Modellierung, Strukturverbesserung und sequentielle Zuordnung als vollautomatische Module für die automatisierte Proteinstrukturbestimmung im Softwareprojekt AUREMOL

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER NATURWISSENSCHAFTLICHEN FAKULTÄT III – BIOLOGIE UND VORKLINISCHE MEDIZIN DER UNIVERSITÄT REGENSBURG

vorgelegt von

Konrad Brunner

aus Schwandorf

im Juli 2006

Promotionsgesuch eingereicht am: 26.07.2006
Die Arbeit wurde angeleitet von: Prof. Dr. Dr. Hans Robert Kalbitzer

Prüfungsausschuss:

Vorsitzender:	Prof. Dr. Ralph Witzgall
Erstgutachter:	Prof. Dr. Dr. Hans Robert Kalbitzer
Zweitgutachter:	PD Dr. Rainer Merkl
Drittprüfer:	Prof. Dr. Eike Brunner

Inhaltsverzeichnis

1 Einleitung

1.1 Proteine

Jede Lebensform stellt in ihren Zellen Proteine her, um mit deren Hilfe wichtige Funktionen des Lebens zu regeln. Diese Funktionen sind weit gefächert und regulieren z. B. die Genexpression und den Stoffwechsel, dienen als Transporter für kleinere Moleküle und wirken als Rezeptoren oder Botenstoffe bei der Signalweiterleitung mit. Strukturproteine bilden das Gerüst jeder einzelnen Zelle oder stabilisieren ganze Organismen. Weiterhin unterstützen sie auch die Muskelkontraktion oder die Fortbewegung von Zellen mithilfe von Geißeln, Flagellen oder Cilien. Eine der wichtigsten Aufgaben von Proteinen besteht darin, biochemische Reaktionen zu beschleunigen oder überhaupt ablaufen zu lassen, indem sie als Enzyme fungieren, die als Biokatalysatoren in eine Reaktion eingreifen. Die Proteine bilden mit fast 50 Prozent der trockenen Masse den Hauptbestandteil jeder Zelle.
Aufgebaut sind diese Makromoleküle aus Kohlenstoff, Wasserstoff, Sauerstoff, Stickstoff und Schwefel. Proteine sind unverzweigte Kettenmoleküle (Polymere), die aus 21 Bausteinen, den so genannten proteinogenen Aminosäuren, aufgebaut sind. Die Länge dieser Aminosäureketten reicht dabei von unter 20 bis weit über 1000 Aminosäuren.

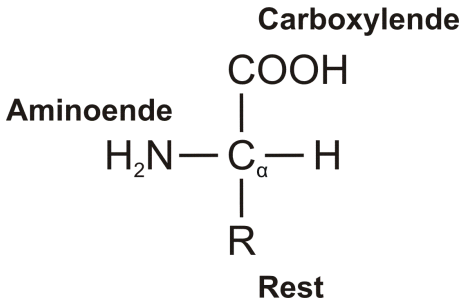

Abbildung 1: Typische Struktur einer Aminosäure.

Die verschiedenen Aminosäuren haben prinzipiell die gleiche Gestalt und unterscheiden sich nur in der spezifischen Seitenkette, die auch Aminosäurerest oder kurz Rest genannt wird. In Abbildung 1 sind die zwei funktionellen Gruppen, Carboxylgruppe -COOH und Aminogruppe -NH_2, zu sehen. Diese Gruppen dienen als Verbindung zu den anderen Aminosäuren. Die Carboxylgruppe der einen Aminosäure geht mit der Aminogruppe der anderen Aminosäure unter Abspaltung von Wasser eine so genannte Peptidbindung ein.

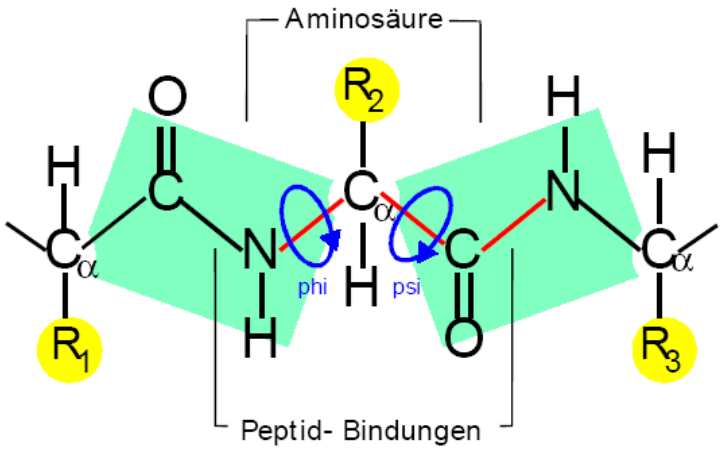

Abbildung 2: Sterische Beschränkungen der Bindungswinkel in einem Protein. Jede Aminosäure trägt drei Bindungen (rot) zu einer Polypeptidkette bei. Die C=O, C-N und H-N- Bindungen sind eben und erlauben keine Drehung. Drehungen in der Hauptkette können nur um die C_α -C und um die N-C_α-Bindung stattfinden. Die R_i bezeichnen die Aminosäuren-Seitenketten. (nach [1]).

Für die Ausübung der Funktion des Proteins ist seine räumliche Struktur von herausragender Bedeutung. Die Struktur eines Proteins wird in vier Klassen eingeteilt (Abbildung 3).

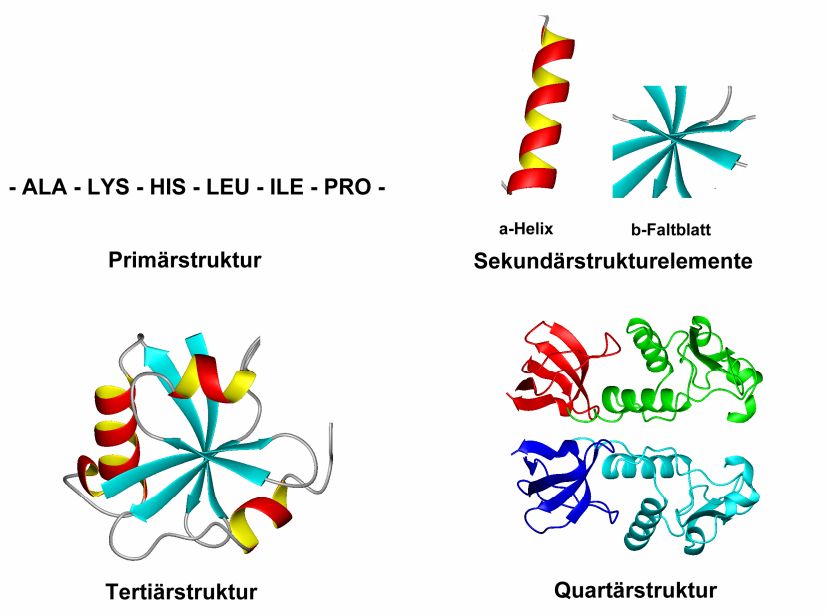

Abbildung 3: Die vier Klassen der Proteinstruktur.

Unter der Primärstruktur versteht man die Reihenfolge der Aminosäuren, die Sequenz. Die Sekundärstruktur beschreibt die räumliche Anordnung nah benachbarter Aminosäuren. Der Carbonyl-Sauerstoff und der Stickstoff der Amino-Gruppe von nicht direkt benachbarten Aminosäuren bilden Wasserstoffbrücken, wenn zwei geeignet weit voneinander getrennte Abschnitte eines Makromoleküls über Wasserstoffatome in Wechselwirkung treten, und es

entstehen vorzugsweise zwei Sekundärstrukturen, das β-Faltblatt und die α-Helix. Als Tertiärstruktur wird die räumliche Anordnung der gesamten Polypeptidkette bezeichnet. Helikale Abschnitte können mit Faltblattstrukturen abwechseln. Dazwischen liegen Teilstücke, in denen die Polypeptid-Kette in unregelmäßigen Schleifen verläuft. Werden Proteine aus mehreren Polypeptid-Ketten oder aus mehreren funktionellen Gruppen gebildet, so gibt die Quartärstruktur an, aus welchen und wie vielen solcher Untereinheiten ein Protein besteht, wie diese räumlich zueinander orientiert sind und ob sie kovalent oder über Wasserstoffbrücken, ionisch oder über *van-der-Waals*-Bindungen verknüpft sind.

Ein Lehrsatz der Biophysik besagt, dass die räumliche Struktur von biologischen Makromolekülen deren biologische Funktion bestimmt. Der Beweis hierfür ist die Feststellung, dass mit der Denaturierung eines Proteins, also mit der Zerstörung der geordneten 3D-Struktur, ein Funktionsverlust einhergeht und es seine Aufgaben nicht mehr oder nur noch teilweise wahrnehmen kann. Ebenso kann ein Fehler in der Faltung zu pathologischen Symptomen führen. Prominente Beispiele dafür sind die Krankheiten BSE, Creutzfeld-Jakob und Altzheimer [2-6]. Die Kenntnis der 3D Struktur ermöglicht also ein detailliertes Verständnis der biologischen Funktion des Proteins.
Im Wesentlichen gibt es zwei experimentelle Methoden zur Strukturaufklärung mit atomarer Auflösung. Zum einen die ältere Methode der Röntgenstrukturanalyse und zum anderen die Kernresonanzspektroskopie (NMR). Jede Methode hat ihre Vor- und Nachteile. So wird bei der Röntgenstrukturanalyse das Protein im kristallinen Zustand benötigt, was ihr folgende Nachteile einbringt:

- Viele Proteine kristallisieren nicht.
- Mögliche Unterschiede der Proteinstruktur im kristallinen Zustand und in Lösung.
- Dynamische Prozesse sind schwer beobachtbar.
- Die gefundenen Strukturen ergeben sich aus einer räumlichen statistischen Mittelung aller vorhanden Strukturen innerhalb des Kristalls.

Für die Röntgenstrukturanalyse spricht, dass Proteine mit bis zu mehreren hundert Kilodalton (kDa) aufgeklärt werden können und dass die Strukturbestimmung sehr schnell geht, wenn die Kristallographiedaten vorliegen. Die Vorteile der Röntgenstrukturanalyse sind gleichzeitig die Nachteile der NMR. Die Größenbeschränkung bei der NMR beträgt zurzeit etwa 30 bis 40 kDa und die Auswertung der Daten ist sehr zeitaufwändig.

Allerdings bietet die NMR entscheidende Vorteile:

- Es wird kein Kristall benötigt.
- Das Protein liegt im nativen Zustand vor.
- Dynamische Prozesse beobachtbar auf einer Zeitskala von Pikosekunden bis Millisekunden.

Das Institut für Biophysik und physikalische Biochemie der Universität Regensburg beschäftigt sich u. a. mit der 3D-strukturbasierten Analyse von Funktionen und Interaktionen biologischer Makromoleküle in Lösung und verwendet dafür vor allem die Methode der NMR-Strukturaufklärung, auf die nun näher eingegangen werden soll.

1.2 NMR-Spektroskopie

Der erste Schritt bei der NMR-Spektroskopie besteht in der Herstellung der Proteinproben entweder *in vivo* (in einem Lebewesen oder einer lebendigen Zelle) oder *in vitro* (im Reagenzglas). Liegt das exprimierte Protein in genügend hoher Konzentration und Reinheit vor und befindet es sich in der gewünschten Konformation, kann mit der Spektroskopie begonnen werden.

Die Wechselwirkung des magnetischen Moments μ eines Atomkerns mit einem äußeren magnetischen Feld bildet die Grundlage der NMR-Spektroskopie. Deswegen wird das in einem Lösungsmittel gelöste Protein in ein starkes magnetisches Feld B_0, das in z-Richtung zeigt, gebracht. Die Protonen der Aminosäuren des Proteins besitzen einen Spin I mit dem Wert ½ und daraus ergibt sich für das magnetisches Dipolmoment in z-Richtung $\mu_z = m\gamma\hbar$. γ bezeichnet das gyromagnetische Verhältnis, das für jede Kernart charakteristisch ist, und $\hbar = \frac{h}{2\pi}$ das Plancksche Wirkungsquantum. Die magnetische Quantenzahl m nimmt die ganzzahligen Werte von $-I$ bis $+I$, bei den hier verwendeten Kernen ^{1}H, ^{15}N und ^{13}C -½ und +½ an. Dies führt zu einer Energieaufspaltung in zwei Energieniveaus $E_{\pm 1/2} = -m\gamma\hbar B_0$. Die Population der der Zustände $N_{\pm 1/2}$ im thermischen Gleichgewicht folgt der *Boltzmann-Statistik*

$$\frac{N_{1/2}}{N_{-1/2}} = \exp\left(\frac{-\gamma\hbar B_0}{k_B T}\right) = \exp\left(\frac{\Delta E}{k_B T}\right), \tag{1.1}$$

wobei ΔE die Energiedifferenz der beiden Zustände, k_B die *Boltzmannkonstante* und T die absolute Temperatur angibt.

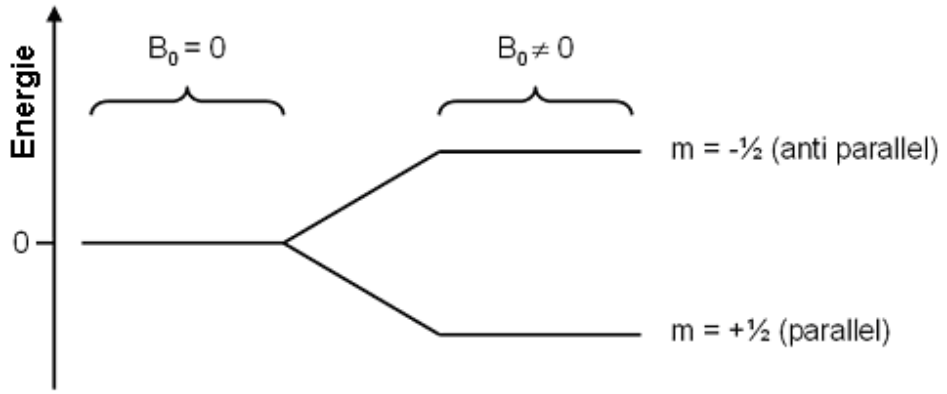

Abbildung 4: Energieniveaus eines Kerns mit der Spinquantenzahl ½ in einem magnetischen Feld B_0.

Das magnetische Moment jedes Kerns präzessiert unter dem Einfluss des Magnetfelds B_0 um die z-Achse mit der Larmorfrequenz

$$\nu_0 = \frac{\gamma}{2\pi} B_0 = \frac{\Delta E}{h}. \tag{1.2}$$

Sie entspricht der Resonanzfrequenz des Kerns und damit der Übergangsfrequenz zwischen den beiden Energieniveaus. Unterschiedliche chemische Umgebungen eines jeden einzelnen Kernes führen dazu, dass die einzelnen Kerne wegen Abschirmungseffekten der umgebenden Elektronen leicht unterschiedlichen magnetischen Feldern unterliegen. Dies macht sich in unterschiedlichen Larmorfrequenzen für die Kerne bemerkbar. Erst dieses Phänomen macht NMR-Spektroskopie überhaupt möglich, denn so können die verschiedenen Kerne mit ihren daraus resultierenden unterschiedlichen chemischen Verschiebungen unterschieden werden.
Strahlt man ein elektromagnetisches Wechselfeld bei der Resonanzfrequenz ν_0 ein, so wird aus diesem Feld Energie absorbiert und es werden Übergänge zwischen den Energieniveaus der Spins induziert. Während der Relaxation zurück ins Gleichgewicht, wird eine Strahlung emittiert, die sich aus den gedämpften Schwingungen aller beteiligten Spins zusammensetzt. Diese Strahlung bezeichnet man als FID (*Free Induction Decay*) und wird während des NMR-Experiments aufgezeichnet. Um die so erhaltenen Daten der Zeitdomäne auszuwerten, ist eine so genannte Prozessierung notwendig. Hierbei werden mit Hilfe der Fourier-Transformation [7] die Daten in die Frequenzdomäne übertragen. Durch geeignete, vorher auf den FID angewandte mathematische Filter kann u. a. die Linienform, die Auflösung oder das

Signal-Rausch-Verhältnis optimiert werden. Eine anschließende Phasen- und Basislinienkorrektur [8] macht die Prozessierung vollständig und die Auswertung der Daten kann beginnen.

Komplexe organische Substanzen, wie es Proteine sind, liefern aufgrund ihrer Größe eine Vielzahl an Signalen, die bei eindimensionalen Spektren zu erheblichen Überlappungen führen, so dass eine eindeutige Zuordnung der Signale und damit eine Strukturbestimmung unmöglich ist.

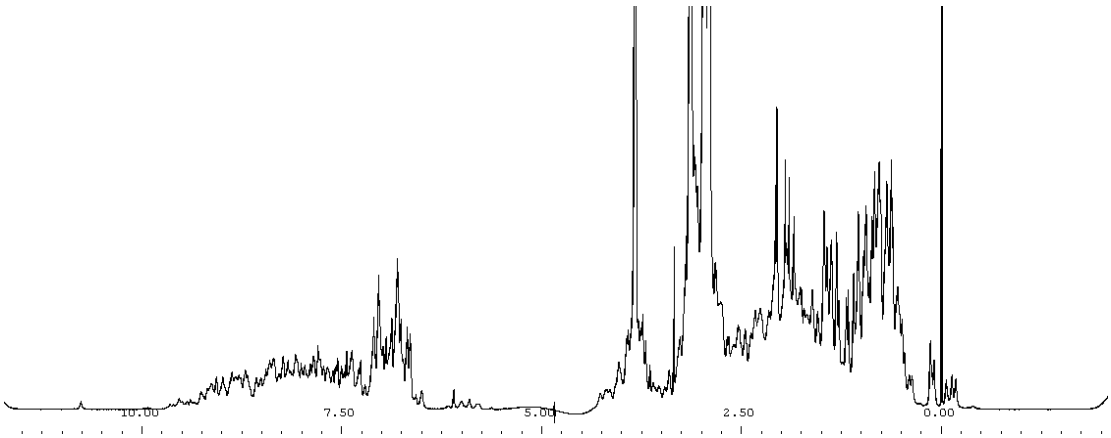

Abbildung 5: 1D-Spektrum von Ras (T35A). Die Überlappung der Resonanzen verhindert eine eindeutige Signalidentifizierung.

Deshalb wurde zur Durchführung mehrdimensionaler NMR-Experimente übergegangen.

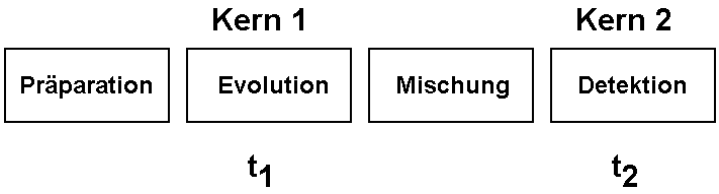

Abbildung 6: Schematische Darstellung eines 2D NMR-Experiments

Das 2D-NMR-Experiment entspricht im Prinzip vielen, hintereinander aufgenommenen, 1D-Spektren, wobei die Evolutionszeit t_1 inkrementiert wird. Während der Mischung erfolgt der Magnetisierungsübertrag von Kern 1 auf Kern 2 mittels skalarer oder dipolarer Kopplung. Sind am Experiment nur Kerne gleicher Art, z. B. H, beteiligt, so spricht man von homonuklearen Experimenten. Bei heteronuklearen Experimenten kommen weitere NMR-

aktive Isotope wie ^{13}C und ^{15}N, die ebenfalls Spin ½ besitzen, zum Einsatz. In Abbildung 7 ist ein typisches 2D-NOESY-Spektrum zu sehen.

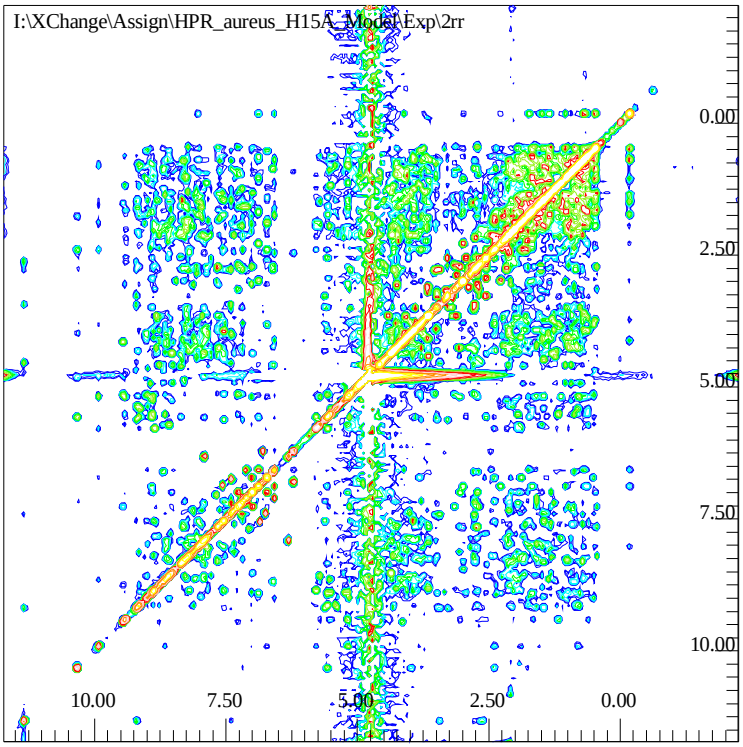

Abbildung 7: Typisches 2D-NOESY Spektrum (HPr *S. aureus*).

Mehrere Mischprozesse hintereinander führen zu zusätzlichen Kreuzsignalen. Fügt man zwischen die Mischprozesse weitere Phasen freier Evolution ein, kommt man zu höherdimensionalen (3D-, 4D-, etc.) Spektren. Auf eine weitere Vertiefung wird hier nicht eingegangen und auf entsprechende Literatur verwiesen [9] [10].

Für die Strukturbestimmung von Proteinen ist es wichtig, zuerst die chemischen Verschiebungen der beteiligten Kerne zu finden und die sequentielle Zuordnung zu erhalten. Hierzu dienen bei kleineren Proteinen (< 15 kDa) u. a. das 2D-TOCSY und das 2D-COSY Experiment. Im 2D-TOCSY Experiment (*Total Correlation Spectroscopy*) wird die Magnetisierung durch die skalare Kopplung aufeinander folgender Kerne über das komplette Spinsystem einer Aminosäure transferiert. Deshalb werden hier Kreuzsignale zwischen allen am Spinsystem beteiligten Kernen beobachtet. Dieses charakteristische Muster von Signalen ist für jede Aminosäure unterschiedlich und ermöglicht es, die Aminosäuren zu identifizieren. Beim 2D-COSY-Experiment (*Correlation Spectroscopy*) wird die Magnetisierung durch eine skalare Kopplung (J -Kopplung) übertragen. Protonen, die mehr als drei chemische

Bindungen voneinander entfernt sind, ergeben keine Kreuzsignale, denn die Kopplungskonstante $^{>3}J$ hierfür ist nahezu Null. Besonders wichtig sind dabei die Signale zwischen den H^N- und H^α-Protonen, weil der Torsionswinkel Φ der Protein-Hauptkette aus der ^3J-Kopplungskonstante über die Karplus-Beziehung [11] zwischen diesen beiden Protonen bestimmt werden kann.

Den mit der Größe des Proteins wachsenden Überlappungen der Signale kann mit HSQC-Experimenten (*Heteronuclear Single Quantum Coherence*) entgegengewirkt werden. Diese Experimente messen den Übertrag der Magnetisierung auf einen Heterokern und wieder zurück. Mit Hilfe dieser Technik ist es möglich, in 3D-Experimenten überlappende Resonanzen zu entzerren.

Bei größeren Proteinen (>15kDa) wird die Zuordnung des Rückgrates wegen Überlappungen und fehlenden Signalen immer schwieriger. Abhilfe schaffen hier Tripleresonanzexperimente. Diese werden an doppelt markierten ($^{15}N/^{13}C$)-Proteinen durchgeführt. Auf jeder Frequenz gibt es nur wenige Signale und das Problem der Überlappungen tritt somit seltener auf. Zudem wird der Magnetisierungstransfer durch starke ^1J bzw. ^2J Kopplungen zwischen den Kernen übertragen, was zu einer hohen Empfindlichkeit führt.

Hat man die sequentielle Zuordnung gefunden, so kann man mit ihr die aus dem NOESY-Experiment (*Nuclear Overhauser and Exchange Spectroscopy*) gewonnenen Spektren zuordnen. Diese Spektren sind unverzichtbar, um Proteinstrukturen zu bestimmen. Die physikalische Grundlage ist der Kern-Overhauser-Effekt, der NOE (*Nuclear Overhauser Effect*). Er besagt, dass ein Kernspin mittels dipolarer Wechselwirkung Magnetisierung auf einen räumlich benachbarten Kernspin übertragen kann. Im 2D-NOESY-Spektrum wird für ein räumlich benachbartes Protonenpaar ein Kreuzsignal (*Peak*) erwartet. Das Volumen des Signals ist dabei in erster Näherung proportional zu r^{-6}, wobei r der Abstand der Protonen ist. Im Normalfall erhält man ab einem Abstand größer als 0,5 nm keine Signale mehr. Es ist klar, dass für die Strukturbestimmung NOE-Signale von Protonen, die in der Aminosäuresequenz weit auseinander liegen, in der Tertiärstruktur aber benachbart sind, besonders wichtig sind.

Sind die Signale des NOE-Spektrums zugeordnet, können daraus Atomabstände bestimmt werden, die dann zur Einschränkung des Konformationsraums in Moleküldynamik-Simulationen dienen.

Der aufwändigste Schritt bei der Bestimmung der Proteinstrukturen ist die manuelle Auswertung der NMR-Spektren. Deswegen werden große Bemühungen dahingehend

unternommen, eine Automatisierung dieses Prozesses oder zumindest von Teilprozessen zu erreichen. Obwohl schon lange an der automatischen Auswertung von NMR-Spektren gearbeitet wird, ist das Problem bis jetzt noch nicht vollständig gelöst.

1.3 Strategien zur automatische Strukturbestimmung aus NMR-Spektren

Es existieren zwei Strategien zur automatischen Strukturbestimmung aus NMR-Spektren. Im *Bottom-Up*-Ansatz soll mit Hilfe möglichst vieler experimenteller Daten die sequentielle Zuordnung gewonnen und dann die räumliche Struktur bestimmt werden. Zahlreiche Programme, die in Kapitel 3.3 noch genauer beschrieben werden, sind dazu veröffentlicht. Diese Programme stützen sich primär auf eine NMR-zentrierte Auswertung der Spektren und nutzen wenig Information über das verwendete Protein.

Die andere Strategie ist der so genannte *Top-Down*-Ansatz. Hier wird bereits vor der Analyse möglichst viel Wissen gesammelt und bei der anschließenden Auswertung benutzt. Ausgehend von der bekannten bzw. vermuteten oder einer homologen Molekülstruktur werden die NMR-Parameter möglichst genau vorhergesagt. Dazu zählen z. B. chemische Verschiebungen und Torsionswinkel. Neben der Struktur werden zusätzlich statistische Analysen von Proteinstruktur-Datenbanken genutzt. Die Startstruktur wird iterativ solange verfeinert, bis sie optimal mit den experimentellen Daten übereinstimmt. Im Extremfall werden die experimentellen Daten nur benötigt, um die Startstruktur zu verifizieren. Ein Vorteil ist, dass nicht notwendigerweise eine vollständige sequentielle Zuordnung benötigt wird. 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. Das in dieser Arbeit weiter entwickelte Computerprogramm AUREMOL nutzt genau diesen *Top-Down*-Ansatz.

1.4 AUREMOL

Die am Institut für Biophysik und physikalische Biochemie der Universität Regensburg entwickelte Software AUREMOL wird in Kooperation mit dem Spektrometerhersteller Bruker BioSpin GmbH entwickelt. 1999 wurde die Zusammenarbeit gestartet, bei der in Anlehnung an AURELIA [12] ein Programmpaket entwickelt wird, das eine automatischen Strukturbestimmung von biologischen Makromolekülen in Lösung zum Ziel hat und die 3D-Strukturvorhersage von Proteinen erlaubt. Bruker BioSpin stellt den AMIX^{TM}-Viewer als

Gerüst zur Verfügung, der die graphischen und betriebssystemspezifischen Anforderungen bereits enthält. Module zur wissenschaftlichen Analyse, Automatisierung und Evaluation werden vom Institut erarbeitet. Dabei ist das Ziel, so wenig wie möglich an experimentellen Daten zu verwenden und manuelle Eingriffe so weit wie möglich zu reduzieren. Wie oben erwähnt, nutzt AUREMOL den *Top-Down*-Ansatz. Eine Datenbank für statistische Analysen ist bereits in AUREMOL enthalten und ein Modul zur automatischen Generierung homologer Startstrukturen wurde im Rahmen dieser Arbeit erstellt. Lediglich für die Spektrenprozessierung und für die Strukturberechnung werden externe Programme benötigt. In Kapitel 2 werden einige wichtige Funktionen näher erläutert, die auch im Rahmen dieser Arbeit Verwendung finden. Für eine detaillierte Übersicht der implementierten Funktionen wird auf den Review-Artikel von Gronwald und Kalbitzer [13] verwiesen.
Testversionen können von der AUREMOL-Homepage (www.auremol.de) herunter geladen werden.

1.5 Ziele der vorliegende Arbeit

Eine schnelle und genaue Strukturbestimmung von Proteinen ist eines der wichtigsten Ziele im Zeitalter der Proteomik. Deshalb beschäftigt sich die vorliegende Arbeit mit neuen Routinen, aber auch mit Korrekturen und Pflege bereits bestehender Module für das Softwarepaket AUREMOL, das zur manuellen, halbautomatischen und vollautomatischen Auswertung von NMR-Spektren dient.
Ein Überblick über die wichtigsten Module in AUREMOL wird in Kapitel 2.1 gegeben. Ein Flussdiagramm zeigt, wie die einzelnen Module ineinander greifen und eine kurze Beschreibung der Module macht deren Arbeitsweise klar. Damit diese Module reibungslos zusammenarbeiten können, ist im Rahmen dieser Arbeit eine Standardisierung der verwendeten Datenformate mittels einer durchgängigen Einführung der IUPAC-Nomenklatur entstanden. Zu diesem Zweck wurden mehrere Konverter entwickelt.
Kapitel 2.2 beschäftigt sich mit den in dieser Arbeit verwendeten Strukturberechnungsprogrammen. Es wird die grundsätzliche Arbeitsweise derartiger Programme erläutert und welche Eingabedaten Verwendung finden. Alle im Rahmen dieser Arbeit entstandenen Proteinstrukturen wurden mit diesen Programmen berechnet.
Weil es sinnvoll ist, die erhaltenen Strukturen auch hinsichtlich ihrer Qualität zu bewerten, werden in Kapitel 2.3 Strukturvalidierungsmechanismen vorgestellt. Anhand der Ergebnisse dieser Methoden wird die Qualität von Proteinstrukturen beurteilt.

Kapitel 3 gibt die theoretischen Grundlagen der drei für diese Arbeit entwickelten Hauptroutinen wieder. Wie oben beschrieben, verwendet AUREMOL einen *Top-Down*-Ansatz, bei dem es u. a. wünschenswert ist, eine bereits gut definierte Struktur als Basis zu verwenden. Aus diesem Grund wurde ein neuer Ansatz entwickelt, der es erlaubt, mit Hilfe homologer Proteine geeignete Startstrukturen zu erzeugen.
Um bereits bestehende Strukturen zu verbessern, wurde der ISIC-Algorithmus entwickelt, der Informationen aus verschiedenen Quellen nutzt und konsistente Daten in einer wohl definierten Weise so verbindet, dass keine Verzerrungen hin zu falschen Strukturen auftreten können. Dies ist besonders sinnvoll, da in den PDB-Datenbanken eine Vielzahl von Strukturen mit mittlerer Qualität enthalten ist und zusätzlich auf verschiedene Arten erstellt wurde. Durch Kombination der Daten der verschiedenen Strukturaufklärungsmethoden wird so ein Maximum an Information erzeugt, die zu einer neuen Strukturbestimmung herangezogen wird.
Das Problem zur Bestimmung der sequentiellen Zuordnung wird im dritten Hauptmodul mit ASSIGN behandelt. Ziel des Moduls ist die sequentielle Zuordnung vollautomatisch zu bestimmen. Dazu dient ein Vergleich eines simulierten NOESY-Spektrums mit einem experimentell gemessenen NOESY-Spektrum, bei dem nicht nur die chemischen Verschiebungen der einzelnen Atome Verwendung finden. Die oft vernachlässigte Linienform und die Volumina der Signale sind entscheidende Faktoren beim Vergleich. Hinzu kommt die Nutzung statistischer Vorhersagen für die chemischen Verschiebungen. Mit Hilfe heuristischer Optimierungsverfahren werden die chemischen Verschiebungen der einzelnen Atome aus der Simulation so lange verändert, bis die wahrscheinlichste Übereinstimmung zwischen Simulation und Experiment erreicht ist.
In Kapitel 4 werden die Ergebnisse der neu entwickelten Module vorgestellt. Bei der Homologie-Modellierung wird anhand von HPr *aureus* aus *Staphylococcus* gezeigt, dass der Ansatz zum Erfolg führt und die gewünschten Ergebnisse liefert. Weiterhin wird gezeigt, dass diese Methode verwendet werden kann, um vorhandene Strukturbündel zu optimieren.
Die Ergebnisse von ISIC zeigen, wie eine NMR-Struktur mittlerer Qualität (Ras Bindedomäne Byr2) mit Hilfe seiner Röntgenstruktur, die ebenfalls eine mittlere Qualität hat und bei der sogar 13 Aminosäuren fehlen, verbessert werden kann. NMR-Strukturen, die mit wenig experimentellen Daten erstellt wurden und somit eine schlechte Qualität haben, können ebenfalls mit ISIC durch Zuhilfenahme von weiteren Quellen verbessert werden. Ein Beispiel mit der Immunoglobulin-Binde-Domäne zeigt auf, dass bei dem verwendeten Ansatz keine

Gefahr besteht, die zu verbessernde Struktur zu verschlechtern, selbst wenn die zur Verbesserung herangezogene Struktur eigentlich ungeeignet ist.

Anhand eines idealen künstlichen Datensatzes wird gezeigt, dass der Ansatz, der bei ASSIGN Verwendung findet, zum Erfolg führt. Durch Einführung von Rauschen nähert man sich experimentellen Zuständen an. Schließlich wird mit experimentellen Datensätzen der Punkmutante H15A von HPr *S. aureus* gezeigt, dass ASSIGN auch unter realen Bedingungen verwendet werden kann.

Die anschließende Diskussion in Kapitel 5 zeigt auf, in welchem Stadium sich die neu entwickelten Module befinden und in welchen Richtungen diese weiterentwickelt werden können. In Kapitel 6 ist eine Zusammenfassung der vorliegenden Arbeit gegeben.

2 Material und Methoden

2.1 Wichtige Module in AUREMOL

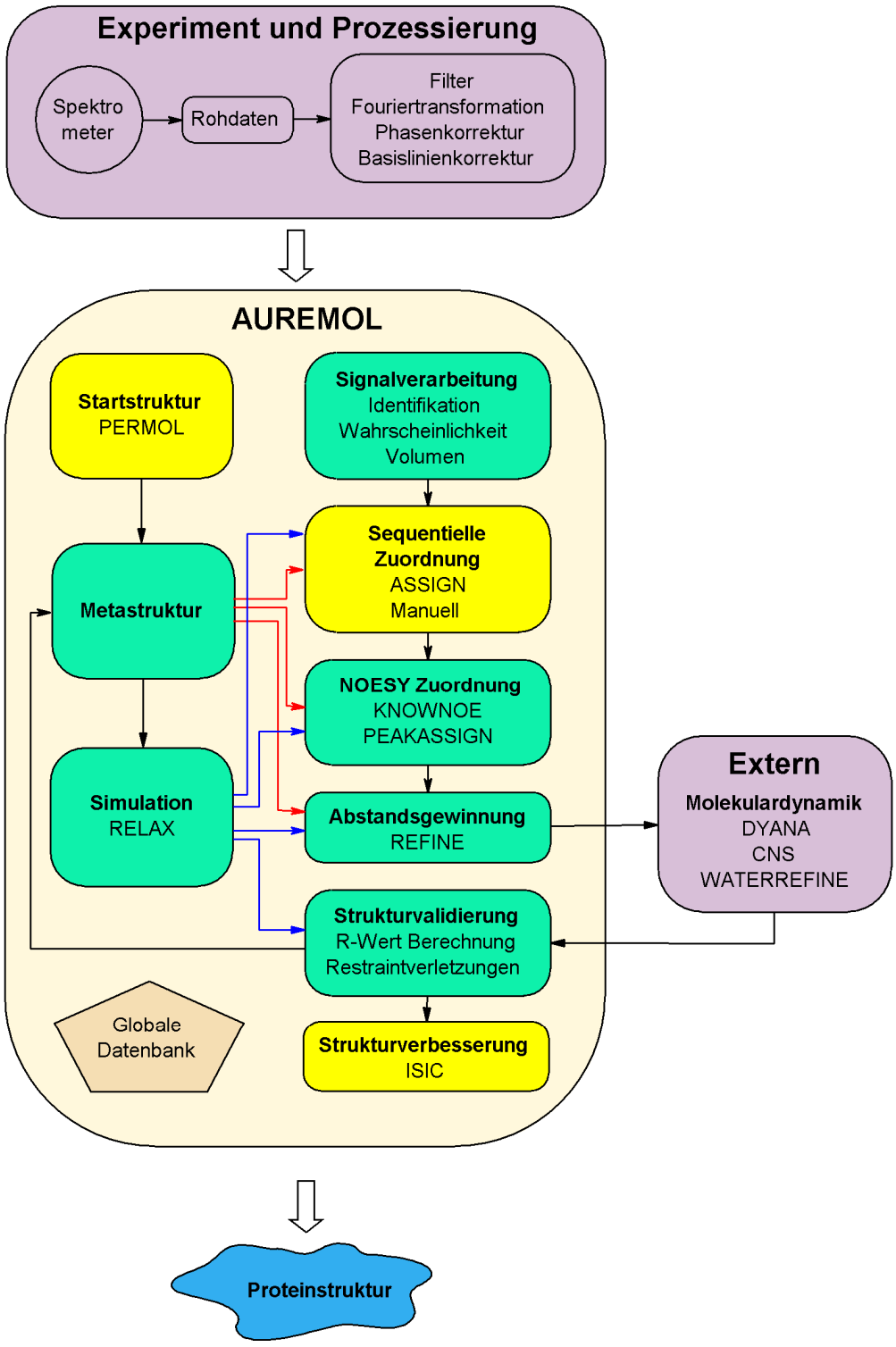

Abbildung 8: Wichtige Module in AUREMOL und ihre Beziehungen untereinander. Die gelb hervorgehobenen Module sind im Rahmen dieser Arbeit entstanden.

Abbildung 8 zeigt die wichtigsten Module von AUREMOL und veranschaulicht ihr Zusammenwirken. Die globale Datenbank ist in [14] beschrieben. Die gelb eingefärbten Module sind im Zuge dieser Arbeit entstanden und werden weiter unten ausführlich beschrieben. Die grün hinterlegten Module waren bereits vorhanden. Da auf diese Module oft zurückgegriffen wird, werden diese im nächsten Abschnitt kurz beschrieben.

2.1.1 Signalidentifikation und Signalwahrscheinlichkeit

Nachdem NMR-Spektren aufgenommen und prozessiert wurden, besteht der erste Schritt, sowohl bei der manuellen als auch bei der automatischen Auswertung, in der Identifizierung der Signale. Dabei wird die Intensitätsverteilung des Spektrums ausgewertet und alle lokalen Maxima bzw. Minima, herausgesucht. Man nennt diese Maxima (Minima) *Peaks* und den beschriebenen Prozess *Peakpicking*. Die *Peaks* werden also gekennzeichnet und in einer *Peakliste* gespeichert. In AURMOL gibt es zwei Verfahren zum *Peakpicking*. Das erste Verfahren beruht in der Identifizierung aller lokalen Maxima, die oberhalb eines bestimmten Grenzwertes liegen. Der Grenzwert, der auch als *Threshold* bezeichnet wird, kann dabei vom Benutzer frei eingestellt werden. Und genau das ist auch der große Nachteil. Ist der Grenzwert zu gering, erhält man sehr viele Signale, bei denen der Anteil der der *Rauschpeaks* ziemlich hoch ist. Wählt man den Grenzwert zu hoch, gehen viele echte Signale verloren. Aus diesem Grund wurde in AUREMOL erst kürzlich ein adaptives *Peakpicking* eingeführt [15]. Dieses Verfahren bestimmt automatisch anhand von lokalen Rauschniveaus [16] einen lokalen Grenzwert. Zusätzlich wurde die zuvor manuelle bayessche Berechnung der Signalwahrscheinlichkeiten [17;18] automatisiert angeschlossen. Dadurch erhält jeder *Peak* einen Wahrscheinlichkeitswert für seine Gültigkeit. Vorteile des Verfahrens bestehen darin, dass schwache Signale nicht automatisch ausgeschlossen werden, starke Wassersignale nicht von vornherein dabei sind und Symmetriebedingungen sowie Linienform mit berücksichtigt werden. Dies führt zu einer deutlich besseren Signalidentifikation in den Spektren.

Neben automatischem *Peakpicking* besteht in AUREMOL natürlich auch die Möglichkeit, *Peaks* manuell zu picken oder zu löschen. Ebenso ist es möglich, Gebiete in einem Spektrum zu definieren, in denen entweder alle *Peaks* innerhalb oder alle *Peaks* außerhalb entfernt werden sollen. Schließlich hat der Benutzer noch die Möglichkeit, *Peaks*, die unterhalb eines bestimmten Wahrscheinlichkeitsgrenzwertes, liegen zu eliminieren.

2.1.2 Signalvolumen

Nach dem *Peakpicking* schließt sich eine Volumenberechnung mit Hilfe einer Signalintegration des identifizierten *Peaks* an. Das *Peakvolumen* wird durch iterative Segmentierung des betrachteten *Peaks* berechnet, wobei die Segmentierung jeweils auf dem höchsten Punkt des Signals beginnt und sich rekursiv nach unten fortsetzt bis Datenpunkte benachbarter Signale oder der Segmentierungsschwellwert erreicht werden. Dieser Schwellwert, gemessen in Prozent vom *Peakmaximum*, muss vom Benutzer vorgegeben werden. Zusätzlich muss die maximale Integrationsbreite in Hz für alle Frequenzdimensionen angegeben werden, was sinnvoller Weise die maximal zu erwartende experimentelle Linienbreite (auf halber Höhe) ist [19]. Überdies ist in AUREMOL eine Volumenfehlerberechnung integriert, welche den Volumenfehler für jeden *Peak* auf der Basis von lokalen Rauschlevel-Abschätzungen und Signalüberlappungen berechnet [15].

2.1.3 NOESY Zuordnung

Die Zuordnung der NOESY-Spektren ist ein sehr wichtiger Schritt auf dem Weg der Strukturbestimmung. In AUREMOL sind hierzu zwei verschiedene Methoden implementiert. Für beide Methoden ist eine sequentielle Zuordnung notwendig. PEAKASSIGN [20] (früherer Name des Moduls: AUNOAS) wird verwendet, wenn die Struktur bereits gut bekannt ist und die einzelnen Signale aufgrund von verschiedenen experimentellen Bedingungen verschoben sind. Mit KNOWNOE [21] kann man in einem iterativen Prozess Zuordnungen aus schlechteren Strukturen bis hin zum ausgestreckten Strang bestimmen.

PEAKASSIGN

Ziel dieses Moduls ist, zuerst eine gegebene sequentielle Zuordnung an ein neues Spektrum anzupassen und dann möglichst viele NOE-Signale mit Hilfe dieser sequentiellen Zuordnung und mit Hilfe von Strukturinformationen zu benennen. In der Regel decken sich die chemischen Verschiebungen der verschiedenen Atome aus der sequentiellen Zuordnung nicht ganz genau mit den in den NOESY-Spektren gefundenen chemischen Verschiebungen. Das kommt daher, dass die chemischen Verschiebungen eines Atoms zusätzlich u. a. von Parametern wie Temperatur, Druck, pH-Wert und Salzgehalt der Probe abhängen und diese Parameter bei jeder Spektrenaufnahme variieren können. In manchen Experimenten will man ja gerade diese Parameter gezielt verändern, um das Proteinverhalten unter den veränderten Bedingungen zu studieren.

Im Prinzip funktioniert der Algorithmus nach folgendem Schema: Ausgehend von sicheren Zuordnungen isolierter *Peaks* wird versucht, die noch unsicheren Zuordnungen zu erhalten. Dieser Prozess wird iterativ auf sich selbst rückgekoppelt ausgeführt. In einem ersten Schritt werden isoliert liegende Signale mit nur einer eindeutigen Zuordnungsmöglichkeit benannt, im zweiten Schritt Signale, die in mindestens einer Koordinate eindeutig sind, und im letzten Schritt werden alle noch übrigen Signale mittels eines Qualitätsbeurteilungskriteriums überprüft und gegebenenfalls benannt. Dabei werden nicht nur chemische Verschiebungen genutzt, sondern auch Volumeninformationen, die aus der Simulation von Spektren mittels RELAX (siehe unten) mit geeigneten Strukturen stammen.

KNOWNOE

Die automatische Zuordnung mit KNOWNOE gliedert sich in zwei Teile. Zuerst wird eine Zuordnung von NOE-Signalen nur mit Hilfe der chemischen Verschiebungen ermittelt. Danach schließt sich im Falle von Mehrdeutigkeiten die Berechnung der wahrscheinlichsten Zuordnung an. Zunächst wird das experimentelle NOESY-Spektrum automatisch vorverarbeitet. Mit den oben beschriebenen Modulen werden *Peaks* gepickt und die Wahrscheinlichkeiten und Volumen derselben berechnet. Die Zuordnung der NOE-Signale anhand der sequentiellen Zuordnung führt aufgrund der Entartung von chemischen Verschiebungen meist zu Mehrdeutigkeiten. Deswegen wurden aus der Analyse von 326 Proteinstrukturen statistische Tabellen erzeugt, die Wahrscheinlichkeitsverteilungen der Volumina von Atompaaren repräsentieren. Für ein NOE-Signal, das N verschiedene Zuordnungen A_i besitzt, kann so die bedingte Wahrscheinlichkeit $P\left(A_i, a \middle| V_0\right)$ berechnet werden, dass die Zuordnung A_i mindestens $a \cdot V_0$ des Volumens V_0 erklärt. Ist für eine Zuordnung A_i die Wahrscheinlichkeit $P\left(A_i, a \middle| V_0\right) \geq P_{\min}$, wird diese als eindeutig erklärt.

Weiterhin werden Informationen aus wechselseitigen Beziehungen zu anderen sicheren NOE-Signalen verwendet, ähnlich wie beim *Network Anchoring* [22]. Mit einer Liste von eindeutig zugeordneten Signalen wird danach ein Satz von Strukturen berechnet, die als Input für weitere Iterationsschritte verwendet werden. Im Gegensatz zu PEAKASSIGN benötigt KNOWNOE nicht notwendigerweise eine bekannte 3D-Struktur.

2.1.4 Abstandsberechnung mit REFINE [15]

Sind die NOE-Spektren richtig benannt, ist der nächste Schritt, daraus Abstandsinformationen für die Strukturberechnung zu gewinnen. In der Arbeitsgruppe wurde hierzu REFINE entwickelt, das diese Aufgabe mit Hilfe von Spektrensimulation durch die oben beschriebene Rückrechung und den Volumenintegralen der zugeordneten NOE-*Peaks* bewerkstelligt. Benötigt werden ein Strukturbündel oder eine einzelne Struktur, die zugehörigen Simulationsparameter für RELAX und die chemischen Verschiebungen der Signale. Zuerst werden auf Basis des Relaxationsmatrixformalismus die NOE-Volumen der Eingabestruktur(en) berechnet, die dann für einen Vergleich mit den experimentellen Volumen herangezogen werden. In einem iterativen Prozess werden für alle Signale, bei denen keine gute Übereinstimmung aufgrund dieses Vergleiches herrscht, die Kreuzrelaxationsraten angepasst und die NOE-Volumen neu simuliert. Dies geschieht solange bis eine genügend hohe Übereinstimmung zwischen Simulation und Experiment erzielt ist. Danach werden die Abstandsinformationen und zugehörige Fehler aus den Relaxationsraten berechnet und als *Restraint*-Dateien für Moleküldynamik-Simulationen geschrieben. Für den Fall, dass während der Iteration für ein Signal keine befriedigende Übereinstimmung erzielt wurde, wird der Abstand einfach aus dem experimentellen Volumen mittels einer einfachen *Isolated Spin Pair Approach* (ISPA)-Näherung berechnet. Die Fehler werden für jeden Peak explizit aus einer Volumenfehlerabschätzung basierend auf lokalen Rauschlevels und Überlappfehler bestimmt. Für weitere Details sei auf [15] verwiesen.

2.1.5 Spektrensimulation mit RELAX [23]

Bei vielen Modulen in AUREMOL (Abbildung 8) wird ein simuliertes Spektrum benötigt, um über die Eingabestruktur Rückschlüsse auf das experimentelle Spektrum zu ziehen. In analoger Weise wie die Abstandsinformationen des NOESY-Spektrums die 3D-Struktur des Proteins festlegen, bestimmt umgekehrt die Proteinstruktur die Abstandsmatrix der Atome. Es lässt sich vorhersagen, welches NOESY-Spektrum man für eine gegebene Proteinstruktur mit bekannten chemischen Verschiebungen erwartet. Für derartige Spektrensimulationen wurde in unserer Arbeitsgruppe das Programm RELAX entwickelt. RELAX basiert auf der Relaxationsmatrixanalyse, die im Gegensatz zum ISPA den Magnetisierungstransfer nicht nur zwischen zwei als isoliert betrachteten Kernen betrachtet, sondern Spindiffusionprozesse zwischen allen Protonen des Systems berücksichtigt. Man erhält ein Spektrum, das mit einem fouriertransformierten, prozessierten experimentellen Spektrum korrespondiert und

qualitative Vergleiche erlaubt. Absolute Vergleiche der Volumen der Signale sind von vornherein nicht möglich, da das Maximum der Startmagnetisierungen aller Spins in der Simulation auf 1 gesetzt wird, wohingegen bei experimentellen Daten die Signalintensitäten von mehreren Faktoren abhängen, allen voran von der Probenkonzentration. In RELAX ist seit der Arbeit von Ried et al. [24] die Berechnung der Linienform, die hauptsächlich von der transversalen Relaxation abhängt, und die Aufspaltung der Signale durch J-Kopplung hinzugekommen. Dies ist insbesondere für diese Arbeit wichtig, da das später beschriebene Modul ASSIGN eben genau diese Linienformen zu einer automatischen sequentiellen Zuordnung von NOESY-Spektren nutzt. Für eine ausführliche Darstellung und Funktionsweise von RELAX sei auf [23] und [24] verwiesen.

2.1.6 Daten Konversion

In dieser Arbeit sind viele neue Werkzeuge entstanden, die das Arbeiten mit den vielen vorhandenen Datenformaten erleichtern.

PDB-Konverter

Zu allererst wurde sich in AUREMOL auf das Datenformat der IUPAC verständigt. D. h. alle Routinen innerhalb AUREMOL laufen mit IUPAC-konformen Atomnamen ab. AUREMOL ist aber auch auf externen Input angewiesen. Dazu gehören vor allem PDB-Dateien, die die räumliche 3D-Struktur von Proteinen beschreiben. Wenn man diese Dateien genauer betrachtet, stellt man fest, dass viele, manchmal sogar inkonsistente Formate für Atomnamen existieren. Hinzu kommt, dass oftmals die Definitionen für das Format der PDB-Dateien von *Brookhaven* [25] nicht beachtet worden sind. Hier ist klar definiert, in welcher Spalte welcher Wert stehen muss. All dies kann in konkreten Fällen zu Problemen führen. Deshalb wurde ein PDB-Konverter entwickelt, der diese Probleme in den meisten Fällen löst. Der PDB-Konverter kann eine oder mehrere Dateien auf einmal von einem definierten Format in ein Zielformat konvertieren. Weiß man das Ausgangsformat nicht genau, erlaubt es die Option *unknown* die richtigen IUPAC-Atomnamen zu finden. Weiterhin können mit diesem Tool terminale Wasser- und Sauerstoffatome gelöscht werden, da AUREMOL diese Atome noch nicht verwendet. Es kann ein Offset zur Aminosäurensequenz addiert oder subtrahiert werden und nach Bedarf kann eine neue Nummerierung der Sequenz erstellt werden. In den Fällen, in denen die *Brookhaven* Spaltendefinitionen verletzt sind, ist es möglich, die Datei getrennt

durch Leerzeichen einzulesen. Der PDB Konverter hat sich im praktischen Einsatz sehr bewährt, da viel Zeit auf die Präparation der PDB-Dateien eingespart werden konnte.

***Restraint*-Datei-Konverter**

Ein weiterer Konverter ist der *Restraint*-Datei-Konverter. Mit ihm können Atomnamen in *Restraint*-Dateien, die bei Moleküldynamik-Programmen benötigt werden, konvertiert werden. Hauptsächlich unterstützt der Konverter die weiter unten beschriebenen Moleküldynamik Programme DYANA [26] und CNS [27]. Dabei ist das globale Format der Dateien gleich, da DYANA *Restraint*-Dateien im CNS Format einlesen kann. Es ist lediglich eine Anpassung der Atomnamen nötig. Wie beim PDB-Konverter können hier Quell- und Zielformat frei vom Benutzer eingestellt werden.

Masterlist Konverter

In AUREMOL werden die *Peaklisten* in einer so genannten *Masterliste* gespeichert. Die Benennung der *Peaks* sollte, damit AUREMOL die Listen verarbeiten kann, im IUPAC-Format erfolgen. Sollten dennoch Benennungen in einem anderen Format vorliegen, wie es z. B. bei einer manuellen Zuordnung möglich ist, so können diese mit diesem *Tool* ins IUPAC-Format umbenannt werden.

r-Datei Konverter

Falls ein prozessiertes Spektrum vorliegt, das nicht im Bruker r-Dateiformat vorliegt, wurde im Rahmen dieser Arbeit ein Konverter entwickelt, der ein beliebiges binär vorliegendes Dateiformat in eine Bruker r-Datei (1r, 2rr, 3rrr, …) konvertiert, da AUREMOL auf diese Bruker Dateien angewiesen ist. Dabei können die Quelldateien auch als mehrere Dateien vorliegen, wie dies z. B. bei NMRPipe [28] der Fall sein kann.

Von der Quelle muss einiges bekannt sein. Die *Headerlänge* und der Datentyp sowie das verwendete Bitformat, wie auch die Dimension der einzelnen Eingabedatei(en) des Spektrums dienen als Startparameter. Zu den Zieleinstellungen gehören wiederum der Datentyp, das Bitformat sowie Dateipfad und Dateiname. Um eine konvertierte r-Datei anzeigen zu können, müssen auch noch einige Prozessierungsparameter bekannt sein. Die Anzahl der Datenpunkte, der gewünschte Bruker spezifische XDIM-Parameter, der die Art der Datenablage in der Datei zu Optimierungszwecken beschreibt, die Resonanzfrequenzen, Spektrenbreiten und der Offset müssen in der Eingabemaske eingegeben werden.

2.2 Berechnung von Proteinstrukturen

Ab initio Proteinfaltungssimulationen können bis heute aufgrund des hohen Rechenaufwandes trotz der heute zur Verfügung stehenden Computerkapazitäten nur ansatzweise simuliert werden [29-35]. Um den enorm hohen Rechenaufwand aufgrund der vielen möglichen Konformationen zu verringern, nutzt man experimentell gewonnene Informationen für Einschränkungen des Konformationsraums. Derartige Informationen sind u. a. interatomare Abstände, Diederwinkel, Wasserstoffbrücken und Orientierungen.

Im Prinzip gibt es zwei verschiedene Verfahren, die in Computerprogrammen zur Berechnung von Proteinstrukturen zum Einsatz kommen. Das eine ist die Methode der Distanzgeometrie, bei der aus allen experimentell gewonnenen Einschränkungen und aus der kovalenten Struktur Abstandsgrenzenmatrizen für die Atompaare erstellt werden. Diese Matrizen beinhalten einen Satz von Abständen in einem N-dimensionalen Abstandsraum, der dann in den Raum eines dreidimensionalen kartesischen Koordinatensystems projiziert wird. In ihm sind dann alle Atomkoordinaten bestimmt und somit auch die Proteinstruktur.

Die andere Methode ist die Simulation der Moleküldynamik. Hier wird von vornherein in den Raumkoordinaten (kartesische Koordinaten oder Torsionswinkel) gerechnet. Es werden alle experimentell gewonnenen Informationen in Form von Pseudopotentialen in ein empirisches Kraftfeld des Moleküls eingebaut. Durch das Lösen der Newtonschen Bewegungsgleichungen aller Atome versucht man die Trajektorie der Atome, die zur Faltung führt, zu simulieren. Um den Prozess weiter zu beschleunigen, werden den Atomen durch „Erhitzung“ höhere Energien verliehen, um den Konfigurationsraum schneller abzutasten und die Gefahr zu verhindern, in eventuellen Energienebenminimas hängen zu bleiben. Danach wird das System langsam abgekühlt und in einen stabilen, energiearmen Zustand übergeführt.

Da man nicht sicher davon ausgehen kann, dass wirklich immer das globale Energieminimum gefunden wird, werden bei derartigen Strukturrechenprogrammen immer viele (500-1000) Strukturen berechnet und anhand einer Zielfunktion, die die Energie der einzelnen Strukturen beschreibt, selektiert.

In dieser Arbeit wurde hauptsächlich DYANA 1.5 [26], CNS [27] und als speziellen Fall für Optimierungen im expliziten Lösungsmittel (in dieser Arbeit Wasser) NMR-WATERREFINEMENT [36] als Anwendung von XPLOR-NIH [37] verwendet, die jetzt kurz vorgestellt werden sollen.

2.2.1 DYANA

Das Programm DYANA (**DY**namics **A**lgorithm for **N**MR **A**pplications) [26] wurde 1997 von Güntert et al. entwickelt. Es benutzt ein *Simulated-Annealing*-Protokoll und rechnet vollständig im Torsionswinkelraum, d. h. es sucht nach numerischen Lösungen für die Bewegungsgleichungen der klassischen Mechanik mit Torsionswinkel als generalisierte Koordinaten. Die Zielfunktion ist die potentielle Energie des Systems, welches in einem „Temperaturbad“ langsam von seiner am Anfang hohen Temperatur abgekühlt wird und so lokale Energieminima überspringen kann, um zum globalen Minimum zu finden. Der große Vorteil von DYANA ist, dass die Rechnungen im Torsionswinkelraum ablaufen und so die kovalenten Strukturparameter wie Bindungslängen, Bindungswinkel, Chirialität und Planarität immer auf ihren optimalen Werten festgehalten werden. Daraus resultiert eine einfachere Energiefunktion für die potentielle Energie als bei konventionellen Moleküldynamikprogrammen, die im kartesischen Raum rechnen. Folge davon ist, dass der Algorithmus effizienter ist, weil er längere Zeitschritte für die numerische Integration der Bewegungsgleichungen erlaubt.

In DYANA wird das Molekül als eine Art Baumstruktur verstanden, die aus einem starren, im Raum fixierten, Basiskörper (N-terminus) und aus weiteren n starren Körpern besteht, die mit untereinander über n rotierenden Bindungen verbunden sind und mit der Seitenketten des C-Terminus aufhört. Die einzigen Freiheitsgrade sind die Torsionswinkel. Jeder der starren Körper besteht aus einem oder mehreren Massenpunkten (Atome), deren relative Positionen nicht variabel sind. Die Konformation ist eindeutig festgelegt durch die Werte aller Torsionswinkel.

Die verwendete potentielle Energie $V\left(V \geq 0\right)$ ist genau dann 0, wenn alle experimentellen Distanz-*Restraints* und Torsionswinkel-*Restraints* unverletzt sind und alle ungebundenen Atompaare keine sterischen Überlappungen aufweisen. Eine genaue Definition ist in [26] zu finden. Die kinetische Energie wird rekursiv aus Winkelgeschwindigkeiten, linearen Geschwindigkeiten bezüglich eines Referenzpunktes und dem Trägheitstensor berechnet. Die Bewegungsgleichungen sind die Lagrange-Gleichungen

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_k}\right) - \frac{\partial L}{\partial \theta_k} = 0 \quad (k = 1,...n), \tag{2.1}$$

wobei $L = E_{kin} - E_{pot}$ die Differenz der kinetischen und potentiellen Energie, θ die Torsionswinkel und k die einzelnen starren Körper bezeichnet. Die Lösung dieser Bewegungsgleichungen führt dann zur Proteinstruktur.

Für die im Verlauf der Arbeit durchgeführten Strukturberechnungen wurde das „Standard DYANA *Simulated-Annealing*-Protokoll“ der DYANA Version 1.5 benutzt.
Zur Rechnung benötigt DYANA die Sequenz des Proteins und Distanz- und/oder Diederwinkel-*Restraints*.

2.2.2 CNS

Die Software Suite CNS 1.1 (**C**rystallography & **N**MR **S**ystem) [27] wurde 1998 von Brünger et al. veröffentlicht. Es wurde entwickelt zur Strukturbestimmung mittels Röntgenkristallographie oder NMR. Die Architektur ist hochflexibel, da es auch Schnittstellen zu anderen Strukturbestimmungsmethoden wie z. B. zu Elektronenmikroskopie oder zur Festkörper-NMR bereitstellt. Der Benutzer hat zahlreiche Möglichkeiten mittels Steuerdateien in die Strukturrechnung einzugreifen. Speziell für die Strukturbestimmung aus NMR-Daten ist das Rechenprotokoll in vier Hauptabschnitte unterteilt: Dateneingabe, *Annealing-Protokoll*, Akzeptanz-Test und die Analyse aller NMR-Strukturen. Die Dateneingabe umfasst aus NOEs abgeleitete Atomabstände, NOE-Intensitäten, Torsionswinkel, Kopplungskonstanten, chemische Verschiebungen von ^{1}H, $^{13}C_{\alpha}$, $^{13}C_{\beta}$ Atome, dipolare Kopplung und heteronukleare T_1/T_2 Verhältnisse. Zu Beginn einer Simulation wird eine Anfangsstruktur, die aus allen vorhandenen Informationen über die Molekularstruktur, wie Bindungslängen und Bindungswinkel, generiert. Das kann sowohl ein ausgestreckter Strang (*extended strand*) oder eine bereits gefaltete Kette sein. Ausgehend davon simuliert das Programm die thermische Dynamik der einzelnen Atome und Atomgruppen unter Berücksichtigung der experimentellen Einschränkungen, indem die Newtonschen Bewegungsgleichungen integriert werden. Vor Beginn des ersten Simulationsschrittes werden allen Atomen der Startstruktur zufällig gewählte Anfangsgeschwindigkeiten aus einer der Temperatur entsprechenden Maxwellverteilung zugeordnet. Während der Simulation werden in bestimmten Zeitintervallen, die im Femtosekunden-Bereich liegen, die Positionen der Atome berechnet. Dabei gehen die Startpositionen und die mittleren, durch die Maxwellverteilung bei einer bestimmten Temperatur gegebenen, Geschwindigkeiten ein.

$$F_i(t) = m_i a_i(t) = -grad_i E\left(r(t)\right) \tag{2.2}$$

$F_i(t)$ bezeichnet die Kraft auf Teilchen i zum Zeitpunkt t, m_i die Masse dieses Teilchens und a_i seine Beschleunigung. Hinzu kommen die Beschleunigungen, welche man aus den Newtonschen Bewegungsgleichungen für die angenommenen Kraftfelder ("force fields")

erhält, unter deren Einfluss sich die Struktur entwickelt. Das typische molekulare Kraftfeld oder effektive Potential eines solchen Systems lässt sich in empirische und effektive Energieterme trennen [38]:

$$E = E_{emp} + E_{eff} = \sum \left(E_{Bindung} + E_{Winkel} + E_{Diederwinkel} + E_{vdW} + E_{elek} \right) + \sum \left(E_{NOE} + E_{Torsion} \right) \quad (2.3)$$

E_{emp} beinhaltet die ganze Information über die Primärstruktur des Proteins, sowie die Daten über die Topologie und Bindungen des Proteins allgemein. Die Beiträge der kovalenten Bindungen, der Bindungswinkel und der Diederwinkel bezüglich E_{emp} werden näherungsweise durch eine harmonische Funktion beschrieben. Im Gegensatz dazu werden die nichtkovalenten *van-der-Waals*-Kräfte und die elektrostatischen Wechselwirkungen durch ein Lennard-Jones- oder Coulomb-Potential simuliert. E_{eff} berücksichtigt die experimentell gefundenen Abstände. Für Distanz und Winkelbeschränkungen verwendet CNS empirische Funktionen, welche wie physikalische Potentiale wirken. Damit werden die Beiträge der unterschiedlichen Wechselwirkungen modelliert. Für die NOE-Einschränkungen wird ein *soft-square* Potential herangezogen, dessen flacher Bereich durch die vorgegebenen Distanzen und deren Fehlergrenzen bestimmt ist. Zu großen Entfernungen hin wächst das Potential nur linear an, damit einzelne unerfüllbare Einschränkungen (z. B. aus fehlerhaften Zuordnungen) nicht dominant werden.

Für die im Verlauf der Arbeit durchgeführten Strukturberechnungen wurden ein Standard-Annealing-Protokoll der CNS- Version 1.1, sowie die Standardwerte für alle Kraftkonstanten verwendet.

2.2.3 Strukturberechnung im expliziten Lösungsmittel

Die oben beschriebenen Strukturrechenprogramme benutzen aus Geschwindigkeitsgründen oft vereinfachte Annahmen. Interaktionen ungebundener Atome und eine unrealistische Behandlung der Elektrostatik und *van-der-Waals*-Bindungen führen oft zu nichtoptimalen Packungseffekten und unvollständigen Wasserstoffbrücken. Abhilfe schaffen hier Kraftfelder, die elektrostatische und Lennard-Jones-Potentiale berücksichtigen. Ebenso wirkt sich die Berücksichtigung der Wechselwirkung mit einem expliziten Lösungsmittel positiv aus.

Das hier in dieser Arbeit verwendete CNS-Protokoll zur Verbesserung von Strukturen im expliziten Lösungsmittel [36] wurde entwickelt, um die Strukturen den physikalischen

Gegebenheiten näher zu bringen. Das in der vorliegenden Arbeit verwendete Lösungsmittel ist Wasser, um möglichst physiologische Strukturen in nativer Umgebung zu erhalten und das Protokoll hierzu heißt *Water Refinement*. Um die zu verbessernden Strukturen wird dabei eine Wasserbox gelegt und mit einem S*imulated-Annealing*-Verfahren energieminimiert. Das Protokoll beinhaltet mehrere Tools zur Strukturvalidierung, die eine Analyse der verbesserten Strukturen erleichtert.

2.3 Validierung von Strukturen

Hat man Proteinstrukturen bestimmt, will man wissen, wie gut diese das Experiment beschreiben oder wie gut die geometrischen Eigenschaften mit der Theorie und den Erfahrungswerten zusammenpassen. Im Anschluss finden sich die in dieser Arbeit benutzten Strukturvalidierungsmethoden.

2.3.1 R-Wert Berechnung

Um in AUREMOL die Qualität von den berechneten Strukturen zu bestimmen, wurde in dieser Arbeitsgruppe in Analogie zur Röntgenkristallographie eine R-Wert Berechnung entwickelt [39]. Die R-Werte geben Aufschluss darüber, wie gut die jeweilige Strukturvorstellung mit den tatsächlichen experimentellen Daten übereinstimmt und werden sowohl für die ganze Struktur (global) als auch für unterschiedliche Distanzklassen berechnet. Der R-Wert ist definiert als normierte Standardabweichung zwischen berechneten und experimentellen Strukturparametern

$$R = \sqrt{\frac{\sum\left(\left|F_{sim} - F_{ex}\right|\right)^2}{\sum\left|F_{ex}\right|^2}} \tag{2.4}$$

wobei F_{sim} und F_{ex} beliebige strukturabhängige Größen darstellen. Ist $R = 0$, liegt eine perfekte Übereinstimmung zwischen den simulierten und experimentellen Daten vor, andernfalls ist $R > 0$.

Die Berechnung der R-Werte besteht im Prinzip im Vergleich vom experimentellen und simulierten NOESY-Spektrum. Das simulierte Spektrum erhält man mittels RELAX aus der zu untersuchenden Struktur sowie der sequentiellen Zuordnung. Man erhält so eine Liste mit simulierten *Peaks* (B-Liste). Da experimentelle Spektren eine limitierte Qualität haben und

viele *Rauschpeaks* und Artefakte enthalten, ist es sinnvoll die Qualität eines *Kreuzpeaks* durch seine Wahrscheinlichkeit $p_{ex,i}$ auszudrücken und diese mit in die Berechnung einfließen zu lassen. Diese Wahrscheinlichkeitsberechnung geschieht in AUREMOL automatisch mit dem adaptiven *Peakpicking*. In einem ersten Schritt müssen die *Kreuzpeaks* des experimentellen NOESY-Spektrums mit Hilfe der sequentiellen Zuordnung zugeordnet werden. Dies kann mit PEAKASSIGN erledigt werden und es wird eine Liste zugeordneter *Peaks* (A-Liste) erhalten.

Stellt sich beim Vergleich heraus, dass für ein experimentelles Signal kein zugehöriges simuliertes vorhanden ist, kommt dieses in eine Liste für unzugeordnete *Peaks* (U-Liste). Nun wird versucht diese U-Liste weiter zu verringern, indem ein Gitteralgorithmus darauf anwendet wird, der mit Hilfe der sequentiellen Zuordnung und einem benutzerdefinierten Suchradius (bei 2D-Spektren z. B. 0,01 ppm) unzugeordnete Signale z. B. in den Randbereichen des Spektrums zuordnet und mit einer entsprechenden Wahrscheinlichkeit versieht. Signale, die auf Grund einer unvollständigen sequentiellen Zuordnung bei der Simulation nicht vorhanden sind und sich deshalb in der U-Liste befinden, werden mit Hilfe einer Simulation eines Spektrums mit *random coil shifts* versucht, zuzuordnen. Dieser Teil wurde im Rahmen dieser Arbeit entwickelt.

Als experimentelle Größen werden die Volumen der NOE-Signale verwendet. Würde man nun die reinen Volumen V zur R-Wert Berechnung heranziehen, würden die kurzreichweitigen Signale wegen der Beziehung $V \propto r^{-6}$ dominieren. Deswegen weicht man auf einen Distanzabhängigkeit aus, die mit V^{α}, $\alpha = -1/6$ hergestellt werden kann. Weiterhin ist eine Skalierung der experimentellen oder simulierten Daten notwendig, um experimentelle Parameter wie z.B. die Probenkonzentration, pH-Wert, etc. zu berücksichtigen. Dies geschieht mit der Maximum-Likelihood-Methode, mit der sich der Skalierungsfaktor sf_{α} zu

$$sf_{\alpha} = \frac{\sum_{i \in A} \left(V_{ex,i} \cdot V_{sim,i} \right)^{\alpha}}{\sum_{i \in A} V_{sim,i}^{2\alpha}} \tag{2.5}$$

ergibt. $V_{ex,i}$ bezeichnet das experimentelle und $V_{sim,i}$ das simulierte Volumen des zugeordneten Signals i. A ist die Menge aller zugeordneter Signale.

Als globaler R-Wert resultiert

$$R_5(\alpha) = \sqrt{\frac{\sum_{i\in A}\left(V_{ex,i}^{\alpha} - sf_{\alpha}\cdot V_{sim,i}^{\alpha}\right)^2 \cdot p_{ex,i}^2 + \sum_{i\in U}\left(V_{ex,i}^{\alpha} - sf_{\alpha}\cdot V_{noise,i}^{\alpha}\right)^2 \cdot p_{ex,i}^2}{\sum_{i\in A} V_{ex,i}^{2\alpha}\cdot p_{ex,i}^2 + \sum_{i\in U}\left(V_{ex,i}^{\alpha} - sf_{\alpha}\cdot V_{noise,i}^{\alpha}\right)^2 \cdot p_{ex,i}^2}}, \qquad (2.6)$$

wobei $V_{ex,i}$ das experimentelle und $V_{sim,i}$ das simulierte Volumen beschreiben und sf_{α} für den oben beschriebenen Skalierungsfaktor steht. $V_{noise,i}$ wird im Falle der unzugeordneten Signale als Vergleichsvolumen eines typischen Rauschsignals verwendet. $p_{ex,i}$ steht für die Wahrscheinlichkeit, dass ein experimenteller *Peak* ein echtes Signal oder Rauschen ist. A ist die Menge aller zugeordneter, U die Menge aller unzugeordneter Signale.

Mit der R-Wert-Berechnung kann nicht nur die globale Übereinstimmung gemessen werden. Es ist möglich, bestimmte Regionen innerhalb des Proteins zu untersuchen. So können z. B. Sekundärstrukturelemente und einzelne Aminosäuren untersucht werden. Diese R-Werte können, je nachdem welche *Peaklisten* A, B und U man verwendet, berechnet werden. Für ein detaillierte Beschreibung sei auf [39] verwiesen.

2.3.2 Weitere Methoden zur Strukturvalidierung

RMSD

Der RMSD-Wert (**R**oot **M**ean **S**quare **D**eviation) wird verwendet, um mehrere Konformationen eines Moleküls an bestimmten Stellen, typischerweise an Atompositionen, zu vergleichen. Er gibt ein Maß für den Abstand zwischen den zu vergleichenden Atompositionen an. Es wird eine Superposition, die den RMSD minimiert, angestrebt.

Die Qualität eines Strukturbündels wird durch die mittlere Abweichung jeder einzelnen Struktur von der mittleren Struktur dieses Bündels beschrieben. Man spricht hier vom RMSD zum Mittelwert. Weiterhin kann ein paarweiser RMSD berechnet werden, bei dem der RMSD zwischen allen Strukturen berechnet und anschließend der Mittelwert gebildet wird.

Sind zwei Sätze von N Atompositionen $\boldsymbol{v}$ und $\boldsymbol{w}$ gegeben, errechnet sich der RMSD der beiden Strukturen zu

$$RMSD(\boldsymbol{v},\boldsymbol{w}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|v_i - w_i\right\|} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(v_{ix} - w_{ix}\right)^2 + \left(v_{iy} - w_{iy}\right)^2 + \left(v_{iz} - w_{iz}\right)^2}\,. \qquad (2.7)$$

Der RMSD-Wert eines Strukturbündels spiegelt dessen Präzision wieder und ist für die verschiedenen Bereiche einer Proteinstruktur unterschiedlich. Regionen mit beweglichen Strukturen oder ohne Sekundärstruktur (Loops) zeigen eine größere Abweichung als solche mit einer starren und gut definierten Sekundärstruktur. Strukturbündel mit einem RMSD des Protein-Rückgrades kleiner als 1Å kann man als gute Strukturen bezeichnen. Der RMSD-Wert kann z. B. mit dem Programm MOLMOL [40] berechnet werden.

Stereochemische Qualität

Zur stereochemischen Prüfung der Qualität von Proteinstrukturen wurde bereits 1993 das Softwarepaket PROCHECK-NMR [41] von Laskowski et al. vorgestellt. Das Paket besteht aus mehreren Einzelprogrammen, die die Eingabestrukturen aufbereiten, bewerten und schließlich in Form von *Postscript*-Dateien die Analyseergebnisse zusammenstellen. Wichtige Ausgaben des Programms sind der *Ramachandran-Plot* [42], der die Diederwinkel-Verteilung ($\Phi, \Psi, \chi, \ldots$) berechnet, und Abweichungen von Bindungslängen, Bindungswinkel und Planaritäten. Basis für diese Berechnungen sind statistische Auswertungen von Proteinstrukturen hoher Auflösung.

3 Theoretische Grundlagen

3.1 Homologie-Modellierung (PERMOL)

Üblicherweise sind experimentelle Daten die Grundlage für die Bestimmung der 3D-Tertiärstruktur von Proteinen auf atomarer Ebene. Es kommen hier hauptsächlich NMR-Untersuchungen und Röntgenkristallographie zum Einsatz. Doch haben beide Methoden auch Nachteile. Das Hauptproblem der Röntgenkristallographie liegt bei der oft schwierigen Protein-Kristallisation, während in der NMR-Spektroskopie Schwierigkeiten entstehen, wenn das Protein nicht löslich genug oder zu groß ist. Beide Methoden haben gemein, dass sie relativ komplex und deswegen sehr zeitintensiv sind. Was liegt also näher als über schnellere Methoden der 3D-Strukturgewinnung nachzudenken. Eine entscheidende Entdeckung war, dass Proteine, die eine ähnliche Primärsequenz besitzen, sich auch in der räumlichen Faltung ähneln [43-47]. Diese Tatsache gilt bereits ab einer Sequenzidentität von etwa 25 %. Nach der SCOP-Klassifikation [48;49] zeigen die 18946 gespeicherten Proteine in der Proteindatenbank nur 765 verschiedene Faltungstypen. Dies ist eine wichtige Grundlage für die *Human Proteomics Initiative* [50]. Mit den etablierten experimentellen Methoden soll aus jeder Strukturfamilie eine repräsentative Tertiärstruktur ermittelt werden und dann als Basis zur Modellierung aller sequenzverwandten Moleküle benutzt werden [51].

In der Vergangenheit sind bereits viele verschiedene Ansätze zur Homologie-Modellierung vorgestellt worden. Diese reichen von stark interaktiven Methoden (manuelle Modellierung) bis hin zu voll automatischen Methoden [51;52].

Bereits bestehende Softwarepakete zur Homologie-Modellierung sind u. a. SWISS-MODEL [53] und MODELLER [54]. Das Paket SWISS-MODEL wird von der Universität Lausanne im Internet als ein Werkzeug zur automatischen Homologie-Modellierung angeboten. Hierbei geht das Programm wie folgt vor: Nach der Auswahl sequenzhomologer Proteine mit bereits bekannter Struktur werden diese nach Generierung eines multiplen *Alignments* übereinander gelagert. Aus diesen übereinander gelagerten Strukturen werden gemittelte Koordinaten für die einzelnen Atome errechnet, woraus dann eine Peptidkette ermittelt wird. Fehlende Bereiche werden mit Hilfe von Daten aus einer Pentamere enthaltende Koordinatenbibliothek rekonstruiert, womit dann das gesamte Proteinrückgrat vervollständigt werden kann. Im nächsten Schritt werden die entsprechenden Seitenketten eingebaut und dann anhand ihrer chemischen Umgebung möglichst optimal eingepasst. Danach wird noch einmal die gesamte ermittelte Struktur durch einen Vergleich der Seitenkettenumgebung mit einer

aminosäurespezifischen Rotamer-Datenbank und einer Analyse der Packung des Proteins verifiziert.

Das Paket MODELLER [54] respektive MODWEB (Internetversion) wird von der Gruppe von Andrej Šali an der Universität von Kalifornien in San Francisco entwickelt und zur Verfügung gestellt. Aus einer an die PDB.org angelehnten internen Datenbank wird ein *Sequenzalignment* zur zu modellierenden Struktur generiert. Aus den Daten der aus der Datenbank ausgewählten Strukturen werden dann Beschränkungen für intramolekulare Distanzen sowie Diederwinkel erzeugt. Als Ausgabe generiert MODELLER eine dreidimensionale Struktur des Proteins, welche die Beschränkungen so gut wie möglich erfüllt. Die Optimierung der Struktur erfolgt über eine *Simulated-Annealing*-Methode, die eine Zielfunktion minimiert.

Ab initio Moleküldynamik-Simulationen, bei denen neben dem verwendeten Kraftfeld nur die Primärsequenz des Proteins als Information dient, führen nur bei kleinen Molekülen zum Erfolg [29-35;55]. Moleküldynamik-Simulationen basierend auf *Simulated-Annealing*-Protokollen, bei denen der Konformationsraum durch experimentell gewonnene Beschränkungen (*Restraints*) eingeschränkt wird, werden dagegen schon lange und sehr erfolgreich in der NMR-Strukturbestimmung eingesetzt. In dieser Arbeit wird ein neuer Ansatz vorgestellt, der Informationen aus bereits vorhandenen homologen 3D-Strukturen gewinnt. Das in AUREMOL implementierte Modul extrahiert aus PDB-Dateien, die die 3D-Struktur der Proteine enthalten, intramolekulare Distanzen und Diederwinkel vom Rückgrat und von den Seitenketten. Zusätzlich werden Wasserstoffbrücken identifiziert. Die so gewonnen Daten werden als *Restraints* für Moleküldynamik-Programme wie DYANA [26], CNS, usw. eingesetzt.

Das Modul Homologie-Modellierung (PERMOL) setzt auf den gleichnamigen Vorläufer PERMOL [56] auf und wurde bereits in einer Diplomarbeit von Josef Scheiber [57] in AUREMOL implementiert. In dieser Arbeit wurde das Projekt übernommen, erweitert und fertig gestellt.

3.1.1 Homologe Proteine

Die erste Aufgabe für den vorgestellten Ansatzes ist das Auffinden homologer Proteine, deren Struktur bereits bekannt ist. Die Quelle hierfür ist die PDB-Datenbank [58]. Mit Hilfe heuristischer Algorithmen, wie z. B. FASTA [59], BLAST [60] und PSI-BLAST [61], ist es möglich, diese Datenbank auf zu einer Zielsequenz ähnliche Sequenzen zu untersuchen. Die

Algorithmen liefern i. a. Score-Werte, die eine Bewertung der Übereinstimmung der Sequenzen im weitesten Sinne anzeigen. Unter den nun vorhandenen potentiellen Modellen müssen nun die am besten geeigneten ausgewählt werden. Dabei kann man sich im einfachsten Fall an den *Score*-Werten orientieren [52] oder aber man lässt eine nachgewiesene evolutionäre Verwandtschaft zweier Organismen in die Auswahl mit einfließen. Den besten Hinweis auf ein gutes Modell bekommt man dann, wenn zusätzliche Moleküle die gleichen Bindungspartner aufweisen und die gleiche Reaktion katalysieren. Die Qualität der Modellstrukturen kann ebenfalls von entscheidender Bedeutung sein. So ist bei mittels Röntgenkristallographie ermittelten Strukturen die Auflösung ein entscheidender Faktor und bei NMR-Strukturen der RMSD-Wert sowie die Anzahl der *Restraints* pro Aminosäure.

3.1.2 Sequenzalignment

Hat man sich für Modellstrukturen entschieden, ist die Erstellung eines guten *Alignments* der nächste Schritt. Zwar wurde auch schon bei der Suche nach den Modellstrukturen ein *Alignment* erstellt, doch waren die verwendeten Algorithmen auf Geschwindigkeit optimiert. Um gute Modellierungsergebnisse zu erhalten, ist ein bestmöglichstes *Alignment* notwendig. PERMOL [56] hat dabei auf das externe Programme ClustalX [62] zurückgegriffen. In AUREMOL ist zu diesem Zweck ein neues Modul, das ein paarweises globales *Alignment* ausführt, implementiert worden. Dabei werden die Sequenzen über ihre gesamte Länge ausgerichtet. Es werden gegebenenfalls Leerstellen (gaps) eingefügt, so dass beide Sequenzen möglichst gut übereinstimmen. Danach werden die durch die Leerstellen erweiterten Sequenzen übereinander gelegt, so dass eine Korrespondenz zwischen den Residuen und Leerstellen der einen Sequenz und den Zeichen und Leerstellen der anderen Sequenz entsteht. Zusätzlich wird vereinbart, dass keine Leerstelle in der einen Sequenz mit einer Leerstelle in der anderen Sequenz korrespondieren darf und Leerstellen sowohl am Anfang als auch am Ende einer Sequenz eingefügt werden können. Die Anzahl der Möglichkeiten eines globalen *Alignments* bei einer Sequenzlänge n ist gegeben durch

$$\binom{2n}{n} = \frac{(2n)!}{(n!^2)} \sim \frac{2^{2n}}{\sqrt{\pi \cdot n}}. \tag{3.1}$$

Das entspricht bei einer Sequenzlänge von $n = 100$ Aminosäuren ca. $9 \cdot 10^{58}$ Möglichkeiten. Ein systematisches Aufzählen ist somit nicht vertretbar. Probleme dieser Art werden mittels

dynamischer Programmierung – ein algorithmisches Verfahren zum Lösen von Optimierungsproblemen – gelöst. Speziell im Falle des Sequenzvergleichs wurde die dynamische Programmierung von Saul Needleman und Christian Wunsch [63] in die Bioinformatik eingebracht. Das Prinzip hinter den dynamischen Algorithmen zur Auffindung der besten Lösung ist die Aufspaltung des eigentlichen Problems in kleinere, leichter lösbare Teilprobleme. Die Teillösungen werden in der richtigen Reihenfolge und mit einer Bewertung in einer Tabelle abgespeichert. Nun muss nur noch die richtige Abfolge der Teilprobleme gewählt werden, die nach der Addition der Einzelbewertung die höchste Gesamtpunktzahl ergibt.

Beim Sequenzvergleich ist es notwendig, jede einzelne Aminosäure einer Sequenz mit der Aminosäure der anderen Sequenz zu paaren. Dabei stellt sich die Frage, ob man die beiden Reste paart oder eine Lücke in eine der Sequenzen einfügt. Formell ausgedrückt ergibt sich: Welche Möglichkeiten gibt es, ein angefangenes *Alignment* der Sequenz s bis zum Zeichen $i-1\left(s[1..i-1]\right)$, mit der Sequenz t bis zum Zeichen $j-1\left(t[1..j-1]\right)$ fortzusetzen? Die drei Möglichkeiten sind:

1. Ordne dem $t[j]$ eine Lücke zu
2. Ordne dem $s[i]$ das $t[j]$ zu
3. Ordne dem $s[i]$ eine Lücke zu.

Weitere Möglichkeiten gibt es nicht, da die Zuordnung einer Lücke der einen Sequenz zu einer Lücke der anderen Sequenz ausgeschlossen wurde.

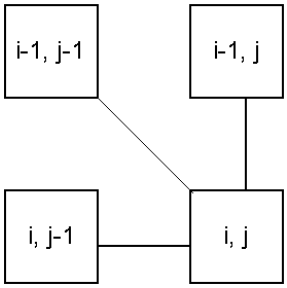

Abbildung 9: Die drei Möglichkeiten beim *Sequenzalignment*:

1, Paare $s[1..i]$ mit $t[1..j-1] \Rightarrow$ Lücke bei $t[j]$

2, Paare $s[1..i-1]$ mit $t[1..j-1] \Rightarrow$ Übereinstimmung $s[i]$ mit $t[j]$

3, Paare $s[1..i-1]$ mit $t[1..j] \Rightarrow$ Lücke bei $s[i]$

Die *Score*funktion sc ergibt sich zu

$$sc(s[1..i], t[1..j]) = \max \begin{cases} sc(s[1..i], t[1..j-1]) + gp \\ sc(s[1..i-1], t[1..j-1]) + m(i, j) + mm \\ sc(s[1..i-1], t[1..j]) + gp \end{cases} \tag{3.2}$$

wobei gp (*gap penalty*) eine Strafe für das Einführen einer Lücke angibt und einen negativen Wert hat. Der Wert m (*match award*) steht für eine Belohnung für eine Übereinstimmung und mm (*mismatch penalty*) für eine Strafe für Nichtübereinstimmung.

Als Beispiel sei ein globales Alignment von `s=AAAC` mit `t=AGC` gezeigt.

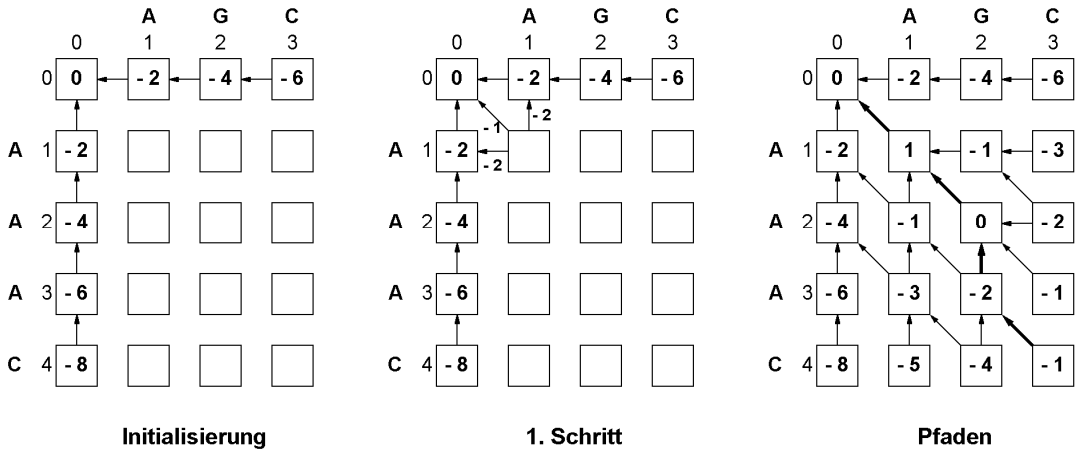

Abbildung 10: *Sequenzalignment* nach Needleman-Wünsch mit den Parametern $gp = -2$ (*gap penalty*), $m(i, j) = +1$, falls $s[i] = t[j]$ (*match*), $mm = -1$, falls $s[i] \neq t[j]$ (*mismatch*). Nach der Initialsierung werden die Matrixfelder schrittweise aufgefüllt. Durch das Pfaden (dicke Pfeile) zum Schluss wird das optimale Alignment erhalten.

Die erste Spalte und die erste Zeile werden mit Vielfachen der $gp = -2$ gefüllt. Danach werden schrittweise die restlichen Matrixfelder von links nach rechts und von oben nach unten gefüllt. Für jedes Feld wird sich der bestmöglichste Wert gemerkt und aus welchen Richtungen dieser Wert Zustande gekommen ist. Am Ende kann im Feld [4,3] die Ähnlichkeit der beiden Sequenzen abgelesen werden (= -1). Zum Schluss wird durch das Pfaden das optimale Alignment der beiden Sequenzen bestimmt. Dabei wird schrittweise wieder zurückgegangen, um die Alignments, die diesen Ähnlichkeitswert ergeben, von rechts nach links aufzubauen. Die fettgedruckten Pfeile geben das optimale Alignment an. Die Lösung ist AAAC,AG-C.

Dem *match award* kommt noch eine besondere Rolle zu. Wie oben ersichtlich, handelt es sich um eine Funktion der beiden zu paarenden Aminosäuren. Die Funktion gibt die Wahrscheinlichkeit an, dass zwei Aminosäuren ausgetauscht werden. Es wurde beobachtet, dass im Laufe der Evolution bestimmte Aminosäuren häufig gegen andere ausgetauscht werden. Solche Austausche lassen die Proteine funktionsfähig und sind somit kompatibel mit der Funktion und der räumlichen Struktur der Proteine. Besonders häufig sind Austausche von Aminosäuren, die auch chemisch ähnliche Eigenschaften haben: hydrophobe durch hydrophobe, große durch große, polare durch polare, etc. Austausche von chemisch unterschiedlichen Aminosäuren sind dagegen selten.

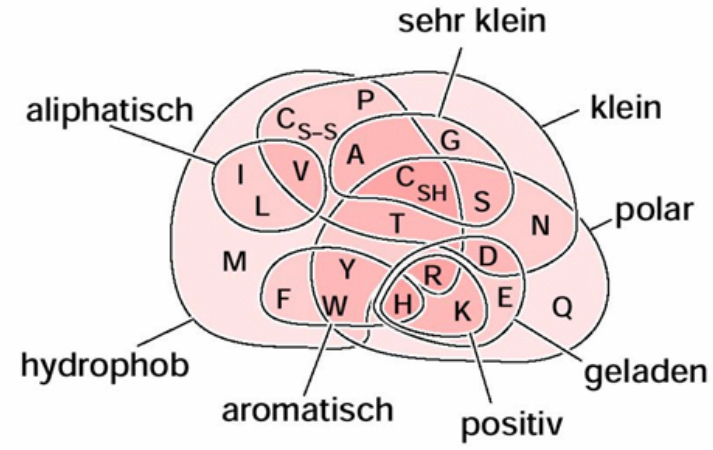

Abbildung 11: Einteilung der Aminosäuren in ihre Eigenschaften nach [64].

Diese Wahrscheinlichkeiten stehen in Substitutionsmatrizen. Die Werte des Wahrscheinlichkeitsmaßes (keine echten Wahrscheinlichkeiten) werden als Logarithmus des Verhältnisses zweier Wahrscheinlichkeiten berechnet. Dabei entspricht der Nenner dieses Quotienten der Wahrscheinlichkeit des zufälligen Auftretens eines Aminosäurepaares im

Alignment, ist also das Produkt der unabhängigen Häufigkeiten jeder einzelnen Aminosäure. Der Wert im Zähler gibt die Wahrscheinlichkeit des nicht-zufälligen gemeinsamen Auftretens beider Reste zusammen an. Diese Wahrscheinlichkeiten beruhen auf Erfahrungswerten aus einer großen Zahl von tatsächlichen *Sequenzalignments*, deren Gültigkeit zuvor überprüft wurde. Der Logarithmus wird positiv, wenn die Bedeutung für ein biologisches bedeutungsvolles Zusammentreffen überwiegt, und negativ, wenn das gemeinsame Auftreten durch Zufall wahrscheinlicher ist. AUREMOL benutzt die BLOSUM (**Blo**ck **Su**bstitution **M**atrix) [65], genauer die BLOSUM62 (Abbildung 12). Die Daten für die Matrizen stammen aus der Blocks-Datenbank [65], die lückenlose *Alignments* von ausgewählten Sequenzregionen aus verschiedenen Proteinfamilien enthält. Die Zahl 62 bei BLOSUM62 bedeutet, dass nur Sequenzen, die untereinander höchstens 62 % Sequenzidentität aufwiesen, für die Berechnung der Wahrscheinlichkeitswerte herangezogen wurden. Je kleiner die Zahl in der BLOSUM-Matrix, desto besser ist die Matrix geeignet um weiter entfernte Verwandtschaftsbeziehungen zu untersuchen. Die BLOSUM62-Matrix hat sich für den Bereich des Sequenzvergleichs als die Optimale erwiesen und wird überall zu diesem Zweck standardmäßig eingesetzt [66;67].

Res	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	O	°
A	4																								
R	-1	5																							
N	-2	0	6																						
D	-2	-2	1	6																					
C	0	-3	-3	-3	9																				
Q	-1	1	0	0	-3	5																			
E	-1	0	0	2	-4	2	5																		
G	0	-2	0	-1	-3	-2	-2	6																	
H	-2	0	1	-1	-3	0	0	-2	8																
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4															
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4														
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5													
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5												
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6											
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7										
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4									
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5								
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11							
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7						
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4					
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4				
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4			
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1		
O	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	6	-1	-1	-4	-3	-2	-2	-1	-2	7	
°	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6	-6
#	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	O	°

Abbildung 12: BLOSUM62-Matrix, wie sie in AUREMOL verwendet wird.

3.1.3 Erzeugung der *Restraints*

Nachdem das *Sequenzalignment* durchgeführt wurde, werden nun Informationen, die in den PDB-Dateien der homologen Strukturen kodiert sind, gewonnen. Die wichtigsten Informationen in den PDB-Dateien sind die Raumkoordinaten eines jeden Atoms des Proteins. Mit ihrer Hilfe können Abstands-, Diederwinkel- und Wasserstoffbrücken-*Restraints* bestimmt werden.

Zur Homologie-Modellierung werden n verschiedene Strukturfamilien mit jeweils m_i Strukturen herangezogen. Dies führt dazu, dass für einen zu bestimmendes *Restraint* mehrere Werte aus den einzelnen Modellen vorliegen und es können daraus Mittelwerte und zugehörige Fehlergrenzen berechnet werden. Diese Mittelwerte dienen bei den typischen Anwendungen der Moleküldynamik, die für diese Homologie-Modellierung eingesetzt werden, als Erwartungswerte für einen *Restraint* und die Fehlergrenzen spiegeln die zugehörigen oberen und unteren Grenzen wieder.

Eine weiterhin wichtige Information bei Proteinstrukturen, die mit Hilfe von Röntgenkristallographie bestimmt wurden, ist die Auflösung und der B-Faktor. Bei Röntgen-Strukturen liegt, im Gegensatz zu NMR-Strukturen, im Normalfall immer nur eine Struktur vor. Daraus kann kein Mittelwert mit entsprechendem Fehler berechnet werden. Die Auflösung und die B-Faktoren, die normalerweise bei Röntgen-Strukturen veröffentlicht sind, spiegeln die Unsicherheit der Atompositionen wieder, die in AUREMOL für die Fehlergrenzenberechnung herangezogen wird. Eine genaue Beschreibung ist im Kapitel 3.2 zu finden.

3.1.4 Distanz-*Restraints*

Distanz-*Restraints* werden für jedes Modell, das im *Alignment* enthalten ist, berechnet. Vorsicht ist geboten bei der Auswahl der Atome, da die Anzahl der Abstände schnell wächst:

$$anz = \underbrace{a \cdot \frac{(n-1)\cdot n}{2}}_{gleiche\ Atome} + \underbrace{\frac{(a-1)\cdot a}{2} \cdot n^2}_{verschiedene\ Atome} \tag{3.3}$$

Dabei ist *anz* die Anzahl der *Restraints*, die man erhält, wenn a der Anzahl der Atome und n der Sequenzlänge entspricht. Würde man 10 Atome wählen bei einer Sequenzlänge von 100 Aminosäuren, erhielte man bereits 500000 Distanz-*Restraints*, eine Menge, die jedes Moleküldynamikprogramm überfordert. Obendrein wäre dies auch nicht sinnvoll, da so kein Spielraum zur Energieminimierung vorhanden wäre.

Neben der sorgfältigen Auswahl der Atome ist eine Limitierung der Distanzen ratsam. So können Rückgratatomen und Seitenkettenatomen unterschiedliche Reichweiten zugeordnet werden und die Anzahl der *Restraints* kann über die Auswahl der Atome und Limitierung der Reichweiten sinnvoll eingeschränkt werden.

Die Distanzen eines jeden Modells j aus einer Strukturfamilie i werden nach Pythagoras unter zu Hilfenahme der Koordinaten $x_{ij,1}, y_{ij,1}, z_{ij,1}$ des einen Atoms und $x_{ij,2}, y_{ij,2}, z_{ij,2}$ des anderen Atoms mit

$$d_{ij} = \sqrt{(x_{ij,1} - x_{ij,2})^2 + (y_{ij,1} - y_{ij,2})^2 + (z_{ij,1} - z_{ij,2})^2} \tag{3.4}$$

berechnet.

Der Erwartungswert für den Abstand d ergibt sich zu

$$\langle d \rangle = \frac{1}{\sum_n m_i} \sum_{i=1}^{n} \sum_{j=1}^{m_i} d_{ij} \tag{3.5}$$

und die zugehörige Standardabweichung s zu

$$s = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \left(d_{ij} - \langle d \rangle \right)^2}{\sum_n m_i - 1}}. \tag{3.6}$$

Dabei steht n für die Anzahl der verschiedenen Strukturfamilien, die jeweils m_i Strukturen enthalten.

Um auch bei einer niedrigen Anzahl von Werten vernünftige Grenzen für die *Restraints* zu erhalten, werden diese über das Konfidenzintervall c berechnet. Das Konfidenzintervall gibt das Intervall an, das den wahren Mittelwert $\langle d \rangle$ mit einer bestimmten Wahrscheinlichkeit P einschließt und wird folgendermaßen berechnet

$$c = s \cdot \frac{t_\alpha}{\sqrt{\sum_n m_i}}, \tag{3.7}$$

wobei c die Konfidenzgrenze, s die Standardabweichung, t_α den t-Wert der t-Verteilung wiedergibt. α bezeichnet die Vertrauenswahrscheinlichkeit und kann frei gewählt werden. Typische Werte hierfür sind 0,95 bis 0,99. Je höher die Vertrauenswahrscheinlichkeit gewählt wird, desto breiter wird das Konfidenzintervall, um auch wirklich sicher zu gehen, dass der wahre Mittelwert eingeschlossen wird. Der t-Wert mit der zugehörigen Vertrauenswahrscheinlichkeit α kann dabei in einer Tabelle nachgeschlagen werden oder wie in AUREMOL implementiert mit Hilfe eines Näherungsverfahrens [68] berechnet werden. Dies hat zum Vorteil, dass α frei gewählt werden kann.

3.1.5 Diederwinkel-*Restraints*

Die Berechnung der Diederwinkel-*Restraints* erfolgt nach [69] unter Zuhilfenahme folgender Methode: Für die vier Atome A, B, C und D mit den Ortsvektoren $\vec{A}$, $\vec{B}$, $\vec{C}$ und $\vec{D}$ ist der Torsionswinkel Θ definiert als der Winkel zwischen der Projektion $\vec{x}$ und $\vec{y}$ der Vektoren $\overrightarrow{ba} = \vec{A} - \vec{B}$ und $\overrightarrow{cd} = \vec{C} - \vec{D}$ auf eine Ebene senkrecht zum Bindungsvektor $\vec{z} = \vec{C} - \vec{B}$. Das Vorzeichen von Θ ist positiv, wenn in $\vec{z}$-Richtung $\vec{x}$ in $\vec{y}$ durch eine Rechtsdrehung übergeht. Für den Winkel Θ gilt also mit $\vec{n} = \vec{x} \times \vec{y}$

$$\Theta = \arccos\left(\frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|}\right) \tag{3.8}$$

und

$$sign(\Theta) = sign(\vec{z} \cdot \vec{n}). \tag{3.9}$$

Die so erhaltenen Winkel decken einen Bereich von -180° bis +180° ab.

Da es sich bei Winkeln um zyklische Größen handelt und hierfür nicht eindeutig definiert ist, wie man Mittelwerte und Standardabweichungen berechnet, wird der Ansatz nach Döker et al. [70] gewählt, der sich ausgiebig mit dem Problem beschäftigte und sehr gute Werte liefert.

In AUREMOL besteht die Möglichkeit, die Diederwinkel der Haupt- und Seitenketten ψ, ϕ, ω, χ_1, χ_2, χ_{21}, χ_{22}, χ_3, χ_{31}, χ_{32}, χ_4, χ_5, χ_6 zu berechnen.

3.1.6 Wasserstoffbrücken-*Restraints*

Die Wasserstoffbrückenbindung ist die stärkste Form unter den zwischenmolekularen nichtkovalenten Wechselwirkungen, welche bei organischen Verbindungen sowohl intra- als

auch intermolekular auftreten kann. Die Wasserstoffbrückenbindung beschreibt im Allgemeinen die schwache Wechselwirkung zwischen einem Wasserstoffatom, welches kovalent an ein elektronegatives Element X gebunden ist, und einem elektronegativen Element Y, das ein freies Elektronenpaar besitzt. Zu den elektronegativen Elementen X gehören in Aminosäuren die Atome N, O und S, zu den elektronegativen Elementen Y die Atome O und N. Die X-H-Gruppe bezeichnet man in der Regel als Wasserstoffbrückendonor (D) und das Y-Atom als Wasserstoffbrückenakzeptor (A).

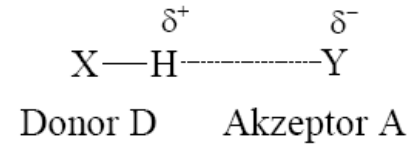

Abbildung 13: Die schematische Darstellung einer Wasserstoffbrückenbindung.

Bei der Wasserstoffbrückenbindung handelt es sich um eine mäßig schwache Bindung mit zumeist weniger als 10 % der Bindungsstärke einer gewöhnlichen kovalenten Bindung. Verglichen jedoch mit den *van-der-Waals*-Kräften zwischen Molekülen mit einer Energie von einigen wenigen kJ/mol, ist die Wasserstoffbrückenbindung eine relativ starke zwischenmolekulare Wechselwirkung. Um diesem Effekt Rechnung zu tragen, wurde eine automatische Wasserstoffbrückenerkennung hinzugefügt. Für Wasserstoffbrücken-*Restraints* wird jedes einzelne Modell untersucht. Zur Identifizierung dient dabei eine sehr schnelle Methode, die einen maximalen Donatoren-Akzeptoren-Abstand von 0,24 nm und einem Wasserstoffbrückenwinkel a_{DHA} von $180^o \pm 35^o$ verwendet. Die Donatoren (N (H^N), O^γ (H^γ), S^γ ($H^{\eta 11}$), $N^{\eta 1}$ ($H^{\eta 12}$), $N^{\eta 1}$ ($H^{\eta 22}$), $N^{\eta 2}$ ($H^{\eta 21}$), $N^{\eta 2}$ ($H^{\eta 22}$), N^ζ ($H^{\zeta 1}$), N^ζ ($H^{\zeta 2}$), N^ζ ($H^{\zeta 3}$), N^γ ($H^{\gamma 1}$)) und Akzeptoren (O, $O^{\delta 1}$, $O^{\delta 2}$, $O^{\varepsilon 2}$, N, $N^{\eta 1}$, $N^{\eta 2}$, $N^{\delta 2}$) können in AUREMOL frei gewählt werden. Ebenso kann durch den Benutzer die Entscheidung getroffen werden, in wie vielen Modellen die Bedingung für eine Wasserstoffbrücke erfüllt sein muss, um daraus dann Mittelwerte und Fehlergrenzen auszugeben. Mittelwerte und Fehlergrenzen werden dabei so behandelt wie Abstand-*Restraints*.

3.1.7 Strukturrechnung

Wenn nun Abstände, Diederwinkel und Wasserstoffbrücken als *Restraints* vorliegen, kann die Strukturrechnung gestartet werden. Die oben beschriebenen Standardprogramme hierfür sind

CNS [27] und DYANA [26], für die AUREMOL automatisch die richtigen Dateiformate bereitstellt.

3.2 Verbesserung von Proteinstrukturen (ISIC [71])

Das wichtigste Ziel bei der Strukturaufklärung eines biologischen Makromoleküls ist eine möglichst genaue und fehlerfreie Struktur schnell und mit möglichst wenig experimentellen Daten zu erhalten. Ein möglicher Weg die NMR-Strukturaufklärung zu beschleunigen, ist die Zahl der experimentellen *Restraints* zu reduzieren oder nur *Restraints* zu verwenden, die leicht gewonnen werden können wie z. B. Diederwinkel und chemische Verschiebungen von Rückgratatomen, Restdipolkopplungen, Wasserstoffbrücken oder H^N-H^N-NOEs. Wenn aber die experimentellen Daten reduziert sind, ist es empfehlenswert, zusätzliche Informationen zu nutzen. Diese können z. B. von homologen oder aus mit anderen Methoden bestimmten Proteinstrukturen stammen. In den letzten Jahren sind immer mehr Daten in den Proteindatenbanken abgelegt worden. Diese Daten können eine wichtige Rolle bei neuen Strukturbestimmungen einnehmen, wenn sie richtig genutzt werden. Die Schlüsselfrage hierbei ist, die vorhandenen Datenbankinformationen mit experimentellen Daten so zu kombinieren, dass nur relevante Daten Einfluss auf die Struktur haben und keine Verfälschungen auftreten. Aus diesem Grund wurde im Rahmen dieser Arbeit ein neuer vollautomatischer Algorithmus basierend auf dem Bayeschen Prinzip entwickelt, der es erlaubt Strukturinformationen aus verschiedenen Quellen konsistent zu kombinieren. Ziel ist es, qualitativ hochwertige Strukturen zu erhalten ohne dabei die Menge an experimentellen Daten zu erhöhen. Dieser neue ISIC- (*I*ntelligent *S*tructural *I*nformation *C*ombination) Algorithmus ist in AUREMOL eingebettet. Die Ergebnisse werden automatisch mit Hilfe experimenteller Daten verifiziert.

Die meisten bisher in der Literatur zu findenden Methoden zielen darauf ab, nur die globale Faltung zu bestimmen [31;72-79]. Im einfachsten Fall kann man abklären, ob eine Zielstruktur eine bereits bekannte Faltung annehmen kann [72]. Bowers et al. beschreiben eine Ansatz, der die ROSETTA *ab initio* Proteinstrukturvorhersage mit wenigen NMR-Daten kombiniert [73]. Die ROSETTA-Methode baut Proteinstrukturen aufgrund von Fragmenten bereits bekannter Strukturen auf, die eine gewisse Sequenzähnlichkeit in diesen Fragmenten zur Zielstruktur aufweisen [74;75]. Liegen genauere chemische Verschiebungen vor, kann man den TALOS-Algorithmus [80] dazu verwenden, um Diederwinkel des Rückgrats zu bestimmen. Zusammen mit NOE-Daten kann diese Information zusätzlich zur Fragmentauswahl dienen. Beim MFR-Ansatz von Delagio et al. [76] wird eine Startstruktur erzeugt, indem eine Datenbank mit 3D-Strukturen nach bestimmten Fragmenten durchsucht wird. Es werden solche Fragmente ausgewählt, deren vorhergesagten Restdipolkopplungen

am besten zu den gemessenen passen. Ebenso wird, aber mit geringerer Gewichtung, mit chemischen Verschiebungen verfahren. Einen ähnlichen Ansatz gibt es auch von Andrec et al. [77]. Haliloglu et al. [31] benutzen eine relativ geringe Anzahl von Restdipolkopplungen zusammen mit einem vereinfachten Proteinmodell, das nur die Schwerpunkte der Seitenketten beinhaltet, und erhielten damit Strukturen mittlerer Auflösung. In einem anderen Vorgehen konnte gezeigt werden, dass Faltungsvorhersagen mittels *Proteinthreading* verbessert werden können, wenn man experimentelle Abstandsbeschränkungen, die z. B. aus Massenspektroskopie oder NOE- Messungen gewonnen werden, mit berücksichtigt [78]. Von Li et al. [79] wurden einige wenige NOE-*Restraints* mit dem Strukturvorhersage-Algorithmus TOUCHSTONE kombiniert. Weitere Methoden nutzen gemeinsame Informationen aus verschiedenen Quellen für eine verbundene Strukturverbesserung um hochqualitative Strukturen zu erhalten. Der erste Testfall war Interleukin-1b [81], bei dem das so gewonnene Modell sowohl kristallographisch vergleichbare R-Werte aufwies als auch seine geometrische Qualität ähnlich der war, die man aus Röntgen-Daten allein erhält. Die wenigen verbliebenen NOE-Verletzungen zeigten dann die wirklichen Unterschiede zwischen Lösungsstrukturen und Kristallstrukturen auf. Eine ähnliche Vorgehensweise wurde auch bei der Verbesserung der Struktur von BPTI (Bovine Pancreatic Trypsin Inhibitor) [82] benutzt, bei der Strukturen erzeugt wurden, die sowohl den NMR-Daten wie auch den Röntgen-Daten Rechnung trugen. Die Struktur des Ribosomal Proteins L9 [83] wurde gelöst, indem NMR-*Restraints* von den 39 N-terminalen Resten, die von den Röntgen-Daten sehr schlecht beschrieben waren, verwendet wurden. Für die Struktur der Oligomerisation des p53-Tumor-Suppressor-Proteins [84] konnte gezeigt werden, dass das Modell sowohl mit den NMR-Daten also auch den Röntgen-Daten verträglich ist. Zudem wurde das Modell im Vergleich zur reinen Röntgen-Struktur stereochemisch verbessert. Während der Verbesserung des HU-Protein-Dimers konnten die Unterschiede zwischen den NMR- und Röntgen-Daten durch Kristallpackungseffekte und durch Spindiffusionseffekte erklärt werden. Bei dem Ansatz von Chao et al. [85] wird zuerst eine Struktur auf Röntgen-Basis erzeugt und dann mit aus NMR-Daten erhaltenen Abstands- und Diederwinkel-Restraints verfeinert. In den Fällen, wo die NMR und die Röntgen-Daten nicht übereinstimmen, wurden die NMR-Daten manuell entfernt. Dies führte zu Strukturen mit verbesserter Stereochemie und geringeren R-Werten. Alle diese Methoden haben gemeinsam, dass Diskrepanzen zwischen NMR und Röntgen manuell korrigiert werden müssen. Dies beinhaltet u. a. das Entfernen verletzter NOEs, Neuzuordnungen von NOEs oder Wasserstoffbrücken und die Notwendigkeit, Spindiffusionseffekte bei der NMR zu berücksichtigen.

Im Gegensatz dazu soll der in dieser Arbeit entwickelte Algorithmus hochqualitative Strukturen automatisch bestimmen. Mit ISIC (*I*ntelligent *S*tructural *I*nformation *C*ombination) wird ein neuer, allgemeiner und vollautomatischer Ansatz für die Kombination von Strukturinformationen aus verschiedenen Quellen vorgestellt. ISIC berücksichtigt dabei die unvermeidlichen Diskrepanzen der Eingabestrukturen und stellt sicher, dass sowohl die Eingabedaten richtig gewichtet werden als auch fehlerhafte Eingabedaten auf das Ergebnis keinen negativen Einfluss haben. Mit den so kombinierten Informationen werden neue hoch aufgelöste Strukturen berechnet und das Ergebnis wird automatisch an experimentellen Daten verifiziert. Ein mögliche Anwendung des ISIC-Algorithmus liegt beispielsweise darin, relativ schnell gewonnene Daten aus der Flüssigkeits-NMR wie Diederwinkel und chemische Verschiebungen vom Rückgrat, Restdipolkopplungen, Wasserstoffbrücken und H^N-H^N-NOEs zu verwenden um gering oder mittelmäßig aufgelöste Strukturen zu berechnen und diese dann mit Daten aus Homologie-Modellierung oder einer homologen Röntgen-Struktur zu ergänzen oder zu verbessern.

3.2.1 Überlegungen

Bei der Strukturverbesserung unter Verwendung von Informationen aus anderen Datenquellen müssen vor allem zwei Fälle unterschieden werden. Im ersten Fall beschreibt die zusätzliche Information dasselbe Strukturbündel wie z. B. eine Lösungsstruktur eines Proteins bei gegebenen pH, gegebener Temperatur und verschiedenen Probenzusammensetzungen. Hierbei ist es wichtig, die zusätzlichen Informationen richtig zu gewichten, um die wahre Struktur zu erhalten. Im anderen Fall geht man davon aus, dass die zusätzliche Information von ähnlichen Strukturen stammt, aber doch unterschiedlich ist, wie es zum Beispiel bei einer Lösungsstruktur und einer Kristallstruktur verschiedener Komplexe vorkommt. Hier tritt allerdings eine zusätzliche Schwierigkeit auf. Es muss entschieden werden, wie gut die Information aus der zusätzlichen Struktur zur fraglichen Struktur passt. Andernfalls erhält man eine durch falsche Daten verzerrte Struktur.

Formal beschrieben ergibt sich also das Ziel, die wahrscheinlichste Struktur oder das wahrscheinlichste Strukturbündel S_0 mit der bedingten Wahrscheinlichkeit $P\left(S_0 \middle| A, I_i, i=1,N\right)$ zu erhalten, die einen bestimmten Grenzwert P_t übersteigt. Die Kombination von Informationen von N verschiedenen Quellen I_i ist ein Problem, das sehr häufig in der Strukturbiologie auftritt. Dabei ist S_0 ein Strukturbündel, das mit Hilfe von

NMR gewonnen wurde, und A ist das allgemeine Wissen über das System als physikalisches Modell. Dies beinhaltet die kovalente Struktur und die sich gegenseitig beeinflussenden Potentiale, wie sie in der Moleküldynamik-Berechnung vorkommen. Die aus der NMR abgeleitete Information I_1 liegt normalerweise als Satz von experimentellen *Restraints* $R_1 = \left\{R_1^1, ..., R_1^M\right\}$ vor. Dieser Satz enthält M *Restraints*, welcher den Konformationsraum im Wesentlichen auf die wahrscheinlichsten Lösungen einschränkt. Die experimentellen *Restraints* sind von Natur aus inhomogen, weil sie Abstands-Informationen aus NOESY-Spektren, Diederwinkel aus J-Kopplungen oder chemischen Verschiebungen und intramolekulare *Restraints* aus Restdipolkopplung, die die Orientierung der Aminosäuren einschränken, umfassen.

Ein sehr eleganter Weg, die wahrscheinlichste Lösung zu finden, ist ein *Simulated-Annealing*-Protokoll [86], bei der die Information A bereits in die Moleküldynamik-Routinen eingebettet ist.

Im zweiten Fall (ähnliche Strukturen) wird die Situation viel komplexer, weil Strukturinformationen anderer Quellen herangezogen werden, die nicht exakt mit den Bedingungen des aktuellen Experiments übereinstimmen. Werden diese Informationen als Sätze von Beschränkungen R_i ausgedrückt, so müssen nun Strukturen S_0^p (p=1,..., L_0, mit L_0 gleich der Gesamtzahl der Strukturen im Bündel S_0) mit hoher Wahrscheinlichkeit $P\left(S_0^p \middle| A, R_i\, i = 1, ..., N\right)$ gefunden werden. Wird dabei eine auf Beschränkungen basierende Moleküldynamik mit *Simulated-Annealing-Protokoll* verwendet, ist das physikalische Modell wiederum bereits implizit, d. h. $P\left(S_0^p \middle| A, R_i\, i = 1, ..., N\right)$ kann durch $P\left(S_0^p \middle| R_i\, i = 1, ..., N\right)$ ersetzt werden. Mit Ausnahme der Menge an Beschränkungen R_1, die aus der Führungsstruktur (Struktur, die verbessert werden soll) S_1 stammt, stimmen die anfänglichen Beschränkungen $R_i^*\left(i = 2, ..., N\right)$, die aus den anderen Quellen abgeleitet werden, nicht unbedingt direkt mit den Bedingungen überein, mit denen das Führungsstrukturbündel gewonnen wurde. Das kann zum Beispiel schon auftreten, wenn die experimentellen Bedingungen unterschiedlich sind. Deswegen müssen aus diesen neue Beschränkungen R_i berechnet werden, die dann direkt mit dem wahren Strukturbündel S_0 übereinstimmen. D. h. $R_1 = R_1^*$, aber für $R_i^*\left(i = 2, ..., N\right)$ muss bestimmt werden, inwieweit ihre einzelnen *Restraints* mit der wahren Struktur S_0 übereinstimmen.

$$P\left(S_0 \left| R_i \;\; i=1,\; ...,\; N\right.\right) \;=\; P\left(S_i \left| R_1^* = R_1, R_i^*, i=2,...,N\right.\right) \tag{3.10}$$

Wie oben bereits erwähnt, muss die Menge der Beschränkungen R_i als mehrdimensionale Wahrscheinlichkeitsverteilung $P\left(R_i, i=1,...,N\right)$ beschrieben werden. Die verschiedenen Beschränkungsmengen und die einzelnen Beschränkungen selbst sind über das physikalische Modell gekoppelt, weil sie von verwandten Strukturen abgeleitet sind. Die Wahrscheinlichkeitsverteilung eines *Restraint*-Satzes R_i in den Führungsstrukturen wird aus dem bekannten R_i^* berechnet:

$$P\left(R_i\right) = P\left(R_i \left| R_i^*, i=1,...,N\right.\right) P\left(R_i^*, i=1,N\right) \tag{3.11}$$

Gleichung (3.11) zeigt, dass R_i von einer mehrdimensionalen Wahrscheinlichkeitsverteilung abhängig ist. Eine Vereinfachung des Problems ist zwingend erforderlich.

Beim Standardansatz des *Simulated-Annealings* wird angenommen, dass die einzelnen Beschränkungen R_i^k im Wesentlichen unabhängig sind. Erst der Algorithmus, der konsistente Lösungen sucht, bewirkt eine indirekte Kopplung. Solange die gleichen *Restraints* R_i^k berücksichtigt werden, kann man die Wahrscheinlichkeit für einen neu erzeugten *Restraint* R_0^k berechnen, der mit der wahren Lösungsstrukturen S_0 übereinstimmt und der einen gegeben Wert im Bündel S_0 hat. Die *Restraints* R_0^k werden dann später benutzt um das wahre Strukturbündel S_0 zu berechnen.

$$P\left(R_0^k\right) = P\left(R_0^k \left| R_i^{k*}, i=1,...,N\right.\right) P\left(R_i^{k*}, i=1,...N\right) \tag{3.12}$$

Der Index i steht für den verwendeten Datensatz und k für einen bestimmten *Restraint*. Dabei wird angenommen, dass in erster Näherung die einzelnen *Restraints* R_0^k und R_0^1 unabhängig sind für $k \neq 1$. Zur Berechnung von $P\left(R_0^k\right)$ ist es nützlich, wenn gleiche *Restraints* aus den verschiedenen Datensätzen vorliegen.
Gleichung (3.12) kann auf zwei verschiedene Weisen genutzt werden. Wenn eine gute Bestimmung der bedingten Wahrscheinlichkeit möglich ist, kann sie direkt genutzt werden.

Wenn nicht, kann die Hypothese, dass $P\left(R_0^k \middle| R_1^{k*}\right)$ nahe 1 ist, für einen Datensatz i getestet werden. Da angenommen wird, dass der experimentelle Datensatz 1 das wahre Ensemble repräsentiert, kann geprüft werden, ob die *Restraints* R_i^k und R_1^k zum selben Ensemble gehören. Alle *Restraints*, die diese Bedingung nicht erfüllen, werden verworfen. $P\left(R_i^{k*}, i=1,...,N\right)$ in Gleichung (3.12) beschreibt die Wahrscheinlichkeit, dass ein Ersatz-*Restraint* R_i^{k*} einen gegebenen Wert im Strukturbündel S_i hat. Diese Wahrscheinlichkeit ist natürlich abhängig von Faktoren wie den zugehörigen σ-Werten der *Restraints* im Strukturbündel S_i.

3.2.2 Implementierung des Algorithmus

In ISIC werden die Strukturinformationen der verschiedenen Quellen i bestehend aus den einzelnen Bündeln S_i mit $i=1,...,N$ und Anzahl der benutzen Quellen $N \geq 2$ verwendet um Strukturen des Bündels S_1, z. B. NMR-Strukturen, mit einer homologen Röntgen-Struktur S_2 zu verbessern. Bei diesem Vorgehen sind die verschiedenen strukturellen Quellen S_i nicht identisch, z. B. im Fall von Lösungs- und Kristallstrukturen. Hinzu kommen aber auch noch andere Unterschiede wie z. B. abweichende Aminosäurensequenzen oder das Fehlen bzw. Vorhandensein eines wechselwirkenden Moleküls. Ein wichtiges Konzept ist, die verfügbaren strukturellen Informationen verschiedener Quellen anfangs in ein dichtes Netzwerk von daraus abgeleiteten Ersatz-*Restraints* R_i^{k*} zu konvertieren, die dann direkt verglichen werden können (Gleichung (3.12)). Sie werden aus Strukturbündeln berechnet und liegen in Form von Diederwinkel-, Distanz- und Wasserstoffbrücken-*Restraints* vor. Die Erwartungswerte und Standardabweichungen s werden direkt vom gegebenen Bündel mit Hilfe des PERMOL-Algorithmus [56;87] berechnet. Wenn die Führungsstruktur S_1 aus NMR-Strukturen besteht, ist eine solches Bündel bereits vorhanden. Liegt kein Strukturbündel vor, muss dieses erst in wohl definierter Weise erzeugt werden (siehe unten). Die *Restraints* $R_1^{k*} = R_1^k \left(k=1,...,M\right)$ werden dann mit den *Restraints* $R_i^{k*}\left(i=2,...N,\ k=1,...,M_i,\ M_i \leq M\right)$ kombiniert um die finalen *Restraints* $R_0^k \left(k=1,...,M\right)$ zu erhalten, mit deren Hilfe dann ein neues Strukturbündel S_0 berechnet wird. Die Qualität des neuen Strukturbündels kann an den originalen experimentellen Daten validiert werden, um die Verbesserung der Struktur abzuschätzen. Dieser Schritt wird vom Algorithmus aber nicht zwingend benötigt.

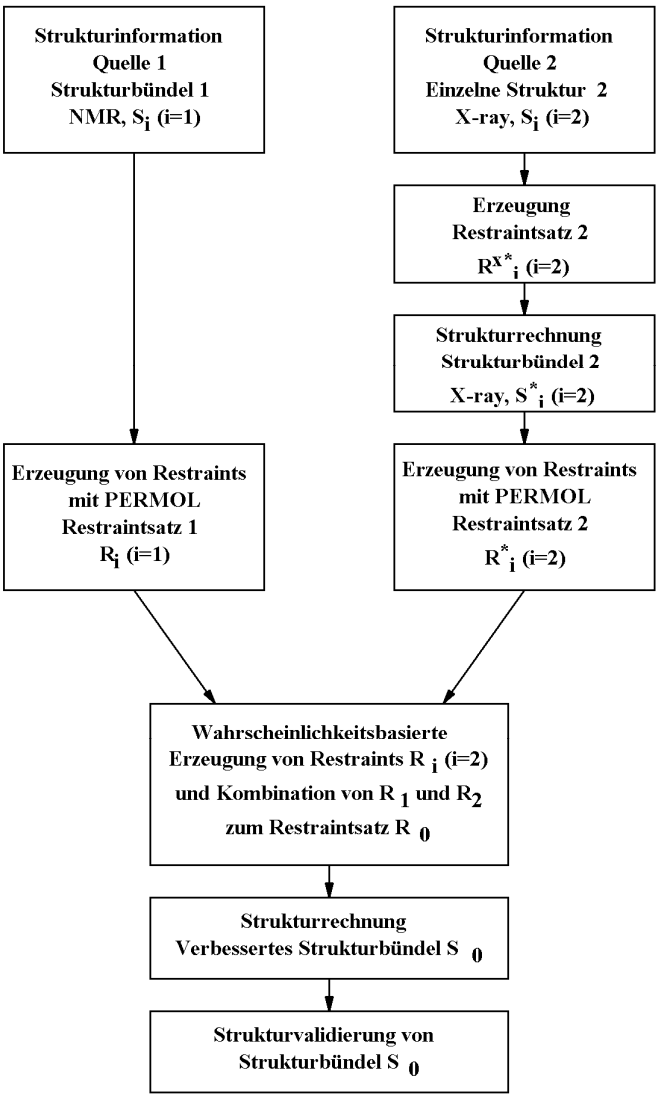

Abbildung 14: Schematische Darstellung des ISIC-Algorithmus. Im oberen Beispiel werden zwei Eingabequellen benutzt. Die eine Quelle repräsentiert das NMR-Strukturbündel S_1, die andere Quelle die einzelne Röntgen-Struktur (S_2).

3.2.3 Berechnung des Netzwerkes der Ersatz-*Restraints*

Die Berechnung eines dichten Netzwerkes von Diederwinkel- und Abstand-*Restraints* mit Hilfe des PERMOL-Algorithmus aus Strukturbündeln ist weiter oben (Kapitel 3.1) beschrieben worden. Im Wesentlichen werden hier Erwartungswerte und Standardabweichungen aus Strukturbündeln berechnet. Die Fehlerbereiche der einzelnen *Restraints* werden aus der Standardabweichung auf der Basis des *t*-Tests, genauer mit Bestimmung des Konfidenzintervalls, berechnet. Wenn der Original-Datensatz nur aus einer einzigen Struktur besteht, muss erst noch ein zugehöriges Bündel berechnet werden. Im

Folgenden wird nur der wichtigste Fall für Kristallstrukturen diskutiert, das Prinzip kann jedoch genau so auf andere Daten angewendet werden kann. Kristallstrukturen werden normalerweise von einer einzelnen Struktur $S_i^p \left(p = 1 \right)$ beschrieben. Manchmal kommt es auch vor, dass aufgrund verschiedener Kristallsymmetrien oder verschiedener *Refinement*-Methoden doch einige wenige Strukturen in der Datenbank $\left(p > 1 \right)$ vorhanden sind. Dennoch ist selbst dann das statistische Ensemble zu klein, um daraus verlässlich Restraints berechnen zu können. In Analogie zur Berechnung von NMR-Strukturen können systematische Koordinatenunsicherheiten dazu verwendet werden, Strukturbündel zu berechnen. Mit Hilfe von Atomabständen und Diederwinkeln aus der Kristallstruktur zusammen mit den Koordinatenunsicherheiten wird ein Satz von *Restraints* R_i^{X*} gewonnen, der die Original Röntgen-Struktur repräsentiert. Aus diesen *Restraints* werden dann die Strukturen für das Bündel S_i^X erzeugt und aus diesem dann mittels PERMOL die Ersatz-*Restraints* R_i^* abgeleitet. Um nun den Satz R_i^{X*} zu generieren, werden zwei Faktoren, die normalerweise mit der Kristall-Struktur veröffentlicht werden, herangezogen, die für eine konservative Abschätzung der strukturellen Streuung verwendet werden. Erstens ergibt sich in einer ersten Näherung der durchschnittliche erwartete Fehler in den einzelnen Atompositionen $\sigma\left(r_i \right)$ zu 1/3 der Auflösung R [88]. Bei einer umfassenderen Analyse wird $\sigma\left(r_i \right)$ bei Atomen mit einem kleinen B-Faktor oft mit Hilfe des Luzzati-Plots bestimmt. Zweitens wird der lokale B-Faktor herangezogen um zusätzliche individuelle Fehler für Atompositionen zu errechnen. Statistische und thermische Unordnung verbreitern die Elektronendichte eines Atoms at_j, was zu einer Erhöhung seines B-Faktors führt. Der B_j-Faktor steht mit dem rms (*root mean square*) Fehler der Position eines Atoms in Beziehung und ist gegeben durch

$$\sigma(U_j) = \sqrt{\frac{B_j}{8 \cdot \pi^2}} \tag{3.13}$$

B_j steht für den B-Faktor eines gegebenen Atoms at_j und $\sigma(U_j)$ ist der zugehörige durchschnittliche Fehler der Atomsposition. Da für unsere Berechnungen eine konservative Bestimmung von Distanzbereichen sehr nützlich ist, wird das Quadrat der Standardabweichung $\sigma^2\left(d_{jk} \right)$ des Abstandes d_{jk} zwischen zwei Atomen at_j und at_k angenähert durch

$$\sigma^2(d_{jk}) = \sigma(U_j)^2 + \sigma(U_k)^2 + 2\sigma(r_i)^2 \tag{3.14}$$

Eine tiefer gehende Beschreibung der Genauigkeit von Proteinstrukturen kann im Artikel von Cruickshank [89;89] nachgeschlagen werden. Sollte mehr als eine Struktur des gleichen Kristalls in der Datenbank vorliegen, können diese als getrennte Strukturen S_i betrachtet und analog behandelt werden. Wie oben bereits erwähnt, wird mit diesem vorläufigen *Restraint*-Satz R_i^{X*} ein Strukturbündel S_i^X mittels Molekül-Dynamik-Programmen wie DYANA [26], XPLOR-NIH [37] oder CNS [27] erzeugt. Aus diesem Bündel wird dann ein *Restraint*-Satz R_i^* auf die gleiche Weise wie der *Restraint*-Satz R_1 der Führungsstruktur S_1 gewonnen.

3.2.4 *Restraint* Kombination

Wie bereits oben hergeleitet (Gleichung (3.11), (3.12)) wird nun aus den *Restraint*-Sätzen $R_1\left(R_1 = R_1^*\right)$ und $R_i^*\left(i = 2,...,N\right)$ ein neuer *Restraint* Satz R_0 berechnet, der dann für die endgültige Strukturrechnung verwendet wird. Obwohl der Algorithmus für alle Datensätze die zum Führungs-*Restraint*-Satz R_1 passenden *Restraint*-Sätze R_i^* produziert, kommt es in einzelnen Fällen vor, dass kein zu R_1^k passender *Restraint* R_i^{k*} aus dem Datensatz i erzeugt werden kann. Dieser Fall tritt ein, wenn ein Atom oder gar eine Aminosäure des Datensatzes R_1 in den Daten, die aus dem Datensatz i gewonnen wurden, nicht existiert. In diesem Fall wird R_0^k gleich R_1^k gesetzt. In allen anderen Fällen wird der endgültige *Restraint* R_0^k nach Gleichung (3.12) berechnet. Weil aber $P\left(R_0^k \middle| R_i^{k*}, i>1\right)$ für Abstände und Winkel schwer zu bestimmen ist, wird ein paarweiser Null-Hypothesen-Test $P\left(R_1^k \middle| R_i^{k*}, i>1\right)$ angewandt, der bestimmt, ob zwei korrespondierenden *Restraints* der beiden Datensätze dasselbe Ensemble beschreiben. Ist die Bedingung erfüllt, wird eine neue Wahrscheinlichkeitsverteilung für den *Restraint* berechnet. Wenn nicht, wird der *Restraint* R_i^{k*} verworfen und nur R_1^k benutzt.

Wurden große Strukturbündel erstellt, kann die Wahrscheinlichkeitsverteilung direkt aus dem Bündel erhalten werden. Da kein a priori Wissen über den Verteilungstyp der einzelnen *Restraints* vorhanden ist, kann man bekannte statistische Tests wie den Rangdispersionstest von Siegel und Tukey [90] oder den Vergleich zweier unabhängiger Stichproben nach Kolmogoroff und Smirnoff [90] anwenden. Wenn die untersuchten *Restraints* die gleiche oder wenigstens nahezu die gleiche Verteilung haben, kann der so genannte U-Test nach

Wilcoxon, Mann und Whitney [90] verwendet werden. Das ist das verteilungsfreie Gegenstück zum parametrischen Student t-Test, der nur bei normal verteilten Daten zu Anwendung kommen darf.

An einer Vielzahl von Datensätzen wurde mit Hilfe des Kolmogoroff-Smirnoff-Tests geprüft, ob die verwendeten Daten einer Normalverteilung unterliegen. Das Ergebnis war, dass die Daten in allen Testfällen innerhalb kleiner Fehlergrenzen normal verteilt sind. Deswegen kann angenommen werden, dass die Verteilung ausreichend gut durch die Gauß-Verteilung angenähert wird. Es wird daher auf die Null-Hypothese getestet, indem ein paarweiser doppelseitiger t-Test durchgeführt wird, der die individuellen Abstands- und Winkel-*Restraints* aus allen *Restraint*-Sätzen $R_i^* \left(i>1\right)$ mit den zugehörigen *Restraints* aus dem Satz R_1^* vergleicht. Aus den Strukturbündeln wird der Mittelwert der Abstände $< d_i^{k*} >$ und der Winkel $< a_i^{k*} >$ zusammen mit den zugehörigen Standardabweichungen $s\left(d_i^{k*}\right)$ und $s\left(a_i^{k*}\right)$ berechnet. Die t-Werte $t_1^k \left(i>1\right)$ für Abstände und Winkel werden wie folgt berechnet

$$t_1^k = \frac{\left| < R_1^k > - < R_i^{k*} > \right|}{\sqrt{\frac{s^2(R_1^k)}{L_1} + \frac{s^2(R_i^{k*})}{L_i}}}. \tag{3.15}$$

L entspricht dabei der Anzahl der Messwerte für einen *Restraint*.

Danach werden die einzelnen t-Werte t_1^k mit dem kritischen t-Wert t_c verglichen. Der kritische t-Wert bei gegebenen Signifikanzniveau und bekannten Freiheitsgrad df (mit $df = L_1 - L_i - 1$) wird mit Hilfe eines Näherungsverfahrens [68] berechnet. Ist der berechnete t-Wert t_1^k größer als der kritische t-Wert t_c, muss die Null-Hypothese abgelehnt werden und der *Restraint* R_i^{k*} wird nicht weiter verwendet. *Restraints* mit $t_1^k \le t_c$ werden behalten und der Mittelwert $\left\langle R_0^k \right\rangle$ des *Restraint* R_0^k wird wie folgt berechnet

$$< R_0^k >= \frac{\frac{< R_1^k >}{s^2(R_1^k)} + \sum_{i \in A_i} \frac{< R_i^k * >}{s^2(R_i^k *)}}{\frac{1}{s^2(R_1^k)} + \sum_{i \in A_i} \frac{1}{s^2(R_i^k *)}}, \tag{3.16}$$

wobei $A_i = \{\text{für alle } i \text{ mit } t_1^k \le t_c\}$ gilt. Die gewichtete Standardabweichung $s\left(R_0^k\right)$, die für die Berechnung der Fehlergrenzen der Abstand-*Restraints* verwendet wird erhält man folgendermaßen:

$$s\left(R_0^k\right) = \sqrt{\frac{\frac{1}{s^2\left(R_1^k\right)} \cdot \sum_{p=1}^{L_1} (d_{1p}^k - < R_0^k >)^2 + \sum_{i \in A_i} \frac{1}{s^2\left(R_i^k\right)} \cdot \sum_{p=1}^{L_i} (d_{ip}^k - < R_0^k >)^2}{\left(L_1 \cdot \frac{1}{s^2\left(R_1^k\right)} + \sum_{i \in A_i} L_i \cdot \frac{1}{s^2(R_i^k)} \right) \cdot \frac{L_1 + \sum_{i \in A_i} L_i - 1}{L_1 + \sum_{i \in A_i} L_i}}} \tag{3.17}$$

Die d_{ip}^k beschreiben die p einzelnen Messwerte für den *Restraint* R_i^k.

Die Berechnung der Mittelwerte und Standardabweichungen für Winkel-*Restraints* erfolgt analog nach dem Ansatz von Döker [70].

3.2.5 Filterung der Winkel-*Restraints*

Wenn Diederwinkel kombiniert und gemittelt werden kann es vorkommen, dass das berechnete Mittel in unerlaubte Regionen des *Ramachandran-Plots* fällt. Deshalb wurde ein Filter eingebaut, der es dem Benutzer erlaubt, Diederwinkel von Rückgrat und Seitenketten zu ignorieren, abhängig davon in welche Regionen (*disallowed, generously allowed, additional allowed, favored regions*) des *Ramachandran-Plots* sie fallen. Die Regionen sind klassifiziert in seltene (Wert 0) and übliche (Wert 9) Kombinationen und wurden aus den PROCHECK [41] abgeleitet.

3.2.6 Wasserstoffbrücken-*Restraints*

Der ISIC-Algorithmus verwendet zusätzlich zu den kombinierten Abstands- und Diederwinkel-*Restraints* ebenso Wasserstoffbrücken-*Restraints* H_i^k. Im Prinzip können Wasserstoffbrücken ähnlich behandelt werden wie die Abstand-*Restraints*, indem die Verteilungen der Wasserstoffbrückenenergien als Parameter verwendet werden. Die Wasserstoffbrückenenergien werden dabei nach Freund [91] bestimmt. Aus Geschwindigkeitsgründen wurde in ISIC eine schnellere Methode zur Bestimmung von Wasserstoffbrücken verwendet. Sie werden definiert durch einen maximalen H^N-O-Abstand

von 0,24 nm und einem Wasserstoffbrückenwinkel a_{NHO} von $180^o \pm 35^o$. In ISIC wird nun die Häufigkeit X_i^{k*} der Wasserstoffbrücken in den verschiedenen Strukturbündeln S_i bestimmt und als Wasserstoffbrückenwahrscheinlichkeit $P\left(H_i^{k*}\right)$ benutzt. Anhand der bedingten Wahrscheinlichkeit $P\left(H_0^k \middle| H_1^k, H_i^{k*}, i = 2,...,N\right)$ wird bestimmt, ob eine Wasserstoffbrücke in der Lösungsstruktur vorliegt

$$P(H_0^k \mid H_1^k, H_i^k*, i = 2,...,N) = \frac{P(H)(P(H_1^k, H_i^k*, i = 1,...,N)}{P(H)(P(H_1^k, H_i^k*, i = 2,...,N) + (1 - P(H)(1 - P(H_1^k, H_i^k, i = 2,...,N))} . \quad (3.18)$$

Mit der Annahme, dass *Restraints* aus verschiedenen Strukturätzen statistisch unabhängig sind und zusammen mit Gleichung (3.11) kann die Wahrscheinlichkeit $P\left(H_i^k\right)$, dass eine Wasserstoffbrücke auch unter den Bedingungen der wahren Lösungsstruktur existiert, geschrieben werden als

$$P\left(H_i^k\right) = P\left(H_i^k \middle| H_i^{k*}, i = 1,...,N\right) P\left(H_i^{k*}, i = 1,...,N\right). \quad (3.19)$$

Aus Gleichung (3.18) und (3.19) erhält man

$$P(H_0^k \mid H_1^k, H_i^k*, i = 2,...,N) = \frac{P(H)(P(H_1^k \cdot \prod_{i=2}^{N} P(H_i^k \mid H_i^k*) P(H_i^k*)}{P(H)(P(H_1^k \cdot \prod_{i=2}^{N} P(H_i^k \mid H_i^k*) P(H_i^k*)) + (1 - P(H)(1 - P(H_0^k)(P(H_1^k \cdot \prod_{i=2}^{N} P(H_i^k \mid H_i^k*) P(H_i^k*))} . \quad (3.20)$$

Für die bedingte Wahrscheinlichkeit, dass eine Wasserstoffbrücke $P\left(H_0^k \middle| H_i^{k*}\right)$ auch in Lösung existiert, wenn sie in der Kristallstruktur vorhanden ist, wurde ein plausibler Wert von 0,9 angenommen. Genauere Werte können durch eine statistische Auswertung der bestehenden Strukturdatenbank erhalten werden. Die *a priori* Wahrscheinlichkeit $P\left(H\right)$, dass

eine Wasserstoffbrücke zwischen einem gegebenen Atompaar besteht ist ziemlich klein, ein plausibler Wert wäre $\frac{1}{Q}$, wobei Q die Anzahl der Aminosäuren des betrachteten Proteins ist.

Für den Fall, dass $P(H_0^k \mid H_1^k, H_i^k *, i = 2,..., N)$ einen benutzerdefinierten Schwellwert (z. B. 0,75) überschreitet, wird die zugehörige Wasserstoffbrücke als vorhanden definiert und in einen geeigneten Abstand-*Restraint* umgewandelt.

3.3 Automatische sequentielle Zuordnung (ASSIGN)

Die manuelle Auswertung von NMR-Spektren ist ein sehr zeitaufwändiger und fehleranfälliger Prozess, der sich im Extremfall über Jahre hinziehen kann, bis die Struktur bestimmt ist. Daher wird seit einiger Zeit verstärkt an der Automatisierung der experimentellen NMR-Strukturbestimmung gearbeitet. Die einfachste Klasse der Computerprogramme dient zur Anzeige von NMR-Spektren. Mit ihnen können Signale identifiziert, benannt und in so genannten *Peaklisten* festgehalten werden. Zu diesen Programmen zählen u. a. Felix [92], Xeasy [93], NMRPipe [28], ANSIG [94] und AURELIA [12]. Weiterhin wurden zahlreiche Programme veröffentlicht, die eine *Bottom-Up*-Strategie verwenden und so eine sequentielle Zuordnung mittels einer teilweise automatischen Auswertung von Tripleresonanzexperimenten liefern. Zu diesen gehören GARANT [95], PASTA [96], CONTRAST [97], AUTOASSIGN [98] und Ansätze von Lukin et al. [99], Buchler et al. [100] und Li et al. [101]. Die Vorgehensweise ist bei allen Programmen ähnlich [102]. Vorausgesetzt werden *Peaklisten* von heteronuklearen NMR-Spektren. Die Signale werden zu Gruppen bzw. Spinsystemen zusammengefasst, die dann durch statistische Analysen von chemischen Verschiebungen identifiziert werden. Gelingt dies, müssen sequentielle Nachbarschaften ermittelt werden. Zu diesem Zweck kommen deterministische Methoden oder energieminimierende Verfahren wie *Simulated-Annealing* [86] zum Einsatz. Die so erhaltenen Segmente sequentiell benachbarter Spinsysteme müssen in einem letzten Schritt mit der Primärsequenz des Proteins abgeglichen werden. Hier werden wiederum deterministische oder energieminimierende Methoden eingesetzt. Die Vorgehensweise ist der eines menschlichen Experten ähnlich. Die Qualität der Ergebnisse der genannten Programme hängt entscheidend von den verwendeten *Peaklisten* ab. Sind diese nicht vollständig, d. h. fehlen wichtige Signale in den Spektren, so scheitert die automatische Zuordnung. Auch zusätzliche *Rauschpeaks* tragen oft zu einer Verschlechterung der Performance dieser Algorithmen bei.

Wie bereits oben beschrieben ist für die Strukturbestimmung nicht nur die sequentielle Zuordnung entscheidend. Erst mit einer richtigen Zuordnung der NOESY-Spektren können strukturrelevante Informationen gewonnen werden. Die Ansätze zur Zuordnung von NOESY-Spektren, die in NOAH [103] und ARIA [104] beschrieben sind, benötigen die komplette sequentielle Zuordnung in hoher Genauigkeit. Bei fehlerhafter oder unvollständiger sequentieller Zuordnung und bei zu vielen überlappenden Signalen in den NOESY-Spektren scheitern diese Methoden, weil sich ausgehend von der sequentiellen Zuordnung nur wenige

Signale eindeutig zuordnen lassen. Deshalb verwendet ARIA bei der Strukturberechnung mehrdeutige Zuordnungen als gewichtete Summe in den Pseudopotentialen. NOAH berechnet für jede Zuordnungsmöglichkeit Strukturen. Anschließend wird bewertet, wie die Zuordnungen durch die erhaltenen Strukturen erklärt werden. Gemeinsam ist diesen Programmen, dass nur die Positionen der Signale in den NMR-Spektren verwendet werden, die Linienform der Signale aber vernachlässigt wird.

In dieser Arbeit wurde ein neuer Ansatz, angelehnt an TWOSTEP [14], für eine sequentielle Zuordnung mit Hilfe von NOE-Spektren entwickelt. Der erste Schritt in TWOSTEP wurde durch das in Kapitel 2.1.3 beschriebene PEAKASSIGN [20] ersetzt, das bereits bekannte chemische Verschiebungen an das experimentelle NOESY-Spektrum anpasst.
Ziel des neuen ASSIGN-Algorithmus ist, fehlende chemische Verschiebungen durch einen Vergleich von einem simulierten und einem experimentellen 2D-NOESY-Spektrum zu finden. Dabei werden Linienformen, Volumen und statistische Vorhersagen für chemische Verschiebungen herangezogen. Durch Variation der chemischen Verschiebungen wird bei jedem Schritt ein neues simuliertes Spektrum aufgebaut und dann mit dem experimentellen Spektrum verglichen. Über empirisch ermittelte Wahrscheinlichkeitsverteilungen werden Wahrscheinlichkeiten für Linienform und Volumen in den Spektren berechnet. Zusätzlich werden statistische Vorhersagen chemischer Verschiebungen benutzt, um Wahrscheinlichkeiten für die einzelnen Verschiebungen zu erhalten. Mittels eines *Threshold-Accepting*-(TA)-Algorithmus werden die Wahrscheinlichkeiten maximiert und die wahrscheinlichste Lösung wird als sequentielle Zuordnung festgehalten. In Abbildung 15 ist der ASSIGN-Algorithmus dargestellt, der im Folgenden genau beschrieben wird.

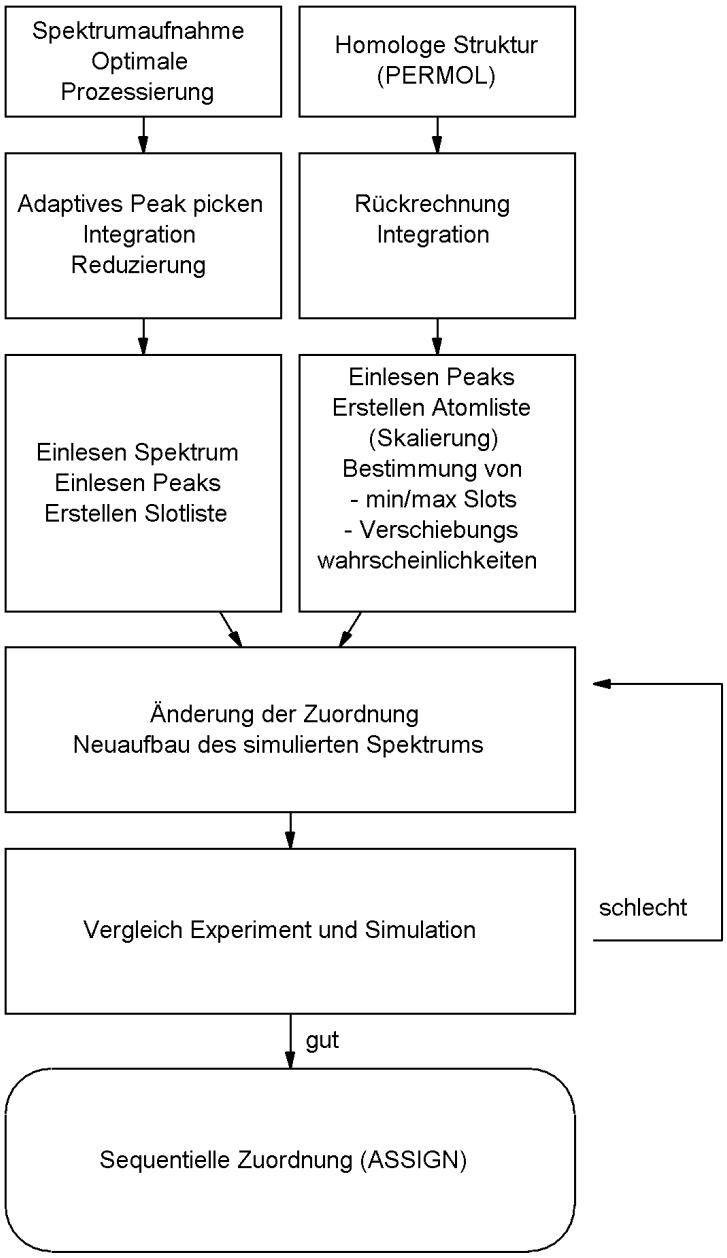

Abbildung 15: Eine Übersicht des ASSIGN-Algorithmus. Vor dem eigentlichen Optimierungslauf werden das experimentelle und das simulierte Spektrum automatisch vorverarbeitet bzw. erzeugt.

3.3.1 Aufbereitung des experimentellen Spektrums

Der erste Schritt besteht in der Präparation der experimentellen Daten. Nach der Aufnahme des Spektrums muss zunächst der FID prozessiert werden. Ein wichtiges Kriterium beim Vergleich von simulierten und experimentellen Daten ist die Multiplettstruktur der NOE-Signale, denn durch sie wird die Linienform der *Peaks* sehr stark beeinflusst. Deswegen muss bei der Fourier-Transformation darauf geachtet werden, dass diese Information nicht durch

ungeeignete Filterfunktionen verloren geht. Häufig werden NMR-Spektren so transformiert, dass das Signal-Rausch-Verhältnis maximal wird, wobei jedoch meistens die Multiplettstruktur verloren geht. Es muss also ein geeigneter Kompromiss gefunden werden, der die Multiplettstruktur erhält und aber auch leider ein schlechteres Signal-Rausch-Verhältnis liefert.

Jetzt folgt die automatische Datenauswertung in AUREMOL. Hierzu werden die *Peaks* des experimentellen 2D-NOESY-Spektrums mit Hilfe der *„adpativen Peakpicking"*-Routine aus AUREMOL (Kapitel 2.1.1) gepickt und anschließend integriert. Im letzten Teilschritt erfolgt eine Reduktion des experimentellen Spektrums. Die verbliebenen Signale werden vom *Peakmaximum* ausgehend in Richtungen parallel zu den Frequenzachsen des Spektrums segmentiert. Wird ein vom Benutzer vorgegebener Grenzwert der Intensitätswerte, der in Prozent vom *Peakmaximum* angegeben wird, erreicht, wird die Segmentierung abgebrochen. Ein zweites Abbruchkriterium ist durch eine Begrenzungsbox um das *Peakmaximum* ebenfalls vom Benutzer anzugeben. Die Box sollte dabei mindestens die doppelte maximal zu erwartende experimentelle Linienbreite (auf halber Höhe des Signals) in Hz haben, um die komplette *Peakform* für alle *Peaks* zu erhalten. Die so erhaltenen Intensitätswerte werden nun als reduziertes Spektrum in einer Datei festgehalten. Weiterhin werden in dieser Datei die Spektrenparameter wie die Anzahl der Datenpunkte, die Spektrenbreite, der *Offset*, die Auflösung und die Resonanzfrequenzen abgespeichert. Zudem wird jeder einzelne *Peak* mit seinen Koordinaten, Volumen und durch die Reduzierung erhaltenen effektiven Ausmaße abgespeichert.

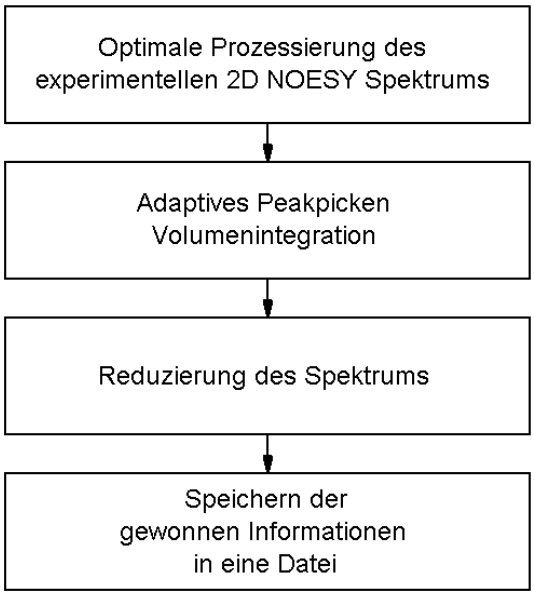

Abbildung 16: Arbeitsschritte zur Aufbereitung der Daten aus dem experimentellen 2D-NOESY-Spektrum. Die Aufgaben werden mit den in AUREMOL bereitgestellten Routinen automatisch erledigt.

3.3.2 Simulation des Vergleichspektrums

Die Simulation des Vergleichspektrums erfolgt mit der in AUREMOL enthaltenen Rückrechungsroutine RELAX. Als erstes benötigt man dafür eine Proteinstruktur. Wenn die Struktur bereits bekannt ist, kann auf diese zurückgriffen oder aber eine homologe Strukturen verwendet werden. Die weiteren Parameter wie maximale Distanz, Mischzeit, Relaxationszeit und Resonanzfrequenz sollten analog zu den experimentellen Daten eingestellt werden. Abhängig von der Prozessierung der experimentellen Daten kann die Linienform (Gauß oder Lorentz), die zusätzliche Linienverbreiterung und die Berücksichtigung von Multipletts festgelegt werden. Die chemischen Verschiebungen sind in AUREMOL in einer so genannten meta-Datei gespeichert, die zusätzlich noch Informationen zur Temperatur und Korrelationszeit des Moleküls enthalten, welche für die Simulation ebenfalls wichtig sind. Dazu wurde im Rahmen dieser Arbeit in AUREMOL eine Korrelationszeitberechnung eingebaut, die sich an den Routinen in [57] orientiert. Der bekannte Teil der Zuordnung ist in der meta-Datei gespeichert. Für den unbekannten Teil der Zuordnung, die von ASSIGN bestimmt werden, genügen Zufallswerte im Bereich der Spektrenbreite. ASSIGN bietet hier die Funktion, die fehlende Zuordnung automatisch durch zufällige Werte zu ersetzen.

Das für ASSIGN wichtige Ergebnis der Rückrechnung ist eine Datei, die alle simulierten *Peaks* mit ihren Linienformen, Volumen und Ausmaßen enthält. Wie in Kapitel 2.1.5 erwähnt, sind jedoch die Volumen der Rückrechnung nicht mit den Volumen des Experiments

vergleichbar. Deswegen werden die simulierten Volumen durch die Volumen ersetzt, die mit der AUREMOL Volumenintegrationsroutine aus den simulierten Daten berechnet werden, die ebenfalls beim Experiment angewendet wurde.

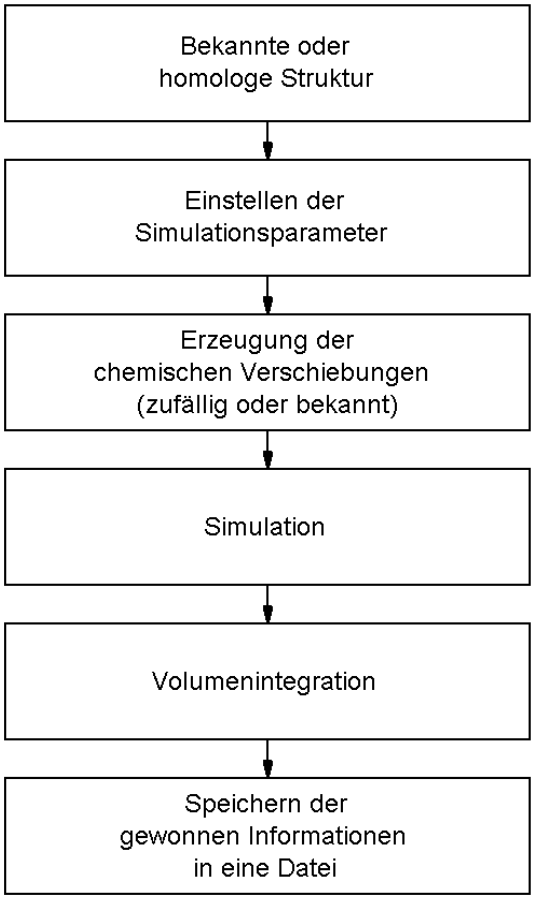

Abbildung 17: Ablauf zur Erzeugung des simulierten Vergleichspektrums. Die einzelnen Routinen sind in AUREMOL eingebettet.

3.3.3 Vorbereitungen

Als erstes müssen die oben gewonnenen experimentellen und simulierten Daten eingelesen und ausgewertet werden. Dazu werden alle *Peaks* eingelesen und das reduzierte Spektrum im Speicher aufgebaut. Dann werden alle vorkommenden *Peakpositionen* sowohl in der direkten wie auch in der indirekten Richtung in einer Liste festgehalten und nach ppm-Werten sortiert. Diese Liste wird als *Slotliste* bezeichnet und gibt die Positionen wieder, an welche die chemischen Verschiebungen der simulierten Atome verschoben werden können. Da beim *Peakpicken* alle Maxima selektiert werden, liegen bei Multipletts mit einer geraden Anzahl N_{MP} immer N_{MP} *Subpeaks* vor. In der *Slot*-Liste sind die ppm-Werte dieser N_{MP} gepickten

Subpeaks vorhanden. Nicht aber der ppm-Wert des *Hauptpeaks*, der genau in der Mitte aller *Subpeaks* liegt. Aus diesem Grund wird die Verschiebeposition der *Hauptpeaks* nachträglich auf folgende Weise z.B. für ein Dublett hinzugefügt: Liegen zwei *Peaks* nicht weiter als 12 Hz auseinander, kann man nach der Karplus-Beziehung [11] annehmen, dass es sich um ein Dublett handeln könnte und es wird in die Mitte der beiden *Subpeaks* eine künstliche *Peakposition* hinzugefügt, der der *Hauptpeak* sein könnte.

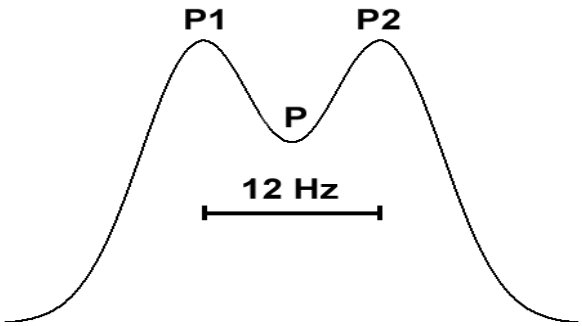

Abbildung 18: Behandlung von Multipletts mit gerader Anzahl von *Subpeaks* am Beispiel eines Dubletts: Liegen zwei *Peaks* P1 und P2 näher als 12 Hz beieinander, so wird in die Mitte ein weiter künstlicher *Peak* P eingefügt, der dem wahren *Peak* entspricht. Die so erhaltene Peakposition dient als *Slot* für die chemischen Verschiebungen.

Handelt es sich wirklich um ein höheres Multiplett, so kann die chemische Verschiebung des Atoms, das an dem *Multiplettpeak* beteiligt ist, an die richtige Position geschoben werden. Wenn nicht, stellt das für den Algorithmus kein Problem dar, da er leere *Slots* zulässt.

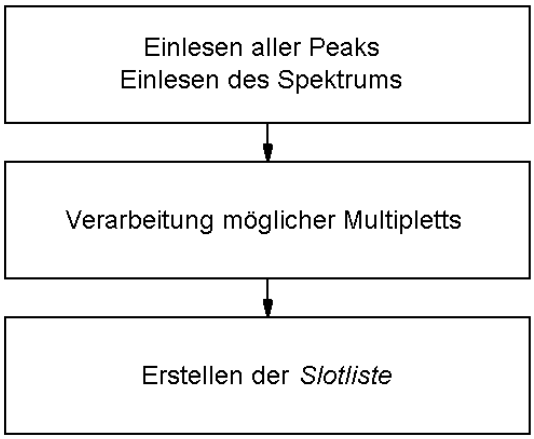

Abbildung 19: Das Einlesen und Verarbeiten der experimentellen Daten.

Als nächstes werden alle simulierten *Peaks* mit den zugehörigen *Peakformen* eingelesen. Danach werden die einzelnen Atome aus der *Peakliste* extrahiert und nun zu diesen Atomen eine Liste mit den zugehörigen *Peaks* erstellt.
Ist bereits eine Teilzuordnung bekannt, werden die *Peakintensitäten* auf das experimentelle Spektrum skaliert. Dabei werden die Volumen von eindeutig zugeordneten experimentellen *Peaks* mit den entsprechenden simulierten *Peaks* über die *Maximum-Likelihood*-Methode (siehe R-Wert [39]) skaliert. Die simulierten *Peakintensitäten* werden anschließend mit dem so erhaltenen Skalierungsfaktor multipliziert.

Um den ppm-Bereich der erlaubten Verschiebungen für ein Atom einzuschränken, gibt es für jedes Atom einen minimalen und maximalen *Slot* und damit einen minimalen und maximalen ppm-Wert. Diese Werte werden wie folgt erhalten: Zuerst wird die in der Datenbank von AUREMOL enthaltene Tabelle für chemische Verschiebungen (Chemical Shift Table: BMRB 06-30-2004) für die einzelnen Atome ausgewertet. In der Datenbank sind hierfür die Mittelwerte der Verschiebungen und die zugehörige Standardabweichung abgelegt. Der minimale und maximale *Slot* wird erhalten, indem man die Minima und Maxima der chemischen Verschiebungen verwendet, die ebenfalls in der Datenbank abgelegt sind. Danach wird die Ausgabe des Programms SHIFTS [105] verwertet. Dieses Programm sagt für eine vorgegebene Struktur die zu erwartenden chemischen Verschiebungen voraus. In unsrer Arbeitgruppe wurde von Baskaran et al. (nicht veröffentlicht) Voraussagen von SHIFTS von bereits gut bekannten Proteinstrukturen und deren tatsächlichen chemischen Verschiebungen analytisch untersucht. Als Ergebnis wurde für jede Atomsorte eine Standardabweichung erhalten, die besagt, wie weit die wahren Verschiebungen von den Vorhersagen abweichen.
Unter der Annahme, dass die chemischen Verschiebungen gaußverteilt um die vorhergesagte Verschiebung liegen, kann man mit Hilfe dieser Standardabweichung Wahrscheinlichkeiten für eine Zuordnung berechnen, je nachdem wie weit die Zuordnung vom vorhergesagten Wert entfernt ist.

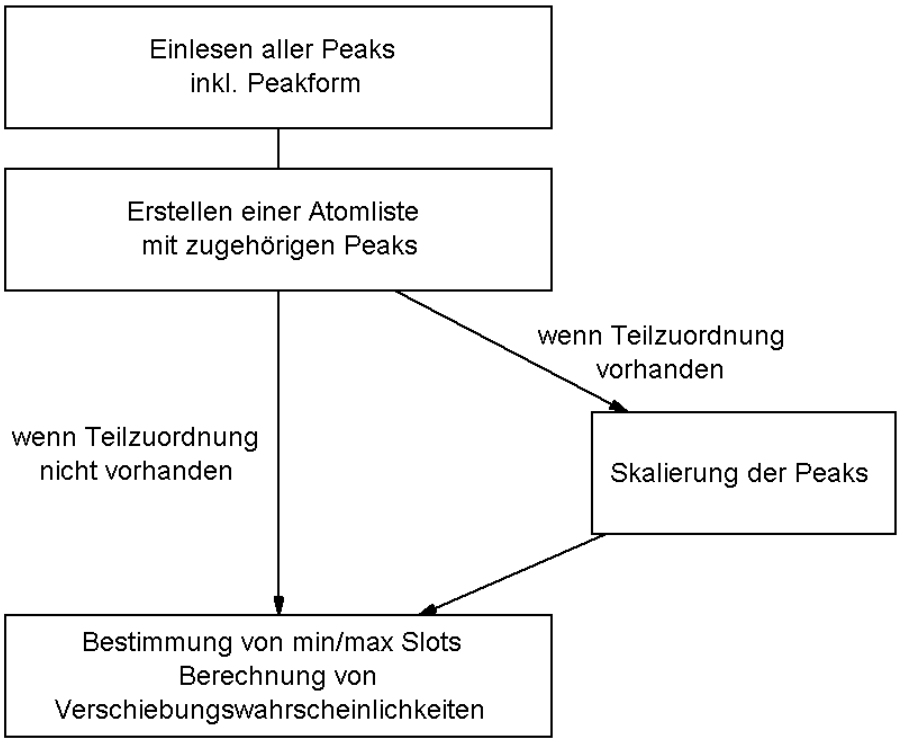

Abbildung 20: Das Einlesen und die Weiterverarbeitung der simulierten *Peaks*.

3.3.4 Änderung der Zuordnung

In ASSIGN werden die chemischen Verschiebungen der einzelnen simulierten Atome solange variiert, bis eine bestmöglichste Übereinstimmung zwischen experimentellen und simulierten Spektrum erreicht ist. Verschoben wird auf den oben genannten *Slots*. In der *Slotliste* sind alle möglichen ppm-Werte des experimentellen Spektrums enthalten. Zusätzlich gibt es einen *Dummy-Slot*, der mit einer einstellbaren Wahrscheinlichkeit angewählt wird. Wird für die chemische Verschiebung eines Atoms dieser *Slot* als Ziel gewählt, hat dies zur Folge, dass alle *Peaks*, an dem das Atom beteiligt ist, aus dem simulierten Spektrum verschwinden. Dies ist wichtig, da bei experimentellen Spektren Signale fehlen können, was u. a. auf Austauscheffekte zurückzuführen ist. Die Größe dieser Wahrscheinlichkeit kann mit Hilfe eines TOCSY-Spektrums abgeschätzt werden, bei dem zu erkennen ist, wie viele Spinsysteme zu sehen sind. Weiterhin kann die Sequenzlänge als Abschätzung dienen.

Wird die chemische Verschiebung eines Atoms fälschlicherweise in diesen Dummy-*Slot* verschoben, so impliziert dies eine Strafe, da beim späteren Vergleich der beiden Spektren die entsprechenden Peaks im simulierten Spektrum fehlen, was zu einer schlechten Bewertung führt. Weiterhin sind Mehrfach- und Nichtbelegung von *Slots* erlaubt.

Für die Verschiebungen gibt es zwei Mechanismen, den *Move* und den *Swap*. Beim *Move* wird nur das einzelne Atom verschoben, beim *Swap* werden zwei Atome gegeneinander ausgetauscht. Die Auswahl des Atoms, welches verschoben wird, ist zufällig. Das Ziel, wohin es verschoben wird, wird über die oben beschriebene Wahrscheinlichkeit der chemischen Verschiebungen des Atoms bestimmt. Der ausgewählte *Zielslot* hat dadurch eine Wahrscheinlichkeit PS zwischen 0 und 1. Ein Zufallszahlengenerator erzeugt eine Zahl Z_{rnd} zwischen 0 und 1. Ist nun $Z_{rnd} > PS$, wird der *Slot* ausgewählt, andernfalls ein neuer *Slot* ausgesucht und in der eben beschriebenen Weise geprüft.
Nach jedem *Move* oder *Swap* wird das simulierte Spektrum neu aufgebaut und mit dem experimentellen Spektrum in unten beschriebener Art und Weise verglichen und bewertet.

3.3.5 Bewertung der Zuordnung

Ist ein *Move* oder *Swap* ausgeführt, wird zur Bewertung der aktuellen Zuordnung das neu gebildete simulierte Spektrum mit dem immer konstanten experimentellen Spektrum an definierten Messbereichen verglichen. Die Mitte dieser Messbereiche sind die Positionen der experimentellen *Peaks*. Die Abmessungen der Bereiche sind verschieden und entsprechen den horizontalen und vertikalen *Peakbreiten* nx und ny. nx und ny sind dabei einfach nur die Anzahl der Datenpunkte in der direkten und indirekten Richtung des Spektrums.
Es werden mehrere Pseudoenergieterme und daraus Wahrscheinlichkeiten berechnet, die im Folgenden genau beschreiben werden.

Linienform

Der erste Term ist eine Maß für die Übereinstimmung des experimentellen Spektrums mit dem simulierten Spektrum im jeweiligen Bereich der Messstellen bzgl. der Linienform. Die Prozedur ist für alle Messbereiche dieselbe. Die Intensitätswerte der beiden Spektren werden an den Messbereichen jeweils in einen N-Dimensionalen Vektor $\vec{s}_p^{ex}$ und $\vec{s}_p^{sim}$ festgehalten. Die Dimension N ergibt sich aus den Abmessungen der Box nx und ny des experimentellen *Peaks* zu $N = nx \cdot ny$. Aus den beiden Vektoren wird dann mit Hilfe des Cosinuskriteriums berechnet, wie gut die Übereinstimmung ist

$$ES_p = \cos\left(\vec{s}_p^{ex}, \vec{s}_p^{sim}\right) = \frac{\vec{s}_p^{ex} \cdot \vec{s}_p^{sim}}{\left|\vec{s}_p^{ex}\right| \cdot \left|\vec{s}_p^{sim}\right|}, \tag{3.21}$$

wobei ES_p für ***E**nergy **S**hape* des Messbereichs *p* steht.

Stimmen beide Formen perfekt überein, so erhält man den Wert 1, bei schlechter Übereinstimmung einen Wert nahe 0. Aus dieser Pseudoenergie ist es nun möglich, über entsprechende Häufigkeitsverteilungen der *ES* -Energien eine Wahrscheinlichkeit für jede Messstelle zu berechnen. Dabei werden einmal eine Häufigkeitsverteilung einer völlig zufälligen Zuordnung und einmal die Häufigkeitsverteilung einer partiellen Lösung aufgestellt. Aus beiden Verteilungen wird dann die Bayessche Wahrscheinlichkeit für die Messstelle berechnet

$$PS_p = \frac{P_p^{S,ok}}{P_p^{S,ok} + P_p^{S,rnd}}, \tag{3.22}$$

wobei PS_p für ***P**robability **S**hape*, $P_p^{S,ok}$ für die Wahrscheinlichkeit der Linienform aus der partiellen Lösung und $P_p^{S,rnd}$ für die Wahrscheinlichkeit aus der zufälligen Lösung des *Peaks* *p* steht.

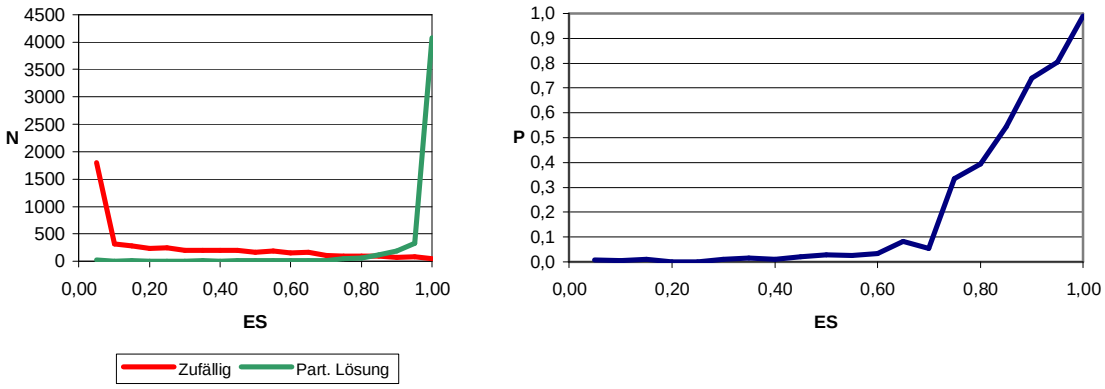

Abbildung 21: Beispiel für eine Häufigkeitsverteilung der Pseudoenergie *ES* (links) mit daraus resultierender Wahrscheinlichkeitsfunktion (rechts).

Volumen

Analog zur Berechnung der Pseudoenergie der Linienform kann nun ebenfalls eine Pseudoenergie für die Volumen berechnet werden. Auch hier werden bei allen Messbereichen die Volumen V_p^{ex} für das experimentelle und V_p^{sim} für das simulierte Spektrum durch Addition

der Intensitätswerte berechnet. Da bei den Volumen starke Differenzen aufgrund der r^{-6}-Abhängigkeit auftreten, werden diese in Distanzen umgerechnet. Man erhält mit

$$\begin{aligned} D_p^{ex} &= \left(V_p^{ex}\right)^{-\frac{1}{6}} \\ D_p^{sim} &= \left(V_p^{sim}\right)^{-\frac{1}{6}} \end{aligned} \tag{3.23}$$

die Distanzen D_p^{ex} und D_p^{sim} für den jeweiligen Messbereich p. Aus beiden Distanzen wird nun die Differenz berechnet

$$EV_p^* = D_p^{ex} - D_p^{sim} \tag{3.24}$$

und man erhält mit EV_p^* (***E**nergy **V**olume*) einen Wert, der angibt, wie gut die Volumen an den jeweiligen Stellen im Spektrum zusammenpassen. Ist der Wert 0, stimmen die Volumen perfekt überein, andernfalls ist die Abweichung linear von der Differenz abhängig. Um wie bei der Linienform einen Wert zwischen 0 und 1 für die Qualität der Übereinstimmung zu erhalten, wird EV_p^* folgendermaßen normiert und umgekehrt

$$EV_p = \begin{cases} 1 - \dfrac{\left|D_p^{ex} - D_p^{sim}\right|}{D_p^{ex}}, & \text{wenn } D_p^{sim} < 2 \cdot D_p^{ex} \\ 0 & \text{sonst} \end{cases}. \tag{3.25}$$

Wenn also die aus dem simulierten Volumen gewonnene Distanz D_p^{sim} größer ist als das 2-fache der aus den experimentellen Volumen gewonnene Distanz D_p^{ex}, kann davon ausgegangen werden, dass die beiden Stellen im Spektrum nicht übereinstimmen und EV_p wird auf 0 gesetzt. Sind beide Distanzen gleich, ist $EV_p = 1$.

Analog zu der Pseudoenergie der Linienform kann nun aus dieser Pseudoenergie für das Volumen über entsprechende Häufigkeitsverteilungen der EV-Energien eine Wahrscheinlichkeit für die Übereinstimmung an jedem Messbereich berechnet werden. Die Bayessche Wahrscheinlichkeit aus beiden Verteilungen berechnet sich zu

$$PV_p = \frac{P_p^{V,ok}}{P_p^{V,ok} + P_p^{V,rnd}}. \tag{3.26}$$

Hier steht PV_p für ***P**robability **V**olume*, $P_p^{V,ok}$ für die Wahrscheinlichkeit gewonnen aus der Volumeninformation der partiellen Lösung und $P_p^{V,rnd}$ für die Wahrscheinlichkeit aus der zufälligen Lösung des Messbereichs p steht.

Gesamtwahrscheinlichkeit der Messbereiche

Beide Wahrscheinlichkeiten der Messbereiche können nun kombiniert werden und man erhält die Gesamtwahrscheinlichkeit, dass das experimentelle und das simulierte Spektrum an den Messbereichen bzgl. der Linienform und des Volumens übereinstimmen. Diese lässt sich als

$$PSV_p = \frac{P_p^{S,ok} \cdot P_p^{V,ok}}{P_p^{S,ok} \cdot P_p^{V,ok} + P_p^{S,rnd} \cdot P_p^{V,rnd}} \tag{3.27}$$

schreiben. PSV_p steht für ***P**robability **S**hape **V**olume* am Messbereich p.

Die Gesamtwahrscheinlichkeit für alle Messbereiche erhält man, wenn man alle Wahrscheinlichkeiten miteinander multipliziert.

$$PSV = \prod_{p=1}^{N_{ex}} PSV_p \tag{3.28}$$

wobei N_{ex} die Anzahl der Messbereiche ist.

Sind jedoch beispielsweise 5000 Messbereiche vorhanden und wird für jeden Messbereich eine Wahrscheinlichkeit von $PSV_p = 0,95$ gesetzt, errechnet sich die Gesamtwahrscheinlichkeit zu $PSV = 0,95^{5000} \approx 4,15 \cdot 10^{-112}$. Da dieser Wert sehr klein ist, kann er numerisch schlecht optimiert werden. Aus diesem Grund werden die natürlichen Logarithmen der Einzelwahrscheinlichkeiten gebildet:

$$\ln\left(PSV\right) = \ln\left(\prod_{p=1}^{N_{ex}} PSV_p\right) = \sum_{p=1}^{N_{ex}} \ln\left(PSV_p\right) \tag{3.29}$$

mit der Bedingung, dass $PSV_p = 0,05$ wenn $PSV_p < 0,05$ ist. Dies ist notwendig, da der Logarithmus für kleine Werte gegen $-\infty$ geht und sonst einzelne Messbereiche mit geringer Wahrscheinlichkeit die Gesamtwahrscheinlichkeit dominieren würden.

Ziel des Algorithmus ist, diese logarithmische Wahrscheinlichkeit durch Änderungen der chemischen Verschiebungen zu maximieren. Dazu wird in dieser Arbeit das weiter unten beschriebene *Threshold-Accepting*-Verfahren verwendet, das eine Pseudoenergie minimiert. Die verwendetet Pseudoenergiefunktion ESV (***E**nergy **S**hape **V**olume*) ist abhängig von den Wahrscheinlichkeiten der Übereinstimmung an den Messbereichen und von der Anzahl der Messstellen selbst

$$ESV = \sum_{p=1}^{N_{ex}} \left| \ln\left(PSV_p\right) \right| \tag{3.30}$$

Ist für jede Messstelle die Wahrscheinlichkeit 1, so erhält man $ESV = 0$, ist sie weit unter 1 und damit sehr schlecht, wird ESV maximal $\ln(0,05) \cdot N_{ex} \approx 3 \cdot N_{ex}$. Der Bereich für ESV ist also das Intervall $[0,\ 3 \cdot N_{ex}]$. Je niedriger ESV, desto besser stimmen das experimentelle und das simulierte Spektrum überein und umgekehrt. Das obige Beispiel liefert nun $ESV = 5000 \cdot \left|\ln(0,95)\right| \approx 256,5$ – ein Wert der numerisch gut zu erfassen ist.

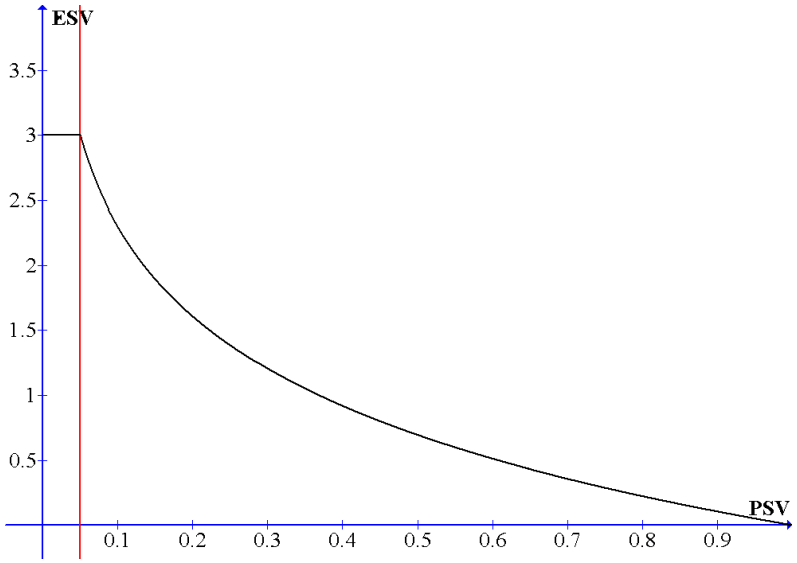

Abbildung 22: Gezeigt ist die Pseudoenergiefunktion *ESV*, die logarithmisch aus der Wahrscheinlichkeiten *PSV* abgeleitet ist. Ist PSV < 0,05, ist der Wert von *ESV* 3.

Wahrscheinlichkeit der chemischen Verschiebungen

Wie oben bei den Vorbereitungen beim Einlesen der simulierten *Peaks* beschrieben, wurde für jede chemische Verschiebung eine Wahrscheinlichkeitstabelle erstellt, die besagt, mit welcher Wahrscheinlichkeit die chemische Verschiebung eines Atoms richtig ist bzgl. der Vorhersage von SHIFTS.

Mit Hilfe dieser Wahrscheinlichkeiten wird nun ein zweiter Pseudoenergieterm in analoger Weise zu den Wahrscheinlichkeiten an den Messstellen erstellt

$$ECS = \sum_{s=1}^{N_{CS}} \left| \ln\left(PCS_s\right) \right|, \tag{3.31}$$

wobei PCS_s (***P**robability **C**hemical **S**hift*) die Wahrscheinlichkeit der chemischen Verschiebung des Atoms s und ECS (***E**nergy **C**hemical **S**hift*) die gesamte Abweichung aller chemischer Verschiebungen von den vorhergesagten Verschiebungen beschreibt. N_{CS} ist die Anzahl der Atome s mit den zu variierenden chemischen Verschiebungen. Liegen alle Atome auf den vorhergesagten Verschiebungen, ist diese Energie 0 und das theoretische Optimum wäre erreicht. Dies ist allerdings nicht zu erwarten, da in Realität die richtigen

Verschiebungen nicht komplett mit den Vorhersagen übereinstimmen werden. Vielmehr ist zu erwarten, dass sich ES einem bestimmten Wert annähert.

Die beiden Pseudoenergieterme ESV und ECS werden auf N_{ex} normiert, damit beide Terme die gleiche Größenordnung haben. Danach werden sie gewichtet addiert und die totale Pseudoenergiefunktion E_{tot} ist

$$E_{tot} = w_{ESV} \cdot ESV + w_{ECS} \cdot ECS \frac{N_{ex}}{N_{CS}}. \tag{3.32}$$

Dabei sind w_{ESV} und w_{ECS} Gewichtungsterme der Einzelpseudoenergien.

Die Aufgabe des Algorithmus ist die Minimierung von E_{tot}, was über das im folgenden beschriebene Optimierungsverfahren realisiert ist.

Threshold-Accpeting

In der Praxis ist die oben beschriebene Problemstellung derart komplex, dass eine *brute force* Methode nicht alle Kombinationsmöglichkeiten der Verschiebungen aller Atome innerhalb einer endlichen Zeitspanne evaluieren kann. Um das globale Minimum zu finden, wird in dieser Arbeit der *Threshold Accepting*-Algorithmus (TA) eingesetzt, welcher zu den heuristischen Optimierungsverfahren zählt. Er stellt eine Weiterentwicklung des S*imulated-Annealing*-Algorithmus [86] dar und wurde von Dück und Scheuer [106] bei IBM entwickelt, um die Anordnung der Leiterbahnen auf Computerchips zu optimieren. Dieser Algorithmus ist einfach zu implementieren und zu dem sehr schnell, was besonders wichtig ist um die Laufzeit des Algorithmus so klein wie möglich zu halten. Weiterhin wären genetische Algorithmen [107] und Hidden-Markov-Modelle [107] denkbar. ASSIGN ist so programmiert, dass ein Austausch des Optimierungsalgorithmus relativ einfach zu bewerkstelligen ist.

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. die Verschiebungen der Atome liegen auf zufällig ausgewählten Slots. Dieser Zustand des Systems soll mit α bezeichnet werden. Durch eine kleine Änderung wird eine neue Konfiguration β herbeigeführt. Dies entspricht bei ASSIGN einem *Mov*e oder einem

Swap, d.h. es wird oder werden die chemische Verschiebung(en) eines oder zweier Atome geändert.
Im Anschluss daran wird die Änderung der Pseudoenergie berechnet

$$\Delta E = E(\beta) - E(\alpha). \tag{3.33}$$

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

$$P(\alpha \to \beta) = \Theta(Th - \Delta E) \tag{3.34}$$

gegeben, wobei Θ die *Heaviside*-Funktion bzw. Stufenfunktion ist.
Ähnlich wie die Temperatur beim *Simulated-Annealing* langsam erniedrigt wird, wird beim TA der *Threshold Th* langsam erniedrigt, um das System einzufrieren. Wichtig hierbei ist, dass abhängig von der Wahl des *Threshold Th*, auch schlechtere Konfigurationen akzeptiert werden und so nicht die Gefahr besteht, in lokalen Minima stecken zu bleiben, wie das bei einfachen Gradientenabstiegsverfahren der Fall wäre. Im Idealfall kann so das globale Minimum gefunden werden.

4 Ergebnisse

4.1 Homologie-Modellierung (PERMOL)

4.1.1 Modellierung von HPr aus *Staphylococcus aureus*

Um den in dieser Arbeit vorgestellten Ansatz der Homologie-Modellierung zu testen, wurde die Struktur des Histidin enthaltenden Phosphorüberträgerproteins (*histidine containing phosphorcarrier proteins*, HPr) aus *Staphylococcus aureus* modelliert. Das HPr-Protein ist ein integraler Bestandteil des bakteriellen Phosphotransferase-Systems (PTS) [108]. Das PTS ist für die spezifische Zuckeraufnahme in die Zellen zuständig, wobei die Zucker gegen einen Konzentrationsgradienten bei gleichzeitiger Phosphorylierung transportiert werden.
HPr-Moleküle mehrerer verschiedener Organismen wurden ausführlich studiert und bereits verschiedene 3D-Strukturen davon aufgeklärt, insbesondere auch HPr aus *Staphylococcus aureus* mittels NMR [109] (PDB Code 1KA5). Deswegen eignet es sich in hervorragender Weise zur Überprüfung der entwickelten Modellierungsstrategie.
Als Modellstrukturen wurden drei vorher bestimmte HPr-Strukturen aus zwei verschiedenen Organismen (*Escherichia faecalis* und *Bacillus subtilis*) verwendet (PDB Codes: 1PTF, 1QFR und 2HID).

PDB Code	Organismus	Methode	Referenz	Auflösung [A]	Sequenz-Identität [%]
1PTF	*E. faecalis*	Röntgen	[110]	1,6	64,8
1QFR	*E. faecalis*	NMR	[111]	2,7	64,0
2HID	*B. subtilis*	NMR	[112]	1,9	60,2

Tabelle 1: Die verwendeten Modellstrukturen. Die Auflösung der NMR-Strukturen wurde mit PROCHECK-NMR [41] bestimmt.

Als erstes wurde mit AUREMOL ein *Alignment* der Sequenzen der Modellstrukturen an die Zielsequenz durchgeführt. Es zeigte sich eine Sequenzidentität zwischen 60 % bis 65 % für die einzelnen Modellstrukturen, welche somit für eine Homologie-Modellierung geeignet sind, da bereits ab einer Sequenzidentität von mindestens 20 % der globale Faltung erhalten bleibt (siehe Kapitel 3.1).

```
AUREMOL alignment file

OFFSET: 0

E:\HM_HPr_Au\Target\HPr_au.seq|TARGET      M   E   Q   N   S   Y   V   I   I   D   E   T   G
                                           1   2   3   4   5   6   7   8   9  10  11  12  13
E:\HM_HPr_Au\Models\1PTF_IUPAC.pdb|MODEL   M   E   K   K   E   F   H   I   V   A   E   T   G
                                           1   2   3   4   5   6   7   8   9  10  11  12  13
E:\HM_HPr_Au\Models\1QFR_IUPAC.pdb|MODEL   M   E   K   K   E   F   H   I   V   A   E   T   G
                                           1   2   3   4   5   6   7   8   9  10  11  12  13
E:\HM_HPr_Au\Models\2HID_IUPAC.pdb|MODEL   A   Q   K   T   F   K   V   T   A   D   -   S   G
                                           1   2   3   4   5   6   7   8   9  10   0  11  12



# E:\HM_HPr_Au\Models\1PTF_IUPAC.pdb|MODEL
# MEQNSYVIIDETGIHARPATMLVQTASKFDSDIQLEYNGKKVNLKSIMGVMSLGVGKDAEITIYADGSDESDAIQAISDVLSKEGLTK
# MEKKEFHIVAETGIHARPATLLVQTASKFNSDINLEYKGKSVNLKSIMGVMSLGVGQGSDVTITVDGADEAEGMAAIVETLQKEGLA-
# Sequence Identity: 64.77%
# Score: 848.00

# E:\HM_HPr_Au\Models\1QFR_IUPAC.pdb|MODEL
# MEQNSYVIIDETGIHARPATMLVQTASKFDSDIQLEYNGKKVNLKSIMGVMSLGVGKDAEITIYADGSDESDAIQAISDVLSKEGL-TK
# MEKKEFHIVAETGIHARPATLLVQTASKFNSDINLEYKGKSVNLKSIMGVMSLGVGQGSDVTITVDGADEAEGMAAIVETLQKEGLAEQ
# Sequence Identity: 64.04%
# Score: 849.00

# E:\HM_HPr_Au\Models\2HID_IUPAC.pdb|MODEL
# MEQNSYVIIDETGIHARPATMLVQTASKFDSDIQLEYNGKKVNLKSIMGVMSLGVGKDAEITIYADGSDESDAIQAISDVLSKEGLTK
# AQKTFKVTAD-SGIHARPATVLVQTASKYDADVNLEYNGKTVNLKSIMGVVSLGIAKGAEITISASGADENDALNALEETMKSEGLGE
# Sequence Identity: 60.23%
# Score: 796.00
```

Abbildung 23. Das Sequenzalignment für die drei Modelle
an die Zielstruktur HPr *S. aureus* (durchgeführt mit AUREMOL).

Danach wurden *Restraints* für Atomabstände, Diederwinkel und Wasserstoffbrücken erzeugt.

Parameter zur Erzeugung der *Restraints*	
Konfidenzniveau	99,9 %
Abstände	
Abstandsbereich	0,18 nm – 1,00 nm
Ausgewählte Atome	N, C
Anzahl	3151
Winkel	
Ausgewählte Winkel	ψ, ϕ, χ_1, χ_2, χ_{21}, χ_{22}, χ_3, χ_{31}, χ_{32}, χ_4, χ_5, χ_6
Anzahl	369
Wasserstoffbrücken	
Donatoren	HN, H_γ, $H_{\eta 11}$, $H_{\eta 12}$, $H_{\eta 22}$, $H_{\zeta 1}$, $H_{\zeta 2}$, $H_{\zeta 3}$,
Akzeptoren	O, $O_{\delta 1}$, $O_{\delta 2}$, $O_{\varepsilon 2}$, N, $N_{\eta 1}$, $N_{\eta 2}$, $N_{\delta 2}$
Schwellwert	50 % der Modelle müssen W'brücke haben
Anzahl	32

Tabelle 2: Die verwendeten Parameter für PERMOL zur
Erzeugung der *Restraints* zur Modellierung von HPr *S. aureus*.

Die Parameter zur Erzeugung der *Restraints* aus den drei Modellen sind in Tabelle 2 zu finden. Es wurden die Rückgratatome N und C bis 1 nm und alle verfügbaren Diederwinkel von Rückgrat und Seitenketten und Wasserstoffbrücken, die in mindestens aller 50 % der Modelle vorkommen, verwendet. Für die Wasserstoffbrücken wurden alle Donatoren und Akzeptoren zugelassen. Es wurden nach der Mittelung und nach der Anwendung des Thresholds aber nur H^N-O-Wasserstoffbrücken erhalten, die hauptsächlich für die Sekundärstrukturbildung verantwortlich sind. Insgesamt wurden 3151 Abstand-*Restraints*, 369 Diederwinkel-*Restraints* und 32 Wasserstoffbrücken-*Restraints* gewonnen.
Die so erhaltenen *Restraints* wurden anschließend zur Strukturberechnung in DYANA 1.5 [26] verwendet. Es wurden 1000 Strukturen berechnet, von denen die zehn Besten zur weiteren Untersuchung herangezogen wurden.

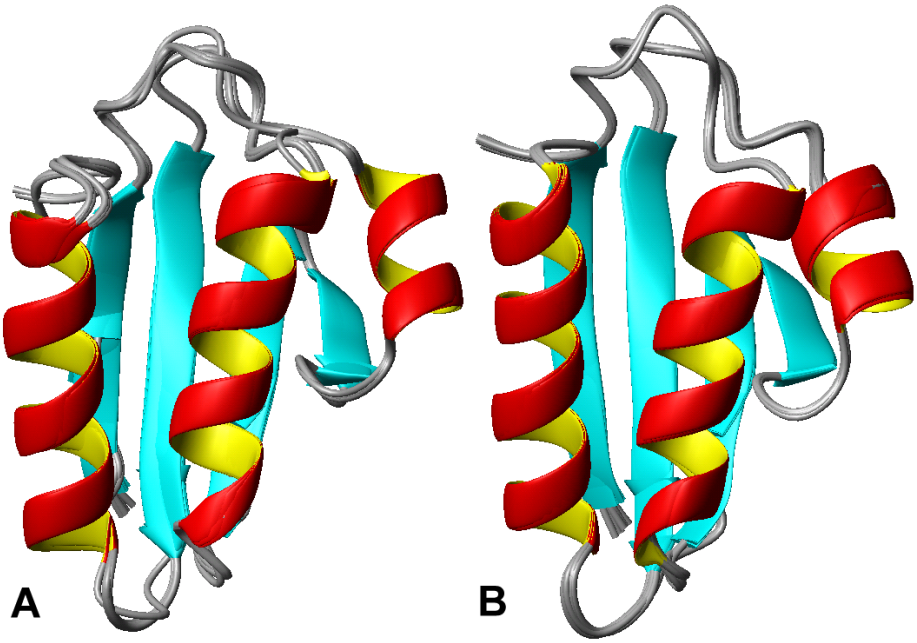

Abbildung 24: Die modellierten Strukturen (A) und die Originalstruktur (B) von HPr *S. aureus*. In der Modellstruktur sind alle Sekundärstrukturelemente wohl definiert.
Der mittlere RMSD zur Originalstruktur beträgt 0,143 nm.

Diese zehn besten Strukturen zeigten eine gute Konvergenz mit einem RMSD innerhalb des Bündels von 0,018 nm. Der mittlere RMSD der Modellstrukturen zu den Originalstrukturen beträgt 0,143 nm. Die wohlbekannten Sekundärstrukturelemente (ein viersträngiges β-Faltblatt und drei α-Helizes), die allen bisher untersuchten HPr-Molekülen gemeinsam sind, zeigen sich auch in den modellierten Strukturen sehr gut bestimmt. Eine Analyse mit

PROCHECK-NMR [41] ergab, dass die meisten Diederwinkel (94,8 %) des Rückgrats in die *most favored* und *additionally allowed* Regionen des *Ramachandran-Plots* fallen.

Modellierung von HPr *S. aureus*	**Original (1KA5)**	**Modell**
RMSD innerhalb der Bündel (N) [nm]	0,016	0,018
RMSD zur Originalstruktur [nm]		0,143
Ramachandran m.f. + a. [%]	98,7	94,8
Most favored [%]	75,6	67,9
Additional allowed [%]	23,1	26,9
Generously allowed [%]	1,3	5,1
Disallowed [%]	0,0	0,0

Tabelle 3: Qualitätswerte für die modellierte Struktur von HPr *S. aureus*.
RMSD-Werte aus MOLMOL und *Ramachandran-Plot* aus PROCHECK-NMR.

4.1.2 Modellierung der Punktmutante HPr *S. aureus* (H15A)

Ein weiteres Beispiel ist die Modellierung der Punktmutante HPr *Staphylococcus aureus* (H15A) aus der bereits mittels NMR gelösten Struktur von HPr *Staphylococcus aureus* (PDB-Code: 1KA5) [109]. In der PDB-Datei sind 16 Strukturen vorhanden, deren RMSD-Wert (N-Rückgrat-Atome) bei 0,016 nm lag und damit sehr gut bestimmt ist. In der Zielsequenz der Mutante ist HIS 15 gegen ALA getauscht. Bei der *Restraint*-Berechnung werden deshalb nur die Aminosäuren 1-14 und 16-88 verwendet.

Parameter zur Erzeugung der *Restraints* aus dem NMR-Bündel HPr *S. aureus*	
Konfidenzniveau	99,90 %
Aminosäuren NMR	1-14,16-88
Abstände	
Abstandsbereich Rückgrat	0,18 nm – 1,00 nm
Atome Rückgrat	N,C
Anzahl NMR	1669
Winkel	
Ausgewählte Winkel	ψ, ϕ, χ_1, χ_2, χ_{21}, χ_{22}, χ_3, χ_{31}, χ_{32}, χ_4, χ_5, χ_6
Anzahl NMR	405
Wasserstoffbrücken	
Donatoren	HN, H_γ, $H_{\eta 11}$, $H_{\eta 12}$, $H_{\eta 22}$, $H_{\zeta 1}$, $H_{\zeta 2}$, $H_{\zeta 3}$, $H_{\gamma 1}$

Akzeptoren	O, $O_{\delta 1}$, $O_{\delta 2}$, $O_{\varepsilon 2}$, N, $N_{\eta 1}$, $N_{\eta 2}$, $N_{\delta 2}$
Schwellwert	80 % der Modelle müssen W'brücke haben
Anzahl NMR	40

Tabelle 4: Parameter für PERMOL zur Erzeugung des Modells von HPr *S. aureus* (H15A) aus dem Strukturbündel von HPr *S. aureus* (wt).

Es wurden 1669 Abstands-, 405 Diederwinkel- und 40 Wasserstoffbrücken-*Restraints* gewonnen. Mit diesen Restraints wurden in DYANA 1000 Strukturen gerechnet und die 10 Besten zur weiteren Analyse verwendet.

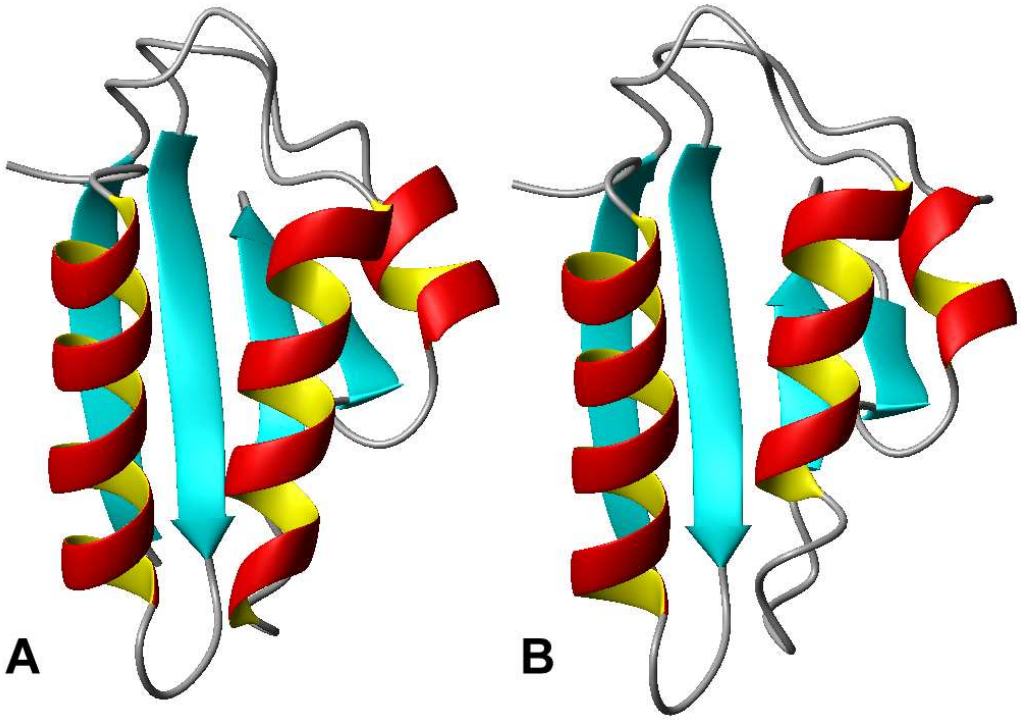

Abbildung 25: Die modellierte (A) und die Original-Struktur (B) von HPr *S. aureus* (H15A).

In Abbildung 25 ist das Ergebnis der Modellierung im Vergleich zur Original-Struktur von HPr *S. aureus* (H15A) gezeigt. Deutlich ist zu sehen, dass alle Sekundärstrukturelemente ausgebildet wurden. Insbesondere hat die Stellung der Seitenkette bei der Mutation ALA 15 die gleiche Ausrichtung wie beim Original (Abbildung 26). In Tabelle 5 sind die Qualitätswerte der modellierten Struktur zu finden. Der RMSD-Wert zur Originalstruktur beträgt im Mittel 0,098 nm. 98, 7 % der Diederwinkel liegen in den erlaubten Bereichen.

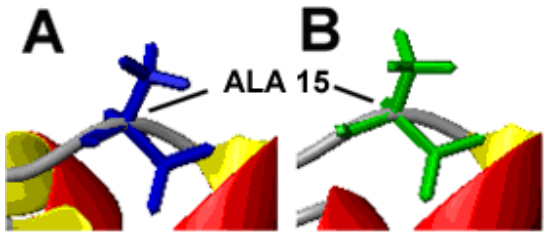

Abbildung 26: Die Stellung der Seitenketten bei der Mutation ALA 15 wurde bei der Modellierung (A) richtig gefunden im Vergleich zur Original-Struktur (B) von HPr *S. aureus* (H15A).

Modellierung von HPr *S. aureus* (H15A)	**Modell**
RMSD innerhalb der Bündel (N) [nm]	0,006
RMSD zur Originalstruktur [nm]	0,098
Ramachandran m.f. + a.	98,7
Most favored [%]	79,5
Additional allowed [%]	19,2
Generously allowed [%]	1,3
Disallowed [%]	0,0

Tabelle 5: Qualitätswerte für die modellierte Struktur von HPr S. aureus (H15A).
RMSD-Werte aus MOLMOL und Ramachandran-Plot aus PROCHECK-NMR.

Weitere Ergebnisse des Moduls PERMOL sind in den nachfolgenden Kapiteln zu sehen. Bei der Strukturverbesserung von Proteinen (Kapitel 4.2) ist PERMOL maßgeblich bei der Gewinnung der notwendigen *Restraints* beteiligt. Bei der Vorstellung der Ergebnisse des ASSIGN Algorithmus in Kapitel 4.3 wird die oben modellierte Mutante HPr *S. aureus* (H15A) verwendet.

4.2 Verbesserung von Proteinstrukturen (ISIC [71])

4.2.1 Verbesserung der Lösungsstruktur von Byr2

Der AUREMOL-ISIC-Algorithmus wurde zur Strukturverbesserung der Ras-Bindedomäne Byr2 eingesetzt. Verfügbar ist je ein Satz von 10 NMR-Strukturen in Lösung [113] und eine einzelne Röntgen-Struktur von Byr2 im Komplex mit Ras [114]. Das sequentielle *Assignment* der NMR-Signale von Byr2 und die experimentellen Parameter sind beschrieben in [115;115]. Im vorliegenden Fall wurden die Informationen der Röntgenstruktur als Quelle benutzt, um die NMR-Struktur S_1 zu verbessern.

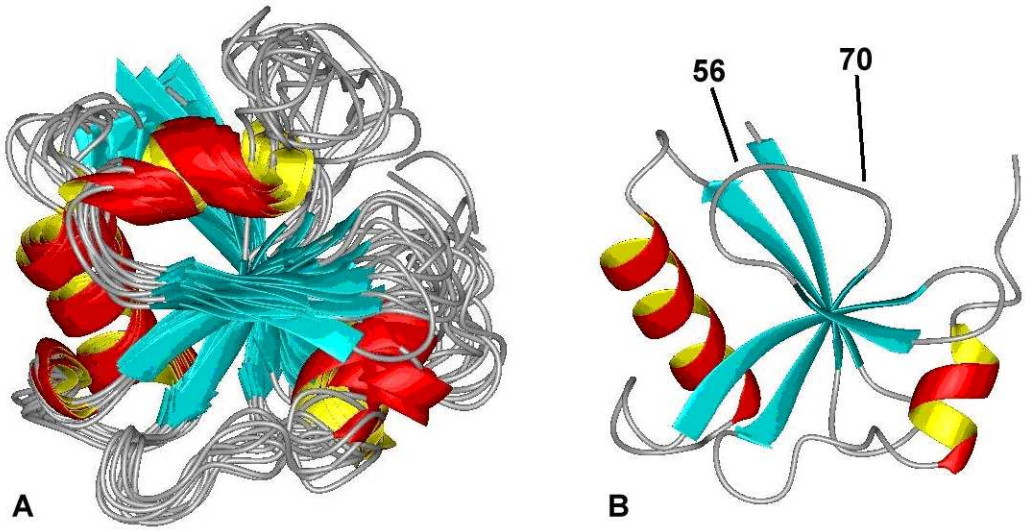

Abbildung 27: (A) 10 NMR-Strukturen der Ras-Bindedomäne Byr2 (S_1). Deutlich zu sehen ist, dass die mittlere Qualität der Strukturen die globale Faltung zwar erkennen lassen, das Bündel aber weit streut. Die C-terminale α-Helix ist sehr schlecht definiert. (B) Die Röntgen-Struktur von Byr2 (S_2). Die C-terminale α-Helix ist besser ausgebildet als beim NMR-Bündel. Zu beachten: Für die Reste 57-69 lagen keine Röntgenstrukturdaten vor.

Wie oben beschrieben und mit den Parametern in Tabelle 6 wurden aus der einzelnen Röntgenstruktur Abstand- und Winkel-*Restraints* erzeugt.

Parameter zur Erzeugung der *Restraints* aus der Röntgen-Struktur (S_2)	
Konfidenzniveau	99,00 %
Abstände	
Abstandsbereich	0,18 nm – 1,00 nm
Ausgewählte Atome	N, C
Abstandsbereich	0,18 nm – 0,60 nm

Ausgewählte Atome	C^{α}, C^{β}, C^{γ}, C^{δ}, C^{ε}, C^{ζ}, O
Anzahl	5248
Winkel	
Ausgewählte Winkel	ψ, ϕ, χ_1, χ_2, χ_{21}, χ_{22}, χ_3, χ_{31}, χ_{32}, χ_4, χ_5, χ_6
Anzahl	321

Tabelle 6: PERMOL-Parameter zur Erzeugung der Distanz- und Winkel-*Restraints* aus der Röntgen-Struktur (S_2) von Byr2, die dann in einer MD-Rechnung verwendet wurde um das Röntgen-Bündel (S_2^X) zu erhalten.

Insgesamt wurden 5248 Abstand-*Restraints* und 341 Winkel-*Restraints* erhalten, die den *Restraint*-Satz R_2^{X*} definieren. Zu beachten ist hierbei, dass es für die Aminosäuren 57-69 keine *Restraints* gibt, weil diese Aminosäuren bei der Röntgenstrukturanalyse nicht sichtbar waren. Mit diesen *Restraints* wurden dann mit DYANA 1.5 [26] 1000 Strukturen berechnet und die zehn Besten hinsichtlich der DYANA Zielfunktion ausgewählt. Diese zehn Strukturen stellen das Strukturbündel S_2^X dar und repräsentieren nun die Röntgenstrukturdaten. Abbildung 28 (A) und (B) zeigen im Vergleich die originale Röntgenstruktur und das zugehörige Strukturbündel S_2^X.

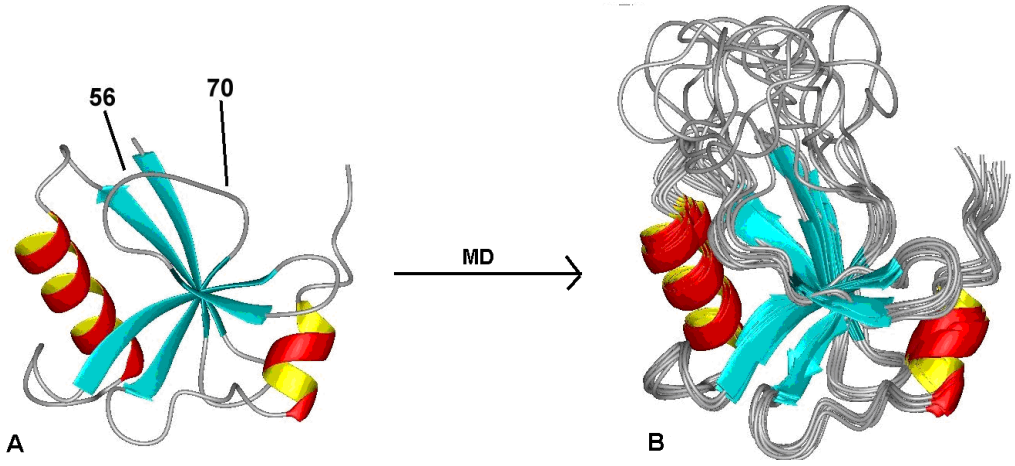

Abbildung 28: (A) Die einzelne Röntgen-Struktur von Byr2 (S_2). (B) Das mit dem *Restraint*-Satz R_2^{X*} mittels MD berechnete Strukturbündel S_2^X.

Wie weiter oben beschrieben, wurde aus dem Bündel S_2^X der *Restraint*-Satz R_2^* generiert, der aus 5600 Abstand-*Restraints*, 396 Diederwinkel-*Restraints* und 53 Wasserstoffbrücken-*Restraints* besteht. Die zugehörigen Parameter sind in Tabelle 7 aufgelistet. Das Bündel der

10 NMR-Strukturen in Lösung bildet S_1 (Abbildung 27(A)). Hieraus wurden 6642 Abstand-*Restraints*, 453 Diederwinkel-*Restraints* und 106 Wasserstoffbrücken-*Restraints* gewonnen. Diese definieren den Führungs-*Restraint*-Satz $R_1 = R_1^*$. Die zugehörigen Parameter sind ebenfalls in
Tabelle 7 Tabelle 7 zu finden.

Parameter zur Erzeugung der *Restraints* aus dem NMR-Bündel (S_1) und aus dem Röntgen-Bündel (S_2^X)	
Konfidenzniveau	99,90 %
Aminosäuren NMR	1-95
Aminosäuren Röntgen	1-56, 70-95
Abstände	
Abstandsbereich Rückgrat	0,18 nm – 1,00 nm
Atome Rückgrat	N, C
Abstandsbereich Seitenketten	0,18 nm – 0,60 nm
Atome Seitenketten	HN, H^{α}, $H^{\alpha 2}$, $H^{\alpha 3}$, H^{β}, $H^{\beta 1}$, $H^{\beta 2}$, $H^{\beta 3}$, H^{γ}, $H^{\gamma 2}$, $H^{\gamma 3}$, $H^{\gamma 1}$, H^{δ}, $H^{\delta 1}$, $H^{\delta 2}$, $H^{\delta 3}$, H^{ε}, $H^{\varepsilon 2}$, $H^{\varepsilon 3}$, $H^{\varepsilon 1}$
Anzahl NMR	6642
Anzahl Röntgen	5600
Winkel	
Ausgewählte Winkel	ψ, ϕ, χ_1, χ_2, χ_{21}, χ_{22}, χ_3, χ_{31}, χ_{32}, χ_4, χ_5, χ_6
Anzahl NMR	453
Anzahl Röntgen	396
Wasserstoffbrücken	
Donatoren	HN, H^{γ}, $H^{\eta 11}$, $H^{\eta 12}$, $H^{\eta 22}$, $H^{\zeta 1}$, $H^{\zeta 2}$, $H^{\zeta 3}$, $H^{\gamma 1}$
Akzeptoren	O, $O^{\delta 1}$, $O^{\delta 2}$, $O^{\varepsilon 2}$, N, $N^{\eta 1}$, $N^{\eta 2}$, $N^{\delta 2}$
Anzahl NMR	106
Anzahl Röntgen	53

Tabelle 7: PERMOL-Parameter zur Erzeugung der Distanz-, Winkel- und Wasserstoffbrücken-*Restraints* vom NMR-Bündel (S_1) und vom Röntgen-Bündel (S_2^X) von Byr2, die dann für die Kombination verwendet wurden.

Im nächsten Schritt wurden die *Restraint*-Sätze R_1^* und R_2^* kombiniert. Die Parameter hierzu finden sich in Tabelle 8. Für den Fall, dass die *Restraints* nicht zusammenpassen wurde nur der NMR-*Restraint* behalten. Nach der Kombination definierten 6642 Abstand-*Restraints*, 338 Diederwinkel-*Restraints* und 26 Wasserstoffbrücken-*Restraints* den *Restraint*-Satz R_0.

Parameter für die *Restraint*-Kombination	
Winkelfilter	*Favoured regions*, GLY, PRO, CHI1-CHI2: < level 2
W'brücken Grenzwert	0,75 %
W'brücken Austausch	0,90 %
Signifikanzniveau	0,2 %
Anzahl der erhaltenen *Restraints*	
Abstände	6642
Winkel	338
Wasserstoffbrücken	26

Tabelle 8: Parameter für die Restraint-Kombination von R_1^* und R_2^* und die jeweilige Anzahl der erhaltenen Restraints für Byr2.

Mit Hilfe dieses *Restraint*-Satzes R_0 wurden wieder 1000 Strukturen mit DYANA 1.5 berechnet und wieder die zehn Besten hinsichtlich der DYANA-Zielfunktion selektiert. Das Resultat war das Strukturbündel S_0 (Abbildung 29 (A)), das nun zur weiteren Analyse verwendet wurde. Die Strukturen S_0 wurden in explizitem Lösungsmittel (Wasser) [36;116] verfeinert und das Ergebnis war einen Satz von zehn Strukturen (S_{0_WR}) der Byr2-RBD (Abbildung 29 (B)).

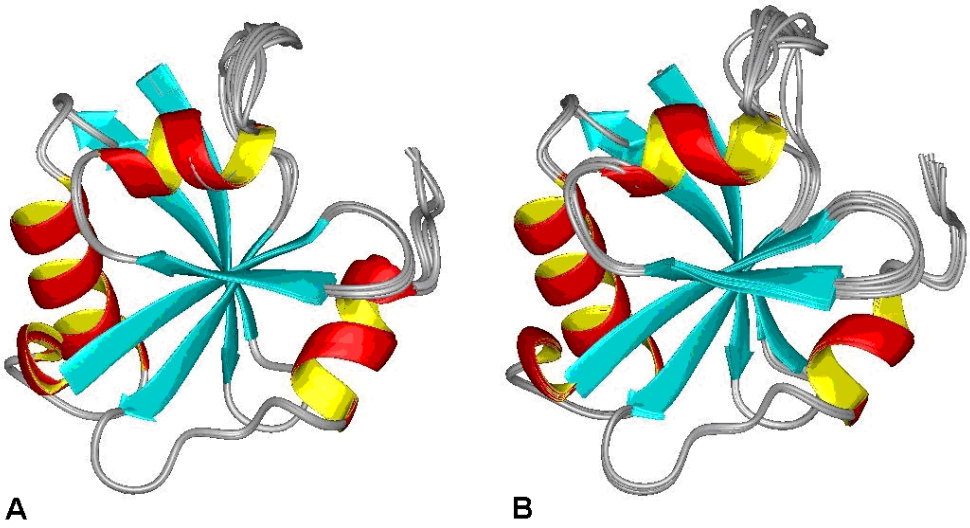

Abbildung 29: Die verbesserten Strukturen S_0 (A) und S_{0_WR} (B). Die C-terminale α-Helix ist gut definiert und die Strukturen der Bündel liegen deutlich näher zusammen.

Bei den Strukturen sind sämtliche Sekundärstrukturelemente wohl definiert. Insbesondere ist die C-terminale α-Helix, die in den originalen NMR-Strukturen schlecht charakterisiert ist, nun sehr gut definiert. Zusätzlich wurde die Qualität der erhaltenen Strukturen mit den originalen NMR- und Röntgenstrukturen verglichen (Tabelle 9).

Hierzu wurden RMSD-Werte, *Ramachandran-Plots* und AUREMOL-R-Werte herangezogen. Die Ergebnisse zeigen eine deutliche Verbesserung der Werte für die verfeinerte Struktur. Der RMSD-Wert (N-Rückgrat-Atome) der neu berechneten Strukturen ist drastisch von 0,144 nm auf 0,033 nm gesunken. Der prozentuale Anteil der Aminosäuren in den *most favored* und *additional allowed* Regionen des *Ramachandran-Plots* ist im Vergleich zu den zwei Eingabestrukturen (S_1 and S_2) gestiegen. Das Ziel war, verbesserte Strukturen in Lösung zu erhalten. Deshalb wurden die erhaltenen Strukturen daraufhin untersucht, ob sie die experimentellen Daten tatsächlich besser erklären als die Ausgangsstrukturen. Ein geeigneter Test hierfür ist die Berechnung des AUREMOL R-Wertes (R_5) [39], der ein experimentelles NMR-NOESY-Spektrum direkt mit dem zugehörigen zurückgerechneten Spektrum vergleicht. Es wurde ein 2D-^{1}H-NOESY Spektrum verwendet, das mit einer Mischzeit von 100 ms gemessen wurde. Bei der Rückrechnung dient die zu untersuchende Struktur als Eingabe (Im vorliegenden Fall waren das eben die Originalstruktur und die verbesserte Struktur). Für die Berechnungen, deren Ergebnisse in Tabelle 9 gezeigt sind, wurde der AUREMOL R-Wert (R_5), wie in Kapitel 2.3.1 beschrieben, verwendet, der Unterschiede in der Struktur zuverlässig beschreibt [39]. Die AUREMOL R-Werte (R_5) zeigen ebenfalls eine signifikante Verbesserung für die neuen Strukturen. Das zeigt klar, dass verbesserte Strukturen erhalten werden können, wenn Daten aus zusätzlichen Quellen in der oben beschriebenen Weise zu Hilfe genommen werden.

Verbesserung von Byr2	**S_1 (NMR)**	**S_2 (Röntgen)**	**S_0**	**S_{0_WR}**
AUREMOL R-Wert (R_5)	0,534	-	0,455	0,451
RMSD MOLMOL (N) [nm]	0,144	0,067	0,026	0,033
Ramachandran m.f. + a. [%]	87,3	88,5	94,3	90,8
Most favored [%]	67,8	70,1	71,3	78,2
Additional allowed [%]	19,5	18,4	23,0	12,6
Generously allowed [%]	11,5	8,0	4,6	8,0
Disallowed [%]	1,1	3,4	1,1	1,1

Tabelle 9: Qualitätswerte von AUREMOL und Procheck für Eingabestrukturen S_1 und S_2 und für die mit ISIC verbesserten Strukturen von Byr2 S_0 und S_{0_WR}.

In Abbildung 30 ist die Häufigkeitsverteilung der Gesamtenergie aller NMR-Strukturen von Byr2, aus der das Bündel S_1 stammt, aufgetragen. Es wurden 1000 Strukturen mit den Original-Restraints in CNS [27] gerechnet.

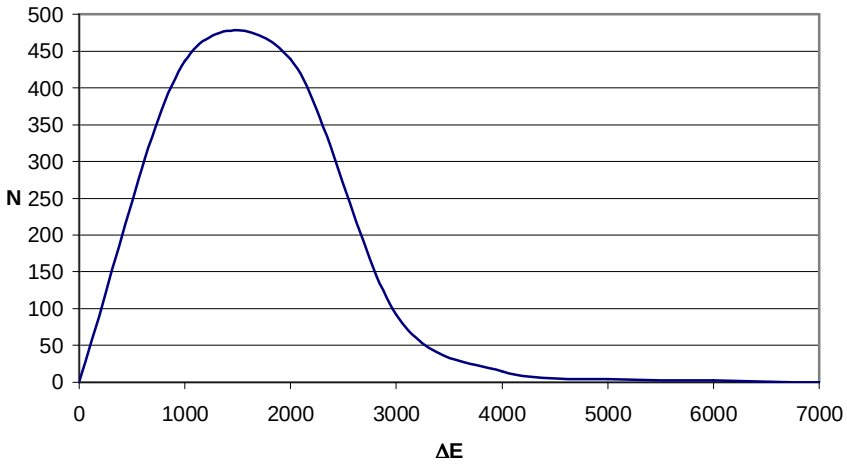

Abbildung 30: Die Verteilung der Gesamtenergien aller NMR-Strukturen von Byr2, aus der auch die Eingabestruktur S_1 stammt.

Mit Hilfe der Wahrscheinlichkeitsbeziehung nach [117]

$$P(\Delta E) = e^{-\frac{\Delta E}{k_B T}} \tag{4.1}$$

wurden alle Strukturen mit einer Wahrscheinlichkeit kleiner als 0,1 verworfen. Zudem wurden alle Strukturen, in denen mehr als 10 % der Abstand-*Restraints* verletzt sind, weggelassen. Die restlichen Strukturen stellen das Grundbündel der NMR für Byr2 dar. In Abbildung 31 ist dieses Grundbündel als *Sausage-Plot* in blauer Farbe dargestellt. In grün bzw. rot sind die mit ISIC verbesserten Strukturbündel S_0 bzw. S_{0_WR} ebenfalls als *Sausage-Plot* gezeigt. Es ist deutlich zu sehen, dass die ISIC-Strukturen sehr gut im Grundbündel der NMR liegen und zudem viel besser definiert sind. Dies zeigt, dass die mit ISIC erhaltenen Lösungen sich im Lösungsraum der von der NMR vorgegebenen Konformation befindet.

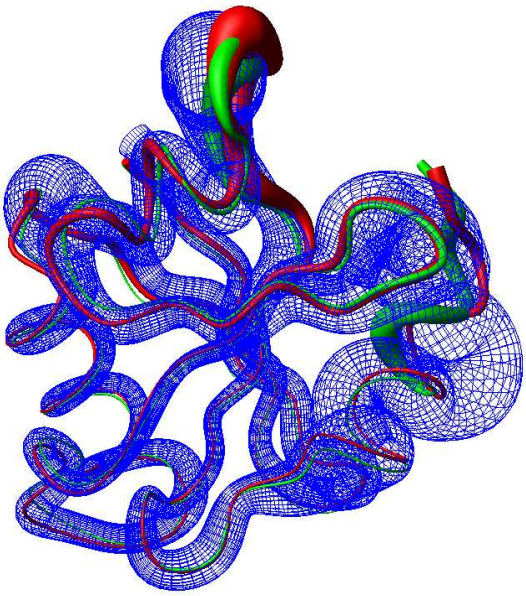

Abbildung 31: Blau dargestellt ist das NMR Grundbündel, die Lösungsstrukturen von ISIC sind in grün (S_0) und rot (S_{0_WR}) gezeigt.

4.2.2 Strukturverbesserung der Ras-Bindedomäne RalGDS-RBD

Als zweiter Testfall wurde die Ras-Bindedomäne von RalGDS (RalGDS-RBD) des Menschen verwendet. Die Lösungsstruktur (Aminosäuren 1-97, entsprechend den Aminosäuren 788-884 des Gesamtproteins, Swiss prot accession code: Q12967) wurde bereits publiziert [118]. Für den Test wurde ein kürzeres Teilstück (Aminosäuren 11 bis 97) einer schlecht aufgelösten NMR Struktur verwendet, bei der nur relativ leicht zu gewinnende NMR Daten wie Wasserstoffbrücken, Diederwinkel und Rückgrat-Atomabstände verwendet wurden. Zur Verbesserung wurde eine Röntgenstruktur mit mittlerer Qualität (Auflösung 3,4 Å) von RalGDS im Komplex mit Ras [119] herangezogen.

Die schlecht aufgelösten NMR-Strukturen von RalGDS-RBD (Aminosäuren 11-97) wurden mit den leicht zu erhaltenen NMR-Daten neu berechnet. Dies waren 25 Wasserstoffbrücken, 102 Φ- und Ψ- Diederwinkel und 232 Rückgrat-Abstände zwischen H^N- und H^α-Atomen.

Mit diesen *Restraints* wurde mit DYANA 1.5 300 Strukturen berechnet und die zehn besten im Hinblick auf die DYANA Zielfunktion zur weiteren Verwendung als NMR Eingabestrukturen S_1 herangezogen.

Wie oben beschrieben und mit den Parametern aus Tabelle 10 wurden aus der originalen Röntgen-Struktur 2001 Abstand- und 263 Diederwinkel-*Restraints* erzeugt, die den *Restraint*-Satz R_2^{X*} bildeten. Zu beachten ist hierbei, dass für die Aminosäuren 11, 50-55, 78-89 und 97

keine *Restraints* gewonnen werden konnten, da diese in der originalen Röntgenstruktur nicht zu sehen waren. Mit Hilfe dieser *Restraints* wurden nun mit DYANA 1.5 1000 Strukturen berechnet und wiederum die zehn Besten ausgewählt, die dann das Strukturbündel S_2^X darstellten.

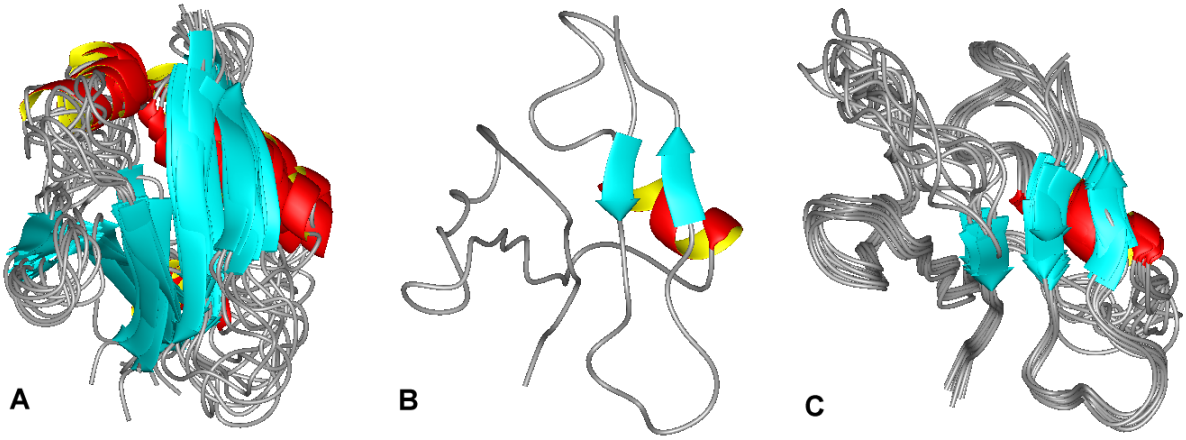

Abbildung 32: Die mit leicht zu erhaltenden NMR Daten neu berechnete NMR Struktur (A), die originale Röntgen Struktur (B) und das daraus erzeugte Bündel (C) von RalGDS.

Parameter zur Erzeugung der *Restraints* aus der Röntgen-Struktur (S_2)	
Konfidenzniveau	99,90 %
Abstände	
Abstandsbereich	0,18 nm – 1,00 nm
Ausgewählte Atome	N, C
Abstandsbereich	0,18 nm – 0,60 nm
Ausgewählte Atome	C_β, C_γ, C_δ, C_ε, C_ζ
Anzahl	2001
Winkel	
Ausgewählte Winkel	ψ, ϕ, χ_1, χ_2, χ_{21}, χ_{22}, χ_3, χ_{31}, χ_{32}, χ_4, χ_5, χ_6
Anzahl	263

Tabelle 10: PERMOL-Parameter zur Erzeugung der Distanz- und Winkel-*Restraints* aus der Röntgen-Struktur (S_2) von RalGDS, die dann in einer MD-Rechnung verwendet wurde um das Röntgen-Bündel (S_2^X) zu erhalten.

Mit den Parametern aus Tabelle 11 wurden aus Strukturbündel S_2^X der *Restraint*-Satz R_2^* erzeugt, der 1784 Abstand-, 326 Diederwinkel- und 13 Wasserstoffbrücken-*Restraints* enthielt. In gleicher Weise wurde aus dem Strukturbündel S_1 2344 Abstands-, 417 Diederwinkel und 70 Wasserstoffbrücken-*Restraints* gewonnen, die nun den führenden *Restraint*-Satz $R_1 = R_1^*$ bildeten.

Parameter zur Erzeugung der *Restraints* aus dem NMR-Bündel (S_1) und aus dem Röntgen-Bündel (S_2^X)	
Konfidenzniveau	99,90 %
Aminosäuren NMR	11-97
Aminosäuren Röntgen	12-49,56-77,90-96
Abstände	
Abstandsbereich Rückgrat	0,18 nm – 1,00 nm
Atome Rückgrat	N
Abstandsbereich Seitenketten	0,5 nm – 1,5 nm
Atome Seitenketten	$H_{\delta 2}$, $H_{\delta 21}$, $H_{\delta 22}$, $H_{\delta 3}$, H_{ε}, $H_{\varepsilon 2}$, $H_{\varepsilon 3}$, $H_{\varepsilon 1}$
Anzahl NMR	2344
Anzahl Röntgen	1784
Winkel	
Ausgewählte Winkel	ψ, ϕ, χ_1, χ_2, χ_{21}, χ_{22}, χ_3, χ_{31}, χ_{32}, χ_4, χ_5, χ_6
Anzahl NMR	417
Anzahl Röntgen	326
Wasserstoffbrücken	
Donatoren	HN, H_{γ}, $H_{\eta 11}$, $H_{\eta 12}$, $H_{\eta 22}$, $H_{\zeta 1}$, $H_{\zeta 2}$, $H_{\zeta 3}$, $H_{\gamma 1}$
Akzeptoren	O, $O_{\delta 1}$, $O_{\delta 2}$, $O_{\varepsilon 2}$, N, $N_{\eta 1}$, $N_{\eta 2}$, $N_{\delta 2}$
Anzahl NMR	70
Anzahl Röntgen	13

Tabelle 11: PERMOL-Parameter zur Erzeugung der Distanz-, Winkel- und Wasserstoffbrücken-*Restraints* vom NMR-Bündel (S_1) und vom Röntgen-Bündel (S_2^X) von RalGDS, die dann für die Kombination verwendet wurden.

Im nächsten Schritt wurden die Sätze R_1^* und R_2^* kombiniert (Tabelle 12). Im Falle von nicht zusammenpassenden *Restraints* wurde nur der zur NMR gehörige *Restraint* weiterverwendet. Nach der Kombination lagen im *Restraint*-Satz R_0 2344 Abstands-, 285 Diederwinkel- und 27 Wasserstoffbrücken-*Restraints* vor. Mit diesem Satz wurden nun wiederum mit DYANA 300 Strukturen berechnet und die zehn Besten (S_0) zur Analyse ausgewählt.

Parameter für die *Restraint*-Kombination	
Winkelfilter	*Favored regions*, GLY, PRO, CHI1-CHI2: < level 2
W'brücken Grenzwert	0,75 %
W'brücken Austausch	0,90 %
Signifikanzniveau	0,2 %
Anzahl der erhaltenen *Restraints*	
Abstände	2344
Winkel	285
Wasserstoffbrücken	27

Tabelle 12: Parameter für die *Restraint*-Kombination von R_1^* und R_2^* und die jeweilige Anzahl der erhaltenen *Restraints von RalGDS*.

Alle Sekundärstrukturelemente der ISIC-Strukturen sind wohl definiert. Insbesondere die beiden α-Helizes, die in den NMR Eingabedaten sehr schlecht definiert sind, sind wesentlich besser bestimmt.

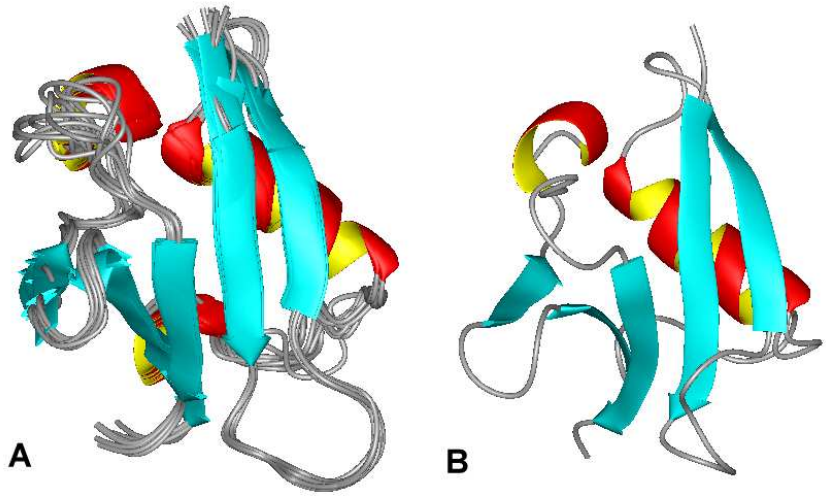

Abbildung 33: Die mit ISIC verbesserte Struktur (A) und die hoch aufgelöste NMR Struktur (B) im Vergleich. Es ist zu sehen, dass die ISIC-Strukturen gut definierte Sekundärstrukturelemente ausbilden, die im Vergleich zur hoch aufgelösten NMR-Struktur gleiche Orientierungen einnehmen.

Die Qualität des Ramachandran-Plots blieb gleich. Der RMSD-Wert der N Rückgratatome (0,06 nm) konnte drastisch im Vergleich zum NMR Eingabestrukturbündel (0,21 nm) reduziert werden. Ebenso konnte der AUREMOL R-Wert (R_5) von 0,383 auf 0,353 verbessert werden.

Verbesserung von RalGDS	S_1 (NMR)	S_2 (Röntgen)	S_0
AUREMOL R-Wert (R_5)	0,383	-	0,353
RMSD innerhalb der Bündel N [nm]	0,21	0,13	0,06
Ramachandran m.f. + a. [%]	91,3	74,4	88,8
Most favored [%]	72,8	36,7	72,8
Additional allowed [%]	18,5	38,0	16,0
Generously allowed [%]	6,2	16,5	7,4
Disallowed [%]	2,5	8,9	3,7

Tabelle 13: Qualitätswerte von AUREMOL und PROCHECK-NMR für Eingabestrukturen S_1 und S_2 und für die mit ISIC verbesserten Strukturen S_0 von RalGDS.

4.2.3 Stabilität am Beispiel der Immunoglobulin-Binde-Domäne

Die größte Gefahr bei derartigen Methoden ist, dass eine mögliche Beeinflussung zu Gunsten der Struktur auftritt, die zum Verbessern herangezogen wird. Im schlimmsten Fall erhält man statt einer verbesserten Struktur genau die Struktur, die man zum Verbessern herangezogen hat. Um genau diesen Effekt zu untersuchen, wurde ISIC dahingehend überprüft, dass zwei Strukturen verwendet wurden, die klare strukturelle Unterschiede besitzen. Die Lösungsstruktur der B2 Immunglobulin-Binde-Domäne von *Streptococcal* Protein G [120;120] unterscheidet sich deutlich von der Röntgen-Struktur [121]. Die NMR-Struktur wurde aus einer Dimer-Form des Proteins erhalten. Vier *Core*-Mutationen führten zu der Dimerisierung des Proteins und zu einem Austausch eines β-Faltblatts. Abbildung 34 zeigt eine Hälfte der Dimer NMR Struktur (Abbildung 34 (A)) im Vergleich zur Monomer Röntgen-Struktur (Abbildung 34 (B)) der B2 Domäne. Klar zu sehen ist, dass die Orientierung der beiden letzten β-Faltblätter völlig unterschiedlich ist. Ein einfacher Mittelungsprozess zwischen diesen beiden Strukturbündeln führt zu einem absolut falschen Strukturbündel und nicht zu einer Verbesserung. Der ISIC-Algorithmus berücksichtigt derartige Strukturunterschiede automatisch und wurde wie oben beschrieben verwendet. Die Parameter zur *Restraint*-Gewinnung sind die gleichen wie in Tabelle 6 und Tabelle 7. Aus der zu verbessernden NMR-Struktur (Abbildung 34 (B)) wurden 2948 Distanz-, 260 Winkel- und 41 Wasserstoffbrücken-*Restraints* gewonnen (R_1^*). Der RMSD beträgt 0,022 nm. Der *Ramachandran-Plot* liefert 90 % in den *most favored* und 10 % in den *additional allowed* Regionen. Weiterhin wurde ein Strukturbündel erzeugt, das die Röntgen-Daten interpretiert. Hierzu wurden 1888 Distanz- und 243 Winkel-*Restraints* aus der einzelnen monomeren Röntgen-Struktur (Abbildung 34 (B)) gewonnen und zur Berechnung des Röntgen-

Strukturbündels (Abbildung 34 (C)) verwendet. Aus diesem Bündel wurden dann 2762 Distanz-, 241 Winkel- und 45 Wasserstoffbrücken-*Restraints* gewonnen (R_2^*). Die *Restraint*-Kombination von R_1^* und R_2^* ergab 2948 Distanz-, 224 Winkel- und 26 Wasserstoffbrücken-*Restraints*. Mit diesen *Restraints* wurden nun die endgültigen zehn durch ISIC verbessertet Strukturen berechnet (Abbildung 34(D)). Ihr RMSD für die Rückgratatome ist 0,008 nm. Der *Ramachandran-Plot* liefert 92 % in den *most favored* und 8 % in den *additional allowed* Regionen. Wie leicht zu erkennen ist, sehen die verbesserten Strukturen den originalen NMR-Strukturen sehr ähnlich. Der RMSD-Wert und die *Ramachandran*-Qualität konnten sogar leicht verbessert werden. Zu beachten ist hierbei, dass die originalen NMR Strukturen bereits sehr gut definiert waren.

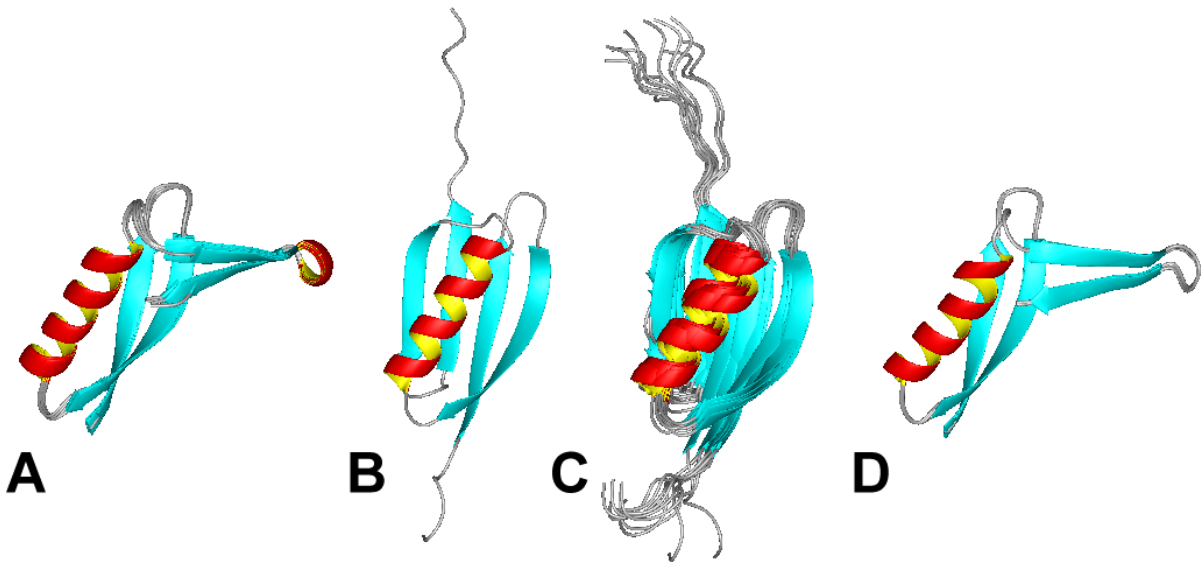

Abbildung 34: Verzerrungsfreie Verbesserung der Imunoglobulin-Binde-Domäne von Protein G.

(A) NMR Struktur (Zehn Monomere der Dimerstruktur)

(B) Einzelne monomere Röntgen-Struktur

(C) Röntgen-Strukturbündel

(D) Zehn endgültige durch ISIC verbesserte Strukturen

4.3 Automatische sequentielle Zuordnung (ASSIGN)

Referenzzuordnung

Für alle Testfälle wurde, um die Qualität der von ASSIGN erstellten Zuordnung zu überprüfen, eine so genannte Referenzzuordnung erstellt. Die *Peaks* des experimentellen NOESY-Spektrums wurden mit der jeweils richtigen Struktur und der richtigen sequentiellen Zuordnung mit Hilfe von PEAKSASSIGN zugeordnet. Aus dieser NOESY-Zuordnung wurde dann wiederum eine auf das NOESY-Spektrum angepasste sequentielle Zuordnung gewonnen. Ein Vergleich der aktuellen Zuordnung von ASSIGN mit dieser Referenzzuordnung gibt ein Maß für die Qualität der mit ASSIGN erhaltenen Zuordnungen.

Parameter zur Erzeugung der Referenzzuordnung	
Spektrum	HPr *S. aureus* (H15A)
Sequentielle Zuordnung	HPr *S. aureus* (H15A)
3D-Struktur	HPr *S. aureus* (H15A)
Max. Radius	0,06 ppm
Min. Radius	0,03 ppm
Rückrechnung	
Max. Distanz	0,5 nm
Mischzeit	0,3 s
Relaxationszeit	1,54 s
Larmorfrequenz	600,13 MHz

Tabelle 14: Parameter für das Modul PEAKASSIGN, mit dem die Referenzzuordnung erstellt wurde. Die sequentielle Zuordnung und die 3D-Struktur von HPr *S. aureus* (H15A) sind in der Arbeitsgruppe gelöst worden (Munte et al., nicht veröffentlicht).

4.3.1 Idealer Datensatz

Um zu zeigen, dass der Algorithmus vom Prinzip her funktioniert, wurde er auf einen idealen Datensatz angewandt. Anstelle eines experimentellen 2D-NOESY-Spektrums wurde ein künstlich erzeugtes Spektrum verwendet. Das verwendete Protein war die in der Arbeitsgruppe bereits gelöste HPr Mutante aus *S. aureus* (H15A). Es lag eine Lösungsstruktur und eine sequentielle Zuordnung vor (Munte et al., nicht veröffentlicht). Unter Verwendung der Parameter des entsprechenden experimentellen Spektrums wurde der ideale Datensatz

simuliert. Die Verwendung eines idealen Datensatzes hat den Vorteil, dass Fehler z. B. durch Artefaktsignale, Basislinienschwankungen, etc. ausgeschlossen werden können und damit die Resultate nur vom Algorithmus und nicht vom verwendeten Spektrum abhängen.

Simuliertes Spektrum als experimentelles Spektrum

Mit Hilfe der Lösungsstruktur von HPr *S. aureus* (H15A) und der entsprechenden sequentiellen Zuordnung wurde ein ^{1}H-2D-NOESY-Spektrum mit einer maximalen Distanz von 0,5 nm simuliert und die dazu gehörende *Peakliste* gelöscht. Für das Spektrum wurde ein Auflösung von 1024×4096 Datenpunkten gewählt. Zudem wurde eine Mischzeit von 0,08 s, ein Relaxationszeit von 1,54 s, eine Resonanzfrequenz von 600,13 MHz und gaußförmige Linienformen verwendet. Ebenso wurde die J-Kopplung und damit die Multiplettstruktur berücksichtigt. Die Parameter wurden entsprechend dem später verwendeten experimentellen Datensatz gewählt. Nach der Simulation wurde das so simulierte Spektrum als experimentelles Spektrum verwendet. Anschließend wurden die *Peaks* des Spektrums neu identifiziert und die entsprechenden Signalvolumen berechnet. Das so erhaltene Spektrum wurde dann mit einer Segmentierungstiefe von 0,01 und eine Segmentierungsbreite von 30 Hz reduziert (siehe Kapitel 3.3.1). Insgesamt waren 4881 experimentelle *Peaks* vorhanden. Die Anzahl der möglichen ppm-Werte (*Slots*) aus den *Peakmaxima* ergab sich zu 1123.

Parameter zur Erzeugung des simulierten Spektrums	
Sequentielle Zuordnung	HPr S. aureus (H15A)
3D-Struktur	HPr S. aureus (H15A)
Max. Distanz	0,5 nm
Mischzeit	0,3 s
Relaxationszeit	1,54 s
Larmorfrequenz	600,13 MHz
Datenpunkte w1, w2	1024, 4096
Linienform	Gauß
Linienverbreiterung w1, w2	9,4 Hz, 5,4 Hz
J-Kopplung	Ja

Tabelle 15: Parameter des Moduls RELAX zur Erzeugung des künstlichen Spektrums. Für den idealen Datensatz wurde das so erzeugte Spektrum als experimentelles Spektrum verwendet. Die Parameter ab „Max. Distanz“ wurden für alle Simulationen gleich verwendet.

Simulierte *Peaks*

Mit derselben sequentiellen Zuordnung und derselben Struktur, mit deren Hilfe das experimentelle Spektrum erstellt wurde, wird nun wiederum ein Spektrum simuliert und die simulierten *Peaks* mit ihren *Peakformen* gespeichert. Es wurden 8473 *Peaks* von 455 Atomen erhalten. Die Anzahl der Peaks ist höher als beim simulierten experimentellen Spektrum, da bei der Signalidentifikation im simulierten experimentellen Spektrum nur Maxima identifiziert werden und so auf Grund von Überlappungseffekten nicht alle Peaks gefunden werden können.

Im Anschluss daran wurde die Optimierung mit ASSIGN gestartet. Die Parameter für den TA wurden empirisch bestimmt und waren 1,5 für den Start-*Threshold*, 0,99 für den Kühlfaktor und für die Kühlschrittweite 50000. Der Lauf wurde ohne einer partiellen Startzuordnung gestartet, d. h. die Startkonfiguration für die Werte aller chemischen Verschiebungen der einzelnen Atome war zufällig, d. h. sie wurden mit einem Zufallsgenerator auf die Slots des experimentellen Spektrums verteilt. Nur die Vorhersage der chemischen Verschiebungen hatte am Anfang einen Einfluss, weil dadurch *Slots*, die näher an den jeweiligen Vorhersagen liegen, bevorzugt ausgewählt werden.

Idealer Datensatz ohne Rauschen	
Exp. Spektrum	Künstlicher Datensatz
3D-Struktur für Simulation	HPr *S. aureus* (H15A)
Zuordnung für Simulation	HPr *S. aureus* (H15A)
Rauschniveau	0 %
Anzahl exp. *Peaks*	4881
Anzahl sim. *Peaks*	8473
Anzahl chem. Verschiebungen	455
Partielle Startzuordnung	0 %
Gefundene Zuordnung	99,34 %

Tabelle 16: Ergebnisse für den idealen Datensatz. Es wurden 99,34 % der sequentiellen Zuordnung richtig gefunden, die partielle Startzuordnung dabei war 0 %.

In Abbildung 35 ist die Häufigkeitsverteilung der Pseudoenergiefunktion der Linienformen *ES* zu sehen. Deutlich ist zu sehen, dass bei der richtigen Zuordnung die über die experimentellen *Peaks* selektierten Messbereiche in den Spektren einen sehr guten Energiewert besitzen. Genauso klar verhält sich eine zufällig verteilte Zuordnung, die meisten Messbereiche zeigen eine schlechte Energie. Aus den beiden Verteilungen wurde dann die

zugehörige Wahrscheinlichkeitsfunktion, wie in Kapitel 3.3.5 beschrieben, gebildet, welche in Abbildung 36 zu sehen ist. Messbereiche, bei denen eine sehr gute bis perfekte Übereinstimmung gefunden wird, haben eine Wahrscheinlichkeit von 1, Pseudoenergiewerte kleiner als 0,85 werden bereits mit einer Wahrscheinlichkeit von weniger als 0,5 bewertet.

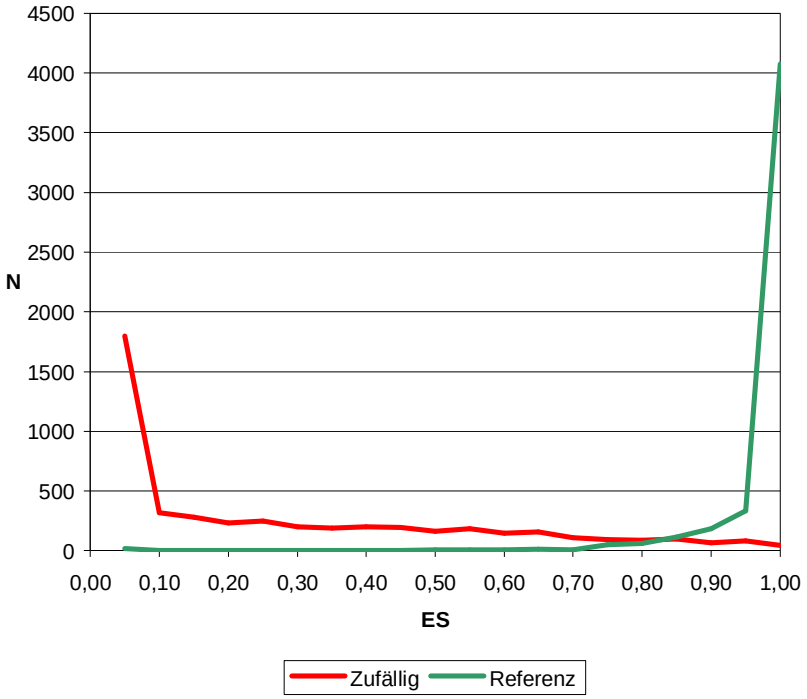

Abbildung 35: Die Häufigkeitsverteilung der Pseudoenergie der Linienformen *ES*. Grün bezeichnet die Referenz, Rot eine zufällige Zuordnung.

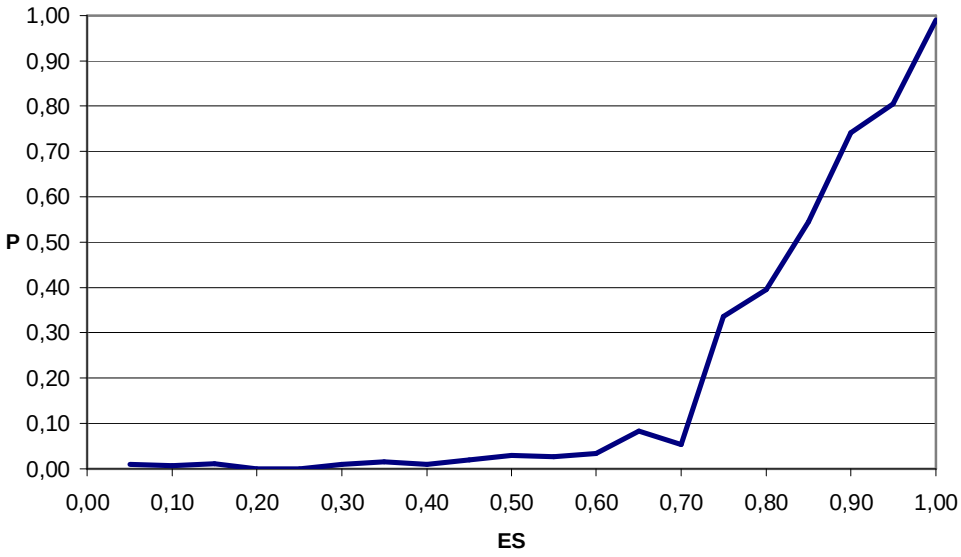

Abbildung 36: Die aus den Häufigkeitsverteilungen resultierende Wahrscheinlichkeitsfunktion.

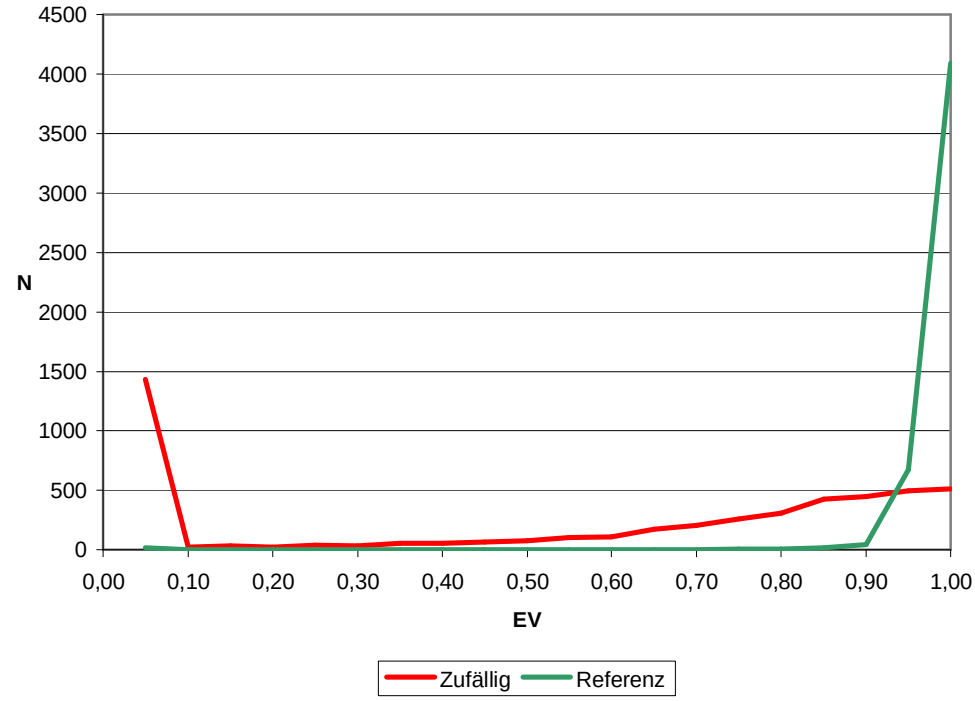

Abbildung 37: Die Häufigkeitsverteilung der Pseudoenergie der Volumen *EV*.
Grün bezeichnet die Referenz, Rot eine zufällige Zuordnung.

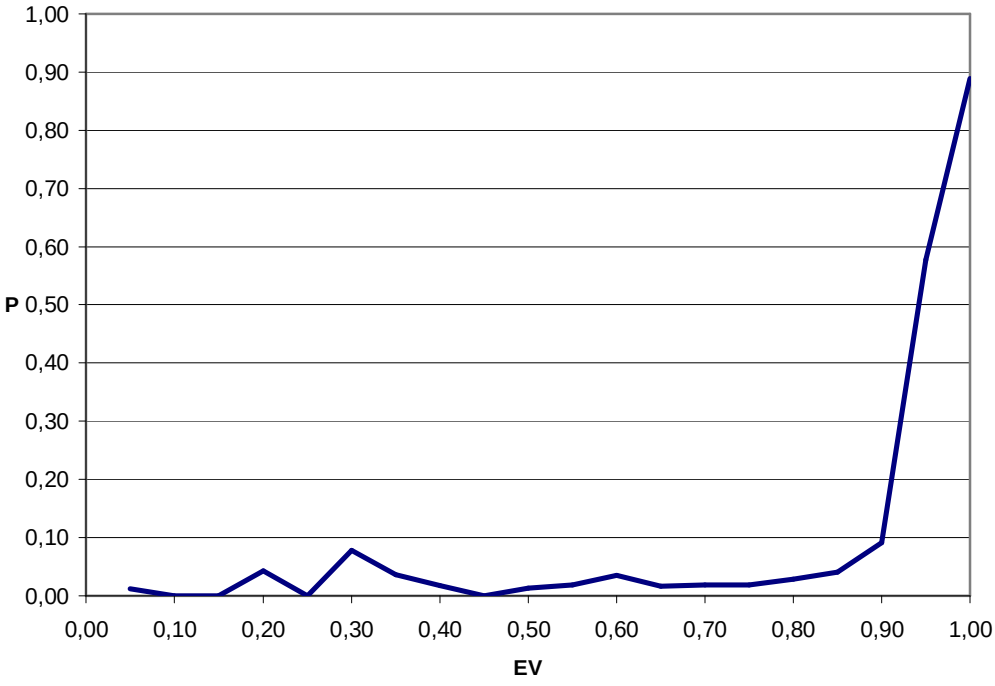

Abbildung 38: Die aus den Häufigkeitsverteilungen resultierende Wahrscheinlichkeitsfunktion.

In Abbildung 37 ist die Häufigkeitsverteilung der Pseudoenergiefunktion der Volumen *EV* dargestellt. Auch hier ist leicht zu erkennen, dass bei der richtigen Zuordnung sehr gute

Energiewerte an den Messbereichen erhalten werden. Die zufällige Zuordnung zeigt, dass schlechte Übereinstimmungen am häufigsten vorkommen. Allerdings gibt es mehr gute Übereinstimmungen als bei der Linienform. Dies führt dazu, dass in der zugehörigen Wahrscheinlichkeitsfunktion, welche in Abbildung 38 zu sehen ist, die Selektivität empfindlicher ist, d. h. dass selbst gute Übereinstimmungen mit einer Pseudoenergie von 0,9 nur eine Wahrscheinlichkeit von 0,1 haben. Selbst eine perfekte Übereinstimmung liefert eine Wahrscheinlichkeit von nur knapp 0,9.

Das Ergebnis des Optimierungslaufes war eine richtige Zuordnung von 99,34 % der chemischen Verschiebungen, d. h. es wurden die chemischen Verschiebungen von 450 (455 insgesamt) Atomen in einem Fehlerbereich von $\pm 0,03$ *ppm* richtig gefunden.

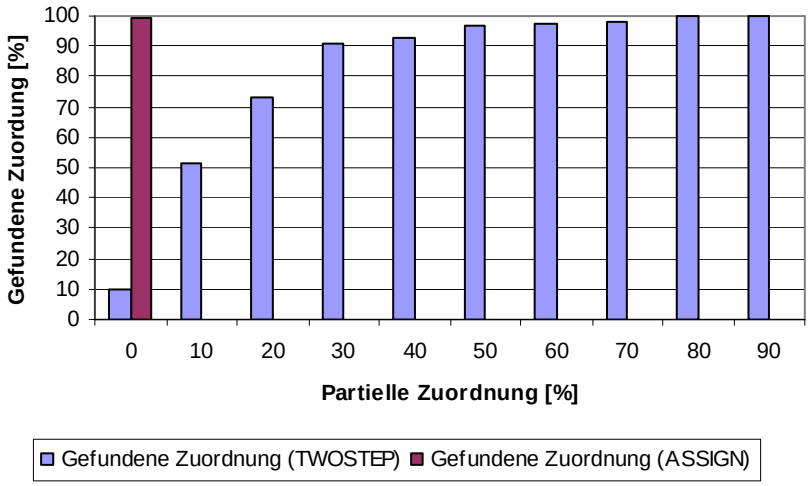

Abbildung 39: Die gefundenen Zuordnungen beim idealen Datensatz von ASSIGN verglichen mit TWOSTEP. ASSIGN kommt ohne partielle Startzuordnung aus.

Der Vergleich mit der früheren Version TWOSTEP von Ganslmeier et al. [14] zeigt sehr deutlich, dass die Neuerungen in ASSIGN sehr effektiv sind. TWOSTEP fand bei einer partiellen Startzuordnung von 0 % gerade einmal 10 % der Zuordnungen richtig, wohingegen ASSIGN bereits hier seine maximale Zuordnung (99,34 %) findet. Weiterhin war bei TWOSTEP in der Simulation ein maximale Distanz von lediglich 0,2 nm eingestellt, was zu einer weitaus geringeren Anzahl der Peaks und damit zu viel geringeren Überlappungen der Peaks führte.

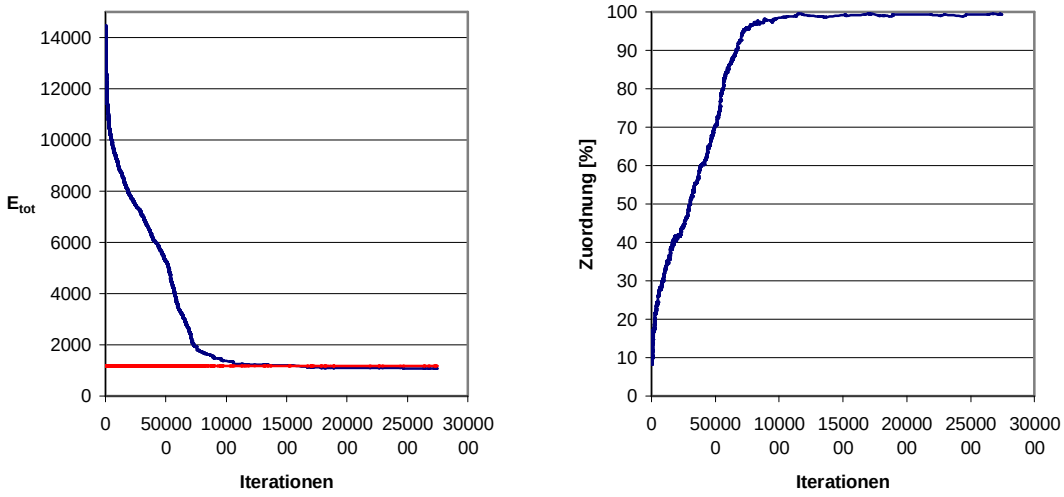

Abbildung 40: Links: Der Verlauf der totalen Pseudoenergie E_{tot} nach Formel (3.32) in Abhängigkeit von den Iterationen. Rot eingezeichnet ist die Referenz der totalen Pseudoenergiefunktion. Rechts: Die gefundene Zuordnung in Abhängigkeit von den Iterationen. Die partielle Startzuordnung war immer 0 %.

In Abbildung 40 ist der Verlauf der totalen Pseudoenergiefunktion E_{tot} und die gefundene Zuordnung in Abhängigkeit von der Anzahl der Iterationsschritte gezeigt. Es zeigt sich, dass das System sich sehr schnell dem globalen Minimum nähert. Die totale Pseudoenergie der Referenz lag bei 1183,8. Der Optimierungslauf lieferte eine Energie von 1080,5. Das bedeutet, dass die Referenzenergie bei der vollständigen und richtigen Zuordnung nicht ganz der Minimalenergie des Systems entspricht (siehe Diskussion in Kapitel 5.3.1). Die gefundenen Zuordnungen wachsen analog zur Abnahme der totalen Pseudoenergie. Dies zeigt klar, dass das System durch die gewählten Pseudoenergieterme gut beschrieben wird.

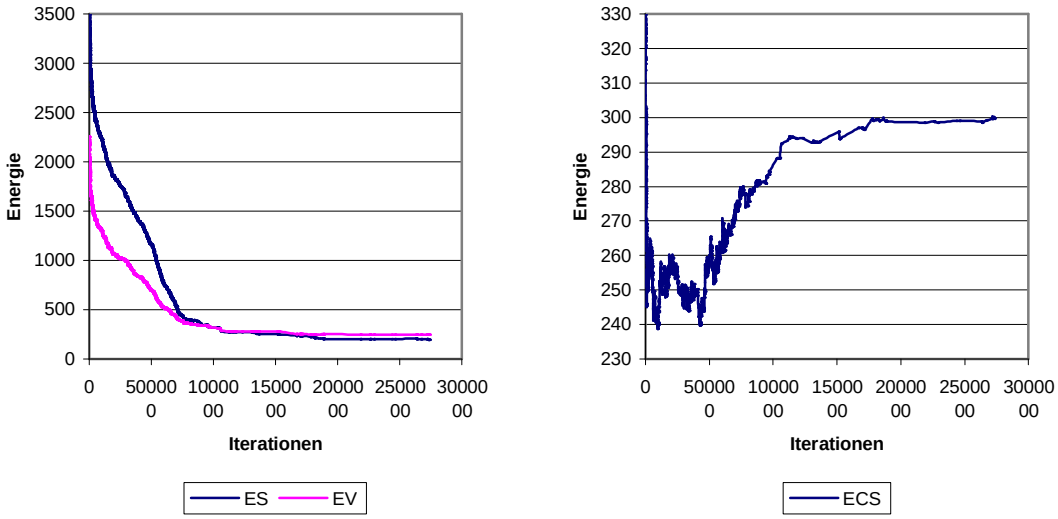

Abbildung 41: Der Verlauf der Einzelterme der Pseudoenergien in Abhängigkeit von den Iterationen.

Abbildung 41 zeigt den Verlauf der einzelnen Pseudoenergie-Terme, die zur totalen Pseudoenergie beitragen. *ES* (*Energy Shape*) und *EV* (*Energy Volume*) zeigen ein ähnliches Verhalten wie die totale Pseudoenergie E_{tot}. *ECS* (*Energy Chemical Shift*) dagegen schwankt in einem sehr kleinen Bereich zwischen 240 und 300 und pendelt sich bei ca. 300 ein, was auch dem Wert aus der Referenzzuordnung entspricht. Dieser Term strebt also nicht einem Minimum, sondern, wie in Kapitel 3.3.5 beschrieben, einem bestimmten Wert entgegen. Dieser Wert ist für jedes Protein verschieden und ist abhängig von der Zuordnung und der statistischen Vorhersage der chemischen Verschiebungen. Der Grund dafür, dass *ECS* zu Beginn des Optimierungslaufs kleiner ist als am Ende liegt darin, dass die *Slot*-Auswahl nicht zufällig, sondern, wie in Kapitel 3.3.4 beschrieben mit Hilfe der Wahrscheinlichkeiten der chemischen Verschiebungen erfolgt. Deswegen liegen am Anfang die chemischen Verschiebungen näher an der Vorhersage. Deswegen wurden auch die Gewichtungsterme w_{ESV} und w_{ECS} aus Formel (3.32) empirisch auf 1 bzw. 0,25 gesetzt, um dem Term der Verschiebungswahrscheinlichkeiten keinen zu großen Einfluss auf die totale Pseudoenergiefunktion zu geben. Diese *Slot*-Auswahl wurde, wenn die Energiedifferenz der totalen Pseudoenergie E_{tot} vor und nach erfolgreichen *Move* oder *Swap* der chemischen Verschiebungen längere Zeit klein war, automatisch auf Zufall umgestellt. Das hat den Vorteil, dass wenn schon fast alle Verschiebungen richtig zugeordnet sind, alle möglichen *Slots* für ein Atom gleichwahrscheinlich freigegeben werden und somit auch Ausreißer, die weit weg von statistischen Vorhersagen liegen, schneller gefunden werden können.

4.3.2 Idealer Datensatz mit Rauschen

In der Realität wird man nie ideale Datensätze vorfinden. Die Datensätze sind aufgrund von z.B. thermischem Rauschen in der Empfängerspule verrauscht. Im Folgenden wird das Verhalten des Algorithmus darauf näher untersucht. Zu dem oben verwendeten idealen Datensatz wird weißes Gaußsches Rauschen hinzugefügt. Dazu wird zuerst die mittlere Intensität I_{mean} über alle Intensitätswerte des Spektrums, die ungleich 0 sind, ermittelt

$$I_{mean} = \frac{1}{N_{I\neq 0}} \sum_{i=1}^{N} I_i \,, \tag{4.2}$$

wobei $N_{I\neq 0}$ die Anzahl der Datenpunkte im Spektrum mit Intensität ungleich 0, $N = N_1 \times N_2$ und I_i die Intensität des Datenpunktes i ist. N_1 und N_2 entsprechen der Anzahl der digitalen Datenpunkte pro Frequenzachse. Danach werden alle Datenpunkte wie folgt verändert

$$I_i = I_i + G(s) \,. \tag{4.3}$$

$G(s)$ entspricht einer Funktion, die gaußverteilte Zufallszahlen im Bereich von $\pm 10 \cdot s$ liefert, wobei $s = n \cdot I_{mean}$ als Standardabweichung zu Grunde liegt.

I_{mean} war im verwendeten Datensatz 75807. Bei $n = 10\%$ wird also als Standardabweichung zur Erzeugung der Gaußverteilten Zufallszahlen 7580,7 verwendet. Die Intensität des Signals von H^{α} 47/H^{N} 50 (~ 0,3 nm) ist 46988. Wird nun z. B. $n = 10\%$ gesetzt, bedeutet das, dass gaußverteilte Zufallszahlen zwischen ± 75807 addiert werden, was ca. 161 % des Signals H^{α} 47/H^{N} 50 ausmacht. Im folgendem wird nun das Rauschniveau n durch den maximalen Rauschwert *mr* ersetzt, der sich wie oben dargestellt berechnet.

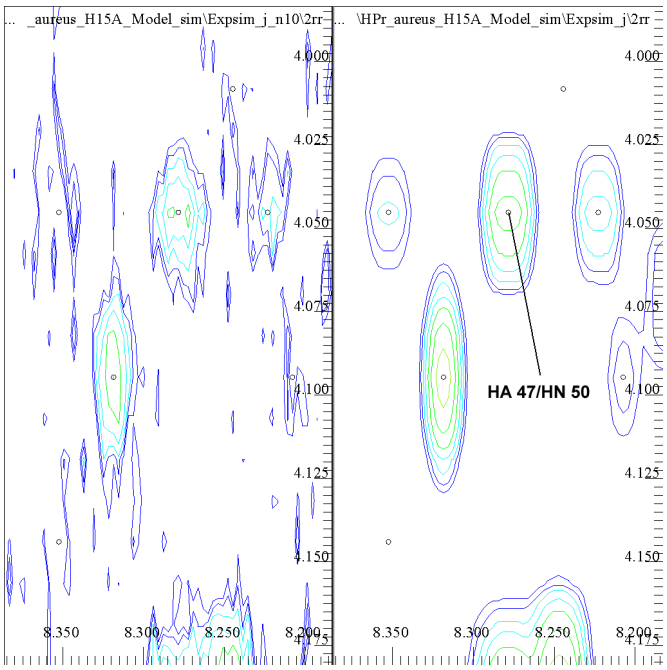

Abbildung 42: Links das mit n=10% verrauschte Spektrum. Die Normierung an den das Signal H^{α} 47/H^{N} 50 liefert hierfür einen Maximal-Rauschwert *mr* von 161 %. Rechts das Spektrum ohne Rauschen

Im Folgenden wurde nun unterschiedlich starkes Rauschen zum experimentellen Spektrum hinzugefügt. Die Häufigkeitsverteilungen der Pseudoenergie der Linienform *ES* ist in Abbildung 43 für drei max. Rauschwerte *mr* (0 %, 81 % und 161 %) dargestellt.

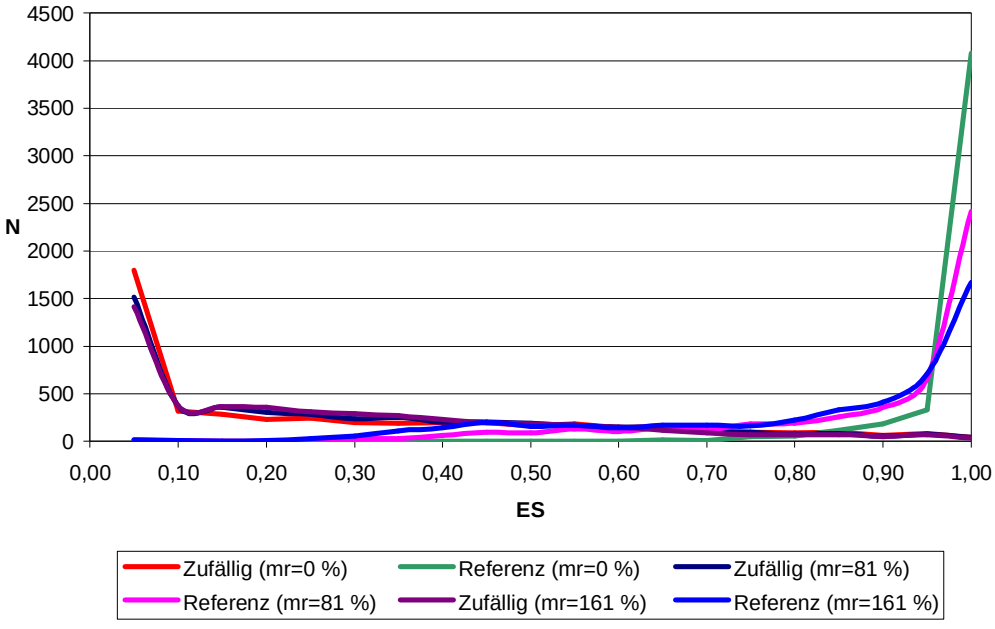

Abbildung 43: Die Häufigkeitsverteilungen der Pseudoenergie der Linienform für die Rauschniveaus *mr* =0 %, *mr* =81 % und *mr* =161 %.

Mit zunehmendem Rauschniveau werden die Verteilungen der Referenz breiter. Bei der zufälligen Zuordnung ändert sich kaum etwas. Die Auswirkungen auf die Wahrscheinlichkeitsfunktion sind in Abbildung 44 zu sehen. Je höher das Rauschniveau, desto unspezifischer wird diese Funktion. Stimmen die Linienformen nur mittelmäßig überein, so ergibt sich schon eine Wahrscheinlichkeit von 0,5 im Falle von mr =161 %. Das gleiche Verhalten ist in Abbildung 45 und Abbildung 46 für die Volumina zu sehen.

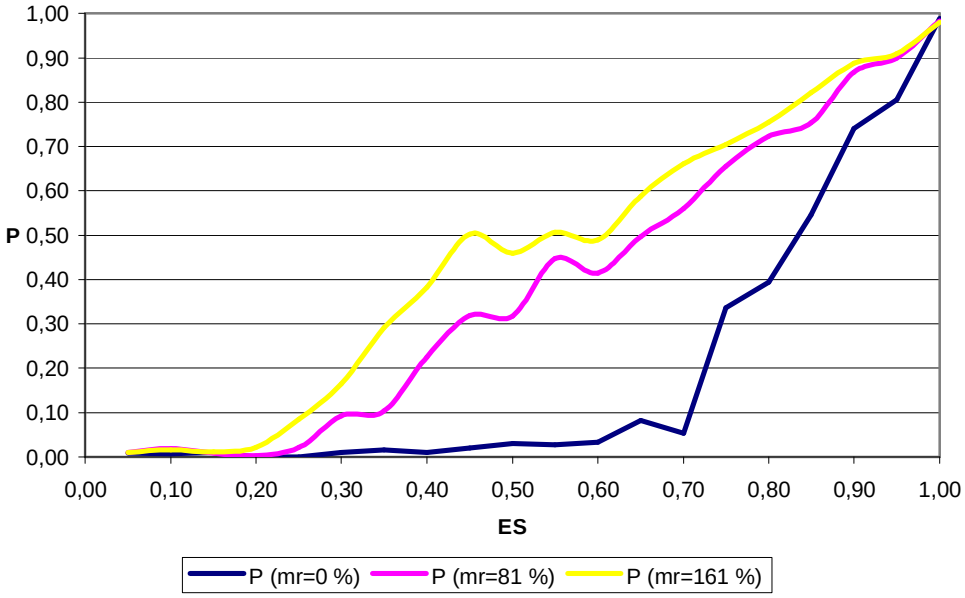

Abbildung 44: Die Wahrscheinlichkeitsfunktion der Linienform für die max. Rauschwerte mr =0 %, mr =81 % und mr =161 %.

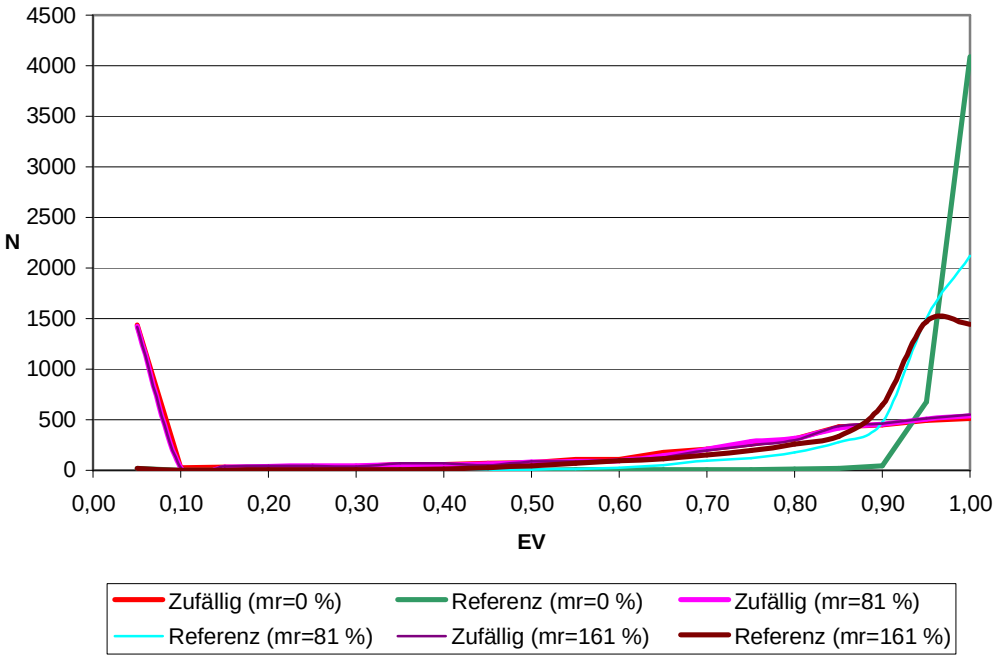

Abbildung 45: Die Häufigkeitsverteilungen der Volumen für die max. Rauschwerte mr =0 %, mr =81 % und mr =161%.

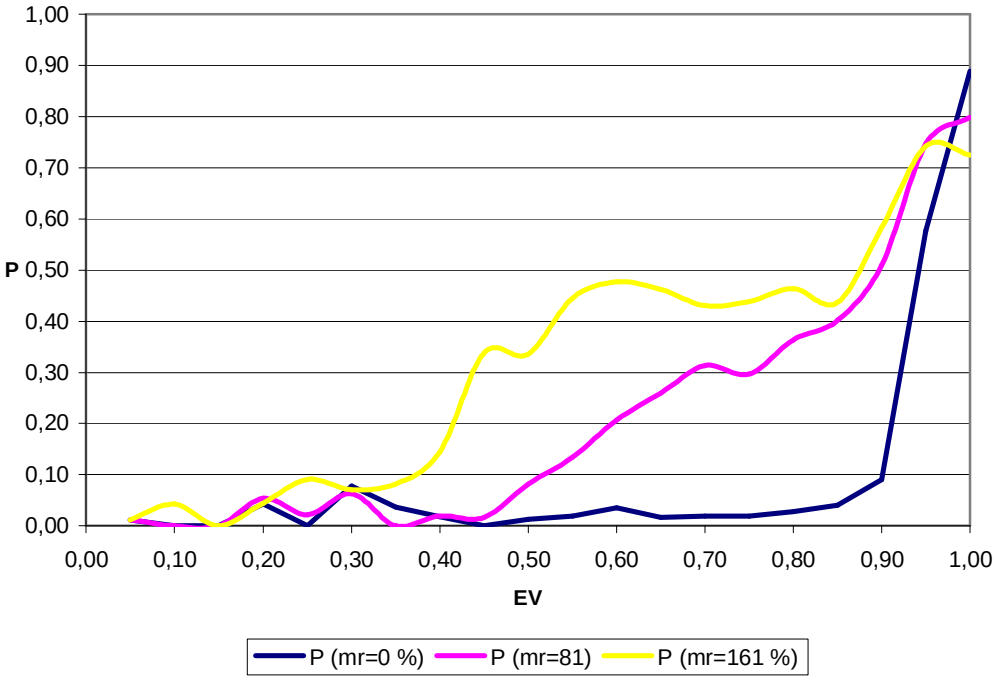

Abbildung 46: Die Wahrscheinlichkeitsfunktion des Volumens für die Rauschniveaus mr =0, mr =81 und mr =161.

Bei einer Rauschrate von 0 % wurden mit unbekannter Startzuordnung 99,34 % korrekte Zuordnungen gefunden. Es wurde nun eine Reihe von Optimierungsläufen gestartet, die alle

mit unbekannter Zuordnung begannen und bei denen das Rauschniveau erhöht wurde. Abbildung 47 zeigt die gefundenen Zuordnungen in Abhängigkeit vom max. Rauschwert *mr* .

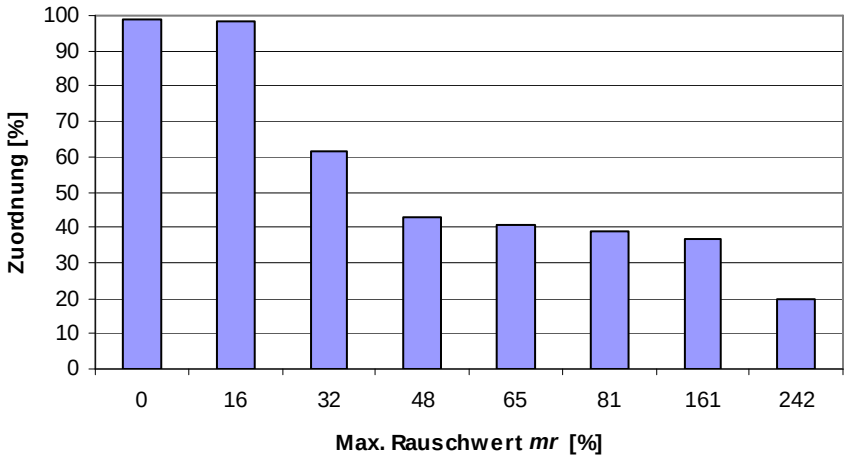

Abbildung 47: Die gefundenen Zuordnungen in Abhängigkeit vom max. Rauschwert mr .
Alle Läufe wurden mit unbekannter Zuordnung gestartet.

Kleine Rauschraten von $mr \leq 16\,\%$ bereiten noch keine Probleme, es wird trotzdem die bestmögliche Zuordnung gefunden. Mit zunehmenden Rauschraten nimmt die Zahl der gefundenen Zuordnungen ab.

Idealer Datensatz mit Rauschen	
Exp. Spektrum	Künstlicher Datensatz
3D-Struktur für Simulation	HPr *S. aureus* (H15A)
Zuordnung für Simulation	HPr *S. aureus* (H15A)
Max. Rauschwert *mr*	0 % bis 242 %
Anzahl exp. *Peaks*	4881
Anzahl sim. *Peaks*	8473
Anzahl Verschiebungen	455
Partielle Startzuordnung	0 %
Gefundene Zuordnung	99,34 % – 20,00 %

Tabelle 17: Ergebnisse für den idealen Datensatz mit verschiedenen max. Rauschwerten *mr*. Es wurden 99,34 % (*mr*=0%) bis 20 % (*mr*=242%) der sequentiellen Zuordnung richtig gefunden, die partielle Startzuordnung dabei war immer 0 %.

4.3.3 Experimenteller Datensatz

Es wurde der gleiche Test wie beim idealen Datensatz durchgeführt. Als experimenteller Datensatz diente nun ein reales experimentelles Spektrum von HPr *S. aureus* (H15A), das an einem 600 MHz Bruker Spektrometer aufgenommen wurde. In der indirekten Dimension w_1 wurden 1024 Datenpunkte aufgenommen, in der direkten Dimension w_2 4096. Um die Multiplettstruktur aufzulösen, wurde bei der Prozessierung 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. Prozessiert wurde mit einer digitalen Auflösung von 1024×4096 Punkten.
Das anschließende *Peakpicking* wurde mit den AUREMOL-Modulen *Auto Peak Pick (adaptive)* und *Peak Probability 2D/3D* durchgeführt. Nach einer kurzen manuellen Nachbearbeitung waren 2777 Peaks vorhanden. Das Spektrum wurde dann mit einer Segmentierungstiefe von 90 % und einer maximalen Segmentierungsbreite von 30 Hz in beiden Dimensionen w_1 und w_2 segmentiert. Nachdem das segmentierte Spektrum eingelesen wurde, ergaben sich 1401 *Slots*.

Bei den Häufigkeitsverteilungen ergibt sich im Prinzip ein zum idealen Datensatz analoges Verhalten. Perfekte Übereinstimmungen in der Linienform bei der Referenz sind jedoch nicht so häufig zu sehen. Die meisten Peaks haben eine Pseudoenergie im Bereich von 0,95. Analog sind die Volumenübereinstimmungen breiter verteilt als beim idealen Datensatz.
Die resultierenden Wahrscheinlichkeitsfunktionen zeigen ein Verhalten vergleichbar zum idealen Datensatz mit Rauschen.

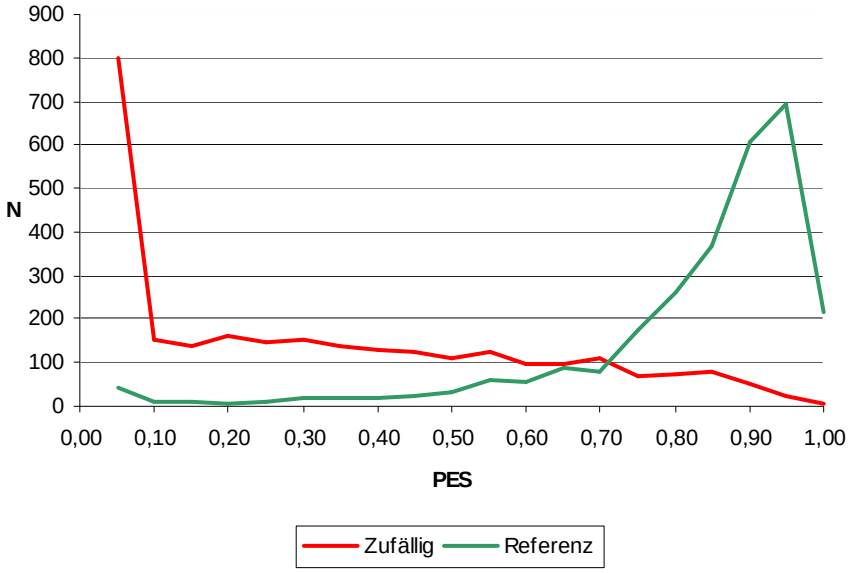

Abbildung 48: Die Häufigkeitsverteilungen der Linienform bei experimentellen Daten. Bei der Referenz ist eine Verschiebung hin zu 0,95 zu beobachten.

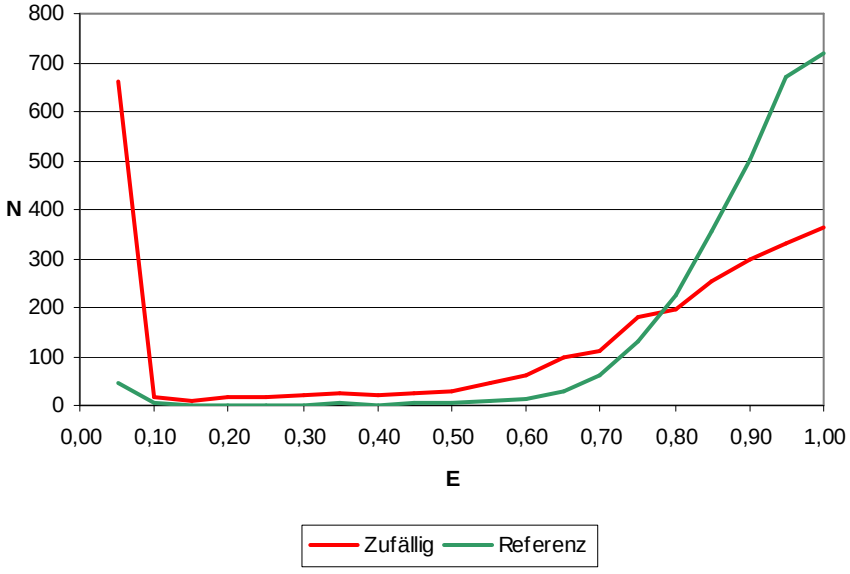

Abbildung 49: Die Häufigkeitsverteilungen der Volumen bei experimentellen Daten.

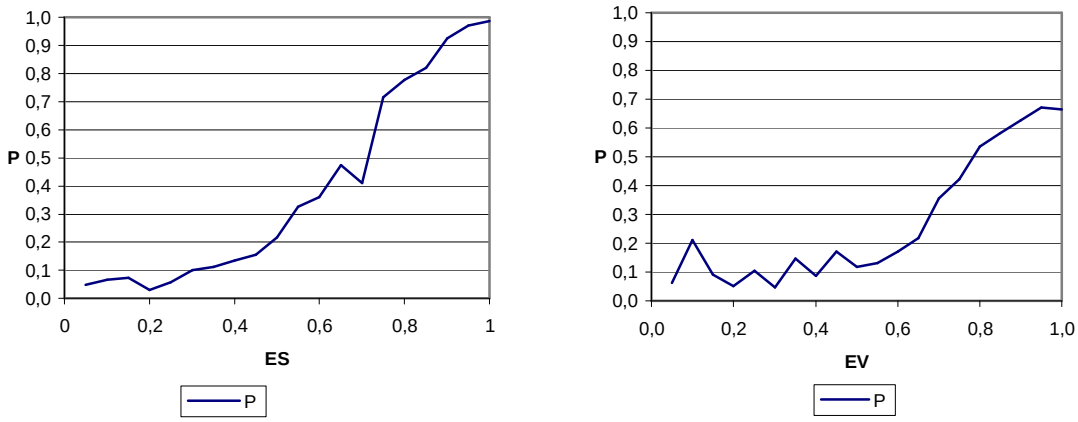

Abbildung 50: Die Wahrscheinlichkeitsfunktionen der Linienform (ES) und des Volumens (EV).

Startzuordnung aus zufälligen Zuordnungen

Es wurde die Abhängigkeit der gefundenen Zuordnung von der Startzuordnung untersucht, d. h. ein bestimmter Anteil der Zuordnung wird fest vorgegeben. Welcher Anteil dies ist, wird mit einem Zufallsgenerator ausgewählt.

Aus Abbildung 51 geht hervor, dass bei einer partiellen Startzuordnung von 0 % etwa 31 % der Zuordnung richtig gefunden wird. Bereits bei einer partiellen Startzuordnung von 20 % werden über 90 % der Zuordnungen richtig gefunden. Bei 90 % Startzuordnung werden mit 98,66 % (441 von 447) fast alle Zuordnungen richtig gefunden.

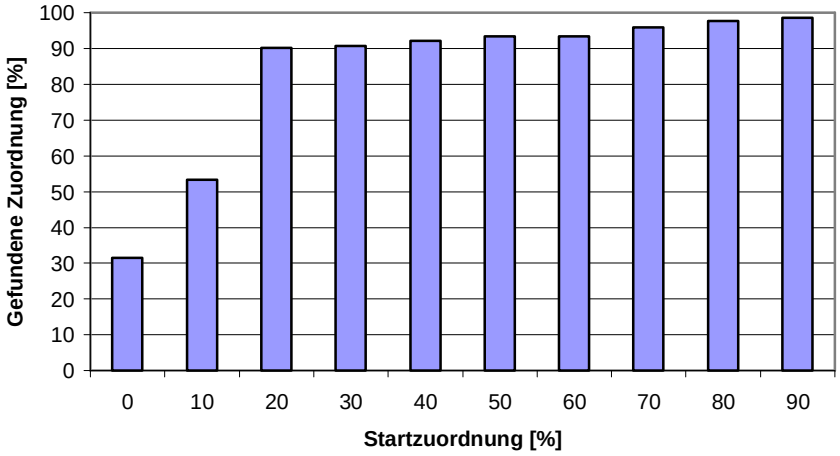

Abbildung 51: Die gefundene Zuordnung in Abhängigkeit von der Startzuordnung beim experimentellen Datensatz.

Experimenteller Datensatz, Startzuordnung aus zufälligen Zuordnungen	
Exp. Spektrum	HPr *S. aureus* (H15A), 600,13 MHz
Datenpunkte w1, w2	1024 x 4096
3D-Struktur für Simulation	HPr *S. aureus* (H15A)
Zuordnung für Simulation	HPr *S. aureus* (H15A)
Anzahl exp. *Peaks*	2777
Anzahl sim. *Peaks*	8271
Anzahl chem. Verschiebungen	447
Partielle Startzuordnung	0 % - 90 %
Gefundene Zuordnung	31,54 % – 98,66 %

Tabelle 18: Ergebnisse für den experimentellen Datensatz mit verschiedenen partiellen Startzuordnungen. Es wurden 31,54 % (0 % part. Startzuordnung) bis 98,66 % (90 % part. Startzuordnung) der sequentiellen Zuordnung richtig gefunden. Bereits ab 20 % partieller Startzuordnung wurden über 90 % der Zuordnungen richtig gefunden.

In Abbildung 52 sind die erreichten Energien im Vergleich zur Referenzenergie in Abhängigkeit der Startzuordnungen aufgetragen. Es ist zu sehen, dass mit 0 % Startzuordnung die Optimierung in einem lokalen Minimum stecken bleibt. Ab 20 % Startzuordnung ist das Energieminimum unterhalb der Referenzenergie. Mit zunehmender Startzuordnung nähert sich die Energie dem Referenzwert an. Es zeigt sich, dass die totale Pseudoenergiefunktion das System nicht perfekt beschreibt. D. h., dass das Minimum der Energie nicht dem Maximum der Zuordnung entspricht. Es gibt also für das System Lösungen, die energieärmer sind als die richtige Lösung. Dieser Effekt ist aber sehr klein und stellt die Richtigkeit des Ansatzes nicht in Frage. Vielmehr zeigt er, dass der Ansatz noch nicht ganz vollständig ist.

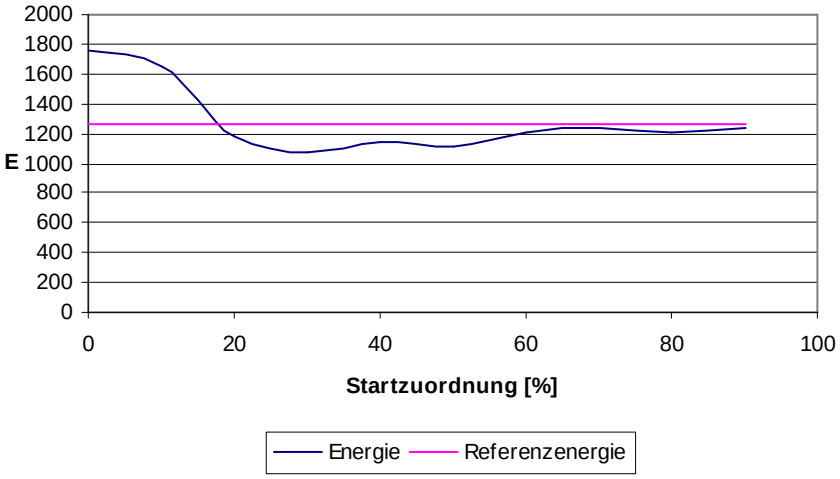

Abbildung 52: Die erreichten Energien im vergleich zur Referenzenergie.

Startzuordnung aus H^N- und H^α-Atomen

Ein weiterer Test bestand darin, nur leicht zu gewinnende NMR-Daten als Startzuordnung zu verwenden. Es wurden nur die 163 (36,5 %) chemischen Verschiebungen der H^N- und H^α-Atome verwendet.

Experimenteller Datensatz, Startzuordnung aus H^N- und H^α-Atomen	
Exp. Spektrum	HPr *S. aureus* (H15A), 600,13 MHz
Datenpunkte w1, w2	1024 x 4096
3D-Struktur für Simulation	HPr *S. aureus* (H15A)
Chem. Versch. für Simulation	HPr *S. aureus* (H15A)
Anzahl exp. *Peaks*	2777
Anzahl sim. *Peaks*	8271
Anzahl chem. Verschiebungen	447
Partielle Startzuordnung	36,5 % (H^N H^α-Atome)
Gefundene Zuordnung	85,2 %

Tabelle 19: Ergebnisse für den experimentellen Datensatz mit der partiellen Startzuordnung aus den H^N- und H^α-Atomen. Es wurden 85,2 % der sequentiellen Zuordnung richtig gefunden.

Mit dieser Zuordnung wurde ein Strukturbündel berechnet. Dazu wurden mit PEAKASSIGN [20] die NOE-Signale zugeordnet und mit REFINE Abstand-*Restraints* erzeugt. 50 Diederwinkel-*Restraints* wurden aus dieser partiellen sequentiellen Zuordnung mit TALOS [80] gewonnen. Dabei wurden Winkel als *good* markiert, wenn alle Vorschläge in derselben Region des *Ramachandran-Plots* zu finden waren. Weiterhin wurden 26 Wasserstoffbrücken-*Restraints* über ein H(N)CO-Spektrum identifiziert. Mit den so gewonnenen Restraints wurden dann mit DYANA 500 Strukturen berechnet und die zehn Besten hinsichtlich der DYANA-Energiefunktion zur weiteren Analyse verwendet. Das so erhaltene Bündel ist in Abbildung 53 (B) gezeigt.

Mit ASSIGN wurden nun die fehlenden chemischen Verschiebungen der Seitenkettenatome zugeordnet. Der Optimierungslauf wurde mit denselben Parametern wie oben gestartet. ASSIGN fand 381 (85,2 %) korrekte chemische Verschiebungen. Mit dieser von ASSIGN gefundenen sequentiellen Zuordnung, die auch die 14,8 % fehlerhaften Zuordnungen enthielt, wurden nun mit Hilfe von PEAKASSIGN 2266 Peaks zugeordnet. Mit REFINE wurden im Anschluss 1606 Abstand-*Restraints* generiert. Die Anzahl der Abstände ist geringer als die Anzahl der Peaks, weil für symmetrische Signale nur ein Abstand berechnet wird.

Sequentielle Zuordnung	H15A	$H^N H^\alpha$	$H^N H^\alpha$ + ASSIGN
Richtige seq. Zuordnung [%]	100	36,5	85,2
Zugeordnete Peaks	2289	2035	2266
Abstand-*Restraints*	1606	276	1606
Diederwinkel-*Restraints*	50	50	50
W'brücken-*Restraints*	26	26	26

Tabelle 20: Die erhaltenen Restraints, die bei der Strukturrechnung Verwendung fanden.

Mit den so gewonnenen Restraints wurden dann mit DYANA 500 Strukturen berechnet und die zehn Besten hinsichtlich der DYANA-Energiefunktion zur weiteren Analyse verwendet. Das erhaltene Bündel ist in Abbildung 53 (C) gezeigt. Man erkennt eine deutliche Strukturverbesserung, die mit Hilfe der von ASSIGN gefundenen chemischen Verschiebungen der Seitenkettenatome erreicht wurde. Sie sieht der Originalstruktur (Abbildung 53 (A)) sehr ähnlich.

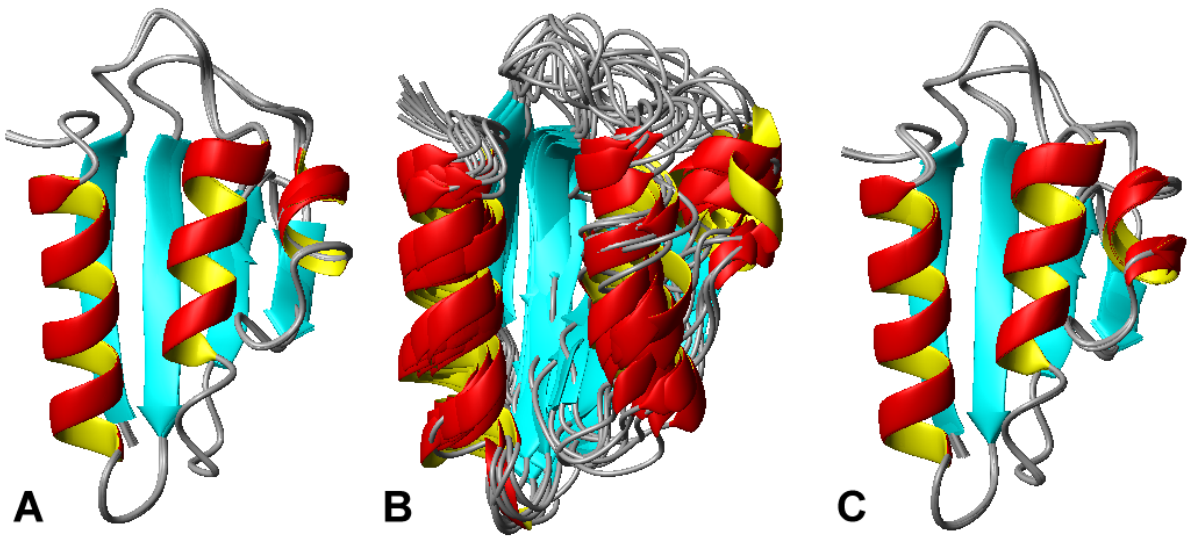

Abbildung 53: (A) Originalstruktur, (B) Struktur nur mit den chemischen Verschiebungen der H^N H^α-Atome, (C) Struktur mit den chemischen Verschiebungen der H^N H^α-Atome und den von ASSIGN gefundenen Zuordnungen der Seitenketten. Die Sekundärstrukturelemente sind in den Strukturen von (B) schlecht bestimmt. Nimmt man die Seitenketten-Zuordnungen mit hinzu, so kommt das Bündel aus (C) dem Originalbündel (A) recht nahe.

Alle Sekundärstrukturelemente sind wohl definiert. Der RMSD-Wert mit 0,015 nm und der AUREMOL R-Wert (R_5) von 0,328 sind sehr nah am Original (0,017 nm bzw. 0,320). In Tabelle 21 sind die Qualitätswerte zusammengefasst.

Sequentielle Zuordnung	**H15A**	**$H^N H^\alpha$**	**$H^N H^\alpha$ + ASSIGN**
AUREMOL R-Wert (R_5)	0,320	0,589	0,328
RMSD MOLMOL N [nm]	0,017	0,151	0,015
Ramachandran m.f. + a. [%]	94,9	93,6	94,8
Most favored [%]	73,1	60,3	76,9
Additional allowed [%]	21,8	43,0	17,9
Generously allowed [%]	3,8	6,4	1,3
Disallowed [%]	1,3	0,0	3,8

Tabelle 21: Qualitätswerte aus AUREMOL, MOLMOL und PROCHECK für die drei Strukturbündel, gerechnet mit den entsprechenden Zuordnungen.

4.3.4 Punktmutante mit experimentellen Daten

Nachdem gezeigt wurde, dass der ASSIGN Algorithmus mit simulierten Daten sehr gut funktioniert und auch bei experimentellen Spektren gute Ergebnisse liefert, soll nun gezeigt werden, dass ASSIGN in einem realistischen Beispiel benutzt werden kann. In diesem Testfall soll die sequentielle Zuordnung der Punktmutante HPr *S. aureus* (H15A) ausgehend von der gelösten Struktur von HPr *S. aureus* (wt) gefunden werden. Es liegt die sequentielle Zuordnung und die 3D-Struktur von HPr *S. aureus* (wt) vor. Aus der 3D-Struktur von HPr *S. aureus* (wt) wurde mit dem AUREMOL-Modul PERMOL eine Modellstruktur für HPr *S. aureus* (H15A) erstellt (siehe Kapitel 4.1.2). Diese 3D-Struktur und die sequentielle Zuordnung von HPr *S. aureus* (wt) dienen als Eingabe für das simulierte Spektrum.

Das experimentelle Spektrum war dasselbe wie in Kapitel 4.3.3. Für das simulierte Spektrum wurden das mit PERMOL erzeugte Modell der Mutante und die sequentielle Zuordnung des Wildtyps verwendet. Aus den vorhandenen 425 Atomen wurden so die Peaks mit den zugehörigen Linienformen berechnet. Bei der Simulation des Spektrums wurde eine max. Distanz von 0,5 nm, eine Mischzeit von 0,08 s, eine Relaxationszeit von 1,54 s und eine Resonanzfrequenz von 600,13 MHz gewählt. Diese Parameter gingen aus dem Experiment hervor. Ebenso wurde die J-Kopplung und damit die Multiplettstruktur berücksichtigt. Auf diese Weise wurden 8435 *Peaks* erzeugt.

Punktmutante mit experimentellen Daten	
Exp. Spektrum	HPr *S. aureus* (H15A), 600,13 MHz
Datenpunkte w1, w2	1024 x 4096
3D-Struktur für Simulation	HPr *S. aureus* (H15A) Modell (PERMOL)
Chem. Versch. Für Simulation	HPr *S. aureus* (wt)
Anzahl exp. *Peaks*	2777
Anzahl sim. *Peaks*	8435
Anzahl chem. Verschiebungen (Ref.)	425
Partielle Startzuordnung	79,8 % (mittels PEAKASSIGN)
Gefundene Zuordnung	96,7 %

Tabelle 22: Ergebnisse für die Punktmutante mit dem experimentellen Datensatz. Ausgehend von einer partiellen Startzuordnung von 79,8 %, die von PEAKASSIGN ermittelt wurde, wurden 96,7 % korrekte sequentielle Zuordnungen gefunden.

Mit PEAKASSIGN wurde in einem weiteren Schritt bestimmt, wie viele chemische Verschiebungen als bereits bekannt vorausgesetzt werden können. Dazu wurden die sequentielle Zuordnung des Wildtyps und die Modellstruktur der Mutante verwendet. PEAKASSIGN lieferte 339 von 425 chemische Verschiebungen (79,8 %), die in einem Toleranzbereich von $\pm 0,03$ ppm richtig waren. Diese 339 chemischen Verschiebungen wurden sodann als feste Zuordnung in ASSIGN verwendet.

Die Aufgabe von ASSIGN war nun, die fehlenden 86 chemischen Verschiebungen zu finden. Nach dem Optimierungslauf wurden 72 richtige chemische Verschiebungen gefunden. 14 chemische Verschiebungen lagen auf falschen *Slots*. Insgesamt wurden also 96,7 % der richtigen Zuordnung erreicht.

Anschließend wurde dann das experimentelle NOESY-Spektrum mit PEAKASSIGN zugeordnet. Als Eingabe wurden die aus ASSIGN stammende sequentielle Zuordnung und die Modellstruktur der Mutante verwendet. Es konnten 2266 Peaks zugeordnet werden. Mit dem Modul REFINE wurden dann 1559 Abstand-Restrains berechnet. Mit TALOS [80] wurden 50 Diederwinkel-*Restraints* gewonnen. Dabei wurden Winkel als *good* markiert, wenn alle Vorschläge in derselben Region des *Ramachandran-Plots* zu finden waren. 26 Wasserstoffbrücken-*Restraints* wurden über ein H(N)CO-Spektrum identifiziert.

Mit den so gewonnenen Restraints wurden dann mit DYANA 500 Strukturen berechnet und die zehn Besten hinsichtlich der Energiefunktion zur weiteren Analyse verwendet.

Das gleiche Verfahren wurde mit der sequentiellen Zuordnung der Mutante und mit der sequentiellen Zuordnung, die aus dem Wildtyp hervorgegangen ist (d. h. es werden nur die

chemischen Verschiebungen verwendet, die sich gegenüber dem Wildtyp kaum verändert haben), praktiziert (Tabelle 23).

Sequentielle Zuordnung	**H15A**	**wt**	**wt + ASSIGN**
Richtige seq. Zuordnung [%]	100	79,8	96,7
Zugeordnete Peaks	2289	2207	2266
Abstand-*Restraints*	1606	981	1559
Diederwinkel-*Restraints*	50	28	50
W'brücken-*Restraints*	26	26	26

Tabelle 23: Die erhaltenen Restraints, die bei den Strukturrechnungen Verwendung fanden.

Auch hier wurden jeweils 500 Strukturen berechnet. Anschließend wurde die Qualität der drei Strukturbündel untersucht. Es wurde der AUREMOL R-Wert (R_5), der RMSD und der *Ramachandran-Plot* herangezogen. Tabelle 24 zeigt eine Übersicht der Ergebnisse.

Sequentielle Zuordnung	**H15A**	**Wt**	**wt + ASSIGN**
AUREMOL R-Wert (R_5)	0,320	0,384	0,356
RMSD MOLMOL N [nm]	0,017	0,046	0,027
Ramachandran m.f. + a. [%]	94,9	92,3	94,8
Most favored [%]	73,1	56,4	76,9
Additional allowed [%]	21,8	35,9	17,9
Generously allowed [%]	3,8	6,4	3,8
Disallowed [%]	1,3	1,3	1,3

Tabelle 24: Qualitätswerte aus AUREMOL, MOLMOL und PROCHECK
für die drei Strukturbündel gerechnet mit den entsprechenden Zuordnungen.

Die Qualitätswerte zeigen eine deutliche Verbesserung mit der mit ASSIGN gefundenen sequentiellen Zuordnung gegenüber der festen Teilzuordnung aus dem Wildtyp. Der AUREMOL R-Wert (R_5) konnte von 0,384 auf 0,356 verbessert werden. Ebenso der RMSD-Wert, der von 0,046 nm auf 0,027 nm verbessert wurde. Die *Ramachandran*-Qualität liefert auch wesentlich bessere Werte. Der prozentuale Anteil in den *most favored* Regionen konnte

von 56,4 % auf 76,9 % gesteigert werden und ist mit dem Strukturbündel, das aus der richtigen Zuordnung der Mutante entstanden ist vergleichbar.

In Abbildung 54 sind die drei Strukturbündel zu sehen. Abbildung 54 (A) zeigt das Bündel mit der richtigen Zuordnung der Mutante. Abbildung 54 (B) zeigt das Bündel, das mit der Zuordnung aus dem Wildtyp entstanden ist. In Abbildung 54 (C) ist das Bündel, das mit der von ASSIGN gefundenen Zuordnung erstellt wurde, abgebildet. Deutlich ist zu sehen, dass mit der Zuordnung aus dem Wildtyp der Bereich der Mutation unterbestimmt ist. Die α-Helix und das dahinter liegende β-Faltblatt sind nicht bestimmt. In der ASSIGN-Struktur sind diese Bereiche wieder vorhanden und gut definiert.

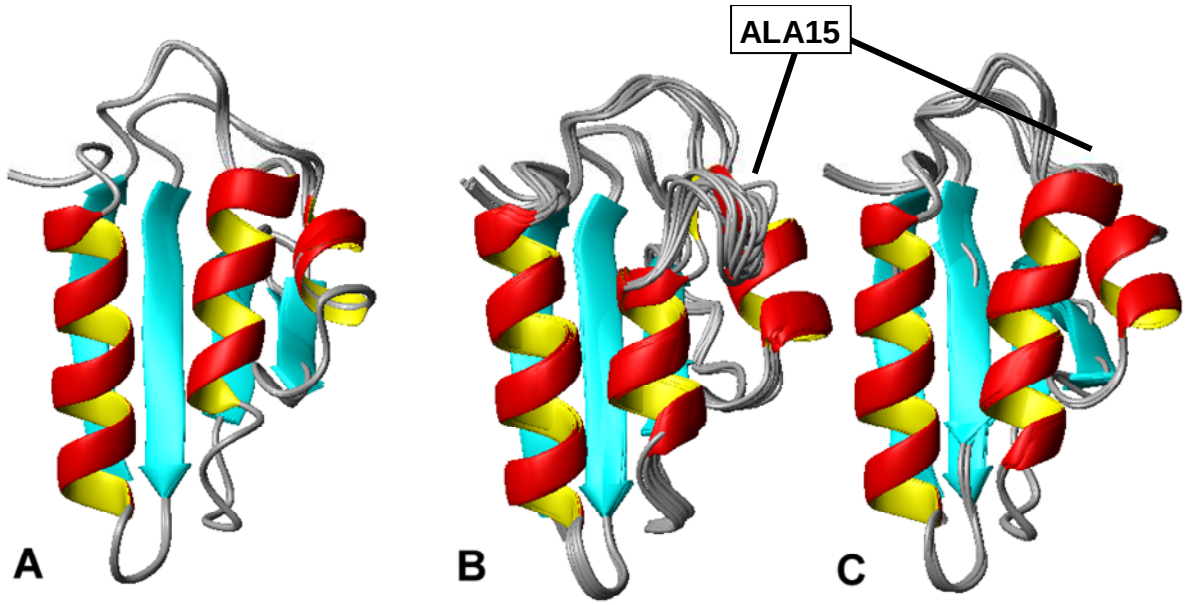

Abbildung 54: (A) Die Struktur von HPr *S. aureus* (H15A) mit sequentieller Zuordnung aus der Mutante. (B) Die Struktur von HPr *S. aureus* (H15A) mit sequentieller Zurodung aus dem Wildtyp. (C) Die Struktur von HPr *S. aureus* (H15A) mit sequentieller Zurodung aus ASSIGN. Deutlich ist zu sehen, dass der Bereich der Mutation im mit ASSIGN gewonnenen Strukturbündel gut definiert ist.

5 Diskussion

In der vorliegenden Arbeit wurden verschiedene Module von AUREMOL weiter- oder neu entwickelt, um die automatische Strukturbestimmung von Proteinen in Lösung voranzutreiben.

Zuerst wurde eine einheitliche IUPAC-Nomenklatur für die Atomnamen innerhalb aller AUREMOL-Module eingeführt und Routinen entwickelt, die eine Konvertierung zwischen den verschiedenen Datenformaten erlaubt. So können PDB-Dateien, Signallisten und Listen chemischer Verschiebungen leicht in für AUREMOL brauchbare Formate umgewandelt und *Restraint*-Dateien, die eine Schnittstelle zu den Moleküldynamik-Programmen DYANA [26], CNS [27] und XPLOR-NIH [37] darstellen, einfach erzeugt werden.

Mit dem in dieser Arbeit weiterentwickeltem und in AUREMOL implementierten Modul PERMOL ist eine neue Art der Homologie-Modellierung möglich, die sehr genau und auch bei geringer Sequenzidentität äußerst brauchbare 3D-Modelle von Proteinen erzeugt, die dann als Startstruktur für den *Top-Down*-Ansatz in AUREMOL dienen. PERMOL dient auch als wichtige Unterroutine für das in dieser Arbeit völlig neu entwickelte Modul ISIC [71]. Mit ISIC können sehr effizient bestehende Proteinstrukturen verbessert werden. Dabei werden die Informationen verschiedener Quellen über ein Protein auf nicht triviale Weise so kombiniert, dass nur relevante Informationen zu Verbesserung herangezogen werden. Das Problem, eine sequentielle Zuordnung automatisiert zu erhalten, wurde mit dem Modul ASSIGN angegangen. Durch Vergleich von simulierten und experimentellen Spektren unter Einbeziehung von Linienformen, Volumina der beteiligten Signale und statistischen Vorhersagen der chemischen Verschiebungen wird die vollautomatische Erlangung der sequentiellen Zuordnung möglich.

5.1 Homologie-Modellierung (PERMOL)

Die sehr gute Qualität der mit PERMOL gewonnenen Strukturen zeigt, dass Moleküldynamik-Programme wie DYANA oder CNS, die in der Regel zur experimentellen Strukturbestimmung verwendet werden, auch zur Homologie-Modellierung eingesetzt werden können. Dabei werden von PERMOL die Restraints aus bereits bekannten, homologen Strukturen abgeleitet. Aus jedem einzelnen Modell werden Atomabstände, Diederwinkel und Wasserstoffbrücken gewonnen und anschließend Mittelwerte mit entsprechenden Fehlergrenzen berechnet.

Ein Vorteil dieser Methode ist ihre Flexibilität, so dass spezifische Probleme in einer effizienten Weise angepasst und optimiert werden können. Die Qualität der so erhaltenen Strukturen ist außerordentlich zufrieden stellend. Zudem bietet der Ansatz enorme Geschwindigkeitsvorteile gegenüber experimentellen Strukturbestimmungsmethoden. Distanz- und Winkel-Restraints müssen nicht mehr zwingend durch die Zuordnung von Spektren ermittelt werden, vielmehr werden sie aus strukturhomologen Proteinen mit bereits bekannten Tertiärstrukturen extrahiert. Da bei hoher Sequenz- oder Funktionshomologie von Proteinen davon ausgegangen werden kann, dass daraus modellierte Tertiärstrukturen qualitativ an experimentell bestimmte Strukturen heranreichen, ist es nur sinnvoll, an Stelle des sehr langwierigen Vorganges der experimentellen Strukturbestimmung die schnelle Vorhersage durch PERMOL treten zu lassen. Viele Module in AUREMOL, wie z. B. KNOWNOE [21], RELAX [23] und das für diese Arbeit entwickelte ASSIGN profitieren davon, wenn die dabei verwendete 3D-Struktur bereits eine gute Qualität hat und nicht vom ausgestreckten Strang begonnen werden muss.

Ein weiterer, entscheidender Vorteil ist die Berechnung lokaler Fehlergrenzen. Dadurch ist es möglich, bestimmte Bereiche sehr genau zu modellieren, wenn diese in den Modellstrukturen bereits genau bestimmt sind. Im Vergleich zu den Ergebnissen anderer Homologie-Modellierungs-Pakete zeigt sich noch ein weiterer Vorteil: Sowohl MODELLER als auch SWISS-MODEL sind auf eine sehr hohe Sequenzhomologie (MODELLER auf 30% [122], SWISS-MODEL auf 30% [123]) angewiesen. PERMOL dagegen kann beim Nachweis der Funktions- und somit der Struktur-Homologie auch bei verschwindender Sequenzähnlichkeit brauchbare Modelle liefern. Es können hierzu entweder die Parameter des Sequenzalignments so eingestellt werden, dass die beiden Sequenzen trotz geringer Sequenzidentität ohne größere Lücken übereinander gelegt werden, oder das *Sequenzalignment* kann manuell erfolgen.

Mit dem Ansatz von PERMOL ist es weiterhin möglich, immer die am aktuellsten entwickelten Molekül-Dynamik-Programme zu verwenden. So ist erreichbar, dass das aktuellste Wissen auf dem Gebiet der Molekül-Dynamik-Simulationen auch in PERMOL genutzt wird. Die Möglichkeit, die mit PERMOL aus den Strukturen ermittelten *Restraints* mit experimentell gewonnenen Daten zu ergänzen und dann mit den neuesten Molekül-Dynamik-Programmen Strukturen zu berechnen, macht PERMOL zu einem sehr flexiblen und leistungsstarken Modellierungs-Werkzeug.

PERMOL dient dem in dieser Arbeit entwickelten Ansatz zur Verbesserung von Proteinstrukturen (ISIC) als Teilmodul zur Erzeugung der Restraints. In Kombination mit den Modulen zur Strukturaufklärung in AUREMOL stellt PERMOL ein sehr wichtiges Glied dar.

In dieser Arbeit wurde PERMOL zur Modellierung von HPr aus *Staphylococcus aureus* herangezogen. Die Ergebnisse zeigen eine sehr gute Konvergenz mit einem RMSD von 0,018 nm für die Rückgrat-Atome des erhaltenen Bündels der Modellstrukturen. Dies zeigt, dass durch die Wahl von individuellen Fehlergrenzen und mit der Berechnung des zugehörigen Konfidenzniveaus die Strukturen sehr gut bestimmt werden können. Ebenso zeigt der *Ramachandran-Plot* mit 94,8 % in den *most favored* und *additional allowed* Regionen einen guten Wert. Die modellierten 3D-Strukturen (Abbildung 24) zeigen gut definierte Sekundärstrukturelemente.

Am zweiten Beispiel der Modellierung der Punktmutante HPr (H15A) (Abbildung 25) ist zu erkennen, wie auf sehr schnellem Weg ein Modell aus dem Wildtyp erstellt werden kann. Der RMSD von 0,098 nm zur Original-Struktur und 0,006 nm innerhalb des Strukturbündels (Rückgrat-Atome N) sowie 98,7 % der Diederwinkel in den *most favored* und *additional allowed* Regionen des *Ramachandran-Plots* zeigen in diesem Fall wiederum eine sehr gute Qualität. Diese Modellierung der Punktmutation wurde dann auch im Modul ASSIGN eingesetzt.

5.2 Verbesserung von Proteinstrukturen (ISIC [71])

Die Bestimmung von Strukturen in Lösung aus experimentellen Daten besteht nicht, wie manchmal fälschlicherweise angenommen wird, in der Berechnung einer einzigen existierenden Lösung, sondern ist vielmehr die Suche nach einem Satz von Strukturen, die mit dem Experiment und dem bereits vorhandenen Vorwissen vereinbar sind. In diesem Zusammenhang sei auch auf das *Paper* von Rieping et. al. verwiesen [124]. Hier wird gezeigt, dass Proteinstrukturen, die mit herkömmlichen Methoden aus NMR-Daten berechnet werden, nicht ausschließlich durch die die experimentellen Daten bestimmt werden, sondern auch durch eine subjektive Datenverarbeitung unter Verwendung von empirischen Parametern, wie z.B. die Wichtung der beteiligten Kraftkonstanten.

Die in dieser Arbeit vorgestellte Verwendung von Ersatz-*Restraints* zusammen mit der Methode der *restrained* Moleküldynamik ist eine effiziente Methode, um stark gekoppeltes Wissen aus verschiedenen Quellen zu kombinieren. Eine angemessene Ausrichtung hin zur gewählten Zielstruktur kann mit Hilfe von Bayesschen Schlussfolgerungen erreicht werden. Auf diese Weise wird die zusätzliche Information zur Erhöhung der Wahrscheinlichkeit verwendet, die wahren Strukturen des Grundzustands, die mit bestimmten experimentellen Bedingungen einhergehen, zu finden. Nach Anwendung der vorgestellten Methode zeigen Validierungs-Tools ganz klar, dass die verbesserten Strukturen weiter an die wahren Strukturen heranreichen als die Ausgangstrukturen.

Im Falle der Byr2-RBD ist deutlich zu sehen, dass die mit ISIC verbesserten Strukturen die experimentellen Daten wesentlich besser erklären. Der AUREMOL NMR R-Wert (R_5) [39] ist von 0,534 auf 0,451 gesunken. Mit 0,033 nm ist der RMSD zum Mittelwert im Vergleich zur Ausgangsstruktur (0,144 nm) drastisch gesunken. Der *Ramachandran-Plot* zeigt, dass die Diederwinkelverteilung mit einer Steigerung um 10,4 Prozentpunkte auf 78,2 % in den *most favored* Regionen deutlich besser ist. Die Darstellung der Sekundärstruktur mit MOLMOL zeigt (Abbildung 29), dass die Sekundärstrukturelemente der ISIC-Strukturen sehr gut bestimmt sind. Dies gilt vor allem für die C-terminale α-Helix, die in der originalen NMR-Struktur sehr schlecht bestimmt ist. Der Vergleich mit dem Grundbündel der NMR für Byr2 zeigt, dass die mit ISIC erhaltenen, verbesserten Strukturen sehr gut in diesem Grundbündel liegen und somit eine gute Lösung darstellen.

Beim zweiten Testfall wurde die Struktur der Ras-Bindedomäne von RalGDS des Menschen verbessert. Dabei wurde Wert darauf gelegt, NMR-Strukturen, die mit experimentell leicht zu erhaltenden Restraints bestimmt wurden, zu verbessern. Für die Bestimmung der

Proteinstruktur von RalGDS wurden deshalb nur 232 Rückgrat-Abstände zwischen den H^N- und H^α-Atomen aus dem NOESY-Spektrum herangezogen. Zudem 25 Wasserstoffbrücken und 102 Diederwinkel aus TALOS. Das daraus resultierende Strukturbündel war weit aufgefächert (RMSD-Wert der Rückgrat-Atome N: 0,21 nm). Zur Verbesserung wurde eine Röntgenstruktur mittlerer Qualität herangezogen. Die mit ISIC erzeugten Strukturen zeigten, wie im ersten Beispiel, mit 0,353 einen gesunkenen und damit besseren AUREMOL R-Wert (R_5) als die NMR-Eingabestruktur (0,383). Die experimentellen Daten werden also auch bei diesem Beispiel nach Anwendung des ISIC-Algorithmus besser erklärt. Ebenso konnte der RMSD-Wert von 0,21 nm auf 0,06 nm reduziert werden. Die Sekundärstrukturelemente sind wesentlich besser bestimmt (Abbildung 33). Die Qualität des *Ramachandran-Plots*, die mit 72,8 % (*most favored*) und 18,5 % (*additional allowed*) ohnehin schon gut war, blieb mit 72,8 % in den *most favored* und 16,0 % in den *additional allowed* Regionen nahezu gleich. Das Beispiel zeigt, dass NMR-Strukturen, die mit leicht und schnell zu gewinnenden Restraints erzeugt wurden, mit ISIC auf eine schnelle und effiziente Weise verfeinert werden können.

Selbst wenn man völlig ungeeignete Daten zur Verbesserung heranzieht, wie das im dritten Beispiel im Fall der Immunoglubulin-Binde-Domäne von Protein G gezeigt ist, führt das nicht zur Zerstörung oder Verzerrungen der Originalstruktur (Abbildung 34). Hier wurden zwei Strukturbündel des gleichen Proteins verwendet, die sich aber auffällig unterscheiden. Die NMR-Struktur zeigte im Gegensatz zur Röntgenstruktur eine Dimerisierung, hervorgerufen durch vier Punktmutationen und einen Austausch eines β-Faltblattes. Eigentlich ist hier die Röntgenstruktur nicht geeignet, die NMR Struktur zu verbessern. Um zu zeigen, dass selbst bei Auswahl einer ungeeigneten Struktur keine Verzerrung der Originalstruktur erfolgen kann, wurde dennoch eine Verbesserung mit ISIC durchgeführt. Das Ergebnis waren Strukturen, die den Originalstrukturen sehr ähnlich sind. RMSD-Wert und *Ramachandran*-Qualität konnten sogar leicht verbessert werden, da an den Stellen, an denen bei den beiden Quell-Strukturen Übereinstimmung zu finden ist, diese Regionen tatsächlich zu einer Verbesserung führten. Wo keine Übereinstimmung herrscht, tritt auf Grund der nicht trivialen Kombination der Restraints in diesem Bereich auch keine Verzerrung auf.

Insgesamt lässt sich sagen, dass der ISIC-Algorithmus ein gutes Werkzeug darstellt, um bessere Lösungen zu finden und unverzerrte Strukturverbesserungen zu erreichen. Es wurden hier Röntgen-Strukturen verwendet, um NMR-Strukturen zu verbessern. Die Qualität der beiden Strukturen im Falle von Byr2-RBD ist mittelmäßig und stellt somit einen idealen

Anwendungsfall für den ISIC-Algorithmus dar. Eine ähnliche Anwendung ist die Bestimmung von NMR-Strukturen von sehr großen Proteinen, wobei man experimentelle Daten nur in begrenztem Maße bestimmen muss. Weiterhin kann man anstelle der Röntgen-Struktur homologe Strukturen verwandter Proteine zur Verbesserung verwenden. Ebenso kann ISIC dazu verwendet werden, Röntgen-Strukturen mit Hilfe von NMR-Strukturen zu verbessern, wenn Teile der Elektronendichteverteilung schlecht zu bestimmen sind.

5.3 Automatische sequentielle Zuordnung (ASSIGN)

Mit dem Modul ASSIGN wird eine automatische sequentielle Zuordnung von NOE-Spektren durch Vergleich von experimentellen und simulierten 2D-NOESY-Spektren angestrebt. Wichtig bei diesem Verfahren ist, dass nicht nur die *Peakpositionen* als Punkte berücksichtigt werden, sondern die Information in den Linienformen und den Signalvolumina eine wichtige Rolle spielt. Da sich durch Überlappung nah beieinander liegender Signale auch die Linienform und das Volumen im jeweiligen Messbereich des Spektrums ändert, ist dies ein wichtiges Maß, um Aussagen für die Übereinstimmung zwischen dem experimentellen und simulierten Spektrum zu treffen. Dabei werden Linienformen und Volumen von Spektrenausschnitten an den definierten Messbereichen der experimentellen Signale sowie die statistische Nutzung von Vorhersagen für chemische Verschiebungen für den Vergleich herangezogen. Durch die iterative Veränderung der chemischen Verschiebungen wird in jedem Schritt das simulierte Spektrum neu aufgebaut. Alle Signale, die zur veränderten chemischen Verschiebung gehören, werden zuerst von der aktuellen Position im simulierten Spektrum subtrahiert und dann an die neue Stelle im Spektrum addiert. Alle Signale, die durch diese Operation wegen Überlappung betroffen sind, werden dann bei der Berechnung des Vergleiches mit dem experimentellen Spektrum, das nie verändert wird, berücksichtigt. Es ist also nicht notwendig, die Wahrscheinlichkeiten für alle Messbereiche immer neu zu berechnen. Auf diese Weise wird sehr viel Rechenzeit eingespart, so dass ein Optimierungslauf in zwischen fünf bis zehn Stunden auf der in Anhang 7.1 genannten Umgebung fertig ist. Um den Bereich für die einzelnen chemischen Verschiebungen einzuschränken, wurde die strukturbasierte statistische Vorhersage für die chemischen Verschiebungen mit Hilfe von SHIFTS genutzt. Der Vorteil hierin ist, dass wahrscheinlichere Positionen der jeweiligen chemischen Verschiebungen öfter angewählt werden und so die richtige Zuordnung wesentlich schneller gefunden wird.

Für die Optimierung der Übereinstimmungen in den Messbereichen ausgedrückt durch *Bayessche* Wahrscheinlichkeiten wird in dieser Arbeit ein *Threshold-Accepting*-(TA)-Algorithmus verwendet. Die Einfachheit, Robustheit und die Geschwindigkeit des Verfahrens gaben hierzu den Ausschlag. Für die Optimierung wären auch genetische Algorithmen oder *Hidden-Markov*-Modelle denkbar, die in ASSIGN einfach zu implementieren wären.

5.3.1 Idealer Datensatz

Als erstes wurde anhand eines idealen Datensatzes getestet, ob der ASSIGN-Algorithmus prinzipiell vom Ansatz her funktioniert. Artefaktsignale, Basislinienschwankungen, etc. werden so ausgeschlossen. Es wurde das simuliertes Spektrum von HPr *S. aureus* (H15A) mit Hilfe der entsprechenden Lösungsstruktur und der zugehörigen sequentiellen Zuordnung berechnet und als experimentelles Spektrum verwendet. Danach wurden alle Peaks gelöscht und mit AUREMOL ein neues *Peakpicking* durchgeführt. Dies führt natürlich zu einer abweichenden *Peakliste*, da Signalüberlappungen die Identifikation aller Peaks verhindern. Aus diesem Grunde wurde auch eine Mehrfachbelegung der möglichen Positionen für die chemischen Verschiebungen (*Slots*) erlaubt. Es waren 4881 Signale und 1123 Slots (mögliche chemische Verschiebungen) vorhanden.

Die Simulation des Vergleichspektrums wurde mit denselben Eingabedaten durchgeführt. 8473 Peaks mit 455 chemischen Verschiebungen wurden auf diese Weise erzeugt. Begonnen wurde der Optimierungslauf mit einer zufälligen *Slot*-Auswahl für die einzelnen chemischen Verschiebungen, die allerdings durch die Vorhersage der chemischen Verschiebungen mittels der zugehörigen Wahrscheinlichkeit beeinflusst ist. Die Rechnung lieferte eine richtige Zuordnung von 99,34 %. Fast alle chemischen Verschiebungen (450 von 455) wurden von ASSIGN in einem Toleranzbereich von $\pm 0,03$ ppm richtig gefunden. Dies zeigt, dass die Vergleichskriterien Linienform, Volumen und Vorhersage der chemischen Verschiebungen sich gut eignen, um die sequentielle Zuordnung mit Hilfe eines 2D-NOESY-Spektrums zu finden. Die Häufigkeitsverteilungen der Pseudoenergieen der Linienformen *ES* (Abbildung 35) und der Volumen *EV* (Abbildung 37) zeigen einen zu erwartenden Verlauf. Bei der richtigen Zuordnung sind gute Energiewerte am häufigsten zu sehen. Umgekehrt sind schlechte Energiewerte bei einer zufälligen Anordnung der chemischen Verschiebungen am meisten anzutreffen. Die daraus resultierenden Wahrscheinlichkeitsfunktionen (Abbildung 36 und Abbildung 38) zeigen eine hohe Selektivität bzgl. guter Energiewerte. Dadurch ist gewährleistet, dass die richtige Zuordnung einen hohen Wahrscheinlichkeitswert besitzt und durch die Maximierung dieses Wertes die richtige sequentielle Zuordnung gefunden werden kann. Der Verlauf des Pseudoenergieterms *ECS* (*Energy Chemical Shifts*) (Abbildung 41) zeigt ebenfalls ein zu erwartendes Verhalten. Am Anfang liegt der Wert bei 324,63, wird dann sehr schnell minimiert und fällt auf Werte von ca. 240 ab. Allerdings entspricht hier ein Minimum nicht unbedingt der richtigen Zuordnung. Das wäre nur dann der Fall, wenn die statistische Vorhersage zu 100 % richtig wäre. In der Realität ist dies aber nicht der Fall. So hat jede chemische Verschiebung einen bestimmten Abstand zu seiner Vorhersage, und die

Summierung dieser Abstände ergibt den Wert, der der richtigen sequentiellen Zuordnung zueigen ist. In diesem Beispiel ist das ca. 300 und mit fortschreitenden Iterationen wird dieser Wert erreicht. Im Vergleich zu den Pseudoenergietermen ES und EV wird deshalb ECS mit nur 25 % gewichtet, damit er bei der Optimierung einen kleineren Einfluss hat. Trotzdem ist ECS ein wichtiger Term, der garantiert, dass die wahrscheinlichsten Lösungen öfter ausgewählt werden und somit den Lauf enorm beschleunigt.

In Abbildung 40 ist der Verlauf der totalen Pseudoenergie E_{tot} und die gefundene Zuordnung in Abhängigkeit von den Iterationen gezeigt. Der direkte Zusammenhang zwischen E_{tot} und der gefundenen Zuordnung ist sehr deutlich zu sehen. Bei bereits $1 \cdot 10^6$ Iterationen liegt E_{tot} im Bereich des Minimums und die gefundene Zuordnung ist zu nahezu 100 % korrekt.

Der Vergleich mit der früheren Version TWOSTEP (Abbildung 39) zeigt deutlich, dass die Neuerungen in ASSIGN sehr effektiv sind. Während ASSIGN ohne partielle Startzuordnung (0%) auskommt, benötigt TWOSTEP mindestens 50 % Startzuordnung, um in die Nähe der richtigen Zuordnungen von ASSIGN zu kommen. Zudem benutzte TWOSTEP bei der Simulation des experimentellen Spektrums eine maximale interatomare Protonendistanz von nur 0,2 nm. Der Vorteil für TWOSTEP war eine viel geringere Anzahl der Signale und damit auch weniger Überlappungen der einzelnen Peaks.

Im nächsten Fall wurde getestet, inwieweit thermisches Rauschen im experimentellen Spektrum ASSIGN beeinflusst. Das maximale Rauschniveau wurde sukzessive von 0 % auf 242 % gesteigert. Je mehr Rauschen, desto weniger selektiv werden die Wahrscheinlichkeitsfunktionen (Abbildung 43-46). Schlechtere Pseudoenergiewerte treten jetzt immer öfter auf. In Abbildung 47 ist die gefundene Zuordnung in Abhängigkeit von der Rauschrate aufgetragen. Es wurde immer mit einer partiellen Startzuordnung von 0 % begonnen. Bis zu einem maximalen Rauschwert $mr = 16\%$ ist fast keine Beeinträchtigung zu erkennen. Bei $mr = 32\%$ werden nur noch 60 % der Zuordnungen gefunden. Trotzdem werden bei $mr = 242\%$ immerhin noch 20 % der richtigen Zuordnung gefunden.

5.3.2 Experimenteller Datensatz

Nachdem ASSIGN beim idealen Datensatz ein gutes Verhalten zeigte, wurde nun das experimentelle Spektrum, das beim idealen Datensatz künstlich erzeugt wurde, durch das entsprechende reale Spektrum ersetzt. Die restliche Umgebung mit Lösungsstruktur und zugehöriger sequentieller Zuordnung blieb gleich. Das Spektrum wurde mit einem 600 MHz Bruker Spektrometer aufgenommen und mit XWINNMR [125] prozessiert. Danach folgte mit

Hilfe der bereits in Kapitel 2.1.1 und Kapitel 2.1.2 vorgestellten AUREMOL Routinen ein *Peakpicking* und eine *Peakintegration*. Eine kurze manuelle Nachbearbeitung stellte sicher, dass keine optisch leicht zu erkennenden Signale fehlten oder *Artefaktpeaks* vorhanden waren. Insgesamt waren 2808 Peaks und 1401 Slots für die chemischen Verschiebungen vorhanden.
Die Häufigkeitsverteilungen der Pseudoenergie-Terme ES und EV sind in Abbildung 48 bzw. Abbildung 49 zu sehen. Es fällt auf, dass für ES (Linienform) das Maximum nicht mehr bei Pseudoenergien im Bereich von $]0,95;1,00]$ liegt, sondern sich auf $]0,90;0,95]$ verschoben hat. Dies ist klar, da sich die Linienformen in realen Spektren durch das thermische Rauschen in den Spulen nicht ideal verhalten. Auch eine Überlagerung von Signalen mit Artefakten aus z.B. Wasser oder anderen Molekülen beeinflussen die realen Linienformen und Volumina.
Hinzu kommt, dass die für die Simulation verwendete Struktur nicht 100 % bekannt ist. Auch die Simulation an sich ist nicht völlig genau. So sind z. B. die Orderparameter, die die Beweglichkeiten der einzelnen Atome beschreiben, unbekannt und der chemische Austausch wird ebenfalls nicht exakt berücksichtigt.
Die Häufigkeitsverteilung der zufälligen Zuordnung verläuft wie erwartet. Das Maximum lieg bei den schlechtesten Energien $[0,00;0,05]$, das Minimum bei $]0,95;1,00]$. Die daraus resultierende Wahrscheinlichkeitsfunktion (Abbildung 50) zeigt ein durchaus selektives Verhalten mit fast glattem Verlauf. Bei der Pseudoenergie EV für die Volumina verhält es sich so, dass das Maximum bei der richtigen Zuordnung bei $]0,95;1,00]$ liegt. Bei der zufälligen Zuordnung liegt das Maximum zwar bei $[0,00;0,05]$, jedoch ist der Anteil bei $]0,95;1,00]$ fast halb so groß, wie bei der richtigen Zuordnung. Dies hat zur Folge, dass mit der zugehörigen Wahrscheinlichkeitsfunktion (Abbildung 50 rechts) Fälle von perfekter Volumenübereinstimmung nur noch eine Wahrscheinlichkeit von knapp 0,7 erhalten. Ansonsten ist der qualitative Verlauf ähnlich zu der Wahrscheinlichkeitsfunktion der Linienform. Dies zeigt, dass das Volumenkriterium nicht so diskriminierend ist wie das Linienformkriterium, was auch der Realität entspricht, wenn betrachtet wird, dass bei einem Vergleich mit falscher Zuordnung Volumenübereinstimmungen durchaus häufiger vorkommen als Linienformübereinstimmungen. Durch den in ASSIGN gewählten Ansatz zur Bestimmung der Wahrscheinlichkeiten von Übereinstimmungen in den Messbereichen der zu vergleichenden Spektren wird der Realität also insofern Rechnung getragen, dass die Linienform ein schärferes Kriterium als das Volumen ist.

Startzuordnung aus zufälligen Zuordnungen

Für diesen Testfall wurden die gefundenen Zuordnungen in Abhängigkeit von der partiellen Startzuordnung bestimmt. Diese partiellen Startzuordnungen wurden erzeugt, indem von der Referenzzuordnung eine entsprechende Anzahl von zufällig ausgewählten chemischen Verschiebungen weggelassen wurde, die dann von ASSIGN wieder zu finden war. Beginnend mit einer partiellen Zuordnung von 0 % bis hin zu 90 % wurden 9 Testläufe ausgeführt. Dabei stellt der Fall mit der Startzuordnung von 0 % einen Sonderfall dar. Da hier keine partielle Zuordnung bekannt ist, werden die Wahrscheinlichkeitsfunktionen aus der richtigen Zuordnung bestimmt. Für alle anderen Fälle werden die Wahrscheinlichkeitsfunktionen aus der partiellen Zuordnung berechnet. Abbildung 51 zeigt das Ergebnis der Testläufe. Für 0 % Startzuordnung werden knapp über 30 % der Zuordnungen richtig gefunden. Bereits ab 20 % Startzuordnung werden über 90 % der Zuordnungen richtig gefunden. ASSIGN findet also auch bei experimentellen Spektren und einem geringen Prozentsatz der Startzuordnung eine gute Lösung.

In Abbildung 52 sind die erreichten totalen Pseudoenergien E_{tot} in Abhängigkeit von der Startzuordnung aufgetragen. Bei einer Startzuordnung von weniger als 20 % liegt E_{tot} mit Werten von 1762,08 (0 %) und 1645,38 (10 %) deutlich über der Referenzenergie von 1264,3, die für die richtige Zuordnung bestimmt wurde. Dies könnte darauf hindeuten, dass der Optimierungs-Algorithmus in einem lokalen Minimum hängen bleibt. Um dieses Problem zu umgehen, wurde versucht, während des Optimierungslaufes den Threshold anzuheben (*Bouncing*), damit es dem System möglich wird, dieses lokale Minimum zu überwinden. Allerdings brachte das keinen Erfolg. Auch andere Änderungen der Optimierungsparameter führten zu keiner Verbesserung. Eventuell können hier die oben genannten Alternativen zur Optimierung (genetischer Algorithmus, *Hidden-Markov*-Modelle) den gewünschten Erfolg liefern. Betrachtet man jedoch den weiteren Verlauf von E_{tot}, so erkennt man bei Startzuordnungen von mehr als 20 %, dass die Referenzenergie in einem geringen Maße unterschritten wird. Dies wiederum spricht für die Optimierungsmethode mit Threshold-Accepting, der es schafft, das globale Minimum zu finden. Das Unterschreiten der Referenzenergie jedoch bedeutet, dass das Minimum der totalen Pseudoenergie E_{tot} nicht dem Maximum der gefundenen Zuordnung entspricht. Allerdings ist dieser Effekt relativ klein und weist darauf hin, dass das System mit dem verwendeten E_{tot} noch nicht vollständig beschrieben ist. Es ist außerdem zu beachten, dass für den Vergleich von Simulation und Experiment nur ein einziges Spektrum (2D-NOESY) verwendet wurde. Abhilfe würde hier

die Hinzunahme von z.B. einem 3D-NOESY-Spektrum, 2D/3D-TOCSY-Spektren, HSQC-Spektrum, 3D-heteronukleare Tripleresonanzspektren schaffen.

Startzuordnung aus H^N H^α-Atomen

Bei diesem Test ging es darum zu zeigen, inwieweit ASSIGN in einem realen Anwendungsfall zum Einsatz kommen kann. Es wurden nur leicht zu gewinnende chemische Verschiebungen der Rückgratatome H^N und H^α als Startzuordnung verwendet. Diese Startzuordnung kann z. B. aus heteronuklearen Tripleresonanzspektren erhalten werden. Somit betrug die partielle Startzuordnung 36,5 % (163 von 447). Die Aufgabe von ASSIGN war es, die fehlenden chemischen Verschiebungen der Seitenkettenatome zu finden. Die sequentielle Zuordnung konnte mit ASSIGN auf 85,2 % gesteigert werden, d.h. es waren 381 von 447 chemische Verschiebungen richtig zugeordnet. 66 Zuordnungen wurden falsch gefunden. Um zu zeigen, welche Fortschritte mit ASSIGN erzielt werden können, wurden nun mit Hilfe der entsprechenden Zuordnungen Strukturen berechnet. Dazu wurde zuerst mit der sequentiellen Zuordnung der Rückgratatome und danach mit der sequentiellen Zuordnung, die ASSIGN gefunden hat, die Signale des 2D-NOESY-Spektrums mit Hilfe des in Kapitel 2.1.3 beschriebenen PEAKASSIGN [20] zugeordnet. Anschließend wurde mit dem in Kapitel 2.1.4 beschriebenen REFINE Abstand-*Restraints* gewonnen. Tabelle 20 zeigt die Anzahl der so erhaltenen Restraints. Für die Berechnung wurden in allen Fällen mit TALOS erhaltene Diederwinkel und Wasserstoffbrücken-*Restraints* eingesetzt. Die Referenzstruktur wurde mit der originalen sequentiellen Zuordnung erstellt. Abbildung 53 zeigt die drei Strukturbündel im Vergleich. Es ist sehr deutlich zu sehen, dass die Strukturen, in denen nur die Zuordnungen der Rückgratatome Verwendung fanden, relativ schlecht bestimmt sind. Hingegen zeigen sich die Strukturen, die mit der sequentiellen Zuordnung von ASSIGN generiert wurden, sehr gut bestimmt. Der RMSD konnte von 0,151 nm auf 0,015 nm verbessert werden. Auch der AUREMOL R-Wert (R_5), der ein Maß für die Übereinstimmung der Strukturen mit den experimentellen Daten liefert, konnte von 0,589 auf 0,328 deutlich verbessert werden und reicht fast an die Referenzstrukturen (0,320) heran. Der *Ramachandran-Plot* liegt mit 94,8 % (*most favored, additional allowed*) ebenfalls im Bereich der Referenzstrukturen. Dies zeigt deutlich, dass mit ASSIGN die sequentielle Zuordnung erfolgreich vervollständigt werden kann. Es ist zu beachten, dass die Strukturen mit einer sequentiellen Zuordnung berechnet wurden, die 14,8 % (66) falsche chemische Verschiebungen enthielt.

Wünschenswert wäre, wenn entweder die Ausbeute von ASSIGN gesteigert werden könnte oder aber wenn chemische Verschiebungen eindeutig als richtig oder falsch identifizieren werden könnten. So könnten mehrere Läufe mit unterschiedlichen *Seeds* für den Zufallsgenerator verschiedene Zuordnungen liefern, bei denen aber immer wieder ein bestimmter Teil gleich zugeordnet würde. Mit Hilfe statistischer Verfahren wie z. B. den χ^2-Test [90] könnten so richtige von falschen Zuordnungen unterschieden werden. Die so als richtig eingestuften Zuordnungen könnten dann in einem nächsten Lauf zur partiellen Startzuordnung hinzugenommen werden und damit ein höhere Zuordnungs-Ausbeute mit ASSIGN erzielt werden.

5.3.3 Punktmutante mit experimentellen Daten

Ein weiterer realer Testfall ist die Bestimmung der sequentiellen Zuordnung der Punktmutante HPr *S. aureus* (H15A) ausgehend von der Struktur und der sequentiellen Zuordnung des Wildtyps von HPr *S. aureus* (wt). Mit PEAKASSIGN wurde die sequentielle Zuordnung des Wildtyps an das Spektrum der Punktmutante angepasst. Auf diese Weise konnten 79,8 % (339 von 425) der chemischen Verschiebungen gefunden werden. Mit ASSIGN konnte diese sequentielle Zuordnung auf 96,7 % gesteigert werden. Die aus den Zuordnungen berechneten Strukturen sind in Abbildung 54 zu sehen. Es ist deutlich zu erkennen, dass die Strukturen im Bereich der Mutation mit der von PEAKASSIGN gelieferten Zuordnung schlecht definiert sind. Die α-Helix und das dahinter liegende β-Faltblatt sind nicht vorhanden. Dagegen zeigen die Strukturen, die mit den Zuordnungen von ASSIGN erstellt wurden, diese Bereiche wieder gut definiert. Der RMSD und der AUREMOL R-Wert (R_5) konnten deutlich verbessert werden, wie Tabelle 24 zeigt. Entscheidend hierbei ist, dass der Bereich der Mutation, insbesondere die chemischen Verschiebungen von ALA 15 richtig gefunden wurden.

5.3.4 Ausblick

Die Ergebnisse von ASSIGN zeigen, dass der Ansatz des Vergleichs von experimentellen mit simulierten Spektren viel versprechend ist. Bereits zum gegenwärtigen Entwicklungszeitpunkt können reale Testfälle erfolgreich bearbeitet werden. Nichtsdestotrotz gibt es einige Ansatzpunkte, die die Zuordnungs-Ausbeute von ASSIGN weiter steigern könnten. ASSIGN benutzt zurzeit nur ein experimentelles 2D-NOESY-Spektrum. Durch Hinzunahme von

weiteren experimentellen Spektren, insbesondere eines 3D-NOESY-Spektrums, könnten ähnliche Wahrscheinlichkeiten bestimmter chemischer Verschiebungen besser differenziert werden. Es kommt auch vor, dass chemische Verschiebungen auf falschen *Slots* sogar bessere Pseudoenergien liefern. Eine Erweiterung der totalen Pseudoenergie E_{tot} durch Terme, die sich auf das 3D-NOESY-Spektrum beziehen, wäre hier sinnvoll. Mit Hilfe von 2D/3D-TOCSY-Spektren, die Signale nur innerhalb einer Aminosäure zeigen, könnte die Lage der chemischen Verschiebungen validiert werden. Mit Hilfe von 3D-heteronuklearen Tripleresonanzspektren könnte die sequentielle Zuordnung des Rückgrates abgeglichen werden. Ein HSQC-Spektrum wäre zur Überprüfung der Lage der $H^{N/C}$-Signale hilfreich.
Doch zur Nutzung dieser weiteren Spektren wäre eine Erweiterung von RELAX notwendig. RELAX kann nur 2D/3D-NOESY-Spektren simulieren. Eine Rückrechnung von den oben genannten Spektren wäre wünschenswert. Zudem wäre an eine Verbesserung der Spektrensimulation zu denken. So sind die Orderparameter, die die Beweglichkeit der einzelnen Atome beschreiben speziell für die Seitenketten weitgehend unbekannt. In RELAX werden zwar verschiedene Bewegungsmodelle für die unterschiedlichen Atomtypen benutzt, jedoch handelt es sich hier um Näherungen. Weiterhin wird in RELAX der chemische Austausch nicht exakt berücksichtigt, so dass immer alle Signale simuliert werden.
Die Wahrscheinlichkeitsfunktionen, die für den Vergleich der Linienformen und Volumen herangezogen werden, werden auf Grund einer bereits bekannten Startzuordnung erstellt. Wenn keine partielle Startzuordnung bekannt ist, ist es im Moment nicht möglich, diese Wahrscheinlichkeitsfunktionen zu erstellen. Eine Lösung hierfür könnte sein, diese Wahrscheinlichkeitsfunktionen durch eine e-Funktion, Cosinus-Funktion oder eine Polynomfunktion anzunähern. Hierzu müssten die Spektren verschiedener Proteine ausgewertet werden, um so eine Tendenz dieser Wahrscheinlichkeitsfunktionen zu erkennen.
Anhand der Beispiele mit HPr *S. aureus* (H15A) ist zu sehen, dass die Pseudoenergieen möglicherweise gaußverteilt sind (Abbildung 50). Ein weiterer Ansatz wäre, am Anfang (ohne partielle Zuordnung) mit einer genäherten Wahrscheinlichkeitsfunktion zu beginnen, die dann während des Optimierungslaufes automatisch optimiert wird, wenn z. B. 10 %, 20 %, usw. der Zuordnung gefunden ist.

5.4 Gegenwärtiger Stand von AUREMOL

Das Softwareprojekt AUREMOL hält zum gegenwärtigen Entwicklungszeitpunkt bereits eine große Menge von automatischen Modulen zur Proteinstrukturbestimmung bereit. Die einzelnen Module benutzen nun alle die gleiche IUPAC-Nomenklatur, was ein Ineinanderschachteln der einzelnen Module ermöglicht. Verschiedene Datenkonverter sind in AUREMOL implementiert, damit auch Fremddaten in AUREMOL genutzt werden können.
Alle Module sind so gestaltet, dass der Benutzer so wenig wie möglich eingreifen muss. Nur beim Start der Module müssen Parameter festgelegt werden. Weiterhin ist nun möglich, dass die Module sich untereinander aufrufen können und so der Automatisierungs-Prozess weiter vereinfacht werden kann.
Für den verwendeten *Top-Down*-Ansatz wurde mit PERMOL in dieser Arbeit eine Homologie-Modellierung erstellt, die Startstrukturen für eine Vielzahl von Modulen bereitstellt. In der Hauptsache profitiert davon RELAX, das zur Simulation von NOESY-Spektren dient. Diese simulierten Spektren finden in den Modulen ASSIGN (sequentielle Zuordnung), KNOWNOE und PEAKASSIGN (NOESY-Zuordnung), REFINE (Atomabstandsgewinnung) und in der R-Wert-Berechnung Verwendung. Dadurch, dass es nun möglich ist, bereits gut definierte Strukturen zu verwenden, kann eine Menge Zeit eingespart werden im Gegensatz zu vorher, wo mit einem ausgestreckten Strang begonnen wurde.
Mit ISIC ist ein Werkzeug entstanden, vorhandene Proteinstrukturen durch Zuhilfenahme anderer Datenquellen weiter zu verbessern. So kann entweder die Startstruktur oder die finale Lösungsstruktur weiter verfeinert werden. Zudem ist es möglich, mit möglichst wenig experimentellen Daten bereits gute Strukturen zu generieren. Sollte zu einem Protein kein homologes Modell vorhanden sein, so kann auf diese Weise eine Startstruktur erzeugt werden.
Die Kombination der vorhandenen Module zu einer vollautomatischen Strukturbestimmung, die nur wenig NMR-Daten und eine Startstruktur benutzt, ist das Ziel von AUREMOL. Ein entscheidender Schritt besteht in der Entwicklung eines flexiblen und verlässlichen Moduls zu Gewinnung der sequentiellen Zuordnung. In dieser Arbeit wurde zu diesem Zweck ASSIGN entwickelt, das in realen Anwendungen bereits gute Ergebnisse liefert. Dass der richtige Weg eingeschlagen ist, zeigen die Ergebnisse klar. Natürlich müssen die in Kapitel 5.3.4 genannten Überlegungen weiterentwickelt und implementiert werden, um für die Vielzahl der

verschiedenen Anwendungen verlässliche Ergebnisse zu liefern. Sobald dies erledigt ist, steht der Vision einer vollautomatischen Strukturbestimmung von Proteinen in Lösung nichts mehr im Weg.

6 Zusammenfassung

In der vorliegenden Arbeit wurde die Entwicklung des Softwareprojekts AUREMOL fortgesetzt. Das Ziel von AUREMOL ist die Bereitstellung von Modulen, die eine automatische Strukturbestimmung von Proteinen in Lösung mit möglichst wenigen experimentellen Daten ermöglichen. Grundsätzliche Neuerungen sind die Einführung einer einheitlichen IUPAC-konformen Nomenklatur der Atomnamen der einzelnen Aminosäuren in den verschiedenen Modulen von AUREMOL und die Bereitstellung entsprechender Konverter, um Fremddaten leicht und schnell nutzen zu können.

Es wurden drei spezielle Module entwickelt, mit deren Hilfe die Proteinbestimmung mittels AUREMOL weiter automatisiert wurde. Der *Top-Down*-Ansatz von AUREMOL verlangt u. a. eine Startstruktur. Im Extremfall kann hier ein ausgestreckter Strang eingesetzt werden. Um jedoch mehr Strukturinformationen gleich von Anfang an nutzbar zu machen und so die Strukturbestimmung zu beschleunigen, wurde als erstes Modul eine Homologie-Modellierung (PERMOL) in AUREMOL implementiert und weiterentwickelt. Da bereits ab einer Sequenzidentität von mehr als 20 % eine Homologie bzgl. Funktion und somit Struktur vorliegen kann, wird die Strukturinformation derartiger, bereits gelöster, Proteine dazu genutzt, Atomabstände, Diederwinkel und Wasserstoffbrücken aus den PDB-Dateien zu berechnen. Diese Daten werden dann als Beschränkungen (*Restraints*) in Moleküldynamik-Programmen verwendet. Die Berechnung von Fehlergrenzen folgt unmittelbar aus den mit Hilfe der jeweiligen Standardabweichung berechneten Konfidenzintervallen. Ebenso können auf diese Weise aus einzelnen Röntgenstrukturen mit Hilfe der für jedes Atom gegebenen B-Faktoren und der Auflösung Strukturbündel gewonnen werden. Die so gewonnenen homologen Strukturen können sehr vielfältig in den verschiedenen Modulen von AUREMOL als Startstruktur eingesetzt werden.

Das zweite Modul (ISIC [71]) dient zur Strukturverbesserung von Proteinen unter Verwendung von Informationen aus anderen Datenquellen. Während der letzten Jahre ist die in der Proteindatenbank hinterlegte Datenmenge enorm angewachsen. Durch die richtige Nutzung können diese Daten eine wichtige Rolle bei künftigen Strukturbestimmungen einnehmen. Mit ISIC wird ein neuer, allgemeiner und vollautomatischer Ansatz für die Kombination von Strukturinformationen vorgestellt. Besonders wichtig hierbei ist, dass die Daten nicht auf einfache Weise gemittelt werden, sondern dass nur relevante Daten zur

Strukturverbesserung beitragen. Auf diese Weise ist es möglich, hoch aufgelöste Strukturen zu berechnen.

Mit ASSIGN wurde in der vorliegenden Arbeit ein drittes Modul entwickelt, mit dessen Hilfe die sequentielle Zuordnung erstellt wird. Der Vergleich eines simulierten und experimentellen 2D-NOESY-Spektrums in definierten Bereichen ermöglicht es ASSIGN, fehlende chemische Verschiebungen zu finden. Dazu werden die fehlenden chemischen Verschiebungen iterativ variiert und jedes Mal das simulierte Spektrum neu aufgebaut. Ein Vergleich von Linienform und Volumina sowie die Nutzung einer statistischen Vorhersage für chemische Verschiebungen liefern eine Bewertung des Vergleichs. Durch Optimierung wird die wahrscheinlichste Lösung als sequentielle Zuordnung festgehalten.

7 Anhang

7.1 Programmierumgebung

Das Softwarepaket AUREMOL ist in der Programmiersprache ANSI C geschrieben. Als Programmierumgebung wurde die IDE (**I**ntegrated **D**evelopment **E**nviroment) Visual Studio 6.0 von Microsoft verwendet. Für die Qualitätskontrolle kam DevPartner Studio 7.1 zum Einsatz.

Als Entwicklungsrechner wurde ein DELL Optiplex GX260 (Intel ® Pentium ® 4 1,8 GHz, 768 MB RAM) unter dem Betriebssystem Microsoft Windows XP verwendet. Die Optimierungsläufe für ASSIGN wurden auf einem DELL Optiplex GX280 (Intel ® Pentium ® 3,2 GHz, 2048 MB RAM) ausgeführt.

Die Moleküldynamik-Simulationen wurden auf dem Linux-Cluster des Rechenzentrums der Universität Regensburg ausgelagert.

7.2 Danksagung

An dieser Stelle möchte ich mich bedanken bei:

- Meinem Doktorvater Prof. Dr. Dr. Hans Robert Kalbitzer für die interessanten Themen, den aufschlussreichen Diskussionen und den vielen Ideen, die zur Erstellung dieser Arbeit wesentlich beitrugen.
- PD Dr. Wolfram Gronwald für die konkreten Hilfestellungen bei allen Problemen, insbesondere NMR spezifischer Natur, die während der Dissertation auftraten.
- Dr. Jochen Trenner für die fachlich kompetente und äußerst angenehme Zusammenarbeit im Softwareprojekt AUREMOL.
- Dr. Jochen Trenner, Dr. Alexander Fink, Dr. Thorsten Graf, Dr. Barbara Domogalla und Ralph Elsner für die angenehme Atmosphäre sowohl innerhalb als auch außerhalb des Lehrstuhls.
- Meinen Zimmerkollegen Dr. Alexander Fink, Gerald Bäuml und Kumaran Baskaran für eine freundliche Arbeitsumgebung.
- Allen Mitgliedern des Lehrstuhls für ein sehr gutes Arbeitsklima.

Besonders bedanken möchte ich mich bei:

- Meiner lieben Ehefrau Julia, die mein Leidklagen geduldig ertrug und mir auch in schweren Zeiten immer Hilfe gab und für mich da war.

- Meinem Sohn Jonas, der mir zur rechten Zeit neue Kraft gab.

- Meiner Familie, insbesondere meinen Eltern, die mir das alles ermöglicht haben.

- Meinem Onkel, Pfarrer Adalbert Brunner, der mir seine Unterstützung im hohen Maße zu Teil werden ließ.

Die vorliegende Arbeit wurde unterstützt von:

- Bruker Biospin GmbH
- Deutsche Forschungsgemeinschaft (DFG)
- SPINE (**S**tructural **P**roteomics **in E**urope)
- Extend-NMR (Europäische Union)

8 Literaturverzeichnis

[1] Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K., & Watson,J.D. (1997) *Molekularbiologie der Zelle*. Wiley-VCH Verlag.

[2] Dobson,C.M. (2003) Protein folding and misfolding. *Nature* **426,** 884-890.

[3] Selkoe,D.J. (2003) Folding proteins in fatal ways. *Nature* **426,** 900-904.

[4] Stefani,M. (2004) Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world. *Biochim. Biophys. Acta* **1739,** 5-25.

[5] Jeyashekar,N.S., Sadana,A., & Vo-Dinh,T. (2005) Protein amyloidose misfolding: mechanisms, detection, and pathological implications. *Methods Mol. Biol.* **300,** 417-435.

[6] Ramos,C.H. & Ferreira,S.T. (2005) Protein folding, misfolding and aggregation: evolving concepts and conformational diseases. *Protein Pept. Lett.* **12,** 213-222.

[7] Ernst,R.R. & Anderson,W.A. (1966) Application of Fourier Transform Spectroscopy to Magnetic Resonance. *Rev. Sci. Instr.* **37,** 93-102.

[8] Cavanagh,J., Fairbrother,W.J., Palmer III,A.G., & Skelton,N.J. (1996) *Protein NMR Spectroscopy Principles and Practice*. Academic Press Inc., San Diego.

[9] Hausser,K.H. & Kalbitzer,H.R. (1989) *NMR für Mediziner und Biologen*. Springer Verlag, Berlin Heidelberg.

[10] Wüthrich,K. (1986) *NMR of proteins and Nucleic Acids*. John Wiley & Sons, New York.

[11] Karplus,M. (1959) Contact electron-spin coupling of nuclear magnetic moments. *The Journal of Chemical Physics* **30,** 11-15.

[12] Neidig,K.-P., Geyer,M., Görler,A., Antz,C., Saffrich,R., Beneicke,W., & Kalbitzer,H.R. (1995) AURELIA, a program for computer-aided analysis of multidimensional NMR spectra. *J. Biomol. NMR* **6,** 255-270.

[13] Gronwald,W. & Kalbitzer,H.R. (2004) Automated structure determination of proteins by NMR spectroscopy. *Prog. NMR Spectrosc.* **44,** 33-96.

[14] Ganslmeier,B. AUREMOL - Softwareprojekt zur automatischen Auswertung von multidimensionalen NMR-Spektren. 2002. Universität Regensburg.
Dissertation

[15] Trenner,J.M. Accurate proton-proton distance calculation and error estimation from NMR data for automated protein structure determination in AUREMOL. 2006. Universität Regensburg.
Dissertation

[16] Koradi,R., Billeter,M., Engeli,M., Guntert,P., & Wuthrich,K. (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J. Magn Reson.* **135,** 288-297.

[17] Antz,C., Neidig,K.-P., & Kalbitzer,H.R. (1995) A general bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis. *J. Biomol. NMR* **5,** 287-296.

[18] Schulte,A.C., Gorler,A., Antz,C., Neidig,K.P., & Kalbitzer,H.R. (1997) Use of global symmetries in automated signal class recognition by a bayesian method. *J. Magn Reson.* **129,** 165-172.

[19] Geyer,M., Neidig,K.-P., & Kalbitzer,H.R. (1995) Automated Peak Integration in Multidimensional NMR Spectra by an Optimized Iterative Segmentation Procedure. *J. Magn Reson. B* **109,** 31-38.

[20] Kirchhöfer,R. Computergestützte Analyse von NMR-Spektren. 2005. Universität Regensburg.
Dissertation

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

[22] Herrmann,T., Guntert,P., & Wuthrich,K. (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319,** 209-227.

[23] Görler,A. & Kalbitzer,H.R. (1997) Relax, a Flexible Program for the Back Calculation of NOESY Spectra Based on Complete-Relaxation-Matrix Formalism. *J. Magn Reson.* **124,** 177-188.

[24] Ried,A., Gronwald,W., Trenner,J.M., Brunner,K., Neidig,K.-P., & Kalbitzer,H.R. (2004) Improved Simulation of NOESY spectra by RELAX-JT2 Including Effects of J-Coupling, Transverse Relaxation and Chemical Shift Anisotropy. *J. Biomol. NMR* **30,** 121-131.

[25] Sussman,J.L., Abola,E.E., Lin,D., Jiang,J., Manning,N.O., & Prilusky,J. (1999) The protein data bank. Bridging the gap between the sequence and 3D structure world. *Genetica* **1-2,** 149-158.

[26] Güntert,P., Mumenthaler,C., & Wüthrich,K. (1997) Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA. *J. Mol. Biol.* **273,** 283-298.

[27] Brunger,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., & Warren,G.L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D. Biol. Crystallogr.* **54,** 905-921.

[28] Delagio,F., Grzesiek,S., Vuister,G.W., Zhu,G., Pfeifer,J., & Bax,A. (1995) NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277-293.

[29] Bonneau,R. & Baker,D. (2001) Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30,** 173-189.

[30] Bonneau,R., Ruczinski,I., Tsai,J., & Baker,D. (2002) Contact order and ab initio protein structure prediction. *Protein Sci.* **11,** 1937-1944.

[31] Haliloglu,T., Kolinski,A., & Skolnick,J. (2003) Use of Residual Dipolar Couplings as Restraints in *Ab Initio* Protein Structure Prediction. *Biopolymers* **70,** 548-562.

[32] Hardin,C., Pogorelov,T.V., & Luthey-Schulten,Z. (2002) Ab initio protein structure prediction. *Curr. Opin. Struct. Biol.* **12,** 176-181.

[33] Pedersen,J.T. & Moult,J. (1997) Ab initio protein folding simulations with genetic algorithms: simulations on the complete sequence of small proteins. *Proteins* **Suppl 1,** 179-184.

[34] Pedersen,J.T. & Moult,J. (1995) Ab initio structure prediction for small polypeptides and protein fragments using genetic algorithms. *Proteins* **23,** 454-460.

[35] Simons,K.T., Bonneau,R., Ruczinski,I., & Baker,D. (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **Suppl 3,** 171-176.

[36] Linge,J.P., Williams,M.A., Spronk,C.A.E.M., Bonvin,A.M.J.J., & Nilges,M. (2003) Refinement of protein structures in explicit solvent. *Proteins* **50,** 496-506.

[37] Schwieters,C.D., Kuszewski,J., Tjandra,N.L., & Clore,G.M. (2003) The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160,** 65-73.

[38] Brünger,A.T. (1992) *X-PLOR Version 3.1*. Yale Univ. Press, New Haven/London.

[39] Gronwald,W., Kirchhofer,R., Gorler,A., Kremer,W., Ganslmeier,B., Neidig,K.P., & Kalbitzer,H.R. (2000) RFAC, a program for automated NMR R-factor estimation. *J. Biomol. NMR* **17,** 137-151.

[40] Koradi,R., Billeter,M., & Wüthrich,K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics* **14,** 51-55.

[41] Laskowski,R.A., Rullmann,J.A.C., MacArthur,M.W., Kaptein,R., & Thornton,J.M. (1996) AQUA and PROCHECK-NMR Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8,** 477-486.

[42] Laskowski,R.A., MacArthur,M.W., & Thornton,J.M. (1998) Validation of protein models derived from experiment. *Curr. Opin. Struct. Biol.* **8,** 631-639.

[43] Browne,W.J., North,A.C.T., Phillips,D.C., Brew,K., Vanaman,T.C., & Hill,R.C. (1969) A possible three-dimensional structure of bovine lactalbumin based on that of hen's egg-white lysosyme. *J. Mol. Biol.* **42,** 65-86.

[44] Chothia,C. & Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823-826.

[45] Sander,C. & Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9,** 56-68.

[46] Martin,A.C.R., MacArthur,M.W., & Thornton,J.M. (1997) Assessment of comparative modeling in CASP2. *Proteins* **29,** 14-28.

[47] Vitkup,D., Melamund,E., Moult,J., & Sander,C. (2001) Completeness in structural genomics. *Nat. Struc. Biol.* **8,** 559-566.

[48] Murzin,A.G., Brenner,S.E., Hubbard,T.J.P., & Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536-540.

[49] Lo Conte,L., Brenner,S.E., Hubbard,T.J.P., Chothia,C., & Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **30,** 264-267.

[50] O'Donovan,C., Apweiler,R., & Bairoch,A. (2001) The human proteomics initiative (HPI). *TRENDS Biotechnol.* **19,** 178-181.

[51] Al-Lazikani,B., Jung,J., Xiang,Z., & Honig,B. (2001) Protein structure prediction. *Curr. Opin. Chem. Biol.* **5,** 51-56.

[52] Martin-Renom,M., Stuart,R.C., Fiser,A., Sanchez,R., Melo,F., & Sali,A. (2000) Comparitive Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291-325.

[53] Guex,N. & Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18,** 2714-2723.

[54] Sali,A., Fiser,A., Sanchez,R., & Marti-Renom,M.A. Modeller, A Protein Structure Modeling Program, Release 6v2. 2002.
Computer Program

[55] Bradley,P., Misura,K.M., & Baker,D. (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* **309,** 1868-1871.

[56] Möglich,A., Weinfurtner,D., Gronwald,W., Maurer,T., & Kalbitzer,H.R. (2005) PERMOL: Restraint-Based Protein Homology Modeling Using DYANA or CNS. *Bioinformatics* **21,** 2110-2111.

[57] Scheiber,J. Entwicklung einer auf Moleküldynamik basierenden Methode zur Homologie - Modellierung der Tertiärstruktur von Proteinen. 2003. Universität Regensburg.
Thesis

[58] Westbrook,J., Feng,Z., Chen,L., Yang,H., & Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.* **31,** 489-491.

[59] Pearson,W.R. & Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A* **85,** 2444-2448.

[60] Altschul,S.F., Gish,W., Miller,W., Myers,E.W., & Lipman,D.J. (1990) Basic local alignmnet search tool. *J. Mol. Biol.* **215,** 403-410.

[61] Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., & Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389-3402.

[62] Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F., & Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25,** 4876-4882.

[63] Needleman,S.B. & Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443-453.

[64] Taylor,W.R. (1986) The classification of amino acid conservation. *J Theor Biol.* **119,** 205-218.

[65] Henikoff,S. & Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A* **89,** 10915-10919.

[66] Forster,M.J. (2002) Molecular modelling in structural biology. *Micron.* **33,** 365-384.

[67] Lottspeich,F. & Zorbas,H. (1998) *Bioanalytik*. Spektrum-Verlag Heidelberg-Berlin.

[68] Heinecke,A. & Köpcke,W. Übungen zur medizinischen Biometrie. http://www.mh-hannover.de/institute/biometrie/JUMBO/bio/script7.html. 2006.
Computer Program

[69] Ried A. Simulation von Linienbreiten,T2-Zeiten und indirekter Spin-Spin-Kopplung in multidimensionalen NOESY-Spektren. 2001. Universität Regensburg.
Thesis

[70] Döker,R., Maurer,T., Kremer,W., Neidig,K.-P., & Kalbitzer,H.R. (1999) Determination of Mean and Standard Deviation of Dihedral Angles. *BBRC* **257,** 348-350.

[71] Brunner,K., Gronwald,W., Trenner,J.M., Neidig,K.P., & Kalbitzer,H.R. (2006) A General Method for the Unbiased Improvement of Solution NMR Structures by the Use of Related X-Ray Data, the AUREMOL-ISIC Algorithm. *BMC. Struct. Biol.* **6,** 14.

[72] Annila,A., Aito,H., Thulin,E., & Drakenberg,T. (1999) Recognition of protein folds via dipolar couplings. *J. Biomol. NMR* **14,** 223-230.

[73] Bowers,P.M., Strauss,C.E.M., & Baker,D. (2000) De novo protein structure determination using sparse NMR data. *J. Biomol. NMR* **18,** 311-318.

[74] Simons,K.T., Kooperberg,C., Huang,E., & Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* **268,** 209-225.

[75] Simons,K.T., Ruczinski,I., Kooperberg,C., Fox,B.A., Bystroff,C., & Baker,D. (1999) Improved Recognition of Native-Like protein Structures Using a Combination of Sequence-Dependent and Sequence-Independent Features of Proteins. *Proteins* **34,** 82-95.

[76] Delagio,F., Kontaxis,G., & Bax,A. (2000) Protein Structure Determination Using Molecular Fragment Replacement and NMR Dipolar Couplings. *J. Am. Chem. Soc.* **122,** 2142-2143.

[77] Andrec,M., Harano,Y., Jacobson,M.P., Friesner,R.A., & Levy,R.M. (2002) Complete protein structure determination using backbone residual dipolar couplings and sidechain rotamer prediction. *J. Struct. Funct. Genomics* **2,** 103-111.

[78] Albrecht,M., Hanisch,D., Zimmer,R., & Lengauer,T. (2002) Improving fold recognition of protein threading by experimental distance constraints. *In Silico Biology* **2,** 1-12.

[79] Li,W., Zhang,Y., Kihara,D., Huang,Y.J., Zheng,D., Montelione,G., Kolinski,A., & Skolnick,J. (2003) TOUCHSTONEX: Protein Structure Prediction With Sparse NMR Data. *Proteins* **53,** 290-306.

[80] Cornilescu,G., Delagio,F., & Bax,A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13,** 289-302.

[81] Shaanan,B., Gronenborn,A.M., Cohen,G.H., Gilliland,G.L., Veerapandian,B., Davies,D.R., & Clore,G.M. (1992) Combining Experimental Information from Crystal and Solution Studies: Joint X-ray and NMR refinement. *Science* **257,** 961-964.

[82] Schiffer,C.A., Huber,R., Wüthrich,K., & Gunsteren,W.F. (1994) Simultaneous Refinement of the Structure of BPTI Against NMR Data Measured in Solution and X-ray Diffraction Data Measured in Single Crystals. *J. Mol. Biol.* **241,** 588-599.

[83] Hoffman,D.W., Cameron,C.S., Davies,C., White,S.W., & Ramakrishnan,V. (1996) Ribosomal Protein L9: A Structure Determination by the Combined Use of X-ray Crystallography and NMR Spectroscopy. *J. Mol. Biol.* **264,** 1058-1071.

[84] Miller,M., Lubkowski,J., Rao,K.K.M., Danishefsky,A.T., Omichinski,J.G., Sakaguchi,K., Sakamoto,H., Apella,E., Gronenborn,A.M., & Clore,G.M. (1996) The Oligomerization Domain of p53: Crystal Structure of the Trigonal Form. *FEBS Lett.* **399,** 166-170.

[85] Chao,J. & Williamson,J.R. (2004) Joint X-Ray and NMR Refinement of the Yeast L30e-mRNA Complex. *Structure* **12,** 1165-1176.

[86] Kirkpatrick,S., Gelatt,C.D., & Vecchi,M.P. (1983) Optimization by Simulated Annealing. *Science* **220,** 671-680.

[87] Möglich,A., Weinfurtner,D., Maurer,T., Gronwald,W., & Kalbitzer,H.R. (2005) A Restraint Molecular Dynamics and Simulated Annealing Approach for Protein Homology Modeling Utilizig Mean angles. *BMC-Bioinformatics* **6,** 91.

[88] Holton,J. & Alber,T. (2004) Automated Protein Crystal Structure Determination using ELVES. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 1537-1542.

[89] Cruickshank,D.W.J. (1999) Remarks About Protein Structure Precision. *Acta Cryst. D* **55,** 583-601.

[90] Sachs,L. (1997) *Angewandte Statistik*. Springer Verlag, Berlin.

[91] Freund,J. 1994. University of Heidelberg.
Dissertation

[92] FELIX. 2003. San Diego CA, Accelrys Inc.
Computer Program

[93] Bartels,C., Xia,Y., Billeter,M., Güntert,P., & Wüthrich,K. (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **5,** 1-10.

[94] Kraulis,P.J. (1989) ANSIG: A Program for the Assignment of Protein ^{1}H 2D NMR Spectra by Interactive Computer Graphics. *J. Magn. Reson.* **84,** 627-633.

[95] Bartels,C., Güntert,P., Billeter,M., & Wüthrich,K. (1997) GARANT-A General Algorithm for Resonance Assignment of Mutidimensional Nuclear Magnetic Resonance Spectra. *J. Comput. Chem.* **18,** 139-149.

[96] Leutner,M., Gschwind,R.M., Liermann,J., Schwarz,C., Gemmecker,G., & Kessler,H. (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *J. Biomol. NMR* **11,** 31-43.

[97] Olson,J.B., Jr. & Markley,J.L. (1994) Evaluation of an algorithm for the automated sequential assignment of protein backbone resonances: a demonstration of the connectivity tracing assignment tools (CONTRAST) software package. *J. Biomol. NMR* **4,** 385-410.

[98] Zimmerman,D., Kulikowski,C., Wang,L., Lyons,B., & Montelione,G.T. (1994) Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods from artificial intelligence. *J. Biomol. NMR* **4,** 241-256.

[99] Lukin,J.A., Gove,A.P., Talukdar,S.N., & Ho,C. (1997) Automated probabilistic method for assigning backbone resonances of (13C,15N)-labeled proteins. *J. Biomol. NMR* **9,** 151-166.

[100] Buchler,N.E.G., Zuiderweg,E.R.P., Wang,H., & Goldstein,R.A. (1997) Protein Heteronuclear NMR Assignments Using Mean-Field Simulated Annealing. *J. Magn. Reson.* **125,** 34-42.

[101] Li,K.B. & Sanctuary,B.C. (1997) Automated resonance assignment of proteins using heteronuclear 3D NMR. 2. Side chain and sequence-specific assignment. *J. Chem. Inf. Comput. Sci.* **37,** 467-477.

[102] Moseley,H.N. & Montelione,G.T. (1999) Automated analysis of NMR assignments and structures for proteins. *Curr. Opin. Struct. Biol.* **9,** 635-642.

[103] Mumenthaler,C. & Braun,W. (1995) Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *J. Mol. Biol.* **254,** 465-480.

[104] Linge,J.P., Habeck,M., Rieping,W., & Nilges,M. (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **19,** 315-316.

[105] Xiao-Ping,X. & Case,D.A. Structure-based chemical shift predictions: Shifts 4.1.1. [4.1.1]. 2002. Department of Molecular Biology, The Scripps Research Institute. Computer Program

[106] Dueck,G. & Scheuer,T. (1990) *Threshold accepting: A general purpose algorithm appearing superior to simulated annealing*. *J. Comput. Phys.* **90,** 161-175.

[107] Merkl,R. & Waack,S. (2003) *Bioinformatik Interaktiv Algorithmen und Praxis.* WILEY-VCH Verlag, Weinheim.

[108] Postma,P.W., Lengeler,J.W., & Jacobson,G.R. (1993) Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. *Microbiol. Rev.* **57,** 543-594.

[109] Maurer,T., Meier,S., Kachel,N., Munte,C.E., Hasenbein,S., Koch,B., Hengstenberg,W., & Kalbitzer,H.R. (2004) High-resolution structure of the histidine-containing phosphocarrier protein (HPr) from Staphylococcus aureus and characterization of its interaction with the bifunctional HPr kinase/phosphorylase. *J. Bacteriol.* **186,** 5906-5918.

[110] Jia,Z., Vandonselaar,M., Hengstenberg,W., Quail,J.W., & Delbaere,L.T. (1994) The 1.6 A structure of histidine-containing phosphotransfer protein HPr from Streptococcus faecalis. *J. Mol. Biol.* **236,** 1341-1355.

[111] Maurer,T., Döker,R., Görler,A., Hengstenberg,W., & Kalbitzer,H.R. (2001) Three-dimensional structure of the histidine containing phosphocarrier protein (HPr) from *Enterococcus faecalis* in solution. *Eur. J. Biochem.* **268,** 635-644.

[112] Jones,B.E., Rajagopal,P., & Klevit,R.E. (1997) Phosphorylation on histidine is accompanied by localized structural changes in the phosphocarrier protein, HPr from Bacillus subtilis. *Prot. Sci.* **6,** 2107-2119.

[113] Gronwald,W., Huber,F., Grünewald,P., Spörner,M., Wohlgemuth,S., Herrmann,C., & Kalbitzer,H.R. (2001) Solution Structure of the Ras binding Domain of the Protein Kinase Byr2 from *Schizosaccharomyces pombe*. *Structure* **9,** 1029-1041.

[114] Scheffzek,K., Grünewald,P., Wohlgemuth,S., Kabsch,W., Tu,H., Wigler,M., Wittinghofer,A., & Herrmann,C. (2001) The Ras-Byr2RBD Complex: Structural Basis for Ras Effector Recognition in Yeast. *Structure* **9,** 1043-1050.

[115] Huber,F., Gronwald,W., Wohlgemuth,S., Herrmann,C., Geyer,M., Wittinghofer,A., & Kalbitzer,H.R. (2000) Letter to the Editor: Sequential NMR Assignment of the Ras-Binding Domain of Byr2. *J. Biomol. NMR* **16,** 355-356.

[116] Nabuurs,S.B., Nederveen,A.J., Vranken,W., Doreleijers,J.F., Bonvin,A.M.J.J., Vuister,G.W., Vriend,G., & Spronk,C.A.E.M. (2004) DRESS: a Database of REfined Solution NMR Structures. *Proteins* **55,** 483-486.

[117] Sippl,M.J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge based prediction of local structures in proteins. *J. Mol. Biol.* **213,** 859-883.

[118] Geyer,M., Herrmann,C., Wohlgemuth,S., Wittinghofer,A., & Kalbitzer,H.R. (1997) Structure of the Ras-binding domain of RalGEF and implications for Ras binding and signalling. *Nat. Struc. Biol.* **4,** 694-699.

[119] Vetter,I.R., Linnemann,T., Wohlgemuth,S., Geyer,M., Kalbitzer,H.R., Herrmann,C., & Wittinghofer,A. (1999) Structural and biochemical analysis of Ras-effector signaling via RalGDS. *FEBS Lett.* **451,** 175-180.

[120] Byeon,I.L., Louis,J.M., & Gronenborn,A.M. (2003) A protein Contortionist: Core mutations of GB1 that Induce Dimerization and Domain Swapping. *J. Mol. Biol.* **333,** 141-152.

[121] Achari,A., Hale,S.P., Howard,A.J., Clore,G.M., Gronenborn,A.M., Hardman,K.D., & Whitlow,M. (1992) 1.67-Å X-ray Structure of the B2 Immunoglobulin-Binding Domain of Strptococcal Protein G and Comparison to the NMR Structure of the B1 Domain. *Biochemistry* **31,** 10449-10457.

[122] Sanchez,R. MODELLER: Low sequence identity. http://salilab.org/archives/modeller_usage/2002/msg00483.html. 2002. Computer Program

[123] Schwede,T., Peitsch,M.C., & Guex,N. Reliability of models generate by SWISS-MODEL. http://swissmodel.expasy.org//SWISS-MODEL.html. 2006. Computer Program

[124] Rieping,W., Habeck,M., & Nilges,M. (2005) Inferential Structure Determination. *Science* **309,** 303-306.

[125] XWINNMR. 2003. Ettlingen, Bruker, Biospin GmbH. Computer Program