

AUREMOL

-

Softwareprojekt zur automatischen Auswertung von multidimensionalen NMR-Spektren

Dissertation
zur Erlangung des Doktorgrades der
Naturwissenschaften (Dr. rer. nat.) der
Naturwissenschaftlichen Fakultät III - Biologie und
Vorklinischen Medizin
der Universität Regensburg



vorgelegt von
Bernhard Ganslmeier aus Landshut
im Juli 2002

Promotionsgesuch eingereicht am: 17.07.2002
Die Arbeit wurde angeleitet von: Prof. Dr. Dr. H. R. Kalbitzer

Prüfungsausschuss:

Vorsitzender: Prof. Dr. R. Witzgall
1. Gutachter: Prof. Dr. Dr. H. R. Kalbitzer
2. Gutachter: Prof. Dr. E. Lang
3. Prüfer: Prof. Dr. U. Krey

Zusammenfassung

In dieser Arbeit wurde das Softwareprojekt AUREMOL zur automatischen Auswertung von multidimensionalen NMR-Spektren entwickelt.

Es wurde zum einen die grafische Benutzeroberfläche erstellt, um ein möglichst komfortables Arbeiten bei der Auswertung zu gewährleisten, zum anderen wurden neue Methoden zur automatischen Auswertung entwickelt. In AUREMOL wurde ein neuer molekülorientierter Ansatz implementiert, der darauf basiert, dass unter Verwendung von *a priori* Wissen und einer möglichst guten Vorhersage der Molekülstruktur NMR-Parameter vorhergesagt werden, die als Leitfaden bei der Auswertung der experimentellen NMR-Spektren dienen. Dazu wurde die Simulation von NOESY-Spektren verbessert, die in diesem Zusammenhang eine wichtige Rolle spielt. Die Berechnung der Linienbreiten der NOE-Signale und der Einfluss der Anisotropie der chemischen Verschiebung und der skalaren Kopplung wurden implementiert.

Zwei neue Methoden zur automatischen NOE-Zuordnung bzw. zur Bestimmung fehlender Zuordnungen wurden entwickelt. Der Algorithmus KNOWNOE ermittelt automatisch Proteinstrukturen, falls die sequentielle Zuordnung der Resonanzen bekannt ist. Dabei werden mit Hilfe eines statistischen Verfahrens die Signale der experimentellen 2D- und 3D-NOESY-Spektren automatisch zugeordnet. Die Struktur des Kälteschock-Proteins *TmCsp* wurde von KNOWNOE ermittelt. Die Qualität der automatisch bestimmten Struktur ist von vergleichbarer Qualität wie die manuell bestimmte Struktur, wie die berechneten R-Faktoren zeigen, die in allen Fällen bei etwa 0,35 liegen.

Das zweite Verfahren ermittelt anhand der partiellen sequentiellen Zuordnung und der Strukturinformation fehlende chemische Verschiebungen in den NOESY-Spektren. Dabei wird ein 2-Stufen Optimierungsalgorithmus eingesetzt, der auf einem Vergleich von simulierten und experimentellen NOESY-Spektren beruht. Anders als bei KNOWNOE werden keine Kreuzsignale zugeordnet, sondern fehlende chemische Verschiebungen (Resonanzlinien) bestimmt. Der Algorithmus benötigt im Falle künstlicher Datensätze eine 30%-ige partielle Zuordnung, um mehr als 90% aller chemischen Resonanzlinien zuzuordnen. Die Methode ist robust gegen Rauschen, Artefaktsignale und fehlende Kreuzsignale in den NOESY-Spektren. Ergebnisse mit einem experimentellen 2D-NOESY-Spektrum der HPr Mutante H15A zeigen, dass auch bei realen Datensätzen fehlende chemische Verschiebungen korrekt bestimmt werden können.

Inhaltsverzeichnis

Zusammenfassung	iii
1 Einleitung	1
1.1 Proteine	1
1.2 NMR-Spektroskopie	2
1.3 Automatische Auswertung von NMR-Spektren	4
1.4 Überblick über die vorliegende Arbeit	7
2 Theoretische Grundlagen	10
2.1 Das Konzept von AUREMOL	10
2.2 Verbesserte Simulation von NOESY-Spektren	14
2.2.1 NOESY-Spektren	15
2.2.2 Longitudinale Relaxation mit J-Kopplung und CSA	16
2.2.3 Transversale Relaxation mit J-Kopplung und CSA	17
2.2.4 Bewegungsmodelle	19
2.2.5 Signalaufspaltung aufgrund der J-Kopplung	20
2.3 NOE-Zuordnungsmethoden	22
2.3.1 Zuordnung mit Bayes (KNOWNOE)	22
2.3.1.1 Die allgemeine Strategie von KNOWNOE	23
2.3.1.2 Der Algorithmus von KNOWNOE	24
2.3.1.3 Berechnung der wahrscheinlichsten Zuordnung	26
2.3.1.4 Wahrscheinlichkeitstabellen	27

2.3.2	Zuordnung durch Vergleich von simulierten und experimentellen Spektren (2-Stufen Algorithmus)	28
2.3.2.1	<i>Threshold Accepting</i>	30
2.3.2.2	Der Kernalgorithmus	31
2.3.2.3	Vorverarbeitung der experimentellen Spektren	34
2.3.2.4	Variation der Zuordnungen	37
2.4	R-Werte zur Qualitätsbeurteilung von NMR-Strukturen	42
2.4.1	Algorithmus von RFAC	43
2.4.2	R-Faktoren	44
3	Ergebnisse	48
3.1	AUREMOL	48
3.1.1	Allgemeine Vorbemerkungen	48
3.1.2	Grafik von AUREMOL	49
3.1.3	Methoden zur Auswertung von NMR-Spektren	50
3.1.4	<i>Homologie Modelling</i> mit PermOL	54
3.1.5	Automatische Artefaktentfernung in NMR-Spektren	55
3.2	Datenbank und Datenstruktur von AUREMOL	57
3.2.1	Globale Datenbank	57
3.2.2	Interne Datenstruktur	59
3.2.2.1	<i>Compounddatei</i>	59
3.2.2.2	<i>Masterliste</i>	60
3.2.2.3	<i>MasterMasterdatei</i>	62
3.2.2.4	<i>Metadatei</i>	62
3.3	Verbesserte Simulation von NOESY-Spektren	65
3.3.1	Implementierung in AUREMOL	65
3.3.2	Beispiele für die verbesserte Simulation	69
3.4	NOE-Zuordnungsmethoden	71
3.4.1	KNOWNOE	71
3.4.1.1	Implementierung in AUREMOL	71

3.4.1.2	Automatische Strukturberechnung von $TmCsp$	73
3.4.2	2-Stufen Algorithmus	74
3.4.2.1	Implementierung in AUREMOL	75
3.4.2.2	Beispiele mit künstlichen Datensätzen	78
3.4.2.3	Beispiele mit experimentellen Datensätzen	86
3.5	Qualitätsbeurteilung von NMR-Strukturen mit R-Werten	91
3.5.1	Implementierung in AUREMOL	91
3.5.2	R-Wert Berechnung am Beispiel des HPr	93
4	Diskussion	95
4.1	Konzept von AUREMOL	95
4.2	Automatische NOE-Zuordnung mit KNOWNOE	97
4.3	Automatische Zuordnung mit 2-Stufen Algorithmus	98
4.3.1	Zuordnung der künstlich erzeugten 2D- und 3D-NOESY-Spektren .	98
4.3.1.1	Zuordnung im 2D-NOESY-Spektrum	98
4.3.1.2	Zuordnung im 3D-NOESY-Spektrum	100
4.3.2	Zuordnung des experimentellen 2D-NOESY-Spektrums	102
4.3.3	Vergleich mit anderen Verfahren zur automatischen Zuordnung . . .	107
	Abbildungsverzeichnis	109
	Tabellenverzeichnis	110
	Abkürzungsverzeichnis	111
	Literaturverzeichnis	112
	Anhang	118
	A Relaxationstheorie	119
	B Kommandos von AUREMOL	121

Kapitel 1

Einleitung

1.1 Proteine

Proteine spielen in biologischen Systemen eine wichtige Rolle. Die Funktion von Proteinen reicht von enzymatischer Katalyse chemischer Reaktionen über Immunabwehr bis hin zur Signaltransduktion [64], [44]. Ähnlich wie die DNA sind Proteine lineare Polymere, aufgebaut aus aneinandergereihten Untereinheiten, den Aminosäuren. Die Sequenz der Aminosäuren, die spezifisch für jedes Protein ist, ist in den Genen codiert.

Das **Humangenomprojekt** (HGP) ist ein internationales Wissenschaftsprojekt mit dem Ziel, die gesamte menschliche Erbinformation zu entziffern. Zur Zeit sind mehr als 90% des menschlichen Genoms entschlüsselt und es hat sich gezeigt, dass statt der erwarteten 100.000 Gene, das menschliche Genom nur über etwa 30.000 Gene verfügt [6]. Das wären nur etwa doppelt so viele Gene wie bei der Fliege. Allerdings produziert das menschliche Gen im Schnitt drei verschiedene Proteine, während das der Fliege nur ein Eiweiß pro Gen produziert. Die vom menschlichen Genom kodierte Gesamtheit der Proteine (Proteom) ist also komplexer als bei einfacheren Organismen. Die Erforschung des Proteoms ist erklärtes Ziel der **Human Proteome Organization** (HUPO) [51], die bis zum Ende des Jahres 2005 alle menschlichen Proteine und deren Interaktionen bestimmen will. Ein fundamentaler Grundsatz der Molekularbiologie besagt, dass man die räumliche Struktur von Molekülen kennen muss, um deren Funktionen und Wechselwirkungen miteinander verstehen zu können. Die Bestimmung der 3D-Struktur von Proteinen liefert also wichtige Erkenntnisse über deren Funktionen und möglichen Wechselwirkungen.

Seit Jahren gibt es Bemühungen, den Faltungsprozess eines Proteins am Computer zu simulieren, d.h. ausgehend von der Primärsequenz des Proteins (Aminosäuresequenz) die räumliche Struktur bzw. die Faltung des Proteins zu berechnen. Durch die Komplexität des Problems ist man allerdings bisher nicht in der Lage, dieses „Protein-Faltungs-

Problem“ zu lösen. Von Bedeutung sind deshalb im Wesentlichen zwei experimentelle Methoden zur Strukturaufklärung von Proteinen in atomarer Auflösung. Das ist zum einen die Röntgendiffraktion bzw. Röntgenstrukturanalyse, und zum anderen die Kernresonanzspektroskopie (NMR), die in den letzten 15-20 Jahren an Bedeutung gewonnen hat. Der Grund für die Entwicklung einer Alternative zur etablierten Röntgenstrukturanalyse liegt darin, dass die Röntgenspektroskopie drei entscheidende Nachteile hat:

- Viele Proteine kristallisieren nicht, was sie nicht zugänglich für die Röntgenstrukturanalyse macht.
- Es existieren mögliche Unterschiede der Proteinstrukturen im kristallinen Zustand und in Lösung.
- Dynamische Prozesse wie die innere Bewegung von Proteinen sind nicht beobachtbar.

Wie bereits erwähnt sind ca. 90% des menschlichen Genoms entschlüsselt, d.h. dass die Sequenz der DNA bekannt ist. Dieser Prozess ist durch Sequenzierungsroboter weitgehend automatisiert, so dass mittlerweile eine große Zahl an Sequenzen bekannt ist, die Proteine kodieren. Im Vergleich dazu sind bisher nur wenige NMR-Strukturen gelöst worden¹, so dass man im Hinblick auf das Proteomics Projekt versucht, mehr Strukturen in kürzerer Zeit zu bestimmen.

1.2 NMR-Spektroskopie

Die NMR-Spektroskopie (*Nuclear Magnetic Resonance*) weist nicht die oben genannten Nachteile der Röntgenstrukturanalyse auf. Allerdings hat auch die NMR limitierende Faktoren. So muss das untersuchte Protein löslich sein, es darf bis zu einer bestimmten Konzentration nicht aggregieren, und die Abhängigkeit der Linienbreite der Signale von der rotatorischen Korrelationszeit des Moleküls liefert eine obere Grenze der mit der aktuellen NMR-Technik messbaren Proteine von etwa 50 bis 60 kDa, was ungefähr 500 Aminosäuren entspricht. Das zur Zeit größte Protein, dessen Struktur mit NMR-Methoden gelöst wurde, ist das Maltodextrin-binding Protein MBP mit einer Molekülmasse von 42 kDa und 370 Aminosäuren [80].

Mit homonuklearen 2D-NMR-Experimenten lassen sich nur Proteinstrukturen von bis zu 100 Aminosäureresten aufklären [105]. Erst vor etwa 10 Jahren wurde es mit der Entwicklung der heteronuklearen 3D- und 4D-NMR-Experimente [10], [23] möglich, die Strukturen von Proteinen mit mehr als 100 Aminosäuren zu lösen.

¹In der PDB Datenbank sind 2236 NMR-Strukturen enthalten, Stand: 07.03.2002

Auf die Grundlagen der NMR-Spektroskopie wird hier nicht näher eingegangen und auf entsprechende Literatur verwiesen [50], [105]. Die Strategie zur manuellen Strukturbestimmung von Biomolekülen mittels NMR-Spektroskopie, die darauf beruht, dass jeder NMR-aktive Kern eine spezifische Resonanzlinie im Spektrum verursacht, gliedert sich in folgende Schritte auf:

1. Sequentielle Zuordnung der Signale
2. Sekundärstrukturbestimmung
3. Tertiärstrukturbestimmung

Zu 1.

Die Identifizierung von Spinsystemen (\cong Aminosäureresten) ist der erste Schritt bei der Strukturbestimmung. Dazu werden geeignete Korrelationsexperimente gemacht, die für die verschiedenen Aminosäuren charakteristische chemische Verschiebungsmuster bzw. Kopplungsmuster erzeugen. Im Wesentlichen versucht man also, bestimmte Muster von Signalen bzw. Resonanzlinien in den NMR-Spektren Aminosäureresten zuzuordnen.

Das Problem bei der Zuordnung liegt darin, dass keine einfache Abhängigkeit zwischen den Resonanzfrequenzen und den Kernen in den Aminosäuren existiert, d.h. die möglichen Werte für die chemischen Verschiebungen eines Kerns einer Aminosäure können aufgrund der Strukturabhängigkeit stark streuen. An welcher Position in der Primärsequenz des Proteins die identifizierten Aminosäuren liegen, erhält man in diesem Stadium der Zuordnung noch nicht. Bei kleineren Proteinen (≤ 100 Aminosäuren) genügt es homonukleare 2D-Experimente wie das COSY- oder TOCSY-Experiment aufzunehmen, um die Spinsysteme eindeutig zu identifizieren. Bei größeren Proteinen ist der Signalüberlapp und die Linienverbreiterung in diesen Spektren zu groß, um eine eindeutige Zuordnung zu erhalten. Die Anfang der 90er Jahre entwickelten heteronuklearen 3D- und 4D-Experimente [10], [23] ermöglichen aber auch bei größeren Molekülen, die Spinsysteme eindeutig zu identifizieren. Durch weitere heteronukleare NMR-Experimente lassen sich sequentiell benachbarte Spinsysteme identifizieren. Es werden Spin-Spin-Wechselwirkungen zwischen sequentiell benachbarten Protonen gemessen, wobei Magnetisierung durch die Bindungselektronen übertragen wird. Auf entsprechende Details sei auf die einschlägige Literatur verwiesen [50] und [105]. Man versucht zuerst, die Zuordnung für das Rückgrat des Proteins zu bestimmen und dann mit weiteren Experimenten, die Zuordnung für die Seitenketten zu ergänzen. Die stereospezifische Zuordnung, also z.B. die Unterscheidung zwischen HB2 und HB3 einer Methylengruppe, ist in diesem Stadium der Auswertung noch nicht möglich. Im Idealfall erhält man für jeden Kern der Aminosäuren mit Spin $\neq 0$ eine chemische Verschiebung, die man sequenzspezifische Zuordnung aller Resonanzlinien nennt.

Zu 2.

Die Bestimmung des chemical shift index (CSI), d.h. die Abweichung der chemischen Verschiebungen der NMR-aktiven Kerne gegenüber den *random coil*² Werten, liefert bereits erste Informationen über die Sekundärstruktur des Proteins, d.h. in welchen Bereichen der Sequenz α -Helizes, β -Faltblätter oder loop-Regionen zu finden sind.

Zu 3.

Die Tertiärstruktur, also die räumliche Struktur des Biomoleküls oder Proteins, wird mit Hilfe weiterer Spektrenparameter, die langreichweitige strukturelle Informationen liefern, bestimmt. Dazu gehören langreichweitige NOE-Kontakte, Wasserstoffbrückenbindungen, Disulfidbrücken und Restdipolkopplungen. NOE-Kontakte zwischen räumlich benachbarten Kernen beruhen auf deren dipolaren Wechselwirkung und sind deshalb abstandsabhängig. Sie liefern damit interatomare Abstände. Gemessen wird dieser Effekt im NOESY-Experiment (**N**uclear **O**verhauser **E**nhancement **S**pectroscopy). Dabei kann man zeigen, dass das Volumen des NOE-Signals bzw. Kontakts zweier Kerne A und B in erster Näherung $\propto r_{AB}^{-6}$ ist, wobei r_{AB} der Abstand der beiden Kerne A und B ist. Durch die Identifizierung vieler Signale in den NOESY-Spektren mit Hilfe der sequentiellen Zuordnung erhält man eine Liste interatomarer Abstände im Protein, die auch Einschränkungen oder *restraints* genannt werden, da sie die Zahl der möglichen Konformationen des Proteins im Raum beschränken. Mit Hilfe dieser Einschränkungen werden Moleküldynamikrechnungen durchgeführt, die den Faltungsprozess simulieren. Die *restraints* werden als Pseudopotentiale in das empirische Kraftfeld des Moleküls eingebaut. Durch Lösen der Newtonschen Bewegungsgleichungen aller Atome versucht man nun, die korrekte Faltung bzw. die Trajektorie der einzelnen Atome nachzuvollziehen. Um den Faltungsprozess künstlich zu beschleunigen, erhitzt man das System, d.h. die Atome erhalten eine höhere kinetische Energie. Dadurch kann der Konfigurationsraum für die möglichen Proteinstrukturen schneller abgetastet werden und die Gefahr, im nächstgelegenen lokalen Energieminimum zu enden, vermieden werden. In einem *simulated annealing* [56] Optimierungsalgorithmus wird dann das System langsam abgekühlt und so in einen stabilen energieminierten Zustand übergeführt. Beispiele für solche Strukturberechnungsprogramme sind CNS [20] oder DYANA [47].

1.3 Automatische Auswertung von NMR-Spektren

Die manuelle Auswertung von NMR-Spektren ist ein sehr zeitaufwendiger und fehleranfälliger Prozess. Abhängig von der Größe des untersuchten Proteins und der damit

²*random coil* bezeichnet den vollständig denaturierten Zustand eines Proteins, dessen Konformation ein statistisches Knäuel darstellt [50].

zusammenhängenden Zahl an verschiedenen NMR-Experimenten kann es einige Jahren dauern, bis die Struktur des Proteins gelöst ist. Daher scheint es sinnvoll zu sein, diese Arbeit zu automatisieren.

Es gibt eine Reihe von Computerprogrammen, die im Wesentlichen nur ein graphisches Interface zur Auswertung zur Verfügung stellen, aber nahezu keine Automatisierung bieten. Sie unterstützen den Benutzer hauptsächlich durch eine automatische Buchführung aller Signale in den Spektren mit Hilfe von *Peaklisten*, leisten aber meist keine weitergehende automatische Zuordnung. Beispiele dafür sind Felix [1], Xeasy [8], NMRPipe [30], ANSIG [59] und AURELIA [85]. Als automatische Auswertung wird z.B. in AURELIA [85] eine Multipletterkennung [84] geboten, mit der Spinsysteme in COSY-Spektren identifiziert werden können. Diese Methode ist auf homonukleare 2D-Spektren beschränkt und liefert nur für Proteine kleiner als 10 kDa brauchbare Ergebnisse. Die Linienverbreiterung bei zunehmender Proteingröße führt zu starken Signalüberlappungen und zu einer Abnahme der Signalintensitäten in den Spektren, was die automatische Multipletterkennung bei größeren Proteinen scheitern lässt.

Die bereits erwähnte Entwicklung neuerer 3D- und 4D-Tripleresonanz NMR-Experimente, s. [10] und [23], hat die Identifizierung von Spinsystemen durch Multipletterkennung abgelöst und die Bestimmung der sequentiellen Zuordnung von größeren Proteinen ermöglicht. Diese Experimente reduzieren die Signalüberlappungen in den Spektren und erhöhen die Empfindlichkeit. Zahlreiche Programme wurden veröffentlicht, die eine teilweise automatische Auswertung dieser Tripleresonanz Experimente vornehmen, um die sequentielle Zuordnung von Resonanzen zu bestimmen. Dazu gehören Programme wie GARANT [9], PASTA [65], CONTRAST [89], AUTOASSIGN [108] oder Softwaretools, entwickelt von Lukin et al. [72], Buchler et al. [21] und Li et al. [67]. Die Vorgehensweise ist bei nahezu allen gleich [78]. Der Input sind *Peaklisten* der oben genannten heteronuklearen NMR-Spektren, die in einem ersten Schritt gefiltert werden, d.h. auf Konsistenz überprüft werden. Danach werden die Resonanzen oder Signale zu Gruppen bzw. Spinsystemen zusammengefasst, die dann identifiziert werden. Bei der Identifizierung werden meist statistische Analysen von chemischen Verschiebungen verwendet. Sind die Spinsysteme identifiziert, müssen sequentielle Nachbarschaften ermittelt werden. Eingesetzt werden hierbei oft deterministische Methoden oder energieminimierende Verfahren wie *simulated annealing* [56]. Damit erhält man Teilstücke oder Segmente sequentiell benachbarter Spinsysteme, die nun in einem letzten Schritt mit der Primärsequenz des Proteins abgeglichen werden müssen. Auch bei dem letzten Schritt werden deterministische oder energieminimierende Methoden eingesetzt. Diese Programme verwenden zur Auswertung der NMR-Spektren also eine ähnliche Strategie wie die eines menschlichen Experten. Das Programm Jigsaw [4] benötigt keine Tripleresonanz Experimente, sondern verwendet zur Bestimmung der sequentiellen Zuordnung der Hauptkettensignale nur vier NMR-Experimente, was zu einer erheblichen Zeiteinsparung bei der Messung der Spektren

und der anschließenden Datenauswertung führt. Das Protein, dessen Struktur bestimmt werden soll, muss auch nicht ^{13}C markiert sein. Die Qualität der Ergebnisse der genannten Programme hängt entscheidend vom Input, also den *Peaklisten* ab. Sind diese nicht vollständig, d.h. fehlen wichtige Signale in den Spektren, so scheitert die automatische Zuordnung. Auch zusätzliche *Rauschpeaks* können die Performance dieser Algorithmen verschlechtern.

Für die Strukturbestimmung ist allerdings nicht nur die sequentielle Zuordnung entscheidend, sondern unter anderem die richtige Zuordnung in den NOESY-Spektren, da hieraus die bereits erwähnten strukturelevanten Informationen gewonnen werden können. NOAH [81] und ARIA [88] sind Programme, die die Zuordnung in NOESY-Spektren ermitteln, wobei die komplette sequentielle Zuordnung als Input in hoher Genauigkeit vorliegen muss. Auch hier gilt wieder, ist die sequentielle Zuordnung fehlerhaft oder unvollständig, und finden sich zu viele überlappende Signale in den NOESY-Spektren, so scheitern auch diese Methoden bei der automatischen Analyse. Das große Problem ist, dass ausgehend von der sequentiellen Zuordnung, nur wenige Signale eindeutig zuzuordnen sind, für die meisten aber mehrere Möglichkeiten existieren. NOAH und ARIA lösen dieses Problem, indem alle mehrdeutigen Zuordnungen mit in die Strukturberechnungen einbezogen werden. ARIA integriert die mehrdeutigen Zuordnungen als gewichtete Summe in die Pseudopotentiale und NOAH berechnet für jede Zuordnungsmöglichkeit Strukturen. Anschließend wird bewertet, wie die Zuordnungen durch die erhaltenen Strukturen erklärt werden. SANE [34] ist ein Tool zur halbautomatischen Zuordnung von NOESY-Spektren, in der der Benutzer bei der Analyse fehlerhafter Zuordnungen interaktiv beteiligt ist. Der verwendete Ansatz ist ähnlich wie bei ARIA.

Gemeinsam ist diesen Programmen, dass nur die Positionen der Signale in den NMR-Spektren, also Informationen über die chemischen Verschiebungen, für die automatische Auswertung verwendet werden, die Linienform der Signale aber vernachlässigt wird. Zum anderen wird eine sogenannte bottom-up Strategie verwendet, die sich primär auf eine NMR-zentrierte Auswertung der Spektren stützt und wenig Information über das Protein oder allgemeiner die NMR-Probe, in der sich das Protein befindet, benutzt. In dieser Arbeit soll ein neuer Ansatz gezeigt werden (top-down Strategie), der auf einem molekülorientierten Vorgehen beruht. Man versucht dabei, anhand von der bekannten oder vermuteten Proteinstruktur, *a priori* Annahmen und statistischen Analysen von NMR-Datenbanken die NMR-Parameter vorherzusagen und diese als Leitfaden zur Auswertung der NMR-Spektren zu verwenden. Wichtige bereits bekannte Informationen über das Protein oder die Zusammensetzung der NMR-Probe fließen auch bei der Analyse mit ein und werden nicht wie bei allen bisherigen Ansätzen ausgeklammert. Der Vorteil dieser Methode ist, dass man, anders als bei der bottom-up Strategie der herkömmlichen Programme, zur automatischen Zuordnung der NOESY-Spektren nicht die komplette sequentielle Zuordnung zur Verfügung haben muss. Dies erspart die Aufnahme und Auswertung zahlreicher

Korrelationsspektren und führt zu einer wesentlichen Beschleunigung der Strukturbestimmung, da im Idealfall nur noch die Aufnahme und Auswertung von 2D- und 3D-NOESY-Spektren nötig ist.

Ein von der Idee her ähnlicher Ansatz wurde mit dem Programm ANSRS [60] entwickelt. Die sequentielle Zuordnung, die die Aufnahme und Analyse vieler Korrelationsspektren erfordert, wird hier nicht als Voraussetzung für die Zuordnung der NOESY-Spektren verwendet. Eine 3D-Struktur wird mit Hilfe nicht zugeordneter NOE-Signale in einem ersten Schritt berechnet, wobei die kovalente Struktur keine Rolle spielt. Anschliessend wird mit einem *simulated annealing* Algorithmus versucht, NOE-Signale zuzuordnen. Der Algorithmus basiert auf Wahrscheinlichkeitsverteilungen chemischer Verschiebungen von ^1H , ^{13}C und ^{15}N Kernen, räumlichen Einschränkungen aufgrund der berechneten Struktur und der bekannten Primärsequenz des Proteins. Weitere Anwendungen, die vor der NOE-Zuordnung eine 3D-Struktur aus nicht zugeordneten Signalen berechnen, wurden von Oshiro und Kuntz [90] sowie Llinas [68] mit dem Programm *Clouds* entwickelt.

Keines der erwähnten Programme ist perfekt. Die Fälle, in denen damit automatisch Strukturen gelöst wurden bzw. sequentielle Zuordnungen bestimmt wurden, sind oftmals mit nahezu idealen Datensätzen durchgeführt worden, die aber in der Praxis in der Regel nicht auftreten. Die Tatsache, dass bisher noch keine Software existiert, die die Auswertung der NMR-Spektren ohne jegliche Einschränkungen automatisiert, rechtfertigt die Entwicklung neuer Programme und Konzepte, die die NMR-Strukturbestimmung ähnlich weitgehend automatisieren sollen, wie es heute schon bei der DNA-Sequenzierung der Fall ist.

1.4 Überblick über die vorliegende Arbeit

In dieser Arbeit wird das neue Softwarepaket AUREMOL vorgestellt, das der manuellen, semi-automatischen und vollautomatischen Auswertung von multidimensionalen NMR-Spektren dienen soll. AUREMOL setzt dabei einen neuen Ansatz ein, die sog. top-down Strategie, die anders als die bisher entwickelten Analyseprogramme auf einem molekülbasierten Vorgehen beruht. Es wird in Zusammenarbeit mit der Firma Bruker BioSpin GmbH in Karlsruhe und dem Lehrstuhl für Biophysik an der Universität Regensburg entwickelt.

Die Arbeit gliedert sich in drei Teile. In Kapitel 2 werden die theoretischen Grundlagen besprochen. Zunächst wird das zugrundeliegende Konzept von AUREMOL dargestellt, wobei besonders auf die top-down Strategie eingegangen wird. Dem molekülbasierten Vorgehen und der Entwicklung einer neuen Datenstruktur und Datenbank kommt dabei eine entscheidende Bedeutung zu. Die Simulation bzw. Rückrechnung von NOESY-Spektren spielt eine wichtige Rolle in diesem Konzept und wird in Kapitel 2.2 näher erklärt. Das am

Lehrstuhl für Biophysik der Universität Regensburg entwickelte Programmpaket RELAX [40], das bereits die Volumina aller auftretenden Signale in 2D- und 3D-NOESY-Spektren mit dem Relaxationsmatrix-Formalismus berechnen konnte, wurde in AUREMOL integriert und um wichtige Punkte verbessert. So wird nun zum einen die Linienbreite der Protonensignale berechnet und zum anderen die Effekte der Anisotropie der chemischen Verschiebung (CSA) und der indirekten Spin-Spin-Kopplung (J-Kopplung) in dem Formalismus berücksichtigt, was zu deutlich realistischeren Simulationen von NOESY-Spektren führt. Die bei den Rechnungen verwendeten Bewegungsmodelle werden kurz erwähnt. Die rückgerechneten NOESY-Spektren dienen als Leitfaden zur Auswertung der experimentellen NMR-Spektren.

Zwei neue Methoden zur automatischen Auswertung von NOESY-Spektren werden in Kapitel 2.3 beschrieben. Das erste Verfahren KNOWNOE, das in Abschnitt 2.3.1 beschrieben wird, führt eine automatische NOE-Zuordnung durch. KNOWNOE setzt ein statistisches Verfahren ein, um ausgehend von der sequentiellen Zuordnung, in einer iterativen Prozedur die Signale in den NOESY-Spektren zuzuordnen. Dabei muss keinerlei Vorkenntnis über die Struktur des Proteins vorliegen, d.h. man kann im Extremfall mit einem ausgestreckten Strang beginnen und iterativ die endgültige Struktur ermitteln. Die zweite Methode in Kapitel 2.3.2 setzt auf den Vergleich von simulierten und experimentellen NOESY-Spektren und benötigt nicht wie KNOWNOE eine möglichst komplette sequentielle Zuordnung, sondern es genügt, wenn eine partielle Zuordnung vorliegt. Dabei wird ein 2-Stufen Modell eingesetzt, wobei die erste Stufe eine erste Grobzuordnung liefert und die zweite Stufe durch die Möglichkeit, beliebig viele Spektren gleichzeitig zu handhaben, eine verfeinerte Zuordnung liefert. Es handelt sich dabei um ein Optimierungsverfahren, wobei zur Optimierung ein *Threshold Accepting* Algorithmus [33] verwendet wird. Neu ist hierbei, dass nicht wie bei allen anderen Programmen, die im vorherigen Abschnitt aufgeführt wurden, nur *Peaklisten* verglichen und optimiert werden, die die Position der Signale und deren Volumina enthalten, sondern als zusätzliches Kriterium auch die Signalformen in den Optimierungsprozess mit einfließen. Dies wird erst durch die verbesserte Simulation von NOE-Signalen möglich, da dadurch die Linienformen exakter berechnet werden können. Ein großes Problem bei der automatischen Auswertung stellen überlappende Signale dar. Alle bisherigen Programme sind durch das *Peaklisten* basierende Vorgehen nicht in der Lage dieses Problem zu lösen, im Gegensatz zu dem neu entwickelten 2-Stufen Modell.

In Kapitel 2.4 wird eine Methode beschrieben, die die Qualität von NMR-Strukturen beurteilt [45]. Dies ist ein wichtiger Punkt in der automatischen Strukturbestimmung, da man damit zu jeder Zeit ein Maß für die Qualität der erreichten Struktur angeben kann. Ein Maß dafür sind die R-Faktoren, die beschreiben, wie gut experimentelle NMR-Parameter, in diesem Falle NOESY-Spektren, durch simulierte Parameter erklärt werden können. Ein neuer R-Faktor wurde entwickelt, der zusätzlich nicht zugeordnete Signale

berücksichtigt und damit genauere Aussagen über die Qualität von Strukturen erlaubt. Alle vorgestellten Methoden sind feste Bestandteile des Konzepts von AUREMOL und bilden das Grundgerüst für die automatische Auswertung der NMR-Spektren.

Die Ergebnisse der Arbeit sind in Kapitel 3 dargestellt. Zunächst wird das Programm AUREMOL in Abschnitt 3.1 beschrieben. Technische Details, die Grafik und Werkzeuge zur manuellen Auswertung von NMR-Spektren werden erklärt, vom Laden der Spektren über das *Peak Picking* bis zur Volumenintegration der Signale. Die Möglichkeit homologe Strukturen zu modellieren, und wichtige Features zur automatischen Vorverarbeitung von experimentellen NMR-Spektren werden beschrieben. Dazu gehören eine automatische Artefaktentfernung.

In Kapitel 3.2 wird die interne Datenstruktur von AUREMOL ausführlich beschrieben und die verwendete Datenbank vorgestellt. Die Ergebnisse der verbesserten Simulation von NOESY-Spektren werden in Abschnitt 3.3 gezeigt. Die neu entwickelten Zuordnungsmethoden, die entscheidend für die automatische Auswertung der NMR-Spektren sind, werden in Kapitel 3.4 dargestellt, wobei Ergebnisse von beiden Algorithmen aufgeführt werden. Bei KNOWNOE wird auf die Implementierung in AUREMOL eingegangen und kurz ein Beispiel für die automatische Strukturbestimmung des Proteins *TmCsp* gezeigt. Für den 2-Stufen Algorithmus werden Ergebnisse von künstlichen und experimentellen Datensätzen dargestellt und damit die Anwendbarkeit und Robustheit des Verfahrens gegenüber unvollständigen Daten gezeigt. Resultate der Qualitätsbeurteilung von NMR-Strukturen mittels R-Faktoren sind in Abschnitt 3.5 beschrieben, wobei besonders auf die Implementierung in AUREMOL eingegangen wird.

Die anschließende Diskussion in Kapitel 4 soll zeigen, in welchem Stadium sich die voll-automatische Analyse von NMR-Spektren in AUREMOL befindet und wie gut die bisherigen Methoden zur automatischen Auswertung arbeiten. Anstehende Projekte, die in naher Zukunft in AUREMOL implementiert werden sollen, sowie ein Vergleich mit bereits existierenden Verfahren zur automatischen Auswertung von NMR-Spektren werden diskutiert.

Kapitel 2

Theoretische Grundlagen

2.1 Das Konzept von AUREMOL

AUREMOL stellt ein Programmpaket zur Auswertung multidimensionaler NMR-Spektren dar, das auf der einen Seite die Möglichkeit der manuellen bzw. interaktiven Analyse von NMR-Spektren bietet, auf der anderen Seite seinen Schwerpunkt aber bei der semi-automatischen bzw. vollautomatischen Auswertung hat. Im Gegensatz zum traditionellen Ansatz, d.h. der NMR-zentrierten Auswertung (bottom-up Strategie), basiert AUREMOL auf einem molekülorientierten Ansatz, der sogenannten top-down Strategie. Der Vorteil ist, dass man vor der Analyse bereits möglichst viel Wissen sammelt und bei der anschließenden Auswertung nutzt. Man versucht dabei, ausgehend von der bekannten bzw. vermuteten Molekülstruktur, die NMR-Parameter möglichst genau vorherzusagen und als Leitfaden bei der Auswertung der NMR-Spektren zu benutzen. Das Konzept, das in AUREMOL benutzt wird, ist schematisch in Abbildung 2.1 dargestellt. In einem ersten Schritt wird möglichst viel Information über das Molekül bzw. Protein gesammelt, was in der Abbildung 2.1 als Definition des Moleküls umschrieben ist. Dazu gehört die Primärsequenz des Proteins, also die Aminosäuresequenz, die Zusammensetzung der NMR Probe, d.h. aus welchen Bestandteilen sich die Pufferlösung zusammensetzt und externe physikalische Parameter, wie die Temperatur, der Druck oder der pH-Wert während der Messung.

Zusätzlich wird in AUREMOL eine interne Datenbank verwendet, die möglichst viel *a priori* Wissen über Proteine und NMR enthält. Der Inhalt der Datenbank, der im Kapitel 3.2 noch näher dargestellt wird, beinhaltet z.B. die chemische Struktur der Aminosäuren, statistische Erwartungswerte chemischer Verschiebungen aller NMR-aktiven Kerne der Aminosäuren, Definition verschiedener NMR-Experimente und die Definition dihedraler Winkel, um J-Kopplungen berechnen zu können. Diese Datenbank wurde möglichst all-

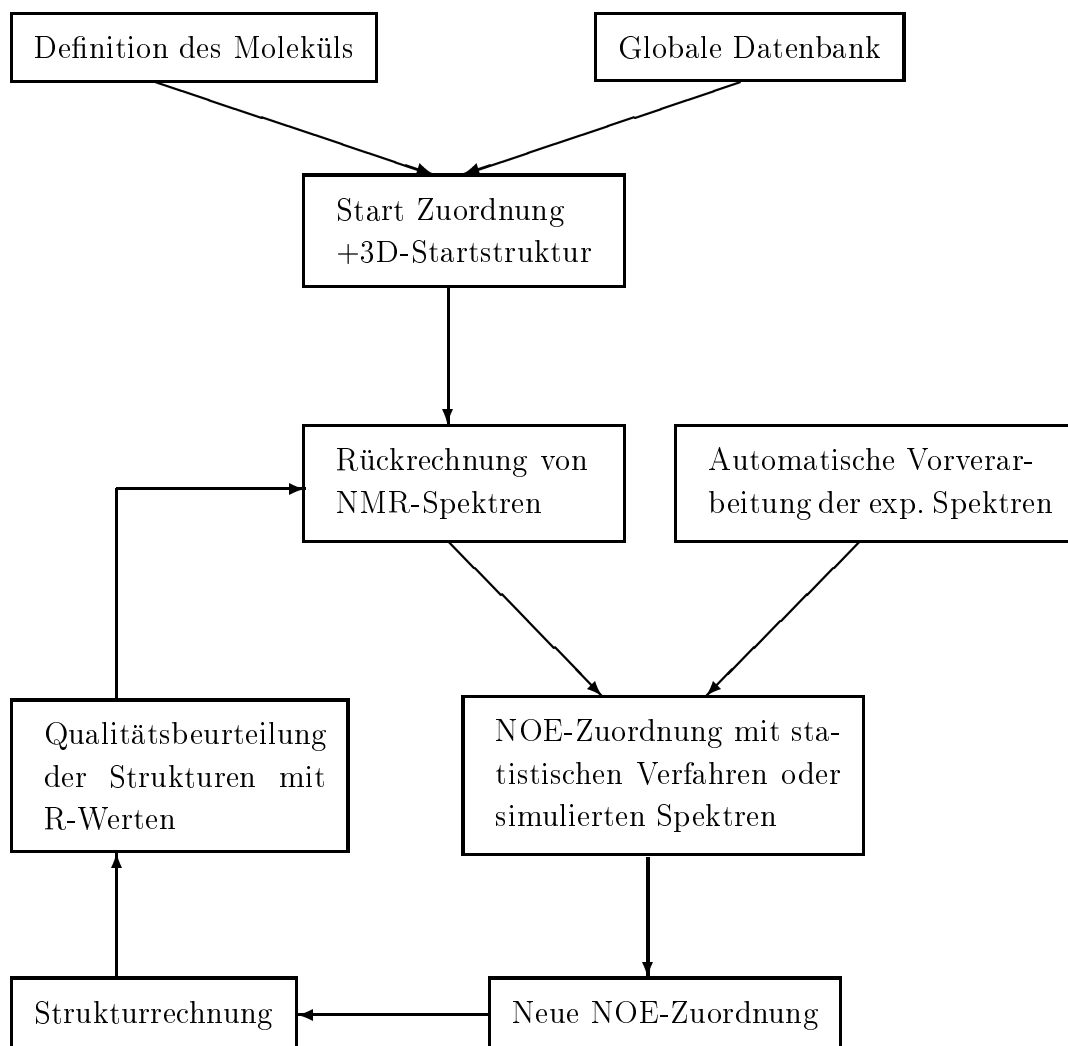


Abbildung 2.1: Schematische Darstellung des molekülorientierten Konzepts von AUREMOL.

gemein gehalten, so dass sie leicht erweitert werden kann. Der Benutzer ist z.B. in der Lage, neue NMR-Experimente zu definieren oder die Datenbank um die chemische Struktur der vier Nukleotiden Cytosin, Guanin, Adenin und Thymin zu erweitern, falls man NMR-Spektren von Nukleinsäuren analysieren will. Darauf wird im Kapitel 3.2 näher eingegangen.

Mit diesen Informationen lässt sich nun z.B. eine erste Zuordnung festlegen, d.h. jedem Kern kann eine chemische Verschiebung mit Hilfe der Datenbank zugeordnet werden. Mit einer Startstruktur, die man z.B. durch *Homologie Modelling* erhalten kann, welches in Kapitel 3.1 näher erläutert wird, oder, falls keinerlei Strukturinformation vorliegt, als *random coil* oder ausgestreckten linearen Strang annehmen kann, lassen sich nun mit dem Relaxationsmatrix-Formalismus NMR-Spektren rückrechnen. Die theoretischen Grundla-

gen werden im Kapitel 2.2 vorgestellt. Bisher ist in AUREMOL die Simulation von 2D- und 3D-NOESY-Spektren implementiert.

Andererseits werden die experimentellen NMR-Spektren einer automatischen Vorverarbeitung unterzogen. Dies beinhaltet ein automatisches *Peak Picking*¹, die Berechnung der Signalvolumina durch Segmentierung und eine automatische Artefaktentfernung, die Signalpeaks und Rauschpeaks diskriminiert. Diese Teilschritte werden in Kapitel 3.1 näher erklärt. Manuelle Änderungen in den experimentellen Spektren durch einen menschlichen Experten sind dabei jederzeit möglich, z.B. die Annotation eines *gepickten Peaks* zu ändern oder *Peaks* in den Spektren hinzuzufügen oder zu entfernen.

Die automatische Zuordnung der experimentellen NOESY-Spektren ist der nächste und auch wichtigste Schritt bei der automatisierten Auswertung. Dabei wurden zwei neue Methoden entwickelt, die auf komplett unterschiedlichen Ansätzen beruhen. Der Algorithmus mit dem Namen KNOWNOE, der in Kapitel 2.3.1 erklärt wird, beruht auf einem statistischen Verfahren, wobei die Faltung des Proteins zu Beginn nicht bekannt sein muss, d.h. es ist möglich, mit einem ausgestreckten Strang zu beginnen, wobei die sequentielle Zuordnung möglichst komplett vorliegen muss. Die zweite Methode, ein 2-Stufen Algorithmus, beruht auf dem Vergleich von simulierten und experimentellen Spektren. Ausgehend von einer bereits gut definierten Anfangsstruktur werden NOESY-Spektren simuliert und mit experimentellen Spektren verglichen. Die Neuerung dieses Verfahrens ist, dass die sequentielle Zuordnung nicht komplett sein muss, sondern eine partielle Zuordnung von etwa 30-40% ausreichend ist. Das Ergebnis sind partiell zugeordnete NOESY-Spektren, aus denen AUREMOL automatisch *restraint files* erzeugt, die für externe Strukturberechnungsprogramme wie DYANA [47] oder CNS [20] als Input dienen. Die Qualität der berechneten Strukturen wird mit Hilfe der R-Faktoren bewertet. Mit den neuen Strukturen werden neue NOESY-Spektren simuliert, es können neue *Peaks* in den NOESY-Spektren zugeordnet werden und der Prozess wiederholt sich solange, bis die Strukturen die gewünschte Güte erreicht haben.

Das vorgestellte Konzept von AUREMOL erfordert eine neu entwickelte interne Datenstruktur, die auf die Anforderungen der automatischen Auswertung von NMR-Spektren abgestimmt ist. In Abbildung 2.2 ist die Datenstruktur schematisch dargestellt. Das Ziel bei der Entwicklung der Datenstruktur war, das *a priori* Wissen aus der Datenbank, die Informationen über das Protein und die NMR Probe, sowie für die Auswertung relevante Daten von den experimentellen Spektren möglichst konsistent und auf das Konzept von AUREMOL abgestimmt auf eine interne Datenstruktur abzubilden. Dazu werden im Wesentlichen vier verschiedene Dateitypen angelegt, die in der Abbildung 2.2 als *Compounddatei*, *Masterliste*, *MasterMasterdatei* und *Metadatei* bezeichnet sind und im Kapitel 3.2 im Detail vorgestellt werden.

¹*Peak Picking* bedeutet die automatische Suche nach lokalen Maxima in den NMR-Spektren, wobei nur oberhalb eines definierten Schwellwerts detektiert wird

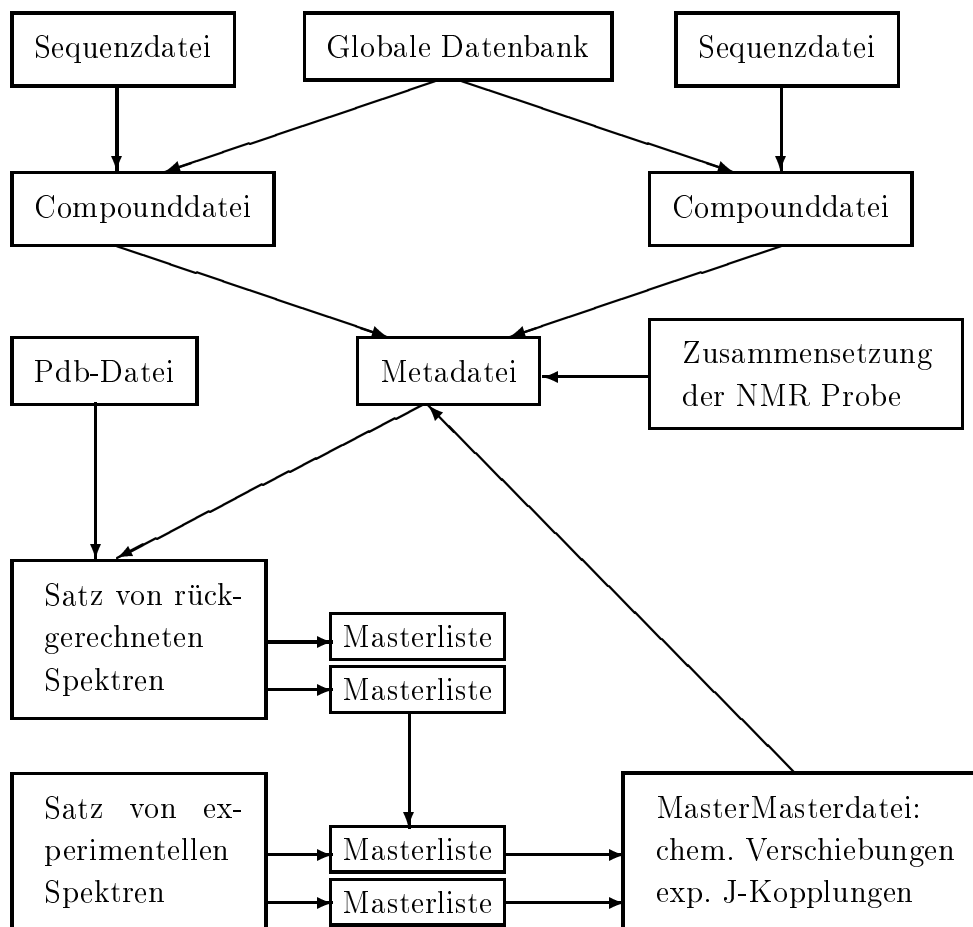


Abbildung 2.2: Interne Datenstruktur, die an das Konzept von AUREMOL angepasst ist, s. Abbildung 2.1.

In der *Compounddatei* wird ein Bestandteil der NMR Probe definiert, das kann das interessierende Protein sein oder andere Moleküle in der Pufferlösung, wie z.B. Wasser, das durch seine Wechselwirkung mit den Proteinmolekülen die NMR-Spektren maßgeblich beeinflusst. Theoretisch können so viele *Compounddateien* angelegt werden, wie verschiedene chemische Stoffe in der NMR Probe vorhanden sind. AUREMOL erzeugt die *Compounddateien* nur im Falle von Proteinen automatisch, dazu wird die Primärsequenz des Proteins und das *a priori* Wissen aus der Datenbank benötigt. *Compounddateien* anderer Bestandteile müssen bis jetzt noch vom Anwender manuell erzeugt werden, d.h. mit einem externen Texteditor.

Die *Masterliste* beinhaltet die komplette Information eines NMR-Spektrums. Das umfasst allgemeine Prozessierungsparameter bis hin zu den *gepickten Peaks* des Spektrums.

Die spektrenunabhängigen Daten werden aus den *Masterlisten* aller zur Auswertung vorhandenen NMR-Spektren extrahiert und in einer neuen Datei, der *MasterMasterdatei*, gespeichert. Spektrenunabhängige Daten sind z.B. die Zuordnung der Resonanzfrequenzen.

Um eine Verknüpfung zwischen den einzelnen Bestandteilen der NMR Probe, also den *Compounddateien*, und den aus den NMR-Spektren gewonnenen Daten, also der *MasterMasterdatei*, zu gewinnen, wird eine neue Datei, die *Metadatei*, erzeugt. In ihr bzw. über die *Metadatei* sind alle für die Auswertung relevanten Daten und Parameter abrufbar. Zusätzlich fließen noch Informationen über die NMR Probe mit ein, s. Abbildung 2.2.

Während der automatischen Zuordnung, die einen iterativen Prozess darstellt und in Abbildung 2.1 zu sehen ist, werden, wie bereits weiter oben erwähnt, immer mehr Signale in den experimentellen NOESY-Spektren zugeordnet. Das bedeutet, die Information in den *Masterlisten* und damit der *MasterMasterdatei* wird laufend erweitert, wodurch auch die *Metadatei* automatisch aktualisiert wird, die wiederum für die Simulation der NOESY-Spektren als Input verwendet wird. Die *Pdb-Datei* in der Abbildung 2.2 enthält die Information über die Proteinstruktur und wird bei der Simulation der Spektren ebenfalls benötigt. Die Struktur der erwähnten Dateien und Listen, sowie die interne Datenbank wird ausführlich im Ergebnisteil im Kapitel 3.2 aufgeführt.

2.2 Verbesserte Simulation von NOESY-Spektren

Die Simulation oder Rückrechnung von NOESY-Spektren ist ein zentraler Bestandteil des Konzepts von AUREMOL, wie man anhand der Abbildung 2.1 sehen kann. Die Berechnungen beruhen auf der vollen Relaxationsmatrix-Analyse und schon seit Jahren gibt es Programme wie CORMA [16], BCKCALC [7], BIRDER [107] und RELAX [40] [39] [38], die damit die Volumina der NOE-Signale berechnen. Andere Softwaretools wie MARDIGRAS [15], NO2DI [103] oder das Programm von Kim und Reid [54] berechnen umgekehrt aus NOE-Volumina die dazugehörigen interatomaren Abstände. Auch für die iterative Verfeinerung von Strukturen (*Refinement*) gibt es Programme wie IRMA [13] und DINOSAUR [14], die den Relaxationsmatrix-Formalismus verwenden.

Anders als bei der ISPA Methode ², bei der das Volumen V_{ij} eines NOE-Signals und der Abstand r_{ij} der zwei wechselwirkenden Spins i und j über die Formel

$$V_{ij} = \alpha r_{ij}^{-6} \quad (2.1)$$

gegeben ist, werden bei der Relaxationsmatrix-Analyse alle dipolar wechselwirkenden Spins als Netzwerk behandelt und die Volumina der NOEs über die Matrix der Kreuzre-

²ISPA steht für *isolated spin-pair approximation* und bedeutet, dass man die dipolar wechselwirkenden Spins als isoliert betrachtet

laxationsraten berechnet.

In AUREMOL werden simulierte NOE-Signale bei der automatischen Zuordnung experimenteller NOESY-Spektren mit Hilfe des 2-Stufen Algorithmus verwendet, s. Abschnitt 2.3.2. Auch die R-Wert Berechnung im Kapitel 2.4 zur Beurteilung der Güte von NMR-Spektren setzt die simulierten *Peaks* ein. Bei den oben genannten Programmen wurden bisher nur die Volumina der NOE-Signale berechnet, andere Merkmale wie Linienbreiten oder die Berücksichtigung weiterer Wechselwirkungen wurden nicht mit einbezogen in die Berechnungen. Die Simulation von NOESY-Peaks wurde in AUREMOL um folgende Punkte erweitert [93]

- Berechnung der Protonen Linienbreiten.
- Berücksichtigung der Spin-Spin Kopplung (J-Kopplung).
- Berücksichtigung der Anisotropie der chemischen Verschiebung (CSA).

Die Linienform der simulierten NOE-Signale wird damit noch exakter bestimmt, was für die automatische Zuordnung in Kapitel 2.3.2 wichtig ist. Dort wird ein Algorithmus vorgestellt, der auf dem Vergleich von rückgerechneten und experimentellen Signalen beruht.

2.2.1 NOESY-Spektren

Der Kernoverhausereffekt (NOE) [91] resultiert aus der Änderung der Dipol-Dipol Kopplung zweier Spins aufgrund der Brownschen Bewegung des Proteins in Lösung [105]. Es findet dabei ein Magnetisierungstransfer während der Mischzeit τ_m zwischen den wechselwirkenden Kernen statt, dessen Größe im NOESY-Experiment gemessen werden kann. Der Magnetisierungstransfer ist dem Volumen des entsprechenden NOE-Signals proportional und durch die dipolare Wechselwirkung abstandsabhängig.

Relevant für die Strukturbestimmung von Proteinen sind die zweidimensionalen bzw. dreidimensionalen NOESY-Experimente. Beim 2D-NOESY-Experiment [52] wird der Magnetisierungstransfer zwischen räumlich benachbarten Protonen gemessen. Das heteronukleare 3D-NOESY-Experiment setzt voraus, dass das Protein isotopenmarkiert ist, d.h. mit ^{13}C oder ^{15}N editiert ist. Es besteht aus einer HSQC- und einer NOESY-Pulssequenz und man spricht von ^{13}C oder ^{15}N NOESY/HSQC-Experimenten. Für weitergehende Informationen sei auf die einschlägige Literatur [50], [105] und [23] verwiesen.

Der Vorteil des 3D-NOESY-Experiments gegenüber dem 2D liegt in einer Verminderung von Signalüberlapp in den Spektren durch Einführung einer weiteren Dimension. Auch 4D-NOESY-Experimente werden bereits bei der Strukturbestimmung sehr großer Proteine eingesetzt [80]. In AUREMOL werden bisher 2D- und 3D-NOESY-Spektren simuliert,

es ist aber geplant, in naher Zukunft auch höherdimensionale NOESY-Spektren zu berechnen.

2.2.2 Longitudinale Relaxation mit J-Kopplung und CSA

Die bereits von Adrian Görler [40] in unserer Arbeitsgruppe entwickelten Methoden zur Berechnung der longitudinalen Relaxationsrate R_1 mit Hilfe des Relaxationsmatrix-Formalismus wurden um die J-Kopplung und CSA erweitert. Die longitudinale Relaxationsrate bestimmt die Änderung der Magnetisierung in Richtung des äußeren Magnetfeldes und damit den Magnetisierungstransfer räumlich benachbarter Spins. Im Folgenden sollen kurz die wichtigsten Ergebnisse angegeben werden, für eine ausführliche Darstellung sei auf die Diplomarbeit von Andreas Ried [93] verwiesen.

Um die Berechnungen zu vereinfachen werden nur zwei Grenzfälle betrachtet und zwar *unlike* und *like* Spins. Zwei Spins I und S nennt man *unlike*, wenn $\omega_I \neq \omega_S$, ansonsten *like* Spins.

In Anhang A ist die Herleitung zur Berechnung der longitudinalen Relaxation eines Systems bestehend aus zwei *unlike* Spins $\frac{1}{2}$ I und S zu sehen. Nach Gleichung A.7 erhält man nun

$$\frac{d}{dt} \langle I_z \rangle = -\rho (\langle I_z \rangle - 2I_0) - \iota \langle 2I_z S_z \rangle - \sigma (\langle S_z \rangle - 2S_0) \quad (2.2)$$

$$\frac{d}{dt} \langle S_z \rangle = -\rho' (\langle S_z \rangle - 2S_0) - \sigma' (\langle I_z \rangle - 2I_0) \quad (2.3)$$

mit

$$\begin{aligned} \rho &= \rho_D + \rho_{CSA} \\ &= \frac{2\pi}{5} \gamma_I^2 \gamma_S^2 \hbar^2 \left(\frac{\mu_0}{4\pi} \right) [J(\omega_I - \omega_S) + 3J(\omega_I) + 6J(\omega_I + \omega_S)] \end{aligned} \quad (2.4)$$

$$\begin{aligned} &+ \frac{2}{15} \omega_I^2 (\sigma_{\parallel} - \sigma_{\perp})^2 \frac{\tau_c}{1 + \omega_I^2 \tau_c^2} \\ \iota &= -\frac{2}{5} \frac{\gamma_I \gamma_S \hbar \omega_I \left(\frac{\mu_0}{4\pi} \right) (\sigma_{\parallel} - \sigma_{\perp})}{r^3} \frac{\tau_c}{1 + \omega_I^2 \tau_c^2} \end{aligned} \quad (2.5)$$

$$\sigma = \frac{2\pi}{5} \gamma_I^2 \gamma_S^2 \hbar^2 \left(\frac{\mu_0}{4\pi} \right)^2 [-J(\omega_I - \omega_S) + 6J(\omega_I + \omega_S)] \quad (2.6)$$

$$\rho' = \frac{2\pi}{5} \gamma_I^2 \gamma_S^2 \hbar^2 \left(\frac{\mu_0}{4\pi} \right)^2 [J(\omega_I - \omega_S) + 3J(\omega_S) + 6J(\omega_I + \omega_S)] \quad (2.7)$$

$$\sigma' = \frac{2\pi}{5} \gamma_I^2 \gamma_S^2 \hbar^2 \left(\frac{\mu_0}{4\pi} \right)^2 [6J(\omega_I + \omega_S) - J(\omega_I - \omega_S)] \quad (2.8)$$

wobei $J(\omega)$ die spektrale Dichte ist, die im Kapitel 2.2.4 näher erläutert ist. ρ bzw. σ bezeichnet man als die Autorelaxations- bzw. Kreuzrelaxationsrate von Spin I. ρ' und

σ' analog von Spin S. Die Autorelaxationsrate ρ in Gleichung 2.4 setzt sich aus einem dipolaren Anteil und einem Anteil, der aus der Anisotropie der chemischen Verschiebung herrührt, zusammen. Bei Letzterem wurde als spektrale Dichte die eines starren Körpers verwendet. Der Interferenzterm ι wird im Folgenden vernachlässigt, da er nur einen kleinen Effekt auf die Relaxationsrate hat, aber einen enormen Anstieg der Rechenzeit zur Folge hätte.

Ein System mit zwei *like* Spins I und S mit $\omega_I = \omega_S = \omega_0$ wird ähnlich gerechnet, nur mit dem Unterschied, dass man keine Aufspaltung erwartet aufgrund der J-Kopplung, d.h. man betrachtet nur die dipolare Wechselwirkung und den CSA Wechselwirkungsterm. Unter Vernachlässigung des oben genannten Interferenzterms ι in Gleichung 2.5 erhält man für die longitudinale Relaxation

$$\frac{d}{dt} \langle I_z \rangle = -\rho (\langle I_z \rangle - 2I_0) \quad (2.9)$$

mit

$$\rho = \frac{2\pi}{5} \gamma_I^2 \gamma_S^2 \hbar^2 \left(\frac{\mu_0}{4\pi} \right) [J(\omega_0) + 4J(2\omega_0)] + \frac{2}{15} \omega_0^2 (\sigma_{\parallel} - \sigma_{\perp})^2 \frac{\tau_c}{1 + \omega_0^2 \tau_c^2} \quad (2.10)$$

In diesem Fall erhält man keine Kreuzrelaxationsanteile für die longitudinale Relaxation und die Autorelaxationsrate ρ ist wieder die Summe aus dipolaren Term und CSA Term. Die longitudinale Relaxationsrate R_1^I eines Spins I wird wegen der nichtverschwindenden Kreuzrelaxation mit dem Relaxationsmatrix Formalismus berechnet, während der Beitrag der anisotropen chemischen Verschiebung nur zur Autorelaxationsrate addiert wird.

2.2.3 Transversale Relaxation mit J-Kopplung und CSA

Die transversale Relaxation R_2 bestimmt die Linienbreite $\Delta\nu$ des Signals im Frequenzspektrum. Wie man leicht zeigen kann [22] gilt

$$\Delta\nu \propto R_2 \quad (2.11)$$

Die Berechnung der transversalen Relaxationsrate R_2 läuft in analoger Weise ab wie die Rechnung der longitudinalen Relaxationsrate, wobei nicht die zeitliche Entwicklung der z-Komponente der Magnetisierung I_z bzw. S_z interessiert, sondern die dazu senkrechten Komponenten I_+ oder S_+ ³. Betrachtet man wieder zwei *unlike* Spins $\frac{1}{2}$, so erhält man aufgrund der J-Kopplung eine Aufspaltung in ein Dublet und muss demnach zwei Komponenten unterscheiden

$$\frac{d}{dt} \langle I_+^{(1)} \rangle = -i\pi J \langle I_+^{(1)} \rangle - (\lambda + \eta) I_+^{(1)} - \mu I_+^{(2)} \quad (2.12)$$

$$\frac{d}{dt} \langle I_+^{(2)} \rangle = i\pi J \langle I_+^{(2)} \rangle - (\lambda - \eta) I_+^{(2)} - \mu I_+^{(1)} \quad (2.13)$$

³ $I_+ = I_x + iI_y$ und $S_+ = S_x + iS_y$

mit

$$\begin{aligned} \lambda = & \frac{2\pi}{10} \gamma_I^2 \gamma_S^2 \hbar^2 \left(\frac{\mu_0}{4\pi} \right) [4J(0) + J(\omega_I - \omega_S) + 3J(\omega_I) + 3J(\omega_S) \\ & + 6J(\omega_I + \omega_S)] + \frac{1}{15} \omega_I^2 (\sigma_{\parallel} - \sigma_{\perp}) \frac{\tau_c}{1 + \omega_I^2 \tau_c^2} + \frac{4}{45} \omega_I^2 (\sigma_{\parallel} - \sigma_{\perp})^2 \tau_c \end{aligned} \quad (2.14)$$

$$\eta = \frac{1}{15} \left(\frac{\mu_0}{4\pi} \right) \frac{\gamma_I \gamma_S \hbar \omega_I (\sigma_{\parallel} - \sigma_{\perp})}{r^3} \left[4\tau_c + \frac{3\tau_c}{1 + \omega_I^2 \tau_c^2} \right] \quad (2.15)$$

$$\mu = \frac{6\pi}{10} \gamma_I^2 \gamma_S^2 \hbar^2 \left(\frac{\mu_0}{4\pi} \right)^2 J(\omega_S) \quad (2.16)$$

Die Autorelaxationsrate λ zerfällt in einen dipolaren und einen anisotropen Anteil der chemischen Verschiebung. Der Interferenzterm η wird wieder vernachlässigt aufgrund der oben genannten Gründe, ansonsten würde man für die einzelnen Komponenten unterschiedliche Linienbreiten erhalten. Aufgrund der Kreuzrelaxation μ zwischen den aufgespaltenen *Peaks* müsste man auch hier den Relaxationsmatrix-Formalismus anwenden. Da aber der Beitrag der Kreuzrelaxation zur Autorelaxationsrate λ sehr klein ist, kann man den Kreuzrelaxationsterm vernachlässigen, wenn man indirekt koppelnde Spins hat. Ist keine J-Kopplung vorhanden, so wird der Kreuzrelaxationsterm zur Autorelaxationsrate addiert:

$$R_2^{(1,2)} = \lambda + \mu \quad , \text{ wenn } J = 0 \quad (2.17)$$

$$R_2^{(1,2)} = \lambda \quad , \text{ wenn } J \neq 0 \quad (2.18)$$

Man sieht, dass die Linienbreite indirekt gekoppelter Spins kleiner wird.

Wenn man wiederum ein System von zwei *like* Spins I und S mit $\omega_I = \omega_S = \omega_0$ betrachtet, erhält man folgendes Ergebnis

$$\frac{d}{dt} \langle I_+ \rangle = -(\lambda + \mu) \langle I_+ \rangle \quad (2.19)$$

$$\frac{d}{dt} \langle S_+ \rangle = -(\lambda + \mu) \langle S_+ \rangle \quad (2.20)$$

mit

$$\lambda = \frac{2\pi}{10} \gamma_I^2 \gamma_S^2 \hbar^2 \left(\frac{\mu_0}{4\pi} \right)^2 [5J(0) + 9J(\omega_0) + 6J(2\omega_0)] \quad (2.21)$$

$$+ \frac{1}{15} \omega_0^2 (\sigma_{\parallel} - \sigma_{\perp}) \frac{\tau_c}{1 + \omega_0^2 \tau_c^2} + \frac{4}{45} \omega_0^2 (\sigma_{\parallel} - \sigma_{\perp})^2 \tau_c \quad (2.22)$$

$$\mu = \frac{2\pi}{10} \gamma_I^2 \gamma_S^2 \hbar^2 \left(\frac{\mu_0}{4\pi} \right) [4J(0) + 6J(\omega_0)] \quad (2.23)$$

Aufgrund der nicht verschwindenden Kreuzrelaxation müsste man den Relaxationsmatrix-Formalismus verwenden, um R_2 zu berechnen. Da aber die Spins I und S magnetisch

äquivalent (*like*) sind, ist R_2 einfach die Summe aus Autorelaxationsrate λ und Kreuzrelaxationsrate μ .

Die Ergebnisse sind leicht auf ein System von N Spins erweiterbar. Man erhält für die transversale Relaxationsrate R_2^I eines Spins I

$$R_2^I = \sum_{\substack{S, S \neq I \\ \text{like}}} R_2^{IS} + \sum_{\substack{S, S \neq I \\ \text{unlike}}} R_2^{IS} \quad (2.24)$$

wobei $S = 1, \dots, N$ und $S \neq I$ ist.

2.2.4 Bewegungsmodelle

Wie bereits erwähnt, hängen die Relaxationsraten von der spektralen Dichtefunktion $J^{IS}(\omega)$ ab. Die spektrale Dichte im Falle dipolar wechselwirkender Spins beschreibt, wie sich der Verbindungsvektor zwischen Spin I und S relativ zum äusseren Magnetfeld bewegt. Bei den Simulationen werden die folgenden Bewegungsmodelle verwendet, um analytische Ausdrücke für $J^{IS}(\omega)$ zu erhalten [40].

- Das Modell des starren Körpers

Man nimmt an, dass das Protein starr ist und sich isotrop mit einer Korrelationszeit τ_c in der Lösung bewegt. Das führt zu einer spektralen Dichte

$$J^{IS}(\omega) = \frac{1}{4\pi} \frac{1}{r_{IS}^6} \frac{\tau_c}{1 + \omega^2 \tau_c^2} \quad (2.25)$$

Dieses Modell wird für die Protonen im Rückgrat des Proteins verwendet und stellt dort eine gute Näherung dar. Für die Berechnung der anisotropen chemischen Verschiebung in Gleichung 2.4 wird dieser Ansatz auch verwendet.

- Das Modell von Lipari

Lipari und Szabo [69], [70] haben gezeigt, dass die spektrale Dichte eines Moleküls, das eine Überlagerung einer langsamen isotropen Bewegung und einer schnellen anisotropen internen Bewegung aufweist, sich schreiben lässt

$$J^{IS}(\omega) = \frac{1}{4\pi} \frac{1}{r_{IS}^6} \left(\frac{S^2 \tau_c}{1 + \omega^2 \tau_c^2} + \frac{(1 - S^2) \tau}{1 + \omega^2 \tau^2} \right) \quad (2.26)$$

mit $\frac{1}{\tau} = \frac{1}{\tau_c} + \frac{1}{\tau_e}$

τ_e : Korrelationszeit der internen Bewegung

S : Ordnungsparameter, ein Maß für die räumliche Einschränkung der Bewegung

- Vereinfachtes Modell von Lipari

Mit der Annahme, dass $\tau_e \ll \tau_c$ gilt

$$J^{IS}(\omega) = \frac{1}{4\pi} \frac{1}{r_{IS}^6} \frac{S^2 \tau_c}{1 + \omega^2 \tau_c^2} \quad (2.27)$$

- Sprungmodelle

Angenommen, die Verbindungslinie r_{IS} zwischen Spin I und S kann N diskrete Zustände einnehmen und die Korrelationszeit der Sprungbewegung sei τ_1 , dann kann man drei Arten von Sprüngen unterscheiden

- Slow Jump

Die Korrelationszeit der Sprungbewegung ist viel größer als die Korrelationszeit τ_c des Moleküls, d.h. $\tau_1 \gg \tau_c$ und man erhält

$$J^{IS}(\omega) = \frac{1}{4\pi N} \frac{\tau_c}{1 + \omega^2 \tau_c^2} \sum_{\mu=1}^N \frac{1}{r_{IS,\mu}^6} \quad (2.28)$$

Dieses Modell beschreibt die δ - und ϵ -Protonen in aromatischen Ringen gut.

- Fast Jump

Hier ist die Korrelationszeit der Sprungbewegung sehr viel kleiner als τ_c , d.h. $\tau_1 \ll \tau_c$ und es gilt

$$J^{IS}(\omega) = \frac{1}{8\pi N^2} \frac{\tau_c}{1 + \omega^2 \tau_c^2} \sum_{\mu\nu=1}^N \frac{1}{r_{IS,\mu}^5 r_{IS,\nu}^5} \left(3 (\vec{r}_{IS,\mu} \cdot \vec{r}_{IS,\nu})^2 - r_{IS,\mu}^2 r_{IS,\nu}^2 \right) \quad (2.29)$$

Das Modell beschreibt in guter Näherung die Bewegung von Methylgruppen.

- Average_3

$$J^{IS}(\omega) = \frac{1}{4\pi N^2} \frac{\tau_c}{1 + \omega^2 \tau_c^2} \left| \sum_{\mu=1}^N \frac{1}{r_{IS,\mu}^3} \right|^2 \quad (2.30)$$

Dieses Modell ist eine gute Näherung von Gleichung 2.29, falls Spin I und Spin S weit voneinander entfernt sind.

2.2.5 Signalaufspaltung aufgrund der J-Kopplung

Die J-Kopplung bewirkt in der NMR eine Aufspaltung der Resonanzlinien und beruht auf einer Wechselwirkung, die durch das bindende Elektronenpaar der koppelnden Kerne übertragen wird. Die Größe der J-Kopplung ist durch eine Kopplungskonstante ${}^n J_{AX}$ gegeben, wobei n die Zahl der Bindungen zwischen den Kernen A und X bezeichnet.

Wichtig sind die 2J - und 3J -Kopplungen, die man auch geminale und vicinale Kopplungen nennt.

Bei den Berechnungen wurde eine schwache Kopplung angenommen, d.h.

$$|\omega_A - \omega_X| \gg J_{AX} \quad (2.31)$$

Die Aufspaltung der Signale unterliegt dadurch einem einfachen Schema [50].

Die Kopplung eines Kerns A mit einem Spin $\frac{1}{2}$ Teilchen X führt zu einem Dublet, mit Abstand J_{AX} und halber Intensität. Koppelt A nun mit einem weiteren Spin Y , so werden die beiden Linien auch wieder aufgespalten in jeweils zwei Linien, halber Intensität und Abstand J_{AY} . Dieses Verfahren lässt sich iterativ auf mehrere koppelnde Partner anwenden.

Bei einer Kopplung von A mit N äquivalenten Kernen X_0, X_1, \dots, X_{N-1} erhält man nach der eben beschriebenen Regel eine Aufspaltung in $N+1$ Linien, wenn es sich um Spin $\frac{1}{2}$ Teilchen handelt. Die dazugehörigen Intensitäten sind binomialverteilt, d.h. die K -te Linie hat die Intensität

$$I_K \propto \binom{N}{K} = \frac{N!}{K!(N-K)!} \quad (2.32)$$

Vicinale Kopplungen, also Kopplungen über drei Bindungen, hängen von der 3D-Struktur des Proteins ab. Bei der 3J -Kopplung ist die Kopplungskonstante eine Funktion des dihedralen Winkels Θ , wie er in den IUPAC-Konventionen [73] definiert ist

$$^3J_{AX} = A \cos^2 \Theta + B \cos \Theta + C \quad (2.33)$$

Dies ist die semiempirische Karplusgleichung [53], deren Parameter A , B und C für homonukleare Kopplung den Arbeiten von [29], [31], [71], [95], [97], [99], [104] und [106] entnommen sind. Tabellarisch zusammengefasst sind sie auch in der Diplomarbeit von A. Ried [93] zu finden. Die entsprechenden Koeffizienten im heteronuklearen Fall sind in [97] nachzulesen. Damit ist es möglich, anhand der Proteinstruktur die vicinalen Kopplungskonstanten für die Simulation zu berechnen.

2.3 NOE-Zuordnungsmethoden

Nach der Rückrechnung von NOESY-Spektren folgt in der Abbildung 2.1 die automatische Zuordnung von experimentellen NOESY-Spektren. Dafür wurden zwei neue Methoden entwickelt, die sich in ihrer Vorgehensweise komplett unterscheiden.

Der Algorithmus KNOWNOE [46] ist auf einem wahrscheinlichkeitsbasierten Ansatz aufgebaut. Die sequentielle Zuordnung wird hierbei als Leitfaden verwendet, d.h. es findet eine Zuordnung aufgrund gegebener chemischer Verschiebungen statt und bei mehrdeutigen Zuordnungen wird die wahrscheinlichste Zuordnung berechnet. Über die 3D-Struktur des Proteins muss keinerlei *a priori* Information vorliegen, d.h. man kann im schlechtesten Fall als Startstruktur einen ausgestreckten Strang oder eine *random coil* Struktur annehmen.

Die andere Methode, der 2-Stufen Algorithmus, basiert auf dem Vergleich von simulierten und experimentellen *Peakformen*. Da die *Peakformen* bis zu einem gewissen Grade strukturabhängig sind (Linienbreiten, Volumina), muss die Proteinstruktur genauer bekannt sein als bei KNOWNOE. Die Stärke dieses Algorithmus ist, dass man mit einer partiellen sequentiellen Zuordnung auskommt und die restlichen unbekannten Zuordnungen vom Algorithmus automatisch bestimmt werden.

2.3.1 Zuordnung mit Bayes (KNOWNOE)

Die automatische Zuordnung mit KNOWNOE gliedert sich in zwei Teile:

1. Zuordnung von NOE-Signalen allein mit Hilfe der chemischen Verschiebungen
2. Im Falle von Mehrdeutigkeiten Berechnung der wahrscheinlichsten Zuordnung

Zunächst wird das experimentelle NOESY-Spektrum automatisch vorverarbeitet, d.h. es findet ein automatisches *Peak Picking* statt, die *Peaks* werden integriert und es wird eine automatische Signal- und Artefakterkennung durchgeführt. Diese Teilschritte werden in Kapitel 3.1.3 und 3.1.5 näher erläutert. Die Zuordnung der NOE-Signale anhand der sequentiellen Zuordnung führt aufgrund der Entartung von chemischen Verschiebungen meist zu Mehrdeutigkeiten, d.h. in den wenigsten Fällen findet eine eindeutige Zuordnung statt. Aus der Analyse von 326 Proteinstrukturen wurden statistische Tabellen erzeugt, die Wahrscheinlichkeitsverteilungen der Volumina von Atompaaren repräsentieren, im Folgenden kurz VPD⁴ genannt. Es wurden 1238 VPDs für möglichst alle relevanten Zuordnungsmöglichkeiten berechnet. Für ein NOE-Signal, das N verschiedene Zuordnungen A_i besitzt, kann die bedingte Wahrscheinlichkeit $P(A_i, a|V_0)$ berechnet werden, dass

⁴volume probability distributions, über $V \propto r^{-6}$

die Zuordnung A_i mindestens $a \cdot V_0$ des Volumens V_0 erklärt. Eine Zuordnung A_k mit einer Wahrscheinlichkeit $P(A_k, a|V_0) \geq P_{min}$ wird als eindeutig bestimmt. Mit einer Liste von eindeutig zugeordneten Signalen wird dann ein Satz von Strukturen berechnet, die als Input für weitere Iterationsschritte verwendet werden.

2.3.1.1 Die allgemeine Strategie von KNOWNOE

In Abbildung 2.3 wird die Strategie von KNOWNOE gezeigt. Wichtig ist eine bestmögliche

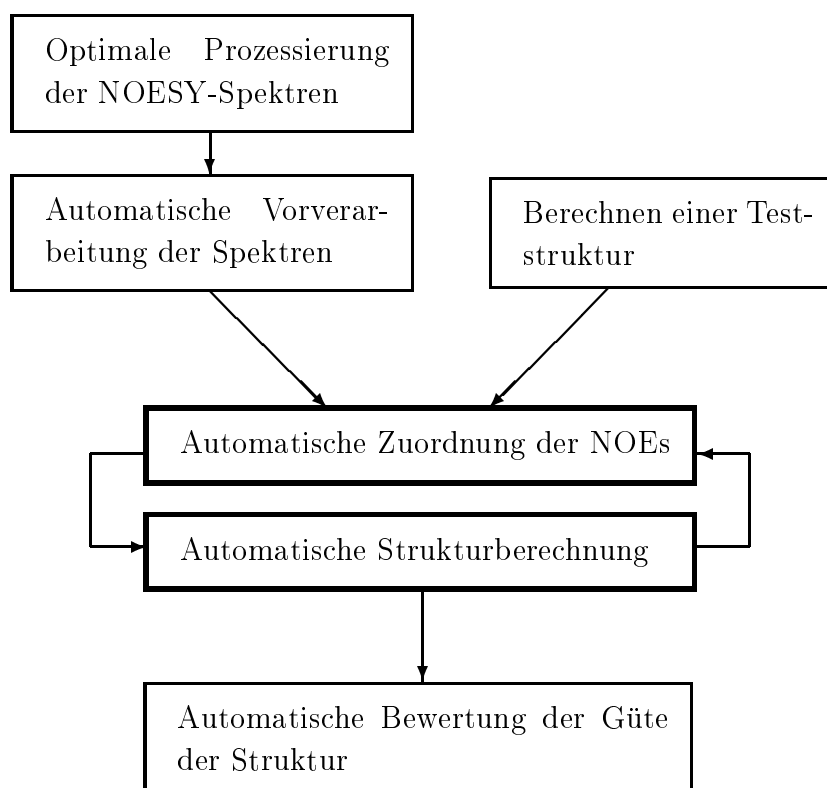


Abbildung 2.3: Die allgemeine Strategie von KNOWNOE. Die hervorgehobenen Teilschritte, die NOE-Zuordnung und die Strukturrechnung, werden in der Abbildung 2.4 detaillierter erklärt.

Prozessierung der experimentellen NOESY-Spektren. Darauf soll hier nicht näher eingegangen werden und auf entsprechende Literatur verwiesen werden [37], [76], [96] und [57], da diese Verfahren nicht in AUREMOL implementiert wurden. Anschließend wird ein automatisches *Peak Picking* durchgeführt, eine Trennung von Signalen und Artefakten erfolgt und eine automatische Volumenintegration der NOE-Signale wird gemacht, was in Abbildung 2.3 als automatische Vorverarbeitung bezeichnet wird. Diese Teilschritte wer-

den wie bereits erwähnt in Kapitel 3.1.3 und 3.1.5 vorgestellt. In einem nächsten Schritt wird eine Startstruktur benötigt, was z.B. ein ausgestreckter Strang sein kann. Hier wird also keinerlei Vorwissen über die Proteinstruktur benötigt. Die nächsten beiden Schritte in der Abbildung 2.3, die automatische Zuordnung von NOE-Signalen und die Strukturrechnung, werden im folgenden Kapitel näher ausgeführt. Der letzte Schritt besteht in der Beurteilung der Güte der Struktur mit Hilfe von R-Faktoren und wird in Kapitel 2.4 dargestellt.

2.3.1.2 Der Algorithmus von KNOWNOE

Der Algorithmus von KNOWNOE ist in Abbildung 2.4 zu sehen. In einem ersten Schritt werden die experimentellen NOE-Signale anhand der gegebenen chemischen Verschiebungen zugeordnet. Dies führt in den meisten Fällen zu Mehrdeutigkeiten, was folgende Gründe hat

- Überlagerung mehrerer Signale aufgrund einer Entartung chemischer Verschiebungen
- Maximum eines experimentellen Signals kann nur mit einer begrenzten Genauigkeit ermittelt werden
- Unvollständige Zuordnungsliste

Da die chemischen Verschiebungen in verschiedenen Spektren leicht variieren, werden Toleranzwerte in allen Dimensionen definiert, innerhalb derer im experimentellen NOESY-Spektrum Signale zugeordnet werden. Dies führt zu einer noch geringeren Zahl an eindeutig zugeordneten Signalen.

Um die Zahl an Mehrdeutigkeiten zu reduzieren, verwendet man strukturelle Informationen, falls eine gefaltete Struktur vorliegt. Man benutzt im ersten Schritt nur die Zuordnungen, deren zugehörige Kerne in der Startstruktur folgendes Kriterium erfüllen

$$r_{ij} \leq D_{max} \quad (2.34)$$

wobei r_{ij} der Abstand der Kerne i und j ist und D_{max} der Distanzschwellwert, also die maximale Distanz der beiden Kerne. Diese Zuordnungen werden zusammen mit dem Volumen in der sogenannten U-Liste gespeichert, die alle nicht zugeordneten Signale enthält. Wenn es nur eine Zuordnung für einen NOE gibt, so wird diese Zuordnung in die A-Liste transferiert, die nur eindeutig zugeordnete Signale enthält.

Nun wird für alle mehrdeutigen Zuordnungen die bedingte Wahrscheinlichkeit $P(A_i, a|V_0)$ berechnet, dass mindestens $a \cdot V_0$ des experimentellen Volumens V_0 nur von einer Zuordnung erklärt wird. Man nutzt dabei die Tatsache aus, dass das Volumen eines NOEs und die richtige Zuordnung nicht voneinander unabhängig sind.

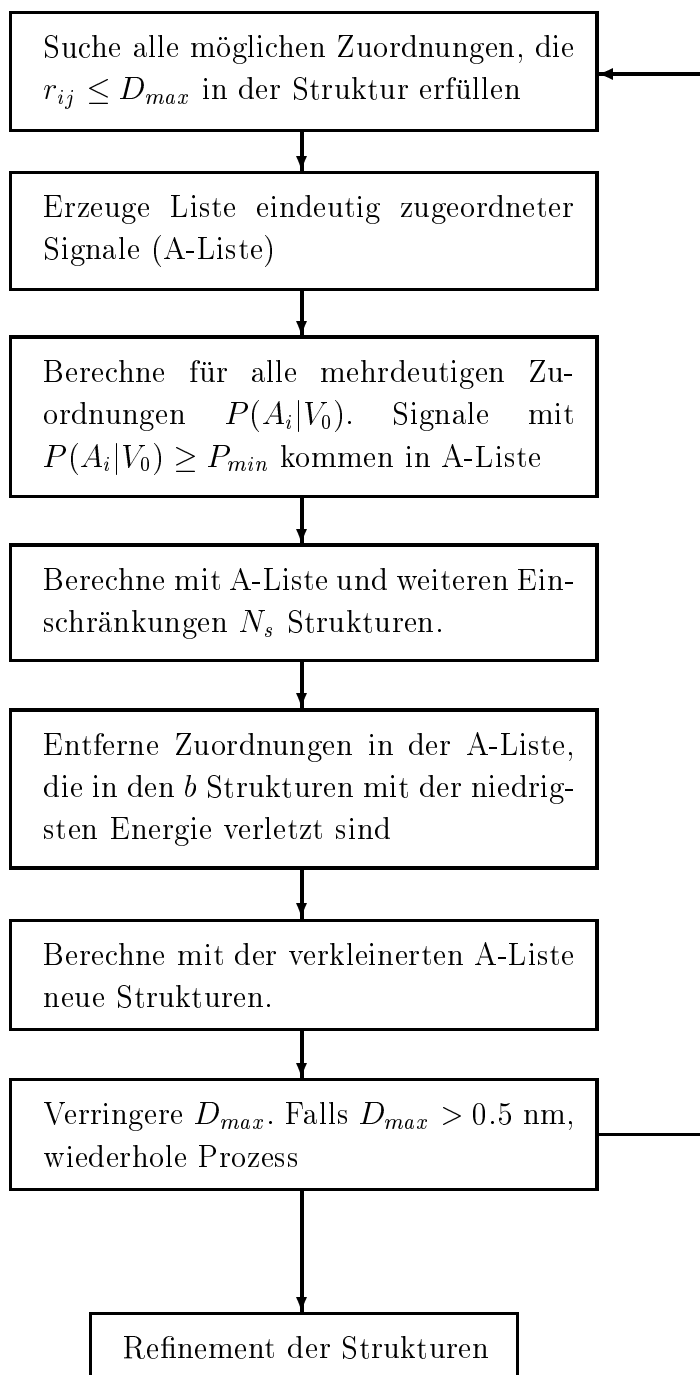


Abbildung 2.4: Schematische Darstellung des Algorithmus von KNOWNOE

Wenn die Ungleichung

$$P(A_i, a|V_0) \geq P_{min} \quad (2.35)$$

erfüllt ist, so wird die Zuordnung A_i in die A-Liste kopiert.

Mit den so erhaltenen eindeutigen Zuordnungen aus der A-Liste und weiteren experimentellen Einschränkungen, wie z.B. J-Kopplungen, werden nun N_s Strukturen berechnet. Davon werden die b Strukturen mit der niedrigsten Gesamtenergie ausgewählt und überprüft, ob einige NOE Einschränkungen systematisch verletzt sind. Diese werden dann aus der A-Liste entfernt. Mit diesen Zuordnungen werden neue Strukturen berechnet. Der Distanzschwellwert D_{max} wird verringert und eine neue A-Liste wird erstellt, wie oben beschrieben. Dieser Vorgang wird solange iteriert, bis $D_{max} = 0.5$ nm ist. Dies stellt die Detektionsgrenze des Kernoverhausereffekts im Experiment dar. Anschließend werden die Strukturen noch verfeinert (*refinement*).

2.3.1.3 Berechnung der wahrscheinlichsten Zuordnung

Hier soll kurz gezeigt werden wie die bereits erwähnte bedingte Wahrscheinlichkeit $P(A_i, a|V_0)$ berechnet wird. Mit Hilfe des Bayesschen Theorems [26] und [27] lässt sich die Wahrscheinlichkeit, dass mehr als $a \cdot V_0$ des experimentellen Volumens V_0 durch die Zuordnung A_i erklärt wird, schreiben

$$P(A_i, a|V_0) = \frac{P(A_i, a) P(V_0|A_i, a)}{\sum_{i=1}^{N_{ab}} P(A_i, a) P(V_0|A_i, a)} \quad (2.36)$$

Im einfachsten Fall, falls nur eine Zuordnung A_1 existiert, gilt für die *a priori* Wahrscheinlichkeit $P(A_1, a) = 1$ und $P(A_i, a, i > 1) = 0$. Man erhält damit

$$P(A_i, a|V_0) = 1 \quad (2.37)$$

Für den Fall, dass zwei mögliche Zuordnungen A_1 und A_2 existieren, werden die *a priori* Wahrscheinlichkeiten wie folgt berechnet

$$P(A_1, a) = \int_{V_0=0}^{\infty} \int_{V_1=aV_0}^{V_0} p_1(V_1) p_2(V_0 - V_1) dV_1 dV_0 \quad (2.38)$$

$$P(A_2, a) = \int_{V_0=0}^{\infty} \int_{V_2=aV_0}^{V_0} p_1(V_0 - V_2) p_2(V_2) dV_2 dV_0 \quad (2.39)$$

wobei $p_1(V)$ und $p_2(V)$ auf 1 normierte Wahrscheinlichkeitsdichten sind, dass man für ein Atompaar mit den Zuordnungen A_1 und A_2 das Volumen V findet. Die Wahrscheinlich-

keiten $P(V_0|A_i, a)$ werden berechnet mit

$$P(V_0|A_1, a) = \int_{V_1=aV_0}^{V_0} p_1(V_1) p_2(V_0 - V_1) dV_1 \quad (2.40)$$

$$P(V_0|A_2, a) = \int_{V_2=aV_0}^{V_0} p_1(V_0 - V_2) p_2(V_2) dV_2 \quad (2.41)$$

Der Fall für drei Zuordnungsmöglichkeiten kann ganz analog erweitert werden. Die Möglichkeit von mehr als drei Zuordnungen wurde in diesen Algorithmus nicht integriert.

2.3.1.4 Wahrscheinlichkeitstabellen

Notwendig für die Berechnung sind daher die bereits erwähnten Wahrscheinlichkeitsverteilungen p_i . Dazu wurden statistische Analysen an 326 Proteinstrukturen aus der PDB Datenbank [11] durchgeführt. Es wurden nur NMR-Strukturen wasserlöslicher Proteine ohne paramagnetisches Zentrum oder größerer Kofaktoren verwendet. Ebenso durften die Proteine nicht mit DNA oder RNA komplexiert sein. Eine Liste der verwendeten Strukturen findet sich bei [79].

Da die Zahl der möglichen Zuordnungen sehr groß ist und nur eine begrenzte Zahl an Strukturen zur Verfügung stand, wurden die Zuordnungen in Klassen zusammengefasst. Diese Klassen sind nicht mehr abhängig von der absoluten Position der Kerne in der Sequenz. Die Zuordnungen HA 2/HN 3 bzw. HA 5/HN 6 werden z.B. zur Zuordnungsklasse HA i/HN i+1 integriert. Dadurch kann die Wahrscheinlichkeitsverteilung des Volumens einer Zuordnung durch die Volumenwahrscheinlichkeitsverteilung der zugehörigen Klasse genähert werden.

Für 1238 verschiedene Zuordnungsklassen wurden Distanzwahrscheinlichkeitsverteilungen berechnet und diese auf 1 normiert. Die Distanzen r_{ij} zwischen den Kernen i und j wurden über die Beziehung

$$V_{ij} = c_v r_{ij}^{-6} \quad (2.42)$$

in Volumen umgerechnet. Der Proportionalitätsfaktor c_v wurde auf 0.047 nm^6 gesetzt, wodurch eine Distanz von 0.6 nm einem relativen Volumen von 1.0 entspricht. Zwei Beispiele für die so erhaltenen VPDs sind in Abbildung 2.5 zu sehen. Man sieht, dass sich die Wahrscheinlichkeitsverteilungen der beiden Zuordnungsmöglichkeiten deutlich unterscheiden. Dadurch ist gewährleistet, dass man bei mehrdeutigen Zuordnungen eine eindeutige Präferenz bei der Berechnung der Wahrscheinlichkeiten erhält.

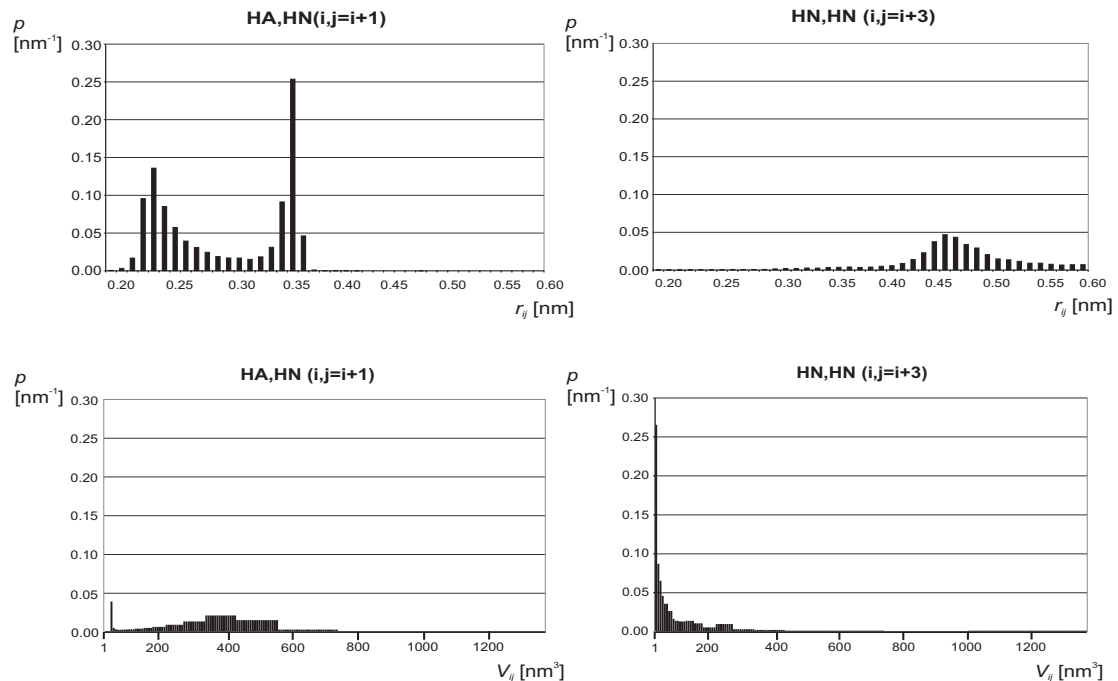


Abbildung 2.5: Oben: Wahrscheinlichkeitsverteilungen der Abstände zwischen HA der Aminosäure an der Stelle i in der Sequenz und HN an der Stelle $i+1$ in der Sequenz, bzw. HN an der Stelle i in der Sequenz und HN an der Stelle $i+/-3$. Unten: Die entsprechenden Wahrscheinlichkeitsverteilungen der nach Gleichung 2.42 aus den Distanzen berechneten Volumina.

2.3.2 Zuordnung durch Vergleich von simulierten und experimentellen Spektren (2-Stufen Algorithmus)

Anders als bei dem Algorithmus KNOWNOE ist hier die Zielsetzung, fehlende chemische Verschiebungen in den NOESY-Spektren zu finden, also primär keine Zuordnung von NOE-Signalen. Ausgehend von der bekannten bzw. vermuteten Proteinstruktur, entsprechend des molekülorientierten Konzepts von AUREMOL, versucht der Algorithmus durch den Vergleich von simulierten und experimentellen NOESY-Spektren Resonanzfrequenzen zuzuordnen. Dabei wird, anders als bei den herkömmlichen Verfahren, die Kenntnis der sequentiellen Zuordnung der Resonanzen nicht zwingend vorausgesetzt. Im Idealfall sollte es möglich sein ohne dieses Wissen unbekannte chemische Verschiebungen zu bestimmen, in der Praxis hat sich herausgestellt, dass wenigstens ein Teil der sequentiellen Zuordnung bekannt sein muss. Das heißt, man verwendet als Leitfaden zur Auswertung der NOESY-Spektren eine vermutete Struktur des Proteins und eine partielle sequentielle Zuordnung. Der entwickelte Algorithmus, der schematisch in Abbildung 2.6 zu sehen ist, besteht aus zwei Stufen. In der ersten Stufe wird mit Hilfe der sequentiellen Zuordnung versucht, die-

se Zuordnungen an das experimentelle NOESY-Spektrum anzupassen. Das Problem bei

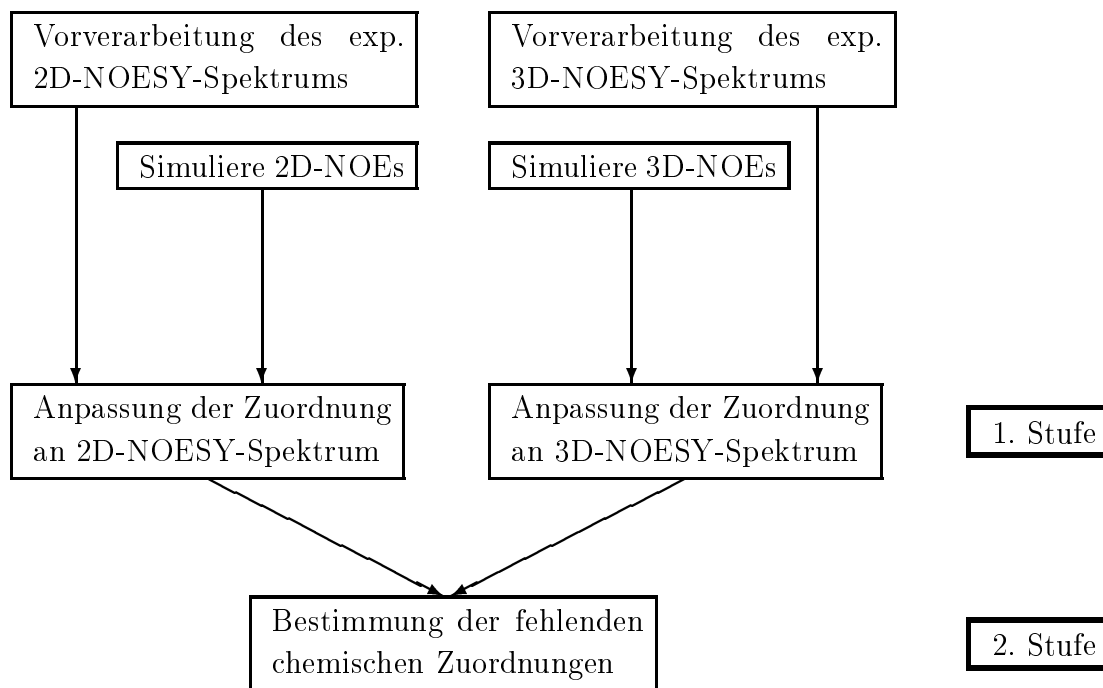


Abbildung 2.6: Schematische Darstellung des 2 Stufen Algorithmus. Nach einer automatischen Vorverarbeitung der NOESY-Spektren werden die aus der sequentiellen Zuordnung bekannten chemischen Verschiebungen an das experimentelle NOESY-Spektrum angepasst (1. Stufe). In der 2. Stufe werden die fehlenden Zuordnungen durch eine parallele Optimierung der NOESY-Spektren ermittelt.

der Analyse von NMR-Spektren liegt unter anderem auch darin, dass sich die chemischen Verschiebungen der Kerne abhängig vom Experiment unterscheiden. Das bedeutet, daß die chemischen Verschiebungen, die man aus der Auswertung verschiedener Korrelations-experimente gewonnen hat, sich teilweise deutlich von denen in den NOESY-Spektren unterscheiden können. Deshalb wird in einem ersten Schritt versucht, die chemischen Verschiebungen der partiellen Zuordnung auf die NOESY-Spektren anzupassen. Dies geschieht für das 2D- und 3D-NOESY-Spektrum getrennt. Die zweite Stufe versucht dann, die restlichen fehlenden chemischen Zuordnungen zu finden, wobei eine parallele Verarbeitung der beiden NOESY-Spektren stattfindet. Als Input für die zweite Stufe wird die an die NOESY-Spektren angepasste Zuordnung verwendet, die als Ergebnis der ersten Stufe ermittelt wurde. In beiden Stufen wird zur Optimierung ein *Threshold Accepting* Algorithmus verwendet, der im Folgenden kurz erklärt wird.

2.3.2.1 *Threshold Accepting*

Der *Threshold Accepting* Algorithmus (TA) ist ein heuristisches Optimierungsverfahren und stellt eine Weiterentwicklung des *simulated annealing* Algorithmus [56] dar. Er wurde von Dück und Scheuer [33] bei IBM entwickelt, um die Anordnung der Leiterbahnen auf Computerchips zu optimieren. Mittlerweile wurde dieser Algorithmus für die Lösung zahlreicher Problemstellungen erfolgreich eingesetzt, als Beispiel seien nur das *Traveling salesman* Problem oder die Optimierung von Produktionsabläufen in Unternehmen erwähnt.

Der Ausgangspunkt ist ein System mit zahlreichen freien Parametern. Der Zustand des Systems, der durch die Wahl der Parameter gegeben ist, soll nun in Bezug auf eine bestimmte Größe, meist Energiefunktion genannt, optimiert werden. Beim *Travelling salesman* Problem ist die Energiefunktion einfach die Länge der Wegstrecke, die der Kaufmann auf seiner Route zurücklegt. Das Ziel ist es, die Energiefunktion in Abhängigkeit der freien Parameter zu optimieren bzw. zu minimieren. In der Praxis sind die Problemstellungen derart komplex, dass eine *brute force* Methode nicht alle Kombinationsmöglichkeiten innerhalb einer endlichen Zeitspanne durchprobieren kann, um das globale Minimum zu finden.

Die Idee des TA ist recht einfach und ähnlich der Idee, die beim *simulated annealing* Algorithmus eingesetzt wird. Man startet mit einer zufälligen Anfangskonfiguration des Systems, d.h. einer zufälligen Vorgabe der Parameter, die als σ bezeichnet werden soll. Nun wird eine neue Konfiguration τ als kleine Änderung der alten Konfiguration σ gewählt. Das bedeutet, dass τ in einer kleinen Nachbarschaft von σ liegt. Deshalb nennt man den TA auch einen lokalen Suchalgorithmus⁵. Nun wird für beide Konfigurationen die Energiefunktion berechnet und die Differenz

$$\Delta E = E(\tau) - E(\sigma) \quad (2.43)$$

gebildet. Ist nun diese Energiedifferenz kleiner als ein bestimmter Schwellwert oder *Threshold* Th , so wird die neue Konfiguration akzeptiert, ansonsten verworfen. Die Wahrscheinlichkeit, dass die neue Konfiguration τ akzeptiert wird, ist durch

$$p(\sigma \rightarrow \tau) = \Theta(Th - \Delta E) \quad (2.44)$$

gegeben, wobei Θ die Heavisidefunktion bzw. Stufenfunktion ist. Man sieht, dass abhängig von der Wahl des *Threshold* Th , auch schlechtere Konfigurationen akzeptiert werden. Ähnlich wie die Temperatur beim *simulated annealing* langsam erniedrigt wird, wird beim TA der *Threshold* Th langsam erniedrigt, um das System einzufrieren, d.h. in ein meist lokales Minimum zu führen. Es gibt verschiedene Möglichkeiten, den *Threshold* zu reduzieren. Die lineare Methode adaptiert den *Threshold* nach einer gewissen Zahl von Iterationen

⁵auf Neudeutsch auch *refined local search algorithm*

folgendermaßen an

$$Th_{i+1} = Th_i - a \quad (2.45)$$

während die logarithmische Methode Th folgendermaßen anpasst

$$Th_{i+1} = a \cdot Th_i \quad (2.46)$$

wobei a im Intervall $[0, 1]$ liegen muss. Der Vorteil des TA und des *simulated annealing* Algorithmus gegenüber einem einfachen Gradientenabstiegsverfahren, der nur Verbesserungen akzeptiert⁶, liegt darin, dass man ausgehend vom Startpunkt nicht im nächstgelegenen Minimum steckenbleibt, sondern abhängig vom *Anfangsthreshold* und der Zahl der Iterationen tiefer gelegene lokale Minima und im Idealfall auch das globale Minimum finden kann.

2.3.2.2 Der Kernalgorithmus

Wie bereits erwähnt zerfällt der Algorithmus in zwei Stufen, die im Prinzip denselben Programmablauf haben. In Abbildung 2.7 ist eine schematische Darstellung gezeigt. Der Kernalgorithmus ist auch unabhängig von der Dimension des NOESY-Spektrums. In einem ersten Schritt werden mit der Berechnung der Relaxationsmatrix, s. Kapitel 2.2, NOE-Signale simuliert. Es werden die Volumina und Linienbreiten der Protonen berechnet und damit die Linienform der einzelnen NOE-Signale bestimmt. Die Relaxationsraten, die diese Größen bestimmen, sind strukturabhängig und setzen voraus, dass die vermutete Struktur des Proteins bereits relativ genau bekannt ist. Mit Hilfe der partiellen Zuordnung wird ein NOESY-Spektrum simuliert, wobei die *Peakformen* durch die Rückrechnung bekannt sind und die Position der Signale durch die Zuordnung gegeben ist. Die fehlenden Zuordnungen werden anfangs zufällig auf chemische Verschiebungen gesetzt, die aus der Vorverarbeitung der experimentellen NOESY-Spektren ermittelt wurden, näheres in 2.3.2.4.

Die Erzeugung des simulierten Spektrums ist dimensionsabhängig. Im Fall eines 2D-NOESY-Spektrums wird sowohl das experimentelle als auch das simulierte Spektrum als 1:1 Abbild, also nicht komprimiert, im Speicher gehalten. Bei 3D-Spektren ist dies nicht mehr möglich. Das experimentelle Spektrum wird in komprimierter Form im Speicher gehalten, s. Abschnitt 2.3.2.3, während für das simulierte Spektrum für jeden simulierten *Peak* Überlappbereiche definiert werden, in denen die *Peakform* des simulierten *Peaks* und evtl. Überlagerungen mit anderen Signalen berechnet werden.

Nun startet ein iterativer Prozess, in dem die Zuordnung leicht variiert wird, s. Abschnitt 2.3.2.4. Das bedeutet, dass sich die Position der entsprechenden NOE-Signale

⁶wird oft auch *Greedy* genannt

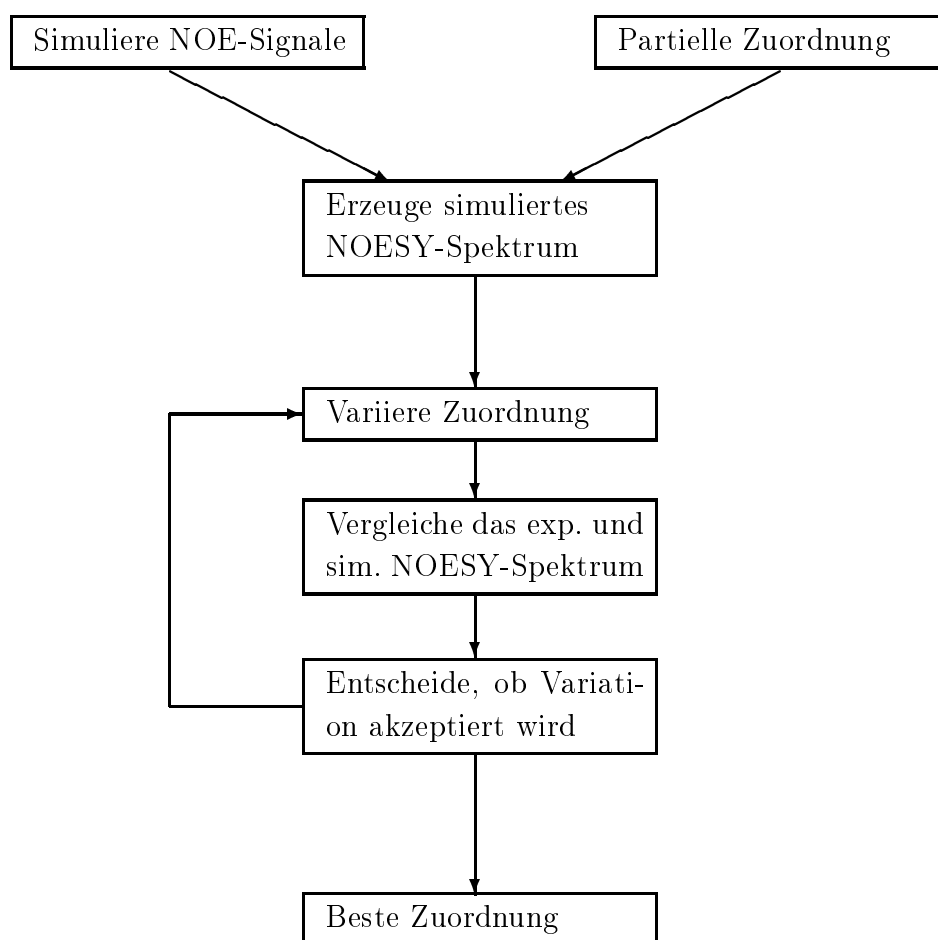


Abbildung 2.7: Schematische Darstellung des Kernalgorithmus der 2-Stufen Optimierung

ändert. Im simulierten 2D-NOESY-Spektrum werden die davon betroffenen Signale verschoben, im 3D-Spektrum müssen dazu die Überlappbereiche der simulierten *Peaks* neu berechnet werden. Anschließend wird ein Vergleich mit dem experimentellen NOESY-Spektrum durchgeführt. Dazu wird eine Energiefunktion berechnet, die angibt, wie ähnlich die beiden Spektren sind, d.h. es wird für alle simulierten Signale ein Matchfaktor ermittelt, der die Ähnlichkeit mit den im experimentellen Spektrum korrespondierenden Flächen bzw. Würfel im 3D Fall angibt. Um Rechenzeit zu sparen, werden nur die lokalen Veränderungen im Spektrum neu berechnet, also die Bereiche in denen die Signale verschoben worden sind und sich Überlappregionen geändert haben. Dieser Matchfaktor ist

gegeben durch

$$M_p = \cos(\vec{p}_p^{sim}, \vec{p}_p^{exp}) = \frac{\vec{p}_p^{sim} \cdot \vec{p}_p^{exp}}{|\vec{p}_p^{sim}| \cdot |\vec{p}_p^{exp}|} \quad (2.47)$$

wobei \vec{p}_p^{sim} die *Peakform* des *Peaks* p im simulierten Spektrum mit der Grundfläche des isolierten *Peaks* p ist und \vec{p}_p^{exp} die Intensitätsverteilung der korrespondierenden Fläche im experimentellen Spektrum⁷. Bei einer großen Übereinstimmung zwischen der Linienform des simulierten *Peaks* und der Intensitätsverteilung im experimentellen Spektrum liegt M_p nahe 1, bei schlechter Übereinstimmung nahe 0. An dieser Stelle unterscheiden sich die beiden Optimierungsstufen. In der 1. Stufe wird dieser Matchfaktor nur dann berechnet, wenn an exakt derselben Stelle im experimentellen Spektrum ein *gepickter Peak* zu finden ist, ansonsten ist $M_p = 0$, d.h. der Matchfaktor lautet allgemein

$$M_p = a_p \cos(\vec{p}_p^{sim}, \vec{p}_p^{exp}) \quad (2.48)$$

mit

$$a_p = \begin{cases} 0 & \text{falls experimenteller Peak nicht vorhanden, 1. Stufe.} \\ 1 & \text{falls experimenteller Peak vorhanden, 1. Stufe.} \\ 1 & \text{2. Stufe.} \end{cases} \quad (2.49)$$

In der 2. Stufe wird der Cosinus also immer berechnet, unabhängig davon, ob an der Position des simulierten Signals im experimentellen Spektrum ein *Peak* ist. Der Unterschied rührt daher, dass in der 1. Stufe die sequentielle Zuordnung an die NOESY-Spektren angepasst werden soll, d.h. die *Peaks* im Idealfall an exakt denselben Stellen im simulierten und experimentellen Spektrum liegen sollen. Statt des berechneten Cosinus könnte man sich auch überlegen, einen Intensitätsvergleich der *Peakformen* zu machen, allerdings hätte das den Nachteil, dass die Absolutwerte der Intensitäten den Matchfaktor beeinflussen würden. Durch das Cosinuskriterium ist der Matchfaktor unabhängig von den Absolutwerten der Intensitäten und damit auch unabhängig vom strukturabhängigen Volumen des Signals.

Die Gesamtenergie berechnet sich jetzt zu

$$E = N_{sim} - \sum_{p=1}^{N_{sim}} M_p \quad (2.50)$$

wobei N_{sim} die Anzahl der simulierten NOE-Signale ist. Die Energie E liegt im Intervall $[0, 2N_{sim}]$, wobei eine niedrige Energie eine gute Übereinstimmung zwischen dem experimentellen und simulierten Spektrum darstellt und umgekehrt. Nun wird entschieden, ob die Variation der Zuordnung akzeptiert oder verworfen wird. Dabei wird der TA

⁷Im 3D-Spektrum spricht man dann nicht von einer zweidimensionalen Fläche sondern von einem dreidimensionalen Würfel

Algorithmus eingesetzt, d.h. es wird geprüft, ob die Energiedifferenz der beiden Konfigurationen, einmal vor der Variation und einmal danach, kleiner als ein Schwellwert ist oder nicht. Wenn ja, wird die Variation akzeptiert, ansonsten verworfen. Dieser Prozess wird solange iteriert, bis es zu keiner Verbesserung mehr kommt, d.h. die bestmögliche Übereinstimmung zwischen dem experimentellen und simulierten Spektrum erreicht ist. Der Schwellwert wird dabei nach einer einstellbaren Anzahl von Iterationen nach der im Kapitel 2.3.2.1 erwähnten linearen Methode reduziert.

2.3.2.3 Vorverarbeitung der experimentellen Spektren

Wie in der schematischen Darstellung des 2-Stufen Algorithmus in Abbildung 2.6 zu sehen ist, müssen die experimentellen 2D- und 3D-NOESY-Spektren automatisch vorverarbeitet werden. Die einzelnen Teilschritte sind in Abbildung 2.8 dargestellt. Beim

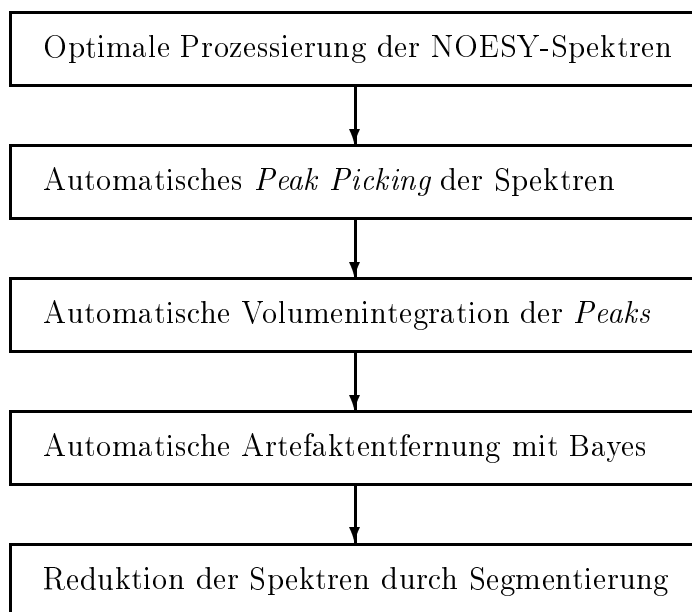


Abbildung 2.8: Die automatische Vorverarbeitung der experimentellen NOESY-Spektren beim 2-Stufen Algorithmus. Wichtig ist die optimale Prozessierung, um die Multipllettstruktur der Signale sichtbar zu machen und die Komprimierung der Spektren im letzten Teilschritt.

Vergleich zwischen simulierten und experimentellen NOESY-Spektren ist ein wichtiges Kriterium die Multipllettstruktur der NOE-Signale. Sie beeinflusst in starker Weise die *Peakform* und geht in den oben erwähnten Matchfaktor ein, der die Ähnlichkeit verschiedener *Peakformen* misst. Außerdem ist die Multipllettstruktur, also die Aufspaltung der

Signale aufgrund von J-Kopplungen mit anderen Kernen, nur wenig strukturabhängig⁸, d.h. unabhängig von der Modellstruktur, die für die Simulation der NOE-Signale verwendet wird. Üblicherweise werden NMR-Spektren so prozessiert, dass ein möglichst großes Signal-Rausch-Verhältnis erzielt wird. Dabei werden die Signale durch Multiplikation des FIDs mit einer geeigneten Fensterfunktion verbreitert, was zu einer Entfernung der Multipllettstruktur im Spektrum führt. Für den 2-Stufen Algorithmus bedeutet das, dass durch die Anwendung einer anderen Fensterfunktion die Multipllettstruktur im fouriertransformierten Spektrum sichtbar gemacht werden muss. Der Nachteil ist, dass sich das Signal-Rausch-Verhältnis verschlechtert.

Anschliessend werden in AUREMOL die üblichen Vorverarbeitungsschritte vorgenommen, wie das automatische *Peak Picking*, die automatische Volumenintegration der Signale und die automatische Artefaktentfernung durch die Bayessche Analyse. Diese Teilschritte sind im Kapitel 3.1.3 bzw. 3.1.5 näher erklärt. Im letzten Teilschritt erfolgt eine Reduktion der experimentellen Spektren. Es werden alle verbliebenen Signale vom *Peakmaximum* ausgehend in Richtungen parallel zu den Frequenzachsen des Spektrums nach unten segmentiert, d.h. die Intensitätswerte ausgelesen. Ein Abbruchkriterium für das Weitergehen in einer Richtung ist das Unterschreiten der Intensität eines vom Benutzer vorzugebenen Grenzwertes. Dieser Grenzwert ist der prozentuale Teil der *Peakintensität* am Maximum. Das andere Kriterium ist das Erreichen einer Begrenzungsbox, die auch vom Benutzer vorzugeben ist. Die Dimensionen der Box werden auch vom Benutzer vorgegeben und sollten mindestens der doppelten maximal erwarteten experimentellen Linienbreiten⁹ in Hz entsprechen, um in jedem Fall die komplette *Peakform* zu erhalten. Die so erhaltenen Intensitätswerte werden nun als reduzierte Spektren für die 2D- und 3D-NOESY-Spektren in unterschiedlicher Form weggeschrieben. Im Falle der 2D-Spektren ist eine Komprimierung der Spektrendaten aufgrund der in der Praxis oft maximalen Spektrengroße von etwa 16-32 MB Speicherbedarf¹⁰ nicht erforderlich, d.h. es wird eine Datei *peakshapes2D_exp.dat* angelegt, die neben den *Peakinformationen*, wie den *Peakpositionen*, auch ein komplettes 1:1 Abbild des reduzierten 2D-Spektrums enthält.

Für die 3D-Spektren, die bis zu 100 oder 200 MB groß werden, ist es nicht praktikabel ein 1:1 Abbild des reduzierten Spektrums zu erzeugen, sondern es werden die segmentierten Intensitätswerte mit den dazugehörigen drei Koordinaten für jede Frequenzachse des Spektrums in der Datei *peakshapes3D_exp.dat* weggeschrieben. Die Intensitäten sind dabei nach den beiden Achsen mit den höchsten digitalen Auflösungen in aufsteigender Reihenfolge sortiert. Beispiele für die Dateien *peakshapes2D_exp.dat* und *peakshapes3D_exp.dat* sind im Ergebnisteil 3.4.2.1 gezeigt. Der schnelle Zugriff auf die experimentellen Intensitätswerte eines 3D-Spektrums bei der automatischen Zuordnung läuft über eine zweidimensionale

⁸Die ³J-Kopplungen sind abhängig vom dihedralen Winkel Θ

⁹Linienbreite bei halber Höhe des Signals

¹⁰32 MB würde einem Spektrum von 4096×2048 oder 8192×1024 Punkten Auflösung entsprechen

Hashtabelle ab, die eine Indizierung der beiden Frequenzachsen mit der größten digitalen Auflösung vornimmt. Das Prinzip soll in Abbildung 2.9 veranschaulicht werden. Das

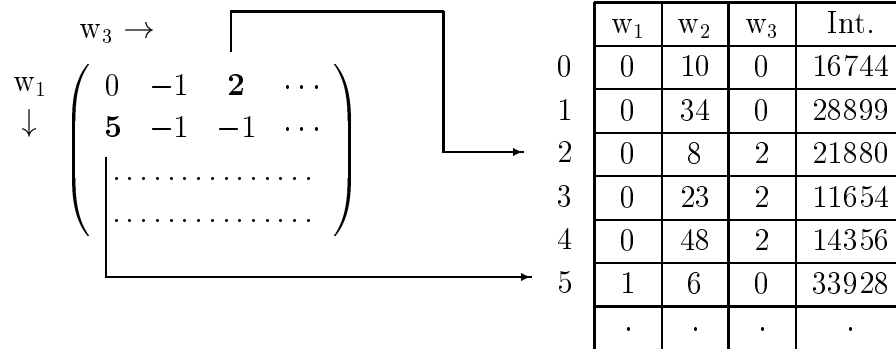


Abbildung 2.9: Das Prinzip der Hashtabellen. Links ist eine zweidimensionale Hash-tabelle zu sehen, aufgebaut aus den Achsen w_1 und w_3 mit der höchsten digitalen Auflösung. Die Einträge verweisen als Index auf den Array der experimentellen Intensitätswerte, dargestellt in der rechten Tabelle. Die Koordinaten in w_1 , w_2 und w_3 sind in aufsteigender Reihenfolge sortiert.

Beispiel ist für den Fall gezeigt, dass w_1 und w_3 eine digitale Auflösung von N_1 bzw. N_3 besitzen. Es wird eine zweidimensionale Hashtabelle H mit $N_1 \times N_3$ Einträgen angelegt. Diese verweisen auf den Array E der experimentellen Intensitätswerte, in welchem die drei Koordinaten und die Intensität für jeden Pixel des reduzierten Spektrums abgelegt sind. Möchte man z.B. die Intensität an der Stelle $(w_1, w_2, w_3) = (0, 8, 2)$ wissen, genügt es, in der Hashtabelle $H(0, 2)$ nachzusehen. Diese verweist in diesem Fall auf den Index 2 im Array E , s. Abbildung 2.9. Das bedeutet, dass ab dieser Stelle im Array E alle Intensitäten mit den Koordinaten $(w_1, w_2, w_3) = (0, x, 2)$ abgelegt sind. Nun genügt noch ein Durchlauf bis zur Stelle $(0, y, 2)$ im Array E mit $y > 8$, um zu sehen, ob in E ein Eintrag mit den Koordinaten $(0, 8, 2)$ zu finden ist. Für den Fall, dass man die Intensität an einer Position $(w_1, w_2, w_3) = (a, b, c)$ wissen will, deren $w_1 = a$ und $w_3 = c$ Koordinaten im Array E nicht enthalten sind, ist in der Hashtabelle $H(a, c) = -1$ hinterlegt. Der Vorteil der Hashtabellen ist, dass man nicht bei jeder Abfrage den gesamten Array E durchlaufen muss. Eine kleine Abschätzung mag dies verdeutlichen. Nehmen wir an, es handelt sich um ein 3D-Spektrum mit einer Auflösung von $256 \times 64 \times 512$ Punkten und das reduzierte Spektrum besteht aus 100000 Intensitätswerten. Dann müsste man also im schlimmsten Fall alle 100000 Einträge in E absuchen, im Fall der Hashtabellen maximal 64.

2.3.2.4 Variation der Zuordnungen

Die Aufgabe des Algorithmus besteht darin, fehlende Zuordnungen in den experimentellen NOESY-Spektrum zu finden. Dazu werden die Zuordnungen solange variiert, bis eine bestmögliche Übereinstimmung zwischen experimentellem und simuliertem Spektrum erreicht ist. Entweder wird ein Atom oder Spin zufällig ausgewählt, welches eine neue Zuordnung zugewiesen bekommt oder es werden zwei verschiedene Atome selektiert, deren Zuordnungen ausgetauscht werden. Die Zuweisung einer neuen Zuordnung ist abhängig von der Stufe der Optimierung und ist für die 1. Stufe schematisch in Abbildung 2.10 dargestellt und für die 2. Stufe in Abbildung 2.11. Die Parameter in diesen Abbildungen

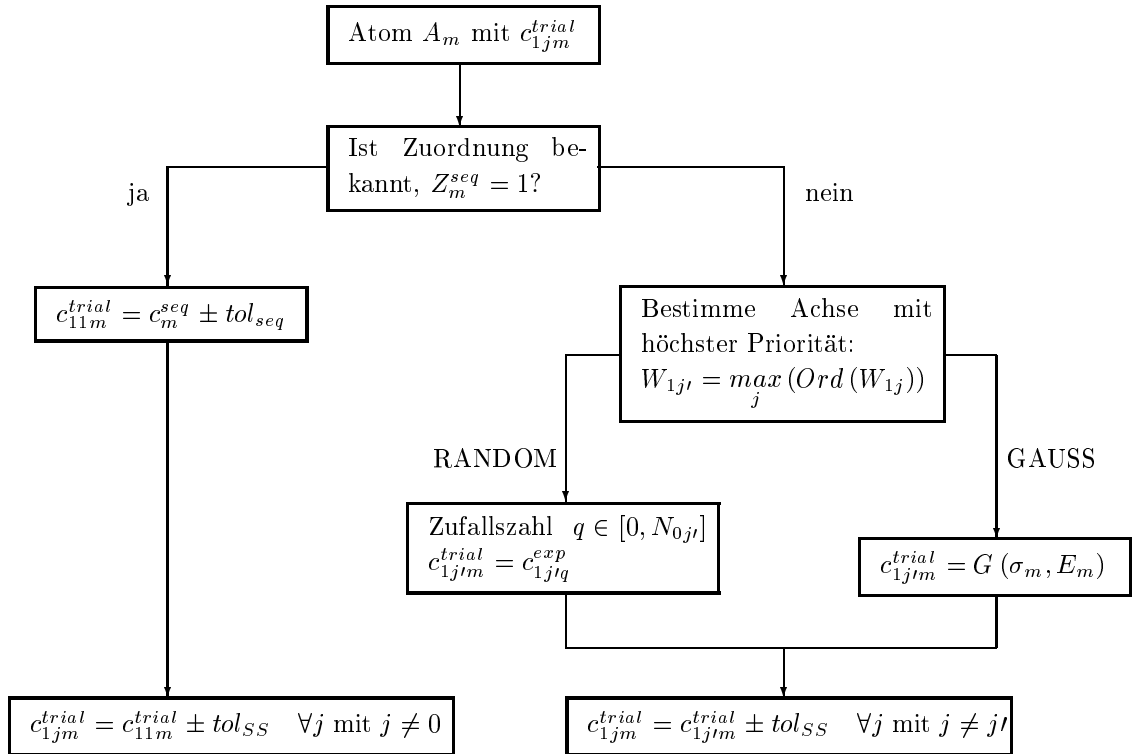


Abbildung 2.10: Das Flussdiagramm zeigt, wie die Zuweisung einer neuen Zuordnung eines Atoms A_m in der 1. Stufe der Optimierung abläuft.

haben folgende Bedeutung:

- A_m bezeichnet das Atom bzw. den Spin, dessen Zuordnung geändert werden soll, mit $m = 1, \dots, M$ und M der Gesamtzahl der Spins.

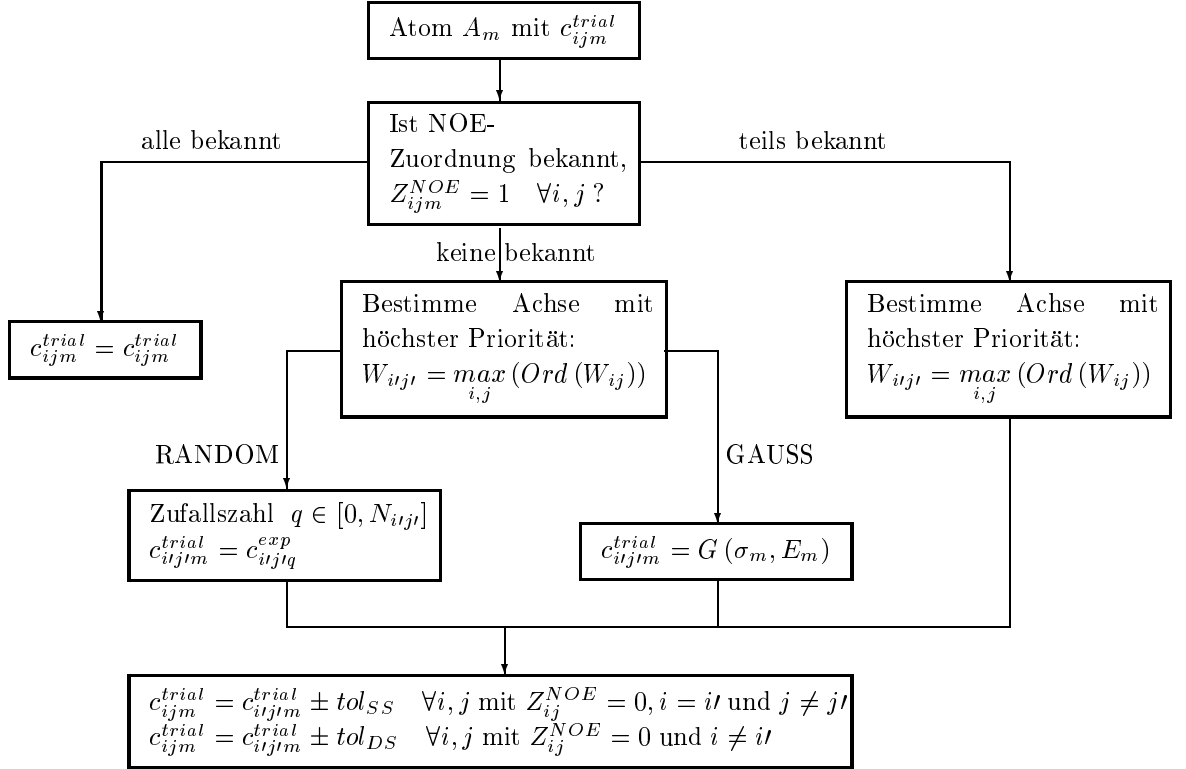


Abbildung 2.11: Das Flussdiagramm zeigt, wie die Zuweisung einer neuen Zuordnung eines Atoms A_m in der 2. Stufe der Optimierung abläuft.

- W_{ij} nummeriert die Frequenzachsen des Spektrums i und dessen Dimension j durch, mit $i = 1, \dots, N$, $j = 1, \dots, J_i$. N ist die Zahl der Spektren und J_i ist die Dimension des Spektrums i .
- c_{ijq}^{exp} sind die aus dem experimentellen Spektrum i der Frequenzachse j ermittelten chemischen Verschiebungen in ppm, mit $q = 1, \dots, N_{ij}$. N_{ij} ist die Anzahl der experimentellen chem. Verschiebungen der Frequenzachsen j der Spektren i .
- c_m^{seq} ist die von der sequentiellen Zuordnung bekannte chemische Verschiebung des Atoms A_m in ppm.
- c_{ijm}^{trial} ist die Zuordnung des Atoms A_m im Spektrum i der Achse j in ppm.
- $Z_m^{seq} = \begin{cases} 0 & \text{falls Zuordnung des Atoms } A_m \text{ nicht bekannt ist,} \\ 1 & \text{falls Zuordnung des Atoms } A_m \text{ bekannt ist.} \end{cases}$

- $Z_{ijm}^{NOE} = \begin{cases} 0 & \text{falls Zuordnung von } A_m \text{ im Spektrum } i \text{ der Achse } j \text{ unbekannt ist,} \\ 1 & \text{falls Zuordnung von } A_m \text{ im Spektrum } i \text{ der Achse } j \text{ bekannt ist.} \end{cases}$
- σ_m ist die Standardabweichung der chemischen Verschiebung des Atoms A_m in der Datenbank von AUREMOL, s. Abschnitt 3.2.
- E_m ist die mittlere chemische Verschiebung des Atoms A_m in der Datenbank von AUREMOL, s. Abschnitt 3.2. Es ist geplant, E_m und σ_m optional anhand der Teststruktur in AUREMOL berechnen zu lassen und diese Werte statt der Datenbankwerte zu verwenden.
- tol_{SS} ist die maximal erlaubte Abweichung in ppm der Zuordnung verschiedener Frequenzachsen innerhalb desselben Spektrums.
- tol_{DS} ist die maximal erlaubte Abweichung in ppm der Zuordnung in verschiedenen Spektren.
- tol_{seq} ist die maximal erlaubte Abweichung in ppm der sequentiellen Zuordnung zur Zuordnung in den NOESY-Spektren.

In der 1. Stufe wird getestet, ob die sequentielle Zuordnung des Atoms A_m bekannt ist, d.h. ob $Z_m^{seq} = 1$ ist. Falls ja, wird die bekannte chemische Verschiebung c_m^{seq} der sequentiellen Zuordnung leicht variiert, maximal um den Wert tol_{seq} , und A_m neu zugeordnet. Damit trägt man der Tatsache Rechnung, dass die sequentielle Zuordnung in der Datentabelle und die Zuordnung im NOESY-Spektrum leicht unterschiedlich sind. Da sich die chemische Verschiebung eines Signals auch innerhalb eines Spektrums in Abhängigkeit von der Frequenzachse ändern kann, wird dies im Algorithmus durch den Toleranzwert tol_{SS} berücksichtigt, der die chemischen Verschiebungen um maximal diesen Wert variieren kann.

In der 2. Stufe können mehrere Spektren gleichzeitig optimiert werden. Es wird für jedes Spektrum der Energieterm aus Gleichung 2.50 berechnet und die Gesamtenergie

$$E = \sum_{i=1}^N \left(N_{sim}^i - \sum_{p=1}^{N_{sim}^i} a_p \cos(\vec{p}_p^{sim}, \vec{p}_p^{exp}) \right) \quad (2.51)$$

minimiert. N ist die Zahl der Spektren und N_{sim}^i ist die Zahl der simulierten Signale im Spektrum i . Die Berechnung der Energiefunktion wurde optional noch um einen Bestrafungsterm weit entfernter Zuordnungen erweitert. Der Matchfaktor eines simulierten Signals p , dessen Zuordnung(en) mehr als das Doppelte der Standardabweichung σ_m von der mittleren chemischen Verschiebung E_m abweicht, wird um einen Faktor p_v verringert,

so dass die Energiefunktion nun lautet

$$E = \sum_{i=1}^N \left(N_{sim}^i - \sum_{p=1}^{N_{sim}^i} a_p \cos(\vec{p}_p^{sim}, \vec{p}_p^{exp}) \right. \\ \left. \times \left[1 - p_v \sum_{j=1}^{J_i} \Theta(|c_{ijQ(p,j)}^{trial} - E_{Q(p,j)}| - 2\sigma_{Q(p,j)}) \right] \right) \quad (2.52)$$

wobei $p_v \in [0, 1]$, J_i die Dimension von Spektrum i und Θ die Heavisidefunktion ist. $Q(p, j)$ zerlegt den *Peaknamen* des simulierten Signals p entsprechend der Dimension j und liefert einen Index m zurück, der auf das korrespondierende Atom A_m verweist. Ist $p_v = 0$, so vereinfacht sich Gleichung 2.52 zu dem Ausdruck in Gleichung 2.51. Wird eine Bestrafung $p_v > 0$ eingeführt, so wird der berechnete Matchfaktor im Falle weit entfernter Zuordnungen verringert.

Ein weiterer Toleranzwert tol_{DS} wurde implementiert, der angibt, um wieviel sich die chemische Verschiebung eines Signals in verschiedenen Spektren maximal ändern darf. Dadurch löst man das Problem, dass unterschiedliche Spektren in der Praxis nie unter absolut identischen Bedingungen aufgenommen werden können und aufgrund unterschiedlichster Ursachen Variationen der chemischen Verschiebungen auftreten können. Falls nur ein Spektrum optimiert wird, entfällt die Verwendung dieses Toleranzwertes.

Bei beiden Optimierungsstufen wird für den Fall, dass Zuordnungen unbekannt sind, die Frequenzachse W_{ijl} mit der höchsten Priorität bestimmt, also $W_{ijl} = \max_{i,j} (Ord(W_{ij}))$.

Die Achse mit dem kleinsten Frequenzintervall in ppm besitzt die höchste Priorität und umgekehrt. Dies vermeidet folgendes Problem bei der Zuordnung:

Angenommen in Stufe 1 bzw. 2 wird ein 3D-NOESY-Spektrum optimiert. Die detektierten Kerne im 3D-Spektrum seien ^1H , ^{15}N und HN. Üblicherweise liegen die chemischen Verschiebungen der HN Signale im Bereich von 6 bis 11 ppm. Falls nun für die Zuordnung eines HN Atoms eine experimentelle chemische Verschiebung aus dem Bereich der Frequenzachse des ^1H Kerns, der üblicherweise zwischen 0 und 11 ppm liegt, gewählt wird, so wäre es denkbar, dass eine chemische Verschiebung von 2.4 ppm dem HN zugewiesen wird, was in diesem Fall keinen Sinn macht, da die möglichen experimentellen Verschiebungen nur im Bereich von 6 bis 11 ppm liegen. Dieses Verfahren wurde im Hinblick auf die Tests mit künstlichen Datensätzen verwendet. Im Falle experimenteller Datensätze muss diese Methode nicht notwendigerweise zum Erfolg führen, da z.B. durch spezielle Aufnahmetechniken Signale in 3D-Spektren durch Rückfaltung an eine andere Position im Spektrum wandern können. Außerdem können ganze Spinsysteme in dem einen experimentellen Spektrum fehlen, im anderen aber nicht. Dies alles könnte zur Folge haben, dass man durch die oben erklärte Auswahl der Achse mit der höchsten Priorität das erlaubte Intervall an chemischen Verschiebungen zu stark einschränkt.

Nach der Bestimmung der Achse mit der höchsten Priorität gibt es zwei Möglichkeiten,

dem Atom neue Verschiebungen zuzuordnen. Im ersten Fall, der in den Abbildungen 2.10 und 2.11 mit *RANDOM* bezeichnet wird, wird dem Atom zufällig eine der experimentell ermittelten Verschiebungen zugeordnet, im zweiten Fall, mit *GAUSS* bezeichnet, wird eine experimentelle Verschiebung zugeordnet, die gaussverteilt [92] um den Erwartungswert E_m mit der Standardabweichung σ_m streut.

Das Problem, dass je nach Aufnahmebedingungen der experimentellen Spektren, zwischen 10 und 20% der Spinsysteme in den Spektren komplett fehlen, wird mit dem beschriebenen 2-Stufen Algorithmus noch nicht abgefangen. Der Algorithmus wird versuchen, diese Spinsysteme experimentellen Verschiebungen zuzuordnen, was zu mehrdeutigen Zuordnungen führt. Es ist geplant, dafür eine weitere 3. Optimierungsstufe zu entwickeln, die nach einem erstmaligen Durchlauf des 2-Stufen Algorithmus versucht, mehrdeutige Zuordnungen, d.h. diese fehlenden Spinsysteme, zu finden. Diesen könnten dann *Dummy*-Zuordnungen, also in den Spektren nicht vorhandene experimentelle Verschiebungen, zugewiesen werden. Daran anschliessen würde sich ein weiterer Optimierungslauf der 2. Stufe, wobei nun nicht mehr versucht wird, die gefundenen, in den experimentellen Spektren fehlenden Spinsysteme, neu zuzuordnen.

2.4 R-Werte zur Qualitätsbeurteilung von NMR-Strukturen

Es gibt verschiedene Methoden, die Qualität von NMR-Strukturen zu messen. RMSD-Werte¹¹ geben an, wie stark die berechneten Strukturen des Strukturbündels von der gemittelten NMR-Struktur abweichen und sind nur ein Maß für die Genauigkeit nicht aber für die Richtigkeit der Struktur. Im Ramachandran Plot, der die Verteilung der diedralen Winkel Θ und Φ [73] jeder Aminosäure des Moleküls darstellt, können verbotene Winkelkonformationen sichtbar gemacht werden. Als weiteres Qualitätsmerkmal kann man die Zahl der verletzten Winkel- oder Abstandseinschränkungen in der berechneten Struktur heranziehen. Nachteil der genannten Methoden, die detaillierter in [63] nachzulesen sind, ist, dass man damit kein Maß dafür hat, wie gut die berechneten Strukturen die experimentellen NMR-Daten erklären.

Ähnlich wie in der Röntgenstrukturanalyse [18] kann man auch in der NMR-Spektroskopie einen R-Faktor¹² einführen, der experimentelle und berechnete NMR-Parameter miteinander vergleicht und damit ein direktes Maß für die Qualität der NMR-Struktur liefert. R-Faktoren werden neben der Qualitätsbeurteilung von Strukturen sowohl zum Verfeinern (*refinement*) von Strukturen und Dynamikuntersuchungen¹³ verwendet als auch zum Testen wie gut ein Strukturmodell die experimentellen Daten erklärt. Einen Überblick findet man bei [66], [48], [86], [62], [5], [15], [14], [43], [87], [100], [74], [19], [24], [28] und [25]. Bis auf [25], in der R-Faktoren von dipolaren Kopplungen berechnet werden, werden in all diesen Fällen experimentelle NOE-Signale mit den entsprechenden rückgerechneten NOE-Signalen verglichen.

In AUREMOL wurde das Softwaretool RFAC [45], [55] integriert, das eine vollautomatische Bestimmung der R-Faktoren berechneter Strukturen ermöglicht. Im Gegensatz zu den zitierten Beispielen findet dabei vor der R-Wert Berechnung eine automatische Zuordnung der experimentellen NOE-Signale statt. Ein neuer R-Faktor wurde entwickelt, der zusätzlich nicht zugeordnete Signale berücksichtigt, was eine deutlich bessere Bewertung von NMR-Strukturen ermöglicht. Im Folgenden soll der Algorithmus von RFAC vorgestellt werden, sowie die verschiedenen R-Faktoren, die implementiert wurden.

¹¹root mean square deviation

¹²residual indices

¹³Vorteilhaft, wenn keine ¹⁵N editierte Probe verfügbar ist und somit keine T_1 -, T_2 - und hetNOE-Messungen möglich sind.

2.4.1 Algorithmus von RFAC

Die automatische R-Faktor Bestimmung gliedert sich in zwei Teile

- Vergleich des experimentellen mit dem rückgerechneten NOESY-Spektrum
- Berechnung der R-Faktoren

In Abbildung 2.12 ist der Programmfluss schematisch dargestellt. Im ersten Teil wird

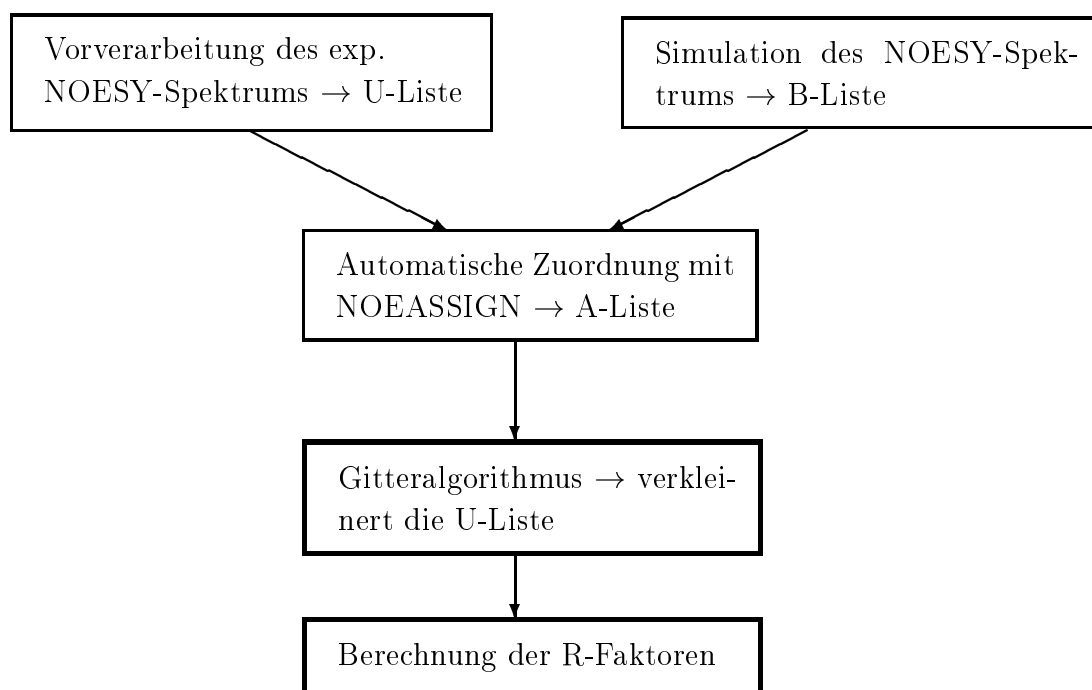


Abbildung 2.12: Das Konzept von RFAC, die beiden hervorgehobenen Teilschritte am Ende stellen den Algorithmus von RFAC dar.

ausgehend von der berechneten Struktur, deren Qualität gemessen werden soll, und der sequentiellen Zuordnung ein NOESY-Spektrum simuliert, s. Kapitel 2.2. Dies liefert eine Liste der Positionen und Volumina der simulierten *Peaks* (B-Liste). Beim experimentellen NOESY-Spektrum wird automatisch ein *Peak Picking* durchgeführt und die erhaltenen Signale integriert, d.h. die Volumina bestimmt. Nähere Informationen finden sich im Kapitel 3.1.3. Zusätzlich wird eine automatische Artefaktbestimmung durchgeführt. Dabei werden Wahrscheinlichkeiten p_i berechnet, ob ein *Peak* ein echtes Signal oder ein Rauschsignal ist, s. Abschnitt 3.1.5. Diese Wahrscheinlichkeiten werden als Gewichtungsfaktoren bei

der R-Wert Berechnung benötigt. Damit erhält man eine zweite Liste, die U-Liste, mit experimentellen *Peaks*, die die Positionen, Volumina und Wahrscheinlichkeiten dieser *Peaks* enthält.

Die experimentellen Signale der U-Liste werden nun automatisch mit dem Zuordnungsprogramm NOEASSIGN [39], das Bestandteil von RELAX [40] und damit auch AUREMOL ist, den simulierten Signalen der B-Liste zugeordnet. Der Algorithmus versucht, eine Zuordnung anhand der sequentiellen Zuordnung zu finden. Dabei wird für jedes simulierte Signal geprüft, ob an der entsprechenden Stelle innerhalb eines Suchradius von 0.01 ppm ein experimentelles Signal existiert. Wenn es mehrere Zuordnungsmöglichkeiten gibt, so wird anhand eines Maximum Likelihood Kriteriums entschieden, welche Zuordnung die wahrscheinlichste ist. Die zugeordneten experimentellen Signale werden in der A-Liste abgelegt.

Mit diesen drei Listen (A-Liste, U-Liste und B-Liste) werden nun die R-Faktoren berechnet. Dazu wird zuerst mittels eines Gitteralgorithmus versucht, die Zahl der unzugeordneten Signale weiter zu verringern. Man nimmt dabei an, dass die sequentielle Zuordnung richtig und komplett ist. Es wird an Stellen innerhalb eines Suchradius von 0.01 ppm im experimentellen Spektrum nach unzugeordneten Signalen gesucht, an denen simulierte Signale liegen. Anschliessend wird untersucht, ob einem experimentellen Signal mehrere simulierte zugeordnet wurden, falls ja, werden die dazugehörigen experimentellen und simulierten Volumina gemittelt [55]. Nach diesen Teilschritten werden die R-Werte berechnet.

2.4.2 R-Faktoren

R-Werte bzw. R-Faktoren stellen ein Maß dar, wie exakt die experimentellen NMR-Parameter durch simulierte oder rückgerechnete NMR-Parameter erklärt werden können. Der R-Faktor ist definiert als normierte Standardabweichung zwischen berechneten und experimentellen Strukturparametern

$$R = \sqrt{\frac{\sum (|F_{sim} - F_{exp}|)^2}{\sum |F_{exp}|^2}} \quad (2.53)$$

wobei F_{sim} und F_{exp} eine beliebige strukturabhängige simulierte bzw. experimentelle Größe darstellen kann. Im Fall von RFAC handelt es sich um das Volumen eines NOE-Signals, das abstandsabhängig und damit strukturabhängig ist. Im Fall einer idealen Übereinstimmung der experimentellen und simulierten Daten ist $R = 0$, andernfalls ist $R > 0$.

In der NMR-Spektroskopie hängen die Absolutwerte der experimentellen Größen von nicht genau bekannten Parametern ab, d.h. man muss die experimentellen oder simulierten Daten skalieren, um sie vergleichen zu können. Wie bei [39] nachzulesen, ist ein sinnvoller

Skalierungsfaktor

$$sf = \frac{\sum F_{exp} \cdot F_{sim}}{\sum F_{sim}^2} \quad (2.54)$$

Verwendet man nun die zugeordneten experimentellen NOE-Signale und als experimentelle Größen deren Volumina, so ergibt sich als R-Faktor

$$R_1 = \sqrt{\frac{\sum_{i \in A} (V_{exp,i} - sf \cdot V_{sim,i})^2}{\sum_{i \in A} V_{exp,i}^2}} \quad (2.55)$$

Der Nachteil dieses R-Faktors ist, dass er durch die kurzreichweitigen NOEs, die für die Strukturinformation weniger wichtig sind, dominiert wird. Dies liegt daran, dass kurzreichweitige NOE-Signale durch die in erster Näherung gültige Beziehung $V \propto r^{-6}$ stärker zu R_1 beitragen als langreichweitige Signale. Deshalb ist es günstiger, das Volumen V durch die Größe V^α zu ersetzen, was im Falle von $\alpha = -1/6$ zu einem distanzabhängigen R-Faktor führt [43]. Man erhält einen neuen R-Faktor

$$R_2 = \sqrt{\frac{\sum_{i \in A} (V_{exp,i}^\alpha - sf_\alpha \cdot V_{sim,i}^\alpha)^2}{\sum_{i \in A} V_{exp,i}^{2\alpha}}} \quad (2.56)$$

Ebenso muss der Skalierungsfaktor sf geändert werden

$$sf_\alpha = \frac{\sum_{i \in A} V_{exp,i}^\alpha \cdot V_{sim,i}^\alpha}{\sum_{i \in A} V_{sim,i}^{2\alpha}} \quad (2.57)$$

Die Bestimmung der Wahrscheinlichkeit $p_{exp,i}$, dass ein experimenteller *Peak* ein echtes Signal oder Rauschen ist, ermöglicht es, einen neuen R-Faktor R_3 zu definieren

$$R_3(\alpha) = \sqrt{\frac{\sum_{i \in A} (V_{exp,i}^\alpha - sf_\alpha \cdot V_{sim,i}^\alpha)^2 \cdot p_{exp,i}^2}{\sum_{i \in A} V_{exp,i}^{2\alpha} \cdot p_{exp,i}^2}} \quad (2.58)$$

Dadurch werden Rauschsignale, die ein kleines $p_{exp,i}$ haben, weniger stark gewichtet als echte Signale mit großem $p_{exp,i}$. Der R-Faktor R_3 misst bisher nur, wie gut die bereits zugeordneten Signale durch das Strukturmodell erklärt werden. Er ist aber noch unabhängig davon, wieviele Signale korrekt zugeordnet sind. Damit R_3 ein Maß für die Qualität der Struktur ist, müsste er kleiner werden, wenn mehr Signale zugeordnet werden.

Eine mögliche Erweiterung des R-Faktors um die Liste der nicht zugeordneten Signale (U-Liste) sieht folgendermaßen aus

$$R_4(\alpha) = \sqrt{\frac{\sum_{i \in A} (V_{exp,i}^\alpha - sf_\alpha \cdot V_{sim,i}^\alpha)^2 \cdot p_{exp,i}^2 + \sum_{i \in U} (V_{exp,i}^\alpha - sf_\alpha \cdot V_{noise}^\alpha)^2 \cdot p_{exp,i}^2}{\sum_{i \in A} V_{exp,i}^{2\alpha} \cdot p_{exp,i}^2 + \sum_{i \in U} V_{exp,i}^{2\alpha} \cdot p_{exp,i}^2}} \quad (2.59)$$

Die erste Summe im Zähler geht über alle zugeordneten Signale (A-Liste), während die zweite Summe über die Liste aller nicht zugeordneten Signale läuft (U-Liste). Man sieht, dass der R-Faktor kleiner wird, je mehr Signale korrekt zugeordnet sind, da dann der Anteil, der von der Summe über alle nicht zugeordneten Signale aus der U-Liste stammt, kleiner wird. Im Falle der nicht zugeordneten Signale wird als Vergleichsvolumen das Volumen V_{noise} eines typischen Rauschsignals verwendet. RFAC bietet dabei die Option an, das Rauschvolumen V_{noise} automatisch zu berechnen oder vom Benutzer vorzugeben. Bei der automatischen Berechnung wird das niedrigste simulierte Volumen als Rauschvolumen verwendet, dessen zugehöriger Abstand nicht die Detektionsgrenze überschreitet. Als Detektionsgrenze sind 0.5 nm festgelegt.

Da in R_4 die großen Distanzen und damit die langreichweitigen Signale dominieren, führt die Normierung in Gleichung 2.59 zu einer starken Abhängigkeit vom genauen Wert des Rauschvolumens V_{noise} . Dies kann man umgehen durch den leicht abgeänderten R-Faktor

$$R_5(\alpha) = \sqrt{\frac{\sum_{i \in A} (V_{exp,i}^\alpha - sf_\alpha \cdot V_{sim,i}^\alpha)^2 \cdot p_{exp,i}^2 + \sum_{i \in U} (V_{exp,i}^\alpha - sf_\alpha \cdot V_{noise}^\alpha)^2 \cdot p_{exp,i}^2}{\sum_{i \in A} V_{exp,i}^{2\alpha} \cdot p_{exp,i}^2 + \sum_{i \in U} (V_{exp,i}^\alpha - sf_\alpha V_{noise}^\alpha)^2 \cdot p_{exp,i}^2}} \quad (2.60)$$

Die R-Faktoren R_4 und R_5 in Gleichung 2.59 bzw. 2.60 sind ein geeignetes Maß dafür, wie gut die experimentellen Signale durch die simulierten Signale erklärt werden. R_5 stellt den bereits im vorherigen Kapitel erwähnten neuen R-Faktor dar, der anders als R_3 auch die nicht zugeordneten Signale mit berücksichtigt.

Auf der anderen Seite ist man vielleicht auch daran interessiert, wie gut die simulierten Daten durch die experimentellen erklärt werden. Ganz analog zum R-Faktor R_5 kann man einen neuen definieren

$$R_6(\alpha) = \sqrt{\frac{\sum_{i \in A} (V_{exp,i}^\alpha - sf_\alpha \cdot V_{sim,i}^\alpha)^2 \cdot p_{exp,i}^2 + \sum_{i \in U_I} (V_{noise}^\alpha - sf_\alpha \cdot V_{sim,i}^\alpha)^2 \cdot p_{exp,i}^2}{\sum_{i \in A} V_{exp,i}^{2\alpha} \cdot p_{exp,i}^2 + \sum_{i \in U_I} (V_{noise}^\alpha - sf_\alpha V_{sim,i}^\alpha)^2 \cdot p_{exp,i}^2}} \quad (2.61)$$

Die Summe über die nicht zugeordneten Signale läuft nun über eine neue Liste U_I , die alle simulierten nicht zugeordneten Signale mit $V_{sim,i} \geq V_{noise}$ enthält.

Die R-Faktoren R_3 , R_4 , R_5 und R_6 werden verwendet, um globale Merkmale der Struktur

zu bewerten, wie man bei den Ergebnissen im Kapitel 3.5 sehen wird. Es ist aber auch möglich, bestimmte Regionen innerhalb des Moleküls zu untersuchen, um z.B. zu sehen, wie gut bestimmte Sekundärstrukturelemente definiert sind. Dabei scheint R_3 dort gut geeignet zu sein, wo nur die zugeordneten Signale in die Berechnung mit einfließen.

R_3 wird auch verwendet, um R-Faktoren für jede einzelne Aminosäure zu berechnen. Dies kann bei der Suche nach falsch zugeordneten Signalen von Nutzen sein. Ausserdem kann man distanzabhängige R-Faktoren berechnen. Dabei werden die NOEs nach Distanzen sortiert und in 10 Klassen eingeteilt, beginnend mit 0.15 nm und einer Intervallgröße von 0.05 nm. Für jede Klasse an Signalen wird dann ein eigener R-Wert berechnet. Dies kann benutzt werden, um herauszufinden, ob Signale, die zu einer bestimmten Distanzklasse gehören, überproportional verletzt werden.

Kapitel 3

Ergebnisse

3.1 AUREMOL

3.1.1 Allgemeine Vorbemerkungen

Das Softwarepaket AUREMOL entstand in einer Kooperation mit der Firma Bruker Bio-Spin GmbH in Karlsruhe und dem Lehrstuhl Biophysik an der Universität Regensburg. Die Erstellung des grafischen Interfaces und zahlreicher Basisroutinen, wie dem Lesen und Visualisieren von n-dimensionalen NMR-Spektren wurden von der Fa. Bruker geleistet. Sie sind Teil eines allgemeinen Konzepts namens VIEWER, das die Basis zahlreicher kommerzieller Programme wie dem AMIX-VIEWER oder AURELIA-VIEWER von Bruker darstellt. Die dem VIEWER zugrundeliegende Philosophie ist es, einen graphischen Standard und eine möglichst konsistente Benutzerbedienung aller Softwarepakete von Bruker, einschließlich AUREMOL, bereitzustellen.

Die wissenschaftliche Komponente von AUREMOL, wie z.B. die Automatisierung der NMR-Strukturbestimmung, wurde am Lehrstuhl für Biophysik entwickelt. Zusätzlich wurden zahlreiche Routinen integriert, die bereits erfolgreich im Auswertungsprogramm AURELIA [85] implementiert waren. Als Beispiele seien die Routinen für das *Peak Picking* oder die Volumen Integration zu nennen, die in Kapitel 3.1.3 kurz vorgestellt werden. Die Kooperation mit Bruker sieht vor, AUREMOL als Nachfolger des erfolgreichen Programms AURELIA am Markt zu etablieren. Die bisherigen Beta Versionen von AUREMOL sind nicht kostenpflichtig, aber auf ein Jahr Laufzeit begrenzt. Ein kostenloser Download [3] ist möglich.

AUREMOL wurde in ANSI-C entwickelt und ist für folgende Plattformen verfügbar

- Windows NT/2000/XP

- IRIX 6.5 (SGI)
- Linux (Red Hat bzw. SuSE)

3.1.2 Grafik von AUREMOL

Ein *Screenshot* der aktuellen Beta Version von AUREMOL ist in Abbildung 3.1 zu sehen. Die grafische Oberfläche von AUREMOL ist frei konfigurierbar, die Kommandos und Be-

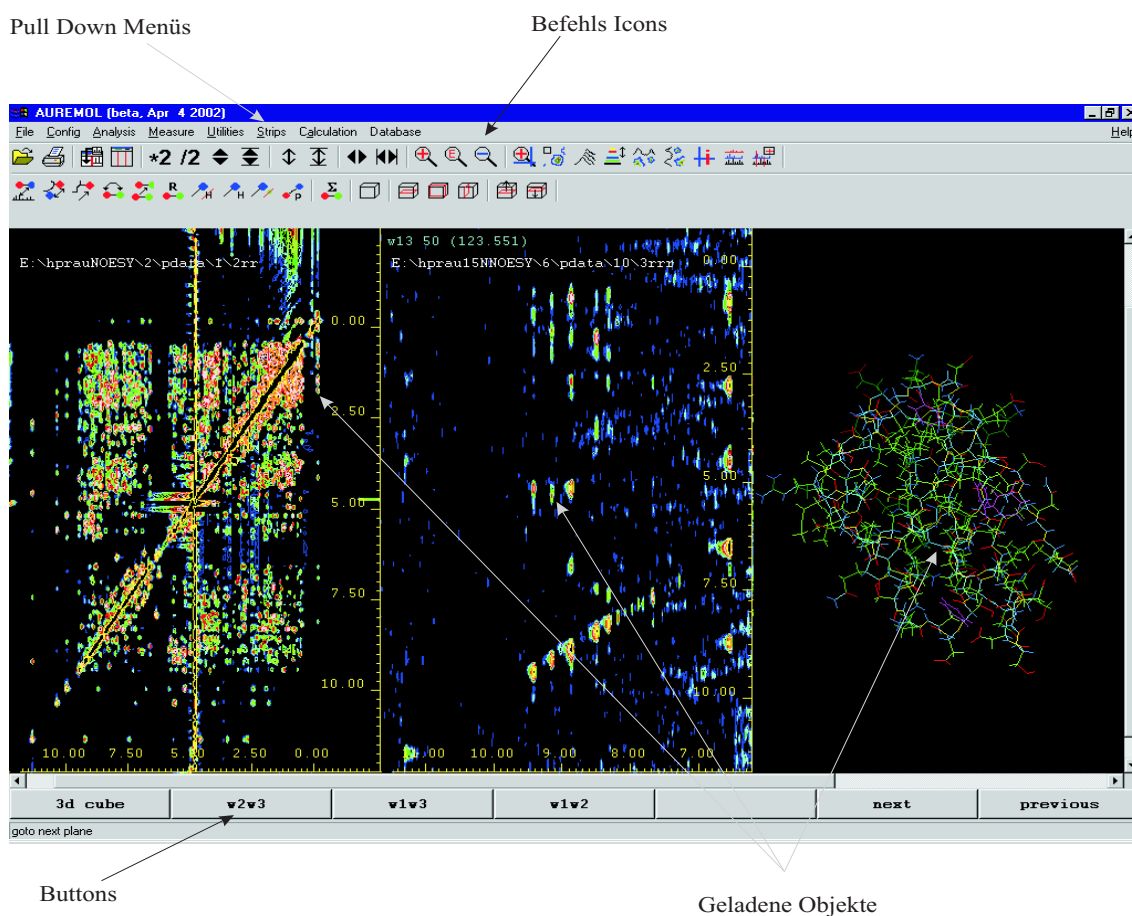


Abbildung 3.1: *Screenshot* von AUREMOL. Die grafische Oberfläche ist durch den Benutzer frei konfigurierbar, d.h. die Kommandos oder Befehle sind in den *Pulldown* Menüs enthalten oder als *Icons* oder *Buttons* abrufbar. Links: 2D-NOESY-Spektrum von HPr aureus, Mitte: Ebene oder *Plane* eines 3D-NOESY-Spektrums von HPr aureus, Rechts: Räumliche Struktur von HPr aureus.

fehle sind über *Icons* oder *Buttons* abrufbar oder als Einträge in den *Pulldown* Menüs

enthalten. Es können 1D-, 2D- und 3D-NMR-Spektren sowie Molekülstrukturen, die entweder im PDB [11] oder im MolMol [58] Format vorliegen, geladen und angezeigt werden. Das Format der NMR-Spektren muss dabei dem der Fa. Bruker entsprechen [17], welches die binären Spektrendaten im Submatrizenformat ablegt. Zusätzlich werden noch sogenannte Prozessierungsdateien benötigt, die spektrenspezifische Informationen wie z.B. die Referenzierung oder die Größe in jeder Dimension des Spektrums enthalten. Diese Dateien werden automatisch bei der Prozessierung der NMR-Spektren erzeugt [17].

Es können bis zu 128 verschiedene Objekte¹ gleichzeitig geladen und angezeigt werden. Ausserdem ist es möglich, NMR-Spektren gleicher Dimension zu überlagern. Diese Overlay-Technik ist z.B. nützlich beim Vergleich eines 2D-NOESY- und 2D-TOCSY-Spektrums, um manuell die sequentielle Zuordnung zu bestimmen oder auch beim Vergleich eines experimentellen und simulierten NOESY-Spektrums, um manuell experimentelle Signale zuzuordnen.

Grafische Kommandos wie das Vergrößern (*Zoomen*) oder das *Scrollen* innerhalb von Spektren können auf alle Spektren gleichzeitig oder nur auf das gerade aktive Spektrum angewandt werden. Dies soll hier nicht näher ausgeführt werden, sondern es wird auf den Eintrag im *Pulldown* Menü *Config* → *Correlation* im Anhang B verwiesen. In diesem Anhang sind alle Befehle von AUREMOL aufgelistet. Nähere Erläuterungen finden sich im Folgenden oder im Handbuch von AUREMOL, das der Beta Version des Programms beiliegt.

Die Icons in Abbildung 3.1 enthalten meist zusätzliche Kommandos, die oft benutzt werden. So sind dort die Zoomfunktion, das Setzen der Levels² in einem Spektrum oder das Anzeigen einer 1D Reihe oder Spalte in einem 2D-Spektrum für einen einfachen Linienfit zu finden. Ein weiterer nützlicher Befehl ist der Multicursor, der in mehreren gleichzeitig geöffneten NMR-Spektren miteinander korrelierte Fadenkreuze darstellt. Diese Funktion erleichtert den manuellen Vergleich zweier Spektren, um z.B. schnell Signale an gleichen Positionen zu finden. Schließlich sind bei den Icons auch noch alle Funktionen untergebracht, die sich auf geladene Molekülstrukturen beziehen, wie z.B. das Drehen der Struktur in Echtzeit, das farbige Markieren bestimmter Atomgruppen oder die Anzeige des Abstands zweier mit der Maus selektierter Atome.

3.1.3 Methoden zur Auswertung von NMR-Spektren

Zahlreiche Methoden zur manuellen Auswertung wurden in AUREMOL implementiert. Einige davon, wie z.B. das *Auto peak pick*, *Integration* und *Peak Quality*, sind auch für die automatische Vorverarbeitung von experimentellen NMR-Spektren einsetzbar. Im Folgenden werden die wichtigsten Routinen in der Reihenfolge beschrieben, in der sie bei der

¹Mit Objekten sind nD-NMR-Spektren oder Strukturen gemeint

²Levels sind die Höhenlinien, denen ein bestimmter Farbwert zugeordnet werden kann

Auswertung verwendet werden.

Auto Peak Pick:

Der erste Schritt bei der manuellen und automatischen Analyse ist das sog. *Auto Peak Pick* [85], das alle lokalen Maxima³ oberhalb eines bestimmten Grenzwertes, oft auch *Threshold* genannt, in einem n-dimensionalen NMR-Spektrum automatisch ermittelt. Diese Maxima, im Folgenden auch kurz *Peaks* genannt, sind Datenpunkte des Spektrums, deren Intensität größer ist als die der nächsten Nachbarpunkte. Der Grenzwert kann dabei manuell vom Benutzer vorgegeben werden oder, was derzeit noch in der Entwicklung ist, automatisch ermittelt werden.

Manipulate Peaks:

In den NMR-Spektren können interaktiv einzelne *Peaks* hinzugefügt oder entfernt werden. Ebenso ist es möglich, Gebiete in einem Spektrum zu definieren, in denen entweder alle *Peaks* innerhalb oder alle *Peaks* ausserhalb des Gebietes entfernt werden sollen. Die Aufgabe bei der Auswertung von NMR-Spektren besteht darin, die Signale richtig zuzuordnen. Daher ist es notwendig, die *Peaks* zu beschriften. In AUREMOL kann der Benutzer bis zu 10 Varianten für die Annotation eines Signals und für Kommentare vergeben. Da es mittlerweile eine Reihe an verschiedenen Nomenklaturen für die Atomnamen gibt, wurde eine Konvertierungsroutine, s. Anhang B unter *Utilities* → *Convert masterlist*, implementiert, die verschiedene Formate ineinander umwandeln kann. Als Beispiele seien nur das offizielle IUPAC Format [73] oder das Format des Strukturberechnungsprogramms CNS [20] erwähnt.

Eine weitere Methode erlaubt automatische Konsistenzüberprüfungen der vom Benutzer vergebenen *Peaknamen* bzw. *Peaklabel*. Prüfkriterien sind unter anderem die Existenz des Atoms in der zugehörigen Aminosäure und Konsistenz der koppelnden Atome mit dem zugrundeliegenden NMR-Experiment. Die Primärsequenz des Proteins muss dabei vom Benutzer vorgegeben werden, ebenso das NMR-Experiment, das wie andere spektrentypischen Parameter gesetzt werden kann. Die entsprechenden Kommandos sind im Anhang unter *Utilities* → *Load primary sequence*, *Utilities* → *Set processing parameter* und *Analysis* → *Annotate peaks* → *Check all peaklabels* zu finden.

Integration:

Die Integration der Volumina von NOE-Signalen liefert wichtige Strukturinformationen. Dazu wurde eine iterative Segmentierungsroutine [36] verwendet, wie sie bereits in AURELIA enthalten war. Diese Routine ist für 2D-, 3D- und 4D-Spektren ausgelegt. Die Segmentierung eines Signals startet am *Peakmaximum* und bewegt sich nach unten bis Datenpunkte benachbarter Signale oder der Segmentierungsschwellwert erreicht werden. Dieser Grenzwert, gemessen in Prozent vom *Peakmaximum*, muss vom Benutzer vorgegeben werden. Zusätzlich muss die maximale Integrationsbreite in Hz für alle Frequenzdimensionen angegeben werden, was sinnvollerweise die maximal zu erwartende experi-

³Es ist auch einstellbar, dass Minima *gepickt* werden.

mentelle Linienbreite ist. Die beschriebene Methode ist über das Kommando *Analysis* \rightarrow *Integrate* \rightarrow *Integrate peaks by shape*, s. Anhang B ausführbar. Eine vereinfachte Integration ist mit dem Befehl *Analysis* \rightarrow *Integrate* \rightarrow *Integrate peaks by area* möglich. Hier wird vom Benutzer das Integrationsgebiet interaktiv vorgegeben und die Datenpunkte einfach aufsummiert. Dies mag in Fällen extrem überlappender Signale sinnvoll sein, wenn die automatische Integrationsroutine diese nicht mehr sauber trennen kann.

Peak quality:

Ein Problem, das bei der Analyse von NMR-Spektren auftritt, sind Rausch- und Artefaktsignale. Speziell bei den automatischen Zuordnungsroutinen kann dies zu falschen Zuordnungen führen. Eine automatische Klassifizierung von 2D- [2], [98] und 3D-NMR-*Peaks* [94] in Signal- und Rauschpeaks wurde in AUREMOL implementiert. Diese Routine wird im Abschnitt 3.1.5 kurz vorgestellt.

Strips:

Strips sind schmale zweidimensionale Streifen in 3D-NMR-Spektren. Sie können entweder interaktiv durch den Benutzer oder automatisch aus dem Würfel eines 3D-Spektrums erstellt werden. *Strips* stellen ein geeignetes Hilfsmittel dar, die sequentielle Zuordnung eines Moleküls manuell zu bestimmen. Der Nutzen liegt darin, die in den 3D-Spektren enthaltene Information auf zweidimensionale schmale Streifen zu reduzieren und mit Hilfe dieser, die sequentielle Zuordnung zu erhalten.

Die definierten *Strips* werden in einem *Pool* gesammelt und können in beliebiger Reihenfolge in sogenannten *Slots* angeordnet bzw. sortiert werden. *Strips* aus mehreren 3D-Spektren können gleichzeitig verwendet werden. Das Sortieren erfolgt durch den Vergleich von *Peak Positionen* verschiedener *Strips* von verschiedenen Korrelationsspektren. In Abbildung 3.2 ist zu sehen, wie man durch den Vergleich von *Strips* aus zwei verschiedenen 3D-NMR-Spektren sequentielle Informationen erhält.

Der Umgang mit *Strips* ist im Handbuch von AUREMOL bzw. in der Diplomarbeit von Adel Nasser [83] näher beschrieben und ist eng an das *Strip Tool* von AURELIA angelehnt. Dieses Verfahren ist auch zu Dokumentationszwecken geeignet, da bei richtiger Anordnung der *Strips* der sequentielle Pfad visualisiert werden kann.

Strukturinformationen:

AUREMOL bietet die Möglichkeit, die Volumeninformation zugeordneter Signale aus 2D- bzw. 3D-NOESY-Spektren in Abstandsinformationen, sog. *restraints*, zu konvertieren, die von externen Strukturrechnungsprogrammen verwendet werden können. Unterstützt wird bisher das Format von CNS [20]. Die zugeordneten Signale müssen bereits integriert sein, d.h. die Volumina bekannt sein. Um die Abstände zu kalibrieren, muss interaktiv ein Referenzsignal mit bekanntem Abstand angegeben werden. Die Distanzen werden dann über die Beziehung

$$d = V^{-\frac{1}{x}} \quad (3.1)$$

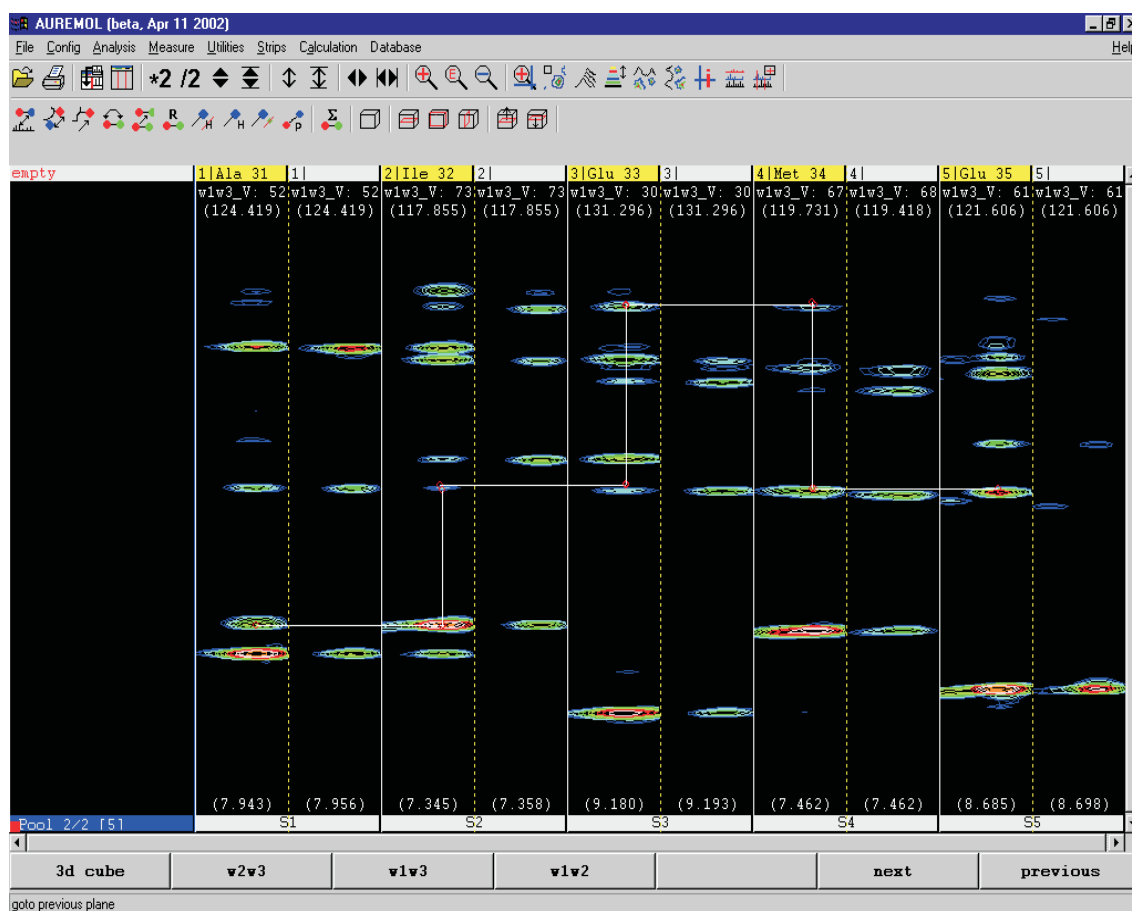


Abbildung 3.2: Anwendung der *Strips*, um die sequentielle Zuordnung zu erhalten. Dargestellt sind 5 *Strips* aus dem 3D-NOESY- (gelb) bzw. 3D-TOCSY- (weiß) Spektrum von HPr. Durch richtiges Sortieren der *Strips* ist es möglich, den sequentiellen Pfad durch die Aminosäuresequenz zu bestimmen, was in der Abbildung durch die weißen Linien angedeutet ist.

berechnet, wobei x üblicherweise auf 6 zu setzen ist. Die Zahl der beitragenden Protonen beim Referenzsignal muss ebenfalls gesetzt sein. Dieser Wert ist bei einem Referenzsignal, das von den Protonen HA und HN herrührt, eins, im Falle einer Methylgruppe oder Methylengruppe aber drei bzw. zwei. Zusätzlich können noch obere und untere Fehlergrenzen für die Abstände angegeben werden, entweder in Absolutwerten oder prozentual zur berechneten Distanz. Im Falle von Methyl- oder Methylengruppen werden die Volumina entsprechend reduziert, d.h. durch drei bzw. zwei dividiert.

Weitere Methoden:

Strukturen, die mittels Röntgenstrukturanalyse gewonnen werden, enthalten in den entsprechenden *Pdb*-Dateien meist keine Protonen. Eine Routine berechnet automatisch die Position der fehlenden Protonen und ergänzt diese in der Strukturdatei, s. *Calculation* →

Add protons to pdb.

Ebenso wie das Konvertieren der Atomnamen in den *Peaknamen*, ist es mit dem Befehl *Calculation* \rightarrow *Convert pdb-file* möglich, die Atomnamen in *Pdb*-Dateien zwischen verschiedenen Formaten zu konvertieren.

Wie im Kapitel 3.3 gezeigt, benötigt man für die Simulation von NOESY-Spektren entweder eine *Metadatei*, s. Abschnitt 3.2, oder eine *Spt*-Datei. Diese Dateien enthalten die sequentielle Zuordnung sowie zusätzliche Informationen zur Rückrechnung, s. Abschnitt 3.3. Mit dem Kommando *Calculation* \rightarrow *Pdb to Spt* ist es möglich, aus einer Strukturdatei eine *Spt*-Datei zu erstellen, die für die chemischen Verschiebungen der Kerne statistische Werte aus der globalen Datenbank verwendet. Der Benutzer kann dabei auswählen, ob *random coil* Werte verwendet werden sollen, oder Werte aus unterschiedlichen Sekundärstrukturelementen, wie z.B. α -Helix oder β -Faltblatt.

Für den Fall, dass nach der Rückrechnung automatisch eine Zuordnung des experimentellen NOESY-Spektrums erfolgen soll, benötigt die Rückrechnungsroutine Informationen der experimentellen NOE-Signale, wie z.B. deren Volumina und Positionen. Mit dem Kommando *Calculation* \rightarrow *User SPT* werden diese Daten der experimentellen Signale als neuer Abschnitt an die *Spt*-Datei angehängt. Zusätzlich ist es mit dem Befehl *Database* \rightarrow *Create metafile from sptfile* möglich, aus einer *Spt*-Datei und einer *Compounddatei*, eine *Metadatei* für die verbesserte Simulation von NOESY-Spektren zu erzeugen.

Der magnetische Suszeptibilitätstensor und dessen Eigenvektoren einer Proteinstruktur können mit dem Kommando *Utilities* \rightarrow *Magnetic susceptibility* laut Anhang B berechnet werden. Der Algorithmus basiert auf der Methode von Grziesek und Bax [102].

Der aktuelle Zustand von AUREMOL, welche Spektren bzw. Strukturen geladen und evtl. übereinandergelegt wurden, kann in einer Sitzungsdatei abgelegt werden. Der Benutzer kann dann zu einem späteren Zeitpunkt durch Öffnen dieser Datei den alten Zustand wiederherstellen. Die Kommandos sind laut Anhang B unter *File* \rightarrow *Open session* bzw. *File* \rightarrow *Save session* abrufbar.

3.1.4 Homologie Modelling mit Permол

Das von Andreas Möglich [77] entwickelte Programm Permол wurde in AUREMOL integriert. Permол ermittelt für ein Protein unbekannter 3D-Struktur mit Hilfe von homologen Proteinen, deren räumlichen Strukturen bekannt sind, Informationen über interatomare Distanzen, dihedrale Winkel und Wasserstoffbrückenbindungen. Diese Daten werden in das Eingabeformat der zwei Strukturberechnungsprogramme CNS [20] und DYANA [47] umgewandelt, so dass damit die Berechnung einer Homologiestruktur des Proteins möglich ist.

Das Programm unterteilt sich in 5 Einzelschritte:

1. Eingabe der Daten
2. Alignment der Sequenzen
3. Berechnung interatomarer Distanzen
4. Berechnung dihedraler Winkel
5. Berechnung der Wasserstoffbrückenbindungen

Im ersten Schritt muss die Primärsequenz des Proteins, dessen Struktur unbekannt ist, eingegeben werden. Die Sequenz muss dabei als Textdatei im Ein-Buchstabencode vorliegen. Zusätzlich müssen die räumlichen Strukturen der homologen Proteine als *Pdb*-Dateien vorliegen, sowie der Name der Eingabedatei für das anschließende *Alignment* angegeben werden.

Nun wird das externe Programm ClustalX [101] aufgerufen, das ein komplettes *Alignment* der Sequenzen der Modellstrukturen und des interessierenden Proteins durchführt. Dabei werden für jede Aminosäure in der Proteinsequenz Scorewerte berechnet, die angeben, wie gut diese in den Homologiestrukturen konserviert sind. Die von ClustalX erzeugte Ausgabedatei wird nun für die Extraktion der strukturelevanten Parameter verwendet. In den folgenden Berechnungen können zum einen Aminosäuren, die gleichen Scorewert besitzen, zum anderen aber auch einzelne Aminosäuren interaktiv selektiert werden. Bei der Berechnung der Distanzen kann angegeben werden, ob die Distanzen mit dem Scorewert gewichtet und bis zu welchem maximalen Wert Distanzen gespeichert werden sollen. Bei der Winkelberechnung werden auch die Mittelwerte und Standardabweichungen mit dem Programm CYCLIST [32] bestimmt. Auch hier kann man die Ergebnisse wieder gewichten. Auf Einzelheiten sei auf das Handbuch von AUREMOL verwiesen. Im letzten Schritt können Wasserstoffbrückenbindungen berechnet werden. Die Kriterien für eine Bindung, also der Abstand zwischen dem Amidproton und dem Carbonylsauerstoff, sowie der Winkel zwischen dem Verbindungsvektor dieser Atome, sind interaktiv einstellbar. Ausserdem kann der Benutzer angeben, bei wievielen der Modellstrukturen die Wasserstoffbrückenbindung auftreten muss, damit sie verwendet werden.

Die erzeugten Ausgabedateien dienen als Eingabe für die eigentliche Strukturrechnung mit den externen Programmen CNS und DYANA.

3.1.5 Automatische Artefaktentfernung in NMR-Spektren

In AUREMOL wurde eine Routine zur automatischen Artefakterkennung in NMR-Spektren implementiert, die für 2D-Spektren bereits in AURELIA vorhanden war. Sie wurde in unserer Arbeitsgruppe auf 3D-NMR-Spektren erweitert [94].

Die Methode separiert echte NMR-Signale von Rauschsignalen und Artefakten und beruht auf der Annahme, dass verschiedene Klassen⁴ verschiedene Verteilungen spezifischer lokaler Merkmale, wie Linienformen, und globaler Merkmale, wie der Symmetrie bezüglich der Diagonalen im Falle von 2D-Spektren, besitzen. Dazu werden, basierend auf dem Bayes'schen Theorem [26], [27], Wahrscheinlichkeiten berechnet, dass ein *Peak* zu einer dieser Klassen gehört. Die theoretischen Grundlagen sind in [2] und [98] nachzulesen.

In einem ersten Schritt müssen für die verschiedenen Klassen Wahrscheinlichkeitsverteilungen der verwendeten Merkmale berechnet werden. In der Implementierung von AUREMOL werden zwei Klassen verwendet, eine Klasse echter NMR-Signale und eine Klasse von Rauschsignalen. Die verwendeten Merkmale sind:

- Signalintensität und 3 verschiedene Volumenverhältnisse [2] als lokale Merkmale
- Symmetrie bzgl. der Diagonalen im 2D Fall [98] als globales Merkmal

Dazu müssen Trainingsgebiete für jede Klasse definiert werden, die möglichst nur *Peaks* dieser Klasse enthalten. Diese Gebiete werden interaktiv vom Benutzer festgelegt, eine automatische Suche dieser Trainingsgebiete wird derzeit von Hubert Riepl [94] entwickelt, ist aber noch nicht in AUREMOL eingebaut.

Mit Hilfe der so erhaltenen Wahrscheinlichkeitsverteilungen kann nun mit dem Entscheidungskriterium von Bayes die Wahrscheinlichkeit p_i berechnet werden, dass ein *Peak* i aus einem benutzerdefinierten Testgebiet mit seinen spezifischen Merkmalen zur Klasse der echten NMR-Signalen gehört. Je größer p_i ist, desto wahrscheinlicher ist *Peak* i ein Signal. In Abbildung 3.3 ist ein Ausschnitt aus einem 2D-NOESY-Spektrum nach der Berechnung dieser Wahrscheinlichkeiten zu sehen. Man sieht anhand der vergrößerten Ausschnitte rechts, dass der Algorithmus die beiden Rauschsignale mit den Wahrscheinlichkeiten $p_i = 0.21$ und $p_i = 0.19$ erkennt und gut von den echten NMR-Signalen separiert ($p_i = 0.99$ bzw. $p_i = 1.00$). Es wurde eine weitere Routine implementiert, die automatisch alle *Peaks* mit Wahrscheinlichkeiten unterhalb eines vom Benutzer vorzugebenden *Thresholds* entfernt. In diesem Beispiel wurden alle *Peaks* mit $p_i < 0.8$ als Rauschsignale festgesetzt und automatisch entfernt, was in der Abbildung 3.3 im rechten unteren Ausschnitt zu sehen ist.

⁴Damit sind z.B. die Klasse echter NMR-Signale oder die Klasse von Rausch- und Artefaktsignalen gemeint.

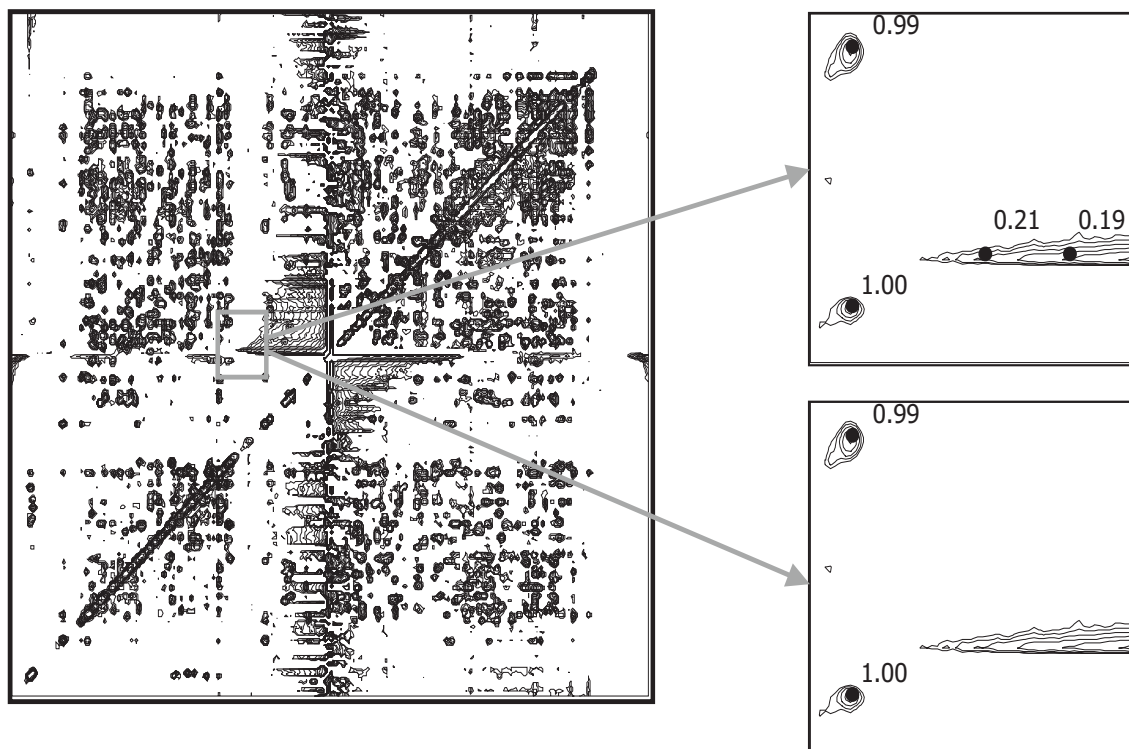


Abbildung 3.3: Trennung von Signal- und Artefaktpeaks. Rechts sind zwei Ausschnitte aus dem linken NOESY-Spektrum vergrößert dargestellt. Der obere Ausschnitt zeigt 4 *Peaks* mit den berechneten Wahrscheinlichkeiten. Man sieht, dass der Algorithmus die beiden Artefaktpeaks erkennt ($p_i = 0.21$ und $p_i = 0.19$), die im unteren Ausschnitt automatisch entfernt wurden.

3.2 Datenbank und Datenstruktur von AUREMOL

Im Folgenden soll näher auf die globale Datenbank und interne Datenstruktur von AUREMOL, wie sie in der Abbildung 2.2 gezeigt war, eingegangen werden. Wie im Kapitel 2.1 erläutert, ist die Datenstruktur eng auf das Konzept von AUREMOL abgestimmt.

3.2.1 Globale Datenbank

Die globale Datenbank besteht aus zahlreichen ASCII Textdateien, die editierbar sind und damit vom Benutzer beliebig erweitert werden können.

1. `as_def.txt` enthält die Definition der 20 natürlich vorkommenden Aminosäuren in der IUPAC Konvention [73]. Die chemische Struktur jeder Aminosäure und mögliche

Aliasnamen für Pseudoatome⁵ sind hier festgelegt. Dies wird z.B. bei der Erstellung der *Compounddateien* verwendet.

2. *classes.txt* definiert verschiedene Klassen von Atomen, wie z.B. Amidprotonen, für die Spektren simuliert werden. Es sind 4 Klassen definiert: Backbone Protonen, Protonen in Methylgruppen, Protonen in Ringen und Amidprotonen. Neue Klassen für Heterokerne sind hierbei leicht zu ergänzen. Durch diese Klassen sind zahlreiche Merkmale, wie Korrelationszeiten und Bewegungsmodelle s. Abschnitt 2.2.4, für die Simulation von NOESY-Spektren festgelegt.
3. *cs_table.txt* wird beim 2-Stufen-Algorithmus in Abschnitt 2.3.2 verwendet. Für jeden Kern aller Aminosäuren ist die mittlere chemische Verschiebung mit der Standardabweichung angegeben. Die Daten sind dabei der BioMagResBank [12] entnommen und entstammen statistischen Analysen gelöster Proteinstrukturen.
4. *csa.txt* enthält die Anisotropiewerte für die chemische Verschiebung von Wasserstoff, Stickstoff und Kohlenstoff, die für die Simulation von NOESY-Spektren verwendet werden.
5. *experiments.txt* definiert zahlreiche NMR-Experimente, was bei der Verifizierung manuell vergebener *Peaknamen* benutzt wird. Der Name, die Dimension des Experiments und welche Kerne in den einzelnen Frequenzachsen detektiert werden, ist dabei festgelegt. Der Benutzer kann neue Experimente unter Berücksichtigung der verwendeten Nomenklatur zusätzlich in dieser Datei ergänzen.
6. *IUPAC.txt* enthält verschiedene Konventionen für die Bezeichnung von Atomnamen. Beispiele sind die IUPAC Konvention, aber auch das CNS oder DYANA Format. Dies wird bei der Konvertierung der *Peaknamen* bzw. *Pdb*-Dateien benutzt. Die Datei kann vom Benutzer durch eigene Formate erweitert werden.
7. *shifts.txt* beinhaltet chemische Verschiebungen der Atome von jeder Aminosäure, unterteilt in *random coil* Werte, sowie Werte für β -Faltblatt und α -Helix. Das Kommando *Calculation* \rightarrow *Pdb to Spt* verwendet diese Werte.
8. *susc.txt* enthält die magnetische Anisotropie der Peptidbindung und der Seitenketten aromatischer Aminosäuren [102]. Diese werden bei der Berechnung des magnetischen Suszeptibilitätstensors benutzt.
9. *topo.jcc* definiert dihedrale Winkel im Protein. Ausserdem sind J-Kopplungskonstanten für jede Aminosäure enthalten, die bei der Simulation von NOESY-Spektren verwendet werden.

⁵Atome, deren chemische Verschiebungen entartet sind.

3.2.2 Interne Datenstruktur

Wie in Abbildung 2.2 zu sehen ist, besteht die interne Datenstruktur aus zahlreichen Komponenten, die im Folgenden beschrieben werden sollen. Diese Struktur enthält die komplette Information, die für die Auswertung der NMR-Spektren von Bedeutung ist und ist an die Notation des STAR Formats angelehnt [12]. Das STAR Format wurde von der BioMagResBank als Datenformat für die NMR-Spektroskopie vorgeschlagen. Es findet eine Unterteilung in einzelne Sektionen statt. Jede Sektion beginnt mit einer Definition der verwendeten physikalischen Größen⁶, woran sich zeilenweise die Werte für diese Parameter anschließen.

3.2.2.1 *Compounddatei*

Die *Compounddatei* wird aus den Parametern, die in der globalen Datenbank enthalten sind, und einer *Sequenzdatei* erzeugt. Die *Sequenzdatei* enthält die Primärsequenz des Biomoleküls im Ein- bzw. Drei-Buchstaben-Code.

Die *Compounddatei* definiert einen Bestandteil der NMR-Probe, was entweder das interessierende Protein sein kann oder ein anderes Molekül der Lösung. In Abbildung 3.4 ist ein Auszug aus einer solchen Datei zu sehen. Sie gliedert sich in drei Abschnitte bzw. Sektionen.

Im ersten Abschnitt werden alle Atome in sequentieller Reihenfolge spezifiziert. Es beginnt mit der Nummer der Aminosäure in der Sequenz und der Atomnummer in der Aminosäure. Dann folgt der Name der Aminosäure im Drei-Buchstaben-Code, der Atomname im IUPAC-Format und der Atomtyp. Der Atomtyp unterscheidet zwischen ¹H, ¹³C und ¹⁵N Kernen. Zusätzlich finden sich Informationen, ob ein Aliasname existiert⁷, zu welchen anderen Atomen dieser Kern magnetisch äquivalent ist und über die Anisotropiekonstanten $\sigma_{\parallel} - \sigma_{\perp}$ der chemischen Verschiebung. Anschließend wird die chemische Struktur angegeben, d.h. für jedes Atom ist festgelegt mit welchen anderen Atomen es kovalent gebunden ist. Der letzte Abschnitt definiert dihedrale Winkel und zwar sowohl vom Rückgrat als auch von den Seitenketten des Proteins. Zusätzlich sind die Karpluskonstanten A, B und C angegeben. Anschließend werden für alle Atompaare, die über maximal drei Bindungen miteinander gekoppelt sind, J-Kopplungskonstanten bzw. einer der oben definierten dihedralen Winkel angegeben, um mit der Karplusgleichung bei der Simulation von NOESY-Spektrum die J-Kopplungsstärke berechnen zu können. In der *Compounddatei* sind keinerlei probenspezifische Informationen enthalten, wie z.B. chemische Verschiebungen.

⁶zu erkennen am vorangehenden Unterstrich.

⁷Beispielsweise werden die Kerne HB1, HB2 und HB3 einer Methylgruppe oft zu dem Aliasatom HB nach der IUPAC-Konvention zusammengefasst.

```

COMPOUND: HPRC

section_sequenzdefinition
  _Residue_seq_code
  _Atom_num_code
  _Residue_label
  _Atom_name
  _Atom_type
  _Atom_alias
  _Atom_equivalent
  _Atom_CSA
  1 1 MET HN H - - 8.95
  1 2 MET N N - - 157.00
  ...
  88 16 GLU HE2 H - - 8.95
end_section

section_bond
  _Bond_start
  _Bond_atom1
  _Bond_atom2
  _Bond_atom3
  _Bond_atom4
  1,1 1,2 - - -
  1,2 1,1 1,3 - -
  ...
  88,15 88,13 88,16 - -
  88,16 88,15 - - -
end_section

...

section_j-coupling
DEFINE
  _Angle_name
  _Angle_type
  _Karplus_constantA
  _Karplus_constantB
  _Karplus_constantC
  _Karplus_thetainc
  phi C'-N-CA-CB 6.51 -1.76 1.6 +60
  chi_1 N-CA-CB-X 9.5 -1.6 1.8 +60
  ...
  chi_52 PHE CD2-CE2-CZ-CE1 9.5 -1.6 1.8 +60
END_DEFINE
DATA
  _Coupled_atom1
  _Coupled_atom2
  _Coupling_value
  _Angle_name
  1,1 1,2 -92.00 -
  1,1 1,4 - phi
  ...
  88,12 88,9 8.00 chi_2
  88,12 88,10 130.00 -
  88,12 88,11 13.00 -
END_DATA
end_section

```

Abbildung 3.4: Beispiel für eine *Compounddatei*

3.2.2.2 Masterliste

In AUREMOL wurde ein neues Listenformat entwickelt, das die für die Auswertung relevanten Daten aus den Spektren zusammenfasst. Dies umfasst sowohl experimentelle als auch simulierte Spektren. Die *Masterliste* eines Spektrums enthält wichtige Informationen aller gepickten *Peaks*. In Abbildung 3.5 ist ein Beispiel zu sehen.

Es beginnt mit einem Header, der das NMR-Experiment, die Dimension des Spektrums und andere wichtige Prozessierungsparameter, wie z.B. die Spektrenbreite in Hz oder ppm enthält. Dann werden alle *gepickten Peaks* aufgelistet, beginnend mit dem *Peaknamen* und evtl. Kommentaren. Die chemischen Verschiebungen in jeder Dimension, die Intensität, das Volumen des *Peaks* und der Qualitätswert, der angibt, mit welcher Wahrscheinlichkeit es sich um ein echtes NMR-Signal oder um ein Rauschsignal handelt, folgen.

Zu erwähnen ist, dass AUREMOL im Bezug auf das Listenformat abwärtskompatibel zu seinem Vorgänger AURELIA ist. AUREMOL kann automatisch aus den alten AURELIA Listen eine *Masterliste* erzeugen. Ebenso ist es möglich aus der aktuellen *Masterliste* die AURELIA Listen zu generieren. Damit ist im Prinzip ein problemloses Wechseln zwischen den beiden Applikationen möglich.

```

Masterlist (Volumes rescaled with NC_Proc)
=====
HEADER
  EXPERIMENT:          NOESY->H/H
  DIMENSION:           2
  MIXINGTIME_SEC:      0.10
  AQUISITIONDELAY:     *
  RESONANCEFREQUENCY_1: 600.142105
  RESONANCEFREQUENCY_2: 600.142105
  SPECTRAL_WIDTH_HZ_1: 7644.891443
  SPECTRAL_WIDTH_HZ_2: 7644.891443
  SPECTRAL_WIDTH_PPM_1: 12.738469
  SPECTRAL_WIDTH_PPM_2: 12.738469
  OFFSET_PPM_1:        11.897447
  OFFSET_PPM_2:        11.897447
  FILTER_1:            gaussian
  FILTER_2:            gaussian
  LINEBROAD_HZ_1:      25.000
  LINEBROAD_HZ_2:      25.000
  AQ_MODE_1:           simultaneous
  AQ_MODE_2:           simultaneous
  SCALING:             -6
END_HEADER

Total amount of peaks: 1728

PEAKLABEL: HG2 57/H 57
COMMENT:
COMPOUNDS: 1 1
PEAKDESCRIPTION:
coord: 1      coord: 2      intensity      volume      quality
  850 1.324    320 7.917      32015        235161      0.97

...

PEAKLABEL: HA 4/HB2 4
COMMENT:
COMPOUNDS: 1 1
PEAKDESCRIPTION:
coord: 1      coord: 2      intensity      volume      quality
  613 4.272    803 1.908      12499        282592      0.95

...

PEAKLABEL: HA 58/HB3 58
COMMENT:
PEAKLABEL: HA 58/HB2 58
COMMENT: first variant
COMPOUNDS: 1 1
PEAKDESCRIPTION:
coord: 1      coord: 2      intensity      volume      quality
  613 4.272    726 2.866      16677        197706      0.98

```

Abbildung 3.5: Beispiel für eine *Masterliste*

3.2.2.3 MasterMasterdatei

Für den Automatisierungsprozess ist es nötig, viele verschiedene NMR-Spektren zu analysieren. Dazu kann aus mehreren *Masterlisten* eine *MasterMasterdatei* erzeugt werden, die die Informationen aus den selektierten Spektren bündelt. In Abbildung 3.6 ist ein Auszug aus einer solchen Datei zu sehen.

```
# MasterMasterList

section_spectra
/x/user/gab03899/auremol_assignment/ttt_1024x1024/Arr.ml
end_section

# ATOM atom identifier
#       averaged chem. shift in ppm of atom of used spectra
#       j-coupling in Hz to atoms of spectra 1
#       ...
#       j-coupling in Hz to atoms of spectra n

# Example:
# ATOM HN 15
#       7.352  7.367  7.330
#       -
#       HA3 15 4.7 HA2 15 4.5
#       -

section_atoms compound 1
ATOM HD1 81
        -0.262
        *

...

ATOM HN 39
        8.731
        *

end_section
```

Abbildung 3.6: Beispiel für eine *MasterMasterdatei*

Sie besteht aus mehreren Abschnitten. Im ersten Abschnitt sind alle NMR-Spektren bzw. *Masterlisten* aufgeführt, die in dieser Datei zusammengeführt wurden. In der nächsten Sektion sind Informationen der Atome aus den verschiedenen Bestandteilen der NMR-Probe, also den einzelnen *Compounddateien*, aufgeführt, die nicht mehr spektrenabhängig sind. Das heißt, es sind z.B. keine *Peakvolumina* oder Wahrscheinlichkeiten mehr zu finden, sondern nur chemische Verschiebungen in ppm und experimentell bestimmte J-Kopplungen. Für jeden Bestandteil, also für jede *Compounddatei*, wird eine entsprechende Sektion angelegt. In diesen Sektionen befinden sich für jeden Bestandteil bzw. *Compound* der Lösung die gemittelten chemischen Verschiebungen in ppm der einzelnen zugeordneten Signale. Zusätzlich sind hier auch experimentell ermittelte J-Kopplungen mit anderen Atomen aufgeführt.

3.2.2.4 Metadatei

Die *Metadatei* fasst nun alle bisherigen Informationen zusammen. Sie wird aus den einzelnen *Compounddateien* und der *MasterMasterdatei* erzeugt. Diese besteht wieder aus mehreren Sektionen und in Abbildung 3.7 ist ein Ausschnitt zu sehen.

Im ersten Abschnitt werden die einzelnen Bestandteile der NMR-Probe angegeben. Es sind die *Compounddateien* aufgeführt sowie Informationen über die Konzentration der Bestandteile in der Probe und evtl. Isotopenmarkierungen. Anschließend sind einige wei-

```

section_probedefinition
COMPOUNDS:
_Mol_label
_Compound_file
_Concentration_value
_Concentration_value_units
_Isotopic_labeling
HPRC /x/user/gab03899/auremol_assignment/compound - - -
END_COMPOUNDS
PARAMETER:
_Variable_type
_Variable_value
_Variable_value_unit
pH 7.0 -
Temperature 300 K
Pressure 100000 N/m²
END_PARAMETER
end_section

section_compounddefinition 1
NAME: HPRC
CLASSDEF:
DEFINE CLASSES

NAME Default
NUCLEUS 1H
OCCUPANCY 1

NAME Methyl
NUCLEUS 1H
OCCUPANCY 1

NAME Ring
NUCLEUS 1H
OCCUPANCY 1

NAME HN
NUCLEUS 1H
OCCUPANCY 1

END_DEFINE

DEFINE MATRICES

DENSITY
RIGID FAST_JUMP SLOW_JUMP RIGID
FAST_JUMP FAST_JUMP FAST_JUMP FAST_JUMP
SLOW_JUMP FAST_JUMP SLOW_JUMP SLOW_JUMP
RIGID FAST_JUMP SLOW_JUMP RIGID

ORDER
1 1 1 1
1 1 1 1
1 1 1 1
1 1 1 1

CORRELATION
3.0e-9 3.0e-9 3.0e-9 3.0e-9
3.0e-9 3.0e-9 3.0e-9 3.0e-9
3.0e-9 3.0e-9 3.0e-9 3.0e-9
3.0e-9 3.0e-9 3.0e-9 3.0e-9

INTERNAL
1.0e-10 1.0e-10 1.0e-10 1.0e-10
1.0e-10 1.0e-10 1.0e-10 1.0e-10
1.0e-10 1.0e-10 1.0e-10 1.0e-10
1.0e-10 1.0e-10 1.0e-10 1.0e-10

CALIBRATION
0 0 0 0
0 0 0 0
0 0 0 0
0 0 0 0

END_DEFINE
END_CLASSDEF
SHIFTS:
_Residue_seq_code
_Atom_num_code
_Atom_alias
_Chem_shift_value
_Chem_shift_value_error
_Chem_shift_ambiguity_code
_Atom_class
_Linewidth
81 13 HD1 -0.262 0.20 2 1 -
...
39 1 - 8.731 0.20 2 3 -
END_SHIFTS
J_COUPL:
_Coupled_atom_1
_Coupled_atom_2
_Coupling_value
END_J_COUPL
end_section

```

Abbildung 3.7: Beispiel für eine *Metadatei*

tere physikalische Parameter zu setzen, wie z.B. der pH-Wert oder die Temperatur bei der Messung, die bisher bei der Auswertung keine Rolle spielen und daher nur Defaultwerte gesetzt sind. Die weiteren Sektionen definieren die einzelnen Bestandteile näher. Zunächst sind Klassen für verschiedene Atomtypen definiert, die z.B. die Bewegungsmodelle von Abschnitt 2.2.4 für die Simulation der NOESY-Spektren näher beschreiben. Dann werden alle Zuordnungen aus der *MasterMasterdatei* aufgelistet, allerdings nicht mit den entsprechenden Atomnamen sondern mit zwei Indizes, die auf das korrespondierende Atom in der *Compounddatei* verweisen. Darauf folgt der über alle experimentellen Spektren gemittelte Wert für die chemische Verschiebung und die Genauigkeit mit der dieser Wert gegeben ist. Weiter ist eine der oben definierten Atomklassen und eine evtl. bestimmte experimentelle Linienbreite angegeben. Daran schließt sich ein weiterer Abschnitt über experimentell bestimmte J-Kopplungen an, die aus der *MasterMasterdatei* extrahiert wurden. Mit der Information, die in der *Metadatei* und den zugehörigen *Compounddateien* enthalten ist, ist es nun möglich, NOESY-Spektren zu simulieren.

Für die automatische Auswertung ist es nötig, zu Beginn die verschiedenen Bestandteile der NMR-Probe zu definieren, also die einzelnen *Compounddateien* zu generieren. Nach der Messung mehrerer experimenteller NMR-Spektren und einer automatischen Vorverarbeitung, wird daraus eine *MasterMasterdatei* erzeugt. Die *Metadatei*, die die einzelnen *Compounddateien* mit der *MasterMasterdatei* verbindet, wird angelegt.

Nun startet ein iterativer Prozess. NMR-Spektren werden simuliert und manuell oder automatisch mit den experimentellen Spektren verglichen. Dies liefert neue Informationen, d.h. Zuordnungen in den experimentellen Spektren, was eine Modifikation der *Masterlisten* bzw. der *MasterMasterdatei* bedeutet. Mit der modifizierten *MasterMasterdatei* wird auch die *Metadatei* neu generiert und ein neuer Iterationsschritt kann beginnen.

3.3 Verbesserte Simulation von NOESY-Spektren

Eine exakte Simulation von 2D- und 3D-NOESY-Spektren stellt eine wichtige Voraussetzung für die automatische Zuordnung von NOESY-Spektren mit Hilfe des 2-Stufen Algorithmus dar. Bisher wurden in RELAX [40], [39], [38] vom Benutzer fest vordefinierte Linienbreiten gesetzt. Die Linienbreiten von Protonen liegen aber meist in einem Bereich zwischen 6 und 9 Hz [22], so dass deren Berechnung in AUREMOL zu einer deutlich genaueren Simulation der *Peakformen* führt. Ebenso war die Aufspaltung der *Peaks* durch die J-Kopplung nicht berücksichtigt. Mit der Entwicklung neuerer Spektrometer mit 800 und 900 MHz⁸ und dem Einsatz von Kryoprobenköpfen, die durch Heliumkühlung das Signal-Rausch-Verhältnis deutlich verbessern, werden J-Kopplungseffekte auch in den NOESY-Spektren sichtbar. Die Aufspaltung, die zwischen 2 und 15 Hz liegt, ist meist nicht richtig aufgelöst, kann aber die *Peakform* bereits beeinflussen. Die geminale Kopplung der Protonen in Methylengruppen ist mit 13 Hz bei Signalen mit geringer Linienbreite und entsprechender Auflösung des Spektrums bereits sichtbar. Die Verwendung der J-Kopplung verbessert die Qualität der simulierten *Peakformen*, was für den Vergleich von experimentellen und simulierten NOESY-Spektren ein wichtiges Kriterium darstellt. Die CSA wird in AUREMOL sowohl bei der Berechnung der longitudinalen als auch der transversalen Relaxationszeiten berücksichtigt.

3.3.1 Implementierung in AUREMOL

Für die Simulation von NOESY-Spektren wurde das Programm RELAX um die oben erwähnten Punkte erweitert. RELAX benötigt als Eingabedaten eine *Pdb*-Datei und eine *Spt*-Datei. Die *Pdb*-Datei enthält die Strukturinformation des Proteins und die *Spt*-Datei die sequentielle Zuordnung der Kerne und deren Klassenzugehörigkeit, wodurch die einzelnen Bewegungsmodelle für die Relaxationsberechnungen festgelegt sind. Die Berechnung der J-Kopplung und der CSA erfordert aber die Kenntnis zusätzlicher Parameter, wie z.B. dihedraler Winkel. Dazu wurde die interne Datenstruktur von AUREMOL, s. Kapitel 3.2, an RELAX angepasst, d.h. eine interne Schnittstelle entwickelt, die aus der *Compounddatei* und der *Metadatei* eine *Spt*-Datei erzeugt, welche als Eingabe für die Relaxationsberechnungen dient. Damit wurde eine Abwärtskompatibilität erreicht, die die Simulation von NOESY-Spektren entweder mit der *Metadatei* oder mit einer *Spt*-Datei als Eingabedaten möglich macht. Im letzteren Fall wird allerdings keine J-Kopplung und kein CSA-Effekt berücksichtigt, da die nötigen Parameter nicht in der *Spt*-Datei enthalten sind.

In Abbildung 3.8 ist der Dialog zu sehen, wo die freien Parameter bei der Rückrechnung

⁸entspricht der Resonanzfrequenz eines Protons.

durch den Benutzer zu setzen sind:

Zuerst muss angegeben werden, mit welcher Eingabedatei die Berechnung gestartet wer-

The screenshot shows a dialog box titled "Parameters / 2D3D-NOESY backcalculation". It contains the following parameters and their values:

- Select filetype: Use Metafile (dropdown)
- name of meta-file: meta_HPr (dropdown)
- no. of compound (metafile): 1 (text input)
- name of spt-file: (empty text input)
- name of pdb-file/ensemble (IUPAC): HPr.pdb (dropdown)
- cutoff distance (nm): 5 (text input) nm
- mixing time (s): 0.1 (text input) Seconds
- larmor (Mhz): 600.13 (text input) Mhz
- name of output text file: output.txt (text input)
- name of log file: relax.log (text input)
- path of sim. 2D/3D spectrum: d:\HPr\sim (dropdown)
- sample points in 1: 1024 (text input)
- sample points in 2: 4096 (text input)
- sample points in 3 (3D): 256 (text input)
- linewidth 15N/13C: 25 (text input) Hz
- type of experiment: NOESY->H/H (dropdown)
- lineshapes: gaussian (dropdown)
- phase increment/F1 (3D): F1: simultaneous acquisition (dropdown)
- phase increment/F2 (3D): F2: simultaneous acquisition (dropdown)
- phase increment/F3 (3D): F3: simultaneous acquisition (dropdown)
- Line broadening in F1 (2D): 3 (text input) Hz
- Line broadening in F2 (2D): 3 (text input) Hz
- consider hetero atoms ? (3D): ☐
- Assign: ☐
- J-correction for T2: ☐
- path of exp. 2D/3D spectrum (procs): (empty dropdown)

At the bottom, there are "OK" and "Cancel" buttons.

Abbildung 3.8: Einzustellende Parameter bei der Simulation von NOESY-Spektren in AURE-MOL

den soll, d.h. mit einer *Metadatei* oder mit einer *Spt-Datei*. Falls *Metadatei* selektiert wurde, muss die Nummer des *Compounds* bzw. Proteins der NMR-Probe, das simuliert

werden soll, angegeben werden⁹. Zusätzlich wird die Strukturinformation des Proteins in Form einer *Pdb*-Datei benötigt. Der *cutoff* in Nanometer gibt an, bis zu welchem Abstand Wechselwirkungen zwischen Protonen simuliert werden sollen. Die Mischzeit und Larmorfrequenz sind für die Berechnung der Relaxationszeiten nötig. In der Ausgabedatei sind alle simulierten Signale unter Angabe der Positionen im Spektrum, des berechneten Volumens und der Linienbreiten aufgelistet. Die Log-Datei beinhaltet automatisch zugeordnete experimentelle Signale, falls die Option *Assign* weiter unten im Dialog aktiviert wurde. Diese beiden Textdateien werden bei der Berechnung der R-Werte benötigt, mehr dazu im Kapitel 3.5. Optional kann ein Verzeichnis angegeben werden, in welches das simulierte Spektrum geschrieben werden soll. Die Spektrendaten, Prozessierungsdateien im Brukerformat und die *Masterliste* werden erzeugt und in diesem Verzeichnis abgelegt. Die nächsten drei Parameter spezifizieren die digitale Auflösung des simulierten Spektrums in den Dimensionen w_1 , w_2 und im Falle eines 3D-Spektrums zusätzlich w_3 . Die Linienbreite der Heterokerne ^{13}C und ^{15}N wird als fester Parameter verwendet. Es hat sich gezeigt, dass die Berechnung der Linienbreiten der Heterokerne aufgrund der geringen Auflösung in Richtung der ^{13}C - bzw. ^{15}N -Dimension zu keiner qualitativen Verbesserung der *Peakformen* führt. Als weiterer Parameter ist das zu simulierende Spektrum anzugeben, also entweder das homonukleare 2D-NOESY-Experiment oder heteronukleare 3D-NOESY-Experimente mit ^{13}C bzw. ^{15}N editierten Proteinen. Die Kurvenform der Signale ist einzustellen, zur Auswahl stehen Gaußskurven oder Lorenzkurven. Bei fouriertransformierten experimentellen Spektren hat man es im allgemeinen mit Lorenzkurven zu tun, durch die Verwendung einer sogenannten Fensterfunktion bei der Prozessierung erhalten die Signale jedoch in der Regel Gaußform. Der Benutzer kann zusätzliche Angaben über die Faltung von Signalen in 3D-Spektren machen, was mit den Parametern *phase increment* festzusetzen ist. Die Berechnung der Linienbreiten hat ergeben, dass diese meist etwas schmaler sind als die experimentellen. Ursachen dafür sind nicht ganz homogene Magnetfelder bei der Messung und die Glättung der experimentellen Signale bei der Prozessierung durch die Multiplikation der FIDs mit einer Fensterfunktion. Ausserdem wird bei der Berechnung der Linienbreiten die Wechselwirkung des Proteins mit der Wasserhülle nicht berücksichtigt. Um diese Effekte etwas auszugleichen, kann man für jede Dimension einen additiven Beitrag *Linebroadening* setzen, mit der die Linienbreiten verbreitert werden sollen, im Falle der Abbildung 3.8 werden alle *Peaks* um 3.0 Hz verbreitert. Optional kann angegeben werden, ob bei der Berechnung der Relaxationszeiten auch der Einfluss der Heterokerne mit berücksichtigt werden soll. Dies spielt bei 2D-NOESY-Spektren nur dann eine Rolle, wenn man eine gelabelte Probe hat. Die Möglichkeit der automatischen Zuordnung mit RELAX mit Hilfe der sequentiellen Zuordnung, einstellbar über die Option *Assign*, soll hier nicht näher ausgeführt werden, stattdessen wird auf [39] verwiesen. Wie im Kapitel 2.2 in Gleichung 2.18 gezeigt wurde, führt die skalare Kopp-

⁹Die einzelnen *Compounds* sind in der *Metadatei* durchnummeriert

lung zu einer Verschmälerung der Linienbreiten der koppelnden Protonen. Dieser Effekt kann optional in der Berechnung berücksichtigt werden. Der letzte zu setzende Parameter gibt an, ob bereits vorhandene Prozessierungsdateien bei der Erzeugung des simulierten Spektrums verwendet werden sollen. Dies ist dann sinnvoll, wenn man ein experimentelles und ein simuliertes Spektrum vergleichen will, da dann prozessierungsspezifische Werte wie der Offset oder die Breite der Spektren identisch sein sollten.

Falls ein NOESY-Spektrum auf der Festplatte gespeichert werden soll, d.h. wenn ein Verzeichnis angegeben wurde, so wird in diesem Verzeichnis zusätzlich die Datei *peakshapes2D.dat* bzw. *peakshapes3D.dat* angelegt¹⁰, die die *Peakformen* aller simulierten Signale enthält. In Abbildung 3.9 sind Beispiele für diese Dateien, die für die automatische Zuordnung experimenteller NOESY-Spektren verwendet werden, gezeigt. Im linken Teil-

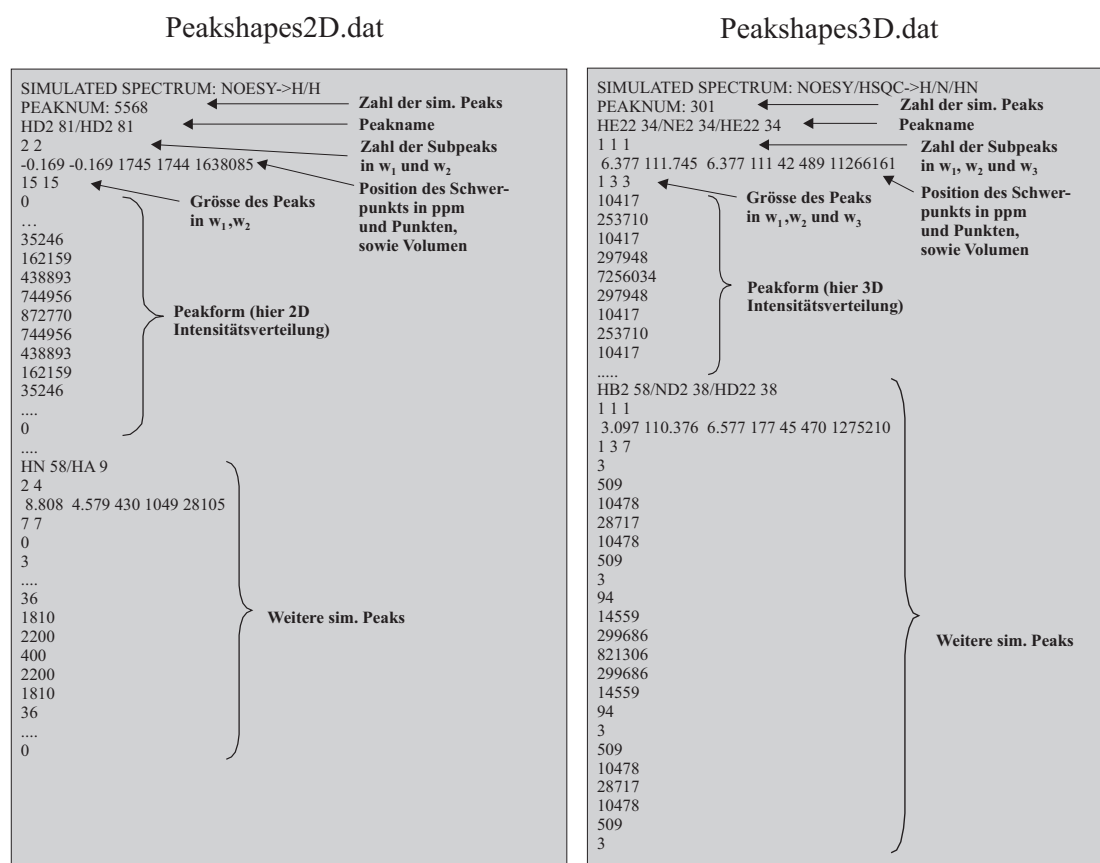


Abbildung 3.9: Beispieldateien mit simulierten *Peakformen*. Links: Enthält die berechneten *Peakformen* eines 2D-NOESY-Spektrums. Rechts: Enthält die berechneten *Peakformen* eines 3D-NOESY/HSQC-Spektrums.

¹⁰je nachdem ob ein 2D- oder 3D-NOESY-Spektrum simuliert wurde

bild sind die *Peakformen* eines 2D-NOESY-Spektrums gezeigt, im rechten die eines 3D-NOESY/HSQC-Spektrums. Die berechneten NOE-Signale werden sequentiell abgelegt, wobei für jedes Signal der *Peakname*, die Multipllettstruktur, falls bekannt die chemischen Verschiebungen, das berechnete Volumen, die Ausdehnung der *Peakform* in den zwei bzw. drei Dimensionen und die Intensitätsverteilung angegeben werden.

3.3.2 Beispiele für die verbesserte Simulation

Im Folgenden sollen simulierte und experimentelle NOE-Linienformen verglichen werden. Das verwendete experimentelle 2D-NOESY-Spektrum der HPr Mutante H15A¹¹ wurde bei einer Temperatur von 303 K und mit einer Mischzeit $\tau_m = 80\text{ms}$ an einem 600 MHz Bruker Spektrometer aufgenommen. Um die Multipllettstruktur sichtbar zu machen, wurde das Spektrum mit einer Exponentialfunktion als Fensterfunktion prozessiert, wobei der Parameter LB in der direkten Dimension auf 3.0 Hz und in der indirekten auf 6.0 Hz gesetzt wurde. Die digitale Auflösung betrug 1024×4096 Punkte.

Um nun Linienformen vergleichen zu können, wurde das simulierte NOESY-Spektrum unter denselben Rahmenbedingungen berechnet. Die rotatorische Korrelationszeit τ_c wurde wie in [50] beschrieben, mit Hilfe der Stokes-Einstein-Beziehung abgeschätzt

$$\tau_c = \frac{4\pi\eta a^3}{3k_B T} \quad (3.2)$$

wobei T die Messtemperatur, η die Viskosität des Lösungsmittels¹² und a der Radius des Moleküls ist. Unter der Annahme, dass das Protein kugelsymmetrisch ist und ein Molekulargewicht M besitzt, vereinfacht sich Gleichung 3.2 zu

$$\tau_c = \frac{MV'\eta}{k_B T} \quad (3.3)$$

Das Partialvolumen V' beträgt bei Proteinen $7.3 \times 10^{-4} \frac{\text{m}^3}{\text{kg}}$. Bei einer Temperatur von 303 K und einem Molekulargewicht $M = 9.5 \text{ kDa}$ ergibt sich eine Korrelationszeit τ_c von 2.8 ns. Mit dieser Korrelationszeit, einer Mischzeit von 80 ms und einer additiven Linienverbreiterung von 3.0 Hz bzw. 6.0 Hz wurde ein 2D-NOESY-Spektrum berechnet. Als Struktur wurde die Lösungsstruktur der HPr Mutante und die Zuordnungsliste aus der Doktorarbeit von Claudia Munte [82] verwendet, die bereits an das experimentelle NOESY-Spektrum angepasst war.

In Abbildung 3.10 sind identische Signale aus einem experimentellen und einem simulierten 2D-NOESY-Spektren des Proteins zu sehen.

Im linken Teilbild werden ein experimenteller *Peak* und das korrespondierende simulierte

¹¹Das His 15 im Wildtyp von HPr ist durch ein Alanin ersetzt.

¹²für Wasser: $\eta = 1.025 \times 10^{-3} \text{ Nm}^{-2}\text{s}$

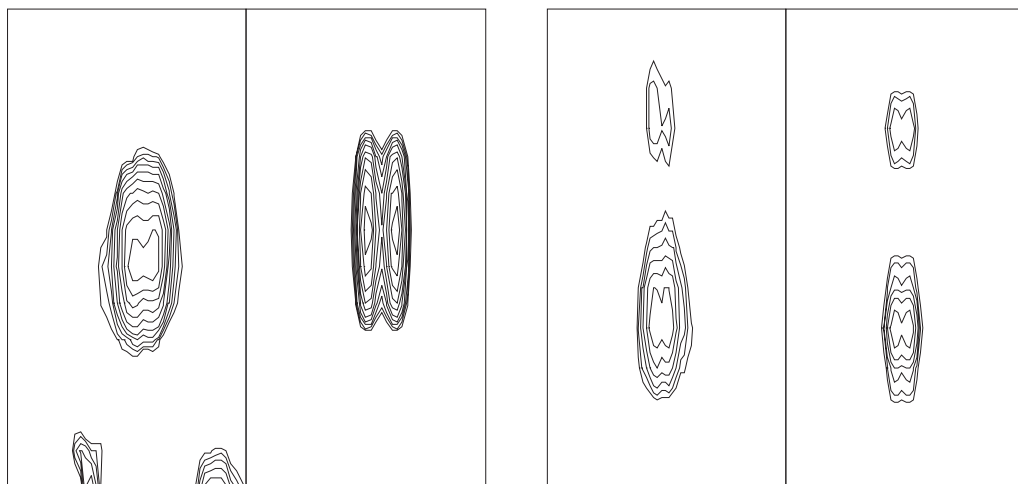


Abbildung 3.10: Vergleich der Linienformen von experimentellen und simulierten Signalen. Links: Im linken Ausschnitt ist das experimentelle Signal zu sehen, im rechten Ausschnitt der korrespondierende simulierte *Peak*. Die Linienbreiten und die Aufspaltung aufgrund skalarer Kopplung sind korrekt berechnet worden. Rechts: Im linken Ausschnitt sind zwei experimentelle Signale zu sehen, im rechten Ausschnitt die dazugehörigen berechneten. Die berechneten Linienbreiten und Intensitätsverhältnisse der beiden *Peaks* zeigen ein korrektes Verhalten.

Signal gezeigt. Der Vergleich der experimentellen und berechneten Linienform zeigt, dass die Linienbreiten und Multipllettstruktur richtig berechnet worden sind. Ähnliches gilt für das rechte Teilbild. Hier sind zwei experimentelle Signale und die beiden dazugehörigen berechneten Signale zu sehen. Die Linienbreiten und die Intensitätsverhältnisse sind denen der experimentellen *Peaks* sehr ähnlich.

3.4 NOE-Zuordnungsmethoden

Die automatische Zuordnung von experimentellen NOESY-Spektren spielt im Konzept von AUREMOL, s. Abbildung 2.1, eine zentrale Rolle. Für diesen Schritt, der bei der manuellen Auswertung von NMR-Spektren, der zeitaufwendigste und fehleranfälligste ist, wurden zwei neue Verfahren entwickelt, die im Theorieteil 2.3 dargestellt worden sind.

3.4.1 KNOWNOE

Die Implementierung von KNOWNOE in AUREMOL wird im Folgenden gezeigt. Im übernächsten Abschnitt wird kurz anhand eines Beispiels dargestellt, wie KNOWNOE anhand einer kompletten sequentiellen Zuordnungsliste und eines iterativen Prozesses die korrekt gefaltete Struktur eines Proteins ermittelt. In diesem Prozess wird wiederholt eine automatische NOE-Zuordnung mit anschließender Strukturrechnung durchgeführt.

3.4.1.1 Implementierung in AUREMOL

Über das Kommando *Calculation* \rightarrow *KnowNOE* \rightarrow *Calculate Assignment*, s. Anhang B, wird KNOWNOE gestartet, wonach sich ein Dialog öffnet, der in Abbildung 3.11 dargestellt ist.

Für die automatische Zuordnung muss mindestens ein 2D- oder 3D-NOESY-Spektrum geladen sein. Folgende Parameter sind vom Benutzer einzustellen:

- Experimentelle Spektren, welche durch KNOWNOE automatisch zugeordnet werden sollen: Dazu wird eine Liste aller in AUREMOL geladener NOESY-Spektren gezeigt, aus der der Benutzer auswählen muss.
- Sequentielle Zuordnungsliste: Kann entweder als *Spt*-Datei oder als *Metadatei* angegeben werden. Die Formate sind in Kapitel 3.3 bzw. 3.2 beschrieben.
- Primärsequenz: Die Aminosäuresequenz des Proteins
- Strukturinformation: Eine *Pdb*-Datei muss angegeben werden, die die räumliche Struktur des Proteins enthält, mit der die automatische Zuordnung gestartet werden soll. Dies wird zu Beginn des iterativen Prozesses im schlechtesten Fall ein ausgestreckter Strang sein oder, falls vorhanden, eine Modellstruktur. Für die späteren Zuordnungsschritte wird die jeweils zuvor berechnete Struktur verwendet.
- Ausgabedatei: In diese Datei schreibt KNOWNOE z.B. die berechneten Wahrscheinlichkeiten für die einzelnen Zuordnungen.

Parameters / KNOWNOE assignment

Select 2D spectra (only 1)

D:\gab03899\TestSpektren\exp\2rr ☒

D:\gab03899\TestSpektren\meta\2rr ☐

Select filetype Use Metafile ▾

name of meta-file meta_HPr ▾

no. of compound (metafile) 1

name of spt-file ▾

name of as-sequenz HPr.seq ▾

name of pdb-file HPr.pdb ▾

name of output-file output.txt

known distance (2D) 0.24 nm

known volume (2D) 13500

lower probability limit 0.8 %

distance limit 1 nm

known distance (3D) 0 nm

known volume (3D) 0

assign limit F1-3D 0.01 ppm

assign limit F2-3D 0.01 ppm

assign limit F3-3D 0.01 ppm

assign limit F1-2D 0.01 ppm

assign limit F2-2D 0.01 ppm

assign all peaks ☐

assignments to masterlist(s) ☒

keep assignments keep no previous assignments ▾

OK Cancel

Abbildung 3.11: Dialog von KnowNOE. Dargestellt sind die für die automatische Zuordnung zu setzenden Parameter.

- Kalibrierung der Volumina: Für jedes experimentelle Spektrum muss für einen bekannten Abstand das zugehörige Volumen angegeben werden. Man erhält einen Skalierungsfaktor, mit dem die experimentellen Volumina mit den Datenbank-Volumina verglichen werden können.
- Grenzwert P_{min} : Eine Zuordnungsmöglichkeit wird als richtig angenommen, falls gilt $P(A_i, a|V_0) \geq P_{min}$. P_{min} wird im Dialog als *lower probability limit* bezeichnet.
- Distanzgrenzwert: D_{max} gibt die maximale Distanz zweier Spins an, die KNOWNOE als Zuordnungsmöglichkeiten berücksichtigt. D_{max} wird im Dialog als *distance limit* bezeichnet.
- Toleranzwerte: Für jede Dimension der verwendeten NOESY-Spektren kann der Benutzer Toleranzwerte für die chemischen Verschiebungen angeben. Damit trägt man der Tatsache Rechnung, dass die sequentielle Zuordnung und die Zuordnung in den NOESY-Spektren unterschiedlich sind.
- Es besteht die Möglichkeit, die von KNOWNOE ermittelten Zuordnungen automatisch auf die selektierten, geladenen NMR-Spektren zu übertragen, d.h. mit in die *Masterlisten* zu übernehmen. Dabei kann der Benutzer angeben, ob alle bisherigen Zuordnungen in den Spektren überschrieben werden, ob nur manuell bestimmte Signale oder alle Zuordnungen behalten werden.

Nach der automatischen Zuordnung von KNOWNOE und der Übertragung der Zuordnungen auf die *Masterlisten* der zugeordneten Spektren, kann man nun mit dem Kommando *Utilities* \rightarrow *X-Plor output*, s. Abschnitt 3.1.3, eine Abstandsdatei erzeugen, die als Eingabedatei für eine externe Strukturberechnung mit CNS dient. Die neu berechnete Struktur wird wiederum als Teststruktur von KNOWNOE verwendet, d.h. eine neue automatische Zuordnung erfolgt. Daran schließt sich die Erstellung einer neuen Abstandsdatei an und eine weitere Strukturberechnung folgt, bis die Strukturen korrekt gefaltet sind.

3.4.1.2 Automatische Strukturberechnung von *TmCsp*

In diesem Abschnitt soll kurz die automatische Bestimmung der Lösungsstruktur von *TmCsp* gezeigt werden. Als Startstruktur wurde ein ausgestreckter Strang verwendet, sowie ein experimentelles 2D-NOESY-Spektrum in H_2O . Die sequentielle Zuordnungsliste von *TmCsp* stammt von Harrieder bzw. Kremer [49], [61].

Die mit KNOWNOE automatisch bestimmte Struktur von *TmCsp* ist in Abbildung 3.12 im Vergleich mit der durch manuelle Zuordnung erhaltenen Struktur zu sehen. Es wurden 5 Iterationen durchgeführt, wobei sukzessive der Distanzgrenzwert D_{max} von 30 nm auf 0.5 nm reduziert wurde. Die Toleranzwerte für die chemischen Verschiebungen wurden jeweils

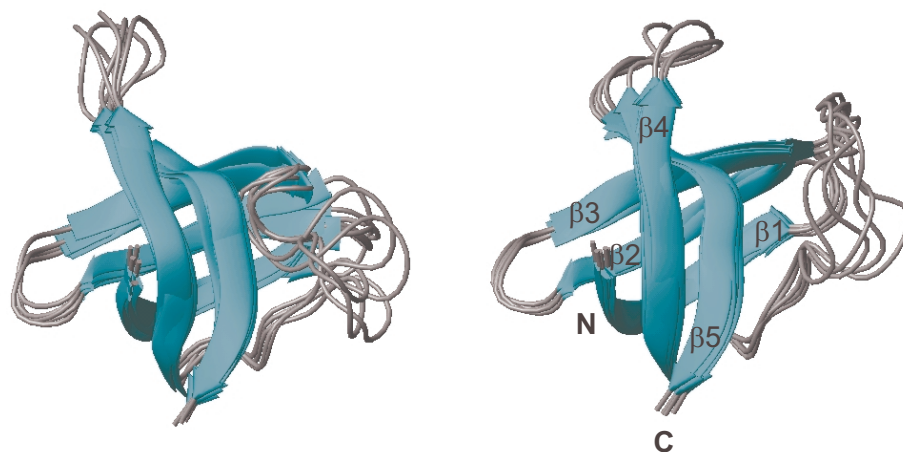


Abbildung 3.12: Vergleich der durch manuelle bzw. automatische NOE-Zuordnung erhaltenen Strukturen von *TmCsp*. Links: KNOWNOE Struktur, rechts: manuell bestimmte Struktur. Zu sehen ist jeweils die Überlagerung der 5 besten Strukturen, d.h. Strukturen mit der geringsten Gesamtenergie.

auf 0.01 ppm gesetzt. P_{min} wurde auf 0.8 und $a = 0.9$ gesetzt. Mit Hilfe der von KNOWNOE zugeordneten Signale und der damit erhaltenen Abstandsbeschränkungen wurden mit CNS 50 Strukturen berechnet. Die 5 Strukturen mit der geringsten Gesamtenergie wurden auf NOE-Verletzungen untersucht und diese verletzten Einschränkungen aus der Abstandsdatei entfernt. Es wurden wieder 50 Strukturen berechnet und die energetisch günstigste ausgewählt. Diese Struktur wurde als Teststruktur für die nächste Iteration verwendet. Insgesamt wurden nach diesem Schema 5 Iterationen durchgeführt.

3.4.2 2-Stufen Algorithmus

Im Folgenden soll die Implementierung des 2-Stufen Algorithmus in AUREMOL dargestellt werden. Mit Beispielen von künstlich erzeugten NOESY-Spektren wird die Robustheit des Verfahrens gegenüber Rauschen, Artefaktsignalen und fehlenden NOE-Signalen getestet. Versuche mit experimentellen Spektren der HPr Mutante H15A zeigen die allgemeine und erfolgreiche Anwendbarkeit des Algorithmus zur Bestimmung fehlender Zuordnungen in den NOESY-Spektren.

3.4.2.1 Implementierung in AUREMOL

In AUREMOL wurden drei Funktionen eingebaut, die im Anhang B aufgeführt sind:

- *Segmentate exp. spectrum*: Es findet, wie in Kapitel 2.3.2.3 ausgeführt, eine Segmentierung des experimentellen Spektrums statt, was im Falle eines 3D-NOESY-Spektrums zu einer erheblichen Reduktion bzw. Kompression führt. In Abbildung 3.13 sind Beispiele für die segmentierten *Peakformen* eines 2D-NOESY- bzw. 3D-NOESY/HSQC-Spektrums zu sehen.

Der Benutzer hat dazu den Namen der Datei anzugeben, in der die segmentier-

Peakshapes2D_exp.dat	Peakshapes3D_exp.dat
EXPERIMENTAL SPECTRUM: NOESY->H/H PEAKNUM: 5472 SW1: 13.979074 SW2: 13.979074 OFF1: -2.240754 OFF2: -2.245805 RES1: 0.013651 RES2: 0.003413 174 11.438 3.327 22 2463 0 194 11.397 3.880 25 2301 0 195 11.397 3.822 25 2318 0 230 11.329 2.696 30 2648 0 245 11.301 3.450 32 2427 0 247 11.301 2.785 32 2622 0 10481 -0.671 0.932 909 3165 0 10500 -0.725 2.068 913 2832 0 10515 -0.753 2.713 915 2643 0 10521 -0.794 0.696 918 3234 0 N1: 1024 N2: 4096 0 0 ... 4737 7676 13130 19925 25518 27665 26572 24601 24378 25099 24660 20925 15169 9008 0 0 ... 0	EXPERIMENTAL SPECTRUM: NOESY/HSQC->H/N/HN PEAKNUM: 301 SW1: 12.731400 SW2: 28.946503 SW3: 5.425200 OFF1: -0.843699 OFF2: 102.001251 OFF3: 6.130400 RES1: 0.049732 RES2: 0.452289 RES3: 0.010596 25 6.367 111.499 6.374 111 43 489 11266161 43 7.859 111.499 6.374 81 43 489 6179051 66 6.517 123.259 6.565 108 17 471 13468255 68 8.755 123.259 6.565 63 17 471 2305962 93 2.836 110.142 6.575 182 46 470 2356357 887 10.296 118.284 10.316 32 28 117 8508873 943 4.478 125.520 11.301 149 12 24 2390028 999 9.152 125.520 11.301 55 12 24 4019651 1033 11.291 125.520 11.301 12 12 24 6845830 N1: 256 N2: 64 N3: 512 PIXEL_ANZ: 5083 12 9 21 2 12 9 22 36 12 9 23 95 12 9 24 36 12 9 25 2 12 10 20 362 244 21 322 798750 244 21 323 2796 244 22 321 79958 244 22 322 22843994 244 22 323 79958 244 23 321 2796 244 23 322 798750 244 23 323 2796
← Zahl der exp. Peaks ← Prozessierungsparameter des exp. Spektrums ← Exp. Peaks: Peaknummer, Position in ppm und Punkten sowie das exp. Volumen ← Auflösung in w_1 und w_2 ← 1:1 Abbild des 2D Spektrums, Intensitäten sind zeilenweise abgelegt	← Zahl der exp. Peaks ← Prozessierungsparameter des exp. Spektrums ← Exp. Peaks: Peaknummer, Position in ppm und Punkten sowie das exp. Volumen ← Auflösung in w_1 , w_2 und w_3 ← Zahl der Intensitätswerte im Spektrum != 0 ← Pixel im Spektrum mit Intensität != 0: Position in w_1 , w_2 und w_3 in Punkten, sowie Intensitätswert

Abbildung 3.13: Beispiele für experimentelle *Peakformen*. Links: Enthält die durch Segmentierung erhaltenen *Peakformen* der experimentellen Signale eines 2D-NOESY-Spektrums. Rechts: Enthält die durch Segmentierung erhaltenen *Peakformen* der experimentellen Signale eines 3D-NOESY/HSQC-Spektrums.

ten *Peakformen* abgelegt werden sollen. Zusätzlich kann ein Verzeichnis gewählt

werden, in welches das segmentierte Spektrum geschrieben wird. Die im Theorie-
 teil 2.3.2.3 erwähnten Parameter zur Angabe der maximalen Segmentierungstiefe in
 Prozent sowie einer Begrenzungsbox, die die maximale Segmentierung in den zwei
 bzw. drei Dimensionen des Spektrums in Hz angibt, sind vom Benutzer festzulegen.
 Üblicherweise werden diese Werte der doppelten maximal erwarteten experimen-
 tellen Breite der Lorenzlinie entsprechen. Die Datei, die nun erzeugt wird, enthält
 neben der Zahl der experimentellen Signale auch prozessierungsspezifische Para-
 meter, wie den Offset und die Spektrenbreite in ppm. Diese Werte werden beim
 Optimierungsalgorithmus zur Erzeugung des simulierten Spektrums benötigt. An-
 schließend sind alle *Peaks* mit den Positionen in ppm bzw. Punkten und dem exper-
 imentell bestimmten Volumen aufgeführt. Mit N1, N2 und im 3D Fall auch N3 ist
 zusätzlich die prozessierte Auflösung des Spektrums angegeben. Im Falle des 2D-
 NOESY-Spektrums folgt dann ein 1:1 Abbild des Spektrums, während im Falle des
 3D-NOESY/HSQC-Spektrums nur die Pixel des Spektrums aufgeführt sind, deren
 Intensitätswerte ungleich Null sind. Im 2D Fall wird ein 1:1 Abbild abgespeichert,
 da der 2-Stufen Algorithmus aus Geschwindigkeitsgründen mit der unkomprimierten
 Form des 2D Spektrums rechnet.

- *Match spectrum (step 1)*: Dies ist die 1. Stufe der Optimierung, in der die partielle
 Zuordnung an jedes der verwendeten experimentellen Spektren einzeln angepasst
 wird. Folgende Angaben müssen vom Benutzer gemacht werden:
 - Datei mit den experimentellen *Peakformen* (*peakshapes_exp.dat*)
 - Datei mit den berechneten *Peakformen* (*peakshapes.dat*)
 - Startwert für *Threshold Th* des TA-Algorithmus
 - Maximale Zahl an Iterationen
 - Primärsequenz des Proteins
 - Datei mit der partiellen Zuordnung. Das Format ist in Tabelle 3.1 gezeigt.

```
Atom:  HA 1   Shift:  4.708
Atom:  HB3 1  Shift:  2.209
Atom:  HB2 1  Shift:  2.133
.....
```

Tabelle 3.1: Eingabedatei mit partieller Zuordnung für die
 1. Stufe der Optimierung. Die chemische Verschiebung in
 ppm für einen Teil der Atome des Proteins ist zeilenweise
 aufgeführt.

- Ausgabedatei mit den an das NOESY-Spektrum angepassten Zuordnungen in dem Format, wie es in Tabelle 3.2 angegeben ist.

Peak:	HD2 81/HD2 81	-0.166	-0.164	872	3486	0.908790
Peak:	HD2 81/HD1 81	-0.166	0.495	872	3293	0.914907
Peak:	HD2 81/HD2 86	-0.166	0.495	872	3293	0.914919
.....						
Peak:	HN 30/HA 29	11.301	4.484	32	2124	0.000000
Peak:	HN 30/HN 31	11.301	9.177	32	749	0.749257

Tabelle 3.2: Nach der 1. Optimierungsstufe wird diese Datei gespeichert, die gleichzeitig als Input für die 2. Stufe der Optimierung dient. Für jeden simulierten NOE wird die Position in ppm und Punkten angegeben, sowie in der letzten Spalte der Matchfaktor, der beschreibt, wie gut das simulierte Signal an dieser Stelle im experimentellen Spektrum erklärt wird.

- Angabe, ob zusätzliche chemische Verschiebungen eingefügt werden sollen. Die Aufspaltung der Signale durch die J-Kopplung kann dazu führen, dass beim *Peak Picking* des experimentellen Spektrums die Schwerpunkte der Signale nicht *gepickt* werden. Der Grund liegt darin, dass bei entsprechender Aufspaltung der Schwerpunkt eines Multipletts in einem Minimum liegt. Diese für die Optimierung fehlenden experimentellen chemischen Verschiebungen können mit dieser Option bestimmt werden. Dazu werden alle chemischen Verschiebungen, die in aufsteigender Reihenfolge sortiert sind, durchlaufen und falls die Differenz zwei aufeinanderfolgender Verschiebungen genau zwei Punkte beträgt, wird eine neue Verschiebung mit dem Mittelwert der beiden eingefügt.
- Anzahl der Iterationsschritte, nach denen überprüft wird, ob der *Th* abgesenkt werden soll. Entscheidungskriterium dafür ist, ob die Gesamtenergie des Systems gesunken ist oder nicht. Ist die Energie nicht gesunken, so wird der *Threshold* abgesenkt.
- *Match spectrum (step 2)*: In der 2. Stufe der Optimierung können nun mehrere NOESY-Spektren gleichzeitig optimiert werden. Die benötigten Eingabeparameter sind dieselben wie in Stufe 1, bis auf folgende Ausnahmen:
 - Für jedes verwendete NOESY-Spektrum muss eine Datei mit den experimentellen (*peakshapes_exp.dat*) bzw. berechneten *Peakformen* (*peakshapes.dat*) angegeben werden.
 - Statt der partiellen Zuordnung aus Tabelle 3.1 werden als Input die angepassten Zuordnungen aus Tabelle 3.2 von Stufe 1 verwendet. Dabei gilt wieder, dass

für jedes Spektrum die angepassten Zuordnungen vorliegen müssen, d.h. jedes NOESY-Spektrum muss zuerst die 1. Stufe der Optimierung durchlaufen.

3.4.2.2 Beispiele mit künstlichen Datensätzen

Die Optimierungsroutine wurde mit künstlichen Datensätzen getestet, d.h. statt experimenteller NOESY-Spektren wurden zurückgerechnete Spektren verwendet. Wie im Kapitel 3.3 dargestellt, wurden dazu eine *Metadatei* und eine *Pdb*-Datei benutzt. Die *Metadatei* enthält unter anderem die sequentielle Zuordnung des verwendeten Proteins HPr, die Strukturinformation ist in der *Pdb*-Datei enthalten. Es wurden 460 ^1H -Verschiebungen, 92 ^{15}N -Verschiebungen und ein *cutoff* von 0.2 nm für die Rückrechnung angesetzt.

Damit wurde ein 2D-NOESY-Spektrum mit 1728 NOE-Signalen und einer Auflösung von 1024×1024 Datenpunkten erzeugt, sowie ein 3D-NOESY/HSQC-Spektrum mit 301 Signalen und einer Auflösung von $256 \times 64 \times 512$ Datenpunkten berechnet. Für das 3D-Spektrum wurden aufgrund des geringen *cutoff* von 0.2 nm von den 460 ^1H -Verschiebungen nur 242 Verschiebungen für die Berechnung benötigt. Mit diesen künstlich erzeugten NOESY-Spektren wurde ein *Peak Picking* durchgeführt, was im 2D-Spektrum zu 2122 *gepickten* Signalen und im 3D-Spektrum zu 283 *Peaks* führte. Da im 2D-Spektrum durch die hohe Auflösung die Signale aufgrund der J-Kopplung in mehrere *Subpeaks* aufspalten, führt dies nach dem *Peak Picking* dazu, dass die *Subpeaks* als eigenständige Signale erkannt werden und die Zahl der Signale sich deshalb von 1728 auf 2122 erhöht hat. Diese Spektren wurden nun mit dem Kommando *Segmentate exp. spectrum*, s. vorheriges Kapitel 3.4.2.1, segmentiert und reduziert. Dabei wurde eine Segmentierungstiefe von 100% und eine maximale Segmentierungsbreite in den Dimensionen w_1 und w_2 von jeweils 30 Hz gewählt. Die Segmentierungsbreite ergab sich aus Abschätzungen zur maximalen Linienbreite der Lorenzkurven. Das so erzeugte künstliche 2D-NOESY-Spektrum ist in Abbildung 3.14 zu sehen.

Das *Peak Picking* hat im 2D-Spektrum zu 429 verschiedenen chemischen Verschiebungen pro Frequenzachse geführt. Für die Simulationen wurde nur das künstlich erzeugte 2D-NOESY-Spektrum und die Zuordnungsliste verwendet, die bereits für die Erzeugung des Spektrums gedient hat, d.h. die Zuordnungen waren bereits exakt auf das experimentelle NOESY-Spektrum angepasst, so dass die 1. Stufe der Optimierung nicht verwendet wurde. Der im Theorieteil 2.3.2.4 eingeführte Toleranzwert tol_{SS} konnte aus diesem Grund und aus der Tatsache, dass die digitale Auflösung der beiden Frequenzachsen identisch war, auf 0 gesetzt werden.

In einer ersten Testreihe wurde gemessen, wieviele Zuordnungen vom Algorithmus in Abhängigkeit von der Zahl der bekannten Zuordnungen (=partielle Zuordnung) gefunden wurden. Mit gefundenen Zuordnungen sind nachfolgend immer Resonanzfrequenzen gemeint und keine gefundenen Kreuzsignale bzw. NOEs, also z.B. HN 30 mit einer che-

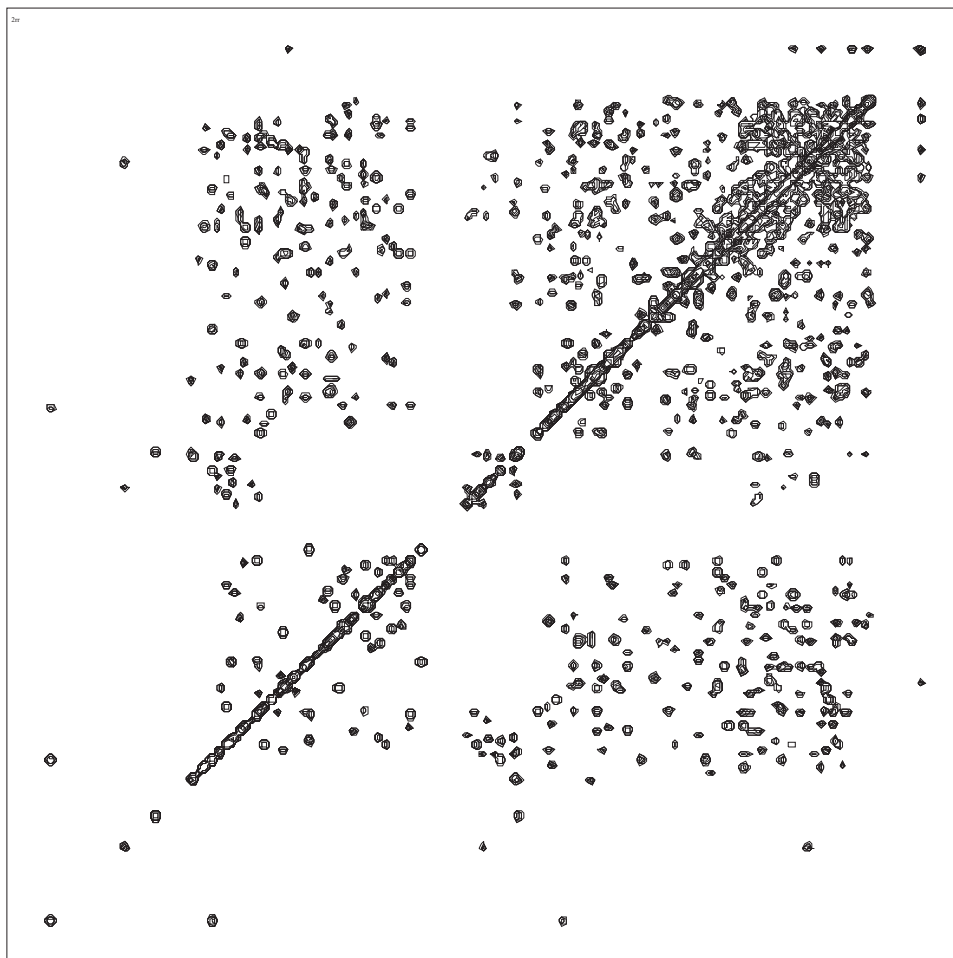


Abbildung 3.14: Beispiel für ein künstlich erzeugtes 2D-NOESY-Spektrum. Es wurden 1728 Signale mit ihrer Multipllettstruktur berechnet, nach dem *Peak Picking* waren es 2122 Signale.

mischen Verschiebung von 8.112 ppm. Die Ergebnisse sind in Abbildung 3.15 zu sehen.

Wie bei allen Simulationen mit künstlichen Datensätzen wurde die Energiefunktion von Gleichung 2.51 minimiert, d.h. der Bestrafungsfaktor p_v auf 0 gesetzt. Die Rechnungen wurden mit einem *Threshold* $Th = 2$ gestartet, wobei nach 5000 Iterationsschritten jeweils geprüft wurde, ob abgekühlt werden sollte. Für die allmähliche Reduzierung des *Thresholds* wurde die logarithmische Methode von Gleichung 2.46 eingesetzt und für den Parameter $a = 0.98$ gesetzt. Die maximale Zahl an Iterationsschritten wurde auf 1000000 festgesetzt. Zusätzlich wurde die Option aktiviert, vor der Optimierung zusätzliche chemische Verschiebungen zu berechnen und diese dann bei der Variation der Zuordnungen während der Simulation zu berücksichtigen. Der Grund liegt darin, wie bereits im vorherigen Kapitel ausgeführt wurde, dass durch die Multiplettaufspaltung der NOE-Signale

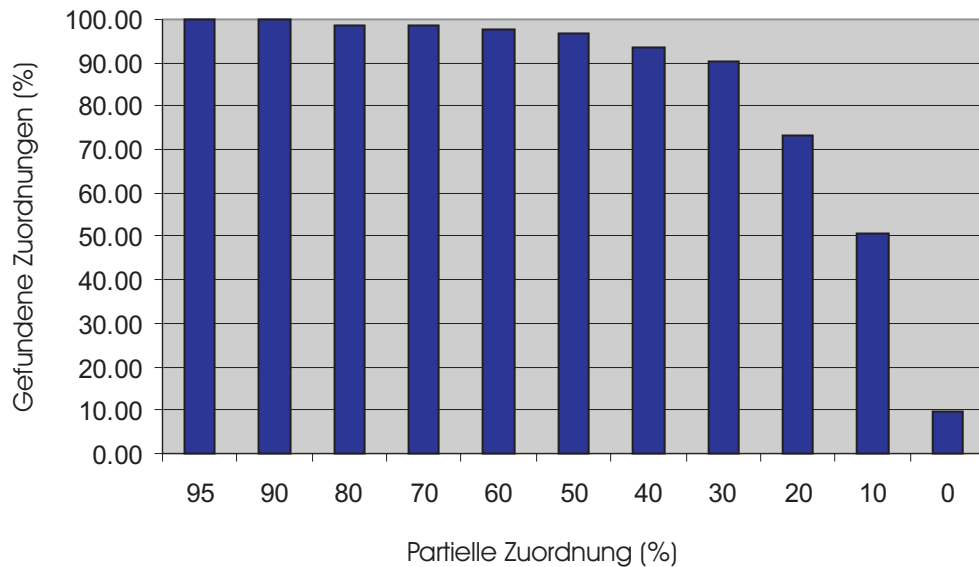


Abbildung 3.15: Zahl der gefundenen Zuordnungen in Abhängigkeit von der partiellen Zuordnung.

und dem nachfolgenden *Peak Picking* zu wenige experimentelle Verschiebungen gefunden werden. Dies hat zu einer Zunahme der experimentellen chemischen Verschiebungen von 429 auf 475 pro Frequenzachse geführt. Die partiellen Zuordnungen wurden vor den Simulationsläufen durch zufälliges Entfernen von Zuordnungen aus dem vollständig bekannten Set an Zuordnungen erzeugt.

Die Rechnungen wurden auf Silicon Graphics Workstations durchgeführt und benötigten je nach Systemlast zwischen 6 und 12 Stunden Rechenzeit. In Abbildung 3.16 ist das Verhalten der Energiefunktion eines typischen Optimierungslaufes gezeigt. Zusätzlich ist die spezifische Wärmekapazität des Systems abgebildet, die sich für den *Threshold Accepting* Algorithmus wie folgt berechnet:

$$C = \frac{1}{Th^2} Var(E) = \frac{1}{Th^2} (\langle E^2 \rangle - \langle E \rangle^2) \quad (3.4)$$

Die übliche Definition enthält statt Th die Temperatur des Systems, was im Falle des TA-Algorithmus aber identisch zum *Threshold* Th ist. Wie man anhand der Energiekurve und der spezifischen Wärmekapazität sieht, zeigt das System ein korrektes Verhalten. Oberhalb eines gewissen *Thresholds* führt das System einen *Random Walk* durch, bevor es einzufrieren beginnt. In der Abbildung 3.16 im rechten Teilbild ist dies bei etwa $Th = 3$ der Fall. Die Temperatur oder *Threshold*, bei der die spezifische Wärmekapazität maximal wird, nennt man Einfriertemperatur.

Wie man in Abbildung 3.15 sieht, benötigt man eine partielle Zuordnung von 30%, um

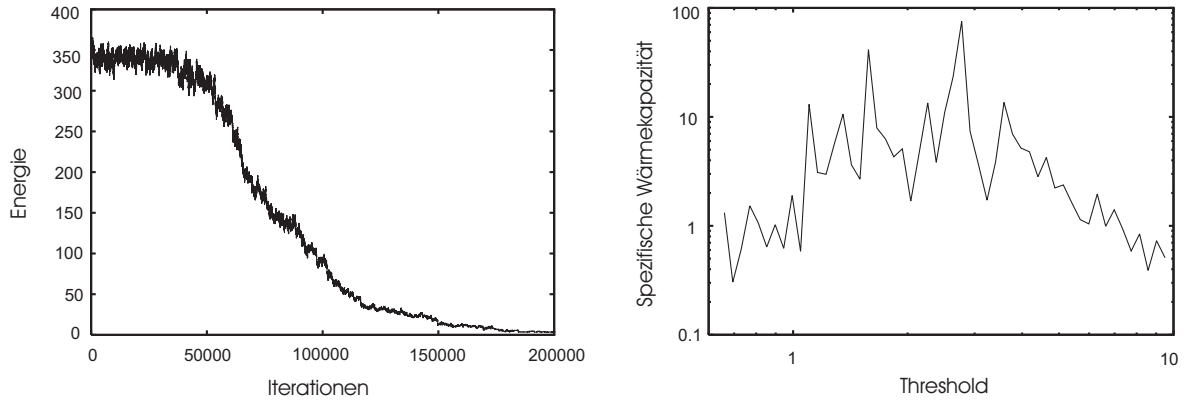


Abbildung 3.16: Links: Energie als Funktion der Iterationen eines typischen Optimierungslaufes, rechts: Spezifische Wärmekapazität des Systems in Abhängigkeit vom Threshold in doppelt-logarithmischer Darstellung.

mehr als 90% aller Zuordnungen zu finden. Das bedeutet, dass in diesem Beispiel mit 460 chemischen Verschiebungen, mindestens 140 Zuordnungen bekannt sein müssen, um mehr als 90%, also ca. 415 Zuordnungen, vom Algorithmus finden zu lassen. Bei 20% bzw. 10% partieller Zuordnung werden 70% bzw. 50% aller Verschiebungen korrekt zugeordnet. Nimmt man aber an, dass keine partielle Zuordnung vorhanden ist, so werden nur etwa 10% aller Zuordnungen bzw. Resonanzfrequenzen richtig gefunden.

In der Realität wird man nie ideale Datensätze vorfinden. Meistens sind die Datensätze verrauscht, es sind Artefaktsignale enthalten oder Signale fehlen. Im Folgenden wird das Verhalten des Algorithmus darauf näher untersucht.

1. Rauschen im künstlichen Datensatz

Der verwendete künstliche 2D-Datensatz wurde folgendermaßen verrauscht. Zuerst wurde eine mittlere Intensität I_{mean} über alle Pixelwerte des 2D-Spektrums ermittelt, die ungleich 0 waren, d.h.

$$I_{mean} = \frac{1}{N_I} \sum_{i=0}^N I_i \quad (3.5)$$

wobei N_I die Zahl der Pixelwerte im Spektrum ist, deren Intensitäten ungleich 0 sind, $N = N_1 \times N_2$ ist und I_i die Intensität des Pixels i ist. N_1 und N_2 sind die Anzahl der digitalen Punkte pro Frequenzachse. Anschließend wurden alle Pixelwerte¹³ des Spektrums folgendermaßen verändert

$$I_i = I_i + n_r I_{mean} r \quad (3.6)$$

¹³unabhängig davon ob die Intensität 0 war

$n_r \in [0, 100]$ ist dabei ein extern vorgegebener Parameter, der angibt wie stark die Spektren verrauscht werden und r ist eine Zufallszahl im Intervall $[-0.01, \dots, 0.01]$. Das so erzeugte Rauschen verrauscht die Spektren stärker als gaussverteiltes oder thermisches Rauschen. Allerdings soll auch nur das generelle Verhalten des Algorithmus mit verrauschten Spektren untersucht werden. In Abbildung 3.17 werden zwei Ausschnitte aus einem künstlich erzeugten 2D-NOESY-Spektrum gezeigt.

Im linken Teilbild ist das nicht verrauschte Spektrum zu sehen, im rechten Teilbild

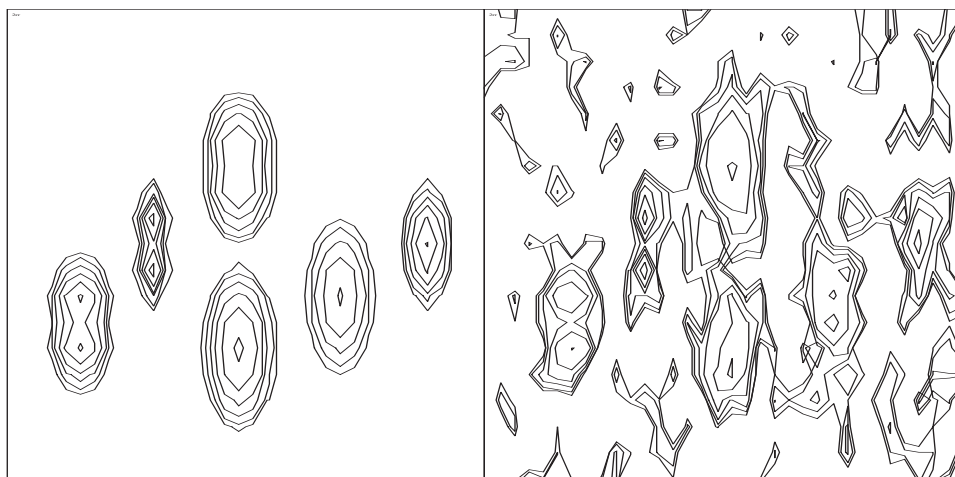


Abbildung 3.17: Links: Ausschnitt aus einem nicht verrauschten Spektrum, rechts: Gleicher Ausschnitt aus selbem Spektrum, aber mit $n_r = 50$ verrauscht.

ein mit einer Rauschrate von $n_r = 50$ verrauschtes künstliches NOESY-Spektrum. In Abbildung 3.18 ist die Abhängigkeit der Zahl der gefundenen Verschiebungen von der Rauschrate n_r zu sehen.

Es wurde dabei eine partielle Zuordnung von 80% angenommen. Die Simulationen wurden mit einer maximalen Zahl an Iterationen von 300000 durchgeführt, auch hier wurde mit einem *Threshold* $Th = 2$ gestartet, allerdings wurde bereits nach 2000 Iterationsschritten geprüft, ob abgekühlt werden musste. Die durchschnittliche Rechenzeit auf einer SGI Workstation betrug für einen Optimierungslauf 2 Stunden.

2. Zusätzliche Artefaktsignale

Im künstlichen Datensatz wurden zusätzliche Signale eingebaut. Zuerst wurde die Liste der experimentellen chemischen Verschiebungen um N_c Stück erweitert. Es wurden nur solche neuen Verschiebungen akzeptiert, die innerhalb des durch die experimentellen Verschiebungen erzeugten Intervalls lagen. Anschließend wurden N_a Artefaktsignale in den künstlichen Datensatz eingebaut. In einer Schleife, die N_a mal durchlaufen wurde, wurde zufällig ein NOE-Signal ausgewählt. Zwei chemische

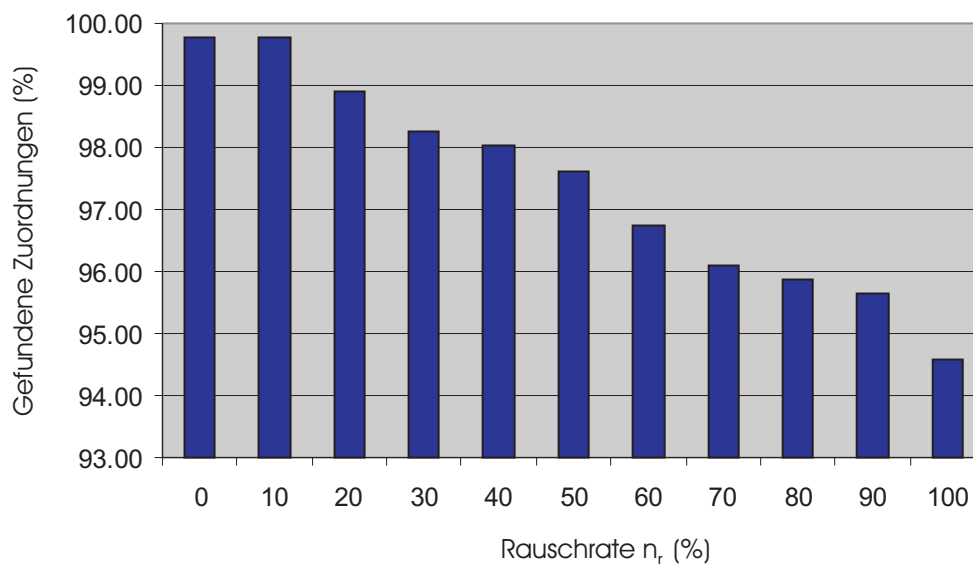


Abbildung 3.18: Zahl der gefundenen Zuordnungen in Abhängigkeit von der Rauschrate n_r . Die partielle Zuordnung zu Beginn der Simulation war 80%.

Verschiebungen wurden diesem Signal zugewiesen, wobei mindestens eine neu sein musste, und dieses Signal wurde dann im Spektrum an dieser Position zusätzlich eingebaut. Das obere Teilbild der Abbildung 3.19 zeigt die Zahl der gefundenen Verschiebungen in Abhängigkeit von der Zahl der Artefaktsignale.

Analog der Variation der Rauschrate wurde auch hier eine partielle Zuordnung von 80% verwendet. Die Parameter für die maximale Zahl an Iterationen, der Startwert für den *Threshold* und das Abkühlschema wurden gleich gewählt wie oben. Bei 100-600 zusätzlichen Artefaktsignalen wurden 5 neue chemische Verschiebungen für jede Dimension des Spektrums hinzugefügt, für 800-1800 Artefakte wurden 10 neue Verschiebungen gewählt.

3. Fehlende NOE-Signale

Aus dem künstlichen 2D-NOESY-Datensatz wurden nun echte NOE-Signale entfernt. Das bedeutet, es wurde für jedes zufällig ausgewählte Signal getestet, ob es sich dabei um ein zusätzlich hinzugefügtes Artefaktsignal handelt. Falls es kein Artefaktsignal war, wurde dieses Signal im Datensatz entfernt, d.h. die isolierte *Peakform* an der Position des Signals aus dem Spektrum subtrahiert. Im unteren Teilbild von Abbildung 3.19 ist die Zahl der gefundenen Verschiebungen in Abhängigkeit von der Zahl der entfernten NOE-Signale dargestellt.

Auch hier wurde eine partielle Zuordnung von 80% und die gleichen Optimierungsparameter wie in den ersten beiden Fällen verwendet.

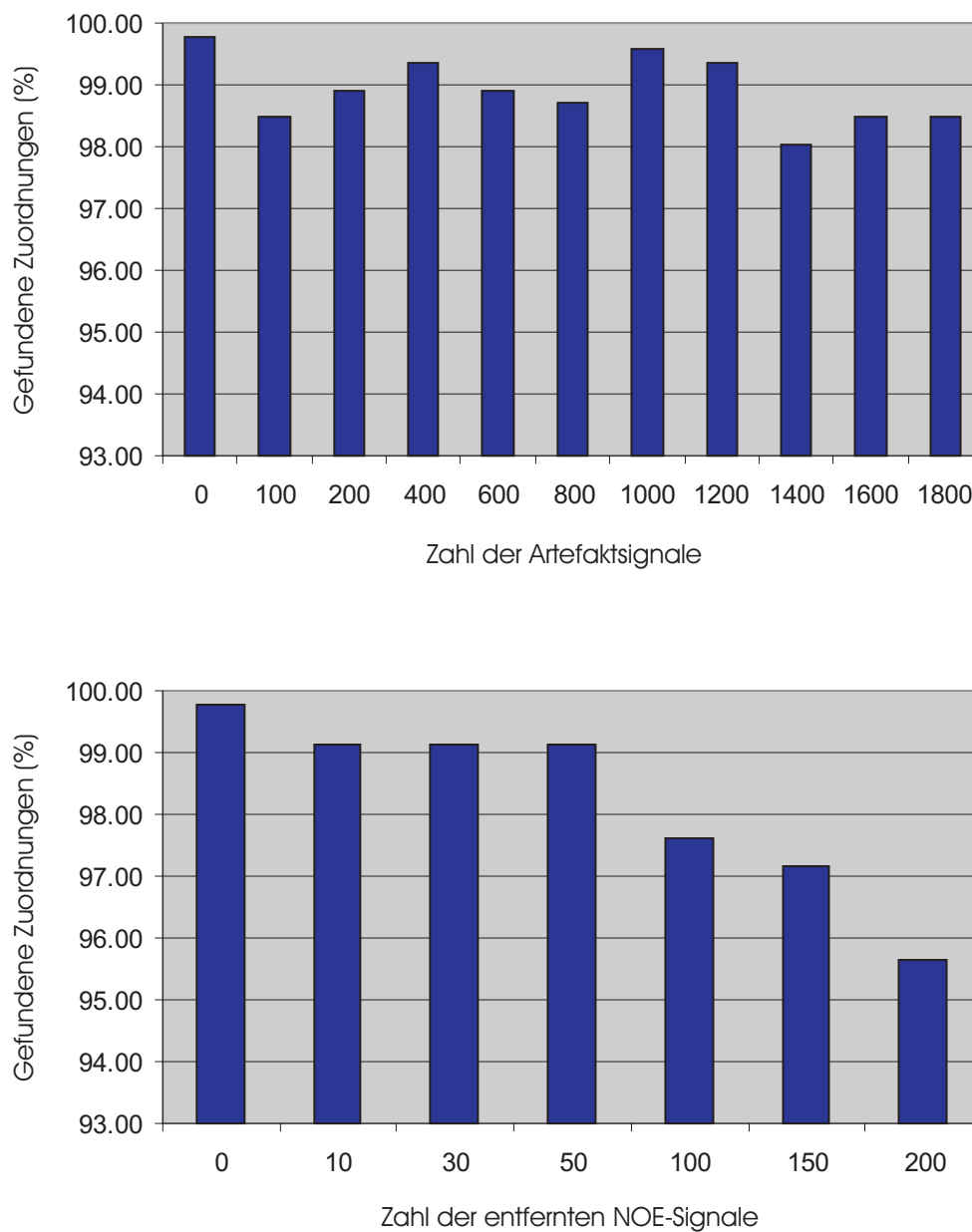


Abbildung 3.19: Oben: Zahl der gefundenen Zuordnungen in Abhängigkeit von der Anzahl der Artefaktsignale. Unten: Zahl der gefundenen Zuordnungen in Abhängigkeit von der Anzahl der entfernten NOE-Signale. In beiden Fällen wurde eine partielle Zuordnung von 80% angenommen.

Die Korrelation dieser drei Effekte wurde im Folgenden untersucht. In einem Simulationslauf wurde der künstliche Datensatz mit $n_r = 20$ verrauscht, es wurden 100 Artefakt-signale hinzugefügt, wobei 5 neue chemische Verschiebungen ermittelt wurden. Zusätzlich wurden 10 NOE-Signale aus dem Spektrum entfernt. Mit einer 80%-igen partiellen Zuordnung startend, wurden vom Algorithmus 455 von 460 Zuordnungen gefunden, also 99%. Es wurden 500000 Iterationsschritte durchgeführt, beginnend mit einem *Threshold* $Th = 2$ und einer Rechenzeit von 3,5 Stunden. Bei den bisherigen Simulationsläufen wurde bei der ersten Hälfte der Iterationsschritte die GAUSS-Methode für die Variation der Zuordnungen verwendet und bei der zweiten Hälfte die RANDOM-Methode, s. Kapitel 2.3.2.4.

Nun werden die Ergebnisse mit dem künstlich erzeugten 3D-NOESY/HSQC-Spektrum gezeigt. Das *Peak Picking* hat zu 130, 45 bzw. 82 verschiedenen chemischen Verschiebungen für die Frequenzachsen w_1 , w_2 bzw. w_3 geführt. Ähnlich dem 2D-Spektrum wurde für die Simulationen die Zuordnungsliste verwendet, die für die Erzeugung des 3D-Spektrums eingesetzt wurde. Das heißt, die 1. Stufe der Optimierung wurde nicht verwendet, der Toleranzwert tol_{SS} wurde aber nicht auf Null gesetzt, da die beiden Protonenachsen w_1 und w_3 eine unterschiedliche spektrale Breite und unterschiedliche digitale Auflösung gehabt haben. Das segmentierte und reduzierte 3D-Spektrum bestand noch aus 4414 Intensitätswerten ungleich Null. Damit wurde die Größe des ursprünglichen 3D-Spektrums von ca. 32 MB auf ca. 40 KB reduziert, da für jeden Intensitätswert die Position im Spektrum und der Intensitätswert selbst gespeichert wird. Die Position ist durch drei 16-Bit Zahlen festgelegt.

In einem ersten Test wurden verschiedene bekannte chemische Verschiebungen aus der Zuordnungsliste entfernt und es wurde untersucht, wieviele der Zuordnungen wieder gefunden wurden. In Tabelle 3.3 sind die Ergebnisse in Abhängigkeit von den Optimierungsparametern aufgelistet.

Für den Fall, dass alle ^{15}N - bzw. HN-Verschiebungen fehlen, wurden diese alle gefunden. Wurden alle ^1H - bzw. die ^1H -Verschiebungen von w_1 entfernt, so hat der Algorithmus 78% bzw. knapp 90% wieder gefunden. Das Entfernen aller bekannten Verschiebungen aus der Zuordnungsliste, also eine partielle Zuordnung von 0%, ergab, dass ca. 50% der Verschiebungen gefunden wurden.

Schließlich wurde noch die gleichzeitige Optimierung der künstlich erzeugten 2D- und 3D-NOESY-Spektren getestet. In einem Fall war das 2D-Spektrum bereits vollständig zugeordnet und das 3D-Spektrum komplett unzugeordnet, d.h. alle 92 ^{15}N - bzw. 242 ^1H -Verschiebungen unbekannt. Es wurde dabei ein Start-Treshold $Th = 1.0$ und eine maximale Zahl an Iterationen von 1000000 verwendet. Nach 10000 Schritten wurde abgekühlt, tol_{SS} betrug 0.05 ppm und tol_{DS} 0.1 ppm. Damit wurden vom Algorithmus 314 der insgesamt 334 Verschiebungen gefunden, was mehr als 94% entspricht. In einem letzten Test wurden aus den Zuordnungslisten für das 2D- und 3D-NOESY-Spektrum alle Amidproto-

Entfernte Verschiebungen	gefunden/von	in %	Th	MaxIter	Steps	tol_{SS}/ppm
alle ^{15}N	92/92	100%	2	300000	2500	0.01
alle ^1H	189/242	78%	2	500000	3000	0.01
	188/242	78%	1.5	500000	10000	0.01
	189/242	78%	1.5	500000	4000	0.05
alle ^1H von w_1	217/242	90%	1.5	500000	5000	0.01
	210/242	87%	1.0	500000	5000	0.01
alle ^1H von w_3	98/98	100%	1.5	500000	5000	0.01
	98/98	100%	1.0	500000	5000	0.01
alle ^1H und ^{15}N	162/334	49%	1.5	1000000	15000	0.05
	174/334	53%	1.0	1000000	15000	0.05

Tabelle 3.3: Zahl der gefundenen chemischen Verschiebungen im 3D-Spektrum in Abhängigkeit von der Zahl der anfangs unbekannten Verschiebungen und der verwendeten Optimierungsparmetern. Der Parameter Steps bezeichnet die Zahl der Schritte nach denen frühestens gekühlt wird, MaxIter die Anzahl der maximalen Iterationen.

nen der Aminosäuren 30 bis 50 entfernt, d.h. die chemischen Verschiebungen der Protonen HN 30 bis HN 50 waren unbekannt. Mit $Th = 1.0$, einer maximalen Zahl an Iterationen von $MaxIter = 3000000$, $Steps = 20000$ und denselben Werten für tol_{SS} und tol_{DS} wie oben, wurde eine Simulation durchgeführt. Bis auf das Amidproton HN 49 wurden alle 21 Protonenverschiebungen richtig ermittelt.

3.4.2.3 Beispiele mit experimentellen Datensätzen

Der 2-Stufen Algorithmus wurde nun mit experimentellen Datensätzen getestet. Es wurde ein 2D-NOESY-Spektrum von der HPr Mutante H15A verwendet, aufgenommen an einem 600 MHz Bruker Spektrometer. Die Primärsequenz des 88 Aminosäuren großen Proteins ist in Abbildung 3.20 zu sehen.

In der indirekten Dimension w_1 wurden 1024 Punkte aufgenommen, in der direkten Di-

MET GLU GLN ASN SER TYR VAL ILE ILE ASP GLU THR GLY ILE ALA ALA ARG PRO ALA THR MET LEU
VAL GLN THR ALA SER LYS PHE ASP SER ASP ILE GLN LEU GLU TYR ASN GLY LYS LYS VAL ASN LEU LYS
SER ILE MET GLY VAL MET SER LEU GLY VAL GLY LYS ASP ALA GLU ILE THR ILE TYR ALA ASP GLY SER
ASP GLU SER ASP ALA ILE GLN ALA ILE SER ASP VAL LEU SER LYS GLU GLY LEU THR LYS

Abbildung 3.20: Primärsequenz der HPr Mutante H15A im Drei-Buchstabencode

mension w_2 4096. Um die Multiplettstruktur aufzulösen, wurde in der direkten Dimension als Fensterfunktion eine Exponentialfunktion verwendet mit einem Parameter $LB=0.3$

Hz. In der indirekten Dimension wurde ein quadrierter Sinus mit $SSB=2$ benutzt. Nach dem Prozessieren des Spektrums mit einer Auflösung von 4096×1024 hat sich ein sehr schlechtes Signal-Rausch-Verhältnis gezeigt, die Ursache wurde im Theorieteil erläutert. Die anschließende Bayessche Analyse zur automatischen Artefaktentfernung lieferte mangelhafte Ergebnisse, da dieses Verfahren keine Multipletts erkennt. Das heißt, es fand keine ordentliche Unterscheidung zwischen NMR-Signalen und Rauschsignalen statt.

Daher wurde das Spektrum neu prozessiert, wobei als Fensterfunktion eine Gaussfunktion mit $LB=-8.0$ Hz und $GB=0.1$ in der indirekten Dimension und $LB=-6.0$ Hz und $GB=0.12$ in der direkten Dimension verwendet wurde. Damit wurde ein besseres Signal-Rausch-Verhältnis erzielt. Prozessiert wurde mit einer digitalen Auflösung von 1024×4096 Punkten. Das anschließende *Peak Picking* lieferte 10307 Signale. Mit der Bayesschen Analyse wurden ca. 3000 Artefaktensignale automatisch entfernt, zusätzlich wurden alle Signale im Bereich zwischen 4.5 ppm und 5.0 ppm in der direkten Dimension entfernt. In dieser Region liegt die Wasserlinie. Dadurch reduzierte sich die Zahl der Signale auf 5472.

Dieses Spektrum wurde nun mit dem Kommando *Segmentate exp. spectrum*, s. vorheriges Kapitel 3.4.2.1, segmentiert. Dabei wurde eine Segmentierungstiefe von 90% und eine maximale Segmentierungsbreite in den Dimensionen w_1 und w_2 von 30 Hz gewählt. In w_1 bzw. w_2 ergaben sich 727 bzw. 1534 verschiedene chemische Verschiebungen.

Aufgrund der Verbreiterung des experimentellen Spektrums wurde bei der Berechnung der simulierten *Peakformen* die J-Kopplung nicht berücksichtigt, d.h. die Multiplettstruktur wurde nicht berechnet. Für die Rückrechnung wurde die Lösungsstruktur der HPr Mutante aus der Doktorarbeit von Claudia Munte [82] verwendet. Als *cutoff* wurden 0.4 nm gewählt, alle anderen spektrenspezifischen Parameter wurden analog zum experimentellen NOESY-Spektrum gesetzt. Es wurden 5568 NOEs simuliert und deren *Peakformen* berechnet.

Für die nachfolgenden Simulationen wurde die Energiefunktion aus Gleichung 2.52 verwendet, wobei $p_v = 0.5$ gesetzt wurde. In einem ersten Schritt wurde die sequentielle Zuordnungsliste in der 1. Stufe des Algorithmus an das experimentelle NOESY-Spektrum angepasst. Da die Liste bereits gut an das verwendete 2D-NOESY-Spektrum angepasst war, wurde der Toleranzwert tol_{seq} auf 0.01 ppm gesetzt. Der Toleranzwert tol_{SS} , der die maximal erlaubte Abweichung der Zuordnung verschiedener Frequenzachsen eines Spektrums angibt, wurde ebenfalls auf 0.01 ppm festgelegt. Wie im Theorieteil 2.3.2 bereits erwähnt, wird der Matchfaktor M_i eines simulierten Signals i in der 1. Stufe nur dann berechnet, falls im experimentellen Spektrum an dieser Stelle ein *gepicktes* Signal liegt, ansonsten wird $M_i = 0$ gesetzt. Es wurden 100000 Iterationen durchgeführt und mit einem *Threshold* $Th = 2$ gestartet. Das Ergebnis, eine an das experimentelle NOESY-Spektrum angepasste Zuordnungsliste, wurde nun im Folgenden als Eingabe für die 2. Stufe der Optimierung übernommen.

Es wurde getestet, inwieweit der Algorithmus in der Lage ist, ausgehend von den be-

kannten an das NOESY-Spektrum angepassten Zuordnungen fehlende chemische Verschiebungen wieder zu finden. Zu Beginn wurden nur einzelne Zuordnungen entfernt, also beispielsweise die Verschiebung von HN 40, alle anderen wurden als bekannt vorausgesetzt. In der Tabelle 3.4 sind die Simulationsergebnisse für alle Spins der Zuordnungsliste aufgeführt.

Die Spins zwischen 5.5 und 4.5 ppm wurden nicht untersucht, da das reduzierte experi-

Entfernte Verschiebungen	Gefunden	Total	Δ in ppm	MaxIter	Th
alle Amidprotonen	58	86	± 0.05	30000	0
Seitenketten 11.0 bis 6.0 ppm	14	31	± 0.05	5000	0
Seitenketten 4.5 bis 1.0 ppm	103	326	± 0.05	5000	0
Seitenketten 1.0 bis -0.2 ppm	16	46	± 0.05	5000	0

Tabelle 3.4: Aufgelistet sind die Spins, deren bekannte Zuordnungen einzeln aus der angepassten Zuordnungsliste von der 1. Stufe entfernt wurden, d.h. für jeden Simulationslauf wurden bis auf eine Zuordnung alle anderen als bekannt vorausgesetzt. Unterschieden wurden die Amidprotonen, deren Verschiebungen im Bereich zwischen 11.0 und 6.0 ppm liegen und die Seitenkettenprotonen, deren Verschiebungen im Bereich zwischen -1.0 und 11.0 ppm liegen. Mit Δ ist die Genauigkeit angegeben, mit der die Zuordnungen im Vergleich zur sequentiellen Zuordnung übereinstimmen müssen, um als korrekt gefunden zu gelten.

mentelle Spektrum dort aufgrund der oben genannten Vorverarbeitung keine Signale mehr enthielt. In diesem Bereich, in dem sich die Wasserlinie befindet, liegen oft die Signale der HA Protonen. In der Tabelle ist die Anzahl der Spins angegeben, deren ermittelte chemische Verschiebung innerhalb eines Toleranzwertes Δ mit der sequentiellen Zuordnung übereinstimmt. Es wurde dabei $\Delta = \pm 0.05$ ppm gesetzt. Die Rechnungen wurden auf einem Windows NT PC mit einem 400MHz Prozessor durchgeführt und betrugen je nach Iterationsanzahl zwischen 10 und 70 min. Von den Amidprotonen wurden etwa 70% aller chemischen Verschiebungen gefunden, von den Seitenketten, deren chemische Verschiebungen links der Wasserlinie (11.0 bis 6.0 ppm) liegen, ca. 50% und von den Seitenketten, deren Verschiebungen rechts der Wasserlinie (4.5 bis -0.2 ppm) liegen, etwa 30%.

Mit den Spins, deren chemische Verschiebungen im experimentellen Spektrum korrekt ermittelt wurden, s. Tabelle 3.4, wurden weitere Simulationen durchgeführt. Dazu wurden diese Spins zu größeren Gruppen zusammengefasst und weitere Optimierungsläufe gestartet, wobei deren chemische Verschiebungen als unbekannt vorausgesetzt wurden. Die Ergebnisse sind in Tabelle 3.5 zu sehen.

Von den Verschiebungen der Amidprotonen der Aminosäuren 6 bis 86 wurden knapp 80% korrekt ermittelt für den Fall, dass alle 58 chemische Verschiebungen als nicht bekannt vorausgesetzt wurden. Von den Verschiebungen der 16 entfernten Amidprotonen

Entfernte Gruppen	Gefunden	Total	Δ in ppm	MaxIter	Th	Steps
Amidprotonen HN 6 - HN 29	14	16	± 0.05	50000	1.0	4000
Amidprotonen HN 6 - HN 86	45	58	± 0.05	150000	1.0	5000
Amidprotonen HN 6 - HN 86	43	58	± 0.05	100000	1.0	3000
Amidprotonen HN 6 - HN 86	45	58	± 0.05	50000	1.0	3000

Tabelle 3.5: Aufgelistet sind die Gruppen von Spins, deren bekannte Zuordnungen aus der angepassten Zuordnungsliste aus der 1. Stufe entfernt wurden, in diesem Fall die Amidprotonen. Mit Δ ist die Genauigkeit angegeben, mit der die Zuordnungen im Vergleich zur sequentiellen Zuordnung übereinstimmen müssen, um als korrekt gefunden zu gelten.

der Aminosäuren 6 bis 29 wurden sogar fast 90% vom Algorithmus richtig zugeordnet. Die Simulationen, die auf einem 866 Mhz Windows 2000 Rechner durchgeführt wurden, dauerten je nach Anzahl der Iterationsschritte zwischen 28 und 75 Minuten.

Als nächstes wurden die Signale der Seitenkettenprotonen näher untersucht. Wieder wurden die in Tabelle 3.4 gefundenen Spins gruppiert und Verschiebungen dieser Gruppen aus der Liste der bekannten angepassten Zuordnungen entfernt. Die vom Algorithmus erhaltenen Ergebnisse sind in Tabelle 3.6 aufgeführt.

Die Seitenkettenprotonen wurden in 3 Gruppen eingeteilt. Links von der Wasserlinie

Entfernte Gruppen	Gefunden	Total	Δ in ppm	MaxIter	Th	Steps
Zwischen 11.0 und 6.0 ppm	12	15	± 0.05	100000	1.0	2500
Zwischen 4.5 und 1.0 ppm	75	103	± 0.05	200000	1.0	5000
Zwischen 1.0 und -0.2 ppm	13	16	± 0.05	100000	1.0	3000

Tabelle 3.6: Aufgelistet sind die Gruppen von Spins, deren Zuordnungen aus der angepassten Zuordnungsliste aus der 1. Stufe entfernt wurden, in diesem Fall die Seitenkettenprotonen. Mit Δ ist die Genauigkeit angegeben, mit der die Zuordnungen im Vergleich zur sequentiellen Zuordnung übereinstimmen müssen, um als korrekt gefunden zu gelten.

sind die Spins mit Verschiebungen im Bereich von 11.0 und 6.0 ppm zu finden, die Spins rechts der Wasserlinie wurden in einen Bereich zwischen 4.5 und 1.0 ppm eingeteilt, sowie zwischen 1.0 und -0.2 ppm. Die letzten beiden Regionen sind geprägt durch eine sehr starke Konzentration an NOE-Signalen, d.h. durch einen sehr großen Überlappbereich, der generell zu Problemen bei der Zuordnung führt. Von den Seitenkettenprotonen, deren Verschiebungen links der Wasserlinie liegen, wurden 80% korrekt zugeordnet, rechts der Wasserlinie zwischen 70 und 80%.

Schließlich wurde der Algorithmus auf das Fehlen von Zuordnungen kompletter Spinsysteme getestet. Es wurden 88 Optimierungsläufe durchgeführt, in denen jeweils die che-

mischen Verschiebungen einer Aminosäure gefehlt haben, also z.B. alle Verschiebungen von Gly 85: HN 85, HA2 85 und HA3 85 waren unbekannt.

Die 88 Optimierungsläufe wurden dreimal ausgeführt, wobei $Th = 1.0$, $MaxIter = 50000$ und $Steps = 3000$ gesetzt war. In allen 3 Fällen wurden 11 Spinsysteme richtig zugeordnet: Thr 20, Thr 25, Ala 26, Ser 46, Ser 52, Leu 53, Gly 67, Asp 72, Ala73, Ser 78, Ser 82. 16 weitere Spinsysteme wurden bis auf eine falsche Zuordnung richtig zugeordnet. Pro Spinsystem dauerte ein Simulationslauf etwa 80 min, was bei einem kompletten Durchlauf aller 88 Aminosäuren etwa 120 h bedeutet. Diese Rechnungen wurden auf einem 400 MHz schnellen Windows NT PC durchgeführt.

3.5 Qualitätsbeurteilung von NMR-Strukturen mit R-Werten

Das Konzept von AUREMOL, s. Abbildung 2.1, beruht auf einer iterativen Verfeinerung der gesuchten Proteinstruktur, wobei der automatische Zuordnung von NOE-Signalen eine externe Strukturberechnung folgt. Anschließend wird mit Hilfe der R-Wert Berechnung die Qualität der berechneten Struktur beurteilt und als Abbruchkriterium für den iterativen Kreisprozess verwendet.

Wie im Theoriekapitel 2.4 erläutert, gibt es zahlreiche Definitionen für verschiedene R-Werte. Verwendet wurden die R-Werte R_3 und R_5 . Sie unterscheiden sich dadurch, dass einmal die nicht zugeordneten Signale berücksichtigt werden, das andere mal nicht.

3.5.1 Implementierung in AUREMOL

Die Berechnung der R-Werte erfordert folgende Eingabedateien:

- Liste der zugeordneten Signale: Dies ist die im Theorieteil bezeichnete A-Liste, in der bereits zugeordnete experimentelle Signale enthalten sind. Ein Auszug aus einer Beispieldatei ist in Tabelle 3.7 zu sehen.

Das Format der A-Liste entspricht dem Format der Log-Datei, die AUREMOL

#	3.971	8.996	HB2 68/HN 68	0.149455	0.000000	2009826
#	0.776	8.276	HG2 50/HN 50	0.320681	0.000000	1463970
#	3.070	7.628	HA2 13/HN 14	0.199749	0.000000	1540491
#	0.640	6.873	HD3 61/HD3 6	0.149444	0.000000	2037202
#	2.797	8.194	HB2 32/HN 32	0.306699	0.000000	4348013
#	1.882	6.710	HB2 36/HE2 64	0.237719	0.000000	5531436
#	2.155	7.856	HB 80/HN 80	0.162221	0.000000	3112717
.....						

Tabelle 3.7: Liste mit zugeordneten experimentellen Signalen (A-Liste), beginnend mit den *Peakpositionen* in ppm, dem *Peaknamen* und als letztem Parameter dem Volumen des Signals.

bei der Rückrechnung von NOESY-Spektren anlegt, s. Abschnitt 3.3. In dieser Datei werden bei der anschließenden automatischen Zuordnung [39] die zugeordneten experimentellen NOE-Signale eingetragen.

- Liste der zurückgerechneten Signale: In dieser Liste (B-Liste) sind alle simulierten NOE-Signale enthalten. Bei der Rückrechnung in AUREMOL wird eine solche Liste

automatisch angelegt, s. Abschnitt 3.3. Ein Auszug aus einer B-Liste ist in Tabelle 3.8 zu sehen.

Für die R-Wert Berechnung verwendet werden hieraus der *Peakname*, die chemi-

name 1	name 2	shift 1	shift 2	dist [Å]	rate [Hz]	noe	width 1 [Hz]	width 2 [Hz]
HD2 81	HD1 81	-0.169	0.492	3.363	-0.588	1.1e-01	13.359	14.041
HD2 81	HD2 86	-0.169	0.492	3.960	-0.220	4.0e-02	13.359	13.154
HD2 81	HD1 8	-0.169	0.492	5.455	-0.032	6.5e-03	13.359	14.098
HD2 81	HD1 86	-0.169	0.635	6.086	-0.017	6.8e-03	13.359	14.976
HD2 81	HD1 61	-0.169	0.649	4.693	-0.079	1.5e-02	13.359	14.116
HD2 81	HG2 77	-0.169	0.703	5.348	-0.036	1.7e-02	13.359	14.289
HD2 81	HD1 22	-0.169	0.710	2.992	-1.185	1.7e-01	13.359	14.375
HD2 81	HD2 22	-0.169	0.710	3.966	-0.218	4.8e-02	13.359	13.221
.....								

Tabelle 3.8: Liste mit rückgerechneten NOE-Signalen. Für die R-Wert Berechnung benötigt wird der *Peakname*, die chemischen Verschiebungen in ppm sowie die berechneten Volumina in der Spalte noe.

schen Verschiebungen in ppm sowie die berechneten Volumina aus der Spalte noe in der Tabelle 3.8.

- Masterliste der experimentellen Signale: Dazu muss das entsprechende experimentelle Spektrum in AUREMOL geladen und als aktives Objekt markiert sein. Im Spektrum muss zuvor ein *Peak Picking* und eine automatische Artefaktentfernung mit der Bayesschen Analyse durchgeführt werden. Anschließend muss für die restlichen Signale das Volumen mittels Integration bestimmt werden. In der Masterliste müssen demnach neben den chemischen Verschiebungen auch die Volumina und die Wahrscheinlichkeiten enthalten sein, so dass ein *Peak* ein NMR-Signal oder ein Rauschsignal ist. Diese Schritte sind in der Abbildung 2.1, die das Konzept von AUREMOL beschreibt, als automatische Vorverarbeitung der experimentellen NMR-Spektren zusammengefasst.
- Zwei Ausgabedateien: In der einen Ausgabedatei werden die berechneten R-Faktoren weggeschrieben, in der anderen die nicht zugeordneten Signale.

Wie bereits erwähnt, werden verschiedene R-Werte mit den Definitionen von R_3 und R_5 berechnet. Mit der Definition von R_5 wird ein globaler R-Wert berechnet, d.h. es werden alle zugeordneten und unzugeordneten Signale verwendet. Für weitere spezifische R-Werte wird die Definition von R_3 verwendet:

- R_{Intra} : R-Wert für die intraresidualen Signale (Signale innerhalb einer Aminosäure)

- R_{Seq} : R-Wert für die sequentiellen Signale (Signale in der Sequenz direkt benachbarter Aminosäuren)
- R_{Medium} : R-Wert für *medium-range* Signale (Signale zwischen Aminosäuren i und j , mit $|i - j| \leq 4$)
- R_{Long} : R-Wert für *long-range* Signale (Signale zwischen Aminosäuren i und j , mit $|i - j| \geq 4$)
- R_{Inter} : R-Wert für die interresidualen Signale (Signale zwischen Aminosäuren i und j , mit $i \neq j$)

3.5.2 R-Wert Berechnung am Beispiel des HPr

Im Folgenden soll ein Beispiel für die Berechnung der R-Faktoren am Beispiel des HPr gezeigt werden. Als experimentelles Spektrum wurde ein 2D-NOESY-Spektrum der H15A Mutante des HPr von *S. aureus* verwendet, gemessen an einem 600 MHz Spektrometer der Firma Bruker. Das *Peak Picking*, das Integrieren der Signalvolumina und die Bayesche Analyse wurde mit AUREMOL automatisch durchgeführt. Signale unterhalb der Diagonalen und rechts der Wasserlinie im Bereich zwischen -1.0 ppm und 5.0 ppm wurden manuell entfernt, s. Kapitel 3.1.3, d.h. es wurde nur der Bereich der Amidprotonen verwendet, da in den Regionen der Seitenketten starke Überlappungen der Signale zu Fehlern und Ungenauigkeiten bei der Volumenberechnung und automatischen Zuordnung mit NOEASSIGN [39] führen können. Dadurch hat sich die Zahl der experimentellen Signale von ca. 8000 auf 1100 reduziert.

Es wurden zwei Teststrukturen verwendet, zum einen ein ausgestreckter, energieminimierter Strang¹⁴ und zum anderen die korrekte NMR-Struktur der HPr Mutante. Mit diesen beiden Strukturen wurden 2D-NOE-Signale rückgerechnet, s. Abschnitt 3.3. Das in der Rückrechnung enthaltene Tool NOEASSIGN [39] führt dabei eine automatische Zuordnung des experimentellen 2D-NOESY-Spektrums anhand der simulierten Signale und der sequentiellen Zuordnung durch. Die berechneten R-Werte sind in Tabelle 3.9 zu sehen. Man beobachtet, dass alle R-Werte der NMR-Struktur deutlich kleiner sind als die des ausgestreckten Strangs. Das bedeutet, dass die NMR-Struktur die experimentellen Daten besser erklärt und die Qualität der NMR-Struktur somit besser ist. Für den ausgestreckten Strang gibt es keine messbaren *long-range* NOEs und damit keinen entsprechenden R-Wert R_{Long} , was den fehlenden Eintrag in der Tabelle erklärt. Die relativ kleinen R-Werte der intraresidualen, sequentiellen, *medium-range*, *long-range* und interresidualen Signale für den ausgestreckten Strang erklären sich dadurch, dass dort die R_3 -Definition

¹⁴Ein ausgestreckter Strang einer Aminosäuresequenz lässt sich mit dem Strukturberechnungsprogramm CNS erzeugen.

	Ext. Strang	NMR-Struktur
R_{Global}	1.00	0.58
R_{Intra}	0.16	0.09
R_{Seq}	0.22	0.10
R_{Medium}	0.29	0.15
R_{Long}	-	0.13
R_{Inter}	0.22	0.12

Tabelle 3.9: Vergleich der R-Werte eines ausgestreckten Strangs und der korrekten NMR-Struktur der H15A Mutante des HPr von *S. aureus*.

verwendet wird und damit die unzugeordneten Signale nicht berücksichtigt werden. Ein Vergleich der Zahl der unzugeordneten Signale der beiden Teststrukturen zeigt, dass sie im Falle des ausgestreckten Strangs mehr als doppelt so hoch sind.

Kapitel 4

Diskussion

Mit dem Programm AUREMOL wurde ein Softwarepaket zur manuellen bzw. automatischen Auswertung von multidimensionalen NMR-Spektren zur Strukturbestimmung von Makromolekülen geschaffen. Wie im Kapitel 3.1 ausgeführt wurde, bietet AUREMOL die Möglichkeit, NMR-Spektren manuell zu bearbeiten und auszuwerten. Dies reicht von der einfachen Visualisierung n-dimensionaler Spektren über die Annotation von Signalen bis zum Einsatz von *Strips*, mit deren Hilfe man die sequentielle Zuordnung eines Makromoleküls manuell ermitteln kann. Methoden, die sowohl bei der manuellen als auch bei der automatischen Auswertung benötigt werden, sind z.B. das automatische *Peak Picking*, das Integrieren von Signalen und die automatische Bestimmung von Rausch- und Artefaktsignalen in einem NMR-Spektrum.

4.1 Konzept von AUREMOL

Das große Ziel der Arbeit war es, den sehr zeitaufwendigen und fehleranfälligen Prozess der Spektrenauswertung vollständig zu automatisieren. Dazu wurde in AUREMOL ein neuer molekülorientierter Ansatz entwickelt, der vorab möglichst viele Informationen bündelt und bei der anschließenden Auswertung benutzt, s. Kapitel 2.1. Dies hat die Entwicklung einer neuen Datenstruktur vorausgesetzt, die die für die Strukturbestimmung relevanten Parameter in geeigneter Form enthält. Die Datenstruktur, die in Kapitel 3.2 und in der Abbildung 2.2 dargelegt wurde, wurde vollständig auf das Konzept von AUREMOL abgestimmt. Der iterative Prozess, der aus der automatischen NOE-Zuordnung und anschließenden Strukturrechnung mit Qualitätsbeurteilung besteht, und die Datenstruktur, bestehend aus den *Compounddateien*, *Masterlisten*, *MasterMasterdatei* und *Metadatei*, sind dabei eng miteinander verzahnt. Die *Compounddateien* definieren die einzelnen Bestandteile der NMR-Probe und enthalten keine proben- oder spektrenspezifischen Para-

meter. Dies hat den Vorteil, dass dort alle Daten abgelegt werden, die unveränderlich für ein Protein gelten, wie z.B. die chemische Struktur oder die Definition dihedraler Winkel. Die Bündelung aller relevanten Daten geschieht dann über die *Metadatei*, die die Informationen aus den einzelnen *Compounddateien* und den spektrenspezifischen Daten aus der *MasterMasterdatei* sammelt. Während des iterativen Auswertungsprozesses werden die *MasterMasterdatei* und damit auch die *Metadatei* aktualisiert, die *Compounddateien* ändern ihren Inhalt nicht. Mit dieser neuen Datenstruktur ist zum einen eine Trennung der spektrenspezifischen und molekülspezifischen Daten möglich, zum anderen ist es möglich, das für die Auswertung relevante Wissen dort zusammenzufassen. Die komplette Beschreibung der NMR-Probe, also der einzelnen Bestandteile (*compounds*) und deren spektrenspezifischen Parameter, vereinfacht die Weiterentwicklung von AUREMOL. Das Problem, dass verschiedene Konformationen eines Proteins in Lösung vorliegen können, was zu einer Zunahme der Signale in den experimentellen Spektren und damit zu Schwierigkeiten bei der Zuordnung führt, ist im Prinzip mit diesem molekülorientierten Ansatz bereits gelöst. Die Simulation von NOE-Signalen muss dazu auf mehrere Bestandteile, also Proteine der NMR-Probe, erweitert werden. Die automatische Zuordnung mit Hilfe des 2-Stufen Algorithmus würde mit den simulierten NOE-Signalen und zwei partiellen Zuordnungslisten¹ wie bisher arbeiten.

Die verbesserte Simulation der NOESY-Spektren, die einen zentralen Punkt in AUREMOL darstellt, hat zu einer realistischeren Berechnung der Linienformen von NOE-Signalen geführt. Wie die Abbildung 3.10 in Kapitel 3.3 zeigt, spielt die Berechnung der Linienbreiten und der Multiplettstruktur aufgrund der J-Kopplung eine wichtige Rolle, um experimentelle NOE-Signale korrekt zu simulieren. Diese verbesserte Simulation der Linienform ist bei der automatischen Zuordnung mit Hilfe des 2-Stufen Algorithmus von entscheidender Bedeutung, wie später noch diskutiert wird.

Der neue molekülbasierte Ansatz von AUREMOL bietet den Vorteil, dass für die Strukturbestimmung kleinerer Moleküle, d.h. etwa 100-120 Aminosäuren, die Messung und Auswertung von 2D- und 3D-NOESY-Spektren ausreichend ist, um die Proteinstruktur zu lösen. Das bedeutet, dass für die Strukturbestimmung nur noch eine kleine Zahl an Spektren gemessen werden muss und nicht, wie bisher, zahlreiche Korrelationsexperimente, deren Auswertung zeitaufwendig und sehr fehleranfällig ist.

Die interne Datenbank von AUREMOL, die *a priori* Wissen über Proteine oder NMR-Experimente enthält, kann beliebig ergänzt werden. Dadurch ist AUREMOL nicht auf die Auswertung von NMR-Spektren von Proteinen beschränkt, sondern kann auf DNA-, RNA- oder andere Biomoleküle erweitert werden.

Mit dem molekülbasierten und wissensbasierten Ansatz von AUREMOL ist es möglich, Strukturinformationen bereits vor der Auswertung der NMR-Spektren zu nutzen. Dies

¹Für jeden Bestandteil der Probe benötigt man für die automatische Zuordnung eine partielle Zuordnungsliste.

stellt einen großen Vorteil bei der Strukturbestimmung von homologen Proteinen dar, z.B. bei Mutationsstudien. Mit dem integrierten *Homologie Modelling* ist es möglich, die Strukturen von punktmutierten Proteinen relativ genau vorherzusagen und mit dieser vermuteten Struktur, die automatische Auswertung der NMR-Spektren mit Hilfe des 2-Stufen Algorithmus oder KNOWNOE zu unterstützen.

Im Folgenden sollen die Ergebnisse der automatischen Strukturbestimmung mit KNOWNOE aus Abschnitt 3.4.1 und der automatischen Bestimmung fehlender chemischer Verschiebungen in NOESY-Spektren mit dem 2-Stufen Algorithmus aus Kapitel 3.4.2 diskutiert werden.

4.2 Automatische NOE-Zuordnung mit KNOWNOE

Die Ergebnisse im Kapitel 3.4.1 zeigen, dass für das Protein *TmCsp* die richtige Struktur ermittelt worden ist. In Abbildung 3.12, die einen Vergleich der automatisch bestimmten Struktur und der manuell bestimmten zeigt, ist deutlich zu sehen, dass die Sekundärstrukturelemente in beiden Fällen identisch sind. Das 5-strängige β -Fass zeigt jeweils die gleiche Topologie. Unterschiede sind hauptsächlich in den *loop*-Bereichen zu finden.

Der RMSD-Wert für die abgebildeten Strukturen beträgt im Fall der automatisch bestimmten Struktur 0.034 nm und im Fall der manuell bestimmten Struktur 0.104 nm. Dies zeigt, dass KNOWNOE eine ausreichende Zahl an NOEs gefunden hat, die zu einer präzisen Strukturbestimmung geführt hat. Wie bereits in Kapitel 2.4 erwähnt, führt die Berechnung von R-Werten zu einer besseren Beurteilung der Güte von Strukturen. Für die beiden Strukturen und die bei der PDB Datenbank eingereichten Struktur von *TmCsp* [61] wurden R-Werte berechnet. Es wurde der globale R-Faktor R_5 von Gleichung 2.60 berechnet, wobei die nicht zugeordneten Signale mit berücksichtigt wurden. Die berechneten R-Werte, die in Tabelle 4.1 zu sehen sind, sind in allen Fällen ähnlich. Dies bedeutet, dass die experimentellen Daten durch die Strukturen gut erklärt werden.

Die automatisch bestimmte Struktur von *TmCsp* ist von ähnlich guter Qualität wie die

	KNOWNOE	Manuell	PDB Datenbank
R-Faktor	0.35	0.35	0.36

Tabelle 4.1: Vergleich der R-Werte automatisch und manuell bestimmten Struktur von *TmCsp*. In allen Fällen wurde ein ähnlicher R-Wert berechnet, was bedeutet, dass die Strukturen die experimentellen Daten gut erklären.

manuell bestimmten Strukturen. Allerdings wurden diese Strukturen mit einem wesentlich geringeren Zeitaufwand und einer kleineren Zahl an Spektren, in diesem Fall einem einzigen experimentellen 2D-NOESY-Spektrum, gelöst als die manuellen.

Die Strukturen wurden ohne den Einsatz einer Homologiestruktur gelöst, d.h. die Strukturbestimmung wurde mit einem ausgestreckten Strang als Startstruktur begonnen. Voraussetzung für die erfolgreiche automatische Strukturbestimmung ist, dass die sequentielle Zuordnungsliste möglichst komplett ist, also nicht nur die Hauptketten-, sondern auch die Seitenkettensignale zugeordnet sind. Zusätzlich sollte diese Liste bereits gut an die NOESY-Spektren angepasst sein, um die Toleranzwerte für die chemischen Verschiebungen klein zu halten. Ansonsten führt dies in vielen Fällen zu einer großen Zahl an Zuordnungsmöglichkeiten (mehr als drei) pro Signal, was dazu führt, dass KNOWNOE diese Signale nicht eindeutig zuordnen kann. Dies führt zu einer geringen Zahl an eindeutigen NOE-Zuordnungen und kann zu Beginn des iterativen Prozesses zu einer falschen Faltung des Proteins beitragen.

Zusammenfassend kann man sagen, dass KNOWNOE eine leistungsfähige Methode darstellt, unter Verwendung experimenteller NOESY-Spektren und der chemischen Zuordnungsliste die Struktur eines Proteins zu lösen.

4.3 Automatische Zuordnung mit 2-Stufen Algorithmus

Der 2-Stufen Algorithmus, der eng an das molekülorientierte Konzept von AUREMOL angelehnt ist, versucht, durch Vergleich von simulierten und experimentellen 2D- und 3D-NOESY-Spektren fehlende chemische Verschiebungen zu finden. Da eine partielle sequentielle Zuordnung der Resonanzen benötigt wird, werden die bekannten Zuordnungen in der 1. Stufe an die verwendeten experimentellen NOESY-Spektren angepasst. In der 2. Stufe werden dann die fehlenden Zuordnungen bzw. Verschiebungen bestimmt.

4.3.1 Zuordnung der künstlich erzeugten 2D- und 3D-NOESY-Spektren

4.3.1.1 Zuordnung im 2D-NOESY-Spektrum

Die ersten Simulationen wurden mit der 2. Stufe der Optimierung und einem künstlich erzeugten 2D-NOESY-Spektrum durchgeführt. Es wurde getestet, inwieweit der Algorithmus in der Lage ist, Resonanzfrequenzen richtig zuzuordnen, falls folgende Parameter variiert werden:

1. Partielle sequentielle Zuordnung:

Es hat sich gezeigt, dass mit einer partiellen Zuordnung von mindestens 30%, mehr als 90% aller Resonanzfrequenzen bzw. chemischen Verschiebungen richtig zugeordnet werden, s. Abbildung 3.15 in Kapitel 3.4.2.2. Falls die partielle sequentielle Zuordnung mindestens 50% beträgt, lassen sich sogar mehr als 97% aller chemischen Verschiebungen korrekt zuordnen. Durch eine Vergrößerung der maximalen Zahl an Iterationen und daher einhergehend mit einer langsameren Abkühlung des Systems, d.h. Verringerung des *Thresholds* Th , könnte man vermutlich mit einer noch kleineren partiellen Zuordnung als 30% mehr als 90% aller Resonanzfrequenzen korrekt zuordnen. Dies würde allerdings zu Lasten der Rechenzeit gehen, die bei diesen Simulationen auf den SGI Workstations bereits 6-12 Stunden betragen hat.

2. Stärke des Rauschens im künstlichen Datensatz:

Je stärker das künstliche NOESY-Spektrum verrauscht wird, desto weniger Resonanzfrequenzen werden richtig zugeordnet. Bei einer partiellen Zuordnung von 80% und einer Rauschrate $n_r = 50\%$ werden mehr als 97% aller Resonanzfrequenzen richtig zugeordnet, s. Abbildung 3.18 in Kapitel 3.4.2.2. Erhöht man die Rauschrate n_r , so sinkt die Zahl an richtig zugeordneten Resonanzen. Eine gewisse Robustheit des Algorithmus gegenüber verrauschten Spektren konnte damit gezeigt werden.

3. Zahl der hinzugefügten Artefaktsignale im 2D-NOESY-Spektrum:

Wie die Simulationen zeigen, gibt es keinen signifikanten Einfluss der Anzahl der Artefaktsignale auf die Zahl der korrekt zugeordneten chemischen Verschiebungen. Bei einer Anzahl von 1728 Signalen im künstlichen Datensatz, einer partiellen Zuordnung von 80% und zusätzlichen 1800 Artefaktsignalen werden mehr als 98% der Resonanzfrequenzen richtig zugeordnet, s. Abbildung 3.19 in Kapitel 3.4.2.2. Auch hier zeigt sich eine sehr gute Robustheit des Algorithmus.

4. Zahl der entfernten Kreuzsignale im 2D-NOESY-Spektrum:

Wie in Abbildung 3.19 in Kapitel 3.4.2.2 zu sehen ist, kann der Algorithmus eine gewisse Zahl an fehlenden NOE-Signalen kompensieren. Bis zu einer Zahl von 50 entfernten Kreuzsignalen² zeigt der Algorithmus keine schlechteren Ergebnisse, es werden über 99% der Resonanzfrequenzen richtig zugeordnet. Erst ab 100 fehlenden Signalen sinkt die Zahl der korrekt zugeordneten Verschiebungen signifikant. Zusammenfassend kann man sagen, dass der Algorithmus in der Lage ist, fehlende Kreuzsignale in den NOESY-Spektren zu kompensieren.

Ebenso konnte gezeigt werden, dass bei gleichzeitigem Zusammenwirken der drei Störeffekte (Rauschen, Artefaktsignale und fehlende NOEs) das Verfahren gute Optimierungser-

²Insgesamt besteht das Spektrum aus 1728 NOE-Signalen, davon sind 1268 Kreuzsignale und 460 Diagonalsignale.

ergebnisse liefert. Das heißt, der 2-Stufen Algorithmus arbeitet nicht nur mit idealen Datensätzen, sondern auch unter realen Bedingungen. Allerdings bleibt zu erwähnen, dass in realen Datensätzen die Zahl der Signale nicht bei etwa 1700 liegt, wie bei den bisherigen Simulationen, sondern je nach Proteingröße bei bis zu 10000-15000 Signalen. Das bedeutet, dass die Dauer der Optimierungsläufe stark mit der Zahl der Signale anwachsen wird, wenn man, wie bisher, mehr als 97 oder 98% richtig zugeordnete Resonanzfrequenzen als Endergebnis haben möchte.

Das Kriterium, welche Resonanzfrequenzen richtig zugeordnet worden sind und welche nicht, liefern die berechneten Matchfaktoren M_p , s. Gleichung 2.48 in Kapitel 2.3.2. Im Falle falscher Zuordnungen sind den entsprechenden simulierten NOE-Signalen kleine Matchfaktoren zugeordnet, d.h. Werte signifikant kleiner als der bei optimaler Übereinstimmung ergebende Matchfaktor von 1. Die 2. Stufe der Optimierung liefert als Ergebnis eine Liste aller simulierten NOE-Signale und deren Matchfaktoren mit den entsprechenden Intensitätsverteilungen im experimentellen Spektrum. Anhand dieser Matchfaktoren kann entschieden werden, ob Resonanzfrequenzen richtig oder falsch zugeordnet worden sind. Zusammenfassend kann man sagen, dass nur eine geringe Zahl an Zuordnungen bekannt sein muss ($\sim 50\%$), um mehr als 97% aller Resonanzfrequenzen korrekt zuzuordnen. Die automatische Auswertung von NOESY-Spektren homologer Proteine ist daher mit diesem Verfahren möglich.

4.3.1.2 Zuordnung im 3D-NOESY-Spektrum

In einem nächsten Schritt wurde die 2. Stufe des Optimierungsalgorithmus mit einem künstlich erzeugten 3D-NOESY-Spektrum untersucht. Als Vorverarbeitung wurde eine Segmentierung und damit eine Reduzierung des Spektrums durchgeführt, was in dem erwähnten Beispiel eines $256 \times 64 \times 512$ großen 3D-Spektrums zu einer erheblichen Kompression geführt hat. Der Komprimierungsfaktor betrug 1:1000. Generell ist es möglich, die Kompression auch auf 2D-Spektren anzuwenden, allerdings wird der Kompressionsfaktor nicht so hoch sein wie bei 3D-Spektren. Bisher wird ein 1:1 Abbild des 2D-Spektrums für die weitere Verwendung des 2-Stufen Algorithmus in der Datei *peakshapes.dat* abgelegt, s. Kapitel 3.4.2.1. Es stellt aber kein Problem dar, das bei den 3D-Spektren verwendete Verfahren auch auf 2D-Spektren zu erweitern. Diese neu entwickelte Form der Komprimierung ermöglicht es, die weitere Auswertung der Spektren mobiler zu gestalten, da bei diesen hohen Komprimierungsfaktoren 2D- und 3D-Spektren auf einer handelsüblichen 1.44 MB Diskette Platz hätten und somit eine Auswertung am Laptop zu Hause oder auf einer Konferenz möglich wäre. Ein weiterer Vorteil stellt die Verringerung der Netzlast des bei der Auswertung von NMR-Spektren verwendeten Computer-Clusters dar, was die manuelle Arbeit an den Spektren effizienter macht. Die internen Lese- und Schreibroutinen von NMR-Spektren werden deshalb in naher Zukunft um die Fähigkeit erweitert,

diese komprimierten Spektren lesen und visualisieren zu können.

Die Simulationen wurden mit unterschiedlichen partiellen Zuordnungen gestartet:

- Die Resonanzfrequenzen der ^{15}N -Kerne oder der Amidprotonen wurden aus der partiellen Zuordnungsliste entfernt. Der Algorithmus war in der Lage, alle fehlenden Resonanzfrequenzen richtig zuzuordnen, s. Tabelle 3.3 in Kapitel 3.4.2.2.
- Das Fehlen der chemischen Verschiebungen aller 242 ^1H -Kerne in der Zuordnungsliste wurde untersucht. 80% aller Resonanzfrequenzen, d.h. 189 von 242 Verschiebungen, wurden korrekt zugeordnet, s. Tabelle 3.3 in Kapitel 3.4.2.2.
- Simulationen mit einer partiellen Zuordnung von 0% ergaben, dass immerhin 50% aller Resonanzfrequenzen richtig zugeordnet wurden, s. Tabelle 3.3 in Abschnitt 3.4.2.2. Im 2D-Fall konnten bei einer partiellen Zuordnung von 0% nur etwa 10% der Resonanzfrequenzen korrekt zugeordnet werden, s. Abbildung 3.15 in Kapitel 3.4.2.2. Wie man sieht, ist auch hier eine partielle Zuordnungsliste nötig, um eine genügend hohe Zahl (mindestens 90%) an Resonanzfrequenzen richtig zuzuordnen.

Die Ergebnisse zeigen, dass der 2-Stufen Algorithmus auch in 3D-NOESY-Spektren zuverlässig Resonanzfrequenzen zuordnen kann. Selbst eine hohe Entartung von chemischen Verschiebungen, wie bei den ^{15}N -Kernen³, hat zu sehr guten Ergebnissen des Algorithmus geführt.

Das Kriterium, welche Zuordnungen falsch oder richtig sind, wird wie im vorherigen Kapitel durch die berechneten Matchfaktoren M_p geliefert. Der für die Praxis wichtige Fall, dass Zuordnungen im 2D- und 3D-Spektrum unbekannt sind, wurde ebenfalls untersucht. Es wurden die Resonanzfrequenzen der 21 Amidprotonen der Aminosäuren 30 bis 50 aus der partiellen Zuordnungsliste entfernt. Die gleichzeitige Optimierung des 2D- und 3D-NOESY-Spektrums ergab, dass bis auf die Zuordnung von HN 49 allen anderen Amidprotonen die korrekten Resonanzfrequenzen zugeordnet wurden. Der Fall, dass viele Zuordnungen gleich bleiben und nur einige wenige sich ändern, tritt auf, wenn man Mutationsstudien macht. Die Resonanzfrequenzen und die Struktur eines Proteins sind bekannt und man möchte wissen, wie sich die Zuordnung des Proteins bei Punktmutationen⁴ ändert. Erfahrungsgemäß werden sich nur wenige Zuordnungen ändern, die meisten, die keiner strukturellen Änderung unterliegen, werden gleich bleiben. Das bedeutet, dass der 2-Stufen Algorithmus fehlende Zuordnungen von punktmutierten Proteinen prinzipiell ermitteln kann, wenn die Zahl der geänderten Zuordnungen klein ist.

³Die 92 Verschiebungen der ^{15}N -Kerne bestehen aus 45 verschiedenen Resonanzfrequenzen, d.h. im Mittel haben zwei Kerne dieselbe Frequenz.

⁴Eine Punktmutation ist der Austausch einer Aminosäure in der Primärsequenz des Proteins.

4.3.2 Zuordnung des experimentellen 2D-NOESY-Spektrums

Die Ergebnisse mit dem experimentellen 2D-NOESY-Spektrum der HPr Mutante H15A sollen nun diskutiert werden. Es wurde im Ergebnisteil 3.4.2 erwähnt, dass wegen der schlechten Artefakterkennung der Bayesschen Analyse das experimentelle Spektrum verbreitert prozessiert wurde. Dadurch hat man die für die Zuordnung wichtige Information der Multiplettstruktur verloren, obwohl auch dann noch gute Ergebnisse bei der Zuordnung erzielt werden, wie im Folgenden noch diskutiert wird. Prinzipiell bestehen zwei Möglichkeiten, dieses Problem zu lösen.

1. Man modifiziert die Bayessche Analyse dahingehend, dass der Algorithmus Multipletts erkennt und diese als ein Signal zusammenfasst. In NOESY-Spektren, in denen man im Bereich der Wasserlinie und der Seitenketten einen extrem hohen Signalüberlapp hat, ist eine zuverlässig funktionierende Multipletterkennung aber nur schlecht zu realisieren. Zusätzlich sind fehlende *Subpeaks* in den Multipletts, sowie das verschlechterte Signal-Rausch-Verhältnis durch die Prozessierung zu berücksichtigen. All dies wird dazu führen, dass die Bayessche Analyse wiederum nur eine mangelhafte Trennung von Rauschsignalen und echten NOE-Signalen liefern wird.
2. Die andere Alternative stellt eine zweistufige Analyse dar. Das experimentelle Spektrum wird auf zwei Arten prozessiert, einmal verbreitert ohne Multiplettaufspaltung und einmal mit Multiplettaufspaltung. Im verbreiterten Spektrum wird die automatische Artefakterkennung wie bisher sehr gute Resultate erzielen. Die Ergebnisse kann man nun auf das Spektrum übertragen, das die Multiplettstruktur aufweist. Als Kriterium, welche Signale in den beiden Spektren zusammengehören, könnte man die Fläche der *Peaks* im verbreiterten Spektrum heranziehen, d.h. die *Subpeaks* im aufgespaltenen Spektrum, die innerhalb der Fläche des entsprechenden *Peaks* im verbreiterten Spektrum liegen, sind identisch.

Die 2. Möglichkeit scheint die erfolgversprechendere zu sein und soll demnächst in AUREMOL implementiert werden, damit die wichtige Information der Multiplettstruktur beim Vergleich der simulierten und experimentellen Linienformen auch berücksichtigt werden kann.

Zu Beginn wurde untersucht, wieviele der insgesamt 524 Resonanzfrequenzen vom Algorithmus prinzipiell zugeordnet werden konnten. Dazu wurde jeweils eine Zuordnung, z.B. HB2 60, aus der partiellen Zuordnungsliste entfernt⁵ und getestet, ob sie korrekt wiedergefunden wurde. Bis auf die HA Protonen im Bereich der Wasserlinie sind die Ergebnisse,

⁵Das bedeutet zum Beispiel, dass die korrekte Resonanzfrequenz 2.011 ppm von HB2 60 aus der Liste entfernt wird.

unterteilt in vier Gruppen, in Tabelle 3.4 im Kapitel 3.4.2.3 dargestellt. Es fällt auf, dass für einige Kerne, speziell die Seitenkettenprotonen, falsche Resonanzfrequenzen zugeordnet werden. Das Problem liegt darin begründet, dass in experimentellen Spektren bis zu 20% der Signale fehlen, was z.B. auf Austauscheffekte zurückzuführen ist. Ein anderer Grund liegt darin, dass es oft falsche Zuordnungen gibt, die einen höheren Matchfaktor liefern als die richtige Lösung. Letzteres soll in Abbildung 4.1 veranschaulicht werden.

Die Ursache für die Zuordnung einer falschen Resonanzfrequenz ist dort für die Spins HN 85 und HG2 24 grafisch dargestellt. Im linken Teilbild ist der Fall für HN 85 abgebildet, im rechten Teilbild für HG2 24. Für jedes Teilbild sind drei Streifen abgebildet, die mit A, B und C durchnummeriert sind. Der Streifen A zeigt jeweils die berechneten NOE-Kontakte von HN 85 bzw. HG2 24 zu anderen wechselwirkenden Spins. Streifen B und C zeigen die entsprechenden Ausschnitte aus dem experimentellen Spektrum an der Stelle $w_1 = 7.843$ ppm und $w_1 = 8.122$ ppm im Falle von HN 85 bzw. $w_1 = 2.396$ ppm und $w_1 = 1.894$ ppm im Falle von HG2 24. Der Streifen B steht für die richtige Zuordnung, der Streifen C für die vom Algorithmus gelieferte falsche Zuordnung. Deutlich zu sehen ist, dass die Kontakte HB3 84/HN 85 und HN 87/HN 85 in Streifen B, hervorgehoben durch Ellipsen in Abbildung 4.1, fehlen, wohingegen sie im Streifen C vorhanden sind. Auch die berechneten Matchfaktoren zeigen, dass es für den Algorithmus mehr Sinn macht, HN 85 die falsche Resonanzfrequenz 8.122 ppm anstatt 7.843 ppm zuzuordnen. Ähnlich lässt sich auch bei HG2 24 argumentieren, wo z.B. das Signal HE21 24/HG2 24 nur in Streifen C zu sehen ist.

Das Problem lässt sich möglicherweise durch eine Bestrafung von Mehrfachbelegungen von Resonanzfrequenzen lösen, solange noch nicht alle chemischen Verschiebungen bzw. Resonanzen bei der Zuordnung verwendet werden. Allerdings darf man nicht generell jede Entartung bestrafen, da Mehrfachbelegungen von Verschiebungen in der Realität durchaus vorkommen können. In den beiden genannten Fällen hätte eine Bestrafung keinen Einfluss gehabt, da die dem Spin HN 85 falsch zugeordnete Resonanzfrequenz von 8.122 ppm noch nicht belegt war und somit keine Bestrafung erfolgt wäre. Die falsch zugeordnete Resonanzfrequenz von 1.894 ppm bei HG2 24 war zwar bereits einmal belegt, allerdings auch die richtige Verschiebung von 2.396 ppm.

Für die weiteren Simulationen wurden die Resonanzfrequenzen der Spins, die bei den ersten Tests zu falschen Zuordnungen geführt haben, als bekannt gesetzt, d.h. nicht aus der partiellen Zuordnungsliste entfernt.

1. Von den insgesamt 58 entfernten Resonanzfrequenzen der Amidprotonen aus der Zuordnungsliste konnten etwa 45 vom Algorithmus korrekt ermittelt werden, was einer Quote von 78% entspricht. Die partielle Zuordnung hat damit 89%⁶ betragen, nach der Optimierung waren etwa 98% aller Resonanzfrequenzen korrekt zugeordnet.

⁶Ergibt sich daraus, dass 58 der insgesamt 524 Zuordnungen aus der Liste entfernt wurden.

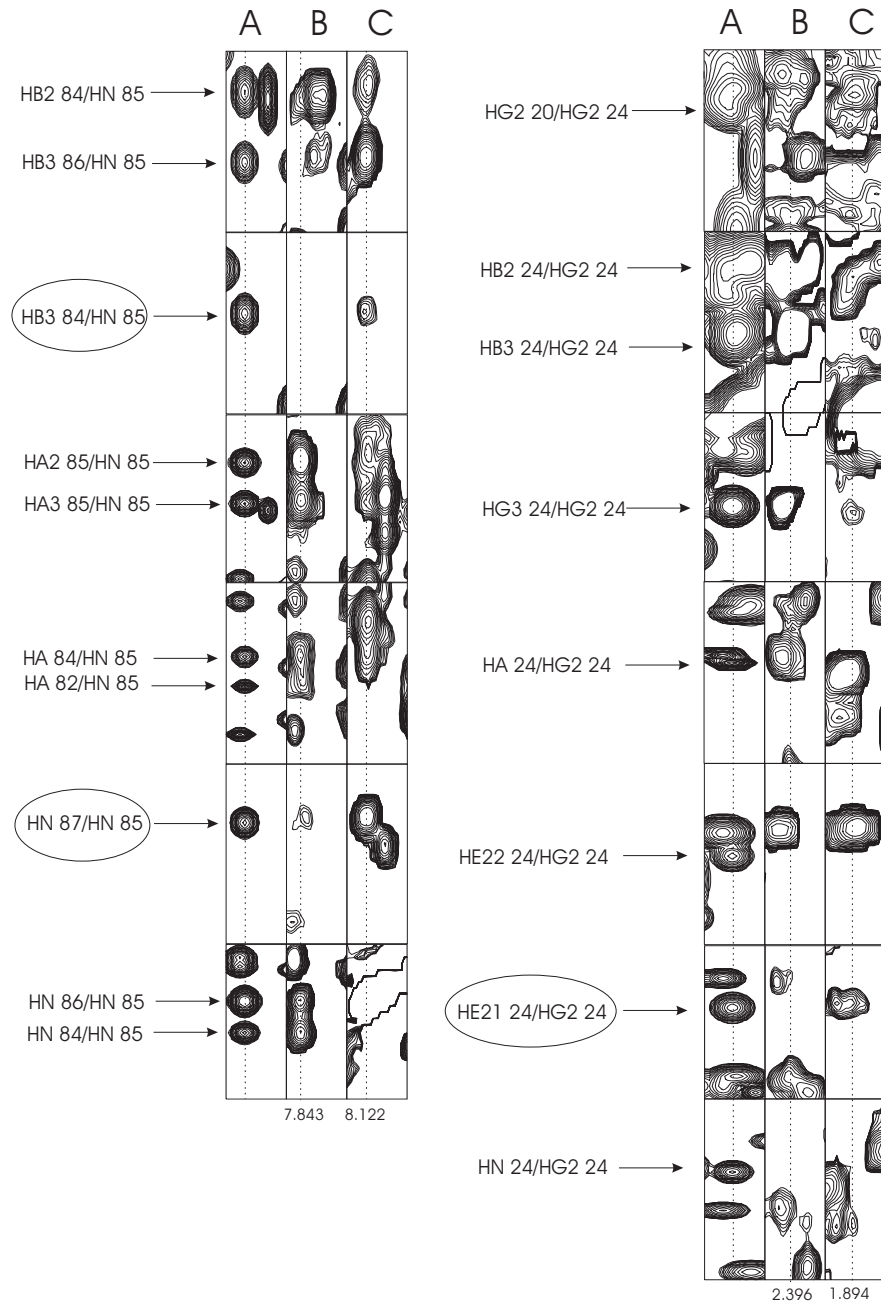


Abbildung 4.1: Fehlerhafte Zuordnung von HN 85 (links) und HG2 24 (rechts). Links: Der Streifen A zeigt das simulierte Spinsystem HN 85, Streifen B das experimentelle Spektrum an der Position $w_1 = 7.843\text{ppm}$ und Streifen C das segmentierte Spektrum an der falsch zugeordneten Position $w_1 = 8.122\text{ppm}$. Streifen B ist laut Zuordnungsliste die korrekte Position des Spinsystems. Rechts: Analog für das Spinsystem HG2 24 an den Positionen $w_1 = 2.396\text{ppm}$ (richtig) und $w_1 = 1.894\text{ppm}$ (falsch).

2. Die Resonanzfrequenzen der Seitenkettenprotonen im Bereich zwischen 4.5 und 1.0 ppm wurden aus der partiellen Zuordnungsliste entfernt. Dies entsprach 103 nicht zugeordneten Resonanzfrequenzen bzw. einer partiellen Zuordnung von 80%. Der Algorithmus war in der Lage, 75 der 103 Resonanzfrequenzen wieder richtig zuzuordnen. Damit wurden 95% aller Resonanzlinien korrekt zugeordnet.

Man sieht, dass der 2-Stufen Algorithmus in der Lage ist, auch in experimentellen Datensätzen fehlende Zuordnungen zuverlässig zu finden unter der Voraussetzung, dass diese Zuordnungen anhand des Spektrums möglich sind. Auf die Problematik wurde bereits in Abbildung 4.1 hingewiesen.

Das Problem, dass einigen Spins falsche Resonanzfrequenzen zugeordnet werden, wie es in Abbildung 4.1 dargestellt wurde, soll nun im Folgenden diskutiert werden. Wie bereits erwähnt, liegen die Ursachen in fehlenden Kreuzsignalen der NOESY-Spektren und in Zuordnungen, die einen höheren Matchfaktor liefern als die richtige Zuordnung. Die Bestrafung der Entartung von Resonanzfrequenzen wurde als möglicher Lösungsansatz bereits erwähnt. Eine weitere Möglichkeit, die Zahl an falschen Zuordnungen zu verringern, könnte darin bestehen, die Volumina und damit die Absolutintensitäten der NOE-Signale mit zu berücksichtigen. Dadurch wird zwar die automatische Zuordnung strukturaabhängiger, es lassen sich aber vermutlich mehr Resonanzfrequenzen korrekt zuordnen. Ein vielversprechender Ansatz wäre, nicht wie bisher, Matchfaktoren für die einzelnen NOE-Signale zu bilden und aufzuaddieren, sondern Matchfaktoren für eine ganz Reihe bzw. Spalte an Signalen zu berechnen und aufzuaddieren. Die Energiefunktion für die Optimierung eines Spektrums würde sich dann in folgender Weise schreiben lassen

$$E = M - \sum_{m=1}^M \cos \left(\vec{T}_m^{sim}, \vec{T}_m^{exp} \right) \quad (4.1)$$

wobei $\vec{T}_m^{sim} = (\vec{p}_1^{sim}, \dots, \vec{p}_k^{sim})$ ist, M die Gesamtzahl der Spins ist und die verwendeten *Peakformen* \vec{p}_q^{sim} die Bedingung

$$\sum_{j=1}^{J_i} \delta(Q(p, j) - m) \neq 0 \quad (4.2)$$

erfüllen müssen. J_i ist die Dimension des Spektrums und die Funktion Q ist im Theorieteil in der Gleichung 2.52 erklärt. Gleichung 4.2 liefert 1 zurück, falls der simulierte *Peak* p einen NOE-Kontakt zwischen solchen Spins darstellt, von denen mindestens einer das Atom A_m ist. *Diagonalpeaks* werden im 2D-NOESY-Spektrum doppelt gezählt, d.h. Gleichung 4.2 liefert 2 zurück.

Mit diesem Matchfaktor hätte man auch wieder ein sicheres Kriterium, welche Resonanzfrequenzen richtig bzw. falsch zugeordnet worden sind, was in einigen Fällen bei den

bisherigen Simulationen mit dem experimentellen 2D-NOESY-Spektrum nicht eindeutig war. Als Beispiel sei nur auf die Abbildung 4.1 und die entsprechenden Ausführungen im Text verwiesen. Der neue Matchfaktor, der komplette Spinsysteme unter Berücksichtigung der Absolutintensitäten untersucht, wird die Zahl der falschen Zuordnungen dadurch verringern, dass es weniger Fälle geben wird, in denen falsche Zuordnungen einen besseren Matchfaktor liefern als die richtige Zuordnung.

Dieser neu eingeführte Matchfaktor ist wesentlich strukturabhängiger und sollte nur dann verwendet werden, wenn die Proteinstruktur bereits gut bekannt ist. Falls man keinerlei Strukturinformationen hat, also mit einer *random coil* Struktur startet, müsste man zu Beginn der automatischen Zuordnung den strukturunabhängigen Matchfaktor aus Gleichung 2.50 verwenden. Im weiteren Verlauf des iterativen Prozesses, wenn man durch die Auswertung der NOESY-Spektren Strukturinformationen gewinnt, kann man den neuen Matchfaktor aus Gleichung 4.1 einsetzen.

Bei den Simulationen der experimentellen Datensätzen wurde der in Gleichung 2.52 eingeführte Bestrafungsterm p_v verwendet. Dadurch werden Zuordnungen bestraft, die weit von der mittleren chemischen Verschiebung E_m der internen AUREMOL Datenbank entfernt sind, d.h. mehr als die doppelte Standardabweichung σ_m . Es hat sich gezeigt, dass bei der Verwendung von $p_v = 0$ die Zahl der falschen Zuordnungen gestiegen wäre. Beispielsweise wurden einige Amidprotonen nicht in der Region zwischen 11.0 und 6.0 ppm sondern im Bereich der Seitenkettensignale, d.h. zwischen 3.0 und 0.0 ppm zugeordnet, was nur in sehr seltenen Fällen vorkommt. Durch den Term p_v werden diese Möglichkeiten der Zuordnung nicht verboten, aber etwas unwahrscheinlicher stattfinden.

Zusammenfassend lässt sich sagen, dass der 2-Stufen Algorithmus ein leistungsfähiges Verfahren darstellt, fehlende chemische Verschiebungen in den experimentellen NOESY-Spektren zu finden. Als Input benötigt man Informationen über die Proteinstruktur und eine partielle Zuordnungsliste. Wie stark die Proteinstruktur die Ergebnisse des Algorithmus beeinflusst, muss noch untersucht werden. Es lässt sich aber vermuten, dass das Verfahren nicht allzu strukturabhängig ist, da z.B. die Multipllettstruktur der NOE-Signale nicht oder nur gering strukturabhängig ist. Da in den NOESY-Spektren die Zahl der *long-range* NOEs, die die Strukturinformation tragen, gering ist und die Zahl der sequentiellen NOEs, die strukturunabhängig sind, groß ist, kann man vermuten, dass der 2-Stufen Algorithmus durch die gezeigte Robustheit gegen fehlende NOE-Signale auch in dieser Hinsicht relativ strukturunabhängig arbeitet. Die Tatsache, dass die Simulationen mit einem verbreiteten experimentellen NOESY-Spektrum, in dem die Multipllettstruktur aufgrund der erwähnten Probleme entfernt war, durchgeführt wurden, zeigt, dass der Algorithmus relativ stabil gegenüber veränderten *Peakformen* ist. Dies ist ein weiteres Indiz, dass die verwendete Strukturinformation beim 2-Stufen Algorithmus keine sehr grosse Rolle spielt. Quantitative Aussagen fehlen aber und müssen noch durchgeführt werden.

4.3.3 Vergleich mit anderen Verfahren zur automatischen Zuordnung

Die meisten auf dem Gebiet der automatischen Auswertung von NMR-Spektren entwickelten Programme und Methoden setzen eine bottom-up Strategie ein. Dabei wird versucht, anhand verschiedener Korrelationsexperimente und der gegebenen Primärsequenz des Proteins, Resonanzfrequenzen zuzuordnen und eine sequentielle Zuordnung des Proteins zu erhalten. Programme wie PASTA [65] oder AUTOASSIGN [108] wurden bereits erfolgreich mit experimentellen Datensätzen getestet. Ein großer Nachteil dieser Methoden ist die mangelhafte Robustheit gegen Rauschen und fehlende Kreuzsignale in den Spektren. Ausserdem liefern sie mit der sequentiellen Zuordnung noch keinerlei Information über die Tertiärstruktur des Proteins. Das Problem der Zuordnung der NOESY-Spektren ist damit noch nicht gelöst.

Die in AUREMOL eingesetzte top-down Strategie ist in ähnlicher Form bereits in dem Programm ANSRS [60] vorhanden. Kraulis hat gezeigt, dass es prinzipiell möglich ist, ohne Kenntnis der sequentiellen Zuordnung und ohne Strukturinformation, das Zuordnungsproblem in experimentellen NOESY-Spektren zu lösen. Allerdings wurde das Verfahren nur anhand simulierter Datensätze getestet. Vorhandenes Wissen über das Protein, wie z.B. eine vermutete Struktur, wurde bei dieser Methode nicht verwendet.

In AUREMOL wird auf der einen Seite die top-down Strategie und ein molekülorientierter Ansatz eingesetzt, d.h. es wird vorhandenes Wissen über das Protein bei der Auswertung der NOESY-Spektren benutzt. Auf der anderen Seite wird als wichtige Zusatzinformation die Linienform der NOE-Signale als Kriterium bei der automatischen Zuordnung verwendet. Dies stellt eine wesentliche Verbesserung gegenüber den konventionellen Verfahren wie ANSRS [60] dar, die nur auf Basis der Peaklisten arbeiten, da so z.B. Signalüberlagerungen in den NOESY-Spektren prinzipiell kein Problem mehr bei der Zuordnung der Resonanzfrequenzen darstellen. Im Gegensatz zum Programm von Kraulis wurde AUREMOL auch mit einem experimentellen 2D-NOESY-Spektrum der HPr Mutante H15A erfolgreich getestet. Fehlende Resonanzfrequenzen sind richtig zugeordnet worden. Ein Kriterium zu finden, das ein quantitatives Maß für die Richtigkeit einer Zuordnung liefert, ist ein Problem, dass bei keinem der erwähnten Verfahren zur automatischen Zuordnung gelöst worden ist. Ähnlich wie bei den R-Faktoren, die ein unabhängiges Maß für die Güte einer NMR-Struktur darstellen, muss bei der automatischen Zuordnung noch ein Kriterium entwickelt werden, das wie bei dem Algorithmus KNOWNOE⁷ ein Maß für die Richtigkeit einer Zuordnung liefert.

Zusammenfassend kann man sagen, dass mit AUREMOL eine effiziente Auswertung von NOESY-Spektrum mit Hilfe des 2-Stufen Algorithmus und KNOWNOE möglich ist. Beim 2-Stufen Algorithmus werden nur experimentelle 2D- und 3D-NOESY-Spektren benötigt,

⁷KNOWNOE berechnet für jede Zuordnungsmöglichkeit eine Wahrscheinlichkeit.

d.h. es müssen keine langwierigen und teuren Korrelationsexperimente gemessen und ausgewertet werden. AUREMOL automatisiert und beschleunigt die Zuordnung von homologen Proteinen und stellt eine wichtige Entwicklung im Bereich des Proteom-Projekts dar, welches eine möglichst effiziente und automatisierte Strukturbestimmung von Proteinen als Ziel hat.

Abbildungsverzeichnis

2.1	Konzept von AUREMOL	11
2.2	Interne Datenstruktur von AUREMOL	13
2.3	Strategie von KNOWNOE	23
2.4	Algorithmus von KNOWNOE	25
2.5	Zwei Beispiele für Distanz- bzw. Volumenwahrscheinlichkeitsverteilungen .	28
2.6	2-Stufen Algorithmus	29
2.7	Kernalgorithmus der 2-Stufen Optimierung	32
2.8	Automatische Vorverarbeitung beim 2-Stufen Algorithmus	34
2.9	Zugriff auf Pixel im 3D-Spektrum	36
2.10	Variation einer Zuordnung in der 1. Stufe	37
2.11	Variation einer Zuordnung in der 2. Stufe	38
2.12	Algorithmus von RFAC	43
3.1	<i>Screenshot</i> von AUREMOL	49
3.2	Anwendung der Strips	53
3.3	Trennung von Signal- und Artefaktpeaks	57
3.4	Beispiel für eine <i>Compounddatei</i>	60
3.5	Beispiel für eine <i>Masterliste</i>	61
3.6	Beispiel für eine <i>MasterMasterdatei</i>	62
3.7	Beispiel für eine <i>Metadatei</i>	63
3.8	Dialog für Simulation von NOESY-Spektren	66
3.9	Beispiel für Dateien, die simulierte <i>Peakformen</i> enthalten	68
3.10	Vergleich sim. und exp. Linienformen	70

3.11	Dialog von KnowNOE	72
3.12	Vergleich der durch manuelle bzw. automatische NOE-Zuordnung erhaltenen Strukturen von <i>TmCsp</i>	74
3.13	Beispiel für Dateien, die experimentelle <i>Peakformen</i> enthalten	75
3.14	Beispiel für künstlich erzeugtes 2D-NOESY-Spektrum	79
3.15	Zahl der gefundenen Zuordnungen in Abhängigkeit von der partiellen Zuordnung	80
3.16	Energiefunktion und Wärmekapazität eines Optimierungslaufes	81
3.17	Vergleich eines idealen und eines verrauschten Datensatzes	82
3.18	Zahl der gefundenen Zuordnungen in Abhängigkeit von der Rauschrate n_r .	83
3.19	Zahl der gefundenen Zuordnungen in Abhängigkeit von der Anzahl der Artefaktsignale und entfernten NOE-Signale	84
3.20	Primärsequenz der HPr Mutante H15A	86
4.1	Fehlerhafte Zuordnung von HN 85 und HG2 24	104

Tabellenverzeichnis

3.1	Datei mit partieller Zuordnung	76
3.2	Datei mit angepassten Zuordnungen	77
3.3	Gefundene chemische Verschiebungen im künstlichen 3D-Spektrum	86
3.4	Gefundene Zuordnungen im experimentellen Datensatz I	88
3.5	Gefundene Zuordnungen im experimentellen Datensatz II	89
3.6	Gefundene Zuordnungen im experimentellen Datensatz III	89
3.7	Auszug aus einer A-Liste	91
3.8	Auszug aus einer B-Liste	92
3.9	Vergleich der R-Werte zweier Strukturen	94
4.1	Vergleich der R-Werte der automatisch und manuell bestimmten Struktur von <i>TmCsp</i>	97

Abkürzungsverzeichnis

2D	zweidimensional
3D	dreidimensional
4D	vierdimensional
nD	n-dimensional
ANSI	American National Standards Institute
ASCII	American Standard Code for Information Interchange
CSA	Chemical Shift Anisotropy
HPr	Histidin Containing Protein
kDa	kilo Dalton
KB	Kilobyte
MB	Megabyte
MHz	Megahertz
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect oder Nuclear Overhauser Enhancement
NOESY	Nuclear Overhauser Enhancement Spectroscopy
PDB	Protein Data Base
ppm	parts per million, $1 : 10^6$
<i>TmCsp</i>	Kälteschock-Potein von <i>Thermotoga maritima</i>

Literaturverzeichnis

- [1] Accelrys, San Diego, CA
- [2] Antz, C., Neidig, K.-P., Kalbitzer, H.R. (1995) *J. Biomol. NMR*, **5**, 287-296.
- [3] <http://www.biologie.uni-regensburg.de/Biophysik/Kalbitzer/software/index.html>
- [4] Bailey-Kellogg, C., Widge, A., Kelley, J.J., Berardi, M.J., Bushweller, J.H., Donald, B.R. (2000) *J. Comp. Biol.*, **7**, 537-558
- [5] Baleja, J.D., Moulton, J., Sykes, B.D. (1990) *J. Magn. Reson.*, **87**, 375-384
- [6] Baltimore, D. (2001) *Nature*, **409**, 814-816
- [7] Banks, K.M., Hare, D.R., Reid, B.R. (1989) *Biochemistry*, **28**, 6996-7010
- [8] Bartels, C., Xia, T., Billeter, M., Güntert, P., Wüthrich, K. (1995) *J. Mol. Biol.*, **6**, 1-10
- [9] Bartels, C., Güntert, P., Billeter, M., Wüthrich, K. (1997) *J. Comp. Chem*, **18**, 139-149
- [10] Bax, A., Grzesiek, S. (1993) *Acc. Chem. Res.*, **26**, 131-138
- [11] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535-542
- [12] BioMagResBank, www.bmrb.wisc.edu
- [13] Boelens, R., Koning, T.M.G., van der Marel, G.A., Boom, J.H., Kaptein, R. (1989) *J. Magn. Reson.*, **82**, 290-308
- [14] Bonvin, A.M.J.J., Boelens, R., Kaptein, R. (1991) *J. Biomol. NMR*, **1**, 305-309
- [15] Borgias, B.A., James, T.L. (1990) *J. Magn. Reson.*, **87**, 475-487

-
- [16] Borgias, B.A., Thomas, P.D., James, T.L. (1987,1989) *CComplete Relaxation Matrix Analysis (CORMA)*, University of California, San Francisco
- [17] Bruker (1997) *Software Manual: XWINNMR processing*
- [18] Brünger, A.T., Campbell, R.L., Clore, G.M., Gronenborn, A.M., Karplus, M., Petsko, G.A., Teeter, M.M. (1987) *Science*, **235**, 1049-1053
- [19] Brünger, A.T., Clore, G.M., Gronenborn, A.M., Saffrich, R., Nilges, M. (1993) *Science*, **261**, 328-331
- [20] Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. and Warren, G.L. (1998) *Acta Cryst.* **D54**, 905-924.
- [21] Buchler, N.E.G., Zuiderweg, E.R.P., Wang, H., Goldstein, R.A. (1997) *J. Magn. Reson.*, **125**, 34-42
- [22] Cavanagh J., Fairbrother W. J., A. G. Palmer III, Skelton N. J. (1995) *Protein NMR Spectroscopy: Principles and Practice*, Academic Press
- [23] Clore, G.M., Gronenborn, A.M. (1991) *Science*, **252**, 1390-1399
- [24] Clore, G.M., Robien, M.A., Gronenborn, A.M. (1993) *J. Mol. Biol.*, **231**, 82-102
- [25] Clore, G.M., Garrett, D.S. (1999) *J. Am. Chem. Soc.*, **121**, 9008-9012
- [26] Cornfield, J. (1967) *Rev. Int. Statist. Inst.*, **35**, 34-49
- [27] Cornfield, J. (1969) *Biometrics*, **25**, 617-642
- [28] Cullinan, D., Korobka, A., Grollman, A.P., Patel, D.J., Eisenberg, M., Santos, C. (1996) *Biochemistry*, **35**, 13319-13327
- [29] Deber, C.M., Torchia, D.A., Blout, E.R. (1971) *J. Am. Chem. Soc.*, **93**, 4893-4897
- [30] Delaglio, F., Grzesiek, S., Vuister, G., Zhu, G., Pfeifer, Bax, A. (1995) *J. Biomol. NMR*, **6**, 277-293.
- [31] DeMarco, A., Llinas, M., Wüthrich, K. (1978) *Biopolymers*, **17**, 617-636
- [32] Dölker, R., Maurer, T., Kremer, W., Neidig, K.P., Kalbitzer, H.R. (1999) *Biochem. Biophys. Res. Com.*, **257(2)**, 348-350
- [33] Dueck, G., Scheuer, T. (1990) *J. Comp. Phys.*, **90**, 161-175

-
- [34] Duggan, B.M., Legge, G.B., Dyson, H.J., Wright, P.E. (2001) J. Bio. NMR, **19**, 321-329
- [35] Fushman, D., Tjandra, N., Cowburn, D. (1998) J. Am. Chem. Soc., **120**, 10947-10952.
- [36] Geyer, M., Neidig, K.-P., Kalbitzer, H.R. (1995) J. Magn. Reson. Ser. B, **109**, 31-38
- [37] Glaser, S., Kalbitzer, H.R. (1986) J. Magn. Reson., **68**, 350-354
- [38] Görler, A. (1994) *Quantitative Auswertung von NOESY-Spektren unter Berücksichtigung der Spindiffusion*, Diplomarbeit, Heidelberg
- [39] Görler, A. (1998) *Entwicklung neuer Methoden zur Auswertung von Kern-Overhauser-Effekt-Spektren und ihre Anwendung zur Bestimmung der dreidimensionalen NMR-Struktur des Proteins HPr aus S. carnosus*, Dissertation, Heidelberg
- [40] Görler, A., Kalbitzer, H.R. (1997) J. Magn. Reson. **124**, 177-188.
- [41] Goldman, M. (1984) J. Magn. Reson. **60**, 437-452
- [42] Goldman, M. (1998) *Quantum description of high-resolution NMR in liquids*, Oxford, Clarendon
- [43] Gonzalez, C., Rullmann, J.A.C., Bonvin, A.M.J.J, Boelens, R., Kaptein, R. (1991) J. Magn. Reson., **91**, 659-664
- [44] Goody, R.S. (1995) *Proteine Spektrum*, Akad. Verlag Heidelberg
- [45] Gronwald, W., Kirchhöfer, R., Görler, A., Kremer, W., Ganslmeier, B., Neidig, K.-P., Kalbitzer, H.R. (2000) J. Biomol. NMR, **17**, 137-151.
- [46] Gronwald, W., Moussa, S., Elsner, R., Jung, A., Ganslmeier, B., Kremer, W., Fischer, C., Neidig, K.-P., Kalbitzer, H.R. (2002) J. Biomol. NMR, accepted
- [47] Güntert, P., Mumenthaler, C., Wüthrich, K. (1997) J. Mol. Biol. , **273**, 283-298
- [48] Guota G., Sarma, M.H., Sarma, R.H. (1988) Biochemistry, **27**, 7909-7919
- [49] Harrieder, S. (1998) *Entwicklung von Programmen zur automatische Zuordnung n-dimensionaler NMR-Spektren von biologischen Makromolekülen und Bestimmung der 3D-Struktur eines Kälteschock-Proteins*, Diplomarbeit, Regensburg
- [50] Hausser, K.H., Kalbitzer, H.R. (1989) *NMR für Mediziner und Biologen; Strukturbestimmung, Bildgebung, In-vivo-Spektroskopie.*, Springer Verlag, Berlin Heidelberg

- [51] www.hupo.org
- [52] Jeener, J., Meier, B.H., Bachmann, P., Ernst, R.R. (1979) J. Chem. Phys., **71**, 4546
- [53] Karplus, M. (1963) J. Am. Chem. Soc., **85**, 2870-2871
- [54] Kim, S., Reid, B.R. (1992) J. Magn. Reson., **100**, 382-390
- [55] Kirchhöfer R. (2001) *Validierung von Proteinstrukturen*, Diplomarbeit, Regensburg
- [56] Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P. (1983) Science, **220**, 671-680
- [57] Koradi, R., Billeter, M., Engeli, M., Güntert, P., Wüthrich, K. (1998) J. Magn. Reson., **135**, 288-297
- [58] Koradi, R., Billeter, Wüthrich, K. (1996) J. Mol. Graph., **14**, 51-55
- [59] Kraulis, P.J. (1989) J. Magn. Reson., **84**, 627-633
- [60] Kraulis, P.J. (1994) J. Mol. Biol., **243**, 696-718
- [61] Kremer, W., Harrieder, S., Geyer, M., Gronwald, W., Welker, C., Schuler, B., Jänicke, R., Kalbitzer, H.R. (2001) Eur. J. Biochem., **268**, 2527-2539
- [62] Lane, A.N. (1990) Biochim. Biophys. Acta, **1049**, 189-204
- [63] Laskowski, R.A., MacArthur, M.W., Thornton, J.M. (1998) Curr. Opin. Struct. Biol., **8**, 631-639
- [64] Lehninger A.L. (1975) *Biochemie* Verlag Chemie Weinheim
- [65] Leutner, M., Gschwind, R., Liermann, J., Schwarz, C., Gemmecker, G., Kessler, H. (1998) J. Biomol. NMR, **11**, 31-43.
- [66] Lefevre, J.F., Lane, A.N., Jardetzky, O. (1987) Biochemistry, **26**, 5076-5090
- [67] Li, K.B., Sanctuary, B.C. (1997) J. Chem. Inform. Comp. Science **37**, 467-477
- [68] Llinas, M., to be published
- [69] Lipari, G., Szabo, A. (1982) J. Am. Chem. Soc., **104**, 4546-4559
- [70] Lipari, G., Szabo, A. (1982) J. Am. Chem. Soc., **104**, 4559-4570
- [71] Ludvigsen, S., Andersen, K.V., Poulsen, F.M. (1991) J. Mol. Biol., **217**, 731-736
- [72] Lukin, J.A., Gove, A.P., Talukdar, S.N., Ho, C. (1997) J. Biomol. NMR **9**, 151-166

-
- [73] Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E., Wüthrich, K. (1998) *Pure & Appl. Chem.*, **70**, 117-142
- [74] Mertz, J.E., Güntert, P., Wüthrich, K., Braun, W. (1991) *J. Biomol. NMR*, **1**, 257-269
- [75] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. (1953) *J. Chem. Phys.*, **21**, 1087-1092
- [76] Mitschang, L., Neidig, K.P., Kalbitzer, H.R. (1990) *J. Magn. Reson.*, **90**, 359-362
- [77] Möglich, A. (2001) *Strukturbestimmung der Ile14→Ala-Austauschmutante des Histidine-Containing Phosphocarrier Protein aus Staphylococcus carnosus unter Verwendung von Restdipolkopplungen*, Diplomarbeit, Regensburg
- [78] Moseley, H., Montelione, G. (1999) *Structural Biology*, **9**, 635-642
- [79] Moussa, S. (2001) *NMR-spectroscopy of polypeptides and a new statistical method for the assignment of nuclear overhauser effect signals*, Dissertation, Regensburg
- [80] Mueller, G.A., Choy, W.Y., Yang, D., Forman-Kay, J.D., Venters, R.A., Kay, L.E. (2000) *J. Mol. Biol.*, **300**, 197-212
- [81] Mumenthaler, C., Braun, W. (1995) *J. Mol. Biol.*, **254**, 465-480.
- [82] Munte, C. (2001) *Ressonância Magnética Nuclear na Determinação de Estrutura de Proteínas: Aplicação à Mutante His15Ala de HPr de Staphylococcus aureus, e ao Peptídeo-C da Proinsulina Humana*, Dissertation, São Paulo
- [83] Nasser, A. (2001) *Neue Methoden zur Auswertung von Protein-NMR-Daten*, Diplomarbeit, Regensburg
- [84] Neidig, K.P., Saffrich, R., Lorenz, M., Kalbitzer, H.R. (1990) *J. Mag. Reson.*, **89**, 543-552
- [85] Neidig, K.-P., Geyer, M., Görler, A., Antz, C., Saffrich, R., Beneicke, W., Kalbitzer, H.R. (1995) *J. Mol. Biol.*, **6**, 255-270.
- [86] Nikonowicz, E.P., Meadows, R.P., Gorenstein, D.G. (1990) *Biochemistry*, **29**, 4193-4204
- [87] Nilges, M., Habazettl, J., Brünger, A.T., Holak, T.A. (1991) *J. Mol. Biol.*, **219**, 499-510
- [88] Nilges, M. (1995) *J. Mol. Biol.*, **245**, 645-660.

-
- [89] Olson, J., Markley, J. (1994) J. Mol. Biol., **4**, 385-410.
- [90] Oshiro, C.M., Kuntz, I.D. (1993) Biopolymers, **33**, 107-115
- [91] Overhauser, A.W. (1953) Physical Review, **89**, 689
- [92] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. (1992) *Numerical Recipes in C - The Art of Scientific Computing*, Cambridge University Press
- [93] Ried, A. (2001) *Simulation von Linienbreiten, T2-Zeiten und indirekter Spin-Spin-Kopplung in multidimensionalen NOESY-Spektren*, Diplomarbeit, Regensburg
- [94] Riepl, H., to be published
- [95] Pardi, A., Billeter, M., Wüthrich, K. (1984) J. Mol. Biol., **180**, 741-751
- [96] Saffrich, R., Beneicke, W., Neidig, K.P., Kalbitzer, H.R. (1993) J. Magn. Reson. B, **101**, 304-308
- [97] Schmidt, J., Blümel, M., Löhr, F., Rüterjans, H. (1999) J. Biomol. NMR, **14**, 1-12
- [98] Schulte, A.C., Görler, A., Antz, C., Neidig, K.P., Kalbitzer, H.R. (1997) J. Magn. Reson., **129**, 165-172
- [99] Smith, L.J., Sutcliffe, M.J., Redfield, C., Dobson, C.M. (1991) Biochemistry, **30**, 986-996
- [100] Thomas, P.D., Basus, V.J., James, T.L. (1991) Proc. Natl. Acad. Sci., **88**, 1237-1241
- [101] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G. (1997) Nucl. Acids Res., **24**, 4876-4882
- [102] Tjandra, N., Grzesiek, S., Bax, A. (1996) J. Amer. Chem. Soc., **118**(26), 6264-6272
- [103] van de Ven, F.J.M., Blommers, M.J.J., Schouten, R.E., Hilbers, C.W. (1991) J. Magn. Reson., **94**, 140-151
- [104] Vuister, G.W., Bax, A. (1993) J. Am. Chem. Soc., **115**, 7772-7777
- [105] Wüthrich, K. (1986) *NMR of proteins and nucleic acids*, John Wiley and Sons
- [106] Wüthrich, K. (1976) *NMR in biological Research: Peptides and Proteins*, North Holland, American Elsevier
- [107] Zhu, L., Reid, B.R. (1995) J. Magn. Reson. B, **106**, 227-235
- [108] Zimmermann, D., Kulikowski, C., Wang, L., Lyons, B., Montelione, G. (1994) J. Mol. Biol., **4**, 241-256.

Anhang A

Relaxationstheorie

Um die Berechnung der Relaxationsraten zu vereinfachen werden nur zwei Grenzfälle betrachtet und zwar die *unlike* bzw. *like* Spins. Zwei Spins I und S nennt man *unlike*, wenn $\omega_I \neq \omega_S$, ansonsten nennt man sie *like* Spins.

Betrachte hier nur ein System zweier *unlike* Spins $\frac{1}{2} I$ und S . Der zeitabhängige Hamiltonian, der dieses System beschreibt, lautet

$$H(t) = H_0 + H_1(t) \quad (\text{A.1})$$

H_0 beschreibt die Zeeman Wechselwirkung und die schwache Spin-Spin Kopplung zwischen den beiden Spins I und S :

$$H_0 = \omega_I I_z + \omega_S S_z + 2\pi J I_z S_z \quad (\text{A.2})$$

I_z und S_z sind die Magnetisierung von Spin I bzw. S in Richtung des äusseren Magnetfeldes und J ist die J-Kopplungskonstante zwischen I und S .

Der stochastische, zeitabhängige Hamiltonian $H_1(t)$ in Gleichung A.1 beschreibt die schwache Kopplung zwischen den Spins und dem Gitter, also in unserem Fall die dipolare Wechselwirkung und die CSA. $H_1(t)$ lässt sich als Linearkombination von Basisoperatoren schreiben

$$H_1(t) = H_{Dipolar} + H_{CSA} = \sum_i H_1^i(t) = \sum_i \sum_{\alpha} V_{\alpha}^i F_{\alpha}^i(t) \quad (\text{A.3})$$

$F_{\alpha}^i(t)$: Zufallsfunktionen

V_{α}^i : Spinoperatoren

Wenn man nun annimmt, dass die anisotrope chemische Verschiebung von Spin S die Relaxation von Spin I nicht beeinflusst (und umgekehrt), dass der Tensor der chemischen Verschiebung von Spin I bzw. S symmetrisch ist und reale Matricelemente besitzt und man in ein Koordinatensystem wechselt, in der die z-Achse parallel zum Verbindungsvek-

tor von Spin I und S steht, so kann man $H_1(t)$ schreiben

$$H_1(t) = H_{Dipolar} + H_{CSA} \quad (A.4)$$

$$= -\sqrt{\frac{24\pi}{5}} \frac{\gamma_I \gamma_S \hbar \mu_0}{r^3} \frac{1}{4\pi} \sum_{\alpha=-2}^2 Y_2^{-\alpha}(t) V_2^{\alpha}(S, I) \quad (A.5)$$

$$+ \frac{1}{3} \gamma_I (\sigma_{\parallel} - \sigma_{\perp}) \sqrt{\frac{24\pi}{5}} \sum_{\alpha=-2}^2 Y_2^{-\alpha}(t) V_2^{\alpha}(H, I) \quad (A.6)$$

$Y_2^{-\alpha}$: Kugelflächenfunktionen

$\sigma_{\parallel} = \sigma_z, \sigma_{\perp} = \sigma_x = \sigma_y$: Komponenten des anisotropen Tensors der chemischen Verschiebung

H : magnetisches Feld

Die zeitliche Entwicklung einer physikalischen Observablen Q ist gegeben durch [42]

$$\frac{d}{dt} \langle Q \rangle = Tr \left\{ Q \frac{d\sigma^T(t)}{dt} \right\} = - \sum_{i,j} \sum_{\alpha} J_{\alpha}(\omega_{\alpha}) \left\{ \langle [[Q, V_{\alpha}^i], V_{\alpha}^{j+}] \rangle - \langle [[Q, V_{\alpha}^i], V_{\alpha}^{j+}] \rangle_{eq} \right\} \quad (A.7)$$

Für die Berechnung der longitudinalen Relaxationsrate ist die Änderung der Magnetisierung in Richtung des äusseren Feldes wichtig, d.h. man muss $\frac{d}{dt} \langle I_z \rangle$ bzw. $\frac{d}{dt} \langle S_z \rangle$ für Spin I bzw. S berechnen.

Anhang B

Kommandos von AUREMOL

Die Kommandos in den Pull Down Menüs:

1. File

- 1.1. Open
- 1.2. Read XWIN-NMR 1D-file
- 1.3. Read XWIN-NMR 2D-file
- 1.4. NMR 3D-file
 - 1.4.1. open 3D spectrum
 - 1.4.2. read w2-w3 plane
 - 1.4.3. read w1-w2 plane
 - 1.4.4. read w1-w3 plane
 - 1.4.5. next plane
 - 1.4.6. previous plane
- 1.5. Read molecule
- 1.6. Close objects
- 1.7. Print
- 1.8. Copy to clipboard
- 1.9. Open session → Kapitel 3.1.3
- 1.10. Save session → Kapitel 3.1.3
- 1.11. Reread processing parameters
- 1.12. Aurelia export → Kapitel 3.2

1.13. Exit

2. **Config**

2.1. Input mode

2.2. Visibility

2.3. Geometry

2.4. Display

2.5. Arrange

2.5.1. Collect

2.5.2. Take apart

2.5.3. Combine 1D NMR

2.5.4. Change object order

2.5.5. Exchange 2 objects

2.6. Correlation

2.7. User layout

2.7.1. Redefine all windows

2.7.2. Redefine current window

2.7.3. Reposition current window

2.7.4. Save layout

2.7.5. Restore layout

2.8. User interface

2.8.1. Configure mouse sensitivity

2.8.2. Configure zoom handling

2.8.3. Configure menu

2.8.4. Change background color

3. **Analysis** → Kapitel 3.1.3

3.1. Reread intensities

3.2. Set peak options

3.3. Auto peak pick

3.4. Manual peak pick

3.5. Manipulate peaks

3.5.1. Keep peaks in area

- 3.5.2. Remove peaks in area
- 3.5.3. Remove non-annotated peaks
- 3.5.4. Remove all peaks
- 3.5.5. Remove low quality peaks
- 3.5.6. Center of gravity
- 3.5.7. Keep symmetrical peaks (homonuclear 2D)
- 3.5.8. Keep symmetrical peaks (J-res)

3.6. Annotate peaks

- 3.6.1. Annotate peaks
- 3.6.2. Annotate from molecule
- 3.6.3. Renumber peaks
- 3.6.4. Check all peaklabels

3.7. Transfer peaks

- 3.7.1. Transfer peaks (NMR)
- 3.7.2. Transfer annotated peaks (NMR)
- 3.7.3. Read XWIN-NMR peaks (NMR)
- 3.7.4. Read external peaks (NMR)

3.8. Integrate

- 3.8.1. Integrate peaks by shape → Kapitel 3.1.3
- 3.8.2. Integrate peaks by area → Kapitel 3.1.3
- 3.8.3. Normalize peak integrals
- 3.8.4. Integrate area(s)
- 3.8.5. Integrate patterns (NMR)

3.9. Output

- 3.9.1. Store peaks to text file

4. Measure

- 4.1. Distances
- 4.2. Intensities
- 4.3. J-Fit (1D, 2D NMR)
- 4.4. Calibrate (1D, 2D NMR)
- 4.5. Line manager (1D, 2D NMR)
- 4.6. Zoom manager

5. Utilities

- 5.1. X-Plor output → Kapitel 3.1.3
- 5.2. Load primary sequence → Kapitel 3.1.3
- 5.3. Show primary sequence
- 5.4. Magnetic susceptibility → Kapitel 3.1.3
- 5.5. Set processing parameters → Kapitel 3.1.3
- 5.6. Convert masterlist → Kapitel 3.1.3

6. Strips → Kapitel 3.1.3

- 6.1. Open Strip Window
- 6.2. Connect 3D spectra to strip window
- 6.3. Quit strip session
- 6.4. Show/hide lines
- 6.5. Show/hide peaks
- 6.6. Show/hide spectra
- 6.7. Change number of visible slots
- 6.8. Define new v-strip
- 6.9. Define new h-strip
- 6.10. Autodefine strips
- 6.11. Label strips
- 6.12. Strip handling
 - 6.12.1. Add strip to slot
 - 6.12.2. Move strip
 - 6.12.3. Move slot
 - 6.12.4. Insert slot
 - 6.12.5. Remove slot
 - 6.12.6. Remove single strip from slot
 - 6.12.7. Remove all strips from slot
- 6.13. Remove strip in pool
- 6.14. Set/remove horizontal help-line
- 6.15. Draw line

- 6.16. Remove line
- 6.17. Add peaks mode on/off
- 6.18. Change update mode
- 6.19. Scroll left
- 6.20. Scroll right
- 6.21. Next strip (+)
- 6.22. Previous strip (-)
- 6.23. Next pool
- 6.24. Previous pool

7. Calculation

- 7.1. Backcalculation → Kapitel 3.3
- 7.2. Pdb to spt → Kapitel 3.3
- 7.3. User spt → Kapitel 3.3
- 7.4. Add protons to pdb → Kapitel 3.1.3
- 7.5. Convert pdb-file → Kapitel 3.3
- 7.6. Peak quality → Kapitel 3.1.5
- 7.7. R-factors → Kapitel 3.5
- 7.8. Calculate angles
- 7.9. Permole → Kapitel 3.1.4
- 7.10. Knownoe → Kapitel 3.4.1
 - 7.10.1. Calculate assignment
- 7.11. 2-Step Algorithm → Kapitel 2.3.2 und 3.4.2
 - 7.11.1. Segmentate exp. spectrum
 - 7.11.2. Match spectrum: step 1
 - 7.11.3. Match spectra: step 2

8. Database → Kapitel 3.2

- 8.1. Create compoundfile
- 8.2. Create metafile
- 8.3. Create metafile from sptfile
- 8.4. Create MasterMasterfile
- 8.5. Update MasterMasterfile

Danksagung

Sehr herzlich bedanken möchte ich mich bei Prof. Dr. Dr. Hans Robert Kalbitzer, der mir die Möglichkeit gab, eine sowohl wissenschaftliche als auch industrienähe Doktorarbeit durchzuführen. Trotz der vielen Kommissions- und Ausschuss-Sitzungen hatte er immer Zeit für (oftmals) lange und fruchtbare Diskussionen.

Für seine Unterstützung bei der Entwicklung des Programms AUREMOL und seinen unermüdlichen Einsatz, AUREMOL bei uns in der Gruppe als Nachfolger von AURELIA zu etablieren, möchte ich meinem Betreuer Dr. Wolfram Gronwald danken. Seine Bereitschaft, zu jeder Zeit NMR-spezifische Fragen zu beantworten, habe ich sehr geschätzt.

Dr. Peter Neidig von der Firma Bruker BioSpin GmbH hat es erst ermöglicht, dass in relativ kurzer Zeit ein lauffähiger Prototyp von AUREMOL erstellt werden konnte. Die Bereitstellung der grafischen Benutzeroberfläche des VIEWERS und der rege Informationsaustausch per E-Mail oder bei Meetings in Karlsruhe hat die Softwareentwicklung vorangetrieben. Dafür vielen Dank!

Bedanken möchte ich mich bei Jochen Trenner, der sich (freiwillig?) bereit erklärt hat, AUREMOL weiterzuentwickeln, mich im letzten Jahr tatkräftig beim Software-Merge unterstützt hat und seitdem den Support und die Userbetreuung von AUREMOL übernommen hat.

Ein besonderer Dank gebührt Markus Puchta, der als alter Hase auf dem Gebiet der Optimierung mir bei der Implementierung des *Threshold Accepting* Algorithmus wichtige Tipps gab.

Rolf Döker, Rudi Füchsl, Astrid Jung und Andreas Möglich möchte ich für die angenehme Zimmeratmosphäre danken.

Guido Steiner hat mir mit zahlreichen Tipps zu Linux und L^AT_EX oftmals sehr geholfen. Ich möchte mich ausserdem bei allen in unserer Gruppe bedanken, die AUREMOL benutzt und getestet haben, um den zahlreichen Bugs und Fehlern auf die Schliche zu kommen. Für evtl. geschrottete Masterlisten übernehme ich allerdings keine Verantwortung. ;-)

Ein besonderer Dank geht an meinen Schatz Bärbel, die tapfer meine komplette Arbeit gelesen und auf Rechtschreib- bzw. Kommafehler überprüft hat. Schliesslich danke ich noch meinen Eltern, die es mir durch ihre finanzielle Unterstützung ermöglicht haben, hier in Regensburg zu studieren und zu promovieren.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig angefertigt habe und keine ausser den angegebenen Hilfsmittel verwendet habe.

Regensburg, im Juli 2002

(Bernhard Ganslmeier)