AUTOMATED SIGNAL IDENTIFICATION AND A STRUCTURAL INFORMATION CONTENT MEASURE FOR BIOMOLECULAR NMR DATA

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

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

von
Julia Maren Würz
geb. Weber
aus Weilburg

Frankfurt am Main 2016
(D 30)

vom Fachbereich Biochemie, Chemie und Pharmazie der
Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Michael Karas

1. Gutachter: Prof. Dr. Peter Güntert

2. Gutachter: Prof. Dr. Volker Dötsch

Datum der Disputation: ....

Contents

Summary **7**

Zusammenfassung **15**

1 Introduction **23**

- 1.1 Protein NMR structure determination 23
  - 1.1.1 Outline 23
  - 1.1.2 NMR spectroscopy 25
  - 1.1.3 Nuclear Spin Interactions 30
  - 1.1.4 Restraint sources for NMR structure calculation 31
- 1.2 Automation of protein NMR structure determination 33
  - 1.2.1 Signal identification 35
  - 1.2.2 Chemical shift assignment 35
  - 1.2.3 NOE assignment 36
  - 1.2.4 NMR structure calculation 40
  - 1.2.5 Alternative NMR structure determination approaches 43
  - 1.2.6 Fully automated protocols 44

2 Peak picking in multidimensional NMR spectra with CYPICK **47**

- 2.1 Introduction 48
- 2.2 Contour Peak Picker Algorithm 56
  - 2.2.1 Determination of the global noise level 57
  - 2.2.2 Determination of the local noise level 58
  - 2.2.3 Determination of the local extremum 58
  - 2.2.4 Scaling of contour lines 59
  - 2.2.5 Creation of contour lines 60
  - 2.2.6 Filtering of contour lines 61
  - 2.2.7 Analysis of contour lines 62

2.2.8 Interpolation of the local extremum 65

2.2.9 Deconvolute overlapping peaks 66

2.2.10 Restricted peak picking 68

2.3 Materials and methods 69

2.3.1 Evaluation dataset 69

2.3.2 Peak list comparison 70

2.3.3 Automated peak picking 72

2.3.4 Automated chemical shift assignment 74

2.3.5 Structure calculation 74

2.4 Results and discussions 76

2.4.1 Dependence of peak picking scores on the number of peaks 76

2.4.2 Evaluation of CYPICK with manually picked ENTH peak lists . . 78

2.4.3 Comparison with reference chemical shift assignment 86

2.4.4 Comparison with reference combined NOE assignment and structure calculation 90

2.4.5 Comparison with other automated peak picking procedures 93

2.4.6 Structure calculation of CASD-NMR proteins using NOESY peak lists from CYPICK 100

2.5 Conclusions 104

2.6 Implementation in CYANA 106

2.6.1 CYANA commands 106

2.6.2 CYANA macros 112

3 Information content of NMR distance restraints 115

3.1 Introduction 116

3.2 Information content algorithm 121

3.3 Material and Methods 125

3.3.1 Experimental data set 125

3.3.2 Radius of gyration estimation 128

3.3.3 Maximal distance estimation 129

3.3.4 Work flow 130

3.4 Results and Discussion 135

3.4.1 General characteristics of the I 135

3.4.2 Agreement between structure dependent and independent calculation mode . . . 142
3.4.3 Scaling of I . . . 145
3.4.4 Correlation between I and structural precision . . . 146
3.4.5 Computation time . . . 149
3.5 Conclusions . . . 151
3.6 Implementation in CYANA . . . 153
3.6.1 Command **distances information** . . . 153

4 Structure-based drug design by NMR . . . 155
4.0 Contributions . . . 156
4.1 Introduction . . . 157
4.2 Materials and methods . . . 160
4.2.1 Generation of reference assignment . . . 160
4.2.2 NMR measurements . . . 160
4.2.3 Peak picking . . . 161
4.2.4 FLYA assignment transfer . . . 162
4.2.5 Generation of CYANA library entries . . . 164
4.2.6 CYANA calculations . . . 164
4.3 Results and discussion . . . 166
4.3.1 Strategy for enabling SBDD . . . 166
4.3.2 The Protein-Ligand NOESY . . . 167
4.3.3 FLYA assignment transfer . . . 168
4.3.4 Application to a diverse set of protein-ligand complexes . . . 169
4.4 Conclusions . . . 179
4.5 Usage of the FLYA assignment transfer protocol in CYANA . . . 180
4.5.1 FLYA parameters . . . 180
4.5.2 FLYA macros . . . 180

5 Conclusions and Outlook . . . 185

Appendices . . . 189

A CYPICK . . . 193

B Information content . . . 203

C Structure-based drug design by NMR 205

Bibliography 209

List of abbreviations 225

Publications 227

Curriculum Vitae 228

Danksagung 231

Summary

The emphasis of this dissertation was to improve the process of automated protein structure determination by nuclear magnetic resonance (NMR). The classical approach towards NMR structure determination can be organized into the following steps:

- Data collection and processing
- Signal identification ('Peak picking')
- Chemical shift assignment
- Collection of conformational restraints
- Structure calculation, refinement and validation

The abovementioned steps are explained in detail in chapter 1.2. Several alternative, non-classical approaches have been reviewed in (Guerry & Herrmann, 2011) and are briefly summarized in chapter 1.2.5.

Some of the sequential steps of NMR protein structure determination have already been successfully automated, which makes them independent from the individual expertise of the user, more time efficient and more objective. CYANA (Güntert *et al.*, 1997; Güntert & Buchner, 2015) is a software package that is routinely used for automated chemical shift assignment (Schmidt & Güntert, 2012), automated *Nuclear Overhauser Enhancement* (NOE) assignment and structure calculation of proteins. However, the major goal is to automate all the steps listed above (with the exception of data collection). Therefore, it is necessary to develop novel algorithms which perform the individual steps more reliable.

One of these steps, which has not yet been successfully automated represents the fully automated signal identification of NMR spectra. The step of peak picking is a very crucial stage in the whole process because NMR signal lists serve as basis for all subsequent steps. Errors in these lists propagate into all subsequent steps of data evaluation and ultimately

result in incorrect structures. A fully automated approach to NMR peak picking which is used on a regular basis is lacking. Therefore one of the major aims of this thesis was to develop a robust peak picking algorithm and implement it in CYANA. This peak picking procedure should be combined with the existing FLYA algorithm for resonance assignment, automated NOE assignment and structure calculation with CYANA. A detailed description of the peak picking algorithm CYPICK can be found in chapter "Peak picking in multidimensional NMR spectra with CYPICK" (chapter 2).

The CYPICK peak picking algorithm mimics the human approach towards peak picking as authentically as possible. Manual peak picking is performed by analyzing two-dimensional contour line representations of the spectrum. The scientist decides with the aid of geometry and similarity criteria if the contour line at hand belongs to a real protein signal or represents a spectral artifact. Protein signals resemble concentric ellipses and comply with specific geometric criteria which artifacts do not fulfill, e.g. approximate circular shape, after appropriate scaling of the spectral axes, and entire convexity. CYPICK evaluates the contour lines of local extrema by means of these conditions and decides whether the signal at hand is real or not.

In the context of the evaluation of CYPICK several input parameters were extensively tested on various NMR spectra of the ENTH-VHS domain At3g16279-(9-135) of *Arabidopsis thaliana* (ENTH). The performance of CYPICK was evaluated and compared by calculating scores with the help of manual established reference peak lists. These scores were also calculated for other prominent programs, i.e. AUTOPSY, NMRViewJ, CCPN, and CV-Peak Picker, and compared to CYPICK. CYPICK peak lists showed a higher reliability which is manifested by peak picking more true peaks and less artifacts than the other programs. Automated peak picking of NOESY spectra was performed particularly well by CYPICK in comparison to the other programs. CYPICK and the abovementioned programs were also used to pick the complete data set of NMR spectra belonging to the Src homology domain 2 from the human feline sarcoma oncogene FES (SH2) and the *Arabidopsis thaliana* rhodanese domain At4g01050 (RHO). In these cases no manual peak lists were available that could be used for score calculation. Automatically generated ENTH, RHO, and SH2 peak lists were then used in automated chemical shift assignment by FLYA, automated NOE assignment and structure calculation by CYANA. Results achieved at these stages were compared to the reference resonance assignment and the reference structure bundle. CYPICK peak lists led to resonance assignments and structure bundles of high accuracy. The results were superior when compared to the other programs. Even in the challenging case of solely NOESY-based chemical shift assignment the results of CYPICK stood out for their reliability throughout the data set. CYPICK was further applied to ten CASD-NMR protein data sets which included NOESY spectra. The peak lists obtained by CYPICK were used in automated NOE assignment and structure calculation together with the reference chemical shift assignment. In 50% of the cases the structures could be recalculated with an RMSD bias ≤ 1 Å. For the remaining proteins correctly folded structures were found, although with slightly higher RMSD biases ≤ 3 Å. The quality of the results obtained through CYPICK lie between the results from manually edited and 'raw' peak lists.

However, these investigations also revealed that certain functionalities of CYPICK can be improved or implemented in future projects. The requirements for the presence of a local extremum should be relaxed in order to identify overlapping signals which do no possess a local extremum. In this context, it would also be desirable to have a stable deconvolution method. Additionally, the algorithm should become completely independent from noise level values. Thereby, very weak signals buried below the global threshold could be identified. A tool for discarding peaks arising from noise bands would also improve the overall performance of CYPICK. Furthermore, it would be worthwhile to have the opportunity to guide the peak picking by additional information such as a 3D structure or a chemical shift assignment. Contour-based quality factors Q_{rad} and Q_{con} can in principle be used to direct automated chemical shift and NOE assignment.

At the end of each NMR structure calculation the question of quality, accuracy and precision of the calculated structure bundle arises. These features are directly accessible in X-ray crystallography from the measured data. In NMR spectroscopy no fully well-established method is available for the determination of this measure. In this context, the knowledge about the structural information of individual restraints used for the calculation becomes of interest. Some algorithms that quantify the structural information of restraints exist, but are not commonly used due to varying reasons. Therefore developing a method to meaningfully and readily quantify the information content of distance restraint data sets has been a further target of this thesis. This measure gives an estimation of the structural precision, the so called RMSD radius, of the resulting structure solely based on the input restraints. Such a measure is comparable to the resolution and B-factors in X-ray crystallography. A description of the algorithm and the results are presented in the chapter "Information content of NMR distance restraints" (chapter 3).

The information content implementation in CYANA (referred to as I) is mostly based on probability theory. I is defined as the negative logarithm of the probability that a set of restraints will be fulfilled by an ensemble of random structures. I is a sum of quotients of a conditional probability logarithm and a redundancy. The conditional probability quantifies the probability of an individual restraint to be fulfilled by a random structure. The redundancy of an individual restraint reflects the probability that an individual restraint will be fulfilled by a random structure that also fulfills another distance restraint.

The I calculation is performed completely automatic and demands no user input.

Initially, several properties of I have been demonstrated on the basis of distance restraints from 27 NorthEast Structural Genomics consortium (NESG) (Wunderlich *et al.*, 2004) proteins, which varied in size from 5.2 kDa to 22 kDa. The following characteristics have been observed:

1. The different structural information of varying restraint types, i.e. long-range distance restraints carry the most structural information, whereas short-range distance restraints contribute to a significant lesser amount of information.
2. Structurally equivalent restraints lead to an increase in restraint redundancy and a concomittant individual restraint information loss, such that the overall information of the data set remains approximately the same.
3. I depends on the upper and lower distance limit, i.e. I increases with a smaller available conformational space and decreases with a larger available conformational space.

Furthermore, it was observed that the value of I also depends on the protein size. It was shown that the information content per residue, I_r, is a measure of information content that is independent from the size of the protein by scaling the overall I by the number of ordered residues included in the structure. The information content can be calculated alternatively with the help of a structure bundle or on the basis of the covalent geometry of the polypeptide chain. It was demonstrated that both calculation modes are in good agreement. Finally, it could be observed that I_r of a distance restraint data set is correlated to the structural precision of a structure bundle that has been determined on the basis of this data set. It was determined that an $I_r > 3$ is needed in order to obtain a structure bundle with an RMSD radius < 2 Å.

In principle, 3D protein structures can be determined by NMR spectroscopy and X-ray crystallography. However, in the pharmaceutical industry X-ray crystallography is sometimes preferred over NMR spectroscopy. Reasons therefore are its potential towards high-throughput which is in principal enabled by "molecular replacement". Structure-based drug design (SBDD) is therefore mostly performed by X-ray crystallography. NMR, however, has several advantages over X-ray crystallography which are crucial for SBDD, such as the possibility to study complexes with low binding affinities. Therefore a strategy to perform SBDD by NMR spectroscopy was developed in collaboration with Dr. Alvar Gossert and co-workers from Novartis and Dr. Elena Schmidt. Methods and strategies developed in this context are described in the chapter "Structure-based drug design by NMR" (chapter 4).

The two major developments within this work were the 'Protein-ligand NOESY' which includes all relevant data for structure calculation of a protein ligand complex, and the 'FLYA assignment transfer' which enables obtaining assignments of an unknown protein ligand complex only by the usage of signal lists from the protein-ligand NOESY' and a reference chemical shift assignment. The FLYA assignment transfer has two calculation modes: a 1-step and a 2-step protocol. The 2-step protocol uses information from the reference complex assignment and additional information from an available structure.

The SBDD by NMR approach was tested on the basis of the protein MDM4 (Wade & Li, 2013), which regulates the cancer suppressor protein p53 (Shvarts *et al.*, 1996). Studies showed that MDM4 is over-expressed in 10–20% of 800 very diverse tumors, which led to a massive reduction of p53. Accordingly, MDM4 is an interesting target for the development of anti-cancer strategies. In this study, diverse ligands (peptide-1, peptide-2, nutlin-3a, fragment) in terms of size, affinity, binding kinetics and chemistry, were analyzed in complex with MDM4 and 3D structures were determined. The chemical shift assignment of MDM4 in complex with peptide-1 was utilized as reference for the assignment of further MDM4 complexes.

The chemical shift assignments of a set of MDM4 complexes was determined on the basis of peak lists from the protein-Ligand NOESY which includes all necessary data for assignment and structure calculation in a single spectrum. Automated chemical shift assignments were performed with the FLYA assignment transfer protocol. Resonance assignments showed a high percentage of correctness ranging from 79–95%. Assignments on the basis of manual and CYPICK peak lists showed comparable results, whereas the number of strong peaks was usually slightly higher when using manual peak lists. The 2-step protocol led to a significant improvement in chemical shift assignment of approximately 10% in case of peptide-2 and approximately 5% in case of the fragment and nutlin. Peptide-2 is chemically very similar to peptide-1, accordingly the knowledge of the structure has a higher impact as in case of the fragment and nutlin-3a, which are chemically very different from peptide-1. Structure calculation results of peptide-1 and peptide-2 led to structure bundles of high quality, with an RMSD radius of approximately 1.0 Å and RMSD bias < 1.8 Å. Manual and CYPICK peak lists yielded structure bundles of similar quality. The structure calculation results of nutlin-3a were inferior to peptide-1 and peptide-2 results. When using manual peak lists, structures achieved had RMSD radius values < 1.0 Å and RMSD bias values of < 2.00 Å. Structure bundles achieved by CYPICK were even less reliable, i.e. the RMSD radius was 0.8–1.7 Å and the RMSD bias varied from 2.5–3.3 Å, when using the 1-step and 2-step protocol, respectively. Structure calculation on MDM4 in complex with the fragment could not be performed and presented within this thesis due to confidentiality reasons.

The SBDD by NMR protocol has several drawbacks which are in line with the general limitations of protein NMR. The protein has to be uniformly isotope labeled and the size should be less than 30 kDa. Within this method ligand signals are measured in a two-dimensional plane of the Protein-Ligand NOESY, which can quickly suffer from signal overlap depending on the chemical composition. In this study, the method was successfully applied to ligands ranging in size from 1–2 kDa.

Zusammenfassung

Der Schwerpunkt dieser Dissertation lag in der Verbesserung einzelner Aspekte im Prozess der automatischen Proteinstrukturbestimmung mittels NMR. Der klassische Ansatz zur NMR Proteinstrukturbestimmung gliedert sich in folgende Schritte:

- Messung und Prozessierung der NMR Daten
- Signalidentifizierung (‘Peak picken’)
- Zuordnung der chemischen Verschiebungen
- Zuordnung der ‘NOE’ Signale
- Strukturrechnung, -verfeinerung und -validierung

Die oben genannten Schritte werden in Kapitel 1.2 ausführlich beschrieben. Alternative nicht klassische Ansätze werden detailliert in Guerry & Herrmann (2011) diskutiert und kurz in Kapitel 1.2.5 zusammengefasst.

Die Zuordnung der chemischen Verschiebungen, der NOEs und die Strukturrechung wurden bereits erfolgreich automatisiert, was die Ergebnisse unabhängig von der individuellen Expertise des Anwenders, zeit-effizienter und objektiver gestaltet. CYANA (Güntert *et al.*, 1997; Güntert & Buchner, 2015) ist ein Programmpaket, welches routinemäßig zur automatischen Zuordnung der chemischen Verschiebungen (Schmidt & Güntert, 2012), der automatischen Zuordnung von *Nuclear Overhauser Enhancement* Signalen (NOEs) und der Strukturrechnung von Proteinen verwendet wird. Das Hauptziel jedoch ist es, alle oben gezeigten Schritte (mit Ausnahme der Messung der Daten) zu automatisieren. Daher ist es notwendig neue Algorithmen zu entwickeln, die die individuellen Schritte robust und verlässlich durchführen.

Einer der Schritte, die noch nicht erfolgreich automatisiert wurden, stellt die Signalidentifizierung von NMR Spektren dar. Dieser Schritt ist besonders wichtig, da Listen von

NMR-Signalen Grundlage aller Folgeschritte sind. Fehler in den Signallisten pflanzen sich in allen Folgeschritten der Datenauswertung fort und können am Ende in falschen Strukturen resultieren. Ein komplett automatisierter Ansatz zur Signalidentifizierung, welcher standardmäßig verwendet wird, existiert bisher noch nicht. Daher war das Hauptziel dieser Arbeit, einen robusten und verlässlichen Algorithmus zur Signalidentifizierung von NMR Spektren in CYANA zu implementieren. Dieser Algorithmus sollte mit dem in FLYA implementierten Ansatz zur automatischen Resonanzzuordnung, der automatischen NOE-Zuordnung und der Strukturrechnung mit CYANA kombiniert werden. Eine detaillierte Beschreibung des CYPICK Algorithmus zur Signalidentifizierung von NMR Spektren erfolgt im Kapitel ‘Peak picking in multidimensional NMR spectra with CYPICK’ (Kapitel 2).

Der in CYANA implementierte CYPICK Algorithmus imitiert den von Hand durchgeführten Ansatz. Bei der manuellen Methode schaut sich der Wissenschaftler zweidimensionale Konturliniendarstellungen der NMR Spektren an. Er entscheidet anhand verschiedener Geometrie- und Ähnlichkeitskriterien, ob es sich um ein Signal des Proteins oder um einen Artefakt handelt. Proteinsignale sind ähnlich zu konzentrischen Ellipsen und erfüllen bestimmte geometrische Kriterien, wie zum Beispiel ungefähr kreisförmiges Aussehen nach entsprechender Skalierung der spektralen Achsen und gänzlich konvexe Formen, die Artefakte nicht aufzeigen. CYPICK bewertet die Konturlinien lokaler Extrema nach diesen Bedingungen und entscheidet anhand dieser, ob es sich um ein echtes Signal handelt oder nicht.

Im Rahmen der Evaluation von CYPICK wurden zunächst Eingabeparameter des Anwenders intensiv an einer Vielzahl von Spektren der ENTH-VHS Domäne At3g16270-(9-135) von *Arabidopsis thaliana* (ENTH) (López-Méndez & Güntert, 2006; López-Méndez *et al.*, 2006) getestet. Die Leistung von CYPICK wurde basierend auf Score-Werten mit manuell erstellten Listen verglichen und bewertet. Diese Score-Werte wurden ebenfalls für Peaklisten, die von den bekannten Programmen AUTOPSY, NMRViewJ, CCPN und CV-Peak Picker erzeugt wurden, berechnet. CYPICK-Peaklisten zeigten ein höheres Maß an Robustheit, welches sich durch die konsistente Identifikation von mehr echten Signalen und weniger Artefakten gegenüber den anderen Programmen manifestierte. Automatisches Peak Picken von NOESY Spektren wurde besonders gut von CYPICK durchgeführt. CYPICK und die oben genannten Programme wurden außerdem zur Analyse der Datensätze der Src Homologie Domäne des humanen Katzensarkoma Onkogens (SH2) (Scott *et al.*, 2004, 2005) und der *Arabidopsis thaliana* Rhodanase Domäne At4g01050

(RHO) (Pantoja-Uceda *et al.*, 2004, 2005) verwendet. In diesen Fällen waren keine manuellen Peaklisten verfügbar. Zum Vergleich der Leistung der Programme anhand der Score-Werte. ENTH, RHO und SH2 Peaklisten wurden dann zur automatischen Zuordnung der Resonanzen, zur automatischen NOE-Zuordnung und zur Strukturrechnung verwendet. Resultate, die in diesen Schritten erzielt wurden, konnten dann mit einer Referenzzuordnung und einem Referenzstrukturbündel verglichen werden. Von CYPICK erstellte Peaklisten führten zu Resonanzzuordnung und Strukturbündeln, die sich durch eine hohe Genauigkeiten auszeichneten. Die mit CYPICK-Listen erzielten Resultate waren besser als die von den übrigen Programmen produzierten Ergebnisse. Sogar im Fall der ausschließlich NOE basierten Resonanzzuordnung, zeichneten sich die Resultate, die mit CYPICK-Peaklisten erzeugt wurden, durch eine hohe Robustheit über den kompletten Datensatz hinweg aus. CYPICK wurde weiter auf zehn Datensätze des CASD-NMR Projekts (Rosato *et al.*, 2012, 2009) angewandt, welche nur NOESY Spektren beinhalteten. Die von CYPICK erzielten Peaklisten wurden dann zusammen mit den Referenzresonanzzuordnungen zur automatischen NOE-Zuordnung und Strukturrechnung verwendet. In 50% der Fälle konnten die Strukturen mit einem RMSD Bias von ≤ 1 Å wieder berechnet werden. In den übrigen 50% der Fälle wurden Proteine mit korrekter Faltung berechnet, wenngleich die RMSD Bias Werte bei ≤ 3 Å lagen. Die auf diesem Datensatz mit CYPICK erstellten Ergebnisse konnten zwischen den manuell editierten und nicht editierten Listen angesiedelt werden.

Die an CYPICK durchgeführten Evaluationsstudien zeigten auf, dass bestimmte Funktionalitäten von CYPICK noch wesentlich verbessert werden oder in zukünftigen Projekten adressiert werden können. Im Moment benötigt CYPICK ein lokales Extremumkriterium, das erfüllt werden muss, damit ein Datenpunkt als Signal berücksichtigt werden kann. Diese Bedingung sollte gelockert werden, um sogenannten ‘Peakschultern’, die an der Steigung eines anderen lokalen Extremums lokalisiert sind, berücksichtigen zu können. Schwache Signale, die nicht genug Konturlinien aufweisen, werden derzeit ausgeschlossen. Eine Verfeinerung der Konturlinienkriterien sollte die Identifikation solcher Signale ermöglichen. Viele mit identifizierte Artefakte stammen von sogenannten *ridges* oder Rauschbändern. Eine Funktionalität zur Erkennung dieser Regionen sollte die Anzahl der Artefakte deutlich reduzieren. CYPICK nutzt keine externen Informationen über das zugrunde liegende System. Die Signalidentifizierung könnte durch die Verwendung zusätzlicher Information, wie zum Beispiel der 3D Struktur oder der Resonanzzuordnung verbessert werden. Eine fortschrittlichere Methode zur Dekonvolution überlappender

Regionen ist notwendig und sollte die Ergebnisse der Signalidentifizierung von CYPICK weiter verbessern. Es ist denkbar, die von den Konturlinien abgeleiteten Peak-Qualitätstfaktoren in der automatischen Resonanz und NOE-Zuordnung zu verwenden, um verlässliche und unsichere Peaks unterschiedlich zu behandeln.

Das zweite Ziel dieser Arbeit war es ein Maß zur Quantifizierung der strukturellen Information von NMR Datensätzen (strukturelle Distanzeinschränkungen) zu entwickeln, welches schnell und einfach zu berechnen ist. Am Ende jeder Strukturrechnung eröffnet sich die Frage nach der Qualität, der Genauigkeit und der Präzission des berechneten Strukturbündels. In der Röntgenkristallographie sind diese Merkmale direkt aus den Messdaten zugänglich. In der NMR Spektroskopie gibt es jedoch noch keine vollständig etablierten Methoden zur Bestimmung dieser Größen. In diesem Kontext ist es daher interessant, die strukturelle Information der Datensätze, die zur Strukturrechnung verwendet wurden, zu quantifizieren. Einige Algorithmen zur Quantifizierung der strukturellen Information von NMR Datensätzen existieren zwar, werden allerdings aus unterschiedlichen Gründen nicht standardmäßig angewandt. Dieses Maß korreliert mit der Präzission der resultierenden Strukturen, dem sogenannten RMSD Radius. Ein solches Maß ist vergleichbar mit der Auflösung in der Röntgenkristallographie. Eine Beschreibung des Algorithmus und eine Übersicht der Ergebnisse folgt im Kapitel ‘Information content of NMR distance restraints’ (Kapitel 3).

Der Informationsgehalt (I) wurde in CYANA implementiert und basiert in erster Linie auf der Wahrscheinlichkeitstheorie. I wurde definiert als der negative Logarithmus der Wahrscheinlichkeit, dass ein Datensatz aus Distanzeinschränkungen von einem Ensemble aus Zufallsstrukturen erfüllt wird. I berechnet sich über den Quotienten einer bedingten Wahrscheinlichkeit und einer Redundanz. Die bedingte Wahrscheinlichkeit quantifiziert, wie genau eine individuelle Distanzeinschränkung durch eine Zufallsstruktur erfüllt wird. Die Redundanz der individuellen Distanzeinschränkung spiegelt die Wahrscheinlichkeit wider, dass diese von einer Zufallsstruktur erfüllt wird, die auch eine weitere Distanzeinschränkung erfüllt. Die Berechnung von I wird vollständig automatisch durchgeführt und bedarf keiner Eingaben des Anwenders.

Zunächst wurde anhand eines Datensatzes bestehend aus 27 *NorthEast Structural Genomics consortium* (NESG) (Wunderlich *et al.*, 2004) Proteinen, welche in Größe von 5,2 bis 22 kDa variierten eine Vielzahl von Eigenschaften des I aufgezeigt. Folgende Charakteristika konnten gezeigt werden:

1. Es konnte der Unterschied in der strukturellen Information von verschiedenen Distanzeinschränkungstypen gezeigt werden. Langreichweitige Distanzeinschränkungen wiesen die höchste strukturelle Information auf, wobei kurzreichweitige Distanzeinschränkungen einen deutlich niedrigeren Informationsgehalt zeigten.
2. Strukturell äquivalente Distanzeinschränkungen führten zu einer Erhöhung der Redundanz und einem Verlust an individueller Information, wobei die Gesamtinformation des Datensatzes gleich blieb oder sich erhöhte.
3. I zeigte eine starke Abhängigkeit von den oberen und unteren Distanzschranken. Der Informationsgehalt erhöhte sich, wenn der zugängliche Konformationsraum verkleinert wurde, wobei sich der Informationsgehalt erniedrigte, wenn ein größerer Konformationsraum zur Verfügung stand.

Außerdem konnte gezeigt werden, dass I neben der Abhängigkeit von der strukturellen Information von der Größe des Proteins abhängt. Daher wurde I_r definiert, welcher unabhängig von der Proteingröße ist. Dabei wird der gesamte Informationsgehalt des Datensatzes mit der Anzahl der geordneten Reste in der Struktur skaliert. Die Anzahl der geordneten Reste kann vom Anwender vorgegeben werden, mit Hilfe einer vorhandenen Struktur mittels CYRANGE (Kirchner & Güntert, 2011) bestimmt werden oder basierend auf der Anzahl der Reste, die lang- und mittelreichweitige Distanzeinschränkungen aufzeigen, gemessen werden. Es ist möglich, den Informationsgehalt eines Datensatzes entweder anhand eines Ensembles von Strukturen zu bestimmen oder unter zur Hilfe nahme der kovalenten Geometrie der Polypeptidkette. Beide Berechnungsmodi stimmten sehr gut miteinander überein und wiesen eine Korrelation von > 0.99 auf. Schließlich wurde herausgefunden, dass der I_r eines Datensatzes mit der Präzission einer aus dem Datensatz resultierenden Struktur korreliert. Es wurde demonstriert, dass I_r größer als 3 sein muss, damit man mit dem zugrunde liegenden Datensatz eine Struktur mit einem RMSD Radius < 2.0 Å erzielen kann.

Grundsätzlich stehen zwei experimentelle Methoden zur Verfügung, um die 3D Struktur von Proteinen auf atomarer Ebene aufzuklären: die NMR Spektroskopie oder die Röntgenkristallographie. In der Pharmaindustrie wird tweilweise die Röntgenkristallographie. gegenüber der NMR Spektroskopie bevorzugt. Gründe dafür sind das Potenzial zum Hochdurchsatzverfahren, welches hauptsächlich durch die Methode des ‘molekularen Ersatzes’ ermöglicht wird. NMR hat jedoch einige Vorteile gegenüber der Röntgenkristallographie. Zum Beispiel können mit NMR auch Komplexe mit niedrigen Bindungsaffinitäten untersucht werden. Daher wurden Strategien entwickelt, die NMR für SBDD zugänglicher machen sollen. Dieses Projekt wurde in Zusammenarbeit mit Dr. Alvar Gossert und Dr. Elena Schmidt entwickelt. Neue Methoden und Strategien, die in diesem Kontext entwickelt wurden, werden im Kapitel 'Structure-based drug design by NMR' dargestellt (Kapitel 4).

Die zwei wichtigsten Entwicklungen in diesem Projekt sind das 'Protein-Liganden NOESY', welches alle relevanten Daten zur Strukturrechnung beinhaltet und der 'FLYA Zuordnungstransfer', welcher es ermöglicht, einen unbekannten Protein-Ligand Komplex nur unter Verwendung von Signalen des 'Protein-Liganden NOESYs' und einer Resonanzzuordnung eines anderen Komplexes zuzuordnen. Der FLYA Zuordnungstransfer hat zwei Berechnungsmodi: den Ein-Stufen Prozess und den Zwei-Stufen Prozess. Der Zwei-Stufen Prozess unterscheidet sich vom Ein-Stufen Prozess darin, dass er zusätzlich zur Resonanzzuordnungsinformation noch Information aus einer bereits gelösten Struktur verwendet.

Der SBDD mit NMR Ansatz wurde an dem Beispiel des Proteins MDM4 (Wade & Li, 2013) getestet. MDM4 reguliert das Tumorsuppressor Protein p53 (Shvarts *et al.*, 1996). Studien an MDM4 zeigten, dass MDM4 in 10-20 % von 800 sehr diversen Tumoren überexprimiert wird. Diese Überexpression hat eine massive Reduktion von p53 Aktivität zur Folge. Folglich ist MDM4 ein interessantes Zielmolekül zur Entwicklung von Anti-Tumor Strategien. In dieser Studie wurde ein vielfältiger Datensatz an Liganden für MDM4 (Petid-1, Peptid-2, Nutlin-3a und Fragment) getestet und die Komplexstrukturen bestimmt. Die Diversität des Datensatzes spiegelte sich in der Größe, der Affinität, den Bindungskinetiken und der chemischen Komposition wider. Die Zuordnung von MDM4 in Komplex mit Peptid-1 galt als Referenz für die Zuordnung aller weiteren MDM4 Komplexe.

Die Resonanzzuordnung von einer Reihe von MDM4 Komplexen wurde mit Peaklisten des 'Protein-Ligand NOESYs' und einer manuell erstellten Referenzzuordnung automatisch mit dem 'FLYA Zuordnungstransfer' erzeugt. Die auf diese Art generierten Resonanzzuordnungen zeigten eine Genauigkeit, die zwischen 79 und 95 % variierte. Die Zuordnungsergebnisse, welche mit manuell und mit automatisch erzeugten Signallisten von CYPICK erstellt wurden, zeigten vergleichbare Ergebnisse, wobei die Anzahl an 'starken' Zuordnungen mit den manuellen Listen meist etwas höher war. Der Zwei-Stufen Zuordnungsprozess führte zu einer Verbesserung in der Resonanzzuordnung um etwa 10 % für MDM4 in Komplex mit Peptid-2 und um etwa 5 % für MDM4 in Komplex mit dem Fragment und Nutlin-3a. Das Fragment und Nutlin-3a sind chemisch jeweils sehr verschieden von Peptid-1, wobei Peptid-2 chemisch sehr ähnlich zu Peptid-1 ist. Folglich hat die Benutzung einer Referenzstruktur in diesem Zusammenhang einen größeren Einfluss. Strukturrechnungen von MDM4 in Komplex mit Peptid-1 und Peptid-2 führten zu Strukturbündeln mit hoher Genauigkeit und Präzission mit einem RMSD Radius von etwa 1.0 Å und einem RMSD Bias von < 1.80 Å. In diesen Beispielen erzielten manuelle und CYPICK-Peaklisten Strukturbündel ähnlicher Qualität. Die Strukturrechnungsergebnisse von Nutlin-3a waren deutlich schlechter als die von Peptid-1 und Peptid-2. Mit manuellen Peaklisten konnten Strukturen mit einem RMSD Radius < 1.0 Å und einem RMSD Bias von < 2.0 Å erzielt werden. Strukturrechnungen mit CYPICK-Peaklisten führten zu weniger verlässlichen Ergebnissen. Der RMSD Radius lag in etwa bei 0.8–1.7 Å und der RMSD Bias variierte von 2.5–3.3 Åmit dem Zwei-Stufen und Ein-Stufen Protokoll respektive. Strukturrechnungsergebnisse zu MDM4 in Komplex mit dem Fragment konnten aus vetraulichen Gründen nicht in dieser Arbeit durchgeführt und gezeigt werden.

Das SBDD mit NMR Protokoll hat Limitierungen und Beeinträchtigungen, die den allgemeinen Limitierungen von Protein NMR entsprechen. Das Protein muss uniform isotopenmarkiert werden und die Größe sollte 30 kDa nicht überschreiten. Bei der Verwendung des ‘Protein-Liganden NOESYs’ werden die Liganden Signale in einer zweidimensionalen Ebene aufgenommen, welche schnell unter Signalüberlappung, die von der chemischen Komposition abhängig ist, leiden kann. In dieser Studie, wurde die Methode erfolgreich auf Liganden der Größe 1–2 kDa angewandt.

Chapter 1

Introduction

1.1 Protein NMR structure determination

1.1.1 Outline

Proteins represent a versatile group of macromolecules in living organisms. They satisfy a multitude of biological tasks, e.g. the catalysis of chemical reactions, transport or storage of molecules, cell mobility, give structure, grant immunity, transmit signal and control cell growth and differentiation. Proteins are linear polymers which mainly take an three-dimensional (3D) active fold. The 3D fold of a protein determines its function and sheds light on potential roles of the molecule in the cell or displays possibilities to interact with other molecules. Knowledge of the 3D structure leads for example to possibilities to understand the protein in a mechanistic manner, since the specificity of the activity center and the binding pocket depend on the 3D structure. Especially in the case of a pathogenic germ, the structure gives perspectives to block this molecule in an artificial manner and thereby inactivating it. Most of the target molecules in the pharmaceutical industry are proteins, i.e. enzymes, ion channels and membrane receptors. Especially membrane receptors pose a central field of research in the business of pharmaceuticals because they control which molecules dock to a cell and pass through the membrane. This is in particular relevant for drugs which need to act on the inside of a cell and therefore have to pass the membrane. Consequently, the knowledge of the 3D structure is an essential component in the field of protein analytic.

In principle, there are two major experimental methods which enable the study of proteins, nucleic acids or other biomolecules on an atomic level, namely nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography. These methods do not exclude other each in terms of information that can be achieved from them but are complementary in many ways forming a clearer picture of the molecular structure at hand. There are significant differences between these two methods in terms of requirements towards the sample and information which can be deduced from the experiments (Wüthrich, 1986). Molecule samples for NMR spectroscopy are usually isotopically labeled and in aqueous or detergent solutions, whereas the size of the protein is in general less than 50 kDa. In X-ray crystallography the molecule must be crystallized and the crystal must also diffract. The samples are usually in a solid-frozen state and have virtually no limitation in terms of size. Sample preparation and data acquisition time is usually short in NMR spectroscopy, whereas data analysis takes a tremendous amount of time. In X-ray crystallography it is the other way around, a long time exposure for screening and optimization, but a short amount of time for data processing is required. A huge advantage of NMR over X-ray crystallography is the physical state of the sample which is usually closer to the native environment and therefore one expects the fold of the protein to be closer to the native fold. Furthermore, the solution conditions can be varied in NMR spectroscopy which renders possibilities to examine samples under native or denaturing conditions. Additionally, NMR enables the study of molecule dynamics on time scales from ns to hours and the investigation of inter-molecular interactions (Billeter, 2015). X-ray crystallography holds the advantage of a higher atomic resolution for well-diffracting crystals over NMR. The main disadvantages of NMR are the need for isotopically labeled samples which can be very expensive and the poor resolution of NMR spectra of large proteins. The main challenge in X-ray crystallography is to solve the phase problem. In addition, one always has to consider that the crystallized structure can differ from the native fold of the protein.

In the year 2000 it was reported that approximately 90% of the structures deposited in the Protein Data Bank (PDB) were solved by X-ray crystallography. Regarding the ratio of structures solved by NMR and X-ray not much has changed up to now. In December 2016 the PDB contained more than 120,000 entries of which 11,500 were solved by NMR spectroscopy, and about ten times as much by X-ray crystallography. In the meantime, both fields have experienced development. X-ray crystallography has been employed for more than 50 years in the field of protein structure determination. Since the structural determination of Myoglobin by Kendrew and co-workers in 1958 (Kendrew *et al.*, 1985) the field has experienced a lot of improvement, e.g. the use of synchrotron sources and cryo-crystallography (Joachimiak, 2009). Even though NMR still holds many advantages over X-ray crystallography, X-ray crystallography is still in many fields preferred over NMR mainly because of practical reasons, i.e. analysis time. Analysis of NMR data is still performed manually in many cases, taking up to several months of analysis time for one protein structure.

CYANA (Güntert *et al.*, 1997; Güntert & Buchner, 2015) is a software package that incorporates many tools for automated NMR data analysis and structure calculation (explained below). However, in order to perform a fully automated protocol of NMR structure determination starting from a processed NMR spectrum, the step of signal identification has to be automated and implemented in the CYANA software package. Therefore, one major goal of this PhD thesis was to develop a new robust peak picking algorithm in CYANA which allows a combination with automated resonance assignment, NOE assignment and structure calculation. The contour geometry based peak picking algorithm CYPICK is introduced in chapter 2. Additionally, the 'information content' a method to help the scientist to quantify the structural information included in NMR restraint data sets was developed. The information content quantifies the structural information of restraint data sets by a single meaningful number, which correlates with the precision of the resulting NMR structure bundle, as presented in chapter 3. Further, we developed new methods in collaboration with Dr. Alvar Gossert from Novartis and Dr. Elena Schmidt which aim to make NMR more accessible towards structure-based drug design, i.e. the development of the 'Protein-Ligand NOESY' experiment that includes all data necessary for structure determination in one spectrum and the development of the 'FLYA assignment transfer protocol' that uses chemical shifts from a known protein complex structure and guides the chemical shift assignment of a new complex structure. Results are presented in chapter 4.

1.1.2 NMR spectroscopy

The following summary is mainly based on the books "Nuclear Magnetic Resonance" by P.J. Hore (Hore, 2007) and "Understanding NMR Spectroscopy" by J. Keeler (Keeler, 2010).

NMR is a physical effect which has first been described in 1938 (Rabi *et al.*, 1938). Principle of NMR is the interaction of the nuclear magnetic moment, μ, with the external magnetic field, B. Source of μ is the intrinsic quantum property of spin resulting from the nucleons, which are protons and neutrons. The overall spin of a nucleus is determined by the spin quantum number, I, which is zero if the number of protons equals the number of neutrons. In this case the accordant nucleus have no magnetic moment, and hence no possibility to interact with the external magnetic field. The angular momentum, $\vec{J}$, associated with nuclear spin is quantized as following:

$$|\vec{J}| = \sqrt{I(I+1)} \cdot \hbar, \tag{1.1}$$

where I is the spin quantum number, which can take integer or half-integer values and $\hbar$ is the reduced Planck constant ($h/2\pi$). The associated magnetic quantum number, m, can take values from $-I$ to $+I$ in integer steps. A nucleus can take $2I+1$ states, i.e. if $I = 1/2$ two states $m_z = \pm 1/2$, else if $I = 1$ three states $m_z = -1, 0, 1$. In an external magnetic field along a defined z-axis the z-component of the magnetic moment is defined as:

$$\mu_z = \gamma I_z = \gamma \hbar m_z. \tag{1.2}$$

The energy of a spin in a magnetic field of strength B_0 can be deviated from the classical formulation of a magnetic dipole in an external magnetic field:

$$E = -\mu_z B_0 = -\gamma I_z = -\gamma \hbar m_z B_0. \tag{1.3}$$

The external magnetic field B_0 is aligned along the z-axis. The energetically favored state for spins with $I = 1/2$ is $m_z = +1/2$ (α-state) and the energetically unfavorable state is $m_z = -1/2$ (β-state). The energy of the α- and β-state, E_α and E_β, and the resulting energy difference, ΔE are defined as:

$$E_\alpha = -\gamma \hbar \left(\frac{1}{2}\right) B_0 \tag{1.4}$$

$$E_\beta = -\gamma \hbar \left(-\frac{1}{2}\right) B_0 \tag{1.5}$$

$$\Delta E_{\alpha \rightarrow \beta} = E_\alpha - E_\beta = \gamma \hbar B_0 \tag{1.6}$$

Absorption will only occur if the frequency of the electromagnetic radiation, hv, matches the energy difference between the nuclear spin levels, ΔE. The Larmor frequency ω_0

(Equation 1.7), expressed in rad·s^{-1}, specifies at which angular frequency resonance occurs, i.e. energy transition between the α- and β-state. ω_0 is characteristic for every spin in an external magnetic field B_0.

$$\omega_0 = \gamma B_0 \tag{1.7}$$

The most important nuclear spins for NMR spectroscopy are $I = 1/2$-spins. Important properties are the gyromagnetic ratio γ and the natural abundance of the isotopes. Popular isotopes for the investigation of proteins are ^{1}H, ^{2}H, ^{15}N and ^{13}C. $I = 1$ spins are less suited for NMR spectroscopy, because they have a quadrupole moment additionally to the magnetic moment which enables a pathway for much faster relaxation.

Each nucleus in a molecule has a characteristic chemical shift value. The chemical shift value depends on the gyromagnetic ratio and the strength of the effective magnetic field B_{eff} that the nucleus experiences. This effective magnetic field B_{eff} is influenced by the chemical environment of the nucleus, i.e. the electron distribution. The electrons around the nucleus have a shielding effect towards the external magnetic field B_0. Therefore, each nucleus in a molecule experiences a slightly different effective magnetic field which leads to individual characteristic chemical shift values.

Due to the dependence of the chemical shift value on the magnetic field it is convenient to use a scale that is independent from the field strength. Tetramethylsilane (TMS) is a referencing compound that is used to define zero on the chemical shift scale for ^{1}H and ^{13}C. All signal positions are expressed by their difference towards the referencing compound. The chemical shift δ, measured in ppm, is defined as:

$$\delta(\text{ppm}) = 10^6 \cdot \frac{v - v_{ref}}{v_{ref}}, \tag{1.8}$$

where v is the Larmor frequency of the signal of interest and v_{ref} is the reference signal.

As already noted before, spin-1/2 nuclei can adopt two states, i.e α and β, in a magnetic field. The population difference between the α-state, N_α, and the β-state, N_β, can be described by the Boltzmann distribution:

$$\frac{N_\alpha}{N_\beta} = e^{-\frac{\gamma \hbar B_0}{kT}}, \tag{1.9}$$

where k is the Boltzmann constant and T is the absolute temperature. For ^{1}H the energy difference between those two states at room temperature is close to the thermal energy.

Accordingly, both states are almost equally occupied. Nevertheless, the alignment along the external magnetic field is energetically favored. Consequently, when considering an ensemble of spins a macroscopic bulk magnetization, M_z, aligned with the external field, can be observed. The size of the equilibrium magnetization, M_z, mainly depends on the strength of the magnetic field B_0 and the temperature of the sample. The achievement of equilibrium magnetization, as well as changes in equilibrium magnetization can be visualized by a vector model. Only the z-component of the spins is definite, the y- and x-component have completely random phases, which often is symbolized by two cones.

Radio frequency (rf) pulses can be used to disturb the thermal equilibrium by applying short rf-pulses B_1, which match the resonance condition, along the x-axis. The M_z vector is rotated around B_1 towards the xy-plane. The duration of the rf-pulse determines the angle through which the magnetization vector turns. It is common to use 90°- and 180°-pulses. For example, a 90°-x-pulse rotates the bulk magnetization vector to the $-y$-axis.

The precession of the magnetization vector at Larmor-frequency is detected by a coil placed in the xy-plane. The magnetization vector induces a voltage in the coil which is amplified and then recorded. The recorded signal is called free induction decay (FID). In order to perform computer based processing, the NMR signal has to be converted from a voltage to data points with certain intensities first. The analogue to digital converter (ADC) samples the FID at regular intervals Δ:

$$\Delta = \frac{1}{2f_{max}}, \tag{1.10}$$

where f_{max} represents the maximal frequency that can be represented correctly. Δ is the dwell time, which specifies the distance between two data points.

The FID is a time-domain signal that oscillates at Larmor-frequency and decays over time due to relaxation processes. There are two main relaxation mechanisms: First, the longitudinal relaxation T_1 (spin-lattice relaxation), that is responsible for restoring the thermal equilibrium distribution of the spin population. T_1 is enabled by small fluctuating magnetic fields of nearby spins which allows an exchange of energy with their neighboring spins (the lattice). Second, the transverse relaxation T_2 (spin-spin relaxation), which is responsible for the exponential decay of the detectable transverse magnetization vector (FID). Transverse relaxation is fundamentally caused by variations in spin precession frequencies, which leads to a loss of their phase coherence.

The FID, recorded as a time-domain signal, is a superposition of all the individual oscillations, at their Larmor-frequencies, from all detected nuclei in the sample. In order to extract the frequency-domain signal, Fourier-transformation (FT) is used. FT is a simple mathematical procedure that multiplies the FID by trial cosine functions. The area under the resulting function represents the intensity of the signal at the corresponding frequency. This frequency intensity information is stored in a data matrix, where each data point corresponds to a specific frequency with certain intensity.

The exponential decay and the length of the recorded FID influences the line width and shape of the resulting peaks in the spectrum, respectively. Therefore, the acquisition time of the FID has to be chosen carefully. In typical NMR experiments the line shapes are in absorption mode, i.e. being entirely positive and symmetrical about the maximum.

A simple 1D NMR experiment starts with a relaxation delay, t_r, of a few seconds that lets the spin come to equilibrium, i.e. bulk magnetization is build up. This is followed by a short rf-pulse that lasts a few μs (preparation phase). As explained above, this leads to a detectable signal in the xy-plane, i.e. FID. The FID is then recorded for a specific time, called acquisition time (t_{acq}), lasting between 50 ms and a few seconds (detection phase). FT of the FID results in a 1D NMR spectrum. Usually, recording a single FID is not sufficient due to sensitivity reason. Therefore, it is common to use time averaging by repeating the above explained procedure N times and adding up the FID, which improves N times thereby. This severely improves the signal-to-noise ratio, since the random noise only adds up by $\sqrt{N}$.

Interpretation of 1D NMR spectra is impractical for complex molecules due to signal overlap. Signal overlap can however be resolved by adding another spectral dimension. This is shortly explained by the example of a 2D NMR experiment. A 2D NMR experiment includes additionally to the preparation and detection phase of the 1D experiment, an indirect preparation (evolution) phase and a mixing period. After the preparation phase, including for example a 90°-x-pulse, spins can precess freely for a fixed time period t_1. During this period the magnetization is labeled with the chemical shift of the first nucleus. In the mixing period the state of the magnetization after t_1 is retrieved and magnetization is transferred from the first nucleus to another nucleus. In essence, two mechanisms for magnetization transfer are available, J-coupling and dipolar coupling (explained below). Final stage of the 2D experiment is the acquisition time, t_{acq}. During this period the magnetization is labeled with the second nucleus. FT of t_{acq} would yield a simple 1D spectrum. Typically, the evolution period t_1 is incremented by time steps Δt_1. Thereby, one can observe the chronological development of the spins. FT of t_1 yields a 2D spectrum.

1.1.3 Nuclear Spin Interactions

In this chapter only nuclear spin interactions with magnetic fields originating from other spins are explained in detail. Interactions arising from the external magnetic field B_0 and fields induced by rf-pulses are explained above.

J-coupling

The origin of J-coupling (also called scalar coupling or spin-spin coupling) is an indirect interaction through the electrons involved in the chemical bond between two atoms. J-couplings hold valuable information on the angle and distance between two atoms, and most importantly indicate atom connectivities in the molecule. NMR experiments that use J-coupling as magnetization transfer mechanism give information about the covalent structure of the molecule. The strength of the coupling decays with the numbers of bonds separating the coupled atoms. A valuable specific coupling is the three-bond J-coupling. It has been identified that these couplings vary with the dihedral torsion angle θ between the participating atoms. This relation is known as Karplus relation (Karplus , 1963):

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

where A, B, and C are empirical determined coefficients.

Dipolar coupling

Nuclei generate magnetic fields and can also receive magnetic fields from other nuclei. This, in contrast to J-couplings, through-space dipolar interaction is crucial for structure determination by NMR. The interaction between two nuclei i and j is defined as:

$$D_{ij}(\theta) = d_{ij}\left(1 - 3\cos^2\theta\right) \tag{1.12}$$

with the coupling constant d_{ij}:

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

where μ_0 represents the permeability of the vacuum, γ_i and γ_j represent the gyromagnetic ratios of nuclei i and j, respectively, r_{ij} represents the inter-nuclear distances between nuclei i and j, and θ describes the angle between the inter-nuclear distance vector and the external magnetic field. The strength of the dipolar coupling d_{ij} depends on (i) the distance between the two contributing spins ($1/r_{ij}^3$), (ii) the gyromagnetic ratio (the larger the gyromagnetic ratio, the larger the magnetic moment, the larger the local field), and (iii) the orientation of the vector between the two spins relative to the external magnetic field B_0, θ, which is usually averaged in isotropic media.

The most relevant experiment for structure determination by NMR spectroscopy is based on dipolar couplings, i.e. the NOESY experiment (Nuclear Overhauser Enhancement SpectroscopY). Basis of this experiment is the Nuclear Overhauser Enhancement (NOE), which can be described by cross relaxation between dipolar interacting spins.

1.1.4 Restraint sources for NMR structure calculation

Nowadays, there is a diversity of experimental information available from NMR spectroscopy which can be used to obtain protein structures, e.g. distance restraints from Nuclear Overhauser Enhancement (NOE), Paramagnetic Relaxation Enhancement (PRE) effect (Battiste & Wagner, 2000), or information about hydrogen bonding (Grzesiek *et al.*, 2004), J-couplings (Wüthrich, 2003), chemical shifts (Cavalli *et al.*, 2007; Shen *et al.*, 2008), and residual dipolar couplings (RDCs) (Bax & Grishaev, 2005; Blackledge, 2005). Nevertheless, distance restraints, especially from NOE experiments, and dihedral restraints obtained from chemical shifts are still the main source of structural information used in the conventional approach of structure determination by NMR spectroscopy (Guerry & Herrmann, 2011). In the following the standard procedures of distance restraints, torsion angle restraints, and the unique information available from RDCs, are explained.

Distance restraints

NOE experiments represent the most important source of information for NMR structure calculation. It is possible to calculate structures of proteins solely based on NOE derived distance restraints (chapter 2.2 and 4). Physical basis of the NOE are dipolar couplings as explained above. A NOE can be observed if two protons come sufficiently close in space, usually $<$ 5-6 Å. If the contributing protons are distant in the underlying primary sequence, the observation of a NOE directly leads to a significant limitation of the accessible conformational space.

A source of long distance information are PRE experiments, which can yield distance restraints from 15-24 Å (Battiste & Wagner, 2000). PRE experiments require the introduction of a spin label, i.e. a molecule which possesses an unpaired electron. The PRE arises from dipolar interactions between the unpaired electron on the spin label and a nucleus within the protein. These dipolar interactions result in an increase of the nuclear relaxation rates. Effects of the spin label on NMR signals with respect to the unlabeled sample can be (i) disappearance of signals, (ii) broadening of signals, or (iii) unaffected signals. A disappearance of the signals implies a distance < 15 Å between the spin label and the nucleus. A signal broadening represents a distance in the range of 15-24 Å between the spin label and the nucleus. If the signal is not affected by the spin label, the distance between spin label and nucleus is > 24 Å. The usage of PRE experiments is an option when one wants to investigate low populated states (Clore & Iwahara, 2007), larger proteins (Battiste & Wagner, 2000) or membrane proteins in detergent micelles (Gottstein *et al.*, 2012b; Reckel *et al.*, 2011).

Hydrogen bonds (H-bonds), which mainly stabilize secondary structural elements, provide another source of distance information relevant for structure calculation. They can be determined by NMR through amide proton exchange rates. Intra-molecular H-bonds have a much slower exchange rate and can therefore easily be distinguished from transient H-bonds to surrounding water molecules (Vuister *et al.*, 2011). Another approach to detect H-bond scalar couplings is the application of COSY-type experiments (Grzesiek *et al.*, 2004).

Dihedral angle restraints

The direct source of angle restraints for protein structure determination is the chemical shift which depends strongly on local structures, i.e. the secondary structure of the protein. Protein backbone dihedral angle restraints can be derived from secondary chemical shifts, i.e. the difference between the observed chemical shift and the random coil chemical shift. This can be performed by the well-established program TALOS+ (Shen *et al.*, 2009), which predicts Φ and Ψ values empirically.

^3J-coupling constants can also be used to derive angular information by the Karplus relation (Equation 1.11). However, this is complicated by several factors, (i) Several dihedral angles can correspond to a single ^3J value, which can be solved by directly incorporating the ^3J-coupling constant in the structure calculation (Vuister *et al.*, 2011); (ii) the Karplus coefficient underlie the experimental uncertainty, because they are obtained by comparing measured values and dihedral angles obtained from X-ray or NMR structures (Vuister *et al.*, 2011); (iii) the magnitude of the ^3J-coupling constant depends on other factors, e.g.

geometric distortions or hydrogen bonding (Vuister *et al.*, 2011).

Orientational restraints

Equation 1.12 shows that the dipolar coupling of two nuclei i and j depends on the distance between the two nuclei and the orientation of the distance vector relative to the external magnetic field. In solution this orientational component, θ, of the dipolar coupling vanishes as a result of isotropic tumbling. However, residual dipolar couplings (RDCs) can be reintroduced by weakly aligning the sample in a suitable medium (Tjandra & Bax, 1997). They give information on the orientation of local groups relative to an alignment tensor. In order to extract structural information, it is necessary to measure the sample under isotropic and aninsotropic conditions. Then, the difference in total coupling within these media is determined. If the dipolar interaction is measured between two directly covalently bonded nuclei, the distance is fixed and the coupling depends only on the orientation of the inter-nuclear vector with respect to the alignment tensor (Chen & Tjandra, 2012). In order to use RDCs in structure determination, refinement or validation, it is necessary to determine the alignment tensor, i.e. the diagonal elements of the Saupe matrix (amplitude and rhombicity) and its Euler angles (Vuister *et al.*, 2011). The extraction of structural information from RDCs is further complicated by their ambiguous nature.

1.2 Automation of protein NMR structure determination

Several NMR spectroscopy experiments give information on inter-atomic distances and specific angles which can be used as input for structure calculation algorithms. In these algorithms a set of random polypeptide chains is generated and folded through MD simulations in Cartesian or torsion angle space by simulated annealing (SA) (Kirkpatrick *et al.*, 1983). The potential energy landscape of the polypeptide is represented by a target function. The goal is to minimize the target function which ideally would become zero if all input constraints were fulfilled. In principle, the value of the target function represents the agreement between the resulting structure and the experimental input restraints. The final result of an NMR structure calculation is an ensemble of structures which represents the experimental input constraints adequately. An overview of some of the most popular NMR structure calculation programs is given in Tab. 1.1. Structure calculation results presented in this thesis have exclusively been performed with CYANA, therefore algorithmic details explained in the following correspond to CYANA if not stated otherwise.

Table 1.1: Prominent programs for NMR structure determination.

Name	**Reference**
ARIA	(Bardiaux *et al.*, 2009)
CNS	(Brünger *et al.*, 1998)
CYANA	(Güntert *et al.*, 1997; Güntert & Buchner, 2015)
J-UNIO	(Serrano *et al.*, 2012)
XPLOR-NIH	(Schwieters *et al.*, 2003)

Structure determination by NMR spectroscopy needs several sequential steps which are summarized in Fig. 1.1. The darker shaded steps are described in more detail in the following. These steps have a high potential towards automation. Principle reasons for automating these steps are; first, the process becomes more objective and independent from individual decisions of the user. Second, automation saves a tremendous amount of time. Third, the method becomes more accessible towards scientists that are not necessarily experts in NMR.

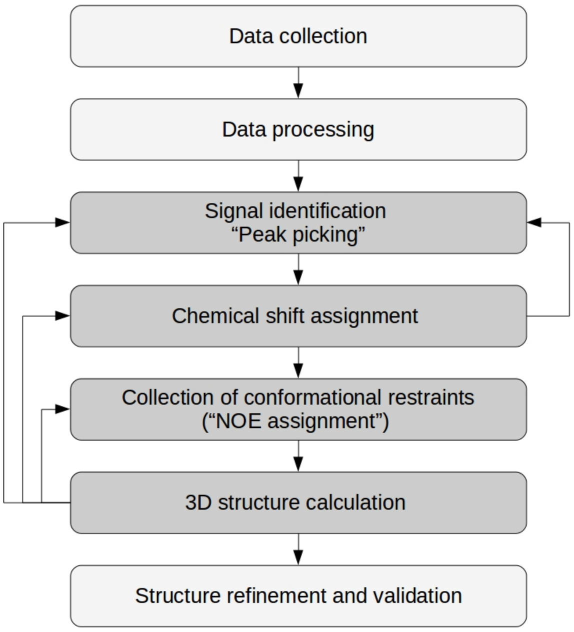


Figure 1.1: Flowchart of the individual steps in structure determination by NMR. Steps presented in dark gray are described in the following in more detail. Additionally, these steps can be performed iteratively, indicated by back-cycling arrows.

1.2.1 Signal identification

Signal identification in NMR, i.e. identification of the exact ppm positions of protein signals and their intensity (in the following called *peak picking*), is one of the crucial steps in NMR data analysis. The identified peak lists serve as input for chemical shift assignment and NOE assignment, which results in distance restraints that are used in structure calculation. Accordingly, peak lists have to be as accurate as possible because errors in peak lists add up with every step presented in Fig. 1.1 and can ultimately result in incorrect structure bundles. When performed manually, peak picking is more or less an iterative process where one picks and assigns unambiguous signals. Based on the knowledge gained from the assigned signals, one searches for further signals which are possibly weaker and/or overlapped. Successful automation of NMR peak picking is still lacking. Reasons therefore are, cross-peak overlap and artifacts which can be attributed to baseline distortions, intense solvent lines, ridges and/or sinc wiggles (Baran *et al.*, 2004). A detailed presentation of prominent peak picking algorithms follows in section 2.1.

1.2.2 Chemical shift assignment

Each NMR sensitive nucleus leads to a peak in a NMR spectrum with a distinct chemical shift value which originates from the unique chemical environment of the nucleus. Assigning each of these unique chemical shift values to the correct atom is called *chemical shift assignment*. The method for chemical shift assignment, which was mainly developed by Kurt Wüthrich is called *sequence specific assignment* (Wüthrich, 1986; Billeter *et al.*, 1982; Wagner & Wüthrich, 1982). Basis of this method is, first, to identify all spin systems (grouping) in a spectrum and to connect these spin systems to a certain type of amino acid (typing), and, secondly, to link the resonances of each amino acid $i + 1$ to a neighboring amino acid i (linking), and finally to map these segments of spin systems to the primary sequence (mapping) (Baran *et al.*, 2004). When using unlabeled proteins, spin-system signals are manifested by through-bond coupling connections whereas sequential links are realized by through-space dipolar couplings. However, when assigning proteins of size $\geq$ 8 kDa it is common practice to use uniformly ^{13}C and ^{15}N isotopically labeled proteins which makes the assignment much more feasible. In this case, a typical set of triple resonance NMR spectra includes experiments for backbone assignment, which generate connectivities between the NH group of residue i and carbon atoms of residue $i - 1$, i.e. HNCO, HN(CO)CA, CBCA(CO)NH, and experiments which yield connectivities between N_i and C^{α}_{i}, as well as N_i and C^{α}_{i-1}, i.e. HN(CA)CO, HNCA, and HNCACB (Lian & Barsukov, 2011). In general cross-peaks which originate from sequential connections are weaker than cross-peaks from intra-residual connections. The assignment strategy is in principle the same for labeled proteins as for unlabeled proteins, first, grouping all HN correlations into spin-systems, and, secondly, assembling the spin-system in a sequential manner that is also manifested by through-bond coupling (Lian & Barsukov, 2011). A typical set for side-chain assignment consists of (H)CC(CO)NH, HCC(CO)NH, HCCH-TOCSY and ^{15}N-edited TOCSY (Shin *et al.*, 2008).

A significant number of algorithms for either backbone or complete chemical shift assignment have been introduced in the last couple of years (reviewed in (Baran *et al.*, 2004; Guerry & Herrmann, 2011)). These algorithms usually implement the above explained strategy of spin-system grouping, typing to specific amino acids, linking into sequential segments and mapping onto the primary sequence (Baran *et al.*, 2004). However, the main difference between these algorithms is in the mapping step which is solved by methods like exhaustive search, best-first, Monte-Carlo or genetic algorithms. Some popular algorithms are namely AutoAssign (Zimmerman, 1997), MATCH (Volk *et al.*, 2008), MARS (Jung & Zweckstetter, 2004), GARANT (Bartels *et al.*, 1996, 1997), PINE (Bahrami *et al.*, 2009), and FLYA (Schmidt & Güntert, 2012).

Noteworthy progress has been accomplished with the development of the automated chemical shift assignment algorithm FLYA (Schmidt & Güntert, 2012). The FLYA algorithm holds advantages over other automated methods in terms of (i) quality of the resulting backbone and side-chain assignment, (ii) the universality of NMR experiments to be utilized, (iii) robustness towards imperfect input lists (Schmidt & Güntert, 2012), and (iv) the versatility of applications, i.e. chemical shift assignment solely based on NOESY spectra (Schmidt & Güntert, 2013a), RNA assignment (Aeschbacher *et al.*, 2013; Krähenbühl *et al.*, 2014) and assignment of solid-state NMR samples (Schmidt *et al.*, 2013b).

1.2.3 NOE assignment

The classical approach towards structure determination by NMR (Williamson & Craven, 2009) relies on a set of assigned distance restraints from NOESY experiments together with dihedral angle restraints predicted from chemical shifts. Source of the NOESY cross-peaks is the magnetization transfer through dipolar couplings (cross-relaxation) between spatial close nuclear spins. The intensity of the resulting signal depends on the inverse of the distance between the two atoms. In order to get conformational information from a NOESY spectrum the signals have to be assigned. This means the atom-pairs which give rise to the NOE cross-peaks have to be identified (Wüthrich, 1986). NOE assignment is a very time-consuming step when performed manually. Thus, several algorithms have been developed to perform this task automatically: NOAH (Mumenthaler & Braun, 1995; Mumenthaler *et al.*, 1997), ARIA (Bardiaux *et al.*, 2009), AutoStructure (Huang *et al.*, 2006), KNOWNOE (Gronwald *et al.*, 2002), CANDID (Herrmann *et al.*, 2002b), and PASD (Kuszewski *et al.*, 2004). However, those algorithms have to have a high tolerance towards ambiguity which is especially high at the beginning of the assignment process (Guerry & Herrmann, 2011).

The CANDID NOE assignment strategy implemented in CYANA (Güntert, 2004) identifies the atom pairs giving rise to the NOESY cross-peaks in an iterative manner. CYANA usually performs seven cycles of NOE assignment and structure calculation plus an additional final structure calculation. In each cycle CYANA is fed with the sequence, the sequence-specific resonance assignment, the list of unassigned NOESY cross-peak positions and volumes, and with the 3D structure from the previous cycle, except in the first cycle. In the first cycle peaks are initially assigned based on the closeness of the chemical shift coordinates and the cross-peak position. This condition however, should not be over-interpreted because the spectra used for chemical shift assignment could be misaligned towards the NOESY spectra or there might be referencing offsets (Guerry & Herrmann, 2011). The assigned cross-peaks are converted into restraints which are used as input in a preliminary structure calculation. This preliminary structure is used as a filter in the following assignment steps among other validation criteria (explained below). In the final cycle all remaining ambiguous assignments (explained below) are transformed to unambiguous distance restraints.

As already indicated before, a difficulty in the NOE assignment process is the distinctive width of individual peaks and the limited accuracy of peak position measurements. Accordingly, usually more than one chemical shift matches the chemical shift closeness condition. Hence, the method of ambiguous distance restraints (Nilges, 1993, 1995) has been introduced, which helps to solve the problem of uncertainty in the NMR data. An ambiguous distance restraint represents all the possible contributions that match the chemical shift of the cross-peak. The assignment is transferred into an upper distance limit based on the peak volume. If the assignment of the cross-peak is unique, the upper bound is calculated from the intensity of the peak into the inter-atomic distance d_{AB} between atoms A and B. For ambiguous assignments the distance has to be refined with the effective r^{-6}-

summed distance over all individual distances d_k (Nilges, 1993), leading to the effective distance d_{eff}:

$$d_{\text{eff}} = \left(\sum_{k=1}^{n} d_k^{-6} \right)^{-\frac{1}{6}}, \tag{1.14}$$

where n represents the number of assignments belonging to one peak. All ambiguous distance restraints belonging to the same cross-peak get the exact same upper limit value which is calculated from the effective distance. The ambiguous distance is shorter than any of the individual distances belonging to the peak. If the correct assignment is among all ambiguous assignments, the restraints will be fulfilled by the correct structure. This concept enhances the probability of finding the correct assignment and reduces the distorting effect of wrong assignment on the structure.

The initial NOE assignments are realized via the matching of the chemical shift and the peak coordinates within an atom-type specific tolerance, as noted above. Those pre-validated assignments are evaluated by three criteria that are visualized in Fig. 1.2: (i) the chemical shift of the atom pair (ω_{A} and ω_{B}) and the peak position (ω_1 and ω_2) need to be within a given tolerance range ($\Delta\omega$). The closeness of the match is expressed by a Gaussian probability (P_{shift}) (Fig. 1.2 **A**). (ii) The pre-validated assignment has to be compatible with the intermediate 3D structure ($P_{structure}$), i.e. atom-pair distances in the structure bundle (d_{AB}) have to be lower than the calibrated upper limit (upl_{AB}) (Fig. 1.2 **C**). Assignments whose $P_{shift} \cdot P_{structure}$ product does not exceed a cycle-dependent threshold are erased from the set of potential assignments. However, the difficulty with using the structure filter is to distinguish between true violations and violations caused by insufficient convergence due to sparse data for example (Guerry & Herrmann, 2011). Accordingly, the native fold of the structure should roughly be achieved in the first cycle. (iii) Assignments which survived the first filtering process are evaluated by their anchoring in a network ($P_{network}$) that is build-up from all the pre-validated assignments. The basic idea of network-anchoring (Herrmann *et al.*, 2002b) (Fig. 1.2 **B**) is that a set of correct distance restraints forms a consistent network of paths. Correct assignments between atoms A and B are incorporated in that network and are supported by paths through a third atom C, e.g. other pre-validated assignments or the covalent polypeptide structure. Erroneous assignments on the other hand, should not be embedded in the network and are thus classified as not being anchored. The network-anchoring score (N_{AB}) is the sum over all indirect paths AB through a third atom C, being not equal to A or B. Network-anchoring serves as replacement of the 3D structure filter in the first CANDID cycle. The concept of Network-anchoring is replaced in alternative NOE assignment protocols like KNOWNOE (Gronwald *et al.*, 2002) by Bayesian statistics. In these algorithms the most probable assignment is determined using information from a database of known structures.

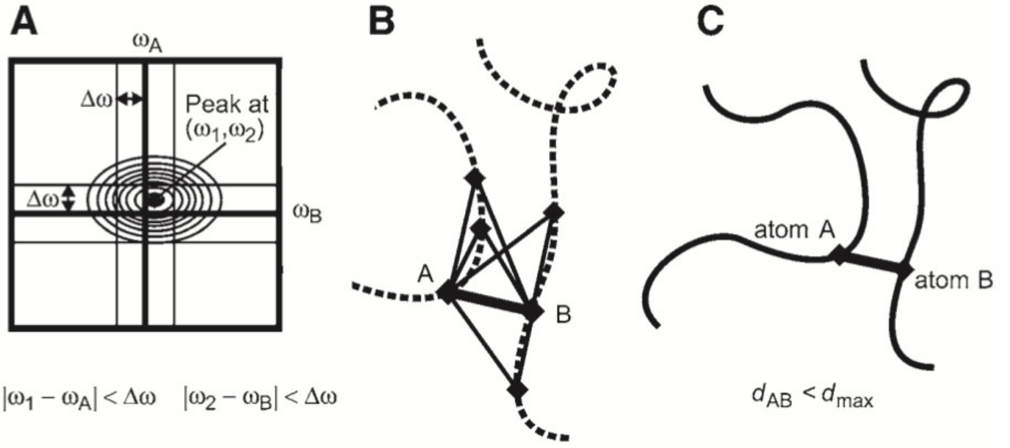

A
ω_A
$\Delta\omega$
Peak at (ω_1,ω_2)
$\Delta\omega$
ω_B
$|\omega_1 - \omega_A| < \Delta\omega$ $|\omega_2 - \omega_B| < \Delta\omega$
B
A
B
C
atom A
atom B
$d_{AB} < d_{max}$

Figure 1.2: Conditions that have to be fulfilled by a valid NOE assignment to two atoms A and B. **A** The chemical shift of the atom pair A and B and the cross-peak position have to be within a given tolerance. **B** A set of correctly assigned distance restraints forms a consistent network of paths. Each assignment is evaluated by a score representing the sum over all indirect paths AB through a third atom C. **C** Assignments have to be compatible with an intermediate 3D structure (Figure from (Güntert, 2004)).

The overall probability of an assignment is calculated as follows:

$$P_{total} = P_{shift} \cdot P_{structure} \cdot P_{network} \tag{1.15}$$

Remaining assignments which survived the P_{total} filter, now called distance restraints, are calibrated into upper limits using the $1/r^6$ dependence of the peak intensity towards the distance.

Additionally, in automated NOE assignment cycles 1 and 2 the concept of constraint combination is employed in order to eliminate artifact assignments. The concept of constraint combination was first introduced in the CANDID algorithm (Herrmann *et al.*, 2002b). As the name implies, virtual constraints are generated by combining assignments of two unrelated peaks into one new distance restraint. The principle idea behind constraint combination is the same as of ambiguous distance restraints: if one correct assignment is among the combined ambiguous distance restraints the resulting structure is not distorted by incorrect assignments.

It has been reported that errors in the automated NOE assignment have a severe impact on the accuracy of the resulting structure bundle (Jee & Güntert, 2003; Buchner & Güntert, 2015b). These errors can be attributed to sparse data sets, incorrect chemical shift assignment. or inaccurate NOESY peak picking. It is therefore possible that automated procedures can converge towards incorrect protein structures. The identification of incorrect NMR protein structures is not trivial due to the lack of a reliable structure quality validation tools. However, in many cases the precision of the structure expressed as the backbone RMSD radius and the agreement between the experimental data and the resulting structure bundle, expressed in the target function (explained below) are used instead as a measure of structural quality (Nabuurs *et al.*, 2006). Even though, those measures are not capable of distinguishing between correct and incorrect results. Due to the fact that fundamental calculation errors are rarely expressed in terms of precision of the final structure bundle, but rather in a precise but inaccurate structure, a new consensus structure bundle method implemented in CYANA has been introduced recently (Buchner & Güntert, 2015a). In case of principle errors in the NOE assignment procedure the final result of a calculation depends to some extent on the input starting structure, that is influenced by a random number generator seed. The new consensus structure bundle method comprises 20 individual combined automated NOE assignments and structure calculations starting from different initial input structures. The final distance restraint sets from each of the 20 individual calculations are combined into one consensus set of distance restraints. The consensus distance restraint set is used as input for a structure calculation which yields the consensus structure bundle (Buchner & Güntert, 2015a).

1.2.4 NMR structure calculation

NMR structure calculation can be performed by molecular dynamics (MD) simulation in torsion angle or Cartesian space. However, MD with torsional angles instead of Cartesian coordinates is more efficient. Reasons for this are, (i) the torsion angles are the only degrees of freedom, (ii) the geometric force field retains only the most important non covalent interactions in a simplified manner, and (iii) the simulation can be performed with longer time steps since bond length and bond angles are kept fix in torsion angle space (Güntert, 2004).

CYANA performs structure calculation by a MD simulation in torsion angle space driven by simulated annealing (SA) (Kirkpatrick *et al.*, 1983). This task is realized via a target function V optimization strategy. The target function serves as the potential energy and can take values $V \geq 0$, whereas $V = 0$ only if all upper and lower distance and torsion angle constraints are fulfilled and all non bonded atom pairs satisfy a check for the absence of steric overlap. The exact definition of the CYANA target function V (Güntert *et al.*, 1991, 1997) is as following:

$$V = \sum_{c=u,l,v} w_c \sum_{(\alpha,\beta)\in I_c} (d_{\alpha,\beta} - b_{\alpha,\beta})^2 + w_a \sum_{i\in I_a} \left[1 - \frac{1}{2}\left(\frac{\Delta_i}{\Gamma_i}\right)^2\right] \Delta_i^2. \tag{1.16}$$

The target function considers upper and lower distance bounds ($b_{\alpha,\beta}$) for the distance ($d_{\alpha,\beta}$) between the atoms α and β, as well as torsion angle constraints (θ_i) in the allowed ranges $[\Theta_i^{max}, \Theta_i^{min}]$, corresponding to constraint set I_a. I_u, I_l, and I_v represent the set of atoms α and β with upper, lower and van der Waals restraints, respectively. w_u, w_l, w_v and w_a demonstrate the weighting factors for different types of constraints. The torsion angle violation is considered in the second part of equation 1.16. Δ_i is the quantity of the violation and Γ_i is the half width of the forbidden range of torsion angle values, which can be computed as:

$$\Gamma_i = \pi - \frac{\Theta_i^{max} - \Theta_i^{min}}{2}. \tag{1.17}$$

It is possible to extent the target function with terms for scalar coupling constants, residual dipolar couplings, pseudocontact shifts, and identity and symmetry restraints for calculation of symmetric multimers.

In oder to calculate the torsional acceleration in the torsion angle dynamics algorithm of CYANA the molecule is represented by a tree structure consisting of $k + 1$ rigid bodies connected by k rotatable bonds. Each rigid body compromises one or several mass points (atoms) for which relative positions are fixed. The rigid bodies are numbered from 0 to k. The tree structure starts at the N-terminus and terminates at the side-chain of the C-terminus. Each rigid body, except the base, has a single nearest neighbor in the direction towards the base, which has a number $p(k)$ lower than k, visualized in Fig. 1.3. The torsion angle between the rigid body $p(k)$ and k is denoted by Φ_k. The conformation of a molecule is uniquely specified by the value of all torsion angles, $\phi = (\phi_1, \phi_2, \phi_3 \ldots \phi_k)$, which are the only degrees of freedom in a torsion angles dynamics calculation (Güntert, 2004).

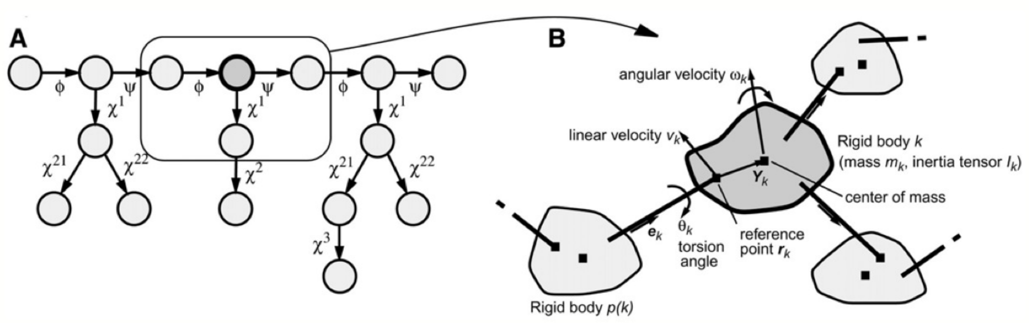


Figure 1.3: Schematic overview of the molecule representation for torsion angle dynamics calculation. **A**: Tree structure of torsion angles for tripeptide the Val-Ser-Ile. The circles represent the rigid bodies (atoms or rings). Rotatable bonds are depicted by arrows which point towards the part of the structure that is rotated upon corresponding torsion angle change. **B**: Excerpt from the tree structure with an overview of all the quantities required by the CYANA torsion angle dynamics algorithm (Figure from (Güntert, 2004)).

Quantities shown in Fig. 1.3 allow to calculate the kinetic energy, E_{kin}, of the entire system in a recursive and efficient way, which is implemented as described in (Jain *et al.*, 1993). Together with the potential Energy, E_{pot}, given by the above introduced target function, it is possible to solve the accordant equations of motion.

Even though using a simplified force field and a target function, the potential energy landscape of a protein is still complex compromising numerous local minima. Therefore, it is desirable to have a method which has the potential to overcome energy barriers in order to reduce the probability of getting caught in a local minimum. SA is a general method to find the global minimum of a given function, in this case the target function V. The temperature is a parameter that is proportional to the kinetic energy. A high kinetic energy allows a molecule to overcome energy barriers and prevents the system from getting trapped in a local minimum. The SA protocol in CYANA starts form a random conformation which is generated from the protein amino acid sequence and characterized by independent torsion angles. The initial minimization stage contains 100 conjugate gradient steps including only restraints that are 3 residues apart form each other. This is followed by another round of 100 conjugate gradient steps including all distance restraints. In the first three stages of the protocol a check for steric overlap is conducted which excludes hydrogen atoms. The repulsive core radii of heavy atoms are enhanced by 0.15 Å. In the second stage the protein is virtually heated to 10,000 K. The high temperature allows to overcome energy barriers and the molecule can adopt every possible conformation. One-fifth of all N torsion angle dynamic steps are performed at that temperature. The remaining torsion angle dynamic steps are performed in the third stage where the temperature is slowly annealed to 0 K. In the fourth stage the hydrogen atoms are added and 100 conjugate gradient steps, followed by 200 torsion angle dynamic steps are performed. The final stage includes 1,000 conjugate gradient steps. This five stage procedure is repeated several times for different initial random structures, which all lead to a final structure with a local energy minimum.

As noted above the target function only includes the physical component in a simplified manner. Therefore it is necessary to refine the final structure using a full physical force field in explicit solvent.

Other prominent structure calculation bundles are Xplor-NIH (Schwieters *et al.*, 2003) and CNS (Brünger *et al.*, 1998). Both Xplor-NIH and CNS can perform structure calculation in Cartesian or torsion angles space and use simulated annealing methods (Baran *et al.*, 2004). Additionally, Xplor-NIH and CNS incorporate a MD simulation in explicit water for energy minimization and structure refinement.

1.2.5 Alternative NMR structure determination approaches

Chemical-shift based methods

Chemical-shifts are NMR parameters which can be determined straightforward and with a high degree of accuracy. They are sensitive towards the conformation of native and non-native molecule states. In structural biology chemical shifts are predominantly used to predict secondary structures, to guide in structure refinement, and to characterize conformational changes. Often, only chemical shifts are available. In this case it is useful to have an approach that can determine the structure directly from chemical shifts.

This can be achieved by using 3D structures and corresponding chemical shifts to extract molecular fragment conformations that match the experimentally determined secondary chemical shifts of the protein under investigation. Molecular modeling approaches are then used to assemble the fragments into 3D structures. Popular methods to perform chemical-shift based structure determination are CHESHIRE (Cavalli *et al.*, 2007), CS-ROSETTA (Shen *et al.*, 2008; van der Schot *et al.*, 2013), and the CS23D web server (Wishart *et al.*, 2008).

RDC-based methods

RDCs can be measured by weakly aligning the molecules which leads to a measurable orientational dependence of the dipolar coupling with respect to the alignment tensor. The molecular fragment replacement (MFR) (Delaglio *et al.*, 2000) approach is used to determine structures based on RDCs. The first steo of this method is to search the PDB for fragments that fit the measured RDCs. In the next step torsion angles from the database matches are used to assemble a protein structure. These structures are then refined in an iterative manner, by adjusting backbone torsion angles in order to minimize differences in the measured and fit RDCs and between the measured and predicted chemical shifts. RDC-ROSETTA (Rohl &Baker, 2002) can be used to improve the database search for relevant fragments in ROSETTA by taking the agreement between RDC data into account analog to the above explained CS-ROSETTA method.

Assignment-free methods

Chemical shift assignment is often considered the most time consuming and tedious step in NMR protein structure determination. Therefore, assignment-free methods have been developed that rely exclusively on the distance information in NOESY spectra. An example is the CLOUDS (Grishaev & Llinás, 2002) protocol where NOESY spectra are used to determine a spatial distribution from proton-proton distances that are calculated based on a relaxation matrix analysis. A MD simulated annealing scheme is used to generate a set of structured proton clouds. A model structure is fit into this spatial distribution. CLOUDS has been further developed to SC-CLOUDS (Bermejo & Llinás, 2008) which can use data from perdeuterated proteins and final structures are assembled by ROSETTA.

1.2.6 Fully automated protocols

The usage of fully automated structure determination protocols for NMR has so far not been completely established. This task is much more demanding than automating individual steps due to cumulative errors (Güntert, 2004). It has been reported that reliable NOE assignment needs at least 90% completeness in chemical shift assignment (Jee & Güntert, 2003). A detailed analysis of combined automated NOE assignment and structure calculation with CYANA (Buchner & Güntert, 2015b) also revealed that 10% missing or erroneous chemical shifts result in inaccurate structures with RMSD bias above 3 Å. It has also been presented in that work that the algorithm is relatively robust towards missing peaks, errors in peak positions and volumes, and lower resolution (Buchner & Güntert, 2015b). Improvements in the usage of automated protocols has been achieved through the application of SAIL amino-acids (Kainosho *et al.*, 2006), which results in sharper lines and reduced signal overlap, hence improved spectral quality (Takeda & Kainosho, 2011).

Prominent programs for fully automated analysis are FLYA (López-Méndez & Güntert, 2006), AUREMOL (Gronwald *et al.*, 2004), and J-UNIO (Serrano *et al.*, 2012).

Chapter 2

Peak picking in multidimensional NMR spectra with CYPICK

This chapter is based on the following publication:

Würz J. and Güntert P. Peak picking in multidimensional NMR spectra with the contour geometry based algorithm CYPICK. *Journal of Biomolecular NMR* (in press).

2.1 Introduction

Identifying real signals in an NMR spectrum also known as peak picking plays a central role in the process of NMR protein structure determination. The resulting peak lists serve as basis for chemical shift and NOE assignment, followed by structure determination. Peak lists provide, among other information, the position and the intensity of the signals and are directly accessible by interactive or automated spectral analysis programs. The quality of peak lists can be described by: first, including as many true signals as possible, second, including as few artifacts as possible, and third, the accuracy of the signal positions and intensities.

It has been reported that peak lists do not have to be flawless to be used for chemical shift assignment, NOE assignment and subsequent structure calculation. For example, the automated resonance assignment algorithm FLYA (Schmidt & Güntert, 2012) can yield more than 90% correct resonance assignments even in extreme cases where 60% of true peaks are missing or for data sets containing 5 times more artifacts than true peaks in the input (Schmidt & Güntert, 2012). Automated NOE assignment and structure calculation with CYANA (Güntert & Buchner, 2015; Herrmann *et al.*, 2002a) can also tolerate imperfections in input resonance assignment and NOESY peak lists (Buchner & Güntert, 2015a; Jee & Güntert, 2003). 10% missing or erroneous resonances can result in structures with an RMSD bias above 3 Å. Missing NOE peaks do not affect NOE assignment as severely as missing or erroneous resonances. It has been further reported that the random deletion of 30% NOESY peaks results in structures with an RMSD bias below 3 Å, whereas the random deletion of 45% NOESY peaks increases the RMSD bias only slightly above 3 Å (Buchner & Güntert, 2015a). This can be explained by the fact that NOE data sets usually include a high amount of redundant and short-range data (Chapter 3), which can be excluded from the data set without a loss in structural information. It has further been reported that the deletion of 30% weak NOESY peaks has a comparable effect to randomly deleting 30% of all peaks. However, a deletion of 45% weak peaks results in a significant increase in RMSD bias to 7 Å (Buchner & Güntert, 2015a). This can be attributed to the fact that weak signals correspond to important structural long-range information.

The task of automating peak picking remains challenging. Reasons for this are: identification of real signals, and in case of NOESY-based spectra, weak signals, is impeded by low signal-to-noise, signal overlap, and artifacts, e.g. baseline distortions, intense solvent lines, ridges or sinc wiggles. Automating the step of peak picking in the process of structure determination saves a great amount of time and makes peak picking much more objective. Therefore, a great deal of work has already been invested in the automation of peak picking in NMR spectroscopy. Existing algorithms can be classified into four groups: (1) threshold-based methods, (2) methods that depend on symmetry criteria, (3) peak-shape based methods, and (4) incorporating peak picking in the design of the experiment. Combinations of this classification has been implemented in many algorithms. Interactive spectrum analysis programs like XEASY (Bartels *et al.*, 1995), SPARKY (Goddard *et al.*, 2005), NMRVIEWJ (Johnson, 2004; Johnson & Blevins, 1994), or CCPNMR ANALYSISASSIGN (Skinner *et al.*, 2016; Vranken *et al.*, 2005) (in the following referred to as CCPN) allow the user to adjust thresholds manually and allow peak picking by identifying all local extrema above the adjusted threshold automatically. These methods are useful as a starting point for semi-automated peak picking which is refined manually afterwards. WAVPEAK (Liu *et al.*, 2012) performs the following sequence of steps: wavelet-based smoothing, identifying all local extrema, and volume-based filtering. PICKY (Alipanahi *et al.*, 2009) is an SVD-based automated peak picking method. Machine learning and computer vision approaches have also been employed for peak picking, e.g. CV-PEAK PICKER (Klukowski *et al.*, 2015). AUTOPSY (Koradi *et al.*, 1998) includes functions to determine a local noise level and to separate overlapping signals from resolved signals by measurements of peak uniformity. Overlapping regions are then resolved with the help of unique line shapes derived from resolved signals. ATNOS (Herrmann *et al.*, 2002b) is an automated peak picking software exclusively for NOESY spectra, that uses chemical shift assignment information to guide peak picking. Peak picking can be performed in an iterative manner during NOE assignment and structure calculation, making use of preliminary structural information. Peak picking can be part of NMR data processing, e.g. MUNIN (Orekhov *et al.*, 2001) that uses a three-way decomposition to decompose a three-dimensional (3D) NMR spectrum into a sum of components defined as the direct product of three 1D shapes. The GAPRO peak identification algorithm (Hiller *et al.*, 2005) produces peak lists for high-dimensional (e.g. 4D, 5D, 6D) APSY-type spectra by picking peaks in the experimentally recorded tilted 2D projections.

The human approach to peak picking can be described as the analysis of shape and regularity of two-dimensional contour lines. Real signals are similar to concentric ellipses and have a variety of common properties which artifacts do not share, e.g. peak width, convexity or similarity. The shape of NMR signals can be described by a mixture of Lorentzian and Gaussian shape. Usually NMR signals are displayed in absorption mode in order to achieve minimal peak width, and positive, symmetric peak shapes. However, real signals can deviate from the proposed 'perfect' shape for a number of reasons, such as limited digital resolution, spectral overlap and improper phasing of the spectrum. An automated peak picking procedure should be able to handle these imperfections and shortcomings. Accordingly, a peak picking approach which tries to mimic the human way of analyzing these criteria of similarity and symmetry in 2D spectral planes is a promising approach for an automated procedure. This approach has first been implemented in the COMMON SENSE APPROACH TO PEAK-PICKING (Garrett *et al.*, 1991), short CAPP. Our aim was to develop an effective and fast automated peak picking procedure in CYANA which can be directly linked to chemical shift assignment and/or NOE assignment, followed by structure calculation. We reduced the requirements with respect to user intervention as far as possible in order to increase objectivity and reproducibility in comparison to manual peak picking. The new peak picking algorithm CYPICK is introduced in this chapter.

In the following sections some of the above introduced peak picking algorithms are explained in more detail.

ATNOS

Atomated NOESY peak picking (ATNOS) (Herrmann *et al.*, 2002b) is a software which is used to obtain NOESY peak list from 2D or 3D homonuclear NOESY spectra. The software is usually combined with automated NOE assignment (Herrmann *et al.*, 2002a) and structure calculation routines to obtain a direct link between intermediate structures and raw NMR data. Accordingly, the used NOESY peak list changes with every cycle of NOE assignment and structure calculation. As input the sequence specific resonance assignment of the protein in XEASY (Bartels *et al.*, 1995) format, the amino acid sequence and the spectra are required. Additional conformational restraints can also be included.

The peak picking algorithm follows several steps: ATNOS determines the local baseline and the local noise level. The baseline determination is based on the FLATT (Güntert & Wüthrich, 1992) algorithm, whereas the local noise determination is derived from AUTOPSY (Koradi *et al.*, 1998). A signal is considered as a peak if a defined signal-to-noise ratio and a minimal local extremum condition are met. A preliminary chemical shift-based NOE assignment is established with CANDID. All signals that have an assignment which can be linked to a covalent NOE, an NOE which arises from the covalent geometry (Güntert *et al.*, 1998), i.e. bond length, bond angles and Karplus relation, are used to determine spectrum-specific threshold values for the minimal signal-to-noise ratio and the minimal peak area. Based on these threshold values the initial NOESY peak list is filtered. In the first cycle no intermediate structure is available. Peaks are validated by classification into two groups: First, signals with and without and structure-imposed NOE-observable upper limit, and second, signals that are within a limit from the diagonal or solvent regions and the remaining peaks. Peaks that are not close to the diagonal and the solvent signal are kept, all others are discarded. In the second and following cycles peaks are only categorized in a set of signals which are compatible with the intermediate structure and those that are not. Further validation criteria applied in all cycles are the compatibility with the resonance assignment, network anchoring and symmetry considerations. Signals that pass these filters are used as input for CANDID (Herrmann *et al.*, 2002a) and DYANA (Güntert *et al.*, 1997).

AUTOPSY

The first step of the AUTOPSY algorithm is the determination of the local noise level. Afterwards the spectrum is divided into connected data points which are above the noise level via a ‘flood fill’ algorithm (Foley *et al.*, 1990). These connected regions are filtered for separated peaks based on symmetry and regular shape criteria. In order to determine symmetry violations, a function which includes a set of data points and a set of symmetrized data points is minimized with the symmetry center as parameter. The uniformity of peaks has proven to be a good measure for well-resolved peaks. In general, a 2D peak can be factorized into the product of two line shapes and an amplitude. The nonlinear system can be solved by an iterative method where starting values of the two line shapes are expected and then one line shape value is kept fixed while the other is modified. The outcome of this is the error in difference of the combination of line shapes and the data points. This error correlates with peak uniformity. The line shapes and the chemical shifts of these peaks are stored and sorted according to their degree of separation. The stored line shapes are then filtered for unique line shapes and shifts which can be found in several peaks. The line shapes are grouped by a clustering algorithm. Regions of the spectra which are strongly overlapping are resolved with the help of unique line shapes and the selected peaks are integrated. A symmetrization step can be performed on symmetric spectra, e.g. NOESY or COSY experiments.

CAPP

CAPP (Garrett *et al.*, 1991) is a peak picking algorithm which tries to imitate the spectroscopist's behavior as realistically as possible by evaluating contour line shapes. The algorithm follows four steps. First, a contour diagram is generated on a logarithmic intensity scale by defining a cutoff level and a level multiplier. Neighboring contour points are linearly interpolated. Second, ellipses are calculated which best fit the contour lines. The center and the radius of each ellipse are determined and optimized by the simplex method as RMS of the deviation of each contour point and the closest point on the ellipse. Third, definition of noise ridges. In order to define ridges in dimension x, peaks are sorted by their y shift. Peaks which are within a defined range are used for the definition of ridges, if the sum of the radii in x exceeds a minimal length or the set of peaks includes a minimal number of peaks with x chemical shifts which exceed a minimal length. Fourth, definition and localization of real peaks. A peak is defined if the conditions: (a) the RMS between contour point and ellipse is less than a predefined value, (b) its radius is within a defined range, (c) the ratio of ellipse radii has to be within a defined range and (d) the circumference ratio between ellipse and contour line should be in a defined range, are all fulfilled. Ellipses which fulfill requirement (a) through (d) are defined as peaks and checked for the following requirements: (i) peak center is not a ridge, (ii) at least 2 ellipses define the peak and (iii) in case of 3D or 4D spectra the peak has to be a local maximum in each dimension.

CCPN

CCPN (Skinner *et al.*, 2016; Vranken *et al.*, 2005) aims to connect computational tools used in NMR spectroscopy, especially in terms of different program data formats. In the CCPN software suite the CcpNmr Analysis program is responsible for the complete interactive analysis of NMR data using a graphical interface. Results from the analysis tool can easily be connected to structure calculation (ARIA) and validation software (e.g. QUEEN). The CcpNmr FormatConverter allows conversion of data in-between all the common used data formats. Peak picking by CCPN can be performed by manually adjusting a spectrum-specific threshold and automatically identifying all local extreme above the defined threshold.

CV-Peak Picker

CV-Peak Picker (Klukowski *et al.*, 2015) is an automated peak picking procedure that uses computer vision and machine learning methods for identifying real signals in NMR spectra. Peaks are identified on the basis of local extrema conditions. The volume of the identified peaks are then calculated and peaks below a user-defined threshold are discarded. CV-Peak Picker uses the 'Histogram of Oriented Gradients'-method to extract features from the peak shapes. A Support Vector Machine is used as a binary classifier, comparing all recognized peak shapes to a trained and manually identified peak shape set.

EASY/XEASY

The program EASY (Eccles *et al.*, 1991) was designed for spectral analysis of biomolecular 2D NMR spectra. It includes routines for automated peak picking, spin-system identification, sequential resonance assignment and cross-peak integration. The successor XEASY (Bartels *et al.*, 1995) includes all the EASY functionalities and is capable of also analyzing 3D and 4D spectra. EASY and XEASY provide an interactive and an automatic peak picking mode. In the interactive mode the user selects peaks with the mouse cursor by means of the 2D contour plot. Automated peak picking can be performed on anti-phase peaks or on in-phase peaks. Anti-phase peaks have a symmetry towards their center which can easily be monitored by an algorithm which uses a symmetry function (Meier *et al.*, 1987). The peak center is identified as the local maximum of the symmetry function above a user-defined threshold. In-phase peak picking is performed by selecting all extrema above a global threshold which is usually one to two times the noise level. Those peaks require to have a peak width in a predefined range in order to exclude noise spikes or artifacts stemming from errors in baseline correction or solvent signals.

NMRView and NMRViewJ

The program NMRView (Johnson & Blevins, 1994; Johnson, 2004) can be used for visualizing and analyzing 2D, 3D or 4D NMR data. The package includes routines for automated peak picking, analysis, and also aids in assignment. Via the Molecular Data Viewer it is possible to correlate the calculated structure to the underlying spectra. The newer NMRView 3.0 version (Johnson, 2004) enables automation of individual steps by incorporating the Tool Command Language (Tcl). Peak picking in NMRView is performed by identifying local extrema and interpolating the exact position of the maximum.

Peaks are valued as 'good' and 'bad' based on the peak width at the threshold. Further peak information, like the peak width a half height and the peak bounds are determined and stored. Afterwards the user can interactively delete or add peaks using the graphical user interface. Further, simple Tcl scripts can be written that filter the peak lists based on certain user-defined criteria, e.g. peak width.

Non-negative matrix factorization

Peak picking NMR spectral data using non-negative matrix factorization (NMF) (Tikole *et al.*, 2014) was developed to automatically decompose overlapping peaks from non-uniform sampling schedules. A basic 2D NMF model is used and sequentially extended to 3D to decompose the corresponding data tensor. Factorization convergence is measured by an Euclidean distance cost function.

Sparky

SPARKY (Goddard *et al.*, 2005) is a program which can be used to display NMR spectra, to pick, assign and integrate peaks via a graphical user interface. It is possible to load spectra with up to 4 dimensions simultaneously. In order to pick peaks the program first represents the spectrum as a contour plot. Characteristics of the plot can be adjusted, e.g. the contour levels can be scaled. Single peaks can be picked manually and it is also possible to pick distinct regions or the whole spectrum based on a threshold which is defined as the lowest positive and/or negative contour level. The line width can also be defined and used as a picking criteria.

STELLA

The program STELLA (Kleywegt & Kaptein, 1989) provides both a fully automatic and a fully manual peak picking algorithm for low-symmetry spectra. The fully automatic picking is performed by the modules LEARN2 and SMART2. LEARN2 is an interactive program which learns to distinguish between real and erroneous peaks based on peak shape representations. The user classifies a set of real and a set of spurious peaks for LEARN2 which in turn creates a peak-definition file. The subsequent program SMART2 can use the peak-definition file as input. Nevertheless, the peak-definition file is not mandatory. The following restrictions can be made by the user: (a) the search area can be defined, (b) definitions of local maximum can be determined, and (c) definitions of a real peak can be set. A data point is considered a local maximum if the intensity exceeds a certain threshold, the data point is not too close to the diagonal, the data point has to be the local maximum in a defined area, and a predefined number of neighboring points have to be higher than a specific threshold. If a peak-definition file has been established via LEARN2, a match factor of the pre-selected local extrema is determined on the basis of the peak shape. Peaks are then sorted by their match factor. The *k*-Nearest-Neighbor algorithm is used to classify the pre-selection into real and spurious peaks. If no peak-definition file is provided all local extrema are treated as peaks. Finally, the peaks are interpolated and all relevant information is stored in a binary peak file. The peak picking algorithm can also be used in an iterative fashion by manual re-evaluation of the output peak file of SMART2 and using an edited input for LEARN2.

2.2 Contour Peak Picker Algorithm

As explained above, the human approach to peak picking can be described as the analysis of shape and regularity of two-dimensional contour lines. Real signals are similar to concentric ellipses and have a variety of common properties which artifacts do not share, e.g. peak width, convexity or similarity. Fig. 2.1 contrasts the 2D contour plots of a real signal and noise.

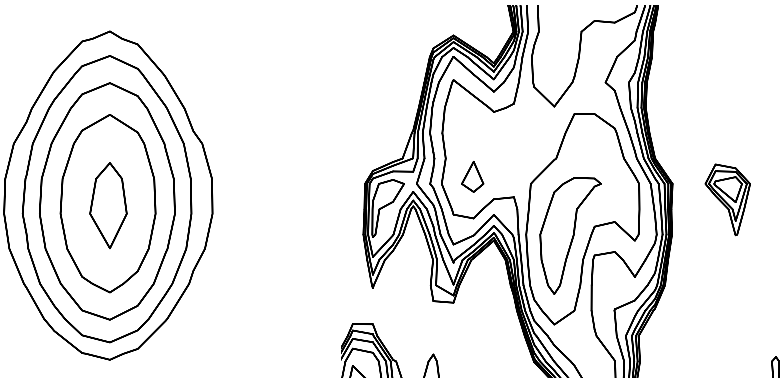

(a) Real signal **(b)** Noise

Figure 2.1: Contrasting juxtaposition of the contour lines of a real signal (a) and a noise region (b). The contour lines of the real signal possess a high degree of symmetry and similarity. In contrast, noise regions do not fulfill these criteria. A decrease in contour level intensity results in more irregular and jagged contour lines.

A simplified graphical overview of the different steps and picking modes available for the contour peak picker are presented in Fig. 2.2. CYPICK is implemented in the CYANA software package.

The first step in the process depicted in Fig. 2.2 is to read the processed NMR spectrum. NMR spectra are stored in the form of an intensity matrix. The position in the matrix corresponds to the chemical shift value, whereas the entry itself corresponds to the intensity of the data point. The routine for reading NMR spectra was implemented by Dr. Donata Kirchner for BRUKER and UCSF format. Within this work, this reading routine was extended to XEASY and AZARA format. After storing the spectrum in memory, an estimate of the intensity of the lowest contour level is required. This can either be the global noise level (CYPICK command: *spec noise*) of the spectrum or the local noise level (CYPICK command: *spec localnoise*) at each data point. The following step is to find local maxima. This can either be done over the complete spectral range (CYPICK command: *spec pick contour*) or by a frequency filter, which can be provided in the form of a peak list (CYPICK command: *spec pick filter*). Then contour lines are created.

The contour lines belonging to the local extremum are then filtered and analyzed. The remaining local maxima are stored in a peak list (CYANA command: *write peaks*). Details are explained in the following subsections.

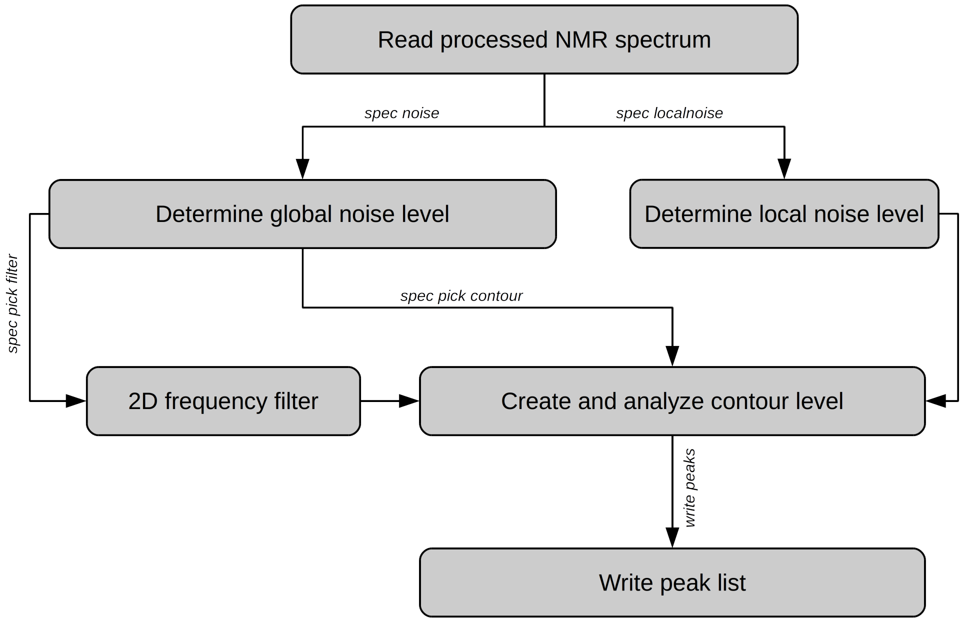


Figure 2.2: Graphical overview of the CYPICK peak picking algorithm implemented in CYANA. The algorithms needs as input a processed NMR spectrum. Three different picking modes are available for the contour approach: First, the global noise level is determined and used as intensity of the first contour line (CYPICK command: *spec pick global*). Second, a restricted peak picking with a 2D frequency filter that can be provided in the form of a peak list (CYPICK command: *spec pick filter*). The position of the 2D peaks is used as a filter for local maxima which are considered in the contour approach. Third, a local noise level is determined for every data point in the spectrum (CYPICK command: *spec pick local*). The local noise level is used as a first filtering step for local maxima and creation of contour lines. In the restricted peak picking mode the global noise level is needed for the selection of local maxima and the estimation of contour level intensities. After appropriate contour line creation, contour lines are analyzed and peak lists are generated.

2.2.1 Determination of the global noise level

The global noise level, L_{global} is approximated by estimating the median of the absolute intensity values of the spectral data points as implemented in the program PROSA (Güntert *et al.*, 1992), which assumes that most of the data points in a multidimensional NMR spectrum are at locations not occupied by signals. The method applied for median approximation is described in the book ‘Numerical Recipes’ (Press *et al.*, 2007).

2.2.2 Determination of the local noise level

The local noise level, $L_{local}(\omega_1, \ldots, \omega_D)$ at a given position $(\omega_1, ..., \omega_D)$ in a D-dimensional spectrum, estimation is a reimplementation of the local noise level estimation used in the program AUTOPSY (Koradi *et al.*, 1998).

Each one-dimensional slice, i.e. each row and column of a 2D spectrum, is subdivided into segments equal in size to 5% of the number of data points in the accordant dimension. The segment with the smallest standard deviation within slice ω_n represents a specific noise level for that slice in dimension n and referred to as δ_{n,ω_n}. Individual noise levels are represented as a base noise level, being a characteristic value for the complete spectrum, plus an additional noise level, being characteristic for a specific one-dimensional slice. The base noise level is defined as the minimal value out of all individual noise levels δ_{n,ω_n}, as follows:

$$\delta_{min} = \min_{n,\omega_n}(\delta_{n,\omega_n}) \quad \text{for} \quad n = 1, \ldots, D \quad \text{and} \quad \omega_n = 1_n, \ldots, I_n, \tag{2.1}$$

where D represents the dimensionality of the spectrum and I_n represent the number of one dimensional slices in the respective dimension n. The additional noise, δ'_{n,ω_n}, is then calculated as:

$$\delta'_{n,\omega_n} = \sqrt{\delta^2_{n,\omega_n} - \delta^2_{min}}. \tag{2.2}$$

The noise level at a data point with coordinates $(\omega_1, \ldots, \omega_n)$ is calculated from the noise level of all the slices that pass through that data point and the base noise level:

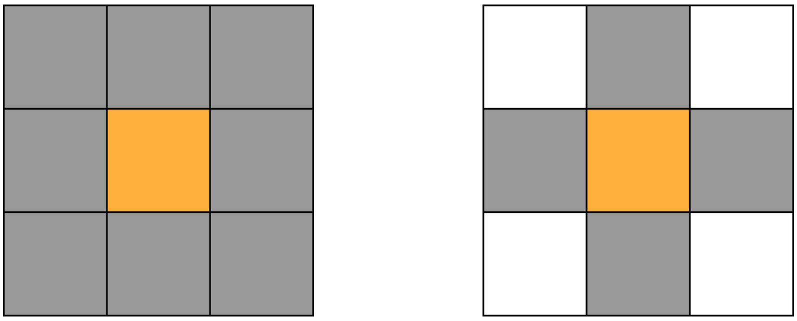
$$L_{local}(\omega_1, \ldots, \omega_n) = \sqrt{\sum_{n=1}^{D} \delta'^2_{n,\omega_n} + \delta^2_{min}} = \sqrt{\sum_{n=1}^{D} \delta^2_{n,\omega_n} - (n-1) \cdot \delta^2_{min}}. \tag{2.3}$$

2.2.3 Determination of the local extremum

A data point is only checked for being a local extremum if its intensity exceeds a spectrum-specific base level. The base level, B, represents the intensity of the lowest contour level c_0 and is defined as:

$$B = c_0 = \beta \cdot L, \tag{2.4}$$

where L denotes either the global noise level L_{global}, or the local noise level L_{local} at this position, depending on the desired picking mode. β is an empirical factor between 2 and 3 (Koradi *et al.*, 1998). If not stated otherwise we used β =3.0. This factor can be adjusted by the user. However, it is strongly recommended to use the default values for reasons of objectivity and reproducibility. A data point is considered a local extremum if all the neighboring data points have an absolute intensity lower than or equal to the investigated data point. Two modes are provided which are shown schematically for a two-dimensional spectrum in Fig. 2.3.



(a) *diagonal* mode **(b)** *no diagonal* mode

Figure 2.3: Comparison of *diagonal* and *no diagonal* local extremum determination. The orange data point represents the data point of interest. The grey neighboring points are the points which are analyzed in the accordant method.

In the *diagonal* mode all $3^D - 1$ neighbors are considered, whereas in the *no diagonal* case only the direct $2 \cdot D$ neighbors are taken into consideration, where D represents the number of dimensions. Generally speaking, the *no diagonal* mode provides a faster scanning of the spectrum, whereas the *diagonal* mode provides more accurate results. The *diagonal* mode was used for calculations presented in this study.

2.2.4 Scaling of contour lines

While creating the contour lines they are scaled by a scaling factor σ_i, the peak width at half height in ppm for the corresponding dimension i which has to be provided by the user for each individual dimension $i = 1, ..., D$ (see chapter 2.2.7.1). For example, in a ^{15}N-HSQC spectrum with a peak width at half height of 0.4 ppm in the ^{15}N dimension and of 0.04 ppm in the ^{1}H dimension, the ^{1}H dimension is scaled with factor 10.0 and the ^{15}N dimension with factor 1.0. The user however only provides $\sigma_1 = 0.4$ and $\sigma_2 = 0.04$, as explained in chapter 2.6.

2.2.5 Creation of contour lines

If the data point of interest has been classified as a local extremum, contour lines are determined in a defined peak area. The peak area can in principle be changed by the user according to the expected peak width in each dimension. However, the algorithm is insensitive towards the size of the peak area as long as it is chosen sufficiently large. Larger peak ares, however, do enhance the calculation time.

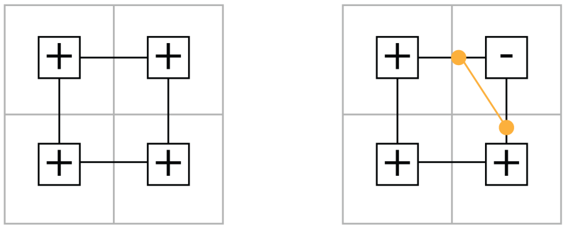
The next higher contour line c_{i+1} is determined by multiplying the preceding contour line, c_i by the factors γ.

$$c_{i+1} = c_i \cdot \gamma. \tag{2.5}$$

$\gamma = 1.4$ approximately doubles the height of contour line i after 2 contour lines. Whereas, when using multiplication factor of $\gamma = 1.2$ the height of contour line i is approximately doubled after 4 contour lines. Accordingly, with lower level multipliers one generates more contour lines within a given intensity range. We extensively tested the usage of different contour level multipliers and therefore recommend using $\gamma = 1.3$.

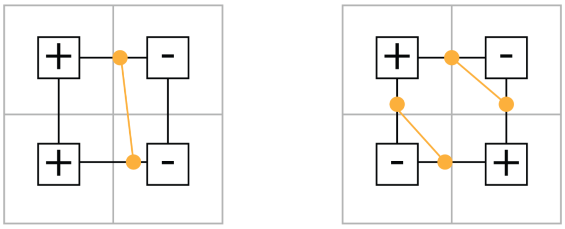
The number of contour lines encircling a local extremum depends on its absolute intensity. After having defined the peak area and the number of contour lines, the actual position of the contour points are determined by an algorithm which is very similar to the marching squares algorithm (Lorensen & Cline, 1987). This algorithm is a reimplementation of the plotting algorithm used in the program PROSA (Güntert *et al.*, 1992). PROSA was developed to generate closed contour lines, which is vital for the analysis in CYPICK.

The marching squares algorithm is perfectly suited for this situation because the spectral data points are already rasterized on a regular grid. The peak area is further subdivided into sets of 4 data points (2 x 2 squares). Each of these sub-squares is checked for having an intensity higher or lower as the contour line intensity. 16 cases can be distinguished which can be further reduced to 4 cases under consideration of symmetry. These four cases are shown in Fig. 2.4. Based on these cases contour points describing the contour line are determined by a simple linear interpolation procedure. Contour points are stored in an array for each local extremum if the contour line is closed and encircles the local extremum. Otherwise the algorithm searches for another closed contour line in the defined search space.



Case 1: All data points are higher than the contour line intensity

Case 2: All data points except one are higher than the contour line intensity



Case 3: Two data points at one edge are higher than the contour line intensity

Case 4: Two data points on opposite sites are higher than the contour line intensity

Figure 2.4: Schmematic representation of the marching squares algorithm. Data points in a NMR spectrum are rasterized on a regular grid. The algorithm then subdivides this regular grid into sets of 4 data points (2 x 2 squares). Each sub-square is checked for having an intensity higher (denoted by '+') or lower (denoted by '−') as the contour line intensity, c_i. The four possible cases are depicted. The position of contour points (denoted by orange circles) is estimated via linear interpolation.

2.2.6 Filtering of contour lines

After having stored all contour lines belonging to all local extrema above the defined base level, the contour lines are subjected to a preliminary filtering process:

1. The local extremum of interest has to be inside the contour line.
2. No other local extremum except the local extremum of interest may be inside the contour line.
3. The number of contour points per contour line has to be greater than or equal to five. If the number of contour points describing an contour line is lower than 5, the shape of the contour line does not resemble a concentric ellipse anymore.
4. The number of contour lines per local extremum has to be greater than or equal to four. This ensures that the contour lines are covering an intensity range that is greater than twice the intensity of the first contour line when using 1.3 as level multiplier.

5. The number of closed contour lines per local extremum has to be greater than or equal to two. This ensures that parts of the peak are within the predefined peak area.

In order to check if the local extremum of interest is within the contour line we used the ray casting algorithm (Shimrat, 1962) which relies on the assumptions of Jordan's polygon theorem. In this procedure, a ray is sent out from the point of interest, in this case the local extremum, and the number of times it intersects with the edges of the contour line are counted. If the ray crosses the boundary of the contour line, it goes from inside to outside, then from outside to inside, and so on. As a result, after every two 'border crossings' the ray goes inside. Accordingly, odd numbers of 'border crossings' imply that the point of interest is inside of the contour line, whereas even numbers imply that the point is outside of the contour line.

2.2.7 Analysis of contour lines

After filtering, the remaining contour lines are analyzed starting from the contour line with the highest absolute intensity. If the highest contour line does not fulfill the requirements, the next lower contour line is analyzed. At least two contour lines have to fulfill the specified conditions.

2.2.7.1 Circular shape

The first condition that has to be fulfilled by a contour line belonging to a real signal, is the circular shape. As mentioned earlier, the shape of contour lines can be described by concentric ellipses. However, the width of overlapping peaks can deviate severely. Therefore, the analysis of overlapping signals is performed independently as explained in chapter 2.2.9.

NMR contour lines consist of contour points. Connecting contour points by a line results in a polygon. To this end, we availed ourselves of polygon equations. The contour line area is determined via Gauss's area formula (Braden, 1986):

$$2\mathrm{A} = \sum_{i=1}^{n} (x_i + x_{i+1}) \cdot (y_{i+1} - y_i), \tag{2.6}$$

where n reflects the number of contour points in the contour line, and x and y represent the coordinates of the contour points. The circumference of a contour line can be determined by summing up all distances between the corresponding contour points:

$$\mathrm{C} = \sum_{i=1}^{n} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}, \tag{2.7}$$

where n reflects the number of contour points in the contour line and x and y represent the coordinates of the contour points. For a circle, the area, A_{circle}, to circumference, C_{circle}, squared ratio is defined as follows:

$$\frac{\mathrm{A}_{circle}}{\mathrm{C}^2_{circle}} = \frac{\pi \cdot r^2}{(2\pi \cdot r)^2} = \frac{1}{4\pi} \approx 0.08 \tag{2.8}$$

The area-to-circumference squared ratio is determined for each contour line that survives the filtering step presented in the previous section. This ratio should equal $\frac{1}{4\pi}$ in case of a perfect circle. We then use this ratio and convert it into a quality factor Q_{rad} within the range of $[0, 1]$. In this case 1 reflects a perfectly circular shape. Q_{rad} was designed in a way that also deviations from the perfect circular shape, which can still be present even though the contour lines were scaled, are considered (see Fig. 2.5). Q_{rad} is defined as:

$$Q_{rad} = e^{-750\cdot\left(\frac{a}{c^2}-\frac{1}{4\pi}\right)^2}. \tag{2.9}$$

Fig. 2.5 (a) shows Q_{rad} as a function of area to circumference square. The tolerated range is shaded in grey ($Q_{rad} \geq 0.7$). In Fig. 2.5 (b) examples of ellipses with varying eccentricities and their corresponding Q_{rad} values are visualized.

2.2.7.2 Convexity

Contour lines around a peak extremum are required to form an approximately convex polygon, i.e. all interior angles are less than 180°. The interior angle α_i is calculated via the scalar product of the two edges forming the vertex i. The convexity of contour points is checked via the cross product of the two edges connected by the vertex. For some of the real signals a slight deviation from the perfect convex shape could be observed. Therefore, a quality factor Q_{con}, similar to Q_{rad}, within the range of $[0, 1]$ was defined. $Q_{con} = 1.0$ corresponds to a convex contour point. $Q_{con,i}$ is calculated for each contour point of the contour line. Q_{con} is the product of all $Q_{con,i}$ values belonging to the same contour line.

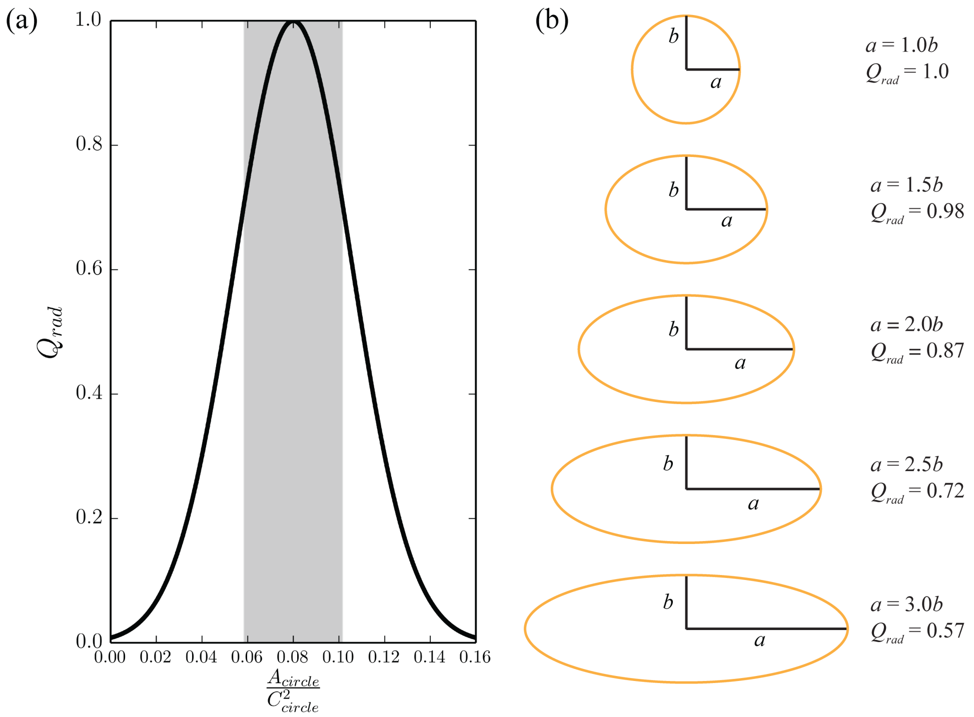


Figure 2.5: (a) Q_{rad} as a function of area to circumference square. (b) Ellipse examples and their accordant Q_{rad} values.

$Q_{con,i}$ and Q_{con} are defined as:

$$Q_{con,i} = \begin{cases} 1, \quad \alpha_i \leq \pi \\ \left(\frac{2\alpha_i}{\pi} - 3\right)^2, \quad \pi < \alpha_i < \frac{3\pi}{2} \\ 0, \quad \alpha_i \geq \frac{3\pi}{2} \end{cases} \tag{2.10}$$

$$Q_{con} = \prod_{i=1}^{n} Q_{con,i}, \tag{2.11}$$

where n reflects the number of contour points in the contour line, and α_i represents the angle at edge i. The dependence of $Q_{con,i}$ on the angle α_i visualized in Fig. 2.6 (a). The tolerated range is shaded in gray ($Q_{con} \geq 0.7$). Fig. 2.6 (b) shows polygons with varying angle α and the accordant Q_{con} values.

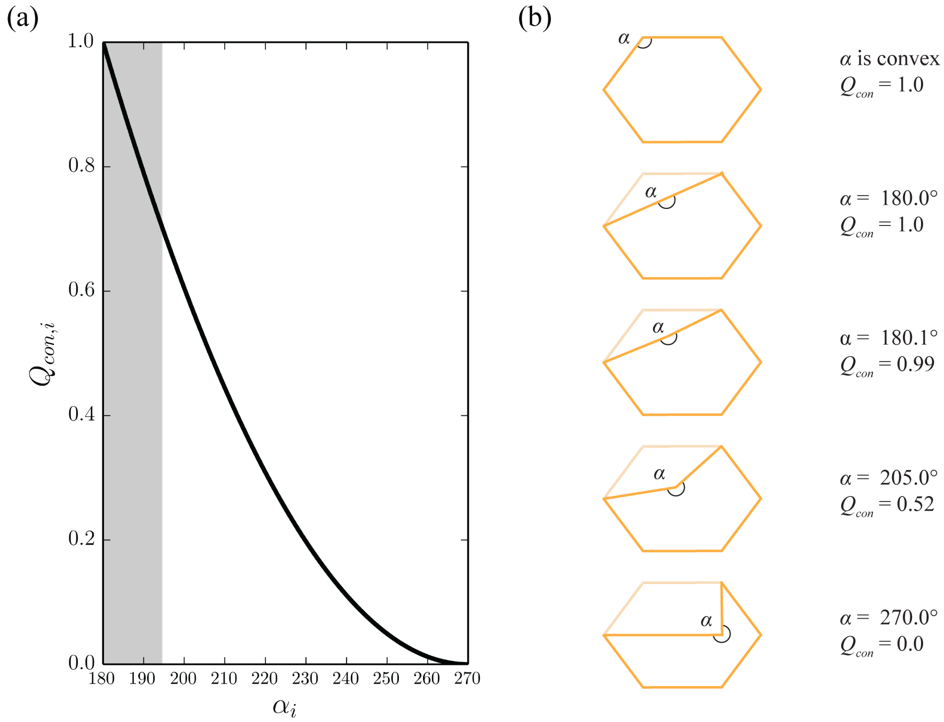


Figure 2.6: (a) $Q_{con,i}$ as a function of α_i. (b) Examples of different convex and concave angles and their accordant Q_{con} values.

2.2.7.3 Absolute quality

Q_{rad} and Q_{con} values were calculated for real and artifact signals of numerous spectra. From our investigations, both, Q_{rad} and Q_{con} were determined to have a value of at least equal to or greater than 0.7. Further, the product of Q_{rad} and Q_{con}, Q_{abs}, was required to be equal to or greater than 0.6. This means that if for example Q_{rad} is somewhere close to the threshold, Q_{con} has to have a quality of approximately 0.9.

2.2.8 Interpolation of the local extremum

A local extremum in a D-dimensional spectrum is accepted as a peak if it fulfills all the contour based criteria (defined above) in one of the $D(D-1)/2$ 2D contour planes. The digital resolution of an NMR spectrum limits the accuracy with which NMR signals can be described. It can be calculated by the quotient of the spectral sweep width and the number of complex data points which have been acquired. Accordingly, the true position of a NMR signal is rather somewhere between data points than exactly on the measured data point. The exact peak position plays a crucial role in chemical shift and NOE assignment. Therefore, it is of utmost importance to estimate a peak location as accurately as possible.

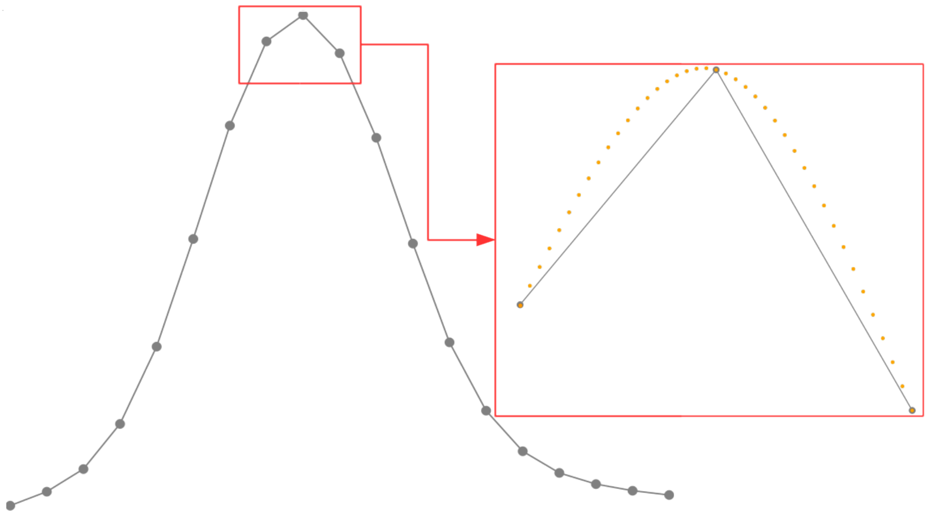


Figure 2.7: Cubic spline interpolation of the ^{15}N dimension of a well resolved peak from the ^{15}N-HSQC spectrum of ENTH. Grey circles represent original data points from the spectral file, connected by straight lines. In the right box the interpolated peak shape is depicted as it has been performed by the cubic spline interpolation implemented in CYANA (orange data points).

For the estimation of the exact position of the local extremum a cubic spline interpolation (Press *et al.*, 2007) was implemented. The interpolant is a piece-wise cubic polynomial, called a spline. Due to the piece-wise definition, splines are more flexible than polynomials and yet relatively simple and smooth. Thereby, splines do not have the disadvantages of polynomials, such as strong oscillations between data points. The cubic spline interpolation is performed along each one-dimensional slice passing though the local extremum of interest.

2.2.9 Deconvolute overlapping peaks

Local maxima which do not fulfill the requirements listed above were analyzed once more under the considerations for overlapping peaks. In order to resolve overlapping peaks peak symmetry criteria described in the AUTOPSY (Koradi *et al.*, 1998) publication were applied. In the program AUTOPSY these criteria were used for determining whether peaks are overlapping or not. This procedure was adjusted in order to resolve overlapping peaks in a simple and efficient way.

First, a symmetry center which in principal matches the coordinates of the local extremum is defined. However, for reasons of accuracy, the exact position was interpolated by a cubic spline interpolation of the local extremum data points. The new coordinates of the symmetry center are then determined from this cubic spline. The intensities of neigh-

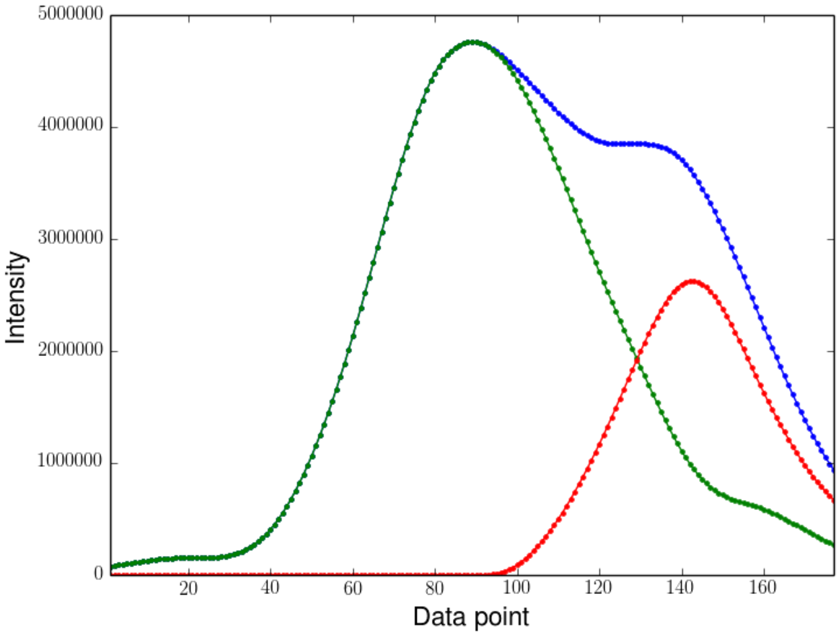


Figure 2.8: Example of resolving two overlapping peaks from a 2D ^{15}N-HSQC spectrum. The blue line represents the original data from the spectrum file; the green line represents the symmetrized local extremum of higher intensity; the red line represents the difference spectrum which is achieved by subtracting the green line from the blue line. Afterwards it is possible to analyze the green and red resolved peaks using the contour approach criteria. The symmetrization procedure is shown exemplary for a one-dimensional slice through the local extremum. The procedure is repeated along each one-dimensional slice, each row and column in case of a two-dimensional spectrum, in the defined peak area.

boring data points at mirror positions k and k', with respect to the symmetry center, are compared. Then the symmetrized intensity for point k, $I_{sym,k}$ is calculated from the minimal value of either the square root of the product of the intensity at the symmetry related position I_k and $I_{k'}$ or the original intensity at point I_k as following:

$$I_{sym,k} = \min\left(\sqrt{Int_k \cdot Int_{k'}}, Int_k\right) \tag{2.12}$$

Symmetrizing all data points with respect to the symmetry center results in the resolved peak shape (Fig. 2.8; green data points). The symmetrized data points can then be subtracted from the original data points (Fig. 2.8; blue data points), leading to the difference data points (Fig. 2.8; red data points) which ideally represent the deconvoluted second peak. It is then possible to create contour lines to the symmetrized slice as well as to the difference slice.

2.2.10 Restricted peak picking

Restricted peak picking can be performed by providing a peak lists, e.g. a 2D ^{15}N-HSQC. The peak list is read with the command **read peaks** and the spectrum of interest, e.g. a 3D ^{15}N-resolved NOESY, with the command **read spectrum**. The algorithm goes step by step though the 2D peak lists and navigates to the exact position in the accordant dimension of the 3D spectrum. This corresponds to the HSQC dimension in the 3D NOESY. The algorithm then searches for local extrema along the 3D NOE dimension at the 2D HSQC position, provided by the peak list, within a user-defined tolerance range (see chapter 2.6 for description of the exact usage). Contour lines are created for the surviving local maxima, and the accordant contour lines are analyzed by the above specified criteria. Local maxima that survive filtering and analysis are stored in a peak list.

2.3 Materials and methods

2.3.1 Evaluation dataset

The performance of CYPICK was first analyzed on the basis of 16 different spectra of the ENTH-VHS domain At3g16270-(9-135) from *Arabidopsis thaliana* (referred to as ENTH, PDB code 1VDY; BMRB code 5928) (López-Méndez & Güntert, 2006; López-Méndez *et al.*, 2006) that are summarized in appendix A in Tab. A.1. Spectra were converted to UCSF (Sparky) format, which can be read by all programs that were used to evaluate CYPICK. Manually picked peak lists and lists picked automatically by AUTOPSY were available from an earlier study (López-Méndez & Güntert, 2006). Manually picked peak lists served as a reference for calculating scores for finding real peaks, artifact peaks, and an overall score combining both. The exact definition of the scores is given in chapter 2.3.2. CYPICK user input was systematically tested and evaluated on the basis of these scores with respect to the manual peak lists. Later those scores were also used to compare CYPICK's performance to other programs, namely AUTOPSY, NMRViewJ, CCPN, and CV-Peak Picker.

We further used spectra from the Src homology domain from the human feline sarcoma oncogene FES (referred to as SH2; PDB code 1VEE; BMRB code 5929) (Scott *et al.*, 2004, 2005) and the *Arabidopsis thaliana* rhodanese domain At4g01050 (referred to as RHO; PDB code 1WQU; BMRB code 6331) (Pantoja-Uceda *et al.*, 2004, 2005), also summarized in appendix A in Tab. A.1, together with ENTH to evaluate CYPICK. Chemical shift assignments and NMR solution structures of the proteins ENTH, RHO, and SH2 have been solved earlier by conventional techniques and their data sets have previously been used to evaluated the automated assignment algorithm FLYA (Schmidt & Güntert, 2012). Existing assignments and structure bundles of these proteins were used to evaluate CYPICK on two more levels: the performance of automated established peak lists by CYPICK in chemical assignment (performed by FLYA) and in combined NOE assignment and structure calculation (performed by CYANA). The performance of CYPICK in these two steps was then compared to the performance of other algorithms (namely AUTOPSY, NMRViewJ, CCPN, and CV-Peak Picker) in these steps.

CYPICK was further tested on 10 data sets from the CASD-NMR project (Rosato *et al.*, 2012, 2009), i.e. the human NFU1 iron-sulfur cluster scaffold homolog, Northeast Structural Genomics Consortium (NESG) target HR2876B (PDB code 2LTM; BMRB code 18489), the CTD domain of the human NFU1 iron-sulfur cluster scaffold homolog, NESG target HR2876C (PDB code 2M5O; BMRB code 19068), the N-terminal domain of the human mitotic checkpoint serine/threonine-protein kinase BUB1, NESG target HR5460A (PDB code 2LAH; BMRB code 17524), the RRM domain of the human RNA-binding protein FUS, NESG target HR6430A (PDB code 2LA6; BMRB code 17504), the homeobox domain of the human homeobox protein Nkx-3.1, NESG target HR6470A (PDB code 2L9R; BMRB code 17484), the SANT domain of human DNAJC2, NESG target HR8254A (PDB code 2M2E; BMRB 18909), a *de novo* designed protein, IF3-like fold, NESG target OR135 (PDB code 2LN3; BMRB code 18145) (Koga *et al.*, 2012), a *de novo* designed protein, P-loop NTPase fold, NESG target OR36 (PDB code 2LCI; BMRB code 17613), TSTM1273 from *Salmonella typhimurium* LT2, NESG target StT322 (PDB code 2LOJ; BMRB code 18214) and the NifU-like protein *Saccharomyces cerevisiae*, NESG target YR313A (PDB code 2LTL; BMRB code 18487). For all proteins, ^{13}C-edited and ^{15}N-edited NOESY spectra were provided. The spectra were automatically picked by CYPICK using default parameters. Resulting peak lists together with reference chemical shifts, available from the specified BMRB codes, were then used in structure calculation. The performance of CYPICK was then analyzed based on the accordant PDB deposited structures, and compared to RMSD bias values achieved by manually established 'raw' and 'refined' peak lists. CYPICK peak lists scores were additionally calculated with respect to ATNOS cycle 7 peak lists that were provided to us by Prof. Dr. Torsten Herrmann[1].

2.3.2 Peak list comparison

The agreement between peak lists is determined with an implementation of the Hungarian algorithm. The Hungarian method (Munkres, 1957; Silver, 1960; Bourgeois & Lasalle, 1971) is an optimization algorithm that solves the assignment problem in combinatorial optimization (not to be confused with the problem of finding chemical shift assignments in NMR). Our implementation has a complexity of $O(n^3)$ (Edmonds & Karp, 1972). The Hungarian algorithm determines the 'cost' of assigning peak i in a *trial* peak list, including N peaks, to a peak j in a *reference* peak lists, including N_0 peaks:

$$C_{ij} = 1 - \exp^{-\min\left(\frac{d_{ij}^2, d_{cut}^2}{2}\right)}, \tag{2.13}$$

---

[1] Research Director CNRS, Insitut des Sciences Analytiques, Lyon

where

$$d_{ij}^2 = \sum_{k=1}^{D} \left(\frac{\omega_{ik} - \omega_{jk}}{\sigma_k} \right)^2 \tag{2.14}$$

is the squared scaled distance between the two peak positions $(\omega_{i1}, ..., \omega_{iD})$ and $(\omega_{j1}, ..., \omega_{jD})$ in a D-dimemsional spectrum, scaled by a chemical shift scaling factor σ_k. The cutoff d_{cut} implements the idea that all deviations larger than a certain value d_{cut} indicate that the two peaks cannot originate from the same atoms. All these peak pairs get the same high cost regardless of the actual deviation $d_{ij} \geq d_{cut}$. We used the default value $d_{cut} = 3$ for all results presented in this study. The algorithm assigns each of the $M = \min(N_0, N)$ peaks in the shorter peak list to a peak in the longer peak list, thereby minimizing the total cost $\sum_{k=1}^{M} C_{i_k j_k}$. The result is a list of $k = 1, ..., M$ pairs (i_k, j_k) of corresponding peaks in the two peak lists. The computation time can be reduced drastically be using C_{ij} from Eq. 2.13 instead of d_{ij} from Eq. 2.14, and thereby avoiding optimization of peaks that cannot belong to the same atom.

The quality of peak correspondence is rated by:

$$H = \sum_{k=1}^{M} \exp^{\left(\frac{-d_{i_k j_k}^2}{2} \right)}, \tag{2.15}$$

H represents the number of corrsponding peak pairs, weighted by the deviation of the peak positions; $0 \leq H \leq M$. H can be used to calculate a *find score* $F = H/N_0$ and an *artifact score* $A = 1 - H/N$. Both scores take values between 0 and 1 (or 0-100%). The find score gives the fraction of ‘true’ peaks in the reference list that have a corresponding peak in the trial peak list. The artifact score gives the fraction of ‘artifact’ peaks in the trial list that do not have a corresponding peak in the reference peak list. If the trial and reference peak list are identical $F = 1$ and $A = 0$. Except for the exact definition of the number H of corresponding peak pairs, F and $1-A$ are identical to ‘recall’ and ‘precision’ as defined by (Alipanahi *et al.*, 2009), respectively.

It is possible to calculate an *overall score* $S = (H - w(N - H))/N_0$, given by the number of found peaks, H, minus the number of erroneous peaks, N_H, weighted by a factor w that specifies the detrimental effect of artifacts in comparison to found peaks. In the following applications, we used $w = 0.2$, assuming that 5 additional artifact peaks are as severe as one missing true peak, as suggested by observations of their effect on automated resonance assignment (Schmidt & Güntert, 2012) and structure calculation (Buchner &

Güntert, 2015a). The overall score combines the find and artifact scores according to $S = F - w(N/N_0)A$ and reaches a maximum value of 1 in the ideal case of identical peak lists. The score can be calculated for a trial and a reference peak list by using the CYANA command **peaks compare** as explained in chapter 2.6.1.

2.3.3 Automated peak picking

CYPICK

Automated peak picking by CYPICK can be performed as explained in chapter 2.6. In order to evaluate CYPICK, input parameters were systematically tested. The tested parameters, β (Eq. 2.4) and γ (Eq. 2.5), were varied from 2.0 to 5.0 in steps of 1.0, and from 1.2 to 1.4 in steps of 0.1, respectively. β is multiplied with the noise level, L_{global} or $L_{local}(\omega_1, ..., \omega_D)$ to give the intensity of the first contour line, c_0, (Eq. 2.4). γ is the multiplication factor for generating the next contour level (Eq. 2.5). The exact input used with the CYPICK algorithm is summarized in Tab. A.2 and Tab. A.3 in appendix A. A detailed documentation of the usage of CYPICK follows in chapter 2.6.

Other well-established peak picking algorithms were applied to compare the results to CYPICK. We employed CCPN and NMRViewJ to ENTH, RHO, and SH2, where a threshold for peak picking has to be defined by the user. All local extrema above the specified threshold are picked and stored in form of a peak lists. NMRViewJ additionally comprises methods to determine local threshold which allow, e.g. exclusion of solvent lines. Automated peak picking of ENTH, RHO, and SH" by CV-Peak Picker was performed by Piotr Klukowski[2] who was responsible for the development of the algorithm. For ENTH, AUTOPSY and manually picked peak lists were available from an earlier study (López-Méndez & Güntert, 2006). In the following peak picking with CCPN, NMRViewJ, and CV-Peak Picker is explained.

The following peak lists needed to be unfolded in the carbon dimensions: ^{13}C-HSQC, ^{13}C-NOESY, HCCH-TOCSY, (H)CCH-TOCSY, HCCH-COSY, HN(CA)CO, HNCA, HN(CO)CA, HNCO, HBHACONH, CBCANH, and CBCA(CO)NH. In addition, peaks in the water region were excluded in some cases which are listed in appendix A, Tab. A.8. Peak lists from different programs were edited exactly the same way in all cases.

---

[2] Institue of Computer Science, Wroclaw University of Technology, Wroclaw

AUTOPSY

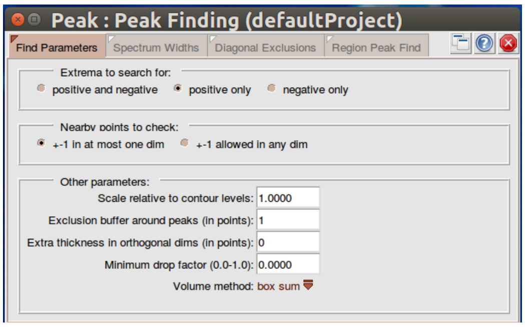
Peak picking parameters used in AUTOPSY are summarized in Tab. A.4 in appendix A.

NMRViewJ

Peak picking by NMRViewJ was performed within this work. Clicking the 'Attributes' button in the upper left corner of the spectrum window opens a new dialog which allows control over the way the spectrum is displayed. Below the icon bar, one finds the file panel. Via the file panel it is possible to control whether positive '+' and/or negative '-' contour levels should be displayed. The contour threshold is controlled via the column 'Level', where desired values can be entered. The levels that we used for peak picking are summarized in Tab. A.5 in appendix A. The 'Peak pick panel' then allows the actual peak picking. 'Local noise thresholding' methods can be activated by changing the 'Noise' value via the slider. We used 'Noise' value 10.0 as recommended on the homepage `http://www.onemoonscientific.com/`.

CCPN

Automated peak picking by CCPN (Version 2.4, Release 2) can be performed by clicking 'Peak' and then 'Peak Finding' in the main panel. This opens a window, displayed in Fig. 2.9 where parameters for peak picking can be specified.



Extrema to search for:
○ positive and negative ● positive only ○ negative only

Nearby points to check:
● +-1 in at most one dim ○ +-1 allowed in any dim

Other parameters:
Scale relative to contour levels: 1.0000
Exclusion buffer around peaks (in points): 1
Extra thickness in orthogonal dims (in points): 0
Minimum drop factor (0.0-1.0): 0.0000
Volume method: box sum

Figure 2.9: 'Peak Finding' window in CCPN. The most relevant peak picking parameters can be set in the 'Find Parameters' tab which is shown here. In the upper panel it is possible to set whether peak picking is performed on positive, negative or both local extrema. In the second panel the permissiveness of peak picking can be set. In the third panel the threshold above which peaks are picked can be set relative to the lowest contour levels. In the example presented here, scale is set to 1.0 which means that peak picking is performed on what is visible on the spectrum. It is possible to specify regions which are not to be picked around an existing picked peak via the exclusion buffer parameter. Remaining parameters were not documented. Accordingly, we set their values to zero.

Additionally to the parameters explained in Fig. 2.9 it is possible to specify minimal line width of peaks to be picked via the ‘Spectrum Widths’ Tab. Furthermore, diagonal regions in a spectrum can be excluded in a user-defined tolerance area (‘Diagonal Exclusion’). Via the ‘Region Peak Find’ panel the peak picking can be started by pushing the button ‘Find Peaks’. Besides, spectral regions for exclusion or inclusion can be specified.

We performed peak picking with CCPN by adjusting the contour level for the individual spectra as listed in Tab. A.6 in appendix A. Furthermore, the sign of the peaks to be picked was adjusted in the window shown in Fig. 2.9. All other parameters were used as depicted in Fig. 2.9.

CV-Peak Picker

Peak Picking with CV-Peak Picker was performed by Piotr Klukowski. Scanning parameters used are summarized in Tab. A.7 in appendix A.

2.3.4 Automated chemical shift assignment

Automated chemical shift assignment is performed by the FLYA algorithm (Schmidt & Güntert, 2012). In all calculations a tolerance of 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N for the chemical shift matching and comparison with a manual reference chemical shift assignment was used. The reference assignment, indicated by the BMRB codes, was exclusively used to evaluate the quality of the assignment results. The population size of the evolutionary algorithm was 50 in most cases. When only backbone spectra were used for chemical shift assignment we increased the population size to 100. When performing solely NOESY-based chemical shift assignment a population size of 200 was used following recommendation discussed in (Schmidt & Güntert, 2013a). The chemical shift assignment was consolidated from 20 independent runs. Only assignment which could be reproduced in at least 80% of the 20 runs with an accuracy that deviated from the consensus values by less than the defined tolerance was classified as ‘strong’, otherwise ‘weak’. The side-chain terminal amide groups of arginine and lysine were excluded from the assignment calculations.

2.3.5 Structure calculation

The chemical shift assignment established by FLYA or the reference chemical shift assignment was used to obtain torsional angles restraints by TALOS+ (Shen *et al.*, 2009). Combined automated NOE assignment and structure calculation was performed by the standard CYANA protocol (Güntert & Buchner, 2015), using as input the protein sequence, assigned chemical shifts, torsional angle restraints, and unassigned NOESY peak lists. Tolerances for chemical shift and peak position matching were set to 0.03 ppm for ^{1}H and 0.4 for ^{13}C and ^{15}N. NOESY peak intensities of the assigned NOESY peaks were converted into upper distance limits using the $1/r^6$ dependence. Structure calculation was performed starting from 200 conformers using 10,000 torsion angle dynamics steps. The 20 best conformers in terms of CYANA target function were selected for structure bundle representation. No energy refinement was performed on the resulting structures. The quality of structure calculation was exclusively evaluated on the basis of RMSD bias (Güntert *et al.*, 1998). Structures were first superimposed within their ordered regions which were either determined by CYRANGE (Kirchner & Güntert, 2011) (in case of ENTH, RHO, and, SH2) or applied as specified in the corresponding CASD-NMR publication (Rosato *et al.*, 2015). Then the average structure is obtained by averaging the coordinates of the atoms in the superimposed conformers in the structure bundle. The backbone RMSD between the average given structure and the reference mean structure yields the RMSD bias. The precision of calculated structure bundles is expressed by the RMSD radius (Güntert *et al.*, 1998), i.e. the average RMSD between individual conformers and the mean coordinates of the structure.

2.4 Results and discussions

In order to evaluate our picking algorithm we used the experimental data sets of ENTH, RHO, SH2, and CASD-NMR (Rosato *et al.*, 2015, 2009). As mentioned previously, the ENTH, RHO, and SH2 data set includes a complete set of NMR spectra for backbone and side-chain chemical shift assignment, as well as NOESY spectra for restraint generation. The data set also includes manually and automatically picked peak lists for ENTH. The CASD-NMR data set is composed of NOESY spectra, which can be used for restraint generation. In all cases reference PDB structures and reference resonance assignments from the BMRB are available.

Different CYPICK peak picking parameters and contour picking modes were systematically tested on the ENTH data set for which manually picked peak lists were available. Manually picked peak lists were used to measure the CYPICK picking accuracy in terms of find, artifact, and overall score. The most promising parameters were then used in automated peak picking of RHO, SH2, and CASD-NMR data sets. The resulting peak lists were further used for chemical shift assignment by FLYA and combined NOE assignment and structure calculation by CYANA. Results of these steps were compared to reference chemical shift assignments and reference NMR structure bundles.

Other well established peak picking programs were also used to compare their performance to CYPICK. We utilized AUTOPSY, NMRViewJ, CCPN, and CV-Peak Picker. Peak lists produced by other programs were also analyzed with the above mentioned score values, and utilized in chemical shift assignment and structure calculation. The results are compared to CYPICK on all three levels, i.e. picking scores, accuracy of chemical shift assignment, and accuracy of resulting structure bundles.

CASD-NMR NOESY spectra were automatically picked by CYPICK. The resulting peak lists were used in NOE assignment and structure calculation. CYPICK performance was assessed by the RMSD bias with respect to a given structure and the score values with respect to ATNOS cycle 7 peak lists that were provided to us by Prof. Dr. Torsten Herrmann.

2.4.1 Dependence of peak picking scores on the number of peaks

In Fig. 2.10 the behavior of the scores, used to evaluate peak picking, upon varying the number of real signals and artifacts is visualized for the ^{13}C-edited NOESY spectrum of the

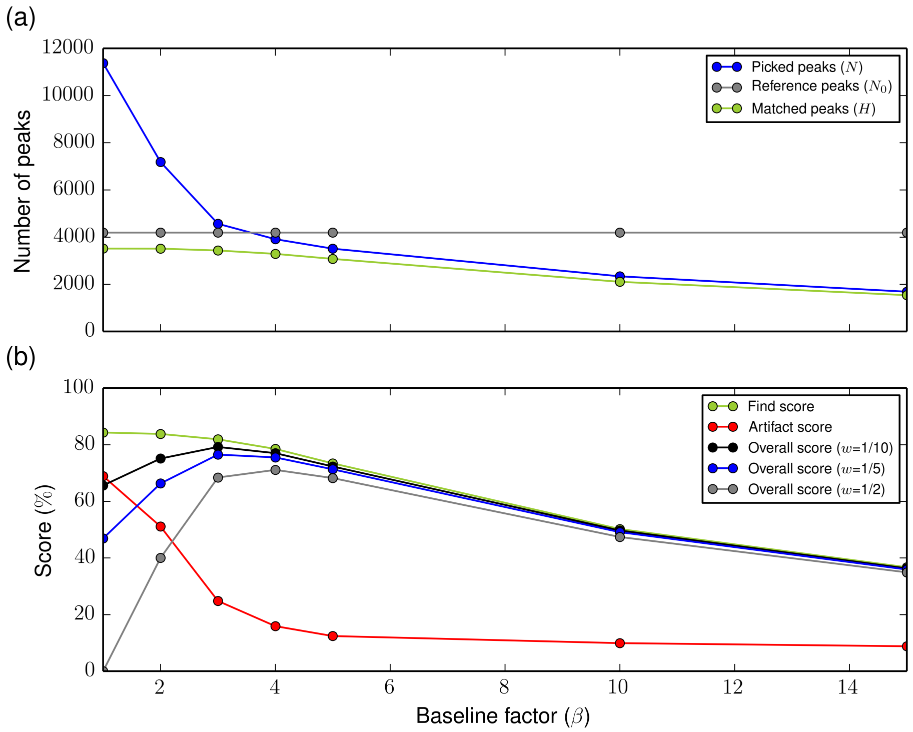


Figure 2.10: Influence of the baseline factor β on the number of peaks picked by CYPICK in the 3D ^{13}C-edited NOESY spectrum of the protein ENTH. (a) Number of peaks picked by CYPICK (N, blue), constant number of manually picked reference peaks (N_0, grey), and weighted number of matches between the two peak lists (H, green). (b) Find score (green), artifact score (red), and overall scores (black, blue, gray). The latter are shown for different values of the weighting factor w.

protein ENTH, exemplary. Peak lists with different fractions of real peaks and artifacts were produced with CYPICK by varying the baseline factor β, which determines the height of the first contour line, from 1.0 to 15.0. Decreasing the lowest contour level increases the number of picked peaks, N, significantly, and to a much lesser degree the number of real signals, H, (Fig. 2.10 (a)). At low baseline factors the number of picked peaks N clearly exceeds the number of reference peaks, N_0, being indicative of artifacts which are accumulated. Consequently, the find score $F = H/N_0$ and artifact score $A = 1 - H/N$ increase with decreasing baseline factor, β, (Fig. 2.10 (b)). Both scores approach but do not reach their ideal values of $A = 0$ and $F = 1$, i.e. some strong artifacts are always picked and a small fraction of the manually identified peaks can never be found by the algorithm. Strong artifacts can be attributed mainly to truncation artifacts of strong peaks, axial peaks, and a few putative real peaks missing in the manually prepared reference peak list.

The overall score $S = (H - w(N - H))/N_0 = F - w(N/N_0)A$ combines find and artifact score in order to evaluate peak list by a single number, considering that one strives towards maximizing the number of real peaks while minimizing the number of artifacts. The overall score includes the weighting factor, w, accounting for the fact that a missing real peak has a more detrimental effect on resonance assignment (Schmidt & Güntert, 2012) and structure calculation (Buchner & Güntert, 2015a) than an additional incorrect peak (Schmidt & Güntert, 2012). Obviously, lower weighting factors reduce the impact of artifacts in the overall score, whereas higher values increase their effect, as visualized in Fig 2.10 (b). In our following studies we used $w = 0.2$ resulting from observations discussed in (Schmidt & Güntert, 2012; Buchner & Güntert, 2015a).

2.4.2 Evaluation of CYPICK with manually picked ENTH peak lists

CYPICK was systematically tested with different user-input parameters. The contour level multiplier, γ, was varied from 1.2 to 1.4 in steps of 0.1, and the base level multiplier, β was varied from 2.0 to 5.0 in steps of 1.0. In our calculations we used the above mentioned parameters and defined the peak width in each dimension. The peak width is used as a scaling factor for the contour lines in the selected dimensions. The exact input parameters are summarized in Tab. A.2 in appendix A. We analyzed the results of our peak picking algorithm with the above explained find and artifact score (chapter 2.3.2) of our peak lists with respect to peak lists that were produced manually. The results are visualized in Fig. 2.17. The computation times for CYPICK varied between 1 s for the ^{15}N-HSQC and 31 s for the ^{13}C-resolved NOESY spectrum on a standard desktop computer.

The performance of CYPICK did not show a strong dependence on the input parameters in terms of find score especially in spectra which tend to have a higher sensitivity or resolution, e.g. ^{13}C- and ^{15}N-HSQC, HNCO, HNCA, HN(CO)CA, CBCA(CO)NH. Most of these spectra are backbone experiments. Side-chain experiments on the other hand, showed a stronger dependence on the input parameters, e.g. (H)CC(CO)NH, H(CCCO)NH. This can be attributed to a higher degree of overlap within the spectrum. The find score is generally higher in case of overlapping peaks, when γ and β are lower. This can be explained by the fact, that the lower these two parameters, the more contour lines are created within the same intensity range. Accordingly, more closed contour lines, which only include the local maximum of interest, can be created for those overlapping peaks. A disadvantage of using lower input parameters is that in general the artifact score is higher. The artifact score on the other hand shows a strong dependence on the input parameters. The lower these two parameters, the higher the score value. In most cases however the artifact score is severely reduced by using $\beta > 2.0$. In order to get the best peak picking results the number of real peaks is supposed to be maximized while simultaneously minimizing the number of artifacts. In the following the results are discussed quantitatively for the performance of CYPICK on the individual spectra.

The ^{15}N-HSCQ spectrum was picked with a 90% find score and an artifact score of approximately 20%. In this case the automated peak picking was relatively independent from the β and γ values. This can be explained by the fact that ^{15}N-HSCQ spectra are usually the most sensitive experiments with well resolved peaks and nearly no artifacts (Kwan *et al.*, 2011). Automatic peak picking of the ^{13}C-HSQC resulted in a find score of approximately 60% and an artifact score of approximately 20%. The automatic peak picking of ^{13}C-HSQC spectra is usually more demanding due to the high degree of overlap.

HNCO and HN(CO)CA spectra were picked with a find score of almost 100%, and artifact scores that varied from 80% to 20%. The HNCO is the most sensitive triple resonance experiments and one observes correlations between the NH groups and the preceding carbonyl group carbon atom. In an HN(CO)CA spectrum one observes correlations between the NH group of one residue and its preceding Cα atom. The HN(CO)CA also has a high sensitivity and almost no overlap. Accordingly, automatic procedures generally achieve results of high quality. In HNCO however, the artifact score was explicitly higher than in HN(CO)CA lists. In the case of the ENTH HNCO spectrum artifacts can mostly be attributed to peaks that showed pronounced sinc wiggles.

In general HNCA and HN(CA)CO spectra were picked with lower find scores than the HNCO and the HN(CO)CA spectra. The HNCA was picked with a find score of approximately 80% whereas the artifact score results were dependent on the used parameters and varied from 80% to 20%. In case of the HN(CA)CO, both found and artifact score showed a strong dependence on the user-input. The find scores varied from 80% to 60%, whereas the artifact scores varied from 80% to 10%. The HN(CA)CO experiment correlates the NH group of the own residues with the carbonyl group of the own and the preceding residue. The peak belonging to the carbonyl group of the preceding residue is in general much weaker and can be buried in noise. The HNCA experiment correlates the NH atom of the own residues with the Cα of the preceding and the own residue. In this case the Cα

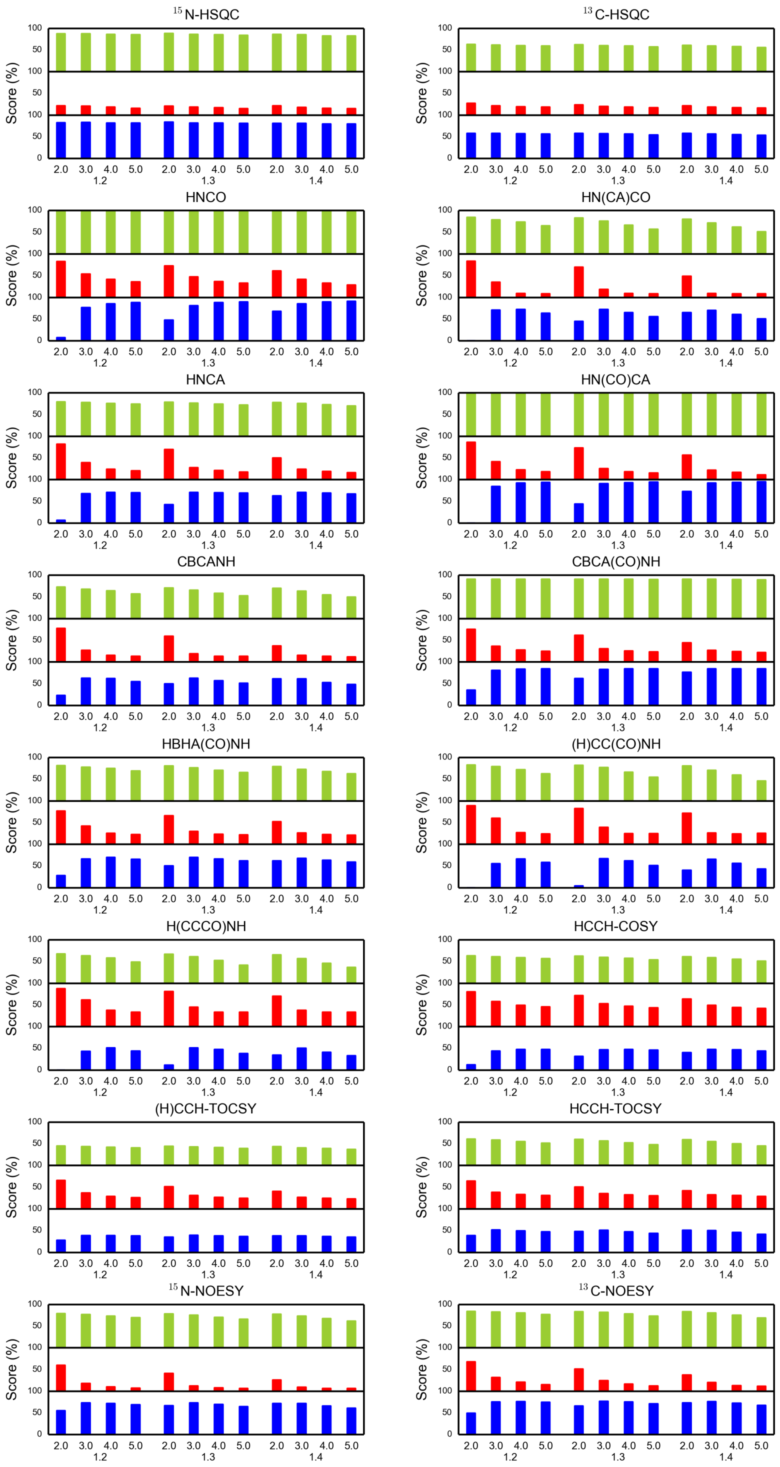


Figure 2.11: Find, artifact and overall score of automatic picked ENTH peak lists with respect to manual prepared peak lists. Peak lists were produced with the global noise estimation by CYPICK. Different user-input parameters were systematically tested for the individual spectra. The x-axis is labeled with β (upper labeling) and γ (lower labeling). β values were varied from 2.0 to 5.0 in steps of 1.0 and γ values were varied from 1.2 to 1.4 in steps of 0.1. The y-axis is labeled with the Score (%). Find scores are presented in green (upper panel within subplot), artifact scores are depicted in red (middle panel within subplot), and overall scores are presented in blue (lower panel within subplot). Subplots are labeled with the associated spectrum name.

peak belonging to the preceding residue is also much weaker in intensity than the peak belonging to the Cα of the own residue. Therefore, an automated peak picking procedure can fail to identify these weaker signals that are often buried in noise. Using a higher contour level multiplier means that the intensity range that has to be covered can exceed the peak's intensity. In case of HN(CA)CO spectra the estimated global noise level itself seemed to be simply too high for the detection of low intensity peaks. Missing peaks in the HNCA and HN(CA)CO, when compared to the manual peak list, can be attributed to peaks with very weak intensities that also displayed rather irregular peak shapes.

The CBCA(CO)NH spectrum was picked with a find score of 90%, whereas the artifact score varied from 60% to less than 30%. This experiment correlates the NH atom with the Cα and the Cβ atom of the preceding residue. In general the resolution is good, but the sensitivity in terms of signal-to-noise ratio can worsen in case of larger proteins. The CBCA(CO)NH spectrum of ENTH has a high resolution. Peaks that are not picked by CYPICK mostly showed a deviation from the regular peak shape. Peak artifacts can mainly be attributed to a couple of peaks with exceptionally high intensities that contain sinc wiggles. The CBCANH spectrum was picked with a find score of 70% to 60% and an artifact score ranging from 80% to 10%. The CBCANH experiment correlates the NH atom of one residue with the Cα and Cβ atoms of the own and the preceding residue. The signals belonging to the Cα and Cβ atoms of the own residue are in general stronger than those belonging to the preceding residue. These atoms were in many cases clearly buried in noise, leading to lower find score values. In addition some of the peaks showed irregular peak shapes. Artifacts can also mainly be attributed to sinc wiggles stemming from peaks with exceptional high intensities.

The HBHA(CO)NH was picked with a find score of approximately 80% to 70% whereas the artifact score was dependent on the applied parameters and varied between 80% to 20%. This experiment is similar to the CBCACONH, only in this case the NH atom is correlated with the Hα and Hβ of the preceding residue. In the case of the HBHA(CO)NH spectrum missing real peaks can mainly be attributed to overlapping signals. Peak artifacts mainly resulted from picking axial peaks.

The C(CO)NH peak list shows the strongest correlation of find and artifact score on the picking parameters. Find scores ranging from 80% to 50% and artifact scores between 90% to 20% were obtained. In this experiment the NH atom is correlated with all the side-chain carbon atoms of the preceding residue. Some of the side-chain carbons are either buried completely in noise or may simply not be visible, making the peak picking especially demanding. This is the case in the ENTH C(CO)NH spectrum. Many real peaks are not picked because they are deeply buried in noise. Additionally, the sensitivity of the spectrum is low and, accordingly, many noise peaks were picked. In addition, also noise ridges of high intensity were present, increasing the artifact score, as some of these signals showed regular shapes fulfilling the conditions for real peaks. The HC(CO)NH spectrum was picked with find scores ranging from 60% to 50% and artifact scores varying from 90% to 30%. This experiment correlates the NH atom of the own residues with all hydrogen atoms of the preceding residue. Missing peaks can mainly be attributed to overlapping peaks and irregular peak shapes that cannot be traced back to overlapping signals. A lot of artifacts were a result of the spectrum's low sensitivity.

In general, automatic peak picking of COSY and TOCSY spectra showed very high artifact scores in the same range as find scores. The (H)CCH-TOCSY was picked with a find score of approximately 40% and artifact scores ranging from 80% to 40%. HCCH-TOCSY and HCCH-COSY were picked with 60% find scores and artifact score varying from 80% to 50%. The HCCH-TOCSY correlates each CH shift with all the hydrogen atoms bound to all other carbons in the same residue. In HCCH-COSY the hydrogen resonances of the own and the neighboring carbons are visible. Accordingly, it is a less crowded version of the HCCH-TOCSY. TOCSY and COSY spectra usually have a high degree of overlap, making automatic peak picking challenging and leads to the omission of real peaks. Sensitivity was the main reason for having such a high amount of artifacts in the case of ENTH.

The ^{13}C-NOESY was picked with a find score of 80% and an artifact score varying from 60% to less than 20%. The results of the ^{15}N-NOESY were similar to ^{13}C-NOESY. In the NOESY experiments magnetization is transferred between nearby hydrogen atoms (< 5 Å) by the NOE. NOESY spectra are probably the most challenging cases because the intensity of the peaks depends on the inverse of the distance between the contributing atoms. As a results, the most interesting peaks for structure calculation are the peaks stemming from long-range contacts which have low intensities close to the noise level or even buried completely in noise. NOESY spectra usually contain a lot of artifacts complicating automated peak picking.

The current results led to the conclusion that using $\gamma = 1.3$ and $\beta = 3.0$ yields the best results. These parameters considerably reduced the number of artifacts, whereas the number of real peaks is not reduced significantly compared to using the lowest contour and base level multiplier which is also reflected in the overall score.

In order to improve the automated peak picking with CYPICK we utilized the local noise peak picking mode. In this picking mode a local noise level is determined for each individual data point. The local noise level intensity is then used to create contour lines. We thereby expect to reduce the number of peak artifacts, whilst not influencing the number of real peaks.

A comparison between the find and artifact score for the global and local noise picking mode is visualized in Fig. 2.12. Generally speaking, the amount of artifacts is reduced by using the local noise functionalities. However, in cases where peaks with lower intensity are present, i.e. HN(CA)CO, HNCA, CBCANH, ^{13}C-NOESY and ^{15}N-NOESY, the find score is also reduced significantly.

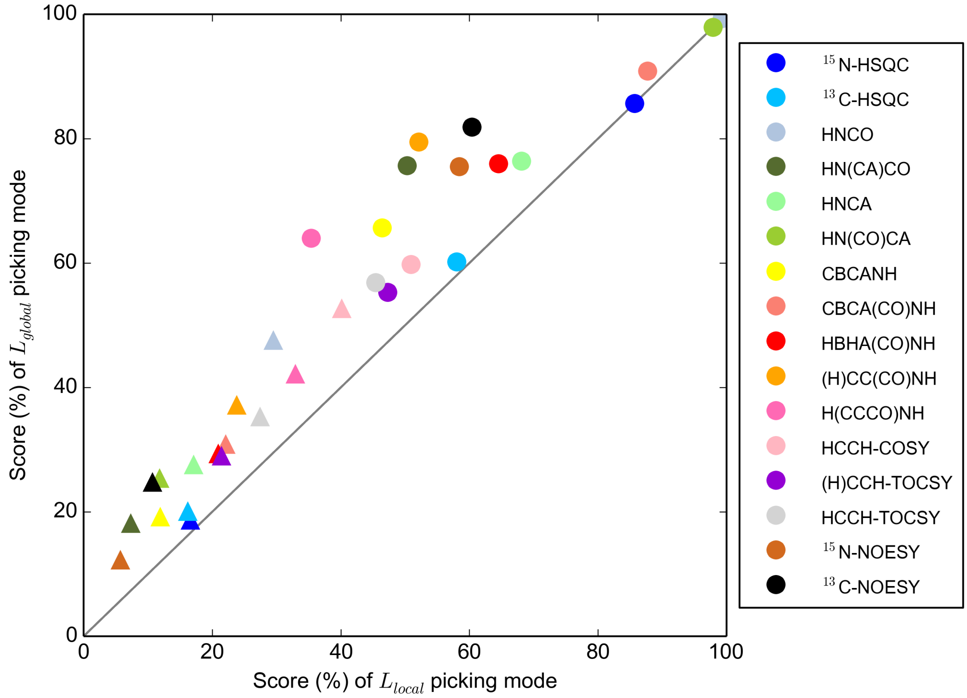


Figure 2.12: Comparison of find and artifact scores achieved by local noise and global noise contour peak picking mode with CYPICK. Find scores are denoted by circles and artifact scores by triangles. The color code is explained on the right side of the plot. Peak lists were determined with $\gamma = 1.3$ and $\beta = 3.0$ in both picking modes.

In the case of ^{15}N-HSQC- and ^{13}C-HSQC-spectra the find score is not affected by using the local noise picking mode, the artifacts however are slightly reduced. As already mentioned, the ^{15}N-HSQC is a well resolved spectrum with no overlap and only very few artifacts. The ^{13}C-HSQC is picked with a relatively low artifact score, therefore the local noise mode does not lead to a large improvement. The HNCO, HN(CO)CA, and CBCA(CO)NH spectrum are picked with a similar or exactly the same find score. However, the artifact score is reduced significantly. These spectra usually do not have peaks with low intensity that are close to the noise level, therefore no real peaks are eliminated. In all remaining spectra the find score is reduced significantly by using the local noise level mode, therefore it is likely that the usage of these peak lists is probably unfavorable to the chemical shift assignment. In general, the global noise level is estimated to be much lower than the actual noise level because the picking itself should mainly depend on contour line characteristics. In most cases, the local noise is higher than the global noise level. Accordingly, the possibility of eliminating real peaks rises when having a spectrum with low sensitivity and signals that are close to the noise level.

Suggestions on the picking mode depend on the quality of the data and the type of experiment. So far, we can record that using the local noise level estimation is advisable whenever the sensitivity of the spectrum is high. Nevertheless, before giving explicit recommendations the influence on chemical shift assignment and structure calculation should be analyzed.

In Fig. 2.13 we compared results from using the global noise picking mode alone and using the global noise picking mode combined with functionalities to resolve overlapping peaks. In general, the amount of correct peaks is enhanced when using functionalities to resolve overlapping peaks. Find and artifact score are both enhanced in most cases. Usage of the resolve overlapping function combined with the global noise peak picking mode does not influence the find score of the HNCO, HN(CO)CA, and the CBCA(CO)NH spectrum. The artifact score, however, is enhanced significantly in all cases. In most cases, using the resolve overlap functionalities is unnecessary for these spectra that characteristically show good resolution. The find score is only slightly enhanced in the remaining spectra, with HBHA(CO)NH having the highest improvement in find score from 76.0% to 80.6%. Nevertheless, the artifact score is also enhanced significantly. The increase in artifact score is lowest in the ^{15}N-HSQC spectrum, from 18.7% to 21.9%. In this case, only an improvement of about 2.0% is observed when using the resolve overlap functionalities.

In Fig. 2.14 we compared results for using only the local noise picking mode and using the local noise picking mode combined with functionalities to resolve overlapping peaks.

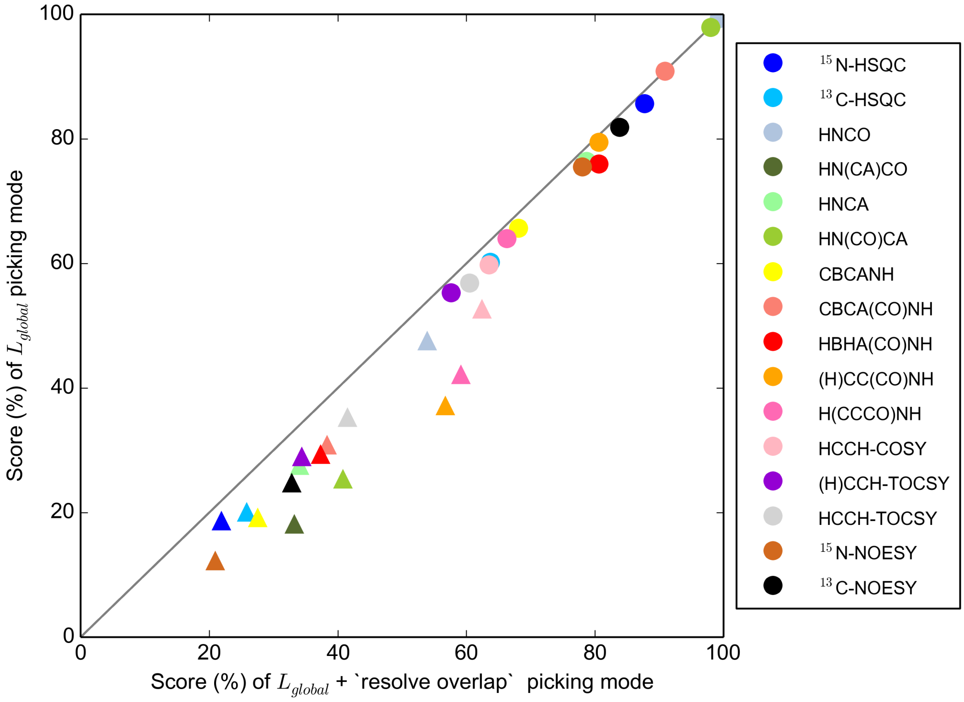


Figure 2.13: Comparison of find and artifact scores achieved by global noise and combined global noise and resolve overlap contour peak picking mode with CYPICK. Find scores are denoted by circles and artifact scores by triangles. The color code is explained on the right side of the plot. Corresponding peak lists were determined with $\gamma = 1.3$ and $\beta = 3.0$ in both picking modes.

In some cases the find score is not affected at all by the resolving overlap functionalities, i.e. HNCO, HNCA, HN(CO)CA, CBCANH, CBCA(CO)NH, and HBHA(CO)NH. In case of HNCA, HBHA(CO)NH, and HN(CO)CA the artifact score is also entirely unaffected. The remaining spectra show a slight increase in find score, but also a slight increase in artifact score, e.g. the ^{13}C-HSQC shows the strongest effect when using resolve overlap functions, i.e. the find score is enhanced from 58.0% to 63.4% and the artifact score also rises from 16.2% to 23.2%.

The results achieved by the resolve overlap functionalities did not show much improvement in terms of find score. The reason for this might be that a more advanced method is needed for identifying overlapping signals. So far no measure is available for the identification of so-called 'peak shoulders' which do not have a local maximum of their own. Accordingly, only overlapping peaks that have a local maximum are considered while others are neglected. When using the resolve overlap functionalities the amount of artifacts also increases significantly. Therefore, one should reconsider a strategy to choose potential

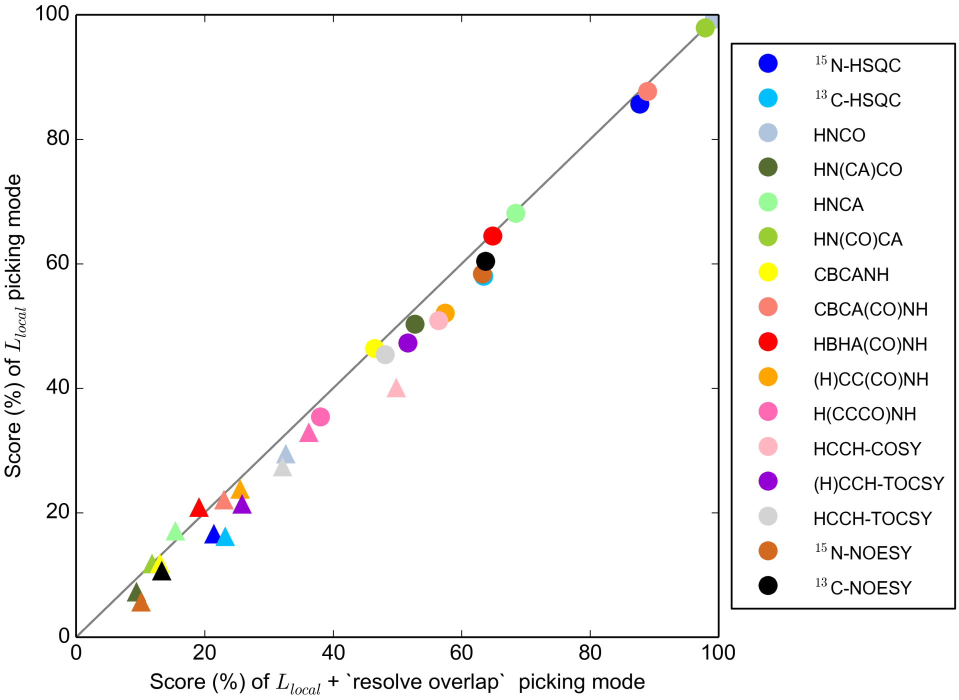


Figure 2.14: Comparison of find and artifact scores achieved by local noise and combined local noise and resolve overlap contour peak picking mode with CYPICK. Find scores are denoted by circles and artifact scores by triangles. The color code is explained on the right side of the plot. Corresponding peak lists were determined with$\gamma = 1.3$ and $\beta = 3.0$ in both picking modes.

peaks from the symmetrized spectrum.

Summarized, the best overall scores could be achieved by using $\beta = 3.0$ and $\gamma = 1.3$. In Fig. 2.12-2.14 it was presented that neither the local noise level calculation and the resolve overlap function, nor a combination of those did bring much improvement in terms of overall scores. In the next sections the influence of the automatically picked peak lists on the accuracy of chemical shift assignment and structure calculation is analyzed and discussed.

2.4.3 Comparison with reference chemical shift assignment

Results of chemical shift assignment correctness performed with CYPICK peak lists, with respect to a reference chemical shift assignment, are presented in Tabs. 2.1 and 2.2. We present results which have been achieved with the global noise and the local noise picking mode combined with functionalities to resolve overlap for the proteins ENTH, RHO, and SH2 using peak lists from all available spectra (Tab. 2.1) and solely NOESY-based chemical shift assignment (Tab. 2.2). An assignment correctness of 100% is equivalent to reproducing the reference chemical shift assignment within the defined tolerance ranges. Missing assignments within the reference assignment are not considered.

Table 2.1: Correctness of automated chemical shift assignment with respect to a reference chemical shift assignment. Assignments are categorized as being correct if the assignment is within the tolerance range of the reference assignment. Results are compared to the global noise estimation and the local noise estimation ($\beta = 3.0$ and $\gamma = 1.3$). All available peak lists were used in the calculations.

	ENTH	RHO	SH2	ENTH	RHO	SH2
	L_{global}			L_{local}		
Backbone	95.4%	96.4%	97.5%	93.0%	96.7%	97.3%
Side-chain	85.2%	86.2%	81.4%	80.4%	86.1%	82.0%
All atoms	89.4%	90.6%	87.9%	85.5%	90.7%	88.2%
	L_{global} + resolve overlap			L_{local} + resolve overlap		
Backbone	95.7%	96.5%	97.5%	94.6%	96.9%	96.9%
Side-chain	84.4%	86.7%	82.2%	81.1%	86.5%	80.7%
All atoms	89.0%	91.0%	88.4%	86.6%	91.0%	87.3%

'Backbone': N, HN, C', C_α, and C_β; 'Side-chain': all atoms except 'Backbone'; 'All atoms': 'Backbone' and 'Side-chain'

The correctness of the chemical shift assignment varies between 93.0–97.5% for the backbone atoms N, HN, C', C_α, and C_β using all available peak lists. The side-chain assignments vary between 80.7–86.2% which results in an overall correctness ranging from 85.5–90.7%. In case of ENTH, results achieved by the local noise picking mode have a lower accuracy when compared to the global noise picking results. However, in case of RHO and SH2, the local noise picking mode achieved similar results for chemical shift assignment compared to the global picking mode. In chapter 2.4.2 we showed that by using the local noise picking mode the number of artifacts was in fact reduced. However, the number of real peaks was also diminished in case of the less sensitive experiments. This leads to the conclusion, that a higher number of artifacts is not necessarily harmful for the chemical shift assignment, provided the real signals are present as complete as possible. Resolving overlapping peaks does not have a significant influence on the accuracy of the chemical shift assignment.

Table 2.2: Correctness of automated chemical shift assignment with respect to a reference chemical shift assignment. Assignments are categorized as being correct if the assignment is within the tolerance range of the reference assignment. Results are compared to the global noise estimation and the local noise estimation ($\beta = 3.0$ and $\gamma = 1.3$). Only ^{13}C-edited- and ^{15}N-edited-NOESY peak lists were used in the calculations.

	ENTH	RHO	SH2	ENTH	RHO	SH2
	L_{global}			L_{local}		
All atoms	79.3%	79.5%	77.0%	74.9%	81.6%	79.7%
	L_{global} + resolve overlap			L_{local} + resolve overlap		
All atoms	76.5%	79.2%	78.4%	77.5%	82.6%	80.0%

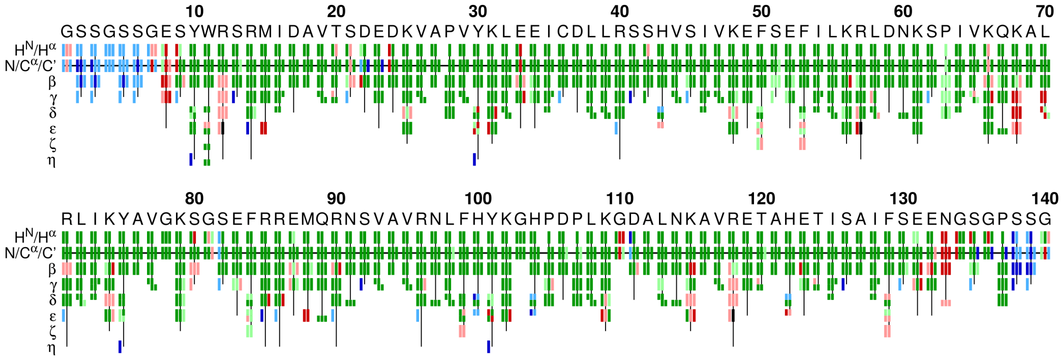
‘All atoms’: all atoms except C’

When applying NOESY-based chemical shift assignment we could achieve overall correctnesses ranging from 74.9–82.6%. In case of ENTH, the global noise picking mode scores displayed a significant higher accuracy compared to the local noise picking mode. Whereas, in case of RHO and SH2 the local noise results had a slightly higher accuracy than the global noise results. Using functionalities to resolve overlapping peaks did not lead to a significant improvement or deterioration, neither in global noise picking mode nor in local noise picking mode, in case of RHO and SH2. In peak picking of ENTH spectra the resolve overlap functionalities led to an decrease and increase of approximately 2.5–3% in resonance assignment accuracy when using the L_{global} and L_{local} picking mode, respectively. Summarized, neither the local noise picking mode nor functionalities to resolve overlap led to improved chemical shift assignment accuracies.

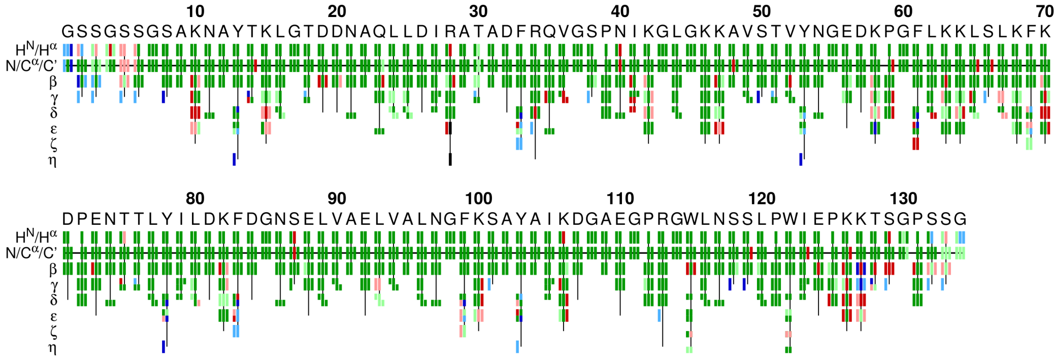
Before giving explicit recommendations on the picking mode, the influence on automated NOE assignment and structure calculation is analyzed.

Figure 2.15 displays the individual assignments of the proteins ENTH, RHO and SH2 established from the L_{global} picking mode using all available peak lists. Erroneous assignments occur especially in the side-chains of lysine, leucine, arginine, and phenylalanine. Reasons for this might be in general the quality of the automatically picked side-chain peak lists which have significantly lower overall scores ($\sim$ 60%) than backbone peak lists ($>$ 70%). The quality of the HCCH-COSY spectrum is especially critical in this case because it is the only spectrum in the data set which yields unambiguous information on side-chain assignments.

(a) ENTH



(b) RHO



(c) SH2

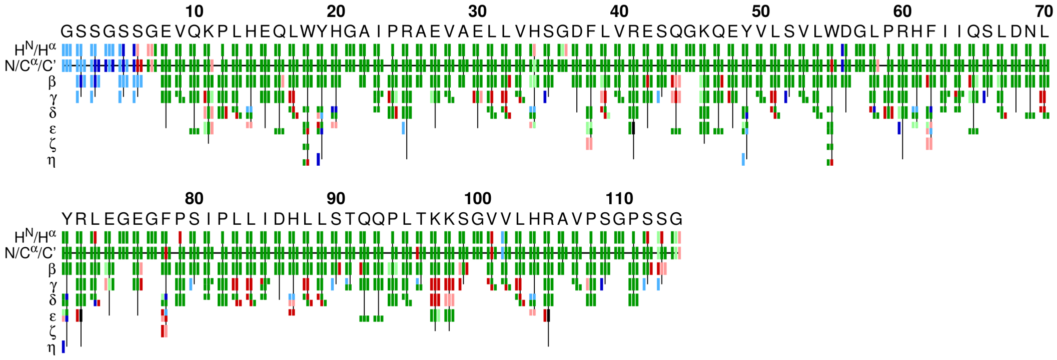


Figure 2.15: Results of chemical shift assignment starting from the complete set of automatically picked peak lists: (a) ENTH, (b) RHO, and (c) SH2. Peaklists were established by using the global noise picking mode in CYPICK($\beta = 3.0$ and $\gamma = 1.3$). The primary sequence of the protein is represented by differently colored rectangles: green, assignment agrees with reference assignment within a defined tolerance; red, assignment deviates from reference; blue, no reference assignment is available; black: only a reference assignment is available. Stronger colors reflect consolidated, safe assignments.

2.4.4 Comparison with reference combined NOE assignment and structure calculation

CYPICK NOESY peak lists and the chemical shift assignment which is retrieved from the automated picked peak lists are used in combined NOE assignment and structure calculation. A detailed overview of the results from using the global noise estimation with $\beta = 3.0$ and $\gamma = 1.3$ are given in Tab. 2.3 and 2.4. Tab. 2.3 shows results from using a chemical shift assignment that has been established on the basis of the complete set of automated picked peak lists, whereas Tab. 2.4 shows results that have been achieved solely based on NOESY spectra.

When using the complete set of automated picked peak lists it was clearly shown that structures can be recalculated with an RMSD bias close to 1.00 Å. Also, in the much more challenging case of using only NOESY peak lists it is possible to reliably recalculated the structure bundles of the three proteins with RMSD bias values < 2.00 Å in case of ENTH and SH2, and close to 2.00 Å in the case of RHO. Structure bundles recalculated on the basis of a chemical shift assignment achieved from the full data set are presented in Fig. 2.16 **(a)-(c)** and solely NOESY based results are presented in Fig. 2.16 **(d)-(f)**.

Structure calculations based on peak lists that were picked using the local noise picking mode showed different results (appendix A, Tab. A.9). The accuracy of the resulting structure bundles is reduced, i.e. average backbone and heavy atom RMSD to mean values are significantly higher and the RMSD bias values are also significantly above 1.0 Å for the backbone and even above 2.0 Å for the heavy atoms. This can be explained by the fact that the number of distance restraints is reduced. Especially medium- and long-range restraints are reduced by approximately 50% in all three calculations, indicative of peaks with lower intensity that are excluded by the local noise picking mode in NOESY peak picking. These peaks do not have a strong influence on the accuracy of the chemical shift assignment but they have an impact on the structure calculation. Therefore, we suggest to not use the local noise picking mode for the automated peak picking of NOESY spectra. In other cases it depends on the sensitivity of the spectrum and should be decided individually. Using functionalities to resolve overlap did not result in a significant improvement, therefore we suggest to not use these functionalities until they are improved. Further, automated peak picking is performed by using the global noise picking mode with $\beta = 3.0$ and $\gamma = 1.3$.

Table 2.3: Results of automated NOE assignment and structure calculation by CYANA using the automated chemical shift assignment and ^{15}N- and ^{13}C-NOESY CYPICK lists. Results are shown for peak lists that were picked by the global noise estimation mode with $\beta = 3.0$ and $\gamma = 1.3$.

	ENTH	**RHO**	**SH2**
NOE assignment[a]			
^{15}N-NOESY	1594	2349	1848
^{13}C-NOESY	4567	5265	5973
Assigned cross peaks	4438(74.1%)	4006(52.6%)	4843(61.9%)
Unassigned cross peaks	1723(25.9%)	3608(47.4%)	2978(38.1%)
Restraints			
NOE distance restraints			
short-range	1393(51.6%)	1241(51.6%)	1381(50.2%)
medium-range	702(26.0%)	426(17.8%)	393(14.3%)
long-range	603(22.3%)	725(30.3%)	976(35.5%)
Dihedral angle restraints (ϕ/ψ)	107	94	80
Structure statistics[a]			
Average CYANA target function [Å^2]	1.12±0.11	2.75±0.16	5.53±0.27
Restraint violations			
Max. distance restraint violations [Å]	0.15	0.22	0.67
Number of violated distance restraints > 0.2 Å	0	1	8
Max. dihedral angle restraint violations (°)	0.25	10.03	11.51
Number of violated dihedral angle constraints > 5 °	-	2	2
Ramachandran plot			
Residues in most favored regions	87.6%	86.3%	81.2%
Residues in additionally allowed regions	13.6%	18.4%	17.6%
Residues in generously allowed regions	0.1%	0.0%	1.2%
Residues in disallowed regions	0.0%	0.0%	0.0%
RMSD			
RMSD range[b]	9..102,113..130	6..125	8..109
Average backbone RMSD radius [Å]	0.44±0.08	0.27±0.06	0.33±0.04
Average heavy atom RMSD radius [Å]	0.88±0.08	0.63±0.05	0.72±0.06
Backbone RMSD bias [Å]	0.91	1.35	1.24
Heavy atom RMSD bias [Å]	1.66	1.79	1.63

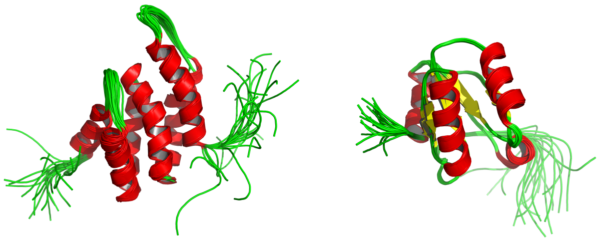
[a] using automated NOE assignment and structure calculation functionalities of CYANA. [b] determined by CYRANGE

Table 2.4: Results of automated NOE assignment and structure calculation by CYANA using automated solely NOESY-based chemical shift assignment from and ^{15}N- and ^{13}C-NOESY CYPICK lists. Results are shown for peak lists that were picked by the global noise estimation mode with $\beta = 3.0$ and $\gamma = 1.3$.

	ENTH	**RHO**	**SH2**
NOE assignment^{a}			
^{15}N-NOESY	1594	2349	1848
^{13}C-NOESY	4567	5265	5973
Assigned cross peaks	4399(71.4%)	4354(57.0%)	4540(58.0%)
Unassigned cross peaks	1762(28.6%)	3260(43.0%)	3281(42.0%)
Restraints			
NOE distance restraints			
short-range	1451(59.4%)	1308(55.4%)	1319(53.2%)
medium-range	516(21.1%)	386(16.4%)	330(13.3%)
long-range	476(19.5%)	665(28.2%)	831(33.5%)
Dihedral angle restraints (ϕ/ψ)	109	110	82
Structure statistics^{a}			
Average CYANA target function [Å^2]	8.58±0.65	16.81±0.43	8.26±0.20
Restraint violations			
Max. distance restraint violations [Å]	0.63	0.45	0.64
Number of violated distance restraints > 0.2 Å	1	6	3
Max. dihedral angle restraint violations (°)	37.86	30.59	27.85
Number of violated dihedral angle constraints > 5 °	11	25	8
Ramachandran plot			
Residues in most favored regions	83.8%	73.7%	76.4%
Residues in additionally allowed regions	16.1%	26.2%	23.0%
Residues in generously allowed regions	0.1%	0.1%	0.6%
Residues in disallowed regions	0.0%	0.0%	0.0%
RMSD			
RMSD range^{b}	9..102,113..130	6..125	8..109
Average backbone RMSD radius [Å]	0.51±0.07	0.29±0.06	0.29±0.04
Average heavy atom RMSD radius [Å]	1.02±0.08	0.75±0.07	0.59±0.06
Backbone RMSD bias [Å]	1.43	2.11	1.56
Heavy atom RMSD bias [Å]	2.10	2.55	2.33

^{a} using automated NOE assignment and structure calculation functionalities of CYANA. ^{b} determined by CYRANGE

(a) ENTH full **(b)** RHO full **(c)** SH2 full



(d) ENTH NOESY **(e)** RHO NOESY **(f)** SH2 NOESY

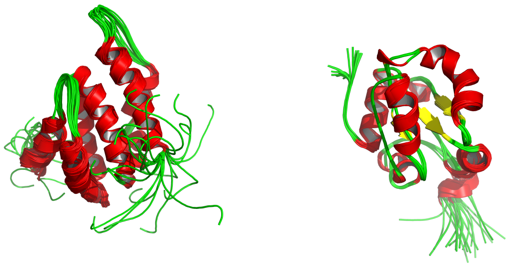


Figure 2.16: ENTH, RHO, and SH2 structure bundles calculated from CYPICK peak lists using chemical shift assignments achieved from all available peak lists (**(a)**-**(c)**) and only on the basis of NOESY peak lists (**(d)**-**(f)**). α-helical regions are presented in red, β-sheets are shown in yellow, and random coil regions in green.

2.4.5 Comparison with other automated peak picking procedures

In order to evaluate the performance of CYPICK, we compared the results from using the global noise contour picking mode ($\beta = 3.0$ and $\gamma = 1.3$) to other well-established automated or semi-automated peak picking algorithms. In case of ENTH we analyzed the accuracy of the peak picking by calculating score values with respect to manual established peak lists, and compared the scores from different programs to CYPICK afterwards. For all three protein data sets the performance in automated chemical shift assignment and structure calculation is evaluated.

Table 2.5: Comparison of peak picking performance in-between the different programs. Mean and standard deviation values of the find, artifact, and overall scores (expressed in %) are calculated for the individual sets of peak lists.

	CYPICK	AUTOPSY	NMRViewJ	CCPN	CV-Picker
All					
Find score	75 ± 14	72 ± 15	76 ± 12	74 ± 14	72 ± 14
Artifact score	29 ± 11	49 ± 15	44 ± 9	49 ± 17	35 ± 19
Overall score	68 ± 14	56 ± 18	63 ± 11	52 ± 20	58 ± 22
2D					
Find score	73 ± 18	65 ± 1	76 ± 16	79 ±13	71±18
Artifact score	1 9± 1	33 ± 8	36 ± 4	54 ± 37	14 ± 5
Overall score	69 ± 17	58 ± 1	67 ± 13	48 ± 46	69 ± 18
Backbone					
Find score	84 ± 14	86 ± 10	87 ± 10	84 ± 13	85 ± 15
Artifact score	27 ± 12	40 ± 12	46 ± 8	49 ± 18	26 ± 9
Overall score	77 ± 10	73 ± 9	71 ± 8	63 ± 11	76 ± 12
Side-chain					
Find score	65 ± 9	61 ± 13	66 ± 8	66 ± 6	60 ± 7
Artifact score	38 ± 9	62 ± 9	45 ± 12	50 ± 11	46 ± 15
Overall score	56 ± 10	39 ± 14	55 ± 11	50 ± 12	47 ± 11
NOESY					
Find score	79 ± 4	72 ± 5	74 ± 10	67 ± 20	73 ± 1
Artifact score	19 ± 9	51 ± 2	41 ± 4	41 ± 29	53 ± 32
Overall score	75 ± 2	56 ± 2	64 ± 7	52 ± 4	46 ± 27

‘ALL’ includes all available peak lists; ‘2D’ only includes the 2D HSQC peak lists; ‘Backbone’ includes CBCANH, CBCA(CO)NH, HNCA, HN(CO)CA, HNCO, HN(CA)CO lists; ‘Side-chain’ includes HBHA(CO)NH, (H)CC(CO)NH, H(CCCO)NH, HCCH-COSY, (H)CCH-TOCSY, HCCH-TOCSY; ‘NOESY’ includes ^{13}C-edited- and ^{15}N-edited-NOESY.

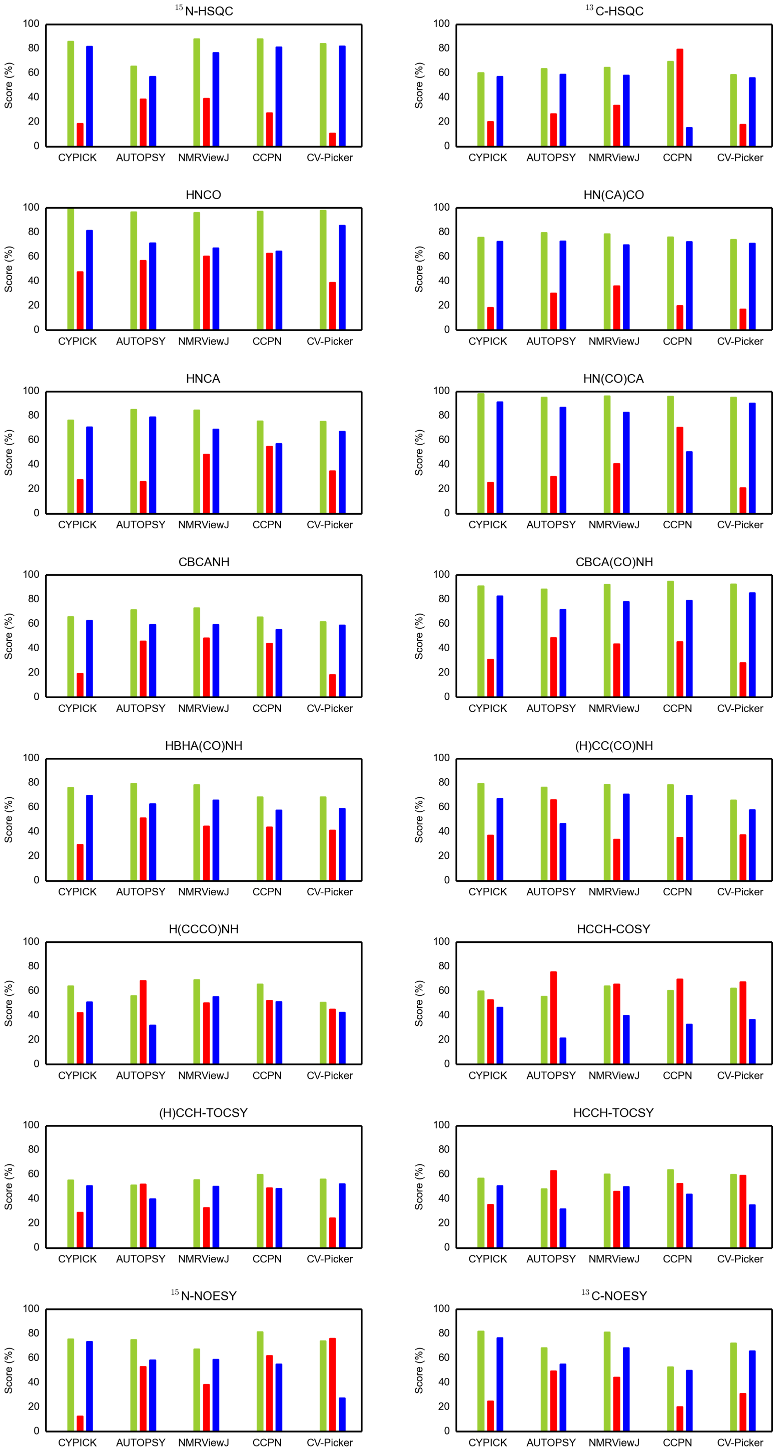


Figure 2.17: Find, artifact and overall scores of automatic picked ENTH peak lists with respect to manual established lists. CYPICK peak lists were picked with the global noise level mode and the default parameters. AUTOPSY, NMRViewJ, CCPN, and CV-Picker results have been achieved as explained in chapter 2.3. The x-axis is labeled with the employed program and the y-axis is labeled with the accordant Score (%). Find scores are presented in green, artifact scores in red, and overall scores are depicted in blue. Subplots are labeled with the associated spectrum name.

A comparison between the average peak-picking performance in-between CYPICK, AUTOPSY, NMRViewJ, CCPN, and CV-Peak Picker on the ENTH spectra is presented in Tab. 2.5 and Fig. 2.17. In Tab. 2.5 average score values and their standard deviations are presented for the respective specified peak lists. Over all spectra, the average find scores for the different algorithms were similar, ranging from 72–76%, whereas the average artifact scores displayed a higher degree of variation, from 29–49% (Tab. 2.5). Also the average overall scores vary appreciably from 55–68% for the different algorithms. Within this data set CYPICK obtained the highest over all score, the second highest find score, and by far the lowest artifact score.

Considering only the two 2D HSQC spectra, CYPICK produced peak lists with an acceptable find score and one of the lowest artifact scores, together with CV-Peak Picker, both share similar overall scores. CCPN peak lists achieved the highest find and artifact scores, indicative of an underestimation of the noise threshold. Consequently, their overall score was lowest. The standard deviations within the '2D' group was relatively high on account of the significant differences in resolution and overlap of the ^{15}N- and ^{13}C-HSQC spectra. Automatic peak picking on the ^{15}N-HSQC spectrum is in general performed much more accurate with find scores around 90% (Fig. 2.17). This can be explained by the fact that the ^{15}N-HSQC is among the most sensitive experiments with well resolved peaks and very few artifacts. Automatic peak picking of ^{13}C-HSQC spectra, on the other hand, is much more demanding due to the high degree of overlap usually being present.

Automatic peak picking of the triple resonance 'backbone' spectra for backbone assignment yielded uniformly high average find scores of 83–87% (Tab. 2.5). Within this group CYPICK and CV-Peak Picker achieved the lowest average artifact score of 28% and 26%, respectively, compared to 40–49% for the other algorithms. Due to their higher sensitivity and better resolution, backbone assignment spectra are in general more straightforward to pick than side-chain experiments. HNCO and HN(CO)CA spectra were picked with find scores close to 100% by CYPICK (Fig. 2.17), which reflects the high sensitivity and resolution of these spectra. In HNCA, HN(CA)CO and CBCANH, CYPICK missed some weak peaks that are buried in noise and show irregular peak shapes. Artifacts within these lists from CYPICK can be attributed mainly to sinc wiggles.

'Side-chain' spectra peak picking was performed with average find scores of 60–66% and relatively high average artifact scores of 38-62%, which in case of AUTOPSY did even exceed the find score (Tab. 2.5). The highest average find score was achieved by the NMRViewJ, CCPN, and CYPICK peak lists. CYPICK peak lists showed the lowest average artifact score among the programs, resulting again in the highest overall score of 56%, closely followed by NMRViewJ, whereas the other algorithms have overall scores that are 5–17% lower. TOCSY- and COSY-type side-chain assignment spectra usually exhibit a high degree of overlap, which makes automatic peak picking challenging and leads to the omission of many real signals by CYPICK because these peaks show deviations from the expected peak shape.

Automatic peak picking of the 3D NOESY spectra of ENTH was best performed by CYPICK which produced the highest mean find score of 79% vs. 67–74% for the other programs, as well as the lowest average artifact score (19% vs. 41–53%), leading to a significantly higher average overall score (75% vs. 46–64%).

Summarized, when comparing automatic peak picking by CYPICK to the other programs, the higher robustness manifested by consistently highest overall scores is mainly due to the fact that CYPICK picks considerably fewer artifacts than other methods (Tab. 2.5, Fig. 2.17). The find scores are more uniform; those from CYPICK are usually among the highest. CYPICK performs particularly well in the automated peak picking of NOESY spectra, which is promising for NOE distance restraint-based structure calculation and for the solely NOESY-based chemical shift assignment procedure in FLYA (Schmidt & Güntert, 2013a). The individual scores for each spectrum and program in Fig. 2.17 show a stable performance of CYPICK without outliers for individual spectra.

Automatically established peak lists of the proteins ENTH, RHO, and SH2 were used as input for automated chemical shift assignment with FLYA, followed by combined NOE assignment and structure calculation with CYANA.

Tab. 2.6 summarizes the assignment and structure calculation results obtained using all available peak lists as input for FLYA. Despite the above described variations in the peak picking scores, the overall correctness of the chemical shift assignments by FLYA was relatively uniform over the different peak picking methods that were used to prepare the input peak list: 88–90% for ENTH, 87–91% for RHO, and 87–88% for SH2. CYPICK peak lists yielded a chemical shift assignment result that does not deviate by more than 1% from the best assignment.

For ENTH, the assignment correctness was best for AUTOPSY and CYPICK, and about 4% lower for CCPN, which is in line with the CCPN peak lists showing the lowest overall score (Tab. 2.5). On the other hand, the fact that AUTOPSY yielded the most correct assignment could not have been discerned from the peak picking score values. The correctness of the resonance assignment was reflected in the structural statistics.

The backbone RMSD to the reference was 0.90 Å for the structures obtained using the CYPICK peak lists, and 0.99 Å for AUTOPSY, whereas NMRViewJ, CV-Peak Picker, and CCPN yielded RMSD bias values well above 1 Å. RMSD radius values were all significantly below 1 Å.

In case of RHO, NMRViewJ, CYPICK and CV-Peak Picker achieved a similar overall chemical shift correctness of 89–91%, whereas CCPN yielded 87%. The resulting structures were closest to the reference for CYPICK with a backbone RMSD to the reference structure of 1.35 Å, followed by NMRViewJ and CV-Peak Picker with RMSD bias below 1.75 Å. In case of CCPN, however, the structure calculation converged to an incorrect structure bundle. This can be explained by a lack of structural information that could be deduced from the NOESY peak lists. Automated NOE assignment based on the CYPICK peak lists led to 2392 distance restraints, of which 725 were long-range. In comparison to that, automated NOE assignment with CCPN peak lists resulted in a significantly lower number of 1192 distance restraints, of which only 214 were long-range.

For SH2, the chemical shift assignment accuracy was essentially the same with the peak lists from all programs, showing only 1.2% variation. The structural accuracy was also very similar. RMSD bias values below 1 Å were achieved with CYPICK and NMRViewJ peak lists, whereas CV-Peak Picker and CCPN yielded RMSD bias values slightly above 1.0 Å.

We also obtained resonance assignments by automated chemical shift assignment with FLYA using as input exclusively the 3D NOESY spectra. This approach is generally challenging for FLYA and requires good input NOESY peak lists (Ikeya *et al.*, 2011; Schmidt & Güntert, 2013a). Using the NOESY peak lists from CYPICK 77–80% correct assignments could be achieved for the three proteins ENTH, RHO, and SH2 (Tab. 2.7). The peak lists from the other programs yielded in general fewer correct assignments, except for CCPN in the case SH2, where 80% correct assignments were achieved, as compared to 77% for CYPICK. This is reflected also in the accuracy in the structures obtained by automated NOESY assignment based on the FLYA chemical shifts. CYPICK yielded backbone RMSDs to the reference structure of 1.4–2.1 Å, i.e. for all three proteins essentially a correct structure, whereas most of the structures obtained for ENTH and RHO using the peak lists from the other programs were incorrect with RMSD bias values of 2.4–10.3 Å (Tab. 2.7). Only for the smaller SH2 protein the peak lists from all programs were sufficient to yield a structure with 1.5–2.2 Å backbone RMSD to the reference. For ENTH, these results can be compared with the NOESY peak list scores of Tab. 2.5. The best overall scores for the NOESY peak lists were achieved with CYPICK (75%), followed by NMRViewJ (64%), and the other programs (46–56%). The different quality of these NOESY peak lists is clearly reflected in Tab. 2.7: CYPICK peak lists yielded the highest assignment correctness (79%) and lowest RMSD bias (1.4 Å), followed by NMRViewJ (75%/2.4 Å), and the other programs (66–74%/3.6–10.3 Å).

Table 2.6: Percentage of correct assignments with respect to a reference chemical shift assignment, RMSD radius RMSD bias. All available lists where used in the calculations.

	CYPICK	AUTOPSY	NMRViewJ	CCPN	CV-Picker
ENTH					
Backbone [%]	95.4	96.0	94.9	92.7	94.9
Side-chain [%]	85.2	85.3	83.5	80.7	82.8
All atoms [%]	89.4	89.7	88.2	85.5	87.7
RMSD radius [Å]	0.48	0.33	0.41	0.77	0.74
RMSD bias [Å]	0.91	0.99	1.20	1.78	1.66
RHO					
Backbone [%]	96.4	-	95.0	92.6	95.3
Side-chain [%]	86.2	-	88.5	85.5	84.3
All atoms [%]	90.6	-	91.3	87.4	89.1
RMSD radius [Å]	0.27	-	0.35	1.49	0.37
RMSD bias [Å]	1.35	-	1.61	6.41	1.74
SH2					
Backbone [%]	96.1	-	91.6	97.1	97.1
Side-chain [%]	81.4	-	83.4	81.4	81.6
All atoms [%]	87.3	-	86.7	87.7	87.9
RMSD radius [Å]	0.21	-	0.22	0.22	0.31
RMSD bias [Å]	0.98	-	0.91	1.23	1.07

'Backbone', 'Side-chain' and 'All atoms' refers to the chemical shift assignment correctness with respect to a manual chemical shift assignment. 'Backbone' includes the atoms N, HN, C', Cα, and Cβ; 'Side-chain' includes all atoms except 'Backbone' atoms, 'All atoms' includes all atoms. RMSD radius is the average backbone RMSD of the 20 individual conformers to their mean coordinates. RMSD bias is the backbone RMSD between the mean coordinates of the structure bundle and the reference structure. Residue ranges for RMSDs calculation, determined with CYRANGE (Kirchner & Güntert, 2011): 9–102 and 113–130 of ENTH, 6–125 of RHO, and 8–109 for SH2

Table 2.7: Percentage of correct assignments with respect to a reference chemical shift assignment, RMSD radius RMSD bias. Only ^{13}C-edited- and ^{15}N-edited-NOESY peak lists were used in the calculations.

	CYPICK	AUTOPSY	NMRViewJ	CCPN	CV-Picker
ENTH					
All atoms [%]	79.3	71.8	75.4	66.0	73.6
RMSD radius [Å]	0.51	0.61	0.44	4.39	2.98
RMSD bias [Å]	1.43	3.58	2.40	10.31	4.86
RHO					
All atoms [%]	79.5	-	76.1	72.2	78.3
RMSD radius [Å]	0.29	-	0.55	4.60	0.43
RMSD bias [Å]	2.11	-	4.46	8.95	3.49
SH2					
All atoms [%]	77.0	-	70.8	80.3	79.0
RMSD radius [Å]	0.29	-	0.31	0.38	0.41
RMSD bias [Å]	1.56	-	1.73	1.50	2.20

'All atoms' refers to the chemical shift assignment correctness with respect to a manual chemical shift assignment and includes all atoms except C'. For RMSD calculation details see Tab. 2.6.

2.4.6 Structure calculation of CASD-NMR proteins using NOESY peak lists from CYPICK

Critical Assessment of automated Structure Determination of proteins by NMR (CASD-NMR) is a project for the blind testing of routine, fully automated determination of protein structures from NMR data (Rosato *et al.*, 2012, 2009). From the most recent round of CASD-NMR, NMR data sets are available for ten proteins (Rosato *et al.*, 2015), comprising NOESY spectra, NOESY peak lists, manually determined reference chemical shift assignments, and reference structures. Automatic peak picking by CYPICK was performed on the CASD-NMR data set using default parameters ($\beta = 3.0$ and $\gamma = 1.3$). Peak lists were then used in combined NOE assignment and structure calculation by CYANA, using as input the protein sequence, the reference chemical shift assignment, the unassigned NOESY peak lists, torsional angles restraints derived from the reference assignment. The accordant results are summarized in Tab. 2.8. Structure bundles achieved on the basis of CYPICK peak lists are presented in Fig. 2.18.

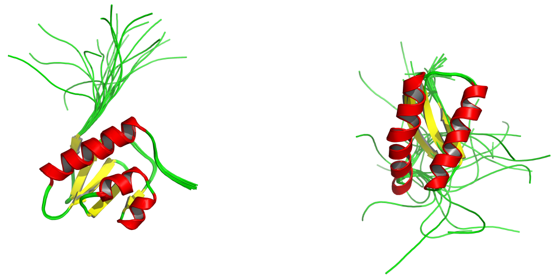
Automatic peak picking of CASD-NMR NOESY spectra led to overall scores ranging from 52–84% with respect to the ATNOS cycle 7 peak lists (Guerry *et al.*, 2015). In most cases, these scores were lower than those observed above for the NOESY peak lists of the protein ENTH (Tab. 2.5 and Fig. 2.17). One reason for this are the significantly higher artifact scores of the CYPICK peak lists. These were computed with respect to final ATNOS peak lists, which were filtered based on the known chemical shift assignment and the 3D structure. Nonetheless, CYPICK and ATNOS peak lists share a considerable set of the same peaks, expressed in high find scores ranging from 70–93%. It is also possible that CYPICK identifies true peaks that ATNOS peak lists lack. Available 'refined' or 'raw' peak lists were not used as a reference in score computation, due to their idealization with respect to the HSQC position and the existence of too many peaks that do not possess a local extremum in the accordant spectrum. In most cases, scores for the ^{15}N-resolved NOESY peak list are better than for the ^{13}C-resolved NOESY, which is complicated by a high degree of signal overlap.

Nevertheless, in five out of ten cases, i.e. HR2876B, HR2876C, HR6430A, HR6470A, and OR135, structure calculation with automatic picked NOESY peak lists by CYPICK was successful, yielding structures with a backbone RMSD to the reference structure of 0.6–1.1 Å (Tab. 2.8). Also for the other proteins correctly folded structures were found, albeit with slightly higher RMSD biases of 2.0–3.2 Å. For comparison, the RMSD bias of the structures obtained by the same approach but based on refined manual peak lists was 0.4–1.6 Å (Tab. 2.8). In addition to the manually refined final peak lists, the CASD-NMR data sets include also uncurated, 'raw' peak lists from earlier stages of the original structure determination. These 'raw' peak lists yielded structures with RMSD bias values of 1.0–7.4 Å (Tab. 2.8). In general, the peak lists from CYPICK thus yielded structures with an accuracy between those obtained from the manually curated and uncurated peak lists provided by CASD-NMR.

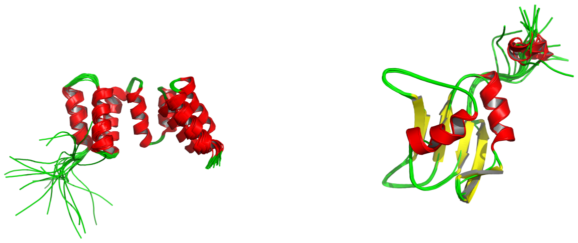
Table 2.8: Peak picking and structure calculation results from CASD-NMR proteins

Protein	Residues	Scores(%) vs ATNOS cycle 7 peak lists (^{13}C-/^{15}N-NOESY)			Backbone RMSD to reference (Å)		
		Find	Artifact	Overall	CYPICK	raw	refined
HR2876B	107	76.7/91.1	46.8/35.0	63.2/81.3	0.89	0.95	0.79
HR2876C	97	85.5/93.4	52.0/68.8	67.0/52.5	0.98	0.88	0.71
HR5460A	160	77.5/88.8	54.2/48.1	59.2/72.4	2.95	3.38	1.38
HR6430A	99	72.3/86.9	47.3/35.5	59.4/77.4	1.07	1.15	0.92
HR6470A	69	70.3/82.5	49.7/42.8	56.4/70.1	0.60	0.61	0.37
HR8254A	73				1.95	7.43	0.77
OR135	83	82.3/87.3	46.8/58.7	67.8/62.5	0.95	1.13	0.89
OR36	134	88.0/87.1	58.7/35.7	63.0/77.5	3.02	1.03	0.98
StT322	63				2.08	6.73	1.49
YR313A	119	73.9/89.7	43.8/24.4	62.4/83.9	3.22	1.64	1.59

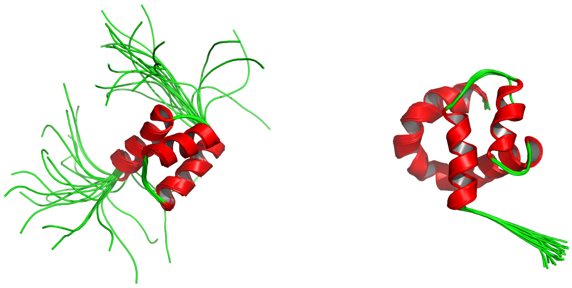
Residues ranges for RMSD calculation (Rosato *et al.*, 2015): 13–105 for HR2876B, 17–91 for HR2876C, 14–25 and 33–158 for HR5460A, 14-99 for HR6430A, 15–56 for HR6470A, 554–608 for HR8254A, 4–74 for OR136, 2–46 and 53–125 for OR36, 23–63 for StT322, and 17–41 and 45–115 for YR313A. ATNOS peak lists are not available for HR8254A and StT322 (Guerry *et al.*, 2015).



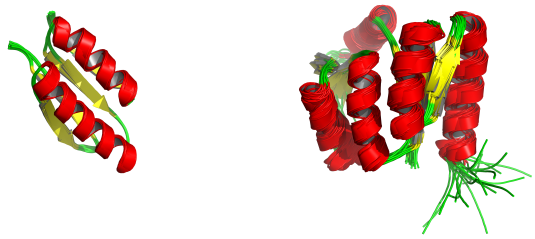
(a) HR2876B **(b)** HR2876C



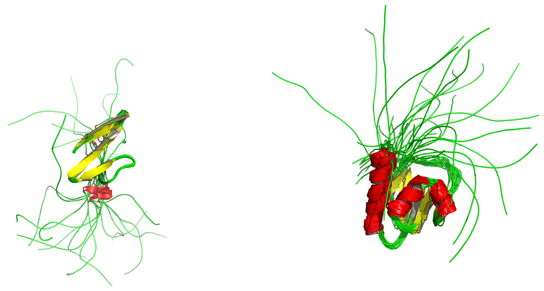
(c) HR5460A **(d)** HR6430A



(e) HR6470A **(f)** HR8254A



(g) OR135 **(h)** OR36



(i) StT322 **(j)** YR313A

Figure 2.18: Structure bundles calculated from CYPICK peak lists. α-helical regions are presented in red, β-sheets are shown in yellow, and random coil regions in green.

2.5 Conclusions

In this chapter, CYPICK, an automated peak picking procedure implemented in CYANA that analyzes geometric criteria of contour line plots was introduced. The CYPICK approach does not use any information about the underlying structural system, .e.g. chemical shift assignment, or experimental information, e.g. symmetry considerations which makes it universally employable to any kind of experimental NMR spectrum. The required user-input is reduced as far as possible, making the approach very objective.

Results presented in chapter 2.4 clearly show that CYPICK peak lists lead to resonance assignments and structure bundles of high accuracy. The results are superior when compared to other programs. Even in the challenging case of solely NOESY-based chemical shift assignment the results achieved by CYPICK stand out for their robustness throughout the complete data set. When compared to other programs, CYPICK achieves a good balance between picking real signals and rejecting artifacts, and the resulting peak lists are sufficiently good to determine the resonance assignments and 3D structures of proteins by a fully automatic approach.

Nevertheless, evaluation studies performed on various data sets revealed so far that certain functionalities of CYPICK can be substantially improved or implemented in future projects to make peak picking even more reliable:

1. Peak picking by CYPICK requires a local extremum condition to be fulfilled. Signals that do not present a local extremum, such as ‘shoulders’ located on the slope of a stronger, overlapping peak, are currently discarded and not further analyzed. Relaxing the requirement for a local extremum can improve the completeness of peak list for crowded spectra, such as ^{13}C-HSQC, side-chain HCCH-TOCSY and NOESY.

2. Very weak signals not possessing enough contour lines are currently discarded. Refining the criteria on the regularity of peak contours may enable identifying very weak but “well-shaped” signals without unduly increasing the number of artifacts.

3. Many of the picked artifacts originate from small regions of the spectrum, typically narrow noise bands. Their number may be reduced significantly by a better recognition and exclusion of these regions.

4. Peak picking by CYPICK does not take into account other information than the local features of the spectrum at and near the location of interest. It has been shown that especially the number of artifact peaks can be reduced by considering

self-consistency with a spectrum or between spectra (Hiller *et al.*, 2005), or by guiding peak picking by external information, such as a known 3D structure (Herrmann *et al.*, 2002b).

5. In situations of strong overlap CYPICK picks significantly fewer real signals than can be identified by visual inspection. Improvements may be achieved by the implementation of more advanced deconvolution methods for overlapping peaks.
6. Contour line based quality factors can in principle be used in automated chemical shift assignment and NOE assignment.

In conclusion, we developed a stable and versatile automated peak picking method that is fully integrated into the CYANA software package for automated resonance assignment, NOESY assignment, and structure calculation.

2.6 Implementation in CYANA

The CYANA software package, written in Fortran90, can be accessed and controlled via the scripting language INCLAN. Individual CYANA commands can be combined into CYANA macros, recognizable by the suffix **.cya**. Literal CYANA input is written in **bold** and other CYANA input is written in *italics*. In order to allow peak picking by CYPICK several functionalities, that can be accessed through the commands summarized in chapter 2.6.1, had to be implemented into CYANA. Macros for performing peak picking by CYPICK have to be individually adjusted to the user's need. Nevertheless, a few examples are given in chapter 2.6.2 (written in `typewriterfont`).

2.6.1 CYANA commands

read spectrum *filename*

- *filename*=string (required)

  Name of the spectrum file. In case of **xeasy** two files will be read; a XEASY parameter file called *filename***.3D.param** and the data file *filename***.3D.16**. In case of **azara** also two files will be read; an azara parameter file called *filename***.spc.par** and the data file *filename***.spc**. In case of **bruker** format a parameter file for each dimension is read. That parameter file has to have the name proc for the first dimension, proc2s for the second dimension, proc3s for the third dimension and so on. The data file usually has no suffix. In case of **ucsf** all the information are stored in a single data file called *filename***.ucsf**.

- **type**=string (required)

  type specifies the format of the spectral data file and can be **azara**, **bruker**, **ucsf**, or **xeasy**.

- **format**=string (required)

  format refers to the experimental definition of the spectrum within the CYANA library file, e.g. a ^{13}C-resolved NOESY spectrum is defined as: *format="C13NOESY H HC C"*. The order of the atoms has to be the same as they are stored in the parameter file of the spectrum or the header of the intensity file.

  Example: A three-dimensional ^{13}C-resolved NOESY spectrum can be read by: *read spectrum C13NOESY type=xeasy format="C13NOESY H HC C"*

write spectrum *filename*

This command converst a spectrum in memory to another spectral format.

- *filename*=string (required)
- **format**=string (required)

  format can either be **ucsf** or **topspin**

  Example: A three dimensional ^{13}C-resolved NOESY spectrum can be written to **ucsf** format by: *write spectrum C13NOESY format=ucsf*

spec noise

Calculates the global noise level L_{global} of the spectrum as explained in chapter 2.2.1.

spec pick local

This command determines the local noise spectrum $L_{local}(\omega_i, ..., \omega_D)$ at each data point of the specified spectrum as explained in chapter 2.2.2.

- **specfile**=string (required) **specfile** equals the name of the spectrum in memory as specified by *filename* in `read spectrum` with the suffix '.spectrum'.

  Example: *speck pick local specfile=C13NOESY.spectrum*

spec pick contour

- **specfile**=string (required) Explanation see `spec pick local` command.
- **method**=string (optional, default **diag**) **method** specifies the condition for local extremum determination as explained in chapter 2.2.3. Provided modes are **diag** (all neighbors of a data point are considered) and **nodiag** (only direct neighbors of a data point are considered).
- **scale**=float list (required) **scale** is used to specify the scaling factor of the spectral axes as explained in chapter 2.2.4. The individual scaling factors have to be provided in the same order as in the **format** specification above. Individual values are separated by ",", e.g. **scale=0.03,0.03,0,4**.
- **contourdim**=string list (required)

  contourdim defines the 2D plane to be analyzed. Contour lines can only be created in 2D. Dimensions for which contour lines should be created have to be provided by the user following CYANA library nomenclature, e.g. **contourdim=HC,C**.

- **peakarea**=float list (optional, default 1.0 for each dimension)

  peakarea defines the size of the box in each dimension around a local maximum within which contour lines are created and analyzed. The size of **peakarea** can be provided by the user. Smaller values lead to a faster scanning of the spectrum. For peak picking it is vital to choose a sufficiently large **peakarea** otherwise it does not influence the outcome of the algorithm. Individual sizes, separated by ",", have to provided in the same order as specified in **format**, e.g. **peakarea=0.06,0.06,0.8**.

- **basefactor**=real (optional, default 3.0)

  basefactor is a multiplication for the L_{global} to give the intensity of the first contour line, as explained in chapter 2.2.3.

- **contourfactor**=real (optional, default=1.3)

  contourfactor is a multiplication factor for determining the intensity of the next higher contour line by multiplying with the intensity of current contour line.

- **userglobal**=real (optional)

  Instead of calculating the global noise level L_{global} via `spec noise` the user can set L_{global} explicitly.

- **range**=float list (optional, default minimal and maximal chemical shifts)

  range specifies the spectral range in ppm that is to be picked. Ranges belonging different dimensions can be separated by ",". The order of the ranges specified has to correspond to the order of the dimensions given in **format**, e.g. **range=1.0..10.0, 2.0..11.0,0.0..75.0**

- **only_pos**,**only_neg** (option) If the option **only_pos** is given only positive signals are considered, else if the option **only_neg** is given only negative signals are considered. Is none of these two options set, both negative and positive signals are included in analysis.

- **half_pixel_cal** (option)

  There are two ways of interpreting the maximal chemical shift of the spectrum, usually given in ppm: (i) interpreting the chemical shift as ppm of the edge of the spectrum, i.e. the left side of the left pixel, or (ii) the center of the first pixel. If the option **half_pixel_cal** is set, the maximal chemical shift is construed as explained in (ii).

- **sol**=string list (optional)

  sol specifies dimension that has a solvent signal which should be excluded. The axis name has to follow the nomenclature of the **format** specification, e.g. **sol=HC,H**. However, the order is irrelevant.

- **waterline**=float list (optional, default 4.7)

  waterline determines the region of the solvent signal to be excluded in ppm. Different positions can be indicated for varying dimensions but have to follow the order given in **sol**, e.g. **waterline=4.7,4.6**.

- **watertol**=float list (optional, default 0.04)

  watertol represents the tolerance range for solvent region to be excluded in ppm, i.e. **waterline**± **watertol**. Different tolerances can be indicated for varying dimensions but have to follow the order given in **sol**, e.g. **waterline=0.5,0.3**.

- **include_diagonal** (option)

  If the option **include_diagonal** is set diagonal peaks of corresponding dimensions are picked.

spec pick global

This command allows automated peak picking above a user specified global cutoff. If the below listed user input is not specified in detail it is used as explained in **spec pick contour**.

- **specfile**=string (required)
- **method**=string (optional, default **diag**)
- **only_pos**,**only_neg** (option)
- **half_pixel_cal** (option)
- **sol**=string list (optional)
- **waterline**=float list (optional, default 4.7)
- **watertol**=float list (optional, default 0.04)
- **include_diagonal** (option)

- **userglobal**=real (required)

  Intensity threshold for peak picking. Only signals that have an intensity above the specified **userglobal** values are considered.

spec pick filter

This command allows peak picking by a frequency filter which is provided in the form of a peak list. If the below listed user input is not specified in detail it is used as explained in **spec pick contour**.

- **specfile**=string (required)
- **method**=string (optional, default **diag**)
- **scale**=float list (required)
- **contourdim**=string list (required)
- **peakarea**=float list (optional, default 1.0 for each dimension)
- **basefactor**=real (optional, default 3.0)
- **contourfactor**=real (optional, default=1.3)
- **userglobal**=real (optional)
- **range**=string (optional, default minimal and maximal chemical shifts)
- **only_pos**,**only_neg** (option)
- **half_pixel_cal** (option)
- **sol**=string list (optional)
- **waterline**=float list (optional, default 4.7)
- **watertol**=float list (optional, default 0.04)
- **include_diagonal** (option)
- **perm**=string (required)

  perm specifies the axis labels of the frequency filter, i.e. in case of using a ^{15}N-HSQC peak lists as frequency filter one has to specify **perm=H,HN**.

- **piktol**=string (optional, default 0.03 for ^{1}H and 0.4 for ^{13}C and ^{15}N)

  Via **piktol** the tolerance range (in ppm) for accepting peaks that are close to the position of the frequency filter can be defined, e.g. **piktol=0.03,0.03,0.4**. The order of the tolerance ranges has to follow the specification in **format**.

peaks compare

This command can be used to calculate scores between a given peak list and a reference peak list as explained in chapter 2.3.2.

- **selection**=string (optional)

  selection follows the syntax of the command **peaks select** as explained in detail on the homepage `http://www.cyana.org/wiki/index.php/Peak_selection`.

- **width**=real (optional, default 1.0)

  width is an additional scaling factor for the chemical shift scaling factor σ_k as explained in chapter 2.3.2.

- **distcut**=real (optional, default 3.0)

  distcut represents the cutoff d_{cut} for matching to peaks as explained in chapter 2.3.2.

- **artifactweight**=real (optional, default 0.2)

  artifactweight represents the weighting factor w for artifact peaks in the overall score S as explained in chapter 2.3.2.

- **info=full** (optional)

  If the option **info=full** is set, complete information on the individual peaks which could be matched and which could not be matched is given. The first column refers to the peak index of the first peak list, the second to the peak index of the reference peak list that could be matched to the corresponding peak (if no reference peak could be matched the value '-1' is set), the third column represents the distance between the two peaks (in case of not finding a match, the distance is set to '1000.00', otherwise a value between 0.0 and **distcut**), and the fourth column represents the match between the two peaks, which is '0.00' in case no match is found, otherwise between 0.00 and 1.0.

2.6.2 CYANA macros

- **PCOM.cya**

  The CYANA macro **PCOM.cya** can be used to calculate a find, artifact and overall score of a trial peak list with respect to a reference peak list. In the example below we used an **artifactweight** value of 0.5 and **info=full** to get detailed information on the individual peaks.

  ```
  tolerance:=0.03,0.03,0.4
  read peaks trial.peaks
  read peaks reference.peaks append
  peaks compare artifactweight=0.5 info=full
  ```

- **PICK_contour.cya**

  The CYANA macro **PICK_contour.cya** can be used to pick peaks from a ^{13}-resolved NOESY spectrum of the ENTH data set.

  ```
  read seq enth
  read spectrum C13NOESY type=xeasy format="C13NOESY HC H C"
  spec noise
  spec pick contour specfile=C13NOESY.spectrum contourdim=HC,C scale=0.05,0.05,0.1
  write peaks C13NOESY format="C13NOESY HC H C"
  ```

- **PICK_filter.cya**

  The CYANA macro **PICK_filter.cya** can be used to pick peaks from a ^{13}C-resolved NOESY spectrum of the ENTH data set using a 2D frequency filter in the form a ^{13}C-HSQC peak list.

  ```
  read seq enth
  read peaks C13HSQC
  read spectrum C13NOESY type=xeasy format="C13NOESY HC H C"
  spec noise
  spec pick contour specfile=C13NOESY.spectrum contourdim=HC,C scale=0.05,0.05,0.1
  write peaks C13NOESY format="C13NOESY HC H C"
  ```

- **PICK_global.cya**

  The CYANA macro **PICK_global.cya** can be used to pick only positive peaks from a ^{13}C-resolved NOESY spectrum of the ENTH data set using an intensity threshold value.

```
read seq enth

read spectrum C13NOESY type=xeasy format="C13NOESY HC H C"

spec noise

spec pick global specfile=C13NOESY.spectrum userglobal=4000.0 only_pos

write peaks C13NOESY format="C13NOESY HC H C"
```

Chapter 3

Information content of NMR distance restraints

3.1 Introduction

Determining the three-dimensional (3D) protein structure is often the first step in the structural and biophysical characterization of a protein. The calculated structures usually serve as basis for further investigations, e.g. structure-based drug design (see chapter 4) or homology modeling. Therefore, it is essential to validate the resulting structure model. This task can in principle be achieved by analysis of the agreement between the experimental data and the resulting structure model (accuracy), by examining the uncertainty in the coordinates of the structure model (precision), or by checking physical and chemical properties of the structure (quality) (Doreleijers *et al.*, 1998). In this chapter, a new method for determining the information content of NMR distance restraint data sets is presented (referred to as I). The information content correlates with the structure's precision and is comparable to the resolution in X-ray crystallography.

In X-ray crystallography the above mentioned accuracy, precision and quality of structure models are determined on the basis of R-factors, B-factors and resolution. In the process of model creation and refinement, one can compare experimental structure factors F_{obs} and back calculated structure factors F_{calc} (from the structure model). The difference between F_{obs} and F_{calc} is minimized in the process of refinement and defined as the R-factor (sometimes called R_{work}). Accordingly, the value of the R-factor reflects the agreement between the structure model and the experimental data (Morris *et al.*, 1992), hence the accuracy of the structure model. Disorders in the protein crystal or the size of the protein can lead to uncertainty in the position of the atoms. Resolution in X-ray crystallography quantifies the resolvability of the electron density map, i.e. resolution limits the precision of the resulting structure model. B-factors (sometimes called temperature factors), measure the uncertainty of each atom.

In structure determination by NMR spectroscopy one does not have a direct measure of structural accuracy, precision, and quality from the experimental data and the resulting structures as in X-ray crystallography. However, several methods which address the aspect of validating NMR structures with the above introduced X-ray measures have been developed, i.e. methods that determined the accuracy of structures, e.g. R-FAC (Gronwald *et al.*, 2002) or DP-Score (Huang *et al.*, 2005); and methods that quantify the structural information of NMR restraints, e.g. NOE completeness (Doreleijers *et al.*, 1999) and QUEEN

(Nabuurs *et al.*, 2003, 2005). The NOE completeness and QUEEN were inspirational for the development of the information content in this thesis.

The NOE completeness is defined as the ratio of the number of matched observed NOEs and the number of expected NOEs. The completeness in itself is a more informative quantity than the number of NOEs alone. However, it does not respect data redundancy in a meaningful way.

QUEEN (Nabuurs *et al.*, 2003, 2005) is a tool for the quantification of NMR distance restraint integrated in the structure validation suite CING (Doreleijers *et al.*, 2012). Basic concepts of QUEEN rely on Shannon's information theory (Shannon, 1948). According to this theory, the uncertainty (H) of a variable (x) with a probability density function $p(x)$ is calculated as:

$$H(x) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx. \tag{3.1}$$

In case of a biomolecular structure the information of a distance restraint (H_{ij}) is presented as the uncertainty of the system minus the uncertainty of the system after adding the restraint. The actual distance of a restraint, D_{ij}, belonging to a pair of atoms i and j can be found between the experimental upper, u_{ij}, and lower, l_{ij}, bounds. The probability of D_{ij} to be found between u_{ij} and l_{ij} is uniformly distributed over the range $[u_{ij}, l_{ij}]$. The uncertainty of a distance restraint can then be calculated as:

$$H_{ij} = -\int_{l_{ij}}^{u_{ij}} \left(\frac{1}{u_{ij} - l_{ij}}\right) \log \left(\frac{1}{u_{ij} - l_{ij}}\right) dD_{ij} = \log(u_{ij} - l_{ij}) \quad . \tag{3.2}$$

Based on Eq. 3.2 it is possible to calculate the uncertainty of one atom by calculating and averaging the uncertainties of the mentioned atom with every other atom in the structure. In order to calculate the overall uncertainty of the structure in the absence of distance restraints, $H_{structure|0}$, the uncertainties of the individual atoms are averaged. The uncertainty of a distance restraint set, R, can be calculated if the uncertainty of the system after adding the restraints, $H_{structure|R}$, is given. According to this, the structural information of R can be computed by:

$$I_{total} = H_{structure|0} - H_{structure|R}. \tag{3.3}$$

Despite the availability of methods for the quantification of the structural information included in distance restraint data sets, it is still common practice to give an overview of the data that was used in the calculation, as shown in Tab. 3.1. Drawbacks of this type of data presentation are, e.g. the lack of quantitative conclusions on the amount of structural information included.

Table 3.1: Overview of restraints used for structure calculation of Proteorhodopsin as published in (Reckel *et al.*, 2011).

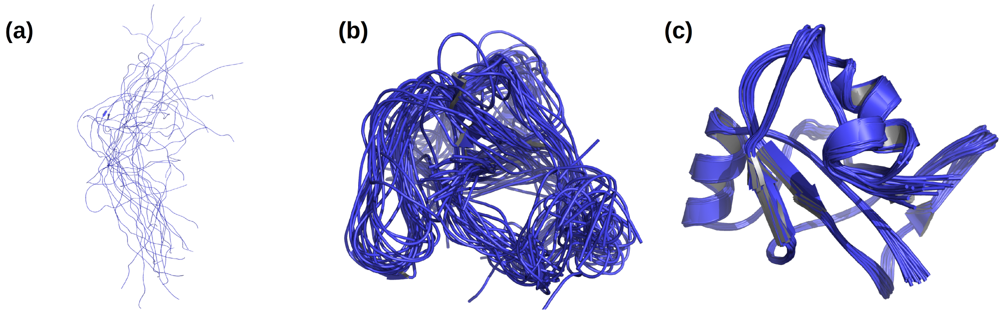
NOE distance restraints	
Sequential ($1 \leq \|i-j\| \leq 2$)	239
Medium-range ($2 < \|i-j\| \leq 4$)	50
Long-range ($\|i-j\| > 4$)	87
Hydrogen Bonds	133
Dihedral angle restraints	
Φ	196
Ψ	196
PRE distance restraints	
Upper limits	290
Lower limits	760
RDC restrains	81
Restraints from biochemical experiments	4

i and j refer to the indices of the residues in which the contributing atoms can be find.

The motivation for developing a new approach can be made clear by the example of a set of structure calculations which were all performed from 1000 distance restraints as shown in Fig. 3.1. In all cases 3D structures of the same protein are shown which have been calculated from a different input data set. The most important characteristic observed is that despite using the same number of distance restraints, the resultant structures display a strong variation in terms of precision. This leads to the conclusion that the number of restraints alone is a poor indicator of structural information included in the underlying data set. Reasons for this are:

1. The information of a data set depends on the type of restraints included.
2. The information depends on the uniqueness of data included.

Generally, distance restraints can be categorized into four groups determined by the distance between indices of the residues in which atoms i and j can be find in the primary sequence: intraresidual-range ($|i-j| = 0$), sequential-range ($|i-j| = 1$), medium-range ($2 \leq |i-j| \leq 4$) and long-range ($|i-j| < 5$). This classification is motivated by the kind of information that is provided to the structure. Short-range restraints (intraresidual- and sequential-restraints) can merely define local features of the conformation, medium-range distance restraints define helical secondary structures, and the all-important long-range distance restraints determine the tertiary structure. Obviously, long-range distance restraints carry more structural information than medium-range restraints, which in turn are more important than short-range distance restraints.



RMSD radius: 28.86 Å	RMSD radius: 6.32 Å	RMSD radius: 1.15 Å
I = 54.16 (I_r = 0.63)	I = 146.35 (I_r = 1.70)	I = 431.71 (I_r = 5.02)
1000 short-range restraints	1000 long-range restraints	1000 long-range restraints

Figure 3.1: Visualization of three structure bundles of the protein 2JQN. Each structure was calculated from a different set of distance restraints by CYANA. For all three calculations the RMSD radius (Å), the information content (denoted by I), the information content per number of residues in ordered region (I_r, explained below), and the type and number of restraints are listed. **(a)** 1000 short-range restraints were used to calculate the structure. The structure itself is completely elongated and displays no 3D globular fold. **(b)** 1000 long-range restraints were used to calculate the structure. The structure is globular but the information contained in the data is not sufficient to calculate a precise structure, i.e. the RMSD radius is still above 6.00 Å. The calculated information content however is much higher than in example **(a)**. **(c)** 1000 long-rang restraints were used to calculate the structure. In this case, the structure is perfectly folded and the information content is much higher than in examples **(a)** and **(b)**. The backbone RMSD radius is close to 1.00 Å.

Fig. 3.1 **(a)** depicts a structure that was calculated from 1000 short-range restraints. The structure is sprawled, no global fold is recognizable and the structure calculation did not converge to the same solution. The main reason for this is that only local structural information was included in the restraint data set. In comparison, the structure bundles shown in Fig. 3.1 **(b)** and **(c)** were calculated from 1000 long-range restraints. Despite differences in the details, both structures show a similar globular fold and converge to a similar solution. Fig. 3.1 **(c)** depicts a perfectly folded ensemble of structures with a low RMSD radius (1.15 Å). Fig. 3.1 **(b)** also displays a globular folded structure but uncertainty remains in many regions underlined by the backbone RMSD radius of 6.32 Å. The fact that structure 3.1 **(c)** is inferior to 3.1 **(b)** even though it has been calculated from 1000 long-range distance restraints can be explained by data redundancy, i.e. multiple restraints for the same or very similar distances and regions of the protein. For clarification: all 1000 restraints used for calculating structure 3.1 **(c)** are unique with respect to the distance they restrain. Whereas, the 1000 restraints used for calculating structure 3.1 **(b)** show a very high degree of data redundancy. Accordingly, when defining a measure that quantifies the structural information of a set of distance restraints the effect of data redundancy and the different structural input of different restraint types should be taken into consideration.

Our approach focuses currently only on the information included in NMR distance restraints. This can be rationalized by that fact that distance restraint and torsion angle restraints remain the main source of structural information (Guerry & Herrmann, 2011), despite the diversity of existing experimental data. The reason for this is the predominantly globular structure of proteins, which leads to a high proton density and in turn yields a dense network of potential NOE distance restraints. The efficiency of the NOE strongly depends on the inverse of the distance between the interacting protons. Accordingly, the observation of an NOE denotes directly two spatially close protons ($<$ 5-6 Å). If the two interacting protons are distant with respect to the primary sequence, the observation of a NOE leads to a significant limitation in terms of available conformational space.

The information content (I) has been implemented in the CYANA software package and is also available as a stand-alone software bundle CYINFO. The ideas of I are mostly based on probability theory. We define the information content of NMR restraint data sets by the negative logarithm of the probability to fulfill the restraints by random structures, considering how much each restraint restricts the conformational space of the structure and how redundant it is with other restraints. The theoretical considerations exposed above are incorporated in the definition of I and explained in section 3.2.

3.2 Information content algorithm

The information content, I, quantifies the structural information included in a distance restraint data set and it is usually monitored against an ensemble of random structures.

I was designed in a way that the information of a set of distance restraints against a perfectly folded structure bundle is zero or at least close to zero. Thus, I reflects the structural information ratio between a random structure and a folded structure. It is therefore possible to link I to the precision of the resulting structure bundle. Hence, the higher I, the lower the spread of the atomic coordinates of the resulting structure ensemble.

The information content of a set of distance restraints is defined as the negative logarithm of the probability that a set of restraints will be fulfilled by an ensemble of random structures:

$$\text{Information content} = -\log P\left(\text{restraints fulfilled by random structure}\right). \tag{3.4}$$

Here, we consider a set of n distance restraints $A = \{A_1, ..., A_n\}$. A distance restraint A_i between atoms a_i and b_i restraints the distance $d_i = d(a_i, b_i)$ with an upper limit u_i and a lower limit l_i: $l_i < d_i < u_i$. The aforementioned probability can then be defined as the conditional probability:

$$P\left(\text{data set}|\text{random structure}\right) = P\left(A|0\right) = \prod_{i=1}^{n} P\left(A_i|0\right)^{\frac{1}{R_i}}. \tag{3.5}$$

$P\left(A_i|0\right)$ is the conditional probability that an individual restraint A_i is fulfilled by a random structure, denoted by "0". The redundancy of restraints is taken into account by the exponent $\frac{1}{R_i}$: if two restraints, A_i and A_j, are redundant, i.e. if they restrain the conformational space in the same or a very similar way, then only one of the two terms $P\left(A_i|0\right)$ and $P\left(A_j|0\right)$ should be fully included in the above product. The information content $I\left(A|0\right)$ of a restraint set A in the context of a random structure "0" is given by:

$$I\left(A|0\right) = -\log P\left(A|0\right) = -\sum_{i=1}^{n} \frac{\log P\left(A_i|0\right)}{R_i}. \tag{3.6}$$

$P\left(A_i|0\right)$ is positive unless the restraint is inconsistent, i.e the lower bound has a higher value than the upper bound. R_i is the sum of the individual redundancies R_{ij} of restraint A_i with all other restraints A_j in the data set. The individual redundancy R_{ij}, defined below, is a quantity that should equal zero for a restraint A_i which restricts the conformation in a unique way that is not enforced by restraint A_j, and one for a restraint whose restriction is completely enforced by A_j. By definition, $R_{ii} = 1$, and hence $R_i \geq 1$. Intermediate values reflect the situation that A_j restricts the conformation in a similar way as A_i.

For a distance restraint A_i with an upper bound u_i and a lower bound l_i, the probability to be fulfilled by a random structure, represented by a bundle of conformers, can be computed as:

$$P\left(A_i|0\right) = \frac{u_i^3 - l_i^3}{U_i^3 - L_i^3}, \tag{3.7}$$

truncated to the range $[0, 1]$. U_i represents the maximum of all corresponding distances in the bundle of random structures and L_i is calculated from the sum of the repulsive core radii. This expression reflects the ratio between two spheres with radii $u_i - l_i$ and $U_i - L_i$, respectively, visualized in figure 3.2. The smaller the radii difference of these two spheres the higher the probability of the corresponding restraint to be fulfilled by a random structure.

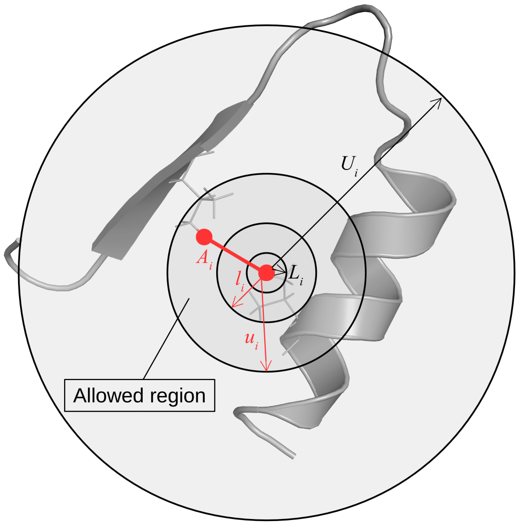


Figure 3.2: Visualization of the allowed region of a structure (depicted in blue) restricted by a distance restraints A_i (depicted in red) with an upper u_i and a lower limit l_i. The minimal distance, L_i, and the maximal distance, U_i, of the two atoms restrained by A_i is also visualized.

The redundancy R_{ij} of two restraints A_i and A_j is only calculated if these restraints share a common set of torsion angles. Let F_i and F_j, respectively, be the sets of torsion angles on which A_i and A_j depend. R_{ij} is zero if the two restraints do not overlap, i.e. if $F_i \cap F_j = \emptyset$. Otherwise, the individual redundancy can be defined as the conditional probability that restraint A_i will be fulfilled by a random structure that also fulfills restraint A_j:

$$R_{ij} = P\left(A_i|A_j,0\right) = \frac{u_i^3 - l_i^3}{\left(u_j + \Delta d_{ij}\right)^3 - \max\left(l_j - \Delta d_{ij}, 0\right)^3}, \tag{3.8}$$

truncated to the range $[0,1]$, where Δd_{ij} is defined as:

$$\Delta d_{ij} = \min\left(U\left(a_i,a_j\right) + U\left(b_i,b_j\right), U\left(a_i,b_j\right) + U\left(b_i,a_j\right)\right), \tag{3.9}$$

where U reflects the maximal distance between the corresponding atom pairs in the ensemble of randoms structures. Δd_{ij} reflects the maximal difference between the distances d_i and d_j, i.e. it is a measure of distance similarity. The redundancy R_{ij} yields values significantly above zero only if Δd_{ij} is reasonably small, i.e. if both atoms of restraint A_i are close to those of restraint A_j. For two perfectly correlated distances Δd_{ij} will be zero, and hence $R_{ij} = 1$. Although the calculation of R_{ij} is based on estimating the upper and lower bounds that restraint A_j imposes on the distances restrained by A_i, it can be calculated in a meaningful way using random structures that do not have to fulfill restraints A_i or A_j. Rather, the random structures are used to estimate the separation of the atoms belonging to restraints A_i and A_j.

The above mentioned redundancy calculation is not suitable if restraint A_i comprises only a lower bound l_i, but no upper bound u_i. In this case, we compute R_{ij} by considering that a lower bound $d_i \geq l_i$ is equivalent to an upper bound on the reciprocal distance, $\frac{1}{d_i} \geq \frac{1}{l_i}$, i.e.

$$R_{ij} = \frac{\frac{1}{l_i}^3 - \frac{1}{u_i}^3}{\frac{1}{\max(l_i - \Delta d_{ij},0)^3} - \frac{1}{(u_j + \Delta d_{ij})^3}} = \frac{\max\left(l_j - \Delta d_{ij}, 0\right)^3}{l_i^3}, \tag{3.10}$$

in the absence of upper bounds ($u_i = u_j = \infty$). If there are multiple restraints for the same distance, the following logic is applied: If a restraint file contains an upper, as well as a lower limit for the same distance, both limits are taken into consideration. The lower limit will later be compared to the theoretical, calculated lower limit based on van der Waals radii. If there are multiple upper limit values for the same distance, only the most restrictive value is further considered and stored. In case of multiple lower limit values it is also the most restrictive value that is further taken into consideration.

So far, the calculation of I has been defined on the basis of a random structure ensemble (in the following referred to as I_s). This structure ensemble is only used to have an estimate of U. Alternatively, U can be defined by expressions that depend only on the covalent structure of the molecule. This makes the redundancy and the probability calculation independent of the bundle of random structures and exactly reproducible. The maximal distance U_i between two atoms a_i and b_i can be estimated by a function which only depends on the number of torsion angles separating the atom pair belonging to the distance restraint. The calculation of I on the basis of the covalent geometry is referred to as I_a. For short-range distances involving up to two torsion angles, the maximal distance can be determined analytically (Güntert *et al.*, 1991). For distances depending on $t > 2$ torsion angles, the empirical relationship:

$$U_i = U_{\max} \cdot \left(1 - e^{\frac{-2.1 \cdot t}{U_{\max}}}\right) \tag{3.11}$$

with $U_{\max} = 2 \cdot R_G$ is employed, where R_G corresponds to the protein's radius of gyration. The determination of the empirical relationship between the number of angles and the correct distance, and the estimation of the radius of gyration are given in chapter 3.3.3 and 3.3.2, respectively.

The presented information content concept can also be generalized to ambiguous distance restraints with the effective distance $d = \left(\sum_{k=1}^{m} d\left(a_k, b_k\right)^{-6}\right)^{-\frac{1}{6}}$, where the sum runs over the m assignments $\left(a_1, b_1\right), \ldots, \left(a_m, b_m\right)$ (Nilges *et al.*, 1997). The probability calculation can be used without a change if the distance between the single pair of atoms is replaced by the corresponding effective distance. The redundancy R_{ij} between two ambiguous distance restraints A_i and A_j with m_i and m_j assignments, respectively, is computed as:

$$R_{ij} = \sum_{k=1}^{m_i} \frac{1}{m_j} \sum_{l=1}^{m_j} R_{i_k j_l}, \tag{3.12}$$

where $R_{i_k j_l}$ is the individual redundancy between the k^{th} assignment of restraint A_i and the l^{th} assignment of restraint A_j. The above defined redundancy calculation is inspired by the assumption that usually one of the individual assignments of an ambiguous distances restraint is correct. Assuming that all possible assignments of restraint A_j are equally probable to be correct, the corresponding probability $R_{i_k j_l} = P\left(A_{i_k} | A_{j_l}\right)$ for an individual assignment A_{j_l} is multiplied by the *a priori* probability $\frac{1}{m_j}$ when incorporating it into the redundancy R_{ij} between the two ambiguous distance restraints.

3.3 Material and Methods

3.3.1 Experimental data set

The information content, I, has been assessed on the basis of NMR distance restraint sets of 27 proteins for which a NMR solution structure has previously been determined. The data sets were obtained from the NorthEast Structural Genomics consortium (NESG) (Wunderlich *et al.*, 2004) which is a project that solves three-dimensional protein structures on a large scale using X-ray and NMR data. Tab. 3.2 gives an overview of some characteristics and data bank accession codes of the utilized proteins. Only monomeric protein structure data sets were selected. The data sets can be obtained directly from the NESG webpage (`http://psvs-1_4-dev.nesg.org/results/rosetta_MR/dataset.html`). The diversity of the set of proteins was convincing. The size of the proteins ranges from 5.2 kDa to 22 kDa. The α-helical fraction ranges from 0.0-40.0 % and the β-strand fraction ranges from 0.0-30.0%. Some proteins are purely α-helical while others display structures dominated by β-strands. Results achieved within this study are demonstrated by using the example of 2JQN. 2JQN is a protein of medium size within this data set and includes both α-helical and β-strand secondary structural elements.

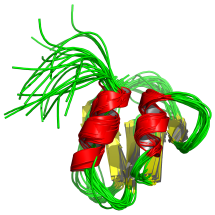
Table 3.2: Overview of the dataset used for the evaluation of the information content.

NESG ID	**PDB Code**	**BMRB code**	**Sequence**	n	α **helical**	β **strands**	**random coil**
BeR31	2K2E	15702	158	1106	4.5	11.6	81.9
CcR55	2JQN	15821	116	1228	16.7	9.0	72.2
DhR29B	2KPU	16570	90	802	0	36.4	63.5
DrR147D	2KCZ	16100	155	1406	5.1	2.5	90.3
ER382A	2JN0	15079	50	468	1.9	14.8	81.2
GR4	1RZW	6058	123	2369	25.1	6.4	66.4
GmR137	2K5P	15844	78	950	12.2	15.4	70.3
HR1958	1XPW	6344	143	976	5.1	5.8	87.0
HR3646E	2KHN	16250	121	1604	32.2	0.0	74.8
HR4435B	2L1P	17092	83	801	27.0	0.0	70.9
HR4527E	2L33	17169	91	2501	12.9	14.1	83.8
HR4694F	2L05	17436	86	1717	10.3	15.4	72.2
HR5546	2KPW	16572	122	1551	0.0	18.0	80.0
LkR112	2KPP	16563	114	2578	3.2	32.2	71.6
MrR110B	2K5V	15849	98	1633	5.1	27.7	65.1
OR8C	2KKZ	16376	134	2212	2.5	27.0	68.3
PfR193A	2KL6	16358	108	2674	0.0	27.0	70.9
PsR293	2KFP	16186	125	1865	14.1	10.9	72.9
SgR209C	2L06	17031	155	1852	23.8	0.0	74.1
SgR42	2JZ2	15604	66	560	1.9	19.3	76.7
SR213	2HFI	16113	123	2161	35.4	0.0	62.5
SR384	2JVD	15476	48	1053	20.0	0.0	78.0
SrR115C	2KCV	16084	100	3142	75.7	0	24.2
StR65	2JN8	15089	109	1274	29.6	0.0	68.3
StR70	2JZT	7178	142	1525	18.0	7.7	72.2
XcR50	1XPV	6363	78	1559	20.6	7.0	70.3
ZR18	1PQX	5844	91	1174	7.7	18.0	72.2

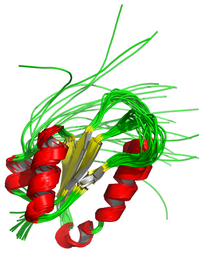
n refers to the number of distance restraints.

Available distance restraint data sets and the protein sequence were used to recalculate the NMR structures of the proteins listed in Tab. 3.2 using CYANA. Structure calculation

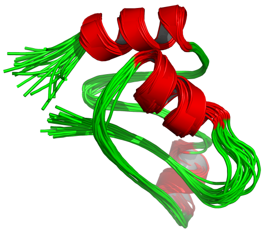
1PQX



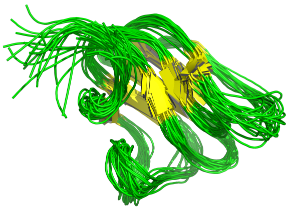
1RZW



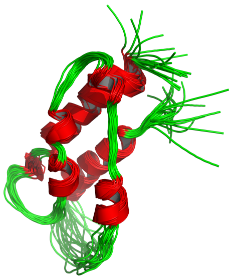
1XPV



1XPW

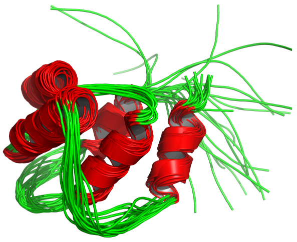


2HFI

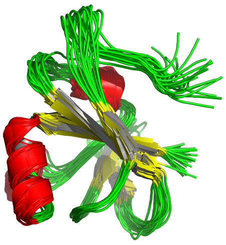


2JN0

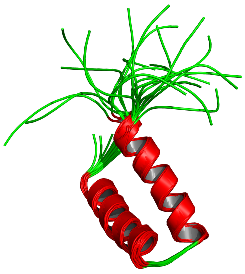
2JN8



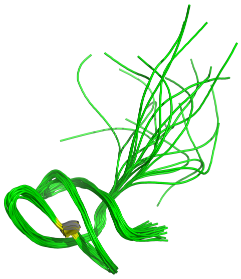
2JQN



2JVD



2JZ2



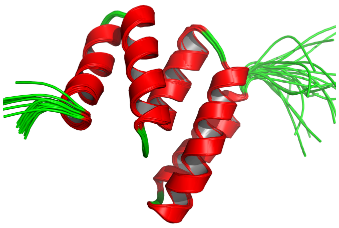
2JZT

2K2E

2K5P

2K5V

2KCV



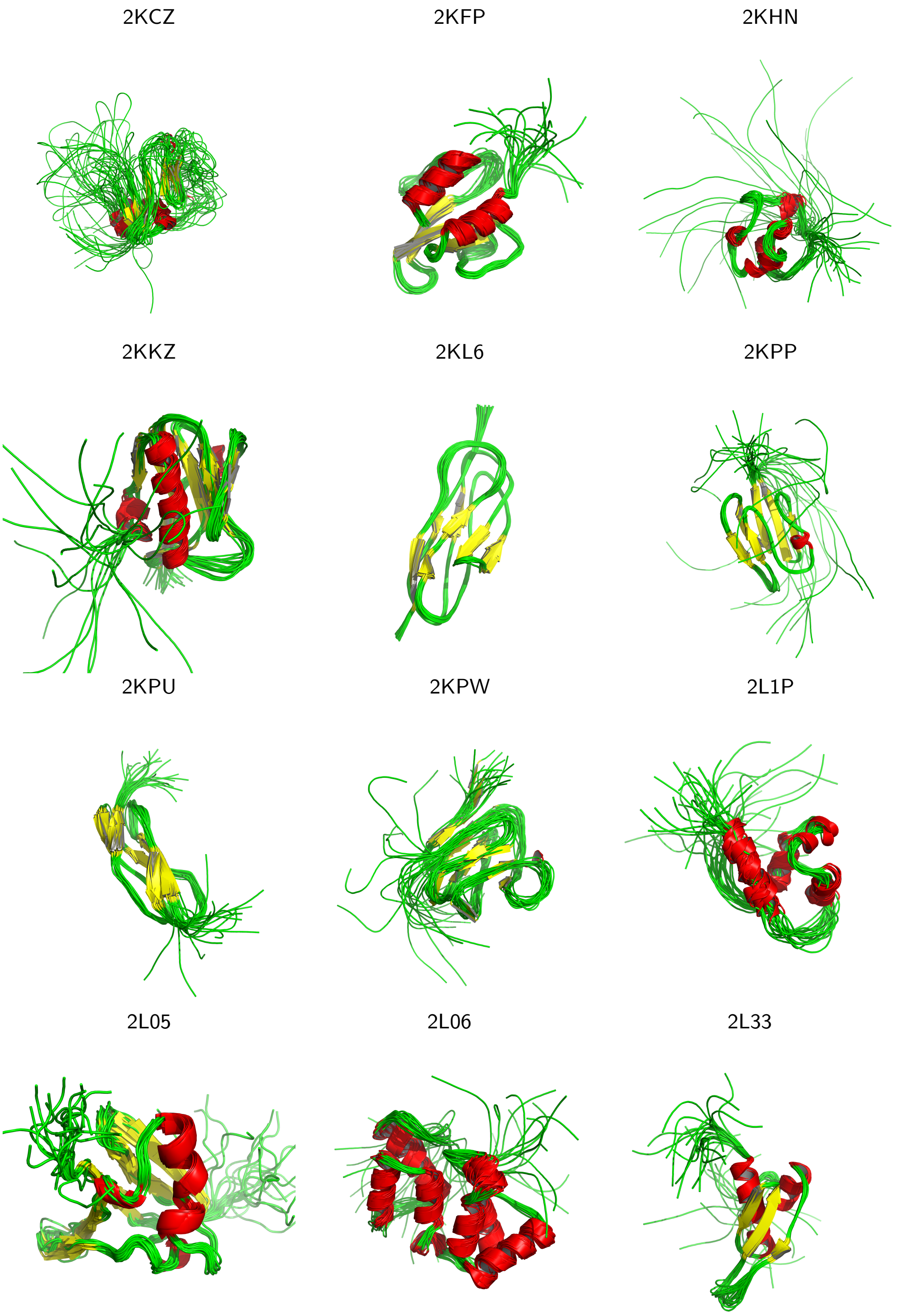


Figure 3.4: Recalculated structure bundles from the proteins of the above introduced data set (Tab. 3.2). Structure calculation was performed with CYANA, as explained, using as input only the protein sequence and the deposited distance restraint file. The twenty best calculations in terms of CYANA target function were used for structure bundle representation. Helical regions are depicted in red, β-sheet regions are shown in yellow, and random coil regions are shown in green. PyMol was used for visualization.

was performed with 10,000 torsion angle dynamic steps using 200 starting structures. The twenty best structures, in terms of CYANA target function, were used for bundle representation. The recalculated structure bundles are presented in Fig. 3.4 (double sided figure).

Table 3.3: RMSD radius of recalculated structure bundles and the underlying RMSD range, which is determined from the ordered regions as explained in chapter 3.3.4.4. Structure bundles are presented in Fig. 3.4.

PDB code	RMSD radius [Å]	RMSD range
1PQX	0.97	2-9,13-20,29-34,41-47,51-57,60-71,74-75,78-85
1RZW	0.70	3-14,22-45,49-82,91-110
1XPV	0.55	3-9,16-27,30-37,40-41,44-45,48-61,64-65,68-72
1XPW	1.38	3-9,12-22,25-33,39-80,83-95,104-124,127-140
2HFI	0.81	9-24,39-65,74-94,98-120
2JN0	0.60	4-7,10-15,19-22,25-32,35-47
2JN8	0.83	11-22,27-28,32-90,96-108
2JQN	1.06	3-28,33-52,55-63,70-75,80-99,105-109
2JVD	0.32	6-17,21-22,26-39
2JZ2	0.76	3-15,18-28,31-38,42-56
2JZT	0.96	13-17,20-42,58-65,68-89,95-107,115-126
2K2E	1.01	12-18,22-37,42-64,89-103,107-147
2K5P	1.05	2-8,11-12,18-24,31-65
2K5V	0.69	2-24,27-29,37-46,49-81,84-94
2KCV	0.28	3-17,21-93
2KCZ	1.71	5-23,32-33,42-48,56-57,64-80,89-97,100-110,138-147
2KFP	0.94	2-20,29-33,36-71,74-83,89-100,104-118
2KHN	1.01	25-35,38-45,48-65,70-92,95-100,106-114
2KKZ	0.73	5-54,58-80,90-121
2KL6	0.45	3-48,50-96,98-108
2KPP	0.45	7-38,41-94
2KPU	1.05	8-28,31-41,49-72,76-92
2KPW	1.42	15-21,25-39,43-48,53-59,62-76,86-94,101-106,109-117
2L1P	1.07	22-30,34-55,68-79
2L05	0.56	9-16,19-54,59-82
2L06	1.01	13-19,22-41,43-45,47-77,79-84,86-104,106-148,150-151
2L33	0.61	12-27,31-38,45-82

3.3.2 Radius of gyration estimation

The concept of I is based on the assumption that structures are folded in a globular manner. The radius of gyration is used to create random structures within the defined radius limit in the I_s calculation mode. If a reference structure is available the radius of gyration can simply be determined from the coordinates of the PDB file using the CYANA command **structure radius**. If the radius of gyration cannot be determined from an available ensemble of structures, an alternative approach, which relies only on the primary sequence, is needed. The expected radius, R_G, of a single domain protein consisting of N amino acid residues in its native conformation follows the empirical relationship $R_G = 2.2N^{0.38}$Å (Skolnick *et al.*, 1997). The radius of gyration limits the value of the potential maximal distance (see chapter 3.3.3) between the atom pair being restricted by restraint A_i in the I_a calculation mode.

3.3.3 Maximal distance estimation

As explained above the information content can be determined on the basis of a random structure bundle (I_s) or on the basis of the covalent geometry of the protein (I_a), which makes it completely reproducible and independent from the bundle of input structures. Both, the probability $P(A_i|0)$ (Eq. 3.7) and the redundancy $P(A_i|A_j, 0)$ (Eq. 3.8) rely on an estimation of the maximal distance, U_i, between the atom pair being restricted by restraint A_i. In the I_s mode, U_i is derived directly from the coordinates of the random structure bundle. If the I_a mode is preferred, U_i has to be determined in an alternative way. We therefore chose to estimate U_i based on the number of torsion angles t which separate the atom pair, i.e. the torsion angles separating the two atoms are summed up. t was then compared to the real maximal distance between the atom pair derived from a structure bundle. The relation between the maximal determined distance from the random structure bundle and t is visualized in Fig. 3.5 exemplary for 2JQN. Remaining proteins are presented in appendix B, Fig. B.1. The resulting plot has been fit for each protein to yield the parameters $a = 2.1$ Å in:

$$U_i = U_{\max} \cdot \left(1 - e^{\frac{-a \cdot t}{U_{\max}}}\right), \tag{3.13}$$

where t represents the number of torsion angles separating the atom pair, U_i represents the maximal distance between the atom pair, and $U_{\max}$ corresponds to $2.0 \cdot R_G$.

The distance estimated on the basis of t correlates with the real distance. The correlation is especially good in the range of lower t values, i.e. $t < 30$. The more long range the distance becomes, the higher the deviation of the real distances from the on t estimated distance. However, in the range of larger t values, the fit passes through the middle of the observed distances. Accordingly, some distances are underestimated while others are overestimated. The deviation between real distance and the estimated distance becomes more severe at higher t values, because t can only take discrete values, whereas within the structure bundle the distance can virtually take every value within $2.0 \cdot R_G$. The influence on the information content, i.e. the correlation between I_s and I_a is discussed in chapter 3.4.2.

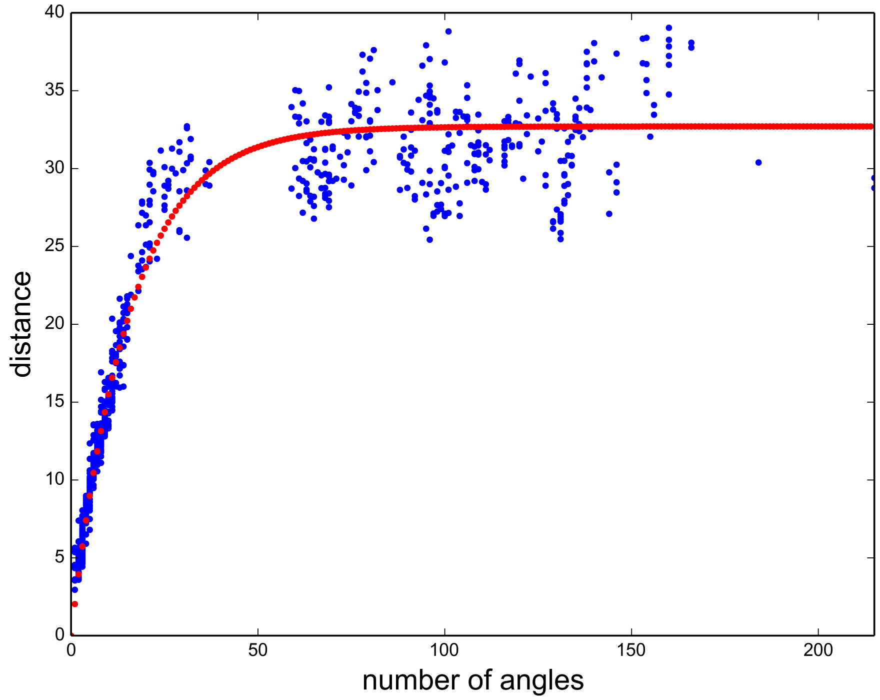


Figure 3.5: Correlation between the maximal distance of an atom pair and the number of torsion angles separating the atom pair using the example of 2JQN. The real maximal distances is derived from a bundle of random structures. The number of torsion angles are determined from the covalent geometry of the protein. Corresponding values are plotted and fit by Eq. 3.13 (red data points). The exponential fit is used to estimate the maximal distance when no structure bundle is available.

3.3.4 Work flow

The information content is calculated as explained in chapter 3.6. In the following, additional commands and tools are explained that were used for performing the experiments that correspond to the results presented in chapter 3.4. Literal CYANA input is written in `typewriterfont`.

3.3.4.1 Generation of restraint type specific data sets

Restraint type specific data sets correspond to data sets that include only short-range, $|i$-$j| \leq 1$, (intraresidual, $|i-j| = 0$ and sequential, $|i-j| = 1$), medium-range, $2 < |i-j| \leq 5$, or long-range, $|i-j| > 5$, distance restraints, where i and j refers to the index of the residues in the primary sequence. The results of I calculated from restraint-type specific data sets are presented in chapter 3.4.1.1. Data sets can be filtered according to the above defined classification of short-, medium- and long-range restraints by using the following CYANA commands:

- Select only **short-range** distance restraints

  ```
  read upl name
  distances select "** level=0..1"
  write upl short
  ```

- Select only **medium-range** distance restraints

  ```
  read upl name
  distances select "** level=2..4"
  write upl medium
  ```

- Select only **long-range** distance restraints

  ```
  read upl name
  distances select "** level=..5 multiple=ifall"
  write upl long
  ```

The above presented examples read the distance restraint file first (line 1), then select restraints of a specific type (line 2), and store the selected distance restraints in a new file (line 3).

3.3.4.2 Changing upper limit values

The upper limit value of a distance restraint has a crucial role in I calculation. It affects the probability $P(A_i|0)$ and the redundancy R_i with the remaining restraints in the data set. Therefore, we investigated the influence of the upper limit bound on I by increasing and decreasing the upper limit value by 1 Å with respect to the original value. All upper limits of a restraint data set can be changed by the following CYANA commands:

```
read upl name
distances set "**" bound=bound+1.0
write upl name-upl.upl
```

The above example, reads the upper limit file first (line 1), then selects all distance restraints, increases the upper limit value by 1 Å (line 2), and the modified restraint file is stored (line 3).

3.3.4.3 Generation of random structure bundles

I_s relies on an ensemble of random structures. In order to estimate the structural information of distance restraint data sets an ensemble of random structures has to be created. The structure bundle has to be random but also globular, because I is based on the assumption that protein structures usually fold in a globular manner. In order to generate globular random structures, the radius of gyration is determined, either as explained in chapter 3.3.2 or by estimating the radius of an available structure bundle, using the CYANA command **structure radius**. For our test data set, we recalculated the NMR structures on the basis of the provided distance restraint data sets, as explained above 3.3.1 and determined the radius of the recalculated structures with the CYANA command. The structure radius is then used as a restraint for random structure generation by CYANA using the following commands:

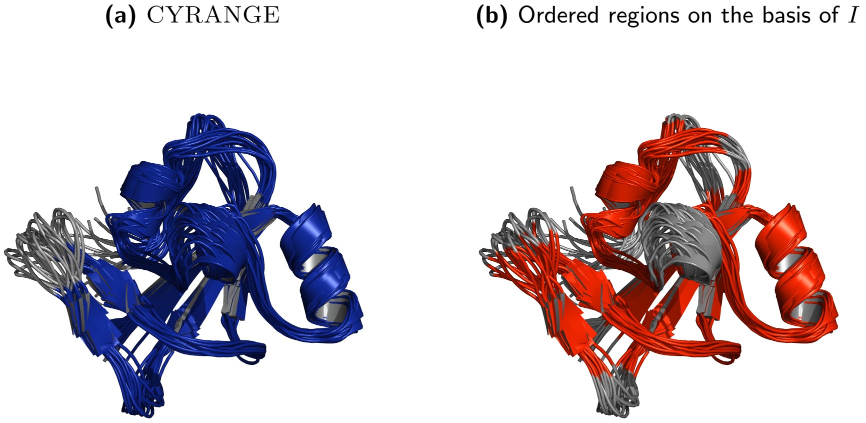
```
read seq name
seed=4290
atom gyr bounds=0.0..14.5
calc_all 200 steps=4000
overview file=random.ovw structures=20 pdb
```

In the example above, the protein radius is set to 14.5 Å (line 3), and the bundle of random structures includes 20 conformers (line 5). We tested varying numbers of input structures S, i.e. 2 and 10, and 10 to 90 in steps of 10. For each number of structures S, included in the bundle, 10 random structure calculations were performed with a different seed for creating starting structures (line 2), leading to 10 different bundles including S structures. The results of the I calculation were averaged over these 10 different structure bundles, including the same number of structures S. The results are presented in chapter 3.4.1.3.

3.3.4.4 Estimation of the ordered protein regions

Ordered regions of the protein can be determined on the basis of the individual I of each residue. In this case only restraints which contain secondary and tertiary structural information, i.e. medium- and long-range distance restraints, were considered. Thereby, one gets a good approximation of the ordered regions. I of residue i has to be > 1 in order to contribute to the ordered regions. Residues with $I > 1$ are summed up and used for scaling I to give I_r. Fig. 3.6 shows the ordered regions determined on the basis of I and ordered regions determined from CYRANGE (Kirchner & Güntert, 2011) in the context of a structure bundle, using the example of 2JQN.

CYRANGE determines a larger range of ordered residues than the I based method. The I based method ignores lonely residues which are either very informative or not informative, CYRANGE on the other side includes small stretches that are flanked by


(a) CYRANGE

(b) Ordered regions on the basis of I

Figure 3.6: Structure bundle of 2JQN is presented in grey, whereas ordered regions determined with CYRANGE are depicted in blue (**(a)**), and ordered regions determined on the basis of the long- and medium-range information content of individual residues are depicted in red (**(b)**).

ordered regions. Accordingly, the small helical region pictured in the middle of Fig. 3.6 (**a**) and (**b**) is included by CYRANGE and is excluded in case of the I based method, respectively. This regions is rather uncertain regarding the atomic coordinates, and accordingly does not bear residues which participate in distance restraints carrying enough long-range information. Summed up, the I based method gives a good estimation of the ordered regions in case no protein structure is available. For our investigations we used the I based method to determined ordered regions on the basis of the original data sets. The ordered regions are summarized in Tab. 3.3 and were kept the same for all RMSD radius calculations within this study.

3.3.4.5 Restraint data set minimization

The original data set of 2JQN is used to perform a data set minimization. In this context minimization is equivalent to deleting distance restraints. Gradually, one distance restraint is removed from the data set and the impact on I is monitored. For our experiment we performed two independent runs. In the first we remove in every cycle the restraint bearing the highest individual information content. In the second run (which is completely independent from the first run), the restraint with the lowest information content is deleted from the data set. With every step only one distance restraint is removed. The resulting restraint data set is used for I and structure calculation. Thereby, we were able to observe the behavior of I and the RMSD radius.

3.3.4.6 Generation of sparse data sets

Sparse distance restraint data sets can be generated by deleting a specified percentage of distance restraints from the original data set randomly. In this experiment 10-90% of the original data set were deleted in steps of 10. Each deletion of $x\%$ was repeated 10 times, in order to get a representative set of thinned out data sets. The results of this study are presented in chapter 3.4.4.2. A specific percentage of a distance restraint data set can be deleted by using the following CYANA commands:

```
read seq name
seed=4290
read upl name
distances delete "** fraction=0.4"
write upl sparse
```

In the shown example a distance restraint file, ‘name’ (line 3), is read and 40% of the distance restraints are randomly deleted (line 4), the resulting set of distance restraints is stored (line 5). The set of deleted distance restraints can be varied by changing the seed (line 2).

I is calculated on the basis of the thinned out distance restraint file and the restraint set is used for structure calculation. I and RMSD radius values are averaged over the 10 deletions and standard deviations are calculated.

3.3.4.7 Structure calculation by CYANA

Structure calculations by CYANA within this project were all performed starting from a set of distance restraints and the protein sequence. Structure calculation was performed with 10,000 torsion angle dynamic steps using 200 starting structures. The precision of calculated structure bundles is expressed by the RMSD radius (Güntert *et al.*, 1998), i.e. the average RMSD between individual conformers and the mean coordinates of the structure.

3.4 Results and Discussion

In the first part of this study, characteristics of the information content, I, e.g. restraint type-specific I, and the dependence on certain calculation parameters, e.g. number of random structures, upper and lower limit value, are analyzed and presented. We then show how I is scaled in order to get a measure that is data set size independent and comparable, I_r. In the last part we correlate I_r to the precision of the resulting structure bundle.

3.4.1 General characteristics of the I

3.4.1.1 Restraint type-specific I

The information content of a distance restraint data set has been assessed for different types of restraints. The type-specific information content results are shown in Tab. 3.4. Restraints were categorized into intraresidual ($|i-j|=0$), sequential ($|i-j|=1$), short-range ($|i\text{-}j| \leq 1$), medium-range ($2 < |i-j| \leq 5$), and long-range($|i-j| > 5$), where i and j refer to the residue indices of the participating atoms. Additionally, the data set size in percentage of the total number of restraints and the backbone RMSD radius of a subsequent structure calculation with the corresponding type-specific restraint sets as input are reported.

Table 3.4: Type-specific I of 2JQN. I and I_r (explained in chapter 3.4.3) are presented. Furthermore, the data set size and the RMSD radius of the resulting structures calculated from the corresponding restraint files is presented.

	I (I_r)	**RMSD radius (Å)**	**Data set size [%] (absolute number)**
Intraresidual restraints	11.70 (0.14)	25.56	14.90 (183)
Sequential restraints	19.21 (0.22)	28.62	29.32 (360)
Short-range restraints	30.91 (0.36)	25.72	44.22 (543)
Medium-range restraints	124.9 (1.45)	17.12	22.15 (272)
Long-range restraints	355.22 (4.13)	1.63	33.63 (413)
All restraints	510.42 (5.94)	1.05	100.00 (1228)

Intraresidual restraints: $|i-j|=0$; Sequential restraints: $|i-j|=1$; Short-range restraints $|i\text{-}j| \leq 1$; Medium-range restraints: $2 < |i-j| \leq 5$; Long-range restraints: $|i-j| > 5$

Tab. 3.4 shows that short-range (intraresidual and sequential) restraints carried the fewest structural information (I=30.91), even though they represent 44% of the data set.

This can be explained by the fact that short-range restraints are mostly already fulfilled by a bundle of random structures due to restraining only local features of the structure that are already restricted by the covalent geometry of the protein. Accordingly, $P(A_i|0) = 1.0$, and the information $I(A_i|0) = 0.0$. Medium-range restraints contributed a considerable higher amount of information (I=124.9), i.e. four times as much as short-range restraints even though they accounted only 22% in numbers to the data set. This can be explained by the fact that medium-range restraints hold more information, especially in terms of secondary structural information. The RMSD radius of a structure calculated from only medium-range restraints however does not improve significantly compared to a structure that was only calculated from short-range restraints. However, when analyzing the resulting structure models in more detail, it can be observed that the structures calculated from medium-range restraints included most of the helices of the native fold. The bulk of the information stemmed from long-range restraints (I=355.22) which was approximately three times as much as the I of medium-range restraints. Long-range restraints are commonly not fulfilled by a random structure, accordingly $P(A_i|0)$ becomes zero. Hence, the information added is comparatively high, assuming the redundancy is in general low and implying the prominence of these restraints for structure calculation. The RMSD radius also reflects the importance of long-range restraints. Structures calculated from a set of exclusively long-range restraints yield significantly lower RMSD radius values ($<$ 2.0 Å) supporting their central role. I of the complete restraint data set is considerable higher than I of only long-range data set (I=510.42) what is also reflected in the resulting structure bundle that has a significantly higher precision (1.05 Å).

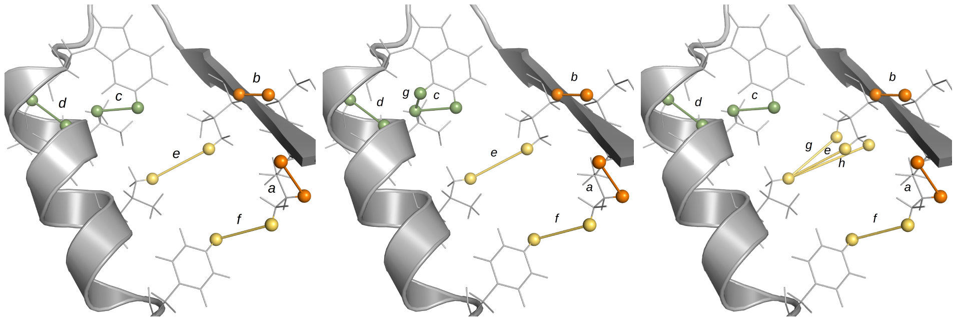
The study of the restraint type-specific I on a single distance restraint data set already reveals its intuitive change with the restraint type and already shows a tendency of the correlation with the RMSD radius. This will further be analyzed in chapter 3.4.4.

3.4.1.2 Effect of redundant restraints

In this section, we want to exemplify the impact of redundant restraints on the individual redundancy R_i and the individual information $I(A_i|0)$. Therefore, we created a small distance restraint data set consisting of six restraints (referred to as data set **(a)**), which are visualized in Fig. 3.7 **(a)**, i.e. two short-range restraints (depicted in orange; a and b), two medium-range restraints (depicted in green; c and d), and two long-range restraints (depicted in yellow; e and f). Tab. 3.5 lists the upper limit value of the corresponding restraint (upl), the individual probability to be fulfilled by a random structure, $P\left(A_i|0\right)$, the individual redundancy, R_i, and the individual information, $I\left(A_i|0\right)$. The overall information content of data set **(a)**, $I\left((a)|0\right)$, can be calculated by summing up the individual $I\left(A_i|0\right)$, to give $I\left((a)|0\right)$ =6.76. The restraint type imposed differences in information was also observed and was in agreement with previous observations.

Table 3.5: Effect of redundant restraints on the individual redundancy R_i and the individual information content $I\left(A_i|0\right)$ of restraint A_i. The distance restraints belonging to set **(a)**, **(b)**, and **(c)** are displayed in Fig. 3.7 **(a)**, **(b)**, and **(c)**, respectively, in the context of a structure.

	(a)				**(b)**				**(c)**			
A_i	upl (Å)	$P\left(A_i\|0\right)$	R_i	$I\left(A_i\|0\right)$	upl (Å)	$P\left(A_i\|0\right)$	R_i	$I\left(A_i\|0\right)$	upl (Å)	$P\left(A_i\|0\right)$	R_i	$I\left(A_i\|0\right)$
a	5.0	0.66	1.00	0.18	5.0	0.66	1.00	0.18	5.0	0.66	1.00	0.18
b	3.5	0.93	1.00	0.03	3.5	0.93	1.00	0.03	3.5	0.93	1.00	0.03
c	5.0	0.04	1.03	1.35	5.0	0.04	1.46	0.95	5.0	0.04	1.03	1.34
d	5.0	0.30	1.03	0.51	5.0	0.30	1.05	0.50	5.0	0.30	1.03	0.51
e	4.4	0.00	1.00	2.43	4.4	0.00	1.01	2.43	4.4	0.00	2.14	1.14
f	4.0	0.01	1.01	2.26	4.0	0.01	1.01	2.26	4.0	0.01	1.02	2.24
g					5.0	0.04	1.46	0.95	4.4	0.00	2.14	1.14
h									4.4	0.00	2.14	1.14



(a) $I = 6.76$ **(b)** $I = 7.30$ **(c)** $I = 7.72$

Figure 3.7: Visualization of distance restraint sets that correspond to the restraints specified in Tab. 3.5.

Then we introduced a medium-range restraint (restraint g in Tab. 3.5) to data set **(a)** (now referred to as data set **(b)**), which was partly redundant with restraint c (see Fig. 3.7 **(b)**). Restraint g enhanced the redundancy of restraint c from 1.03 to 1.46 and thereby reduced I of restraint c from 1.35 to 0.95. Restraint d was also effected by the introduced restraint g, but to a much lesser extent; the redundancy was enhanced from 1.03 to 1.05, and I was lowered from 0.51 to 0.50. The overall I of data set **(b)** was 7.30. Compared to data set **(a)** the overall I of the data set was enhanced by approximately 0.5. This however follows the intuitive behavior because restraint g added structural information on the orientation of the tryptophan side-chain with respect to the valine side-chain atoms.

An overall gain in I can be justified therewith.

Then we introduced two long-range restraints to data set **(a)** which were partly redundant with restraint e (referred to as data set **(c)**), visualized in Fig. 3.7 **(c)**. The two redundant long-range restraints, g and h, led to an enhancement in redundancy of restraint e from 1.00 to 2.14 and thereby a reduction in information of restraint e from 2.43 to 1.14. Restraints g and h also had a minor effect on f, whose redundancy was slightly enhanced from 1.01 to 1.02, and I was reduced from 2.26 to 2.24. The overall I of restraint set **(c)**) was 7.72. Adding the two partly redundant long-range restraints g and h had a similar effect as adding the partly redundant medium-range restraint to data set **(b)**. Restraint g and h provided additional structural information and reduced the available conformational space of the valine side-chain with respect to the second valine side-chain and therefore enhanced the overall I but reduced the individual I values.

3.4.1.3 Dependence of I on the number of random structures

As already noted above, I can be calculated on the basis of a structure bundle, I_s, or based on the covalent geometry of the polypeptide chain, I_a. I_s relies on a bundle of random conformers based on which the maximal distance, U_i, of a specific atom pair, A_i, is estimated. The probability that a given atom pair distance in a random conformer lies within x% of its maximal value U_i is approximately given by $U_i^3 - (1-x)^3 U_i^3)/U_i^3 \approx 3x$. This probability should equal $1/S$ to be expected by at least one of the S conformers of a random structure bundle. Hence, $x \approx 1/3S$. This means that, for instance, with $S = 20$ random conformers, one will approach the true maximal distance on average within less than 2%. The same approach cannot be used for minimal distances because in general a very large number of random structures would be necessary to come sufficiently close to the true minimal distance. However, in the absence of other information the minimal acceptable distance between two atoms is given by the sum of their repulsive core radii.

In order to verify the proposed number of random structures that should be used for the I calculation, the impact on I upon changing the number of random structures S has been assessed. The results, visualized in Fig. 3.8, have been determined on the 2JQN data set. The number of random structures was increased from 2 to 10, and then from 10 to 90 in steps of 10. For each number S, 10 structure bundles, starting from different seeds, were created. The structures were created with a restraint on the radius of gyration, as explained in chapter 3.3.4.3. I was then calculated on the basis of each structure bundle, thereby determining the average and the standard deviation of I over the 10 different structure bundles including the same number of structures S.

I showed an hyperbolic increase at increasing numbers of random structures S. The average I rose significantly when enhancing the number of structures in the bundle from 2 to 10. The standard deviation of I was also reduced significantly when comparing I calculated on the basis of 2 structures and 20 or more structures. As proposed above, a set of 20 random structures was sufficient for the I calculation. Increasing the number of random structures to > 20 does not influence the I significantly. A larger number of random structures prolongs computation time, which additionally justifies the usage of 20 structures. Following I_s calculations were performed on the basis of a bundle consisting of 20 random structures if not stated otherwise.

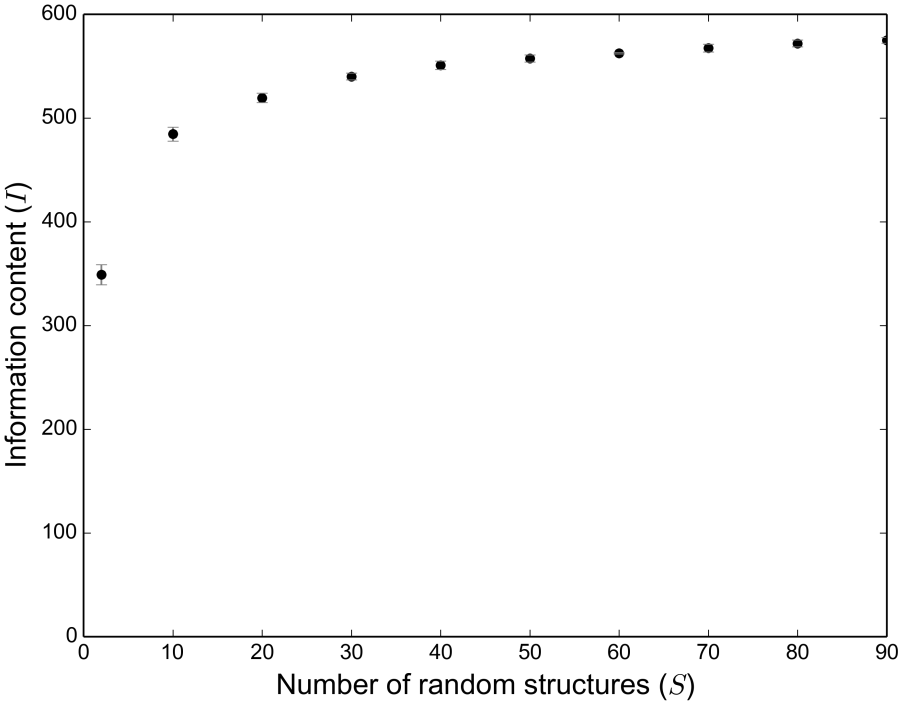


Figure 3.8: Influence of the number of random structures S included in the structure bundle on I calculation. The number of random structures S was enhanced from 2 to 10, and from 10 to 90 in steps of 10. For each number S, 10 different structure bundles on the basis of a different seed, were generated as explained in chapter 3.3.4.3. Average I values and standard deviations were calculated from these 10 calculations. The influence of the number of random structures on I is presented using the example of 2JQN.

3.4.1.4 Multiple restraints in the data set

I has been defined in a way that multiplying single restraints or the whole data set does not have an influence on the absolute I value. As explained earlier, in case of multiple restraints for the same distance only the more restrictive restraint is kept and considered for I calculation. I has been calculated for the original distance restraint data set of 2JQN. This data set was doubled and tripled, and additionally only long-range, short-range or medium-range distance restraints are doubled successively. The I value remained the same in all cases (I=510.42).

3.4.1.5 Dependence of I on upper and lower bound

The information content itself is very sensitive towards the value of the upper and lower limits as already visualized in Fig. 3.2 by the influence of these bounds on the accessible conformational space. Upper limit values are deduced from the volume or intensity of the peak corresponding to the atom pair. During calculation the upper and lower bounds have an influence on the probability of restraint A_i, $P(A_i|0)$ (Eq. 3.7), as well as the redundancy, R_i (Eq. 3.8). If the upper limit value of restraint A_i was increased and every other values was kept fixed, probability and redundancy of A_i also increased. This can be rationalized by the fact that a higher upper limit value corresponds to a larger accessible conformational space, making a restraint more likely to be fulfilled by a given random structure and also more likely to be redundant with other restraints A_j. Accordingly, both, a higher probability and a higher redundancy will yield lower I values comparatively. This supports the idea that higher upper limit values restrain the tertiary structure less and thus contain less information if the lower limit remains unchanged. The same characteristics were monitored when keeping the upper limit of restraint A_i fixed and decreasing the lower limit of A_i. Diminishing the lower limits enhances the accessible conformational space compared to the increased lower limit value and making it more redundant with the remaining restraints. Decreasing the upper limit value has the contrary effect, the probability and the redundancy is lowered, and hence I rises. Keeping the upper limit fixed and only enhancing the lower limit also lowers probability and redundancy, and thereby enhances I. In this case the available conformational space of the restraint is more restricted and therefore more informative. The behavior of I upon increasing and decreasing the upper limit value is illustrated in Tab. 3.6 using the example of 2JQN.

Table 3.6: Dependence of I and I_r on upper limit value (upl) using the example of 2JQN.

upl/lol (Å)	I	I_r
original	510.42	5.94
upl + 1.0 Å	305.75	3.56
upl − 1.0 Å	798.21	9.28

'original': data set remained unmodified; 'upl + 1.0 Å': all upper limits were enhanced by 1 Å; 'upl − 1.0 Å': all upper limits were decreased by 1 Å.

When enhancing all upper limits by 1.0 Å, I was drastically reduced from 510.42 to 305.75. Compared to that, lowering all upper limits by 1.0 Å led to a significant increase in I from 510.42 to 798.21. The behavior exemplified by the data set of 2JQN followed exactly the above proposed properties.

3.4.1.6 Dependence of I on protein size

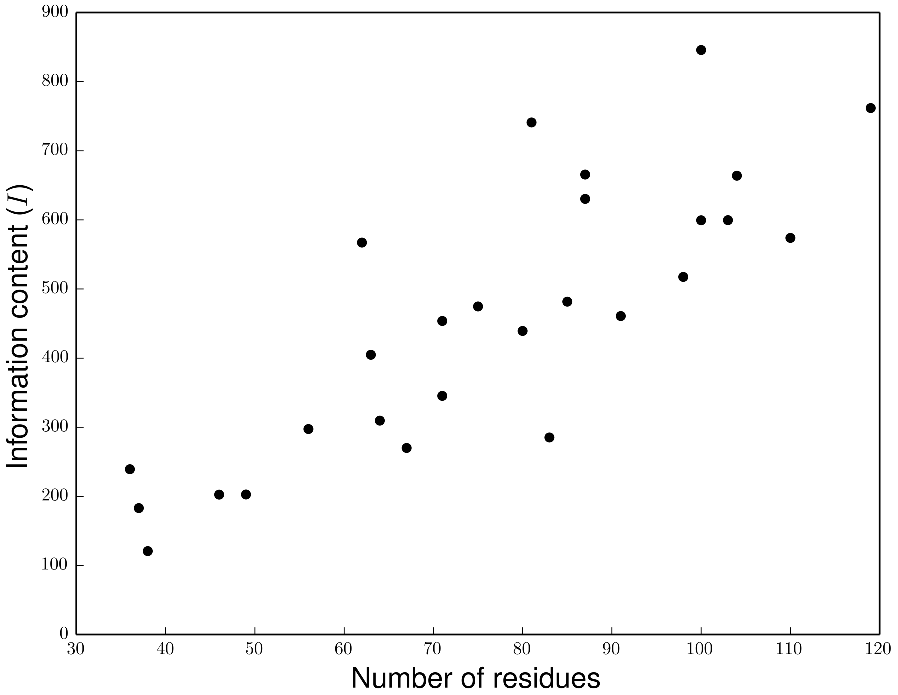


Figure 3.9: Correlation between I and number of ordered residues. Ordered regions were determined on the basis of the original distance restraint data set (Tab. 3.3) using the I based method explained in chapter 3.3.4.4. The underlying distance restraint data sets belong to the validation data set introduced in chapter 3.3.1.

I depends on the one hand on the structural information included in the data set, as presented before. On the other hand, I increases approximately linear with the size of the protein. The absolute I of our test data set, introduced in chapter 3.3.1, Tab. 3.2, has been calculated. The dependence of I on the size of the protein, here expressed in terms of number of residues in ordered regions, is presented in Fig. 3.9. The number of ordered residues was plotted against I. The value of I varied between 100 to 900 for ordered protein ranges from 37 to 119 residues. Deviations from the observed approximate linear behavior can be attributed to differences in the structural information included.

Generally, larger proteins have larger data sets, because more distance restraints are necessary to describe the underlying structure accurately. The structures recalculated from the deposited restraint files, were all characterized by a high structural precision, expressed in RMSD radius (see Tab. 3.3). However, the value of I varied with the size of the system. Consequently, I has to be scaled with respect to the size of the underlying system in order to make it comparable in-between different proteins and to give I an universal meaning. The scaled I is later correlated to the structural precision (chapter 3.4.3).

3.4.2 Agreement between structure dependent and independent calculation mode

As already mentioned, we provide two different calculation modes for I; one that depends on a bundle of random structures (I_s) and another one that is independent from random structures and relies only on the covalent geometry of the protein (I_a). The main difference between those two calculation modes is the estimation of the maximal distance, U_i, between the atom pair belonging to the distance restraint of interest. In case of I_s, U_i is deduced from the ensemble of random structures itself. In I_a mode, U_i does only depend on the number of torsion angles that separate the atom pair. The relationship between the number of torsional angles and U_i has been determined empirically (Eq. 3.11), as explained in 3.3.3. We compared I_s and I_a calculation modes for our test data set. The results are shown exemplary for 2JQN in Fig. 3.10. We created different input distance restraint files by randomly deleting a specific fraction of the original data set, as explained in chapter 3.3.4.6. In Fig. 3.11 I_s and I_a are presented for the remaining proteins of the test data set. The calculation modes correlate very well in most of the data sets, i.e. 1RZW, 2JQN, 2JVD, 2K2E, 2K5V, 2KCZ, 2KFP, 2KHN, 2KPP, 2KPW, 2L1P, 2L05, 2L33, and 2KPU. Minor deviations between the two calculation modes can be observed for 1PQX, 1XPW,

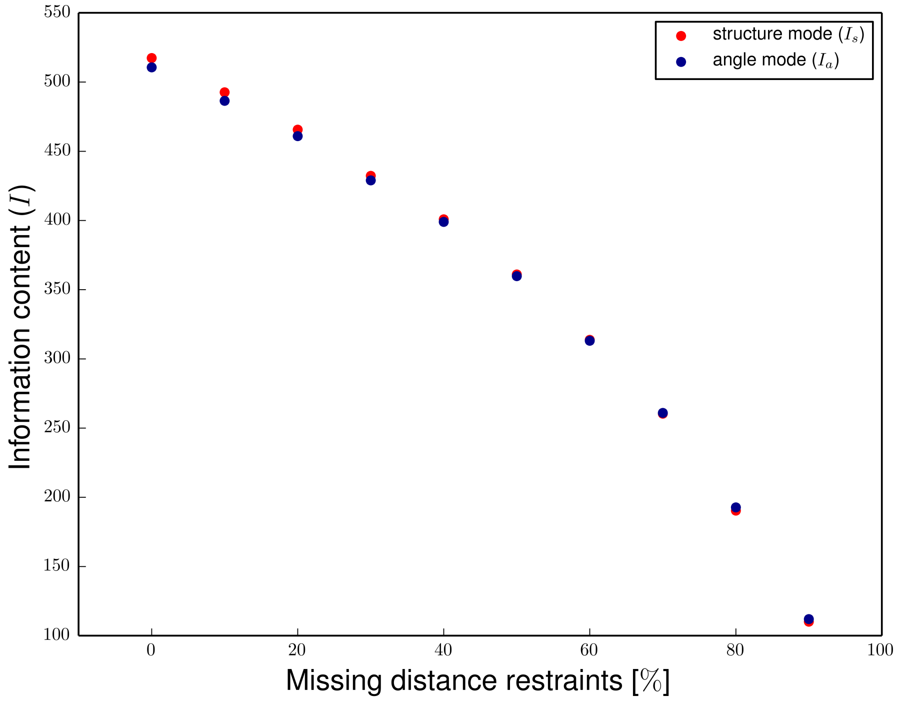


Figure 3.10: Agreement between structure dependent and structure independent information content calculation mode using the example of 2JQN. The percentage of missing distance restraint is plotted against I_s (red) and I_a (blue). Pearson correlation coefficient is 0.99.

2HFI, 2JN8, 2JZ2, 2JZT, 2K5P, 2KCV, 2KKZ, and 2L05. Rather significant deviations can be observed for 1XPV, 2JN0, and 2KL6. Deviation between results achieved in I_s and I_a mode can be attributed to differences in the U_i value. U_i can achieve higher values in the structure based mode, i.e. not all distances in the structure bundle are $< 2 \cdot R_G$ (referred to as the theoretical maximal distance U_{max}. In I_a mode, U_{max} is used as a fixed cutoff for the maximal distance estimation (Eq. 3.11). Whereas in I_s calculation mode this limitation is not given and even though the random structures have been calculated with a constraint on the radius of gyration (chapter 3.3.4.3) it might still be possible that there are some deflected local conformations which yield U_i values that are higher than U_{max}. Furthermore, in the I_a mode U_i takes discrete values for a specific number of angles, corresponding to a set of atom pairs, in contrast to that in I_s mode these atom pairs can adopt a wide range of U_i values. Despite the differences observed, the Pearson correlation coefficient of I_s and I_a is > 0.99 for the complete data set.

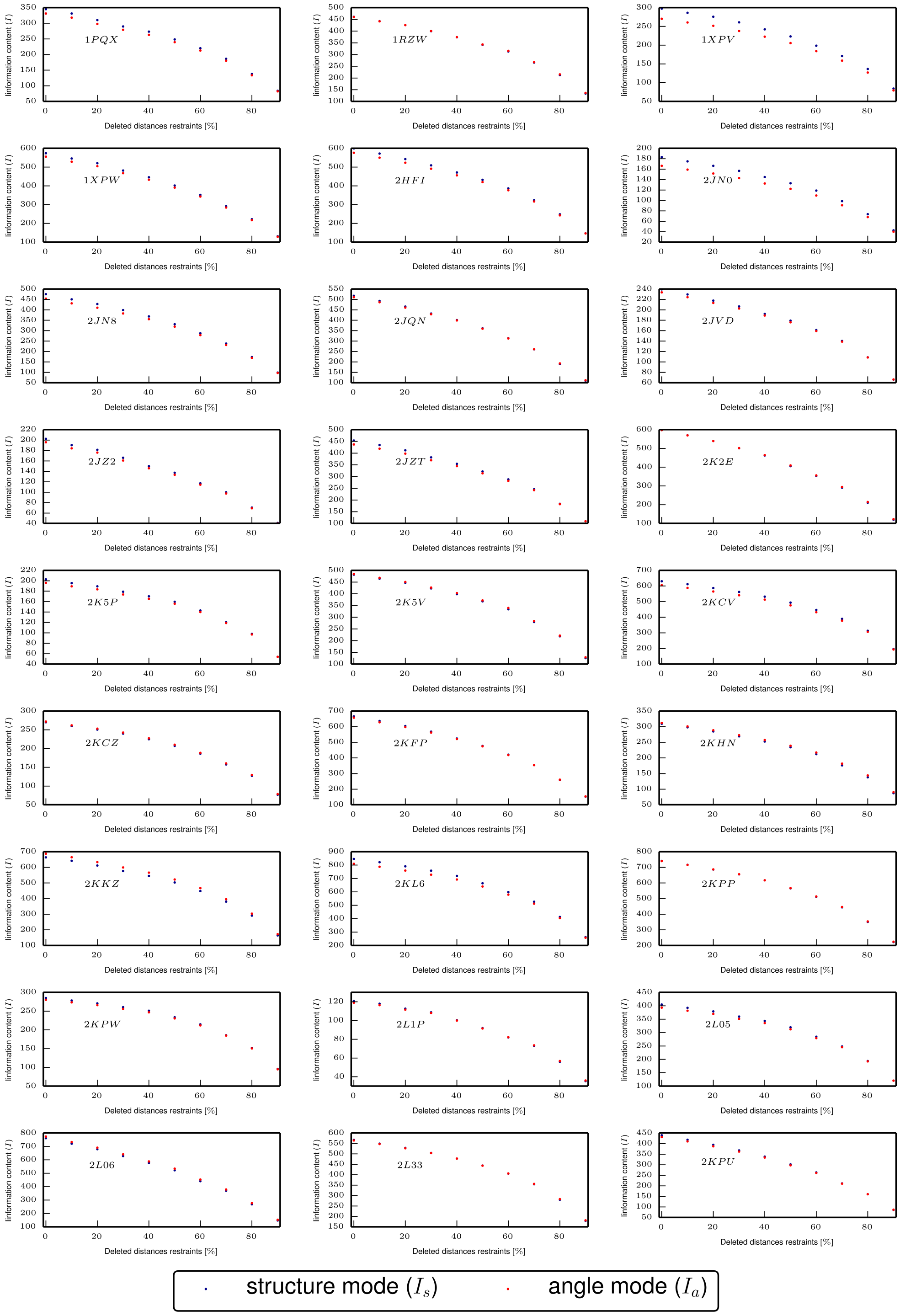


Figure 3.11: Agreement between structure dependent and structure independent information content calculation mode on the basis of the complete data set introduced in chapter 3.3.1. The percentage of missing distance restraint is plotted against I_s (blue) and I_a (red).

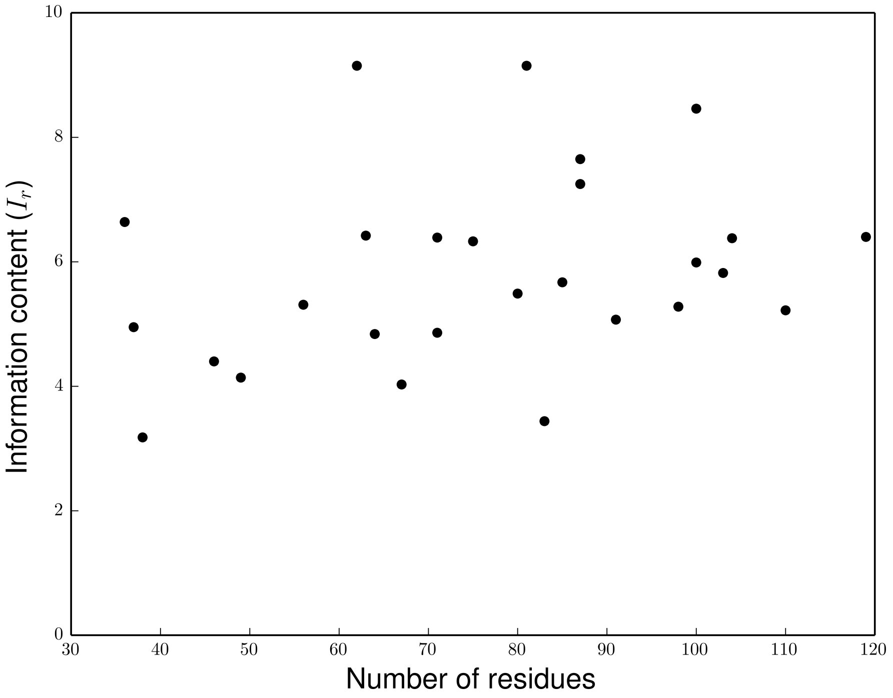


Figure 3.12: Correlation between I_r and number of ordered residues. The absolute I is scaled by the number of residues belonging to ordered regions of the protein. Ordered regions were determined from the original distance restraint data set (Tab. 3.3). The underlying distance restraint data sets belong to the validation data set introduced in chapter 3.3.1.

3.4.3 Scaling of I

As presented above, I strongly depends on the size of the underlying system, i.e. the number of residues in the protein sequence, apart from its dependence on the structural information. Generally, the larger the occupied conformational space by the protein, the more restraints are necessary to describe the structure adequately. On account of this the absolute value of I can be very different, even though the structural precision is exactly the same. The behavior of I is consistent within one system as presented above (chapter 3.4.1). However, our aim is to define a system size independent measure which can be correlated to the structural precision, expressed in RMSD radius. We therefore define I_r, a measure which is independent from the size of the protein and can be directly correlated to the structural precision. In Fig. 3.12 we demonstrate the correlation between I_r and the number of ordered residue regions. In principle, the results presented in Fig. 3.12 are equivalent to the results in Fig. 3.9 with the exception that I is scaled by the number of ordered residues. The number of ordered residues can either be determined by

CYRANGE on the basis of a reference structure bundle, can be provided by the user if ordered regions of the native fold are known, or it can be determined on the basis of the medium- and long-range information presented in the corresponding regions (the exact determination of the ordered regions on the basis of I is explained in chapter 3.3.4.4). For our calculations we used the I based method on the original restraint files to determine the ordered regions. Fig. 3.12 visualizes, in contrast to Fig. 3.9, that I_r is independent from the protein size. I_r varied between 3 and 9, whereas I, visualized in Fig. 3.9, varied between 100 and 900. Accordingly, the I_r gives a direct and comparable measure of structural information, that is in the following chapter connected to the RMSD radius.

3.4.4 Correlation between I and structural precision

3.4.4.1 Data set minimization

Exemplary, the provided 2JQN distance restraint file has been used to minimize the original data set as explained in chapter 3.3.4.5. Gradually, one distance restraint is removed from the data set, I of the reduced data set is determined and a structure bundle is calculated using CYANA. This procedure is repeated until the data set is diminished to one restraint. In each cycle, the distance restraint with the most informative individual I or in a second independent run the least informative individual I is withdrawn from the data set. Thereby, the behavior of I can be monitored and linked to the structural precision, i.e. RMSD radius. We used the individual information $I(A_i|0)$ of restraints as an indicator of non-relevant or relevant structural information.

Fig. 3.13 **(a)** shows the behavior of I upon removing the most informative restraints first. The I value dropped exponentially, whereas the RMSD radius increased exponentially and reached RMSD radius maximal values between 20–30 Å, which corresponds to an elongated structure bundle as shown in Fig. 3.13 **(a)**. The significant drop in I_r and the considerable rise in RMSD radius can be explained by the fact, that when eliminating the most informative restraints first, preferably long-range non-redundant restraints are removed first. These restraints hold the most structural information because of the long-range character and their unique information. Thereby, the structure bundle becomes rapidly imprecise. With this procedure approximately 200 restraints can be removed and the structure obtained is still reliable in terms of overall global fold (RMSD radius < 2 Å). I_r has to be > 3 in this case.

Fig. 3.13 **(b)** illustrates the behavior of I_r upon least informative restraint withdrawal.

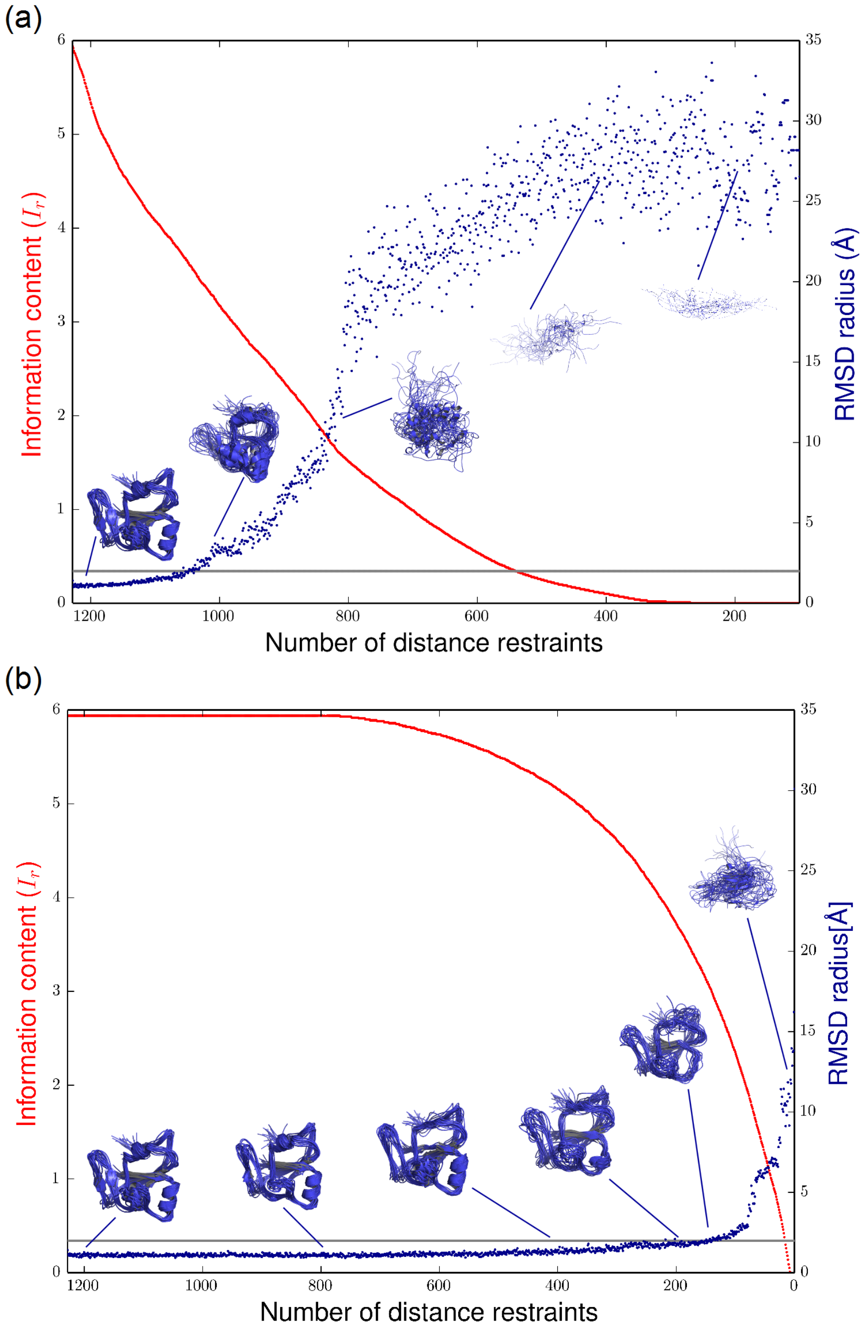


Figure 3.13: Number of distance restraints vs I_r (red, left axis), and number of distance restraints vs backbone RMSD (blue, right axis) using the example of 2JQN. Starting from the original data set, the restraint with the highest I is removed from the data set. The reduced data set is then used in I_r and structure calculation. Again, the most informative restraint is withdrawn from the data set, I_r and structures are calculated again. This is repeated until only a single restraint remains. The results are presented in **(a)**. The same procedure was performed on the least informative restraint, leading to **(b)**. Structure bundles calculated are linked to their accordant data set by blue lines. The gray line in both figures corresponds to a RMSD radius of 2 Å.

As can be seen, I_r remained stable over approximately 400 reduction steps. The RMSD radius was also not affected by the removal of these restraints which are either highly redundant restraints or short-range distance restraints. This underlines the high amount of redundancy which is usually present in experimental NMR data sets (Doreleijers *et al.*, 2009). Additionally 600 more distance restraints were removed from the data set without a significant drop in RMSD radius. The I_r however decreased by about 30%. This loss in information can mainly be attributed to medium-range distance restraints, which is illustrated in Fig. 3.13 **(b)** by the structure bundle corresponding to 200 distance restraints. This structure bundle achieved the approximate globular fold correctly, but secondary structural elements are rarely recognizable any more. As has already been observed in the minimization procedure on the least informative restraints, an $I_r > 3$ is necessary to achieve a precise structure bundle with a RMSD radius < 2.0 Å.

3.4.4.2 Sparse data

The complete protein data set was used to examine the characteristics of I upon random data withdrawal. For each protein 0–90% of the original restraints were randomly deleted. The deletion was performed 10 times for each percentage, using a different random seed. The exact performance of this experiment is explained in chapter 3.3.4.6. The reduced data sets were used to calculate I_r and the RMSD radius of the resulting structure bundle from a subsequent structure calculation on the reduced data set using CYANA. I_r and RMSD radius values are averaged over the 10 calculations and standard deviations are determined. The missing distance restraints (in %) are plotted against the mean value and the standard deviation of I_r and the RMSD radius, visualized in Fig. 3.14 using the example of 2JQN. I_r was reduced with rising percentage of missing distance restraints. Structure bundles presented in Fig. 3.14 refer to one of the 10 calculations, which can be considered as representative for all calculations since the standard deviations are usually relative low, except when deleting 80 or 90%. The conclusions that can be drawn from Fig. 3.14 coincide with observations discussed in chapter 3.4.4.1 and presented in Fig. 3.13: An I_r of approximately > 3 is necessary to obtain a structure bundle from the data set with a precision < 2.0 Å.

Results achieved on the basis of the remaining proteins are present in Fig. 3.15. The results obtained on the complete data set also coincide with previous observations. Generally, an $I_r > 3$ is necessary to obtain a structure bundle with a precision < 2.0 Å in a subsequent structure calculation.

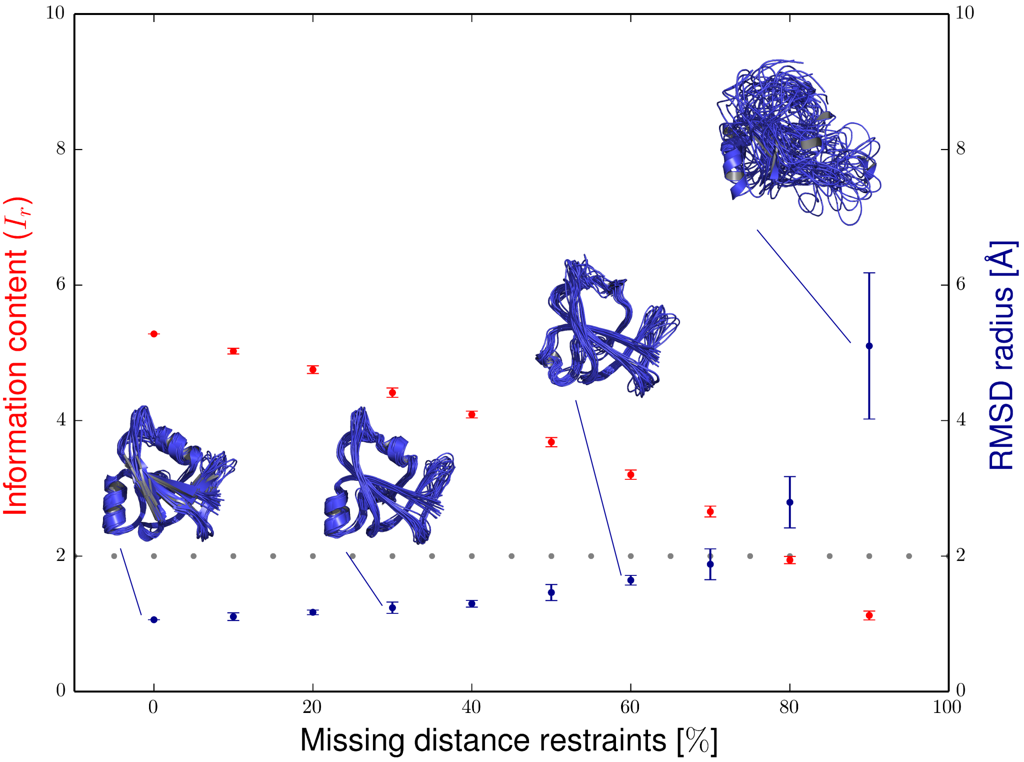


Figure 3.14: Missing distance restraints in dependence of I_r (red) and RMSD radius (Å) (blue) using the example of 2JQN. 10 to 90% of the distance restraint data set were deleted randomly in steps of 10. Each deletion has been performed 10 times. For each of the ten deletions of a specific percentage I_r and the RMSD radius of the structure bundle resulting from a subsequent structure calculation with the reduced restraint data set as input are calculated. Mean and standard deviation values are calculated from these ten calculations. The structure bundles shown refer to one of the ten structure calculations performed. The gray dotted line refers to an RMSD radius of 2 Å.

3.4.5 Computation time

The computation time of I depends on the size of the distance restraint files. A standard information content estimation with a set of about 1100 restraints took about 2 s. Another distance restraint data set which included about 6100 distance restraints took 26 s. Calculations have been performed on an standard desktop PC (Intel Core 2 Quad Q9400 processor).

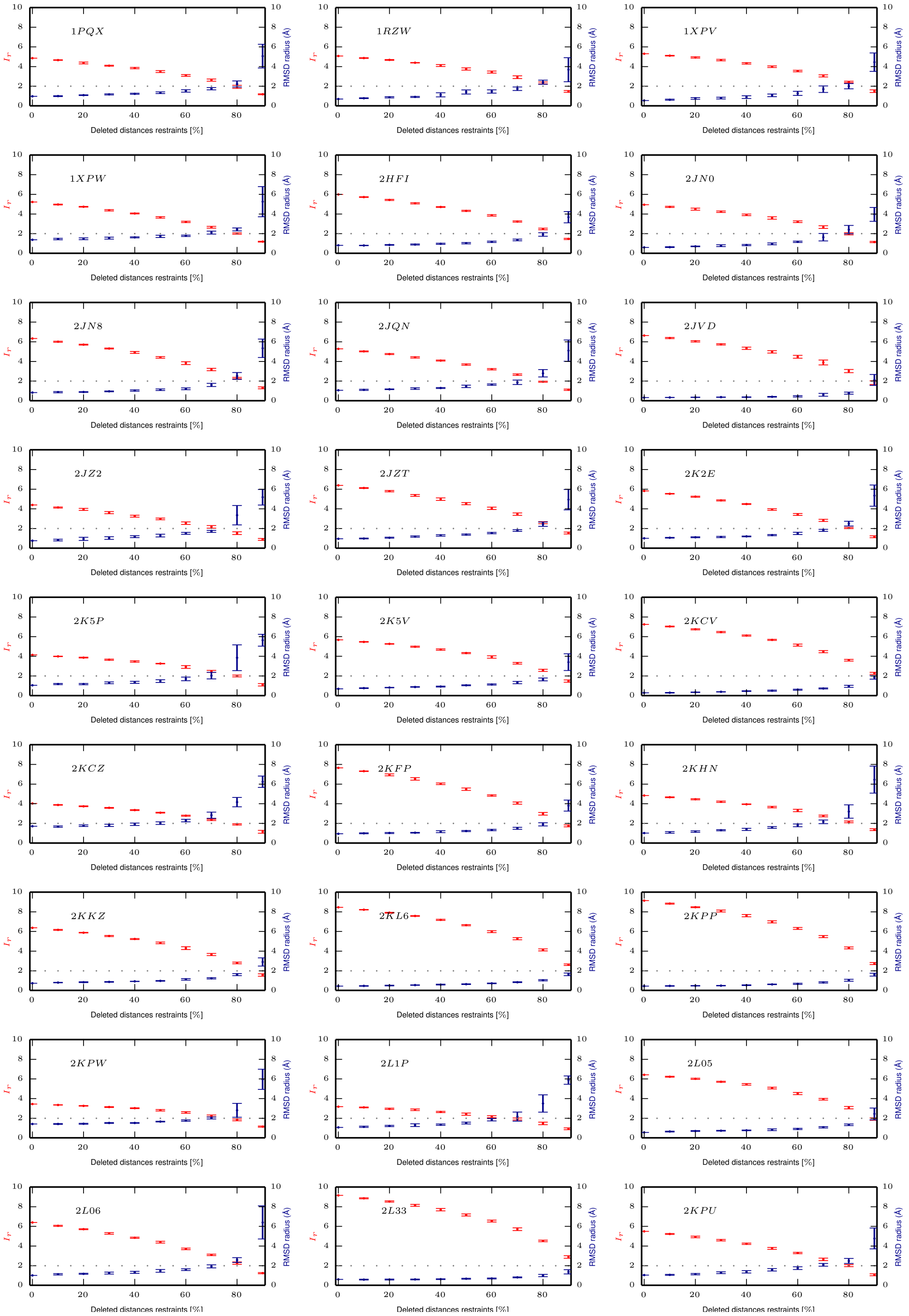


Figure 3.15: Missing distance restraints in dependence of I_r (red) and RMSD radius (Å) (blue) using the complete data set. Subplots are labeled with the PDB code of the protein. 10 to 90% of the distance restraint data set were deleted randomly in steps of 10. Each deletion has been performed 10 times. For each of the ten deletions of a specific percentage I_r and the RMSD radius of the structure bundle resulting from a subsequent structure calculation with the reduced restraint data set as input are calculated. Mean and standard deviation values are calculated from these ten calculations. The gray dotted line refers to an RMSD radius of 2 Å.

3.5 Conclusions

We introduced the information content I, a quantitative measure of structural information contained in NMR distance restraints. I is straightforward and fast to calculate. The calculation is performed completely automatic and does not require user input. However, user input can be provided if desired (see chapter 3.6). Several intuitive characteristics of I have been demonstrated:

1. The difference in structural information of varying restraint types has been presented, i.e. long-range distance restraints carry the most structural information, whereas short-range distance restraints contribute a significant lesser amount of information.
2. Structural equivalent restraints lead to an increase in restraint redundancy and an individual restraint information loss, whereas the overall information of the data set remains the same or is increased.
3. I depends on the upper and lower limit, i.e. I increases with a smaller available conformational space, and decreases with a larger available conformational space.

We showed that the I value depends aside from the structural information on the protein size. Therefore, we presented that I_r is an information content measure that is independent from the size of the protein. The information content can be calculated on the basis of a structure bundle or on the basis of the covalent geometry of the polypeptide chain. We demonstrated that both calculation modes are in good agreement. Finally, we presented that I_r can be correlated to the precision of a structure bundle that has been determined on the basis of this data set. We showed that an $I_r > 3$ is needed in order to obtain a structure bundle with an RMSD radius < 2 Å.

The information content calculation is implemented in CYANA but is also available as a stand-alone program, CYINFO, and can be downloaded from our web server (`http://tsushima/info.html`). It is also possible to upload data (in XEASY format) directly onto our server and get I and additional information on individual restraints instantly. CYINFO does not include any structure calculation routines. If the structure dependent calculation mode I_s is desired a random structure has to be provided otherwise it is possible to use the structure independent calculation mode I_a.

An extension of the information content to other types of structural restraints is convenient, e.g. torsion angle restraints, residual dipolar couplings, or pseudocontact shifts.

The basic ideas of probability and redundancy can be directly transferred to other restraints. Further, it would be interesting to connect I to structure quality parameters and investigate if a correlation is present.

3.6 Implementation in CYANA

In the following literal CYANA input is written in **bold** and other input is written in *italics.*

3.6.1 Command distances information

- **method**=string (default: angles)

  method can either be **angles** or **structure**, referring to I_a and I_s, respectively. **method=angles** determines the information content on the basis of the covalent geometry without using a three-dimensional structure as input. The calculation mode **structure** uses an ensemble of random structures to determine the information content.

- **radius**=real (optional)

  The radius of gyration can either by specified by the user via **radius** or if not specified it is calculated using the empirical relationship $R = 2.2N^{0.38}$Å (Skolnick *et al.*, 1997), where N represents the number of residues, as explained in chapter 3.3.2.

- **inforange**=integer list (optional)

  Ordered residue ranges can either be specified by the user or are calculated based on residues which possess long-range structural information. A region can be specified by the first and last residue number which are separated by '..'. Several ordered regions are separated by ',' (e.g. **inforange=2..48,70..112**).

- **offset**=real (optional, default=0.0) **offset** is a distance offset (Å) that can be subtracted from Δd_{ij}. **offset** values which are > 0.0 make the distance similarity measure Δd_{ij} less sensitive, whereas values < 0.0 make the measure more sensitive. Higher **offset** values lead in general to higher redundancies and accordingly lower I values. Lower **offset** values have the contrary effect.

Chapter 4

Structure-based drug design by NMR

4.0 Contributions

This project is a collaboration with Dr. Alvar Gossert[1] and co-workers (C. Henry and A. Widmer), who were responsible for preparing protein and ligand samples, NMR measurements, manual signal identification, and preparing manual protein and ligand chemical shift assignments. Adjustments regarding the FLYA protocol to enable transfer of chemical shifts have been performed by Dr. Elena Schmidt. Automated peak picking with CYPICK, automated chemical shift assignment with the 1- and 2-step FLYA assignment transfer protocol, and structure calculations on the complexes have been achieved within this work.

---

[1] Institute for BioMedical Research, Novartis, Basel.

4.1 Introduction

Structure-based drug design (SBDD) is a well-established strategy in the field of drug discovery (Anderson, 2003). High-resolution structures of protein-ligand complexes accelerate drug development in the pharmaceutical industry and medicinal chemistry. A simplified flowchart of the iterative process is presented in Fig. 4.1. The first step is cloning, purification and 3D structure determination of the protein (in the following called target) by either X-ray crystallography, NMR or homology modeling (Anderson, 2003). This is followed by a detailed analysis of potential binding sites and possible interactions within the 3D structure and virtual docking of compounds from a data base against the target. The compounds are then scored by their steric and electrostatic interactions. The best of these compounds are synthesized, analyzed and the structure of the 3D complex is determined. Analysis of the complex then reveals potential additional interactions in the binding pocket and sites on the compound that can be optimized. Additional cycles

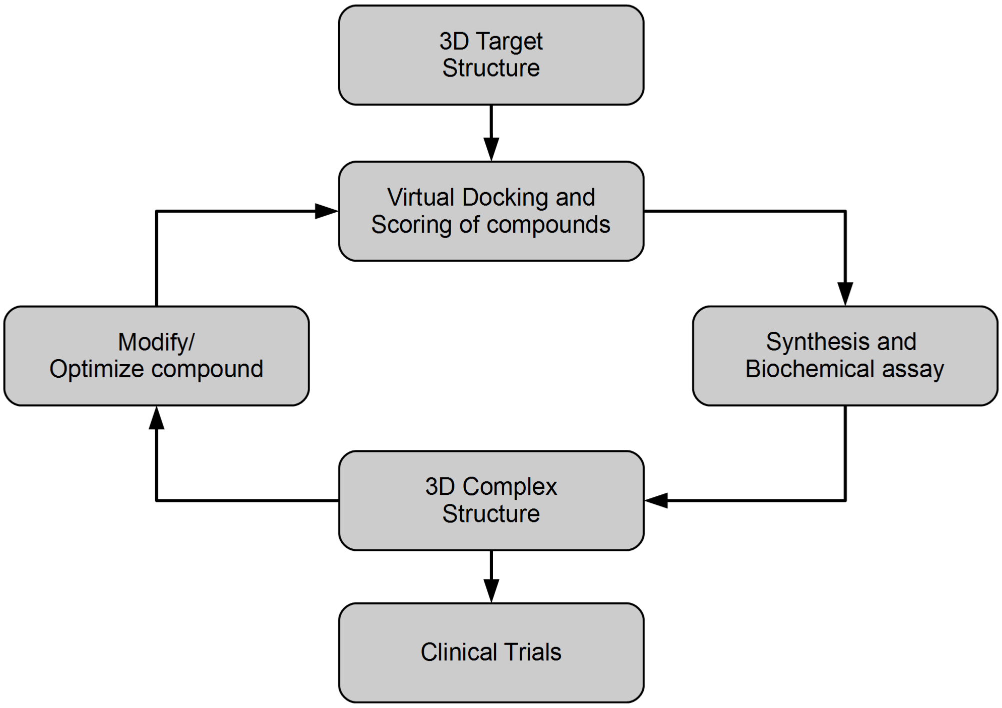

3D Target Structure
Virtual Docking and Scoring of compounds
Modify/ Optimize compound
Synthesis and Biochemical assay
3D Complex Structure
Clinical Trials

Figure 4.1: Simplified flowchart of the iterative process of structure-based drug design. The 3D target structure is determined and potential binding sites are analyzed. Compounds from a database are then docked against the target using computer programs. The compounds are ranked based on their interactions with the target. The best of these compounds are synthesized and analyzed by biochemical assays. This is followed by structure determination of the complex structure. Analysis of the complex structure reveals potential sides within the ligand for optimization. The optimized compound then reenters the described process. This iterative process can be repeated until a compound with desired affinity and specificity is achieved.

of synthesis of the optimized compound, structure determination of the new compound target complex, and further optimization can be performed until an initial weak ligand transforms into a potent drug-like molecule (Anderson, 2003).

At several stages in the SBDD process, visualized in Fig. 4.1, it is necessary to determine a 3D protein or protein-compound structure. This can in principle be achieved by NMR spectroscopy or X-ray crystallography. A contrasting juxtaposition of the advantages and disadvantages of the two methods has already been given in chapter 1.1. However, SBDD is mainly performed by X-ray crystallography. Reasons for this are its potential towards high-throughput, which is among other methodical aspects enabled by molecular replacement (Evans & McCoy, 2008) (MR). MR is a standard procedure for solving the phase problem and thereby solving structures by X-ray crystallography. As the name implies MR needs a 3D structure model. The structure model can for example be a mutant or a homologous structure of the target. In contrast to that, NMR is a low-throughput method (discussed in more detail below) for which an efficient MR approach is not available. Most recently a new method called $N\mathrm{MR}^2$ (Orts *et al.*, 2016), has been developed that aims at fulfilling the same task. This approach does not need a protein chemical shift assignment. A homologous protein structure is needed as starting structure in a CYANA structure calculation. Inter-molecular NOEs are of semi-ambiguous nature because only the chemical shift assignment of the ligand is known. Calculated structures which have the smallest violations are then used for protein-ligand complex representation.

The classical approach towards solving structures by NMR is to perform resonance assignment, NOE assignment, and structure calculation. New ligands change the chemical environment of the protein and induce chemical shift changes within the protein, especially in the binding pocket, with respect to the apo-form or a differently ligated form of the protein. Conformational changes of the protein that are induced by the ligand also affect the chemical shifts for the same reason. Accordingly, the time-consuming steps of chemical shift assignment and NOE assignment have to be performed more or less from scratch. Consequently, NMR has so far not become a true alternative to X-ray crystallography in high-throughput SBDD.

Nevertheless, X-ray crystallography has a few limitations. In some cases, it is not possible to obtain a protein crystal or a protein-ligand co-crystal at all. The success of co-crystallization depends on the binding affinity of the protein-ligand complex. The weaker a ligand binds, the lower the probability of achieving a co-crystal. Weak ligands ($K_D > 100$ μM) are common at the beginning of a SBDD project, especially in fragment-based drug design (Folkers *et al.*, 2006). Additionally, crystal contacts may hindered the accessibility for the ligand in the protein crystal. In modern drug discovery it is important to have an alternative method for determining structures of protein-ligand complexes on a molecular level.

In order to make NMR more accessible towards SBDD the following aspects have to be considered: (i) the amount of isotope labeled protein needed for one protein-ligand complex is rather large and should be reduced (20 mg for a 20 kDa protein), (ii) data collection time should be reduced ($>$ two weeks measurement time), and (iii) there is no MR method in NMR, i.e. it takes nearly the same amount of time to determine a structure of a protein with ligand-2 as it took to determine the structure of the same protein with ligand-1. All these aspects are addressed in the development of the SBDD by NMR protocol. The general strategy developed is explained in chapter 4.3.1.

Background of this work was to find novel inhibitors for the protein MDM4 (Wade & Li, 2013) by fragment-based screening. MDM4 (sometimes called MDMX), is an adaptor protein which specifically binds to the cancer suppressor protein p53 (Shvarts *et al.*, 1996). p53 is a human oncosuppressor gene and plays a central role in many cell functions. MDM4 regulates p53 activity. It has been observed that MDM4 is overexpressed in 10-20% of 800 very diverse tumors (Toledo & Wahl, 2006) leading to a massive reduction of p53 which ultimately allows these cells to proliferate in an uncontrolled manner. Those findings supported the idea that MDM4 could be an interesting target in anticancer strategies (Toledo & Wahl, 2007). It was possible to obtain co-crystal structures for several tight-binding p53-analogous peptides (Kallen *et al.*, 2009) and drugs like nutlin-3a (unpublished). However, for weak ligands, being present at the beginning of fragment-based drug design, no crystal structures could be obtained so far. Several computer-assisted methods are available that can be used to obtain structural models of a protein-ligand complex. However, these computer programs have their limitations and often fail to predict conformational changes within the protein that could be induced upon ligand binding. Using experimental data is therefore much more reliable than using computer predicted complex structures. Consequently, NMR was used to obtain complex structures of MDM4 with weak binding ligands. In order to make the process more efficient the SBDD by NMR strategy was developed within this project which is explained in more detail in the following. To investigate the general applicability of the process, ligands with diverse properties in terms of size ($200 \text{ Da} < \text{MW} < 1 \text{ kDa}$), affinity ($200 \text{ nM} < \text{K}_\text{D} < 1 \text{ mM}$) and binding kinetics ($1 \text{ s}^{-1} > \text{k}_\text{off} < 10^4 \text{ s}^{-1}$) were studied.

4.2 Materials and methods

4.2.1 Generation of reference assignment

The apo-form of MDM4 could not be used as reference assignment, because a lot of signals were missing in the recorded spectra. Therefore, MDM4 was stabilized with peptide-1 (Ac-Phe-Met-Aib-Pmp-Clw-Glu-Ac$_3$c-Leu-NH$_2$) (Kallen *et al.*, 2009). For MDM4 in complex with peptide-1 only a reference backbone assignment was available (Sanchez *et al.*, 2010). Hence, an automated chemical shift assignment by the standard FLYA protocol (Schmidt & Güntert, 2012) was performed. The automated achieved chemical shift assignment was manually checked an edited by Dr. Alvar Gossert and used in all subsequent calculations as reference. As input for FLYA a HNCACB (Wittekind & Müller, 1993), a HC^{ali}C^{aro}H-TOCSY (Kovacs & Gossert, 1987) and a Protein-Ligand NOESY (explained below) were used. Automated resonance assignment by FLYA was performed using tolerances of 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N for chemical shift matching and comparison with the manual refined reference resonance assignment. The reference assignment is only used to evaluate the quality of the results. The population size of the evolutionary algorithm was 200 in all cases. The chemical shift assignment was consolidated from 20 independent runs. Only assignments which could be reproduced in at least 80% of the 20 runs with an accuracy that deviated from the consensus values by less than the defined tolerances was classified as ‘strong’, otherwise ‘weak’. The side-chain terminal amide groups of arginine and lysine were excluded from the assignment calculations.

Within this study, aside from the MDM4-peptide-1 complex, we studied MDM4 in complex with peptide-2, nutlin-3a and the fragment, as well as the apo-form of MDM4.

4.2.2 NMR measurements

The Protein-Ligand NOESY (explained in chapter 4.3.2) spectra were recorded by Dr. Alvar Gossert on a Bruker AV800 spectrometer equipped with a TCI cryoprobe. Measurements were performed at 23°C. No folding of the indirect dimensions was applied and the ligand signals were made to appear at 100 ppm in the ^{13}C-dimension after referencing against DSS. The measurement time was 57 hours.

4.2.3 Peak picking

Manual peak lists were generated by semi-automatically picking the ^{15}N-HSQC, the constant-time ^{13}C-HSQC and the ^{13}C-detected HSQC spectra using CcpNmr Analysis (short CCPN) (Skinner *et al.*, 2016; Vranken *et al.*, 2005). 2D HSQC peak lists were used as root resonances for picking the Protein-Ligand NOESY with ATNOS (Herrmann *et al.*, 2002b). Peak lists were manually edited in Cara (Keller, 2004) to remove artifacts and typically about 30% additional signals were identified: mostly overlapping signals, signals with very low intensity and signals close to the diagonal.

Fully automated peak picking was performed with CYPICK (Würz & Güntert, 2016), by using default parameters (β=3.0 and γ=1.3, see chapter 2.2). The ^{15}N, aliphatic, and aromatic ^{13}C regions were picked separately. Solvent regions were excluded from the peak lists and Peakmatch (Buchner *et al.*, 2013) was used to optimally match NOESY peak lists to the HSQC peak lists. Tab. 4.1 gives an overview of the used picking parameters and the spectral regions that were selected for automated peak picking.

Table 4.1: Spectral regions that were picked from the Protein-Ligand NOESY by CYPICK. The regions are specified in ppm.

MDM4 in complex with	**Peptide-1**	**Peptide-2**	**Nutlin-3a**	**Fragment**	**Apo-form**
^{13}**C**$^{\text{aliphatic}}$	10.8-72.4	10.0-71.0	11.0-71.5	10.2-72.9	10.0-71.0
^{1}**H**$^{\text{HSQC}}$	0.4-5.6	-0.4-5.6	-0.4-5.7	-0.3-5.6	-0.3-5.6
^{13}**C**$^{\text{aromatic}}$	112.2-142.2	110.0-140.0	112.7-142.6	112.7-139.2	112.8-139.4
^{1}**H**$^{\text{HSQC}}$	5.1-7.8	5.0-7.8	5.0-7.8	5.0-7.8	6.2-7.7
^{15}**N**	104.1-127.3	103.4-126.9	103.0-126.1	102.3-126.8	103.0-127.0
^{1}**H**$^{\text{HSQC}}$	6.2-10.1	6.0-10.1	6.4-4.9	6.4-10.1	6.5-10.0

^{13}C$^{\text{aliphatic}}$ and ^{13}C$^{\text{aromatic}}$ regions were picked with **scale=0.05,0.1,1.0**, ^{15}N spectral regions were picked with **scale=0.06,0.1,0.6**. The complete ^{1}H$^{\text{NOE}}$ dimension was picked. In all cases only signals with positive sign were analyzed.

The ligand plane was always picked manually due to the existence of many artifacts. Ligand planes of the Protein-Ligand NOESY are visualized in appendix C Figs. C.1-C.3. All NOESY peaks belonging to one MDM4 complex are picked from the same Protein-Ligand NOESY spectrum. The peaks are separated into two peak lists, i.e. a ^{13}C- and a ^{15}N-NOESY peak list for chemical shift and NOE assignment.

4.2.3.1 Peak list comparison

Peak list comparison was performed with the CYANA command **peaks compare**, explained in chapter 2.3.2. The manual picked peak lists served as a reference in the calculation. Default parameters for **width**, **distcut**, and **artifactweight** were used. The tolerance for peak matching was set to 0.03 in case of ^{1}H and 0.4 in case of ^{13}C and ^{15}N.

4.2.4 FLYA assignment transfer

Automatic chemical shift assignment of the reference complex was performed with the standard FLYA protocol (Schmidt & Güntert, 2012) as explained in chapter 4.2.1. A simplified schematic of the FLYA algorithm is visualized in Fig. 4.2. FLYA generates a network of expected peaks from the magnetization transfer pathways of the corresponding experiment and the protein sequence. The position of the expected peaks is known approximately e.g. from the Biological Magnetic Resonance Bank (BMRB) (Ulrich *et al.*, 2008) statistic mean, $f(a)$, and standard deviation values, $\sigma(a)$, of the atom a that corresponds to the expected peak. The assignment process is reduced to finding a mapping between the expected peaks and the observed peaks, provided in the form of a peak list (see Fig. 4.2). The mapping procedure is scored and optimized by an evolutionary algorithm combined with a local optimization. The algorithm is explained in detail in (Schmidt & Güntert, 2012).

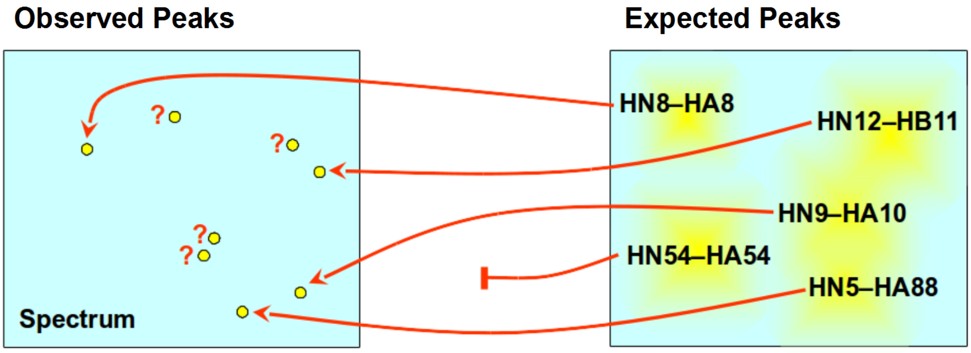


Figure 4.2: Assignment strategy of the FLYA algorithm. A network of expected peaks is deduced from the protein sequence and the underlying experimental specifications. The expected peaks, whose position is only known approximately from the BMRB statistics for example, are then mapped to the observed peaks. The mapping procedure is scored and optimized by an evolutionary algorithm combined with a local optimization. Figure is taken from (López-Méndez & Güntert, 2006).

In the developed assignment transfer protocol by FLYA the mean atom-specific chemical shift value, $f(a)$, usually taken from the BMRB, is replaced by the resonance assignment of the reference structure which is in our application the assignment of MDM4 in complex with peptide-1. The above introduced standard deviation $\sigma(a)$ that defines the chemical shift range is taken from the corresponding atom-specific BMRB statistic. However, we adjusted $\sigma(a)$ by multiplying it with a factor $\alpha \leq 1$.

In the 1-step FLYA assignment transfer the chemical shift range of expected peaks is determined from $f(a)$ and $\alpha \cdot \sigma(a)$, where $f(a)$ equals the chemical shift assignment of the reference complex, $\sigma(a)$ refers to the atom-specific BMRB standard deviation, and $\alpha = 0.5$. The value for α was found empirically as explained below. In the 1-step protocol expected NOESY peaks are generated on the basis of random structures, i.e. connectivity patterns are only obtained for short-range NOEs.

We also established a two-step assignment transfer protocol by FLYA (referred to as 2-step protocol). In the first step of this assignment protocol expected peaks are generated from an available reference structure. In our application a structure of MDM4 in complex with peptide-1. The expected NOESY peaks are determined for ^{1}H-^{1}H distances up to 6 Å that could be observed in all individual structures from the structure bundle. The probabilities of expected NOESY peaks were set to 0.9 to 0.5 in steps of 0.1 for ^{1}H-^{1}H distances of 4.0-6.0 Å in steps of 0.5 Å, respectively. The chemical shift range of the corresponding atoms was determined by $f(a)$ and $\alpha \cdot \sigma(a)$, where $f(a)$ equals the chemical shift assignment of the reference complex, $\sigma(a)$ refers to the atom-specific BMRB standard deviation, and $\alpha = 0.1$. Assignments that were classified as ‘strong’ in the first step are kept fixed in the second step. The second step is otherwise equal to the 1-step protocol, i.e. $f(a)$ equals the chemical shift assignment of the reference complex, $\sigma(a)$ refers to the atom-specific BMRB standard deviation, and $\alpha = 0.5$. Expected peaks are generated on the basis of a random structures. The 2-step protocol aims at identifying atom assignments that do not differ significantly from the reference assignment in the first step and keeping them fixed in the second step. Whereas, the second step strives for finding assignments that change considerable with respect to the reference assignment.

As input for the 1- and 2-step FLYA assignment transfer protocol we used ^{13}C- and ^{15}N-NOESY and -HSQC peak lists, the protein sequence, the reference chemical shift assignment, and in case of the two step protocol the reference structure. In both protocols, tolerances for chemical shift matching and comparison with a reference assignment, if at hand, were set to 0.03 ppm in case of ^{1}H and 0.4 ppm in case of ^{13}C and ^{15}N. The reference assignment is only used to evaluate the quality of the results. The population size of the evolutionary algorithm was 200 in all cases. Chemical shift assignments were consolidated from 20 independent runs. Only assignments which could be reproduced in at least 80% of the 20 runs with an accuracy that deviated from the consensus values by less than the defined tolerances was classified as 'strong', otherwise 'weak'. The side-chain terminal amide groups of arginine and lysine were excluded from the assignment calculations. Examples and explanations of the exact usage of the FLYA assignment transfer protocol are given in chapter 4.5.

4.2.5 Generation of CYANA library entries

For non-standard amino acids and organic compounds, library files for CYANA (Yilmaz & Güntert, 2015) being consistent with the covalent geometry of the AMBER force field (Cornell *et al.*, 1995) were generated with the program WIT!P, using PDB files as input.

4.2.6 CYANA calculations

Structure calculations were performed by CYANA (Güntert *et al.*, 1997; Güntert & Buchner, 2015) using the recently introduced consensus structure bundle method (Buchner & Güntert, 2015a). Based on twenty individual structure calculations a consensus distance restraint set is generated, which is used for a final consensus structure calculation. The twenty conformers of the final consensus structure calculation with the lowest CYANA target function were selected for representation of the consensus NMR structure bundle. The input consisted of a sequence file, containing protein and ligand sequence, WIT!P library files were appended to the standard library, and if needed, restraints for correct closure of aliphatic rings were generated with the WIT!P CYANA macro and included in the calculation. The ^{13}C- and ^{15}N-NOESY peak lists that were also used for generating resonance assignments by FLYA, were used in structure calculation. Additionally intra-molecular and inter-molecular peak lists were manually prepared for the ligand and used in the calculation. Intra-molecular ligand and inter-molecular protein-ligand peaks can be distinguished from one another by the presence and absence of diagonal-symmetric peaks, respectively (explained below). Inter-molecular peaks from small ligands often do not obtain enough weight in the network anchoring procedure of automated NOE assignment and sometimes are treated as artifacts by the algorithm. The automated NOE assignment strategy of CYANA is explained in detail in chapter 1.2.3. Therefore, a set of clearly visible inter-molecular NOE cross peaks (visualized in appendix C Figs. C.1-C.3) were defined as true signals, i.e. they have to be assigned by the algorithm and fulfilled in structure calculation (CYANA command **assign_noartifact**). In this way, there is no user bias, that is, no assignment is explicitly defined for these signals by the user, but the algorithm is not allowed to ignore the signals.

The quality of structure calculation was exclusively evaluated on the basis of RMSD bias (Güntert *et al.*, 1998). Structures were first superimposed within their ordered regions which were either determined by CYRANGE (Kirchner & Güntert, 2011). Then the average structure is obtained by averaging the coordinates of the atoms in the superimposed conformers in the structure bundle. The backbone RMSD between the average given structure and the reference mean structure yields the RMSD bias. The precision of calculated structure bundles is expressed by the RMSD radius (Güntert *et al.*, 1998), i.e. the average RMSD between individual conformers and the mean coordinates of the structure.

4.3 Results and discussion

4.3.1 Strategy for enabling SBDD

The high-throughput strategy developed for solving structures of protein-ligand complexes starts with carefully determining reference assignments of the protein by the classical approach. The chemical shift information is then used to guide the automated assignment of all subsequent protein-ligand complexes. Reference assignments can be established on the apo-protein or on a well-behaved complex with a ligand. For subsequent protein-ligand complexes a single sample is prepared and only one Protein-Ligand NOESY (see chapter

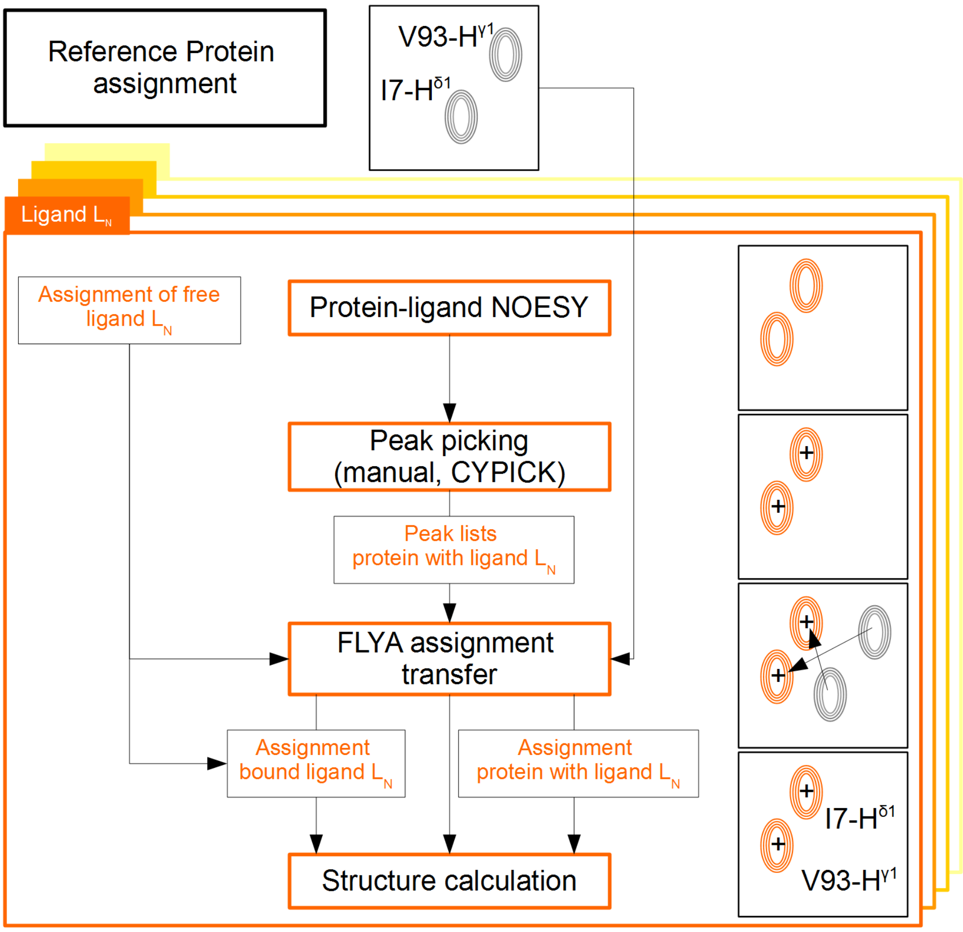

Reference Protein assignment
V93-Hγ1
I7-Hδ1
Ligand L_N
Assignment of free ligand L_N
Protein-ligand NOESY
Peak picking (manual, CYPICK)
Peak lists protein with ligand L_N
FLYA assignment transfer
Assignment bound ligand L_N
Assignment protein with ligand L_N
Structure calculation
+
I7-Hδ1
V93-Hγ1

Figure 4.3: Strategy for enabling SBDD by NMR. The serial process of determining complex structures of a target protein and several ligands (orange boxes) is depicted. The reference protein assignment, which is usually obtained manually, is used in the assignment process of each individual complex. First, the Protein-Ligand NOESY is recorded. Second, peaks are picked manually or automatically by CYPICK for example. Third, the peak lists are used as input in the FLYA assignment transfer protocol together with the reference protein assignment, ligand peaks can optionally also be included. The FLYA assignment transfer protocol produces protein assignment and ligand assignments of peptide ligands. Fourth, these assignments can be used in structure calculation together with the protein and ligand NOESY peak lists using CYANA. The figure was adapted from (Gossert *et al.*, 2016).

4.3.2) spectrum is recorded. Assignments are automatically obtained by the FLYA assignment transfer using the reference assignment as additional input. Structure-based drug design by NMR was enabled by two main developments: The 'Protein-Ligand NOESY' explained in chapter 4.3.2, a more efficient way of recording all necessary NOESY spectra in one single spectrum, and the 'FLYA assignment transfer' explained in chapters 4.2.4 and 4.3.3, an automated way of assigning spectra of the target protein in complex with different ligands based only on NOESY data from the Protein-Ligand NOESY spectrum.

4.3.2 The Protein-Ligand NOESY

The Protein-Ligand NOESY was exclusively developed by Dr. Alvar Gossert and coworkers. Main principles are visualized in Fig. 4.4. The Protein-Ligand NOESY is a combined 3D HSQC-NOESY, which includes all signals necessary for the structure determination of a uniformly ^{13}C and ^{15}N labeled protein and an unlabeled ligand. Usually this task would need the recording of five NOESY spectra, i.e. (i) ^{15}N-edited 3D NOESY, (ii) ^{13}C$^{\text{aliphatic}}$-edited 3D NOESY, (iii) ^{13}C$^{\text{aromatic}}$-edited 3D NOESY ((i)-(iii) are needed to obtain intra-molecular protein NOEs), (iv) ^{15}N/^{13}C-filtered, ^{15}N/^{13}C-filtered NOESY (for intra-molecular ligand NOEs), (v) ^{15}N/^{13}C-edited, ^{15}N/^{13}C-filtered NOESY (for inter-molecular protein and ligand NOEs). ^{15}N-edited ^{1}H signals, aliphatic and aromatic ^{13}C-edited ^{1}H signals are integrated in one NOESY experiment as explained in (Boelens *et al.*, 1994). The inter-molecular and intra-molecular signals of the ligand are recorded by using a modified time-proportional phase incrementation (TPPI) procedure on the unlabeled ligand signals (States *et al.*, 1982; Marion & Wüthrich, 1983; Bodenhausen *et al.*, 1983; Keeler & Neuhaus, 1985). TPPI allows the unlabeled signals to appear at an arbitrarily chosen position in the combined ^{15}N/^{13}C dimension. It is advisable to select a region for the ligand plane where no natural protein signals occur in order to reduce ambiguity. In our application the ligand signals were shifted to an artificial ^{13}C shift of 100 ppm, which corresponds to an ^{15}N shift of 126 ppm.

The Protein-Ligand NOESY is an HSQC-NOESY and not a standard NOESY-HSQC spectrum (Mishra *et al.*, 2014). In a 3D NOESY experiment, one has two indirect dimensions with lower resolution and one direct dimension with higher resolution. Two nuclei give rise to an NOE, that is a sending and a receiving nucleus. The sending nucleus can be described by two chemical shifts (^{13}C and ^{1}H$^{\text{HSQC}}$). The receiving nucleus however can only be described by one chemical shift (^{1}H$^{\text{NOE}}$). Therefore, the ^{1}H$^{\text{NOE}}$ dimension was chosen to be recorded with the highest possible resolution. The sending nucleus obtains a lower resolution which in turn can be compensated by the second chemical shift.

Technical details of this procedure will be explained in a future publication by Dr. Alvar Gossert and co-workers. The full parameter set, including pulse program and detailed description of how to set up the experiment has been deposited in the Bruker user library (`https://www.bruker.com/service/information-communication/nmr-pulse-program\discretionary{-}{}{}lib/bruker-user-library.html`).

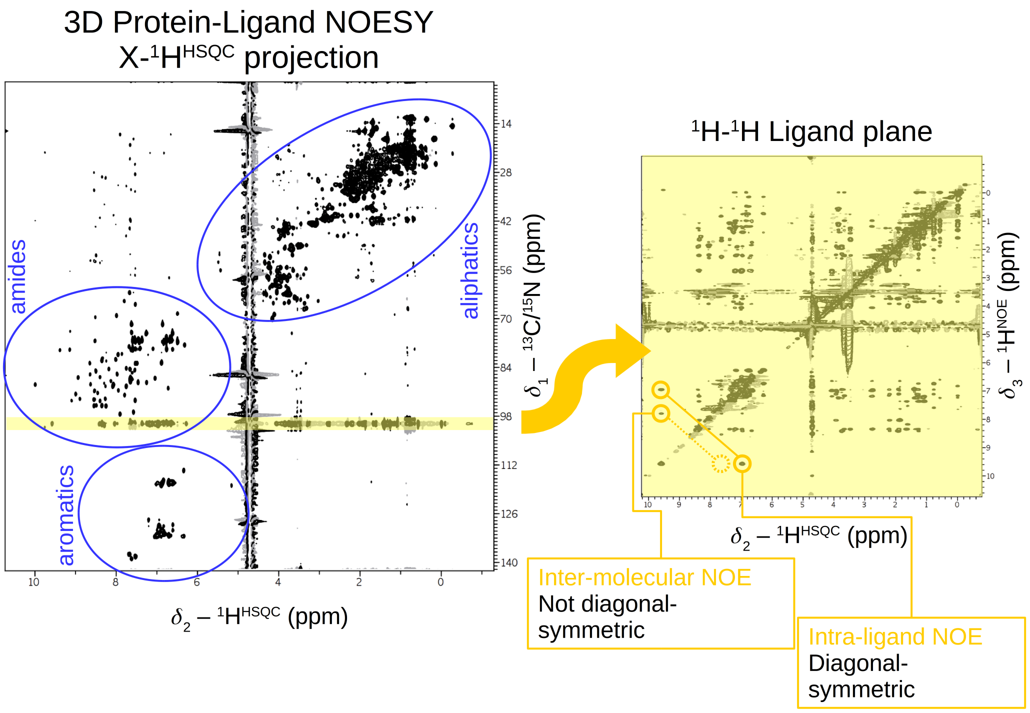


Figure 4.4: Schematic representation of the Protein-Ligand NOESY. On the left, the X-^{1}H HSQC projection is visualized, specifying the aromatic and aliphatic ^{13}C region and the ^{15}N region. The ligand plane is indicated by the yellow rectangle and the ^{1}H-^{1}H ligand plane is presented on the right side. Inter-molecular and intra-ligand examples are depicted.

4.3.3 FLYA assignment transfer

A detailed explanation of the strategy for enabling the automated chemical shift assignment transfer by FLYA is given in chapter 4.2.4. As noted above, the FLYA assignment transfer uses as input the protein sequence, the Protein-Ligand NOESY peak lists together with the reference chemical shifts. In general, resonance assignments are less reliable when only deduced from NOESY spectra (Ikeya *et al.*, 2011; Schmidt & Güntert, 2013a). Therefore, the loss in information can be compensated by adding chemical shift information from a reference complex or the apo-form of the protein. We observed that the complex chemical shifts did not change dramatically with varying ligands in case of MDM4. Approximately, 50% of the resonances remained at the same position while others did not change by more than 1 ppm in the ^{1}H dimension. The reference chemical shift assignment replaces the BMRB statistic in the algorithm which is used to estimate expected peak positions, $f(a)$.

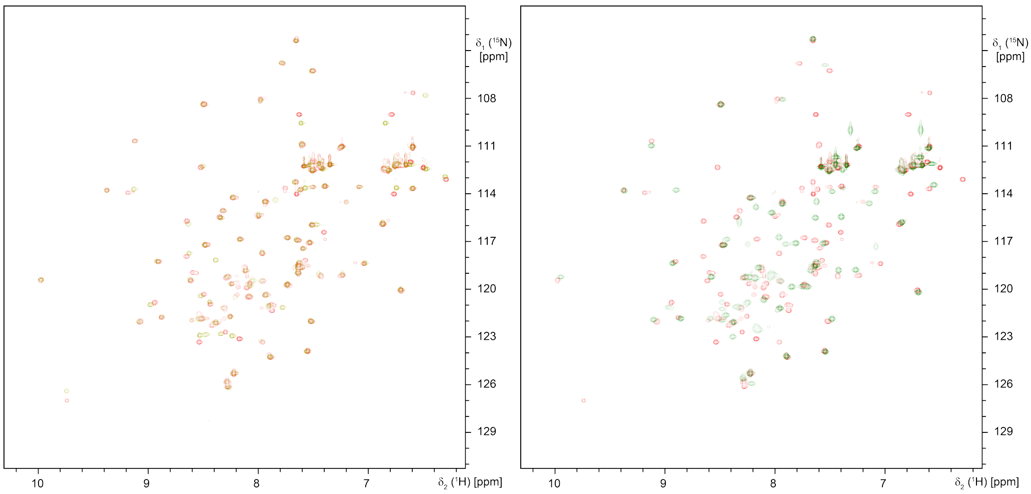
As explained above, we developed a 1-step and a 2-step assignment transfer protocol. In both protocols the chemical shift range for expected peaks is determined from the standard deviation of the BMRB, $\sigma(a)$, belonging to the corresponding atom a and the provided reference assignment, $f(a)$. In order to achieve the best possible assignment of the new complex we had to optimize the allowed chemical shift range of an expected peak, i.e. the multiplication factor α for $\sigma(a)$. We therefore extensively tested values between 0.05 and 1.0 for α. Best results could be achieved with $0.5 \cdot \sigma(a)$. Higher values made the obtained assignment too independent from the reference chemical shift assignment, whereas lower values did not allow sufficient deviations in the chemical shift assignment, which can for example be attributed to conformational changes upon varying ligands. In the first step of the 2-step protocol we used $\alpha = 0.1$ because we wanted to reduce the chemical shift range as far as possible and obtain only assignment that did not change significantly with respect to the reference assignment.

4.3.4 Application to a diverse set of protein-ligand complexes

In order to demonstrate the range of application of the SBDD by NMR method a diverse set of ligands was chosen: MDM4 in complex with peptide-1 was used as the reference in the FLYA assignment transfer protocol (Kallen *et al.*, 2009). The complexes consisted of MDM4 with: peptide-2, a ligand that is chemically very similar to peptide-1; nutlin-3a, a drug-like compound; a fragment that is typical for the beginning of a SBDD project; and the apo-form which exhibits considerably more flexibility.

The diversity of this test set is expressed in terms of affinity and chemical diversity. Some of the ligands are tight binding compounds, i.e. peptide-2 and nutlin-3a with $K_D <$ 1 μM, and a weak binding small compound, i.e. the fragment with $K_D >$ 100 μM. Tight binding compounds bind in the slow exchange regime which implies that two distinct free and bound states of protein and ligand can be observed on the spectrum. Weak binding compounds bind in the fast exchange regime, meaning that an average between free and bound state can be observed. The SBDD by NMR method needs to work in both cases and should be able to cope with complications arising from both cases. The chemical diversity is expressed by choosing ligands that are similar to the reference ligand and others that

(a) peptide-1 vs. peptide-2 **(b)** peptide-1 vs. nutline-3a



(c) peptide-1 vs. fragment **(d)** peptide-1 vs. apo form

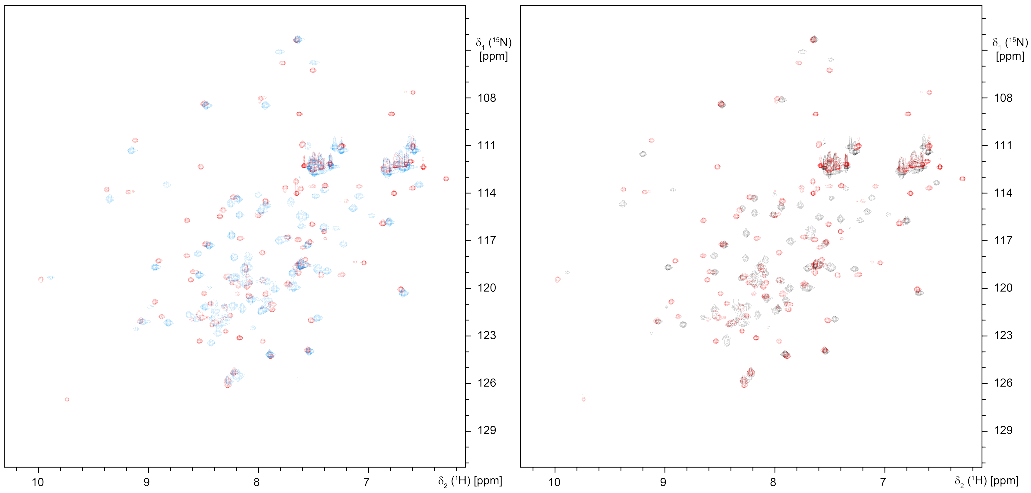


Figure 4.5: Demonstration of chemical shift changes that are induced by chemically equal or diverse ligands with respect to the reference complex, using the example of the ^{15}N-HSQC: **(a)** peptide-1 and peptide-2 are chemically very similar and accordingly the chemical shift changes induced are very small; **(b)**,**(c)** nutlin-3a and the fragment are chemically rather diverse with respect to peptide-1, as a result chemical shift changes induced are rather large; **(d)** the apo-form of MDM4 is rather unstable and accordingly many chemical shifts are missing while others are shifted with respect to peptide-1. Spectral overlays were kindly provided by Dr. Alvar Gossert.

vary extremely. The consequence of the chemical diversity are the chemical shift changes that are induced on the NMR spectrum with respect to the reference complex. Peptide-2 induces only small changes, whereas the other ligands and the apo-state lead to rather severe changes in the spectrum. The changes induced are visualized in Fig. 4.5, using the example of a ^{15}N-HSQC spectrum.

4.3.4.1 Automated peak picking of the Protein-Ligand NOESY with CYPICK

In order to reduce spectral analysis time we picked the Protein-Ligand NOESY automatically using CYPICK (Würz & Güntert, 2016) (explained in detail in chapter 2), and compared automated assignment and structure calculation results in-between manual established and automatically picked CYPICK results. The main focus was laid on judging whether is is useful to thoroughly prepare manual peak lists or use automatically determined peak lists for high-throughput analysis.

Automated CYPICK peak lists are compared to manual peak lists by means of find, artifact, and overall scores (explained in chapter 2.3.2) summarized in Tab. 4.2. Manual peak lists served as reference peak list.

Table 4.2: Comparison of CYPICK scores with respect to manual peak lists

MDM4 in complex with	^{13}C-NOESY scores (%)			^{15}N-NOESY scores (%)		
	Find	Artifact	Overall	Find	Artifact	Overall
Peptide-1	80.3	47.4	65.8	81.1	38.6	70.9
Peptide-2	76.9	44.2	64.7	75.0	34.5	67.1
Nutlin-3a	68.4	29.8	62.6	71.5	35.2	63.8
Fragment	60.7	27.5	56.1	66.2	17.6	63.4
Apo-form	39.4	36.7	34.9	41.7	53.9	31.9

Excluding peak picking results achieved on the Protein-Ligand NOESY of the apo-form, CYPICK peak lists obtained overall scores in the range of 56–71%. The spectrum of the apo-form is picked with significantly lower overall scores in the range of 32–35%. One reason therefore might be that the apo-form behaves more flexible than the ligand-bound forms of the protein. The flexibility may lead to a loss in signals or signals with severely lower sensitivity that can only be obtained by manual inspection of the spectrum. Find scores of peptide-1 are highest with > 80% what can be explained by the high quality of the Protein-Ligand NOESY for this complex, which is the main reason it was chosen to serve as reference complex in this study.

4.3.4.2 FLYA assignment transfer of MDM4

The FLYA assignment transfer step is the crucial step in the process of SBDD. Therefore, samples were optimized to create the best possible experimental conditions. In all samples, ligand was added in great excess to saturate MDM4. For all samples only one set of protein signals could be observed, where the fragment, binding in fast exchange, yielded a single set of averaged signals.

The ligand can either be assigned manually or automatically by FLYA, if the ligand is peptide-based, providing the assignment of the free ligand and peak lists of the bound ligand from the Protein-Ligand NOESY. The protein and ligand resonance assignment can then be used in structure calculation. In our study ligand assignments were exclusively obtained manually.

Results achieved by the 1- and 2-step FLYA assignment transfer for the individual complexes are summarized in Tab. 4.3. In general, assignments obtained from the 2-step FLYA assignment transfer are more accurate in terms of percent correct assignments and absolute number of strong assignment compared to the 1-step protocol assignments.

Assignments from peptide-1 were achieved by a standard FLYA protocol since these shifts were used for transfer to other complexes. Peptide-1 was assigned with an accuracy of 88-89% by the standard FLYA protocol. Comparison was performed with respect to a manual assignment with an overall completeness of 75.3%. CYPICK and manual peak lists achieved very similar values in 1-step and 2-step FLYA assignment protocol, with respect to percentage correct assignments and number of strong assignments.

Peptide-2 was assigned with an accuracy of 85% and 91–95% with the 1-step and 2-step protocol, respectively, using a manual assignment. In case of peptide-2 the usage of a known structure had the highest impact within this data set.

For nutlin-3a the reference assignment included only N and HN shifts, accordingly the percentage of correct assignments corresponds only to these atoms. Using the 1-step protocol 82–86% of these shifts were assigned. Chemical shift assignments obtained in the 1-step protocol by manual peak lists and CYPICK peak lists showed an agreement of more than 80%.

For the fragment assignments achieved by the 1-step and 2-step protocol with manual and CYPICK peak lists showed similar assignment correctness with respect to a manual chemical shift assignment (75% completeness). Approximately 78% were correctly assigned when using the 1-step protocol, whereas approximately 83% were correctly assigned when using the 2-step protocol. Assignments achieved on the basis of manual and CYPICK peak lists also showed a high percentage of consistency, which is reflected in a resonance assignment correspondence of $\sim$ 90% in case of the 1-step protocol.

Summarized, assignments achieved on the basis of manual and automatically obtained peak lists by CYPICK lead to protein resonance assignments that show a high degree of correspondence and correctness with respect to the reference chemical shift assignment. In general, the knowledge of a structure improves the correctness of resonance assign-

Table 4.3: Results of the FLYA assignment transfer

MDM4 in complex with	Manual peak lists precentage of correct assignments (number of strong assignments)		CYPICK peak lists precentage of correct assignments (number of strong assignments)	
	1-step	2-step	1-step	2-step
Peptide-1[a]	88.5 (914)	-	87.5 (920)	-
Peptide-2[b]	84.5 (908)	95.0 (1032)	84.7 (881)	91.4 (990)
Nutlin-3a[c]	82.2 (895)	90.7 (976)	85.6 (846)	90.7 (965)
Fragment[d]	78.6 (904)	83.1 (977)	77.0 (882)	82.3 (962)
Apo-form[e]	52.9 (785)	57.3 (918)	50.0 (663)	57.3 (902)

[a] reference assignment is a manual assignment of MDM4 in complex with peptide-1 (75.3% completeness); peptide-1 was assigned with a standard FLYA protocol

[b] reference assignment is a manual assignment of MDM4 in complex with peptide-2 (75.1% completeness)

[c] reference assignment includes only backbone N and HN shifts (9.7% completeness)

[d] reference assignment is a manual assignment of MDM4 in complex with fragment (75.0% completeness)

[e] reference assignment includes only backbone N,HN,C$^{\alpha}$, and C$^{\beta}$ shifts (23.3% completeness)

ments significantly in case of peptide-2 (improvement of $\sim$ 10%), which exhibits a very similar complex structure as peptide-1, and considerable improvements in case of the fragment and nutlin-3a (improvement of $\sim$ 5%), where the complex structures deviate with a higher degree from the reference complex structure. The actual correctness of the nutlin-3a assignment is difficult to predict since no complete reference assignment is available. Nutlin-3a showed the lowest correspondence in chemical shift assignment from manual and CYPICK peak lists.

4.3.4.3 Structure calculation results

Chemical shift assignments obtained from the FLYA assignment transfer were then used in structure calculation together with peak lists from the Protein-Ligand NOESY. Structure calculation statistics of MDM4 complexes determined by the SBDD by NMR protocol are summarized in Tab. 4.4. The information content, I, (see chapter 3) of the distance restraints files that were used for the final consensus structure calculation is summarized in Tab. 4.5. The fragment is a classified compound, therefore structure calculations on the MDM4-fragment complex could not be performed within this work.

Structures achieved from manual peak lists, with the exception of the apo-form, can be characterized by a high precision which is reflected in RMSD radius values < 1.2 Å. MDM4 in complex with peptide-1 structures were calculated from the manual reference assignment using either manual peak lists or CYPICK peak lists. Structure bundles from the consensus structure bundle calculation are presented in Fig. 4.6 **(a)** and **(b)**. The calculations from manual and CYPICK results show similar results in terms of RMSD bias and radius. The information contents of the distance restraint data sets were also very similar.

Table 4.4: Structure calculation results of the complexes in terms of RMSD bias and RMSD radius (Å). Structure calculation has been performed with CYANA as explained.

MDM4 in complex with	Manual peak lists		CYPICK peak lists	
	1-step RMSD bias (Å) (RMSD radius)	2-step RMSD bias (Å) (RMSD radius)	1-step RMSD bias (Å) (RMSD radius)	2-step RMSD bias (Å) (RMSD radius)
Peptide-1[a]	1.58 (1.02)	-	1.67 (0.86)	-
Peptide-2[b]	1.52 (1.15)	1.78 (0.82)	1.21 (1.01)	1.28 (1.00)
Nutlin-3a[c]	1.72 (0.91)	1.89 (0.80)	3.30 (1.74)	2.48 (0.80)
Apo-form[d]	3.39 (1.81)	4.47 (2.18)	6.02 (7.74)	5.40 (2.66)

[a,b] RMSD bias is calculated with respect to 3FEA residue ranges: 28–102; RMSD radius residue ranges: 28–102 and 133–140;

[a]Structure calculation was performed on the basis of a reference chemical shift assignment and the usage of either manual peak lists or CYPICK peak lists.

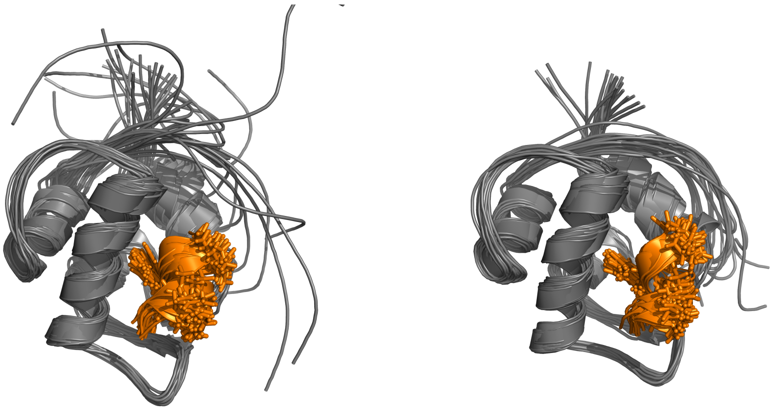
[c] RMSD bias is calculated with respect to 4HG7 residue ranges: 26–108; RMSD radius residue ranges: 26–108 and 130.

[d] RMSD bias is calculated with respect to 4HG7 residue ranges: 28–102; RMSD radius residue ranges: 28–102.

Table 4.5: Information content of the distance restraints achieved from the consensus structure calculation.

MDM4 in complex with	Manual peak lists		CYPICK peak lists	
	1-step I (I_r)	2-step I (I_r)	1-step I (I_r)	2-step I (I_r)
Peptide-1	274.94 (4.58)	-	274.84 (4.58)	-
Peptide-2	204.72 (3.42)	231.22 (3.85)	228.11 (3.80)	233.94 (3.90)
Nutlin-3a	248.43 (4.14)	276.61 (4.61)	169.22 (2.82)	253.18 (4.22)
Apo-form	99.95 (1.67)	116.04 (1.93)	64.4 (1.07)	91.49 (1.52)

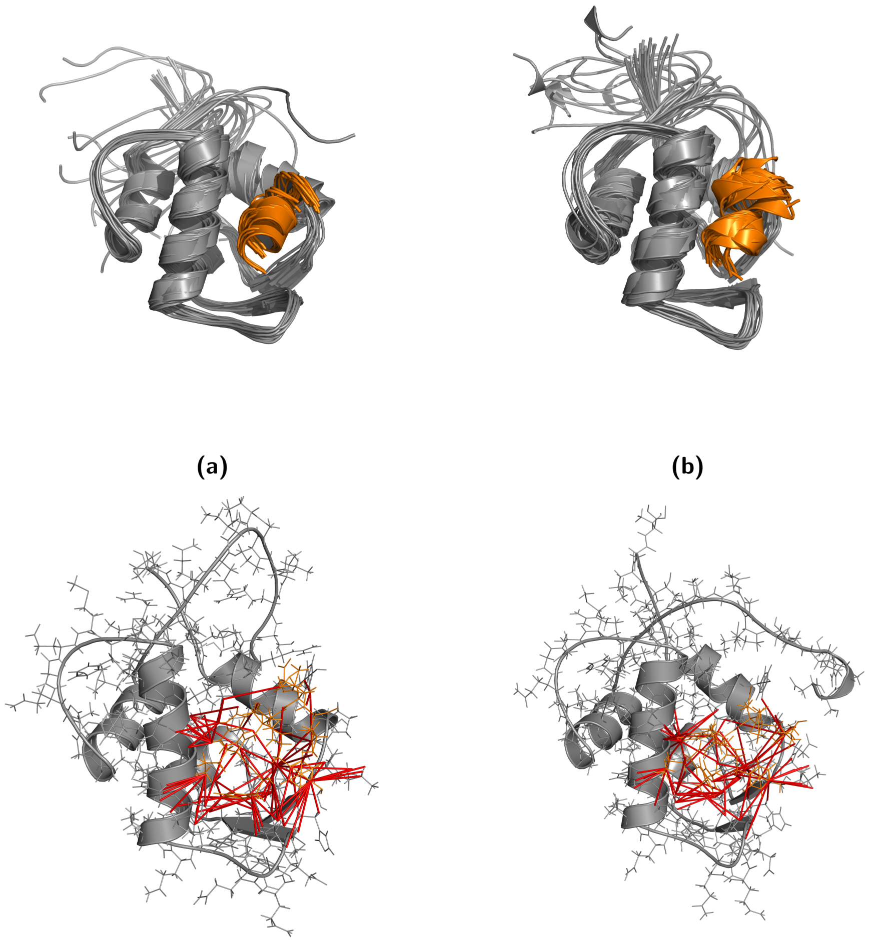
Information content calculation was performed with R_G =14.3 Å.



(a) **(b)**

Figure 4.6: Visualization of MDM4 in complex with peptide-1 structure bundles: **(a)** calculated from the manual chemical shift assignment and manual NOESY peak lists; **(b)** calculated from the manual chemical shift assignment and CYPICK peak lists. MDM4 is visualized in gray and peptide-1 in orange.

Structure calculations of MDM4 in complex with peptide-2 were characterized by a high precision of approximately 1.00 Å. The RMSD bias of these calculations with respect to the reference X-ray structure (PDB code 3FEA) is $<$ 1.80 Å. Structure bundles of



(a) **(b)**

(c) **(d)**

Figure 4.7: Structure bundles of MDM4 in complex with peptide-2: **(a)** calculated from an automated assignment by 2-step protocol and manual peak lists; **(b)** calculated from an automated assignment by 2-step protocol and manual peak lists and CYICK; **(c)** inter-molecular distance restraints of the structure presented in **(a)**; **(d)** inter-molecular distance restraints of the structure presented in **(b)**. MDM4 is visualized in gray, peptide-2 in orange, and distance restraints are depicted in red.

MDM4 in complex with peptide-2 are visualized in Fig. 4.7 **(a)** and **(b)**, using the 2-step assignment protocol in both cases and manual and CYPICK peak lists, respectively. The consensus structure bundle was selected for representation. The assigned inter-molecular distance restraints are depicted in Fig. 4.7 **(c)** and **(d)**, for a structure calculation from the 2-step assignment protocol using manual and CYPICK peak lists, respectively. The information content values of the distance restraint files are also in similar ranges, however slightly lower as those from the MDM4-peptide-1 complex. However, I_r is > 3.0, which was presented before as a threshold value for obtaining a reliable structure bundle. Differences in the RMSD radius values can be explained by the differences in the I values. When

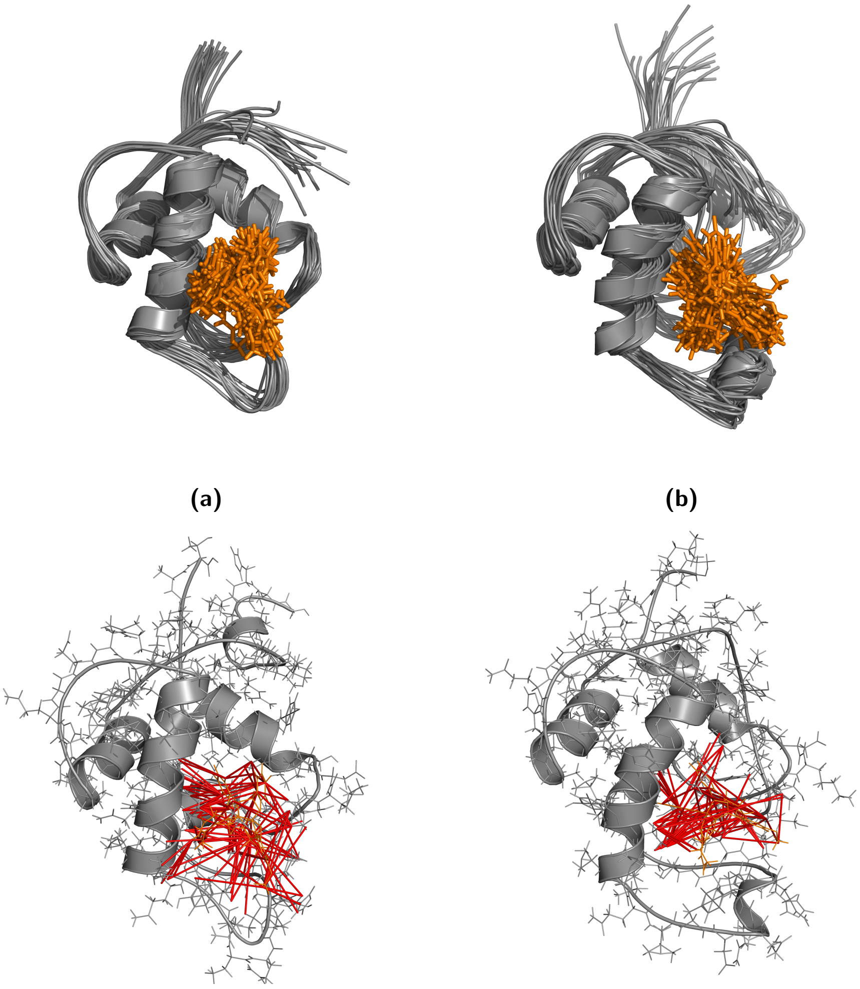

(a) **(b)** **(c)** **(d)**

Figure 4.8: Structure bundles of MDM4 in complex with nutlin-3a: **(a)** calculated from an automated assignment by 2-step protocol and manual peak lists; **(b)** calculated from an automated assignment by 2-step protocol and manual peak lists and CYICK; **(c)** inter-molecular distance restraints of the structure presented in **(a)**; **(d)** inter-molecular distance restraints of the structure presented in **(b)**. MDM4 is visualized in gray, nutlin-3a in orange, and distance restraints are depicted in red.

comparing the structure bundles depicted in Fig. 4.7 **(a)** and **(b)**, the difference in precision of the ligand can be noticed. The position of the ligand is more precise when calculating from manual peak lists compared to CYPICK peak lists. Peptide-1 is also less buried in the binding pocket when using CYPICK peak lists. Also the network of inter-molecular distance restraints is more dense when using manual peak lists compared to CYPICK peak lists. CYPICK peak lists probably lack too many real long-range signals which affects the position of peptide-1. The chemical shift assignments obtained before were very similar for the two cases and can consequently not be responsible for the differences in the resulting complex structures.

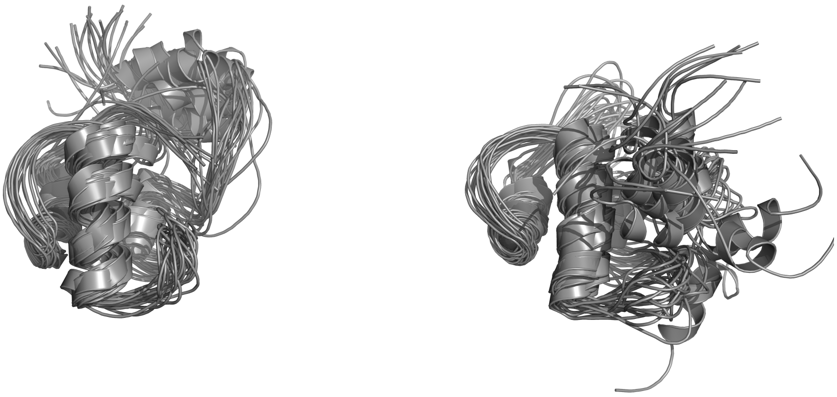


Figure 4.9: Structure bundles of the MDM4 apo-form: **(a)** calculated from an automated assignment by 1-step protocol and manual peak lists; **(b)** calculated from an automated assignment by 1-step protocol and manual peak lists and CYICK.

The structure calculation of the MDM4-nutlin-3a complex was characterized by RMSD radius values < 1.0 Å and RMSD bias values of approximately 1.8 Å when using manual peak lists. Results achieved through CYPICK peak lists are less reliable then those achieved with manual peak lists, i.e. RMSD radius values ranged from 0.80-1.73Å and RMSD bias values ranged from 2.48-3.30 Å, when using assignments from the 2-step and 1-step protocol, respectively. I_r values support the RMSD radius values. Structure bundles of MDM4 in complex with nutlin-3a achieved from the 2-step protocol assignment using the manual and CYPICK peak lists are depicted in Fig. 4.8 **(a)** and **(b)**, respectively. Fig. 4.8 **(c)** and **(d)** shows the inter-molecular assigned distance restraints from the calculation of the structures presented in **(a)** and **(b)**, respectively. Structure calculations of the MDM4-nutlin-3a complex were in general more complicated than the peptide-2 calculation. Nutlin-3a was in many calculations not attached to the binding pocket of MDM4. Therefore, four inter-molecular distance restraints were defined and used in the structure calculation of the complex. Nevertheless, the position of nutlin-3a was still rather imprecise. The ligand was flipped in many cases by 180°, especially when using CYPICK peak lists. Generally, the position of nutlin-3a became more defined when using manual peak lists. Also the distance restraint network is more dense when compared to the calculation with CYPICK peak lists. The reason for the imprecision of nutlin-3a is not completely clear. The chemical shift assignment could be responsible for the resulting imprecise complex structure.

Structure calculation of apo-MDM4 was in general performed with a much lower precision, i.e. ranging from 1.8–7.7 Å. The RMSD bias ranged from 3.4–6.0 Å. The accordant structure bundles achieved from assignments of the 2-step protocol are depicted in Fig. 4.9 **(a)** and **(b)**, using manual and CYPICK peak lists, respectively. Apo-MDM4 was not selected as reference assignment within this study because of the absence of many signals and the unstable structure. As can be seed from Fig. 4.9 the coordinates of the atoms are rather uncertain. Especially, the C-terminal helix is rather flexible. However, automated peak picking by CYPICK was performed poorly. In this case the uncertainty in the structure could also be retraced to the missing signals in the CYPICK peak lists.

4.4 Conclusions

In this chapter a new method to allow SBDD by NMR in a high-throughput manner comparable to X-ray crystallography has been presented. Major developments were: (i) the Protein-Ligand NOESY which allows recording all data necessary for assignment and structure calculation of a protein-ligand complex within a couple of hours, and (ii) the FLYA assignment transfer which guides the automatic assignment of a new complex structure based on a known structure (complex or apo). With these two developments the measurement time for obtaining the necessary data was significantly reduced and the time spend on spectral analysis and obtaining a complete chemical shift assignment was also reduced significantly.

The FLYA assignment transfer led to correct chemical shift assignments with respect to manual resonance assignment if available. Differences in chemical shift assignment obtained with manual and CYPICK peak list are negligible in most cases. The 2-step assignment protocol however brought some improvement in all studied examples. In case of nutlin-3a however, the chemical shift assignment is rather uncertain. In this case no complete manual assignment was available for comparison. Nutlin-3a has further proven itself to be difficult in structure calculation, also when using manually picked NOESY peak lists and when adding structural information in the form of a set of manual assigned inter-molecular distance restraints. The remaining structure calculations were performed well and their was no need for explicitly defining inter-molecular NOEs. Structure bundles achieved on the basis of manual peak lists were superior to those obtained from CYPICK peak lists. Manual refinement of automatically picked peak lists has proven itself worthwhile within this study. The performance of the SBDD by NMR protocol can probably be improved when adding additional information for chemical shift assignment, i.e. the measurement of one additional triple-resonance spectrum or one spectrum which gives information on side-chain assignments.

SBDD by NMR has some additional limitations which are equal to conventional NMR methods: One has to obtain a uniformly isotope labeled protein which should have a size below ~ 30 kDa (in case of non-symmetric proteins). The ligand plane can quickly suffer from crowding of signals, depending on the chemical composition of the ligand. Ligands up to 1–2 kDa could in general be analyzed in this study.

4.5 Usage of the FLYA assignment transfer protocol in CYANA

In the following literal CYANA input is written in **bold** and other input is written in *italics*. CYANA macros are recognizable by the suffix **.cya**. Example macros are written in `typewritterfont`.

4.5.1 FLYA parameters

- **shiftref**=string (optional)

  Name of the reference chemical shift list for comparison with FLYA results. Has to be provided in XEASY format.

- **shiftassign_statistic**=string (optional, default **bmrb**)

  Name of the chemical shift statistic that is used for estimating the chemical shift position of the expected peaks, $f(a)$. In the FLYA assignment transfer protocol the name of the chemical shift file to be used as statistic can be specified.

- **shiftassign_iterations**=integer (optional, default **15000**)

  Number of local iterations for improving assignments.

- **shiftassign_population**=integer (optional, default **50**)

  Population size of the evolutionary algorithm.

- **shiftassign_sdfac**=real (optional, default **1.0**)

  Multiplication factor α for the atom-specific chemical shift standard deviation from the BMRB statistics, $\sigma(a)$.

- **shiftassign_fix**=string (optional)

  A chemical shift list can be defined which includes shifts that are kept fixed during automated chemical shift assignment (Only relevant for the 2-step FLYA assignment transfer).

4.5.2 FLYA macros

- **1-step FLYA assignment transfer**

  The 1-step FLYA assignment transfer protocol is described in chapter 4.2.4. Expected peak positions are derived from the reference chemical shift assignment, which is referred to as `stat.prot` in the following example. The chemical shift range is determined from the chemical shift value in the reference assignment, $f(a)$, and a fraction of the atom-specific standard deviation specified in the BMRB, $\alpha \cdot \sigma(a)$. The fraction of the standard deviation can be specified via **shiftassign_sdfac**. We used the factor 0.5 in the 1-step protocol, which we found empirically.

```
tolerance:=0.02,0.03,0.4
aspeaks:=15NHSQC,13CHSQC,15NNOESY,13CNOESY
read seq protein.seq

shiftref:=reference.prot
shiftassign_statistics:=stat.prot
shiftassign_iterations:=15000
shiftassign_population:=200
shiftassign_sdfac:=0.5

command select atoms
    select atoms "* - CZ ?H* @ARG - ?Z @LYA"
end

flya runs=20 shiftreference=$shiftref stages=0 assignpeaks=$aspeaks
```

- **2-step FLYA assignment transfer**

  The 2-step FLYA assignment transfer protocol is described in chapter 4.2.4. In the first step expected peaks are derived from a reference structure in the following example named `ref.pdb`. The position of the expected peaks is determined from the reference chemical shift assignment, which is referred to as `stat.prot` in the following example. The chemical shift range is determined from the chemical shift value in the reference assignment, $f(a)$, and a fraction of the atom-specific standard deviation specified in the BMRB, $\alpha \cdot \sigma(a)$. The fraction of the standard deviation can be specified via **shiftassign_sdfac**. We used the factor 0.1 in the first step of the 2-step protocol. The second step of the 2-step protocol is equal to the 1-step protocol, expected peaks are generated from a random structure, the chemical shift range of the atoms linked to the expected peaks is determined via the reference chemical shift assignment and 0.5-fold the standard deviation from the BMRB of the accordant atom.

```
tolerance:=0.02,0.03,0.4
aspeaks:=15NHSQC,13CHSQC,15NNOESY,13CNOESY
read seq protein.seq

shiftref:=reference.prot
shiftassign_statistics:=stat.prot
shiftassign_iterations:=15000
shiftassign_population:=200

command select atoms
  select atoms "* - CZ ?H* @ARG - ?Z @LYA"
end

#------------------------------------Flya run 01------------------------------------#
shiftassign_sdfac:=0.1
flya runs=20 shiftreference=$shiftref stages=0 assignpeaks=$aspeaks structure=ref.pdb

#write Flya results to folder flya_01

if (master) then
  system "mkdir flya_01;cp -f *.* flya_01/.;cp -rf details flya_01/."
  #write prot retaining only strong assignments form first run
  read prot flya_01/flya.prot
  atom select "* tolerance=0.00..0.009"
  write prot fix_01.prot
end if

#------------------------------------Flya run 02------------------------------------#
shiftassign_sdfac:=0.5
shiftassign_fix:=fix_01.prot

flya runs=20 shiftreference=$shiftref stages=0 assignpeaks=$aspeaks

# write flya results to folder flya_02
if (master) then
  system "mkdir flya_02;cp -f *.* flya_02/.;cp -rf details flya_02/."
end if
```

Chapter 5

Conclusions and Outlook

Projects addressed in this dissertation aimed at improving the process of automated structure determination by NMR. Signal identification in NMR is a time-consuming process which is performed faster and more objectively by an automated procedure. In order to perform automated signal identification of NMR spectra within CYANA (Güntert *et al.*, 1997; Güntert & Buchner, 2015), we developed CYPICK (Würz & Güntert, 2016), a contour geometry based algorithm. We further developed the information content, a tool for helping the scientist evaluating the information of distance restraints and anticipating the precision of the resulting structure bundle. The structural information is expressed by a single number and can be compared to resolution in X-ray crystallography. Additionally, a new approach for structure-based drug design (SBDD) by NMR was developed. SBDD by NMR is a procedures that has the aim of performing SBDD in a high-throughput manner by providing information for the chemical shift assignment in the form a resonance assignment of a reference structure and/or a reference structure. This procedure also aims at reducing the analysis time.

Results achieved with the new peak picking algorithm CYPICK in fully automated protein structure determination starting from NMR spectra were promising. The large scale study on several proteins and a very diverse set of NMR spectra showed that it is possible to obtain correct chemical shift assignments from CYPICK peak lists and calculate accurate structures. CYPICK yielded peak lists that are more favorable than those obtained by other automated peak picking programs with respect to finding more true peaks, rejecting more artifacts, the correctness of chemical shift assignments, and the accuracy of 3D protein structures. CYPICK identifies peaks efficiently, i.e. computation times varied between 1 s for a ^{15}N-HSQC spectrum and 31 s for a ^{13}C-resolved NOESY spectrum on a standard desktop computer. CYPICK analyzes only local properties of the spectrum and is therefore applicable to any type of multidimensional NMR spectrum.

However, these investigations also revealed that certain functionalities of CYPICK can be improved or implemented in future projects. The requirements for the presence of a local extremum should be relaxed in order to identify overlapping signals which do no possess a local extremum. In this context, it would also be desirable to have a stable deconvolution method. Additionally, the algorithm should become completely independent from noise level values. Thereby, very weak signals buried below the global threshold could be identified. A tool for discarding peaks arising from noise bands would also improve the overall performance of CYPICK. Furthermore, it would be worthwhile to have the opportunity to guide the peak picking by additional information such as a 3D structure or a chemical shift assignment. Contour-based quality factors Q_{rad} and Q_{con} can in principle be used to direct automated chemical shift and NOE assignment.

CYPICK is also employed in the protein validation tool CYVAL (Kirchner & Güntert, 2016) developed by Dr. Donata Kirchner. CYVAL calculates a structure validation score on the basis of a peak match between predicted peaks from the structure and picked peaks by CYPICK from a NOESY spectrum. CYPICK peak lists were also used in the structure-based drug design (SBDD) by NMR project and yielded chemical shift assignments comparable to those obtained from manual peak picking.

In another project the information content of NMR distance restraints was developed. The information content is a quantitative measure of structural information included in NMR distance restraints, which is straightforward and fast to calculate. The computation time of a data set including 1100 distance restraints took 2 s. For the calculation of the information content no explicit user input is required, only the protein amino acid sequence and a distance restraint list have to be provided. The information content can be calculated within CYANA and is also available as a stand-alone program.

Within this thesis, several characteristics of the information content have been demonstrated. The information content correlates with the precision of the resulting structure bundle and is comparable to resolution in X-ray crystallography. We showed that a certain minimal information content is necessary to obtain a structure bundle that has a precision better than 2.0 Å.

The information content has not been published yet, but it has been used within our group for several tasks. The information content measure has proven itself to be not only interesting for structure calculation projects, but also represents a direct measurement of the degree of data sparseness with respect to a reference data set. A measurement of data sparseness is very useful for studies that investigate the performance of certain methods with respect to data sets of varying quality. The information content was also used in the new structure validation tool CYVAL developed by Dr. Donata Kirchner. In this tool the information content is used in the weighting factor calculation of missing and matching peaks.

Generally, the concept of the information content can be extended to other types of structure restraints, e.g. torsion angles constraints, residual dipolar couplings, and pseudo contact shifts. The investigation of a correlation of the information content with structure quality parameters would also be interesting in future projects.

In a collaborative project with Dr. Alvar Gossert and Dr. Elena Schmidt the SBDD by NMR method was developed. The method aims at performing structure calculations of different ligated forms of a protein in a high-throughput manner, comparable to X-ray crystallography. Major developments were the 'Protein-Ligand NOESY' and the 'FLYA assignment transfer'. With these two developments the measurement time for obtaining the necessary data was significantly reduced and the time spent for spectral analysis and obtaining a complete chemical shift assignment was also reduced significantly.

The 'FLYA assignment transfer' yielded chemical shift assignments of high accuracy for the new complexes. However, if no complete reference chemical shift assignment is available, it is difficult to evaluate the accuracy of the chemical shift assignment. A score for evaluating the automated chemical shift assignment by FLYA would be very helpful. Structure calculations led to results of varying quality. The structure of MDM4 in complex with peptide-1 was determined with a high accuracy and precision, whereas the MDM4-nutlin-3a structure was less reliable. Within this study the helpfulness of the consensus structure calculation (Buchner & Güntert, 2015a) was demonstrated. The combined and also the consensus structure bundle directly gave an indication of the quality of the structure calculation. Improvements in the SBDD by NMR protocol can probably be achieved by adding additional information for chemical shift assignment, i.e. the measurement of an additional triple-resonance spectrum or a spectrum which gives information on side-chain assignments.

SBDD by NMR has, however, some additional limitations which are equal to conventional NMR methods: One has to obtain a uniformly isotope labeled protein which should have a size below $\sim$ 30 kDa (in case of non-symmetric proteins). The ligand plane can quickly suffer from crowding of signals, depending on the chemical composition of the ligand. Ligands up to 1-2 kDa could in general be analyzed in this study.

Appendices

Appendix A

CYPICK

Table A.1: Available multidimensional NMR spectra and reference peak lists for ENTH, RHO and SH2

	ENTH		RHO		SH2	
spectrum	points[a]	widths (kHz)[b]	points[a]	widths (kHz)[b]	points[a]	widths (kHz)[b]
^{15}N-HSQC	512 x 128	11.2, 1.8	512 x 64	11.2, 2.7	512 x 46	11.2, 2.7
^{13}C-HSQC	512 x 128	11.2, 8.7	512 x 128	11.2, 15.1	512 x 40	11.2, 7.9
			512 x 64	5.4, 4.8		
HNCO	27 x 70	8.4,1.4,3.3	46 x 50	8.4, 2.0, 3.3	46 x 50	8.4, 2.0, 3.3
HN(CA)CO	27 x 70	8.4,1.4,3.3	46 x 50	8.4, 2.0, 3.3	46 x 50	8.4, 2.0, 3.3
HNCA	29 x 70	8.4,1.4,4,8	46 x 50	8.4, 2.0, 4.8	46 x 50	8.4, 2.0, 4.8
HN(CO)CA	27 x 70	8.4,1.4,4.8	46 x 50	8.4, 2.0, 4.8	46 x 50	8.4, 2.0, 4.8
CBCANH	32 x 75	8.4,1.4,11.3	46 x 64	8.4, 2.0, 11.3	46 x 64	8.4, 2.0, 11.3
CBCA(CO)NH	32 x 70	8.4,1.4,11.3	46 x 64	8.4, 2.0, 11.3	46 x 64	8.4, 2.0, 11.3
HBHA(CO)NH	26 x 60	8.4,1.4,6.8	46 x 64	8.4, 2.0, 7.5	46 x 64	8.4, 2.0, 8.4
(H)CC(CO)NH	24 x 60	8.4,1.4,11.3	46 x 64	8.4, 2.0, 11.3	46 x 64	8.4, 2.0, 11.3
H(CCCO)NH	27 x 77	8.4,1.4,6.8	46 x 64	8.4, 2.0, 7.5	46 x 64	8.4, 2.0, 6.7
HCCH-COSY[c]	32 x 85	7.8,6.5,6.8	16 x 80	5.4, 3.9, 5.4	50 x 100	8.4, 11.3, 8.4
	17 x 85	5.2,4.0,5.2			16 x 80	6.1, 4.6, 6.1
(H)CCH-TOCSY	44 x 80	8.4,6.5,13.9				
HCCH-TOCSY	32 x 120	7.8,6.5,6.8	64 x 100	8.4, 11.3, 8.4	64 x 100	8.4, 11.3, 8.4
^{15}N-edited NOESY	36 x 128	11.2,1.8,10.1	46 x 128	11.2, 2.7, 11.2	46 x 128	11.2, 2.7, 11.2
^{13}C-edited NOESY[c]	46 x 150	11.2,8.7,8.8	34 x 116	11.2, 7.7, 11.2	40 x 150	11.2, 8.0, 11.2
			32 x 128	11.2, 5.1, 11.2		

[a] Points represent the number of complex time domain data points in the indirect dimension. The first number refers to ^{15}N, if present, or ^{13}C. The second number refers to ^{1}H, if present, or ^{13}C. The direct ^{1}H dimension was in case of 3D spectra recorded with 512 complex time domain data points (López-Méndez & Güntert, 2006). [b] Width represents the spectral width in the direct and the indirect detected dimension(s) (López-Méndez & Güntert, 2006). [c] The two sets of values refer to the separately recorded aliphatic and aromatic carbon regions (López-Méndez & Güntert, 2006).

Table A.2: Input parameters for picking of ENTH, RHO, and SH2 spectra using the CYPICK algorithm. Column labels are explained in chapter 2.6.

Spectrum	*type*	ENTH				RHO				SH2			
		σ_1	σ_2	σ_3	*option*	σ_1	σ_2	σ_3	*option*	σ_1	σ_2	σ_3	*option*
^{15}N-HSQC	"N15HSQC H N"	0.04	0.2		±	0.04	0.3		±	0.04	0.4		±
^{13}C-HSQC	"C13HSQC C H"	0.4	0.04		±	0.04	0.4		+	0.04	0.6		±
HNCO	"HNCO HN C N"	0.05	0.2	0.8	+	0.05	0.2	0.8	+	0.09	0.3	0.8	+
HN(CA)CO	"HNcaCO HN C N"	0.04	0.2	0.8	+	0.04	0.3	0.7	+	0.04	0.2	1.3	+
HNCA	"HNCA HN C N"	0.04	0.6	0.8	+	0.04	0.5	0.8	+	0.2	0.4	0.8	+
HN(CO)CA	"HNcoCA HN C N"	0.05	0.5	0.8	+	0.04	0.5	0.8	+	0.05	0.5	0.8	+
CBCANH	"CBCANH HN C N"	0.07	0.6	0.7	±	0.04	0.7	0.8	±	0.04	0.8	0.8	±
CBCA(CO)NH	"CBCAcoNH HN C N"	0.07	0.6	0.7	+	0.04	0.6	0.7	+	0.04	0.9	0.8	+
HBHA(CO)NH	"HBHAcoNH HN H N"	0.04	0.2	0.4	+	0.04	0.2	0.3	+	0.04	0.2	0.4	+
(H)CC(CO)NH	"CcoNH HN C N"	0.04	1.3	0.5	+	0.04	1.2	0.4	+	0.04	1.2	0.4	+
H(CCCO)NH	"HCcoNH HN H N"	0.04	0.1	0.4	+	0.04	0.2	0.3	+	0.04	0.2	0.3	+
HCCH-COSY[a]	"HCCHCOSY HC H C"	0.1	0.04	0.7	+					0.1	0.06	0.7	+
	"HCCHCOSY HC H C"	0.1	0.03	0.8	+	0.1	0.04	0.8	+	0.1	0.04	1.0	+
(H)CCH-TOCSY	"CCHTOCSY H1 C1 C2"	0.05	1.2	1.2	+								
HCCH-TOCSY	"HCCHTOCSY HC H C"	0.1	0.05	1.5	+	0.2	0.05	1.3	+	0.1	0.07	1.2	+
^{15}N-edited NOESY	"N15NOESY HN H N"	0.04	0.1	0.6	±	0.04	0.1	0.7	+	0.04	0.1	0.4	+
^{13}C-edited NOESY	"C13NOESY HC H C"	0.05	0.08	1.0	±	0.04	0.1	0.7	±	0.04	0.1	0.8	±
					0.04	0.1	0.7	+					

[a] The first entry refers to the aliphatic region, the second to the aromatic region.

'+': only positive signals were picked with the option **only_pos**, '±': positive and negative signals were picked, in this case CYPICK does not need an explicit specification.

Table A.3: Input parameters for picking CASD-NMR data sets using the CYPICK algorithm. Column labels are explained in chapter 2.6.

Data set	^{13}C-edited NOESY aliphatic region				^{13}C-edited NOESY aromatic region				^{15}N-edited NOESY			
	σ_1	σ_2	σ_3	*option*	σ_1	σ_2	σ_3	*option*	σ_1	σ_2	σ_3	*option*
HR2876B	0.05	1.2	0.1	+	0.04	1.0	0.4	+	0.05	0.6	0.1	−
HR2876C	0.06	1.2	0.1	+	0.06	1.2	0.1	+	0.04	0.6	0.1	−
HR5460A	0.07	1.2	0.1	+	0.07	1.2	0.1	+	0.07	0.6	0.1	−
HR6430A	0.04	1.2	0.1	+	0.04	1.2	0.1	+	0.04	0.6	0.1	−
HR6470A	0.03	0.7	0.06	+	0.04	1.2	0.1	+	0.05	0.5	0.11	−
HR8254A	0.05	1.2	0.06	+	0.04	1.2	0.1	+	0.05	0.5	0.12	−
OR135	0.05	1.2	0.1	+	0.04	1.2	0.1	+	0.04	0.6	0.1	−
OR36	0.04	1.2	0.1	+	0.05	1.4	0.1	+	0.04	0.6	0.1	−
StT322	0.04	0.8	0.05	+	0.03	0.9	0.07	+	0.03	0.4	0.08	+
YR313A	0.04	1.2	0.1	+	0.06	1.2	0.1	+	0.04	0.6	0.1	−

'+': only positive signals were picked with the option **only_pos**, '−': only negative signals were picked with the option **only_neg**.

Table A.4: AUTOPSY picking parameters for ENTH spectra.

Spectrum	Threshold	Extrema to search for
^{15}N-HSQC	$0.7\cdot10^4$	±
^{13}C-HSQC	$0.4\cdot10^3$	±
HNCO	$0.8\cdot10^4$	+
HN(CA)CO	$0.7\cdot10^4$	+
HNCA	$0.9\cdot10^4$	+
HN(CO)CA	$0.3\cdot10^4$	+
CBCANH	$0.5\cdot10^4$	±
CBCA(CO)NH	$0.5\cdot10^4$	+
HBHA(CO)NH	$0.1\cdot10^5$	+
(H)CC(CO)NH	$0.3\cdot10^4$	+
H(CCCO)NH	$0.6\cdot10^4$	+
HCCH-COSY[a]	$0.4\cdot10^3$	+
	$0.4\cdot10^3$	+
HCCH-TOCSY	$0.5\cdot10^4$	+
(H)CCH-TOCSY	$0.2\cdot10^3$	+
^{15}N-edited NOESY	$0.9\cdot10^4$	+
^{13}C-edited NOESY	$0.4\cdot10^3$	±

[a] The first entry refers to the aliphatic region, the second to the aromatic region.

'+': only positive signals were picked, '±': positive and negative signals were picked.

Table A.5: NMRVIEWJ picking parameters for ENTH, RHO, and SH2 spectra.

Spectrum	ENTH		RHO		SH2	
	Autolevel	Extrema to search for	Autolevel	Extrema to search for	Autolevel	Extrema to search for
^{15}N-HSQC	0.6	±	0.6	±	0.9	±
^{13}C-HSQC	0.8	±	0.3	+	0.3	±
HNCO	0.7	+	0.2	+	0.4	+
HN(CA)CO	0.7	+	0.3	+	0.8	+
HNCA	0.7	+	0.2	+	0.6	+
HN(CO)CA	0.4	+	0.2	+	0.3	+
CBCANH	0.6	±	0.4	±	0.5	±
CBCA(CO)NH	0.4	+	0.2	+	0.3	+
HBHA(CO)NH	1.0	+	0.2	+	0.4	+
(H)CC(CO)NH	0.3	+	0.2	+	0.4	+
H(CCCO)NH	0.5	+	0.2	+	0.4	+
HCCH-COSY[a]	0.3	+			0.3	+
	0.7	+	0.1	+	42.3	+
HCCH-TOCSY	0.3	+	0.2	+	0.3	+
(H)CCH-TOCSY	0.2	+				
^{15}N-edited NOESY	1.5	+	1.1	+	16.3	+
^{13}C-edited NOESY[a]	0.2	±	0.3	±	0.4	±
			0.2	+		

[a] The first entry refers to the aliphatic region, the second to the aromatic region.

'+': only positive signals were picked, '±': positive and negative signals were picked.

Table A.6: CCPN picking parameters for ENTH, RHO, and SH2 spectra.

Spectrum	ENTH		RHO		SH2	
	Threshold	Extrema to search for	Threshold	Extrema to search for	Threshold	Extrema to search for
^{15}N-HSQC	7,000	±	1,000	±	50,000	±
^{13}C-HSQC	400	±	2,000	+	2,000	±
HNCO	8,000	+	3,000	+	3,000	+
HN(CA)CO	7,000	+	4,000	+	7,000	+
HNCA	9,000	+	7,000	+	6,000	+
HN(CO)CA	3,000	+	3,000	+	3,000	+
CBCANH	5,000	±	4,000	±	4,000	±
CBCA(CO)NH	5,000	+	6,000	+	3,000	+
HBHA(CO)NH	10,000	+	2,000	+	5,000	+
(H)CC(CO)NH	3,000	+	2,000	+	5,000	+
H(CCCO)NH	6,000	+	2,000	+	5,000	+
HCCH-COSY[a]	400	+			5,000	+
	400	+	1,000	+	4,000	+
HCCH-TOCSY	5,000	+	9,000	+	7,000	+
(H)CCH-TOCSY	200	+				
^{15}N-edited NOESY	9,000	+	100,000	+	9,000	+
^{13}C-edited NOESY[a]	400	±	70,000	±	3,000	±
			2,000	+		

[a] The first entry refers to the aliphatic region, the second to the aromatic region.

'+': only positive signals were picked, '±': positive and negative signals were picked.

Table A.7: CV-Peak Picker scanning parameters for ENTH, RHO and SH2.

Spectrum	ENTH				RHO				SH2			
	Sensitivity	Volume	Exclude negative peaks	Window size	Sensitivity	Volume	Exclude negative peaks	Window size	Sensitivity	Volume	Exclude negative peaks	Window size
^{15}N-HSQC	0.2	20,000	False	27x23	0.5	20,000	False	17x15	0.5	20,000	False	15x15
^{13}C-HSQC	0.06	20,000	False	31x15	0.4	21,000	True	15x9	0.3	20,000	False	15x17
HNCO	0.1	52,000	True	37x11	0.5	20,000	Tue	21x9	0.75	30,000	False	9x9
HN(CA)CO	0.3	51,000	False	17x11	0.3	25,000	False	17x11	0.9	55,000	False	9x9
HNCA	0.1	65,000	True	23x13	0.8	20,000	True	19x11	0.9	40,000	False	11x11
HN(CO)CA	0.25	20,000	True	25x15	0.7	0	True	21x13	0.75	30,000	False	11x9
CBCANH	0.36	40,000	False	17x11	0.2	15,000	False	15x7	0.5	31,000	False	11x9
CBCA(CO)NH	0.15	55,000	True	23x7	0.9	20,000	False	17x7	0.9	30,000	True	9x7
HBHA(CO)NH	0.5	79,000	False	21x15	0.8	33,000	True	11x21	0.8	20,000	True	11x11
(H)CC(CO)NH	0.15	45,000	True	19x13	0.2	11,000	True	19x11	0.5	25,000	True	9x11
H(CCCO)NH	0.15	16,000	True	11x23	0.15	16,000	True	11x23	0.5	30,000	True	9x9
HCCH-COSY[a]	0.5	20,000	False	33x31	0.45	20,000	False	12x11				
	0.15	27,000	False	39x37	0.14	10,000	True	25x33	0.7	20,000	False	11x13
(H)CCH-TOCSY	0.25	20,000	True	25x13								
HCCH-TOCSY	0.05	20,000	False	45x39	0.05	10,000	True	37x31	0.5	20,000	True	11X11
^{15}N-edited NOESY	0.25	70,000	False	21x21	0.55	20,000	True	17x15	0.75	30,000	True	17x19
^{13}C-edited NOESY[a]	0.05	20,000	False	33x31	0.75	17,000	False	14x19	0.4	30,000	False	13x23
					0.55	10,000	True	21x17				

[a] The first entry refers to the aliphatic region, the second to the aromatic region.

Table A.8: Excluded water signal regions

Spectrum	ENTH [ppm]	RHO [ppm]	SH2 [ppm]
^{13}C-HSQC	4.65-4.78		4.65-4.78
HCCH-COCSY	4.62-4.78		4.62-4.78
HCCH-TOCSY	4.62-4.78	4.62-4.78	4.62-4.78
(H)CCH-TOCSY	4.62-4.78		
^{15}N-edited NOESY	4.62-4.78		4.62-4.78
^{13}C-edited NOESY	4.50-4.92	4.50-4.92	4.50-4.92
	4.62-4.78	4.62-4.78	4.62-4.78

Table A.9: Results of automated NOE assignment and structure calculation by CYANA using the automated chemical shift assignment and ^{15}N- and ^{13}C-NOESY CYPICK lists. Results are shown for peak lists that were picked by the local noise estimation mode with $\beta = 3.0$ and $\gamma = 1.3$.

	ENTH	**RHO**	**SH2**
NOE assignment[a]			
^{15}N-NOESY	1177	1523	1164
^{13}C-NOESY	2836	3556	3619
Assigned cross peaks	3132(78.0%)	3076(60.6%)	3729(78.0%)
Unassigned cross peaks	881(22.0%)	2003(39.4%)	1054(22.0%)
Restraints			
NOE distance restraints			
short-range	1029(62.2%)	966(60.1%)	953(53.4%)
medium-range	304(18.4%)	221(13.8%)	244(13.7%)
long-range	321(19.4%)	420(26.1%)	589(33.0%)
Dihedral angle restraints (ϕ/ψ)	110	94	82
Structure statistics[a]			
Average CYANA target function [Å^2]	0.5±0.13	0.96±0.08	1.21±0.08
Restraint violations			
Max. distance restraint violations [Å]	-	0.21	0.22
Number of violated distance restraints > 0.2 Å	-	1	1
Max. dihedral angle restraint violations (°)	-	-	-
Number of violated dihedral angle constraints > 5 °	-	-	-
Ramachandran plot			
Residues in most favored regions	86.9%	77.7%	75.2%
Residues in additionally allowed regions	12.9%	22.3%	24.4%
Residues in generously allowed regions	0.1%	0.0%	0.4%
Residues in disallowed regions	0.0%	0.0%	0.0%
RMSD			
RMSD range[b]	9..102,113..130	6..125	8..109
Average backbone RMSD radius [Å]	0.84±0.18	0.49±0.09	0.56±0.11
Average heavy atom RMSD radius [Å]	1.41±0.18	0.97±0.09	0.98±0.11
Backbone RMSD bias [Å]	1.47	2.78	1.55
Heavy atom RMSD bias [Å]	2.12	3.17	2.04

[a] using automated NOE assignment and structure calculation functionalities of CYANA. [b] determined by CYRANGE

Table A.10: Results of automated NOE assignment and structure calculation by CYANA using the chemical shift assignment and the ^{15}N- and ^{13}C-NOESY CYPICK lists. Results are shown for peak lists that were picked by the global noise estimation mode combined with resolve overlap functionalities with $\beta = 3.0$ and $\gamma = 1.3$.

	ENTH	**RHO**	**SH2**
NOE assignment[a]			
^{15}N-NOESY	1813	2848	2276
^{13}C-NOESY	5231	6502	7511
Assigned cross peaks	4787(68.0%)	4293(45.9%)	5120(52.3%)
Unassigned cross peaks	2257(32.0%)	5057(54.1%)	4667(47.7%)
Restraints			
NOE distance restraints			
short-range	1450(52.0%)	1332(52.6%)	1422(49.5%)
medium-range	729(26.1%)	429(16.9%)	403(14.0%)
long-range	611(21.9%)	772(30.5%)	1046(36.4%)
Dihedral angle restraints (ϕ/ψ)	108	99	80
Structure statistics[a]			
Average CYANA target function [Å^2]	2.80±0.21	3.48±0.28	4.95±0.25
Restraint violations			
Max. distance restraint violations [Å]	-	0.4.0	0.74
Number of violated distance restraints > 0.2 Å	8	3	6
Max. dihedral angle restraint violations (°)	10.98	7.22	9.42
Number of violated dihedral angle constraints > 5 °	3	1	2
Ramachandran plot			
Residues in most favored regions	86.6%	78.4%	78.0%
Residues in additionally allowed regions	12.8%	20.6%	20.4%
Residues in generously allowed regions	0.5%	1.0%	1.5%
Residues in disallowed regions	0.0%	0.0%	0.0%
RMSD			
RMSD range[b]	9..102,113..130	6..125	8..109
Average backbone RMSD radius [Å]	0.41±0.18	0.22±0.06	0.16±0.03
Average heavy atom RMSD radius [Å]	0.87±0.18	0.57±0.07	0.53±0.05
Backbone RMSD bias [Å]	1.08	1.46	1.43
Heavy atom RMSD bias [Å]	1.62	2.01	1.87

[a] using automated NOE assignment and structure calculation functionalities of CYANA. [b] determined by CYRANGE

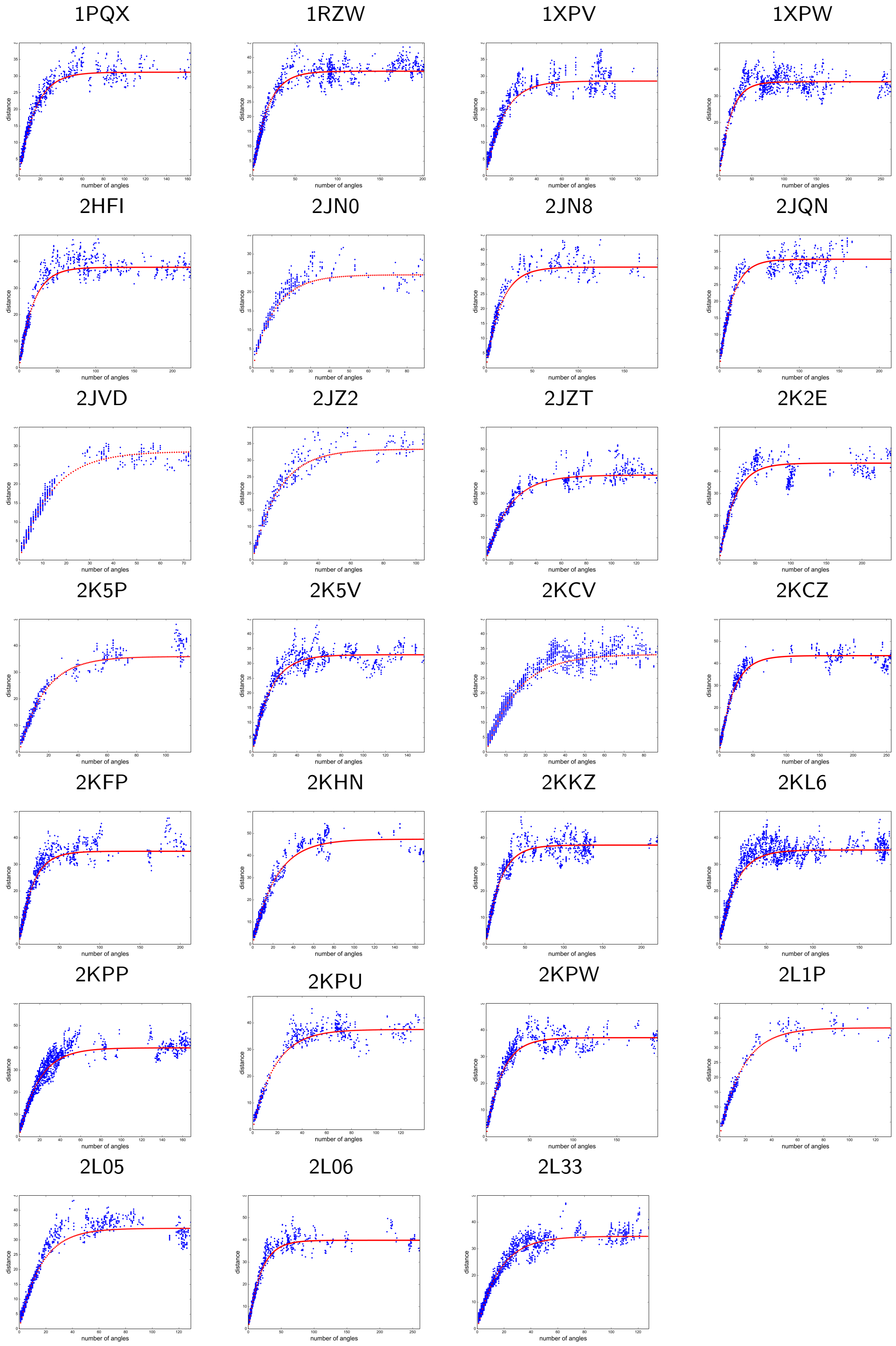
Table A.11: Results of automated NOE assignment and structure calculation by CYANA using the automated chemical shift assignment and the ^{15}N- and ^{13}C-NOESY CYPICK lists. Results are shown for peak lists that were picked by the local noise picking combined with resolve overlap mode with $\beta = 3.0$ and $\gamma = 1.3$.

	ENTH	**RHO**	**SH2**
NOE assignment^{a}			
^{15}N-NOESY	1331	1797	1203
^{13}C-NOESY	3083	4101	4233
Assigned cross peaks	3425(77.6%)	3537(60.0%)	4072(74.9%)
Unassigned cross peaks	989(22.4%)	2361(40.0%)	1364(25.1%)
Restraints			
NOE distance restraints			
short-range	1058(59.4%)	1041(56.9%)	1030(52.4%)
medium-range	388(21.8%)	285(15.6%)	266(13.5%)
long-range	334(18.8%)	503(27.5%)	670(34.1%)
Dihedral angle restraints (ϕ/ψ)	110	99	78
Structure statistics^{a}			
Average CYANA target function [Å^2]	0.61±0.08	1.03±0.08	1.35±0.06
Restraint violations			
Max. distance restraint violations [Å]	-	-	-
Number of violated distance restraints > 0.2 Å	-	-	-
Max. dihedral angle restraint violations (°)	-	-	-
Number of violated dihedral angle constraints > 5 °	-	-	-
Ramachandran plot			
Residues in most favored regions	87.2%	82.8%	76.3%
Residues in additionally allowed regions	12.8%	17.2%	23.6%
Residues in generously allowed regions	0.1%	0.0%	0.0%
Residues in disallowed regions	0.0%	0.0%	0.2%
RMSD			
RMSD range^{b}	9..102,113..130	6..125	8..109
Average backbone RMSD radius [Å]	0.83±0.10	0.38±0.82	0.42±0.07
Average heavy atom RMSD radius [Å]	1.35±0.10	0.82±0.06	0.81±0.05
Backbone RMSD bias [Å]	1.52	1.59	1.20
Heavy atom RMSD bias [Å]	2.23	2.05	1.68

[a] using automated NOE assignment and structure calculation functionalities of CYANA. [b] determined by CYRANGE

Appendix B

Information content



1PQX 1RZW 1XPV 1XPW

2HFI 2JN0 2JN8 2JQN

2JVD 2JZ2 2JZT 2K2E

2K5P 2K5V 2KCV 2KCZ

2KFP 2KHN 2KKZ 2KL6

2KPP 2KPU 2KPW 2L1P

2L05 2L06 2L33

Figure B.1: Correlation between the maximal distance of an atom pair and the number of torsion angles separating the atom pair using the complete evaluation data set. The real maximal distances is derived from a bundle of random structures. The number of torsion angles are determined from the covalent geometry of the protein. Corresponding values are plotted and fit by Eq. 3.13 (red data points). The exponential fit is used to estimate the maximal distance when no structure bundle is available.

Appendix C

Structure-based drug design by NMR

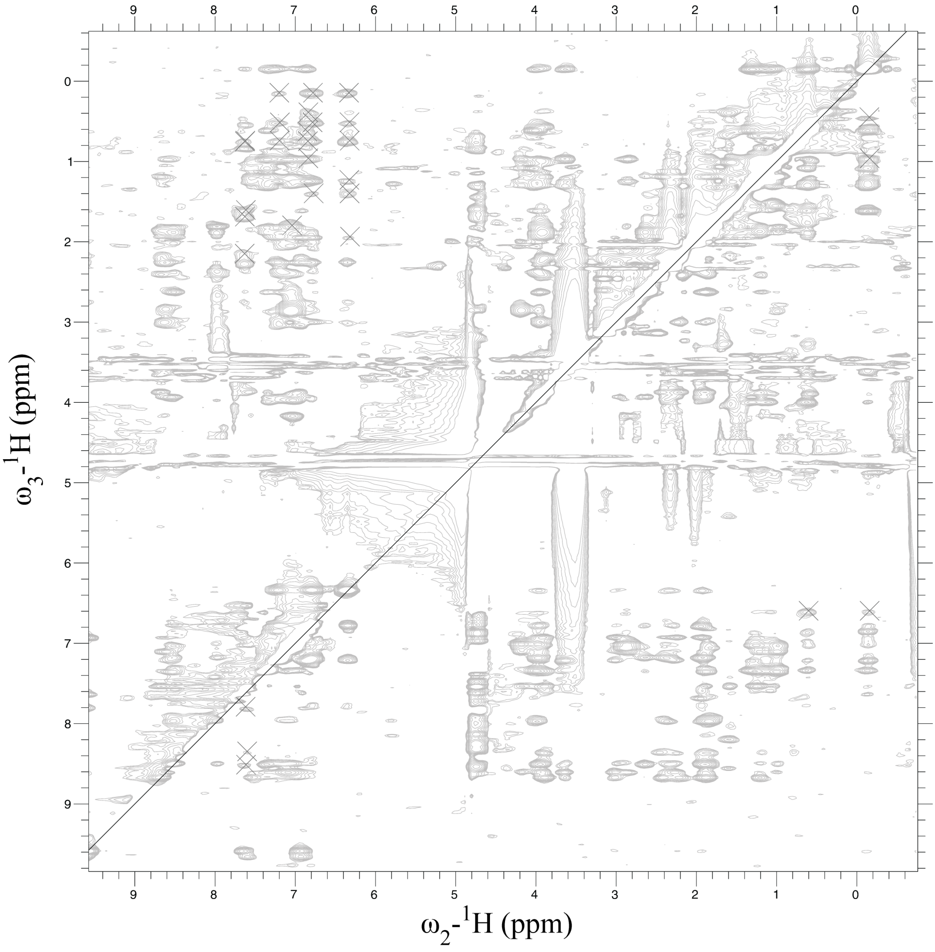


Figure C.1: Peptide-1. Ligand plane of the Protein-Ligand NOESY. Peaks that were selected as **noartifact** in automated NOE assignment are marked with a black cross.

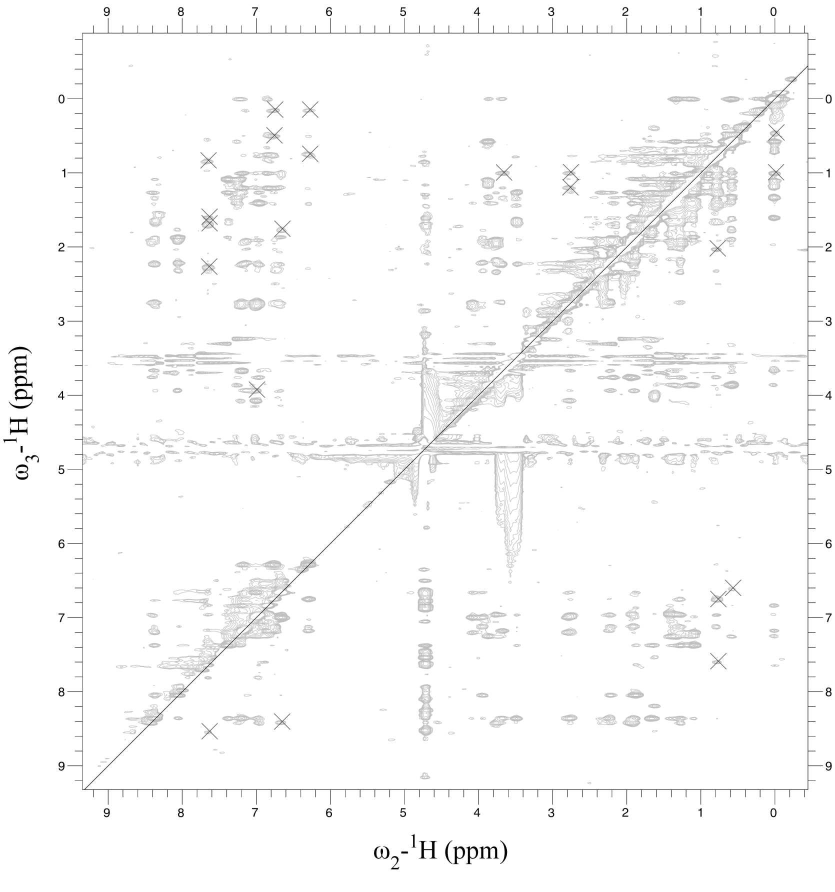


Figure C.2: Peptide-2. Ligand plane of the Protein-Ligand NOESY. Peaks that were selected as **noartifact** in automated NOE assignment are marked with a black cross.

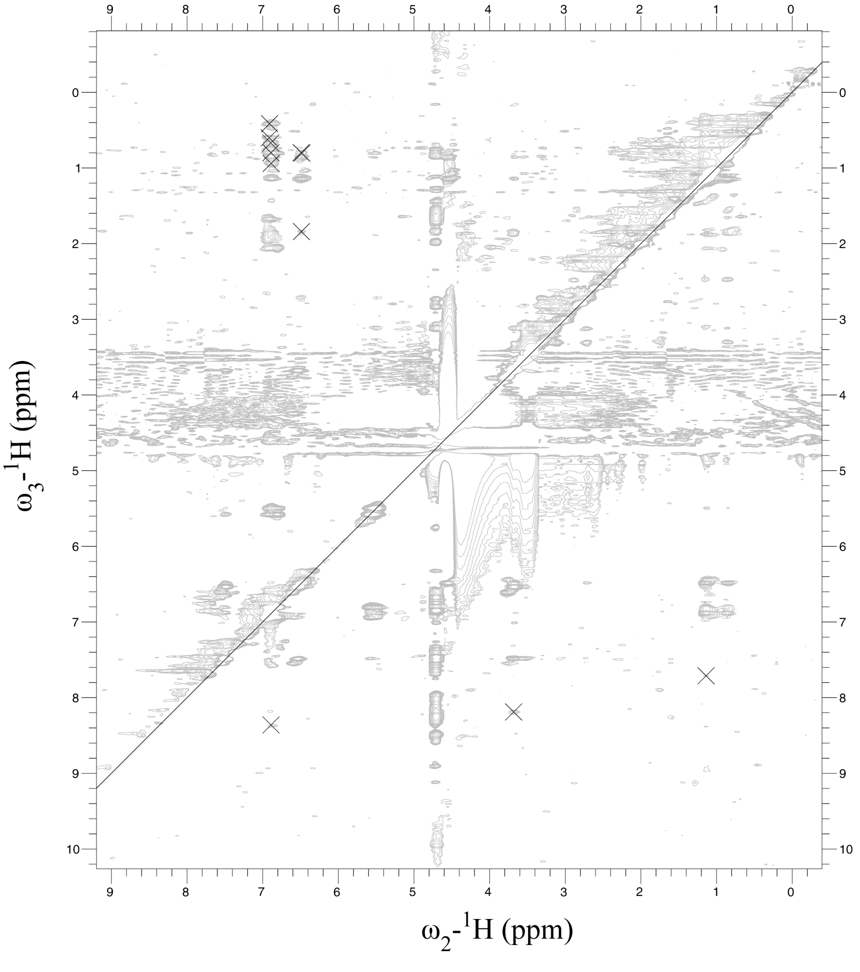


Figure C.3: Nutlin-3a. Ligand plane of the Protein-Ligand NOESY. Peaks that were selected as **noartifact** in automated NOE assignment are marked with a black cross.

Bibliography

Aeschbacher T, Schmidt E, Blatter M, Maris C, Duss O, Allain FHT, Güntert P, Schubert M. Automated and assisted RNA resonance assignment using NMR chemical shift statistics. *Nucleic Acids Research*, 41:172, 2013.

Alipanahi B, Gao X, Karakoc E, Donaldson L, Li M. PICKY: A novel SVD-based NMR spectra peak picking method. *Bioinformatics*, 25:i268-i275, 2009.

Anderson AC. The process of structure-based drug design. *Journal of Chemical Biology*, 10:787-797, 2003.

Bahrami A, Assadi AH, Markley JL, Eghbalnia HR. Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. *PLoS Computational Biology*, 5:1-15, 2009.

Baran MC, Huang YJ, Moseley HNB, Montelione GT. Automated Analysis of Protein NMR Assignments and Structures. *Chemical Reviews*, 104:3541-3555, 2004.

Bartels C, Xia TH, Billeter M, Güntert P, Wüthrich K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *Journal of Biomolecular NMR*, 6:1-10, 1995.

Bartels C, Billeter M, Güntert P, Wüthrich K. Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. *Journal of Biomolecular NMR*, 7:207-213, 1996.

Bartels C, Güntert P, Billeter M, Wüthrich K. GARANT - a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *Journal of Computational Chemistry*, 18:139-149, 1997.

Bardiaux B, Bernard A, Rieping W, Habeck M, Mallavin TE, Nilges M. Influence of different assignment conditions on the determination of symmetric homodimeric structures with ARIA. *Proteins: Structure, Function and Bioinformatics*, 75:569-585, 2009.

Bax A & Grishaev A. Weak alignment in NMR: a hawk-eyed view of bimolecular structure. *Current Opinion in Structural Biology*, 15: 563-750, 2005.

Battiste JL & Wagner G. Utilization of site-directed spin labeling and high-resolution heteronuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data. *Biochemistry*, 39: 5355-5365, 2000.

Bermejo GA & Llinás M. Structure-oriented methods for protein NMR data analysis. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 56:311-328, 2008.

Billeter M, Braun W, Wüthrich K. Sequential resonance assignments in protein ^{1}H nuclear magnetic resonance spectra: computation of sterically allowed proton-proton distances and statistical analysis of proton-proton distances in single crystal protein conformations. *Journal of Molecular Biology*, 155:321-346, 1982.

Billeter M. A consensus on protein structure accuracy in NMR? *Structure Previews*, 23:255-256, 2015.

Blackledge M. Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 46:23-61, 2005.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Research*, 28:235-242, 2000.

Bodenhausen G, Kogler H, Ernst RR. Selection of coherence-transfer pathways in NMR pulse experiments. *Journal of Magnetic Resonance*, 58:370-388, 1984.

Boelens R, Burgering M, Fogh RH, Kaptein R. Time-saving methods for heteronuclear multidimensional NMR of (^{13}C, ^{15}N) doubly labeled proteins. *Journal of Biomlecular NMR*, 4:201-213, 1994.

Bourgeois F & Lasalle JC. An extension of the Munkres algorithm for the assignment problem to rectangular matrices. *Communications of the ACM*, 15:802-804, 1971.

Braden B. The Surveyor's Area Formula. *The College Mathematics Journal*, 17:326-337, 1986.

Buchner L, Schmidt E, Güntert P. Peakmatch: a simple and robust method for peak list matching. *Journal of Biomolecular NMR* 55:267-277, 2013.

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

Buchner L & Güntert P. Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA. *Journal of Biomolecular NMR*, 1:81-95,2015b.

Brünger AT, Adams PD, Clore GM. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica Section D*, 54:905-921, 1998.

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

Chen K & Tjandra N. The use of residual dipolar couplings in studying proteins by NMR. *Topics in Current Chemistry*, 326:47-67, 2012.

Clore M & Iwahara J. Theory, practice,and applications of paramagnetic relaxation enhancement for the characterizations of transient low-population states of biological macromolecules and their complexes. *Chemical Reviews*, 109:4108-4136, 2009.

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, 117:5179-5197, 1995.

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: A multidimensional spectral processing syste, based on UNIX pipes. *Journal of Biomolecular NMR*, 6:277-293, 1995.

Delaglio F, Kontaxis G, Bax A. Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *Journal of the American Chemical Society*, 122:2142-2143, 2000.

Doreleijers JF, Rullmann T, Kaptein R. Wuality assessment of NMR structures: a statistical survey *Journal of Molecular Biology*, 281:149-164.

Doreleijers JF, Raves ML, Rullmann T, Kaptein R. Completeness of NOEs in protein structures: A statistical analysis of NMR data. *Journal of Biomolecular NMR*, 14:123-132, 1999.

Doreleijers JF, Mading S, Maziuk D, Sojourner K, Yin L, Zhu J, Markley JL, and Ulrich EL. BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *Journal of Biomolecular NMR*, 26:139-146, 2003.

Doreleijers JF, Nederveen AJ, Vranken W, Lin J, Bonvin AM, Kaptein R, Markley JL, Ulrich EL. BioMagResBank databases DOCR and FRED with converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. *Journal of Biomolecular NMR*, 32:1-12, 2005.

Doreleijers JF, Vranken WF, Schulte C, Lin J, Wedell JR, Penkett CJ, Vuister GW, Vriend G, Markley JL, Ulrich EL. The NMR restraints grid at BMRB for 5.266 protein and nucleic acid PDB entries. *Journal of Biomolecular NMR*, 45:389-396, 2009.

Doreleijers JF, Sousa da Silva AW, Krieger E, Nabuurs SB, Spronk CA, Stevens TJ, Vranken WF, Vriend G, Vuister GW. CING: an integrated residue-based structure validation program suite. *Journal of Biomolecular NMR*, 54:267-283, 2012.

Dyson HJ & Wright PE. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, 6:197-208, 2005.

Eccles C, Güntert P, Billeter M, Wüthrich K. Efficient analysis of protein 2D NMR spectra using the software package EASY. *Journal of Biomolecular NMR* 1:111-130, 1991.

Edmonds J & Karp R. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the Association for Computing Machinery* 19:248-264, 1972.

Evans P & McCoy A. An introduction to molecular replacement. *Acta Crystallographica Section D* 64:1-10, 2008.

Foley JD, van Dam A, Feiner SK, Hughes JF. Computer Graphics: Principles and Practice. *Addison-Wesley Reading*, 689-693, 1995.

Folkers G, Jahnke W, Erlanson DA, Mannhold R, Kubinyi H. Fragment-based approaches in drug discovery. *Methods and Principles in Medicinal Chemistry*, 2006.

Garrett DS, Powers R, Gronenborn AM, Clore GM. A Common Sense Approach to Peak Picking in Two-, Three-, and Four-Dimensional Spectra Using Automatic Computer Analysis of Contour Diagrams. *Journal of Magnetic Resonance,* 95:214-220, 1991.

Goddard TD, Kneller DG. SPARKY 3. University of California, San Francisco, 2005.

Gossert AD, Würz JM, Schmidt E, Henry C, Widmer, Güntert, P. Structure-based drug design by NMR. *In preparation* 2016.

Gottstein D, Kirchner DK, Güntert, P. Simultaneous single-structure and bundle representation of protein NMR structures in torsion angle space. *Journal of Biomolecular NMR*, 52:351-364, 2012.

Gottstein D, Reckel S, Dötsch V, Güntert, P. Requirements on Paramagnetic Relaxation Enhancement Data for Membrane Protein Structure Determination by NMR. *Structure*, 6:1019-1027, 2012.

Gronwald W, Kirchhofer G, Gorler A, Kremer W, Ganslmeier B, Neidig KP, Kalbitzer HR. Title. *Journal of Biomolecular NMR*, 17:137-151, 2000.

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

Gronwald W, Brunner K, Kirchöfer R, Nasser A, Trenner J, Ganslmeier B, Riepl H, Ried A, Scheiber J, Elsner R, Neidig KP, Kalbitzer HR. AUREMOL, a new program for the automated structure elucidation of biological macromolecules. *Bruker Reports*, 154/155:11-14, 2004.

Grishaev A & Llinás M. CLOUDS, a protocol for deriving a molecular proton density via NMR. *Proceedings of the National Academy of Sciences*, 44:6707-6712, 2002.

Grzesiek S, Cordier F, Jaravine VA, Barfiel M. Insights into biomolecular hydrogen bonds from hydrogen bond scalar couplings. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 45:275-300, 2004.

Güntert P, Braun W, 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. *Journal of Molecular Biology*, 217:515-530, 1991.

Güntert P & Wüthrich K. FLATT - A new procedure for high-quality baseline correction of two- and higher-dimensional NMR spectra. *Journal of Magnetic Resonance*, 96:403-407, 1992.

Güntert P, Dötsch V, Wider, G, Wüthrich K. Processing of multi-dimensional NMR data with the new software PROSA. *Journal of Biomlecular NMR*, 2:619-629, 1992.

Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *Journal of Molecular Biology*, 273:283-298, 1997.

Güntert P, Billeter M, Ohlenschläger O, Brown LR, Wüthrich K. Conformational analysis of protein and nucleic acid fragments with the new grid search algorithm FOUND. *Journal of Biomlecular NMR*, 12:543-548, 1998.

Güntert, P. Automated NMR protein structure calculation. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 43:105-125, 2003.

Güntert, P. Automated NMR Structure Calculation with CYANA. *Methods in Molecular Biology*, 278:353-378, 2004.

Güntert, P. Automated Structure determination from NMR spectra. *European Biophysics Journal*, 38:129-143, 2009.

Güntert P & Buchner L. Combined automated NOE assignment and structure calculation with CYANA. *Progress in Nuclear Magnetic Resonance*, 62:453-471, 2015.

Guerry P & Herrmann T. Advances in automated NMR protein structure determination. *Quarterly Reviews of Biophysics*, 44:257-309, 2011.

Guerry P, Duong VD, Herrmann T. CASD-NMR 2: robust and accurate unsupervised analysis of raw NOESY spectra and protein structure determination with UNIO. *Journal of Biomolecular NMR*, 62:473-480, 2015.

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

Herrmann T, Güntert P, Wüthrich K. Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *Journal of Biomolecular NMR*, 24:171-189, 2002.

Hiller S, Fiorito F, Wüthrich K, Wider G. Automated projection spectroscopy (APSY). *Proceedings of the National Academy of Sciences*, 102:10876-10881, 2005.

Hore PJ. Nuclear Magnetic Resonance. *Oxford Chemistry Primers*, No. 32, 2007.

Huang YJ, Powers R, Montelione GP. Protein NMR Recall, Precision, and F-measure Scores (RPF Scores): Structure Quality Assessment Measure Based on Information Retrieval Statistics. *Journal of the American Chemical Society*, 127:1665-1674, 2002.

Huang YJ, Tejero R, Powers R, Montelione GP. A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins*, 62:587-603, 2006.

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

Jain A, Vaidehi N, Rodriguez G. A fast and recursive algorithm for molecular dynamics simulation. *Journal of Computational Physics*, 106:258-268, 1993.

Jee JG & Güntert P. Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. *Journal of Structural and Functional Genomics*, 4:179-189, 2003.

Joachimiak A. High-troughput Crystallography for Structural Genomics. *Current Opinion in Structural Biology*, 5:573-584, 2009.

Johnson BA. Using NMRView to Visualize and Analyze the NMR Spectra of Macromolecules. *Methods in Molecular Biology*, 278:313-352, 2004.

Johnson BA & Blevins RA. NMRView: A computer program for the visualization and analysis of NMR data. *Journal of Biomolecular NMR*, 4:603-614, 1994.

Jung YS & Zweckstetter M. MARS - robust automatic backbone assignment of proteins. *Journal of Biomolecular NMR*, 30:11-23, 2004.

Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Ono AM, Güntert P. Optimal isotope labelling for NMR protein structure determination. *Nature*, 440:52-57, 2006.

Kallen J, Goepfert A, Blechschmidt A, Izaac A, Geiser M, Tavares G, Ramage P, Furet P, Keiichi M, Lisztwan J. Crystal Structures of Human MdmX (HdmX) in Complex with p53 Peptide Analogues Reveal Surprising Conformational Changes. *The Journal of Biological Chemistry*, 284:8812-8821, 2009.

Karplus M. In Vicinal Proton Coupling in Nuclear Magnetic Resonance. *Journal of the American Chemical Society*, 85:2870-2871, 1963.

Keeler J & Neuhaus D. Comparison and Evaluation of Methods for Two-Dimensional NMR Spectra with Absorption-Mode Lineshapes. *Journal of Magnetic Resonance*, 63:454-472, 1985.

Keeler J. Understanding NMR Spectroscopy. *Wiley*, 2nd Edition, 2010.

Keller R. *CARA: computer aided resonance assignment.* `http://cara.nmr.ch/`, 2004.

Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181:662-666, 1958.

Kirchner DK & Güntert P. Objective identification of residue ranges for the superposition of protein structures. *BMC Bioinformatics*, 12:170-180, 2011.

Kirchner DK & Güntert P. CYVAL. *In preparation*, 2016.

Kirkpatrick S, Gelatt C Jr, Vecchi MP. Optimization by simulated annealing. *Science*, 220:671-680, 1983.

Kleywegt GJ & Kaptein R. A Versatile Approach toward the Partially Automatic Recognition of Cross Peaks in 2D ^1H NMR Spectra. *Journal of Magnetic Resonance*, 88:601-608, 1989.

Klukowski P, Walczak MJ, Gonczarek A, Boudet J, Wider G. Computer vision-based automated peak picking applied to protein NMR spectra. *Bioinformatics*, 31:2981-2988, 2015.

Koga N, Tatsumi-Koga R, Liu GH, Xiao R, Acton TB, Montelione GT, Baker D. Principles for designing ideal protein structures. *Nature*, 491:222-227, 2012.

Koradi P, Billeter M, Engeli M, Güntert P, Wüthrich K. Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *Journal of Magnetic Resonance*, 135:288-297, 1998.

Koradi P, Billeter M, Güntert P. Point-centered domain decomposition for parallel molecular dynamics simulation. *Computer Physics Communnications*, 124:139-147, 2000.

Kovacs H & Gossert AD. Improved NMR experiments with ^{13}C-isotropic mixing for assignment of aromatic and aliphatic side chains in labeled proteins. *Journal of Biomolecular NMR*, 58:101-112, 2014.

Krähenbühl B, Bakkali IE., 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. *Journal of Biomolcular NMR*, 59:87-93, 2014.

Kuszewski J, Schwieters CD, Garett DS, Byrd RA, Tjandra N, Clore GM. Completely automated, highly error-tolerant macromolecular structure determination from multi-dimensional nuclear Overhauser enhancement spectra and chemical shift assignment. *Journal of the American Chemical Society*, 126:6258-6273, 2004.

Kwab AH, Moblie M, Gooley PR, King GF, Mackay JP. Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS Journal*, 278:687-703, 2011.

Lian LY & Barsukov IL. In *Protein NMR SPectroscopy: Practical Techniques and Applications*, pages 55-77. Wiley, 1st edition, 2011.

Liu Z, Abbas A, Jing BY, Gao X. WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering. *Bioinformatics*, 28:914-920, 2012.

Lorensen WE & Cline HE. Marching Cubes: A high resolution 3D surface construction algorithm. *Computer Graphics*, 21:163-169, 1987.

López-Méndez B & Güntert P. Automated Protein Structure Determination from NMR Spectra. *Journal of the American Chemical Society*, 128:13112-13122, 2006.

López-Méndez B, Pantoja-Uceda D, Tomizawa T, Koshiba S, Kigawa T, Shirouzo 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, Güntert P. Letter to the Editor: NMR assignment of the hypothetical ENTH-VHS domain At3g16270 from Arabidopsis thaliana. *Journal of Biomolecular NMR*, 29:205-206, 2006.

Luginbühl P, Güntert P, Billeter M, Wüthrich K. The new program OPAL for molecular dynamics simulations and energy refinements of biological macromolecules. *Journal of Biomolecular NMR*, 8:136-146, 1996.

Marion D & Wüthrich K. Application of phase sensitive two-dimensional correlated spectroscopy (COSY) for measurements of ^{1}H-^{1}H spin-spin coupling constants in proteins. *Biochemical and Biophysical Research Communications*, 113:967-974, 1983.

Meier BU, Madi ZL, Ernst RR. Computer analysis of nuclear spin systems based on local symmetry in 2D spectra. *Journal of Magnetic Resonance*, 74:565-673, 1987.

Mishra SH, Jarden BJ, Frúh DP. A 3D time-shared NOESY experiment designed to provide optimal resolution for accurate assignment of NMR distance restraints in large proteins. *Journal of Biomolecular NMR*, 60:265-274, 2014.

Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. *Proteins*, 12:345-364, 1992.

Mumenthaler C, Braun W. Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *Journal of Molecular Biology*, 254:465-480, 1955.

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

Munkres J. Algorithms for the assignment and transportation problems. *SIAM Journals*, 5:32-38, 1957.

Nabuurs SB, Spronk CAEM, Krieger E, Maassen GV, Vuister GW. Quantitative Evaluation of Experimental NMR Restraints. *Journal of the American Chemical Society*, 125:12026-12034, 2004.

Nabuurs SB, Krieger E, Spronk CAEM, Nederveen AJ, Vriend G, Vuister GW. Definition of a new information-based per-residue quality parameter. *Journal of Biomolecular NMR*, 33:123-134, 2005.

Nabuurs SB, Spronk CAEM, Vuister GW, Vriend G. Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. *PLoS Computational Biology*, 2:1-9, 2006.

Nilges M. A calculation strategy for the structure determination of symmetric dimers by ^1H NMR. *Proteins*, 17:297-309, 1993.

Nilges M. Calculation of protein structures with ambiguous distance restraints: automated assignment of ambiguous NOE cross-peaks and disulphide connectivities. *Journal of Molecular Biology*, 245:645-660, 1995.

Nilges M, Macias MJ, ODonghue SI, Oschkinat H. Automated NOESY interpretation with ambiguous distance restraints: The refined NMR soluction structure of pleckstrin homology domain from beta-spectrin. *Journal of Molecular Biology*, 269:408-422, 1997.

Orekhov VY, Ibraghimov IV, Billeter M. MUNIN: A new approach to multi-dimensional NMR spectra interpretation. *Journal of Biomolecular NMR*, 20:49-60, 2001.

Orts J, Wältli MA, Marsh M, Vera L, Gossert A, Güntert P, Riek R. NMR-based determination of the 3D structure of the ligand-protein interaction site without protein resonance assignment. *Journal of the American Chemical Society*, 138:4393-4400, 2016.

Pantoja-Uceda D, López-Méndez B, Koshiba S, Kigawa T, Shirouzo 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, Güntert P. Letter to the Editor: NMR assignment of the hypothetical rhodanese domain At4g01050 from Arabidopsis thaliana. *Journal of Biomolecular NMR*, 29:207-208, 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, Güntert P. Solution structure of the rhodanese homology domain At4g01050(175-295) from Arabidopsis thaliana. *Protein*, 14:224-230, 2005

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes. The art of scientific computing. *Cambridge University Press, Cambridge*, third edition, 2007.

Rabi II, Zacharias, JR, Millman S, Kusch P. A New Method of Measuring Nuclear Magnetic Moment. *Physical Review*, 53: 318-327, 1938.

Reckel S, Gottstein D, Stehle J, Löhr F, Verhoefen MK, Takeda M, Silvers R, Kainosho M, Glaubitz C, Wachtveitl J, Bernhard F, Schwalbe H, Güntert P, Dötsch V. Solution NMR structure of proteorhodopsin. *Angewandte Chemie International Edition*, 50:11942-11946, 2011.

Rohl CA & Baker D. De nove determination of protein backbone structures from residual dipolar coupling using Rosetta. *Journal of the American Chemical Society*, 124:2723-2729, 2002.

Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, Doreleijers JF, Eletsky A, Giachetti A, Guerry P, Güntert P, He YF, Herrmann T, Huang YJ, Jaravine V, Jonker HRA, Kennedy MA, Lange OF, Liu GH, Malliavin TE, Mani R, Mao BC, Montelione GT, Nilges M, Rossi P, van der Schot G, Schwalbe H, Szyperski TA, Vendruscolo M, Vernon R, Vranken WF, de Vries S, Vuister GW, Wu B, Yang YH, Bonvin AMJJ. Blind testing of routine, fully automated structure determination by NMR data. *Structure*, 20:227-236, 2012.

Rosato A, Bagaria A, Baker D, Bardiaux B, Cavalli A, Doreleijers JF, Giachetti A, Guerry P, Güntert P, Herrmann T, Huang YJ, Jonker HRA, Mao B, Malliavin TE, Montelione GT, Nilges M, Raman S, van der Schot G, Vranken WF, Vuister GW, Bonvin AMJJ. CASD-NMR: critical assessment of automated structure determination by NMR. *Nature Methods*, 6:625-626, 2009.

Rosato A, Vranken W, Fogh RH, Ragan TJ, Tejero R, Pederson K, Lee HW, Prestegard JH, Yee A, Wu B, Lemak A, Houliston S, Arrowsmith CH, Kennedy M, Acton TB, XIAO R, Liu GH, Montelione GT, Vuister GW. The second round of Critical Assessment of Automated Structure Determination of Proteins by NMR: CASD-NMR-2013. *Journal of Biomolecular NMR*, 62:413-424, 2015.

Sanchez MC, Renshaw JG, Davies G, Barlow PN, Vogtherr M. MDM4 binds ligands via a mechanism in which disordered regions become structured. *FEBS Letters*, 584:3035-3041,2010.

Schmidt E & Güntert P. A new algorithm for reliable and general NMR resonance assignment. *Journal of the American Chemical Society*, 134:12817-12829, 2012.

Schmidt E & Güntert P. Reliability of exclusively NOESY-based automated resonance assignment and structure determination of proteins.. *Journal of Biomolecular NMR*, 57:193-214, 2013.

Schmidt E, Gath J, Habenstein B, Ravotti F, Szekely K, Huber M, Buchner L, Böckmann A, Meier BH, Güntert P. Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. *Journal of Biomolecular NMR*, 56:234-54, 2013.

Schmidt E & Güntert P. In *Structural Proteomics: High-Throughput Methods, Methods in Molecular Biology*, chapter 16, pages 303-329, Springer, 2nd edition, 2015.

Takeda M & Kainosho M. In *Protein NMR Spectroscopy: Practical Techniques and Applications*, pages 23-53. Wiley, 1st edition, 2011.

Schwieters CD, Kuszweski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. *Journal of Magnetic Resonance*, 160:65-73, 2003.

Scott A, Pantoja-Uceda D, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzo M, Tanaka A, Sugano S, Yokoyama S, Güntert P. NMR assignment of the SH2 domain from the human feline sarcoma oncogenes FES. *Journal of Biomolecular NMR*, 30:463-464, 2004.

Scott A, Pantoja-Uceda D, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzo M Tanaka A, Sugano S, Yokoyama S, Güntert P. Solution structure of the Src homology 2 domain from the human feline sarcoma oncogenes FES. *Journal of Biomolecular NMR*, 31:357-361, 2005.

Serrano P, Pedrini B, Mohanty B, Geralt M, Herrmann T, Wüthrich K. The J-UNIO protocol for automated protein structure determination by NMR in solution. *Journal of Biomolecular NMR* 53:341-354, 2012.

Shannon CE. A mathematical theory of communication. *The Bell System Technology Journal*, 27:279-309, 1948.

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A. Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences*, 105:4685-4690, 2008.

Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of Biomolecular NMR*, 44:213-223, 2009.

Shimrat M. Position of point relative to polygon. *Communications ACM*, 5:434, 1962.

Shin J, Lee W, and Lee W. Structural proteomics by NMR spectroscopy. *Expert Review of Proteomics*, 5:589-601, 2008.

Shvarts A, Steegenga WT, Riteco N, van Laar T, Dekker P, Bazuine M, van Ham RC, van der Houven, van Oordt W, Hateboer G, van der Eb AJ, Jochemsen AG. MDMX: a novel p53-binding protein with some functional properties of MDM2. *EMBO J*, 1:5349-5357, 1996.

Silver R. An Algorithm for the assignment problem. *Comm ACM*, 3:605-606, 1960.

Skinner SP, Fogh RH, Boucher W, Ragan TJ, Mureddu LG, Vuister GW. CcpNmr AnalysisAssign: a flexible platform for integrated NMR analysis. *Journal of Biomolecular NMR*, 2016.

Skolnick J, Kolinksi A, Ortiz AR. MONSSTER: A Method for Folding Globular Proteins with a Small Number of Distance Restraints. *Journal of Molecular Biology*, 256:217-241, 1997.

States DJ, Haberkorn RA, Ruben DJ. A two-dimensional nuclear Overhauser experiment with pure absorption phase in four quadrants. *Journal of Magentic Resonance*, 48:286-292, 1982.

Tjandra N & Bax A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science*, 278:1111-1114, 1997.

Takeda M & Kainosho M. In *Protein NMR Spectroscopy: Practical Techniques and Applications*, pages 23-53. Wiley, 1st edition, 2011.

Theodorou DN & Suter UW. Shape of Unperturbed Linear Polymers: Polypropylene. *Macromolecules*, 18:1206-1214, 1985.

Tikole S, Jaravine V, Rogov V, Dötsch V, Güntert, P. Peak picking NMR spectral data using non-negative matrix factorization. *BMC Bioinformatics*, 15:46, 2014.

Toledo F & Wahl GM.. Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nature Reviews Cancer*, 6:909-923, 2006.

Toledo F & Wahl GM. MDM2 and MDM4: p53 regulators as targets in anticancer therapy. *The International Journal of Biochemistry and Cell Biology*, 39:1476-1482, 2007.

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livnby M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao HY, Markley JL. BioMagResBank. *Nucleic Acids Research*, 36:D402-D408, 2008.

van der Schot G, Zhang Z, Vernon R, Shen Y, Vranken WF, Baker D, Bonvin AMJJ, Lange OF. Improving 3D structure prediction from chemical shift data. *Accounts of Chemical Research*, 42:1545-1553, 2013.

Volk J, Herrmann T, Wüthrich K. Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. *Journal of Biomolecular NMR*, 41:127-138, 2008.

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinás M, Ulrich EL, Markley JL, Ionides J, Laue ED. The CCPN Data Model for NMR Spectroscopy: Development of a Software Pipeline. *PROTEINS: Structure, Functions, and Bioinformatics*, 59:687-696, 2005.

Vuister GW, Tjandra N, Shen Y, Grishaev A, Grzesiek S. In *Protein NMR Spectroscopy: Practical Techniques and Applications*, pages 83-268. Wiley, 1st edition, 2011.

Wade M & Li YC. MDM2, MDMX and p53 in oncogenesis and cancer therapy. *Nature Reviews Cancer*, 13:83-96, 2013.

Wagner G & Wüthrich K. Sequential resonance assignments in protein ^{1}H nuclear magnetic resonance spectra: Basic pancreatic trypsin inhibitor. *Journal of Molecular Biology*, 155:347-366, 1982.

Williamson MP & Craven CJ. Automated protein structure calculation form NMR data. *Journal of Biomolecular NMR*, 43:131-143, 2009.

Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G. CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucleic Acids Research*, 36:496-502, 2008.

Wittekind M & Müller L. HNCACB, a High-Sensitivity 3D NMR Experiment to Correlate Amide-Proton and Nitrogen Resonances with the Alpha- and Beta-Carbon Resonances in Proteins. *Journal of Magnetic Resonance*, B 101:201-205, 1993.

Wüthrich K. NMR of Proteins and Nucleic Acids. *Wiley, New York*, 1986.

Wüthrich K (2003) NMR studies of structure and function of biological macromolecules (Nobel Lecture). *Journal of Biomolecular NMR*, 27:13-39, 2003.

Würz JM & Güntert P. Peak picking in multidimensional NMR spectra with the contour geometry based algorithm CYPICK. *Journal of Biomolecular NMR*, accepted, 2016.

Wunderlich Z, Acton TB, Lliu J, Kornhaber G, Everett J, Carter P, Lan Ning, Chols N, Gerstein M, Rost B, Montelione G. The protein target list of the Northeast Structural Genomics Consortium. *Proteins: Structure, Function, and Bioinformatics*, 56:181-187, 2004.

Yilmaz E & Güntert P. NMR structure calculation for all small molecule ligands and non-standard residues from the PDB Chemical Component Dictionary. *Journal of Biomolecular NMR*, 63:21-37, 2015.

Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien C, Powers R, and Montelione GT. Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology*, 269:592-610, 1997.

List of abbreviations

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
4D	four-dimensional
ATNOS	AuTomated NOeSy peak picking
AURELIA	AUtomated REsonance LIne Assignment
AUTOPSY	AUTOmated Peak picking for NMR SpectroscopY
BMRB	BioMagResBank
CANDID	Combined Automated NOE assignment and structure DeterminatIon moDule
CAPP	Contour Approach to Peak Picking
CASD-NMR	Critical Assessment of automated Structure Determination of proteins from NMR data
CCPN	Collaborative Computing Project for the NMR Community
CYANA	Combined assignment and dYnamics Algorithm for NMR Applications
DOCR	Database Of Converted Restraints
Eq.	Equation
FID	Free Induction Decay
Fig.	Figure
FRED	Filtered Restraints Database
HSQC	Heteronuclear Single-Quantum Coherence
IUPAC	International Union of Pure and Applied Chemistry
MD	Molecular Dynamics
MR	Molecular Replacement
NESG	NorthEast Structural Genomics consortium
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Enhancement
NOESY	Nuclear Overhauser Enhancement SpectroscopY

PDB	Protein Data Bank
ppm	Parts Per Million
PRE	Paramagnetic Relaxation Enhancement
RDC	Residual Dipolar Coupling
RMSD	Root Mean Square Deviation
QUEEN	QUantitative Evaluation of Experimental NMR restraints
STELLA	Synergistic approach toward The Evaluation of Local maxima in Low-symmetry spectrA
SA	Simulated Annealing
SBDD	Structure-Based Drug Desgin
Tab.	Table
Tcl	Tool Command Language
TMS	Tetramethylsilane
TPPI	Time-Proportional Phase Incrementation

Publications

- Würz JM & Güntert P. Peak picking in multidimensional NMR spectra with the contour geometry based algorithm CYPICK. *Journal of Biomolecular NMR*, in press, 2016.
- Würz JM, Buchner L & Güntert P. Information content of NMR distance restraints. *In preparation*, 2016.
- Gossert A, Würz JM, Schmidt E, Henry C, Widmer A & Güntert P. Structure based drug design by NMR *In preparation*, 2016.
- Kirchner D, Würz JM & Güntert P. Spectra-based validation of protein structures determined by NMR. *In preparation*, 2016.

Curriculum Vitae

Julia Maren Würz geb. Weber

Personal information

Date of birth	16 August 1986
Place of birth	Weilburg (Hessen), Germany
Nationality	German

Education

2012-present	PhD candidate in biochemistry, Goethe-University, Frankfurt am Main, supervised by Prof. Dr. Peter Güntert
02/2016-03/2016	Research Period, laboratory of Prof. Dr. Yutaka Ito, Tokyo Metropolitan University, Minami-Osawa, Japan
11/2011	Diploma in biochemistry, Goethe-University, Frankfurt am Main. Thesis: "Robust automated distance restraint assignment for solid-state and solution NMR structure calculation", supervised by Prof. Dr. Peter Güntert
09/2010-12/2010	Internship, laboratory of Prof. Dr. Chad Rienstra, University of Illinois, Urbana Champaign, USA
2006-2011	Studies in Biochemistry and Biophysical Chemistry, Goethe University, Frankfurt am Main
2006	Abitur, Gymnasium Phillippinum, Weilburg
1997-2006	Gymnasium Philippinum, Weilburg

Danksagung

Ich möchte mich bei Prof. Dr. Peter Güntert für die Betreuung meiner Arbeit, die hilfreichen Ratschläge sowie die Möglichkeit komplett unabhängig zu arbeiten bedanken. Mein Dank gilt außerdem Prof. Dr. Volker Dötsch für die Begutachtung dieser Arbeit. Danke auch an Dr. Alvar Gossert und Dr. Elena Schmidt für die Zusammenarbeit an dem SBDD Projekt. Bei Prof. Dr. Yutaka Ito und Dr. Teppei Ikeya möchte ich mich für einen sehr spannenden und lehrreichen Aufenthalt in Japan bedanken. Für das Korrekturlesen meiner Arbeit möchte ich mich bei Dr. Daniel Gottstein, Dr. Sina Kazemi und Dr. Elena Schmidt bedanken. Danke auch für die nette und hilfsbereite Arbeitsatmosphäre und nochmal ein besonderer Dank an Sina für die vielen technischen Hilfestellungen. Mein Dank gilt natürlich auch allen anderen Mitgliedern und ehemaligen Mitgliedern der Güntert Gruppe, die diese Zeit zu einem besonderen Erlebnis gemacht haben. Zum Schluss möchte ich mich auch bei meinen Freunden und meiner Familie für viel Liebe und Unterstützung bedanken.