# Chapter 6 Summary

## 6.1 Summary of the work and future perspectives

In the last few years, protein structure determination by solution NMR has benefited enormously from the introduction of software packages for automatically assigning NOESY spectra of small proteins. However, experience shows that more robust protocols are required to tackle the problem of structure calculation of larger proteins or protein complexes. Also, new challenges such as the automation of protein structure determination by MAS solid-state NMR need to be addressed.

This first part of this work represents a thorough study of the influence of the input conditions on the performance of such programmes, using the ARIA protocol (Ambiguous Restraints for Iterative Assignment) as an example. The second section is focused on automated structure calculations using solid-state NMR data.

In Chapters 3 and 4, the influence of three important parameters, the chemical shift tolerances $\Delta$, the cut-off $n_{max}$ for the assignment possibilities of a peak and the number of simulated annealing cooling steps, was extensively investigated. A large number of structure calculations on datasets from five proteins were performed in which these three parameters were systematically varied.

In the course of this study, the dependence of the average number of assignment options per

peak ($n_{av}$) and of the number of rejected peaks on $\Delta$ and $n_{max}$ was studied and rationalised by mathematical relations. It turned out that these functions might be employed as a diagnostic tool to detect macroscopic anomalies in datasets, and, most importantly, to guide the choice of $\Delta$ and $n_{max}$ prior to structure calculation.

A Python script, named Cesta.py, was created and made freely available for rapidly evaluating these diagnostic functions prior to structure calculations. A comparison of structure calculatins with different parameter settings led to the following conclusions:

i) The correct fold of the protein can be obtained from ambiguously assigned cross peaks only.

ii) An estimation of the lower limit for $\Delta$ can be obtained by observing the dependence of the number of rejected peaks due to a lack of assignment options as a function of $\Delta$.

iii) A slow-cooling protocol is a far more convenient way to support convergence than simply increasing the number of calculated structures per iteration, as previously suggested. Surprisingly large numbers of assignment options can be handled (e.g., an average of approx. 50 per peak in the case of the ArgR protein, compared to approx. 7 with a standard protocol), provided that the number of cooling steps is substantially increased.

iv) Small values for the chemical shift tolerances $\Delta$ introduce the risk of excluding the correct assignment from the assignment possibilities for an accepted peak and should be avoided. On the other hand, the choice of large values of $\Delta$ results in high numbers of assignment options. However, this does not represent a problem for convergence if a slow-cooling protocol is used during simulated annealing, as described in ii).

v) The total number of accepted peaks is a misleading parameter for identifying optimal values for Δ, since often the highest number of accepted peaks is obtained for Δ values smaller than the lower limit as determined in iv).

Taken together, these observations suggest that, in general, large Δ values should be applied in conjunction with high numbers of cooling steps.

In Chapter 5, a modification of the ARIA procedure is described, which is specifically designed to assign cross-peaks of solid-state NMR PDSD spectra. This new software, SOLARIA, accepts also typical solid-state $^{13}C$-$^{13}C$ and $^{13}C$-$^{15}N$ correlations in the input peak lists and exploits the characteristic $^{13}C$-labelling scheme of the samples to simplify the assignment of the cross-peaks. PDSD spectra are notably affected by extensive resonance overlap and large line widths, which prevent an accurate measurement of the chemical shifts. To compensate for this, generous values of the tolerances Δ have to be chosen, which leads to high numbers of assignment options per peak. Furthermore, PDSD spectra lack a clear dependence of measured volumes on inter-nuclear distances. Such a problem was circumvented in SOLARIA by the use of unusually large, uniform boundaries for all distance restraints, which reduces, however, their effectiveness in restraining the structure. In view of this, constraints derived from solid-state data are in general more ambiguous and looser than those derived from solution data, which makes structure calculations particularly cumbersome. Despite these limitations inherent in the use of solid-state data, SOLARIA produced accurate structures, provided that the input-parameters were set as suggested above to enhance the robustness of the protocol. The quality of the structures expressed as root-mean-square deviation to the X-ray reference varied from 1.3 Å, when inter-molecular cross-peaks were manually removed, to 2.2 Å, when they were included in the calculation. This is the first example of protein structure determination by automated assignment of MAS NMR spectra of solid protein samples. The automation of cross-peak assignment resulted in a dramatic speed-up of the whole procedure and, most importantly, provided a way to handle

unassigned, ambiguous cross-peaks in MAS NMR spectra. Finally, SOLARIA contains a useful routine to identify inter-molecular contacts by searching for patterns of interactions among the assignment options of the peaks rejected during the calculation.

These findings point out that ARIA and analogous software, if correctly used, might be far more robust than commonly perceived by the community and that the potential of programmes for the automated cross-peak assignment is still largely unexplored. A possible continuation of this work would be the realisation of an improved programme capable to correctly handle intra-molecular and inter-molecular cross-peaks in the peak lists, which would open the way to automated structural investigations of protein complexes in solution or amyloidal fibrils and virus capsids in the solid-state.

## 6.2 Zusammenfassung

In den letzten Jahren hat die Proteinstrukturbestimmung mittels NMR von der Einführung von Softwarepaketen für die automatische Zuordnung von NOESY Spektren und die Strukturberechnung kleiner Proteine stark profitiert. Robustere Protokolle sind jedoch notwendig, um das Problem der Strukturberechnung von größeren Proteinen und von Proteinkomplexen anzugehen. Zusätzlich ergeben sich mit der Automatisierung von Strukturbestimmung durch MAS NMR neue Herausforderungen, zu deren Bewältigung neue Lösungswege beschritten werden müssen.

Der erste Teil dieser Arbeit stellt eine Untersuchung des Einflusses der Eingabeparameter auf die Leistung solcher Programme anhand des Beispiels des ARIA-Protokolls dar. Der zweite Teil beschäftigt sich mit der Automatisierung von Proteinstrukturbestimmung mittels Festkörper-NMR Daten.

In den Kapiteln 3 und 4 wird der Einfluss dreier wichtiger Parameter, der Toleranzen $\Delta$ für die chemischen Verschiebungen, des Höchstwertes $n_{max}$ für die Zuordnungsmöglichkeiten

eines Kreuzsignals und der Kühlungsgeschwindigkeit während des Simulated Annealing, untersucht. Dazu wurden viele Strukturberechnungen angesetzt, in denen diese Parameter systematisch geändert wurden.

Die gleichen Tests wurden mit Datensätzen von vier Proteinen unterschiedlicher Größe und Sekundärstruktur-Architektur wiederholt. Dabei wurde die Abhängigkeit von $\Delta$ und $n_{max}$ der Durchschnittsanzahl der Zuordnungsmöglichkeiten pro Kreuzsignal und der Anzahl der verworfenen Kreuzsignale untersucht und durch Formeln beschrieben. Diese Untersuchungen zeigten, dass solche Funktionen als diagnostisches Werkzeug benutzt werden können, um makroskopische Anomalien zu erkennen, und, was noch wichtiger ist, bei der Auswahl von $\Delta$ und $n_{max}$ vor der Strukturberechnung zu helfen. Ein frei herunterladbares Python Script, Cesta.py, wurde erstellt, um diese diagnostischen Funktionen vor der Strukturberechnung schnell auszuwerten. Die große Anzahl von gewonnenen Daten ermöglichte die Ableitung von allgemeinen Schlussfolgerungen, die sich wie folgt zusammenfassen lassen:

i) Die korrekte Faltung eines Proteins kann ausschließlich unter Verwendung mehrdeutig zugeordneten Kreuzsignale berechnet werden.

ii) Eine langsame Kühlungsphase während des Simulated Annealing ist eine wesentlich bessere Methode, um Konvergenz zu unterstützen, als die Anzahl der in jeder Iteration berechneten Strukturen zu erhöhen, wie dies in der Literatur empfohlen wird. Überraschend viele Zuordnungsmöglichkeiten können gehandhabt werden ( >50 im Falle des ArgR Proteins, im Vergleich zu ca. 7 mit dem Standardprotokoll), vorausgesetzt, dass die Anzahl von Kühlungsstufen substanziell erhöht wird.

iii) Kleine Werte der Toleranzen $\Delta$ für die chemischen Verschiebungen erhöhen das Risiko, dass die korrekte Zuordnung unter den Zuordnungsmöglichkeiten für ein akzeptiertes Kreuzsignal verworfen wird und sollten deswegen vermieden werden.

Auf der anderen Seite führt die Auswahl von großen Δ-Werten zu einer hohen Anzahl von Zuordnungsmöglichkeiten. Das stellt jedoch kein Problem im Hinblick auf die Konvergenz dar, wenn eine langsame Kühlungsphase angewandt wird, wie unter ii) beschrieben.

iv) Eine optimale Werte für eine Untergrenze von Δ kann durch die Analyse der Abhängigkeit der Anzahl verworfener Kreuzsignale als Funktion von Δ gewonnen werden.

v) Die gesamte Anzahl verworfener Kreuzsignale ist ein irreführender Parameter, um optimale Δ–Werte zu identifizieren, da die höchste Anzahl von angenommenen Kreuzsignalen oft für Δ-Werte gewonnen wird, die kleiner als die nach iv) bestimmte Untergrenze ist.

Zusammengefasst zeigen diese Beobachtungen im Allgemeinen, dass große Δ–Werte im Zusammenhang mit vielen Kühlungsstufen benutzt werden sollten.

Im Kapitel 5 wird eine Abwandlung des ARIA-Verfahrens beschrieben, die besonders dafür geeignet ist, Kreuzsignale in PDSD Festkörper-NMR-Spektren zuzuordnen. Dieses neue Software, SOLARIA, akzeptiert auch typische Festkörper-Korrelationen in den Kreuzsignallisten und nutzt das charakteristische Markierungsmuster der Proben, um die Zuordnung zu vereinfachen. PDSD Spektren zeigen bekanntlich eine höhere Überlagerung von Resonanzen und große Linienbreiten, die eine genaue Messung der chemischen Verschiebungen erschweren. Um Zuordnungsfehler zu vermeiden, müssen große Werte für die Δ-Toleranzen gewählt werden, welche zu einer großen Anzahl von Zuordnungsmöglichkeiten führen. Dazu mangelt es PDSD Spektren an einer deutlichen Abhängigkeit der gemessenen Volumina von internuklearen Abständen. Dieses Problem wurde in SOLARIA dadurch gelöst, dass außergewöhnlich große und gleiche Grenzen für alle

Abstandsschranken benutzt werden, die jedoch ihre Nützlichkeit, die Struktur zu definieren, mindern. Angesichts dieser Tatsachen sind von Festkörper-NMR Daten abgeleitete Abstandsschranken im Allgemeinen mehrdeutiger und lockerer als solche, die durch Lösungs-NMR gewonnenen wurden, was Strukturberechnungen besonders erschwert. Trotzdem lieferte SOLARIA genaue Strukturen, mit mittlerer quadratischer Differenz zu der Röntgen-Referenzstruktur von 1.3 Å, wenn die intermolekularen Kreuzsignale manuell verworfen wurden, oder 2.2 Å, wenn sie in die Rechnung einbezogen wurden. Das ist das erste Beispiel von Strukturbestimmung eines Proteins im Festkörper mittels automatischer Zuordnung von MAS-NMR Spektren. Die Automatisierung der Kreuzsignalzuordnung führte zu einer erheblichen Beschleunigung des gesamten Verfahrens und, was noch wichtiger ist, ermöglichte die Zuordnung von sehr mehrdeutigen Kreuzsignale, die man manuell nicht zuordnen konnte. Schließlich enthält SOLARIA eine nützliche Routine, um intermolekulare Kontakte zu identifizieren, die auf einer Untersuchung von Interaktionsmustern unter den Zuordnungsmöglichkeiten der verworfenen Kreuzsignale basiert. Diese Ergebnisse zeigen, dass ARIA und ähnliche Softwarepakete bei einer korrekten Anwendung viel robuster sind, als allgemein wahrgenommen wird, und dass das Potenzial von Programmen für die automatische Kreuzsignalzuordnung immer noch unerforscht ist. Eine mögliche Weiterentwicklung dieser Arbeit böte die Chance, eine sicherere Interpretation von inter- und intramolekularen Kreuzsignale in den Kreuzsignallisten zu erreichen. Das würde automatische strukturelle Untersuchungen von Proteinkomplexen in wässriger Lösung beziehungsweise Amyloiden und Viruskapsiden im Festkörper erleichtern.