it is equally important to concentrate on the sample preparation in order to obtain sufficiently narrow NMR signals. For the structure calculation itself, it might be helpful to incorporate additional information, for example from sequence-based structure prediction tools.

Appendix

A Evaluation of structure calculation with CYANA – results for the individual proteins

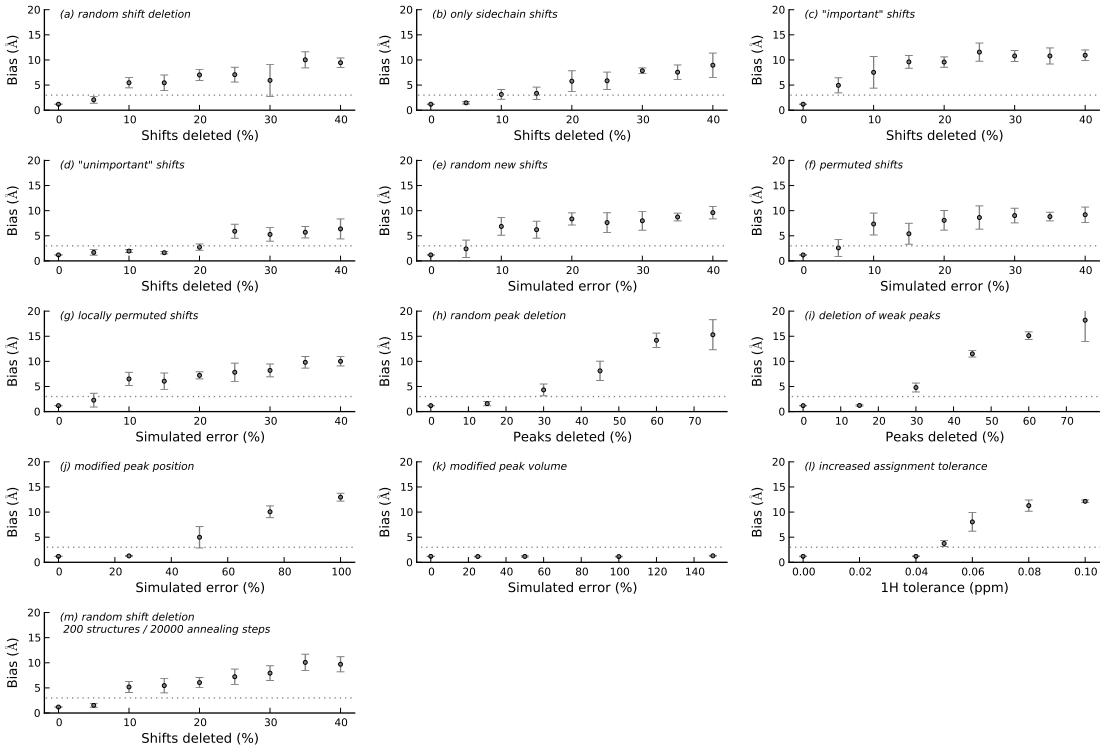


Figure A.1: RMSD to the reference structure for different input data modifications of the protein copz.

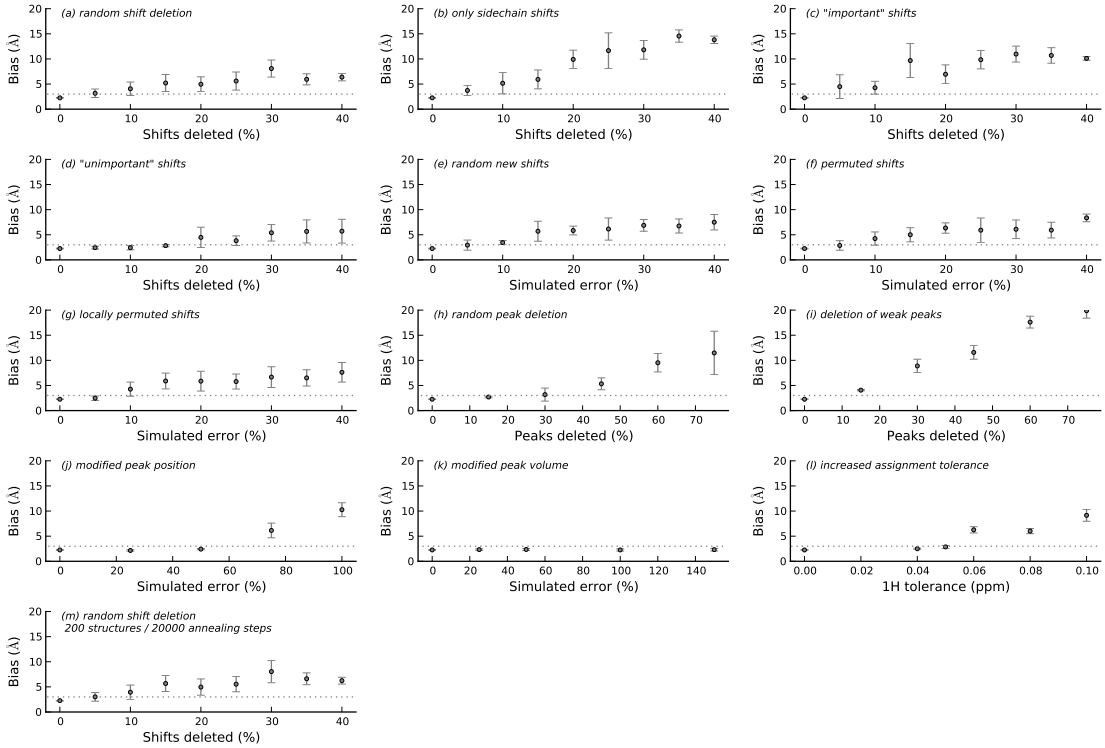


Figure A.2: RMSD to the reference structure for different input data modifications of the protein cprp.

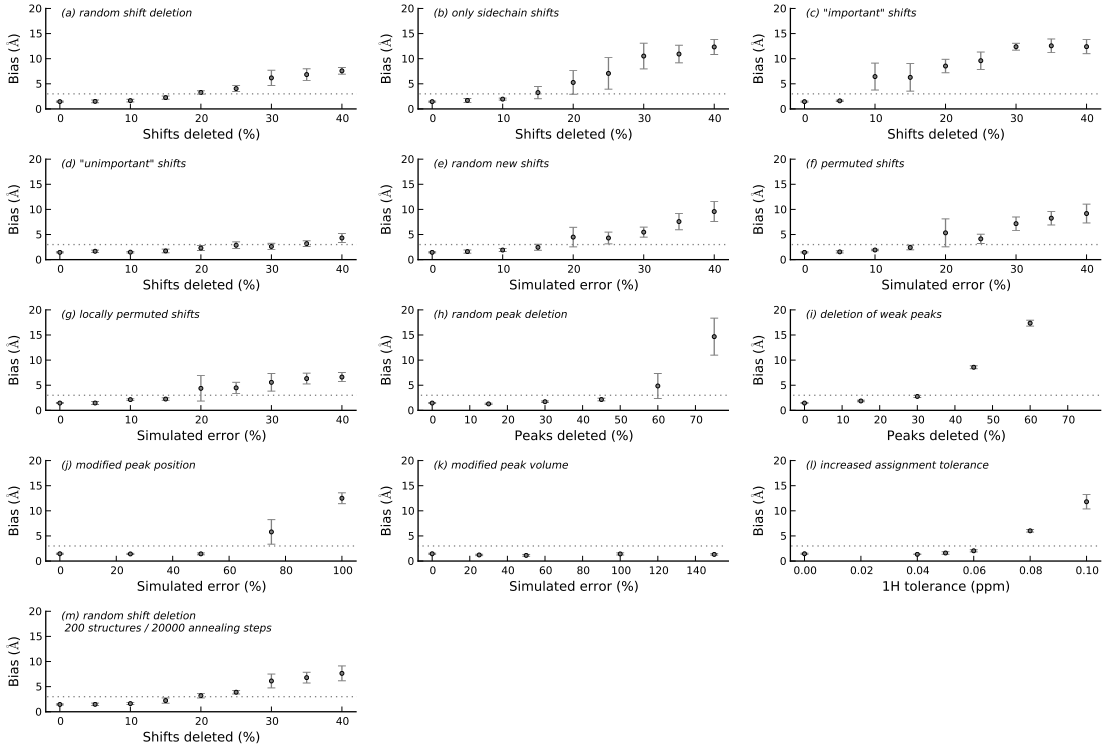


Figure A.3: RMSD to the reference structure for different input data modifications of the protein enth.

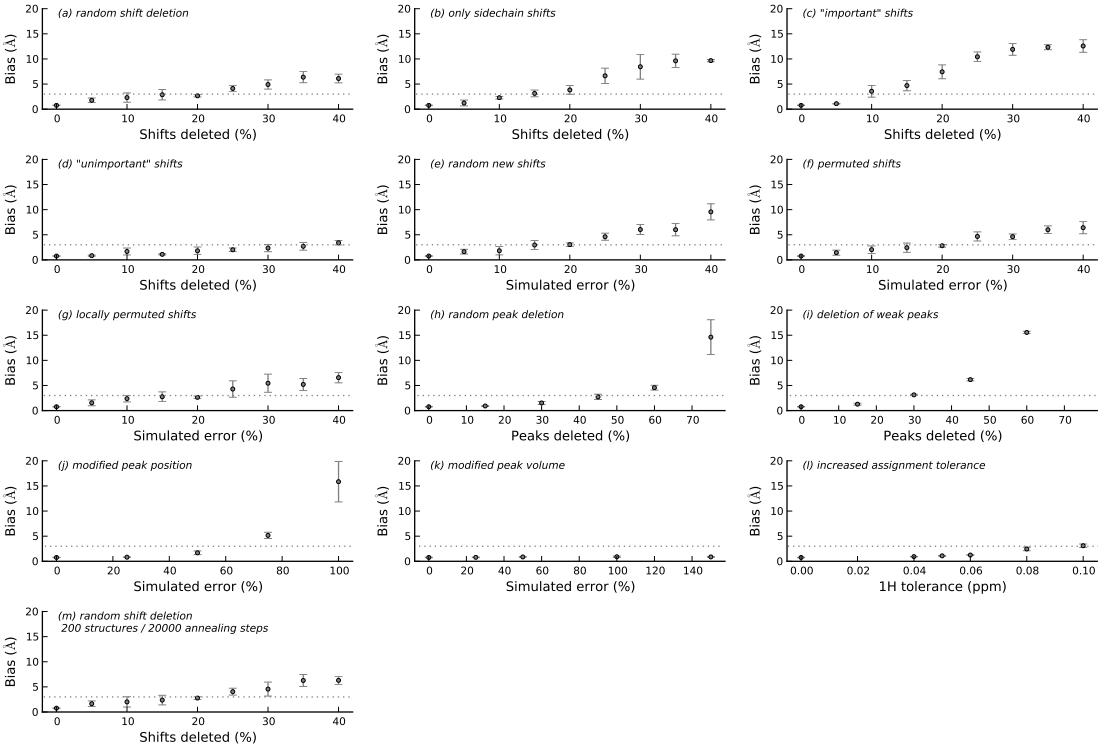


Figure A.4: RMSD to the reference structure for different input data modifications of the protein fsh2.

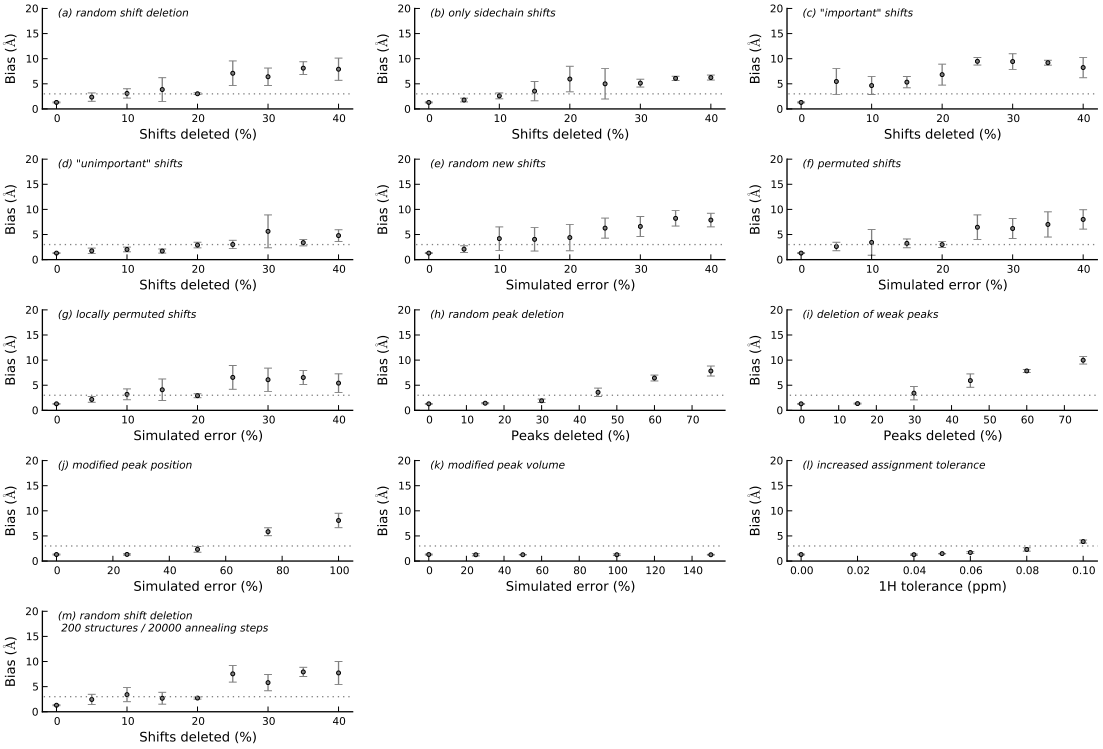


Figure A.5: RMSD to the reference structure for different input data modifications of the protein fspo.

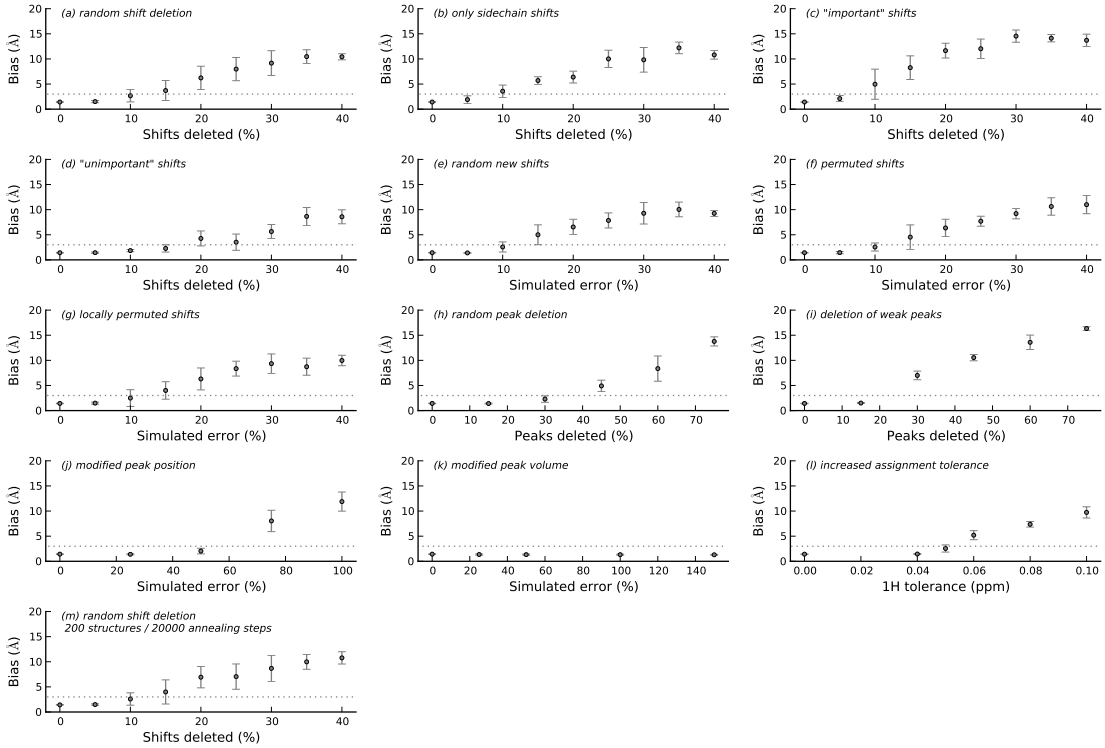


Figure A.6: RMSD to the reference structure for different input data modifications of the protein pbpa.

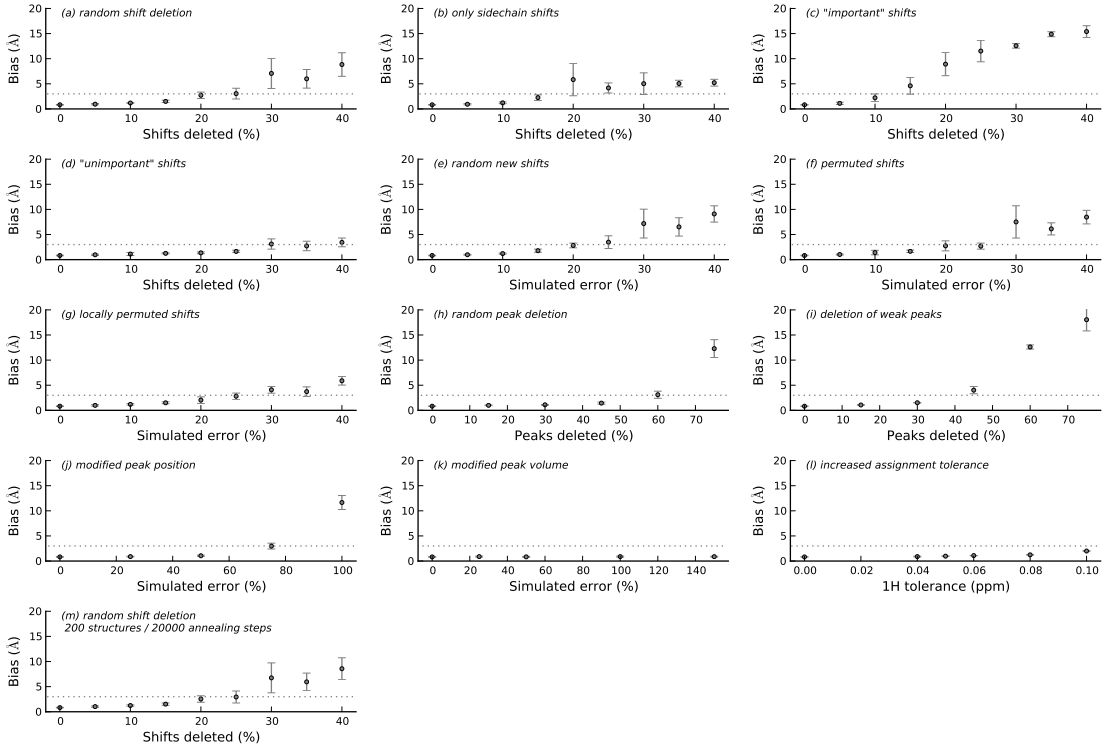


Figure A.7: RMSD to the reference structure for different input data modifications of the protein rhod.

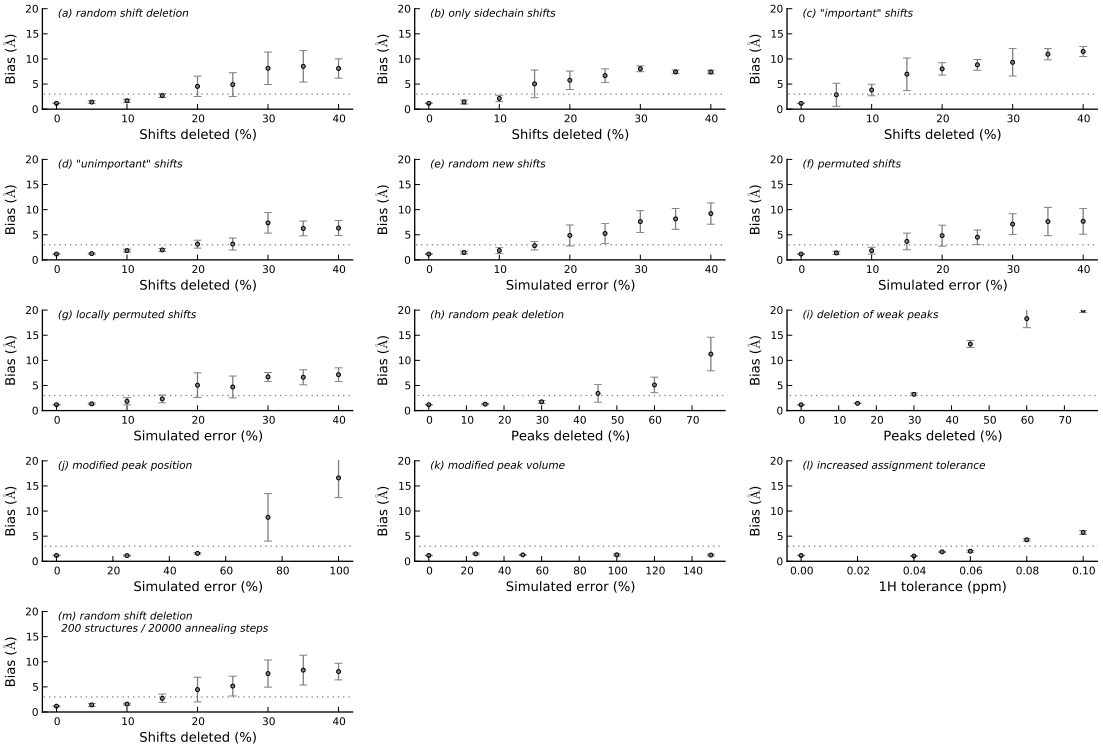


Figure A.8: RMSD to the reference structure for different input data modifications of the protein scam.

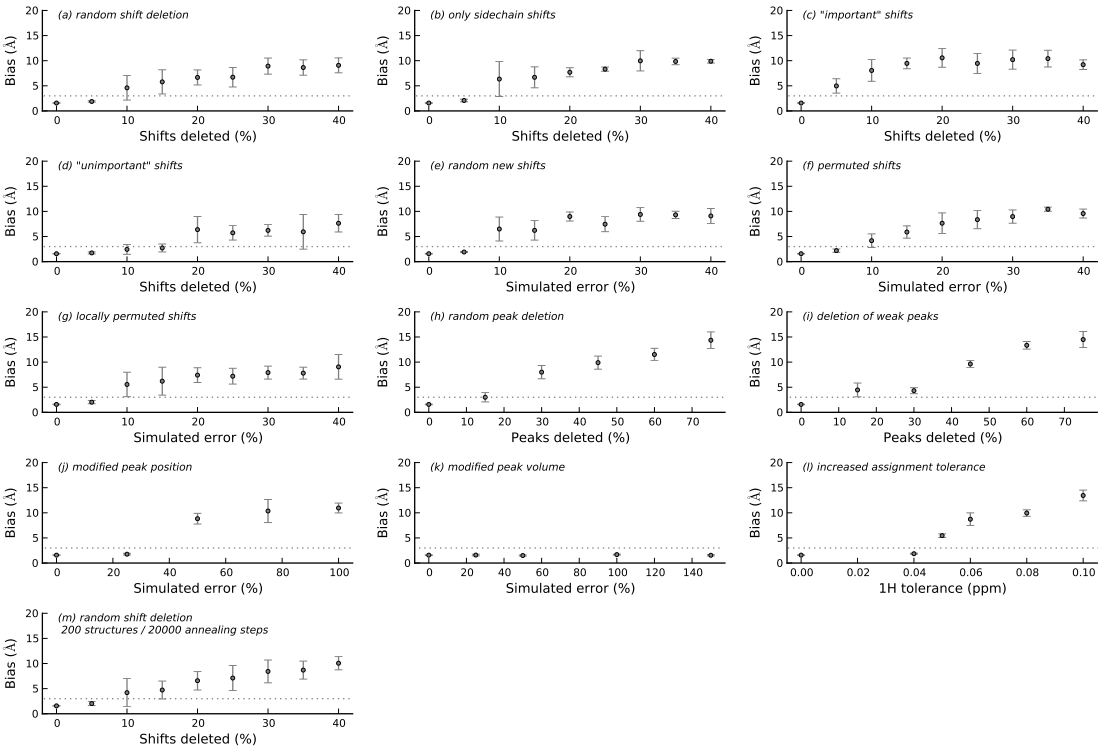


Figure A.9: RMSD to the reference structure for different input data modifications of the protein wmkt.

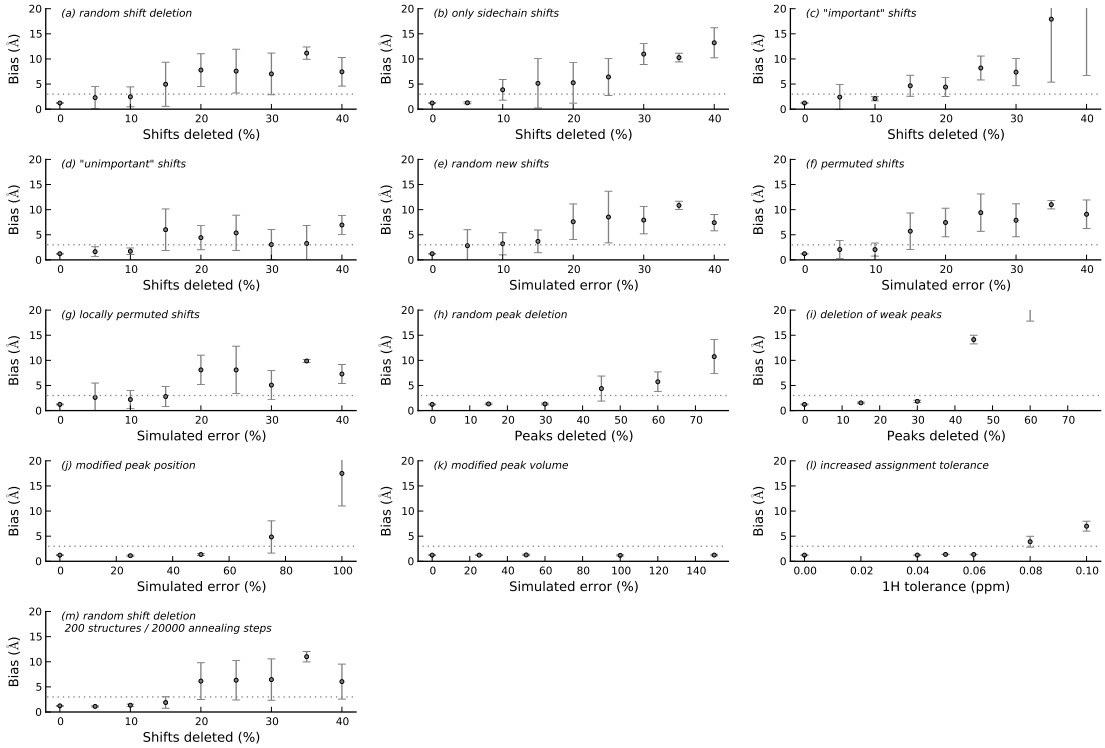


Figure A.10: RMSD to the reference structure for different input data modifications of the protein ww2d.

B GB1 sample preparation

The GB1 gene in a pET21a-vector including ampicillin resistance was kindly provided by Prof. Stephan Grzesiek (Biozentrum, University Basel) and the BL21-DE3 *E. coli* strain was used for recombinant protein expression. Sample preparation was performed based on a modified version of the protocol presented in Schmidt et al. (2007).

Expression

Freshly transformed cells cultivated on LB plates including 100 μg/ml ampicillin are used to inoculate 10 ml LB medium (including 100 μg/ml ampicillin) as an overnight culture per 1 l of expression medium. The overnight culture is partitioned to 15 ml Falcons and centrifuged for 10 min at 5000 rpm. The supernatant is discarded and the cell pellet is resuspended in 5 ml of the freshly prepared M9 Medium and transferred back to the flask. Cells are shaken at 37°C and the optical density at 600 nm is controlled until an OD_{600} of 0.9 is reached. Protein expression is induced using 0.5 mM IPTG (595,2 μl [200 mg/ml = 0,84 M] for 1 l). Expression is carried out for 3-4 h at 37°C. Cells are then harvested by centrifugation for 7 min at 6000 rpm. The cell pellet can be stored at -20°C until purification.

Table 6: M9 Medium for GB1 labeling

Ingredient[a]	u$^{13}C/^{15}N$	1,3$^{13}C/^{15}N$	2$^{13}C/^{15}N$
M9-Saltmix (5x)[b]	200 ml	200 ml	200 ml
$CaCl_2$ [0.01 M]	1 ml	1 ml	1 ml
$MgSO_4$ [1 M]	1 ml	1 ml	1 ml
Thiamine [10 mg/ml]	1 ml	1 ml	1 ml
Biotin [1 mg/ml]	10 ml	10 ml	10 ml
Ampcillin [100 mg/ml]	1 ml	1 ml	1 ml
Vitamin solution[c]	1 ml	1 ml	1 ml
Bioexpress (Euriso-top)	10 ml	10 ml	10 ml
$FeCl_3$ [0.01 M]	1 ml	1 ml	1 ml
u-^{13}C-Glucose	2 g		
1,3-^{13}C-Glycerol		2 g	
2-^{13}C-Glycerol			2 g
^{15}N-NH_4Cl	1 g	1 g	1 g

[a] 740 ml H_2O is autoclaved in an Erlenmeyer flask and the listed ingredients are added under sterile conditions. M9-saltmix (5x), $CaCl_2$ and $MgSO_4$ are autoclaved prior to usage. Thiamine- and Biotin-solutions are sterile filtered instead.

[b] 30 g Na_2HPO_4 (anhydrous), 15 g KH_2PO_4, 2.5 g NaCl filled to 1 l with H_2O and pH adjusted to 7.4.

[c] Vitamine solution from Centrum. 1.5 pills are pestled and filled to 20 ml with H_2O and mixed by vortexing for 20 min. Centrifugation for 5 min at 5,000 rpm and sterile filtration of the supernatant.

Heatshock

GB1 is a thermostable protein enabling a heat shock as a first step of purification. The cell pellet is resuspended in at least 15 ml of PBS buffer (50 mM Phosphate, 150 mM NaCl, pH 7,4) per 1 l of cell culture and the homogeneous cell suspension is incubated for 10 min at 80°C. The heatshock disrupts the cells and denatures all non-thermostable proteins. Cell debris and denatured proteins are separated from the soluble part by ultra-centrifugation at 10000 rpm for 30 min and the supernatant is prepared for gelfiltration. The maximum volume for gelfiltration is 5 ml, therefore the protein solution needs to be concentrated to 5 ml using an Amicon Stirred Cell with a 3 kDa Cutoff Filter or a comparable system for protein concentration.

Gelfiltration

A HiPrep 16/60 Sephacryl S-100 HR Column from GE Healthcare and the Äkta Prime System from Amersham Biosciences is used for gelfiltration. For purification H2O and phosphate buffer (50 mM Na-phosphate, pH 5,5) are needed (ultra-pure chemicals need to be used for labeled samples). All solutions for gelfiltration need to be sterile-filtered and degassed before usage. The system is washed the column is inserted under flow. The column is washed with at least one column volume of H2O (120 ml). After washing the column is equilibrated with one column volume of phosphate buffer. Washing and equilibration steps are performed at a flow rate of 0.5 ml/min. The sample loop is washed using phosphate buffer before inserting the sample. The fraction collector is filled with tubes and the program Prime view is started. The setting “Set Injection Valve Pos” is changed from “Load” to “Inj” for 6 ml until the sample is loaded onto the column. Afterwards it is set back to “Load” and the absorption is reset by pressing “Autozero”. Collection of fractions is started after 60 ml, the GB1 peak is expected between 70 and 90 ml. After purification the column is again washed with at least one column volume of H2O and one column volume of 20 % EtOH for storage at 4°C.

Concentration determination

All fractions of the GB1 peak are combined and the concentration is determined using a Bradford Assay. For the Bradford Assay, 20 μl sample (diluted 1:2 and 1:10) as well as 20 μl of a BSA standard (0 mg/ml, 0,4 mg/ml, 0,8 mg/ml, 1,2 mg/ml, 1,6 mg/ml, 2 mg/ml, 3 mg/ml, 4 mg/ml in H2O) are mixed with 100 μl Reagent A and 800 μl Reagent B and the mixture is incubated for 15 min. The absorption is measured at 750 nm for each sample and the BSA standard is used to calculate the concentration of the samples. The absorption needs to be in the range between 0.1 and 1.0 in order to obtain a reliable result.

Preparation of microcrystals

For precipitation, the GB1 sample needs to have a concentration of 25 mg/ml and no Cl-Ions should be present in the buffer. This is achieved by buffer exchange during gelfiltration. The GB1 solution in phosphate buffer (X ml, depending on the absolute amount of GB1) in a 15 ml-Falcon tube is incubated on ice in order to obtain a temperature of 4°C. X ml of precipitation solution (66.6 % MPD, 33.3 % 2-Propanol) are added, mixed by vortexing and incubated on ice for 7 min. Another X ml of precipitation solution is added, mixed by vortexing and incubated on ice again for 7 min. X/2 ml of precipitation solution are added, mixed by vortexing and incubated on ice for 7 min. At this stage, the solution should turn cloudy. X/2 ml of precipitation solution are added, mixed by vortexing and incubated over night at 4°C. Microcrystals can be centrifuged for 2 min at maximum 5,000 rpm and transferred into the rotor for NMR measurements.

Bibliography

Aeschbacher T., Schubert M., and Allain F. H.-T. A procedure to validate and correct the 13C chemical shift calibration of RNA datasets. *J Biomol NMR*, 52(2):179–190, 2012.

Aeschbacher T., Schmidt E., Blatter M., Maris C., Duss O., Allain F. H.-T., Güntert P., and Schubert M. Automated and assisted RNA resonance assignment using NMR chemical shift statistics. *Nucleic Acids Res*, 41(18):e172, 2013.

Allegrozzi M., Bertini I., Janik M. B. L., Lee Y.-M., Liu G., and Luchinat C. Lanthanide-induced pseudocontact shifts for solution structure refinements of macromolecules in shells up to 40 Å from the Metal Ion. *J Am Chem Soc*, 122(17):4154–4161, 2000.

Altieri A. S. and Byrd R. A. Automation of NMR structure determination of proteins. *Curr Opin Struct Biol*, 14(5):547–553, 2004.

Andrew E., Bradbury A., and Eades R. Nuclear magnetic resonance scpectra from a crystal rotated at high speed. *Nature*, 182(4650):1659, 1958.

Andrew E., Bradbury A., Eades R., and Wynn V. Nuclear cross-relaxation induced by specimen rotation. *Phys Lett*, 4(2):99 – 100, 1963.

Angelis A. A. D., Howell S. C., Nevzorov A. A., and Opella S. J. Structure determination of a membrane protein with two trans-membrane helices in aligned phospholipid bicelles by solid-state NMR spectroscopy. *J Am Chem Soc*, 128(37):12256–12267, 2006.

Antuch W., Güntert P., and Wüthrich K. Ancestral beta gamma-crystallin precursor structure in a yeast killer toxin. *Nat Struct Biol*, 3(8):662–665, 1996.

Bagaria A., Jaravine V., Huang Y. J., Montelione G. T., and Güntert P. Protein structure validation by generalized linear model root-mean-square deviation prediction. *Protein Sci*, 21(2):229–238, 2012.

Bagaria A., Jaravine V., and Güntert P. Estimating structure quality trends in the protein data bank by equivalent resolution. *Comput Biol Chem*, 46:8–15, 2013.

Bahrami A., Assadi A. H., Markley J. L., and Eghbalnia H. R. Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. *PLoS Comput Biol*, 5(3):e1000307, 2009.

Balayssac S., Bertini I., Bhaumik A., Lelli M., and Luchinat C. Paramagnetic shifts in solid-state NMR of proteins to elicit structural information. *Proc Natl Acad Sci U S A*, 105(45):17284–17289, 2008.

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

Bardiaux B., van Rossum B.-J., Nilges M., and Oschkinat H. Efficient modeling of symmetric protein aggregates from NMR data. *Angew Chem Int Ed Engl*, 51(28):6916–6919, 2012.

Bartels C., Xia T. H., Billeter M., Güntert P., and Wüthrich K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J Biomol NMR*, 6(1):1–10, 1995.

Bartels C., Güntert P., Billeter M., and Wüthrich K. GARANT - a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J Comput Chem*, 18(1):139–149, 1997.

Bayro M. J., Huber M., Ramachandran R., Davenport T. C., Meier B. H., Ernst M., and Griffin R. G. Dipolar truncation in magic-angle spinning NMR recoupling experiments. *J Chem Phys*, 130(11):114506, 2009.

Bennett A. E., Rienstra C. M., Auger M., Lakshmi K. V., and Griffin R. G. Heteronuclear decoupling in rotating solids. *J Chem Phys*, 103(16):6951–6958, 1995.

Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., and Bourne P. E. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242, 2000.

Bertini I., Bhaumik A., Paëpe G. D., Griffin R. G., Lelli M., Lewandowski J. R., and Luchinat C. High-resolution solid-state NMR structure of a 17.6 kDa protein. *J Am Chem Soc*, 132(3):1032–1040, 2010.

Billeter M., Wagner G., and Wüthrich K. Solution NMR structure determination of proteins revisited. *J Biomol NMR*, 42(3):155–158, 2008.

Bloch F. Dynamical theory of nuclear induction. II. *Phys Rev*, 102:104–135, 1956.

Bloch F. Theory of line narrowing by double-frequency irradiation. *Phys Rev*, 111:841–853, 1958.

Bowie J. U., Lüthy R., and Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170, 1991.

Brünger A. T., Adams P. D., Clore G. M., DeLano W. L., Gros P., Grosse-Kunstleve R. W., Jiang J. S., Kuszewski J., Nilges M., Pannu N. S., Read R. J., Rice L. M., Simonson T., and Warren G. L. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr*, 54 (Pt 5):905–921, 1998.

Calzolai L., Lysek D. A., Pérez D. R., Güntert P., and Wüthrich K. Prion protein NMR structures of chickens, turtles, and frogs. *Proc Natl Acad Sci U S A*, 102(3):651–655, 2005.

Castellani F., van Rossum B., Diehl A., Schubert M., Rehbein K., and Oschkinat H. Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. *Nature*, 420(6911):98–102, 2002.

Castellani F., van Rossum B.-J., Diehl A., Rehbein K., and Oschkinat H. Determination of solid-state NMR structures of proteins by means of three-dimensional 15N-13C-13C dipolar correlation spectroscopy and chemical shift analysis. *Biochemistry*, 42(39): 11476–11483, 2003.

Chan J. C. C. and Tycko R. Recoupling of chemical shift anisotropies in solid-state NMR under high-speed magic-angle spinning and in uniformly 13C-labeled systems. *J Chem Phys*, 118(18):8378–8389, 2003.

Chen V. B., Arendall W. B., Headd J. J., Keedy D. A., Immormino R. M., Kapral G. J., Murray L. W., Richardson J. S., and Richardson D. C. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*, 66 (Pt 1):12–21, 2010.

Clore G. M., Gronenborn A. M., and Bax A. 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(1):216–221, 1998.

Comellas G. and Rienstra C. M. Protein structure determination by magic-angle spinning solid-state NMR, and insights into the formation, structure, and stability of amyloid fibrils. *Annu Rev Biophys*, 42:515–536, 2013.

Cordier F., Nisius L., Dingley A. J., and Grzesiek S. Direct detection of N-H[...]O=C hydrogen bonds in biomolecules by NMR spectroscopy. *Nat Protoc*, 3(2):235–241, 2008.

Cornilescu G., Delaglio F., and Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR*, 13(3):289–302, 1999.

Coutandin D., Löhr F., Niesen F. H., Ikeya T., Weber T. A., Schäfer B., Zielonka E. M., Bullock A. N., Yang A., Güntert P., Knapp S., McKeon F., Ou H. D., and Dötsch V. Conformational stability and activity of p73 require a second helix in the tetramerization domain. *Cell Death Differ*, 16(12):1582–1589, 2009.

Creuzet F., McDermott A., Gebhard R., van der Hoef K., Spijker-Assink M. B., Herzfeld J., Lugtenburg J., Levitt M. H., and Griffin R. G. Determination of membrane protein structure by rotational resonance NMR: bacteriorhodopsin. *Science*, 251(4995):783–786, 1991.

Crick D., Wang J., Graham B., Swarbrick J., Mott H., and Nietlispach D. Integral membrane protein structure determination using pseudocontact shifts. *J Biomol NMR*, 61 (3-4):197–207, 2015.

Das B. B., Nothnagel H. J., Lu G. J., Son W. S., Tian Y., Marassi F. M., and Opella S. J. Structure determination of a membrane protein in proteoliposomes. *J Am Chem Soc*, 134(4):2047–2056, 2012.

Davis I. W., Murray L. W., Richardson J. S., and Richardson D. C. MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Research*, 32(suppl 2):W615–W619, 2004.

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

Delaglio F., Grzesiek S., Vuister G. W., Zhu G., Pfeifer J., and Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR*, 6 (3):277–293, 1995.

Doreleijers J. F., Nederveen A. J., Vranken W., Lin J., Bonvin A. M. J. J., Kaptein R., Markley J. L., and Ulrich E. L. BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. *J Biomol NMR*, 32(1):1–12, 2005.

Duggan B. M., Legge G. B., Dyson H. J., and Wright P. E. SANE (Structure Assisted NOE Evaluation): an automated model-based approach for NOE assignment. *J Biomol NMR*, 19(4):321–329, 2001.

Dumez J.-N. and Emsley L. A master-equation approach to the description of proton-driven spin diffusion from crystal geometry using simulated zero-quantum lineshapes. *Phys Chem Chem Phys*, 13(16):7363–7370, 2011.

Fossi M., Castellani F., Nilges M., Oschkinat H., and van Rossum B.-J. SOLARIA: a protocol for automated cross-peak assignment and structure calculation for solid-state magic-angle spinning NMR spectroscopy. *Angew Chem Int Ed Engl*, 44(38):6151–6154, 2005.

Franks W. T., Zhou D. H., Wylie B. J., Money B. G., Graesser D. T., Frericks H. L., Sahota G., and Rienstra C. M. Magic-angle spinning solid-state NMR spectroscopy of the beta1 immunoglobulin binding domain of protein G (GB1): 15N and 13C chemical shift assignments and conformational analysis. *J Am Chem Soc*, 127(35):12291–12305, 2005.

Franks W. T., Wylie B. J., Stellfox S. A., and Rienstra C. M. Backbone conformational constraints in a microcrystalline U-15N-labeled protein by 3D dipolar-shift solid-state NMR spectroscopy. *J Am Chem Soc*, 128(10):3154–3155, 2006.

Franks W. T., Wylie B. J., Schmidt H. L. F., Nieuwkoop A. J., Mayrhofer R.-M., Shah G. J., Graesser D. T., and Rienstra C. M. Dipole tensor-based atomic-resolution structure determination of a nanocrystalline protein by solid-state NMR. *Proc Natl Acad Sci U S A*, 105(12):4621–4626, 2008.

Fung B., Khitrin A., and Ermolaev K. An improved broadband decoupling sequence for liquid crystals and solids. *J Magn Reson*, 142(1):97 – 101, 2000.

Gaponenko V., Howarth J. W., Columbus L., Gasmi-Seabrook G., Yuan J., Hubbell W. L., and Rosevear P. R. Protein global fold determination using site-directed spin and isotope labeling. *Protein Sci*, 9(2):302–309, 2000.

Ginzinger S. W., Gerick F., Coles M., and Heun V. CheckShift: automatic correction of inconsistent chemical shift referencing. *J Biomol NMR*, 39(3):223–227, 2007.

Goto N. K., Gardner K. H., Mueller G. A., Willis R. C., and Kay L. E. A robust and cost-effective method for the production of Val, Leu, Ile (delta 1) methyl-protonated 15N-, 13C-, 2H-labeled proteins. *J Biomol NMR*, 13(4):369–374, 1999.

Gottstein D., Kirchner D. K., and Güntert P. Simultaneous single-structure and bundle representation of protein NMR structures in torsion angle space. *J Biomol NMR*, 52 (4):351–364, 2012a.

Gottstein D., Reckel S., Dötsch V., and Güntert P. Requirements on paramagnetic relaxation enhancement data for membrane protein structure determination by NMR. *Structure*, 20(6):1019–1027, 2012b.

Grommek A., Meier B. H., and Ernst M. Distance information from proton-driven spin diffusion under MAS. *Chem Phys Lett*, 427(4-6):404–409, 2006.

Gronwald W. and Kalbitzer H. R. Automated structure determination of proteins by NMR spectroscopy. *Prog Nucl Magn Reson Spectrosc*, 44:33 – 96, 2004.

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

Guerry P. and Herrmann T. Advances in automated NMR protein structure determination. *Q Rev Biophys*, 44(3):257–309, 2011.

Gullion T. and Schaefer J. Rotational-echo double-resonance NMR. *J Magn Reson*, 81 (1):196 – 200, 1989.

Güntert P. Structure calculation of biological macromolecules from NMR data. *Q Rev Biophys*, 31(2):145–237, 1998.

Güntert P., Braun W., and Wüthrich K. Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *J Mol Biol*, 217(3):517–530, 1991.

Güntert P., Mumenthaler C., and Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol*, 273(1):283–298, 1997.

Güntert P. Automated NMR protein structure calculation. *Prog Nucl Magn Reson Spectrosc*, 43:105 – 125, 2003.

Güntert P. Automated structure determination from NMR spectra. *Eur Biophys J*, 38 (2):129–143, 2009.

Güntert P. Calculation of structures from NMR restraints. In *Protein NMR Spectroscopy: Practical Techniques and Applications*, pages 159–192. Wiley, 2011.

Güntert P. and Buchner L. Combined automated NOE assignment and structure calculation with CYANA. *J Biomol NMR*, 2015.

Güntert P., Berndt K. D., and Wüthrich K. The program ASNO for computer-supported collection of NOE upper distance constraints as input for protein structure determination. *Journal of Biomolecular NMR*, 3(5):601–606, 1993.

Hartmann S. R. and Hahn E. L. Nuclear double resonance in the rotating frame. *Phys Rev*, 128:2042–2053, 1962.

Havlin R. H., Laws D. D., Bitter H. M., Sanders L. K., Sun H., Grimley J. S., Wemmer D. E., Pines A., and Oldfield E. An experimental and theoretical investigation of the chemical shielding tensors of (13)C(alpha) of alanine, valine, and leucine residues in solid peptides and in proteins in solution. *J Am Chem Soc*, 123(42):10362–10369, 2001.

Helmus J. J., Nadaud P. S., Höfer N., and Jaroniec C. P. Determination of methyl 13C-15N dipolar couplings in peptides and proteins by three-dimensional and four-dimensional magic-angle spinning solid-state NMR spectroscopy. *J Chem Phys*, 128(5):052314, 2008.

Herrmann T., Güntert P., and Wüthrich K. Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J Biomol NMR*, 24(3):171–189, 2002a.

Herrmann T., Güntert P., and Wüthrich K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol*, 319(1):209–227, 2002b.

Hing A. W., Vega S., and Schaefer J. Transferred-echo double-resonance NMR. *J Magn Reson*, 96(1):205 – 209, 1992.

Hohwy M., Jaroniec C. P., Reif B., Rienstra C. M., and Griffin R. G. Local structure and relaxation in solid-state NMR: accurate measurement of amide N-H bond lengths and H-N-H bond angles. *J Am Chem Soc*, 122(13):3218–3219, 2000.

Hooft R. W., Vriend G., Sander C., and Abola E. E. Errors in protein structures. *Nature*, 381(6580):272, 1996.

Horst R., Damberger F., Luginbühl P., Güntert P., Peng G., Nikonova L., Leal W. S., and Wüthrich K. NMR structure reveals intramolecular regulation mechanism for pheromone binding and release. *Proc Natl Acad Sci U S A*, 98(25):14374–14379, 2001.

Huang Y. J., Moseley H. N. B., Baran M. C., Arrowsmith C., Powers R., Tejero R., Szyperski T., and Montelione G. T. An integrated platform for automated analysis of protein NMR structures. *Methods Enzymol*, 394:111–141, 2005.

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

Huber M., Hiller S., Schanda P., Ernst M., Bökmann A., Verel R., and Meier B. H. A proton-detected 4D solid-state NMR experiment for protein structure determination. *Chemphyschem*, 12(5):915–918, 2011.

Hung L.-H. and Samudrala R. An automated assignment-free Bayesian approach for accurately identifying proton contacts from NOESY data. *J Biomol NMR*, 36(3):189–198, 2006.

Ikeya T., Takeda M., Yoshida H., Terauchi T., Jee J.-G., Kainosho M., and Güntert P. Automated NMR structure determination of stereo-array isotope labeled ubiquitin from minimal sets of spectra using the SAIL-FLYA system. *J Biomol NMR*, 44(4):261–272, 2009.

Ikeya T., Jee J.-G., Shigemitsu Y., Hamatsu J., Mishima M., Ito Y., Kainosho M., and Güntert P. Exclusively NOESY-based automated NMR assignment and structure determination of proteins. *J Biomol NMR*, 50(2):137–146, 2011.

Janik R., Peng X., and Ladizhansky V. (13)C-(13)C distance measurements in U-(13)C, (15)N-labeled peptides using rotational resonance width experiment with a homogeneously broadened matching condition. *J Magn Reson*, 188(1):129–140, 2007.

Jaroniec C. P., Tounge B. A., Rienstra C. M., Herzfeld J., and Griffin R. G. Measurement of 13C-15N distances in uniformly 13C labeled biomolecules: J-decoupled REDOR. *J Am Chem Soc*, 121(43):10237–10238, 1999.

Jaroniec C. P., Tounge B. A., Herzfeld J., and Griffin R. G. Frequency selective heteronuclear dipolar recoupling in rotating solids: accurate (13)C-(15)N distance measurements in uniformly (13)C,(15)N-labeled peptides. *J Am Chem Soc*, 123(15):3507–3519, 2001.

Jaroniec C. P., Filip C., and Griffin R. G. 3D TEDOR NMR experiments for the simultaneous measurement of multiple carbon-nitrogen distances in uniformly (13)C,(15)N-labeled solids. *J Am Chem Soc*, 124(36):10728–10742, 2002.

Jee J. and Güntert P. Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. *J Struct Funct Genomics*, 4(2-3): 179–189, 2003.

Jehle S., Rajagopal P., Bardiaux B., Markovic S., Kühne R., Stout J. R., Higman V. A., Klevit R. E., van Rossum B.-J., and Oschkinat H. Solid-state NMR and SAXS studies

provide a structural basis for the activation of alphaB-crystallin oligomers. *Nat Struct Mol Biol*, 17(9):1037–1042, 2010.

Johnson B. A. Using NMRView to visualize and analyze the NMR spectra of macromolecules. *Methods Mol Biol*, 278:313–352, 2004.

Kainosho M. and Güntert P. SAIL–stereo-array isotope labeling. *Q Rev Biophys*, 42(4): 247–300, 2009.

Kainosho M., Torizawa T., Iwashita Y., Terauchi T., Ono A. M., and Güntert P. Optimal isotope labelling for NMR protein structure determinations. *Nature*, 440(7080):52–57, 2006.

Kalk A. and Berendsen H. Proton magnetic relaxation and spin diffusion in proteins. *J Magn Reson*, 24(3):343 – 366, 1976.

Karplus M. Contact electron-spin coupling of nuclear magnetic moments. *J Chem Phy*, 30(1):11–15, 1959.

Karplus M. Vicinal proton coupling in nuclear magnetic resonance. *J Am Chem Soc*, 85 (18):2870–2871, 1963.

Keeler J. *Understanding NMR Spectroscopy*. Wiley, 2005.

Keepers J. W. and James T. L. A theoretical study of distance determinations from NMR. Two-dimensional nuclear overhauser effect spectra. *J Magn Reson*, 57(3):404 – 426, 1984.

Ketchem R. R., Hu W., and Cross T. A. High-resolution conformation of gramicidin A in a lipid bilayer by solid-state NMR. *Science*, 261(5127):1457–1460, 1993.

Kigawa T., Muto Y., and Yokoyama S. Cell-free synthesis and amino acid-selective stable isotope labeling of proteins for NMR analysis. *J Biomol NMR*, 6(2):129–134, 1995.

Kim Y. and Prestegard J. H. Refinement of the NMR structures for acyl carrier protein with scalar coupling data. *Proteins*, 8(4):377–385, 1990.

Kirchner D. K. and Güntert P. Objective identification of residue ranges for the superposition of protein structures. *BMC Bioinformatics*, 12:170, 2011.

Knight M. J., Pell A. J., Bertini I., Felli I. C., Gonnelli L., Pierattelli R., Herrmann T., Emsley L., and Pintacuda G. Structure and backbone dynamics of a microcrystalline metalloprotein by solid-state NMR. *Proc Natl Acad Sci U S A*, 109(28):11095–11100, 2012.

Koga N., Tatsumi-Koga R., Liu G., Xiao R., Acton T. B., Montelione G. T., and Baker D. Principles for designing ideal protein structures. *Nature*, 491(7423):222–227, 2012.

Krähenbühl B., Bakkali I. E., Schmidt E., Güntert P., and Wider G. Automated NMR resonance assignment strategy for RNA via the phosphodiester backbone based on high-dimensional through-bond APSY experiments. *J Biomol NMR*, 59(2):87–93, 2014.

Kubo A. and McDowell C. A. Spectral spin diffusion in polycrystalline solids under magic-angle spinning. *J Chem Soc Faraday Trans 1*, 84:3713–3730, 1988.

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

Kuszewski J. J., Thottungal R. A., Clore G. M., and Schwieters C. D. Automated error-tolerant macromolecular structure determination from multidimensional nuclear Overhauser enhancement spectra and chemical shift assignments: improved robustness and performance of the PASD algorithm. *J Biomol NMR*, 41(4):221–239, 2008.

Kwan A. H., Mobli M., Gooley P. R., King G. F., and Mackay J. P. Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS J*, 278(5):687–703, 2011.

Lange A., Luca S., and Baldus M. Structural constraints from proton-mediated rare-spin correlation spectroscopy in rotating solids. *J Am Chem Soc*, 124(33):9704–9705, 2002.

Lange A., Seidel K., Verdier L., Luca S., and Baldus M. Analysis of proton-proton transfer dynamics in rotating solids and their use for 3D structure determination. *J Am Chem Soc*, 125(41):12640–12648, 2003.

Lange A., Becker S., Seidel K., Giller K., Pongs O., and Baldus M. A concept for rapid protein-structure determination by solid-state NMR spectroscopy. *Angew Chem Int Ed Engl*, 44(14):2089–2092, 2005.

Laskowski R. A., Rullmann J. A. C., MacArthur M. W., Kaptein R., and Thornton J. M. AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, 8(4):477–486, 1996.

Laws D. D., Bitter H.-M. L., and Jerschow A. Solid-state NMR spectroscopic methods in chemistry. *Angew Chem Int Ed Engl*, 41(17):3096–3129, 2002.

LeMaster D. M. and Kushlan D. M. Dynamical mapping of e. coli thioredoxin via 13C NMR relaxation analysis. *J Am Chem Soc*, 118(39):9255–9264, 1996.

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

Linge J. P., Habeck M., Rieping W., and Nilges M. Correction of spin diffusion during iterative automated NOE assignment. *J Magn Reson*, 167(2):334–342, 2004.

López-Méndez B. and Güntert P. Automated protein structure determination from NMR spectra. *J Am Chem Soc*, 128(40):13112–13122, 2006.

López-Méndez B., Pantoja-Uceda D., Tomizawa T., Koshiba S., Kigawa T., Shirouzu M., Terada T., Inoue M., Yabuki T., Aoki M., Seki E., Matsuda T., Hirota H., Yoshida M., Tanaka A., Osanai T., Seki M., Shinozaki K., Yokoyama S., and Güntert P. NMR assignment of the hypothetical ENTH-VHS domain At3g16270 from Arabidopsis thaliana. *J Biomol NMR*, 29(2):205–206, 2004.

Loquet A., Bardiaux B., Gardiennet C., Blanchet C., Baldus M., Nilges M., Malliavin T., and Böckmann A. 3D structure determination of the Crh protein from highly ambiguous solid-state NMR restraints. *J Am Chem Soc*, 130(11):3579–3589, 2008.

Loquet A., Sgourakis N. G., Gupta R., Giller K., Riedel D., Goosmann C., Griesinger C., Kolbe M., Baker D., Becker S., and Lange A. Atomic model of the type III secretion system needle. *Nature*, 486(7402):276–279, 2012.

Lowe I. J. Free induction decays of rotating solids. *Phys Rev Lett*, 2:285–287, 1959.

Lu J.-X., Qiang W., Yau W.-M., Schwieters C. D., Meredith S. C., and Tycko R. Molecular structure of β-amyloid fibrils in Alzheimer's disease brain tissue. *Cell*, 154(6):1257–1268, 2013.

Luginbühl P., Szyperski T., and Wüthrich K. Statistical basis for the use of 13Cα chemical shifts in protein structure determination. *J Magn Reson B*, 109(2):229 – 233, 1995.

Lüthy R., Bowie J. U., and Eisenberg D. Assessment of protein models with three-dimensional profiles. *Nature*, 356(6364):83–85, 1992.

Manolikas T., Herrmann T., and Meier B. H. Protein structure determination from 13C spin-diffusion solid-state NMR spectroscopy. *J Am Chem Soc*, 130(12):3959–3966, 2008.

Marassi F. M. and Opella S. J. Simultaneous assignment and structure determination of a membrane protein from NMR orientational restraints. *Protein Sci*, 12(3):403–411, 2003.

Marin-Montesinos I., Mollica G., Carravetta M., Gansmüller A., Pileio G., Bechmann M., Sebald A., and Levitt M. H. Truncated dipolar recoupling in solid-state nuclear magnetic resonance. *Chem Phys Lett*, 432(4-6):572 – 578, 2006.

Meadows R. P., Olejniczak E. T., and Fesik S. W. A computer-based protocol for semi-automated assignments and 3D structure determination of proteins. *J Biomol NMR*, 4 (1):79–96, 1994.

Melckebeke H. V., Wasmer C., Lange A., Ab E., Loquet A., Böckmann A., and Meier B. H. Atomic-resolution three-dimensional structure of HET-s(218-289) amyloid fibrils by solid-state NMR spectroscopy. *J Am Chem Soc*, 132(39):13765–13775, 2010.

Michal C. A. and Jelinski L. W. REDOR 3D: heteronuclear distance measurements in uniformly labeled and natural abundance solids. *J Am Chem Soc*, 119(38):9059–9060, 1997.

Minch M. J. Orientational dependence of vicinal proton-proton NMR coupling constants: The Karplus relationship. *Concepts Magn Reson*, 6(1):41–56, 1994.

Morris A. L., MacArthur M. W., Hutchinson E. G., and Thornton J. M. Stereochemical quality of protein structure coordinates. *Proteins*, 12(4):345–364, 1992.

Moseley H. N. and Montelione G. T. Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struct Biol*, 9(5):635–642, 1999.

Mumenthaler C. and Braun W. Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *J Mol Biol*, 254(3):465–480, 1995.

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

Nabuurs S. B., Spronk C. A. E. M., Krieger E., Maassen H., Vriend G., and Vuister G. W. Quantitative evaluation of experimental NMR restraints. *J Am Chem Soc*, 125(39): 12026–12034, 2003.

Nabuurs S. B., Spronk C. A. E. M., Vuister G. W., and Vriend G. Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. *PLoS Comput Biol*, 2(2):e9, 2006.

Nadaud P. S., Helmus J. J., Höfer N., and Jaroniec C. P. Long-range structural restraints in spin-labeled proteins probed by solid-state nuclear magnetic resonance spectroscopy. *J Am Chem Soc*, 129(24):7502–7503, 2007.

Nelder J. A. and Mead R. A simplex method for function minimization. *Comput J*, 7(4): 308–313, 1965.

Nieuwkoop A. J. and Rienstra C. M. Supramolecular protein structure determination by site-specific long-range intermolecular solid state NMR spectroscopy. *J Am Chem Soc*, 132(22):7570–7571, 2010.

Nieuwkoop A. J., Wylie B. J., Franks W. T., Shah G. J., and Rienstra C. M. Atomic resolution protein structure determination by three-dimensional transferred echo double resonance solid-state nuclear magnetic resonance spectroscopy. *J Chem Phys*, 131(9): 095101, 2009.

Nilges M. Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J Mol Biol*, 245(5):645–660, 1995.

Nilges M., Macias M. J., O'Donoghue S. I., and Oschkinat H. Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. *J Mol Biol*, 269(3):408–422, 1997.

Ohnishi S., Güntert P., Koshiba S., Tomizawa T., Akasaka R., Tochio N., Sato M., Inoue M., Harada T., Watanabe S., Tanaka A., Shirouzu M., Kigawa T., and Yokoyama S. Solution structure of an atypical WW domain in a novel beta-clam-like dimeric form. *FEBS Lett*, 581(3):462–468, 2007.

Opella S. J., Marassi F. M., Gesell J. J., Valente A. P., Kim Y., Oblatt-Montal M., and Montal M. Structures of the M2 channel-lining segments from nicotinic acetylcholine and NMDA receptors by NMR spectroscopy. *Nat Struct Biol*, 6(4):374–379, 1999.

Orts J., Vögeli B., and Riek R. Relaxation matrix analysis of spin diffusion for the NMR structure calculation with eNOEs. *J Chem Theory Comput*, 8(10):3483–3492, 2012.

Ottiger M., Delaglio F., and Bax A. Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. *J Magn Reson*, 131(2):373–378, 1998.

Pääkkönen K., Tossavainen H., Permi P., Rakkolainen H., Rauvala H., Raulo E., Kilpeläinen I., and Güntert P. Solution structures of the first and fourth TSR domains of F-spondin. *Proteins*, 64(3):665–672, 2006.

Paëpe G. D., Lewandowski J. R., Loquet A., Böckmann A., and Griffin R. G. Proton assisted recoupling and protein structure determination. *J Chem Phys*, 129(24):245101, 2008.

Pantoja-Uceda D., López-Méndez B., Koshiba S., Kigawa T., Shirouzu M., Terada T., Inoue M., Yabuki T., Aoki M., Seki E., Matsuda T., Hirota H., Yoshida M., Tanaka A., Osanai T., Seki M., Shinozaki K., Yokoyama S., and Güntert P. NMR assignment

of the hypothetical rhodanese domain At4g01050 from Arabidopsis thaliana. *J Biomol NMR*, 29(2):207–208, 2004.

Pantoja-Uceda D., López-Méndez B., Koshiba S., Inoue M., Kigawa T., Terada T., Shirouzu M., Tanaka A., Seki M., Shinozaki K., Yokoyama S., and Güntert P. Solution structure of the rhodanese homology domain At4g01050(175-295) from Arabidopsis thaliana. *Protein Sci*, 14(1):224–230, 2005.

Park S. H., Mrse A. A., Nevzorov A. A., Mesleh M. F., Oblatt-Montal M., Montal M., and Opella S. J. Three-dimensional structure of the channel-forming trans-membrane domain of virus protein "u" (Vpu) from HIV-1. *J Mol Biol*, 333(2):409–424, 2003.

Park S. H., Das B. B., Casagrande F., Tian Y., Nothnagel H. J., Chu M., Kiefer H., Maier K., Angelis A. A. D., Marassi F. M., and Opella S. J. Structure of the chemokine receptor CXCR1 in phospholipid bilayers. *Nature*, 491(7426):779–783, 2012.

Pearson J. G., Le H., Sanders L. K., Godbout N., Havlin R. H., and Oldfield E. Predicting chemical shifts in proteins: structure refinement of valine residues by using ab initio and empirical geometry optimizations. *J Am Chem Soc*, 119(49):11941–11950, 1997.

Peng X., Libich D., Janik R., Harauz G., and Ladizhansky V. Dipolar chemical shift correlation spectroscopy for homonuclear carbon distance measurements in proteins in the solid state: application to structure determination and refinement. *J Am Chem Soc*, 130(1):359–369, 2008.

Prestegard J. H., Bougault C. M., and Kishore A. I. Residual dipolar couplings in structure determination of biomolecules. *Chem Rev*, 104(8):3519–3540, 2004.

Ramachandran R., Ladizhansky V., Bajaj V. S., and Griffin R. G. 13C-13C rotational resonance width distance measurements in uniformly 13C-labeled peptides. *J Am Chem Soc*, 125(50):15623–15629, 2003.

Ramachandran R., Lewandowski J. R., van der Wel P. C. A., and Griffin R. G. Multipole-multimode Floquet theory of rotational resonance width experiments: 13C-13C distance measurements in uniformly labeled solids. *J Chem Phys*, 124(21):214107, 2006.

Reif B., Hohwy M., Jaroniec C., Rienstra C., and Griffin R. NH-NH vector correlation in peptides by solid-statetate NMR. *J Magn Reson*, 145(1):132 – 141, 2000.

Richard R. Ernst G. B. and Wokaun A. *Principles of nuclear magnetic resonance in one and two dimensions*. Oxford University Press, 1987.

Rienstra C. M., Hohwy M., Mueller L. J., Jaroniec C. P., Reif B., and Griffin R. G. Determination of multiple torsion-angle constraints in U-(13)C,(15)N-labeled peptides: 3D (1)H-(15)N-(13)C-(1)H dipolar chemical shift NMR spectroscopy in rotating solids. *J Am Chem Soc*, 124(40):11908–11922, 2002.

Rieping W., Habeck M., and Nilges M. Inferential structure determination. *Science*, 309 (5732):303–306, 2005.

Rieping W., Habeck M., Bardiaux B., Bernard A., Malliavin T. E., and Nilges M. ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics*, 23(3):381–382, 2007.

Rocco J. W. and Ellisen L. W. p63 and p73: life and death in squamous cell carcinoma. *Cell Cycle*, 5(9):936–940, 2006.

Rosato A., Bagaria A., Baker D., Bardiaux B., Cavalli A., Doreleijers J. F., Giachetti A., Guerry P., Güntert P., Herrmann T., Huang Y. J., Jonker H. R. A., Mao B., Malliavin T. E., Montelione G. T., Nilges M., Raman S., van der Schot G., Vranken W. F., Vuister G. W., and Bonvin A. M. J. J. CASD-NMR: critical assessment of automated structure determination by NMR. *Nat Methods*, 6(9):625–626, 2009.

Rosato A., Aramini J. M., Arrowsmith C., Bagaria A., Baker D., Cavalli A., Doreleijers J. F., Eletsky A., Giachetti A., Guerry P., Gutmanas A., Güntert P., He Y., Herrmann T., Huang Y. J., Jaravine V., Jonker H. R. A., Kennedy M. A., Lange O. F., Liu G., Malliavin T. E., Mani R., Mao B., Montelione G. T., Nilges M., Rossi P., van der Schot G., Schwalbe H., Szyperski T. A., Vendruscolo M., Vernon R., Vranken W. F., de Vries S., Vuister G. W., Wu B., Yang Y., and Bonvin A. M. J. J. Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure*, 20(2): 227–236, 2012.

Saccenti E. and Rosato A. The war of tools: how can NMR spectroscopists detect errors in their structures? *J Biomol NMR*, 40(4):251–261, 2008.

Schmidt E. and Güntert P. A new algorithm for reliable and general NMR resonance assignment. *J Am Chem Soc*, 134(30):12817–12829, 2012.

Schmidt E. and Güntert P. Reliability of exclusively NOESY-based automated resonance assignment and structure determination of proteins. *J Biomol NMR*, 57(2):193–204, 2013.

Schmidt E., Gath J., Habenstein B., Ravotti F., Székely K., Huber M., Buchner L., Böckmann A., Meier B. H., and Güntert P. Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. *J Biomol NMR*, 56(3):243–254, 2013.

Schmidt H. L. F., Sperling L. J., Gao Y. G., Wylie B. J., Boettcher J. M., Wilson S. R., and Rienstra C. M. Crystal polymorphism of protein gb1 examined by solid-state nmr spectroscopy and x-ray diffraction. *J Phys Chem B*, 111(51):14362–14369, 2007.

Schmucki R., Yokoyama S., and Güntert P. Automated assignment of NMR chemical shifts using peak-particle dynamics simulation with the DYNASSIGN algorithm. *J Biomol NMR*, 43(2):97–109, 2009.

Schütz A. K., Vagt T., Huber M., Ovchinnikova O. Y., Cadalbert R., Wall J., Güntert P., Böckmann A., Glockshuber R., and Meier B. H. Atomic-Resolution Three-Dimensional Structure of Amyloid β Fibrils Bearing the Osaka Mutation. *Angew Chem Int Ed Engl*, 2014.

Schwieters C. D., Kuszewski J. J., Tjandra N., and Clore G. M. The Xplor-NIH NMR molecular structure determination package. *J Magn Reson*, 160(1):65–73, 2003.

Scott A., Pantoja-Uceda D., Koshiba S., Inoue M., Kigawa T., Terada T., Shirouzu M., Tanaka A., Sugano S., Yokoyama S., and Güntert P. NMR assignment of the SH2 domain from the human feline sarcoma oncogene FES. *J Biomol NMR*, 30(4):463–464, 2004.

Scott A., Pantoja-Uceda D., Koshiba S., Inoue M., Kigawa T., Terada T., Shirouzu M., Tanaka A., Sugano S., Yokoyama S., and Güntert P. Solution structure of the Src homology 2 domain from the human feline sarcoma oncogene Fes. *J Biomol NMR*, 31 (4):357–361, 2005.

Sengupta I., Nadaud P. S., Helmus J. J., Schwieters C. D., and Jaroniec C. P. Protein fold determined by paramagnetic magic-angle spinning solid-state NMR spectroscopy. *Nat Chem*, 4(5):410–417, 2012.

Sengupta I., Nadaud P. S., and Jaroniec C. P. Protein structure determination with paramagnetic solid-state NMR spectroscopy. *Acc Chem Res*, 46(9):2117–2126, 2013.

Serrano P., Pedrini B., Mohanty B., Geralt M., Herrmann T., and Wüthrich K. The J-UNIO protocol for automated protein structure determination by NMR in solution. *J Biomol NMR*, 53(4):341–354, 2012.

Shahid S. A., Bardiaux B., Franks W. T., Krabben L., Habeck M., van Rossum B.-J., and Linke D. Membrane-protein structure determination by solid-state NMR spectroscopy of microcrystals. *Nat Methods*, 9(12):1212–1217, 2012.

Sharma M., Yi M., Dong H., Qin H., Peterson E., Busath D. D., Zhou H.-X., and Cross T. A. Insight into the mechanism of the influenza A proton channel from a structure in a lipid bilayer. *Science*, 330(6003):509–512, 2010.

Shen Y. and Bax A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J Biomol NMR*, 56(3):227–241, 2013.

Shen Y., Lange O., Delaglio F., Rossi P., Aramini J. M., Liu G., Eletsky A., Wu Y., Singarapu K. K., Lemak A., Ignatchenko A., Arrowsmith C. H., Szyperski T., Montelione G. T., Baker D., and Bax A. Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci U S A*, 105(12):4685–4690, 2008.

Shen Y., Delaglio F., Cornilescu G., and Bax A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR*, 44(4):213–223, 2009.

Shin J., Lee W., and Lee W. Structural proteomics by NMR spectroscopy. *Expert Rev Proteomics*, 5(4):589–601, 2008.

Sippl M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17 (4):355–362, 1993.

Solomon I. Relaxation processes in a system of two spins. *Phys Rev*, 99:559–565, 1955.

Spera S. and Bax A. Empirical correlation between protein backbone conformation and C.alpha. and C.beta. 13C nuclear magnetic resonance chemical shifts. *J Am Chem Soc*, 113(14):5490–5492, 1991.

Spronk C. A. E. M., Nabuurs S. B., Bonvin A. M. J. J., Krieger E., Vuister G. W., and Vriend G. The precision of NMR structure ensembles revisited. *J Biomol NMR*, 25(3): 225–234, 2003.

Spronk C. A., Nabuurs S. B., Krieger E., Vriend G., and Vuister G. W. Validation of protein structures derived by NMR spectroscopy. *Progr Nucl Magn Reson*, 45(3-4): 315–337, 2004.

Straasø L. A., Bjerring M., Khaneja N., and Nielsen N. C. Multiple-oscillating-field techniques for accurate distance measurements by solid-state NMR. *J Chem Phys*, 130(22): 225103, 2009.

Straasø L. A., Nielsen J. T., Bjerring M., Khaneja N., and Nielsen N. C. Accurate measurements of ^{13}C-^{13}C distances in uniformly ^{13}C-labeled proteins using multi-dimensional four-oscillating field solid-state NMR spectroscopy. *J Chem Phys*, 141(11):114201, 2014.

Sun H., Sanders L. K., and Oldfield E. Carbon-13 NMR shielding in the twenty common amino acids: comparisons with experimental results in proteins. *J Am Chem Soc*, 124 (19):5486–5495, 2002.

Takeda M. and Kainosho M. Isotope Labelling. In *Protein NMR Spectroscopy: Practical Techniques and Applications*, pages 23–53. Wiley, 1st edition, 2011.

Takegoshi K., Nakamura S., and Terao T. 13C-1H dipolar-assisted rotational resonance in magic-angle spinning NMR. *Chem Phys Lett*, 344(5-6):631 – 637, 2001.

Takegoshi K., Nakamura S., and Terao T. 13C-1H dipolar-driven 13C-13C recoupling without 13C rf irradiation in nuclear magnetic resonance of rotating solids. *J Chem Phys*, 118(5):2325–2341, 2003.

Tang M., Nesbitt A. E., Sperling L. J., Berthold D. A., Schwieters C. D., Gennis R. B., and Rienstra C. M. Structure of the disulfide bond generating membrane protein DsbB in the lipid bilayer. *J Mol Biol*, 425(10):1670–1682, 2013.

Torda A. E., Brunne R. M., Huber T., Kessler H., and van Gunsteren W. F. Structure refinement using time-averaged J-coupling constant restraints. *J Biomol NMR*, 3(1): 55–66, 1993.

Torizawa T., Shimizu M., Taoka M., Miyano H., and Kainosho M. Efficient production of isotopically labeled proteins by cell-free synthesis: a practical protocol. *J Biomol NMR*, 30(3):311–325, 2004.

Tycko R. Stochastic dipolar recoupling in nuclear magnetic resonance of solids. *Phys Rev Lett*, 99(18):187601, 2007.

van der Wel P. C. A., Eddy M. T., Ramachandran R., and Griffin R. G. Targeted 13C-13C distance measurements in a microcrystalline protein via J-decoupled rotational resonance width measurements. *Chemphyschem*, 10(9-10):1656–1663, 2009.

Vögeli B., Segawa T. F., Leitz D., Sobol A., Choutko A., Trzesniak D., van Gunsteren W., and Riek R. Exact distances and internal dynamics of perdeuterated ubiquitin from NOE buildups. *J Am Chem Soc*, 131(47):17215–17225, 2009.

Vögeli B., Orts J., Strotz D., Güntert P., and Riek R. Discrete three-dimensional representation of macromolecular motion from eNOE-based ensemble calculation. *Chimia (Aarau)*, 66(10):787–790, 2012.

Vranken W. F., Boucher W., Stevens T. J., Fogh R. H., Pajon A., Llinas M., Ulrich E. L., Markley J. L., Ionides J., and Laue E. D. The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins*, 59(4):687–696, 2005.

Vuister G. W., Tjandra N., Shen Y., Grishaev A., and Grzesiek S. Measurement of structural restraints. In *Protein NMR Spectroscopy: Practical Techniques and Applications*. Wiley, 1st edition, 2011.

Wang L., Eghbalnia H. R., Bahrami A., and Markley J. L. Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. *J Biomol NMR*, 32(1):13–22, 2005.

Wang S., Munro R. A., Shi L., Kawamura I., Okitsu T., Wada A., Kim S.-Y., Jung K.-H., Brown L. S., and Ladizhansky V. Solid-state NMR spectroscopy structure determination of a lipid-embedded heptahelical membrane protein. *Nat Methods*, 10(10):1007–1012, 2013.

Wang Y. and Wishart D. S. A simple method to adjust inconsistently referenced 13C and 15N chemical shift assignments of proteins. *J Biomol NMR*, 31(2):143–148, 2005.

Wasmer C., Lange A., Melckebeke H. V., Siemer A. B., Riek R., and Meier B. H. Amyloid fibrils of the HET-s(218-289) prion form a beta solenoid with a triangular hydrophobic core. *Science*, 319(5869):1523–1526, 2008.

Weber J. and Güntert P. Information content of distance restraints. In preparation.

Williamson M. P. and Craven C. J. Automated protein structure calculation from NMR data. *J Biomol NMR*, 43(3):131–143, 2009.

Wimmer R., Herrmann T., Solioz M., and Wüthrich K. NMR structure and metal interactions of the CopZ copper chaperone. *J Biol Chem*, 274(32):22597–22603, 1999.

Wylie B. J., Franks W. T., Graesser D. T., and Rienstra C. M. Site-specific 13C chemical shift anisotropy measurements in a uniformly 15N,13C-labeled microcrystalline protein by 3D magic-angle spinning NMR spectroscopy. *J Am Chem Soc*, 127(34):11946–11947, 2005.

Wylie B. J., Franks W. T., and Rienstra C. M. Determinations of 15N chemical shift anisotropy magnitudes in a uniformly 15N,13C-labeled microcrystalline protein by three-dimensional magic-angle spinning nuclear magnetic resonance spectroscopy. *J Phys Chem B*, 110(22):10926–10936, 2006.

Wylie B. J., Schwieters C. D., Oldfield E., and Rienstra C. M. Protein structure refinement using 13C alpha chemical shift tensors. *J Am Chem Soc*, 131(3):985–992, 2009.

Wylie B. J., Sperling L. J., Nieuwkoop A. J., Franks W. T., Oldfield E., and Rienstra C. M. Ultrahigh resolution protein structures using NMR chemical shift tensors. *Proc Natl Acad Sci U S A*, 108(41):16974–16979, 2011.

Zech S. G., Wand A. J., and McDermott A. E. Protein structure determination by high-resolution solid-state NMR spectroscopy: application to microcrystalline ubiquitin. *J Am Chem Soc*, 127(24):8618–8626, 2005.

Zhang Z., Porter J., Tripsianes K., and Lange O. F. Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta. *J Biomol NMR*, 59(3): 135–145, 2014.

Zhao D. and Jardetzky O. An assessment of the precision and accuracy of protein structures determined by NMR. Dependence on distance errors. *J Mol Biol*, 239(5):601–607, 1994.

Zhou D. H., Shea J. J., Nieuwkoop A. J., Franks W. T., Wylie B. J., Mullen C., Sandoz D., and Rienstra C. M. Solid-state protein-structure determination with proton-detected triple-resonance 3D magic-angle-spinning NMR spectroscopy. *Angew Chem Int Ed Engl*, 46(44):8380–8383, 2007.

Zimmerman D. E., Kulikowski C. A., Huang Y., Feng W., Tashiro M., Shimotakahara S., Chien C., Powers R., and Montelione G. T. Automated analysis of protein NMR assignments using methods from artificial intelligence. *J Mol Biol*, 269(4):592–610, 1997.

Zwahlen C., Legault P., Vincent S. J. F., Greenblatt J., Konrat R., and Kay L. E. Methods for Measurement of Intermolecular NOEs by Multinuclear NMR Spectroscopy: Application to a Bacteriophage λ N-Peptide/boxB RNA Complex. *Journal of the American Chemical Society*, 119(29):6711–6721, 1997.

Publications

- Güntert P. and Buchner L. Combined automated NOE assignment and structure calculation with CYANA. *J Biomol NMR*, doi:10.1007/s10858-015-9924-9, 2015.
- Buchner L. and Güntert P. Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA. *J Biomol NMR*, doi:10.1007/s10858-015-9921-z, 2015.
- Buchner L. and Güntert P. Increased reliability of NMR protein structures by consensus structure bundles. *Structure*, 23:425-434, 2015.
- Schmidt E., Ikeya T., Takeda M., Löhr F., Buchner L., Ito Y., Kainosho M., and Güntert P. Automated resonance assignment of the 21 kDa stereo-array isotope labeled thiosulfide oxidoreductase DsbA. *J Magn Res*, 249:88-93, 2014.
- Schmidt E., Gath J., Habenstein B., Ravotti F., Székely K., Huber M., Buchner L., Böckmann A., Meier B.H., and Güntert P. Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. *J Biomol NMR*, 56:243-254, 2013.
- Buchner L., Schmidt E., and Güntert P. Peakmatch: A simple and robust method for peaklist matching. *J Biomol NMR*, 55(3):267-277, 2013.