Analysis of the precision and accuracy of protein structures determined by NMR

Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften

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

von

Donata Katharina Kirchner

aus Langen (Hessen)

Frankfurt am Main 2016

(D 30)

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

Dekan:	Prof. Dr. Michael Karas
1. Gutachter:	Prof. Dr. Peter Güntert
2. Gutachter:	Prof. Dr. Volker Dötsch
Datum der Disputation:	04.08.2016

What you do makes a difference, and you have to decide
what kind of difference you want to make.

Jane Goodall, born 1934

Summary

This thesis is concerned with protein structures determined by nuclear magnetic resonance (NMR), and the text focuses on their analysis in terms of accuracy, gauged by the correspondence between the structural model and the experimental data it was calculated from, and in terms of precision, i.e. the degree of uncertainty of the atomic positions. Additionally, two protein structure calculation projects are described.

Various approaches to the validation of protein structures determined by NMR and to the estimation of their precision have been developed, but, as yet, no method has won unequivocal support of the scientific community. The precision of an ensemble of structures is commonly reported as the root mean square deviation (RMSD) of the conformers, but even so a number of issues remain, most notably the question of which residues or atoms to superimpose for RMSD calculation.

Many structure bundles contain not only well-defined but also less well or even ill-defined segments. Including these regions in the RMSD calculation for reporting the precision of the ensemble may obscure the presence of the well-defined structural features. There are various approaches to how these ordered parts might be defined and identified, but despite—or due to—the existence of a number of software tools that use different methods and produce different results there was no widely employed and at the same time reliable and reproducible method of deciding what parts of a structure to superimpose. In order to fill this gap we have created the CYRANGE software (**?**).

CYRANGE takes a protein structure bundle and identifies its RMSD-stable domains, i.e. the parts of the structure that can be simultaneously superimposed with a low backbone RMSD value. This means that CYRANGE successfully recognizes segments of the structure that exhibit a high degree of order but that cannot be meaningfully superimposed at the same time as another such part within the same ensemble. The software accomplishes this by combining two approaches of identifying ordered regions, the calculation of dihedral angle order parameters and the construction of an inter-atomic distance variance matrix (DVM).

Dihedral angle order parameters, also known as angular order parameters, are a measure of local order. They state how strongly a given dihedral angle differs over a set of conformers—

naturally, small differences indicate a high level of local order. Their core drawback is their inability to identify regions of high-level order, which are not necessarily adjacent within the amino acid sequence but which nonetheless belong to the same RMSD-stable domain. This implies that methods making use of dihedral angle order parameters alone will invariably fail to recognize the presence of different ordered domains within an ensemble.

The DVM, on the other hand, is a tool by which global order can be identified. It contains the variances of the intra-conformer distances between atom pairs within the query structure bundle. Thus it helps pick out atom pairs that lie within the same ordered region of the ensemble, even if those atoms belong to residues lying far apart within the sequence. We noted that a pre-selection of atoms to be included in the DVM through the application of an angular order parameter threshold has a beneficial effect on the domain identification results. Therefore we have combined the two approaches in the CYRANGE algorithm.

The entire domain identification method starts by the computation of angular order parameters, which are used for the selection of *core atoms*. These are defined as the C^α atoms of those residues that contain at least one torsion angle with an order parameter value larger than a cutoff value calculated from the query protein's set of order parameters. The DVM is constructed from distances between the core atoms within the conformers that make up the ensemble. Based on this matrix the core atoms are clustered according to the similarity of the inter-atomic distance variances, with the best clustering stage representing the RMSD-stable domains of the structure bundle.

These domains are further refined in order to yield residue ranges comprising as many residues and as few intra-domain gaps as possible while still maintaining a low RMSD value. To this end the algorithm iteratively removes residues until the group of surviving residues remains stable. In each cycle the residues with the least consistent positions over the set of conformers are identified. CYRANGE calculates the RMSD decrease after the trial elimination of each residue, and the one whose removal will yield the largest RMSD decrease is eliminated if certain conditions are fulfilled. Then the procedure starts anew. Once the set of residues no longer changes small intra-domain gaps are filled, and the resulting residue ranges are output to the user.

CYRANGE is available at `http://www.bpc.uni-frankfurt.de/cyrange.html`. There the user may either upload a structure bundle for online residue range determination or download a stand-alone version of the software. In the case of both online and locally performed domain identification the user requires nothing but a file containing an ensemble of protein conformers because even though CYRANGE takes a number of modifiable parameters no individual adjustment of their values is necessary. The calculation itself takes at most a few seconds. For each domain the software outputs the identified residue ranges, their backbone RMSD value to the mean structure, the number of intra-domain gaps, and the total number of residues making up the domain.

Using a test set of 37 proteins we have shown that CYRANGE reliably and in straightforward as well as challenging cases outputs domains that are free of unnecessary small gaps and comprise the correctly identified ordered regions. The individual superpositions of these domains allow a clear presentation of the structure and provide the user with RMSD values that are a meaningful measure of the precision of the structure bundle. The software successfully handles both single and multi-domain proteins, protein complexes, and even structure bundles that are loosely defined on the global level.

We compared CYRANGE to two other well-known software tools that also identify ordered regions within protein structure bundles, FindCore (**?**) and PSVS (**?**). Due to its reliance on angular order parameters alone the latter is unable to recognize the presence of several distinct domains, a limitation that does not hold true for the FindCore algorithm, which, like CYRANGE, includes the construction of a DVM. In the majority of cases the CYRANGE ranges contained fewer gaps and covered significantly larger parts of the sequence than those determined by FindCore and PSVS. Consequently, the RMSD values of the CYRANGE domains were often slightly higher in comparison, but there were also cases in which CYRANGE's residue ranges both contained a larger number of residues and could be superimposed with a lower RMSD value than the results of the other two programs. On average, CYRANGE-reported ranges attained a sequence coverage of 85 % and led to a backbone RMSD value of 0.77 Å, as compared to 67 % coverage and 1.72 Å RMSD with PSVS, and 58 % coverage and 0.73 Å RMSD with FindCore. This demonstrates that CYRANGE succeeds better than the other two programs in creating a favourable balance between sequence coverage and low backbone RMSD.

CYRANGE has been officially recommended by the NMR-validation task force of the Protein Data Bank as the software of choice for the determination of the RMSD-stable domains of a structure bundle (**?**). Moreover, the program was employed to determine ordered residue ranges in the analysis of the structural quality of 2013's CASD-NMR entries (**??**), and it has been included in a version of the structure validation tool MolProbity (MolProbity-HTC; **?**). All in all, the high quality of the CYRANGE results as well as the program being both actively used and officially recommended clearly demonstrate that CYRANGE serves the purpose it has been designed for, and that it serves it well.

The second core project of this thesis deals with the validation of protein structures determined by NMR, a task that is more challenging than the validation of structures determined using X-ray crystallography. The latter method's diffraction data contain a certain degree of redundancy, hence they can be split into a working set for structure determination and a test set for structure validation without compromising structural quality. The structure can then easily be cross-validated against the test set. However, NOESY data, which are commonly the main source of information for NMR structure determination, do not share this property to the same extent.

This central problem of NMR structure validation has led to the creation of a number of knowledge-based tools that aim to assess the quality of a structure by comparing certain structural features to values derived from the analysis of a large number of known protein structures. The drawback of these methods is their inability to recognize correctly formed but unusual features within a structure as these are unlikely to be represented in the data the tools draw upon. Furthermore, such programs do not take the experimental data into account that were used to determine the structure of the query protein, hence they cannot state how well this model actually represents these data. For this reason, knowledge-based programs cannot be expected to be able to estimate the degree of accuracy of the structure, yet this information would be highly valuable to anyone wishing to use the model for their research. Validation tools that consider the experimental data, on the other hand, are rare.

Users would benefit from a tool that validates the structure against the raw experimental data and that requires as little user interaction and adjustment as possible. This is exactly what CYVAL, the validation method that we have developed, aims to provide.

CYVAL is integrated into the structure calculation software CYANA (**??**), and it assesses structural accuracy via the correspondence between the query structure and its original NOESY spectra. To this end, the CYVAL procedure calculates a validation score (ζ) from weighted distance deviation penalties. To simplify its interpretation ζ has been designed to fall into the range between 0.0 and 1.0, with lower values signifying a higher degree of agreement between structure and data.

The program takes a processed NOESY spectrum, the structural model, and a list of assigned chemical shifts. From the latter two files CYVAL generates a set of expected NOESY peaks, and in the spectrum peaks are picked using the peak picking routine built into CYANA. The two sets of signals, predicted and experimental, are matched using a component of the Peakmatch algorithm (**?**), a comparison that yields three distinct peak classes: matching peaks, which are both predicted based on the structure and retrieved from the spectrum; missing peaks, which were predicted but are absent from the list of picked signals; and unpredicted peaks, which are present in the spectrum but cannot be explained by the structural model.

Each peak has an associated inter-proton distance, which stems from the structure in the case of predicted peaks, or which is calibrated from the signal intensities for the picked peaks. Due to the d^{-6}-relationship between NOESY peak intensity and inter-proton distance CYVAL performs the correspondence check on the distance level, since comparatively large errors in peak heights translate to relatively small distance deviations ($|\Delta d|$). Hence the risk of overestimating errors is lower when distances are compared. In the case of peaks from the *matching* category $|\Delta d|$ is the difference between the experimental and the structure-based distance; distances of peaks from the other two categories are compared with $d^{\max}$, the NOE distance threshold.

To restrict ζ to the interval [0.0,1.0], and also in order not to put too pronounced a stress on either very small or very large distance deviations, a sigmoid penalty function with a maximum value of 1.0 is applied to $|\Delta d|$ before it contributes to ζ. Furthermore, not all peaks are equally relevant for structure validation, consequently each peak is assigned a weight composed of different components. Short-range peaks, for example, which are likely to be present for basically any conformation of a given amino acid sequence, do not contribute much valuable information. Long-range peaks, in contrast, are crucial for determining the overall fold of a protein and thus also provide important information on the degree of correspondence between structure and spectrum. On the other hand, weak peaks, which yield long distances, have a higher probability of being noise peaks, and since noise should not disproportionately influence ζ, peak weights need to balance the different factors.

In order to obtain a per-peak measure of structural accuracy ζ_n can be determined, which is obtained by normalizing ζ by the number of peaks that were used in its calculation.

Five proteins for which a reference structure, a list of assigned chemical shifts, and processed NOESY spectra were available, were employed to test the validation power of ζ and ζ_n. The metric used to gauge the performance of CYVAL was the backbone RMSD value to the reference structure ($\mathrm{RMSD}_\mathrm{ref}$), using only the ordered residues of the reference as determined by CYRANGE. For each of the test proteins a number of structures at different degrees of accuracy were generated from successively pruned peak lists. These incorrect structures were given as input to CYVAL, together with the unmodified list of assigned chemical shifts and the processed spectrum. The resulting values of ζ or ζ_n were plotted against $\mathrm{RMSD}_\mathrm{ref}$ for visual inspection, and the correlation coefficient between the two quantities was calculated. A linear relationship between ζ or ζ_n and $\mathrm{RMSD}_\mathrm{ref}$ is not necessarily to be expected, hence correlation coefficients on their own are not a sufficient basis for evaluating the validation power, and they can even be deceptive as a few outliers may obscure an otherwise promising result. Nevertheless, they can provide a valuable first estimate of the performance of CYVAL, and they may be used to quantitatively compare validation runs if the differences are sufficiently large. Additionally, various penalty functions, weighting factor calculation methods, and their combinations were evaluated, as well as several methods of selecting which peaks would contribute to the final value of ζ.

The test results have been encouraging. Although there was no discernible relationship between ζ and $\mathrm{RMSD}_\mathrm{ref}$ for one of the test proteins, in the case of two other proteins CYVAL was able to successfully differentiate between structures of high (below 2.0 Å $\mathrm{RMSD}_\mathrm{ref}$), medium (between 2.0 and 4.0 Å), and low quality (above 4.0 Å). In the remaining two cases the expected increase in ζ values for structures of diminished accuracy was clearly present, despite a somewhat blurred distinction between the quality classes. On the whole, the validation power of ζ_n was slightly superior to that of ζ, but it was observed that both metrics depended on the $\mathrm{RMSD}_\mathrm{ref}$ value as well as on the query protein and the spectrum type used for validation.

During the evaluation of alternative penalty functions and weight modes no method emerged that was clearly superior to all other approaches. The choice of which groups of peaks to consider in the calculation of ζ, however, proved to have a potentially large influence on the success of the validation attempts. We observed that the inclusion of peaks that, individually, contribute little information to the validation procedure was nevertheless important for the validation power of CYVAL.

We examined the performance of three knowledge-based structure validation tools (ProSa2003 (**?**), Verify3D (**???**), and MolProbity (**??**)) using our protein test sets. Even though these programs had no access to the experimental data of the proteins there were cases in which they, too, were able to distinguish between the three classes of structural accuracy. The validation power of CYVAL was similar to and at times better than the performance of the three external programs.

Some final obstacles still need to be overcome in order to render the CYVAL method equally applicable to all cases. Overall, however, the procedure has clearly demonstrated its ability to distinguish between structures of different degrees of accuracy, and its performance is similar or even superior to that of a number of well-known validation programs. The combined results illustrate that CYVAL has the potential to become a valuable tool for the validation of protein structures determined by NMR.

In addition to the analysis of protein structures I was also involved in their determination using liquid-state NMR data, insofar as I performed three structure calculations. The proteins in question were the wild-type Trp-Trp binding module (WW) domain of the peptidyl-prolyl cis/trans isomerase Pin1 as well as the phospho-mimic Ser16Glu-mutant of this domain ($\mathrm{WW}^{\mathrm{S16E}}$) (**?**), and the Ser45Ala-mutant of TycC3_PCP, which is the third peptidyl carrier domain of the tyrocidine A synthetase subunit C from *B. brevis* (**?**). Pin1 is overexpressed in a number of human cancers, and high levels of this protein phosphorylated at Ser16 have been found in brain tissue samples of Alzheimer's disease patients. TycC3_PCP, whose apo state is mimicked by TycC3_PCP(S45A), is a representative of the ubiquitous group of carrier proteins. Elucidating the structure of these proteins, both on their own and in complex with their reaction partners, can further the understanding of the biosynthetic pathways in which they are involved and may also help to discover new drug targets.

In all three cases the structure calculation yielded a well-defined ensemble, with all backbone dihedral angles lying inside the allowed areas of the Ramachandran plot, and with high percentages of both assigned NOESY cross peaks and long-range distance restraints, which are particularly valuable for the determination of the tertiary structure. No distance restraint violations above 0.2 Å and no dihedral angle restraint violations greater than 5° occurred. The calculated knowledge-based validation scores show that all three structural models exhibit high geometric quality, and the application of CYVAL suggests that they are in good agreement with the experimental data as well.

A comparison of the wild-type WW domain of Pin1 and $\mathrm{WW}^{\mathrm{S16E}}$ showed a highly similar backbone fold. The surface charge distribution at the binding site, however, was significantly

changed due to the introduction of the negatively charged Glu sidechain, and the bulkiness of this residue furthermore led to a steric hindrance for protein-protein interactions. Both observations explain why the regulatory phosphorylation of Pin1 at Ser16 of the WW domain abrogates the ability of the protein to interact with potential binding partners. The NMR structure of TycC3_PCP(S45A) in solution suggests that, unlike previously postulated, apo-TycC3_PCP exists in the same conformation as the holo state. Moreover, a crystal structure, solved by **?**, has shown that this is also the conformation TycC3_PCP(S45A) adopts in complex with the phosphopantetheine transferase (PPT) Sfp. The similarity of this structure with the complex of human acyl carrier protein and PPT suggests similarities in the recognition and posttranslational modification of carrier proteins in various organisms, which could have negative implications on the usability of bacterial PPTs as targets for the development of novel antibiotics.

Zusammenfassung

Diese Dissertation befasst sich mit Proteinstrukturen, die mithilfe von kernmagnetischer Resonanzspektroskopie (*nuclear magnetic resonance*, NMR) bestimmt wurden. Der Text konzentriert sich insbesondere auf die Analyse der Präzision dieser Strukturen, also die Streubreite der Atompositionen, und die strukturelle Genauigkeit. Letztere wird gemessen am Grad der Übereinstimmung des Strukturmodells mit den experimentellen Daten, die zu seiner Bestimmung verwendet wurden. Darüber hinaus werden zwei Strukturbestimmungsprojekte beschrieben.

Es gibt eine Vielzahl von Methoden zur Validierung mittels NMR bestimmter Proteinstrukturen und zur Abschätzung ihrer Präzision, bisher hat sich allerdings keine dieser Methoden fest etablieren können. Die Präzision eines Ensembles wird für gewöhnlich als die Wurzel der mittleren quadratischen Abweichung (*root mean square deviation*, RMSD) der Konformere angegeben, was allerdings nicht alle Probleme löst. Die Frage, welche Reste oder Atome zur Berechnung des RMSD-Werts überlagert werden sollen, ist beispielsweise noch offen.

Viele Strukturbündel enthalten nicht nur wohldefinierte, sondern auch nur mittelmäßig oder gar schlecht definierte Bereiche. Werden diese Regionen bei der RMSD-Berechnung zur Präzisionsangabe berücksichtigt, kann dies das Vorhandensein der wohldefinierten Strukturabschnitte verschleiern. Es gibt unterschiedliche Vorgehensweisen zur Definition und zum Auffinden der geordneten Bereiche, aber trotz—oder wegen—der Existenz verschiedener Computerprogramme, die unterschiedliche Methoden anwenden und unterschiedliche Ergebnisse produzieren, gab es keine weitläufig verwendete und gleichzeitig zuverlässige und reproduzierbare Methode, um die zu überlagernden Strukturbereiche zu bestimmen. Um diese Lücke zu füllen, haben wir die Software CYRANGE (**?**) entwickelt.

CYRANGE identifiziert die RMSD-stabilen Domänen in einem Proteinstrukturbündel, also die Abschnitte der Struktur, die gleichzeitig unter Erzielung eines niedrigen RMSD-Werts des Proteinrückgrats überlagert werden können. Dies bedeutet, dass CYRANGE erfolgreich Teile der Struktur erkennt, die einen hohen Grad an Ordnung aufweisen, die jedoch nicht sinnvoll zur gleichen Zeit wie andere analoge Bereiche des Strukturensembles überlagert werden können. Das Programm erreicht dies durch die Kombination zweier Ansätze zur Auffindung geordneter Regionen:

die Berechnung von Torsionswinkel-Ordnungsparametern und die Aufstellung einer inter-atomaren Distanzvarianzmatrize (DVM).

Torsionswinkel-Ordnungsparameter sind ein Maß für die lokale Ordnung. Sie geben den Grad der Streuung eines bestimmten Torsionswinkels innerhalb des Strukturbündels an. Hierbei weisen niedrige Werte auf eine hohe lokale Ordnung hin. Das Kernproblem dieser Ordnungsparameter ist ihr Unvermögen, auf höherer Ebene geordnete Bereiche zu identifizieren, die nicht notwendigerweise in der Aminosäuresequenz aneinander grenzen, die aber trotzdem zur selben RMSD-stabilen Domäne gehören. Folglich sind Methoden, die ausschließlich Torsionswinkel-Ordnungsparameter verwenden, nicht in der Lage, unterschiedliche geordnete Domänen innerhalb eines Ensembles voneinander zu trennen.

Die DVM hingegen ist ein Werkzeug, mithilfe dessen globale Ordnung identifiziert werden kann. Sie enthält die Varianzen der intra-Konformer-Abstände von Atompaaren innerhalb des Strukturbündels. Daher ist es möglich, mit ihr als Basis Atompaare zu identifizieren, die zur selben geordneten Region des Ensembles gehören, selbst wenn diese Atome von Resten stammen, die einen großen Abstand innerhalb der Sequenz aufweisen. Es zeigte sich, dass sich eine Vorselektion der in die DVM aufzunehmenden Atome mit einem mindestens zu erreichenden Ordnungsparameterwert als Auswahlkriterium positiv auf die Erfolgsrate bei der Domänenbestimmung auswirkt. Aus diesem Grund haben wir im CYRANGE-Algorithmus beide Ansätze vereint.

Die Vorgehensweise zur Domänenidentifizierung beginnt mit der Berechnung der Torsionswinkel-Ordnungsparameter, die zur Auswahl der *zentralen Atome* verwendet werden. Letztere sind die C$^{\alpha}$-Atome derjenigen Reste, die mindestens einen Torsionswinkel enthalten, dessen Ordnungsparameter über einer Schwelle liegt, die auf Basis aller Torsionswinkel-Ordnungsparameter des Proteins berechnet wurde. Die DVM wird aus den Distanzen zwischen den zentralen Atomen innerhalb der Konformere des Ensembles aufgebaut. Basierend auf der Ähnlichkeit der in der Matrize gespeicherten Distanzvarianzen werden die zentralen Atome Clustern zugeordnet. Hierbei gibt die beste Stufe des Clusterings die RMSD-stabilen Domänen des Strukturbündels an.

Die Domänen werden anschließend weiter verfeinert, damit die schlussendlich ausgegebenen Bereiche so viele Reste und so wenige interne Lücken wie möglich enthalten und gleichzeitig einen niedrigen RMSD-Wert aufweisen. Zu diesem Zweck entfernt der Algorithmus schrittweise Aminosäuren, bis die Gruppe der noch vorhandenen Reste stabil bleibt. In jedem Zyklus werden die Reste identifiziert, deren Positionen innerhalb des Ensembles am stärksten variieren. CYRANGE berechnet den RMSD-Abfall nach dem testweisen Ausschluss eines jeden dieser Reste, und derjenige, dessen Entfernung die größte Abnahme des RMSD-Werts bedingt, wird entfernt, sofern bestimmte zusätzliche Bedingungen erfüllt sind. Anschließend beginnt der Prozess von vorn. Sobald sich die Aminosäuregruppe nicht mehr verändert, werden kleine domäneninterne Lücken geschlossen und die endgültigen Sequenzabschnitte werden vom Programm ausgegeben.

CYRANGE kann über `http://www.bpc.uni-frankfurt.de/cyrange.html` verwendet und auch bezogen werden. Nutzerinnen und Nutzer können dort entweder ein Strukturbündel hochladen und online die Domänenbestimmung durchführen lassen oder eine Kopie der Software zur Verwendung auf einem lokalen Rechner herunterladen. Bei der lokalen wie auch bei der Online-Berechnung wird nichts außer einer Datei benötigt, die die Koordinaten des Proteinstrukturensembles enthält. Denn obwohl CYRANGE über mehrere Parameter verfügt, die bei der Verwendung des Programms verändert werden können, ist keine individuelle Einstellung ihrer Werte vonnöten. Die Rechnung selbst dauert nur wenige Sekunden. Für jede gefundene Domäne gibt das Programm die identifizierten Sequenzbereiche aus und darüber hinaus den bei ihrer Überlagerung enthaltenen Rückgrat-RMSD-Wert zur mittleren Struktur, die Anzahl der domäneninternen Lücken, und die Gesamtanzahl der Reste, aus denen die Domäne besteht.

Anhand eines aus 37 Proteinstrukturen bestehenden Testsets konnten wir zeigen, dass CYRANGE zuverlässig und sowohl in einfachen als auch in anspruchsvollen Fällen Domänen identifiziert, die frei von unnötigen kleinen Lücken sind und die gleichzeitig die korrekt erkannten geordneten Bereiche enthalten. Die individuelle Überlagerung dieser Domänen erlaubt eine klare Präsentation der Struktur und bildet eine Basis für RMSD-Werte, die eine Aussage über die Präzision des Ensembles ermöglichen. Das Programm analysiert erfolgreich sowohl Einzel- als auch Multidomänenstrukturen, Proteinkomplexe und sogar auf globaler Ebene nur lose definierte Strukturbündel.

Wir haben CYRANGE mit zwei anderen bekannten Computerprogrammen verglichen, die ebenfalls geordnete Bereiche in Strukturbündeln von Proteinen identifizieren, FindCore (**?**) und PSVS (**?**). Da letzteres ausschließlich Torsionswinkel-Ordnungsparameter verwendet, ist es nicht in der Lage, das Vorhandensein mehrerer unterschiedlicher Domänen zu erkennen. Der FindCore-Algorithmus hingegen teilt diese Schwäche nicht, denn wie CYRANGE stellt auch er eine DVM auf. In den meisten Fällen enthielten die von CYRANGE bestimmten Bereiche weniger Lücken und sie deckten einen signifikant höheren Anteil der Sequenz ab als die von FindCore oder PSVS ausgegebenen. Folglich waren auch die RMSD-Werte der CYRANGE-Domänen oftmals im Vergleich leicht erhöht. Allerdings gab es ebenfalls Fälle, in denen die von CYRANGE identifizierten Bereiche sowohl mehr Reste enthielten als auch mit einem niedrigeren RMSD-Wert überlagert werden konnten als die Domänen der anderen beiden Programme. Im Durchschnitt erzielten die von CYRANGE ausgegebenen Bereiche eine Sequenzabdeckung von 85 % und einen Rückgrat-RMSD-Wert von 0.77 Å, verglichen mit 67 % Abdeckung und 1.72 Å RMSD bei PSVS und 58 % Abdeckung und 0.73 Å RMSD bei FindCore. Dies zeigt, dass CYRANGE besser als die beiden anderen Programme dazu in der Lage ist, eine günstige Balance zwischen Sequenzabdeckung und niedrigem Rückgrat-RMSD zu erzielen.

CYRANGE wurde offiziell durch den NMR-Validierungsausschuss der Protein Data Bank zum Programm der Wahl zur Bestimmung der RMSD-stabilen Domänen eines Strukturbündels erklärt (?). Des Weiteren wurde das Programm verwendet, um im Zuge der Analyse der CASD-NMR-Beiträge von 2013 die geordneten Bereiche der Strukturen zu ermitteln (??), und es wurde in eine Version des Strukturvalidierungsprogramms MolProbity integriert (MolProbity-HTC; ?). Insgesamt unterstreicht nicht nur die hohe Qualität der CYRANGE-Ergebnisse, sondern auch die Tatsache, dass das Programm sowohl aktiv genutzt als auch offiziell empfohlen wird, dass CYRANGE den Zweck, für den es erdacht wurde, mehr als erfüllt.

Das zweite in dieser Dissertation beschriebene Hauptprojekt befasst sich mit der Validierung von Proteinstrukturen, die mithilfe von NMR gelöst wurden. Diese Aufgabe ist komplizierter als die Validierung röntgenkristallographisch bestimmter Strukturen, denn die Beugungsdaten der letztgenannten Methode sind in hohem Maße redundant. Folglich können sie in zwei Datensätze aufgeteilt werden, von denen der eine zur Strukturbestimmung und der andere zur Kreuzvalidierung des Ergebnisses verwendet werden kann, ohne dass die Qualität der erhaltenen Struktur darunter leiden würde. NOESY-Daten hingegen, die für gewöhnlich den Großteil der Informationen bei einer Strukturbestimmung mithilfe von NMR beisteuern, sind nicht in demselben Maße redundant.

Dieses zentrale Problem der Validierung von NMR-Strukturen hatte die Entwicklung einiger wissensbasierter Programme zur Folge, deren Ziel es ist, die Qualität eines Strukturmodells mittels eines Vergleichs bestimmter struktureller Merkmale mit Werten, die durch eine Analyse großer Mengen bekannter Proteinstrukturen gewonnen wurden, abzuschätzen. Der Nachteil solcher Methoden ist ihr Unvermögen, korrekt ausgebildete aber ungewöhnliche Eigenschaften der Struktur zu erkennen, da diese sich aller Wahrscheinlichkeit nach nicht in den Daten wiederfinden, die die Entscheidungsgrundlage dieser Programme bilden. Ferner berücksichtigt diese Art von Software die zur Strukturbestimmung verwendeten experimentellen Daten nicht, folglich kann keine Aussage darüber getroffen werden, wie gut das Strukturmodell die Daten wiedergibt. Aus diesem Grund sind wissensbasierte Validierungsprogramme nicht notwendigerweise befähigt, ein Maß für die Genauigkeit der Struktur zu bestimmen, doch gerade diese Information wäre ausgesprochen wertvoll für all diejenigen, die das Strukturmodell in der eigenen Forschung einsetzen möchten. Validierungsprogramme, die die experimentellen Daten berücksichtigen, sind jedoch selten.

Nutzerinnen und Nutzer würden von einer Software profitieren, die ein Strukturmodell auf Basis der ihm zugrunde liegenden experimentellen Daten validiert und die so wenig individuelle Anpassung und Nutzerinteraktion wie möglich benötigt. Genau dafür ist CYVAL, die von uns entwickelte Validierungsmethode, ausgelegt.

CYVAL ist in die Strukturrechnungssoftware CYANA (??) integriert und es beurteilt die Genauigkeit der Struktur anhand deren Übereinstimmung mit den originalen NOESY-Spektren. Zu diesem Zweck berechnet es ein Validierungsmaß (ζ) aus gewichteten Distanzabwei-chungs-Strafbeiträgen. Um die Deutung von ζ zu vereinfachen, wurde diese Größe derartig gestaltet, dass ihr Wert stets in den Bereich zwischen 0.0 und 1.0 fällt, wobei niedrigere Werte eine höhere Übereinstimmung zwischen Struktur und Daten anzeigen.

Die CYVAL-Methode benötigt ein prozessiertes NOESY-Spektrum, eine Datei des Strukturmodells und eine Liste zugeordneter chemischer Verschiebungen. Die letzteren beiden Dateien werden zur Vorhersage von NOESY-Signalen verwendet, und im Spektrum werden mithilfe der in CYANA integrierten Peak Picking-Prozedur Peaks gepickt. Die beiden Peaklisten, erwartet und experimentell, werden durch eine Komponente des Peakmatch-Algorithmus (**?**) verglichen, woraus drei getrennte Peakklassen hervorgehen: gefundene Peaks, welche sowohl auf Basis der Struktur erwartet als auch im Spektrum identifiziert wurden; fehlende Peaks, die vorhergesagt wurden, aber nicht gepickt werden konnten; und unerwartete Peaks, die im Spektrum vorkommen, deren Vorhandensein jedoch nicht durch die Struktur erklärt werden kann.

Zu jedem Peak gehört eine Distanz zwischen zwei Protonen, die im Falle der erwarteten Peaks aus der Struktur stammt, und die bei unerwarteten Peaks aus den experimentellen Signalintensitäten kalibriert wird. Aufgrund des d^{-6}-Zusammenhangs zwischen NOESY-Signalintensität und Interprotonendistanz führt CYVAL die Überprüfung der Übereinstimmung zwischen Struktur und Daten auf der Ebene der Atomabstände durch, denn vergleichsweise große Fehler bei Peakintensitäten ergeben lediglich relativ kleine Distanzabweichungen ($|\Delta d|$). Folglich sinkt das Risiko der Überbewertung von Fehlern, wenn Abstände verglichen werden. Bei Peaks der *gefundenen* Klasse ist $|\Delta d|$ die Differenz zwischen der experimentell bestimmten und der in der Struktur vorkommenden Distanz; zu Peaks der anderen zwei Kategorien gehörige Abstandswerte werden mit d^{max} verglichen, der oberen NOE-Distanzgrenze.

Um ζ auf das Intervall [0.0,1.0] zu beschränken und auch, damit sowohl auf sehr niedrigen als auch auf sehr hohen Distanzabweichungen kein allzu großes Gewicht liegt, wird eine sigmoidale Fehlerfunktion mit einem Maximalwert von 1.0 auf $|\Delta d|$ angewandt, bevor es in ζ einfließt. Des Weiteren sind nicht alle Peaks in gleichem Maße relevant für die Validierung der Struktur, daher wird jedem Peak ein aus mehreren Komponenten berechneter Gewichtungsfaktor zugeteilt. Kurzreichweitige Peaks beispielsweise, die bei jeder von einem Protein gegebener Sequenz eingenommenen Konformation zu erwarten sind, steuern nur wenig wertvolle Informationen bei. Hingegen sind langreichweitige Peaks entscheidend für die Bestimmung der globalen Faltung eines Proteins und enthalten somit auch wichtige Informationen bezüglich der Übereinstimmung von Struktur und Spektrum. Auf der anderen Seite besteht bei schwachen Peaks, die lange Abstandswerte ergeben, eine höhere Wahrscheinlichkeit, dass es sich bei ihnen de facto um Rauschen handelt. Da dieses ζ nicht übermäßig beeinflussen sollte, müssen die Peakgewichte eine Balance zwischen den unterschiedlichen Faktoren aufrecht erhalten.

Um ein Validierungsmaß pro Peak zu bekommen, kann ζ_{n} berechnet werden, das durch die Normierung von ζ mit der Anzahl der verwendeten Peaks ermittelt wird.

Fünf Proteine, für die eine Referenzstruktur, eine Liste zugeordneter chemischer Verschiebungen sowie prozessierte NOESY-Spektren vorlagen, wurden zum Test des Validierungserfolgs von ζ und ζ_{n} eingesetzt. Als Maß für die Bestimmung der Leistung von CYVAL wurde der Rückgrat-

RMSD zur Referenzstruktur ($\mathrm{RMSD}_{\mathrm{ref}}$), berechnet auf den von CYRANGE ermittelten geordneten Bereichen der Referenz, angewandt. Für jedes Testprotein wurden Strukturen unterschiedlicher Genauigkeit aus schrittweise gekürzten Peaklisten erzeugt. Diese inkorrekten Strukturen wurden CYVAL zusammen mit den unveränderten Listen zugeordneter Resonanzen und dem jeweiligen NOESY-Spektrum als Input gegeben. Die sich daraus ergebenden Werte von ζ oder ζ_{n} wurden zum Zweck der visuellen Überprüfung gegen $\mathrm{RMSD}_{\mathrm{ref}}$ aufgetragen, ferner wurde der Korrelationskoeffizient zwischen den beiden Größen berechnet. Eine lineare Abhängigkeit von ζ oder ζ_{n} und $\mathrm{RMSD}_{\mathrm{ref}}$ ist nicht zu erwarten, daher bilden Korrelationskoeffizienten allein keine ausreichende Grundlage für die Bewertung der Validierungsleistung von CYVAL. Da außerdem einige wenige Ausreißer ein von ihnen abgesehen aussichtsreiches Ergebnis verdecken können, können Korrelationskoeffizienten sogar irreführend sein. Nichtsdestotrotz ist mit ihrer Hilfe eine wertvolle erste Abschätzung der Leistung von CYVAL möglich, und sie können dazu verwendet werden, einen quantitativen Vergleich zwischen zwei Validierungsversuchen zu ziehen, sofern die Unterschiede ausreichend stark ausgeprägt sind. Ferner wurden verschiedene Fehlerfunktionen, Methoden zur Berechnung von Gewichtungsfaktoren, und ihre Kombinationen ausgewertet, ebenso wie unterschiedliche Vorgehensweisen zur Auswahl der Peaks, welche zur Berechnung von ζ herangezogen werden.

Die Testergebnisse waren vielversprechend. Zwar war im Falle eines Proteins kein klarer Zusammenhang zwischen ζ und $\mathrm{RMSD}_{\mathrm{ref}}$ erkennbar, bei zwei weiteren Proteinen hingegen konnte CYVAL deutlich zwischen Strukturen hoher (unter 2.0 Å $\mathrm{RMSD}_{\mathrm{ref}}$), mittlerer (zwischen 2.0 und 4.0 Å) und niedriger (über 4.0 Å) Qualität unterscheiden. Bei den restlichen zwei Proteinen zeigte sich ebenfalls die erwartete Zunahme von ζ bei Verminderung der Strukturgenauigkeit, die Unterschiede zwischen den drei Klassen waren jedoch teilweise verschwommen. Alles in allem lag die Validierungsleistung von ζ_{n} etwas über derjenigen von ζ, aber beide Validierungsgrößen waren neben dem $\mathrm{RMSD}_{\mathrm{ref}}$-Wert auch vom untersuchten Protein und dem in der Validierung eingesetzten Spektrentyp abhängig.

Die Tests der alternativen Fehlerfunktionen und Gewichtungsmodi brachten keine Methode hervor, die allen anderen Ansätzen klar überlegen war. Die Auswahl der Peaks für die Berechnung von ζ hingegen hatte einen deutlichen Einfluss auf den Erfolg der Validierungsversuche. Die Einbeziehung von Peaks, die einzeln wenig Information zum Validierungsvorgang beisteuern, war nichtsdestotrotz wichtig für die Validierungsleistung von CYVAL.

Wir haben drei wissensbasierte Strukturvalidierungsprogramme auf unsere Proteintestsets angewandt: ProSa2003 (?), Verify3D (???) und MolProbity (??). Obwohl diese Programme keinen Zugriff auf die experimentellen Daten der Proteine hatten, gab es Fälle, in denen es auch ihnen gelang, die Klassen struktureller Genauigkeit zu unterscheiden. Die Validierungsleistung von CYVAL war derjenigen der anderen Programme meist ebenbürtig oder sogar überlegen.

Es müssen noch finale Hürden überwunden werden, damit die CYVAL-Methode auf alle möglichen Fälle in gleichem Maße anwendbar ist. Insgesamt hat CYVAL allerdings klar seine Fähigkeit zur Unterscheidung zwischen Strukturen unterschiedlicher Genauigkeitsklassen demonstriert. Auch ist seine Validierungsleistung mindestens so hoch wie diejenige einiger bekannter Validierungsprogramme. Zusammengefasst zeigen die Testergebnisse, dass CYVAL das Potential hat, ein wertvolles Werkzeug zur Validierung mittels NMR bestimmter Proteinstrukturen zu werden.

Zusätzlich zur Analyse von Proteinstrukturen habe ich mich auch mit ihrer Bestimmung mithilfe von Flüssig-NMR-Daten befasst, indem ich drei Strukturrechnungen durchgeführt habe. Die hierbei verwendeten Proteine waren der Wildtyp der Trp-Trp (WW) Bindungsmodul-Domäne der Peptidyl-Prolyl-cis/trans-Isomerase Pin1 sowie die Ser16Glu-Mutante (WW$^{\mathrm{S16E}}$) dieser Domäne (**?**), und die Ser45Ala-Mutante von TycC3_PCP, der dritten Peptidyl-Carrierdomäne der Tyrocidin A-Synthethase-Untereinheit C aus *B. brevis* (**?**). Pin1 wird in verschiedenen menschlichen Krebsarten überexprimiert und hohe Konzentrationen der an Ser16 phosphorylierten Form dieses Proteins wurden in Hirngewebeproben von Alzheimerpatienten nachgewiesen. TycC3_PCP, dessen Apo-Form durch TycC3_PCP(S45A) nachgeahmt wird, ist ein Vertreter der großen Gruppe der Carrierproteine. Die Aufklärung der Strukturen der Mitglieder dieser Proteinklasse, sowohl allein als auch im Komplex mit ihren Reaktionspartnern, kann das Verständnis der Biosynthesewege, in denen sie eingesetzt werden, erweitern, und darüber hinaus zur Identifizierung neuer Angriffspunkte für Medikamente beitragen.

In allen drei Fällen ergab die Strukturrechnung ein wohldefiniertes Ensemble, dessen Rückgrat-Torsionswinkel alle in erlaubten Bereichen des Ramachandranplots angesiedelt waren. Außerdem wurden hohe Anteile sowohl zugeordneter NOESY-Kreuzpeaks als auch langreichweitiger Distanzschranken erzielt. Insbesondere letztere sind für die Bestimmung der Tertiärstruktur relevant. Weder kam es zu Distanzschrankenverletzungen über 0.2 Å noch zu Verletzungen der Torsionswinkelschranken über 5°. Die berechneten wissensbasierten Validierungsgrößen zeigen, dass alle drei Strukturmodelle geometrische Kriterien sehr gut erfüllen. Die Ergebnisse der Anwendung von CYVAL legen nahe, dass die Strukturen auch mit den experimentellen Daten gut übereinstimmen.

WW$^{\mathrm{S16E}}$ und die Wildtyp-WW-Domäne von Pin1 weisen eine sehr ähnliche Faltung des Rückgrats auf. Die Ladungsverteilung an der Oberfläche der Bindetasche hingegen wurde durch die Einführung der negativ geladenen Glu-Seitenkette signifikant verändert, und die Größe dieses Rests führte darüber hinaus zu einer sterischen Hinderung für Protein-Protein-Wechselwirkungen. Beide Beobachtungen erklären, weshalb die regulatorische Phosphorylierung von Pin1 am in der WW-Domäne gelegenen Ser16 die Wechselwirkung des Proteins mit potentiellen Bindungspartnern verhindert. Die NMR-Struktur von TycC3_PCP(S45A) in Lösung lässt den Rückschluss zu, dass apo-TycC3_PCP, anders als bisher angenommen, dieselbe Konformation einnimmt wie die Holo-Form. Eine von **?** gelöste Kristallstruktur hat gezeigt, dass dies ebenfalls die Konformation ist, die TycC3_PCP(S45A) im Komplex mit der Phosphopantetheintransferase (PPT) Sfp einnimmt. Die Ähnlichkeit dieser Struktur mit dem Komplex aus humanem Acylcarrierprotein und PPT lässt Ähnlichkeiten bei der Erkennung und posttranslationalen Modifikation von Carrierproteinen in verschiedenen Organismen vermuten. Dies wiederum könnte negative Auswirkungen auf die Verwendbarkeit bakterieller PPTs als Zielproteine für die Entwicklung neuartiger Antibiotika haben.

List of abbreviations

1D	one-dimensional
3D	three-dimensional
ζ	the validation metric output by CYVAL
ζ_{n}	the per peak-validation metric
ADR	Ambiguous Distance Restraint
CASD-NMR	Critical Assessment of automated Structure Determination of proteins from NMR data
CING	Common Interface for NMR structure Generation
CPU	Central Processing Unit
CYANA	Combined assignment and dYnamics Algorithm for NMR Applications
DVM	Distance Variance Matrix
FID	Free Induction Decay
GDT	Global Distance Test
GDT_TS	GDT Total Score
HSQC	Heteronuclear Single-Quantum Coherence
LCS	Longest Continuous Segments
MD	Molecular Dynamics
NESG	NorthEast Structural Genomics
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Effect SpectroscopY
NRPS	NonRibosomal Peptide Synthetase
PCP	Peptidyl Carrier Protein

PDB	Protein Data Bank
PPI	Peptidyl-Prolyl cis/trans Isomerase
ppm	parts per million
PRE	Paramagnetic Relaxation Enhancement
PSVS	Protein Structure Validation Suite
QUEEN	QUantitative Evaluation of Experimental NMR restraints
RDC	Residual Dipolar Coupling
RMSD	Root Mean Square Deviation
$RMSD_{ref}$	backbone RMSD to the reference structure
SNR	Signal to Noise-Ratio
Vivaldi	VIsualization and VALidation DIsplay
VTF	Validation Task Force
WW	Trp-Trp binding module

Contents

Part I

Introduction

Chapter I.1

Outline

The main aim of this thesis has been to provide better software tools for structure analysis and validation to scientists using nuclear magnetic resonance (NMR) spectroscopy for the determination of protein structures. Additionally, NMR data were directly used in two structure calculation projects.

The determination of the structures of biomolecules has become a field of intense research. In February 2016, the Protein Data Bank (PDB) contained more than 116,000 entries, and over 11,000 of the deposited structures had been solved by NMR spectroscopy. This number is low compared to the nearly 104,000 structural models from X-ray crystallography, yet it illustrates that NMR, too, has become a popular method for the study of the structures of biologically relevant macromolecules. Proteins are by far the most commonly examined molecules, which is illustrated by almost 108,000 PDB entries containing structural data of proteins only.

Obviously, a significant research effort is put into the study of protein structures by NMR. Nevertheless, the field still has a number of issues that need addressing in order to simplify the scientists' work, and to make their findings even more accessible and valuable to the scientific community. One example is the reporting of the precision of structural models derived by NMR. As explained in chapter I.3, NMR structure calculations yield an ensemble of conformers instead of a single structure. Different parts of this ensemble, however, exhibit different degrees of local order, which complicates the determination of a measure of structural precision. We have developed CYRANGE (**?**), a software that addresses this problem by identifying ordered domains in structure bundles. Chapter I.4 provides a more thorough explanation of the problem and an introduction to various approaches to solving it. A detailed description of the CYRANGE method is given in chapter II.1. The results presented in chapter III.1 demonstrate that CYRANGE does not only succeed in carrying out the task for which it was designed but that it even outperforms two other well-known software tools that identify structural order.

When structural models of proteins are used as a basis of further research the models must be as correct as reasonably possible, otherwise erroneous conclusions may be drawn from them. In order to avoid such problems the quality of a structure can be assessed by various structure validation tools. Naturally, information about the degree of accuracy, i.e. the similarity of the structural model to the native structure, is particularly relevant. For the evaluation of structural accuracy experimental data need to be taken into account, but this is not trivial in the case of NMR-derived structures, for reasons outlined in chapter I.5. As a result, there are only few computer programs that perform data-based validation of protein structures calculated from NMR data, and all have their shortcomings. For this reason a new validation method, CYVAL, was developed as part of this thesis. It aims to provide the user with a validation metric that is as objective as possible, hence the software performs its assessment of structural quality on the basis of the NMR spectra that were used in structure determination. A detailed description of the CYVAL method is given in chapter II.2. The results reported in chapter III.2 demonstrate that the CYVAL-approach is a promising one, which performs as well or even better than three other structure validation tools.

The third major part of the work was the determination of three protein structures from NMR data. The structure of the Ser45Ala-mutant of TycC3_PCP, a peptidyl carrier domain of tyrocidine A synthetase, was calculated (**?**), as well as the structures of the wild-type WW domain of the peptidyl-prolyl cis/trans isomerase Pin1 and of this domain's Ser16Glu-mutant (**?**). One reason for TycC3_PCP being of scientific interest is its membership in the ubiquitous class of carrier proteins, which are essential agents in various biosynthetic pathways, but the structure determination results may even have implications on the choice of potential targets for the development of novel antibiotics. Pin1 plays a role in various human cancers and other diseases, therefore learning more about how this protein functions and how it is regulated may be of crucial medical importance. Chapter II.3 reports how the structure calculations were carried out, and chapter III.3 presents the solved structures and explains their scientific significance.

Chapter I.2

Nuclear magnetic resonance spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is a research technique by which the structure and dynamics of molecules, e.g. biomolecules such as proteins or nucleic acids, can be investigated. Molecular structures at atomic resolution may also be determined by X-ray crystallography. However, in contrast to NMR this method requires the production of crystals of the target molecule that provide diffraction data of a suitable quality, but obtaining such crystals is not always possible. Furthermore, the dynamics of molecules generally cannot be examined in a crystalline sample. On the other hand, proteins need to be isotope-enriched so that their structures can be determined by NMR, which is expensive, and typically proteins of more than 30–40 kDa do not yield spectra of a quality suitable for the determination of high-resolution structures (**?**). A size limitation does not exist in X-ray crystallography. Obviously, the two experimental techniques are complementary, and both have a right to exist within the scientific tool box.

NMR spectroscopy may be carried out on liquid or solid samples. A drawback of solid-state NMR spectra is their comparatively low resolution, but the method is useful all the same, especially in those cases in which the query molecule is insoluble or unstable in solution. For example, an atomic-resolution structure of amyloid β-peptide fibrils associated with Alzheimer's disease was determined using distance information from solid-state NMR measurements (**?**). The projects of this thesis, however, were mainly concerned with liquid-state NMR data, consequently it is this method the text focuses on.

The current chapter provides a short introduction to the theoretical background of NMR and NMR spectroscopy. It is largely based on the books by **?**, and **?**. The determination of protein structures using liquid-state NMR data will be discussed in chapter I.3.

I.2.1 Theoretical background

The application of a magnetic field affects certain nuclei because they possess spin angular momentum. The latter gives rise to a magnetic moment $\boldsymbol{\mu}$, which is a vectorial quantity whose z-component is expressed by eq. I.2.1.

$$\mu_z = \gamma m_I \hbar \tag{I.2.1}$$

Here, γ is the *gyromagnetic ratio* of the isotope in question. The value of the magnetic quantum number m of the nucleus depends on the nuclear spin quantum number I. The latter can take positive integer or half-integer values, and $m_I = I, I-1, \ldots, -I$. This means that for nuclei with $I = \frac{1}{2}$, such as ^{1}H, ^{13}C, or ^{15}N, there are two possible spin states as m_I can only be equal to $+\frac{1}{2}$ or $-\frac{1}{2}$. If I is zero, which is the case for ^{18}O, for example, the nucleus has zero magnetic moment and can thus not be observed in NMR experiments.

The energy E of a nuclear spin state inside an external magnetic field of strength B_0 depends on the magnetic moment as shown by eq. I.2.2.

$$E_{m_I} = -\mu_z B_0 = -\gamma \hbar B_0 m_I \tag{I.2.2}$$

Consequently, for nuclei with $I = \frac{1}{2}$ there are two energy states. In the α state, m_I is equal to $+\frac{1}{2}$, and to $-\frac{1}{2}$ in the β state. The energy difference $\Delta E_{\alpha \to \beta}$ between the two states is given by eq. I.2.3. As ΔE is proportional to the external field strength, and as larger differences in energy levels bring about an increase in resolution in spectroscopic methods, the application of stronger magnetic fields leads to a higher resolution in the resulting NMR spectra.

$$\Delta E_{\alpha \to \beta} = E_\beta - E_\alpha = \gamma \hbar B_0 = h\nu \tag{I.2.3}$$

ν is the *Larmor frequency* of the nucleus. It describes the frequency at which the transition between the two spin states can take place. ν is a characteristic value by which the nucleus can be identified as the frequency depends not only on the isotope but also on the local electronic environment of the nucleus. Depending on the molecular structure and conformation nuclei experience different degrees of shielding σ from the external magnetic field. The strength B_{loc} of the local field they experience is given by eq. I.2.4.

$$B_{\text{loc}} = (1-\sigma)\, B_0 = B_0 + \delta B \quad \text{with} \quad \delta B = -\sigma B_0 \tag{I.2.4}$$

δ is the characteristic *chemical shift* of the nucleus. The actual Larmor frequency ν for a nucleus is thus equal to $\frac{\gamma B_{\text{loc}}}{2\pi}$.

The chemical shift is commonly reported in units of ppm (parts per million). It is determined according to eq. I.2.5, where ν^0 is the Larmor frequency of the isotope in a reference compound.

$$\delta = \frac{\nu - \nu^0}{\nu^0} \times 10^6 \tag{I.2.5}$$

Magnetic nuclei themselves give rise to magnetic fields, hence spins can modify each other's resonance frequencies, in a process known as *coupling*. If it is mediated by the bonds connecting the nuclei it is referred to as *scalar coupling*, and the effect of scalar coupling on the combined energy of two coupled spin-$\frac{1}{2}$ nuclei A and X is influenced by the coupling constant J as shown by eq. I.2.6.

$$E = -h\nu_A m_A - h\nu_X m_X + hJm_A m_X \tag{I.2.6}$$

${}^N J$ is the term that denotes the scalar coupling constant of two nuclei separated by N bonds, and subscripts are used to state what nuclei are involved. A ${}^3J_{\mathrm{HH}}$ coupling constant often depends on the torsion angle θ according to the *Karplus equation* (eq. I.2.7; **?**).

$$J = A + B\cos\theta + C\cos 2\theta \tag{I.2.7}$$

A, B, and C are empirical constants.

If the coupling is not mediated by bonds but occurs through space it is referred to as *dipole-dipole coupling*. The z-component B_{nuc} of the magnetic field a spin induces in another location depends on the distance r between the two points, and on the angle θ of the connecting vector to the static external magnetic field. This relationship is expressed by eq. I.2.8.

$$B_{\mathrm{nuc}} = -\frac{\gamma\hbar\mu_0}{4\pi r^3}\left(1 - 3\cos^2\theta\right) m_I \tag{I.2.8}$$

Here, μ_0 is the permeability of the vacuum. In an isotropic liquid the molecule tumbles unrestrictedly, hence θ sweeps over all values and so the term $1 - 3\cos^2\theta$ averages to zero.

I.2.2 NMR spectroscopy

NMR spectroscopy basically detects the energy separation between nuclear spin states. A sample of the target molecule contains a population of spins, and thus potentially a bulk magnetization. Looking at spin-$\frac{1}{2}$-nuclei, in the complete absence of any external magnetic field the α and β spin states would be energetically equivalent and thus equally populated. Consequently, the net magnetization of the sample would be zero. Upon application of a magnetic field the α state becomes energetically favourable for nuclei with a positive value of γ, so the population of the α state increases slightly, which causes the net magnetization of the sample to become non-zero. The net magnetization is proportional in size to the population difference, and it can be represented by a vector pointing along the direction of the external magnetic field B_0, which will henceforth be referred to as the z-axis. Such a constant, strong magnetic field is continuously generated inside an NMR spectrometer.

In order to perform NMR measurements on a sample in an NMR spectrometer a circularly polarized radiofrequency field B_1 is applied in the xy-plane, i.e. in a direction perpendicular to the static field B_0. The magnetic component of B_1 interacts with the nuclear magnetic moments of the sample. If the oscillation frequency of B_1 is at or close to the Larmor frequencies of the spins the resonance condition is met and the alignment of the bulk magnetization with the z-axis is disrupted. The duration of the application of B_1 determines to what extent the net magnetization vector moves away from the z-direction: a so-called *90° pulse*, for example, whose duration is $\frac{\pi}{2\gamma B_1}$, will rotate the bulk magnetization into the xy-plane. There, the magnetization is detected as a time-domain signal of a frequency equal to the Larmor frequency of the spins.

The radiofrequency pulse excites a range of frequencies in the sample, and the detected signal indicates which Larmor frequencies are present in the query molecule, hence the signal is a combination of all individual frequency contributions. As B_1 is not static the net magnetization will once more experience only B_0 once the pulse has ceased, therefore the net magnetization slowly returns to the z-direction in a process known as *relaxation*, and the signal that is recorded in the xy-plane decays. For this reason the signal is referred to by the term *free-induction decay* (FID), a simple example of which is shown in fig. I.2.1. The y-component M_y of the signal is given by eq. I.2.9.

$$M_y\left(t\right) = M_0 \cos\left(2\pi\nu t\right) \mathrm{e}^{-\frac{t}{T_2}}, \tag{I.2.9}$$

Here, t denotes the time, M_0 the original net magnetization before the application of B_1, ν the Larmor frequency, and T_2 the time constant of signal decay, generally referred to as *transverse relaxation time*. A slow decay of the FID, i.e. a large value of T_2, leads to narrow lines with large amplitudes in the final spectrum

In order to extract the individual frequencies from the FID, which can then be combined to produce the frequency-domain spectrum, Fourier transformation is employed. Basically, the FID is multiplied by a number of trial cosine waves, and each of the resulting curves is integrated to yield the spectral intensity at the frequency at which this particular cosine wave oscillated. The entire spectrum is the sum of all such integrals. Eq. I.2.10 shows the mathematical notation of the Fourier transformation.

$$I_{\text{spectrum}}(f) = \int_0^{+\infty} I_{\text{FID}}(t)\cos(2\pi f t)\mathrm{d}t \tag{I.2.10}$$

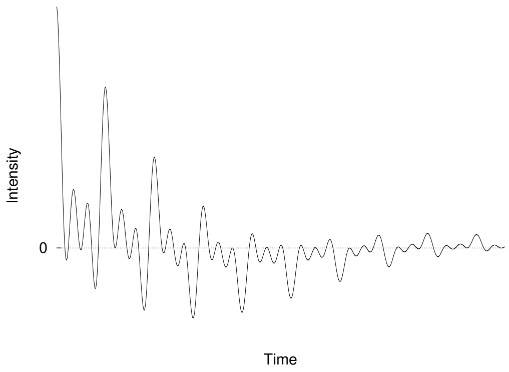


Figure I.2.1: The figure shows a schematic FID, which contains three different frequencies and an exponential term that induces the decay of the signal intensity over time.

I.2.2.1 Multi-dimensional NMR spectroscopy

There are various different NMR experiments, but all of them employ pulse sequences that affect the sample magnetization and allow the spins and their interactions to be detected. In the simplest case, one 90° pulse is applied, and after signal detection and Fourier transformation the signal intensity is plotted against the frequency in ppm, yielding a one-dimensional spectrum. However, as spectral complexity increases with molecular size, multi-dimensional spectra need to be recorded in order to determine the structures of macromolecules.

The general design of a two-dimensional NMR experiment, for example, consists of a series of measurements that all share the same structure. Each measurement begins with a *preparation period* in which the spins return to thermal equilibrium. This is followed by a variable number of radiofrequency pulses, and an *evolution period* of duration t_1. During the subsequent *mixing period* pulses may be used to induce magnetization transfer between spins. After this, the FID is recorded during the *detection period* of duration t_2, and a one-dimensional spectrum of the *direct dimension* with the frequency domain ω_2 is generated via Fourier transformation. t_1 is incremented on each new iteration, which causes the signal intensities in the recorded one-dimensional spectra to oscillate. Therefore, the combination of the final set of one-dimensional spectra yields a time-domain signal in the *indirect dimension.* The Fourier transformation of this time-domain signal produces the final two-dimensional spectrum with frequency domains ω_2 and ω_1.

Nuclear Overhauser effect spectroscopy

The nuclear Overhauser effect (NOE) provides information about inter-nuclear distances. It occurs via dipolar relaxation, which means that the effect is based on the magnetization transfer through space between interacting spins. The transfer rate R between two spin-$\frac{1}{2}$-nuclei A and X in the case of fast molecular rotation is given by eq. I.2.11.

$$R = \left(\frac{\mu_0}{4\pi}\right)^2 \frac{\gamma_A^2 \gamma_X^2 \hbar^2}{r^6} \tau_c \tag{I.2.11}$$

τ_c is the *rotational correlation time*, i.e. the average time the molecule requires to tumble by an angle of 1 rad.

NOE spectroscopy (NOESY) exploits the NOE to obtain distance information that can subsequently be used in the calculation of protein structures, for example. The so-called *cross peaks* in a two-dimensional NOESY spectrum lie at (δ_A, δ_X) and (δ_X, δ_A), where δ_A and δ_X are the chemical shift values of the interacting spins A and X, respectively. The intensities of these peaks depend on the efficiency of the magnetization transfer between the two spins. As the transfer rate R, in turn, depends on the inter-nuclear distance via an r^{-6}-relationship, the NOE is generally only observable for nuclei that are not separated further than around 5.0 Å.

I.2.3 Spectrum file formats

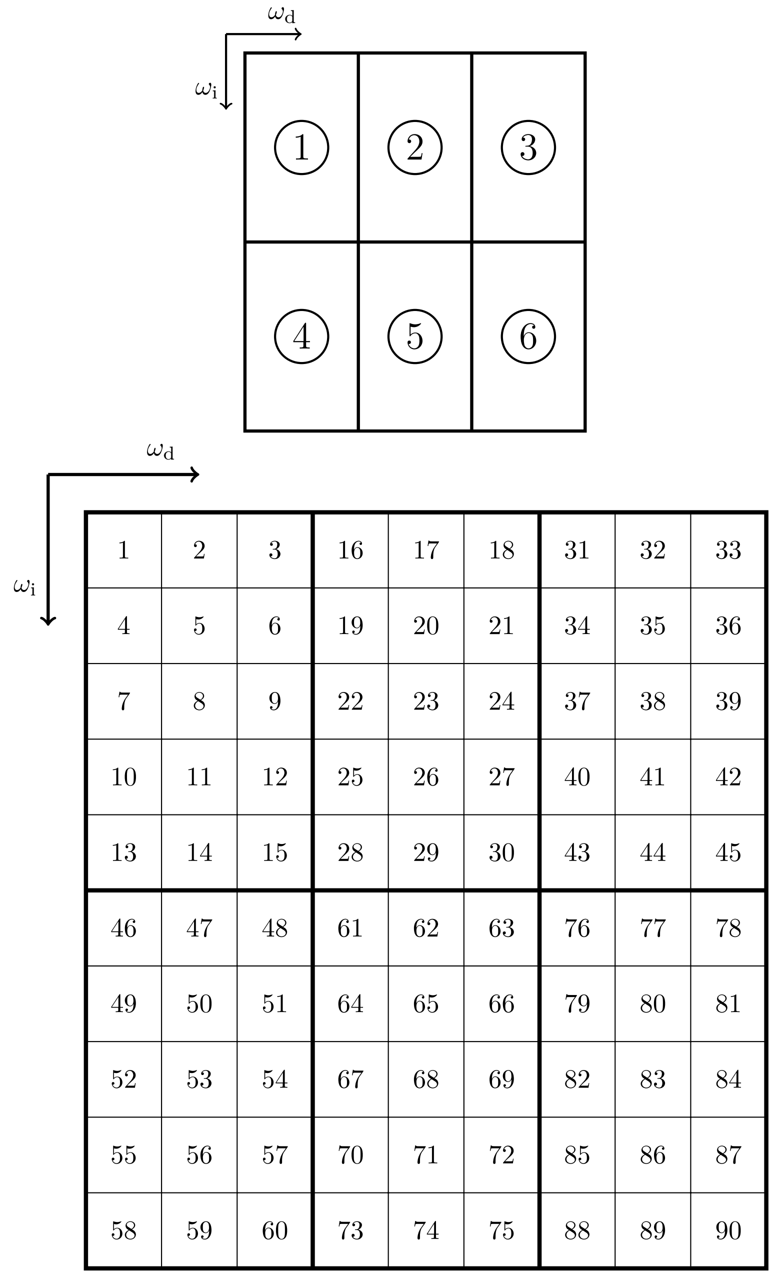
This section contains an overview of the Bruker and UCSF file formats for Fourier transformed NMR spectra. The routines to read such binary files were incorporated into CYANA as part of this thesis. The first sub-section describes the general format that these spectra files use to store intensity data, and the remaining two sub-sections provide specific details regarding the two different file formats.

I.2.3.1 The submatrix format

The *submatrix format* (or *subcube format* for more than two dimensions) is not a file format in itself, it is a way to organize data within a file. It dates back to a time when computer memory was a bottleneck for working with NMR spectra, and so a method of data storage was created that allowed to easily read only a certain section of the spectrum.

As illustrated by fig. I.2.2, in the submatrix format the index changes fastest within the highest dimension (acquisition dimension), and slowest within the lowest dimension. This applies both to the indices of the individual data points and to the indices of the submatrices themselves.

Files in the submatrix format can only be meaningfully read if the dimensions of the submatrices as well as those of the entire spectrum are known. This information is generally made available either via additional parameter files (Bruker) or as part of the data file itself (UCSF format).



1	2	3	16	17	18	31	32	33
4	5	6	19	20	21	34	35	36
7	8	9	22	23	24	37	38	39
10	11	12	25	26	27	40	41	42
13	14	15	28	29	30	43	44	45
46	47	48	61	62	63	76	77	78
49	50	51	64	65	66	79	80	81
52	53	54	67	68	69	82	83	84
55	56	57	70	71	72	85	86	87
58	59	60	73	74	75	88	89	90

Figure I.2.2: The submatrix format in case of a 2D spectrum. ω_d is the direct dimension (acquisition dimension), ω_i is the indirect dimension. The upper graphics illustrates the order of the individual submatrices projected onto the dimensions of the spectrum. The indices change faster along the direct dimension, thus the tiles are first filled along ω_d. The lower graphics shows the indices of the data points within the submatrices. The outer rectangle again symbolizes the entire 2D spectrum. Each square represents one data point; the lower the index the further towards the beginning of the file this point is stored.

I.2.3.2 UCSF format

The byte order of a UCSF file is big-endian. Each file starts with a header section that provides information relevant for reading the intensity data that are stored within the same file. The first header is 180 bytes in length and contains several values, but only the number of spectral dimensions, stored at byte 11, is required by CYANA. This general header is followed by a separate 128 bytes-header for each of the spectral axes. The contents of these headers are listed in tab. I.2.1.

The intensity information is provided directly after the header section. UCSF files use the submatrix format for storage of intensity information, and each individual intensity value is stored as a 32 bit floating-point number.

Table I.2.1: The information contained in each 128 byte-axis header of a UCSF file.

Position[a]	Length [bytes]	Contents	Value type
0	6	nucleus (1H, 13C, ...)	string
8	4	number of data points along this axis	integer
16	4	tile size along this axis	integer
20	4	spectrometer frequency for this nucleus [MHz]	float
24	4	spectral width of this axis [Hz]	float
28	4	centre of data on this axis [ppm]	float

[a] The position is given as the index of the Fortran record at which this data item starts. A Fortran record is 4 bytes in length.

In case the number of intensity data points of an axis is not a multiple of the tile size along this axis, tiles at the far end of the spectrum will be only partially filled. In such a case records within the file need to be skipped so as to avoid reading the filler values that are stored in the non-data slots.

I.2.3.3 Bruker format

To be able to read Bruker intensity files, parameter files are required. These are stored in plain text format. There is one parameter file for each of the spectral axes, named after the number of this dimension from dimension 2 onwards (`procs`, `proc2s`, ...). Each parameter file contains the values listed in tab. I.2.2, amongst other information.

Table I.2.2: Relevant content of a Bruker parameter file for processed spectral data.

Keyword	Contents	Value type
`AXNUC`	nucleus of the axis	string
`SW_p`	spectral width of this axis	float
`OFFSET`	ppm value of the first data point of this axis	float
`SF`	reference frequency of this axis	float
`SI`	number of data points along this axis	integer
`XDIM`	tile width along this axis	integer

Bruker intensity data files, too, are stored in the submatrix format. Each intensity value is stored as a 32 bit integer. Furthermore, each tile size is a multiple of 256 data points (1024 bytes). If the data do not exactly fit such a tile, it is filled as far as possible using intensity information. The remainder of the tile is made up of empty slots, and spectrum data can again be found at the beginning of the following tile.

Chapter I.3

Protein structure calculation from liquid-state NMR data

Data from NMR measurements may be used in the determination of three-dimensional structures of proteins and other biomolecules. Once the data have been appropriately recorded and prepared a structure can be calculated, i.e. one determines the atomic positions that best fulfil the external restraints given the covalent geometry imposed by the amino acid sequence (**?**). The current chapter introduces methods for the determination of protein structures from liquid-state NMR data, with a focus on the use of NOESY data since these are particularly relevant in the context of this thesis.

I.3.1 The classical approach to structure calculation

Typically, the protein sequence, assigned chemical shifts, and NOESY cross-peak positions and intensities constitute the input data for a conventional structure determination from liquid-state NMR data. Additional restraints may be included, e.g. dihedral angle predictions, which are commonly obtained from chemical shift information using computer programs such as TALOS-N (**?**); residual dipolar couplings (RDCs), which provide information regarding the relative orientation of individual groups within the molecule; or distance restraints derived from paramagnetic relaxation enhancement (PRE).

Nowadays, NOESY cross-peaks are generally automatically assigned to proton pairs, and a number of algorithms have been developed for this purpose, e.g. NOAH (**??**), CANDID (**?**), KNOWNOE (**?**), PASD (**?**), AutoStructure (**?**), ARIA (**????**), and ASDP (**?**). Automated chemical shift assignment, too, has become possible (**???**), but it is still significantly less commonly carried out than the automated assignment of NOESY signals (**?**).

The computer programs CYANA (??) and ARIA (????) perform both automated NOESY cross-peak assignment and structure calculation. As CYANA is the most widely used NMR structure calculation software (?) and because this thesis includes the determination of three protein structures from NMR data using CYANA, the structure calculation and NOE assignment approach of this program will be outlined in the following section, which is primarily based on two publications, **?** and **?**.

I.3.1.1 Structure determination with CYANA

The structure calculation routine starts from a set of random structures and it requires experimental restraints, in particular distance restraints, for the determination of the atomic positions within the protein molecule. These distance restraints are obtained from NOESY cross-peak data, therefore the peaks have to be assigned to proton pairs before a structure calculation can be performed. Consequently, the automated NOE assignment procedure must precede the actual structure calculation step.

Combined NOE assignment and structure calculation

CYANA uses seven cycles of combined NOE assignment and structure calculation during which both the peak assignments and the structural model are gradually refined. Afterwards, one final round of structure calculation only is performed.

At the beginning of each cycle the NOESY cross-peaks are assigned to proton pairs based on the assigned chemical shifts provided by the user and, in later cycles, also based on the intermediary three-dimensional structure. Due to chemical shift degeneracy and at times inaccurate peak positions, assigning NOESY peaks is not straightforward, thus the iterative nature of the process becomes mandatory. Between the different cycles information is transferred solely via the original input data and the intermediary three-dimensional structural models.

Chemical shift degeneracy often renders it impossible to unambiguously assign a peak using resonance information alone, hence in most cases each peak will initially have several assignment possibilities. In order to determine which of these are likely to be correct the individual possibilities are weighted by probability factors. The first of these factors, P_{shifts}, evaluates the agreement between the peak position and the resonances of the atoms. Another factor ($P_{\text{structure}}$), which expresses the agreement of the assignment possibility with the structural model, is calculated from the second cycle onwards, i.e. once an intermediary structure exists. Structural information needs to be taken into account since the assignment of a peak to a pair of protons is likely to be erroneous if the preliminary structure suggests that those two atoms are separated further than what is allowed by the upper distance bound of the peak plus an acceptable distance limit violation.

The third probability factor for assignment evaluation, P_{network}, is based on the *network anchoring* of the assignment. It quantifies how well the assignment possibility in question is supported by the entire assignment network of the protein and the constraints imposed by the covalent geometry of the molecule. P_{network} increases if other peaks share this assignment or if the two atoms to which the peak is assigned are indirectly connected via assignments including a third atom. Additionally, it is favourable if atoms located close to one atom of the possible assignment are connected via peaks to atoms in the vicinity of the second atom of the assignment possibility. All these factors raise the likelihood of the query assignment being correct since they suggest that the two protons are indeed close enough in space to be able to satisfy the upper distance bound.

The abovementioned probability factors are combined to give a total assignment probability, P_{tot}, according to eq. I.3.1.

$$P_{\text{tot}} = P_{\text{shifts}} \times P_{\text{structure}} \times P_{\text{network}} \tag{I.3.1}$$

If P_{tot} does not reach a threshold value P_{min} this particular assignment possibility is discarded since its probability of being correct is deemed too low. Nevertheless, especially during the early stages of NOE assignment and structure calculation, many peaks will retain several assignment possibilities. Since the global fold can rarely be determined using solely the initially low number of unambiguously assigned signals (**?**) the structure determination method has to make additional use of the ambiguously assigned peaks. This is done via employing *ambiguous distance restraints* (ADRs) (**?**).

An ADR comprises all of a peak's n probable assignment possibilities i, and the upper distance limit b created for any peak is based on the experimental signal intensity (see below). Whether b is fulfilled by the structure depends on d_{eff}, which is calculated from the structure-based distance values d_i according to eq. I.3.2.

$$d_{\text{eff}} = \left(\sum_{i=1}^{n} d_i^{-6} \right)^{-\frac{1}{6}} \tag{I.3.2}$$

d_{eff} will always be shorter than the shortest contributing distance d_i, therefore d_{eff} will not violate b, provided that the correct assignment has been included. Consequently, an ADR containing the correct assignment of the corresponding peak will not distort the structure.

The presence of erroneous distance restraints stemming from the misinterpretation of spectral artefacts, for example, creates a risk of distortion of the structural model. In order to counteract this potential problem CYANA performs *constraint combination* (**?**) during the first two cycles of NOE assignment and structure calculation. Constraint combination creates ADRs from generally unrelated medium-range and long-range cross-peaks with the aim of blending potentially incorrect restraints with correct ones, since ADRs containing at least one correct restraint will not contort the structure. Constraint combination is restricted to the first two cycles as it entails a loss of information, but its advantages are particularly valuable when no structural information is yet available for the evaluation of NOESY cross-peak assignments.

Once distance restraints have been generated from the NOESY cross-peaks a structure calculation is performed as outlined below. The resulting structure is then used to refine the NOE assignments in the following cycle, but each new structure calculation once more starts from a set of random structures. It is the gradual refinement of the experimental restraints that leads to an increase in structural precision over the course of the whole procedure. After the final cycle one last structure calculation is performed, in which only unambiguously assigned peaks are accepted for restraint generation.

Structure calculation

CYANA performs the actual structure calculation step using a combination of molecular dynamics (MD) simulation and simulated annealing (**?**) in torsion angle space (**?**). Simulated annealing is a probabilistic method for approximating the global minimum of a target function, which is a simplified potential energy (P) function in the case of structure calculation. Eq. I.3.3 defines the CYANA target function (**??**). It is zero only if all restraints are fulfilled and steric overlap does not occur.

$$P = \sum_{c=u,l,v} w_c \sum_{(\alpha,\beta)\in I_c} w^c_{\alpha\beta}(d_{\alpha\beta} - b_{\alpha\beta})^2 + w_a \sum_{i\in I_a} w_i \left[1 - \frac{1}{2}\left(\frac{\Delta_i}{\Gamma_i}\right)^2\right] \Delta_i^2 \tag{I.3.3}$$

$b_{\alpha\beta}$ are upper and lower bounds on distances $d_{\alpha\beta}$ between atoms α and β in the structural model; the upper bounds are determined from the experimental NOESY cross-peak intensities as stated below. I_u, I_l, and I_v are the sets of violated upper bound, lower bound or van der Waals distance restraints, respectively, and I_a is the set of restrained torsion angles. $w_{u/l/v/a}$ are the global weighting factors for the various types of restraints, and both w_i and $w^c_{\alpha\beta}$ are weighting factors for individual restraints. Δ_i denotes the magnitude of the torsion angle restraint violation, and $\Gamma_i = \pi - \left(\theta_i^{\max} - \theta_i^{\min}\right)/2$ is the half-width of the forbidden torsion angle range, with $\theta_i^{\min}$ and $\theta_i^{\max}$ being the lowest and highest allowed torsion angle value for torsion angle i, respectively. Furthermore, the target function may include terms for other types of restraints, e.g. RDCs.

Torsion angle restraints can be obtained from chemical shift values (**?**). The upper distance bounds are derived from signal intensities or volumes via distance calibration, using the simplifying assumption that the two interacting spins are isolated. The relationship between an inter-proton distance d and the corresponding NOESY cross-peak intensity or volume V according to the two-spin approximation is given by eq. I.3.4. Both quantities are connected by a calibration constant C.

$$V = C \times d^{-6} \tag{I.3.4}$$

However, neighbouring spins affect the magnetization transfer, thus NOESY signals of proteins cannot be precisely calibrated using eq. I.3.4. For this reason the obtained distances are not used as exact distance restraints but as upper distance limits b (see eq. I.3.3).

Due to signal overlap, peak intensities cannot be measured with a high degree of accuracy, which introduces additional errors into the calibration process. Nevertheless, on account of the inverse power of six-relationship in eq. I.3.4 errors in peak intensities or volumes do not necessarily give rise to incorrect distance restraints. Therefore the approach works well, despite the abovementioned limitations, as long as a sufficiently large number of experimental restraints is obtained.

As mentioned above, the calculation starts from a set of random conformers, and each of these molecules undergoes the structure calculation procedure separately. Afterwards, the thus obtained structural models are ranked according to their target function values, and the final structure is represented by a bundle of conformers that all satisfy the experimental restraints as well as possible. Typically, 20 conformers are reported.

Since the target function represents a simplified energy term the final structural model will exhibit somewhat inadequate physical properties. Therefore it is recommended to perform a refinement in explicit water subsequent to the structure calculation (**?**).

Quality indicators

There are a number of indicators that help the user judge the reliability of the result of the structure determination process. Ideally, the average CYANA target function value should not exceed 100 Å^2 after the first cycle of NOE assignment and structure calculation, and it should be less than 1 Å^2 after the final round of structure calculation. Less than 20 % of NOESY cross-peaks should remain unassigned, and less than 20 % of long-range NOESY cross-peaks should be discarded. After the first cycle the backbone RMSD value should lie below 3 Å, and the number of upper distance limits obtained from the automatic NOESY cross-peak assignment should increase or at least not show a marked drop during the progression of the calculation.

These criteria ought to be treated as guidelines only—fulfilment of all requirements does not guarantee that the structure is indeed accurate, and neither does the failure to meet one criterion imply that the final result is necessarily erroneous. However, if pronounced deviations from some of the ideal values occur the structural model as well as the input data should be examined carefully.

I.3.2 Alternative structure determination approaches

The classical, semi-manual approach to structure determination by NMR is time-consuming, particularly due to the requirement for assigned chemical shifts, obtaining which is still mostly done manually, but also because a number of different spectra for chemical shift assignment and distance information have to be recorded. Over the years, methods which (partly) make use of other kinds of data or less (processed) data have been developed, as well as fully automated procedures. A selection of these approaches will be introduced in this section, but as the classical method remains the most widely-used one, and as structures determined using NOESY data lie at the focus of this thesis, no detailed descriptions of the individual procedures will be given.

I.3.2.1 Fully automated methods

The FLYA algorithm (**?**) is a fully automated variant of the the traditional structure determination pathway, and an updated version has demonstrated the feasibility of automatically performing resonance assignments and subsequently calculating correct protein structures from high-quality, refined NOESY peak lists alone, i.e. without the use of through-bond spectra, which are traditionally required for chemical shift assignment (**?**). The limiting factor of automated structure determination, however, appears to be peak picking (**?**), since true signals missing from the input to the chemical shift assignment routine may render the assignment of certain atoms impossible, and the inclusion of noise peaks or artefacts may bring about erroneous assignments (**?**).

I.3.2.2 Chemical shift-based methods

Chemical shifts are indicative of the local environment of an atom. They can be used to provide torsion angle predictions (**??????**), and they can inform on secondary structure (**??**), but the dependence on the protein's non-covalent structure is weak and there is no single equation to link chemical shifts to inter-atomic distances or orientation information (**?**). Nevertheless, methods have been developed to predict chemical shifts from three-dimensional protein structures, e.g. SHIFTX/SHIFTX2 (**??**) and SPARTA/SPARTA+ (**??**). As chemical shift information does not provide any long-range restraints the tertiary structure of the target protein needs to be found by molecular modelling approaches (**?**), hence 'structure prediction' may be a term more applicable to the generation of structures from chemical shift data than 'structure determination'.

By now a multitude of tools have been developed that directly employ assigned chemical shift data for structure prediction. Examples include CHESHIRE (**?**), CS-ROSETTA (**??**), the CS23D web server (**?**), and CSI 3.0 (**?**). The latter does not aim to determine the entire protein structure but to accurately locate the elements of secondary and super-secondary structure. All of these tools employ molecular fragment replacement strategies, i.e. they access databases that link chemical shifts to known local conformations. Other approaches that also predict protein structures using chemical shift data aim to fold the protein from an extended conformation without the use of structural homology information. Instead, an energy function is minimized that combines a force field with a term penalizing deviations of predicted chemical shifts from the experimental values (**??**).

I.3.2.3 RDC-based methods

RDCs can be measured if the protein sample is brought into a liquid crystalline medium that brings about a partial alignment of the protein molecules, hence the dipole vectors of the atoms are constrained with respect to the same alignment tensor.

Approaches that make use of RDCs in combination with information from protein structure databases include molecular fragment replacement techniques (e.g. **???**), built into tools such as RDC-Rosetta (**?**). Analogous to the strategies employed by some of the programs that base their structure predictions on chemical shifts, a preliminary structure is constructed from fragments that are selected from a database on the basis of their fit to the RDCs and the primary structure. Some methods also include chemical shift values to guide fragment selection. Subsequent to the assembly of a preliminary structure from the chosen segments this structure is refined to yield the final structural model. The 2D-PDPA software (**?**), on the other hand, uses a library of complete protein structures whose RDC-patterns it back-calculates. These patterns are compared to the RDC-fingerprint of the target protein in order to identify a homologous structure. The program may also be used for the validation of protein folds.

Various other methods do not access structural databases, e.g. MECCANO (**?**), 3P (**?**), and RDC-PANDA (**?**). MECCANO requires RDCs measured in two different alignment media, and from the RDCs the backbone fold is determined. 3P relies on the periodic relationship between peptide plane orientation and coupling strength, and it may also make use of dihedral angle restraints. 3P also determines the backbone structure. RDC-PANDA uses RDCs to create structural elements that are combined with the help of NOEs, and the resulting structure is then employed to refine the assignment of NOESY spectra.

Chapter I.4

CYRANGE

Most proteins comprise structured and unstructured regions. It is important to identify these regions to meaningfully compare or analyze protein structures. The most commonly used similarity measure for three-dimensional structures are atomic root mean square deviation (RMSD) values, which are computed for all or a subset of atoms in two or more structures after their optimal superposition, which is the one that minimizes the RMSD value (**?**). For instance, NMR protein structures are generally represented by a bundle of conformers that have been calculated starting from different randomized initial conformations using identical input data, and it has been proposed to represent also crystallographic structures by an ensemble of conformations (**?**). The precision of an NMR protein structure is measured by the average RMSD value between the individual conformers and their average coordinates, computed after superimposing the conformers onto the first one. Both the superposition and the resulting RMSD values are strongly influenced by the choice of atoms that are included in the fit. Including unstructured parts of a structure yields large RMSD values that may obscure the presence of well-defined structured regions of the protein. It is therefore crucial to identify the structured regions. A (subjective) choice can be made by visually inspecting the structures, but for reasons of consistency, reproducibility, and efficiency an automated method is highly desirable. This chapter introduces a new method for this purpose that has several advantages over existing approaches.

I.4.1 The objective

The CYRANGE algorithm (**?**) yields residue ranges for the superposition of protein structures with the same sequence. The algorithm has been designed (i) to find residue ranges that are suitable for the global superposition of protein domains (rather than detecting local order), (ii) to provide simple residue ranges with no or only a small number of gaps, (iii) to include as many residues as reasonably possible, (iv) to be applicable without change to structure bundles of high and low precision, (v) to be applicable to multi-domain proteins (without input specification of the domain boundaries), (vi) to handle symmetric multimers and protein complexes, and (vii) to work with a single set of parameters for all proteins. CYRANGE requires as input the Cartesian coordinates of two or more structures and consists of two main steps, domain identification and residue range determination for each domain.

I.4.2 Other software tools for domain identification

Various methods have been proposed to identify well-defined regions of a protein on the basis of the atomic coordinates of an ensemble of structures, such as the bundle of conformers resulting from an NMR structure determination (**???????????**). Ordered regions can be identified by analyzing the local structure, for example by applying a cutoff on angular order parameters for the backbone torsion angles ϕ and ψ (**?**). Ordered regions found by these strictly local approaches cannot necessarily be superimposed simultaneously with a low RMSD value.

Other methods aim to find the part(s) of a protein structure that are sufficiently similar to each other within the ensemble. Some algorithms rely on the inter-atomic distance variance matrix (DVM) with elements $D_{ij} = \sigma(d_{ij})^2$, where $\sigma(d_{ij})$ is the standard deviation of the distance between the atoms i and j, computed over all structures in the comparison (**????**). Small matrix elements indicate that the position of the corresponding groups of atoms is similar in all members of the structure bundle.

Another approach, implemented in the molecular graphics program MOLMOL (**?**), superimposes the structures with the current set of atoms (starting with all atoms or a user-defined subset), and then discards in each step the atoms from the residue with the largest global displacement, until either the RMSD falls below a maximally acceptable value or a minimal number of residues is reached.

PSVS (Protein Structure Validation Suite, **?**) and FindCore (**??**) are also tools providing information about the ordered regions of an ensemble. These two tools in particular have been used by some of the NMR community's structural biologists and for this reason the current text includes a comparison of their results and the domains output by CYRANGE (see section III.1.2).

PSVS determines local dihedral angle order parameters only and is thus unable to provide information regarding the division of the structure into different domains that are globally ordered as well as locally. Furthermore, the determination of ordered residue ranges is not its core task; instead, it mainly aims to validate protein structures. FindCore does detect the presence of multiple domains within a protein structure bundle. The software bases its analyses upon the inter-atomic DVM of the query ensemble instead of the dihedral angle information used by PSVS, hence FindCore accesses information on the global structure of the protein that is unavailable to PSVS. FindCore then uses the distance variance information to compile a set of *core atoms*, i.e.

backbone atoms exhibiting a certain degree of global order with respect to one another. Finally, the core atoms are clustered to yield one or more domains, and the RMSD values output by the software are calculated from a superposition of the core atoms making up the individual domains.

In the field of protein structure prediction (**?**), two algorithms, LCS (Longest Continuous Segments) and GDT (Global Distance Test), have been established for detecting regions of local and global structure similarities. The LCS procedure finds the longest continuous segment of residues that can fit under a given RMSD cutoff. The GDT algorithm searches for the largest (not necessarily continuous) set of residues that deviate by no more than a specified distance cutoff (**?**). Results are reported as the percentage of residues under a given distance cutoff. A popular measure is the GDT total score, $GDT\_TS = (P_1 + P_2 + P_4 + P_8)/4$, where P_d is the fraction of residues that can be superimposed under a distance cutoff of d Å, which reduces the dependence on the choice of the cutoff by averaging over four different distance cutoff values.

A number of applications have been developed since the creation of CYRANGE. ENSEMBLATOR (**?**), for instance, is a tool that performs a comparison of structure ensembles on both global and local levels. It aims to facilitate analyses of NMR ensembles or the comparison of an NMR structure bundle with a single crystal structure with the aim of revealing differences at atomic or residue-level instead of merely computing a global RMSD value. A key part of the ENSEMBLATOR routine is the determination of a set of core atoms by employing a user-defined distance cutoff value d_{cut} to iteratively exclude all atoms that cannot be superimposed with an inter-conformer distance $d \leq d_{\text{cut}}$. **?** compared their software to CYRANGE and found that the two programs, as far as their functionalities overlap, produced roughly the same results for the selected query proteins.

The GeoStaS (Geometrically Stable Substructures) software (**?**) takes ensembles of experimental or theoretical origin (molecular dynamics (MD) or Monte Carlo simulations) as input and is designed to detect dynamic domains, i.e. clusters of similarly moving atoms. The application performs its task by comparing the trajectories of the atoms in a pairwise manner and it searches for their optimal superpositions. In contrast to CYRANGE, GeoStaS was developed explicitly to deal with large amounts of data, and it can read binary MD trajectory files as well as files in PDB format. A comparison with CYRANGE by the authors showed that GeoStaS often produced a more complex division of the protein sequence, in most cases yielding a larger number of domains than CYRANGE.

Like GeoStaS, the tool ResiCon (Residue Contacts analysis) (**?**) aims to identify the dynamic domains of a protein—and additionally hinges and interface regions—, and it does so by analyzing the geometric variability of intra-residue contacts between non-neighbours within the query structure bundle.

I.4.3 The need for CYRANGE

Until the creation of CYRANGE there was no one commonly used method for the determination of the ordered domains of a protein, and all existing tools had their limitations and shortcomings. As stated above, PSVS, for example, only takes local order into account, and FindCore often produces domains that visual inspection shows to be rather fragmented and overall too complex.

The aim was to fill the gap by creating a tool that would be easy to use. It should produce straightforward results that intuitively make sense by visual inspection, with domains including as many residues and as few intra-domain gaps as possible while not disproportionately increasing the RMSD values of the domains.

Chapter I.5

Protein structure validation

Protein structures at atomic resolution are required for various scientific purposes, e.g. the investigation of cellular processes on a molecular level, the identification of new drug targets, and the structure-based development of novel drugs. In order to be able to draw valid conclusions from a structural model, however, this model must be correct, hence it needs to be validated before being used as the basis of further enquiry. This chapter gives an overview of a number of approaches to estimating the quality of protein structures. Furthermore, it introduces the structure validation procedure CYVAL, which was developed as a part of this thesis.

I.5.1 Background

A multitude of structural models from both NMR and X-ray crystallography had to be retracted after they were first published because their depositors or other researchers discovered them not to be accurate representations of the underlying experimental data, (e.g. **??????**). In the case of the original NMR structure of the oligomerization domain of p53 (**?**), for instance, the relative orientation of the two dimers was flawed, and the correction of three NOE assignments as well as the inclusion of a number of additional NOE restraints led to the calculation and deposition of a corrected structure (**?**). This example illustrates that a small number of errors can have a potentially drastic effect on the overall accuracy of a structural model (**?**), and that great care needs to be taken to ensure that a correct result is published.

Broadly speaking, structure validation methods can be divided into two categories, knowledge-based and data-based approaches (see section I.5.4 for details). The former make use of sets of known, high-quality X-ray structures from which certain quality indicators are derived (e.g. standard bond lengths or allowed torsion angles) against which the programs judge the query structure. Data-based procedures, on the other hand, aim to assess how well the structural model is in accordance with the underlying experimental data, hence they provide the user with a clearer indicator of the accuracy of the structure.

Knowledge-based validation cannot yield meaningful results in case the criteria against which the validation is performed were already employed and optimised during the structure calculation process. Thus, largely knowledge-based methods of structure determination or structure prediction, e.g. Rosetta and its variants that take into account some experimental data (see section I.3.2), do not produce structures that can be assessed via criteria such as sidechain rotamers, functional group planarity, etc. Structural models arrived at by such methods will often receive excellent knowledge-based quality scores, yet these models may deviate strongly from the native target structure (**??**).

The appropriate method for estimating the quality of a particular structural model is closely linked to how this model was determined. A structure calculated from NMR data via the classical approach (see section I.3.1) can be validated against the NOESY spectra that provided the distance information used during structure calculation; it can also, albeit with less meaningful results, be validated against geometric criteria. Protein structures determined by X-ray crystallography may also be assessed using knowledge-based criteria, but, once again, validation against the actual experimental data leads to more valuable results since it provides information on the accuracy of the structural model.

A multitude of knowledge-based structure validation tools exist, but for NMR-derived structures the number of data-based validation programs is low, for reasons which will be outlined below. It is these tools, however, that would be particularly valuable to the user community since only the degree of correspondence between experimental data and the structural model can give an indication of the latter's accuracy. A useful validation tool for structures calculated from NMR data should meet the following requirements: firstly, it should be as user-independent as possible so as to reduce the risk of the inadvertent introduction of biases or even errors; secondly, it ought to consider experimental data instead of basing its assessment solely on knowledge-based factors. Thirdly, the method should be able to cope with the presence of noise and artefacts, and it should have the ability to analyze ensembles instead of single conformers only. Another important factor is the ease of use of the program and the output of a validation metric whose value is easy to interpret and does not depend on the fine-tuning of various parameters by the user. All these factors were taken into account in the design phase of our data-based protein structure validation method CYVAL, which will be introduced in more detail in sections I.5.5 and II.2.

The following sections give an overview of established structure validation metrics and procedures for X-ray crystallography and, in particular, for NMR. Additionally, the text explains why the quality of structures derived from these two experimental methods cannot be evaluated in a completely analogous fashion if data-based validation is to be performed, and why it is the assessment of NMR-derived protein structures that is more complicated in comparison.

I.5.2 Validation of X-ray structures

In the field of X-ray crystallography there are a number of structure quality indicators, e.g. resolution, reflection completeness, the R-factor, and R_{free} (see below). The resolution can serve as an initial approximation of structure quality as it is a measure of the minimum distance of structural components that are distinguishable on the electron density map (**?**).

Commonly, the quality of a crystal structure is assessed by the R-factor of the structure (eq. I.5.1), which depends on the observed and back-calculated structure factor amplitudes, F_{obs} and F_{calc}, respectively.

$$R = \frac{\sum |F_{\text{obs}} - F_{\text{calc}}|}{\sum |F_{\text{obs}}|} \tag{I.5.1}$$

The R-factor aims to quantify the deviation of the model from reality, and, according to the definition given by eq. I.5.1, a lower value indicates a better agreement between the structural model and the data. However, the R-factor has no built-in safeguard against overfitting by the introduction of parameters that have no basis in the actual experimental data. The addition of an unreasonably large number of water molecules into the noisy parts of the solvent region, for example, will lower the R-factor. This means that a low R-factor may be obtained even at a low information content of the atomic model (**?**).

There is, however, a certain degree of redundancy inherent in diffraction data, therefore the recorded data may be divided into a *working set* and a *test set*. The former is used to determine the structure, whereas the latter is only employed in structure validation using the free R-factor (R_{free}) (**?**). R_{free} is computed just like the standard R-factor, the only difference being that it employs reflections from the test set T only. R_{free} is thus not influenced by overfitting, and it is defined according to eq. I.5.2.

$$R_{\text{free}_T} = \frac{\sum\limits_T |F_{\text{obs}} - F_{\text{calc}}|}{\sum\limits_T |F_{\text{obs}}|} \tag{I.5.2}$$

I.5.3 Validation of NMR structures

Several methods for the calculation of R-factors for NMR-derived structures, analogous to the R-factor or R_{free} from X-ray crystallography, have been proposed (**????????**). In general, a NOESY spectrum is back-calculated from the structural model and compared to the experimental data that were used for structure calculation. However, for a number of reasons the validation of NMR structures in this way is significantly more complicated than for X-ray crystallography. Firstly, a set of NOESY peaks contains far fewer redundancies than reflections from an X-ray experiment, where each data point contains information regarding the entire structure (**?**), whereas

NOE data points only pertain to two atoms each, which cannot be further apart than a comparatively short distance, hence NOE information is local. Secondly, accurate determination of NOE peak intensities can be difficult due to signal overlap, baseline problems, or molecular dynamics, which makes a direct comparison of signal intensities error-prone. Thirdly, reflections in X-ray crystallography can generally be unambiguously assigned, hence only their intensities determine the *R*-factor. In NMR, in contrast, the assignments of a multitude of experimental peaks may be ambiguous, and many signals are in fact artefacts. Furthermore, a structure calculation from NMR data generally uses more than one restraint type, and the inclusion of all possible data types into one factor would be complicated (**?**).

On account of the calculation of an NMR *R*-factor not being straightforward oftentimes knowledge-based (see section I.5.4.1) or restraint-based indicators of structure quality have been used instead of true data-based metrics. Relatively simple NMR-related measures are the completeness of the resonance assignment; the restraint count, either in total (**?**) or per residue; the number and magnitude of restraint violations; and the ratio of observed to expected NOESY signals (**???**). On a more advanced level, the degree of restraint redundancy may be analyzed (**??**). Ensemble precision, too, expressed as the RMSD value of the conformer bundle, is often cited in the context of structural quality; however, the precision of an ensemble is no indicator of its accuracy, and the former generally overestimates the latter (**?**). The precision may even be virtually insensitive to the quality of the data since it can be strongly influenced by the methods of structure refinement and conformer selection (**?**).

A report on the number of restraints per residue may potentially yield a biased picture of the quality of the entire restraint set, depending on how this number was derived (**?**). An inclusion of residues from very flexible regions of the molecule, for which few or no restraints are available, will lower the restraint count per residue and may thus obscure the presence of an extensive network of restraints in other parts of the structure. Furthermore, the total number of restraints and hence also its mean value depends on the shape and amino acid composition of the protein, and, additionally, the degree of redundancy of the experimental data may be high, which lowers the overall information content of the entire set of restraints (**??**).

Violations of experimental restraints are often used as a criterion for conformer selection during structure calculation (see section I.3.1). As the structure determination process also leads to the exclusion of certain restraints the final list of experimental restraints is often a product of manual interpretation and adaptation with the goal of minimizing restraint violations. Consequently, the refined set of restraints is not necessarily a suitable basis for the assessment of the quality of the structural model (**?**).

The ratio of assigned observed to expected NOESY signals, also known as *NOE completeness*, aims to provide an estimate of the information contained in a set of distance restraints, which is more meaningful than a mere count of restraints per residue. However, once again redundancies are not taken into account. Moreover, since the number of expected NOESY signals depends on the structure, it can only be determined once the structure has been calculated, which means that the reported NOE completeness depends on the structural model, but this, in turn, depends on the completeness (**?**).

Redundancies within a set of experimental NMR restraints are very common (**?**). These redundancies can be detected and evaluated using the QUEEN method (QUantitative Evaluation of Experimental NMR restraints), for example (**?**). This tool quantifies the information contained in NMR-derived distance and J-coupling data, and it identifies mutually inconsistent restraints alongside the crucial restraints for a given structure determination, i.e. the restraints that are structurally important as well as unique. However, on its own the degree of restraint redundancy does not give any indication of the accuracy of the resulting structural model: on the one hand, a large percentage of redundant restraints points to the data set having a rather low information content; on the other hand, redundant restraints also confirm each other's validity.

Despite the difficulties inherent in the determination of an NMR R-factor some data-based tools for the quality assessment of NMR structures exist. RFAC and its successor, RFAC-3D (**??**), for example, aim to perform R-factor calculations from processed NOESY spectra, and the tool RPF (**??**) calculates a metric known as DP-score from NOESY peak lists. Both programs will be described in more detail in section I.5.4.2.

I.5.4 Tools for the validation of protein structures

Different structural properties may be used for the estimation of a structure's quality. On the one hand, the structural model can be assessed on the basis of its correspondence with the underlying experimental data; on the other hand, the atomic coordinates can be compared to characteristics based on the study of a large number of high-quality protein structures. The second, knowledge-based approach judges how physically realistic the model is, whereas data-based validation gauges whether the structure calculation process correctly incorporated the experimental data. Data-based methods aim to give an indication of the similarity of the structural model to the native structure, yet it is important to note that only cross-validation, i.e. the validation of the model against independent experimental data, can properly estimate structural accuracy (**?**).

The current section introduces selected knowledge-based and data-based programs that can be used for a quality assessment of NMR-derived protein structures. Some of these tools were originally created for the examination of structural models from X-ray crystallography, while others were developed specifically with NMR-derived structures in mind. Naturally, data-based validation tools must focus on a certain experimental method since the nature of the data is at the centre of the validation procedure.

I.5.4.1 Knowledge-based tools

Various aspects of a protein structure can be used for quality estimation: covalent geometry, H-bond geometry and satisfaction, and core packing, for example. These quantities can easily be analyzed using tools that were originally created for the validation of X-ray structures; however, the existence of ensembles of conformers in the case of NMR-derived structures must be taken into account. In practice, this often means that each conformer needs to be analyzed separately and the user has to evaluate the different results.

In general, a single score calculated by a knowledge-based validation tool is often not sufficient to differentiate between correct and moderately incorrect protein structures as often the score values obtained for correct and incorrect structures of the same protein may overlap (**?**). What is more, models deviating strongly from the correct structure may nevertheless exhibit excellent geometric validation statistics (**??**). Thus, in the absence of a reference structure, several scores that assess different aspects of the structure ought to be computed, but even satisfactory validation results cannot guarantee that the structural model is indeed correct. Besides, knowledge-based validation becomes meaningless if it evaluates parameters that directly influenced the structure determination process. If sidechain rotamer libraries, for example, were used to guide the structure calculation, a quality score based on sidechain rotamer data will invariably classify the structural model as satisfactory.

This section introduces a number of knowledge-based validation tools that were originally developed to be used on X-ray structures but that can also be employed to gauge the structural quality of protein structures determined by NMR.

MolProbity

MolProbity (**????**) evaluates and scores multiple structural characteristics. It is renowned for its all-atom contact analysis, during which it detects close contacts, i.e. non-covalently bonded atom pairs that are not separated far enough. The best-known and most widely used scores this tool computes are the so-called *clashscore*, which reports the number of close contacts per 1,000 atoms, and the *MolProbity-score*, which combines the clashscore, the number of Ramachandran outliers, and the percentage of unfavourable side-chain rotamers. Each of the three constituents is weighted so that the total score represents the crystallographic resolution at which a score value of this magnitude can typically be expected.

The Ramachandran analysis performed by MolProbity is based on current data, and the backbone statistics are subdivided into four groups (glycine, proline, pre-proline, and the remaining common L-amino acids). This allows for a more up to date evaluation of the distribution of the ϕ and ψ dihedral angles than what is possible when using the PROCHECK-NMR software (**?**), which is still often employed for this purpose despite no longer being maintained (**?**).

Verify3D

Verify3D (**???**) analyzes the compatibility of the atomic model of a protein with its own amino acid sequence using the concept of residue environments. To characterize the environment of a residue Verify3D takes three factors into account, namely the area of the residue's side chain that is buried by atoms of other residues, the portion of the side-chain area covered by polar or water atoms, and the local secondary structure. Based on these parameters each position within the protein sequence is assigned an environment class, thus a one-dimensional string of environment values is derived from a three-dimensional structure.

Each amino acid has a statistical preference for certain environments within a folded protein. Using this knowledge Verify3D aligns the sequence of the query protein to the generated string of environment values and calculates a match score based on the preference of each of the residues of the protein for the environment class present at its position within the sequence. Thus an overall match score can be determined that expresses the fit between the query structure and its own sequence. The match scores for each combination of amino acid residue and environment class are referred to as *3D-1D scores*, and as the total match score of the protein depends only on the environment classes the match is independent of sequence similarities between the query protein and proteins that were used in generating the environment class information database.

The total Verify3D validation score is the sum of the 3D-1D scores of all residues within the sequence. A higher score indicates a better fit, although one must bear in mind that large, generally correctly represented proteins will automatically produce higher scores than small proteins as there are more contributions to the overall value. Moreover, the score profile becomes more informative with size as internal residues, which are more numerous in large proteins, provide more information. Furthermore, Verify3D outputs plots of the average 3D-1D score for residues in a sliding window with a length of 21 residues, hence local structure irregularities can be made visible.

WHAT_CHECK

WHAT_CHECK (**?**) combines a large number of checks of the geometric and stereochemical properties of the query structure, such as the Ramachandran plot appearance, backbone conformation, and χ_1-χ_2 rotamer normality, to name but a few. The program expresses its evaluation of the different parameters as Z-scores (eq I.5.3).

$$Z = \frac{x - \overline{x}_{db}}{\sigma_{x_{db}}} \tag{I.5.3}$$

A Z-score describes the relationship between a quantity x and this quantity's mean value from a database ($\overline{x}_{db}$), normalized by the standard deviation $\sigma_{x_{db}}$. Hence the score reports the distance, measured in units of standard deviation, of a data point from the mean. The underlying assumption is that the quantity is normally distributed. Z-scores that are separated from the mean by more than four standard deviations are considered outliers. Outliers are to be expected, and a certain value being flagged as an outlier does not necessarily mean that this value is erroneous. However, the highlighted aspect of the structural model ought to be checked.

ProSa2003

ProSa2003, the successor of ProSa-II, computes a knowledge-based energy value for each of the amino acids making up the query structural model (**??**). Based on these energies the software aims to distinguish physically correct protein structures from misfolded cases.

The energy values it calculates are based on the forces stabilizing known native folds in solution, from which a database of potentials of mean force was created that is accessed by ProSa2003. These potentials are based upon the energies of pair interactions of C^{α} or C^{β} atoms belonging to particular amino acids as a function of the spatial separation of the atoms, and the total pair interaction energy E of the molecule is equal to the sum of the individual contributions. E depends on the protein sequence as well as on the conformation, i.e. ProSa2003 examines the overall arrangement of the protein chain instead of focusing on the violations of basic steric principles.

The program outputs energy Z-scores, where lower score values indicate a more favourable conformation. Additionally, ProSa2003 creates a plot of the residue interaction energies, where the interaction energy of each residue with all other residues within the protein is plotted against the sequence position. This graphical representation of the results highlights strained and thus structurally problematic sections of the query structure.

ProQ

ProQ (**??**) is a method that derives its estimate of the similarity of a protein model to the native fold from a number of structural features. Based on these features the software predicts two metrics that aim to express the accuracy of a structural model, LGscore (**?**) and MaxSub (**?**). These scores may be used as quantifiers of structural similarity in lieu of the $\mathrm{RMSD}_{\mathrm{ref}}$-value. Both measures detect segments that the query model has in common with the known reference structure, but while MaxSub aims to identify the largest subset of C^{α} atoms of a structural model that superimpose well with the true structure, LGscore states the probability of encountering a higher structural similarity by chance. Both scores depend on the length of the query protein; however, LGscore favours long proteins while MaxSub is more likely to take a favourable value for short proteins, hence the two metrics balance each other.

I.5.4.2 Data-based tools

In general, all NMR data that are employed in the structure determination process can be used for validation purposes. An R-factor using RDC data, for example, has been proposed (**?**), and programs such as 2D-PDPA (?) may be used to validate a protein structure via the comparison of the pattern of experimental and back-calculated RDCs. C^{α} chemical shifts, too, have been employed to assess the accuracy of a protein model (?), and for the field of solid-state NMR a procedure has been developed that aims to verify a structure against an unassigned 2D ^{13}C-^{13}C spectrum (?). However, since the traditional approach to NMR structure determination still relies heavily on NOESY data and because our own structure validation software, CYVAL, bases its structure evaluation on NOESY signals, this section will restrict itself to the description of two data-based tools for the validation of NMR-derived protein structures that also employ these data in their analyses.

RPF and the DP-score

The RPF tool (??), which computes the eponymous validation scores *Recall*, *Precision*, and *F-measure*, as well as the *DP-score*, performs an assessment of structural quality that is based on NOESY peak lists. The calculation of its validation metrics requires the construction of two NOE networks: G_{ANOE}, the network of ambiguous NOEs, contains all proton-proton interactions that are feasible based on the unassigned peak lists and the resonance assignments, which are used to create ambiguous NOESY cross-peak assignments; $\bar{G}$ is the target structure's network of inter-proton distances $d \leq d_{\text{NOE\_max}}$, which is the maximum inter-proton distance for which an NOE can still be observed.

The method assumes the existence of four categories of proton-proton interactions for the comparison of G_{ANOE} and $\bar{G}$:

1. *true positive* (TP): an interaction present in $\bar{G}$ (i.e. within the structure) that can also be found within G_{ANOE} (i.e. within the set of experimental signals)
2. *true negative* (TN): an interaction present neither within the structure nor within G_{ANOE}
3. *false positive* (FP): an interaction present within $\bar{G}$ but not supported by G_{ANOE}
4. *false negative* (FN): an interaction missing from $\bar{G}$ but present within G_{ANOE}

An experimental peak p will only be classified as *false negative* if none of the possible interactions from G_{ANOE} can be observed in the structural model.

Recall is defined as the fraction of experimental NOESY cross-peaks that are represented by the structure. It is based on the presence or absence of peaks and inter-proton distances within the distance limit, but disregards distance deviations. The definition is given by eq. I.5.4, where h_1 and h_2 represent the two protons of the interaction in question.

$$\text{Recall}\left(\bar{G}\right) = \frac{\left|\{p|p\left(h_1,h_2\right) \in G_{\text{ANOE}}, d\left(h_1,h_2\right) \in \bar{G}\}\right|}{\left|\{p|p\left(h_1,h_2\right) \in G_{\text{ANOE}}\}\right|} \tag{I.5.4}$$

The Precision metric (eq. I.5.5) is distance weighted, and it is defined as the fraction of NOESY-relevant interactions observed within the structure that are actually supported by the data, represented by G_{ANOE}. Using the inverse sixth power of the distances d reduces the influence of weak signals, stemming from structurally relevant long-range interactions, and for the same reason it also renders the method less sensitive to the exact choice of $d_{\text{NOE\_max}}$. As with the Recall measure, there is no direct comparison of specific inter-proton distances.

$$\text{Precision}\left(\bar{G}\right) = \frac{\sum\limits_{\substack{d(h_1,h_2)\in\bar{G},\\ p(h_1,h_2)\in G_{\text{ANOE}}}} d\left(h_1,h_2\right)^{-6}}{\sum\limits_{d(h_1,h_2)\in\bar{G}} d\left(h_1,h_2\right)^{-6}} \tag{I.5.5}$$

Neither Recall nor Precision on their own provide a very reliable assessment of the accuracy of a protein structure, therefore the two metrics are combined to yield the F-measure (eq. I.5.6). This metric performs better than either of its components in providing an overall evaluation of the degree of correspondence between the structure and the experimental data, assuming the resonance assignment as well as the peak lists are complete, and that the latter contain neither noise nor artefacts.

$$F\left(\bar{G}\right) = \frac{2 \times \text{Recall}\left(\bar{G}\right) \times \text{Precision}\left(\bar{G}\right)}{\text{Recall}\left(\bar{G}\right) + \text{Precision}\left(\bar{G}\right)} \tag{I.5.6}$$

As the F-measure does not take into account the quality of the NMR data provided for structure determination and validation the *Discriminating Power* (DP-score) was introduced. It measures to what extent the structural model is distinguished from a freely rotating polypeptide chain, represented by the interaction network G_{free}. This means that the information content of the data is considered as well, which reduces the strong influence of short-range interactions, which are of little importance for the overall fold. Furthermore, the DP-score requires the construction of a graph G_{ideal} of a hypothetical structure that is in perfect agreement with G_{ANOE}. The score is calculated as given by eq. I.5.7.

$$\text{DP}\left(\bar{G}\right) = \frac{F\left(\bar{G}\right) - F\left(G_{\text{free}}\right)}{F\left(G_{\text{ideal}}\right) - F\left(G_{\text{free}}\right)} \tag{I.5.7}$$

The authors provide data from three different proteins that illustrate that the DP-score performs noticeably better than each of the other three metrics when it comes to assessing the accuracy of a structural model, measured as the RMSD value to the known reference structure, based on all heavy atoms within certain residue ranges. In each case, a DP-score close to 1 signifies a structure that is close to the native structure, and on the whole the score drops with increasing RMSD values, i.e. with a decrease in structural accuracy.

RFAC and RFAC-3D

The RFAC program (?) and its successor RFAC-3D (?) base their automated structure validation on a comparison of simulated and experimental NOESY spectra. RFAC calculates seven different R-factors in total to facilitate a detailed validation of the structural model. The major difference when compared to previously published programs computing NMR R-factors is that RFAC takes unassigned experimental NOESY signals into account, instead of restricting its analysis to assigned peaks.

The software takes a structural model and a complete list of assigned chemical shifts from which it generates a set of back-calculated NOESY peaks and their intensities via relaxation matrix analysis. Furthermore, RFAC automatically picks peaks in the experimental spectrum and assigns each of them a reliability indicator p, calculated using Bayes' theorem. p states the probability of the peak being a true signal, and these probabilities serve as weighting factors during R-factor computation. By comparison with the list of back-calculated peaks RFAC assigns the experimental signals.

? recommend the use of two of their R-factors in particular, R_3 and R_5. R_3 (eq. I.5.8) estimates to what extent the assigned experimental peaks are in accordance with the structural model. However, this R-factor does not provide a complete assessment of structural quality because it does not take into account the number of unassigned peaks, which are potentially incompatible with the query structure.

$$R_3\left(\alpha\right) = \sqrt{\frac{\sum\limits_{i\in A}\left(V^{\alpha}_{\mathrm{exp},i} - \sigma_\alpha V^{\alpha}_{\mathrm{calc},i}\right)^2 p^2_{\mathrm{exp},i}}{\sum\limits_{i\in A} V^{2\alpha}_{\mathrm{exp},i} p^2_{\mathrm{exp},i}}} \tag{I.5.8}$$

Here, i refers to an individual peak from the set of assigned peaks (A), V denotes the peak volume, σ_α is a scaling factor, and α is an exponent typically set to $-\frac{1}{6}$ in order to convert peak volumes to quantities that are proportional to inter-atomic distances, so as not to let short-range, i.e. high-intensity, peaks dominate the R-factor. p is the abovementioned probability factor.

R_5 (eq. I.5.9) states how well the experimental signals are explained by the back-calculated peaks. This R-factor takes the set U of unassigned experimental signals into account as well, and it compares their volumes to the standard noise peak volume V_{noise}.

$$R_5\left(\alpha\right) = \sqrt{\frac{\sum\limits_{i\in A}\left(V^{\alpha}_{\mathrm{exp},i} - \sigma_\alpha V^{\alpha}_{\mathrm{calc},i}\right)^2 p^2_{\mathrm{exp},i} + \sum\limits_{i\in U}\left(V^{\alpha}_{\mathrm{exp},i} - \sigma_\alpha V^{\alpha}_{\mathrm{noise}}\right)^2 p^2_{\mathrm{exp},i}}{\sum\limits_{i\in A} V^{2\alpha}_{\mathrm{exp},i} p^2_{\mathrm{exp},i} + \sum\limits_{i\in U}\left(V^{\alpha}_{\mathrm{exp},i} - \sigma_\alpha V^{\alpha}_{\mathrm{noise}}\right)^2 p^2_{\mathrm{exp},i}}} \tag{I.5.9}$$

According to **?** calculating R_5 using only long-range and non-assigned NOEs yields the most meaningful value of all available R-factors when a global assessment of structure quality is desired. A combination of R_3 with an approapriate subset of peaks is particularly useful when examining specific regions of the protein structure.

RFAC-3D (**?**) extends the functionality of its predecessor by accepting 3D as well as 2D NOESY spectra, which has made the method accessible for larger proteins. Additionally, RFAC-3D can calculate R-factors from a combination of several different NOESY spectra, and a new R-factor is introduced that includes unassigned predicted peaks.

I.5.4.3 Bundle tools

As explained in section I.5.4.1, one single knowledge-based score cannot provide a complete assessment of the quality of a protein structure. Consequently, several different scores ought to be considered together (**?**). A convenient way to obtain the required variety of quality checks is to employ a validation tool bundle that automatically prepares the input required by its constituent programs and both collects their results and formats them for easy accessibility. This section introduces three well-known validation bundles alongside the new CYANA macro `validate.cya`, which was developed as part of this thesis and which allows the user to call a number of knowledge-based validation programs from within CYANA.

PSVS

PSVS (Protein Structure Validation Suite; **?**) can be applied to both NMR and X-ray structures, and it combines access to programs such as MolProbity (**????**), Verify3D (**???**), ProSa (**?**), PROCHECK (**?**), and RPF (**??**), amongst others. The tool allows the user to restrict the analysis of the structural model to its ordered regions or to a residue range of the user's choice. PSVS creates an overview of the results and for NMR structures it computes ensemble averages of scores that are computed separately for each conformer, e.g. the MolProbity clashscore. Chemical shifts as well as their assignments may also be checked, and the tool lists restraint violations if restraint files were uploaded alongside the structural model.

CING

CING (Common Interface for NMR structure Generation; **?**) incorporates 25 different external and internal programs and procedures, and its focus are structural models calculated from NMR data. It tests the experimental restraints for internal consistency and analyzes the completeness and information content of distance restraints using the QUEEN method (**?**); additionally, it validates chemical shift values based on structural and sequence information (**?**), and it invokes programs such as WHAT_CHECK (**?**) and PROCHECK-NMR (**?**) as well as internal routines for assessing the geometrical quality of the query structure. As with PSVS the analyses may be restricted to user-defined or ordered regions of the ensemble. What sets CING apart is its straightforward classification of the various validation results via colour coding: scores shown in green indicate the absence of detected problems, red shows that potentially serious issues with a particular aspect of the structure were discovered, and orange values lie between those two extremes.

Vivaldi

Vivaldi (VIsualization and VALidation DIsplay; **?**) was developed by the Protein Data Bank in Europe for the validation of NMR structures deposited in the PDB. The tool accesses a number of validation scores provided by CING (**?**) and combines them with internal routines to evaluate chemical shifts and various experimental restraints. Vivaldi, too, has the capability to determine the ordered regions within the ensemble. As yet, Vivaldi does not provide any facility for file upload and validation requests by external users, as its purpose is to prepare validation reports for NMR structures already present in the PDB.

`validate.cya`

The new CYANA macro `validate.cya` mediates the invocation of a number of knowledge-based validation tools from within CYANA, thus users can easily perform different knowledge-based validations on their newly calculated protein structure without first having to upload the coordinates to an external web server. All calculations are performed locally, hence not even an internet connection is required. Like the other bundle tools introduced in this section `validate.cya` automatically prepares the input files required by the various external programs, and it collects the results in a concise validation report, while at the same time keeping the detailed original results well-ordered and accessible for the user. By default, the macro invokes ProSa2003 (**?**), MolProbity (**????**), PROCHECK-NMR (**?**), WHAT_CHECK (**?**), Verify3D (**???**), ProQ (**??**), and the PDB validation software (downloadable from `http://sw-tools.pdb.org/apps/VAL/`).

I.5.5 CYVAL, our new data-based protein structure validation method

The accuracy of a protein structure is the most helpful information for researchers who aim to base their analyzes on this model. Knowledge-based quality estimates cannot reliably provide this information, and, as far as protein structures from NMR are concerned, validation tools that perform their assessment of the structure by taking into account experimental data are thin on the ground. RPF (**??**), the current gold standard of data-based validation, uses NOESY peak lists, i.e. processed data, instead of raw spectra. RFAC (**??**) operates on the actual spectra, but its method is not straightforward and the results appear to be highly dependent on user choices and thus not necessarily easy to reproduce or interpret. Furthermore, its main recommended *R*-factor does not take into account the absence of peaks that were expected based on the structure.

Our aim was to develop a procedure that assesses the correspondence between a protein structure and its underlying experimental data, i.e. which gives an indication of the degree of accuracy of the structure. The software should operate without user intervention in order to yield results that are as reproducible and objective as possible, consequently the input data should not have undergone excessive processing and adjustment. Furthermore, the performed analyzes should be as simple as reasonably possible in order to render them easily comprehensible and to allow for a straightforward interpretation of the validation result.

The thus developed method, CYVAL, is integrated into the structure calculation software CYANA. This implies that structure validation can be performed not only as a separate evaluation run but even directly and seamlessly once the structure has been determined. Apart from the atomic coordinates of the structural model the method requires a list of assigned resonances and the NOESY spectra whose peak data were used in structure calculation. CYANA's built-in automatic peak picking functionality (developed by Julia Würz) is employed to extract a list of experimental signals from the spectrum, and CYVAL constructs a set of NOESY peaks that can be expected based on the structural model. Using a component of the Peakmatch method (**?**) CYVAL performs a comparison of these two peak lists to yield sets of matching, missing, and unpredicted peaks. From these peaks it calculates the overall quality indicator, ζ, which is computed from distance deviations and peak reliability factors. The method is described in full detail in section II.2.

What sets CYVAL apart from RPF is CYVAL's direct use of spectra instead of peak lists. Furthermore, the analysis of the correspondence between structure and data that CYVAL performs takes distance deviations into account, which RPF neglects. The latter program also requires the provided peak lists to be as free of noise and artefacts as possible, as it is incapable of separating these from true signals. The general approach of RFAC is similar to that of CYVAL, but RFAC only accepts spectra files in Bruker format and can only work with up to three dimensions, whereas CYVAL accepts three file formats (Bruker, UCSF, and XEASY) and can handle four-dimensional spectra as well. Furthermore, the peak prediction procedure employed by CYVAL, unlike that of RFAC, is not relaxation matrix-based and is thus easy to grasp. Naturally, the simpler approach of CYVAL could lead to a lower quality of the predictions; however, **?** themselves note that choosing a detailed motional model for peak prediction with the relaxation matrix approach has a negligible effect on the *R*-factor value, which suggests that the exact method used for peak prediction may be of minor importance. Another point of difference is the goal of CYVAL to be as independent of user choices as possible, while the end results output by RFAC appear to depend considerably on decisions and selections made by the user. Finally, CYVAL routinely considers all mismatches between the sets of predicted and experimental peaks, i.e. it does not limit itself to either missing or unexpected experimental signals. RFAC-3D has introduced an *R*-factor that also considers all three possibilities, but the authors recommend the usage of an *R*-factor that disregards missing predicted peaks. Nevertheless, the *R*-factors calculated by RFAC and its successor show a good correlation with the degree of accuracy of the test structures, but despite these results neither program has come into general use (**?**). Consequently, there seems to be a spectra-based validation gap, and CYVAL aims to fill it.

Part II

Methods

Chapter II.1

CYRANGE

II.1.1 The CYRANGE workflow

II.1.1.1 Domain identification

Domain identification follows the approach used in the NMRCORE (?) and NMRCLUST (?) algorithms. Similar ideas have been used earlier (????). Dihedral angle order parameter values of torsion angles from all conformers are computed. A cutoff value is calculated from the order parameter list and applied to select the torsion angles that will be used to identify the *core atoms*. These atoms are located in locally well-defined regions of the structure bundle, and are therefore potentially involved in domains that can be superimposed with low RMSD. The variances of the intra-conformer distances between all core atoms are used to cluster the core atoms, which eventually yields the single or multiple domains present in the structure bundle.

Angular order parameter calculation

Dihedral angle order parameters (?) are calculated from all rotatable dihedral angles (except the peptide bond dihedral angle ω). For a given torsion angle, the angular order parameter S with $0 \leq S \leq 1$ is given by

$$S = \frac{1}{N} \left| \sum_{k=1}^{N} e^{i\theta_k} \right| \tag{II.1.1}$$

where the sum runs over the values θ_k of a torsion angle in all N structures in the comparison. The higher the local order the higher the value of S. If angular order parameters are computed from only the dihedral angles ϕ, ψ, and χ^1 the final results are largely identical to those obtained with the approach described above (see III.1). Using all rotatable dihedral angles was given precedence as this approach yields a larger number of core atoms, and thus a larger base for clustering, while still excluding atoms from severely disordered parts of the protein.

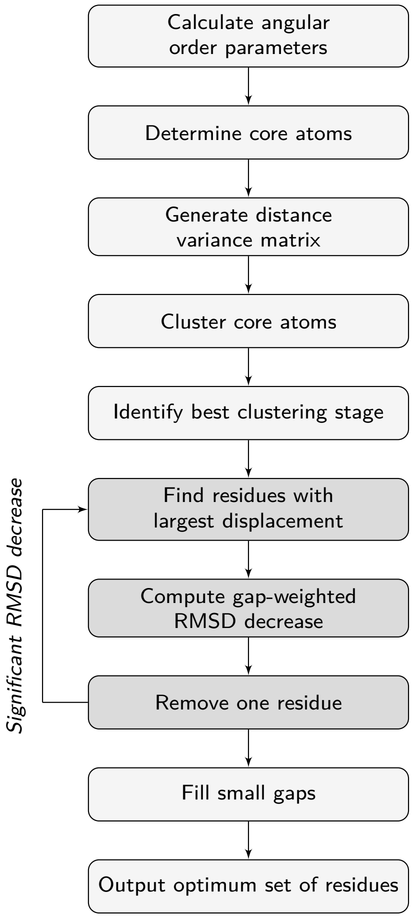

Calculate angular order parameters
Determine core atoms
Generate distance variance matrix
Cluster core atoms
Identify best clustering stage
Find residues with largest displacement
Compute gap-weighted RMSD decrease
Remove one residue
Significant RMSD decrease
Fill small gaps
Output optimum set of residues

Figure II.1.1: Flowchart of the CYRANGE algorithm for finding residue ranges for the global superposition of protein structures.

Core atom determination

The set of core atoms consists of the C^α atoms of those residues that contain at least one well-ordered torsion angle with an angular order parameter $S > S_{\mathrm{cut}}$. We did not want to impose a fixed cutoff value S_{cut} because the degree of order within structure bundles is different in each case. Instead, the cutoff value S_{cut} is chosen as the angular order parameter value S_i of the torsion angle i that maximizes the quantity

$$Q_i = (s-1)\,\frac{S_i - S^{\mathrm{min}}}{S^{\mathrm{max}} - S^{\mathrm{min}}} - r_i \qquad \text{(II.1.2)}$$

Here, s denotes the total number of torsion angles for which angular order parameters are calculated, S^{max} and S^{min} are the maximal and minimal angular order parameter values, and $r_i \in \{1, \ldots, s\}$ is the rank of the torsion angle i in an ordered list of the angular order parameter values (e.g. the torsion angle with the smallest S_i has rank $r_i = 1$, the torsion angle with the largest S_i has rank $r_i = s$). C denotes the number of core atoms. For all examined cases plots of

Q_i as a function of the order parameter rank i show a clear maximum and the absence of distant, comparably high local maxima (fig. A.1).

We found that using only C^αs is sufficient for reliable domain identification. Additional core atoms merely slowed down the calculations. We also attempted to simply use all C^α atoms in the following clustering. This approach, however, was less reliable in domain identification than the present one based on angular order parameters.

Distance variance matrix

The variance V_{ij} of the intra-conformer distance between any two core atoms i and j is calculated over all N members of the structure bundle,

$$V_{ij} = \frac{1}{N}\sum_{k=1}^{N}\left(d_{ijk} - \bar{d}_{ij}\right)^2 \quad \text{with} \quad \bar{d}_{ij} = \frac{1}{N}\sum_{k=1}^{N} d_{ijk} \tag{II.1.3}$$

where d_{ijk} denotes the distance between atoms i and j in conformer k.

Core atom clustering

To determine the residues that belong to the same domain the core atoms are clustered using an agglomerative hierarchical clustering algorithm. At the outset (clustering stage 1) each core atom forms a cluster of its own. At each subsequent stage of clustering two clusters are merged, until at the end there remains a single cluster containing all core atoms. Hence there are as many clustering stages as there are core atoms. In each stage of clustering the two nearest neighbour clusters are identified and merged. All other clusters remain unchanged and are simply propagated to the next stage. The nearest neighbour clusters are defined by Ward's method as the two clusters that yield, after merging, the lowest intra-cluster V-value variance of all possible two-cluster combinations. The intra-cluster V-value variance is computed as the variance of the V_{ij} values for all atoms i and j in the merged cluster, or, if the merged cluster contains only two atoms, by the corresponding single V_{ij} value.

Identification of the best clustering stage

At each of the clustering stages $i = 2, \ldots, C$ the average cluster spread

$$A_i = \frac{1}{c_i}\sum_j \mathrm{RMSD}_j \tag{II.1.4}$$

is calculated, where the sum runs over all clusters with more than one member, c_i is the total number of core atoms from all clusters with more than one member, and RMSD_j is the average over all structures of the RMSD to the mean coordinates for the backbone atoms N, C^α, and C' of the residues given by the core atoms in cluster j. All RMSD values are calculated using singular value decomposition (**?**). Low cluster spreads indicate high intra-cluster homogeneity, i.e. atom pairs with a similar degree of inter-atom distance variation are likely to belong to the same structural unit within the protein. A low number of clusters points to no artificial division of domains having occurred. To determine the optimal clustering stage, the quantity

$$P_i = (C-2)\,\frac{A_i - A^{\min}}{A^{\max} - A^{\min}} + n_i \qquad \text{(II.1.5)}$$

is calculated for the clustering stages $i = 2, \ldots, C$. $A^{\min}$ and $A^{\max}$ are the minimum and maximum average cluster spread values, respectively, and n_i is the number of clusters at stage i, including single-element clusters. The clustering stage with the lowest P value is chosen as the optimal clustering stage, i^*, provided that the average number of core atoms in a cluster at stage i^* exceeds one eighth of the total number of core atoms, rounded up to the nearest integer. In the calculation of this average cluster size only those clusters are considered that contain at least the minimal number of elements (μ; see below) required for a cluster to be considered a domain. If the average cluster size is too small, a new minimum value of P is determined in the restricted range $i = i^* + 1, \ldots, C$. The procedure is repeated until the average number of cluster elements at the clustering stage with minimum P exceeds one eighth of the total number of core atoms. The minimum of P_i as a function of the clustering stage i is sharply defined, usually at or near the last clustering stage (see fig. A.2 of the appendix). Each cluster at the optimum stage of clustering is considered a domain, provided that it contains at least μ elements. By default, μ is equal to 8.

II.1.1.2 Residue range refinement

The residue ranges corresponding to the identified domains are passed to the residue range determination procedure. First, however, these ranges are extended at all boundaries by m residues, so as not to restrict the range determination procedure to perhaps too narrow a starting range. By default, m is equal to 3.

The determination of the residue range for each domain starts with the residues of a previously identified and subsequently extended domain. The algorithm proceeds by iteratively removing residues until the set of residues does not change in an iteration, and it comprises the following seven steps:

1. *RMSD calculation:* Compute the RMSD value, r, for the backbone atoms N, C^{α}, and C' of the current set of residues. If the RMSD value lies below the user–set threshold value (default: 0.1 Å), exit and output the current set of residues.

2. *Removal of isolated residues:* If present, exclude from the set of residues those with two neighbours that do not belong to the current set of residues, and start a new iteration at step 1.

3. *Find residues with largest displacements:* Among the selected residues find the one with the largest average displacement whose removal does not open a new gap in the selected residues. Similarly, find the residue with the largest average displacement whose removal opens a new gap in the selected residues. The average displacement corresponds to the distance between an atom in a given conformer and its mean position after optimal superposition of all conformers onto the first conformer in the RMSD calculation of step 1, averaged over all conformers and over the backbone atoms N, C^α, and C' of a residue.

4. *Compute gap-weighted RMSD decrease:* For the two residues found in step 3, compute the gap-weighted decrease of the RMSD value upon removing the residue from the set, $\Delta r^{\mathrm{nogap}} = r - r^{\mathrm{nogap}}$ and $\Delta r^{\mathrm{gap}} = \gamma(r - r^{\mathrm{gap}})$, where r is the RMSD value from step 1, r^{nogap} and r^{gap} are the RMSD of the selected residues after removing one of the residues from step 3, and γ is a dimensionless parameter that penalizes the opening of new gaps (if $\gamma < 1$).

5. *Residue removal:* If the residue with the larger Δr value fulfils the two conditions $\Delta r \geq \delta^{\mathrm{abs}} \frac{n}{N}$ and $\frac{\Delta r}{r} \geq \delta^{\mathrm{rel}} \frac{n}{N}$, remove it from the set of selected residues and start a new iteration at step 1. Here n and N denote the numbers of atoms included in the RMSD calculation of step 1 from the current residue and from all selected residues, respectively, and δ^{abs} and δ^{rel} are parameters for the minimally required absolute and relative RMSD decrease, respectively.

6. *Retry residue removal:* If no residue was removed in step 5, find among *all* selected residues the residue whose removal yields the largest gap-weighted decrease of the RMSD value. If the conditions of step 5 are fulfilled for this residue, remove it from the set of selected residues and start a new iteration at step 1.

7. *Fill small gaps:* If the set of selected residues contains small gaps of less than g (by default three) residues, fill these gaps by additionally selecting the residues in the gap.

Average displacements are calculated in step 3 to limit the number of RMSD calculations to two in step 4, and to execute step 6 only rarely, as it requires an RMSD calculation for every selected residue. Unless noted otherwise, we used $\mu = 8$, $m = 3$, $\gamma = 0.4$, $\delta^{\mathrm{abs}} = 1.6$ Å, and $\delta^{\mathrm{rel}} = \delta + \frac{3.0}{M}$ with $\delta = 1.2$, where M denotes the current number of selected residues, and $g = 3$. Smaller values of g lead to fewer gaps. The choice of δ^{rel} was motivated by the observation of the relative RMSD decrease values for randomly disordered structures. The increase of δ^{rel} for small numbers of selected residues ensures the termination of the algorithm.

II.1.2 Output

For each of the domains it identified the CYRANGE method states the residue range(s) for superposition, the number of residues therein, and the average RMSD value to the mean coordinates for the backbone atoms in the residue range(s) for superposition. If the input consists of only two structures, the RMSD to the mean coordinates is equal to half the RMSD between the two structures.

II.1.3 Usage

The program is invoked by calling `cyrange [PARAMETERS] <PDB file>`. The available parameters are listed in tab. II.1.1.

Table II.1.1: CYRANGE parameters

Flag	Argument	Symbol	Description	Default value
-a	real number	δ^{abs}	measure of RMSD decrease required for residue removal	1.6
-b	integer	m	number of residues for padding at domain boundaries	3
-d	real number	δ	measure of RMSD decrease required for residue removal	1.2
-g	real number	γ	penalty factor for gap formation	0.4
-gw	integer	g	minimum size of gaps to be retained in the output	3
-h	–	–	display help text and exit	–
-l	file path	–	path to customized library file (CYANA format)	built-in library
-m	integer	μ	minimum size for clusters to be considered domains	8
-r	range	–	range of residues to be considered	all residues
-u	real number	–	RMSD threshold	0.1

II.1.3.1 Examples

```
cyrange -g 0.2 -b 5 -r 20-70,90-130 -l ../local.lib my_struct.pdb
```

This will compute the optimum residue range(s) for `my_struct.pdb`, using only residues 20 to 70 and 90 to 130. The gap parameter γ has a value of 0.2 (instead of its default value of 0.4), and five instead of three residues will be added at each domain boundary before the optimum residue ranges of this domain are determined. CYRANGE will employ the user's customised library file `../local.lib` instead of the standard library.

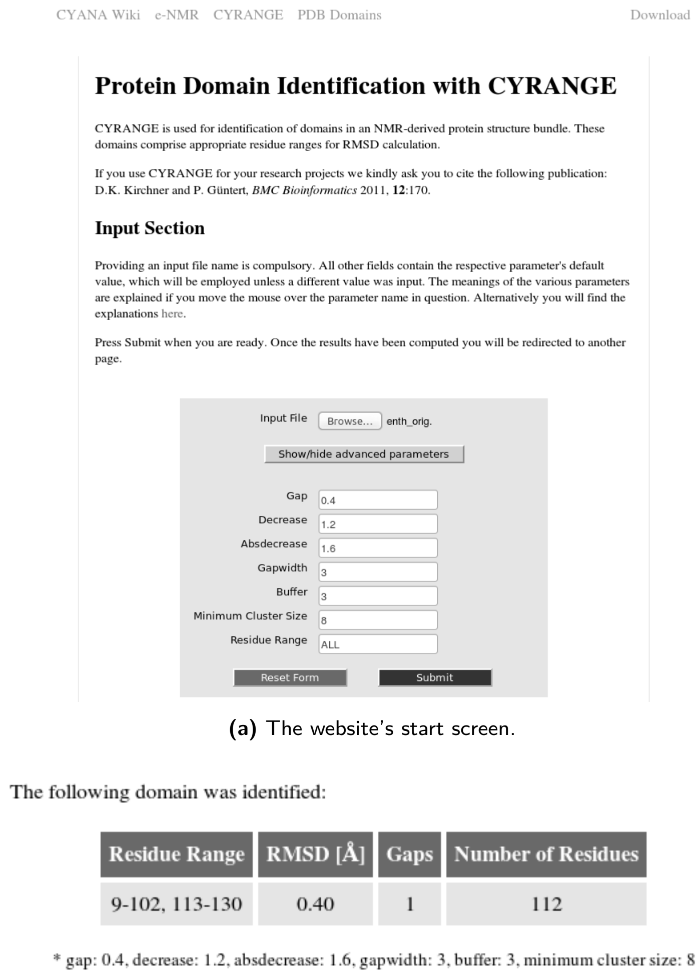
```
cyrange -r A10-A90 my_struct2.pdb
```

If the query protein consists of several chains the user must specify the chain IDs of the residues when stating a range.

II.1.4 The CYRANGE website

At `http://www.bpc.uni-frankfurt.de/cyrange.html` CYRANGE has been made publicly available. It can be both used online and downloaded as a stand-alone software tool for installation on a local machine. Furthermore, the website lists the optimal residue ranges for a large number of proteins from the PDB, and it includes direct links to each of the PDB entries.

When accessing the CYRANGE functionality through the website the user may set the values of most of the parameters listed in tab. II.1.1, but each parameter's default value is employed if the user has not made any changes. The user interface of the website as well as the online presentation of the results is shown in fig. II.1.2.



Protein Domain Identification with CYRANGE

CYRANGE is used for identification of domains in an NMR-derived protein structure bundle. These domains comprise appropriate residue ranges for RMSD calculation.

If you use CYRANGE for your research projects we kindly ask you to cite the following publication: D.K. Kirchner and P. Güntert, *BMC Bioinformatics* 2011, **12**:170.

Input Section

Providing an input file name is compulsory. All other fields contain the respective parameter's default value, which will be employed unless a different value was input. The meanings of the various parameters are explained if you move the mouse over the parameter name in question. Alternatively you will find the explanations here.

Press Submit when you are ready. Once the results have been computed you will be redirected to another page.

Input File: enth_orig.

Gap: 0.4
Decrease: 1.2
Absdecrease: 1.6
Gapwidth: 3
Buffer: 3
Minimum Cluster Size: 8
Residue Range: ALL

(a) The website's start screen.

The following domain was identified:

Residue Range	RMSD [Å]	Gaps	Number of Residues
9-102, 113-130	0.40	1	112

* gap: 0.4, decrease: 1.2, absdecrease: 1.6, gapwidth: 3, buffer: 3, minimum cluster size: 8

(b) Presentation of the results.

Figure II.1.2: Screenshot of the CYRANGE website at `http://www.bpc.uni-frankfurt.de/cyrange.html`. Here, the user can upload a PDB file containing a structure bundle, and the ordered domains will be displayed once the server-side copy of CYRANGE has completed the calculation. In the top right-hand corner there is a link leading to the web page where a stand-alone version of CYRANGE may be downloaded free of charge.

II.1.5 Evaluation of the CYRANGE method

II.1.5.1 NMR protein structures

The performance of CYRANGE was assessed on the basis of the NMR structure bundles of eleven proteins whose NMR solution structures had been determined earlier. These proteins are referred to by four-letter codes: copz (?), PDB 1CPZ; cprp (?), PDB 1U3M; enth (??), PDB 1VDY; fsh2 (??), PDB 1WQU; fspo (?), PDB 1VEX; pbpa (?), PDB 1GM0; rhod (??), PDB 1VEE; scam (?), PDB 1X02; smbp (?), PDB 2D21; wmkt (?), PDB 1WKT; ww2d (?), PDB 2DWV. The proteins copz, cprp, enth, fsh2, pbpa, rhod, and wmkt are proteins with a well-defined single-domain structure. The protein fspo has an unusual, less well-defined fold without regular secondary structure. The proteins scam and smbp are proteins with two domains connected by a flexible linker. The protein ww2d forms a symmetric dimer. Two structure bundles were considered for each of these proteins: the final structure bundle, and the structure bundle obtained in the initial cycle 1 of automated NOE assignment and structure calculation (?) with CYANA (??), i.e. all structures were recalculated using the experimental chemical shift lists, NOESY peak lists, and possible additional torsion angle or hydrogen bond restraints. This enabled comparisons of the CYRANGE ranges for two structure bundles of different precision and quality for each of the proteins. CYRANGE was also applied to a set of 26 NMR protein structures, 23 of which had been used earlier for evaluating the FindCore algorithm (?). Additionally, the protein 2kr6 was included as an example of a protein with a large domain and a flexibly connected small helix that constitutes a separate domain (?) and 2ktf and 2l14 were included as examples of protein-protein complexes. These 26 proteins are labelled by their PDB codes.

II.1.5.2 Application to all NMR structures in the PDB

The entire set of NMR structures from the PDB (?) as available on July 30, 2010 was subjected to domain and residue range determination with the CYRANGE method, provided that the files contained at least five conformers and 15 amino acid residues.

II.1.5.3 Use of FindCore and PSVS

For comparison, PSVS (?) and FindCore (?) were also employed to identify residue ranges and domains in our protein test set. The programs were used through the web portals `http://psvs-1_4-dev.nesg.org` (PSVS) and `http://fps.nesg.org` (FindCore). With PSVS the default option ‘ordered residues’ was selected for the residue selection for analysis. The residues reported as ‘ordered’ were taken as the residue ranges identified by PSVS. Note that with these options PSVS does not identify multiple domains. With FindCore, the ‘average structure’ was selected as the reference structure, the analysis was based on ‘standard amino acids’, and only backbone atoms were used in domain identification. The calculations were also performed using all instead of only backbone atoms, with largely equivalent results (data not shown). FindCore reports the number of domains it identified, yet it does not unambiguously state the boundaries of the domains. Instead, the program provides a list of 'core residues'. When the program reported more than one domain, we manually attributed the core residues to the individual domains.

Analyses of the results

PSVS and FindCore results were downloaded from the internet, and the output residue ranges were extracted from the source code of the downloaded web pages in an automated fashion. All reported RMSD values were calculated with CYANA for the backbone atoms N, C^{α} and, C' in the reported residue ranges, and with respect to the mean coordinates. With FindCore the RMSD value of each identified domain was calculated separately; for PSVS all reported 'ordered residues' were used in the RMSD calculations, as no domain information is provided by PSVS. The program MOLMOL (**?**) was used to visualize structures.

II.1.5.4 Determination of the average GDT_TS value of each structure bundle

The web server on `http://proteinmodel.org/AS2TS/LGA/lga.html` (using the parameter set `-3 -o0 -d:4.0`) was accessed to obtain the *GDT_TS* values whose averages are shown in fig. III.1.4. From each structure bundle all possible conformer pairs consisting of the first conformer and each subsequent conformer were subjected to the calculation. The results were extracted from the web site in an automated fashion, and for each structure bundle the average of the individual *GDT_TS* values was computed.

Chapter II.2

Structure validation

This chapter presents two protein structure validation tools that have been developed for this thesis. The main focus will lie on CYVAL, a fully automated procedure integrated into CYANA (??). CYVAL, in contrast to the vast majority of structure validation tools for NMR, gauges the degree of accuracy of the query structure by assessing the correspondence between this structure and the NOESY spectra that were used for structure determination, instead of evaluating the geometric quality of the structural model. The other tool described in this chapter is a bundle of third party knowledge-based validation programs that can now effortlessly be called from within CYANA, and whose output is automatically processed, formatted, and summarised for the user's convenience.

II.2.1 CYVAL

A schematic representation of the validation method used by CYVAL is depicted in fig. II.2.1. The user must provide a PDB file of the structural model to be validated, a list of assigned resonances, and a processed NOESY spectrum. The latter must be in UCSF, Bruker, or XEASY format. The reading routines for the first two file formats were implemented as part of the validation project.

As explained in detail in the following section, CYVAL generates a set of peaks that can be expected based on the query structure and the provided chemical shifts, and it performs automated peak picking on the NOESY spectrum to produce a list of experimental signals. These two groups of peaks are compared to yield three different peak classes: matching, missing, and unpredicted peaks. Inter-atomic distances are measured within the structure and also derived from experimental peak intensities via distance calibration, and the peaks within the three categories are assigned individual weights and distance deviation contributions, from which ultimately the structure's total validation score, ζ, is computed.

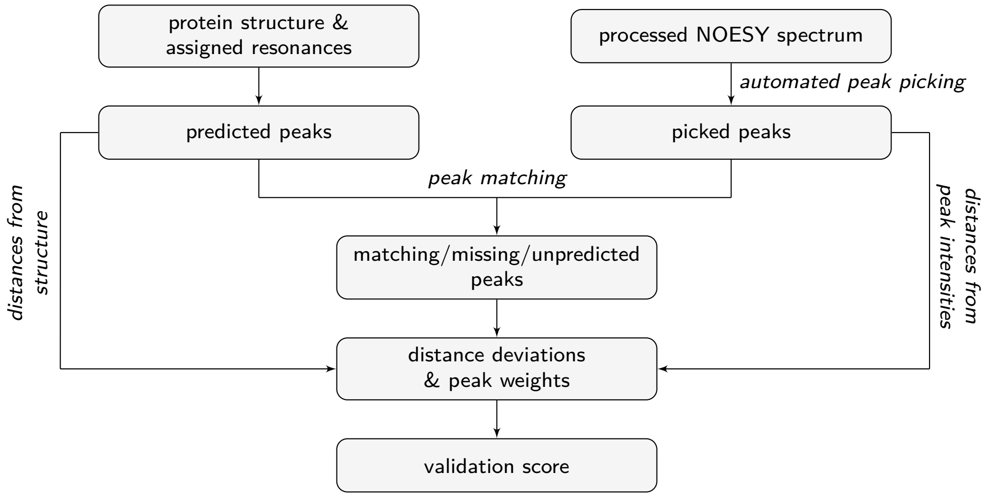

protein structure & assigned resonances
processed NOESY spectrum
automated peak picking
predicted peaks
picked peaks
peak matching
distances from structure
distances from peak intensities
matching/missing/unpredicted peaks
distance deviations & peak weights
validation score

Figure II.2.1: Graphical overview of the validation approach of CYVAL. The essential input data that the user must provide are the query structure, a list of assigned chemical shifts, and a NOESY spectrum. Once the required files are read all information is processed automatically, so there is no need for any kind of user intervention.

II.2.1.1 The validation method in detail

Peak processing and distance calibration

As indicated in fig. II.2.1, CYVAL predicts a set of NOESY peaks from the query structure. To this end the spectrum type (e.g. ^{15}N-NOESY), the maximum inter-proton distance for which an NOE signal can still be expected (default: 4.5 Å), and the assigned chemical shifts are given as input to the `genpik` procedure (developed by Dr. Elena Schmidt). This routine creates a list of all peaks that can be expected based on the input structure bundle and the given spectrum type. A peak will only be predicted in case its corresponding inter-proton distance does not exceed the abovementioned distance threshold within each of the conformers that make up the ensemble. In case the experimental NOESY spectrum contains a solvent signal range whose location the user specified upon invocation of the validation procedure, all peaks that are predicted to fall into this spectral region, as well as peaks with missing chemical shift information in at least one dimension, are subsequently removed from the set of predicted peaks.

In this study the experimental signals from the processed NOESY spectrum were automatically picked using CYANA's internal peak picking routine written by Julia Würz; the value of the noise baseline, which is required for peak picking, must be supplied by the user. The peak picker was invoked using the parameters listed in tab. II.2.1.

Once both peak lists are complete the picked signals are compared to the expected peaks in order to determine which peaks are present in both sets; which are missing from the experimental spectrum (i.e. should be there based on the three-dimensional structure but were not discovered); and which experimental peaks are present but cannot be explained by the structure. The latter group may either be noise peaks or artefacts, but they, alongside the missing peaks, may also be indicators of errors within the structural model.

Table II.2.1: Parameters used when calling the automated peak picking routine (CYANA command `peaks pick`).

Parameter name	Value
`action`	`local_max`
`method`	`diag`
`specfile`	current spectrum
`fmt`	current spectrum's format
`dim`*i*`range`	bounds of spectrum axis i
`sig_cut`	current spectrum's noise level
`only_pos`	-

The CYANA subroutine `pmatch` from the `Peakmatch` tool (?) is used for the purpose of peak list comparison. The advantage of this procedure is the robustness of `Peakmatch`, whose performance is not affected even by a large addition of artefact peaks, nor by a large amount of missing signals, nor by random chemical shift changes of up to twice the chemical shift tolerance for the respective axis (?). Consequently, it is ideally suited for comparing the positions of the automatically picked to those of the predicted peaks.

During matching, the list of experimental peaks is used as the reference peak list. As a result, no, one, or several expected peaks may be mapped onto the same picked peak. If a given experimental peak's number of matches from the predicted list exceeds one, the experimental peak in question is assumed to be the sum of some or all of the predicted peaks that were mapped onto it. I.e. it is estimated that a number of signals have overlapped to form this one observed peak. Picked peaks that were not predicted based on the structure cannot be matched to any of the predicted peaks, and predicted peaks that do not occur in any match are registered as missing peaks because the structure suggests they ought to be present in the experimental spectrum, but they are not.

The mapping of predicted to picked peaks also means that picked signals, if a match from the expected set is found, are assigned to a certain proton pair within the structure. This is a fundamental requirement for the distance comparison that will be conducted later on (see below).

In order to be able to factor in the reliability of a given signal into the final validation score CYVAL calculates a weighting factor for each peak. One component of this factor is the local *peak density* (ρ) of the signal. Within the context of this work the term peak density refers to the number of experimental signals within a given area (for two-dimensional spectra) or volume around the peak in question. ρ is determined for all peaks, both predicted and picked, as the environment of each peak is slightly different, and so is the peak density at its location.

The space or volume used for the computation of ρ is an ellipse or ellipsoid each of whose semi-axes has a length of 10γ, with γ being the half-width at half-maximum height (in ppm) of the peak in this dimension. At a distance of 10γ from the peak maximum on one of the principal axes of the signal the intensity of a peak of a Lorentzian shape has dropped to about 1/100 of its maximum value. This was deemed small enough to be assumed not to greatly distort the intensities of other signals. The elliptical shape was selected as it resembles the outline and hence also the sphere of influence of a NOESY signal. Moreover, determining whether another peak lies within the thus defined neighbourhood of a given peak is straightforward using this shape: all N picked peaks whose maxima meet the condition specified by eq. II.2.1 are registered as neighbours of the peak of interest.

$$\sum_{i=1}^{D} \left(\frac{c_i}{10\gamma_i} \right)^2 \leq 1 \tag{II.2.1}$$

D is the number of dimensions of the spectrum, and c_i is the coordinate difference (in ppm) of the potential neighbour from the peak of interest in dimension i.

ρ, the peak density, is defined by eq. II.2.2. All picked peaks have a minimum ρ value of 1, as each of them is counted towards the number of experimental signals in its own environment. For a predicted peak, however, ρ may also be equal to 0 if no picked peaks lie close enough to its coordinates to be considered its neighbours.

$$\rho = \frac{1}{N} \tag{II.2.2}$$

The true intensity of a peak generally differs from its observed intensity as the residual intensities of neighbouring signals also contribute to the height of the local maximum. Hence, in order to learn about the true inter-atomic distance corresponding to a particular peak its correct intensity must first be determined. CYVAL offers the user the option to perform such a correction of peak heights.

Both the shapes of the signals and their linewidths γ are required for the calculation since these factors determine the residual intensity of a peak at the location of a neighbouring signal. Two peak lineshapes are available, Lorentzian (eq. II.2.3) and Gaussian (eq. II.2.4). If the user does not specify a lineshape CYVAL performs the intensity correction with each in turn and uses the result that yields the best overall fit with the intensity profile of the spectrum.

$$f(x) = \frac{\gamma^2}{(x - x_0)^2 + \gamma^2} \tag{II.2.3}$$

$$f(x) = \mathrm{e}^{-\left(\frac{x-x_0}{2\sigma}\right)^2} \text{ with } \sigma = \frac{\gamma}{\sqrt{2 \ln 2}} \tag{II.2.4}$$

In each dimension, x is the coordinate of the current point and x_0 is the location of the maximum of the peak of interest. Each of the curves described by the above equations is equal to 1 at x_0, hence the function value represents the relative intensity of a signal at a given point in one dimension.

Determination of the true peak heights is accomplished by solving the system of linear equations shown in eq. II.2.5 using singular value decomposition. We assume that peak i contributes to the total intensity $I_{\mathrm{tot},j}$ measured at the location of peak j, but the amount of its contribution depends on both the location and the true intensity I_i of i. Consequently, each measured peak height is the sum of all true peak intensities, multiplied by the residual relative intensity of each peak at this location $(r_{i,j})$. The residual relative intensities are the product of the function values of eq. II.2.3 or eq. II.2.4 for all spectral dimensions.

$$\begin{pmatrix} r_{1,1} & r_{2,1} & \dots & r_{n,1} \\ r_{1,2} & r_{2,2} & \dots & r_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,n} & r_{2,n} & \dots & r_{n,n} \end{pmatrix} \times \begin{pmatrix} I_1 \\ I_2 \\ \vdots \\ I_n \end{pmatrix} = \begin{pmatrix} I_{\mathrm{tot},1} \\ I_{\mathrm{tot},2} \\ \vdots \\ I_{\mathrm{tot},n} \end{pmatrix} \tag{II.2.5}$$

In practice, considering all other signals when calculating the corrected intensity of each peak is both very expensive and unnecessary, as the contribution of peaks that lie far away is negligible. For this reason only the neighbours of a signal, identified as described above, are taken into account. Naturally, this approach means that the corrected heights of most peaks are calculated several times, but only the value obtained when the peak lies at the centre of the examined neighbourhood is stored as the corrected intensity of this signal.

At times, significant peak overlap poses problems to the intensity correction approach that have yet to be solved. In the case of signals that have one coordinate in common and whose intensities are very different the height of the smaller peak will be considerably overestimated. The intensity of the stronger peak is affected as well, but, naturally, far less so. If the peaks do not share any coordinate but their coordinates are very similar, the result will also deviate from the true heights; however, the effect is markedly weaker than in the previous case.

CYVAL assesses the correspondence between structure and spectrum on the basis of interproton distances, which are directly available from the structural model but which also need to be calibrated from the raw or corrected experimental peak intensities. As shown in eq. II.2.6, the calibration constant C connects a reference distance d_{med} to a reference intensity I_{med}, which is is the median intensity of the group of peaks that were both encountered within the spectrum and predicted based on the structural model.

$$C = I_{\mathrm{med}} \times \bar{d}^{6}_{\mathrm{med}} \tag{II.2.6}$$

In the present context $\bar{d}_{\mathrm{med}}$ is the median of the average structure-based distances $\bar{d}$ of those atom pairs whose peaks were both predicted based on the structure and encountered in the spectrum. During the calculation of the average distances $\bar{d}$ each picked peak to which exactly one predicted peak was matched receives the mean distance of the atom pair associated with this predicted peak; if two or more predicted peaks were mapped onto the same picked peak eq. II.2.7 is used,

$$d^{\text{ave}} = \frac{1}{N} \sum_{h} \left(\sum_{i \in A} d_{h,i}^{-6} \right)^{-\frac{1}{6}} \tag{II.2.7}$$

where N is the number of conformers that make up the ensemble, h is an individual conformer, A is the set of predicted peaks matching the picked peak in question, and $d_{h,i}$ is the inter-proton distance associated with peak i in conformer h. Hence the structure-based distance assigned to an experimental peak with two or more matching predicted peaks is the mean value of the r^{-6}-summed distances of all atom pairs belonging to the predicted peaks.

The corresponding distance d_i of experimental peak i is calibrated using the peak intensity I_i according to eq. II.2.8.

$$d_i = \left| \frac{C}{I_i} \right|^{\frac{1}{6}} \tag{II.2.8}$$

Calculation of ζ

Not all signals, neither predicted nor experimental, are equally reliable. Peaks picked within a dense neighbourhood may be affected or distorted by strong peak overlap, very weak long-range peaks may in fact be noise peaks or artefacts, and peaks predicted from protons of a comparatively disordered part of the protein may not be present in the experimental spectrum due to protein flexibility. Furthermore, some signals contribute more valuable information to the validation process than others. Certain short-range peaks, for example, may be present for any conformation as their underlying interactions are based on the amino acid sequence alone, hence their presence in both the sets of predicted and experimental peaks does not necessarily indicate that the structural model is correct. Long-range peaks, on the other hand, are potentially highly valuable for the validation procedure. Experimental long-range peaks are unlikely to be compatible with an incorrect overall fold. Thus they will probably remain unassigned if there is a mismatch between the structural model and the experimental data. Long-range peaks predicted based on an incorrect structure are likely to be classified as missing after the matching step.

Another reason for why certain experimental peaks may remain unassigned is that they could not be predicted from the structure due to locally poorly defined regions, which manifest themselves in low dihedral angle order parameters. Fig. II.2.2 shows the two main reasons for locally low order parameter values, a large degree of protein flexibility leading to a lack of NOESY data for a portion of the structure, or experimental peaks having been discarded during the structure determination process. As the two cases are not easy to tell apart CYVAL does not differentiate between them, even though the second of the abovementioned reasons points to parts of the structural model being potentially problematic.

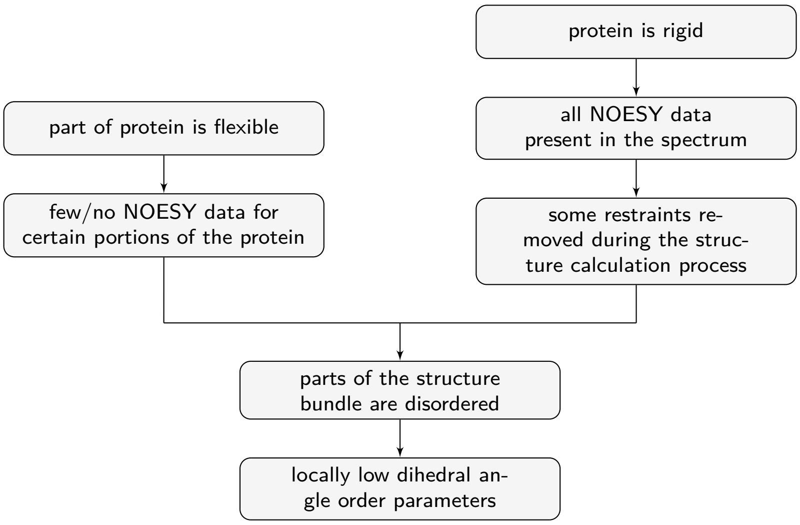

protein is rigid
part of protein is flexible
all NOESY data present in the spectrum
few/no NOESY data for certain portions of the protein
some restraints removed during the structure calculation process
parts of the structure bundle are disordered
locally low dihedral angle order parameters

Figure II.2.2: The two main factors for locally low dihedral angle order parameter values in a protein structure bundle.

To allow for variations in peak reliability and usefulness each peak is attributed a weight. This value takes factors such as local order into account. However, due to different pieces of information being available for different peaks, the traditional weight calculation mode of CYVAL, `trad`, differentiates between unpredicted peaks on the one hand and matching and missing peaks on the other hand. For peaks of the latter two categories the identity of the interacting protons is known since there is a direct link to the structural model. Consequently, the structure-based interproton distance is available as well as angular order parameter data. Furthermore, the information content (see below) of the interaction can be calculated. Unpredicted peaks, however, have no known connection to the protein structure, hence no structure-based data can be accessed. Local peak density data (eq. II.2.2) are available for peaks from any class.

The order parameter list of the structure bundle is set up as described in section II.1.1.1, and all dihedral angles other than ω, which is rigid, contribute to the order parameter determination using eq. II.1.1. Each residue's sum of order parameter values is subsequently divided by the number of its flexible dihedral angles to yield the mean order parameter value of this residue, $\overline{S}_{\text{res}}$, which is given by eq. II.2.9.

$$\overline{S}_{\text{res}} = \frac{1}{N} \sum_{k=1}^{N} S_k \tag{II.2.9}$$

Here, N is total number of order parameters S of the residue, and k represents a flexible dihedral angle.

For each peak i from the missing and matching categories CYVAL calculates a weighting factor w_i according to eq. II.2.10.

$$w_i = \frac{1}{2}\left(\overline{S}_{\mathrm{res}_1} + \overline{S}_{\mathrm{res}_2}\right)\frac{\iota_i}{\max\left(1, \rho_i\right)} \tag{II.2.10}$$

w_i takes into account the $\overline{S}_{\mathrm{res}}$-values of the residues to which the two interacting protons belong—if one or both of the residues are located within ill-defined regions of the structure bundle the order parameters, and hence w_i, will be low. Additionally, the weighting factor includes the information content ι_i (**?**, unpublished; eq. II.2.11) of the peak. ι increases with the degree of structural information conferred by this particular signal.

$$\iota_i = -\frac{\log P\left(r_i|0\right)}{R_i} \tag{II.2.11}$$

$P\left(r_i|0\right)$ is the probability that restraint r_i is fulfilled by a random structure, and R_i is the redundancy of restraint r_i. R_i is a measure of the similarity between the way in which two restraints constrain the structure, thus R_i takes into account the similarity between r_i and all other restraints. The redundancy value for the comparison of r_i with itself is equal to 1 by definition; the comparison of two completely dissimilar restraints yields 0. As R_i is the sum of all individual redundancy values its value will be greater than or equal to 1.

In case the `pmatch` routine mapped several predicted peaks to the same experimental signal the predicted peak with the largest match score is drawn on for its information content ι and its mean order parameters $\overline{S}_{\mathrm{res}}$. Hence, as far as the weighting factor is concerned, an experimental peak that is assumed to be the collection of a number of predicted peaks is treated as the realization of one single predicted peak. Furthermore, the value of the local peak density is not influenced by the matching of several predicted peaks to one picked peak since only experimental signals are taken into account during the calculation of ρ_i.

If a picked peak was not predicted neither an information content value nor average order parameters can be calculated as there is no information regarding what atom pair gave rise to this peak. Consequently, eq. II.2.10 cannot be used to calculate the weight of this signal. In such a case eq. II.2.12 is used instead, where d_i^{calib} is the distance (in Å) calibrated from the experimental peak intensity using eq. II.2.8. The units of the weighting factors are neglected.

$$w_i = \frac{1}{\rho_i d_i^{\mathrm{calib}}} \tag{II.2.12}$$

The weighting factors are employed in the calculation of the final validation score, ζ. Deviations of the experimentally determined inter-proton distances from their structure-based counterparts are the other crucial information contained in ζ. CYVAL compares distances instead of signal intensities since the former are related to the latter via a d^{-6} dependency, which means that relatively large intensity deviations will result in comparatively small differences in distance. It follows that there is a considerably lower risk of overestimating structural errors if the correspondence between structural and experimental data is assessed on the distance level, especially so since accurately estimating NOESY intensities from a three-dimensional structure is both difficult and time-consuming (**??**)[1]. Moreover, performing the comparison on the intensity level is likely to lead to a bias of the comparison towards short-range peaks, as their intensities are particularly high, even though it is the long-range peaks that are more important for the determination and assessment of the overall fold.

The final validation score ζ is calculated according to eq. II.2.13, where the sum of the individual peak contributions F is normalized by the sum of the peak weights w. From the definition of F, which is given below, it follows that a lower value of ζ indicates a greater degree of correspondence between the experimental data and the structural model. N is the number of used peaks.

$$\zeta = \frac{\sum_{i=1}^{N} F_i}{\sum_{i=1}^{N} w_i} \tag{II.2.13}$$

Naturally, different proteins contain different numbers of protons, and thus they will give rise to different numbers of NOESY peaks. The protein fold, too, determines what peaks are to be expected, and the actual quality of the recorded spectra is another factor influencing how many signals will contribute to the final value of ζ. For this reason it was tested whether dividing ζ by the number of peaks used in its calculation would lead to a unified scale that could be employed to assess the accuracy of any query protein. The per peak-version of ζ will henceforth be referred to as ζ_{n} (eq. II.2.14).

$$\zeta_{\mathrm{n}} = \frac{\zeta}{N} \tag{II.2.14}$$

Irrespective of whether a peak i belongs to the class of matching, missing, or unpredicted signals, its contribution F_i to the total value of ζ is given by eq. II.2.15, where w_i is the weighting factor of the peak. δ_i takes into account the distance deviation between structural and experimental data.

$$F_i = w_i \delta_i \tag{II.2.15}$$

δ_i depends on the peak class and it is the result of a distance deviation penalty function, termed `sat1` (fig. II.2.3), which is described by eq. II.2.16. A sigmoid function was selected in order not to put too strong an emphasis on small distance deviations, hence `sat1` was designed to

[1] According to **?** "cross-peak overlaps, effects of spin diffusion, internal and intermolecular dynamics, and differential heteronuclear polarization transfer efficiencies create difficulties in making accurate estimates of NOESY cross-peak intensities from 3D structures, even when using relaxation matrix calculations". Thus, there may be pronounced errors when cross-peak and back-calculated cross-peak intensities are compared for validation purposes.

be rather flat in the range between 0.0 and 0.5 Å. Neither is there a need to strongly distinguish between large and very large distance deviations, since both indicate a clash between the structure and the experimental data. Consequently, the function ought to reach a plateau. Furthermore, no function value should exceed 1.0 in order to ensure that ζ will lie between 0.0 and 1.0, otherwise interpreting its value would not be straightforward. Therefore, the saturation parameter σ was set to 1.

$$f(|\Delta d|) = \frac{\sigma}{1 + \mathrm{e}^{-m\sigma|\Delta d|} \times \left(\frac{\sigma}{s} - 1\right)} \tag{II.2.16}$$

The slope of the curve depends on m, which was set to 3.5 so that the steepest portion of `sat1` would fall into the range of medium distance deviations Δd from around 0.6 to 1.8 Å, where half of all distance deviations were observed. Δd is connected with the class of the peak and will be defined for the matching, missing, and unpredicted category in turn. The starting point s was set to a low value of 0.01 since small distance deviations ought not to be strongly penalized.

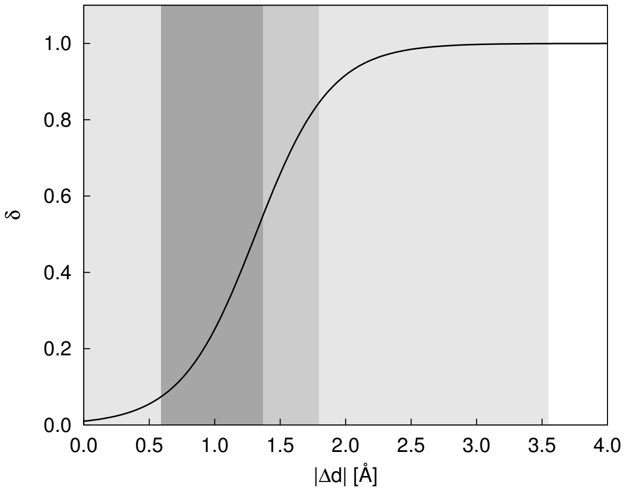


Figure II.2.3: Graph of the distance deviation (Δd) penalty function `sat1` whose function values δ contribute to the final validation score ζ. Each grey area marks 25 % of all Δd values observed during testing of CYVAL, which demonstrates that 75 % of all distance deviations were no greater than 1.8 Å, and that the steepest segment of the function was used on about 50 % of all distance deviations.

The penalty value δ_i^{match} for matching peaks is calculated using eq. II.2.17.

$$\begin{aligned} \Delta d^{\text{match}} &= d_i^{\text{ave}} - d_i^{\text{calib}} \\ \delta_i^{\text{match}} &= \begin{cases} f\left(|\Delta d_i^{\text{match}}|\right), & \text{if } |\Delta d_i^{\text{match}}| > \frac{d_i^{\text{calib}}}{10} \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{II.2.17}$$

d_i^{ave} is the average structure-based distance, and d_i^{calib} is the distance calibrated from the experimental peak intensity. The experimental signal and its best match from amongst the predicted peaks are now treated as one and the same signal, i.e. they do not yield two but only one contribution to ζ. In case two or more predicted peaks were mapped onto the same picked peak, the *secondary matches*, i.e. all matches other than the best one, contribute to the final value of ζ only by entering into d_i^{ave} according to eq. II.2.7. To take into account smalls errors in distance calibration we have introduced a threshold below which any distance deviation will be set to zero. This threshold is equal to 10 % of the experimental distance. Any deviation whose absolute value is below or equal to the threshold value is not taken into consideration. The experimental instead of the structure-based distance is used as the basis of threshold determination since the spectrum provides the actual restraints to which the structure should fit, thus it should be measured to which extent these experimental restraints are violated by the resulting protein structure.

No experimental distance information is available for peaks of the missing category, hence there is no value of d_i^{calib} to which the structure-based distance could be compared. In the calculation of Δd^{miss} the maximum inter-proton distance (in Å) for which an NOE can still be observed in the spectrum, d^{max}, is used instead because the probability of detecting a signal decreases with increasing inter-atomic distance, and long-range peaks in particular may be hidden within the noise. The empirical parameter of 0.1 Å is added to d^{max} for the purpose of the comparison since all peaks up to and including d^{max} are are assumed to be present within the spectrum, so a missing peak of a distance corresponding to d^{max} should not yield a distance deviation of zero. As shown in eq. II.2.18, Δd^{miss} is then used in the calculation of the penalty value δ_i^{miss}.

$$\begin{aligned} \Delta d^{\text{miss}} &= d^{\text{max}} + 0.1 - d_i^{\text{ave}} \\ \delta_i^{\text{miss}} &= f\left(|\Delta d_i^{\text{miss}}|\right) \end{aligned} \tag{II.2.18}$$

In the case of unpredicted peaks d_i^{ave} does not exist as there is no known link between the observed peak and an inter-atomic distance. For this reason Δd^{unpred} contains a comparison of the calibrated distance d_i^{calib} with d^{max}: signals of low intensity, i.e. signals that translate into a distance similar to d^{max}, are more likely to be artefacts or noise peaks. Therefore they ought not be punished as severely as unpredicted true peaks of high intensity, which indicate short-range interactions that are not backed by the structural model. The penalty values of unpredicted peaks are computed using eq. II.2.19.

$$\begin{aligned} \Delta d^{\text{unpred}} &= d_i^{\text{calib}} - (d^{\text{max}} + 0.1) \\ \delta_i^{\text{unpred}} &= \begin{cases} f\left(|\Delta d_i^{\text{unpred}}|\right), & \text{if } \Delta d_i^{\text{unpred}} < 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{II.2.19}$$

δ_i^{unpred} is set to zero if $\Delta d_i^{\text{unpred}} > 0$, as this indicates that the calibrated distance is too large to give rise to a peak stemming from the protein. For this reason the peak in question is assumed to be mere noise, and it is discarded so that it will not distort ζ.

II.2.1.2 The data structure for storing spectra and peak data required for structure validation

All peak and spectrum data required for structure validation are stored in a spectra-based data structure that contains an entry for each experimental spectrum. All other data, e.g. all peaks picked in this spectrum, all peaks predicted for this spectrum type based on the structure bundle, as well as the various peaks' score contributions, etc. are stored in conjunction with the data for each spectrum. As the structure is unified all required information is present at all times, irrespective of the file format in which the spectra themselves were provided. As multi-dimensional spectra generally contain a large number of data points and thus require considerable storage space it is possible to selectively erase the intensity data from memory once they are no longer required, whilst retaining all derived data in case further operations are to be performed that need this information.

II.2.1.3 Peak data file

After each successful validation run the peak data for the spectrum that has just been used for validation are written to a plain-text peak data file, which serves the purpose of providing the user with a detailed overview of which peaks were picked, which were predicted, and which were matched to each other.

Several example entries to illustrate the file format are shown in listing II.2.1.

Listing II.2.1: Peak data file written after a successful structure validation. The first line contains the numbers of 1. combined peaks, 2. picked peaks, 3. predicted peaks, 4. identified peaks, 5. missing peaks, 6. unpredicted peaks, 7. peaks used during the validation. In the following order, each peak entry lists: the peak number (here: 97), the peak coordinates, the raw intensity of the picked peak, the corrected intensity, the local noise, the weighting factor, the information content, the score contribution, the average distance, the calibrated distance, the goodness of fit, $|\Delta d|$, the coordinate indices within the spectrum, the indices of two interacting atoms, the local peak density, the number of peaks predicted in this spot, the penalty value δ, and the mean $\bar{S}_{\text{res}}$-value. A value of -999 occurs if the quantity does not apply to this peak. The 'neigh' line lists the peak numbers of the peak's neighbours. The 'num_match' entry indicates that two peaks were matched to this peak, namely peaks 1656 and 1657. The first peak listed in the 'match' line is the one with the highest match score. The boolean values at the end of the entry state that this peak is picked (T in position 1, F in position 2), not a secondary match (F in position 3, can only be T for predicted peaks), and used for the calculation of ζ (T in position 4).

```
 1973   1216    757    478     16    738   1232
   ...
peak     97
   1.85355   122.65202     9.08459   0.47021E+08   0.47021E+08  -0.99900E+03     0.09408
           0.22092     0.01078     2.32709     3.05579  -999.00000     0.72870   708    53
         141    740    736     2     2     0.11459     0.85173
```

```
neigh     97    121
num_match    2
match   1656   1657
 T  F  F  T
```

II.2.1.4 How to use CYVAL

Before the validation of a structural model against its experimental data can be performed the user has to provide a NOESY spectrum. For this purpose I have implemented the CYANA command `read raw`. Furthermore, CYVAL requires a list of experimental peaks, hence peak picking must be performed on the spectrum before the validation is started. Finally, the CYANA command `structure validate` initiates the actual validation of the structure bundle against the experimental NOESY spectrum. This section explains how to use `read raw` and `structure validate`.

Usage of `read raw`

The `read raw` command reads a processed NMR spectrum in UCSF, Bruker, or XEASY format. The command parameters are given in tab. II.2.2, and detailed descriptions can be found below.

Table II.2.2: Parameters of the CYANA command `read raw`.

Name	Available options	Default value
`append`	`true, false`	`false`
`format`	–	–
`headers`	`true, false`	`false`
`lshape`	`lorentz, gauss`	`lorentz`
`lwidth`	–	–
`type`	`ucsf, bruker, xeasy`	`xeasy`
`water`	–	–

append If this parameter is absent all spectrum information currently in memory will be overwritten, otherwise an additional spectrum entry will be created.

format This parameter takes the spectrum format that provides information regarding the spectrum type and the order of the dimensions, e.g. `format="N15NOESY H N HN"`.

headers If this option is used only the meta-information of the spectrum file is read and output to the screen. The intensity data are ignored.

lshape Lineshape of the peaks.

lwidth Half-width at half-height (in Hertz) of the peaks in the order of the spectral dimensions, e.g. `lwidth="33.0;42.0;64.0"`.

type File format of the processed spectrum.

water Location of the solvent signal (if present), e.g. `water="4.39..4.6,HC"`. The first part specifies the ppm-range of the solvent signal and the second part states for which axis this range is given.

Usage of `structure validate`

Tab. II.2.3 lists the parameters of the `structure validate` command, alongside their available options and default values. Detailed information regarding those parameters is provided below.

Table II.2.3: Parameters of the CYANA command `structure validate`.

Name	Available options	Default value
`corr_intensity`	`true, false`	`false`
`error_function`	`sat1, sat2, sat3, lin4, lin6, log4, pol4, pol6`	`sat1`
`peak_mode`	`all, no_unpred, no_missing, no_match, match_only, long`	`all`
`random`	`true, false`	`false`
`spectrum`	–	last read spectrum
`weight_mode`	`trad, one, no_info, no_OP, equal, density, dist`	`trad`

corr_intensity If this option is present peak intensities from the experimental spectrum are corrected for signal overlap.

error_function Different functions may be used as the distance deviation penalty function. Section II.2.1.6 lists the available functions in detail.

peak_mode This parameter determines which peaks will be used for the calculation of ζ. If `no_unpred`, `no_missing`, or `no_match` is selected, the peaks from the specified class will be discarded. `match_only` leads to only matching peaks being considered, and `long` means that only long-range peaks will be used (see section II.2.1.6 for details).

random This option must be present if the current query structure is a random structure bundle that is to be used solely for preak prediction and peak matching (details are provided in section II.2.1.6).

spectra File name of the spectrum to be used for validation. The spectrum has to have been read using `read raw`.

weight__mode If the default mode (`trad`(itional)) is used the weighting factors are calculated as stated in section II.2.1.1. `one` sets each weighting factor to 1.0, `no_info` and `no_op` mean that information content or order parameter information is not used, `density` sets each peak's weighting factor to the inverse local peak density, and `dist` sets the weighting factor to the inverse inter-atomic distance. `equal` sets all weighting factors to the product of the inverse local peak density and the inverse inter-atomic distance (see section II.2.1.6 for details).

II.2.1.5 Evaluation of the validation power of CYVAL

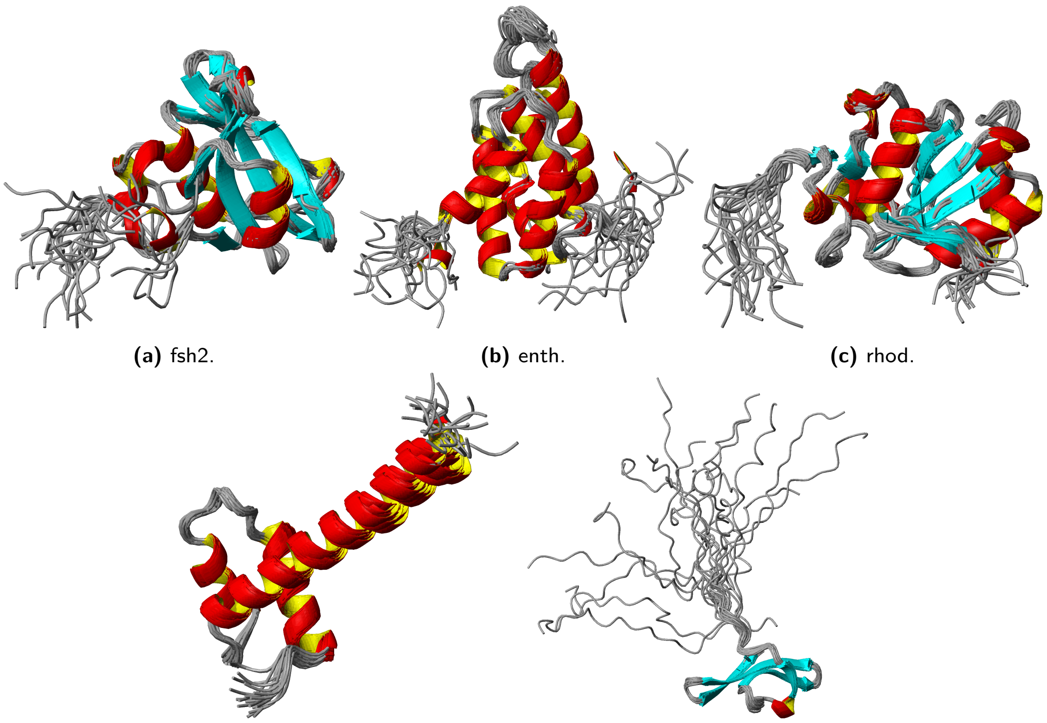
Proteins and spectra used for testing

In total, five proteins were used for evaluating the validation powers of ζ and ζ_n. Two proteins stemmed from the Critical Assessment of Automated Structure Determination of Proteins from NMR Data (CASD-NMR) project: TSTM1273 from *Salmonella typhimurium LT2* (63 residues, PDB ID 2LOJ, Northeast structural genomics (NESG) consortium target StT322) and the SANT 2 domain from human DNAJC2 (73 residues, PDB ID 2M2E, NESG target HR8254a). The three remaining proteins were the ENTH-VHS domain At3g16270 from *Arabidopsis thaliana*, hereafter referred to as enth (140 residues, PDB ID 1VDY; **??**), the rhodanese domain At4g01050 from *Arabidopsis thaliana* (rhod, 134 residues, PDB ID 1VEE; **??**), and the Src homology 2 domain from the human feline sarcoma oncogene Fes (fsh2, 114 residues, PDB ID 1WQU; **??**). For each of these five proteins resonance assignments and experimental spectra as well as refined peak lists for each spectrum were available.

The structure bundles calculated from all available data were used as reference structures. These reference ensembles are shown in fig. II.2.4. The degree of completeness of the resonance assignments and the spectra types available for testing are listed in tab. II.2.4, and tab. II.2.5 contains the noise levels of the spectra and the locations of the solvent signals, if present.

Generation of test structures

Test structures of varying degrees of accuracy were used to evaluate the power of CYVAL to differentiate between correct and incorrect protein folds. For enth, fsh2, and rhod incorrect structures were used whose creation is described elsewhere (**?**). Out of the set of structures from the just mentioned publication a smaller test set for each protein was randomly selected. During the selection process the ordered regions of each reference structure (see above) were determined by CYRANGE (**?**). Only ensembles in which the ordered ranges from the corresponding reference structure exhibited an RMSD value of less than 2.0 Å were included in the subset to be used for testing CYVAL. This filtering procedure was carried out to ensure that effects caused by generally ill-defined structures would not distort the overall evaluation results.



(a) fsh2. **(b)** enth. **(c)** rhod.

(d) HR8254a. **(e)** StT322.

Figure II.2.4: The five test proteins used for evaluation of the validation power of CYVAL. In each case the structure bundle is shown that was calculated using all available experimental data and that was thus used as the reference structure when the degrees of accuracy of newly calculated structures were determined.

Table II.2.4: The proteins used to test CYVAL.

Protein	**PDB ID**	**Number of residues**	**Ordered residue ranges**[a]	**Available spectra**[b]	**Completeness of resonance assignment**
HR8254a	2M2E	73	554-608	^{13}C-NOESY, ^{13}Caro-NOESY, ^{15}N-NOESY	81.4 %
StT322	2LOJ	63	23-63	^{13}C-NOESY, ^{13}Caro-NOESY, ^{15}N-NOESY	77.7 %
enth	1VDY	140	9-102, 113-130	^{13}C-NOESY, ^{15}N-NOESY	76.0 %
rhod	1VEE	134	6-125	^{13}Cali-NOESY, ^{13}Caro-NOESY, ^{15}N-NOESY	72.2 %
fsh2	1WQU	114	8-109	^{13}C-NOESY, ^{15}N-NOESY	71.1 %

[a] Determined by CYRANGE.
[b] X-NOESY: 3D X-resolved NOESY-HSQC spectrum.

The HR8254a and StT322 test structures were created by randomly deleting a variable percentage of peaks from all available refined peak lists, and/or a variable percentage of the assigned resonances. Structure calculations with CYANA were performed using these truncated data files, and the resulting structure bundle was included in the test set if the residues corresponding to the ordered regions of the reference structure had an RMSD value below 2.0 Å. Structures with

Table II.2.5: Noise levels and solvent signal ranges of the test spectra.

Protein	Spectrum type	Noise level[a]	Solvent signal range[a]	Linewidths [Hz][a,b]
HR8254a	^{13}C-NOESY	4.5×10^7	–	72.0, 28.0, 150.0
	^{13}Caro-NOESY	1.5×10^7	–	112.0, 35.0, 220.0
	^{15}N-NOESY	1.3×10^7	–	64.0, 33.0, 42.0
StT322	^{13}C-NOESY	4.9×10^6	4.39–4.9 ppm	48.0, 29.0, 160.0
	^{13}Caro-NOESY	9.0×10^6	–	75.0, 36.0, 160.0
	^{15}N-NOESY	9.3×10^6	–	60.0, 25.0, 30.0
enth	^{13}C-NOESY	3.0×10^3	4.5–4.9 ppm	40.0, 73.0, 210.0
	^{15}N-NOESY	1.0×10^4	–	34.0, 85.0, 47.0
rhod	^{13}Cali-NOESY	4.0×10^3	4.5–4.8 ppm	45.0, 102.0, 138.0
	^{13}Caro-NOESY	3.0×10^3	–	35.0, 96.0, 131.0
	^{15}N-NOESY	4.5×10^3	–	31.0, 96.0, 30.0
fsh2	^{13}C-NOESY	5.0×10^3	4.5–4.9 ppm	43.0, 82.0, 115.0
	^{15}N-NOESY	1.0×10^4	–	30.0, 91.0, 36.0

[a] Manually determined. Linewidths were determined by calculating the median of the linewidth values output by the CCPN software (**?**) for a set of picked peaks. Solvent signal ranges were determined by visual inspection of the spectrum. For noise level determination intensity measurements were carried out in several peak-free areas of the spectrum, and the noise level was set to the mean absolute noise intensity plus five standard deviations.

[b] In the following order of the spectral dimensions: HC/HN, H, C/N.

$\mathrm{RMSD}_{\mathrm{ref}}$ values between 1.3 and 6.8 Å in the case of StT322 and between 1.2 and 8.5 Å in the case of HR8254a were generated.

II.2.1.6 Alternative methods of validation score calculation

A number of different approaches to the calculation of a validation score were tested. In each case the overall structure of the testing routine was the same, i.e. peaks were either picked or read from a peaklist provided by the user, these peaks were matched to those predicted based on the query structure, and the peaks were sorted into different classes depending on the match results. The tested methods differ in how the peak information was then used in the computation of ζ, and which parts of the available information were used at all. The evaluated methods are outlined below.

Different weight modes

In the calculation of ζ the distance deviation penalty associated with a peak is weighted by a factor that, in part, depends on the peak class. The exact method of weight computation is described in section II.2.1.1, but a number of different weight modes were tested in combination with the otherwise unchanged validation procedure:

1. `density`: each peak is weighted only by the inverse local peak density.

2. `dist`: each peak is weighted by the inverse inter-atomic distance. If the peak belongs to the *match* class, either the calibrated or the structure-based distance is used, whichever is smaller. For *missing* peaks the structure-based and for *unpredicted* peaks the calibrated distance is used.

3. `equal`: the weight of each used peak is set to the product of the inverse local peak density and the inverse inter-atomic distance (see weight-mode `dist` for details).

4. `no_info`: information content data are not used to calculate the weighting factors of both missing and matching peaks.

5. `no_OP`: analogous to `no_info`, but in this case order parameter values are excluded.

6. `one`: all weights are set to 1, irrespective of the peak class.

Different error functions

The distance deviation penalty function used during the calculation of ζ is given by eq. II.2.16. However, a number of other penalty functions were evaluated as well. Their graphs are depicted in fig. II.2.5.

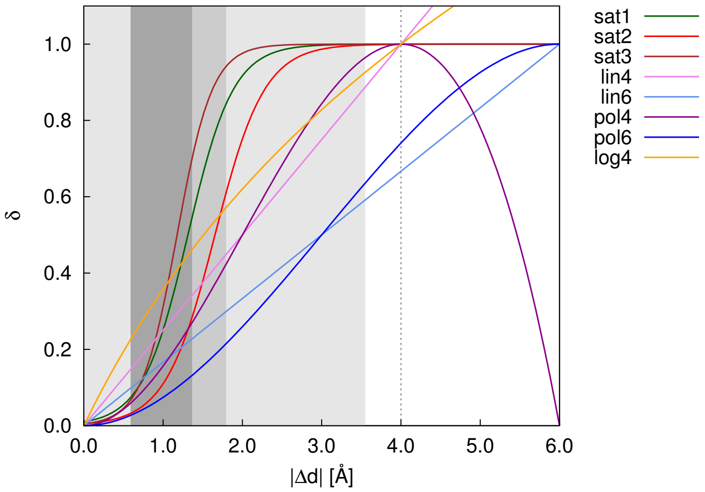


Figure II.2.5: The functions that were evaluated as possible penalty functions to be used for the calculation of ζ. The shaded areas give an indication of the observed $|\Delta d|$ distribution: each grey area contains 25 % of all $|\Delta d|$-values, i.e. 75 % of all observed distance deviations were no larger than 1.8 Å, and no values above 3.55 Å were observed. Beyond 4.0 Å, marked by the dashed vertical line, `lin4`, `pol4`, and `log4` would be artificially set to 1.0 during the calculation of ζ.

Those functions belong to four different classes: saturation functions (`sat1`, `sat2`, `sat3`) (eq. II.2.16), linear functions (`lin4`, `lin6`) (eq. II.2.20), third-order polynomials (`pol4`, `pol6`) (eq. II.2.21), and one logarithmic function (`log4`, eq. II.2.22). The `lin`, `log`, and `pol` functions are distinguished by the absolute value of the distance deviation (in Å) at which their function value is equal to 1.0. `lin4`, `log4`, and `pol4` are equal to 1.0 at 4.0 Å, `lin6` and `pol6` at 6.0 Å. All functions have values of at most 1.0 in the range between 0.0 and either 4.0 or 6.0 Å, in order to ensure that ζ lies within the interval [0.0,1.0].

The individual function parameters can be found in tab. II.2.6.

$$f(|\Delta d|) = m|\Delta d| \tag{II.2.20}$$

$$f(|\Delta d|) = a|\Delta d|^3 + b|\Delta d|^2 \tag{II.2.21}$$

$$f(|\Delta d|) = \ln\left(\frac{\mathrm{e}-1}{q}|\Delta d| + 1\right) \tag{II.2.22}$$

Table II.2.6: Parameters of the evaluated error functions.

Function	σ	m	s	a	b	q
sat1	1.0	3.5	0.010	-	-	-
sat2	1.0	3.2	0.005	-	-	-
sat3	1.0	4.5	0.005	-	-	-
lin4	-	$\frac{1}{4}$	-	-	-	-
lin6	-	$\frac{1}{6}$	-	-	-	-
pol4	-	-	-	$-\frac{1}{32}$	$\frac{3}{16}$	-
pol6	-	-	-	$-\frac{1}{108}$	$\frac{1}{12}$	-
log4	-	-	-	-	-	4

Different peak selections

Different ways of selecting which peaks to include when determining ζ were tested:

1. `no_unpredicted`: CYVAL does not take into account the *unpredicted* class of peaks during the calculation of ζ.

2. `no_missing`: Like `no_unpredicted`, but missing instead of unpredicted peaks are discarded.

3. `match_only`: Both unpredicted and missing peaks are disregarded and the validation is based solely on the class of matched peaks, as using unpredicted and/or missing peaks might render the method too vulnerable to artefacts, peak overlap, and the like, in case the spectral quality is poor.

4. `long_only`: Like `match_only`, with the additional condition of using only long-range peaks, i.e. peaks whose two protons belong to residues that are at least four (i to i+4) positions apart within the sequence, as it is those interactions that provide the most meaningful information about the tertiary structure.

Filtering peaks based on a random structure

The radius of gyration of the query protein is determined. A structure bundle comprising 20 random conformers is created from the query protein's sequence, with the radius of gyration being employed as a restraint to produce an ensemble of compact instead of extended structures. This structure bundle is then used as the query structure during an initial partial validation cycle. The aim of this partial cycle is to identify all picked peaks that do not contribute any meaningful information to the overall validation process, as restraints that are satisfied both by the true query structure and a random structure are assumed to be unlikely to help differentiate between correct and incorrect conformations, and they might even have the potential to mask the presence of indicators of incorrect folds. Hence it was tested whether the validation power of CYVAL would be enhanced by the removal of peaks that receive the same classification for a random structure bundle as for the actual query structure.

The filtering procedure The random structure bundle serves as the query structure in the first partial validation round. This means that peaks predicted from this ensemble are matched with the peaks obtained from automated peak picking of the spectrum. After this step information regarding the matching, missing, and unpredicted peaks is stored, and the validation process is stopped, as nothing beyond peak matching data is required from the random structure.

Once peak matching has been performed for the true query structure all peaks are filtered using the respective sets (matching and unpredicted; alternatively missing peaks as well) from the random structure, and the peaks already observed in the random case within the same category are discarded.

While the reason for deleting redundant matches is straightforward, the rationale for also discarding the set of overlapping unpredicted or even missing peaks needs further explanation. Naturally, on its own the fact that an experimental peak is not predicted based on a random structure has no significance, and the same is true for peaks that were predicted based on a random structure but that could not be retrieved from the spectrum. Furthermore, an experimental peak not predicted based on the actual query structure may be an artefact or a noise peak, but it may also stem from a part of the true structure that is incorrectly represented by the structural model. When discarding signals that were unpredicted both for the random and the query structure the underlying assumption is that these peaks are artefacts or noise peaks and that their exclusion will thus lead to a more reliable validation result.

CYVAL predicts a peak only if the corresponding inter-atomic distance does not exceed the NOE distance threshold in any of the conformers. Therefore the probability of a peak being predicted decreases with an increasing separation of the protons' residues within the sequence. It follows that long-range peaks are unlikely to be expected based on a random structure bundle, and the peak prediction for a random ensemble will yield a set of mostly short-range peaks. In consequence, this set ought to be similar for any conformation of a given amino acid sequence, which, in turn, means that it is probably not due to a mismatch between the structural model and the spectrum if one such peak is missing from the experimental data, but that its absence is caused by other factors. Therefore peaks that are missing for both the random and the query structure may also be filtered out.

II.2.1.7 Comparison of CYVAL to three other validation tools

As described in section II.2.2 the CYANA macro `validate.cya`, accessible within CYANA via the command `validate`, provides access to a number of third-party validation tools. The protein structure bundle is the only required input data, as these tools operate in a knowledge-based fashion.

Three of the tools, ProSa2003 (**?**), Verify3D (**???**), and MolProbity (**????**) were selected for comparison of the validation power of some of their validation scores to the validation power of CYVAL. The chosen scores were:

1. Verify3D score
2. zp-comb (ProSa2003)
3. MolProbity score

Each structure bundle used during the evaluation of CYVAL was subjected to validation with each of the three tools using `validate.cya`. The ensemble coordinates were read by CYANA, which coordinated the invocation of the external applications, and which also collected the output. The latter was subsequently analyzed in an automated fashion.

As all three tools assess the quality of single structures instead of structure bundles each conformer was evaluated separately. For each structure bundle the median value of its conformers' scores was calculated and used for the computation of the correlation coefficient of this particular score against the structure bundles' $\mathrm{RMSD}_{\mathrm{ref}}$-values. Those $\mathrm{RMSD}_{\mathrm{ref}}$-values were the same that were used to evaluate the validation power of CYVAL, i.e. the backbone RMSD to the mean reference structure, calculated for the ordered parts of the reference structure bundle.

II.2.2 Structure validation bundle

The new CYANA command `validate` calls a number of knowledge-based structure validation tools from within CYANA. It collects the results from the external programs in a concise report, which the user can then refer to for an overview of the different validation metrics. Furthermore, all of the output files generated by the programs are stored in a clear directory structure for the user to refer to if a more detailed analysis of the validation results is to be performed.

The software bundled by `validate` are ProSa2003 (**?**), WHAT_CHECK (**??**), Verify3D (**??**), PROCHECK-NMR (**??**), MolProbity (**????**), the PDB validation suite (available from `http://sw-tools.pdb.org/apps/VAL/`), and ProQ (**?**). These tools are described in section I.5.4.1. They were chosen since they have a large user base, they are available for local installation and can thus be easily invoked from within CYANA, and they perform different knowledge-based assessments of the query structure. This presents the user with a more detailed and diverse report of possible structural problems than the output of a single program could provide.

Three scripts in total contribute to the `validate` functionality:

1. `validate.cya` prepares the input files required for the different tools. If the query structure is an ensemble of conformers the script writes a separate PDB file for each individual model, as some of the tools can only analyze single structures. Furthermore, `validate.cya` initiates the creation of secondary structure files in the formats required by Verify3D and ProQ.

2. `validation.sh` is invoked by `validate.cya`. It calls the different validation tools and coordinates processing of their output. `validation.sh` also outputs error messages if problems with one or several of the external programs were encountered, and it deletes all temporary files that will not be required by the user even for a detailed analysis of the validation results. `validation.sh` creates an overview file called `CYANA_Validation_Overview.log`, which is located within the current working directory. Additionally, a main validation directory is created, also within the working directory, with individual sub-directories for the output of the external validation tools. The main directory name is of the format `validation_outputHHMMddmm`, where `HHMMddmm` is a time stamp including the current hour (`HH`) in 24 hour-format, minute (`MM`), day of the month (`dd`), and month (`mm`).

3. `validation_output.pl` is called by `validation.sh` once for each of the validation programs. It reads the output of the program, filters it, and writes the most important information to the overview file. Moreover, it automatically creates a plot visualising the Verify3D results, and it neatly summarizes the output of MolProbity in an HTML file that can be viewed in any web browser.

An example of a validation overview file can be found in the appendix (listing C.1). In the cases of those tools that work on single structures instead of the entire ensemble the validation results of each conformer are reported.

All three scripts are designed in a way that allows the user to easily add, replace, or remove validation programs, but, naturally, routines tailored to the output of a new software need to be included if the user would like this output to be filtered and added to the general validation report.

Chapter II.3

Structure calculation

During the work on this thesis three protein structure calculations from solution NMR data were performed using the program CYANA for automated NOESY cross-peak assignment (**?**) and structure calculation (**??**). During the first project I determined the structures of the wild-type Trp-Trp binding module (WW) domain of the peptidyl-prolyl isomerase Pin1 and of WW$^{\mathrm{S16E}}$, the Ser16Glu-mutant of the WW domain (**?**). The second project focused on TycC3_PCP(S45A), the Ser45Ala-mutant of TycC3_PCP, which is the third peptidyl carrier domain of the tyrocidine A synthetase subunit C from *Bacillus brevis* (**?**). In all cases, sample preparation, data collection, and resonance assignment were performed by my collaborators, thus the current chapter deals solely with the process of structure calculation and automated NOESY assignment. Details regarding all other steps of the structure determination process may be found in the abovementioned publications.

II.3.1 Pin1

The completeness of the chemical shift assignments for the non-labile ^{1}H and the backbone H$^{\mathrm{N}}$ was 95 % for the wild-type protein, and 97 % for the WW$^{\mathrm{S16E}}$ mutant. For each protein, NOE distance restraints were obtained from a 3D ^{15}N-resolved NOE spectroscopy/heteronuclear single-quantum coherence (NOESY-HSQC) and a 3D ^{13}C-resolved NOESY-HSQC peak list by automatic NOESY cross peak assignment and calibration using the program CYANA based on the chemical shift lists. The automatically obtained NOESY assignments were visually inspected, and the peak lists were manually adjusted in case of obvious errors such as artefacts. 90 % of the NOESY cross peaks were assigned in case of the wild-type protein, and 93 % in case of the mutant. The chemical shifts were deposited in the Biological Magnetic Resonance Data Bank (BMRB) with accession numbers 19258 (wild-type WW domain) and 19259 (WW$^{\mathrm{S16E}}$ domain). Backbone dihedral angle restraints were obtained from chemical shift data of the atoms H^N, N, C', C^α, C^β, and H^α using the TALOS+ software (?).

The structure calculation for the mutant protein was performed with CYANA using 100 random starting conformers, and 10,000 torsion angle dynamic steps. The resulting structures were sorted according to their target function values, and the 20 structures with the lowest target function values were selected. Restrained energy refinement was carried out using the OPALp program (?), which employs the AMBER 94 force field (?). As a relatively large number of ^{15}N-resolved NOESY-HSQC cross peaks remained unassigned for the wild-type protein, the combined assignment and structure calculation approach was modified during the final stage of the structure calculation: 20 distinct NOESY assignments and structure calculations, each using a different set of 100 random starting structures, were carried out. Afterwards from each of the structure bundles obtained in the fifth assignment and calculation cycle the conformer with the lowest target function value was extracted and combined with its equivalents from the other 19 calculations. This combined ensemble was used as the basis for two final cycles of structure calculation and NOE assignment, and the obtained final ensemble of 20 conformers and the accompanying list of upper distance limits were used as input to the restrained energy refinement with OPALp. The final structures were deposited in the PDB with the accession code 2m8i for the wild-type WW domain and 2m8j for WW^{S16E} (?).

II.3.2 TycC3_PCP(S45A)

The completeness of the chemical shift assignments for the non-labile ^{1}H and the backbone H^N was 94 %. 3D ^{15}N-resolved NOESY-HSQC and $^{13}C_{aro}$-resolved NOESY-HSQC spectra yielded peak lists from which NOE distance restraints were obtained by automatic NOESY cross-peak assignment and calibration using the program CYANA based on the chemical-shift lists. Backbone dihedral angle restraints were created from chemical-shift data of the atoms H^N, N, C', C^α, C^β, and H^α using the TALOS+ software (?).

The structure calculation was performed with CYANA using 100 random starting conformers and 10,000 torsion-angle dynamics steps. From the resulting structures, the 20 conformers of the lowest target function values were selected. Restrained energy refinement was carried out using the OPALp program (?). The final structure was deposited in the PDB with accession code 2md9 (?).

Part III

Results and Discussion

Chapter III.1

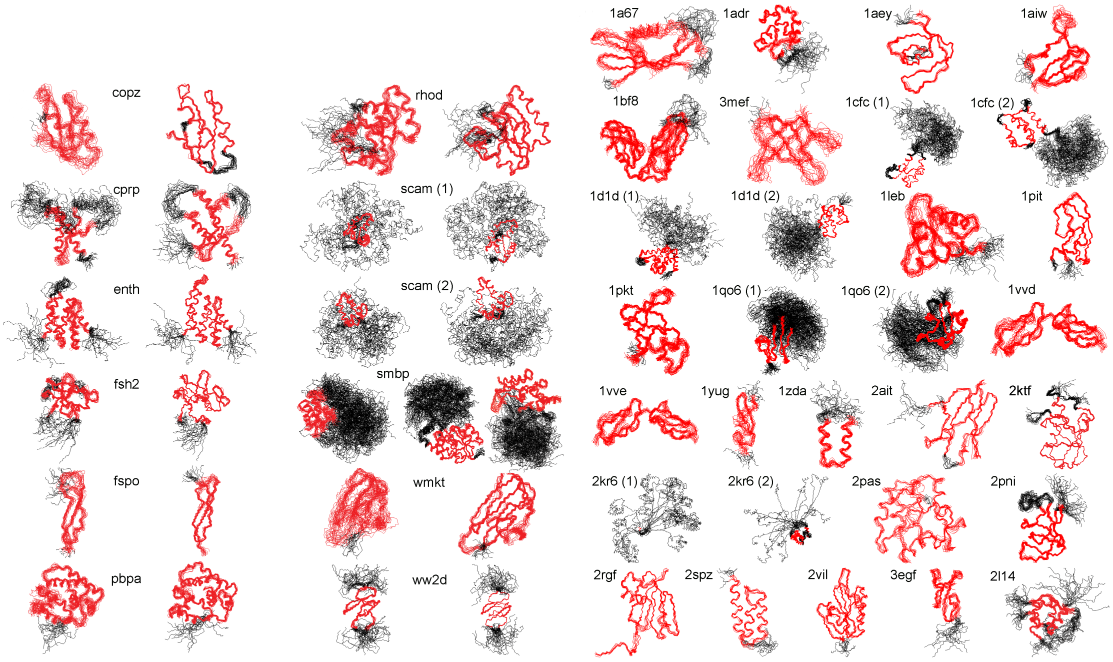
CYRANGE

This chapter presents the results obtained from the determination of RMSD-stable domains for a number of proteins using CYRANGE (**?**). Furthermore, the performance of CYRANGE is compared with that of two other software tools that also aim to detect ordered regions in protein structure bundles. The results show that CYRANGE generally identifies domains that make sense by visual inspection and that typically are both less complex and more complete than the ordered ranges output by the other two programs.

III.1.1 Domain identification in 37 proteins

The residue ranges determined by CYRANGE for 37 proteins are visualized in the 3D structure bundles of fig. III.1.1 and summarized in fig. III.1.2. Details regarding the proteins used are given in section II.1.5.1. The method provides results that make sense by visual inspection in that ordered regions are correctly identified, the global structure superpositions based on the CYRANGE residue ranges allow a clear presentation of the structure, and unnecessary small gaps within the selected ranges are absent.

CYRANGE yielded similar results for the corresponding low- and high-precision structure bundles of the eleven proteins of figs. III.1.1a and III.1.2a, except that in enth and fsh2 additional loops, which are disordered only in the low-precision structure, were excluded, and that in the low-precision structure of smbp the second domain was not identified. For copz, a loop was excluded from the residue range for the high-precision structure that had been included for the low-precision structure because the residues are more uniformly disordered in the low-precision structure than in the high-precision one. The latter exhibited a higher standard deviation of the RMSD per-residue than the low-precision structure. This shows that the CYRANGE method gives meaningful results also in the challenging case of low-quality structures, where the distinction between well-defined and ill-defined regions becomes blurred. For the protein 2kr6, the CYRANGE method correctly identified the two structural domains despite the small size of the isolated helix (figs. III.1.1b and III.1.2b). The two protein-protein complexes, 2ktf and 2l14, did not pose any particular problems.



(a) Eleven proteins for which low-precision (left) and high-precision (right) NMR structures were available (see section II.1.5.1).

(b) 26 proteins, 23 of which had been used earlier for evaluating the FindCore algorithm (?).

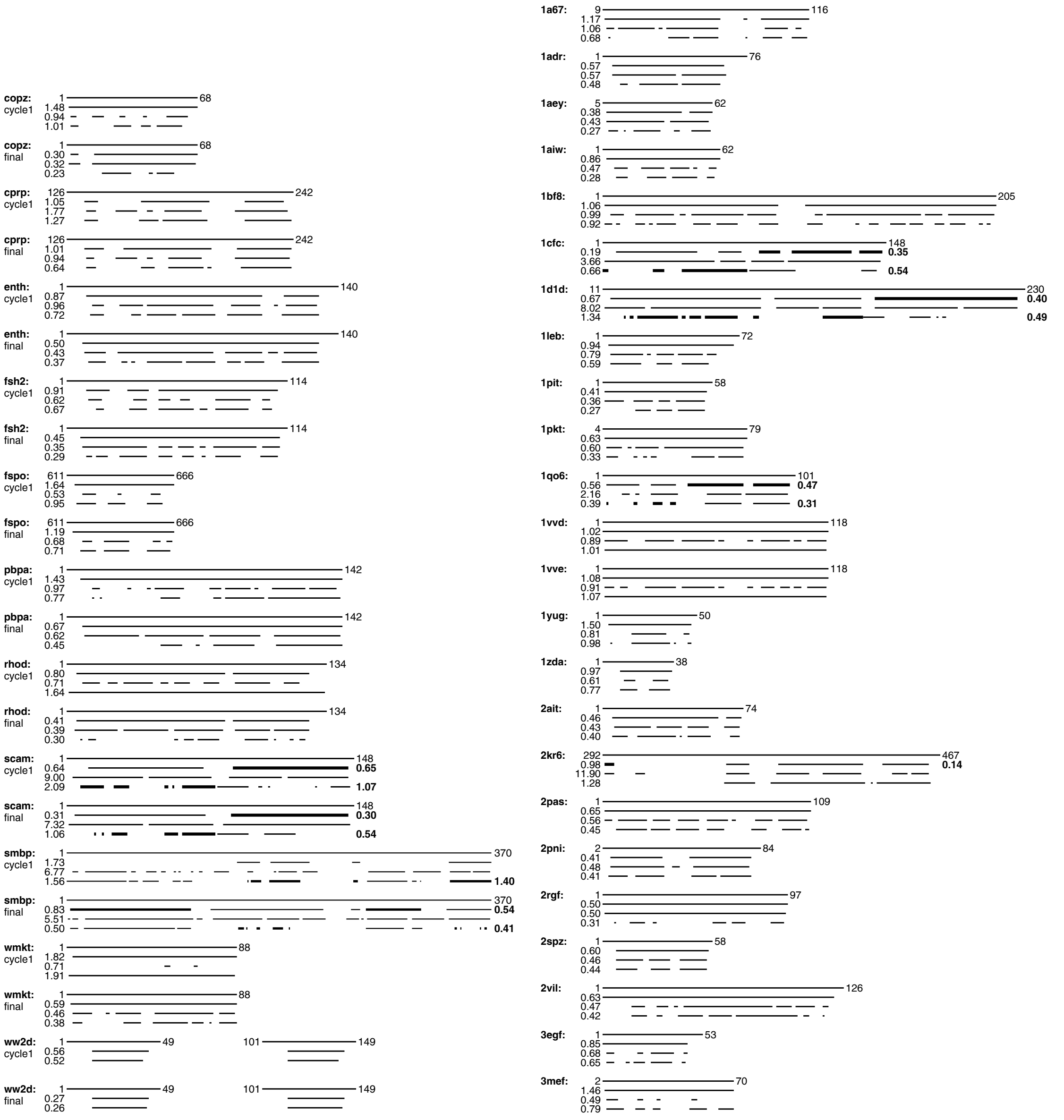
Figure III.1.1: Structure bundles superimposed for minimal backbone RMSD of the CYRANGE-determined residue ranges, indicated in red. Other residues are shown in black. Separate superpositions are presented for each domain.

The CYRANGE results shown in figs. III.1.1 and III.1.2 were obtained using the same parameter set for all proteins. We analyzed the influence of different values of the parameters on the resulting sequence coverage and RMSD for nine different protein structure bundles (see figs. A.3 to A.8 in the appendix). Each of the parameters μ, m, δ, δ^{abs}, γ, and g was varied while keeping the other parameters at their default values.

Varying the minimal cluster size μ in the range of 6-10 residues (fig. A.3), the domain boundary extension m in the range of 1-5 residues (A.4), or the minimal gap width g in the in the range of 1-5 residues (fig. A.5) did not change the CYRANGE residue ranges for these nine proteins, apart from the cases of 1zda and the low-precision structure of copz, where raising m led to minute changes in sequence coverage.

Variation of the parameter δ for the relative RMSD decrease in step 5 of the residue range determination in the range of 0.5-4 led to changes only for the already mentioned low-precision smbp structure (fig. A.6). Higher values of the corresponding absolute RMSD decrease parameter δ^{abs} led to slight increases of the sequence coverage and the RMSD for some of the proteins (fig. A.7).

Similarly, a decrease of the sequence coverage and the RMSD were observed when increasing the gap penalty parameter γ from 0 (no gaps allowed) to 4 (gaps favoured) (fig. A.8). Thus the results from CYRANGE do not critically depend on the choice of these parameters but can, if desired, be guided towards a smaller or larger number of gaps or selected residues.



(a) Low-precision (cycle1) and high-precision (final) NMR structures.

(b) 26 proteins, 23 of which had been used earlier for evaluating the FindCore algorithm.

Figure III.1.2: Comparison of residue ranges found by the CYRANGE, PSVS, and FindCore methods. For each protein, the top line represents the complete amino acid sequence with the first and last residue labelled. Below are the residue ranges obtained from CYRANGE, PSVS, and FindCore. The corresponding RMSD values are indicated in Å. Multiple domains are distinguished by thin and thick lines.

The default values of the parameters appear to be appropriate for almost all proteins. Only the gap penalty parameter γ may occasionally be adapted according to the emphasis put on simple residue ranges with no or few gaps. Meaningful values of γ lie between 0, which yields a residue range without gaps, and 1, which selects residues without concern for the number of gaps.

We also compared the residue ranges obtained by computing the angular order parameters at the outset of the algorithm for all rotatable dihedral angles or only for ϕ, ψ, and χ^1, using default parameter settings. For all but four out of the 37 proteins in the test set the results were identical. In the four other cases the sequence coverage, and usually the RMSD value, were higher when order parameters were calculated for all dihedral angles. The difference in coverage ranged from about 1 to 27 %, and the difference in RMSD amounted to between 0 and about 0.7 Å. The four structure bundles for which differences became apparent were 2kr6, 2spz, the low-precision structure of fspo, and the high-precision structure of smbp. In the case of 2kr6 the small second domain was not identified when only ϕ, ψ, and χ^1 order parameters were used. For smbp the differences amounted to only one or two residues per domain. For 2spz the CYRANGE version using only ϕ, ψ, and χ^1 order parameters unnecessarily identified three small domains instead of the single one reported by standard CYRANGE. For fspo a disordered loop was excluded if only ϕ, ψ, and χ^1 order parameters were employed. Exclusion of this loop does seem sensible; raising the value of γ in the standard CYRANGE will also bring about this result. Overall, both choices of dihedral angle order parameters yielded similar results. Where differences occurred, the results obtained with order parameters for all dihedral angles mostly corresponded better to the conclusions from visual inspection of the structure bundles.

The computation time requirements of the CYRANGE algorithm are insignificant. It took CYRANGE 0.9 s to calculate the residue ranges for pbpa (142 residues, one domain), 1.0 s for scam (148 residues, two domains), and 5.8 s for smbp (370 residues, two domains) on a 2.66 GHz Intel Core 2 processor.

III.1.2 Comparison with PSVS and FindCore

We compared the residue ranges from CYRANGE with those determined by the default algorithm of the Protein Structure Validation Suite PSVS, and by the FindCore algorithm. The algorithm in PSVS (?) is widely used for NMR protein structure validation, and was chosen here as a representative of the straightforward determination of (locally) ordered residues. PSVS does therefore not attempt to identify structural domains. FindCore (?), determines residue ranges for global superposition and is able to identify multiple domains.

The residue ranges obtained with the three algorithms for 37 different proteins are shown in fig. III.1.2, and the differences of the RMSD and sequence coverage relative to the CYRANGE results can be found in fig. III.1.3. In the majority of cases, the residue ranges from CYRANGE contain fewer gaps and cover significantly larger parts of the sequence than those from PSVS and FindCore. Consequently, the RMSD values for the residue ranges identified by CYRANGE are often slightly higher than those from PSVS and FindCore. This, however, does not constitute a general rule. For instance, for the two-domain proteins 1cfc and 1d1d all CYRANGE domains simultaneously comprised more residues and showed lower RMSD values than those obtained from the other two algorithms (fig. III.1.2b). On average, the residue ranges reported by CYRANGE covered 85 % of the sequences of these proteins and led to a backbone RMSD value of 0.77 Å, as compared to 67 % coverage and 1.72 Å RMSD with PSVS, and 58 % coverage and 0.73 Å RMSD with FindCore. In the proteins with multiple domains both CYRANGE and FindCore found all domains except one. CYRANGE missed a domain in the low-precision smbp structure, FindCore in 2kr6.

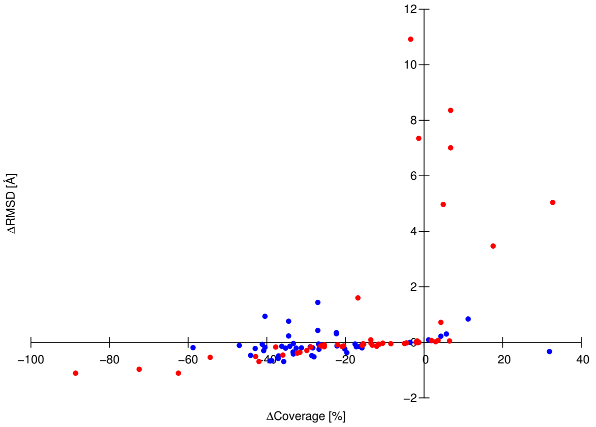
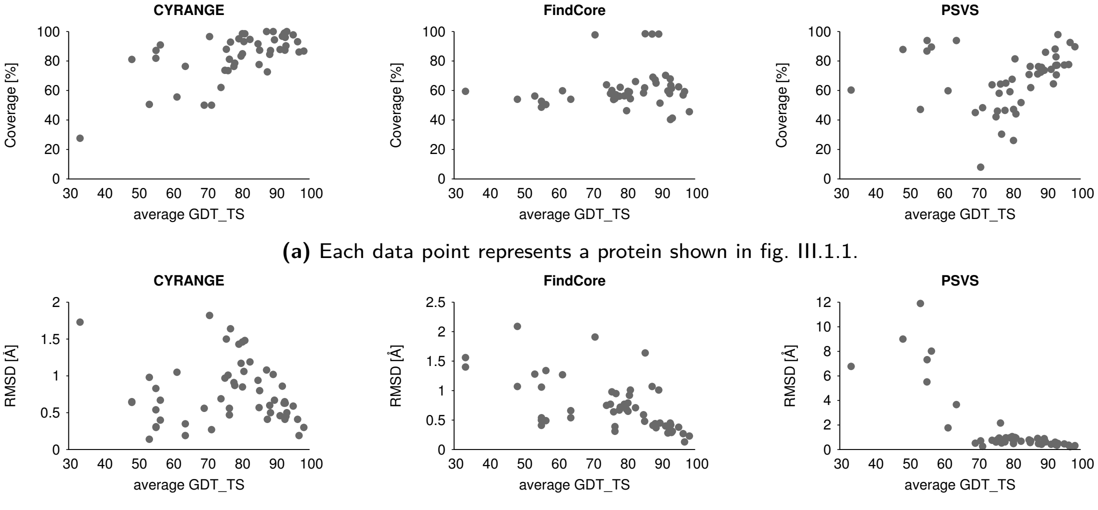


Figure III.1.3: RMSD and sequence coverage differences between the residue ranges found by PSVS (red) or FindCore (blue) and CYRANGE for the proteins shown in fig. III.1.1. The coverage is the percentage of amino acid residues included in the residue ranges found by the different methods.

Fig. III.1.3 compares the RMSD and sequence coverage of the PSVS and FindCore results with those obtained by CYRANGE. Each data point represents one protein domain. Data points above/below the horizontal axis indicate cases where PSVS or FindCore yielded larger/smaller RMSDs than CYRANGE. Data points to the right/left of the vertical axis indicate cases where PSVS or FindCore yielded larger/smaller sequence coverage than CYRANGE. Most data points are found near the horizontal axis in the lower left quadrant. For these proteins CYRANGE covered typically between 10 and 50 % more of the sequence with a small concomitant increase in RMSD. Only a single data point, corresponding to the low-precision smbp structure, is located in the lower right quadrant, indicating a significantly smaller sequence coverage and higher RMSD by CYRANGE (because the algorithm failed to identify the separate domains of this two-domain protein). In all other cases of higher sequence coverage by PSVS or FindCore the greater number of selected residues resulted in larger, often much larger RMSDs. There are also some cases in which, especially with FindCore, simultaneously a smaller sequence coverage and a larger RMSD were found than with CYRANGE.

We also correlated the sequence coverage and the RMSD values obtained by the three methods with the GDT total score, *GDT_TS*, which reports the average percentage of residues that can be superimposed under distance cutoffs of 1, 2, 4, and 8 Å (**?**). Whereas there is a correlation between the sequence coverage by CYRANGE and the *GDT_TS* value, no such correlation is apparent for FindCore or PSVS (fig. III.1.4a). The sequence coverage by FindCore is around 60 % for all proteins except for four cases with nearly 100 % sequence coverage, and independent of the *GDT_TS* value. The RMSD values do not correlate strongly with the *GDT_TS* values for any of the three methods. This is not surprising because the *GDT_TS* measures the fraction of residues that can be superimposed reasonably well, whereas the RMSD reports how well a given subset of residues, which may comprise a smaller or larger part of the entire protein sequence, can be superimposed (fig. III.1.4b).



(a) Each data point represents a protein shown in fig. III.1.1.

(b) Each data point represents one protein domain.

Figure III.1.4: Correlation between the sequence coverage from CYRANGE, FindCore and PSVS, and the GDT total score, *GDT_TS*. The coverage is the percentage of amino acid residues included in the residue ranges found by the different methods. The *GDT_TS* value is defined by $GDT\_TS = (P_1 + P_2 + P_4 + P_8)/4$, where P_d is the fraction of residues that can be superimposed under a distance cutoff of d Å.

More detailed investigations into the differences between FindCore, PSVS, and CYRANGE results were performed, as the results, especially the number of gaps reported by the programs, varied considerably in some cases. Two examples of CYRANGE reporting few or no gaps where FindCore and PSVS reported a high number of gaps are the results obtained for the low-precision structure of pbpa, and for 1bf8. In the first case CYRANGE excluded one disordered chain terminus, but it did not exclude a loop that, by visual inspection, seems slightly less ordered than the rest of the domain. PSVS kept the helices, yet excluded from them small, not considerably disordered-looking segments. It also excluded the chain termini, one of which does not seem highly disordered. The FindCore result was similar. Here we additionally found that some very small portions of otherwise excluded stretches were reported to be part of the domain. From 1bf8

CYRANGE excluded one highly disordered loop. PSVS also excluded this loop, alongside parts that, by visual inspection, did not appear to be considerably disordered. Again, the FindCore result was similar to the one attained by PSVS, but the rather ordered-looking sections excluded by FindCore were often somewhat larger.

III.1.3 Application to all NMR structures in the PDB

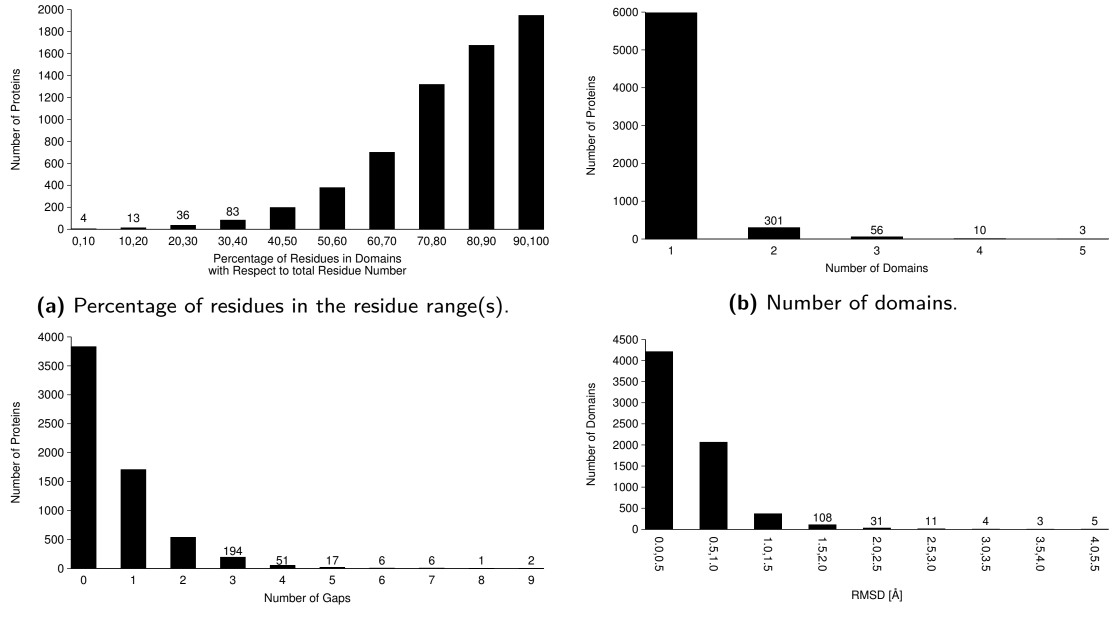
On July 30, 2010 the PDB contained 6373 entries with protein structures determined by NMR that comprised at least 15 residues and for which a bundle of at least five conformers was available. We applied CYRANGE to all of these structure bundles, and obtained results for all but 22 files. In four cases no core atoms could be identified because the PDB files contained C positions only, in one case the residue range determination failed, and in 17 cases no domains were found. 14 of those proteins comprised less than 20 residues and often consisted of a single helix with one or two disordered tails. Of the remaining three structures two were highly disordered, and one, made up of 25 residues, again contained a single helix with disordered tails only. The results for the remaining 6351 files are summarized in fig. III.1.5. A complete list of the residue ranges and RMSD values is available at `http://www.bpc.uni-frankfurt.de/cyrange_pdb.html`. On average, the residue ranges covered 80 % of the residues of a protein, and there were 1.07 domains and 0.59 intra-domain gaps per protein. For 95 % of the proteins the CYRANGE residue ranges comprised more than 50 % of all residues (fig. III.1.5a).

In only four cases (PDB IDs 1r5s, 2fft, 2kes, 2v93) did the residue ranges include less than 10 % of all residues. Visual inspection of these structure bundles revealed that the low percentage of selected residues is correct for the two structures 1r5s and 2fft, which consist of a single helix with long, disordered tails, whereas a larger residue range should have been selected for 2kes, which consists of a single helix. The PDB entry 2v93 could not be handled properly because it combines multiple conformations within the individual conformers.

Fig. III.1.5b shows that in 94 % of structures CYRANGE identified a single domain. Two domains were found in 5 % of the structures, and in 69 cases (1 %) CYRANGE found three or more (at most five) domains. In two cases (1zll and 2hyn) the structure was a pentamer. In the third case (2k27) visual inspection showed that the protein consists of two globular domains, which were correctly identified by CYRANGE with RMSDs of about 1.1 Å, and extended stretches at the chain termini and in the connection between the two globular domains, where CYRANGE identified three small 'domains' of 14-17 residues with RMSDs of 1.23-1.96 Å.

For 60 % of the structures CYRANGE determined a residue range without intra-domain gaps (fig. III.1.5c). Four or more intra-domain gaps were identified in only about 1 % of the structures, i.e. the CYRANGE algorithm selected simple residue ranges with no or only very few gaps whenever this was possible. The distribution of the backbone RMSD values for the domains identified by

CYRANGE (fig. III.1.5d) indicates that 62 % of all domains have an RMSD value below 0.5 Å. Less than 1 % exhibit RMSD values above 2.0 Å, and only five domains reported by CYRANGE are severely disordered with RMSD values in the range of 4.0-5.5 Å, which appear to be largely ill-defined also by visual inspection. This shows that in almost all cases CYRANGE determined residue ranges that can be superimposed well. Considering the large number of domains, it cannot be ruled out that in some cases more appropriate residue ranges could be identified by visual inspection or other methods. Nevertheless, the facts that the algorithm failed to provide a result for only 0.34 % of the PDB entries and that for more than 99 % of the domains CYRANGE yielded residue ranges with RMSDs below 2.0 Å and covering a significant fraction of the sequence indicate the usefulness of CYRANGE as a general tool for objective residue range determination.



(a) Percentage of residues in the residue range(s).

(b) Number of domains.

(c) Number of intra-domain gaps.

(d) RMSD values for the residue range.

Figure III.1.5: Statistics of residue ranges determined with the CYRANGE algorithm for 6351 NMR protein structure bundles in the Protein Data Bank in July 2010.

Chapter III.2

CYVAL

This chapter describes the results obtained from the application of CYVAL to the five protein test sets introduced in section II.2.1.5. In many cases the expected tendency of an increase in ζ and ζ_{n} for rising $\mathrm{RMSD}_{\mathrm{ref}}$ values was observed, and a comparison of CYVAL with three other protein structure validation programs has shown CYVAL's results to be of a similar or even higher quality. Furthermore, this chapter demonstrates how the reliability of CYVAL is affected by the exact method of peak selection and ζ-calculation. While various combinations of distance deviation penalty functions and weight modes yielded results of comparable quality the choice of which peaks contributed to ζ had a strong impact on the validation power of CYVAL.

III.2.1 Expectations

Ideally, all signals within the experimental spectrum should find a counterpart in the list of expected peaks, and vice versa. In such a case the match between structure and spectrum would be considered ideal. However, spectra contain noise and artefacts, and some peaks may not be present or visible due to protein flexibility or overlap with significantly more intense peaks, and some long-range peaks may be of a similar magnitude as the general noise. If certain parts of the data are incorrectly represented by the structure, again there will be mismatches between both sets of peaks.

Based on the definition of ζ as given by eq. II.2.13 it is to be expected that a higher $\mathrm{RMSD}_{\mathrm{ref}}$ value will lead to an increase in ζ because there will most likely be larger distance deviations, and greater numbers of both unpredicted and missing peaks, as the query structure will match the signals from the experimental spectrum less and less well. As can be seen from the definition of the error function (eq. II.2.16 and fig. II.2.3) a larger deviation between a given spectrum-based and its corresponding structure-based inter-proton distance, or between $d_{\max}$ and the spectrum-based/structure-based distance, leads to a larger contribution of this particular peak to the overall value of ζ.

Furthermore, due to the weighting factor of each missing peak containing the information content value of this peak the weights of missing peaks should on average be higher than those of matching or unpredicted peaks. A high percentage of correct short-range interactions will be present even if the query structure exhibits a high $\mathrm{RMSD}_{\mathrm{ref}}$-value as intra-residue interactions and those between neighbouring residues are not greatly influenced by the global fold of the protein. Especially in those cases in which most of the secondary structure elements are formed correctly a large number of expected short-range interactions will most likely have corresponding peaks within the spectrum. Precisely for this reason, however, short-range interactions generally have low information content values, as their restraints do not contribute to defining the tertiary structure of the protein. Long-range interactions, in contrast, are crucial for determining the overall fold, hence their information content values are generally high. Consequently, as missing peaks tend to be those expected from long-range interactions, such peaks typically have high weighting factors. Hence a larger number of missing long-range peaks, which is to be expected for structures of an increasingly incorrect global fold, should significantly contribute to an increase in ζ.

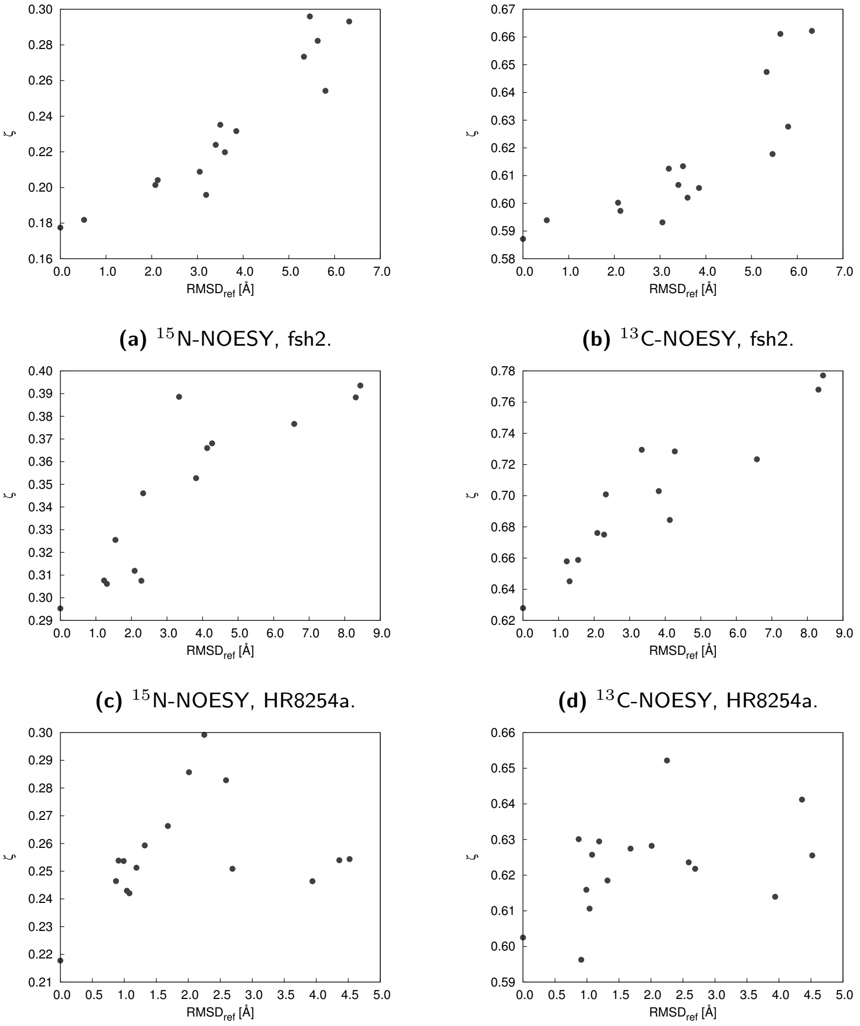
However, $\mathrm{RMSD}_{\mathrm{ref}}$ is just one out of several available metrics for assessing structural accuracy, and there is no direct, linear relationship between $\mathrm{RMSD}_{\mathrm{ref}}$ and ζ. In consequence, correlation coefficients can only be an indicator of the ability of ζ to differentiate between correct and incorrect protein folds, and a correlation coefficient of 1.0 is not necessarily to be expected even in ideal cases. Nevertheless, the correlation coefficient can still provide an idea of how well ζ serves to identify incorrectly folded proteins as the absence of any degree of correlation would clearly demonstrate the inability of a given metric to provide any useful information.

III.2.2 Validation power of CYVAL

As described in section II.2.1.5, five proteins were used in the evaluation of CYVAL. For each of these proteins incorrect structures of different degrees of accuracy were generated, and the collections of these structure bundles were used as test sets for assessing the validation power of ζ and ζ_{n}.

Fig. III.2.1 shows plots of ζ against the $\mathrm{RMSD}_{\mathrm{ref}}$ value, i.e. the structural accuracy, for a number of incorrect structures of the proteins fsh2, HR8254a, and enth. Out of the five test sets enth yielded the poorest validation results. There is no discernible relationship between ζ and $\mathrm{RMSD}_{\mathrm{ref}}$, as illustrated by figs. III.2.1e and III.2.1f. In the cases of HR8254a and fsh2, on the other hand, the behaviour of ζ is clearly in accordance with the expectations outlined above. On average, a lower value of ζ indicates a higher accuracy of the structure bundle. The plots in figs. III.2.1a to III.2.1d illustrate that CYVAL is able to distinguish between structure bundles of high ($\mathrm{RMSD}_{\mathrm{ref}}$ below 2.0 Å), medium (between 2.0 and 4.0 Å), and low (above 4.0 Å) quality.

This observation is corroborated by correlation coefficients between ζ and $\mathrm{RMSD}_{\mathrm{ref}}$ above 0.80 (see tab. III.2.1, *traditional* column). The results listed in tab. III.2.1 also demonstrate that the expected trend of increasing ζ-values with decreasing accuracy is observed for the proteins rhod and StT322 as well, although it is not as pronounced as for HR8254a and fsh2.



(a) ^{15}N-NOESY, fsh2.

(b) ^{13}C-NOESY, fsh2.

(c) ^{15}N-NOESY, HR8254a.

(d) ^{13}C-NOESY, HR8254a.

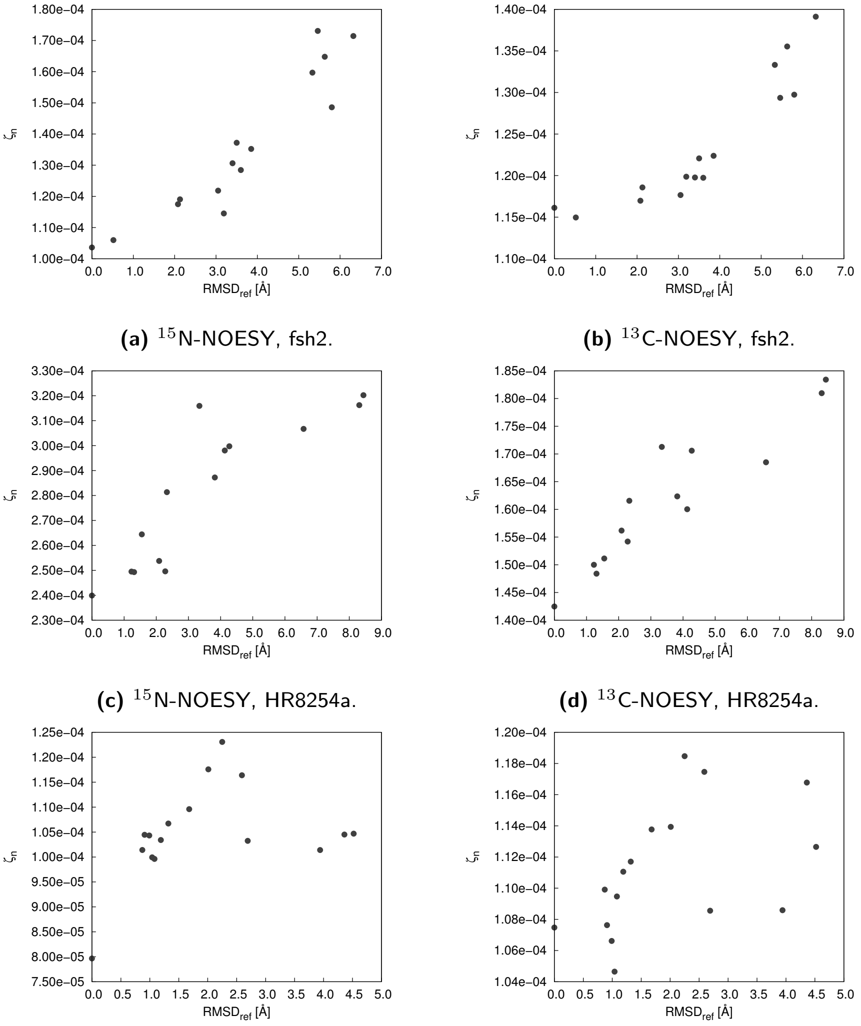
(e) ^{15}N-NOESY, enth.

(f) ^{13}C-NOESY, enth.

Figure III.2.1: Relationship between structural accuracy and ζ for the two most common types of NOESY spectra. For the proteins HR8254a and fsh2 structures with large $\mathrm{RMSD}_{\mathrm{ref}}$ values could clearly be distinguished from those with a high accuracy, i.e. those with low $\mathrm{RMSD}_{\mathrm{ref}}$ values. The examples show data from the proteins fsh2, HR8254a, and enth.

Despite the general trend of increasing ζ with decreasing accuracy that is observed for the HR8254a and fsh2 test sets there are some outliers. In fig. III.2.1c, for example, an ensemble whose $\mathrm{RMSD}_{\mathrm{ref}}$ value is at about 3.5 Å exhibits a value of ζ of the same order of magnitude as structures of $\mathrm{RMSD}_{\mathrm{ref}}$ values above 8.0 Å, and in fig. III.2.1b a structure bundle at an $\mathrm{RMSD}_{\mathrm{ref}}$

value around 3.0 Å yielded roughly the same ζ value as another structure whose $\mathrm{RMSD_{ref}}$ value lies below 1.0 Å.



(a) ^{15}N-NOESY, fsh2.

(b) ^{13}C-NOESY, fsh2.

(c) ^{15}N-NOESY, HR8254a.

(d) ^{13}C-NOESY, HR8254a.

(e) ^{15}N-NOESY, enth.

(f) ^{13}C-NOESY, enth.

Figure III.2.2: Relationship between structure accuracy and ζ_n, i.e. ζ normalized by the number of used peaks, for the two most common types of NOESY spectra. The proteins and spectra used here are the same as those of fig. III.2.1, where the standard value of ζ is plotted.

The results depicted in fig. III.2.2 stem from the same test sets and validation metric calculation approaches as those of fig. III.2.1, the only difference being that now ζ_n instead of ζ is plotted against $\mathrm{RMSD_{ref}}$. A comparison of subfigures (a) and (b) with their counterparts from fig. III.2.1 reveals that, particularly in the case of the ^{13}C-NOESY spectrum, the increase in ζ_n with decreasing accuracy is more pronounced than it is for ζ. Moreover, the outlier observed at around 3.0 Å is no longer present. This leads to the conclusion that the uncharacteristically low ζ value of the structure bundle in question is to do with the number of used peaks not following the general trend. The *traditional* column in tab. III.2.1 shows that ζ_n generally produces a correlation between itself and $\mathrm{RMSD_{ref}}$ at least as good or even slightly better than ζ, but the differences are not pronounced.

The original aim of computing ζ_n was to create a metric that gives a per peak-indication of structural accuracy and that would thus be entirely independent of the individual query protein and spectrum type. However, the range of values of both ζ and ζ_n depends not solely on the accuracy of the structure but also on the protein in question, and on the spectrum used for validation. For fsh2, for instance, the values of ζ_n fall in the range from 1.0×10^{-4} to 1.8×10^{-4} for the ^{15}N-NOESY spectrum and from 1.1×10^{-4} to 1.4×10^{-4} for the ^{13}C-NOESY spectrum. In the case of HR8254a, on the other hand, those ranges are 2.4×10^{-4} to 3.3×10^{-4} and 1.45×10^{-4} to 1.85×10^{-4}, respectively. Moreover, the results show that the weak performance of CYVAL for enth does not stem from different numbers of peaks contributing to ζ in a way that is meaningful in each individual case, but that some unknown features of the enth test set render it immune to a structure quality assessment with the traditional CYVAL approach (however, tab. B.1 shows that this test set is amenable to validation by CYVAL when a different approach to the calculation of ζ is employed; see section III.2.4 for details).

All in all, the results obtained for enth show that the CYVAL method is not yet applicable to all possible cases. The encouraging validation success for the HR8254a and fsh2 test sets, however, as well as the promising tendencies observed for StT322 and rhod illustrate that CYVAL is indeed able to distinguish between structural models of different accuracies.

III.2.2.1 Peak intensity correction

As described in section II.2.1.1 the user may choose to let CYVAL correct experimental peak intensities for peak overlap. Inspection of the results has shown that peak intensity correction only has a negligible influence on the overall quality of the validation results, both in terms of $\mathrm{RMSD_{ref}}$-ζ correlation and in terms of the presence of outliers. Besides, in some cases the intensity correction procedure failed due to issues with a third-party Fortran routine that is unable to handle amounts of data of the size occurring during the intensity correction step. From the available results, however, it can be inferred that raw peak intensities are generally sufficient for the evaluation of the overall protein fold. Furthermore, if the quality assessment is not successful when raw intensities are used, correcting for peak overlap will not necessarily improve the validation power of ζ or ζ_n.

III.2.2.2 Runtimes

All validations were run on a cluster equipped with dualboards (two octacore CPUs, Intel Xeon E5-2690 2.9 GHz; 16 GB memory), and the runtimes were moderate to negligible. Validation of StT322, the smallest test protein with 63 residues, against its $^{13}\mathrm{C_{aro}}$-NOESY, ^{15}N-NOESY, or ^{13}C-NOESY spectrum took 7 s (42 used peaks), 1:02 min (785 used peaks), and 3:32 min (3249 used peaks), respectively. For the largest test protein enth (140 residues) the validation against its ^{15}N-NOESY spectrum took 42 s (2743 used peaks), and 8:30 min (5606 used peaks) against its ^{13}C-NOESY spectrum. Validating fsh2 (114 residues) against its ^{15}N-NOESY spectrum took 32 s (1713 used peaks), and 6:00 min against the ^{13}C-NOESY spectrum (5056 used peaks).

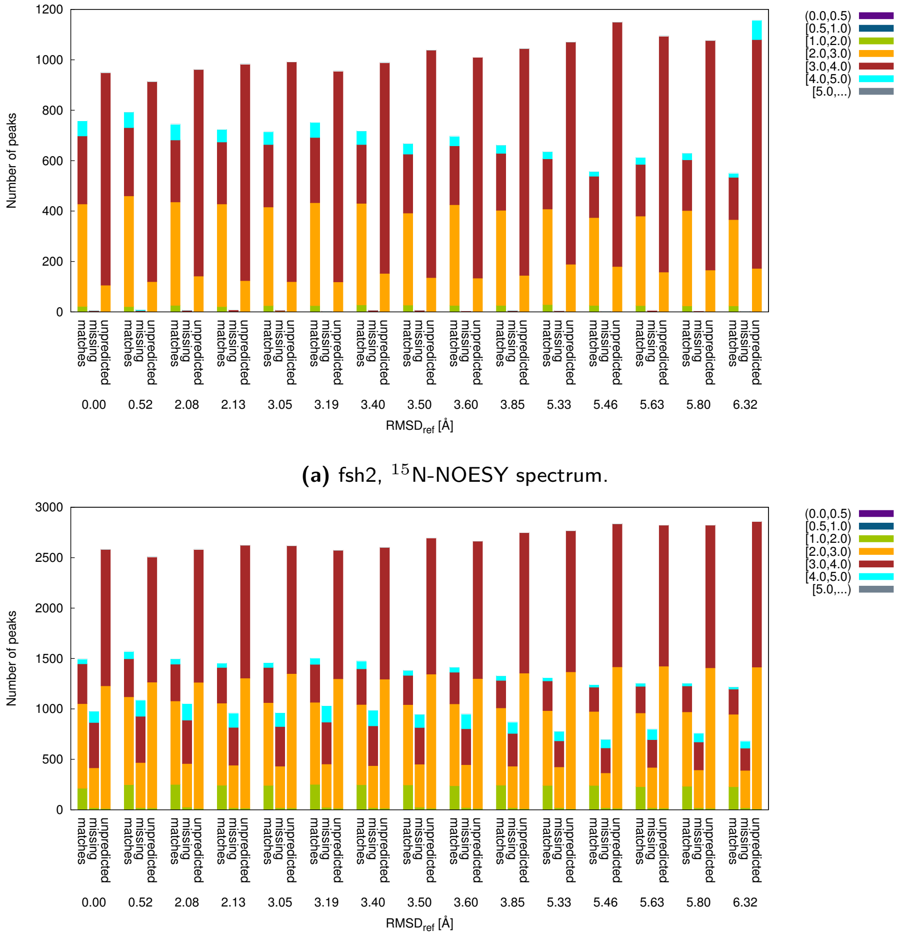
III.2.3 Components of ζ

Validating a structure whose degree of accuracy is low should lead to a large number of unexpected peaks, as comparatively few matches will be found amongst the two sets of picked and predicted peaks. Consequently, when comparing a series of validation results for structures of various RMSD$_{\text{ref}}$ values, an increase in unpredicted and a concomitant decrease in matching peaks is to be expected as the degree of accuracy of the structures goes down. Furthermore, the group of long-range matches should be more severely affected than short-range matches, as the former depend more strongly on the overall fold than the latter, which mainly depend upon the local conformation.

Fig. III.2.3 shows an example distribution of the distance classes within the sets of matching, missing, and unpredicted peaks. The unpredicted peaks always outnumber the primary matches, and, in the case of the ^{15}N-NOESY spectrum (III.2.3a), the number of missing peaks is negligible. The number of unpredicted peaks increases as the degree of accuracy drops, while the number of matching peaks goes down, with the matching peaks in the ranges from 3.0 to 4.0 and from 4.0 to 5.0 Å being disproportionately affected. In the case of the ^{13}C-NOESY spectrum (III.2.3b) both the number and the fraction of missing peaks decrease as well. Furthermore, there are virtually no interactions shorter than 2.0 Å amongst both missing and unpredicted peaks, which illustrates that certain short-range interactions are generally present within a protein of a given amino acid-sequence, irrespective of the overall fold. These observations, apart from the decrease in the number of missing peaks, correspond to the expectations stated above. The development of the numbers of unpredicted, missing, and matching peaks relative to each other should affect the value of ζ in a way that makes the metric reflect the degree of structural accuracy.

The comparatively high number of unpredicted peaks is partly due to the peak matching approach (see section II.2.1.1), where several predicted peaks may be mapped onto the same experimental peak. The predicted peak that yielded the highest match score, the *primary match*, is used to assign the experimental peak to a proton pair, and it is counted towards the number of matching peaks. The remaining predicted peaks that match this particular picked peak are registered as *secondary matches* (and thus do not fall into the category of missing peaks). This makes them unavailable as matches for other experimental peaks, which may thus remain unassigned.

It is not clear why the number of missing peaks in fig. III.2.3b dropped as the degree of accuracy decreased. This behaviour could be explained by a concomitant decrease in ensemble precision, as this would lead to fewer interactions being present in every conformer and thus fewer medium-range and particularly long-range peaks being predicted. However, only test structures with a bundle RMSD below 2.0 Å, superimposed on the ordered ranges obtained from the reference structure, were selected, and additionally no direct link between ensemble precision and RMSD$_{\mathrm{ref}}$ value was observed. Therefore, increasingly ill-defined structure bundles cannot be the reason for the counter-intuitive development of the number of missing peaks.

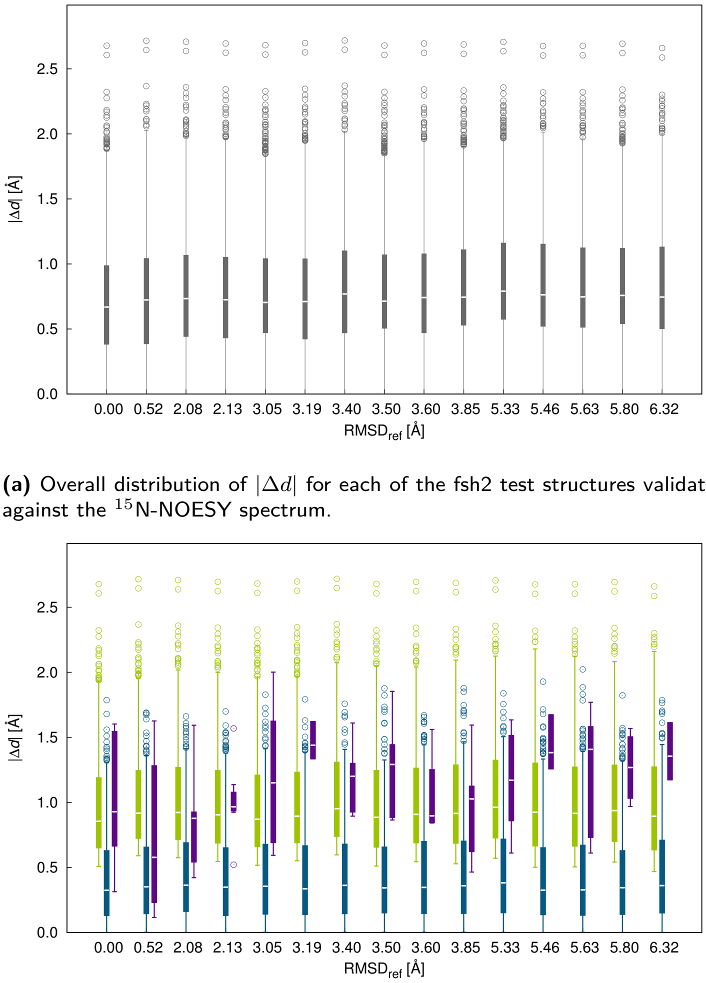


(a) fsh2, ^{15}N-NOESY spectrum.

(b) fsh2, ^{13}C-NOESY spectrum.

Figure III.2.3: Absolute numbers of matching, missing, and unpredicted peaks, split into distance classes. The distance intervals (in Å) are given at the upper right; parentheses mean than the adjacent value is not included in the interval, whereas square brackets indicate that the adjacent value is included. For the matching and missing peaks structure-based distances are shown, and the distances of the unpredicted peaks are spectrum-based, i.e. they result from the calibration of picked peaks. Secondary matches are not shown.

Fig. III.2.4a shows the overall distribution of $|\Delta d|$ obtained from the validation of a number of different fsh2-structures against the ^{15}N-NOESY spectrum of the protein. There appear to be no marked differences across the range of RMSD$_{\mathrm{ref}}$ values. In fig. III.2.4b, $|\Delta d|$ distributions for the same structure bundles are shown; this time, however, the plot differentiates between the three classes of missing, unpredicted, and matching peaks, and only there do differences between the various ensembles become apparent.



(a) Overall distribution of $|\Delta d|$ for each of the fsh2 test structures validated against the ^{15}N-NOESY spectrum.

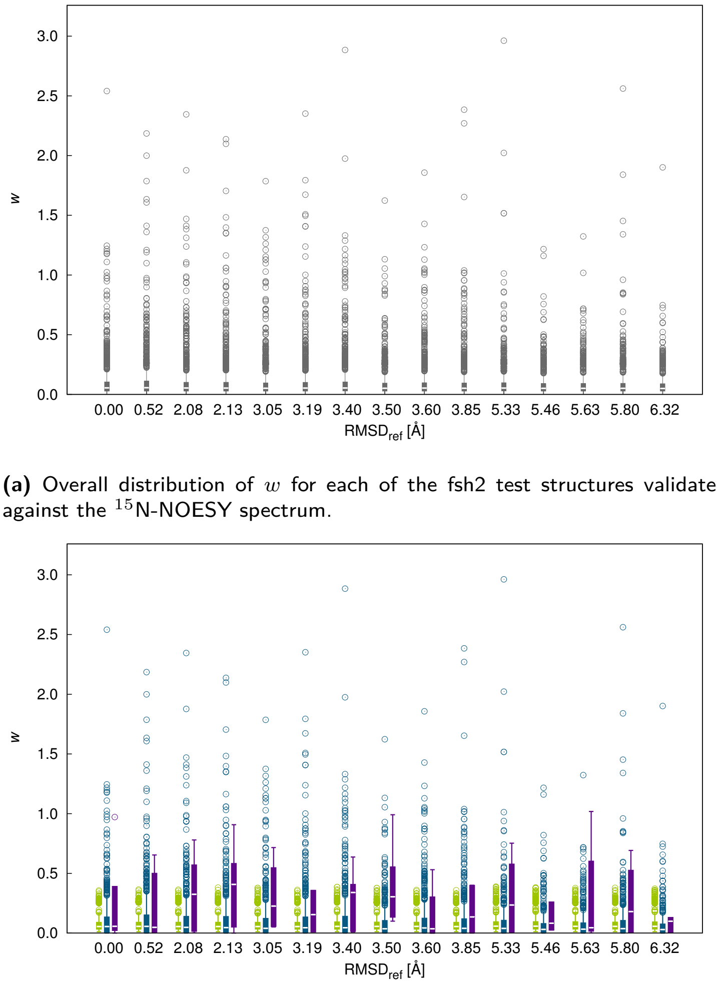
(b) Distribution of the $|\Delta d|$ values within each of the three peak classes, unpredicted (green), matching (teal), and missing (violet).

Figure III.2.4: $|\Delta d|$ distribution obtained from the validation of several different structural models of fsh2 against the ^{15}N-NOESY spectrum of the protein. Each peak used in the calculation of ζ contributes one data point, i.e. one $|\Delta d|$ value, to the distribution. The white line within each box marks the median value of the distribution; the box itself spans the inter-quartile range, i.e. it indicates the width of the distribution. 25 % of all data points are smaller than the lower boundary and 75 % are smaller than the upper boundary of the box. The tails extend to the smallest and largest data points that are located less than twice the inter-quartile range away from the median. All data points beyond this range are classified as outliers and are represented by circles.

On average, the $|\Delta d|$ value of a matched atom pair is low: in each case, irrespective of RMSD_{ref}, the majority is smaller than 0.5 Å, the maximum value is always smaller than 2.0 Å, and only few outliers exceed 1.5 Å. The median value within this category fluctuates around 0.3 Å. As illustrated by fig. II.2.3 this means that the majority of matching peaks will only lead to small error function contributions.

The distribution of $|\Delta d|$ values within the group of unpredicted peaks appears to be the most homogeneous over the range of structural inaccuracies. Here, the median value generally lies at around 0.8 Å, and the largest distance deviation always lies at roughly 2.7 Å. Consequently, the set of unpredicted peaks will lead to significantly higher error function values than the group of matching peaks.

The group of missing peaks exhibits by far the largest variability in its $|\Delta d|$ distribution. Median values range from 0.6 Å to 1.4 Å, and there is no discernible trend that would link the shape of the distribution to the RMSD_{ref} value. In the majority of cases the set of missing peaks exhibits a higher median distance deviation than the other two peak categories for the same structure bundle, but the distribution is generally narrower, which means that neither distance deviations close to 0.0 Å nor above 2.0 Å are present. Nevertheless, peaks from this category will generally contribute relatively large error function values to the overall calculation of ζ. The erratic behaviour of the $|\Delta d|$ distribution of the missing peaks is explained by their low numbers, which prevent the averaging out of fluctuations that consequently manifest themselves as strongly differing sets of (few) values.



(a) Overall distribution of w for each of the fsh2 test structures validated against the ^{15}N-NOESY spectrum.

(b) Distribution of the w values within each of the three peak classes, unpredicted (green), matching (teal), and missing (violet).

Figure III.2.5: Distribution of the weighting factors (w) obtained during the validation of several different fsh2 structure models against the ^{15}N-NOESY spectrum of the protein. The description of this type of plot can be found with fig. III.2.4.

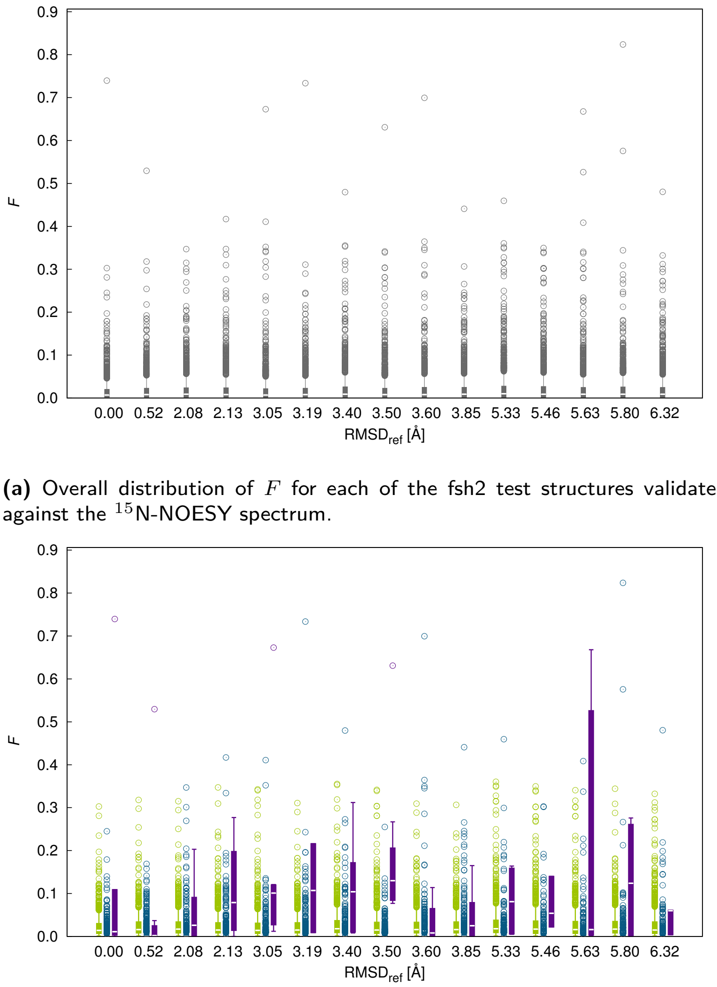
As shown by eq. II.2.15 the true contributions of the peaks also depend upon the weighting factors (w), whose distributions are plotted in fig. III.2.5. Here, as in the case of $|\Delta d|$, the overall distribution of peak weights (III.2.5a) is rather uniform across the set of test structures, and again there is no apparent relationship between the width of the distribution and $\mathrm{RMSD}_{\mathrm{ref}}$. The main difference lies with the maximum weights, i.e. with the outliers. Both the location of the median weight and the inter-quartile range are roughly constant.

The median value lies below 0.1 Å, and 75 % of all values are smaller than 0.1, which means that 75 % of all peaks will contribute at most 1/10 of their error function value to the numerator in the calculation of ζ, and 50 % will contribute even less. The lengths of the tails of the distribution do not greatly vary, either, and they generally extend to around 0.2; the values of the largest outliers, on the other hand, differ widely, from below 1.5 to 3.0. This means that, independent of the degree of accuracy of the structure, a comparatively small number of peaks may contribute strongly to the overall value of ζ, especially if their associated $|\Delta d|$ values are large.

When looking at the three peak categories (fig. III.2.5b), differences between the structural models and the groups of peaks become apparent. The distribution of the weights of the unpredicted peaks is very uniform across the range of $\mathrm{RMSD}_{\mathrm{ref}}$ values. The median value lies always at around 0.06, the upper adjacent value reaches up to about 0.17, and weights above 0.5 do not occur. Consequently, despite the comparatively large $|\Delta d|$ values observed for unpredicted peaks, the contribution of an individual unpredicted peak to ζ will be significantly scaled down on account of the low peak weight. This is desirable as a large portion of unpredicted peaks may well be noise peaks or artefacts, and these contain no information regarding structural accuracy. Comparatively low weighting factors mean that individual unpredicted signals cannot easily dominate ζ, especially so if their calibrated distances are large, which renders the peaks particularly unreliable.

With increasing $\mathrm{RMSD}_{\mathrm{ref}}$ matching peaks exhibit a decrease both in the median peak weight, from 0.06 to 0.03, and in the length of the upper tail, from 0.31 to 0.17. This demonstrates that an individual matching peak becomes less significant as the accuracy of the structure drops. In each case there are also a number of outliers at the upper end of the distribution, but the number of matching peaks with weighting factors above 0.5 also decreases as the $\mathrm{RMSD}_{\mathrm{ref}}$ value rises. As the structural models differ more and more strongly from the reference structure and thus also from the picked peaks, fewer predicted long-range peaks will find a match within the list of experimental peaks. It follows that the percentage of short-range peaks within the group of matches will increase. Short-range interactions generally have a lower information content than long-range peaks, and as the information content of the distance restraint contributes to the weighting factor of a matching peak the falling number of matching long-range peaks accounts for the drop in the weights of matching peaks with increasing $\mathrm{RMSD}_{\mathrm{ref}}$.

The set of the weight distributions of the missing peaks, in contrast, is very inhomogeneous. In most cases the median peak weight by far exceeds the median weight of the other two peak categories, and the inter-quartile range generally extends much further as well, beyond the ends of the tails of the weight distributions of the other classes. Since the weighting factors of missing peaks are calculated in the same way as those of matching peaks, i.e. they also take into account the information content of the peak, and since long-range peaks comprise a large fraction of the missing category, these peaks will on average have high weighting factors. In this case, however, a number of matching peaks corresponded to even longer distances and, accordingly, obtained larger information content values, which explains why there are outliers at high weights for the matching category.



(a) Overall distribution of F for each of the fsh2 test structures validated against the ^{15}N-NOESY spectrum.

(b) Distribution of the F values within each of the three peak classes, unpredicted (green), matching (teal), and missing (violet).

Figure III.2.6: The distribution of the contributions to ζ (F, eq. II.2.15) obtained during the validation of several different structural models of fsh2 against the ^{15}N-NOESY spectrum of the protein. The description of this type of plot can be found with fig. III.2.4.

The effect of the combination of error function values and weighting factors is illustrated in fig. III.2.6, which presents the distribution of contributions to ζ (F) for the fsh2-test set. Once again, there are marked differences between the three peak categories, and also between the different structure bundles, but no clear trends are observable with increasing $\mathrm{RMSD}_{\mathrm{ref}}$.

The median F values are relatively constant across the group of ensembles, as are the ends of the upper tails of the distributions. This can be seen in fig. III.2.6a. In each case there are many outliers, but only comparatively few F values exceed 0.3. Nevertheless, even few such peaks can have a strong influence on the final value of ζ since their contribution may equal the combined contributions of several dozen other peaks. Once more, the $\mathrm{RMSD}_{\mathrm{ref}}$ value of the individual test structure does not appear to influence the shape of the distribution.

Missing peaks, as was predicted from their comparatively high distance deviations and their oftentimes relatively large weighting factors, tend to yield large values of F as well, with their median contribution often significantly surpassing those of the other two peak categories.

The median F value of the matching peaks is 0.00 in each case. Even though the median weight is greater than 0.00 at least half of all matching peaks do not contribute to the numerator during the calculation of ζ as there is threshold filtering in place. This means that matching peaks whose distance deviations fall below said threshold are assigned an F value of 0, irrespective of their individual weighting factors. The upper tails from the matching peaks' distribution extend up to values between 0.006 and 0.010. Once more there are a number of outliers, with higher values of F becoming slightly more common as $\mathrm{RMSD}_{\mathrm{ref}}$ increases. Nevertheless, at least 50 % of all matched peaks contribute nothing to the numerator of ζ, even though some of them may contribute a non-zero value to its denominator. This causes ζ to be most strongly influenced by the classes of unpredicted and missing peaks, an effect that would be the same even if the weighting factors of the matching peaks were somewhat higher, and if threshold filtering were dropped, since matches generally give rise to low penalty function values.

Fig. III.2.7 shows the relative contributions of the three different peak categories to a number of quantities influencing the final value of ζ. The contributions of the unpredicted peaks visibly increase as the degree of accuracy of the structures goes down, and the influence of matching peaks drops due to both their numbers and also their weights declining as $\mathrm{RMSD}_{\mathrm{ref}}$ rises.

Even though matched peaks contribute more strongly to the set of peak weights than what would be predicted based on their numbers alone their fraction of F values is significantly lower, which is explained by their comparatively low error function values. Missing peaks, on the other hand, contribute much relative to their low numbers, which is mainly caused by their large weighting factors (fig. III.2.5b).

The figures within the current section illustrate the relevance of distinguishing the different peak categories. Between the three classes there are clear differences in numbers, and certain subgroups of peaks are disproportionately strongly represented in a certain category (many missing peaks are long-range peaks, for example). Furthermore, ζ would be negatively affected if not all peak classes were considered during the final calculation, as section III.2.4 demonstrates.

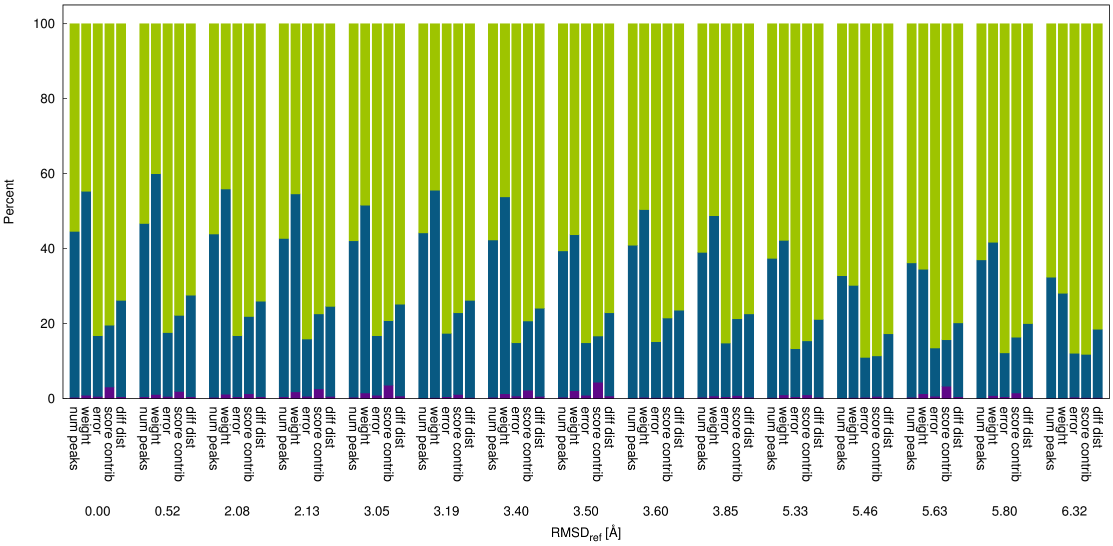


Figure III.2.7: Distribution of various score-influencing quantities for matching (teal), missing (violet), and unpredicted peaks (green) for the fsh2 test set after validation against the ^{15}N-NOESY spectrum of the protein. Each bar shows the relative contributions from the peaks of the three categories to the total of the quantity given on the horizontal axis. *num peaks* is the number of peaks used to calculate ζ, *weight* is the weighting factor w, *error* stands for the penalty function value, *score contrib* denotes the F value, and *diff dist* is the distance deviation ($|\Delta d|$).

III.2.4 Alternative approaches to the calculation of ζ

Section II.2.1.6 outlines a number of alternative approaches to the calculation of ζ that were tested during the development of CYVAL. The current section gives an overview of the results obtained using those methods. Tables listing the correlation coefficients of the validation metrics and structural accuracy for all assessed methods are provided in section B of the appendix.

The results suggest that the combination of error function and weighting factors has comparatively little impact on the ability of ζ and ζ_{n} to differentiate between structures at different levels of accuracy. The choice of peak selection mode, on the other hand, can drastically affect the validation power of CYVAL.

III.2.4.1 Error functions

A number of functions for computing the distance deviation penalties of the individual peaks were evaluated. The functions are described in section II.2.1.6 and their graphs are shown in fig. II.2.5.

Not all error functions perform equally well, as can be seen from the correlation coefficients between RMSD$_{\text{ref}}$ and ζ that are listed in tabs. B.1 to B.5. The steep saturation function `sat3` generally yields results of a similar quality to those obtained using the default error function `sat1` (eq. II.2.16). `sat2`, which has a slope similar to that of `sat1` but whose graph is significantly flatter in the $|\Delta d|$-region below 1.0 Å, is less able to distinguish between different levels of structural accuracy. Nevertheless, its performance is not drastically poorer than that of `sat1` or `sat3`.

pol4 and pol6, the assessed polynomial functions, perform similarly well as the saturation functions, with the results obtained using the protein rhod being a notable example of both of them, but pol6 in particular, being outperformed by the default error function. Here, their results are especially poor for the validation against the ^{13}C-NOESY spectrum, and it is using this type of spectrum, too, that their correlation coefficients are by 0.1 lower for the protein fsh2.

The logarithmic function log4, which yields higher distance deviation penalties than all other assessed functions up to around 1.0 Å and continues to do so over then entire $|\Delta d|$-range when compared to the two linear functions and to pol6, performs about as well as the other assessed functions.

The linear functions lin4 and lin6 yield results that, again, are roughly on a par with the default error function, and with each other, although lin6 appears to be somewhat better suited to the task. In some cases no difference in correlation coefficient can be observed between the two, but on the whole combining lin6 with trad, the traditional way of weight calculation, yielded nine test runs with correlation coefficients of ζ vs. $\mathrm{RMSD}_{\mathrm{ref}}$ above 0.75, three out of which are equal to or greater than 0.90 (four in case the correlation between ζ_{n} and $\mathrm{RMSD}_{\mathrm{ref}}$ is focused on). Going by the numbers only, this makes lin6 the most successful error function.

On the whole, however, the results show that the choice of error function does not appear to have a significant and general influence on the validation power of ζ and ζ_{n}.

III.2.4.2 Weighting factors

From tabs. B.1 to B.5 it can be gathered that the choice of weighting factor, like the choice of error function, is not the one crucial factor determining the quality of the CYVAL results. The two proteins that most easily lend themselves to validation using the traditional method, HR8254a and fsh2, yielded results of a similar quality when different weighting factors were tested. The protein most severely challenging CYVAL, enth, exhibited strong fluctuations in ζ-$\mathrm{RMSD}_{\mathrm{ref}}$ correlation coefficients across the calculations in which various alternative weighting factors were assessed.

A detailed analysis of the results suggests no_info to be the most promising weight mode, as it tends to produce slightly better results than the other methods in a number of cases. In contrast to the traditional approach, no_info does not include information content data. The overall differences between these results and those obtained using the traditional weight mode, however, are neither pronounced nor utterly conclusive. Nevertheless, no_info was used when an alternative combination of error function and weight mode was tested. This is described in the following section.

III.2.4.3 Combination of `lin6` and `no_info`

The results shown in tabs. B.1 to B.5 suggest that combining error function `lin6` and weight mode `no_info` may enhance in the validation power of CYVAL. The combination of `no_info` with the the traditional error function `sat1` produced a total of eight test validation runs with ζ-RMSD$_{\mathrm{ref}}$-correlation coefficients above 0.75, two out of which are at or above 0.90. This result is superior to the combination of `sat1` with any other weight mode. Likewise, combining `lin6` with the traditional way of weight calculation `trad` yielded nine test runs with correlation coefficients above 0.75, three out of which are equal to or greater than 0.90. Consequently, as the tendencies were the same when looking at ζ_{n}, `lin6` and `no_info` were combined to test whether this approach would increase the validation power of CYVAL.

Fig. III.2.8 compares the ζ_{n} correlation results, which according to tab. III.2.1 are slightly superior to those determined for ζ, obtained for the fsh2 and HR8254a test sets. The combinations shown are the traditional approach using `sat1` and the default weight mode `trad`, the combination of `lin6` and `trad`, and the combination of `lin6` and `no_info`, the latter method tending to yield the best results when judged by the ζ_{n}-RMSD$_{\mathrm{ref}}$ correlations of the test sets. Evaluating the individual plots, however, shows that the results are inconclusive as to which combination ought to be preferred. In all cases were there instances where ζ_{n} was able to clearly distinguish between high, medium, and low-quality structures, but each method produced outliers as well, and the same holds true when plots of ζ are examined (data not shown).

Since no method has proven to be clearly superior to the other combinations of error function and weight mode it was decided to retain the traditional approach to the computation of ζ and ζ_{n}, as the default way of weight calculation incorporates information about peak redundancy via the information content ι. Even though excluding this metric appears to have raised ζ-RMSD$_{\mathrm{ref}}$ correlation in some cases it cannot be concluded that ι lowers the validation power of CYVAL—the plots of fig. III.2.8 show that pure numbers can be deceptive, especially in cases, such as this one, where no ideal correlation coefficient of 1.0 can be expected. Moreover, it is not clear whether other query proteins would not benefit from an inclusion of the additional information provided by ι in the validation process.

III.2.4.4 Peak selection

Tabs. B.1 to B.5 list the correlation coefficients of ζ and ζ_{n} versus RMSD$_{\mathrm{ref}}$ for different validation methods. In general, the results obtained from the default method of calculating ζ were as good as or superior to those from other approaches, and even in those individual cases where another approach produced a stronger correlation between RMSD$_{\mathrm{ref}}$ and ζ or ζ_{n} was there no clear trend of a particular alternative method being generally preferable to the traditional approach.

Table III.2.1: Correlation coefficients for $\mathrm{RMSD_{ref}}$ vs. ζ and ζ_n (in parentheses) for three calculation approaches.

^{13}C-NOESY

Protein	**traditional**[a]	**lin6 and no_info**[b]	**lin6 and trad**[c]
enth	0.31 (0.44)	0.45 (0.38)	0.28 (0.45)
HR8254a	0.91 (0.92)	0.84 (0.85)	0.90 (0.91)
rhod	0.75 (0.77)	0.78 (0.80)	0.75 (0.78)
fsh2	0.83 (0.92)	0.93 (0.94)	0.78 (0.93)
StT322	0.70 (0.69)	0.74 (0.74)	0.70 (0.69)

^{15}N-NOESY

Protein	**traditional**	**lin6 and no_info**	**lin6 and trad**[c]
enth	0.10 (0.10)	0.25 (0.25)	0.19 (0.20)
HR8254a	0.84 (0.84)	0.90 (0.90)	0.85 (0.85)
rhod	0.78 (0.79)	0.81 (0.81)	0.82 (0.83)
fsh2	0.93 (0.93)	0.90 (0.90)	0.93 (0.93)
StT322	0.56 (0.56)	0.78 (0.78)	0.77 (0.77)

^{13}C$_\mathrm{aro}$-NOESY

Protein	**traditional**	**lin6 and no_info**	**lin6 and trad**[c]
HR8254a	0.89 (0.89)	0.89 (0.89)	0.90 (0.90)
rhod	0.62 (0.62)	0.66 (0.67)	0.75 (0.77)
StT322	0.51 (0.54)	0.43 (0.47)	0.50 (0.54)

[a] Using sat1 as error function and the traditional weight mode trad.
[b] Using lin6 as error function and the no_info weight mode.
[c] Using lin6 as error function and the trad weight mode.

For three out of the five proteins (HR8254a, fsh2, and rhod) the default approach worked as well or even better if the experimental peaks stemmed from automated peak picking performed by CYANA, instead of them being taken from curated peak lists. In the case of the ^{13}C-NOESY spectrum of fsh2, for example, the poor validation performance upon using a manually refined instead of an automatically picked list of experimental peaks led to a disproportionately large number of missing peaks, whose influence on ζ proved to be too dominant. It appears that during the manual refinement of the peak list in question peaks had been removed—or remained unpicked in case peak picking had also been performed manually—that were actually in accordance with the structure.

In a number of cases discarding peaks based on peak class overlap with a random structure led to a significant deterioration in the ability of ζ to distinguish between correct and incorrect structures. In three cases it even caused a reversal of the sign of the correlation coefficient, i.e. a lower degree of accuracy would now lead to a lower ζ-value, which runs contrary to what would be expected based on the definition of ζ. On the other hand, in each of these cases the absolute values of the negative correlation coefficients were smaller than 0.5, i.e. they cannot point to an actual trend in the value of ζ in relation to $\mathrm{RMSD_{ref}}$.

Low-quality structures were disproportionately strongly affected by peak filtering based on a random structure. Consequently, ζ, which contains the sum of individual contributions in its numerator, roughly decreased with increasing $\mathrm{RMSD}_{\mathrm{ref}}$. This explains the abovementioned reversal of the sign of the correlation coefficient. The fact that ζ_{n} did not reproduce this behaviour and generally exhibited a significantly better correlation result than ζ strengthens this explanation. In the instances where ζ_{n} exhibited a comparable or greater validation power after peak filtering than the 'traditional' ζ_{n} value, visual inspection of the results revealed them not to be a significant improvement on the traditional approach. This renders the additional computing effort required for peak filtering unnecessary.

The abovementioned effect of a strong decrease in the numbers of used peaks with rising $\mathrm{RMSD}_{\mathrm{ref}}$ also explains why neither the approach of taking into account only matching peaks nor that of only considering long-range matches led to an increase in the validation power of CYVAL, and why in these cases some instances of negative correlation coefficients were observed as well. There were only four cases in which using only matching peaks produced a correlation coefficient well above 0.5 for ζ and/or ζ_{n}: validation against the ^{13}C-NOESY spectra of enth and fsh2, the ^{15}N-NOESY spectrum of fsh2, and the ^{15}N-NOESY spectrum of HR8254a. When exclusively long-range peaks were considered only validation against HR8254a's ^{13}C-NOESY, StT322's and rhod's ^{15}N-NOESY, and both fsh2 spectra yielded meaningful values of ζ_{n}.

In the case of enth the exclusive use of matching peaks led to a significant improvement of the CYVAL validation power when compared to the results of the default approach. For the ^{13}C-NOESY spectrum this improvement was substantial enough (from a correlation coefficient of 0.31 to 0.83) to lead to the conclusion that CYVAL was now able to differentiate between enth structures of different degrees of accuracy. In this case it appears that the presence of unpredicted and missing peaks in the calculation of ζ served to obscure the truly meaningful data points. With most other combinations of protein and spectrum type, on the other hand, the matches only-result was inferior to that obtained from the default approach. It seems that, on the whole, both unpredicted and missing peaks contribute valuable information to the validation procedure that ought not to be discarded.

III.2.5 Comparison of CYVAL with other validation tools

III.2.5.1 Knowledge-based tools

Fig. III.2.9 illustrates the validation power of selected knowledge-based validation scores that aim to assess how well a query structure fits certain criteria of structural normality (see section I.5.4.1). The Verify3D score, the MolProbity score, and the zp-comb score from the ProSa2003 software were plotted against the $\mathrm{RMSD}_{\mathrm{ref}}$ values of the fsh2 and HR8254a test sets, and a comparison with figs. III.2.1 and III.2.2 shows that the validation power of ζ and ζ_n is comparable or even superior to that of the other three tools. Since ζ_n performed slightly better than ζ this section will base its detailed comparison on the former metric.

In the case of fsh2, Verify3D yielded similar score values over the $\mathrm{RMSD_{ref}}$-range between 0.5 and 3.0 Å, whereas there was an increase in ζ_n over the same range. Furthermore, the large spread in the values of the Verify3D score in the $\mathrm{RMSD_{ref}}$-range from 5.5 to 6.5 Å led to one of the structure bundles located there receiving a similar score value as structures at around 3.0 Å. This was clearly not the case for ζ_n. The effect of Verify3D score similarity was also observed for HR8254a, where the score value was inconclusive over the $\mathrm{RMSD_{ref}}$-range from 1.0 to 4.5 Å; again, ζ_n did not replicate this effect but displayed a clear difference between structures below and above 2.5 Å $\mathrm{RMSD_{ref}}$.

MolProbity, like CYVAL, yielded more conclusive score values for fsh2 than for HR8254a. Over the fsh2 $\mathrm{RMSD_{ref}}$-range from 2.0 to 4.0 Å MolProbity produced a somewhat clearer correlation between structural accuracy and validation metric than CYVAL, but on the whole both ζ_n and the MolProbity score were able to distinguish between high (below 2.0 Å $\mathrm{RMSD_{ref}}$), medium (between 2.0 and 4.0 Å), and low-quality (above 4.0 Å) structures of fsh2. For HR8254a, on the other hand, MolProbity performed somewhat better in terms of differentiating between medium and low-quality structures than CYVAL, but the latter nevertheless clearly distinguished between high and medium-quality structures, particularly so when the ^{13}C-NOESY spectrum of HR8254a was used for structure validation.

ProSa2003, too, was better at distinguishing between the three quality classes within the fsh2-test set than in the case of HR8254a. Even for fsh2, however, there were two cases in which scores from the set of medium-quality structures overlapped with what was obtained for the set of low-quality structures, an effect not observed for CYVAL. In the case of HR8254a ProSa2003 was unable to clearly differentiate between low and medium-quality structures, and the zp-comb score performed significantly worse than ζ_n.

None of the knowledge-based scores allowed direct conclusions about the $\mathrm{RMSD_{ref}}$ value as all measures yielded rather different values for different proteins at the same degree of accuracy. However, the assessed tools aim to estimate how protein-like the structural model is in terms of criteria derived from known protein structures. Therefore these programs cannot necessarily be expected to be able to differentiate between accurate and inaccurate structures. This fact, however, does not render their validation scores as useful to the user community as a metric that accomplishes this differentiation.

Assuming that the three knowledge-based scores discussed here do indeed provide information about the quality of the query structure in terms of 'protein-ness' the structural models of the test sets that exhibit a low degree of accuracy also appear to be less 'normal' than structures whose $\mathrm{RMSD_{ref}}$ values are comparatively low. Besides, the qualities of the individual conformers seem to fluctuate considerably, particularly those within the HR8254a-test set, as may be concluded from the large spread of the error bars in fig. III.2.9.

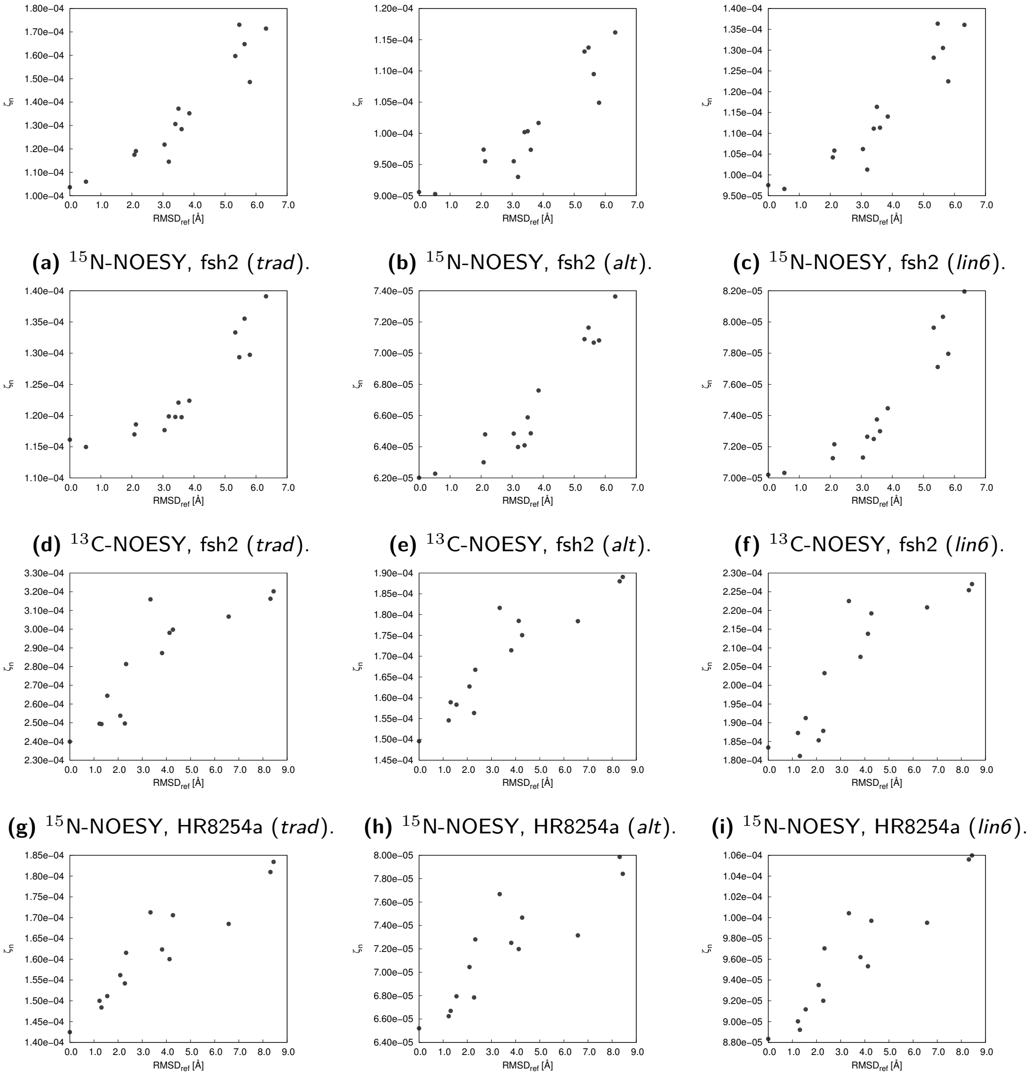
Table III.2.2: Correlation coefficients between the median values of the three validation scores and the $\mathrm{RMSD}_{\mathrm{ref}}$ values of the protein test set members.

Protein	MolProbity score	Verify3D score	zp-comb
enth	0.60	-0.84	0.69
fsh2	0.91	-0.80	0.89
HR8254a	0.83	-0.82	0.94
rhod	0.61	-0.81	0.71
StT322	0.40	-0.36	0.44

The assessed knowledge-based programs can only process single structures, therefore the conformers of each test ensemble were analyzed separately. The median value of each validation metric was then determined for the individual structure bundles. Tab. III.2.2 lists the correlation coefficients obtained for these median values and the $\mathrm{RMSD}_{\mathrm{ref}}$ values of the assessed ensembles. Interestingly, as with CYVAL, the validation attempts for the fsh2 and HR8254a test sets were the most successful. This suggests that the HR8254a and fsh2 test sets exhibit certain features that facilitate the assessment of structural accuracy. As far as the performance of the external validation programs is concerned, these features appear to be lacking in the case of StT322 in particular. There, no absolute correlation values above 0.45 were observed. CYVAL, in contrast, generated promising validation results for this protein, with a maximum ζ-$\mathrm{RMSD}_{\mathrm{ref}}$-correlation coefficient of 0.70 (see tab. B.5).

III.2.5.2 DP-score

Due to the server errors encountered upon all attempts to use RPF via `http://nmr.cabm.rutgers.edu/rpf` it proved impossible to compare the results of CYVAL to the DP-score values of our the protein tests.



(a) ^{15}N-NOESY, fsh2 (*trad*).

(b) ^{15}N-NOESY, fsh2 (*alt*).

(c) ^{15}N-NOESY, fsh2 (*lin6*).

(d) ^{13}C-NOESY, fsh2 (*trad*).

(e) ^{13}C-NOESY, fsh2 (*alt*).

(f) ^{13}C-NOESY, fsh2 (*lin6*).

(g) ^{15}N-NOESY, HR8254a (*trad*).

(h) ^{15}N-NOESY, HR8254a (*alt*).

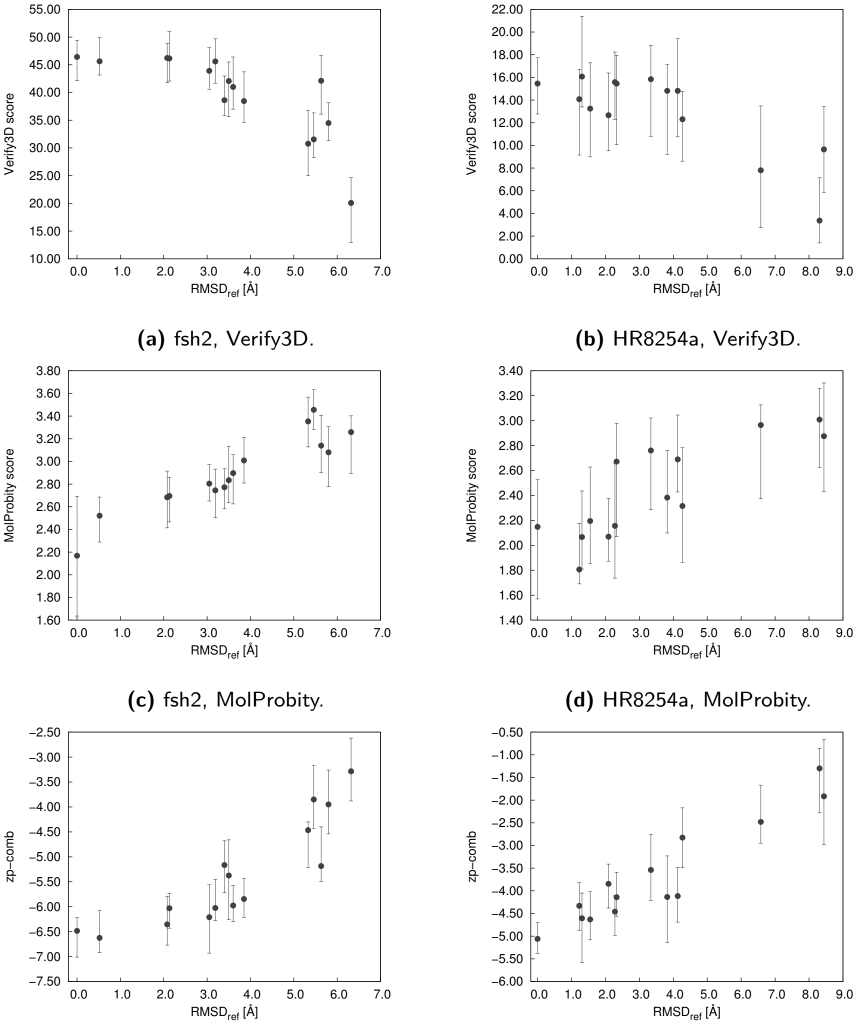
(i) ^{15}N-NOESY, HR8254a (*lin6*).

(j) ^{13}C-NOESY, HR8254a (*trad*).

(k) ^{13}C-NOESY, HR8254a (*alt*).

(l) ^{13}C-NOESY, HR8254a (*lin6*).

Figure III.2.8: Relationship between structure accuracy and ζ_n, calculated in the traditional manner, i.e. weight mode `trad` and error function `sat1` (*trad*, left column), using a combination of error function `lin6` and weight mode `no_info` (*alt*, middle column), and using the traditional weight mode `trad` in combination with error function `lin6` (*lin6*, right column). The examples show data from the proteins fsh2 and HR8254a.



(a) fsh2, Verify3D.

(b) HR8254a, Verify3D.

(c) fsh2, MolProbity.

(d) HR8254a, MolProbity.

(e) fsh2, ProSa2003.

(f) HR8254a, ProSa2003.

Figure III.2.9: Validation of the fsh2 and HR8254a test sets using the programs Verify3D, MolProbity, and ProSa2003 (zp-comb score). With both zp-comb and the MolProbity score a lower value indicates a better-quality structure, whereas the Verify3D score should increase with structure quality. In each sub-figure the median value of the validation score for this particular structure bundle is represented by a dot; the error bars extend to the minimum and maximum score values of the conformers making up this structure bundle.

Chapter III.3

Protein structure calculations

As described in chapter II.3 three protein structures were calculated from solution NMR data: TycC3_PCP(S45A) (?), the wild-type WW domain of Pin1, and the S16E-mutant of this domain ($\mathrm{WW}^{\mathrm{S16E}}$) (?).

The globular protein Pin1 is a peptidyl-prolyl cis/trans isomerase (PPI) regulated by phosphorylation. It interacts with a multitude of phosphorylated proteins and plays a role in various cellular processes such as growth-signal responses and cell-cycle progression (??). Pin1 consists of two domains, the catalytically active PPI domain and WW domain, which mediates protein-protein interaction and is responsible for the subcellular location of Pin1. Phosphorylation of the WW domain at Ser16 prevents Pin1 from binding to its substrates, e.g. cell cycle progression proteins, and it also blocks the ability of the WW domain to determine the cellular localization of Pin1 (??). High levels of Pin1 phosphorylated at Ser16 have been discovered in brain tissue samples of people suffering from Alzheimer's disease (?), and the protein is overexpressed in a number of human cancers (?), where it is known to promote tumour growth (?). In order to better understand the effects of Pin1-phosphorylation at Ser16 the structure of the wild-type WW domain and of its phospho-mimic S16E-mutant were determined by liquid-state NMR.

TycC3_PCP is the third peptidyl carrier domain of the tyrocidine A synthetase subunit C from *Bacillus brevis*, hence it is a component of a nonribosomal peptide synthetase (NRPS). In general, NRPSs are large protein complexes found in microorganisms, where they synthesize functionally diverse products, some of which have important pharmacological properties. Antibiotics and immunosuppresive compounds, amongst others, can be found amongst the nonribosomally produced peptides (??). Carrier proteins such as TycC3_PCP are crucial components of various biosynthetic pathways (?), and within NRPSs peptidyl carrier proteins (PCPs) shuttle intermediates between the individual modules of the enzymatic complex. PCPs and PCP domains share the fold of the acyl carrier protein or domain of fatty acid synthase, which is conserved from bacteria to humans (?): a four-helix bundle and a conserved serine residue near the N-terminus of the second helix, to which a 4'-phosphopantetheine cofactor is attached in the holo state of the PCP. The presence of this cofactor is a prerequisite for the functionality of the PCP as the amino acid substrate, which the PCP passes to the reactive sites, is covalently bound to it (**?**).

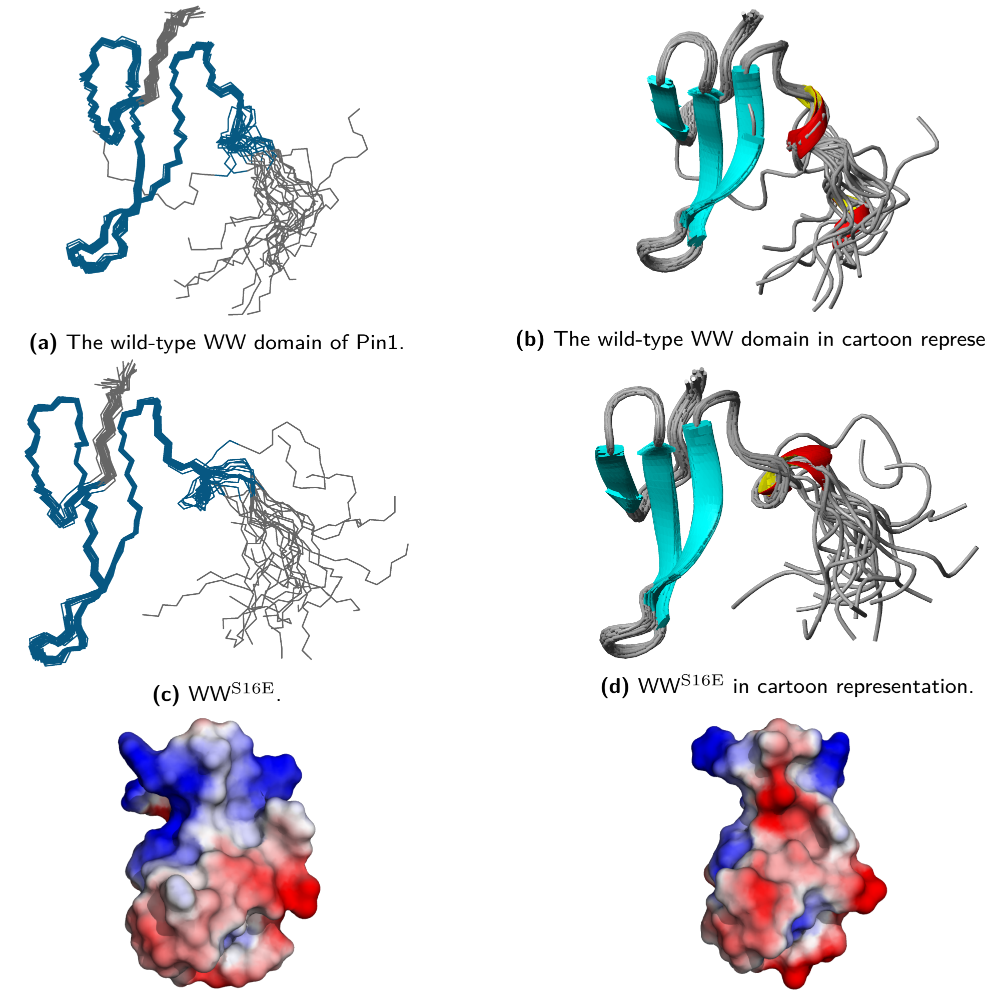
For TycC3_PCP the existence of different conformers was reported (**?**). In its holo state the characteristic four helices are present, which is known as the A/H state. In the absence of the bound cofactor, however, the existence of another conformation was postulated as well. According to **?**, the structure of apo-TycC3_PCP may become more flexible, with the third helix becoming disordered and all other helices being shortened. This was referred to as the A state. However, the current study, which aimed to elucidate the structural basis for carrier protein posttranslational modification, has cast doubt on these findings.

III.3.1 Pin1

Fig. III.3.1 presents the structure bundles of the wild-type and the S16E-mutant WW domain of Pin1. The backbone folds of the two structures are highly similar, which shows that $\mathrm{WW^{S16E}}$ retains a globular three β-strand fold, and that no major conformational rearrangement of the backbone is induced by the introduction of the negative charge into the molecule. However, **?** note that the surface charges have changed from the wild-type to the mutant WW domain (figs. III.3.1e and III.3.1f): the wild-type active site is characterized by the presence of positively charged arginine residues (**?**), an environment that is disturbed by the introduction of the negative charge. Additionally, due to the presence of Glu there is now a steric hindrance to access to the binding pocket. Alongside the results of the remaining experiments performed by **?** the comparison of the two structures presents an explanation for Pin1 being rendered inactive through phosphorylation at Ser16.

The NMR data obtained from the structure determination and the structural statistics of the wild-type WW domain and its S16E-mutant are given in tab. III.3.3. More cross-peak data were available for $\mathrm{WW^{S16E}}$ than for the wild-type WW domain, and a somewhat higher percentage of the peaks could be assigned as well. As stated in section II.3.1 the chemical shift assignment of $\mathrm{WW^{S16E}}$ was more complete by 2 percentage points than in the case of the wild-type domain, which explains the slightly lower percentage of unassigned peaks for $\mathrm{WW^{S16E}}$. In total, 936 NOE distance restraints contributed to the $\mathrm{WW^{S16E}}$ and 703 NOE distance restraints were used for the wild-type WW domain structure. In both cases more than 35 % of these were long-range restraints, which are particularly important for the determination of the overall fold.

According to the CYANA target function the structure of $\mathrm{WW^{S16E}}$ fits its data slightly better than the wild-type structure does. However, even the final target function value of the latter only marginally exceeds 1.0 Å^2, which ideally it should fall below. Both structures exhibit no restraint violations greater than 0.2 Å or 5.0°, and all residues fall into allowed regions of the Ramachandran



(a) The wild-type WW domain of Pin1.

(b) The wild-type WW domain in cartoon representation.

(c) WW^{S16E}.

(d) WW^{S16E} in cartoon representation.

(e) Surface charges of the wild-type WW domain.

(f) Surface charges of WW^{S16E}.

Figure III.3.1: Structures of the wild-type WW domain of Pin1 (PDB ID 2m8i) and of WW^{S16E} (PDB ID 2m8j) after energy refinement with OPALp. Subfigures **a** to **d** show the backbone fold only. In subfigures **a** and **c** the ordered residues (7–39) of the conformers, as determined by CYRANGE, have been superimposed and are shown in teal. Subfigures **e** and **f** show the surface charges of the two proteins (red: negative, blue: positive). The mutation of Ser16 to Glu has introduced an additional negative charge into WW^{S16E}, significantly altering the surface charge distribution.

plot. The precision of the ordered residues of the two structure bundles, measured by the RMSD to the mean structure, is virtually identical and differs by 0.02 Å only, with the WW^{S16E} ensemble being marginally more precise.

The structures of the wild-type WW domain and of WW^{S16E} were subjected to a quality assessment with ProSa2003 (using the zp-comb score), Verify3D, and MolProbity. The median validation scores and the score ranges for the bundles are listed in tab. III.3.1. The evaluation by Verify3D and MolProbity suggests that the quality of the WW^{S16E}-structure is somewhat higher than that of the wild-type structural model, as a higher Verify3D score indicates a better agreement between the sequence and the three-dimensional structure, and a lower MolProbity score points to a better fit to geometric criteria. The zp-comb value from ProSa2003, on the other hand, is lower for the wild-type WW domain than for $\mathrm{WW}^{\mathrm{S16E}}$, which means that this program considers the wild-type structural model to have an energetically more favourable conformation.

Table III.3.1: Knowledge-based validation scores obtained for the wild-type WW domain and for $\mathrm{WW}^{\mathrm{S16E}}$. The median values are given, and the score ranges from all conformers are shown in parentheses.

Score	**wild-type**	$\mathbf{WW^{S16E}}$
MolProbity score	2.43 (1.82 – 3.00)	1.54 (0.84 – 2.46)
Verify3D score	7.62 (3.55 – 11.74)	10.29 (7.91 – 12.41)
zp-comb	-3.42 (-4.60 – -2.43)	-2.70 (-3.19 – -1.60)

Additionally, the structural models were subjected to analysis by CYVAL, and the obtained values of ζ and ζ_{n} are reported in tab. III.3.2. Both metrics are lower for the wild-type WW domain than for $\mathrm{WW}^{\mathrm{S16E}}$ for each of the two available spectra types. Since lower values indicate a better agreement between the structure and the experimental data the structural model of the wild-type protein appears to be slightly more accurate than that of the mutant domain.

Table III.3.2: ζ and ζ_{n} (in parentheses) obtained for the wild-type WW domain and for $\mathrm{WW}^{\mathrm{S16E}}$.

Spectrum	**wild-type**	$\mathbf{WW^{S16E}}$
^{15}N-NOESY	0.24 (4.31×10^{-4})	0.36 (4.67×10^{-4})
^{13}C-NOESY	0.53 (2.13×10^{-4})	0.63 (2.58×10^{-4})

III.3.2 TycC3_PCP(S45A)

TycC3_PCP(S45A) mimics the apo state of TycC3_PCP, the difference being that the mutant cannot be loaded with the 4'-phosphopantetheine cofactor. The structure of TycC3_PCP(S45A) is shown in fig. III.3.2. The four helix-bundle, which is characteristic of PCPs and thus also of the A/H state of TycC3_PCP, is present even in the absence of the cofactor. This, together with the structure of a complex of TycC3_PCP(S45A) with the group II phosphopantetheine transferase Sfp, which **?** solved using X-ray crystallography (fig. III.3.2c), reveals that apo-PCP does not enter the A state but exists in the A/H state, both in solution and within a complex. Hence the performed study has helped correct previous results.

The medical usefulness of various nonribosomally synthesized peptides and the modular structure of the NRPSs that produce them makes these large protein complexes interesting targets for modifications aiming to lead to the production of novel pharmacologically active compounds. Naturally, a detailed understanding of the biosynthetic mechanisms employed by NRPSs is crucial for successfully carrying out such endeavours. It follows that the structure of TycC3_PCP (represented by TycC3_PCP(S45A)), being an NRPS component, can contribute to showing how PCPs function in these environments and interact with other proteins.

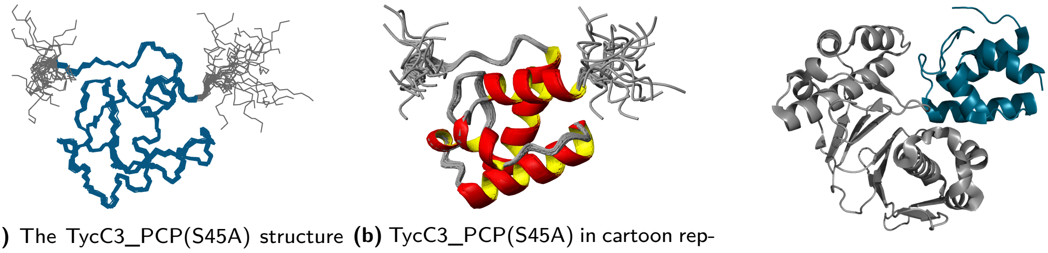
Table III.3.3: NMR structure determination statistics for the wild-type WW domain of Pin1 and its S16E-mutant WW^{S16E}.

	wild-type		**WW^{S16E}**	
NOE assignment[a]				
Total number of NOESY cross-peaks	1428		1920	
Assigned cross-peaks	1291	90 %	1778	93 %
in ^{13}C-resolved NOESY	768	97 %[b]	1194	96 %[b]
in ^{15}N-resolved NOESY	523	82 %[b]	584	87 %[b]
Unassigned peaks	137	10 %	142	7 %
Conformational restraints				
Total NOE distance restraints	703		936	
short range $\|i-j\| \leq 1$	352	50 %	445	48 %
medium range $1 < \|i-j\| < 5$	100	14 %	115	12 %
long range $\|i-j\| \geq 5$	251	36 %	376	40 %
Dihedral angle restraints (ψ/ϕ)	58		56	
Structure statistics[c]				
Average CYANA target function (Å^2)		1.02 ± 0.25		0.55 ± 0.10
AMBER energies (kcal/mol)		-1590 ± 68		-1717 ± 71
Restraint violations[d]				
Max. distance restraint violation (Å)		0.12		0.11
Violated distance restraints > 0.2 Å		0		0
Max. dihedral angle restraint violation (°)		4.83		2.14
Violated dihedral angles > 5°		0		0
Ramachandran plot				
Residues in most favoured regions		71.4 %		78.8 %
Residues in additionally allowerd regions		28.6 %		21.0 %
Residues in generously allowed regions		0.0 %		0.3 %
Residues in disallowed regions		0.0 %		0.0 %
RMSD (residues 7–39)				
Average backbone RMSD to mean (Å)		0.28 ± 0.06		0.26 ± 0.04
Average heavy atom RMSD to mean (Å)		0.82 ± 0.11		0.80 ± 0.09

[a] obtained from the automated NOE assignment and structure calculation functionalities of CYANA.
[b] percentage given relative to the total number of peaks in the respective peak list.
[c] after restrained energy minimization with OPALp.
[d] after energy minimization, calculated with CYANA.

Besides, it had previously been suggested that phosphopantetheine transferases (PPTs) of pathogenic bacteria could be used as targets for the development of novel antibiotics (**??**). However, **?** note a high structural similarity between the TycC3_PCP(S45A)/Sfp crystal structure (fig. III.3.2c) and the complex between human acyl carrier protein and PPT. This suggests both a conserved mechanism of carrier protein posttranslational modification in group II PPTs from various organisms and similarities in carrier protein recognition. These two factors, in turn, indicate a potential for complications when aiming to develop antibiotics targeting bacterial PPTs.

The structural model of TycC3_PCP(S45A), too, was assessed by ProSa2003 (using the zp-comb score), Verify3D, MolProbity, and CYVAL. The obtained validation results are summarized



(a) The TycC3_PCP(S45A) structure bundle.

(b) TycC3_PCP(S45A) in cartoon representation.

(c) TycC3_PCP(S45A) and Sfp.

Figure III.3.2: **a** and **b**: NMR structure of TycC3_PCP(S45A) (PDB ID 2md9) after energy refinement with OPALp. In **a** the ordered residues (7–84), as determined by CYRANGE, have been superimposed and are shown in teal. **c**: Crystal structure (PDB ID 2mrt; **?**) of TycC3_PCP(S45A) (teal) with Sfp (grey) and coenzyme A (not shown) at 2.0 Å resolution. A comparison of **b** and **c** shows that TycC3_PCP(S45A) occurs in the A/H state both in solution and when interacting with a binding partner.

in tab. III.3.4. A rather low MolProbity score points to a high geometric quality of the structure, and both the relatively high Verify3D score and the low value of zp-comb indicate that the structural model is favourably classified by these two programs. As noted in section I.5.4.1 the Verify3D score depends on the size of the protein. TycC3_PCP(S45A) consists of 90 residues, compared to 43 residues for the WW domain of Pin1. This explains why the Verify3D score of TycC3_PCP(S45A) is considerably higher than the scores of both the wild-type WW domain and $\mathrm{WW}^{\mathrm{S16E}}$. The low values of ζ and ζ_n indicate that the structural model of TycC3_PCP(S45A) does not only agree well with geometric criteria, which the knowledge-based validation scores testify to, but that there is also a high correspondence between the structure and the experimental data.

Table III.3.4: Validation scores obtained for TycC3_PCP(S45A).

Score	Value[a]	Score range
MolProbity score	1.91	1.43 – 2.22
Verify3D score	26.37	21.78 – 29.68
zp-comb	-5.33	-5.60 – -5.05
ζ (ζ_n)[b]	0.19 (1.05×10^{-4})	–

[a] For the knowledge-based validation scores (MolProbity score, Verify3D score, zp-comb) the median value is reported, as each conformer of the ensemble had to be evaluated separately. CYVAL analyzes the entire structure bundle at once and outputs only one single value for ζ and ζ_n each.

[b] Obtained from the validation of the structural model against a ^{15}N-NOESY spectrum of the protein.

Tab. III.3.5 lists the relevant NMR statistics and structural data from the structure calculation of TycC3_PCP(S45A). 94 % of all NOESY cross-peaks could be assigned. From this, 1419 NOE distance restraints were derived, 21 % of which were long-range restraints. As noted above, these are crucial for determining the overall structure of the protein. As in the case of the wild-type WW domain of Pin1 the final average CYANA target function only marginally exceeded 1.0 Å^2, and again there were no distance restraint violations above 0.2 Å, and no dihedral angle restraint violations above 5°. All residues are located in allowed regions of the Ramachandran plot, and 90 % lie within the most favoured regions. In its ordered residues the structure bundle is very precise, with a backbone RMSD value to the mean coordinates of 0.27 Å.

Table III.3.5: NMR structure determination statistics for TycC3_PCP(S45A).

NOE assignment[a]		
Total number of NOESY cross-peaks	1693	
Assigned cross-peaks	1584	94 %
in $^{13}C_{aro}$-resolved NOESY	302	95 %[b]
in ^{15}N-resolved NOESY	1282	93 %[b]
Unassigned peaks	109	6 %
Conformational restraints		
Total NOE distance restraints	1419	
short range $\vert i - j\vert \leq 1$	661	47 %
medium range $1 < \vert i - j\vert < 5$	457	32 %
long range $\vert i - j\vert \geq 5$	301	21 %
Dihedral angle restraints (ψ/ϕ)	140	
Structure statistics[c]		
Average CYANA target function (Å^2)		1.19 ± 0.18
AMBER energies (kcal/mol)		-3192 ± 59
Restraint violations[d]		
Max. distance restraint violation (Å)		0.13
Violated distance restraints > 0.2 Å		0
Max. dihedral angle restraint violation (°)		3.63
Violated dihedral angles > 5°		0
Ramachandran plot		
Residues in most favoured regions		89.9 %
Residues in additionally allowerd regions		10.1 %
Residues in generously allowed regions		0.1 %
Residues in disallowed regions		0.0 %
RMSD (residues 7–84)		
Average backbone RMSD to mean (Å)		0.27 ± 0.03
Average heavy atom RMSD to mean (Å)		0.67 ± 0.04

[a] obtained from the automated NOE assignment and structure calculation functionalities of CYANA.
[b] percentage given relative to the total number of peaks in the respective peak list.
[c] after restrained energy minimization with OPALp.
[d] after energy minimization, calculated with CYANA.

Chapter III.4

Conclusions

This thesis has focused on the analysis of the precision and accuracy of protein structures determined by NMR. In order to support scientists in evaluating the precision of a protein structure bundle we have developed CYRANGE (?), a tool that determines the RMSD-stable domains of an ensemble of conformers. For the assessment of the fit of a structural model to the experimental data it was derived from, which can be considered a measure of the accuracy of this model, we have created the CYVAL validation method. Additionally, the structures of the wild-type WW domain of Pin1 and its S16E-mutant (?) were calculated, as well as the structure of TycC3_PCP(S45A) (?).

All projects described in this text have aimed to advance the understanding of the function of proteins via the analysis of their structures. The structure determination projects, of course, have directly yielded information about the query proteins and their ways of interacting with their binding partners. The created software tools, too, contribute to advancing our knowledge of proteins and their modes of action since these programs support researchers in the analysis and evaluation of their findings.

III.4.1 CYRANGE

CYRANGE identifies the parts of a protein structure bundle that can be simultaneously superimposed with a low RMSD value. Hence the software provides structural biologists with a reliable automated and thus reproducible method to determine residue ranges for the measurement of structural precision, without information being obscured by the inclusion of ill-defined segments.

The efficacy of CYRANGE in correctly identifying ordered regions has been demonstrated for a large variety of protein structures including even protein complexes, generally loose structure bundles, and other challenging cases. Global structure superpositions based on the CYRANGE residue ranges allow a clear presentation of the structure, and unnecessary small gaps within the selected ranges are absent. In the majority of cases, the residue ranges from CYRANGE contain fewer gaps and cover considerably larger parts of the sequence than those from other methods without significantly increasing the RMSD values. CYRANGE thus provides an objective and automatic method for standardizing the choice of residue ranges for the superposition of protein structures, which requires no adaptations of parameters for the individual cases. Moreover, the software can easily be incorporated into structure analysis and validation packages, and through its website at `http://www.bpc.uni-frankfurt.de/cyrange.html` it is conveniently and freely accessible.

Since CYRANGE was made publicly available in 2011 the software has been widely used: the website has been accessed more than 500 times for online calculations, and CYRANGE has been officially recommended by the NMR-validation task force of the PDB as the tool of choice for determining the domains of a protein structure bundle (**?**). Furthermore, CYRANGE was used to determine ordered residue ranges in the analysis of the structural quality of 2013's CASD-NMR entries (**??**), and it has been included in a version of the structure validation program MolProbity (MolProbity-HTC) developed for handling large amounts of data (**?**).

Prior to the publication of CYRANGE, and despite the existence of applications with similar functionalities, the manual selection of residue ranges or the use of suboptimal criteria remained commonplace. The favourable reception of CYRANGE by the NMR community demonstrates that CYRANGE has succeeded in filling this gap.

III.4.2 CYVAL

The CYVAL approach, built into CYANA, validates a protein structure against the original NMR spectra that were used for structure determination. CYVAL takes a processed NOESY spectrum in one of three common file formats, a list of resonance assignments, and the protein structure bundle. A prerequisite for the actual validation step is the availability of experimental peak data, and in order to obtain this information the automated peak picking routine of CYANA is used. This means that CYVAL does not require manually generated or curated peak lists, which drastically reduces user intervention and the potential for unintentional bias of the validation result.

The experimental signals are matched to a set of expected peaks predicted from the query structure. This step leads to each signal being assigned to one of three categories: matching, unexpected, and missing peaks. All of these contribute to the final value of the validation score, ζ, and to its per peak-version (ζ_n), but the individual contribution of each signal depends on both its weighted distance deviation penalty and its class affiliation, the latter affecting the weight of the peak.

In order to evaluate the validation power of CYVAL the procedure was applied to a number of protein test sets, each of which contained structure bundles of varying accuracy and different

NOESY spectra types. Apart from one test set for which no relationship between ζ or ζ_{n} and $\mathrm{RMSD}_{\mathrm{ref}}$ was observed, the test results have been encouraging. In two cases the expected increase in ζ for rising $\mathrm{RMSD}_{\mathrm{ref}}$-values was clearly present, despite a somewhat blurred distinction between the main three classes of structural quality. For the remaining test sets in particular the method worked very well, and CYVAL demonstrated its ability to reliably distinguish between structures of high, medium, and low accuracy. A difficulty that we have not yet been able to overcome, however, is that ζ and ζ_{n} depend on the query protein and spectrum type as well as on the degree of structural accuracy. The combined findings illustrate that the performance of the method is not yet applicable to all possible cases but that CYVAL has the potential to reliably assess the correspondence between a structural model and its underlying experimental data.

A number of different approaches were tested in order to evaluate the resulting validation powers of ζ and ζ_{n}. Various error functions for penalizing distance deviations between structure and raw data were explored, alongside diverse ways of calculating the peak weighting factors, and different methods of peak selection. Furthermore, test runs were performed in which the intensities of the automatically picked peaks underwent a correction for peak overlap. A detailed study of the combined results has shown that some approaches appear to be as able as the 'traditional' method to differentiate between structures of different degrees of accuracy, but that no method is clearly superior to all others, while some approaches are markedly less effective at performing the required distinctions.

Arguably, the main weakness of the present CYVAL algorithm is that its validation metrics do not depend solely on the accuracy of the structural model. Nevertheless, the performance of CYVAL is at least equal and at times even superior to the validation power of other approaches to the assessment of structural quality. In fact, none of the three examined tools—Verify3D, ProSa2003, and MolProbity—allow the user to unambiguously deduce the degree of structural quality from the validation score value. This is partly due to variations in the order of magnitude of the validation metrics between the protein test sets. Obviously, other well-known programs for the assessment of structural quality face the same issue as CYVAL. Of course, strictly speaking, a complete comparison of CYVAL and the other three programs is not permissible as these tools operate in a knowledge-based manner and do not take any experimental data into account. Therefore they cannot be expected to provide reliable information about the accuracy of a structural model. This, however, renders them less useful to structural biologists wishing to learn about the quality of the structure they have either calculated or are planning to use in their research. Regardless of this, the three external tools used in this study have a large user base, and not least for this reason it is important to demonstrate that CYVAL is a viable alternative, which the comparison of all four methods has shown it to be.

The currently most widely used data-based validation tool for NMR-derived structures is RPF (**??**). CYVAL has a considerable advantage over this program in that the CYVAL method directly accesses the experimental spectra, whereas RPF requires curated NOESY peak lists due to its inability to distinguish between true signals and noise or artefacts. Consequently, CYVAL depends less strongly on user decisions. Moreover, CYVAL considers deviations between data-derived and structure-based distances, which RPF does not do. This comparison of the two approaches suggests that CYVAL is able to produce results that are more objective, and its evaluation of the structure is likely to be more detailed than the assessment performed by RPF.

Certain features could be incorporated into CYVAL to further enhance the usefulness of the method. For example, the individual contributions to ζ could be mapped onto the protein sequence or the structural model to highlight problematic sites. A range selection mechanism could be introduced, thus allowing the user to restrict the validation procedure to certain areas of the protein. Moreover, since CYVAL is already a part of the NOE assignment and structure calculation software CYANA the method could eventually be incorporated into CYANA's iterative NOESY cross-peak assignment and structure calculation protocol to direct the structure determination process towards a result that is as consistent with the original spectra as possible.

The final obstacles that still need to be overcome once more emphasize the challenging nature of data-based validation of protein structures determined by NMR, but, all in all, the CYVAL method has already displayed its ability to reliably distinguish between structural models of different degrees of accuracy using experimental data. Once the algorithm has fully matured it may well join the group of popular validation tools for NMR-derived protein structures.

III.4.3 Structure calculations

The structures of the WW domain of the peptidyl-prolyl cis/trans isomerase Pin1 and its S16E-mutant WW^{S16E} were calculated (**?**), as well as the structure of the Ser45Ala-mutant of TycC3_PCP, a peptidyl carrier domain of the tyrocidine A synthetase subunit C from *Bacillus brevis* (**?**).

Pin1 plays a role in various cellular processes, e.g. growth-signal responses, hence it has been implicated in several diseases. In numerous human cancers, for example, it is up-regulated (**?**), and its regulation occurs via phosphorylation at various sites, including Ser16. This residue is located in the WW domain of Pin1, which mediates interaction of the protein with its binding partners and is responsible for the subcellular location of Pin1. Phosphorylation at Ser16 down-regulates Pin1 by preventing it from binding to its substrates, which are themselves phosphorylated proteins (**?**). High concentrations of Pin1 phosphorylated at Ser16, i.e. in its down-regulated state, have been encountered in brain tissue samples of Alzheimer's disease patients (**?**). Due to its potentially pivotal role in the abovementioned diseases it is important to gain further information regarding the mechanisms by which the activity of Pin1 is influenced.

The structures of both the wild-type WW domain and its phospho-mimic S16E-mutant were determined in order to learn about the changes induced by phosphorylation at Ser16, hopefully leading to an explanation for how this modification renders the WW domain incapable of mediating protein-protein interactions. While the backbone folds of the wild-type domain and its mutant are highly similar a comparison of the two structures has shown that there is a significant perturbation of the surface charges. This is particularly significant since the Pin1 binding site is characterized by the presence of several arginine residues, i.e. positively charged sidechains (**?**). Additionally, a steric hindrance is caused by the introduction of the somewhat bulky residue. In combination with a number of other experiments performed by **?** these observations offer an explanation for the effects the phosphorylation of Ser16 has on the ability of Pin1 to interact with other proteins.

TycC3_PCP stems from an NRPS, and since these large protein complexes synthesize a multitude of bioactive compounds it is of high pharmacological interest to understand more about their individual components and their reaction mechanisms. Carrier proteins are essential elements of many biosynthetic pathways, and PCPs such as TycC3_PCP generally share the fold of the acyl carrier protein or domain of fatty acid synthases, which may be found in organisms as diverse as bacteria and humans (**?**). Consequently, particularly in the light of efforts to develop new antibiotics that are effective against certain pathogenic bacteria but can be administered to humans without causing severe side-effects, it becomes even more relevant to learn about the differences and similarities between specific proteins and their interactions in different species.

The structure of TycC3_PCP(S45A) in solution has helped correct the understanding of the conformational options of apo-TycC3_PCP. The NMR-derived structure has also strengthened conclusions about the interaction of TycC3_PCP(S45A) with the phosphopantetheine transferase Sfp that could be drawn from an X-ray structure of a complex of these two proteins that was solved by **?** Previously, it was assumed that apo-TycC3_PCP could assume a very loose and flexible conformation (**?**), but the new structural results have shown that both its apo and holo state exhibit the four helix-fold common to all PCPs. Since the aforementioned complex with Sfp is structurally similar to the complex of human acyl carrier protein and phosphopantetheine transferase there may also be similarities between various organisms in how carrier proteins are recognized and posttranslationally modified. This, in turn, may prove to complicate the development of antibiotics targeting bacterial phosphopantetheine transferases.

Part IV

Appendix

Appendix A

CYRANGE

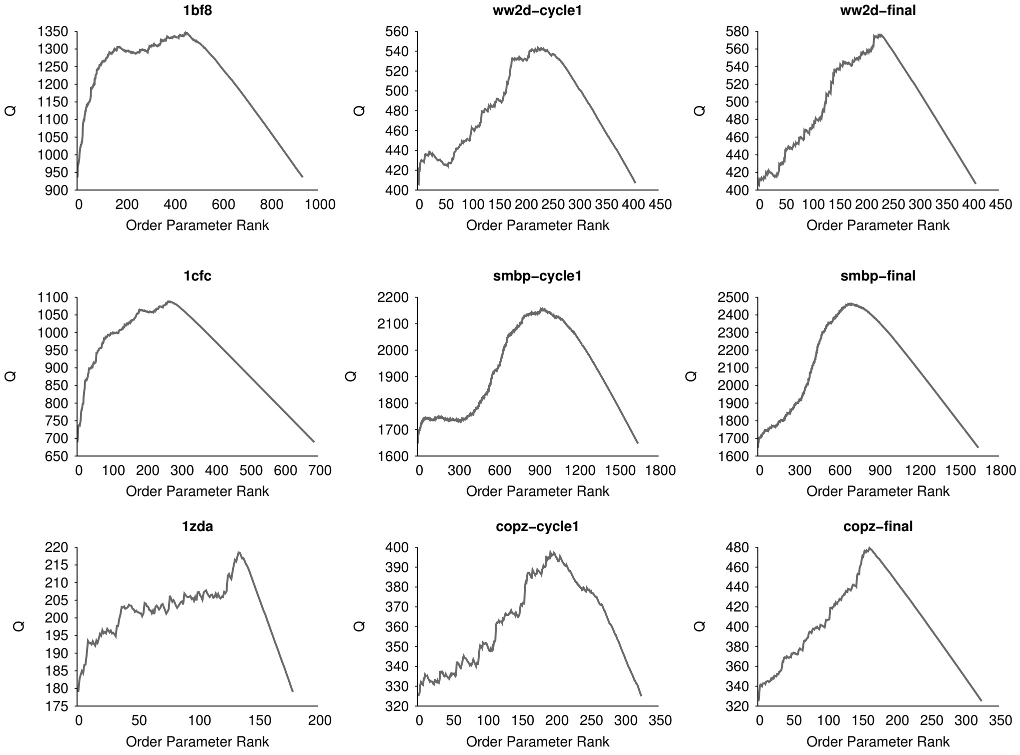


Figure A.1: The quantity Q_i is plotted against the order parameter rank i for nine different protein structure bundles.

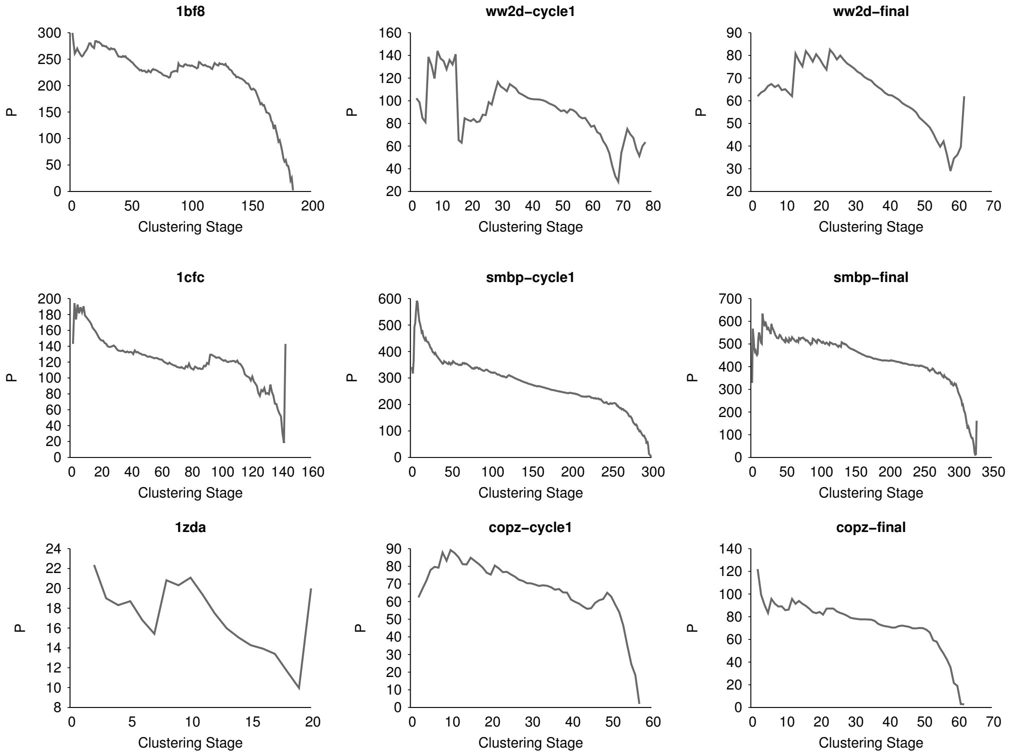


Figure A.2: The quantity P_i is plotted against the order parameter rank i for nine different protein structure bundles.

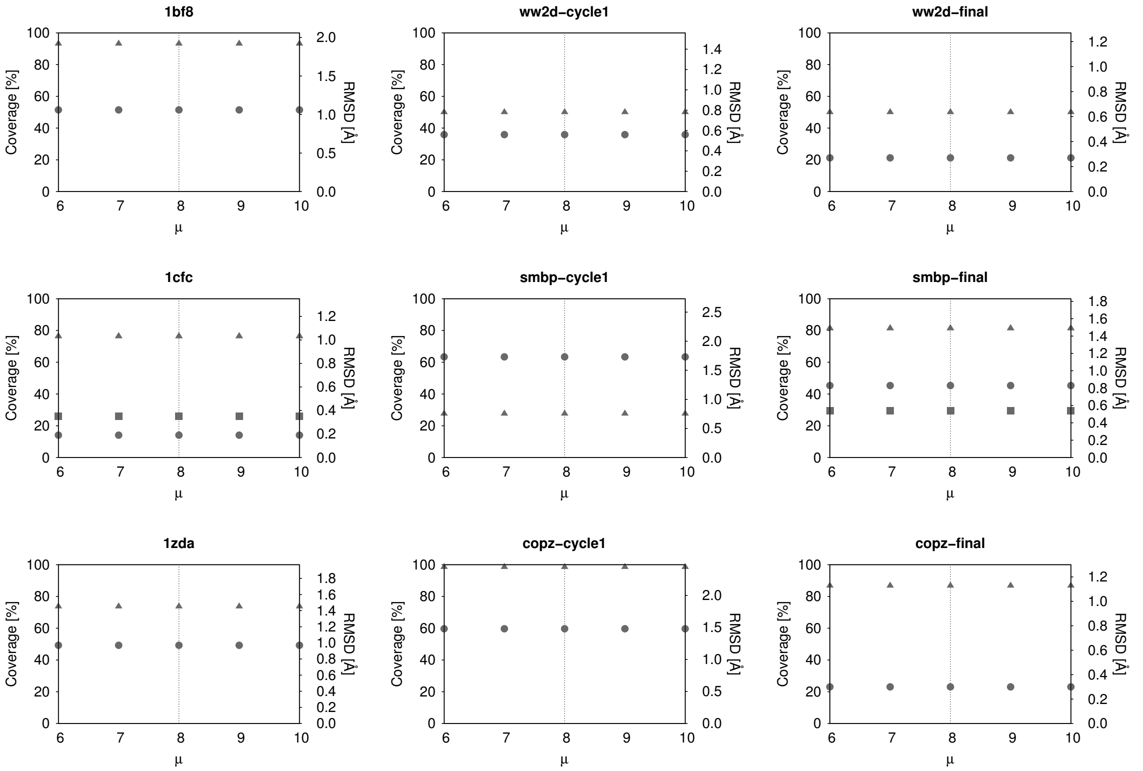


Figure A.3: The sequence coverage (triangles) and RMSD (circles) of the residue ranges determined by CYRANGE were plotted as a function of the minimum cluster size parameter μ for nine different protein structure bundles. The dotted vertical line indicates the default value, $\mu = 8$. Where CYRANGE found two domains, the RMSD values of the second domain are shown as squares.

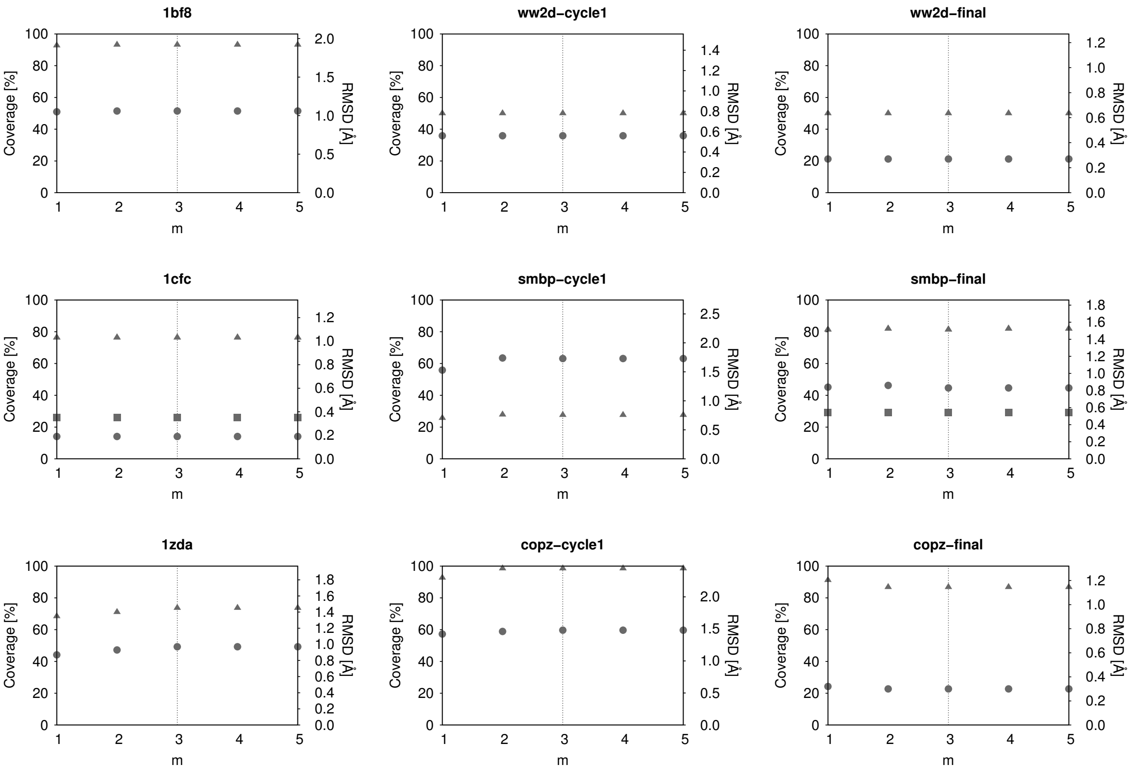


Figure A.4: Dependency of CYRANGE results on the domain boundary extension parameter m. See fig. A.3 for details.

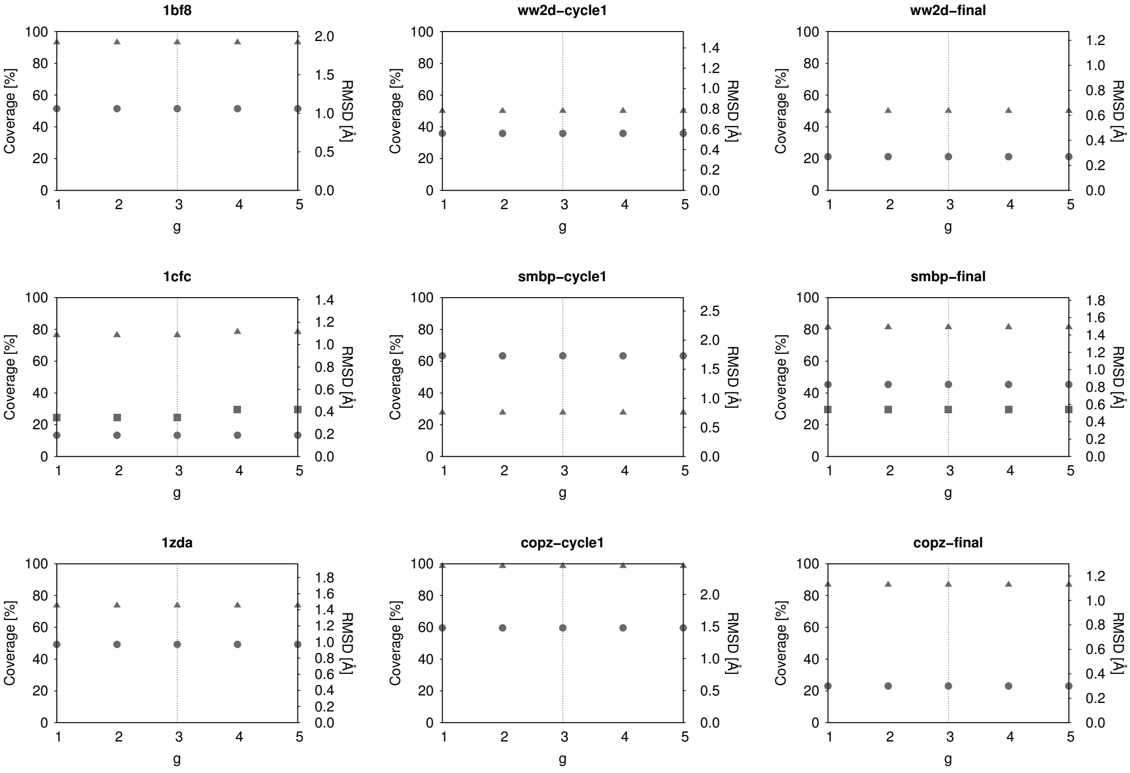


Figure A.5: Dependency of CYRANGE results on the minimal gap width parameter g. See fig. A.3 for details.

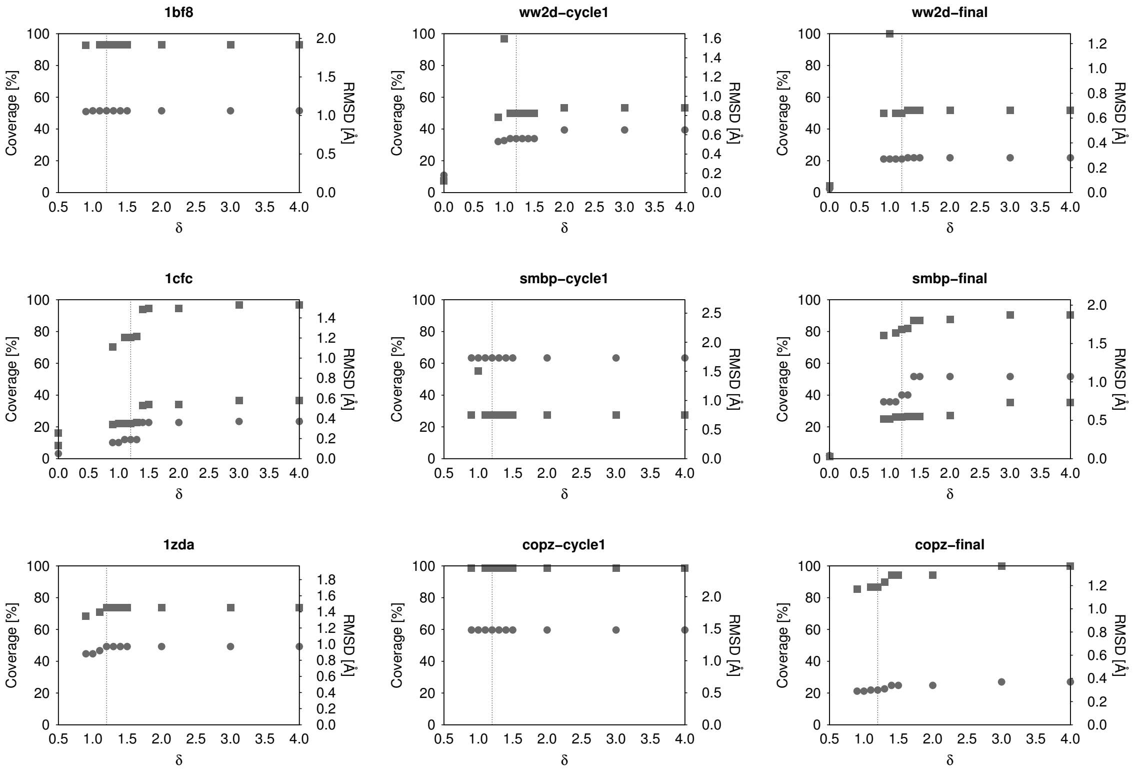


Figure A.6: Dependency of CYRANGE results on the relative RMSD decrease parameter δ. See fig. A.3 for details.

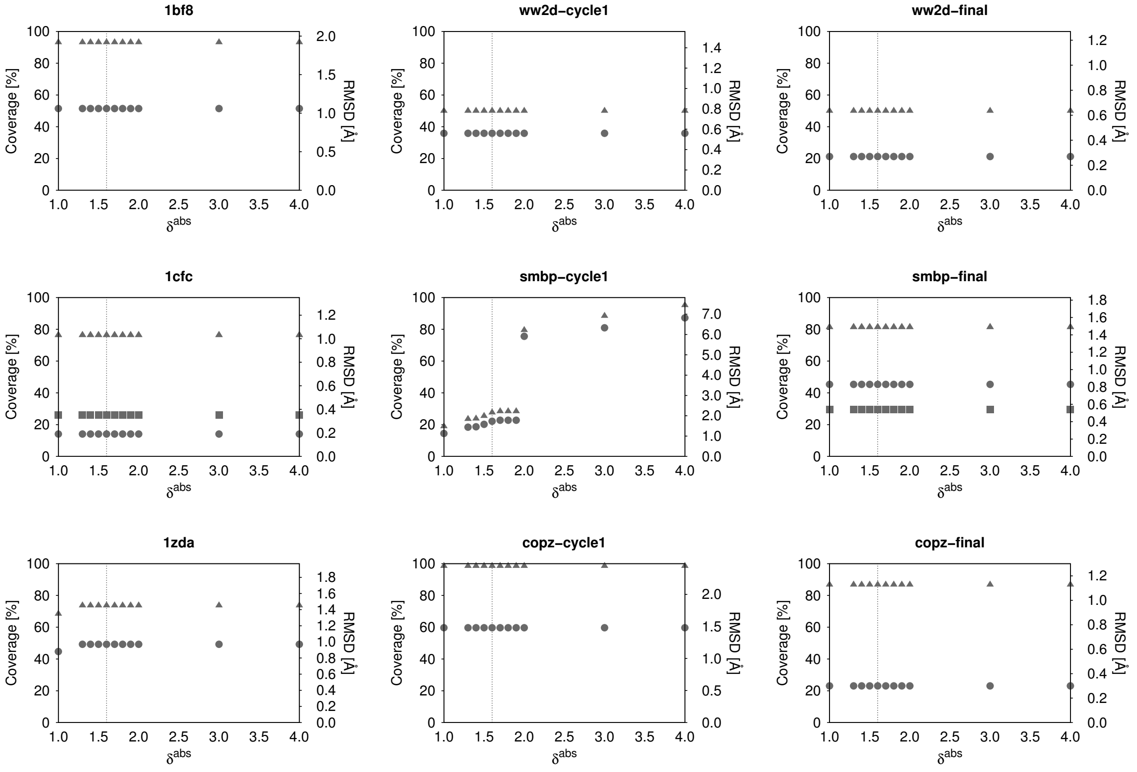


Figure A.7: Dependency of CYRANGE results on the absolute RMSD decrease parameter δ^{abs}. See fig. A.3 for details.

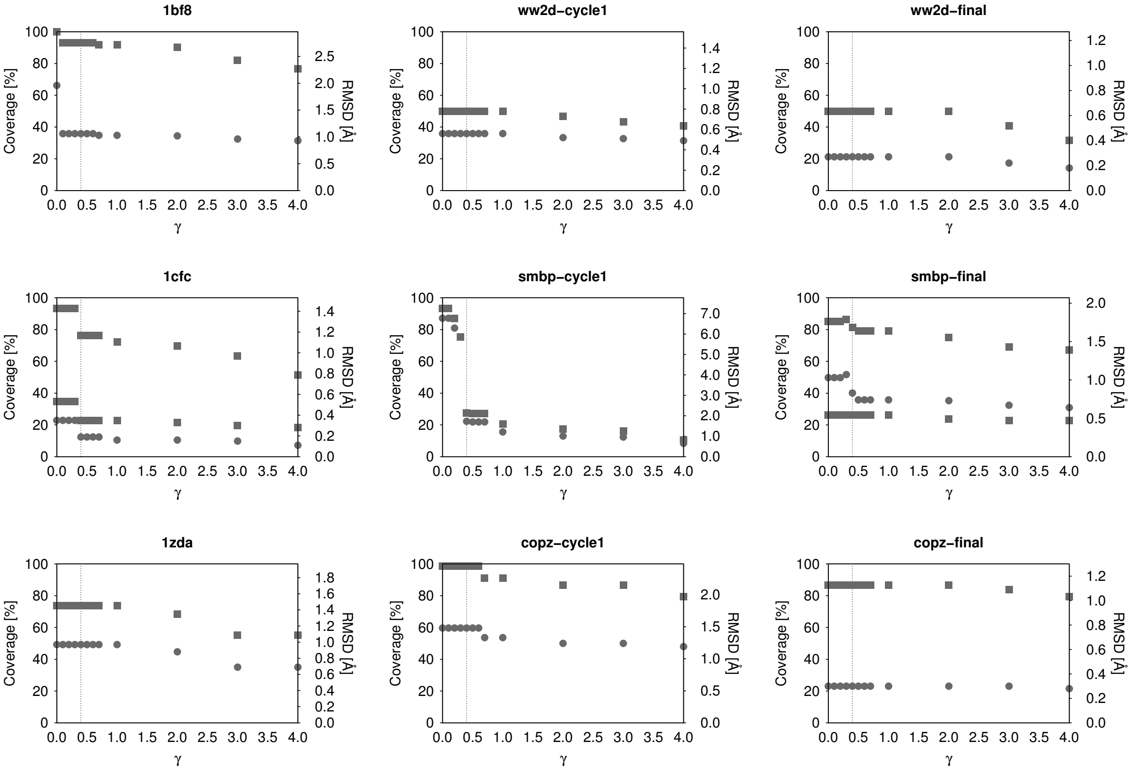


Figure A.8: Dependency of CYRANGE results on the gap penalty parameter γ. See fig. A.3 for details.

Appendix B

CYVAL

Tabs. B.1 to B.5 list the correlation coefficients for ζ and ζ_{n} versus $\mathrm{RMSD}_{\mathrm{ref}}$ for the complete set of test calculations described in section II.2.1.6. See section III.2.4 for the discussion of the results.

Table B.1: Correlation of ζ and ζ_{n} with structural accuracy for the protein enth. The correlation coefficient for ζ_{n} is given in parentheses.

Method	^{13}**C**	^{15}**N**
traditional, corrected intensities	– (–)	– (–)
traditional	0.31 (0.44)	0.10 (0.10)
random structure filtering	-0.11 (0.32)	0.13 (0.23)
matching peaks only	0.83 (0.75)	0.26 (0.36)
long-range peaks only	0.50 (0.24)	-0.01 (0.22)
curated peak list	0.53 (0.53)	0.31 (0.31)
`sat3`, `trad`	0.31 (0.41)	0.08 (0.09)
`sat3`, `one`	0.43 (0.37)	-0.15 (-0.15)
`sat2`, `trad`	0.27 (0.45)	0.09 (0.09)
`sat2`, `one`	0.61 (0.56)	-0.04 (-0.04)
`sat1`, `one`	0.58 (0.49)	-0.09 (-0.09)
`sat1`, `no_info`	0.38 (0.36)	0.16 (0.16)
`sat1`, `no_OP`	0.21 (0.38)	0.06 (0.06)
`sat1`, `equal`	0.49 (0.47)	-0.07 (-0.07)
`sat1`, `dist`	0.56 (0.44)	0.04 (0.04)
`sat1`, `density`	0.41 (0.40)	-0.16 (-0.16)
`pol6`, `trad`	0.25 (0.47)	0.13 (0.13)
`pol6`, `one`	0.62 (0.53)	-0.04 (-0.04)
`pol4`, `trad`	0.26 (0.45)	0.13 (0.14)
`pol4`, `one`	0.64 (0.54)	-0.04 (-0.04)
`log4`, `trad`	0.29 (0.44)	0.21 (0.21)
`log4`, `one`	0.55 (0.48)	0.11 (0.11)
`lin6`, `trad`	0.28 (0.45)	0.19 (0.20)
`lin6`, `one`	0.63 (0.47)	0.07 (0.07)
`lin6`, `no_info`	0.45 (0.38)	0.25 (0.25)
`lin4`, `trad`	0.28 (0.45)	0.19 (0.20)
`lin4`, `one`	0.63 (0.47)	0.07 (0.07)

Table B.2: Correlation of ζ and ζ_n with structural accuracy for the protein HR8254a. The correlation coefficient for ζ_n is given in parentheses.

Method	^{13}**C**	^{15}**N**	^{13}**Caro**
traditional, corrected intensities	0.91 (0.91)	0.82 (0.82)	0.88 (0.88)
traditional	0.91 (0.92)	0.84 (0.84)	0.89 (0.89)
random structure filtering	-0.49 (0.94)	0.15 (0.59)	0.11 (0.89)
matching peaks only	-0.24 (0.02)	0.49 (0.71)	0.07 (0.65)
long-range peaks only	0.62 (0.90)	0.02 (0.68)	– (–)
curated peak list	0.87 (0.90)	0.82 (0.82)	0.90 (0.90)
`sat3`, `trad`	0.92 (0.92)	0.84 (0.84)	0.89 (0.89)
`sat3`, `one`	0.87 (0.86)	0.67 (0.68)	0.71 (0.71)
`sat2`, `trad`	0.89 (0.90)	0.83 (0.83)	0.87 (0.87)
`sat2`, `one`	0.36 (0.39)	0.52 (0.53)	0.67 (0.67)
`sat1`, `one`	0.85 (0.86)	0.66 (0.67)	0.71 (0.71)
`sat1`, `no_info`	0.85 (0.86)	0.87 (0.87)	0.87 (0.87)
`sat1`, `no_OP`	0.91 (0.93)	0.80 (0.80)	0.88 (0.88)
`sat1`, `equal`	0.83 (0.85)	0.82 (0.82)	0.83 (0.83)
`sat1`, `dist`	0.85 (0.86)	0.73 (0.74)	0.75 (0.75)
`sat1`, `density`	0.84 (0.85)	0.81 (0.81)	0.81 (0.81)
`pol6`, `trad`	0.89 (0.90)	0.84 (0.84)	0.89 (0.89)
`pol6`, `one`	0.70 (0.79)	0.63 (0.64)	0.72 (0.72)
`pol4`, `trad`	0.89 (0.90)	0.84 (0.85)	0.89 (0.89)
`pol4`, `one`	0.74 (0.81)	0.68 (0.69)	0.73 (0.73)
`log4`, `trad`	0.91 (0.92)	0.85 (0.85)	0.90 (0.90)
`log4`, `one`	0.25 (0.28)	0.85 (0.85)	0.84 (0.84)
`lin6`, `trad`	0.90 (0.91)	0.85 (0.85)	0.90 (0.90)
`lin6`, `one`	0.77 (0.83)	0.83 (0.83)	0.79 (0.79)
`lin6`, `no_info`	0.84 (0.85)	0.90 (0.90)	0.89 (0.89)
`lin4`, `trad`	0.90 (0.91)	0.85 (0.85)	0.90 (0.90)
`lin4`, `one`	0.78 (0.84)	0.83 (0.83)	0.81 (0.81)

Table B.3: Correlation of ζ and ζ_{n} with structural accuracy for the protein rhod. The correlation coefficient for ζ_{n} is given in parentheses.

Method	**^{13}C**	**^{15}N**	**^{13}Caro**
traditional, corrected intensities	– (–)	– (–)	0.60 (0.61)
traditional	0.75 (0.77)	0.78 (0.79)	0.62 (0.62)
random structure filtering	-0.03 (0.63)	0.11 (0.66)	0.17 (0.33)
matching peaks only	0.35 (0.60)	0.29 (0.60)	– (–)
long-range peaks only	0.08 (0.42)	0.41 (0.81)	– (–)
curated peak list	0.62 (0.75)	0.75 (0.75)	0.10 (0.10)
`sat3`, `trad`	0.75 (0.77)	0.80 (0.80)	0.63 (0.64)
`sat3`, `one`	0.80 (0.80)	0.71 (0.71)	0.65 (0.66)
`sat2`, `trad`	0.25 (0.26)	0.69 (0.70)	0.58 (0.59)
`sat2`, `one`	-0.29 (-0.25)	0.73 (0.73)	0.40 (0.42)
`sat1`, `one`	0.80 (0.81)	0.69 (0.69)	0.58 (0.59)
`sat1`, `no_info`	0.78 (0.79)	0.80 (0.80)	0.18 (0.19)
`sat1`, `no_OP`	-0.06 (-0.01)	0.44 (0.44)	0.57 (0.58)
`sat1`, `equal`	0.21 (0.26)	0.74 (0.75)	0.58 (0.59)
`sat1`, `dist`	0.80 (0.81)	0.76 (0.76)	0.05 (0.06)
`sat1`, `density`	0.81 (0.81)	0.72 (0.73)	0.58 (0.59)
`pol6`, `trad`	0.14 (0.18)	0.28 (0.28)	0.46 (0.46)
`pol6`, `one`	0.80 (0.81)	0.70 (0.70)	0.50 (0.52)
`pol4`, `trad`	0.35 (0.39)	0.77 (0.77)	0.59 (0.60)
`pol4`, `one`	0.80 (0.81)	0.24 (0.24)	0.53 (0.54)
`log4`, `trad`	0.75 (0.78)	0.83 (0.84)	0.65 (0.66)
`log4`, `one`	0.78 (0.79)	0.76 (0.76)	0.67 (0.68)
`lin6`, `trad`	0.75 (0.78)	0.82 (0.83)	0.75 (0.77)
`lin6`, `one`	0.30 (0.33)	0.75 (0.75)	0.62 (0.63)
`lin6`, `no_info`	0.78 (0.80)	0.81 (0.81)	0.66 (0.67)
`lin4`, `trad`	0.79 (0.81)	0.83 (0.83)	0.64 (0.65)
`lin4`, `one`	0.81 (0.82)	0.75 (0.75)	0.62 (0.63)

Table B.4: Correlation of ζ and ζ_{n} with structural accuracy for the protein fsh2. The correlation coefficient for ζ_{n} is given in parentheses.

Method	^{13}**C**	^{15}**N**
traditional, corrected intensities	– (–)	– (–)
traditional	0.83 (0.92)	0.93 (0.93)
random structure filtering	0.40 (0.75)	0.90 (0.87)
matching peaks only	-0.18 (0.72)	0.86 (0.88)
long-range peaks only	-0.27 (0.85)	0.69 (0.80)
curated peak list	0.53 (0.89)	0.86 (0.86)
sat3, trad	0.85 (0.93)	0.93 (0.93)
sat3, one	0.95 (0.94)	0.84 (0.84)
sat2, trad	0.73 (0.88)	0.92 (0.92)
sat2, one	0.94 (0.94)	0.85 (0.85)
sat1, one	0.95 (0.94)	0.84 (0.85)
sat1, no_info	0.93 (0.93)	0.90 (0.91)
sat1, no_OP	0.78 (0.91)	0.92 (0.92)
sat1, equal	0.16 (0.16)	0.87 (0.87)
sat1, dist	0.95 (0.95)	0.86 (0.86)
sat1, density	0.94 (0.94)	0.85 (0.86)
pol6, trad	0.73 (0.89)	0.93 (0.93)
pol6, one	0.95 (0.95)	0.84 (0.85)
pol4, trad	0.70 (0.88)	0.93 (0.93)
pol4, one	0.95 (0.95)	0.84 (0.85)
log4, trad	0.80 (0.93)	0.92 (0.93)
log4, one	0.95 (0.95)	0.87 (0.87)
lin6, trad	0.78 (0.93)	0.93 (0.93)
lin6, one	0.36 (0.45)	0.86 (0.87)
lin6, no_info	0.93 (0.94)	0.90 (0.90)
lin4, trad	0.78 (0.93)	0.93 (0.93)
lin4, one	0.95 (0.95)	0.86 (0.87)

Table B.5: Correlation of ζ and ζ_{n} with structural accuracy for the protein StT322. The correlation coefficient for ζ_{n} is given in parentheses.

Method	^{13}**C**	^{15}**N**	^{13}**Caro**
traditional, corrected intensities	– (–)	0.76 (0.76)	0.51 (0.54)
traditional	0.70 (0.69)	0.56 (0.56)	0.51 (0.54)
random structure filtering	0.07 (0.51)	0.75 (0.84)	0.37 (0.66)
matching peaks only	-0.45 (-0.30)	0.20 (0.60)	0.19 (0.10)
long-range peaks only	-0.10 (0.35)	0.47 (0.76)	0.54 (0.54)
curated peak list	0.74 (0.73)	0.76 (0.76)	0.30 (0.30)
`sat3`, `trad`	0.70 (0.69)	0.73 (0.73)	0.50 (0.54)
`sat3`, `one`	0.66 (0.63)	0.79 (0.79)	0.40 (0.47)
`sat2`, `trad`	0.70 (0.69)	0.75 (0.75)	0.52 (0.55)
`sat2`, `one`	0.45 (0.45)	0.77 (0.77)	0.35 (0.42)
`sat1`, `one`	0.64 (0.62)	0.79 (0.79)	0.38 (0.45)
`sat1`, `no_info`	0.74 (0.73)	0.80 (0.80)	0.27 (0.24)
`sat1`, `no_OP`	0.59 (0.58)	0.73 (0.73)	0.52 (0.55)
`sat1`, `equal`	0.65 (0.63)	0.39 (0.39)	0.39 (0.46)
`sat1`, `dist`	0.61 (0.59)	0.79 (0.79)	0.37 (0.45)
`sat1`, `density`	0.63 (0.61)	0.83 (0.83)	0.41 (0.47)
`pol6`, `trad`	0.68 (0.67)	0.75 (0.75)	0.51 (0.54)
`pol6`, `one`	0.36 (0.37)	0.78 (0.78)	0.35 (0.42)
`pol4`, `trad`	0.68 (0.68)	0.75 (0.75)	0.51 (0.54)
`pol4`, `one`	0.41 (0.42)	0.77 (0.77)	0.36 (0.44)
`log4`, `trad`	0.70 (0.69)	0.78 (0.78)	0.49 (0.54)
`log4`, `one`	-0.68 (-0.68)	0.78 (0.78)	0.38 (0.47)
`lin6`, `trad`	0.70 (0.69)	0.77 (0.77)	0.50 (0.54)
`lin6`, `one`	0.49 (0.49)	0.78 (0.78)	0.37 (0.46)
`lin6`, `no_info`	0.75 (0.74)	0.78 (0.78)	0.43 (0.47)
`lin4`, `trad`	0.70 (0.69)	0.77 (0.77)	0.50 (0.54)
`lin4`, `one`	0.49 (0.49)	0.78 (0.78)	0.37 (0.45)

Appendix C

Structure validation bundle

Listing C.1: Abridged validation report. '. . . ' indicates that entries were removed.

```
***************************************************
*** Validation Overview File generated by CYANA ***
***************************************************

01-Dec-2015, 16:39


*** ProSa2003 (prosa_out001.slp) ***
  conformer   1: molecule     seq-l zp-comb zp-pair zp-surf rk-comb rk-pair rk-surf
       z1-comb z1-pair z1-surf  ep-comb  ep-pair  ep-surf  em-comb  em-pair  em-
      surf es-comb es-pair es-surf
                 protein         68   -6.88   -3.33   -5.61       1       7       1
                        -4.54   -3.67   -4.40   -58.42   -26.52    -6.38    74.75
                         8.55    13.24   19.36   10.52    3.50
  conformer   2: molecule     seq-l zp-comb zp-pair zp-surf rk-comb rk-pair rk-surf
       z1-comb z1-pair z1-surf  ep-comb  ep-pair  ep-surf  em-comb  em-pair  em-
      surf es-comb es-pair es-surf
                 protein         68   -6.73   -3.45   -5.37       1       4       1
                        -4.54   -3.67   -4.40   -55.50   -27.72    -5.56    74.75
                         8.55    13.24   19.36   10.52    3.50
  ...
  conformer  20: molecule     seq-l zp-comb zp-pair zp-surf rk-comb rk-pair rk-surf
       z1-comb z1-pair z1-surf  ep-comb  ep-pair  ep-surf  em-comb  em-pair  em-
      surf es-comb es-pair es-surf
                 protein         68   -6.72   -3.31   -5.44       1       7       1
                        -4.54   -3.67   -4.40   -55.27   -26.30    -5.79    74.75
                         8.55    13.24   19.36   10.52    3.50

*** Verify3D (verify_plot001.out) ***
conformer   1:      Quality:    29.60000
conformer   2:      Quality:    26.51000
               ...
conformer  20:      Quality:    24.88000

*** WhatCheck (pdbout.txt) ***
  -- Warnings and Errors --
    line 81: Warning: Tyrosine convention problem
    line 101: Warning: Phenylalanine convention problem
    line 126: Warning: Aspartic acid convention problem
    line 138: Warning: Glutamic acid convention problem
    ...
    line 1325: Warning: Unusual backbone conformations
    line 1369: Error: Backbone conformation Z-score very low
    line 1400: Error: HIS, ASN, GLN side chain flips
    line 1471: Warning: Buried unsatisfied hydrogen bond donors
    line 1509: Warning: Buried unsatisfied hydrogen bond acceptors

  -- Calculated Values --
  The second part of the table mostly gives an impression of how well
  the model conforms to common refinement constraint values. The
  first part of the table shows a number of constraint-independent
  quality indicators.

   Structure Z-scores, positive is better than average:
    2nd generation packing quality :  -2.988
    Ramachandran plot appearance   :  -4.923 (bad)
    chi-1/chi-2 rotamer normality  :  -4.192 (bad)
    Backbone conformation          :  -6.329 (bad)

   RMS Z-scores, should be close to 1.0:
    Bond lengths                   :   0.116 (tight)
    Bond angles                    :   0.264 (tight)
    Omega angle restraints         :   0.037 (tight)
    Side chain planarity           :   0.033 (tight)
    Improper dihedral distribution :   0.134
    Inside/Outside distribution    :   0.894


*** ProQ (ProQ001.out) ***
   conformer   1: LGscore:             2.099
                  MaxSub:              0.288
   conformer   2: LGscore:             2.356
                  MaxSub:              0.313
            ...
   conformer  20: LGscore:             1.759
                  MaxSub:              0.254

*** ProcheckNMR (tmp.sum) ***
  Ramachandran plot:   78.5% core   21.5% allow    0.0% gener    0.0% disall
  All Ramachandrans:   88 labelled residues (out of1320)
  + Chi1-chi2 plots:      5 labelled residues (out of 780)

*** MolProbity ***
   For raw output see molprobity.out.
   For output in HTML format see molprobity.html.
   For Ramachandran plots see molprobity_Rama.pdf.


*** PDB Validation Suite (PDB_validation.out) ***
   - No suspiciously close contacts between atoms detected.

   - There are missing atoms.

   - There are extra atoms.

   The RMS deviation for covalent bonds relative to the standard
   dictionary is   0.001 Angstroms

   The RMS deviation for covalent angles relative to the standard
   dictionary is    0.2 degrees.
```

Publications

- Lin Y.-J., Ikeya T., Kirchner D. K., and Güntert P. Influence of NMR data completeness on structure determinations of homodimeric proteins. *J Chin Chem Soc-Taip*, 61(12):1297–1306, 2014.
- Tufar P., Rahighi S., Kraas F. I., Kirchner D. K., Löhr F., Henrich E., Köpke J., Dikic I., Güntert P., Marahiel M. A., and Dötsch V. Crystal structure of a PCP/Sfp complex reveals the structural basis for carrier protein posttranslational modification. *Chem Biol*, 21(4):552–562, 2014.
- Luh L. M., Hänsel R., Löhr F., Kirchner D. K., Krauskopf K., Pitzius S., Schäfer B., Tufar P., Corbeski I., Güntert P., and Dötsch V. Molecular crowding drives active Pin1 into non-specific complexes with endogenous proteins prior to substrate recognition. *J Am Chem Soc*, 135(37):13796–13803, 2013.
- Lin Y.-J., Kirchner D. K., and Güntert P. Influence of H-1 chemical shift assignments of the interface residues on structure determinations of homodimeric proteins. *J Magn Reson*, 222:96–104, 2012.
- Gottstein D., Kirchner D. K., and Güntert P. Simultaneous single-structure and bundle representation of protein NMR structures in torsion angle space. *J Biomol NMR*, 52(4):351–364, 2012.
- Kirchner D. K., and Güntert P. Objective identification of residue ranges for the superposition of protein structures. *BMC Bioinformatics*, 12(1):170+, 2011.
- Hirner S., Kirchner D. K., and Somfai P. Synthesis of alpha-amino acids by umpolung of Weinreb amide enolates. *Eur J Org Chem*, 33:5583-5589, 2008.

Eidesstattliche Versicherung

Ich erkläre hiermit an Eides Statt, dass ich die vorgelegte Dissertation über “Analysis of the precision and accuracy of protein structures determined by NMR” selbstständig angefertigt und mich anderer Hilfsmittel als der in ihr angegebenen nicht bedient habe. Insbesondere haben Entlehnungen aus anderen Schriften, soweit sie in der Dissertation nicht ausdrücklich als solche bezeichnet sind, nicht stattgefunden. Ich versichere, nicht die Hilfe einer kommerziellen Promotionsvermittlung in Anspruch genommen zu haben. Ich erkläre weiterhin, dass ich bisher an keiner anderen Universität ein Gesuch um Zulassung zur Promotion eingereicht oder die vorliegende Arbeit als Dissertation vorgelegt habe.

Frankfurt am Main, den