Rekonstruktion von Proteinstrukturen aus unvollständigen NMR-Daten

Dissertation zur Erlangung des Doktorgrades der

Naturwissenschaften (Dr. rer. nat.) der

Naturwissenschaftlichen Fakultät III – Biologie und

Vorklinische Medizin – der Universität Regensburg

vorgelegt von

Michael Habeck aus Remagen

Juli 2004

Promotionsgesuch eingereicht am: 7. Juli 2004
Die Arbeit wurde angeleitet von: Dr. M. Nilges

Prüfungsausschuß:

Vorsitzender: Prof. Dr. R. Sterner
1. Gutachter: Prof. Dr. Dr. H. R. Kalbitzer
2. Gutachter: Dr. M. Nilges
3. Prüfer: Prof. Dr. E. Brunner

Zusammenfassung

Gegenstand dieser Arbeit ist die Anwendung der Bayes'schen Wahrscheinlichkeitstheorie auf das Problem der makromolekularen Strukturbestimmung aus NMR-Daten. Ausgehend vom Prinzip der Inferentiellen Strukturbestimmung (ISD), habe ich wahrscheinlichkeitstheoretische Modelle für Messungen skalarer und dipolarer Kopplungen entwickelt. Es zeigt sich, daß die Regeln der Wahrscheinlichkeitstheorie zusätzliche Parameter wie die Fehler der Datensätze sowie Parameter der Theorie (Karplus-Koeffizienten, Saupe-Matrizen, Kalibrationsfaktoren) direkt festlegen; somit werden die sonst üblichen Heuristiken zur Behandlung solcher Größen überflüssig. Für die in dieser Arbeit entwickelten Modelle ist eine analytische Elimierung der zusätzlichen Parameter aus der A posteriori-Verteilung möglich.

In einem wahrscheinlichkeitstheoretischen Kontext verschiebt sich der Schwerpunkt der Strukturberechnung: Gesucht ist nicht bloß die „wahre" Struktur des Moleküls (die a posteriori wahrscheinlichste Konformation), sondern es gilt nun die A posteriori-Verteilung aller Hypothesenparameter zu simulieren, um neben Schätzwerten auch Abschätzungen für ihre Verlässlichkeit zu erhalten. Dazu werden mittels Monte-Carlo-Methoden Stichproben von der A posteriori-Verteilung gezogen. Es können nun sowohl für die Koordinaten, als auch die zusätzlichen Parameter neben ihren wahrscheinlichsten Werten Fehlerbalken angegeben werden; dies ist im Rahmen der optimierungsbasierten Strukturberechnung nicht möglich.

Ich habe die entwickelten Modelle auf reale Datensätze angewendet und anhand ihrer verschiedene Aspekte der NMR-Strukturbestimmung diskutiert. Die geschätzten Strukturen sind von vergleichbarer Qualität wie durch Minimierung berechnete Strukturen. Je mehr Daten in die Analyse einfließen, umso genauer sind die Strukturen bestimmt und umso mehr ähneln sie der Kristallstruktur des Moleküls. Die gemeinsame Analyse mehrerer Datensätze ist möglich, weil jeder Datensatz mit seinem eigenen Fehler in die Analyse eingeht; dieser wird während der Strukturbestimmung geschätzt. Bei Analyse eines einzelnen Datensatzes ergeben sich diesselben Resultate wie bei

der Kreuzvalidierung. Die Schätzung eines Fehlers für einen Datensatz läßt sich direkt auf mehrere Datensätze verallgemeinern, ohne den Rechenaufwand nennenswert zu vergrößern. Damit werden die Gewichte der Datensätze während der Strukturberechnung relativ zu den anderen Datensätzen und den bekannten physikalischen Eigenschaften des Moleküls angepaßt.

Die wahrscheinlichkeitstheoretische Bestimmung der Karplus-Koeffizienten und der Elemente der Saupe-Matrix läßt sich mit den Heuristiken zur Behandlung dieser Größen vergleichen: jene ergeben sich als Spezialfälle der Bayes'schen Analyse und können aus ihr abgeleitet werden. Die Bayes'sche Behandlung hat den Vorteil, daß eine konsistente Verwendung verschiedener Informationsquellen durch die Grundregeln der Wahrscheinlichkeitsrechnung garantiert ist.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Experimente geben nur indirekt Aufschluß über einen Sachverhalt. Verschiedene Experimente können über denselben Sachverhalt Aufschluß geben. Aber selbst bei Wiederholung des gleichen Experiments können die Bedingungen verändert worden sein.

Wie können wiederholte Messungen oder Messungen aus verschiedenen Experimenten gemeinsam zur Aufklärung eines Sachverhalts genutzt werden?

Laplace [1] hat diese Frage allgemeingültig durch die Entwicklung eines Kalküls zur Verknüpfung und Umkehrung von Wahrscheinlichkeiten beantwortet und seine Methode auf viele Datenanalyseprobleme angewendet.

In dieser Arbeit soll die Bestimmung der dreidimensionalen Struktur biologischer Makromoleküle behandelt werden. Unterschiede zu den von Laplace betrachteten Problemen sind nicht prinzipieller Natur, sondern bestehen lediglich in der Komplexität der Aufgabe.

Der Ausgangspunkt wird sein, Strukturbestimmung als Induktionsproblem aufzufassen [2]. Weil die Wahrscheinlichkeitstheorie der einzige konsistente Formalismus ist, um Schlüsse aus unvollständigen Informationen zu ziehen, sollte die Bestimmung makromolekularer Strukturen demnach auf Wahrscheinlichkeiten gegründet sein. Dies unterscheidet diese Arbeit von allen zur Zeit genutzten Methoden.

Trotz der Allgemeingültigkeit des Prinzips werden hier ausschließlich Mo-

delle zur Analyse von Kernresonanzdaten entwickelt. Meßgrößen der Kernresonanz lassen sich in Beziehung zu geometrischen Parametern setzen; somit kann Kernresonanzspektroskopie der Strukturaufklärung dienen. Im Gegensatz zur Röntgenkristallographie ist die Kernresonanzspektroskopie keine geometrische Methode: Die Röntgen-Streumuster sind das direkte Abbild der Atome und über eine Fourier-Transformation mit ihren Koordinaten verknüpft. Dagegen werden die Observablen der Kernresonanz durch komplexe physikalische Theorien beschrieben.

1.1 Kernresonanzspektroskopie

Ein äußeres Magnetfeld koppelt an die Kernspins eines Makromoleküls und kann so entartete Energiezustände aufspalten. Übergänge zwischen den Zuständen werden durch elektromagnetische Strahlung angeregt, die Kernspins geraten aus dem thermodynamischen Gleichgewicht und relaxieren dahin zurück. Weil die Energieniveaus und die Relaxationsraten von der Konfiguration der Kernspins und ihrer zeitlichen Änderung abhängen, können Kernresonanzexperimente geometrische und dynamische Sachverhalte aufklären.

1.1.1 Theorie der Kernresonanz

Der Spin $\mathbf{I}_k$ eines Kerns k mit gyromagnetischem Verhältnis γ_k erzeugt das magnetische Moment $\boldsymbol{\mu}_k = \gamma_k \mathbf{I}_k$. In einem Kernresonanzexperiment wird die zeitliche Änderung der transversalen Komponente der makroskopischen Magnetisierung

$$M_y(t) = \mathrm{tr}\left[\rho(t) \sum_k \gamma_k I_{ky}\right]$$

mit einer Induktionsspule beobachtet [3]. $\rho(t)$ ist Dichteoperator des Systems, er wirkt auf alle Freiheitsgrade. Weil die Observable M_y jedoch nur von den Kernspins abhängt, genügt es, einen reduzierten Dichteoperator

$$\sigma(t) = \mathrm{tr}_l \rho(t) = \sum_l \langle l|\rho(t)|l\rangle$$

zu kennen, der über alle von den Kernspins verschiedenen Freiheitsgrade l, das „Gitter“ (lattice), gemittelt ist. Die makroskopische Magnetisierung ist

$$\mathbf{M}(t) = \mathrm{tr}\left[\,\rho(t)\sum_k \boldsymbol{\mu}_k\,\right] = \mathrm{tr}\left[\,\sigma(t)\sum_k \boldsymbol{\mu}_k\,\right],$$

wobei die erste Spur über alle, die zweite nur noch über die Spinfreiheitsgrade summiert.

Der gesamte Dichteoperator $\rho(t)$ entwickelt sich gemäß der Liouville-Gleichung. Es existiert keine analoge Differentialgleichung, die bloß die Zeitentwicklung der Spinfreiheitsgrade unter dem Einfluß des äußeren Felds und des Gitters exakt beschreibt. Störungsrechnung liefert eine Master-Gleichung für den reduzierten Dichteoperator [3]

$$\dot{\sigma}(t) = -i\,[H, \sigma(t)] - \hat{\hat{\Gamma}}\;(\sigma(t) - \sigma_0). \tag{1.1}$$

Der Hamiltonoperator wird durch Mittelung über die Freiheitsgrade des Gitters auf den Raum der Kernspins reduziert. Der Relaxationssuperoperator $\hat{\hat{\Gamma}}$ beschreibt die Rückkehr der Kernspins in das thermodynamische Gleichgewicht mit Verteilung σ_0; es handelt sich um Effekte zweiter Ordnung.

Kernspektroskopische Experimente messen die Energieniveaus und die Relaxationsraten der Kernspins. Erstere beruhen auf der kohärenten Übertragung von Magnetisierung, letztere auf der Dissipation magnetischer Energie.

1.1.2 Fourier-Spektroskopie

Eine direkte Spektroskopiemethode würde die Reaktion der Kernspins in Abhängigkeit der Frequenz einer monochromatischen Anregung messen. Dies ist jedoch sehr ineffizient.

In der Fourier-Spektroskopie setzt sich das äußere Feld aus einem konstanten Anteil, der die Zeemann-Niveaus aufspaltet, und einem dazu senkrechten Puls zusammen. Der Puls enthält Schwingungen aus einem großen Frequenzbereich; unter ihnen sind solche, die resonant mit den Übergangsenergien der

Spins sind. Ein Puls regt somit alle Kernspins gleichzeitig an. Die Spins relaxieren unterschiedlich schnell ins Gleichgewicht zurück. Jeder Spin führt eine gedämpfte Schwingung aus; die Detektorspule zeichnet die Überlagerung all dieser Schwingungen auf. Fourier-Transformation des Zeitsignals liefert ein eindimensionales Spektrum.

Die mehrdimensionale Fourier-Spektroskopie [3] bedient sich mehrerer Pulse unterschiedlicher Dauer. In einer solchen Pulssequenz werden die Vorbereitungsphase (Dauer τ_P), die Entwicklungsphase (Dauter t_1), die Mischphase (Dauer τ_m) und die Beobachtungsphase (Dauer t_2) unterschieden. In der Vorbereitungsphase wird durch verschiedene Pulse eine bestimmte Magnetisierung hergestellt, die von der Magnetisierung im Gleichgewicht abweicht. In der Entwicklungsphase präzessieren die Spins und tauschen Magnetisierung aus. Die Zeitentwicklung ist jedoch nicht unitär, weil Wechselwirkungen mit dem Gitter zu Dämpfungseffekten führen. In der Mischphase wird durch weitere Pulse gezielt Magnetisierung übertragen. Schließlich wird für die Dauer t_2 das Signal $s(t_1, t_2) \propto M_y(t_1, \tau_m, t_2)$ beobachtet, welches sich aus freien Induktionszerfällen zusammensetzt. Die Entwicklungszeit t_1 wird inkrementell vergrößert; eine Fourier-Transformation entlang der Zeitdimensionen t_1 und t_2 ergibt das zweidimensionale Spektrum

$$S(\omega_1, \omega_2) = \frac{1}{2\pi} \int \mathrm{d}t_1 \, \mathrm{d}t_2 \; s(t_1, t_2) \, e^{-i(\omega_1 t_1 + \omega_2 t_2)},$$

eine Summe von Resonanzen bei verschiedenen Paaren von Eigenfrequenzen des Systems [4]. Nicht-diagonale Resonanzen (cross peaks) geben Aufschluß über Wechselwirkungen zwischen den Spins.

1.1.3 Spin-Spin-Wechselwirkungen

Der effektive Hamiltonoperator der Kernspins lautet

$$H = -\sum_k \gamma_k \; \mathbf{I}_k^T \left(\mathbf{I} - \boldsymbol{\sigma}_k\right) \mathbf{B} + \frac{1}{2} \sum_{k \neq l} \frac{\mu_0 \hbar \, \gamma_k \gamma_l}{4\pi} \; \mathbf{I}_k^T \left(\mathbf{K}_{kl} + \mathbf{D}_{kl}\right) \mathbf{I}_l.$$

Die Elektronen schirmen das äußere Feld ab, so daß die Kernspins leicht verschoben von ihren Larmorfrequenzen schwingen; die chemische Verschiebung

wird durch Abschirmungstensoren $\boldsymbol{\sigma}_k$ beschrieben.

Die Bindungselektronen modulieren lokal das Magnetfeld und führen zu einer indirekten Übertragung der Magnetisierung zweier Kernspins. Der Spin-Spin-Kopplungstensor $\mathbf{K}_{kl}$ beschreibt diese Wechselwirkung; er ist eine phänomenologische Größe, die durch Mittelung über die Freiheitsgrade der Elektronen berechnet werden kann.

Eine direkte Übertragung von Magnetisierung beobachtet man bei der dipolaren Kopplung. Der Dipoloperator

$$\mathbf{D}_{kl} = \left(3\,\frac{\mathbf{r}_{kl}\mathbf{r}_{kl}^T}{r_{kl}^2} - \mathbf{I}\right) \Big/ r_{kl}^3. \tag{1.2}$$

koppelt die Spins und muß über das Gitter gemittelt werden. Nur in anisotropen Medien können dipolare Kopplungen beobachtet werden; sie geben Aufschluß über die Abstände und Orientierungen der Kernpins.

In isotroper Lösung mittelt sich der anisotrope Anteil der Spin-Spin-Kopplung aus. Der isotrope Beitrag

$$2\pi\, J_{kl}\, \mathbf{I}_k^T \mathbf{I}_l \quad \text{mit} \quad J_{kl} = \frac{\mu_0 \hbar}{8\pi^2} \frac{\gamma_k\, \gamma_l}{3}\, \operatorname{tr} \mathbf{K}_{kl}$$

ist die einzig verbleibende Wechselwirkung zwischen Kernspins. Die indirekte Kopplung über drei chemische Bindungen kann näherungsweise aus dem Dihedralwinkel, den die Bindungsvektoren definieren, berechnet werden [5].

Obwohl sich in isotroper Phase die dipolaren Wechselwirkungen ausmitteln, können sie zur Spinrelaxation beitragen. Beim Kern-Overhauser-Effekt (nuclear Overhauser effect, NOE) [6, 4, 7] überträgt die dipolare Relaxation Magnetisierung direkt von einem auf einen anderen Kernspin.

1.1.4 Strukturbestimmungsproblem

Nach Transformation des Zeitsignals erhält man ein mehrdimensionales Spektrum, das sich aus Resonanzen zusammensetzt, die durch ihre Position und Linienform charakterisiert sind. Wegen der chemischen Verschiebung haben die Resonanzen unterschiedliche Frequenzpositionen, dies erlaubt, sie Atomen zuzuordnen, deren Kernspins Magnetisierung ausgetauscht haben.

Aus den spektralen Parametern lassen sich, je nach Experiment, Meßwerte der skalaren und dipolaren Kopplungen sowie die dipolaren Relaxationskonstanten ableiten: Die Feinstruktur der Energieniveaus wird durch Spin-Spin-Kopplungen bestimmt; Höhe und Breite der Resonanzen hängen von der Relaxation der Kernspins ab.

Das Strukturbestimmungsproblem besteht darin, aus Messungen verschiedener spektraler Parameter die mittlere Konformation des Moleküls im thermodynamischen Gleichgewicht zu bestimmen.

1.2 Strukturberechnung

Berechnung der molekularen Struktur aus Messungen kernspektroskopischer Größen ist ein inverses Problem. Optimierungsalgorithmen versuchen es durch direkte Umkehrung der Daten zu lösen. Das inverse Problem ist jedoch nicht eindeutig lösbar: sowohl die Daten als auch die Theorie sind unvollständig und fehlerhaft. Weil es an Prinzipien mangelt, um solche Probleme zu lösen, muß auf Heuristiken zurückgegriffen werden.

1.2.1 Inversion durch Optimierung

Aus Messungen $D = \{y_1, \ldots, y_n\}$ einer kernspektroskopischen Observable y soll unter Verwendung einer Theorie f die unbekannte Struktur X des Makromoleküls bestimmt werden. Idealerweise gilt das Gleichungssystem

$$\begin{aligned} y_1 &= f_1(X) \\ &\vdots \\ y_n &= f_n(X), \end{aligned} \tag{1.3}$$

so daß man versucht ist, die Struktur durch Umkehrung des Zusammenhangs $D = f(X)$ zu bestimmen. Eine analytische Umkehrung der Daten ist nicht möglich. Optimierungsverfahren lösen das Gleichungssystem numerisch. Dazu wird eine Zielfunktion der Form [8]

$$G(X) = E(X) + \lambda F(X) \tag{1.4}$$

bezüglich der Koordinaten X minimiert. Strukturberechnung wird zu einem nicht-linearen Optimierungsproblem.

Die Zielfunktion (1.4) kann unterschiedlich motiviert werden. Man kann die Daten als Zwangsbedingungen $y_i = f_i(X)$ auffassen, die exakt erfüllt sein müssen. Der Datenterm $F(X)$ faßt diese zusammen. Die unbekannte Struktur wird als Minimum der physikalischen Energie $E(X)$ bei gleichzeitiger Erfüllung der Zwangsbedingungen definiert. Nach Lagrange ist das Minimum des eingeschränkten Systems das Minimum der uneingeschränkten Zielfunktion (1.4). Der Lagrange-Multiplikator λ ist so zu wählen, daß die Zwangsbedingungen erfüllt sind.

Man kann aber auch von der Methode der kleinsten Quadrate ausgehen: Die Funktion $F(X)$ mißt Abweichungen zwischen den gemessenen und theoretischen Werten, beispielsweise $F(X) = \sum_i (y_i - f_i(X))^2$. Die Energie $E(X)$ „regularisiert“ das Problem; sie stellt sicher, daß das Minimum auch physikalisch sinnvoll ist. Hier ist λ ein heuristischer Parameter, der die Daten relativ zur Energie gewichtet. Oft wird sich einer physikalischen Metaphorik bedient und λ als „Kraft-“ oder „Kopplungskonstante“ bezeichnet, G wird zu einer „Pseudo-“ oder „Hybridenergie“.

Seit den Anfängen der kernspektroskopischen Strukturbestimmung sind die dipolaren Relaxationsraten, gemessen im NOESY-Experiment [4], die wichtigsten Meßgrößen. Die isolated spin pair approximation (ISPA) [7] ist eine Näherung für den Zusammenhang zwischen dem Volumen eines Kreuzsignals und der Struktur: Weil die Relaxation ein Effekt zweiter Ordnung ist und die dipolare Wechselwirkung von r^{-3} abhängt, ist die Relaxationsrate und damit näherungsweise das Volumen V proportional zur inversen sechsten Potenz des Abstands r der Kernspins: $V \propto r^{-6}$. Eine Umkehrung dieser Relation liefert Abstände, die in der gesuchten Struktur erfüllt sein müssen. Wegen der Unzulänglichkeiten der Daten und der ISPA wird ein Volumen üblicherweise in eine obere Schranke für den Protonenabstand umgewandelt [9].

Die Distanzgeometrie [10] faßt solche Distanzschranken als Zwangsbedin-

gungen auf. Ein unvollständiger Satz experimenteller oberer Abstandsschranken wird durch untere Schranken, die Summen der van der Waals-Radien, ergänzt und die Struktur in die Abstandsintervalle eingebettet.

Ausgehend von der Methode der kleinsten Quadrate, wurde Strukturberechnung durch Moleküldynamik [11] als Alternative zur Distanzgeometrie entwickelt. Die physikalische Energie wird um einen Datenterm erweitert und durch Integration der Newton-Gleichungen minimiert; der Gradient ∇G ist eine Pseudokraft, welche die Koordinaten auf die gesuchte Struktur zieht.

1.2.2 Schwierigkeiten

Strukturberechnung als inverses Problem aufzufassen und durch Optimierung zu lösen, muß zu Schwierigkeiten führen, weil es sich aus mehreren Gründen um ein mathematisch schlecht gestelltes Problem (ill-posed inverse problem) [12] handelt:

Unterbestimmtheit der Theorie

Die Umkehrung einer Theorie f, welche Messungen mit Atompositionen verknüpft, ist nicht eindeutig möglich, falls unterschiedliche Konformationen zu den gleichen Daten führen. Die Theorie ist dann degeneriert. Die Resonanzen eines NOESY-Spektrums hängen von den Abständen räumlich naher Protonen ab. Deshalb können Strukturen mit derselben Protonenkonfiguration experimentell nicht unterschieden werden; die übrigen Atompositionen bleiben auch in einem vollständigen Spektrum unbestimmt.

Unvollständigkeit der Daten

Zusätzlich sind die Daten in der Regel unvollständig. Das Gleichungssystem (1.3) ist dann unterbestimmt und mehrere Strukturen können zu denselben, unvollständigen Messungen geführt haben. In einem NOESY-Spektrum werden beispielsweise manche Resonanzen aufgrund konformationeller Fluktuationen nicht beobachtet. Es ist auch nicht immer möglich, eine Resonanz

einem Paar von Protonen zuzuordnen.

Fehlerhaftigkeit der Theorie

Die Theorie f ist nur näherungsweise bekannt oder wird aus Gründen der Praktikabilität starken Vereinfachungen unterworfen. Die Beschreibung von NOESY-Resonanzen durch die ISPA berücksichtigt weder Spindiffusion noch interne Dynamik. Da die Theorie fehlerhaft ist, reicht es nicht aus, sie umzukehren. Meist werden die Daten nicht in Zwangsbedingungen für einen Sollwert umgewandelt, sondern ein Intervall an Werten zugelassen, die um den Sollwert streuen.

Fehlerhaftigkeit der Daten

Die Messungen y_i sind aufgrund unkontrollierbarer Einflüsse fehlerhaft. Neben experimentellen Fehlern führen Artefakte der Transformation des Zeitsignals sowie Fehler bei der Interpretation der Spektren zu Ungenauigkeiten. Einem Meßwert muß deshalb eine unbekannte Abweichung zugestanden werden.

Behandlung mehrerer Datensätze

Mehrere Experimente ergeben Datensätze $D_1, \ldots, D_m$; jeder Datensatz erzeugt ein Gleichungssystem (1.3). Wären die Messungen und die Theorie exakt, würde die Umkehrung der einzelnen Datensätze zu individuellen Lösungsmengen führen; die gemeinsame Lösung wäre deren Schnittmenge. Bei realen Daten werden die Verhältnisse komplizierter, weil die Messungen nicht vollkommen konsistent sind. Sie müssen mit unterschiedlichem Gewicht in die Zielfunktion eingehen:

$$G(X) = E(X) + \sum_{j=1}^{m} \lambda_j F_j(X). \tag{1.5}$$

Die Lagrange-Multiplikatoren oder Gewichte $\lambda = \{\lambda_1, \ldots, \lambda_m\}$ legen fest, welchen Daten mehr und welchen weniger vertraut wird. Ein objektives Kri-

terium, um diese Parameter zu wählen, gibt es aber nicht.

Verwendung von Vorwissen

Über makromolekulare Strukturen gibt es reichhaltige Kenntnisse, die in die Interpretation der Daten einfließen sollten. Für ein Protein läßt sich schon vor Durchführung des Experiments seine kovalente Struktur angeben. Die Unterbestimmtheit des inversen Problems wird durch Berücksichtigung solcher Kenntnisse gemildert. Es ist jedoch nicht klar, wie stark sie in die Strukturberechnung eingehen sollen.

Zusätzliche Parameter

Die Theorie ist oft parametrisiert, so daß $D = f(X; \alpha)$ für eine ideale Messung bei bekannter Parametrisierung α gilt. In der Praxis ist die Parametrisierung jedoch unbekannt und kann auch nicht experimentell bestimmt werden. Beispielsweise wird die ISPA durch verschiedene „Kalibrationsalgorithmen“ [13, 14] geeicht. Die Datengewichte λ_i können bestenfalls durch Kreuzvalidierung [15, 16] bestimmt werden. Neben den Koordinaten sind somit auch die Parameter α und λ unbekannt. Die Zielfunktion $G(X)$ ist aber nur auf dem Konfigurationsraum definiert und kann nicht der direkten Bestimmung zusätzlicher Parameter dienen.

Güte der berechneten Strukturen

Weil in die Zielfunktion fehlerhafte Größen eingehen, ist eine Strukturbestimmung nur vollständig, wenn neben den Koordinaten auch ihre Genauigkeit angegeben wird. Es ist darüberhinaus von Interesse, wie sich die Struktur bei Änderungen der Daten verhält, oder welche Konfigurationen sonst von den Daten vergleichbar gut gestützt werden. Optimierung liefert im Idealfall eine Lösung, das globale Minimum der Hybridenergie. Aussagen über deren Genauigkeit sind nicht möglich. Man behilft sich, indem man statt einer einzigen eine Menge von Konfigurationen angibt, die durch mehrmaliges

Ausführen des Algorithmus bei verschiedenen Anfangsbedingungen erhalten wurden. Ein solches „Ensemble“ spiegelt jedoch auch Eigenschaften des Algorithmus wider. Ein Maß für die Güte der Struktur sollte jedoch nur von den Daten, nicht der Berechnungsmethode abhängen. Außerdem beeinflussen die Parameter α und λ die Streuung des Ensembles [17]. Ein großes Datengewicht führt beispielsweise zu engen Strukturensemblen, ohne daß dies die Daten tatsächlich hergeben.

1.2.3 Strukturbestimmung als Induktionsproblem

Ursprung all dieser Schwierigkeiten ist letztlich ein Mangel an Informationen, deren Kenntnis Strukturbestimmung zu einem wohl definierten Problem machen würde: Wir kennen weder die Ungenauigkeiten der Messungen noch der Theorie; wir wissen nicht, wie die Messungen gegenseitig und im Verhältnis zu Vorkenntnissen gewichtet werden müssen; die Theorie ist unvollständig, ihre Parametrisierung muß gemeinsam mit den Koordinaten aus den Daten erschlossen werden.

Diese Schwierigkeiten sind nicht mathematischer Natur, sondern erwachsen aus der Unvollständigkeit der Fragestellung.

Einer Umkehrung der Daten entspräche die Schlußfolgerung $D \to X$. Nur wenn die Struktur vollständig durch die Messungen bestimmt wird, läßt sich ein solcher Schluß eindeutig ziehen. Doch ermöglicht die Theorie bestenfalls den umgekehrten Schluß $X \to D$. Aber auch in dieser Schlußrichtung sind keine eindeutigen Aussagen möglich, denn die Theorie fußt auf Vereinfachungen und enthält unbekannte Größen. Weil es unmöglich ist, eindeutig von den Daten auf die unbekannte Struktur zu schließen, ist Strukturbestimmung ein Induktionsproblem.

Optimierung stellt keine Prinzipien zur Verfügung, um Induktionsprobleme zu lösen. Deshalb greifen traditionelle Strukturbestimmungsmethoden auf Heuristiken zurück. Die Komplexität der Daten und des Systems verhindert zu überschauen, wie sich zusätzlich getroffene Annahmen auswirken und ob sie in sich stimmig sind.

Es bedarf eines Formalismus, der konsistent mit Induktionsproblemen umzugehen vermag. Er muß vollständig in dem Sinn sein, daß zusätzliche Annahmen überflüssig werden.

Hier hilft folgende Einsicht weiter: In Anbetracht bekannter Tatsachen erscheinen manche Schlußfolgerungen schlüssiger als andere, und ein Induktionsproblem wird dadurch gelöst, daß wir angeben, wie *wahrscheinlich* diese oder jene Schlußfolgerung ist (siehe Abschnitt 2.2).

1.3 Zielsetzung

In einer wahrscheinlichkeitsbasierten Behandlung wird dem induktiven Charakter des Strukturbestimmungsproblems unmittelbar Rechnung getragen. Das Ziel, die vermeintlich eindeutig bestimmte Struktur zu berechnen, wird aufgegeben. Stattdessen werden Wahrscheinlichkeiten der möglichen Konfigurationen berechnet (Prinzip der Inferentiellen Strukturbestimmung [2]).

Dieses Prinzip soll in dieser Arbeit auf verschiedene kernspektroskopische Parameter, die Aufschluß über die molekulare Struktur geben, angewendet werden. Neben den traditionell verwendeten dipolaren Relaxationsraten sind dies vor allem skalare und dipolare Kopplungen. In dieser Arbeit werden einfache wahrscheinlichkeitstheoretische Modelle für diese Meßgrößen entwickelt und auf veröffentlichte Datensätze angewendet. Die Vielzahl konformationeller Freiheitsgrade macht eine numerische Behandlung unumgänglich. Dazu lassen sich Algorithmen zur Simulation thermodynamischer Systeme nutzen.

Bei der Interpretation von Kernresonanzmessungen treten zwei grundsätzliche Schwierigkeiten auf: Zum einen enthalten die Theorien, die die Meßgrößen beschreiben, empirische Parameter, die nicht bekannt sind und auch nicht direkt gemessen werden können. Zum anderen ist bei der gemeinsamen Analyse mehrerer Experimente nicht von vornherein klar, wie stark die einzelnen Datensätze in die gemeinsame Analyse eingehen sollen. Weil diese Schwierigkeiten jedoch bloß bestimmte Formen von Unwissen darstellen, löst sie der Wahrscheinlichkeitskalkül während der Strukturberechnung, ohne

weiterer Annahmen zu bedürfen.

Die wahrscheinlichkeitsbasierte Strukturbestimmung bietet somit eine Möglichkeit, während der Strukturberechnung unbekannte Theorieparameter wie die Karplus-Koeffizienten oder die mittlere Ausrichtung des Moleküls, beschrieben durch die Saupe-Matrix, zu bestimmen. In der traditionellen Strukturberechnung werden diese Größen durch verschiedene Heuristiken gewählt. Wie sich herausstellen wird, enthält die probabilistische Strukturbestimmung diese Heuristiken als Spezialfälle.

Dieselbe Argumentation läßt sich auch auf die Wahl der Gewichte λ_j in (1.5) anwenden. Im Falle daß nur ein Datensatz vorliegt, ermöglicht die Kreuzvalidierung [15, 16] eine quantitative Bestimmung von λ. Die Bayes'sche Strukturbestimmung läßt sich mit dieser Vorgehensweise vergleichen; sie liefert ähnliche Ergebnisse. Die Kreuzvalidierung wird jedoch zu rechenaufwendig, wenn mehrere Datensätze gleichzeitig analysiert werden sollen. In der Bayes'schen Strukturbestimmung ist eine solche gemeinsame Analyse ohne nennenswerten Mehraufwand möglich. Eine gemeinsame, nach Qualität des Datensatzes gewichtete Analyse der Daten erlaubt Strukturen zu bestimmen, die genauer festgelegt und näher zur wahren Struktur sind.

Kapitel 2

Materialien und Methoden

2.1 Wahrscheinlichkeit und Induktion

Die Definition der Wahrscheinlichkeit ist umstritten. Gewöhnlich werden Wahrscheinlichkeiten als Häufigkeiten beobachteter Ereignisse aufgefaßt. Dieser „frequentistischen“ Interpretation widerspricht die Ansicht, daß Wahrscheinlichkeiten Wahrheitsgehalte bewerten. Hiernach erweitert die Wahrscheinlichkeitslehre die Logik und bildet einen Formalismus, um induktive Schlüsse zu ziehen.

2.1.1 Wahrscheinlichkeit

Der Sprachgebrauch legt nahe, Wahrscheinlichkeiten zur Lösung von Induktionsproblemen zu verwenden: Aus einem Sachverhalt schließen wir auf das Eintreffen eines Ereignisses; wie sicher wir uns über die Gültigkeit unserer Schlüsse sind, hängt davon ab, wieviel wir wissen. Wir sagen, eine Aussage werde „wahrscheinlicher“, wenn wir zusätzliche Kenntnisse in Betracht ziehen. Selbst wenn ungewiß ist, ob ein Sachverhalt aus einer bekannten Tatsache sicher folgt, können wir ihm eine „Wahrscheinlichkeit“ zuordnen.

Der sich abzeichnende Begriff der Wahrscheinlichkeit unterscheidet sich von ihrer verbreiteten Definition als Häufigkeit von Ereignissen unter „zufälligen“ Bedingungen. Nach dieser Auffassung bestimmt ein realer Vorgang, ein

„Zufallsprozeß“ oder „Rauschen“, die Wahrscheinlichkeit. Wahrscheinlichkeiten werden meßbar. Diese Definition ist jedoch zirkulär: Die Wahrscheinlichkeit wird auf den Zufall zurückgeführt; was jedoch zufällig, also regellos und unvorhersagbar, ist, hängt vom Betrachter ab und bleibt unbestimmt.

Es ist ein Mangel an Wissen, der den Gebrauch von Wahrscheinlichkeiten erfordert: Eine Münze wird geworfen, und für den Betrachter ist das eine wie das andere Ereignis, „Kopf“ oder „Zahl“, gleich wahrscheinlich. Er macht diese Zuweisung, weil seine Kenntnisse nicht zulassen, zwischen beiden Ereignissen zu unterscheiden: eines kann so gut eintreten wie das andere.

Wahrscheinlichkeiten sind somit Ausdruck unseres Wissens, nicht beobachtete Häufigkeiten: „Nur in Ermangelung der Gewißheit gebrauchen wir die Wahrscheinlichkeit.“ [18]

Dennoch können Wahrscheinlichkeiten mit Häufigkeiten in Zusammenhang gebracht werden: Aus Mangel an Kenntnissen folgern wir, daß „Kopf“ und „Zahl“ gleich wahrscheinlich sind, und daß sich somit bei wiederholtem Münzwurf die Häufigkeiten der beiden Ereignisse einander immer mehr angleichen. Nun können jedoch uns unbekannte Umstände vorliegen: Die Münze kann auf beiden Seiten eine Zahl haben, der Werfer kann so geschickt sein, daß er öfter Kopf wirft, usw. Die beobachteten Ereignisse weichen von unserer Erwartung ab und wir bezweifeln, über ausreichende Kenntnisse zu verfügen.

Ein Versuch bestätigt oder widerlegt eine Wahrscheinlichkeitszuweisung. Abweichungen verweisen auf unbekannte Umstände, die bei der Zuweisung der Wahrscheinlichkeit nicht berücksichtigt wurden.

2.1.2 Verknüpfung von Wahrscheinlichkeiten

Wir wägen die Wahrscheinlichkeit gegen das ab, was wir bereits wissen; dies drückt das Symbol

$$P(H|I)$$

aus. Die Zahl $P(H|I)$ entspricht der Wahrscheinlichkeit, daß eine Hypothese H zutrifft, wenn wir uns bekannte Tatsachen I in Betracht ziehen. $P(H|I)$ ist eine *bedingte* oder *konditionelle* Wahrscheinlichkeit. Wahrscheinlichkeiten

müssen immer in ein Verhältnis zu dem, was wir wissen, gesetzt werden; absolute Wahrscheinlichkeiten gibt es nicht. Wahrscheinlichkeiten lassen sich für jedwede Hypothese angeben, betreffen also nicht nur die „zufälligen Ereignisse“ der frequentistischen Auffassung. („Es gibt keinen besonderen Gegenstand, der den Wahrscheinlichkeitssätzen eigen wäre.“ [18])

Wahrscheinlichkeiten sind stets bedingt; was jedoch bedingen und was ausgesagt werden kann, ist offen. Eine Vermutung, deren Wahrscheinlichkeit wir bestimmen wollen, kann sich aus Grundhypothesen zusammensetzen. Die Regeln zur Verknüpfung der Hypothesen sind die Gesetze der Logik. Weil sich jede logische Funktion aus „UND“ und „NICHT“ zusammensetzen läßt, genügt es, das Verhalten von Wahrscheinlichkeiten unter diesen Grundoperationen zu betrachten, um die Gesetze zur Verknüpfung von Wahrscheinlichkeiten abzuleiten.

Dies wurde von Cox erkannt, der aus zwei Forderungen an die Wahrscheinlichkeit $P(AB|I)$ der Konjunktion zweier Hypothesen A und B (AB steht für „A UND B“) und an die Wahrscheinlichkeit $P(\overline{A}|I)$ der Negation einer Hypothese A ($\overline{A}$ bezeichnet „NICHT A“) die Grundregeln der Wahrscheinlichkeitlehre ableiten konnte [19]. Cox' Forderungen sind:

1. „The probability of an inference on given evidence determines the probability of its contradictory on the same evidence.“ ([19], S. 3)

2. „The probability on given evidence that both of two inferences are true is determined by their separate probabilities, one on the given evidence, the other on this evidence with the additional assumption that the first inference is true.“ ([19], S. 4)

Aus diesen beiden Forderungen leitete Cox die Grundgesetzte der Wahrscheinlichkeitsrechnung, die Produktregel

$$P(AB|I) = P(A|BI)\,P(B|I) \tag{2.1}$$

und die Summenregel

$$P(\overline{A}|I) = 1 - P(A|I), \tag{2.2}$$

ab, ohne auf eine Definition der Wahrscheinlichkeit als Häufigkeit zurückgreifen zu müssen. Cox' Theoreme zeigen, daß sich der qualitative und abstrakte Begriff der Wahrscheinlichkeit auf einen geschlossenen quantitativen Formalismus übertragen läßt.

2.1.3 Induktion

Die Wahrscheinlichkeitslehre erweitert die Logik [20]. Unabhängig von den Vorkenntnissen I ist die Wahrscheinlichkeit der falschen Aussage $A\overline{A}$

$$P(A\overline{A}|I) = P(\overline{A}|AI)\,P(A|I) = (1 - P(A|AI))\,P(A|I) = 0.$$

Für die wahre Aussage $A + \overline{A}$ folgt: $P(A + \overline{A}|I) = 1 - P(A\,\overline{A}|I) = 1$. Die Grenzwerte der Wahrscheinlichkeit entsprechen damit den Wahrheitswerten der Logik. Die Logik wird um ein Kontinuum von Wahrheitswerten erweitert, so daß es sinnvoll wird, von mehr oder weniger sicheren Aussagen zu sprechen.

Die Wahrscheinlichkeit $P(B|AI)$ bewertet die Implikation $A \to B$ vor dem Hintergrund des Gewußten I. Falls unsere Vorkenntnisse I darin bestehen, zu wissen, daß B aus A folgt, gilt: $P(B|AI) = P(AB|I)/P(A|I) = P(A|I)/P(A|I) = 1$, weil die Implikation gleichbedeutend mit der Gültigkeit von $AB = A$ ist. Das heißt: „Die Gewißheit des logischen Schlusses ist ein Grenzfall der Wahrscheinlichkeit." [18]

Wie schlüssig wir aus einer Tatsache A einen Sachverhalt B folgern können, gibt die Wahrscheinlichkeit $P(B|AI)$ an, wobei I zusätzliche Kenntnisse zusammenfaßt. Falls dieses Wissen darin besteht, daß A B impliziert, also $AB = A$ gilt, folgt:

$$P(A|BI) = \frac{P(AB|I)}{P(B|I)} = \frac{P(A|I)}{P(B|I)} \geq P(A|I).$$

Die Kenntnis, daß B eingetroffen ist, macht umgekehrt A wahrscheinlicher.

Diese Regel verallgemeinert der Satz von Bayes [21, 1]. Aus der Produktregel (2.1) und der Konsistenzforderung $P(AB|I) = P(BA|I)$ folgt:

$$P(A|BI) = P(A|I)\,\frac{P(B|AI)}{P(B|I)}. \tag{2.3}$$

Anwendung des Bayes'schen Satzes erlaubt, Schlüsse umzukehren. Falls eine Verknüpfung zwischen zwei Tatsachen A und B bekannt ist und in der Wahrscheinlichkeit $P(B|AI)$ dargestellt wird, läßt sich daraus die Wahrscheinlichkeit $P(A|BI)$ des umgekehrten Zusammenhangs ableiten.

Die Wahrscheinlichkeitslehre legt fest, wie Sachverhalte aus unzureichenden Kenntnissen zu folgern sind. Der Gebrauch von Wahrscheinlichkeiten bewahrt vor widersprüchlichen oder voreingenommenen Schlüssen. Cox' Theoreme zeigen, daß die Wahrscheinlichkeitstheorie der einzige Formalismus induktiven Schließens ist.

2.1.4 Parametrisierte Hypothesen

Die Menge aller Elementarhypothesen H_i über einen Sachverhalt bildet einen *Hypothesenraum*. Der einfachste Hypothesenraum ist $\{H, \bar{H}\}$; die Summenregel (2.2) drückt seine Vollständigkeit aus. Bilden einander ausschließende Vermutungen $H_1, \ldots, H_n$ den Hypothesenraum, verallgemeinert sich die Summenregel zu

$$P(H_1|I) + \cdots + P(H_n|I) = 1.$$

Die Vollständigkeit des Aussagensystems – $H_1 + \cdots + H_n$ ist stets wahr – überträgt sich auf die Wahrscheinlichkeiten.

Betreffen die Hypothesen einen Parameter ξ, so bilden die möglichen Werte von ξ ein System einander ausschließender Hypothesen; $P(\xi|I)$ gibt die Wahrscheinlichkeit der Hypothese $H =$ „der Wert des Hypothesenparameters ist ξ“ an. Der Hypothesenraum besteht aus unendlich vielen Elementarhypothesen und die verallgemeinerte Summenregel wird ein Integral über die *Wahrscheinlichkeitsdichte* $p(\xi|I)$:

$$\int \mathrm{d}\xi \; p(\xi|I) = 1.$$

Wahrscheinlichkeiten von ξ können auf andere Parameter η übertragen werden, die über $\eta = T(\xi)$ eineindeutig aus jenen hervorgehen. Die Vollständig-

keit bleibt erhalten:

$$1 = \int \mathrm{d}\xi \; p(\xi|I) = \int \mathrm{d}\eta \; \tilde{p}(\eta|I),$$

wobei $\tilde{p}$ die Dichte in den neuen Parametern ist. Die Substitutionsregel ergibt:

$$\tilde{p}(\eta|I) = p(T^{-1}(\eta)|I) \left| \frac{\partial T^{-1}}{\partial \eta} \right| ; \tag{2.4}$$

die Jacobi-Determinante $|\partial T^{-1} / \partial \eta|$ garantiert, daß die Wahrscheinlichkeit von Bereichen im Hypothesenraum unter Transformationen erhalten bleibt.

2.2 Inferentielle Strukturbestimmung

Die molekulare Struktur durch direkte Umkehrung der Daten zu bestimmen, führt zu den in Abschnitt 1.2 genannten Schwierigkeiten. Deren Ursache ist die Unvollständigkeit der Information, die nach der Messung vorliegt; sie reicht für einen eindeutigen Umkehrschluß nicht aus. Dieser Mangel kann durch eine Erweiterung der Analysemethode ausgeglichen werden. Die aus den Messungen gezogenen Schlüsse bleiben unsicher und werden gemäß ihrer Wahrscheinlichkeit bewertet. Die Wahrscheinlichkeitstheorie ist der gesuchte Formalismus, um das Problem der Strukturbestimmung zu lösen.

2.2.1 Wahrscheinlichkeitstheoretische Lösung

Der idealisierte Zusammenhang $D = f(X)$ entspricht dem Schluß $X \to D$ von der Struktur auf die Daten. Das Gleichungssytem (1.3) direkt zu lösen, bedeutet, den Umkehrschluß zu ziehen. Dies führt zu Schwierigkeiten (siehe Abschnitt 1.2.2), weil ein eindeutiger Zusammenhang $D \to X$ nicht existiert.

Ein bescheidenerer Standpunkt muß eingenommen werden: statt deduktiven, also absolut sicheren Schlüssen begnügen wir uns mit induktiven, nur bis zu einem gewissen Grad sicheren Schlüssen. Den Grad unserer Gewißheit gibt eine Wahrscheinlichkeit an. Statt nach einer einzigen Konfiguration zu suchen und diese mit der wahren Struktur zu identifizieren, fragen wir:

> Wie wahrscheinlich ist X die Konfiguration des Moleküls, wenn die Messungen D und sonstige Kenntnisse I in Betracht gezogen werden?

Zur Beantwortung dieser Frage ordnen wir jeder Struktur ihre Wahrscheinlichkeit

$$P(X|DI)$$

zu. Die Symbole X und D stehen hier für die Aussage „die Konformation des Moleküls ist X“ bzw. „gemessen wurde D“. Alle zusätzlichen Annahmen, werden als bedingende Sachverhalte, zusammengefaßt in I, aufgeführt.

Unser Prinzip zur Lösung eines Strukturbestimmungsproblems ist, die Wahrscheinlichkeit $P(X|DI)$ aufzustellen und aus ihr Aussagen über die Konfiguration des Moleküls abzuleiten; wir nennen diesen Zugang *Inferentielle Strukturbestimmung* (inferential structure determination) [2, 22][1].

Die gesuchte Wahrscheinlichkeit läßt sich nicht direkt angeben. Der Satz von Bayes (2.3)

$$P(X|DI) \propto P(D|XI)\, P(X|I)$$

schafft hier Abhilfe: er zerlegt die Wahrscheinlichkeit $P(X|DI)$ in zwei Faktoren: die *Datenverteilung* $P(D|XI)$ und die *A priori-Verteilung* $P(X|I)$. Die Lösung eines Strukturbestimmungsproblems ist durch die Angabe dieser beiden Wahrscheinlichkeiten vollständig bestimmt.

Die Datenverteilung gibt an, welche Meßwerte wir erwarten, wenn sich das Molekül in einer bestimmten Konfiguration X befindet. Sie verbindet die Messungen mit den Koordinaten. In diese Wahrscheinlichkeit geht die Theorie zur Beschreibung der Messungen ein. Experimentelle Fehler und Unzulänglichkeiten der Theorie führen dazu, daß die beobachteten von den idealen Werten abweichen. Die A priori-Verteilung beschreibt, was wir a priori, also vor Durchführung des Experiments, über die Struktur wissen. Das Produkt der Datenverteilung und der A priori-Verteilung ist die gesuchte *A posteriori-Verteilung* $P(X|DI)$.

[1] Dieses Prinzip wurde gemeinsam mit Wolfgang Rieping erarbeitet.

Der Hypothesenraum ist der Konfigurationsraum des Makromoleküls. Jede Grundhypothese „die Struktur des Moleküls ist X“ ist eindeutig mit den Koordinaten X verknüpft. $P(X|DI)$ ist eine Funktion der Koordinaten:

$$p(X) = P(X|DI).$$

Weil die Koordinaten kontinuierlich sind, ist $p(X)$ eine Wahrscheinlichkeitsdichte.

Die Datenverteilung $P(D|XI)$ hängt sowohl von den Messungen als auch von den Koordinaten ab. Bei gegebenen Daten definiert sie eine Funktion auf dem Konfigurationsraum: die *Likelihood-Funktion*

$$L(X) = P(D|XI).$$

Sie ist nicht-negativ, jedoch nicht bezüglich X normiert.

Unsere Vorkenntnisse über eine unbekannte Struktur ordnen jeder möglichen Konfiguration eine Wahrscheinlichkeit $P(X|I)$ zu und definieren die Wahrscheinlichkeitsdichte

$$\pi(X) = P(X|I).$$

Fortan werden die Symbole $p(X)$, $L(X)$ und $\pi(X)$ verwendet, um die Abhängigkeit der Grundwahrscheinlichkeiten von den Koordinaten anzuzeigen. Als Wahrscheinlichkeitsdichten unterliegen p und π dem Transformationsgesetz (2.4). Der Bayes'sche Satz

$$p(X) \propto L(X)\,\pi(X)$$

stellt den Zusammenhang zwischen den drei Funktionen her.

2.2.2 Beschreibung der Daten

Für eine kernspektroskopische Meßgröße y liefert ein Kernresonanzexperiment einen Datensatz $D = \{y_1, \ldots, y_n\}$. Eine Theorie f beschreibt näherungsweise die Observable als Funktion der Koordinaten X: Wenn sich das

Molekül in Konfiguration X befindet, erwarten wir, den Idealwert $\hat{y} = f(X)$ zu beobachten. Aufgrund von Unzulänglichkeiten des Experiments und der Theorie weichen die tatsächlichen Messungen vom Idealwert ab. Wir beschreiben die Abweichungen durch eine Verteilung

$$p(y|\hat{y}, \sigma, I) = g(y; \hat{y}, \sigma).$$

Das Fehlergesetz $g(\cdot; \hat{y}, \sigma)$ ist normiert; es besitzt gewöhnlich ein Maximum in Nähe des Idealwerts $\hat{y}$. Als einfachstes Modell nehmen wir an, daß eine Variable σ ausreicht, um das Fehlergesetz zu parametrisieren.

Die bedingte Wahrscheinlichkeit des Meßwerts bei gegebener Struktur ist damit:

$$p(y|X, \sigma, I) = g(y; f(X), \sigma). \tag{2.5}$$

Die Kenntnisse I, die in die Wahrscheinlichkeit $p(y|X, \sigma, I)$ eingehen, sind die Theorie f und die Form des Fehlergesetzes g. Die Datenverteilung ist Ausdruck unseres Unwissens über Störungen des Experiment, Artefakten der Datenprozessierung sowie Näherungen der Theorie. Sie beschreibt keine tatsächlichen Schwankungen der Observable, sondern in welchem Bereich wir y zu beobachten erwarten, wenn wir wissen, daß sich das Molekül in Konfiguration X befindet.

Die gemeinsame Wahrscheinlichkeit mehrerer Meßwerte y_i setzt sich aus ihren Einzelwahrscheinlichkeiten zusammen. Die vorausgesagten Werte sind $f_i(X)$. Wir nehmen an, daß die Messungen mit demselben Fehlergesetz

$$p(y_i|X, \sigma, I) = g(y_i; f_i(X), \sigma)$$

beschrieben werden können und σ für alle Messungen gleich ist. Die Messungen sind unabhängig voneinander, somit ist die Gesamtwahrscheinlichkeit der Messung von D das Produkt der einzelnen Wahrscheinlichkeiten:

$$p(D|X, \sigma, I) = \prod_{i=1}^{n} p(y_i|X, \sigma, I). \tag{2.6}$$

In I fließt die Annahme unabhängiger Messungen ein.

Im weiteren werden Datenverteilungen der Form

$$g(y_i; f_i(X), \sigma) = \frac{h(y_i)}{Z(\sigma)} \exp\left\{-\frac{1}{2\sigma^2}\,\chi_i^2(X)\right\}$$

betrachtet; $h(y_i)$ ist der rein datenabhängige Anteil und nicht weiter von Interesse; $Z(\sigma)$ normiert die Wahrscheinlichkeit. Die Abweichung χ_i^2 des Meßwerts y_i vom idealen Wert $f_i(X)$ ist positiv und nimmt ihr Minimum $\chi^2 = 0$ an, wenn beide übereinstimmen. χ_i^2 hat keine absolute Bedeutung, sondern nur in Bezug auf eine Fehlerskala $\sigma > 0$. Erst χ_i^2/σ^2 ist eine einheitenlose Größe und σ ein Maßstab für die Abweichung zwischen gemessenem und berechnetem Wert der Observable. In der Wahrscheinlichkeit aller Daten (2.6)

$$p(D|X, \sigma, I) \propto [Z(\sigma)]^{-n} \exp\left\{-\frac{1}{2\sigma^2}\,\chi^2(X)\right\}$$

ist die Gesamtabweichung $\chi^2(X) = \sum_{i=1}^{n} \chi_i^2(X)$ die Summe der einzelnen Abweichungen zwischen den Messungen y_i und ihren Idealwerten $f_i(X)$.

Falls die Theorie parametrisiert ist, also $y = f(X;\alpha)$, hängt die Abweichung der idealen von den gemessenen Werten nicht nur von der Konfiguration, sondern auch den Parametern ab. Als Datenverteilung ergibt sich:

$$p(D|X, \alpha, \sigma, I) \propto [Z(\sigma)]^{-n} \exp\left\{-\frac{1}{2\sigma^2}\,\chi^2(X, \alpha)\right\}. \tag{2.7}$$

Die Parametrisierung α wird explizit aufgeführt, weil sie bekannt sein muß, um die Wahrscheinlichkeit der Messungen angeben zu können.

2.2.3 Erweiterung des Hypothesenraums

Neben den eigentlichen Hypothesenparametern, den Koordinaten X, mussten zwei weitere Klassen von Größen eingeführt werden: die Theorieparameter α und die Fehlerskala σ. Diese Größen sind nicht bekannt und haben den gleichen Stellenwert wie die gesuchte Struktur. Sie müssen *gemeinsam* mit den Koordinaten aus den Daten geschätzt werden. Eine vollständige Beschreibung des Problems zwingt uns, den ursprünglichen Hypothesenraum zu erweitern. Nur die Wahrscheinlichkeit aller Unbekannten

$$p(X, \alpha, \sigma|D, I)$$

löst das Problem vollständig. Man nennt α und σ auch *nuisance parameter* [20], weil sie sekundäre Hypothesenparameter sind, die lediglich zur Beschreibung der Daten eingeführt werden.

Die Unkenntnis der zusätzlichen Parameter stellt einen Mangel an Wissen dar, der in der Wahrscheinlichkeitstheorie direkt berücksichtigt werden kann: Im erweiterten Hypothesenraum sind alle Unbekannten durch ihre gemeinsame A posteriori-Verteilung bestimmt. Um den Bayes'schen Satz anzuwenden, müssen wir nur die A priori-Verteilung erweitern und erhalten

$$p(X, \alpha, \sigma) \propto L(X, \alpha, \sigma)\, \pi(X, \alpha, \sigma).$$

Die Likelihood-Funktion

$$L(X, \alpha, \sigma) \propto [Z(\sigma)]^{-n} \exp\left\{-\frac{1}{2\sigma^2}\ \chi^2(X, \alpha)\right\}$$

hängt bereits von allen Hypothesenparametern ab.

Durch Integration über die nuisance parameter wird der erweiterte Hypothesenraum wieder auf den molekularen Konfigurationsraum reduziert:

$$p(X) = \int \mathrm{d}\alpha\, \mathrm{d}\sigma\ p(X, \alpha, \sigma). \tag{2.8}$$

Eine solche *Marginalisierung* [20] der zusätzlichen Parameter unterscheidet sich grundsätzlich davon, sie vorab festzulegen. Denn die Parameter werden nicht nur durch das Hintergrundwissen, sondern auch durch die Daten bestimmt. Die Marginalisierungsregel ist eine Verallgemeinerung der Summenregel.

Die Marginalisierung erlaubt, aus Datenverteilungen $p(D|X, \alpha, \sigma, I)$, deren Formulierung die Einführung neuer Hypothesenparameter α und σ erfordert, neue Datenverteilungen $p(D|X, I)$ zu konstruieren, in denen die zusätzlichen Parameter ausintegriert sind:

$$p(D|X, I) = \int \mathrm{d}\alpha\, \mathrm{d}\sigma\ p(D|X, \alpha, \sigma, I)\, p(\alpha, \sigma|X, I).$$

Die reduzierte Datenverteilung $p(D|X, I)$ ist eine Verbundverteilung aller Messungen und läßt sich nicht mehr in ein Produkt von Verteilungen der einzelnen Messungen zerlegen. Die Einführung und Marginalisierung der zusätzlichen Parameter hat die Messungen untereinander gekoppelt. Wir können

$L(X) = P(D|X, I)$ direkt im Bayes'schen Satz verwenden und erhalten eine A posteriori-Verteilung, die nur noch von den Koordinaten abhängt.

2.2.4 Mehrere Datensätze

In der Kernresonanzspektroskopie werden häufig Datensätze für verschiedene Observablen aufgenommen. Man hat beispielsweise eine Liste von zugeordneten NOESY-Signalen und zusätzlich Messungen der skalaren Kopplungskonstanten. Wie sollen verschiedene Datensätze $D_1, \ldots, D_m$ aus m Experimenten gemeinsam zur Strukturbestimmung genutzt werden?

Zur Beantwortung dieser Frage brauchen wir nur die Ableitung der Wahrscheinlichkeit eines Datensatzes zu wiederholen. Jede Observable y_j beschreibt eine Theorie, die ihre eigenen Parameter α_j benötigt. Abweichungen zwischen Theorie und Experiment werden durch eine eigene Fehlerskala σ_j gemessen. Jeder Datensatz $D_j = \{y_{j1}, \ldots, y_{jn_j}\}$ hat eine Wahrscheinlichkeit

$$p(D_j|X, \alpha_j, \sigma_j, I) \propto [Z(\sigma_j)]^{-n_j} \exp\left\{-\frac{1}{2\sigma_j^2}\,\chi_j^2(X, \alpha_j)\right\};$$

die Normierungskonstanten $Z(\sigma_j)$ und die Abweichungen χ_j^2 unterscheiden sich von Datensatz zu Datensatz, weil jeder durch ein eigenes Fehlergesetz und eine eigene Theorie beschrieben wird.

Die Messung der verschiedenen Datensätze ist voneinander unabhängig, so daß analog zu (2.6) die Gesamtwahrscheinlichkeit aller Messungen das Produkt der einzelnen Datenverteilungen ist:

$$p(D|X, \alpha, \sigma, I) = \prod_{j=1}^{m} p(D_j|X, \alpha_j, \sigma_j, I);$$

die Theorieparameter α_j und die Fehler σ_j werden in $\alpha = \{\alpha_1, \ldots, \alpha_m\}$ bzw. $\sigma = \{\sigma_1, \ldots, \sigma_m\}$ zusammengefaßt.

Die gesamte Likelihood-Funktion ist ein Produkt der einzelnen Likelihood-Funktionen, die sich aus den jeweiligen Datensätzen ergeben:

$$L(X, \alpha, \sigma) \propto \left(\prod_{j=1}^{m}[Z(\sigma_j)]^{-n_j}\right) \exp\left\{-\frac{1}{2}\sum_j \chi_j^2(X, \alpha_j) \,/\, \sigma_j^2\right\}. \qquad (2.9)$$

Die Gesamtabweichung zwischen Theorie und Experiment wird durch eine gewichtete Summe der Abweichungen jedes Datensatzes gemessen:

$$\chi^2(X, \alpha, \sigma) = \sum_{i=1}^{m} w_j \, \chi_j^2(X, \alpha_j) \tag{2.10}$$

mit dem Gesamtfehler $\sigma^{-2} = \sum_{i=1}^{m} \sigma_j^{-2}$ und den Gewichten $w_j = \sigma^2/\sigma_j^2$, $\sum_j w_j = 1$. Es besteht keine Unklarheit darüber, wie mit mehreren Datensätzen, welche nie völlig konsistent sein werden, umzugehen ist: Jeder Datensatz wird durch seinen Fehler σ_j in der gemeinsamen Likelihood-Funktion gewichtet; die Gewichtung paßt sich während der Schätzung der Koordinaten und der übrigen Hypothesenparameter an.

Um den Bayes'schen Satz anwenden zu können, muß die A priori-Verteilung aller Unbekannten

$$\pi(X, \alpha, \sigma) = \prod_{i=1}^{m} \pi(X, \alpha_j, \sigma_j) = \Big(\prod_{i=1}^{m} \pi(\alpha_j, \sigma_j | X) \Big) \; \pi(X) \tag{2.11}$$

zugewiesen werden. Der konformationelle Beitrag wurde abgesondert, weil er eine vom Experiment unabhängige Grundwahrscheinlichkeit der Konfigurationen des Moleküls ist.

2.2.5 A priori-Verteilung der Struktur

Kernspektroskopische Messungen an biologischen Makromolekülen werden in der Regel in wässriger Lösung durchgeführt. Wechselwirkungen mit dem Wasser bewirken, daß sich das Molekül in eine stabile globuläre Struktur faltet. Die Beschreibung der Freiheitsgrade des Lösungsmittels ist sehr aufwendig. Deshalb werden wir lediglich die konformationellen Freiheitsgrade des Moleküls betrachten.

Dieser Idealisierung nach befindet sich das Molekül im Vakuum und hat eine potentielle Energie $E(X)$. Das Experiment wird bei Temperatur β^{-1} durchgeführt; meist bei Raumtemperatur. Diese Annahmen können durch das Maximum-Entropie-Prinzip [23] in eine A priori-Verteilung umgewandelt

werden. Man erhält das Boltzmann-Ensemble

$$\pi(X) = \frac{1}{Z(\beta)} \exp\{-\beta E(X)\} \tag{2.12}$$

als A priori-Verteilung der Koordinaten; die Zustandssumme $Z(\beta)$ ist eine Normierungskonstante.

In dieser Vereinfachung bleibt die Konformation unbestimmt. Erst die Analyse von Daten kann die Struktur einigermaßen genau und richtig charakterisieren. Eine vollständigere Beschreibung der Vorkenntnisse würde jedoch erlauben, richtigere Aussagen über die Struktur zu machen.

Biologische Makromoleküle sind lineare Polymere, also Kettenmoleküle, die sich aus kovalent gebundenen Grundeinheiten zusammensetzen. Bei Proteinen sind dies die zwanzig Aminosäuren, im Fall der Nukleinsäuren DNA oder RNA die vier Nukleotide. Wegen der kovalenten Kräfte ist der Konfigurationsraum eines Kettenmoleküls auf eine niederdimensionale Mannigfaltigkeit des Raums der kartesischen Koordinaten beschränkt. Die hierarchische Natur der physikalischen Wechselwirkungen berücksichtigt eine Parameterisierung durch externe und interne Koordinaten (siehe Anhang B).

Die A priori-Verteilung in kartesischen Koordinaten ist das Boltzmann-Ensemble (2.12). Nach (2.4) muß der Übergang zu externen und internen Koordinaten durch die Jacobi-Determinante der Transformation berücksichtigt werden. Weil die potentielle Energie von den Atomabständen abhängt, gehen in die A priori-Verteilung nur noch die internen Koordinaten ein:

$$\pi(\{l_i\}, \{\kappa_i\}, \{\theta_i\}) = \frac{1}{Z(\beta)} \exp\big\{-\beta E(X(l, \kappa, \theta))\big\}\, J(l, \kappa, \theta). \tag{2.13}$$

Dies sind die Bindungslängen $l = \{l_i\}$, die Bindungswinkel $\kappa = \{\kappa_i\}$ und die Dihedralwinkel $\theta = \{\theta_i\}$ mit der Jacobi-Determinante $J(l, \kappa, \theta) = \prod_i \; l_i^2 \sin \kappa_i$ (siehe Abschnitt B.3 und Gleichung (B.7)).

Um die Dimensionalität des Konformationsraums zu verkleinern, werden die Bindungslängen l_i und -winkel κ_i fixiert. Außerdem sind die Planaritäten der Peptidebene sowie der Ringe in den Seitenketten der Aminosäuren starr. Als einzige Freiheitsgrade bleiben die Dihedralwinkel θ_i; sie beschreiben Rotationen um die Bindungen.

In der potentiellen Energie müssen nur noch nicht-kovalente Wechselwirkungen berücksichtigt werden. Dies sind paarweise Beiträge E_{kl}, die von den Atomabständen r_{kl} abhängen:

$$E(\theta) = \sum_{k<l} E_{kl}(r_{kl}(\theta)). \tag{2.14}$$

Die Berechnung elektrostatischer Wechselwirkungen ist nur sinnvoll, wenn auch das Lösungsmittel beschrieben wird, deshalb werden hier nur van der Waals-Abstoßungen zwischen den Atomen berücksichtigt. Eine vereinfachte Version des Lennard-Jones-Potentials [24]

$$E_{kl}(r_{kl}) = \frac{k_{kl}}{2}\,\Theta\big(\hat{r}_{kl} - r_{kl}\big)\,\big(\hat{r}_{kl} - r_{kl}\big)^4 \tag{2.15}$$

nähert die van der Waals-Wechselwirkung an. $\Theta(\cdot)$ ist die Heaviside-Funktion: Abstände, die größer als die Summe ihrer van der Waals-Radien sind, $r_{kl} > \hat{r}_{kl}$, tragen nicht bei. Nähert sich r_{kl} Null, bleiben die Energien endlich; es besteht eine endliche Wahrscheinlichkeit, daß sich die Polypeptidkette durchdringt. Die Gleichgewichtsabstände $\hat{r}_{kl}$ und die Kraftkonstanten k_{kl} werden dem PROLSQ-Kraftfeld [24] entnommen.

2.2.6 A priori-Verteilung der Fehler

Die Theorieparameter α_j und die Fehler σ_j sind zusätzliche Hypothesenparameter. Um sie zusammen mit den Dihedralwinkeln zu schätzen, müssen ihnen A priori-Verteilungen zugewiesen werden. Die gemeinsame A priori-Verteilung zerfällt in

$$\pi(\theta, \alpha, \sigma) = \pi(\alpha, \sigma|\theta)\,\pi(\theta).$$

Wir nehmen an, daß die Fehlerskalen unabhängig von den Dihedralwinkeln, den Theorieparametern und untereinander sind:

$$\pi(\alpha, \sigma|\theta) = \pi(\alpha|\theta)\,\pi(\sigma) \quad \text{mit} \quad \pi(\sigma) = \prod_j \pi(\sigma_j).$$

Aus der Datenverteilung folgt, daß dem Fehler keine absolute Bedeutung zukommt: Eine Skalierung von σ_j kann durch eine Skalierung der Abweichung χ_j^2 aufgehoben werden. Die A priori-Verteilung muß deshalb gegenüber Skalierungen $\sigma \to s\sigma$ invariant sein, d.h. nach (2.4):

$$\pi(s\sigma) = \pi(\sigma)\, s^{-1}.$$

Ableitung nach s liefert an $s = 1$ die Differentialgleichung $\sigma\pi'(\sigma) = -\pi(\sigma)$, welche direkt integriert werden kann:

$$\pi(\sigma) \propto \sigma^{-1}. \tag{2.16}$$

Allein aus der Information I = „σ ist ein Skalenparameter“ ergibt sich Jeffreys' prior (2.16) als A priori-Verteilung [25]. Weil sich Jeffreys' prior nicht normieren läßt, ist er nur als Grenzwert einer Folge von Verteilungen

$$\pi(\sigma|\sigma_{\max}, \sigma_{\min}) = \frac{1}{\log(\sigma_{\max}/\sigma_{\min})}\,\frac{1}{\sigma}, \quad \sigma_{\min} < \sigma < \sigma_{\max}$$

sauber definiert. In der Parameterschätzung kann bereits bei wenigen Messungen von vornherein $\sigma_{\min} = 0$ und $\sigma_{\max} = \infty$ gesetzt werden.

Durch die Definition der A priori-Verteilung $\pi(\sigma)$ haben wir den Hypothesenraum erweitert und sind nun in der Lage, anhand der Daten nicht nur Aussagen über die Konformationen, sondern auch über die Fehler zu machen. Um das Problem zu vervollständigen, müssen noch die A priori-Verteilungen der Theorieparameter aufgestellt werden.

2.2.7 Vergleich mit Optimierung

Üblicherweise werden makromolekulare Strukturen durch Minimierung einer Hybridenergie bestimmt. Der negative Logarithmus der bedingten A posteriori-Verteilung der Koordinaten

$$-\log p(X|\alpha, \sigma) = \frac{1}{2}\sum_j \chi_j^2(X, \alpha)/\sigma_j^2 \;+\; \beta\, E(X)$$

ist analog zur Zielfunktion (1.5):

$$G(X) = E(X) + \sum_j \lambda_j F_j(X).$$

Es ergeben sich die Entsprechungen:

$$\begin{aligned} G(X) &\longleftrightarrow -\log p(X|\alpha,\sigma) \\ E(X) &\longleftrightarrow -\log \pi(X) \\ F_j(X) &\longleftrightarrow \chi_j^2(X,\alpha)/2 \\ \lambda_j &\longleftrightarrow \sigma_j^{-2}. \end{aligned}$$

Die Elemente der Hybridenergie erhalten wahrscheinlichkeitstheoretische Deutungen: Der Datenterm $F_j(X)$ ist der negative Logarithmus der Likelihood-Funktion; seine Gewichtung muß, wegen $\lambda_j \sim \sigma_j^{-2}$, gemäß der Güte der Daten und des Modells erfolgen.

Umgekehrt kann die Zielfunktion $G(X)$ als negativer Logarithmus einer bedingten A posteriori-Verteilung $p(X|\alpha,\sigma)$ interpretiert werden. Optimierung geht also von bekannten Parametern α und σ aus und kommt einer Punktschätzung der Koordinaten gleich. Weil die zusätzlichen Parameter aber nicht bekannt sind, müssen sie mit Heuristiken bestimmt werden. Beispiele sind die Kalibrationsschemata für NOESY-Volumina [13, 14] oder die Kreuzvalidierung [15, 16] zur Bestimmung der Gewichte λ_j.

Im Gegensatz zu Optimierungsverfahren ist die Wahrscheinlichkeitstheorie vollständig. Eine Zielfunktion zur Bestimmung aller Unbekannten ist der negative Logarithmus der gemeinsamen A posteriori-Verteilung:

$$\sum_j \Big[\frac{1}{2\sigma_j^2} \chi_j^2(X,\alpha_j) + n_j \log Z(\sigma_j) \Big] - \log \pi(\alpha,\sigma|X) - \beta E(X).$$

Man könnte zur Bestimmung aller Hypothesenparameter diese Funktion minimieren und bedürfte keiner zusätzlichen Heuristiken. Dies ließe jedoch den Großteil der Information ungenutzt, die in der A posteriori-Verteilung enthalten ist (siehe Abschnitt 2.3).

Normierung der Likelihood-Funktion bezüglich der Daten resuliert in dem σ_j-abhängigen Term $n_j \log Z(\sigma_j)$. Der A priori-Term $\log \pi(\alpha_j,\sigma_j|X)$ fehlt

gänzlich in $G(X)$. Ohne die Wahrscheinlichkeitstheorie wären wir nicht in der Lage, diese zusätzlichen Terme zu motivieren und so die Zielfunktion zu vervollständigen.

2.3 Monte-Carlo-Integration

Kernspektroskopische Messungen an Biopolymeren sind zu komplex, als daß die A posteriori-Verteilungen noch analytisch untersucht werden könnten. Hier schaffen Algorithmen Abhilfe, die sich in der Auswertung der Verteilung auf Bereiche hoher Wahrscheinlichkeit beschränken. Diese Bereiche werden in einer zufälligen Suche aufgespürt.

2.3.1 Markov-Ketten-Monte-Carlo

Anliegen einer Datenanalyse ist, Hypothesen zu formulieren und anhand der Messungen zu bewerten. Unser Unwissen erfordert, alle, nicht bloß die wahrscheinlichsten Werte der Hypothesenparameter in Betracht zu ziehen; also über die möglichen Werte, gewichtet mit ihrer Wahrscheinlichkeit, zu mitteln. Gleich ob es sich um Erwartungswerte, Varianzen, marginale Verteilungen oder Vorhersagen handelt, gilt es stets, hochdimensionale Integrale über die A posteriori-Verteilung auszuwerten. In Anwendungen ist man deshalb mit dem Problem konfrontiert, Integrale der Form

$$I = \int \mathrm{d}x \;\; f(x)\, p(x)$$

möglichst genau zu berechnen, wobei p die Verteilung des Hypothesenparameters x ist und $f(x)$ eine beliebige Funktion sein kann (beispielsweise $f(x) = x$ für Erwartungswerte oder $f(x) = \delta(y - x_i)$ für die marginale Verteilung von x_i). Ein solches Integral wird durch eine endliche Summe angenähert:

$$I \approx \sum_{k=1}^{m} \Delta x^{(k)}\, f\big(x^{(k)}\big)\, p\big(x^{(k)}\big).$$

Die Positionierung der Stützstellen $x^{(k)}$ bestimmt die Güte der Näherung. Die einfachste Wahl wären äquidistante Stützstellen. Doch würde bei zunehmender Zahl von Hypothesenparametern der Rechenaufwand exponentiell wachsen. Desweiteren muß bereits bei der Diskretisierung eines eindimensionalen Problems abgeschätzt werden, welche Bereiche der reellen Achse signifikant zum Integral beitragen.

Bei Wahrscheinlichkeitsintegralen ist aber immer bekannt, wie wichtig der Beitrag einer Stützstelle $x^{(k)}$ ist: ihr (stets positives) Gewicht ist gerade $p\left(x^{(k)}\right)$. Wir müssen also die Stützstellen entsprechend ihrer Wahrscheinlichkeit wählen, d.h. dort dichter, wo die Verteilung Moden aufweist. Die Abstände zwischen den Stützstellen sind dann umgekehrt proportional zur Wahrscheinlichkeitsdichte: $\Delta x^{(k)} \propto 1/p\left(x^{(k)}\right)$. Die Integralsumme wird zu einem Mittel

$$I \approx \frac{1}{m} \sum_{k=1}^{m} f\left(x^{(k)}\right); \tag{2.17}$$

und die Stützstellen $x^{(k)}$ sind *Stichproben* von p: sie sind gemäß der Dichte p verteilt, symbolisiert durch $x^{(k)} \sim p(x)$.

Anders ausgedrückt, wird die Verteilung durch ein unendlich feines Histogramm angenähert:

$$p(x) \approx p_m(x) \equiv \frac{1}{m} \sum_{k=1}^{m} \delta\left(x - x^{(k)}\right).$$

Wie gut diese Näherung ist, hängt von der Anzahl der Stichproben und ihrer Position ab. Die Stichproben müssen so gezogen werden, daß bei unendlich vielen Ziehungen die Dichte reproduziert wird:

$$\lim_{m \to \infty} p_m(x) = p(x).$$

Bei endlichem Aufwand, also fester Anzahl m von Stichproben, sind in erster Linie die Bereiche aufzusuchen, die am meisten Wahrscheinlichkeitsmasse tragen.

Die Näherung des Integrals durch ein Mittel von Stichproben hat den Vorteil, daß ihre Güte weniger von der Dimensionalität des Hypothesen-

raums abhängt, sondern vielmehr von der Topologie der Verteilung. Außerdem können dieselben Stichproben zur Bewertung aller erdenklichen Hypothesenfunktionen f verwendet werden, wobei natürlich auch die Güte der Näherung davon abhängt, wie sehr wichtige Bereiche von p und f sich überdecken. Bei Verwendung klassischer numerischer Verfahren würde sich hingegen mit jeder neuen Hypothese der Integrand pf ändern und eine neuerliche Berechnung erforderlich machen.

Monte-Carlo-Verfahren erzeugen Stichproben durch einen stochastischen Prozeß. Ausgehend von der k-ten Stichprobe $x^{(k)}$ wird die nächste nach einer Zufallsvorschrift gezogen:

$$x^{(k+1)} \sim T\big(x|x^{(k)}\big). \tag{2.18}$$

Solch ein zufälliges Absuchen der Wahrscheinlichkeitsverteilung wird auch sampling genannt. Wiederholte Anwendung erzeugt eine „Kette“ von Stichproben:

$$x^{(1)} \to x^{(2)} \to \ldots \to x^{(m)}.$$

Der Index k bezeichnet nun einen Zeitpunkt in der Simulation und nicht mehr einen Ort, wie zuvor in der Diskretisierung des Integrals. Weil die Übergangswahrscheinlichkeit $T(x|y)$ nur vom Vorgänger in der Simulation abhängt, definiert sie einen Markov-Prozeß erster Ordnung. Wird T zur Erzeugung von Stichproben verwendet, spricht man von Markov-Ketten-Monte-Carlo (MCMC) [26].

Die Übergangswahrscheinlichkeit muß so gewählt werden, daß die Zielverteilung p die stationäre Verteilung der Markov-Kette ist, d.h.

$$p(x) = \int \mathrm{d}x'\; T\left(x|x'\right)\; p(x'),$$

denn dann werden nach einer gewissen Konvergenzphase über die Vorschrift (2.18) nur noch Stichproben von der gewünschten Verteilung gezogen. Ein Spezialfall dieser Forderung ist die Bedingung detaillierten Gleichgewichts (detailed balance):

$$T(x|x')\, p(x') = T(x'|x)\, p(x). \tag{2.19}$$

Integration über x' ergibt p als stationäre Verteilung der Übergangsvorschrift T. Bedingung (2.19) besagt, daß sich die Markov-Kette nach der Konvergenzphase in einem dynamischen Gleichgewicht befindet.

2.3.2 Gibbs sampling

Gibbs sampling [27] ist ein Verfahren, um eine Markov-Kette für mehrdimensionale Verteilungen zu konstruieren. Ist $x = (x_1, \ldots, x_n)^T$ ein n-dimensionaler Hypothesenparameter mit Verteilung $p(x) = p(x_1, \ldots, x_n)$, so erzeugt der Prozeß

$$\begin{aligned}
x_1^{(k+1)} &\sim p\big(x_1 | x_2^{(k)}, \ldots, x_n^{(k)}\big) \\
&\vdots \\
x_i^{(k+1)} &\sim p\big(x_i | x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_{i+1}^{(k)}, \ldots, x_n^{(k)}\big) \\
&\vdots \\
x_n^{(k+1)} &\sim p\big(x_n | x_1^{(k+1)}, \ldots, x_{n-1}^{(k+1)}\big)
\end{aligned}$$

eine Markov-Kette, deren stationäre Verteilung gerade $p(x)$ ist. Die Verteilungen $p(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ sind bedingte Verteilungen von $p(x)$.

Wir können Gibbs sampling verwenden, um die Ziehung von Stichproben der Dihedralwinkel θ, der Theorieparameter α und der Fehler σ zu entkoppeln. Es ergibt sich das Schema

$$\begin{aligned}
\theta^{(k+1)} &\sim p\big(\theta | \alpha^{(k)}, \sigma^{(k)}\big), \\
\alpha^{(k+1)} &\sim p\big(\alpha \,| \theta^{(k+1)}, \sigma^{(k)}\big), \\
\sigma^{(k+1)} &\sim p\big(\sigma \,| \theta^{(k+1)}, \alpha^{(k+1)}\big).
\end{aligned} \tag{2.20}$$

Als bedingte A posteriori-Verteilungen der nuisance parameter α und σ ergeben sich häufig bekannte Verteilungen, für die Zufallszahlengeneratoren existieren. Man kann diese Verfahren verwenden, um Stichproben dieser Parameter zu erzeugen. Die Erzeugung konformationeller Stichproben ist dagegen schwierig.

2.3.3 Hybrid-Monte-Carlo

Die bedingte A posteriori-Verteilung der Dihedralwinkel läßt sich auf keine der gängigen Verteilungen zurückführen. Sowohl die Likelihood-Funktion als auch das Boltzmann-Ensemble prägen den Dihedralwinkeln starke Korrelationen auf. Die Änderung eines Dihedralwinkels kann die Konformation stark ändern und damit sowohl den Wert der Likelihood-Funktion als auch der A priori-Verteilung.

Um Konformationen effizient zu ziehen, muß die Korrelation der Dihedralwinkel in der Erzeugung der Stichprobe berücksichtigt werden. Im Hybrid-Monte-Carlo (HMC) [28] wird die neue Konfiguration durch eine kurze Integration der Bewegungsgleichungen bestimmt, die sich aus dem negativen Logarithmus der bedingten konformationellen A posteriori-Verteilung

$$l(\theta) = -\log p(\theta|\alpha, \sigma)$$

ergeben. Dazu werden zu den Dihedralwinkeln konjugierte Impulse μ_i eingeführt und die erweiterte Verteilung

$$p(\theta, \mu) \propto \exp\left\{-\frac{1}{2}\sum_i \mu_i^2 - l(\theta)\right\}$$

simuliert; Stichproben der Konformationen erhält man wegen

$$p(\theta) \propto \int \mathrm{d}\mu \; p(\theta, \mu)$$

durch numerische Integration über die Impulse: man übernimmt bloß die Dihedralwinkel und vernachlässigt ihre Impluse. Der Ablauf ist:

1. Begonnen wird mit einem Gibbs sampling-Schritt: die Impulse werden von einer Gauß-Verteilung gezogen

   $$\mu_i^{(k)} \sim \exp\Big\{-\mu_i^2/2\Big\}$$

   (damit gehorchen die Geschwindigkeiten einer standardisierten Maxwell-Verteilung).

2. Kandidaten der Dihedralwinkel und der Impulse werden durch eine Integration der Bewegungsgleichungen berechnet, die sich aus der Pseudo-Hamilton-Funktion $H(\theta,\mu) = \frac{1}{2}\sum_i \mu_i^2 + l(\theta)$ ergeben. Ausgehend von den Anfangsbedingungen $\left(\theta^{(k)}, \mu^{(k)}\right)$ werden die Bewegungsgleichungen:

$$\frac{d\theta_i}{dt} = -\mu_i, \quad \frac{d\mu_i}{dt} = \partial_i l(\theta)$$

mit einem Differenzenschema integriert. Nach einer gewissen Zahl von Integrationsschritten erhält man Impulse μ' und Dihedralwinkel θ'.

3. Das Metropolis-Kriterium [29]

$$\min\left\{1\,,\,\exp\left(-\left[H(\theta',\mu') - H(\theta^{(k)},\mu^{(k)})\right]\right)\right\}.$$

entscheidet, ob (θ',μ') als neue Stichprobe $\left(\theta^{(k+1)},\mu^{(k+1)}\right)$ akzeptiert wird.

Damit die mittels HMC erzeugte Markov-Kette detailed balance (2.19) erfüllt, muß das Integrationsschema in Schritt 2 zeitreversibel und volumenerhaltend sein [26]. Der leapfrog-Algorithmus [30] genügt diesen Anforderungen. Idealerweise ändert sich H während der Integration nicht und die Markov-Kette bewegt sich auf Linien konstanter Wahrscheinlichkeit $p(\theta,\mu)$. Wegen numerischer Fehler weicht H jedoch am Ende der Trajektorie mehr oder weniger vom anfänglichen Wert ab. Die Größe des Integrationsfehlers hängt von der Länge der Trajektorie, der Feinheit der Diskretisierung und der Form der Hamilton-Funktion ab. Das Metropolis-Kriterium garantiert detailed balance selbst bei numerischen Fehlern; Abweichungen in H werden exponentiell unterdrückt.

Die Zufallsvariablen können beliebiger Natur sein. Aber selbst wenn keine Daten vorhanden sind und $l(\theta) = \beta\, E(\theta)$ gilt, stimmen die Bewegungsgleichungen nicht mit den echten Hamilton-Gleichungen überein: Beim Wechsel von kartesischen Koordinaten zu Dihedralwinkeln erhält man in den echten Hamilton-Gleichungen eine winkelabhängige kinetische Energie [31]. Beim Hybrid-Monte-Carlo in Dihedralwinkeln ist hingegen die kinetische Energie

winkelunabhängig. Bei der Ziehung von Stichproben muß jedoch die Trajektorie nicht realistisch sein.

2.3.4 Replica-Monte-Carlo

Gibbs sampling wird für hochdimensionale, multimodale Verteilungen mit stark korrelierten Variablen sehr ineffizient. Zur Ziehung einer neuen Stichprobe werden bei n Hypothesenvariablen n Schnitte durch die Verteilung gemacht. Jeder Schnitt zeigt aber nur einen begrenzten Ausschnitt der Verteilung; er zeigt nicht, wie die Variablen gekoppelt sind, und ob für einen benachbarten Schnitt weitere Maxima existieren.

Auch Hybrid-Monte-Carlo durchsucht den Hypothesenraum nicht mehr ergodisch, wenn die Maxima der Verteilung sehr schroff sind. Dann reicht der Impuls nicht aus, um ein Maximum innerhalb der Simulationszeit zu verlassen. Die Nicht-Ergodizität kann aber umgangen werden: Bei höheren Temperaturen ist das Boltzmann-Ensemble weniger schroff und HMC kann es vollständig absuchen; im Grenzfall unendlich hoher Temperatur gehorchen die Dihedralwinkel einer Gleichverteilung, von welcher leicht Stichproben gezogen werden können.

Diese Idee wird verallgemeinert: Eine stetige Transformation bildet die Grundverteilung $p(x)$ auf eine Verteilung $p'(x)$ mit günstigeren Eigenschaften ab. Wenn p' unimodal ist oder nur wenig ausgeprägte Maxima aufweist, können von ihr ergodisch Stichproben gezogen werden. Eine Schar $p(x|\omega)$ von Verteilungen interpoliert zwischen den beiden Verteilungen: $p(x|\omega_{\max}) = p'(x)$, $p(x|\omega_{\min}) = p(x)$. Der Parameter ω entspricht der Temperatur des Boltzmann-Ensembles; Änderung von ω „heizt“ oder „kühlt“ die Verteilung p. Eine mögliche Verteilungsschar ist

$$p(x|\omega) \propto [\, p(x) \,]^{\omega} . \tag{2.21}$$

Die „Hochtemperaturverteilung“ $p(x|\omega = 0)$ ist in diesem Fall eine Gleichverteilung.

Die Schar überträgt Stichproben der Hochtemperaturverteilung p' auf die Zielverteilung p. Für m Stützstellen $\omega = \{\omega_1, \ldots, \omega_m\}$ mit $\omega_1 = \omega_{\min}$ und $\omega_m = \omega_{\max}$ erhält man m „Wärmebäder", die nach und nach in die Zielverteilung übergehen. Die Wärmebäder sind nicht-wechselwirkende Vervielfachungen (replicas) des ursprünglichen Systems. Replica-Monte-Carlo [32] simuliert alle Wärmebäder gleichzeitig:

$$p(x_1, \ldots, x_m|\omega_1, \ldots, \omega_m) = \prod_{i=1}^{m} p(x_i|\omega_i).$$

Benachbarte Wärmebäder tauschen Stichproben aus und ermöglichen so deren Diffusion in die Zielverteilung. Der Austausch wird mit der Wahrscheinlichkeit

$$\min\left\{1,\ \frac{p(x_j|\omega_i)\, p(x_i|\omega_j)}{p(x_i|\omega_i)\, p(x_j|\omega_j)}\right\}$$

akzeptiert oder verworfen.

Die Wahl der Verteilungsschar ist freigestellt. Generell ist es günstig, eine Familie von Verteilungen zu wählen, in denen sich Bereiche hoher Wahrscheinlichkeit überdecken. Das ist für die Familie (2.21) der Fall. Ein weiteres Kritierum ist, die Parametrisierung so zu wählen, daß möglichst wenige Verteilungen nötig sind, um zwischen dem Hochtemperaturbad und der Zielverteilung zu interpolieren

In der Bayes'schen Analyse von Kernresonanzdaten muß das Produkt der Likelihood-Funktion und der A priori-Verteilung simuliert werden

$$p(\theta, \alpha, \sigma) \propto L(\theta, \alpha, \sigma)\, \pi(\theta, \alpha, \sigma).$$

Beide Verteilungen sind in den konformationellen Freiheitsgraden hochkorreliert. Um jeden Faktor einzeln kontrollieren zu können, wird die Idee des Replica-Algorithmus erweitert. Die Parameter ω und q definieren die Verteilungsschar [2]

$$p(\theta, \alpha, \sigma|\omega, q) \propto [L(\theta, \alpha, \sigma)]^{\omega}\ \pi(\theta|q)\ \pi(\alpha, \sigma).$$

Die konformationelle A priori-Verteilung ist hier das Tsallis-Ensemble [33, 34]

$$\pi(\theta|q) \propto [\,1 + \beta\,(q-1)\,(E(\theta) - E_{\min})\,]^{\frac{q}{1-q}}\,. \qquad (2.22)$$

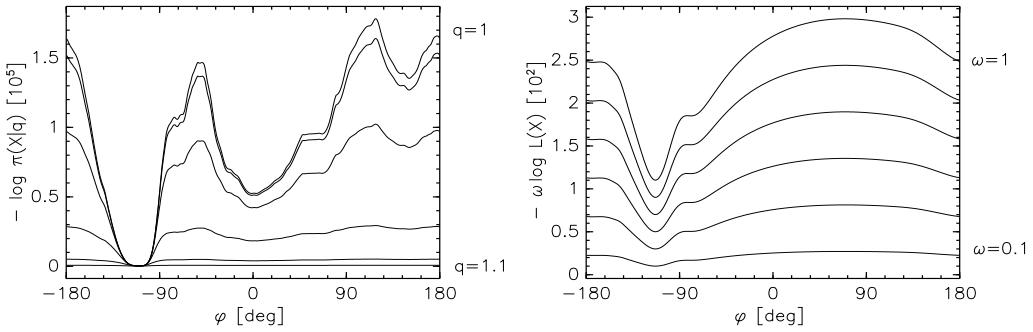


Abbildung 2.1: Einfluß der Replica-Parameter ω und q: Für eine gefaltete Proteinstruktur wurde ein Schnitt durch die Fläche der negativen Logarithmen von $[L(\theta,\alpha,\sigma)]^{\omega}$ und $\pi(\theta|q)$ bei Veränderung eines Dihedralwinkels berechnet. Für wachsende q wird E_q immer flacher (links), für kleinere ω nähert sich L^{ω} einer Gleichverteilung (rechts).

Die Verwendung des Tsallis-Ensembles kann als nicht-lineare Transformation der Energie interpretiert werden: es gilt $\pi(\theta|q) \propto \exp\{-\beta\, E_q(\theta)\}$ mit

$$E_q(\theta) = \frac{q}{\beta\,(q-1)} \log\left[1 + \beta\,(q-1)\,(E(\theta) - E_{\min})\right].$$

Wegen $\log(1+x) = x + \mathcal{O}(x^2)$, ist $E_{q=1}(\theta) = E(\theta)$, also bei $q = 1$ das Tsallis- gleich dem Boltzmann-Ensemble. Für $q \to \infty$ dagegen wird die Verteilung immer flacher. Im Falle $q > 1$ fällt $\pi(\theta|q)$ nur potentiell statt exponentiell ab. Somit haben hochenergetische Konformationen, die das Boltzmann-Ensemble exponentiell unterdrückt, im Tsallis-Ensemble eine höhere Wahrscheinlichkeit. Bei Simulation von $\pi(\theta|q)$ mittels HMC werden die Bewegungsgleichungen integriert, die aus der transformierten Energie E_q resultieren. Weil E_q mit wachsendem q immer flacher wird (siehe Abb. 2.1), kann HMC den Konformationsraum schließlich ergodisch absuchen.

Im Faktor L^{ω} kontrolliert ω den Einfluß der Daten; diese gehen nur mit der effektiven Anzahl $n\omega$ in die Likelihood-Funktion ein:

$$[L(\theta,\alpha,\sigma)]^{\omega} \propto [Z(\sigma)]^{n\omega} \exp\Big\{-\frac{\omega}{2\sigma^2}\,\chi^2(\theta,\alpha)\Big\}.$$

Beim Gibbs sampling muß in den bedingten A posteriori-Verteilungen n durch $n\omega$ ersetzt werden.

Die replicas sind so angeordnet, daß zuerst die Daten durch Verminderung von ω ausgeschaltet werden und anschließend durch Erhöhung von q die nicht-kovalenten Kräfte zwischen den Atomen. Vom „kältesten“ Wärmebad ($\omega = 1, q = 1$) werden Stichproben der A posteriori-Verteilung $p(\theta, \alpha, \sigma)$ gezogen. Das Wärmebad um Übergang $\omega = 0, q = 1$ ist das Boltzmann-Ensemble. Die Hochtemperaturverteilung ($\omega = 0, q \to \infty$) ist eine Gleichverteilung der Dihedralwinkel.

2.4 Datensätze

2.4.1 Daten für eine perdeuterierte SH3-Domäne

Die SH3-Domäne ist eine kleine Proteindomäne, die in vielen Signalübertragungsproteinen vorkommt. Von Mal et al. [35] wurden NOESY-Spektren einer perdeuterierten SH3-Domäne des Proteins Fyn aufgenommen. Alle Wasserstoffe außer den Amidprotonen, die an die Stickstoffatome des Proteinrückgrats gebunden sind, wurden durch Deuteriumatome ersetzt. Dies gestattet, NOESY-Resonanzen bloß zwischen den Amidatomen anzuregen; man erhält saubere und übersichtliche Spektren. Die Technik ist in erster Linie für große Proteine vielversprechend, deren NOESY-Spektren nicht mehr analysiert werden können, weil sich die Resonanzen zu stark überlagern.

Aus Messungen mit unterschiedlichen Mischzeiten haben Mal et al. anhand bekannter, fester Abstände obere Schranken für die H_N-H_N-Abstände abgeschätzt. Die Originaldaten bestanden aus Messungen, die in vier verschiedene Klassen eingeteilt waren. Die oberen Abstandsschranken wurden anhand der Kristallstruktur [36] in Messungen der H_N-H_N-Abstände umgewandelt. So ergaben sich vier Gruppen mit Werten von 2.7 Å, 3.7 Å, 4.7 Å bzw. 6.1 Å. Der Datensatz besteht aus 48 sequentiellen, 36 kurzen ($|i-j| < 4$) und 70 langreichweitigen ($|i - j| \geq 4$) NOEs; insgesamt 154 Meßwerten.

Die Datendichte ist gering: Bei 59 Aminosäuren Länge fallen im Mittel etwas mehr als ein langreichweitiger Abstand auf jede Aminosäure. Hinsichtlich der Meßtechnik ist der Datensatz jedoch vollständig: fast alle theoretisch

beobachtbaren NOE-Signale wurden gemessen (für 188 Paare von Amidprotonen ist der Abstand kleiner 7 Å).

2.4.2 Daten für Ubiquitin

Das Protein Ubiquitin ist ein beliebtes Testsystem bei der Entwicklung neuer experimenteller und datenanalytischer Techniken der biologischen Kernresonanz. Bax und Mitarbeiter haben Daten für alle hier behandelten NMR-Observablen gemessen (PDB-Code 1d3z) [37].

Die NOESY-Daten wurden in einer einzigen Liste in der Datenbank abgelegt, so daß sich nicht mehr rekonstruieren läßt, aus welchen Spektren die Resonanzen stammen. In vielen Fällen sind die NOE-Signale eindeutig zugeordnet; es finden sich aber auch einige mit mehrdeutiger Zuordnung, für die nicht entschieden werden konnte, welches Protonenpaar mit chemischen Verschiebungen nahe den Frequenzpositionen der Resonanz Magnetisierung ausgetauscht hat.

Der Datensatz umfaßt 2727 Messungen; bei einer Länge von 76 Aminosäuren ist er redundant und einige Distanzen sind überbestimmt. Von allen NOEs konnten 2265 eindeutig zugeordnet werden, 462 blieben mehrdeutig. Von den 2265 eindeutig zugeordneten NOEs ist für 850 NOEs ein Meßwert vorhanden; für 454 NOEs gibt es zwei Meßwerte, für 61 drei, für 75 vier, für zwei fünf, für einen NOE sechs und für einen weiteren NOE acht Messungen.

Durch schwache Ausrichtung der Moleküle in einem Medium konnten Bax und Mitarbeiter [37] dipolare Kopplungen für die Bindungsvektoren messen, die die Peptidebene definieren: Dipolare Kopplungen der Bindungsvektoren D(N-H_N), D(C-H_N), D(C-N), D(C_α-C) und D(C_α-H_α) wurden für zwei verschiedene Orientierungen gemessen, die C_α-C_β Kopplung nur für eine Ausrichtung. Die Anzahl der Messungen ist:

	N-H_N	C-H_N	C-N	C_α-C	C_α-H_α	C_α-C_β
Ausrichtung 1	63	61	61	58	66	39
Ausrichtung 2	60	63	63	54	66	–

Die Dihedralwinkel φ_i werden durch die Atome C_{i-1}, N_i, $C_{\alpha,i}$ und C_i entlang des Proteinrückgrats festgelegt. Sechs skalare Kopplungskonstanten hängen direkt von φ ab: ${}^3J$(C-C), ${}^3J$(C-C_β), ${}^3J$(C-H_α), ${}^3J$(H_N-C), ${}^3J$(H_N-C_β) und ${}^3J$(H_N-H_α). Für ${}^3J$(C-C) lagen 55 Messungen vor, für ${}^3J$(C-C_β) 57, für ${}^3J$(C-H_α) 65, für ${}^3J$(H_N-C) 61, für ${}^3J$(H_N-C_β) 60 und für ${}^3J$(H_N-H_α) 63.

2.5 Verwendete Software

Zur Simulation der komplexen A posteriori-Verteilungen wurde ein selbstentwickeltes Programmpaket (siehe Anhang C) verwendet, das mittels dem in Kapitel 2.3 beschriebenen Replica-Monte-Carlo-Algorithmus Stichproben der Hypothesenparameter zieht. Die Moleküle sind in Dihedralwinkeln parametrisiert; zusätzliche Parameter werden ebenfalls geschätzt. Die Software ist in den Programmiersprachen Python und C geschrieben. Die meisten Graphiken wurden mit Hilfe der Python-Bibliothek biggles erstellt. Bloß für die Darstellung der Proteinstrukturen wurde MOLMOL [38] verwendet. Die Bewertung der Qualität der Strukturen erfolgte durch PROCHECK und WHATIF. PROCHECK [39] berechnet die Dihedralwinkel einer Konfiguration und vergleicht sie mit den Ramachandran-Statistiken wie sie sich aus den bekannten Proteinstrukturen ergeben. WHATIF [40] stellt verschiedene Qualitätsindizes für Proteinstrukturen zur Verfügung: QUACHK (quality check) und NQACHK (new quality check) sind zwei Maße, die die Packungsqualität von Proteinstrukturen bewerten. Es handelt sich um Z scores, d.h. bezüglich der Standardabweichung normierte Abweichungen vom Mittelwert für gut aufgelöste Kristallstrukturen. Der Werte sollten also idealerweise bei Null liegen. RAMCHK (Ramachandran check) bewertet die Ramachandran-Statistik; wieder handelt sich um einen Z score. BMPCHK (bump check) mißt die Anzahl der Atomdurchdringungen. Die Werte sind ein Mittel über die Anzahl solcher Zusammenstöße berechnet für die A Posteriori-Konformationen.

Kapitel 3

Ergebnisse

3.1 Datenverteilungen

Skalare und dipolare Kopplungen sowie dipolare Relaxationsraten geben Aufschluß über die Geometrie einer Spinkonfiguration. Einfache Theorien beschreiben die Abhängigkeit dieser Meßgrößen von der molekularen Struktur. Eine unbekannte Fehlerskala parametrisiert Abweichungen zwischen experimentellen und theoretischen Werten. Theorie und Fehlergesetz bestimmen die Wahrscheinlichkeitsverteilung der Observable.

3.1.1 Skalare Kopplungskonstanten

Die Elektronen der chemischen Bindungen vermitteln eine Wechselwirkung zweier Kernspins; dies ist die indirekte Kopplung. In isotroper Lösung wird die Stärke der Wechselwirkung durch die skalare Kopplungskonstante ${}^nJ$ angeben, wobei n die Anzahl der Bindungen ist. Sind die Kerne durch drei Bindungen getrennt, hängt die skalare Kopplungskonstante in erster Linie von dem Dihedralwinkel θ ab, der durch die Bindungen definiert wird: ${}^3J = {}^3J(\theta)$. Weitere Einflüsse, die von der chemischen Umgebung oder lokaler Dynamik herrühren, werden vernachlässigt.

Karplus hat eine einfache Beziehung für den Zusammenhang zwischen der

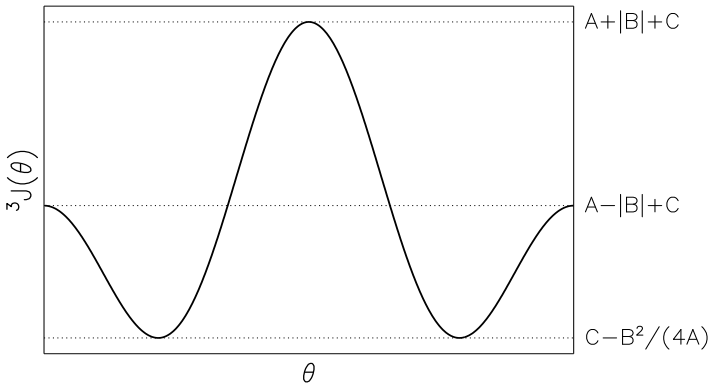


Abbildung 3.1: Karplus-Kurve zur Beschreibung der Abhängigkeit der skalaren Kopplungskonstante ${}^3J$ vom Dihedralwinkel θ.

skalaren Kopplungskonstante und dem Dihedralwinkel abgeleitet [5]:

$$ {}^3J(\theta) = A\,\cos^2\theta + B\,\cos\theta + C. \tag{3.1}$$

Dies sind die ersten drei Gliedern einer Fourier-Entwicklung. Die Koeffizienten A, B, C können quantenmechanisch berechnet werden [41, 42]. In der wahrscheinlichkeitsbasierten Strukturbestimmung sind sie jedoch unbekannte „Materialkonstanten“ die zu Hypothesenparametern werden. Die Maxima liegen bei $\sin\theta = 0$, die Minima bei $\cos\theta = -B/(2A)$ (siehe Abb. 3.1).

Abweichungen zwischen gemessenen und berechneten skalaren Kopplungskonstanten beschreibt ein Gauß'sches Fehlermodell. Diese Verteilung ist vorurteilsfrei, weil sie auf wenigen, plausiblen Annahmen beruht: Erstens soll die Vorhersage nicht systematisch von den Meßwerten abweichen, also

$$ {}^3J(\theta) = \langle {}^3J \rangle,$$

wobei $\langle\cdot\rangle$ ein Mittel über die gesuchte Fehlerverteilung ist. Zweitens soll die mittlere quadratische Abweichung

$$ \sigma^2 = \left\langle \left[{}^3J - {}^3J(\theta) \right]^2 \right\rangle$$

zwischen gemessenen und berechneten Werten bekannt sein. Wir suchen das Fehlergesetz, das außer den beiden genannten keine zusätzlichen Annahmen

trifft. Die Gauß-Verteilung (siehe auch Abschnitt A.1)

$$g\big({}^3J;\, {}^3J(\theta), \sigma\big) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left({}^3J - {}^3J(\theta) \right)^2 \right\} \tag{3.2}$$

hat von allen Verteilungen, die die beiden Bedingungen erfüllen, die größte Entropie; nach dem Maximum-Entropie-Prinzip [23] gehen in sie keine weiteren Annahmen ein; solche würden sich in einer Verminderung der Entropie niederschlagen.

Messungen skalarer Kopplungskonstanten $D = \{{}^3J_1, \ldots, {}^3J_n\}$ resultieren in der Likelihood-Funktion

$$L(X, A, B, C, \sigma) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2}\, \chi^2_{_J}(X, A, B, C) \right\},$$

wobei

$$\chi^2_{_J}(X, A, B, C) = \sum_i \left({}^3J_i - A\, \cos^2\theta_i - B\, \cos\theta_i - C \right)^2$$

die Abweichung zwischen Messungen und Vorhersagen quantifiziert und die Dihedralwinkel θ_i aus Konformation X berechnet werden. Mit der Matrix

$$\mathbf{A}_{_J} = \begin{pmatrix} \cos^2\theta_1 & \cos\theta_1 & 1 \\ \vdots & \vdots & \vdots \\ \cos^2\theta_n & \cos\theta_n & 1 \end{pmatrix} \tag{3.3}$$

und den Daten- und Parametervektoren

$$\mathbf{j} = ({}^3J_1, \ldots, {}^3J_n)^T \quad \text{bzw.} \quad \mathbf{a} = (A, B, C)^T$$

läßt sich die Abweichung als quadratische Form schreiben:

$$\chi^2_{_J}(X, A, B, C) = (\mathbf{j} - \mathbf{A}_{_J}\, \mathbf{a})^T (\mathbf{j} - \mathbf{A}_{_J}\, \mathbf{a}). \tag{3.4}$$

Das Maximum der Likelihood-Funktion $L(X, A, B, C, \sigma)$ bezüglich der Karplus-Koeffizienten ergibt sich durch Minimierung von $\chi^2_{_J}$. Differentiation von (3.4) nach den Koeffizienten liefert das Gleichungssystem

$$(\mathbf{A}^T_{_J}\mathbf{A}_{_J})\, \mathbf{a} = \mathbf{A}^T_{_J}\mathbf{j}.$$

Falls $(\mathbf{A}_J^T\mathbf{A}_J)$ regulär ist, existiert die Pseudo-Inverse $\mathbf{A}_J^\dagger = (\mathbf{A}_J^T\mathbf{A})_J^{-1}\mathbf{A}_J^T$, und als *Maximum-Likelihood-Schätzwert* [20] der Karplus-Parameter erhält man

$$\hat{\mathbf{a}} = \mathbf{A}_J^\dagger \, \mathbf{j}. \tag{3.5}$$

Um die Schätzwerte von X und σ zu bestimmen, wird $\hat{\mathbf{a}}$ in die Likelihood-Funktion eingesetzt und der negative Logarithmus der resultierenden Funktion

$$\frac{1}{2\sigma^2}\,\mathbf{j}^T(\mathbf{I} - \mathbf{A}_J\mathbf{A}_J^\dagger)\,\mathbf{j} + n\,\log\sigma$$

minimiert. Der geschätzte Fehler ist

$$\hat{\sigma} = \sqrt{\frac{\chi_j^2(X,\hat{A},\hat{B},\hat{C})}{n}} = \sqrt{\frac{\mathbf{j}^T(\mathbf{I} - \mathbf{A}_J\mathbf{A}_J^\dagger)\,\mathbf{j}}{n}};$$

er zeigt das übliche $1/\sqrt{n}$ - Verhalten.

Erst die Berücksichtigung der Zustandssumme $Z(\sigma) = \sqrt{2\pi\sigma^2}$ ermöglicht, den Fehler zu bestimmen. Ohne den Beitrag der Zustandssumme in

$$\frac{1}{2\sigma^2}\,\chi_J^2 + n\,\log\sigma$$

würde man unsinnige Schätzwerte erhalten: Bei endlichem Residuum χ_J^2 wird $\chi_J^2/2\sigma^2$ minimal für $\sigma \to \infty$, obwohl intuitiv σ endlich bleiben sollte, je nach Größe von χ_J^2. Falls die Theorie die Daten exakt vorhersagt, $\chi_J^2 = 0$, ist σ unbestimmt, obwohl der Fehler dann verschwinden müßte.

3.1.2 Dipolare Kopplungen

Über die Wechselwirkung ihrer Dipolmomente sind zwei Kernspins k und l direkt gekoppelt. Der Dipoltensor $\mathbf{D}_{kl}$ (1.2) vermittelt diese Wechselwirkung. Weil die Ortskoordinaten dem Gitter angehören, muß im Hamiltonoperator über sie gemittelt werden. In der Säkularapproximation [3] gilt:

$$H_{dipol} \approx \frac{1}{2}\sum_{k\neq l}\frac{\mu_0\hbar\,\gamma_k\gamma_l}{4\pi}\;I_{kz}I_{lz}\;\mathbf{e}_z^T\langle\mathbf{D}_{kl}\rangle\mathbf{e}_z.$$

Die Koordinaten können in äußere und innere Anteile zerlegt werden:

$$\mathbf{r}_k = \mathbf{R}\,\tilde{\mathbf{r}}_k + \mathbf{t},$$

wobei $\tilde{\mathbf{r}}_k$ die Atompositionen in einem fest mit dem Molekül verbundenen Koordinatensystem sind. Die Rotationsmatrix $\mathbf{R}$ beschreibt die Orientierung des Moleküls, $\mathbf{t}$ ist ein Translationsvektor. Für den Dipoltensor $\tilde{\mathbf{D}}_{kl}$ im Koordinatensystem des Moleküls gilt

$$\mathbf{D}_{kl} = \mathbf{R}\,\tilde{\mathbf{D}}_{kl}\,\mathbf{R}^T.$$

Bei einem starren Molekül muß nur noch über die Orientierungen gemittelt werden. Wegen der Orthonormalität der Drehmatrix ist $\mathbf{r} = \mathbf{R}^T\mathbf{e}_z$ ein Einheitsvektor. Die Mittelung vereinfacht sich zu

$$\mathbf{e}_z^T \langle \mathbf{D}_{kl} \rangle \mathbf{e}_z = \mathrm{tr}\left[\langle \mathbf{r}\mathbf{r}^T \rangle \tilde{\mathbf{D}}_{kl}\right] = \tilde{\mathbf{r}}_{kl}^T\,\mathbf{S}\,\tilde{\mathbf{r}}_{kl}/r_{kl}^5.$$

Die Matrix [43]

$$\mathbf{S} = \langle 3\,\mathbf{r}\mathbf{r}^T - \mathbf{I} \rangle = \langle \mathbf{R}^T (3\,\mathbf{e}_z\mathbf{e}_z^T - \mathbf{I})\,\mathbf{R} \rangle$$

charakterisiert die mittlere Ausrichtung des Moleküls und wird Saupe-Matrix genannt. Sie hat nur fünf unabhängige Elemente: die Parametrisierung

$$\mathbf{S} = \begin{pmatrix} s_1 - s_2 & s_3 & s_4 \\ s_3 & -s_1 - s_2 & s_5 \\ s_4 & s_5 & 2s_2 \end{pmatrix}$$

erfüllt die Bedingungen $\mathrm{tr}\,\mathbf{S} = 0$ und $\mathbf{S}^T = \mathbf{S}$. Die dipolare Kopplung berechnet sich zu

$$D_i(X, \mathbf{s}) = \sum_{j=1}^{n_i} \mu_j\ \mathbf{r}_j^T \mathbf{S} \mathbf{r}_j / r_j^5 = \mathbf{a}_i^T \mathbf{s}, \tag{3.6}$$

wobei alle n_i Beiträge mit Kopplungskonstante $\mu_j = \mu_0 \hbar \gamma_{k,j} \gamma_{k,j}/(4\pi)$, die zu derselben Resonanz beitragen, aufzusummieren sind. Bei prochiralen Gruppen hat man beispielsweise zwei Beträge [44]. Außerdem wurden die Vektoren

$$\mathbf{s} = (s_1, s_2, s_3, s_4, s_5)^T$$

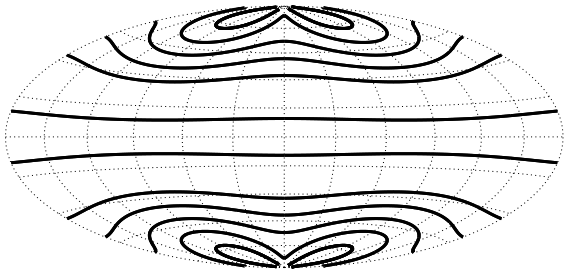


Abbildung 3.2: Linien konstanter dipolarer Kopplungen als Weltkartenprojektion des Polar- und Azimuthalwinkels θ bzw. φ. Der Azimuthalwinkel definiert den Längengrad, der Polarwinkel den Breitengrad.

und

$$\mathbf{a}_i = \sum_j \frac{\mu_j}{r_j^5} \left(x_j^2 - y_j^2,\, 3z_j^2 - r_j^2,\, 2x_j y_j,\, 2x_j z_j,\, 2y_j z_j\right)^T$$

eingeführt.

Oft wird die Diagonaldarstellung $\mathbf{S} = \mathbf{R}\,\mathbf{\Lambda}\,\mathbf{R}^T$ der Saupe-Matrix verwendet. $\mathbf{R}$ ist eine Rotation und $\mathbf{\Lambda}$ die Diagonalmatrix der Eigenwerte λ_1, λ_2, λ_3. Weil die Spur verschwindet, ist die Summe der Eigenwerte Null. Durch eine Permutation der Spalten von $\mathbf{R}$ kann man die Eigenwerte so wählen, daß sie nach der Größe ihres Betrags geordnet sind: $|\lambda_3| \geq |\lambda_2| \geq |\lambda_1|$. Die axiale und die rhombische Komponente des Orientierungstensors werden definiert als:

$$A = \frac{\lambda_3}{2}, \quad R = \frac{2}{3}\,\frac{\lambda_1 - \lambda_2}{\lambda_3}.$$

Die Rhombizität kann nur Werte in $[0, 2/3]$ annehmen. Im Koordinatensystem, das durch die Spalten von $\mathbf{R}$ definiert wird, gilt:

$$\mathbf{r}^T\mathbf{S}\mathbf{r}/r^2 = A\,\left[\,3\cos^2\theta - 1 + 3R\,\sin^2\theta\,\cos(2\varphi)/2\,\right],$$

wobei (r, θ, φ) die Kugelkoordinaten von $\mathbf{r}$ sind. Dipolare Kopplungen sind degeneriert; ein konstanter Wert liefert unendlich viele Einstellungen von θ und φ (siehe Abb. 3.2).

In wässriger Lösung diffundieren die Moleküle ohne Vorzugsrichtung und dipolare Beiträge verschwinden. Aber allein der Einfluß eines äußeren Magnetfelds kann die Moleküle ausrichten (dies ist z.B. bei Hämproteinen der

Fall). Außerdem kann durch Veränderung des Lösungsmittels die Isotropie aufgehoben werden. Das erreichen beispielsweise kristalline Flüssigkeiten [43, 45]: Wechselwirkungen mit dem Lösungsmittel richten das Molekül im Mittel schwach aus.

Die Messung einer dipolaren Kopplung der Stärke D_i wird wieder mit einem Gauß'schen Fehlergesetz beschrieben. Für n dipolare Kopplungen $D = \{D_1, \ldots, D_n\}$ ist die Abweichung

$$\chi^2_{D}(X, s_1, \ldots, s_5) = \sum_i \left(D_i - D_i(X, \mathbf{s})\right)^2 = (\mathbf{d} - \mathbf{A}_{D}\mathbf{s})^T(\mathbf{d} - \mathbf{A}_{D}\mathbf{s}) \quad (3.7)$$

mit $\mathbf{d} = (D_1, \ldots, D_n)^T$ und $\mathbf{A}_{D} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)^T$. Die Saupe-Matrix wird durch das Lösungsmittel bestimmt; sie ist vor Durchführung des Experiments unbekannt. Es ergeben sich die Schätzwerte:

$$\hat{\mathbf{s}} = \mathbf{A}^{\dagger}_{D}\mathbf{d} \quad \text{mit} \quad \mathbf{A}^{\dagger}_{D} = (\mathbf{A}^T_{D}\mathbf{A}_{D})^{-1}\mathbf{A}^T_{D}.$$

Minimierung von

$$\frac{1}{2\sigma^2}\,\mathbf{d}^T(\mathbf{I} - \mathbf{A}_{D}\mathbf{A}^{\dagger}_{D})\mathbf{d} + n\,\log\sigma$$

liefert Schätzwerte der Koordinaten und des Fehlers.

3.1.3 Dipolare Relaxationsraten

Im Gegensatz zur direkten Kopplung können die Relaxationsraten der dipolaren Spin-Spin-Wechselwirkung in isotropen Lösungsmitteln beobachtet werden; dazu dient das NOESY-Experiment [4]. Diagonalisierung des Relaxationssuperoperators liefert die Relaxationsraten [3].

Solomon [6] beschreibt die Zeitentwicklung der longitudinalen magnetischen Momente mit der Ratengleichung

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta\mathbf{M}_z(t) = \mathbf{R}\,\Delta\mathbf{M}_z(t).$$

Aus zeitlicher Störungsrechnung folgt, daß die Elemente der Relaxationsmatrix $\mathbf{R}$ proportional zur sechsten inversen Potenz des räumlichen Abstands der Kernspins sind: Die Dipol-Wechselwirkung hängt von der inversen dritten

Potenz des Abstands ab und geht in eine Störungsrechnung zweiter Ordnung quadratisch ein. Also: $R_{kl} \propto \langle r_{kl}^{-6} \rangle$. Integration der Ratengleichung führt auf einen exponentiellen Abfall der Magnetisierungen. Das Volumen V_{kl} einer Resonanz ist proportional zu $[\exp(-\tau_m \mathbf{R})]_{kl}$ [4]. Die Dauer τ_m des Mischprozesses bestimmt, wie stark die Spindiffusion den direkten Austausch von Magnetisierung beeinflußt.

Bei kurzen Mischzeiten (initial rate regime) erhält man dasselbe Resultat wie für ein Paar isolierter Spins (isolated spin pair approximation, ISPA):

$$V_{kl} \propto \langle r_{kl}^{-6} \rangle.$$

Über die chemischen Verschiebungen kann jeder Resonanz ein Paar von Atomen zugeordnet werden, deren Kernspins k und l im Mischprozeß Magnetisierung ausgetauscht haben. Es sei $d_i = r_{kl}$ der Abstand zwischen den Protonen, die der i-ten Resonanz zugeordnet wurden. Das Volumen ist

$$V_i = \gamma \, d_i^{-6} \tag{3.8}$$

mit einer Proportionalitätskonstante γ. Der gemittelte Abstand wurde durch seinen instantanen Wert angenähert.

Wegen $V > 0$ darf das Fehlergesetz nur positiven Werten eine endliche Wahrscheinlichkeit zuweisen. Ein Spektrum kann beliebig skaliert sein, ohne daß sich sein Informationsgehalt ändert. Abweichungen zwischen Vorhersage $\hat{V}$ und Messung V müssen deshalb gleich bewertet werden, wenn beide um einem Faktor γ skaliert sind. Aus dieser Invarianzforderung folgt nach (2.4):

$$g(\gamma V; \gamma \hat{V}, \sigma) = \gamma^{-1} \, g(V; \hat{V}, \sigma).$$

Für $\gamma = \hat{V}^{-1}$ ergibt sich

$$g(V; \hat{V}, \sigma) = \hat{V}^{-1} \, g(V/\hat{V}; 1, \sigma) = \hat{V}^{-1} \, \hat{g}(V/\hat{V}; \sigma),$$

wobei die Verteilung $\hat{g}(x; \sigma) = g(x; 1, \sigma)$ auf der positiven Achse definiert ist. Die Invarianzforderung schreibt also vor, Abweichungen zwischen experimentellen und theoretischen Volumina über ihr Verhältnis zu messen; das entspricht einem multiplikativen Fehler.

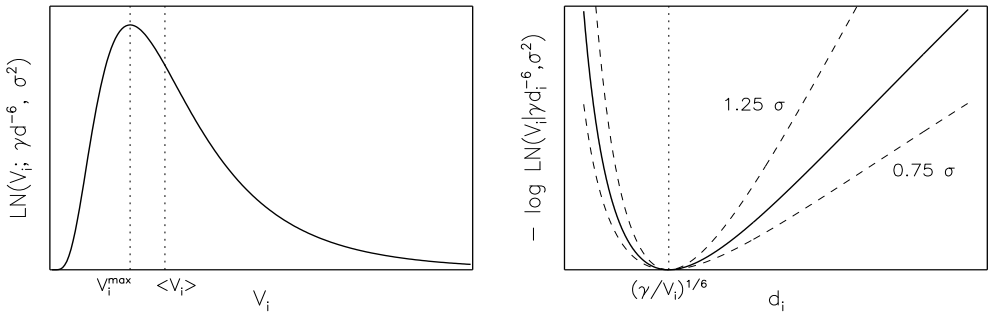


Abbildung 3.3: Links: Datenverteilung der NOESY-Volumina. Rechts: Resultierender Beitrag zum negativen Logarithmus der Likelihood-Funktion.

Der Logarithmus eines multiplikativen Fehlers ist symmetrisch. Ein geeignetes Maß, um multiplikative Abweichungen zwischen theoretischen und experimentellen Volumina zu bewerten, ist deshalb der Logarithmus ihrer Verhältnisse. Wir suchen eine Verteilung, die außer

$$\left\langle \log(V/\hat{V}) \right\rangle = 0 \quad \text{und} \quad \left\langle \log^2(V/\hat{V}) \right\rangle = \sigma^2$$

keine weiteren Bedingungen erfüllt. Aus dem Maximum-Entropie-Prinzip folgt als gesuchtes Fehlergesetz die Lognormal-Verteilung (Anhang A.2 und Abb. 3.3). Die Wahrscheinlichkeit, ein Volumen V_i zu messen, ist somit:

$$p(V_i|X,\gamma,\sigma,I) = \frac{1}{\sqrt{2\pi\sigma^2}\,V_i} \exp\left\{-\frac{1}{2\sigma^2}\,\log^2(\gamma d_i^{-6}/V_i)\right\}, \qquad (3.9)$$

wobei d_i der aus Konfiguration X berechnete Abstand ist. Die Lognormal-Verteilung ist konservativ, weil sie Abweichungen von Größenordnungen bewertet. Sie hat den Vorteil, daß sich ihre Form nicht ändert, wenn man statt der Volumina unkalibrierte experimentelle Distanzen $V_i^{-1/6}$ verwendet.

In Abhängigkeit von d_i betrachtet, ist der negative Logarithmus der Datenverteilung analog zum restraint potential der Minimierungsverfahren:

$$F_i(d_i) = \frac{18}{\sigma^2}\,\log^2(d_i/\hat{d}_i) \quad \text{mit} \quad \hat{d}_i = \gamma^{1/6}\,V_i^{-1/6}.$$

Distanzen aus dem unsymmetrischen Intervall $\left[\hat{d}_i\,e^{-\sqrt{2F_i}\,\sigma}, \hat{d}_i\,e^{\sqrt{2F_i}\,\sigma}\right]$ haben höchstens eine „Energie“ F_i. Die Intervallbreite ist proportional zur gemessenen Distanz $\hat{d}_i$. Mit wachsender Distanz werden die Abstandsschranken

größer. Zusätzlich geht durch σ die Qualität der Daten und des Modells in die Breite des Intervalls ein. Weil γ und σ mit den Koordinaten geschätzt werden, passen sich die Abstandsintervalle automatisch an.

Für n zugeordnete NOESY-Signale mit Volumina $D = \{V_1, \ldots, V_n\}$ erhält man als Abweichung zu den theoretischen Werten

$$\chi^2_V(X,\gamma) = \sum_i \log^2(\gamma\, d_i^{-6}/V_i) = n\,\left[\log^2\left(\gamma\,\overline{d}^{\,-6}/\,\overline{V}\right) \;+\; s^2\right]. \qquad (3.10)$$

Hier wurden die geometrischen Mittel der gemessenen und der unskalierten berechneten Volumina sowie die Varianz der Logarithmen ihrer Verhältnisse eingeführt:

$$\overline{V} = \Big(\prod_i V_i\Big)^{1/n}, \quad \overline{d} = \Big(\prod_i d_i\Big)^{1/n}, \quad s^2 = \frac{1}{n}\sum_i \log^2\left(V_i\, d_i^6\,/\,\overline{V}\,\overline{d}^{\,6}\right).$$

Durch Minimierung von $\frac{1}{2\sigma^2}\;\chi^2_V(X,\gamma) \;+\; n\,\log\sigma$ erhält man Schätzwerte der Skalen γ und σ; es gilt:

$$\hat{\gamma} = \overline{V}\;\overline{d}^{\,6} = \Big(\prod_i V_i\, d_i^6\Big)^{1/n} \quad \text{und} \quad \hat{\sigma} = \sqrt{s^2/n}.$$

3.2 A posteriori-Verteilungen

Die A posteriori-Verteilung ist die vollständige Lösung eines Strukturbestimmungsproblems; sie ist das Produkt der A priori- und der Datenverteilung. Einfache Modelle kernspektroskopischer Meßgrößen wurden im letzten Kapitel abgeleitet. Die A priori-Verteilungen der Dihedralwinkel sowie der Fehler wurden in den Abschnitten 2.2.5 und 2.2.6 eingeführt. Zur vollständigen Charakterisierung der A posteriori-Verteilung müssen noch die A priori-Verteilungen der Theorieparameter gewählt werden.

3.2.1 Skalare Kopplungskonstanten

Die Koeffizienten der Karplus-Kurve können quantenmechanisch berechnet werden [41, 42]. Eine solche Theorie könnte in die A priori-Verteilung

$$\pi(A, B, C|\theta)$$

eingehen. Weil die Berechnung der Karplus-Koeffizienten aufwendig ist, wählen wir jedoch ihre A priori-Verteilung unabhängig von der Konformation. Als einfachste A priori-Verteilung bietet sich eine Gleichverteilung an:

$$\pi(A,B,C)\ \mathrm{d}A\,\mathrm{d}B\,\mathrm{d}C\ = \mathrm{d}A\,\mathrm{d}B\,\mathrm{d}C\,. \tag{3.11}$$

Die Koeffizienten A, B und C können positive und negative Werte annehmen; völliges Unwissen über sie wird durch (3.11) ausgedrückt. Auch diese A priori-Verteilung ist wie Jeffreys' prior (2.16) eine uneigentliche Verteilung. Strenggenommen müssen A, B und C auf endliche Intervalle eingegrenzt und die Grenzen gegen unendlich geschickt werden.

Das Problem ist nun vollständig definiert. Die A posteriori-Verteilung aller unbekannten Größen ist

$$\begin{aligned} p(\theta,A,B,C,\sigma) \quad &\propto \quad \sigma^{-(n+1)} \exp\left\{-\frac{1}{2\sigma^2}\,\chi_J^2(\theta,A,B,C) - \beta\,E(\theta)\right\} \\ &\quad\ \times\ \pi(A,B,C). \end{aligned}$$

Die bedingte A posteriori-Verteilung der Karplus-Koeffizienten ist eine dreidimensionale Gauß-Verteilung:

$$p(A,B,C|\theta,\sigma) \propto \exp\left\{-\frac{1}{2\sigma^2}\,(\mathbf{a}-\hat{\mathbf{a}})^T\,(\mathbf{A}_J^T\mathbf{A}_J)\,(\mathbf{a}-\hat{\mathbf{a}})\right\} \tag{3.12}$$

mit den Abkürzungen aus Abschnitt 3.1.1. Die Karplus-Koeffizienten streuen um die Schätzwerte $\hat{\mathbf{a}} = \mathbf{A}_J^\dagger\mathbf{j}$ mit der Kovarianz-Matrix $\sigma^2(\mathbf{A}_J^T\mathbf{A}_J)^{-1}$. Die bedingte A posteriori-Verteilung des Fehlers ist

$$p(\sigma|\theta,A,B,C) \propto \sigma^{-(n+1)}\ \exp\left\{-\frac{1}{2\sigma^2}\,\chi_J^2(\theta,A,B,C)\right\}.$$

Die inverse Varianz $\lambda = \sigma^{-2}$ folgt einer Gamma-Verteilung (siehe A.3)

$$p(\lambda|\theta,A,B,C) = \mathrm{G}\Big(\lambda; n/2, \chi_J^2/2\Big). \tag{3.13}$$

Der unbekannte Fehler kann in der A posteriori-Verteilung ausintegriert werden (siehe Abschnitt 2.2.3). Die marginale A posteriori-Verteilung der Dihedralwinkel und der Karplus-Koeffizienten ist

$$p(\theta,A,B,C) \propto \left[\,\chi_J^2(\theta,A,B,C)\,\right]^{-n/2}\ \exp\big\{-\beta E(\theta)\big\}\,\pi(A,B,C). \tag{3.14}$$

Weil die A priori-Verteilungen der Hypothesenparameter unabhängig voneinander gewählt wurden, entspricht eine Integration über σ der Verwendung der reduzierten Datenverteilung

$$p(D|\theta, A, B, C) \propto \left[(\mathbf{j} - \mathbf{A}_J \mathbf{a})^T (\mathbf{j} - \mathbf{A}_J \mathbf{a}) \right]^{-n/2},$$

die allein durch θ und A, B, C bedingt wird. $p(D|\theta, A, B, C)$ ist eine Verbundverteilung aller Messungen; sie läßt sich nicht mehr in ein Produkt von Verteilungen $p(J_i|\theta, A, B, C)$ zerlegen. Durch Einführung und Integration von σ wurden die Daten gekoppelt.

Ebensogut kann über die Karplus-Koeffizienten integriert werden. Die marginale A posteriori-Verteilung der Dihedralwinkel und des Fehlers ist

$$p(\theta, \sigma) \propto \sigma^{-(n-2)} \left| \mathbf{A}_J^T \mathbf{A}_J \right|^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} R_J(\theta) - \beta E(\theta) \right\}$$

mit dem Residuum $R_J(\theta) = \mathbf{j}^T (\mathbf{I} - \mathbf{A}_J \mathbf{A}_J^\dagger) \mathbf{j}$; $|\cdot|$ bezeichnet die Determinante einer Matrix. Wieder kann eine Datenverteilung definiert werden, in der nun die Karplus-Koeffizienten entfernt und alle Messungen koppelt sind:

$$p(D|\theta, \sigma, I) \propto \sigma^{-(n-3)} \left| \mathbf{A}_J^T \mathbf{A}_J \right|^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} R_J(\theta, \sigma) \right\}.$$

Der Exponent des Fehlers σ zeigt an, daß durch Integration über die Karplus-Koeffizienten drei Messungen aufgebraucht wurden.

Durch weitere Integration über σ ist es möglich, die gemeinsame A posteriori-Verteilung $p(\theta, A, B, C, \sigma)$ auf den eigentlichen Hypothesenraum, den Konfigurationsraum des Makromoleküls, zu reduzieren:

$$p(\theta) \propto \exp\{-\beta E(\theta)\} \left| \mathbf{A}_J^T \mathbf{A}_J \right|^{-1/2} \left[R_J(\theta) \right]^{-(n-3)/2}. \tag{3.15}$$

Diese Reduzierung kann auch als Verwendung der gekoppelten Datenverteilung

$$p(D|\theta, I) \propto \left| \mathbf{A}_J^T \mathbf{A}_J \right|^{-1/2} \left[R_J(\theta) \right]^{-(n-3)/2}$$

gedeutet werden.

Bei den einfachen hier verwendeten Modellen ist es möglich, die nuisance parameter analytisch auszuintegrieren. Bei komplizierteren Modellen kann

jedoch eine Integration über zusätzliche Hypothesenparameter nur numerisch erfolgen (siehe Abschnitt 2.3).

3.2.2 Dipolare Kopplungen

Zur Berechnung dipolarer Kopplungen muß der Orientierungstensor bekannt sein. Erst die Wahl seiner A priori-Verteilung vervollständigt die Beschreibung des Problems. Die Kenntnis der Struktur und der Eigenschaften des Lösungsmittels erlaubt, den Orientierungstensor abzuschätzen [46], und würde in $\pi(s_1, \ldots, s_5|\theta)$ eingehen. Wir wollen uns jedoch wieder mit der einfachsten Beschreibung begnügen und nehmen für die Elemente der Saupe-Matrix eine von der Struktur unabhängige A priori-Verteilung an. Wenn wir vorab nichts über die Elemente s_i sagen können, wählen wir eine Gleichverteilung

$$\pi(s_1, \ldots, s_5)\, \mathrm{d}s_1 \cdots \mathrm{d}s_5 = \mathrm{d}s_1 \cdots \mathrm{d}s_5\,. \tag{3.16}$$

Die A posteriori-Verteilung der dipolaren Kopplungen hat dieselben Eigenschaften wie die der skalaren Kopplungskonstanten. Es gilt

$$\begin{aligned} p(\theta, s_1, \ldots, s_5, \sigma) \quad \propto \quad & \sigma^{-(n+1)} \exp\left\{ -\frac{1}{2\sigma^2}\, \chi_D^2\,(\theta, s_1, \ldots, s_5) - \beta\, E(\theta) \right\} \\ & \times\, \pi(s_1, \ldots, s_5). \end{aligned}$$

Die bedingte A posteriori-Verteilung der Elemente der Saupe-Matrix ist eine fünfdimensionale Gauß-Glocke:

$$p(s_1, \ldots, s_5|\theta, \sigma) \propto \exp\left\{ -\frac{1}{2\sigma^2}\, (\mathbf{s} - \hat{\mathbf{s}})^T\, (\mathbf{A}_D^T \mathbf{A}_D)\, (\mathbf{s} - \hat{\mathbf{s}}) \right\}. \tag{3.17}$$

Die Elemente der Saupe-Matrix streuen um $\hat{\mathbf{s}} = \mathbf{A}_D^{\dagger}\mathbf{d}$. Die bedingte A posteriori-Verteilung der inversen Varianz $\lambda = \sigma^{-2}$ ist eine Gamma-Verteilung

$$p(\lambda|\theta, s_1, \ldots, s_5) = \mathrm{G}\Big(\lambda; n/2, \chi_D^2/2\Big). \tag{3.18}$$

Integration über den Fehler liefert die marginale A posteriori-Verteilung:

$$p(\theta, s_1, \ldots, s_5) \propto \big[\, \chi_D^2\,(\theta, s_1, \ldots, s_5)\, \big]^{-n/2} \exp\big\{ -\beta\, E(\theta) \big\}\, \pi(s_1, \ldots, s_5).$$

Es kann auch über die Elemente der Saupe-Matrix intergiert werden:

$$p(\theta,\sigma) \propto \sigma^{-(n-4)} \left|\mathbf{A}_D^T\mathbf{A}_D\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2} R_D(\theta) - \beta E(\theta)\right\} \tag{3.19}$$

mit $R_D(\theta) = \mathbf{d}^T(\mathbf{I} - \mathbf{A}_D\mathbf{A}_D^\dagger)\,\mathbf{d}$. Integration über s_i und σ liefert die marginale A posteriori-Verteilung der Konformation:

$$p(\theta) \propto \exp\{-\beta E(\theta)\} \left|\mathbf{A}_D^T\mathbf{A}_D\right|^{-1/2} [\,R_D(\theta)\,]^{-(n-5)/2}. \tag{3.20}$$

Mittelung über die Elemente Saupe-Matrix vermindert die Anzahl der Messungen um fünf.

3.2.3 Dipolare Relaxationsraten

In der Beschreibung der dipolaren Relaxationskonstanten gibt es bloß einen Theorieparameter, den Kalibrationsfaktor γ. Über γ ist lediglich bekannt, daß er ein Skalenparameter ist: Eine Skalierung des gesamten Spektrums kann durch eine Skalierung von γ kompensiert werden und sollte gleich wahrscheinlich sein. Die A priori-Verteilung von γ hat deshalb dieselbe Form wie die des Fehlers:

$$\pi(\gamma)\,\mathrm{d}\gamma \;=\; \mathrm{d}\log\gamma = \gamma^{-1}\,\mathrm{d}\gamma\,.$$

Aus einer Messung von n zugeordneten Resonanzen eines NOESY-Spektrums mit zugehörigen Volumina $V_1,\ldots,V_n$ erhalten wir die A posteriori-Verteilung

$$p(\theta,\gamma,\sigma) \propto \sigma^{-(n+1)}\gamma^{-1} \exp\left\{-\frac{n}{2\sigma^2}\left[\log^2\left(\overline{V}\,\overline{d}^{\,6}/\gamma\right) + s^2\right] - \beta E(\theta)\right\}. \tag{3.21}$$

Die bedingte A posteriori-Verteilung der Skala γ ist eine Lognormal-Verteilung

$$p(\gamma|\theta,\sigma) = \mathrm{LN}\left(\gamma; \overline{V}\,\overline{d}^{\,6}(\theta), \sigma^2/n\right), \tag{3.22}$$

die des inversen quadratischen Fehlers eine Gamma-Verteilung:

$$p(\lambda|\theta,\gamma) = \mathrm{G}\left(\lambda; n/2, \chi_V^2(\theta,\gamma)/2\right). \tag{3.23}$$

Integration über den Kalibrationsfaktor liefert

$$p(\theta,\sigma) \propto \sigma^{-n} \exp\left\{-\beta E(\theta) - \frac{n}{2\sigma^2} s^2\right\}.$$

Wegen $n\,s^2 = \sum_i \log^2(V_i/\hat{\gamma} d_i^{-6})$ mit $\hat{\gamma} = \overline{V}/\,\overline{d}^{\,-6}$ entspricht $p(\theta,\sigma)$ der Verwendung kalibrierter Messungen: In ARIA [47] werden beispielsweise die Volumina durch den Faktor $\sum_i V_i/\sum_i d_i^{-6}$ geeicht; bei genügend vielen Messungen ist dieser Faktor mit $\hat{\gamma}$ nahezu identisch.

Anstatt über γ kann man auch über σ integrieren und erhält als marginale A posteriori-Verteilung

$$p(\theta,\gamma) \propto \gamma^{-1}\,\exp\{-\beta E(\theta)\}\,\left[\log^2\left(\overline{V}\,\overline{d}^{\,6}/\gamma\right) + s^2\right]^{n/2}.$$

Durch eine weitere Integration lassen sich alle zusätzlichen Parameter entfernen. Die marginale A posteriori-Verteilung der Dihedralwinkel ist

$$p(\theta) \propto \exp\bigl\{-\beta E(\theta)\bigr\}\,\left[\,s^2(\theta)\,\right]^{-(n-1)/2}. \tag{3.24}$$

Die Struktur wird vollständig durch die Daten D und das Vorwissen I bestimmt. Es müssen keine zusätzlichen Annahmen getroffen werden.

3.2.4 Allgemeiner Fall

Allgemein liegen Datensätze verschiedener kernspektroskopischer Meßgrössen vor. Es seien q Datensätze der skalaren Kopplungskonstanten gemessen worden, p Datensätze der dipolaren Kopplungen und r zugeordnete NOESY-Spektren. Die vollständige Lösung des Strukturbestimmungsproblems ist die A posteriori-Verteilung

$$\begin{aligned} p\,(\theta,\alpha,\sigma) \;\propto\; & \exp\bigl\{-\beta\,E(\theta)\bigr\} \\ & \times \left(\prod_{i=1}^{p} \sigma_{J,i}^{-(n_i+1)}\,\exp\Bigl\{-\frac{1}{2\sigma_{J,i}^2}\,\chi_J^2(\theta,\mathbf{a}_i)\Bigr\}\;\pi(\mathbf{a}_i)\right) \\ & \times \left(\prod_{j=1}^{q} \sigma_{D,j}^{-(n_j+1)}\,\exp\Bigl\{-\frac{1}{2\sigma_{D,j}^2}\;\chi_D^2\,(\theta,\mathbf{s}_j)\Bigr\}\,\pi(\mathbf{s}_j)\right) \\ & \times \left(\prod_{k=1}^{r} \Bigl(\sigma_{V,k}^{-(n_k+1)}\,\exp\Bigl\{-\frac{1}{2\sigma_{V,k}^2}\,\chi_V^2\,(\theta,\gamma_k)\Bigr\}\,/\gamma_k\right); \end{aligned}$$

in α sind die Theorieparameter $A_i,\ldots,s_{1,j},\ldots,\gamma_k$ zusammengefaßt, in σ die Fehler $\sigma_{V,i}$, $\sigma_{D,j}$, $\sigma_{V,k}$.

Der Hypothesenraum kann auf den makromolekularen Konfigurationsraum reduziert werden:

$$
\begin{aligned}
p(\theta) \;\propto\; & \exp\{-\beta E(\theta)\} \\
& \times \left(\prod_{i=1}^{p} \left| (\mathbf{A}_{J,i}^{T} \mathbf{A}_{J,i} \right|^{-1/2} \left[\mathbf{j}_i{}^{T} (\mathbf{I} - \mathbf{A}_{J,i} \mathbf{A}_{J,I}^{\dagger}) \, \mathbf{j}_i \right]^{-(n_i-3)/2} \right) \\
& \times \left(\prod_{j=1}^{q} \left| \mathbf{A}_{D,j}^{T} \mathbf{A}_{D,j} \right|^{-1/2} \left[\mathbf{d}_j{}^{T} (\mathbf{I} - \mathbf{A}_{D,j} \mathbf{A}_{D,j}^{\dagger}) \, \mathbf{d}_j \right]^{-(n_j-5)/2} \right) \\
& \times \left(\prod_{k=1}^{r} \left[s_k^2(\theta) \right]^{-(n_k-1)/2} \right).
\end{aligned}
$$

Der negative Logarithmus von $p(\theta)$ könnte als Zielfunktion zur Bestimmung der Konformation dienen. In der wahrscheinlichkeitstheoretischen Formulierung wird das Problem der Strukturbestimmung wohl bestimmt: allein die Daten und quantifizierbares Vorwissen legen die Wahrscheinlichkeit $p(\theta, \alpha, \sigma)$ bzw. $p(\theta)$ fest.

3.3 Eine einfache Anwendung

Zur Veranschaulichung wird der Formalismus auf ein einfaches Molekül angewendet. Die Aminosäure Alanin besitzt drei konformationelle Freiheitsgrade: die Dihedralwinkel der Hauptkette φ und ψ, sowie den Dihedralwinkel der Seitenkette χ_1 (siehe Abb. 3.4). Jedoch lediglich die Freiheitgrade des Proteinrückgrats sollen hier variabel sein; χ_1 wird festgehalten.

3.3.1 Ein konformationeller Freiheitsgrad

Zunächst sei der Fall besprochen, daß Messungen aufgenommen wurden, die nur φ, definiert durch C_-, N, C_α und C, betreffen. Jede der hier behandelten Meßgrößen kann seiner Bestimmung dienen: Die skalaren Kopplungskonstanten hängen über die Karplus-Beziehung direkt von φ ab, dipolare Kopplungen über die Orientierung eines Bindungsvektors, dipolare Relaxationsraten über den Abstand zweier Atome.

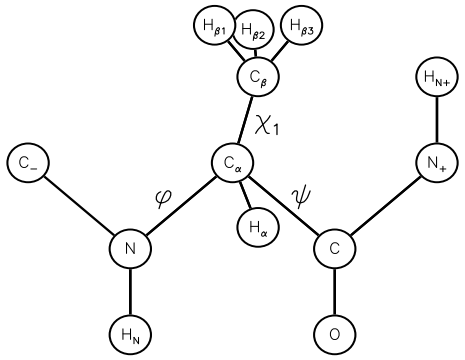


Abbildung 3.4: Konformationelle Freiheitsgrade von Alanin.

Die Daten gehen in der A posteriori-Verteilung durch eine verallgemeinerte χ^2-Funktion ein. Aus n Messungen ${}^3J_i$ der skalaren Kopplungskonstante zwischen den Protonen H_N und H_α folgt (Abschnitt 3.1.1)

$$\chi^2_J(\varphi) = n \left[\left(\overline{J} - {}^3J(\varphi) \right)^2 + s^2_J \right]. \tag{3.25}$$

Die Daten werden in den drei Größen $\overline{J} = \frac{1}{n} \sum_i {}^3J_i$, $s^2_J = \frac{1}{n} \sum_i ({}^3J_i - \overline{J})^2$ und n zusammengefaßt. Der Mittelwert $\overline{J}$ und die Streuung s^2_J sind *hinreichende Statistiken*: Unabhängig von den einzelnen Messungen, sind zwei Datensätze gleichwertig, solange ihre hinreichenden Statistiken dieselben Werte annehmen; die Details der Messungen sind nicht wichtig. Die Anzahl der Daten n bestimmt, wie genau die Parameter geschätzt werden können.

Dipolare Kopplungen, beispielsweise zwischen C_α und H_α, hängen über die Koordinaten des C_α-H_α-Bindungsvektors von φ ab. Messungen dieser Kopplung ergeben

$$\chi^2_D(\varphi) = n \left[\left(\overline{D} - D(\varphi) \right)^2 + s^2_D \right], \tag{3.26}$$

analog zu Messungen der skalaren Kopplungskonstante.

Durch NOESY-Experimente können zusätzlich Abstände zwischen den Protonen H_N, H_α und der Methylgruppe des Kohlenstoffs C_β beobachtet werden. Aus n Messungen des NOE zwischen den Protonen H_N und H_α mit bloß φ-abhängigem Abstand $r(\varphi)$ folgt (Abschnitt 3.1.3)

$$\chi^2_V(\varphi) = n \left[\log^2 \left(\overline{V}\, r^6(\varphi)/\gamma \right) + s^2_V \right] \tag{3.27}$$

mit dem geometrischen Mittel $\overline{V} = (\prod_i V_i)^{1/n}$ und der Streuung der logarithmischen Messungen $s_V^2 = \frac{1}{n}\sum_i \log^2\left(V_i/\overline{V}\right)$ als hinreichenden Statistiken.

Eigentlich hängen χ_J^2, χ_D^2 und χ_V^2 zusätzlich von den Theorieparametern, den Karplus-Koeffizienten, der Saupe-Matrix bzw. dem Kalibrationsfaktor, ab. Weil die Messungen einer Observable y sich in bloß zwei Statistiken, $\overline{y}$ und s_y^2, zusammenfassen lassen, kann entweder der Dihedralwinkel oder einer der Theorieparameter aus einem einzigen Datensatz erschlossen werden. Das Problem ist unterbestimmt. Im Falle der skalaren Kopplungen, beispielsweise, sind die Messungen exakt durch die Wahl $A = B = 0$ und $C = \overline{J}$ erfüllt. Die Statistik $\overline{J}$ wird zur Bestimmung des Theorieparameters aufgebracht und liefert keine neue Information über φ.

Im weiteren seien deshalb die Theorieparameter bekannt. Wird ein einzelner Datensatz analysiert, bleiben als Hypothesenparameter φ und σ, der Fehler der jeweiligen Meßgröße, übrig. Der Hypothesenraum ist zweidimensional und die gemeinsame A posteriori-Verteilung ist

$$p(\varphi,\sigma) \propto \sigma^{-(n+1)} \exp\left\{-\frac{1}{2\sigma^2}\ \chi^2(\varphi) - \beta E(\varphi)\right\}.$$

Wird nun zusätzlich das Vorwissen über die erlaubten Konformationen vernachlässigt ($\beta = 0$), lassen sich die Konturlinien der A posteriori-Verteilung direkt angeben. Sie erfüllen die Gleichung

$$\frac{1}{2\sigma^2}\ \chi^2(\varphi) + (n+1)\log\sigma = \text{const.}$$

mit vorgegebener Konstante größer als

$$\min\{-\log p\} = \frac{n+1}{2}\ \left[\,1 + \log\left(n\,s^2/(n+1)\right)\right].$$

Es wurden synthetische Daten für die Einstellung $\varphi = -60^\circ$ bei vorgegebenen Theorieparametern erzeugt. Für die skalaren Kopplungen war $A = 7\,\text{Hz}$, $B = -1\,\text{Hz}$ und $C = 2\,\text{Hz}$. Die axiale Komponente der mittleren Orientierung betrug $A = -10\,\text{Hz}$, der rhombische Anteil $R = 0.3$. Fünf Messungen je Observable stammen von einer Gauß-Verteilung mit einer Standardabweichung von 2 Hz. NOE-Volumina wurden mit $\gamma = 1$ und $\sigma = 50\,\%$ von einer Lognormal-Verteilung gezogen.

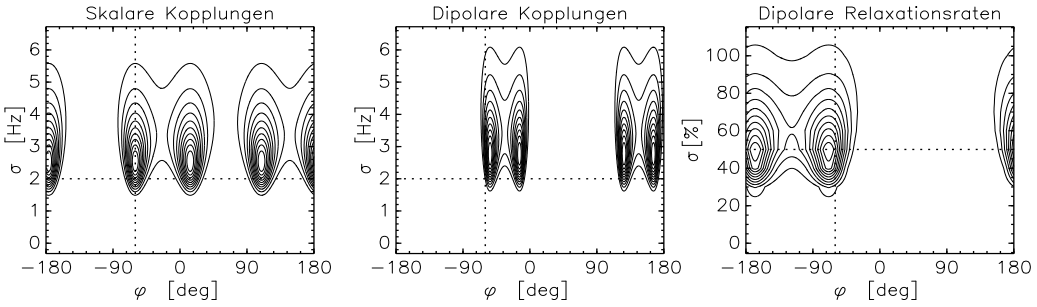


Abbildung 3.5: Konturlinien der A posteriori-Verteilung $p(\varphi, \sigma)$ der verschiedenen Meßgrößen. Die wahren Werte von φ und σ wurden als gepunktete Linien eingezeichnet.

Die Konturlinien der A posteriori-Verteilung $p(\varphi, \sigma)$ sind in Abbildung 3.5 zu sehen. Weil Wahrscheinlichkeiten Unwissen ausdrücken, können sie für beliebige Unbekannte aufgestellt werden; – auch für Größen, die nicht schwanken. Die Abbildung veranschaulicht die Tatsache, daß sowohl der Dihedralwinkel als auch die Fehlerskala unbekannt sind, und daß der Grad des Unwissens über die eine Größe von dem Unwissen über die andere Größe abhängt. Die richtige Darstellung dieser Abhängigkeit ist die Verbundverteilung $p(\varphi, \sigma)$. Ohne eine wahrscheinlichkeitstheoretische Behandlung ist die konsistente Ableitung einer Funktion, die den Fehler und die Koordinaten gemeinsam bewertet und so beide Größen verbindet, nicht möglich.

Wahrscheinlichste Werte $\hat{\varphi}$ und $\hat{\sigma}$ erhält man durch Minimierung von $-\log p(\varphi, \sigma)$. Weil die Meßgrößen durch mehrdeutige Theorien beschrieben werden, ist die A posteriori-Verteilung multimodal in φ. Die wahrscheinlichsten Werte $\hat{\varphi}$ der A posteriori-Verteilung minimieren χ^2 (falls $\beta = 0$). Für jede der drei Meßgrößen $y = J, D, \log V$ ist dies der Fall bei $y(\hat{\varphi}) = \overline{y}$. Die Verhältnisse sind in Abbildung 3.6 dargestellt. Aus der Mehrdeutigkeit des theoretischen Zusammenhangs zwischen der Observable und dem Dihedralwinkel folgt, daß es mehrere Einstellungen mit maximaler Wahrscheinlichkeit gibt.

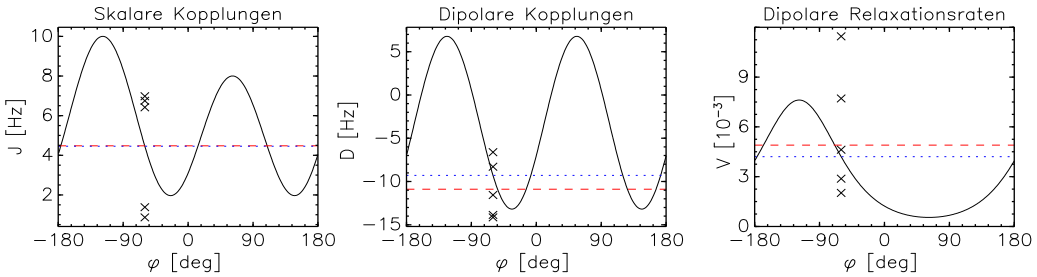


Abbildung 3.6: Theoretischer Zusammenhang zwischen Observable und Dihedralwinkel $y(\varphi)$ (schwarze Linie). Der ideale Meßwert, der sich aus dem wahren Dihedralwinkel ($\varphi = -60°$) ergibt, wurde als blaue gepunktete Linie eingetragen. Von diesem weicht die aus den Messungen berechnete Statistik, der Mittelwert, mehr oder weniger stark ab (rote Linie). Die Messungen sind als schwarze Kreuze über dem wahren Wert von φ eingezeichnet. Die Schätzwerte sind die Schnittpunkte der Linien $y = y(\varphi)$ und $y = \overline{y}$.

3.3.2 Reduzierung des Hypothesenraums

Der Hypothesenraum kann durch Mittelung über den Fehler auf den Konformationsraum reduziert werden. Die reduzierte Likelihood-Funktion ist:

$$L(\varphi) = \int \mathrm{d}\sigma \; L(\varphi, \sigma)\, \pi(\sigma) \; \propto \; \left[\, \chi^2(\varphi) \,\right]^{-n/2} .$$

Durch Multiplikation mit der konformationellen A priori-Verteilung erhält man die marginale A posteriori-Verteilung $p(\varphi) \propto L(\varphi)\, \pi(\varphi)$. Diese Wahrscheinlichkeit drückt aus, was sich anhand der Daten über den unbekannten Winkel sagen läßt, wenn jegliches Vorwissen über den Fehler ignoriert wird. Die Maxima der reduzierten Likelihood-Funktion $L(\varphi)$ liegen bei denselben Werten $\hat{\varphi}$ wie bei der gemeinsamen Likelihood-Funktion $L(\varphi, \sigma)$; nur sind sie weniger scharf ausgeprägt.

Die Aussagen werden ungenauer, wenn weniger Vorwissen in Betracht gezogen wird: Abbildung 3.7 zeigt die marginale und die bedingte A posteriori-Verteilung des Dihedralwinkels, in der die Fehler auf die wahren Werte gesetzt wurden. Dies entspricht dem unrealistischen Fall richtigen Vorwissens über den Fehler. Die marginale A posteriori-Verteilung drückt hingegen völli-

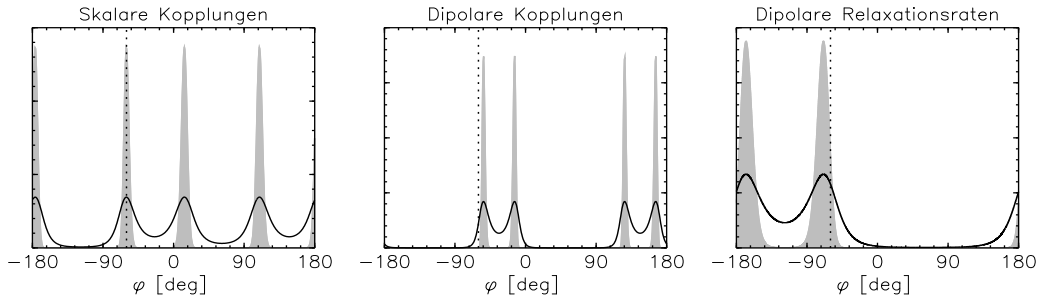


Abbildung 3.7: Vergleich der marginalen A posteriori-Verteilung $p(\varphi)$ (schwarz) und der bedingten A posteriori-Verteilung Verteilung $p(\varphi|\sigma)$ (grau) für $\beta = 0$. In der bedingten A posteriori-Verteilung wurde der Fehler σ auf den wahren Wert gesetzt. Der wahre Wert des Dihedralwinkels ist zusätzlich als horizontale gepunktete Linie eingezeichnet.

ges Unwissen über den Fehler aus. Die Aussagen, die sich aus der bedingten A posteriori-Verteilung ergeben, sind zwar genauer, aber können systematisch falsch sein. So ist im Falle der dipolaren Kopplungen das scharf ausgeprägte Maximum von $p(\varphi|\sigma)$ vom wahren Wert verschoben.

Daß die Berücksichtigung zusätzlicher Information zu genaueren Aussagen führt, zeigt auch die Abhängigkeit der marginalen A posteriori-Verteilung $p(\varphi)$ von der Größe des Datensatzes. Mit zunehmender Datenzahl werden die Verteilungen immer schmaler (Abb. 3.8). Im Grenzfall unendlich vieler Messungen, ist die A posteriori-Verteilung eine Summe von δ-Funktionen. Die Mehrdeutigkeit der Theorie wird freilich selbst durch unendlich oft wiederholte Messung nicht aufgelöst. Dem kann erst entweder die Berücksichtigung von Vorkenntnissen oder die Messung anderer Größen Abhilfe schaffen.

Das Vorwissen über physikalisch zulässige Konformationen des Moleküls löst die Mehrdeutigkeit in φ teilweise auf. Bei Raumtemperatur ist das Boltzmann-Ensemble eine nahezu kastenförmige Verteilung: die van der Waals-Kräfte verhindern, daß sich die Atome durchdringen; die inter-atomaren Abstände können nur in einen beschränkten Bereich fallen. Der Dihedralwinkel darf deshalb a priori nur Werte in einem bestimmten Intervall annehmen. Der Bayes'sche Satz schreibt vor, daß dieses Vorwissen mit den Daten kom-

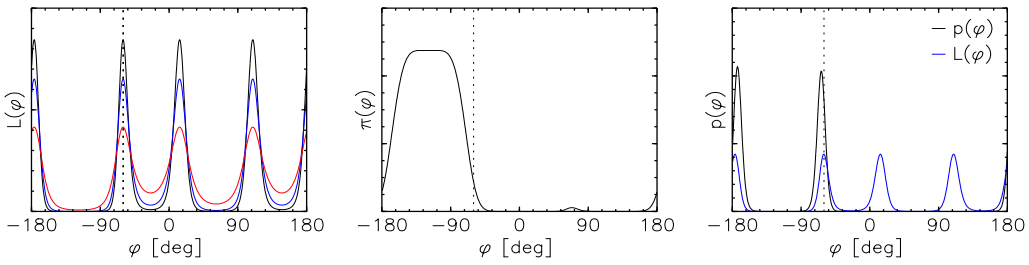


Abbildung 3.8: Links: Reduzierte Likelihood-Funktionen bei verschieden häufigen Messungen der skalaren Kopplungskonstante, wobei sich die Datensätze nicht in ihren hinreichenden Statistiken, sondern bloß in der Anzahl der Messungen unterscheiden: $n = 5, 10, 15$ (rot, blau, schwarz). Mit zunehmender Größe des Datensatzes werden die Verteilungen schmaler und lassen genauere Aussagen zu. Dennoch bleibt die Mehrdeutigkeit der Likelihood-Funktion bestehen. Erst die Berücksichtigung physikalischen Vorwissens, ausgedrückt durch das Boltzmann-Ensemble $\pi(\varphi)$ (Mitte), ermöglicht, die Mehrdeutigkeit teilweise aufzuheben (rechts).

biniert werden muß, indem ihre Wahrscheinlichkeiten multipliziert werden. Weil die A priori-Verteilung hier nahezu kastenförmig ist, bleiben von der Likelihood-Funktion $L(\varphi)$ nur die beiden Maxima übrig, die in den zulässigen Bereich fallen (Abb. 3.8).

3.3.3 Analyse mehrerer Datensätze

Falls mehrere Datensätze D_j, bestehend aus n_j Messungen, vorliegen, ist die A posteriori-Verteilung

$$p(\varphi, \{\sigma_j\}) \propto \exp\Big\{-\beta\, E(\varphi)\Big\} \Big(\prod_j \sigma_j^{-(n_j+1)} \exp\Big\{-\frac{1}{2\sigma_j^2}\,\chi_j^2(\varphi)\Big\}\Big).$$

Die marginale A posteriori-Verteilung von φ erhält man durch Integration über die Fehler σ_j:

$$p(\varphi) \;\propto\; \exp\Big\{-\beta\, E(\varphi)\Big\} \prod_j \big[\,\chi_j^2(\varphi)\,\big]^{-n_j/2}.$$

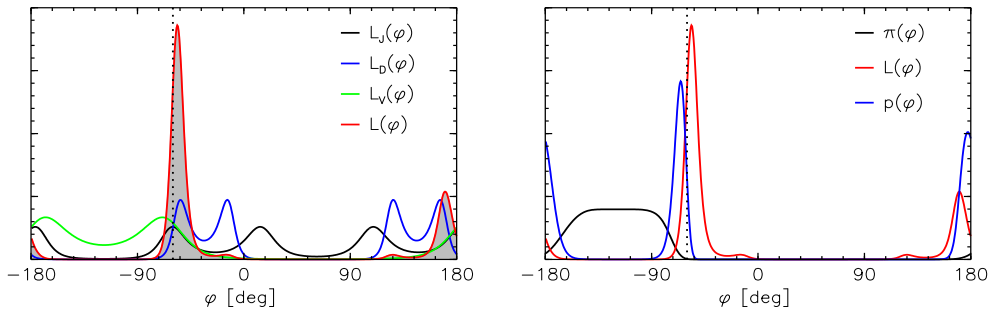


Abbildung 3.9: Die Mehrdeutigkeit der Daten wird durch ihre gemeinsame Verwendung bis zu einem gewissen Grad aufgehoben. Links: Reduzierte Likelihood-Funktionen $L_j(\varphi)$, die sich aus den einzelnen Datensätzen ergeben, und das Produkt $L(\varphi)$ aller Faktoren. Rechts: die gemeinsame Likelihood-Funktion $L(\varphi)$ ergibt durch Multiplikation mit der A priori-Verteilung $\pi(\varphi)$ die A posteriori-Verteilung $p(\varphi)$.

Die Daten lassen sich in der gemeinsamen Likelihood-Funktion

$$L(\varphi) = \prod_j L_j(\varphi) \quad \text{mit} \quad L_j(\varphi) = \left[\chi_j^2(\varphi)\right]^{-n_j/2}$$

zusammenfassen. Der Einfluß eines Datensatzes hängt von der Anzahl der Messungen ab.

In einer gemeinsamen Analyse der künstlichen Datensätze ist die reduzierte Likelihood-Funktion $L(\varphi)$ ein Produkt der bereits gezeigten, einzelnen Verteilungen. In Abbildung 3.9 sind die aus den jeweiligen Datensätzen resultierenden Likelihood-Faktoren und die gemeinsame Likelihood-Funktion gezeigt. Die gemeinsame Verwendung aller Daten ermöglicht, die Mehrdeutigkeit der einzelnen Verteilungen beinah vollständig aufzulösen: Die Likelihood-Funktion $L(\varphi)$ ist zwar noch immer multimodal, das Nebenmaximum ist jedoch bloß ein Viertel so stark bevölkert wie die Hauptmode nahe dem wahren Wert $-60\,°$. Zusätzliche Berücksichtigung von Vorwissen, repräsentiert durch $\pi(\varphi)$, verbessert das Ergebnis in diesem Fall nicht.

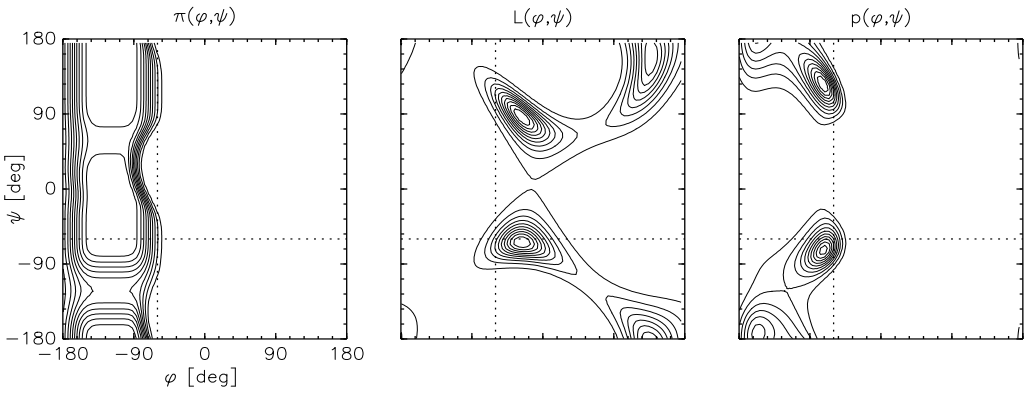


Abbildung 3.10: Konturlinien der konformationellen A priori-Verteilung $\pi(\varphi, \varphi)$, der Likelihood-Funktion $L(\varphi, \sigma)$ und der A posteriori-Verteilung $p(\varphi, \psi)$ der beiden Dihedralwinkel φ und ψ von Alanin.

3.3.4 Zwei konformationelle Freiheitsgrade

Der Dihedralwinkel ψ, definiert durch N, C_α, C und N_+, sei nun ein weiterer konformationeller Freiheitsgrad. Er kann durch Messungen des NOEs zwischen H_N und dem Amidproton H_{N+} einer benachbarten Aminosäure bestimmt werden. Aus Messungen des NOEs zwischen H_N und H_α sowie zwischen H_N und H_{N+} ergibt sich:

$$\chi^2(\varphi, \psi) = \chi^2_{H_N H_\alpha}(\varphi) + \chi^2_{H_N H_{N+}}(\varphi, \psi),$$

wobei n_{kl} Messungen des NOEs zwischen Atom k und Atom l zu dem bekannten Beitrag

$$\chi^2_{kl} = n_{kl} \left[\log^2(\overline{V}_{kl}\, r^6_{kl}) + s^2_{kl} \right]$$

führen. Die marginale konformationelle A posteriori-Verteilung erhält man durch Integration über die Fehlerskala

$$p(\varphi, \psi) = \int \mathrm{d}\sigma \; p(\varphi, \psi, \sigma) \;\propto\; \exp\Big\{ -\beta\, E(\varphi, \psi) \Big\} \; \left[\chi^2(\varphi, \psi) \right]^{-n/2},$$

n ist die Anzahl aller NOE-Messungen.

In Abbildung 3.10 sind die Konturlinien der konformationellen A priori-Verteilung $\pi(\varphi, \psi)$, der Likelihood-Funktion $L(\varphi, \psi)$ und der A posteriori-Verteilung $p(\varphi, \psi)$ zu sehen. Es wurden je fünf Messungen bei $\varphi = -60\,^\circ$ und

$\psi = -60\,^{\circ}$ von einer Lognormal-Verteilung mit einem Fehler $\sigma = 0.75$ gezogen. Die A priori-Verteilung ist die Ramachandran-Karte der Dihedralwinkel φ und ψ, die das Kraftfeld (2.15) impliziert. Die van der Waals-Wechselwirkungen teilen den Konformationsraum in erlaubte und verbotene Bereiche ein. Die Likelihood-Funktion ist multimodal und hat viel Wahrscheinlichkeitsmasse im verbotenen Bereich. Dennoch erlaubt der Bayes'sche Satz, einigermaßen genaue und richtige Aussagen zu machen: Durch Multiplikation der Likelihood-Funktion mit der A priori-Verteilung werden Bereiche ausgeblendet, die in einer der beiden Verteilungen oder in beiden unwahrscheinlich sind. Das Zusammenwirken von Vorwissen und Daten grenzt die gesuchte Konformation auf einen kleinen Bereich ein.

3.4 Strukturberechnung durch Monte-Carlo-Simulation

In realistischen Fällen sind die Betrachtungen des vorangegangenen Abschnitts nicht mehr möglich; es müssen dann die in Abschnitt 2.3 eingeführten Monte-Carlo-Methoden verwendet werden, um die A posteriori-Verteilungen zu analysieren. Zuerst sollen die Komponenten des Monte-Carlo-Algorithmus am Beispiel von Alanin illustiert werden. Der volle Replica-Algorithmus wird dann an einem realistischen System erläutert.

3.4.1 Gibbs sampling und Hybrid-Monte-Carlo

Nach dem Gibbs sampling-Schema (2.20) müssen abwechselnd die bedingten A posteriori-Verteilungen aller Hypothesenparameter simuliert werden. Diese wurden in Abschnitt 3.2 abgeleitet und sind im Falle der nuisance parameter einfache univariate Verteilungen, für die es Zufallszahlengeneratoren gibt: Die inversen quadratischen Fehler $\lambda_i = \sigma_i^{-2}$ folgen einer Gamma-Verteilung ((3.13), (3.18), (3.23)). Die Karplus-Koeffizienten und die Elemente der Saupe-Matrix sind gaußverteilt ((3.12) bzw. (3.17)). Der Kalibrations-

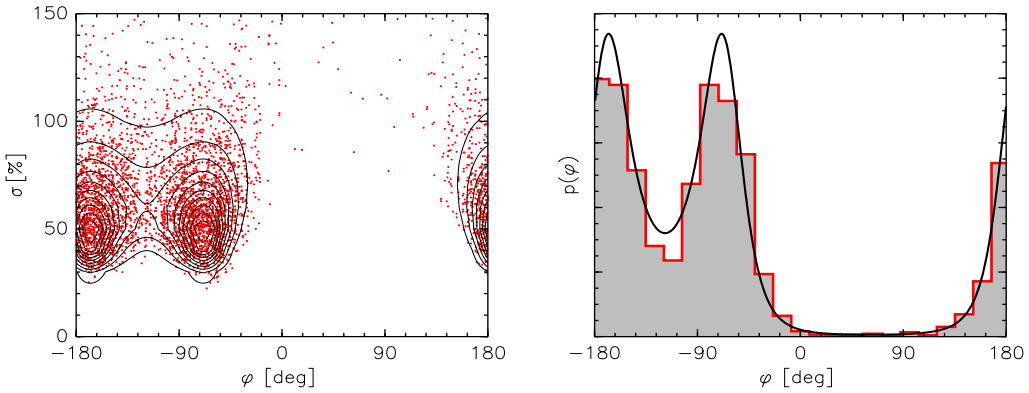


Abbildung 3.11: Konturlinien $p(\varphi, \sigma) = \text{const.}$ der A posteriori-Verteilung, die aus NOESY-Resonanzen resultiert (siehe auch Abb. 3.5), und mit dem Gibbs-Schema erzeugte Stichproben (rot). Rechts ist die marginale A posteriori-Verteilung des Dihedralwinkels gezeigt mit einem Histogramm der Stichproben $\varphi^{(k)}$.

faktor gehorcht der Lognormal-Verteilung (3.22).

Gibbs sampling läßt sich anhand des Beispiels aus Abschnitt 3.3 illustrieren. Für Messungen des H_N-H_α-NOEs sind die bedingten A posteriori-Verteilungen des Dihedralwinkels φ und des Fehlers σ:

$$p(\varphi|\sigma) \;\propto\; \exp\left\{-\frac{n}{2\sigma^2}\,\log^2(\overline{V}\,r^6(\varphi))\right\}$$

bzw.

$$p(\sigma|\varphi) \;\propto\; \sigma^{-(n+1)}\,\exp\left\{-\frac{1}{2\sigma^2}\,\chi^2(\varphi)\right\}.$$

Der inverse quadratische Fehler folgt einer Gamma-Verteilung, für die es Zufallszahlengeneratoren gibt. Ein einfaches Schema, um Stichproben von $p(\varphi|\sigma)$ zu ziehen, ist, einen Kandidaten φ' gleichverteilt zwischen -180° und 180° zu wählen und gemäß dem Metropolis-Kriterium [29] mit der Wahrscheinlichkeit

$$\min\Big\{1, \exp\{-[\chi^2(\varphi') - \chi^2(\varphi)]\,/\,(2\sigma^2)\}\Big\}$$

als neue Stichprobe zu akzeptieren. Die mittels Gibbs sampling gezogenen Stichproben von $p(\varphi, \sigma)$ sind in 3.11 abgebildet. Die marginale A posteriori-Verteilung $p(\varphi)$ stimmt mit dem Histogramm der Stichproben $\varphi^{(k)}$ überein:

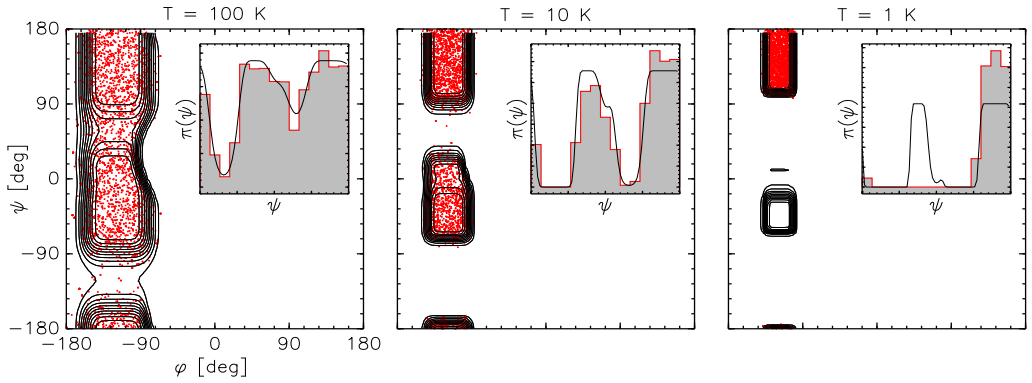


Abbildung 3.12: Simulation des Boltzmann-Ensembles $\pi(\varphi, \psi)$ für verschiedene Temperaturen. Die Bildeinsätze zeigen die marginale Verteilung $\pi(\psi)$ und das Histogramm der ψ-Stichproben.

Man erhält marginale Verteilungen aus Stichproben aller Hypothesenparameter, indem man nur die interessierenden betrachtet und integriert so numerisch über die uninteressanten Parameter.

Hybrid-Monte-Carlo eignet sich, um die Moden einer Verteilung korrelierter Parameter abzusuchen. Bei stark zerklüfteten Verteilungen treten jedoch Probleme auf, wie sich am Beispiel der A priori-Verteilung $\pi(\varphi, \psi)$ von Alanin bei verschiedenen Temperaturen zeigt (siehe Abb. 3.12): Mit abnehmender Temperatur wird das Boltzmann-Ensemble immer schroffer und HMC ist immer weniger in der Lage, die Verteilung vollständig abzusuchen. Die Markov-Kette bleibt schließlich in einem Maximum von $p(\varphi, \psi)$ stecken. Mit abnehmender Temperatur stimmen die marginale A posteriori-Verteilung $p(\psi)$ und das Histogramm der ψ-Stichproben immer schlechter überein.

3.4.2 Demonstration des Replica-Algorithmus

Die vorangegangene Anwendung verdeutlicht die Schwierigkeiten, die bei Verwendung von Hybrid-Monte-Carlo zur Simulation der Dihedralwinkel auftreten können. Diese Schwierigkeiten werden durch den Replica-Algorithmus (Abschnitt 2.3.4) gelöst. Der Replica-Algorithmus wird exemplarisch zur Simulation der A posteriori-Verteilung angewendet, die sich aus NOESY-

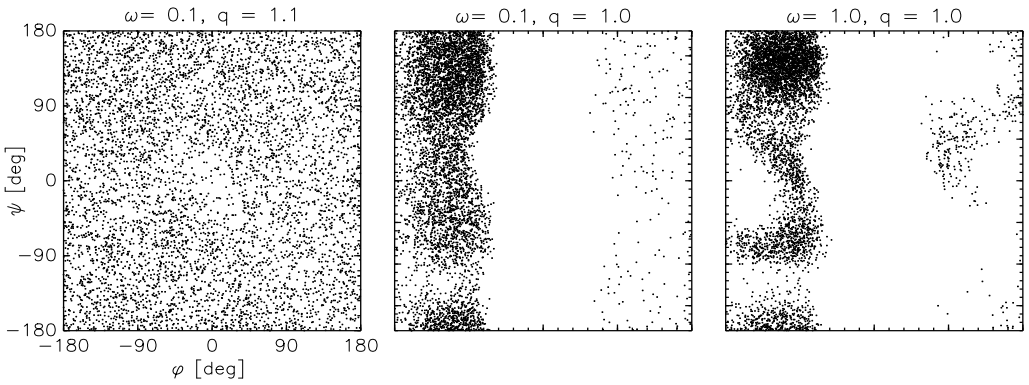


Abbildung 3.13: Stichproben der Dihedralwinkel φ und ψ der Haupkette des Polypeptids für verschiedene Wärmebäder.

Messungen an der Proteindomäne SH3 ergibt (siehe Abschnitt 3.5.1). Die Modellierung der NOESY-Volumina ist in Abschnitt 3.1.3 beschrieben. Die zusätzlichen Parameter γ und σ können analytisch entfernt werden. In der marginalen A posteriori-Verteilung (3.24) werden die Konformationen durch 275 Dihedralwinkel parametrisiert. Zur Simulation wurde die Strategie des vorigen Abschnitts verwendet; ω wurde von 1.0 auf 0.1 herabgesenkt, q von 1.0 auf 1.1 erhöht. 50 Wärmebäder wurden parallel simuliert. In den ersten 23 replicas wird λ mit potenziell von 1.0 auf 0.1 abgesenkt, während q auf 1.0 gesetzt ist: $\lambda_i = [\,1.0 - 0.043\,(i-1)\,]^{0.8}$, $q_i = 1.0$ für $i = 1, \ldots, 22$. In den 27 verbleibenden replicas bleibt λ auf seinem Minimalwert, während q exponentiell von 1.0 auf 1.1 erhöht wird: $\lambda_i = 0.1$, $q_i = 0.993 + 0.007\,\exp(0.1\,(i-23))$ für $i = 23, \ldots, 50$.

Daß eine Simulation nach einer endlichen Anzahl von Schritten wirklich konvergiert ist, kann strenggenommen nicht nachgewiesen werden, jedoch gibt es dafür mehrere Hinweise. In Abbildung 3.13 sind alle Dihedralwinkel der Hauptkette, φ_i und ψ_i, eingetragen. Im Hochtemperaturbad sind die Dihedralwinkel nahezu gleichverteilt. Am Übergang ($\omega = 0.1, q = 1.0$) gehorchen sie dem Boltzmann-Ensemble (vgl. Abb. 3.10). Erst die A posteriori-Verteilung ($\omega = 1.0, q = 1.0$) erzeugt Stichproben, die um die Dihedralwinkel der Kristallstruktur streuen.

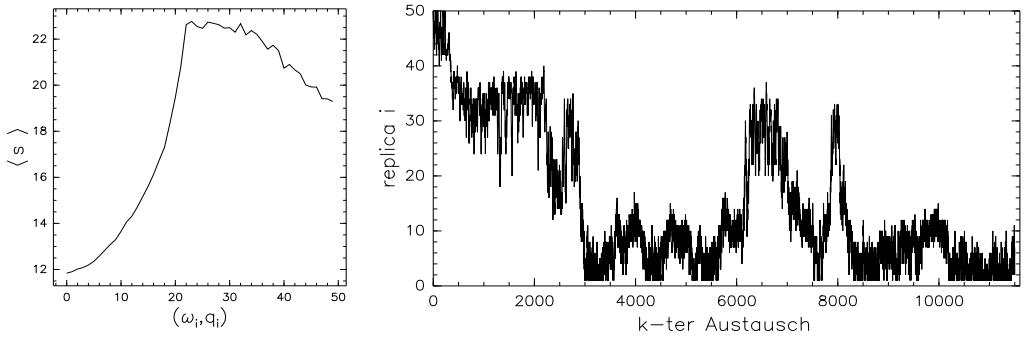


Abbildung 3.14: Mittlerer Gyrationsradius $\langle s \rangle$ in den Wärmebädern der Replica-Kette (links). Diffusion einer Konformation durch die verschiedenen Wärmebäder im Laufe einer Replica-Simulation (rechts).

In Abbildung 3.14 ist der mittlere Gyrations-Radius für alle Wärmebäder aufgetragen. In der ersten Hälfte wird der Einfluß der Daten immer geringer; die Konformationen werden loser, weil das Boltzmann-Ensemble wegen der Vernachlässigung des Lösungsmittels und wegen der Vereinfachung der potentiellen Energie globuläre Strukturen nicht bevorzugt. Am Übergang der Kette ist der Gyrationradius maximal. Die Daten sind nun fast ausgeschaltet, van der Waals-Abstoßungen verhindern jedoch, daß sich die Polypeptidkette zusammenfaltet. Anschließendes Vergrößern von q schaltet die van der Waals-Kräfte aus. Die Konformationen werden wieder kompakter, weil der verbleibende Anteil der Daten die Polypeptidkette stärker zusammenziehen kann.

Desweiteren ist die Diffusion einer Konformation durch die verschiedenen Wärmebäder gezeigt (Abb. 3.14). Beginnend im Hochtemperaturbad erreicht die Konformation schließlich die Zielverteilung. Die Konvergenz und Durchmischung der Konformationen wird von den Austauschraten zwischen benachbarten Wärmebädern bestimmt. Die Austauschrate hängt davon ab, wie sehr sich die Verteilungen überdecken. Dies ist der Fall, wenn die Verteilungen der χ^2 bei benachbarten Wärmebädern ungefähr übereinstimmen. Dasselbe gilt für E_q in dem Abschnitt, in dem q variiert wird. Die Hypothesenparameter gehen nur über diese „Makrovariablen“ in die Verteilungen

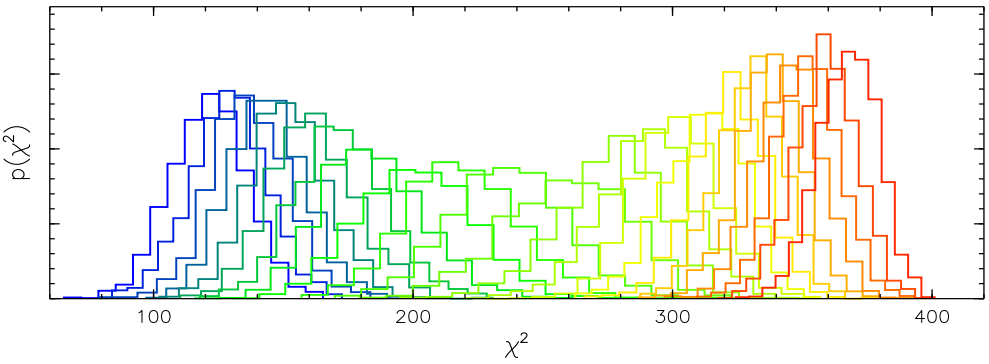


Abbildung 3.15: Histogramme der χ^2-Werte: Die Wärmebäder, in denen ω variiert wird, überdecken sich sehr gut. Die Farbe repräsentiert den Wert des Replica-Parameters ω: Blau entspricht einer „kalten“ Verteilung, d.h. großen ω; zu Rot hin werden die Verteilungen immer „heißer“, d.h. durch Absenken von ω werden die Daten immer mehr ausgeschaltet und die Polypeptidkette kann sich immer freier bewegen.

ein; Zustände mit gleichen Werten in χ^2 und E_q sind nicht unterscheidbar. Exemplarisch sind in Abbildung 3.15 die Histogramme der χ^2-Werte für die Wärmebäder, in denen ω variiert wurde, zu sehen.

3.4.3 Vergleich mit Optimierung

Strukturberechnung durch Optimierung kann probabilistisch als Punktschätzung gedeutet werden: Die Hybridenergie ist der negative Logarithmus einer bedingten A posteriori-Verteilung mit bekannten Theorieparametern und Fehlern (siehe Abschnitt 2.2.7). Neben der unrealistischen Voraussetzung, daß die zusätzlichen Parameter vor der Strukturberechnung bekannt sind, weist die Rechenmethode selbst Mängel auf.

Im Idealfall würde Optimierung das globale Minimum der Hybridenergie, also die wahrscheinlichste Struktur, finden. Begnügen wir uns hiermit, so vernachlässigen wir den Großteil der in der A posteriori-Verteilung enthaltenen Information. Zum einen wissen wir mit Angabe der „besten“ Struktur allein nicht, wie genau diese von den Daten festgelegt wird. Zum anderen können

weitere Strukturen existieren, die die Daten annähernd gut erfüllen, aber bloß lokale Minima sind. Diese Information ist nur in der A posteriori-Verteilung enthalten.

Üblicherweise versucht man diesen Mangel zu beheben, indem man die Lösung eines Strukturbestimmungsproblems durch ein „Ensemble“, d.h. eine Menge von Lösungsmöglichkeiten, darstellt. Ein solches Ensemble erhält man durch wiederholte Minimierung bei zufälliger Variation der Startbedingungen. Die Bezeichnung „Ensemble“ suggeriert eine statistische Deutung. Und tatsächlich werden NMR-Ensemble ähnlich statistischen Ziehungen verwendet, d.h. Mittelwerte und Varianzen von Strukturvariablen berechnet.

Ein solches Vorgehen ist irreführend: Das traditionelle NMR-Ensemble hat keine saubere statistische Grundlage; ein Mittel über ein NMR-Ensemble ist eben nur ein Mittel, aber kein statistischer Schätzwert wie im Falle der Monte-Carlo-Stichproben (2.17). Obwohl die Idee einer Verteilung der Lösungsmöglichkeiten anklingt, können traditionelle Zugänge diese Verteilung nicht angeben.

In der Bayes'schen Strukturberechnung ist die explizite Angabe des Ensembles möglich und nicht in einer Vorschrift wie „Starte den Minimierungsalgorithmus zwanzigmal!“ versteckt: Es ist die A posteriori-Verteilung aller Unbekannten, die in der Modellierung vorkommen. Wie dieses Ensemble praktisch erzeugt wird, ist ein algorithmisches Problem und strikt von der grundsätzlichen Problembehandlung getrennt.

Dagegen ist das traditionelle NMR-Ensemble nicht eindeutig definiert. Es gibt keine anerkannten Kriterien für die Größe des Ensembles und die Selektion der in ihm enthaltenen Konformationen [17, 48]. Die zusätzlichen Parameter α und σ müssen vorab gewählt werden, das Ensemble hängt aber maßgeblich von ihnen ab. Mehrmalige Minimierung bei zufälligen Startbedingungen reproduziert weder die Multiplizitäten der Moden der A posteriori-Verteilung noch sucht es einzelne Moden ergodisch ab. Aussagen über die Genauigkeit der berechneten Struktur, die auf der Analyse eines NMR-Ensembles beruhen, werden von all diesen Faktoren abhängen.

Nur die Bestimmung der wahrscheinlichsten Struktur führt auf ein Optimierungsproblem; weit allgemeiner ist die Aufgabe, über die möglichen Werte der Hypothesenparameter zu summieren, wobei die A posteriori-Verteilung als Gewicht in die Summe eingeht. Diese Regel folgt aus der wahrscheinlichkeitstheoretischen Formulierung: Wenn etwas nicht sicher bekannt ist, muß diese Unsicherheit bei der Bewertung von Hypothesen berücksichtigt werden. Weil wir aus den Messungen die Struktur nicht mit vollständiger Gewißheit schließen können, müssen wir den ganzen Konformationsraum in Betracht ziehen.

Die beschränkte Anwendbarkeit von Optimierungsverfahren tritt auch bei Transformationen der Hypothesenparameter zutage. Falls wir von einer Parametrisierung θ der Struktur zu einer anderen $\eta = T(\theta)$ übergehen, muß dieser Wechsel in der Wahrscheinlichkeit berücksichtigt werden. Der negative Logarithmus der A posteriori-Verteilung $\tilde{p}(\eta|\alpha,\sigma)$ weist in den neuen Variablen einen zusätzlichen Term auf (2.4):

$$-\log\tilde{p}(\eta|\alpha,\sigma) = -\log p(T^{-1}(\eta)|\alpha,\sigma) - \log\left|\frac{\partial T^{-1}}{\partial\eta}\right|.$$

Im allgemeinen stimmen die Minima $\hat{\eta}$ der neuen Zielfunktion $-\log\tilde{p}(\eta|\alpha,\sigma)$ nicht mit den transformierten Minima $\hat{\theta}$ der ursprünglichen Zielfunktion $-\log p(\theta|\alpha,\sigma)$ überein: $\hat{\eta} \neq T(\hat{\theta})$. Dagegen können Stichproben direkt transformiert werden.

Die Mängel von Optimierungsverfahren werden besonders bei gleichverteilten Hypothesenparametern augenfällig. Für jede Verteilung kann eine Parametrisierung gefunden werden, so daß die neuen Variablen über einen multidimensionalen Einheitswürfel gleichverteilt sind [49]. Optimierung ist in diesen Variablen sinnlos: die Hybridenergie ist ein Kastenpotential.

Die Dihedralwinkel φ und ψ von Alanin können auf gleichverteilte Variablen x und y zwischen Null und Eins abgebildet werden:

$$x = f(\varphi) = \int_{-180}^{\varphi} \mathrm{d}\varphi'\; p(\varphi'), \quad y = g(\varphi,\psi) = \int_{-180}^{\psi} \mathrm{d}\psi'\; p(\psi'|\varphi).$$

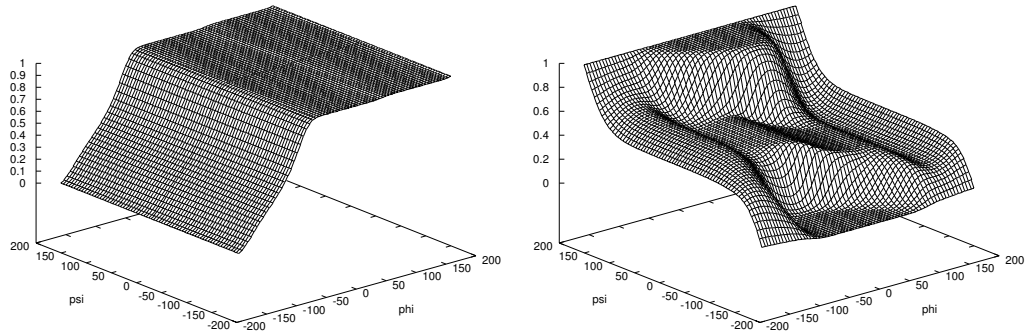


Abbildung 3.16: Transformation der Dihedralwinkel φ und ψ auf neue, gleichverteilte Variablen $x = f(\varphi)$ (links) und $y = g(\varphi, \psi)$ (rechts), $(x, y) \in [0, 1]^2$.

Diese Transformationen lassen sich zwar nicht mehr analytisch berechnen, jedoch in diesem einfachen Beispiel tabellieren (siehe Abb. 3.16). In den neuen Variablen ist die A posteriori-Verteilung $p(x, y)$ flach; Optimierung ist nicht mehr zur Bestimmung eines Schwätzwerts anwendbar; die gesamte Information ist in $f(\varphi)$ und $g(\varphi, \psi)$ enthalten.

Werden nun Stichproben von x und y, die im zweidimensionalen Einheitswürfel gleichverteilt sind, durch Umkehrung der Funktionstabelle wieder auf Dihedralwinkel transformiert

$$\varphi^{(k)} = f^{-1}\big(x^{(k)}\big), \quad \psi^{(k)} = g^{-1}\big(x^{(k)}, y^{(k)}\big),$$

so erhält man Stichproben der Dihedralwinkel, die die ursprüngliche A posteriori-Verteilung $p(\varphi, \psi)$ gut reproduzieren (Abb. 3.17). Dies illustriert auch die Tatsache, daß die Transformation eines Ensembles von Stichproben das richtige Ensemble in den neuen Variablen erzeugt, wohingegen einzelne Schätzwerte nicht direkt transformiert werden können.

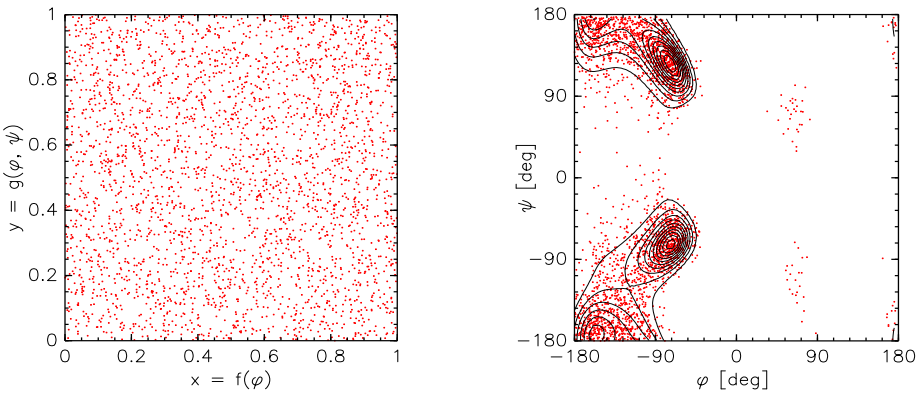


Abbildung 3.17: Durch Umkehrung des Zusammenhangs $x = f(\varphi)$, $y = g(\varphi, \psi)$ lassen sich aus Stichproben $\left(x^{(k)}, y^{(k)}\right)$ Stichproben (links) $\left(\varphi^{(k)}, \psi^{(k)}\right)$ der Dihedralwinkel (rechts) berechnen, die gut mit den Konturlinien der marginalen A posteriori-Verteilung $p(\varphi, \psi)$ aus Abb. 3.10 übereinstimmen.

3.5 Analyse experimenteller Datensätze

Die in den Kapiteln 3.1 und 3.2 entwickelten Modelle werden auf reale Daten angewendet. Mit Hilfe des ISD-Programmpakets (siehe Abschnitt 2.5) habe ich NMR-Messungen an zwei Proteinen analysiert.

3.5.1 Analyse der SH3-Daten

Eine Beschreibung der Daten findet sich in Abschnitt 2.4.1; zur Simulation der A posteriori-Verteilung dient das Replica-Schema. Das Gewicht ω der Daten in der Likelihood-Funktion variierte zwischen 0.1 und 1.0; der q-Parameter des Tsallis-Ensembles zwischen 1.0 und 1.1 (weitere Details finden sich in Abschnitt 3.4.2). Es wurden drei verschiedene Simulationen durchgeführt:

1. Simulation von $p(\theta, \gamma, \sigma)$: Hypothesenparameter sind die 275 Dihedralwinkel θ_i, der Kalibrationsfaktor γ und der Fehler σ;

2. Simulation von $p(\theta, \sigma|\gamma = 1)$: die kalibrierten Distanzen wurden direkt verwendet, also $\gamma = 1$ gesetzt und nicht geschätzt; die Hypothesenpa-

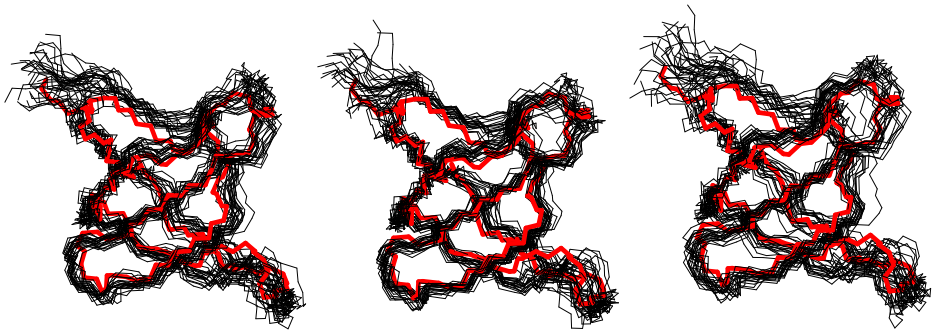


Abbildung 3.18: Hauptketten der zwanzig wahrscheinlichsten Konformationen in den drei Simulationen (links: Simulation 1, Mitte: Simulation 2, rechts: Simulation 3). In Rot wurde die Kristallstruktur eingezeichnet. Zur Überlagerung und Darstellung der Konformationen wurde das Progamm MOLMOL verwendet.

rameter sind θ_i und σ;

3. Simulation von $p(\theta)$: die einzigen Hypothesenparameter sind die Dihedralwinkel θ_i, γ und σ wurden analytisch entfernt (siehe Abschnitt 3.2.3).

Trotz der geringen Anzahl von Messungen ist die Konformation überraschend gut bestimmt (siehe Abb. 3.18). Die zwanzig wahrscheinlichsten Konformationen jeder Simulation wurden überlagert und gezeichnet. Es sind keine wesentlichen Unterschiede zwischen den Strukturen zu erkennen.

Gemessen an der geringen Dichte des Datensatzes sind die berechneten Konformationen der Kristallstruktur [36] (PDB-Code 1shf) recht ähnlich. Die Verteilung der RMSD-Werte (Abb. 3.19) der zweiten Simulation ist systematisch zu niedrigeren Werten verschoben; dies ist zu erwarten, weil in die Berechnung zusätzliches Wissen ($\gamma = 1$) eingeflossen ist, welches in den anderen beiden Simulationen nicht verwendet wurde. Die Verteilungen der Konformationen in der ersten und in der dritten Simulation sollten übereinstimmen: Monte-Carlo-Simulation der A posteriori-Verteilung $p(\theta, \gamma, \sigma)$ integriert numerisch über die zusätzlichen Parameter γ und σ, welche in Simulation 3 analytisch entfernt wurden. Dementsprechend decken sich ihre

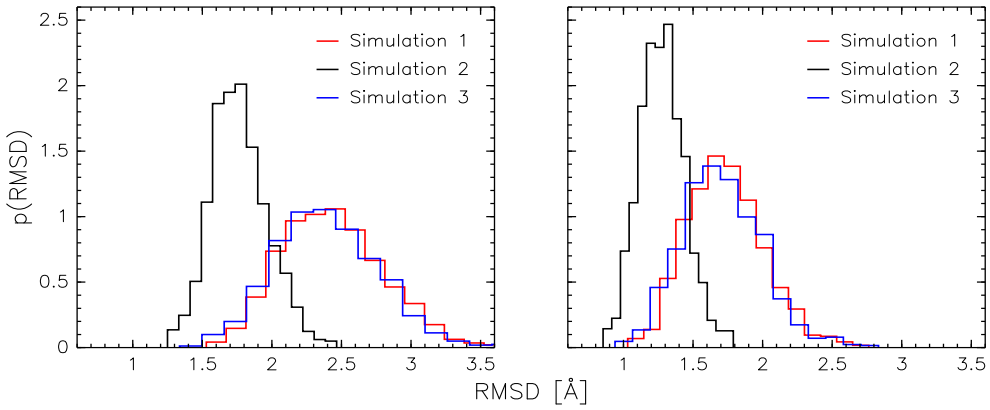


Abbildung 3.19: Histogramme der Abweichungen zwischen den konformationellen Stichproben und der mittels Röntgenkristallographie bestimmten Struktur. Links: Histogramme der RMSD-Werte berechnet für alle schweren Atome des Proteinrückgrats; rechts: nur für Sekundärstrukturelemente berechnete Werte.

Verteilungen der RMSD-Werte.

Obwohl in der Strukturbestimmung von Mal et al. [35] neben den NOESY-Resonanzen der Amidprotonen sieben weitere Resonanzen verwendet wurden, die von den Seitenketten der Tryptophane stammen, sind die Konformationen aus der Bayes'schen Analyse näher zur Kristallstruktur [36]:

	Mal et al. [35]	Simulation 1	Simulation 2	Simulation 3
$RMSD_1$ [Å]	2.86 ± 0.33	1.84 ± 0.20	1.54 ± 0.14	1.84 ± 0.20
$RMSD_2$ [Å]	2.01 ± 0.28	1.36 ± 0.19	1.10 ± 0.13	1.33 ± 0.20

Hier bezeichnet $RMSD_i$ die mittlere Standardabweichung zwischen den kartesischen Koordinaten der Kristallstruktur und der NMR-Strukturen. Zur Berechnung von $RMSD_1$ wurden die Koordinaten der schweren Atome des Proteinrückgrats (N, C_α und C) verwendet. Zur Berechnung von $RMSD_2$ wurden nur Atome herangezogen, die Aminosäuren aus Sekundärstrukturelementen angehören; das sind die Aminosäuren mit Nummern zwischen 86-89, 108-113, 119-124, 130-133 oder 138-140. Mal et al. berechneten mittlere RMSD-Werte für 20 Minimumstrukturen. Um die Ergebnisse der Bayes'schen

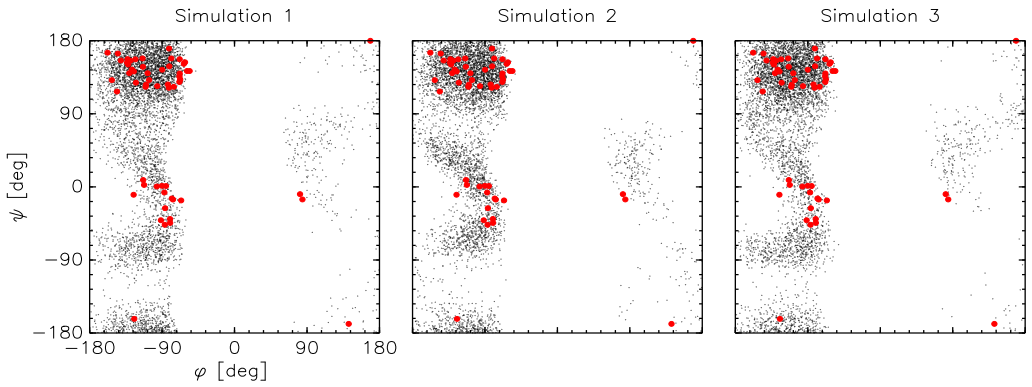


Abbildung 3.20: Dihedralwinkel φ_i und ψ_i der simulierten Konformationen; die entsprechenden Winkel der Kristallstruktur sind als rote Kreise eingetragen.

Analyse damit zu vergleichen, wurden RMSD-Werte für die zwanzig wahrscheinlichsten Konformationen berechnet und gemittelt.

Abbildung 3.20 zeigt eine andere Möglichkeit, die Konformationen mit der Kristallstruktur zu vergleichen. Die Dihedralwinkel φ_i und ψ_i des Proteinrückgrats werden gegeneinander aufgetragen. Die Werte aus der Kristallstruktur liegen in den Bereichen, wo sich die Werte konzentrieren. Die Winkel streuen noch recht stark.

Diese Beobachtung läßt sich quantifizieren. Eine PROCHECK-Analyse der hundert wahrscheinlichsten Strukturen ergibt:

Region	Simulation 1	Simulation 2	Simulation 3
favourisiert [%]	65.64 ± 5.21	67.29 ± 4.62	66.62 ± 5.16
erlaubt [%]	32.28 ± 5.26	30.97 ± 5.09	31.11 ± 5.39
noch erlaubt [%]	2.05 ± 1.76	1.75 ± 2.04	2.26 ± 1.89
verboten [%]	0.04 ± 0.28	0.00 ± 0.00	0.02 ± 0.20

Wie zu erwarten sind die Ergebnisse der zweiten Simulation etwas besser als die der beiden anderen. Die Ramachandran-Statistiken sind eher schlecht; in akzeptablen NMR-Strukturen sollten mindestens 80 % der φ-ψ-Winkel in den erlaubten Bereichen liegen. Dies ist jedoch bei der geringen Anzahl der Messungen nicht verwunderlich.

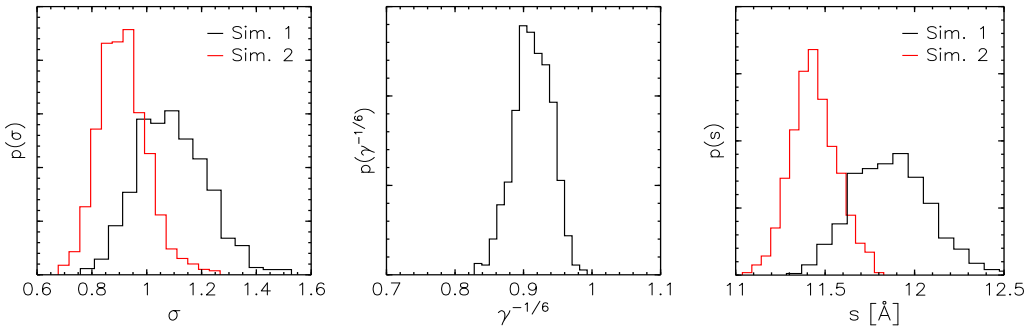


Abbildung 3.21: Links: marginale A posteriori-Verteilung der unbekannten Fehlerskala σ des Lognormal-Modells. In der dritten Simulation wurde σ analytisch ausintegriert. Mitte: Verteilung des Kalibrationsfaktors in der ersten Simulation. Rechts: Verteilungen des Gyrationsradius s der Konformationen aus der ersten und der zweiten Simulation.

Der Bayes'sche Zugang erlaubt, beliebige Parameter, die neben den Dihedralwinkeln eingeführt werden, aus den Daten zu schätzen. Abbildung 3.21 zeigt die marginalen A posteriori-Verteilungen $p(\sigma|D,I)$ der unbekannten Fehlerskala des Lognormal-Modells. Die Verteilung der zweiten Simulation ist gegenüber der der ersten zu kleineren Werten verschoben. Wegen Nichtbeachtung der Information $\gamma = 1$ ist in der ersten Simulation, in welcher neben σ auch der Kalibrationsfaktor γ geschätzt wird, das Problem schlechter bestimmt. Hinsichtlich des eingeflossenen Wissens über den Eichfaktor γ befinden sich Simulation 1 und Simulation 2 an zwei entgegengesetzten Extremen: in der ersten Simulation herrscht völlige Unkenntnis, in der zweiten völlige Kenntnis über γ. Probleme entstünden, wenn fehlerhaftes Vorwissen, bspw. $\gamma = 5$, zur Berechnung der Struktur verwendet würde. Dann wäre es besser, γ an die Daten anzupassen.

Die marginale A posteriori-Verteilung $p(\gamma^{-1/6}|D,I)$ ist in Abb. 3.21 zu sehen. Die inverse sechste Potenz von γ ist ein Korrekturfaktor der Protonenabstände. Das geometrische Mittel der Abstandsskala ist 0.91: die Strukturen der ersten Simulation sind gegenüber denen der zweiten ein wenig aufgebläht. Dies bestätigen die Verteilungen der Gyrationsradien (siehe Abb. 3.21).

Simulation	QUACHK	NQACHK	RAMCHK	BMPCHK
1	-7.50 ± 0.50	-7.32 ± 0.73	-4.42 ± 0.54	5.15 ± 2.38
2	-6.53 ± 0.46	-6.11 ± 0.62	-4.20 ± 0.55	6.58 ± 2.81
3	-7.33 ± 0.59	-7.23 ± 0.75	-4.29 ± 0.67	5.50 ± 2.76

Tabelle 3.1: Qualitätsindizes für die A posteriori-Konformationen der drei Simulationen, berechnet mit WHATIF.

In der ersten Simulation wird $\gamma^{-1/6}$ auf Kosten der Genauigkeit zu klein geschätzt, weil die konformationelle A priori-Verteilung, das Boltzmann-Ensemble, ausgedehnte Strukturen bevorzugt. Sowohl das Lösungsmittel als auch attraktive Terme in der potentiellen Energie (2.15) werden vernachlässigt, so daß wegen der höheren Multiplizität ausgedehnter Konformationen diese häufiger im Boltzmann-Ensemble zu finden sind. Erst durch Berücksichtigung zusätzlicher Informationen, nämlich $\gamma^{-1/6} = 1$, kann dieser Tendenz entgegengewirkt werden.

Die Dichte der berechneten Strukturen wird durch den Packungsindex [40] quantifiziert. Die Werte sind in Tabelle 3.1 aufgelistet. Wieder zeigt die zweite Simulation die besten Ergebnisse. Die Packungseigenschaften sind bei allen Simulationen sehr schlecht; ideale Werte liegen bei Null. Die Strukturen sind zu lose, worauf bereits die Verteilungen der Gyrationsradien hingewiesen haben. Die geringe Dichte der Konformationen führt dazu, daß sich die Atome kaum überlappen (BMPCHK); deshalb ist die Anzahl der Zustammenstöße zwischen Atomen geringer als in der Kristallstruktur (9 bumps).

Es gibt keine freien Parameter; jede Unbekannte kann entweder aus den Daten geschätzt, oder wenn sie nicht interessiert, über ihre möglichen Werte gemittelt werden. Algorithmisch sind beide Zugänge äquivalent: es werden Stichproben von allen Hypothesenparametern gezogen und nur diejenigen betrachtet, die von Interesse sind. Von zweitrangigem Interesse sind in der Strukturberechnung Parameter, die aus der Theorie und dem Fehlergesetz stammen. In Abschnitt 3.2 wurde gezeigt, wie der erweiterte Hypothesenraum durch Integration über die zusätzlichen Parameter auf den Konfigura-

tionsraum reduziert werden kann.

Daß die volle A posteriori-Verteilung $p(\theta, \gamma, \sigma)$ hinsichtlich der unbekannten Struktur dieselbe Information enthält wie die marginale Verteilung $p(\theta) = \int \mathrm{d}\gamma\, \mathrm{d}\sigma\ p(\theta, \gamma, \sigma)$ wird in der Praxis bestätigt. Die Übereinstimmung beider Verteilungen ist zwar nicht sicher nachweisbar, es können jedoch indirekte Hinweise hierfür erbracht werden. Die Ähnlichkeit der RMSD-Histogramme von Simulation 1 und 3 (Abb. 3.19) ist ein erstes, freilich grobes, Indiz für diese Äquivalenz; desweiteren ihre Übereinstimmung in den Qualitätsindizes. Die Verteilungen der Dihedralwinkel θ_i lassen genauere Aussagen zu. Es gilt

$$p(\theta_i) = \int \mathrm{d}\gamma\, \mathrm{d}\sigma \prod_{j \neq i} \mathrm{d}\theta_j \ \ p(\theta_1, \ldots, \theta_{275}, \gamma, \sigma) = \int \prod_{j \neq i} \mathrm{d}\theta_j \ \ p(\theta_1, \ldots, \theta_{275}).$$

Die aus der ersten und dritten Simulation abgeleiteten Verteilungen $p(\theta_i | D, I)$ sollten deshalb idealerweise identisch sein.

Der Abstand zweier Verteilungen $p(\theta_i)$, $q(\theta_i)$ eines Dihedralwinkels θ_i wird durch die relative Entropie oder Kullback-Leibler-Distanz

$$D(p, q) = -\int \mathrm{d}\theta_i \ \ p(\theta_i)\ \log\left[p(\theta_i)\,/\,q(\theta_i)\right] \ \geq \ 0 \tag{3.28}$$

gemessen; sie ist invariant unter Variablentransformation und verschwindet genau dann, wenn die beiden Verteilungen identisch sind. Die Kullback-Leibler-Distanzen der Dihedralwinkelverteilungen sind somit ein Maß für die Übereinstimmung zweier konformationeller A posteriori-Verteilungen.

Abbildung 3.22 zeigt die Kullback-Leibler-Distanzen zwischen den Dihedralwinkelhistogrammen der ersten und der zweiten Simulation, sowie zwischen der ersten und der dritten Simulation. Die Dihedralwinkelverteilungen der ersten und der zweiten Simulation unterscheiden sich weitaus stärker als die Verteilungen der ersten und der dritten Simulation, welche in den Verteilungen der Dihedralwinkel der Hauptkette kaum Unterschiede aufweisen. Die Verteilungen zweier Winkel mit besonders großen Abweichungen sind exemplarisch abgebildet. Während sich die Histogramme der ersten und dritten Simulation gut decken, weicht die zweite Simulation von beiden ab.

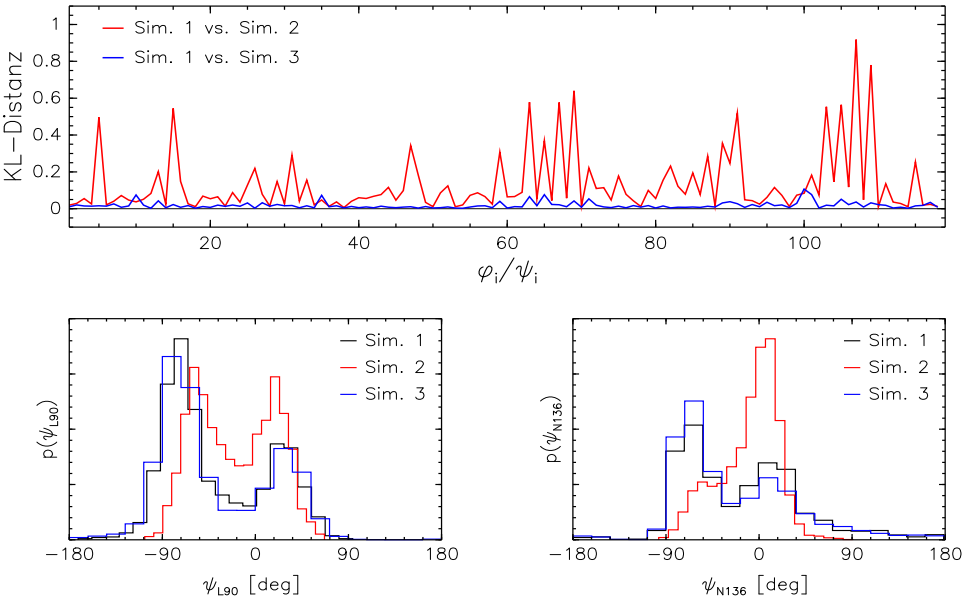


Abbildung 3.22: Links: Kullback-Leibler-Distanzen zwischen den Histogrammen der φ-ψ-Dihedralwinkel. Rechts: Beispiele für Verteilungen, deren Abweichungen besonders groß sind (ψ in LEU90, ψ in ASN136).

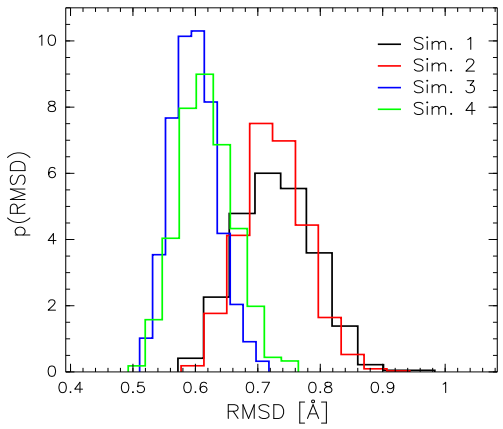
3.5.2 Analyse der Ubiquitin-Daten

Für Ubiquitin liegen Messungen aller hier behandelten Observablen vor (siehe Abschnitt 2.4.2). Jeder skalare Kopplungstyp ${}^3J_j$ wird durch eine eigene Karplus-Kurve

$$ {}^3J_j(\varphi) = A_j \cos^2(\varphi + \delta_j) + B_j \cos(\varphi + \delta_j) + C_j \qquad (3.29) $$

beschrieben. Die Phasen δ_j wandeln den Winkel φ in die Dihedralwinkel um, die durch die Bindungen der jeweiligen Kopplung definiert werden; die Phasen sind bekannt, weil die kovalente Geometrie starr ist.

Dipolare Kopplungen, die in demselben Medium gemessen wurden, werden mit derselben Saupe-Matrix beschrieben, weil die mittlere Ausrichtung der Moleküle im Lösungsmittel unabhängig von der Art des Experiments ist. Die Messungen sind jedoch von unterschiedlicher Güte; jeder Datensatz hat einen eigenen Fehler. Die Wahrscheinlichkeit von Messungen D_1, ..., D_m, die



Simulation	RMSD [Å]
1	0.73 ± 0.06
2	0.73 ± 0.05
3	0.60 ± 0.04
4	0.62 ± 0.04
5	29.49 ± 2.99
6	27.31 ± 5.30
7	20.66 ± 1.41

Abbildung 3.23: Histogramme der Abweichungen zwischen konformationellen Stichproben und der Kristallstruktur. Zum Vergleich wurden die Koordinaten der schweren Atome des Proteinrückgrats herangezogen; die letzten drei Aminosäuren des C-Terminus wurden nicht berücksichtigt, weil die Daten ihre Position nur schlecht bestimmen.

in demselben Medium gemessen wurden, ist also

$$p(D_1, \ldots, D_m|\theta, \mathbf{s}, \sigma_1, \ldots, \sigma_m) = \prod_{j=1}^{m} p(D_j|\theta, \mathbf{s}, \sigma_j).$$

Die NOE-Messungen sind nicht alle eindeutig zugeordnet. Die Die Volumina mehrdeutiger NOEs werden als Summe der Partialvolumina der zugeordneten Protonenpaare berechnet. Das theoretische Volumen einer Resonanz mit c Zuordnungsmöglichkeiten (k_1, l_1), ..., (k_c, l_c) ist

$$\hat{V} = \gamma \sum_{i=1}^{c} r_{k_i l_i}^{-6}.$$

Die inverse sechste Wurzel von $\hat{V}$ entspricht einem ambiguous distance restraint [47]. Für magnetisch äquivalente Protonen, beispielsweise Methylgruppen, ergibt das slow jump model [50] dieselbe Abhängigkeit.

Ich habe Replica-Simulationen mit verschiedenen Daten durchgeführt:

1. Analyse der NOESY-Daten;
2. Analyse der skalaren Kopplungskonstanten und der NOESY-

Sim.	favourisiert [%]	erlaubt [%]	noch erlaubt [%]	verboten [%]
1	74.22 ± 2.73	24.45 ± 2.75	1.30 ± 1.08	0.00 ± 0.00
2	89.08 ± 2.03	9.41 ± 2.02	1.50 ± 0.00	0.00 ± 0.00
3	85.92 ± 2.18	14.05 ± 2.19	0.03 ± 0.21	0.00 ± 0.00
4	91.05 ± 1.72	8.75 ± 1.88	0.20 ± 0.51	0.00 ± 0.00
5	68.25 ± 3.53	29.44 ± 3.76	2.28 ± 1.60	0.01 ± 0.15
6	56.53 ± 5.74	41.38 ± 5.70	2.00 ± 1.76	0.09 ± 0.36
7	72.54 ± 2.98	23.63 ± 3.47	3.76 ± 1.40	0.03 ± 0.21

Tabelle 3.2: Ramachandran-Statistiken der Simulationen von Ubiquitin.

Daten;

3. Analyse der dipolaren Kopplungen und der NOESY-Daten;
4. Analyse aller Daten;
5. Analyse der dipolaren Kopplungen;
6. Analyse der skalaren Kopplungskonstanten;
7. Analyse der skalaren und der dipolaren Kopplungen.

Ein Vergleich der berechneten Konformationen mit der Kristallstruktur zeigt, daß die Übereinstimmung stets besser wird, wenn mehr Daten in die Berechnung eingehen (siehe Abb. 3.23). Dies scheint selbstverständlich, muß aber nicht notwendigerweise gelten, wie Beispiele zeigen, wo Minimierung bei Verwendung zusätzlicher Daten schlechtere Strukturen findet.

Die Güte der berechneten Konformationen entspricht dem, was man für einen Datensatz dieser Qualität und Vollständigkeit erwartet. Die von Bax und Mitarbeitern berechneten Strukturen (PDB-Code 1d3z) sind durchweg näher zur Kristallstruktur. Für die zehn in der Datenbank abgelegten Konformationen ist der RMSD zur Kristallstruktur 0.40 ± 0.02 Å. Zu berücksichtigen ist jedoch, daß zusätzlich zu den hier verwendeten Daten noch 27 Distanzen aus Wasserstoffbrücken und 35 Einschränkungen der χ_1-Dihedralwinkel zur Strukturberechnung verwendet wurden.

Die Verwendung skalarer Kopplungen ändert die Qualität der Strukturen nicht wesentlich: die Verteilungen der RMSD-Werte sind in Simulation 1 und

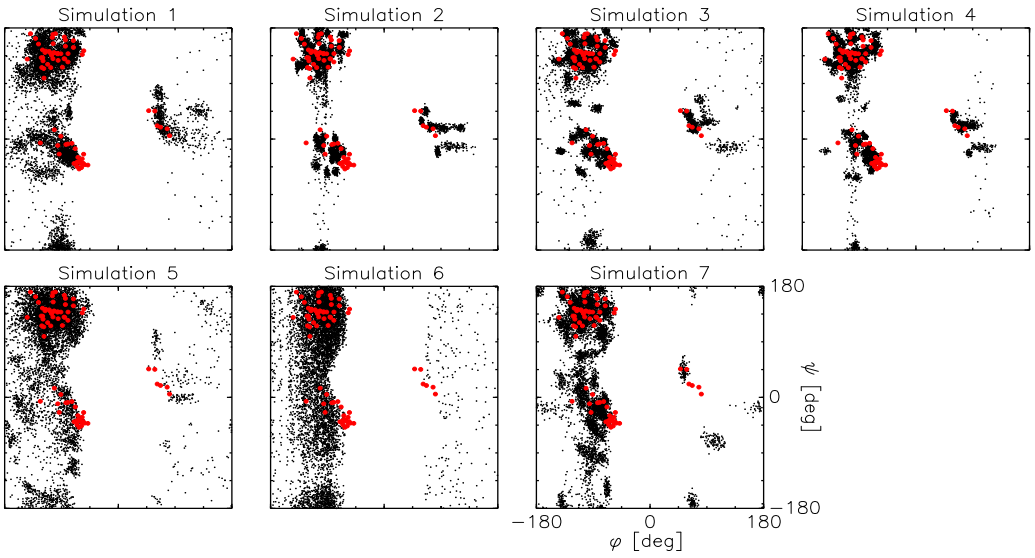


Abbildung 3.24: Stichproben der φ-ψ-Winkel. Die Werte aus der Kristallstruktur sind als rote Kreise eingezeichnet.

2 gleich. Dagegen verbessert sich die Ramachandran-Statistik deutlich (siehe Table 3.2). Durch zusätzliche Analyse dipolarer Kopplungen wird hingegen die Faltung verbessert: die Konformationen aus Simulation 3 und 4 sind systematisch näher zur Kristallstruktur. Auch die Ramachandran-Statistiken verbessern sich. Dipolare Kopplungen hängen nicht bloß von φ_i ab, sondern von allen Dihedralwinkeln, die die Orientierung der Bindungsvektoren beeinflussen. In den verwendeten Datensätzen sind das die Dihedralwinkel des Rückgrats der Polypeptidkette. In Simulation 3 sind die φ-ψ-Stichproben nicht so scharf um die Winkel der Kristallstruktur konzentriert (siehe Abb. 3.24) wie bei Verwendung skalarer Kopplungskonstanten (Simulation 2), die Nähe zur Kristallstruktur ist dennoch größer.

Ohne NOESY-Daten können keine kompakten Strukturen berechnet werden. In Simulation 5 bis 7 sind die konformationellen Stichproben langgestreckt, was sich in sehr schlechten RMSD-Werten niederschlägt. Die skalaren Kopplungskonstanten wurden nur für φ_i gemessen, so daß die übrigen Dihedralwinkel lediglich durch die konformationelle A priori-Verteilung bestimmt werden; die ist jedoch zu ungenau und bevorzugt eher gestreckte

Simulation	QUACHK	NQACHK	RAMCHK	BMPCHK
1	-1.81 ± 0.22	-2.47 ± 0.31	-3.66 ± 0.39	13.90 ± 3.55
2	-1.22 ± 0.17	-2.53 ± 0.28	-1.94 ± 0.30	15.70 ± 2.82
3	-1.13 ± 0.19	-2.22 ± 0.28	-1.82 ± 0.35	10.70 ± 3.86
4	-0.89 ± 0.17	-2.00 ± 0.28	-0.68 ± 0.28	11.90 ± 2.78
5	-9.57 ± 0.22	-11.30 ± 0.23	-4.13 ± 0.42	10.40 ± 2.99
6	-9.49 ± 0.35	-10.90 ± 0.42	-4.96 ± 0.60	7.60 ± 3.43
7	-9.02 ± 0.16	-10.30 ± 0.29	-3.93 ± 0.38	12.10 ± 2.92
PDB 1d3z	-0.03 ± 0.17	-0.25 ± 0.21	1.72 ± 0.20	14.80 ± 4.02

Tabelle 3.3: Qualitätskriterien berechnet mit WHATIF [40]. (Für Erläuterungen der Indizes siehe Tabelle 3.1.)

Konformationen.

Dipolare Kopplungen bestimmen bloß die gegenseitige Orientierung der Bindungsvektoren; auch sie vermögen keine kompakte Struktur festzulegen. Weil dipolare Kopplungen vom Abstand und der Orientierung der Atompaare abhängen, beeinflußt eine räumliche Verschiebung der wechselwirkenden Atompaare die Stärke ihrer Kopplung nicht. Die Daten enthalten nur Messungen dipolarer Kopplungen zwischen Atomen, die durch höchstens zwei Bindungen getrennt sind, und legen deshalb keine Tertiärkontakte zwischen sequentiell entfernten Atomen fest.

Abbildung 3.24 und die Ramachandran-Statistiken zeigen zwar, daß die Dihedralwinkel des Proteinrückgrats selbst ohne Verwendung der NOESY-Daten einigermaßen gut bestimmt sind – Simulation 7 ist hinsichtlich dieser Kriterien vergleichbar gut wie Simulation 1 – , dennoch erhält man keine globulären Konformationen. Dies liegt an der Tendenz des Boltzmann-Ensembles, ausgedehnte Konformationen der Polypeptidkette zu bevorzugen. Beispielsweise weicht die Verteilung der Gyrationsradien stark von der gefalteter Proteine ab. Daß gefaltete Proteine globulär sind, müßte zusätzlich in die konformationelle A priori-Verteilung eingehen.

Weitere Qualitätsmaße sind in Table 3.3 aufgelistet. Auch hinsichlich die-

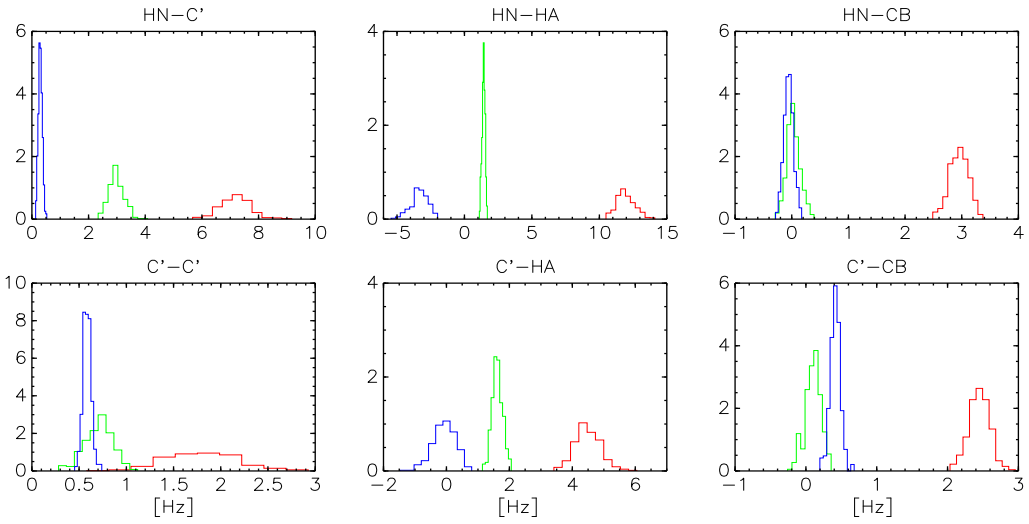


Abbildung 3.25: Marginale A posteriori-Verteilungen der Karplus-Koeffizienten in Simulation 2 (Analyse der NOESY- und der J-Kopplungs-Daten). Rot: $p(A|D,I)$, Grün: $p(B|D,I)$, Blau: $p(C|D,I)$.

ser Kriterien zeigt sich, daß eine Verwendung aller Daten die besten Konformationen liefert. Die Strukturen in 1d3z haben durchweg bessere Werte; dies ist jedoch zu erwarten, weil zusätzlich Distanzen für Wasserstoffbrücken in die Rechnung eingingen.

Gewöhnlich können Messungen skalarer Kopplungskonstanten nur zur Strukturberechnung verwendet werden, wenn bereits eine Struktur vorliegt, an der die Karplus-Kurve geeicht wird (siehe Abschnitt 3.8). In der Bayes' schen Analyse sind die Karplus-Koeffizienten einfach zusätzliche Hypothesenparameter.

Die folgende Tabelle listet die Mittelwerte und Varianzen der Karplus-Koeffizienten auf, die sich aus der gemeinsamen Analyse mit den NOESY-Resonanzen (Simulation 2) ergeben; in runden Klammern stehen die Werte, die in den Originaldaten [37] verwendet wurden:

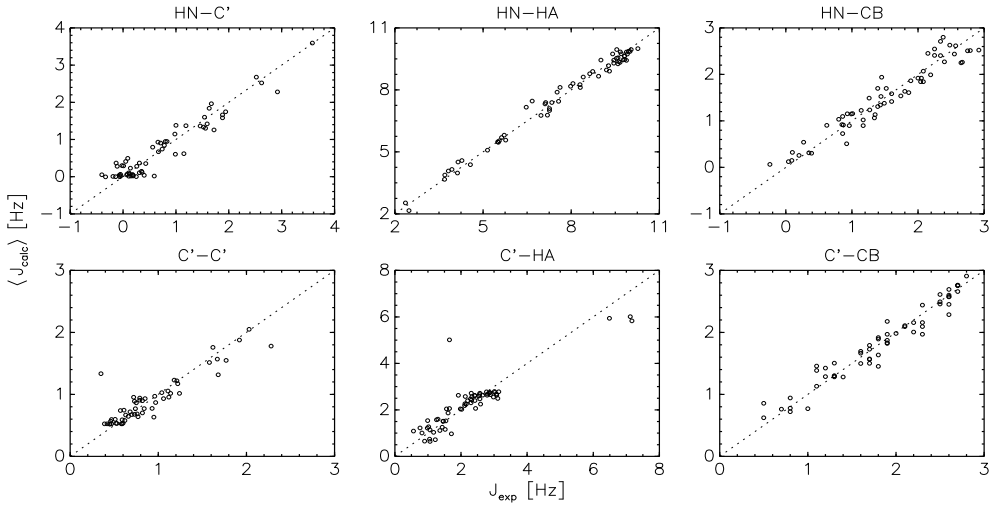


Abbildung 3.26: Gemessene skalare Kopplungskonstanten gegen mittlere berechnete Kopplungen, die sich aus den Stichproben der A posteriori-Verteilung von Simulation 2 ergeben.

${}^3J_i$	A [Hz]	B [Hz]	C [Hz]
H_N-C	7.19 ± 0.53 (4.29)	2.98 ± 0.27 (-1.01)	0.30 ± 0.07 (0.00)
H_N-H_α	11.96 ± 0.68 (7.09)	1.42 ± 0.11 (-1.42)	−3.34 ± 0.64 (1.55)
H_N-C_β	2.95 ± 0.17 (3.06)	0.03 ± 0.12 (-0.74)	−0.05 ± 0.08 (0.13)
C-C	1.80 ± 0.38 (1.36)	0.71 ± 0.15 (-0.93)	0.58 ± 0.04 (0.60)
C-H_α	4.49 ± 0.41 (3.72)	1.62 ± 0.17 (-2.18)	−0.08 ± 0.36 (1.28)
C-C_β	2.45 ± 0.15 (1.74)	0.10 ± 0.11 (-0.57)	0.42 ± 0.07 (0.25)

In einigen Fällen stimmen die Werte innerhalb ihrer Schwankungen mit denen, die sich in den Daten (PDB-code 1d3z) finden, überein. Hierbei ist zu beachten, daß die Vorzeichen von B abweichen können, weil die Phase δ in der Karplus-Kurve (3.29) um 180° verschoben gewählt wurde. Abbildung 3.25 zeigt die marginalen A posteriori-Verteilung der Karplus-Koeffizienten der verschiedenen skalaren Kopplungskonstanten. Auf die Schätzung der Karplus-Kurven wird näher in Abschnitt 3.8 eingegangen.

Ein Vergleich der gemessenen skalaren Kopplungskonstanten mit den mittleren berechneten Kopplungen ist in Abbildung 3.26 gezeigt. Die Karplus-

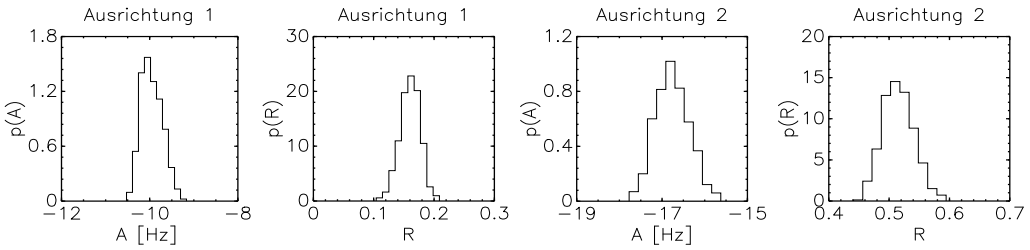


Abbildung 3.27: A posteriori-Verteilungen der axialen und rhombischen Anteile der Orientierungstensoren für beide Ausrichtungen, in denen dipolare Kopplungen gemessen wurden.

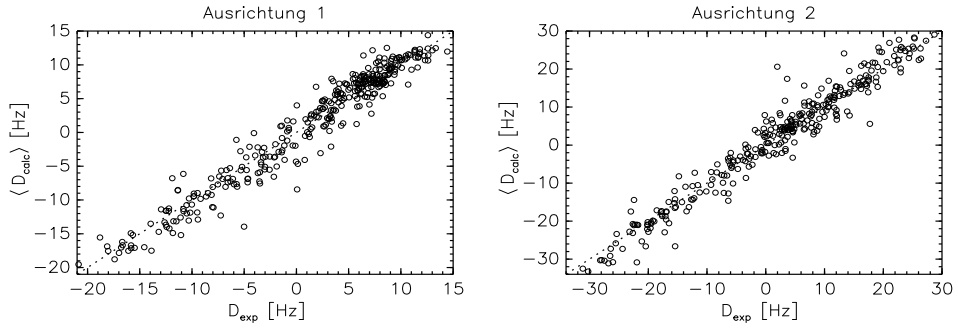


Abbildung 3.28: Gemessene dipolare Kopplungen gegen mittlere berechnete dipolare Kopplungen, die sich aus den Stichproben der A posteriori-Verteilung von Simulation 3 ergeben. Die Werte wurden auf das Maximum der H_N-N-Kopplung normiert.

Koeffizienten und die Dihedralwinkel stellen sich so ein, daß die Messungen innerhalb ihrer Schwankungen gut rekonstruiert werden.

Auch die mittlere Orientierung, beschrieben durch die Saupe-Matrix, wird während der Strukturberechnung geschätzt. In Abbildung 3.27 sind die A posteriori-Verteilungen der axialen und rhombischen Komponente des Orientierungstensors abgebildet.

In Abbildung 3.28 sind die gemessenen gegen die mittleren berechneten dipolaren Kopplungen aufgetragen.

Die A posteriori-Verteilungen der Theorieparameter für sowohl die skalaren Kopplungskonstanten als auch die dipolaren Kopplungen zeigen, daß

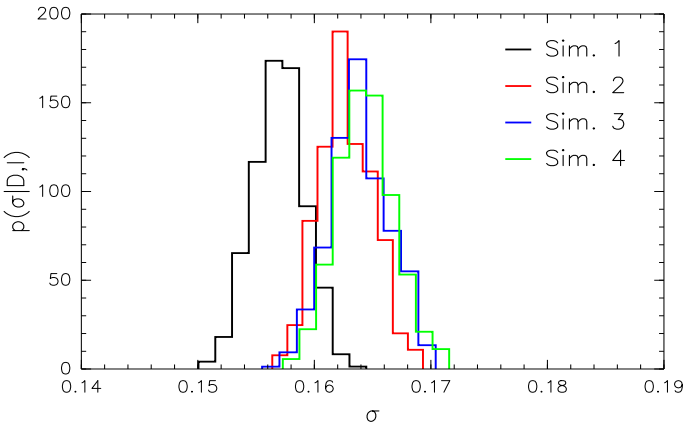


Abbildung 3.29: Marginale A posteriori-Verteilungen der Fehlerskala des Lognormal-Modells für die verschiedenen Simulationen, in denen NOESY-Daten verwendet wurden.

diese Daten allein ausreichen um stabile A posteriori-Verteilungen zu definieren. Nicht weil die Theorieparameter zu schlecht bestimmt wären, sondern weil das strukturelle Vorwissen nur grob und unvollständig dargestellt wird, sind bei alleiniger Verwendung skalarer oder dipolarer Kopplungen die berechneten Strukturen nicht gefaltet.

Neben den Theorieparametern paßt eine Bayes'sche Analyse die unbekannten Fehler σ_j jedes Datensatzes an. Dies erlaubt, die Datensätze individuell zu bewerten und jeden gemäß seiner Güte in die gemeinsame Analyse eingehen zu lassen. Dies ist wichtig, um die Daten optimal zu nutzen und um das Ergebnis nicht aufgrund der Mängel eines Datensatzes unnötig zu verschlechtern. In die Bestimmung der Fehler gehen indirekt durch die Dihedralwinkel sowohl die Informationen der anderen Datensätze ein als auch strukturelles Vorwissen (siehe Abschnitt 3.7). Je nachdem welche Daten zusätzlich in die Analyse einfließen, ergeben sich andere Verteilungen für die Fehler. Abbildung 3.29 zeigt die Verhältnisse für die Fehlerskala der NOESY-Messungen. Bei Verwendung zusätzlicher Daten wird der Fehler größer geschätzt. Simulation 1 liegen nur NOESY-Messungen zugrunde, nur die A priori-Information kann als Korrektiv der Überanpassung der Konformationen an die Messun-

gen dienen. Der Fehler wird kleiner als bei Verwendung zusätzlicher Daten geschätzt, die Struktur wird zu stark an die Messungen angepaßt; d.h. Rauschen wird als Signal interpretiert.

3.6 Gewichtung eines Datensatzes

Größen, die zusätzlich zu den Koordinaten eingeführt werden, bereiten der traditionellen Strukturberechnung grundsätzliche Schwierigkeiten. Zum einen müssen die Messungen gegenseitig und im Verhältnis zur physikalischen Energie gewichtet werden. Zum anderen enthalten die Theorien zur Beschreibung der Observablen empirische Parameter wie die Koeffizienten der Karplus-Kurve oder die Elemente der Saupe-Matrix. Diese Größen sind nicht bekannt und auch nicht direkt meßbar; sie müssen zusammen mit der Struktur geschätzt werden. In den folgenden Abschnitten wird auf solche Problem näher eingegangen.

3.6.1 Kreuzvalidierung

Klassische Strukturberechnung minimiert eine Zielfunktion der Form

$$G(\theta) = E(\theta) + \lambda F(\theta).$$

Das relative Gewicht λ zwischen der physikalischen Energie E und dem Datenterm F ist unbekannt. Dieser zusätzliche Parameter entspricht der Fehlerskala σ eines probabilistischen Modells (Abschnitt 2.2.7). Die richtige Wahl von λ ist entscheidend: Wird zuviel Gewicht auf die Daten gelegt, dann sind die berechneten Strukturen zu stark an die Messungen angepaßt (overfitting) und in unphysikalischer Weise verzerrt; ist jedoch das Gewicht der Daten zu gering, bleiben die Strukturen lose und unbestimmt.

In der orthodoxen Statistik werden unbekannte Parameter wie das Gewicht λ durch eine Kreuzvalidierung (cross-validition) [51] bestimmt. Brünger [15] sowie Brünger et al. [16] haben vorgeschlagen, diese Technik in der makromolekularen Strukturbestimmung mittels Röntgenkristallographie bzw.

Kernresonanzspektroskopie anzuwenden. Kreuzvalidierung kann auch der Bewertung einer berechneten Struktur dienen: in der Kristallographie ist der „R-Faktor“ oft kein verläßliches Maß, sondern erst der kreuzvalidierte „freie R-Faktor“.

Bei der Kreuzvalidierung werden folgende Schritte durchgeführt:

1. Die Daten D werden unterteilt in eine Menge A von Daten (working set) zur Berechnung der Struktur und einer Menge T (test set) zur Bewertung einer Wahl von λ;

2. Für einen bestimmten Wert λ wird die Hybridenergie minimiert, die sich allein aus den Arbeitsdaten A ergibt;

3. Die Werte von λ werden durch eine Funktion R bewertet, die an den Testdaten ausgewertet wird.

Die Bewertungsfunktion R (der „R-Faktor“) ist üblicherweise die normierte Abweichung zwischen experimentellen und aus der Struktur berechneten Testdaten: eine bestimmte Wahl von λ wird daran bewertet, ob die Strukturen die nicht verwendeten Daten vorhersagen können. Man erhält eine Kurve $R(\lambda)$ und als beste Wahl von λ den Wert, für den die Abweichung zwischen Messung und Vorhersage minimal ist.

Den Wert von R, den man für die beste Wahl von λ erhält, nennt man den „freien R-Faktor“. Der freie R-Faktor wird als objektives Maß für die Güte einer Struktur angesehen [15, 16]. Hier zeigt sich, daß die Qualität der Struktur (gemessen durch R) direkt von der Qualität der Daten (gemessen durch λ) abhängt. Dies wurde bereits in Abschnitt 2.2.7 durch die Entsprechung zwischen λ und σ^{-2} deutlich.

Bei K-facher Kreuzvalidierung wird die Rechnung mit K Zerlegungen der Daten durchgeführt. Die Arbeitsdaten setzen sich jeweils aus $K-1$ Teilmengen zusammen (d.h. aus einem Bruchteil $(K-1)/K$ der Daten); getestet wird anhand des verbleibenden Datensatzes. Die R-Werte werden über die verschiedenen Datensätze gemittelt.

3.6.2 Bayes'sche Wahl des Gewichts

Wie verhält sich Kreuzvalidierung zur Bayes'schen Bestimmung von λ? Der negative Logarithmus der A posteriori-Verteilung

$$-\log p(\theta, \lambda) = \frac{\lambda}{2}\, \chi^2(\theta) + \beta\, E(\theta) - \left(\frac{n}{2} - 1\right) \log \lambda$$

bestimmt die wahrscheinlichsten Werte von θ und λ $(= \sigma^{-2})$. Zusätzliche Theorieparameter seien entweder bekannt oder ausintegriert worden; für $Z(\lambda)$ wurde die Abhängigkeit verwendet, die sich aus den Modellen aus Kapitel 3.1 ergibt: $Z(\lambda) \propto 1/\sqrt{\lambda}$.

Weil die Regeln der Wahrscheinlichkeitsrechnung eine *gemeinsame* A posteriori-Verteilung $p(\theta, \lambda)$ der Konformation und des Gewichts festlegen, können beide zusammen bestimmt werden, ohne daß man auf eine Heuristik zurückgreifen müßte. Die Zielfunktion der Optimierungsverfahren ist dagegen nur auf dem Konformationsraum definiert. Es fehlen die aus der Normierung der Datenverteilung stammenden, λ-abhängigen Terme sowie ein Pendant zur A priori-Verteilung $\pi(\lambda)$ (Abschnitt 2.2.7).

Minimierung von $-\log p(\theta, \lambda)$ liefert das wahrscheinlichste Gewicht

$$\hat{\lambda} = (n - 2)/\chi^2(\hat{\theta}),$$

wobei die wahrscheinlichsten Konformationen $\hat{\theta}$ Lösungen von

$$2\beta\, \nabla E(\hat{\theta}) + \hat{\lambda}\, \nabla \chi^2(\hat{\theta}) = 0$$

sind. Einsetzen von $\hat{\lambda}$ und Integration liefern eine reduzierte Zielfunktion

$$\hat{G}(\theta) = \left(\frac{n}{2} - 1\right) \log \chi^2(\theta) + \beta\, E(\theta),$$

in der das Gewicht nicht mehr auftaucht. Diese Zielfunktion ist beinah identisch zum negativen Logarithmus der A posteriori-Verteilung, in der der Fehler ausintegriert wurde: $-\log p(\theta) = \frac{n}{2} \log \chi^2(\theta) + \beta\, E(\theta)$.

Die in der Kreuzvalidierung verwendete Zielfunktion zur Berechnung der Struktur bei vorgegebenem Gewicht ist

$$G_A(\theta) = \lambda\, F_A(\theta) + E(\theta). \tag{3.30}$$

Es gilt $F_A(\theta) \propto \chi^2_A(\theta)$, wobei χ^2_A sich aus der Verteilung der Messungen A ergibt. Durch Minimierung von G_A an verschiedenen Werten λ, erhält man eine Schar $\hat{\theta}(\lambda)$ von Strukturen. λ wird so gewählt, daß der R-Faktor $R(\lambda) \propto \chi^2_T\big(\hat{\theta}(\lambda)\big)$ minimal ist.

3.6.3 Analyse der NOESY-Daten von Ubiquitin

Anhand der NOESY-Messungen von Ubiquitin lassen sich die Ergebnisse einer Kreuzvalidierung mit der Bayes'schen Analyse vergleichen. Als Daten dienten die eindeutig zugeordneten NOESY-Signale (s. Abschnitt 2.4.2). Falls für eine Resonanz mehrere Messungen vorlagen, wurde die mit größtem Volumen gewählt; man erhält insgesamt 1444 Datenpunkte. Es wurde eine vollständige 10-fache Kreuzvalidierung durchgeführt. Die Messungen wurden zufällig in 10 Teilmengen ungefähr gleichen Umfangs eingeteilt und Strukturen bei Fortlassen jeweils eines der zehn Datensätze berechnet. Die Zielfunktion G_A (3.30) wurde mit dem simulated annealing Protokoll [52] von CNS [53] minimiert. Weil der Energieterm, der sich aus der Lognormal-Verteilung ergibt, nicht in CNS implementiert ist, wurden die NOESY-Volumina, entgegen der sonstigen Vorgehensweise, mit einer Gauß'schen Fehlerverteilung beschrieben, um einen Vergleich zwischen Kreuzvalidierung und Bayes'scher Analyse zu ermöglichen. Es wurden Rechnungen mit $\lambda = 0.002$, 0.02, 1.14, 2.28, 4.56, 11.41, 17.11, 22.82, 34.42, 45.64, 228.21 durchgeführt. In der Bayes'schen Analyse wurde $\gamma = 1$ gesetzt.

Der Einfluß von λ auf die physikalischen Eigenschaften der Struktur ist in Abbildung 3.30 dargestellt: Mit zunehmendem Gewicht werden die Strukturen immer stärker verzerrt; Überlappungen der Atome führen zu einem Anstieg der van der Waals-Energie; die kovalenten Parameter werden deformiert. Dagegen sind die Arbeitsdaten, welche zur Strukturberechnung verwendet werden, immer besser erfüllt: χ^2_A fällt monoton mit wachsendem λ. Die Abweichungen in den Testdaten weisen ein schwaches Minimum auf: Die Strukturen werden bei großen λ zu sehr an die Arbeitsdaten angepaßt, so daß die Testdaten schlechter vorausgesagt werden.

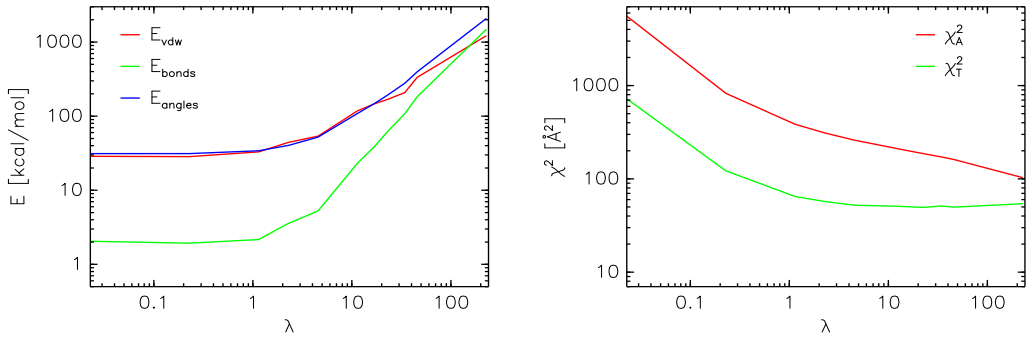


Abbildung 3.30: Links: Änderung der physikalischen Energien bei unterschiedlicher Gewichtung der Daten (E_{vdw}: van der Waals-Energie, E_{bonds} bzw. E_{angles}: Energien der Bindungslängen bzw. -winkel). Rechts: Verlauf der Abweichungen zwischen gemessenen und berechneten Distanzen in den Arbeitsdaten A (χ^2_A) und den Testdaten T (χ^2_T).

Ziel ist es, λ so zu wählen, daß die Minimumstruktur möglichst gut an die Arbeitsdaten angepaßt ist; jedoch nicht so stark, daß ihre kovalente Geometrie verzerrt wird oder sich die Atome durchdringen. Zur Bestimmung des besten Werts von λ wird in der Kreuzvalidierung überlicherweise die normierte Abweichung zu den Testdaten betrachtet:

$$R = \frac{\sum_{i\in T}\left(V_i^{-1/6} - d_i(\theta)\right)^2}{\sum_{i\in T} V_i^{-1/3}} \propto \chi^2_T(\theta),$$

$d_i(\theta)$ sind die Abstände in der Struktur θ, welche bei gegebenem λ die Abweichungen zu den Arbeitdaten minimiert. Ein weiteres Maß ist [16]

$$R_{1/6} = \frac{\sum_{i\in T}\left|V_i^{1/6} - d_i^{-1}(\theta)\right|}{\sum_{i\in T} V_i^{1/6}}.$$

Abbildung 3.31 zeigt die Abhängigkeit dieser beiden Maße von λ. Beide sind eher unspezifisch und weisen nur ein sehr breites, kaum ausgeprägtes Minimum auf; sie legen λ nicht genau fest. In die Graphen ist die marginale A posteriori-Verteilung $p(\lambda|D, I)$ eingezeichnet, welche sich aus einer Bayes'schen Berechnung ergibt. Die Verteilung des Gewichts ist scharf bestimmt und hat sein Maximum in dem Bereich, wo die R-Faktoren minimal werden.

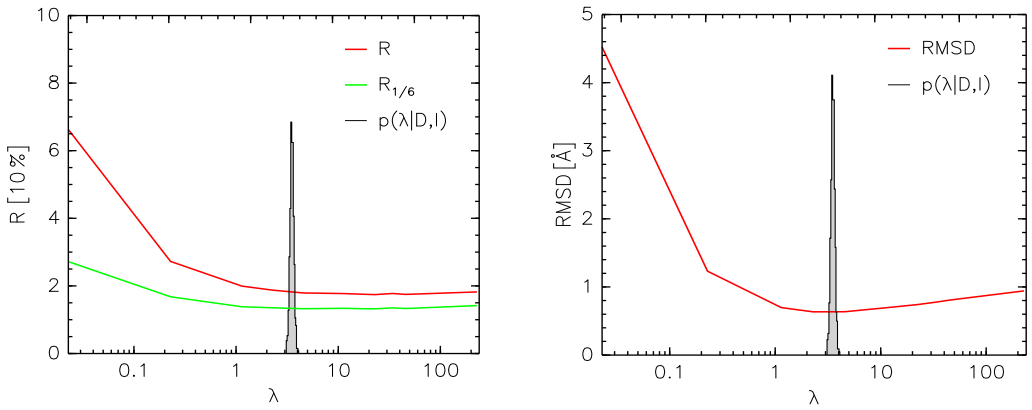


Abbildung 3.31: Verlauf der R-Faktoren (links) und des RMSDs (rechts) zwischen berechneter und Kristallstruktur bei unterschiedlicher Gewichtung der Daten.

Die Wahl des geeigneten Maßes zur Bewertung von λ stellt ein Problem dar: In diesem Beispiel sind die R-Faktoren nicht sehr aussagekräftig; außerdem sollte die physikalische Qualität der Strukturen berücksichtigt werden, R und $R_{1/6}$ hängen jedoch nur von den Testdaten ab.

Idealerweise möchte man das Gewicht so wählen, daß die berechnete Struktur möglichst nahe zur wahren Struktur ist. In dem Beispiel kann der RMSD zur Kristallstruktur als ein solcher Indikator dienen. Abbildung 3.31 zeigt den Verlauf des RMSD zwischen der Kristallstruktur und den bei verschiedener Gewichtung der Daten berechneten Konformationen. RMSDs, berechnet für denselben Wert von λ mit verschiedenen Arbeitsdaten, wurden gemittelt. Der RMSD wird bei den Gewichten minimal, die maximale A Posteriori-Wahrscheinlichkeit haben: die Bayes'sche Analyse paßt λ während der Strukturberechnung an und setzt so die Daten in das richtige Verhältnis zum Vorwissen über die unbekannte Struktur.

Die Ergebnisse einer Kreuzvalidierung sind automatisch in der Bayes'schen Analyse enthalten. Die Verteilung der RMSD-Werte (Abb. 3.32) zeigt, daß die von der A posteriori-Verteilung gezogenen Konformationen so nahe zur Kristallstruktur sind wie die mit idealem Gewicht berechneten Minimumstrukturen.

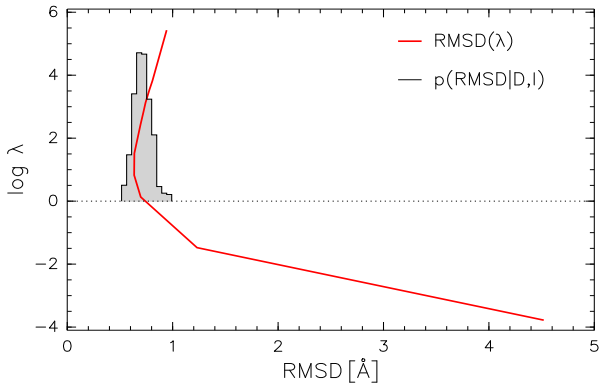


Abbildung 3.32: Zusammenhang zwischen dem Datengewicht und den mittleren RMSDs der minimierten Strukturen im Vergleich zur A posteriori-Verteilung der RMSD-Werte in einer Bayes'schen Strukturberechnung.

3.7 Gewichtung mehrerer Datensätze

Die Gewichte mehrerer Datensätze müßten im traditionellen Zugang durch multidimensionale Kreuzvalidierung bestimmt werden. Der Rechenaufwand wächst exponentiell mit der Anzahl der Datensätze; die Instabilitäten werden gravierender.

3.7.1 Bayes'sche Schätzung der Gewichte

In der Wahrscheinlichkeitsrechnung gibt es solche Schwierigkeiten nicht. Gemäß (2.10) gehen die Datensätze $D_1, \ldots, D_m$ gewichtet in die Gesamtabweichung ein:

$$\chi^2(\theta, \sigma) = \sum_{i=1}^{m} w_j \, \chi_j^2(\theta), \quad w_j \propto \sigma_j^{-2} = \lambda_j,$$

wobei χ_j^2 die Abweichungsfunktion des j-ten Datensatzes ist. Die Gesamtabweichung $\chi^2(\theta, \sigma)$ legt die Verbundverteilung $p(\theta, \sigma_1, \ldots, \sigma_m)$ fest. Die Gewichte passen sich während der Parameterschätzung an. Beim Gibbs sampling (2.20) werden Stichproben der Gewichte von Gamma-Verteilungen gezogen:

$$\lambda_j \sim \mathrm{G}\big(n_j/2, \chi_j^2(\theta)/2 \big).$$

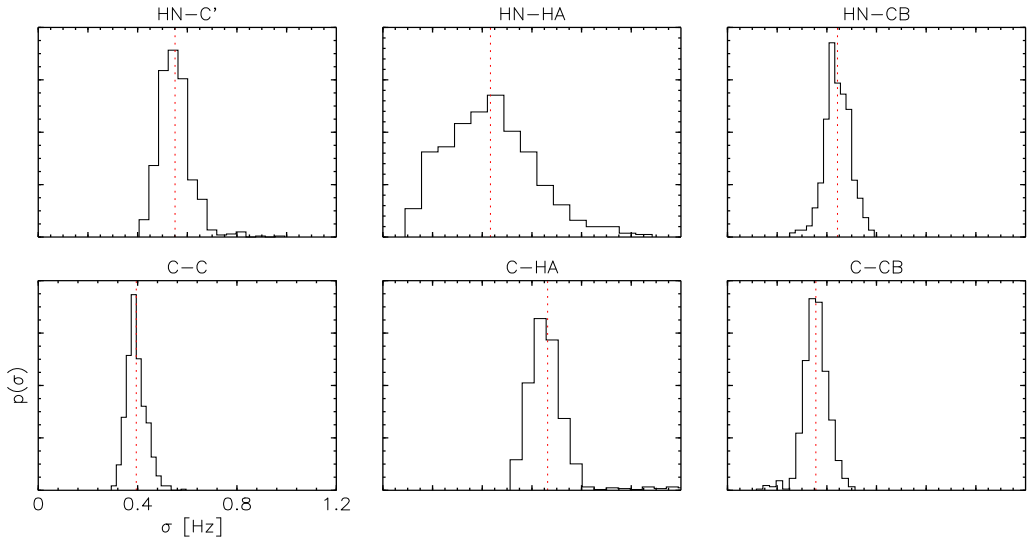


Abbildung 3.33: Marginale A posteriori-Verteilungen der Fehler σ_j (schwarze Kurve) der sechs skalaren Kopplungskonstanten von Ubiquitin (siehe Abschnitt 3.5.2) mit ihren Mittelwerten (rote gestrichelte Linien).

Der bedingte Erwartungswert von λ_j ist (siehe Anhang A.3):

$$\langle \lambda_j | \theta \rangle = n_j \,/\, \chi_j^2(\theta).$$

Der Erwartungswert des Gewichts eines Datensatzes ist also gleich der reziproken mittleren Abweichung zwischen den Messungen und ihren Vorhersagen. Damit sind λ_j und χ_j^2 komplementäre Größen.

Nach jeder Ziehung einer konformationellen Stichprobe mittels Hybrid-Monte-Carlo werden m unabhängige Stichproben der Gewichte von den zugehörigen Gamma-Verteilungen gezogen. Die Analyse jedes weiteren Datensatzes bringt einen neuen unbekannten Fehler σ_j mit sich. Der zusätzliche Rechenaufwand, um diesen Parameter ebenfalls aus den Daten zu schätzen, ist vernachlässigbar.

Dagegen wird der Aufwand bei der Kreuzvalidierung mit jedem zusätzlichen Gewicht potenziert. Strukturen müssen für jede Kombination der Gewichte $\{\lambda_1, \ldots, \lambda_m\}$ berechnet werden. Dazu werden ihre Wertebereiche diskretisiert. Der Aufwand wächst exponentiell mit der Anzahl der Datensätze.

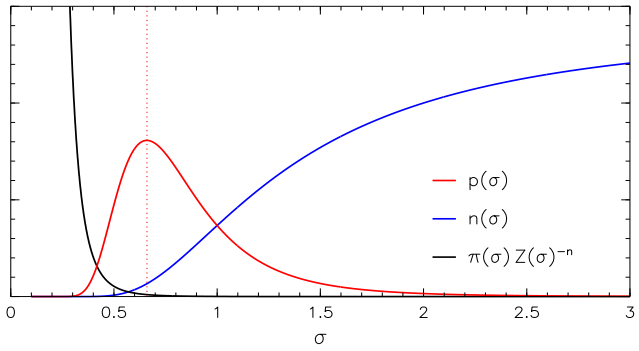


Abbildung 3.34: Beiträge zur A posteriori-Verteilung des Fehlers im Beispiel von Alanin. Als gestrichelte Linie wurde das Maximum der A posteriori-Verteilung eingezeichnet.

Daß die Schätzung mehrerer Gewichte in der Praxis möglich ist, zeigen die Beispiele aus Abschnitt 3.5. Abbildung 3.33 zeigt die marginalen A posteriori-Verteilungen der Fehler für die sechs skalaren Kopplungskonstanten von Ubiquitin.

3.7.2 Beiträge zur A posteriori-Verteilung

Eine Bestimmung der Gewichte ist in der Bayes'schen Analyse möglich, weil sich die marginale A posteriori-Verteilung eines Fehlers aus zwei Faktoren, einem monoton fallenden und einem monoton wachsenden, zusammensetzt:

$$p(\sigma) \propto \pi(\sigma)\ [Z(\sigma)]^{-n}\ n(\sigma),$$

wobei die Daten in die Funktion

$$n(\sigma) = \int \mathrm{d}\theta\ \pi(\theta)\ \exp\left\{-\chi^2(\theta)/(2\sigma^2)\right\} \tag{3.31}$$

eingehen. Wegen $\chi^2(\theta) \geq 0$, ist

$$\exp\left\{-\chi^2(\theta)/(2\sigma_1^2)\right\} \geq \exp\left\{-\chi^2(\theta)/(2\sigma_2^2)\right\} \quad \text{falls} \quad \sigma_1 > \sigma_2$$

und damit $n(\sigma)$ allgemein streng monoton wachsend.

Für das Beispiel von Alanin sind die drei Beiträge in Abbildung 3.34 gezeigt. Die Normierungskonstante $Z(\sigma)$ und die A priori-Verteilung $\pi(\sigma)$

fallen monoton; der Datenbeitrag $n(\sigma)$ wächst monoton; so daß ihr Produkt $p(\sigma)$ ein Maximum hat.

Ohne wahrscheinlichkeitstheoretische Modellierung sind die Terme $Z(\sigma)$ und $\pi(\sigma)$ nicht begründbar: $Z(\sigma)$ folgt aus der Normierung der Datenverteilung bezüglich der Messungen; $\pi(\sigma)$ beschreibt Vorwissen über den Fehler. Die Grundregeln der Wahrscheinlichkeitsrechnung fordern beide Faktoren.

Erst die Gegenterme $Z(\sigma)$ und $\pi(\sigma)$ verhindern ein uneingeschränktes Anwachsen von σ und wirken so einer Überanpassung der Struktur an die Daten entgegen. Der Einfluß von $\pi(\sigma)$ ist bei genügend großer Anzahl von Messungen zwar nur noch schwach; er garantiert aber, daß selbst im Falle weniger Messungen die A posteriori-Verteilung von σ wohl definiert bleibt.

3.7.3 Angepaßte Gewichtung der Daten

Über ihre Fehler werden die Daten individuell gewichtet und damit Aussagen über die Konformation genauer. Die Gewichte λ_j bestimmen in den Likelihood-Termen

$$L_j(\theta, \lambda_j) = \lambda_j^{n_j/2} \exp\left\{-\frac{\lambda_j}{2}\,\chi_j^2(\theta)\right\}$$

die Ausdehnung der Maxima. Jeder Faktor $L_j(\theta, \lambda_j)$ geht in die bedingte A posteriori-Verteilung $p(\theta|\lambda_1, \ldots, \lambda_m)$ ein und schränkt die Dihedralwinkel zusätzlich ein. Wie sehr die Dihedralwinkel eingeschränkt werden, wird durch λ_j festgelegt.

Die Größe der λ_j hängt davon ab, wie verträglich die Daten miteinander sind. Betrachtet man beispielsweise einen bestimmten Dihedralwinkl φ, für den skalare Kopplungskonstanten ${}^3J_1, \ldots, {}^3J_m$ gemessen wurden, so hat seine bedingte A posteriori-Verteilung die Form

$$p(\varphi|\lambda_1, \ldots, \lambda_m) \propto \pi(\varphi) \exp\left\{-\frac{1}{2}\sum_j \lambda_j \left({}^3J_j - {}^3J_j(\varphi)\right)^2\right\},$$

wobei die A priori-Verteilung $\pi(\varphi)$ von der Einstellung der übrigen Dihedralwinkel abhängt. Jeder Likelihood-Term hat sein Maximum $\hat{\varphi}_j$ bei ${}^3J_j =$

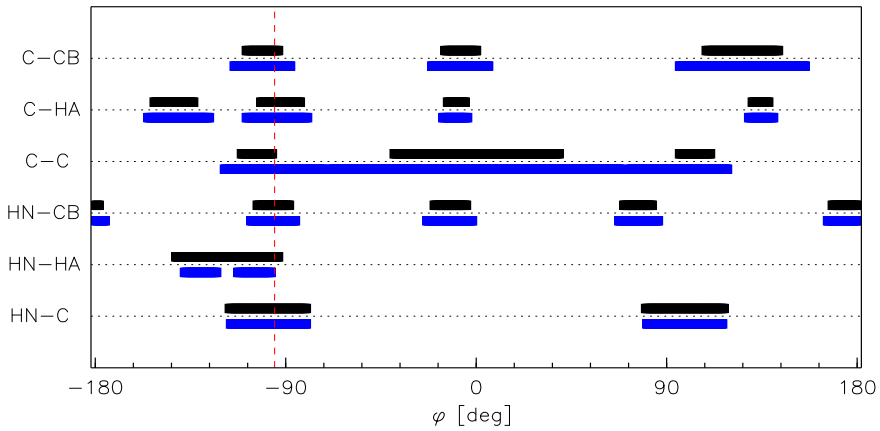


Abbildung 3.35: Bereiche hoher Wahrscheinlichkeit für die verschiedenen Messungen der skalaren Kopplungskonstanten in Ubiquitin. In Schwarz sind die Intervalle hoher Wahrscheinlichkeit für einen Dihedralwinkel (φ in LYS6) bei individueller Gewichtung der Daten gezeigt; die Intervalle, die sich aus einem einzigen gemeinsamen Gewicht ergeben, sind in Blau eingezeichnet. Die rote gestrichelte Linie ist der Wert des Winkels in der Kristallstruktur.

${}^3J_j(\hat{\varphi}_j)$; die Ausdehnung des Maximums ist antiproportional zum zugehörigen Gewicht.

In Abbildung 3.35 ist der Einfluß der Gewichte auf die Bestimmung der Dihedralwinkel für den Fall skalarer Kopplungskonstanten dargestellt. Anhand der Kristallstruktur wurden die Karplus-Koeffizienten und die Gewichte λ_j geschätzt. Die Bereiche mit $\lambda_j({}^3J_j - {}^3J(\varphi))^2/2 \leq 0.5$ wurden für einen der φ-Winkel bestimmt. Es sind die Fälle individueller und gemeinsamer Gewichtung ($\lambda_j = \lambda$) gezeigt. Meist sind die Winkelintervalle bei individueller Gewichtung kleiner und erlauben, den Dihedralwinkel genauer zu bestimmen. Im Falle der ${}^3J(\mathrm{H_N}\text{-}\mathrm{H}_\alpha)$-Kopplungen ist dagegen das Intervall bei individueller Gewichtung größer als bei einem einzigen globalen Gewicht und verhindert hier, eine Überanpassung an die Daten.

Dieselben Argumente lassen sich auch auf andere Meßgrößen anwenden. In der traditionellen Strukturbestimmung werden NOESY-Volumina, die für dasselbe Paar von Atomen gemessen wurden, oft auf eine einzige Messung

reduziert (merging) [14], indem man nur das NOESY-Signal mit dem größten Volumen berücksichtigt, weil dieses die engste Distanzschranke liefert. Dadurch wird aber die Information, die in den verworfenen Messungen enthalten ist, vernachlässigt. In der inferentiellen Strukturbestimmung geht jedes Volumen in die A posteriori-Verteilung ein, gewichtet gemäß der Qualität des Spektrums.

In den einfachen, hier behandelten Modellen ist es möglich, über die Gewichte zu integrieren und sie auf diese Weise zu entfernen. Die reduzierte Likelihood-Funktion des j-ten Datensatzes ist

$$L_j(\theta) = \int \mathrm{d}\lambda_j \; \lambda_j^{n_j/2-1} \exp\left\{-\lambda_j \, \chi_j^2(\theta)/2\right\} = \Gamma(n_j/2) \left[\, \chi_j^2(\theta) \,\right]^{-n_j/2}.$$

Marginalisierung der Gewichte entspricht einer Mittelung über alle Likelihood-Funktionen, die sich bei verschiedenen Werten der Gewichte ergeben.

3.8 Parametrisierung der Karplus-Kurve

Um Messungen der skalaren Kopplungskonstanten in der Strukturberechnung verwenden zu können, muß die Parametrisierung der Karplus-Kurve bekannt sein. Dieses Problem wird gemeinhin behandelt durch eine empirische Parametrisierung anhand einer bekannten Struktur (beispielsweise [54, 55]) oder durch „selbstkonsistente“ Schätzung der Karplus-Koeffizienten [56].

3.8.1 Parametrisierung bei bekannter Struktur

Gewöhnlich werden die unbekannten Parameter der Karplus-Kurve an einer bekannten Struktur geeicht. In einer anschließenden Strukturberechnung wird die geeichte Kurve auf die unbekannte Struktur übertragen. Entweder gehen Abweichungen zwischen gemessenen und theoretischen Kopplungen direkt in die Hybridenergie ein oder die Karplus-Kurve wird invertiert, so daß man zulässige Dihedralwinkelbereiche erhält (siehe [57]).

Liegt bereits eine Kristallstruktur mit Dihedralwinkeln $\hat{\theta}_i$ vor, so werden die Parameter der Karplus-Kurve an die Messungen ${}^3J_i$ angepaßt, indem man

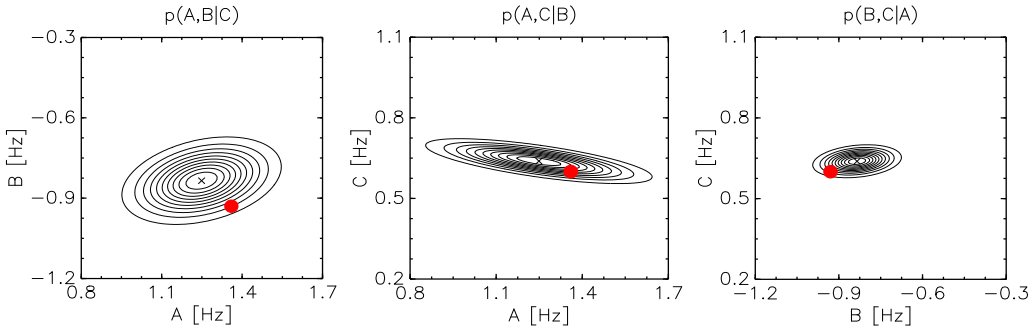


Abbildung 3.36: Schnitte durch die marginale A posteriori-Verteilung der Karplus-Koeffizienten. Die Maxima der Verteilung entsprechen einer Lösung des überbestimmten Gleichungssystems durch eine Singulärwertzerlegung der Matrix der trigonometrischen Basisfunktionen. Die roten Punkte sind die von Bax und Mitarbeitern bestimmten Werte, welche dem Datensatz 1d3z entnommen wurden.

die mittlere quadratische Abweichung

$$\chi^2(A,B,C) = \sum_i \left({}^3J_i - {}^3J(\hat{\theta}_i, A, B, C) \right)^2$$

zwischen gemessenen Kopplungen und Vorhersagen ${}^3J(\hat{\theta}_i, A, B, C)$ minimiert. Das resultierende überbestimmte Gleichungssystem wird durch eine Pseudo-Inverse nach A, B, C aufgelöst. Beispielsweise sind Wang und Bax [55] in ihrer Analyse der Messungen an Ubiquitin so vorgegangen.

Das Modell aus Abschnitt 3.1.1 enthält diese Vorgehensweise: Die Methode der kleinsten Quadrate entspringt der Beschreibung von Fehlern durch eine Gauß-Verteilung. Bei bekannter Struktur und bekanntem Fehler σ ist die bedingte A posteriori-Verteilung der Karplus-Koeffizienten eine dreidimensionale Gauß-Glocke, die um

$$\hat{\mathbf{a}} = \mathbf{A}_J^{\dagger}\mathbf{j} = \mathbf{C}\mathbf{A}_J^T\mathbf{j}$$

zentriert ist mit der Kovarianzmatrix

$$\mathbf{\Sigma} = \sigma^2\,\mathbf{C} = \sigma^2\,(\mathbf{A}_J^T\mathbf{A}_J)^{-1}$$

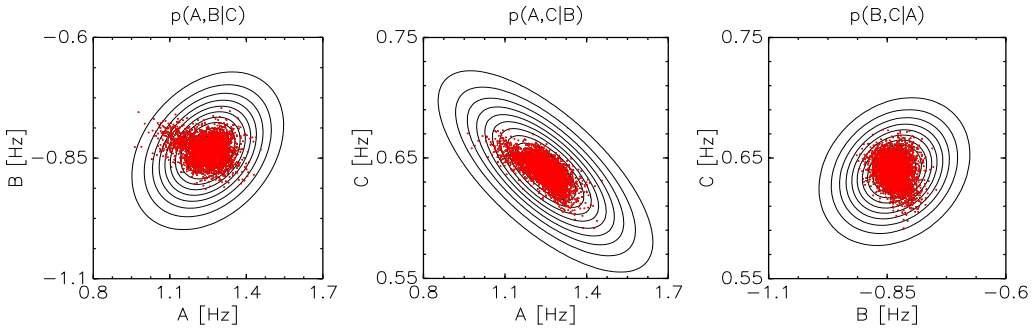


Abbildung 3.37: Konturlinien der marginalen A posteriori-Verteilung der Karplus-Koeffizienten (siehe auch Abb. 3.36). Die roten Punkte sind Schätzwerte, die man durch Weglassen von 10 % zufällig gewählten Messungen erhält.

(siehe Abschnitte 3.1.1 und 3.2.1). $\hat{\mathbf{a}}$ minimiert die Abweichung $\chi^2(A,B,C)$. Aus einer Singulärwertzerlegung der Matrix $\mathbf{A}_J$ der Basisfunktionen $\cos^k\theta$, $k = 0, 1, 2$ können diese Schätzwerte leicht berechnet werden.

In Abbildung 3.36 ist die A posteriori-Verteilung der Karplus-Koeffizienten gezeigt, die sich aus dem Datensatz der ${}^3J$(C-C)-Kopplungen in Ubiquitin aus der Kristallstruktur ergibt. Integration über die Fehlerskala σ liefert eine dreidimensionale t-Verteilung:

$$p(A,B,C|D,I) = \int \mathrm{d}\sigma \; p(A,B,C,\sigma|D,I) \propto \left[\chi^2_J(A,B,C)\right]^{-n/2}$$

Sie ist wie die Gauß-Verteilung um den Schätzwert $\hat{\mathbf{a}}$ zentriert nur etwas weniger scharf bestimmt, weil Integration über den Fehler die Verteilung ausschmiert. Es werden jeweils zweidimensionale Schnitte durch die marginale A posteriori-Verteilung der Karplus-Parameter gezeigt. In diesem Beispiel kann die Verbundverteilung der Theorieparameter analytisch angegeben werden. Die Form der Konturlinien zeigt, daß die Parameter korreliert sind.

Werden die Karplus-Koeffizienten lediglich durch Minimierung von χ^2_J bestimmt, so lassen sich keine Angaben über ihre Genauigkeit machen. Häufig schätzt man die Fehler ab, indem man, ähnlich wie in der Kreuzvalidierung, einen Teil der Messungen wegläßt und die Parameter aufgrund der reduzierten Daten bestimmt. Dies wird viele Male mit anderen, zufällig gewähl-

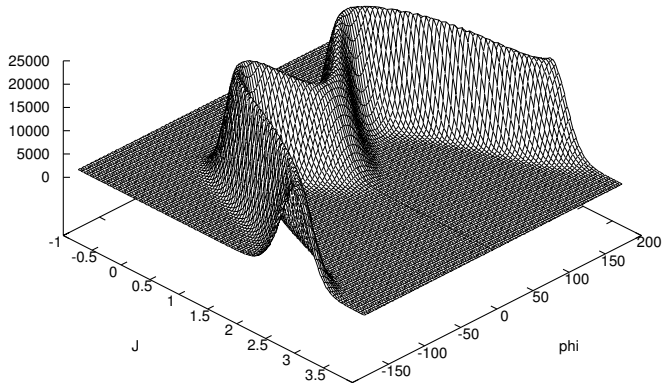


Abbildung 3.38: Aus den C-C-Kopplungen von Ubiquitin und der Kristallstruktur vorhergesagte Karplus-Kurve $p({}^3J|\varphi, D, I)$.

ten Daten wiederholt, so daß man schließlich eine „Verteilung“ der Karplus-Koeffizienten erhält (siehe bspw. [58, 55]). Mit Hilfe dieser Heuristik möchte man ein Analogon zur A posteriori-Verteilung der Karplus-Koeffizienten berechnen (siehe Abb. 3.37).

Die A posteriori-Verteilung ist jedoch eindeutig durch den Bayes'schen Satz festgelegt; Heuristiken wie die beschriebene sind überflüssig. In der Bayes'schen Theorie werden Parameter nicht durch einen Schätzwert bestimmt, sondern durch eine Verteilung ihrer möglichen Werte. Eine Punktschätzung (wie das Minimum von χ^2_J) ergibt lediglich einen Wert und keine Angabe über dessen Genauigkeit. Deshalb versucht man, durch zufällige Ziehung der Messungen eine mehrmalige Wiederholung des Experiments zu simulieren. Dem liegt die Denkweise zugrunde, daß Wahrscheinlichkeiten das Resultat von Zufallsprozessen sind und nicht die quantitative Repräsentation von Wissen. Mit Durchführung des Experiments sind die Daten jedoch festgelegt; es bedarf keiner Modelle über hypothetische Variationen in den Messungen. Das Problem liegt in der Annahme, daß sich mit Anwendung einer Punktschätzung eindeutige Schlüsse aus den Daten ziehen ließen.

Unser Mangel an Wissen führt dazu, daß kein eindeutiger Zusammenhang zwischen Messungen und Hypothesenparametern besteht. Im Falle skalarer Kopplungen manifestiert sich diese Unsicherheit direkt in den Vorhersagen

weiterer Messungen bei gegebenem Dihedralwinkel. Integration über die Koeffizienten und Fehler liefert eine verteilte Karplus-Kurve

$$p({}^3J|\varphi, D, I) = \int \mathrm{d}A\,\mathrm{d}B\,\mathrm{d}C\,\mathrm{d}\sigma\; p({}^3J|\varphi, A, B, C, \sigma)\, p(A, B, C, \sigma|D, I),$$

wobei in den Daten Messungen der skalaren Kopplungskonstante sowie die Koordinaten der Kristallstruktur enthalten sind. In Abbildung 3.38 ist die vorgesagte Karplus-Kurve $p({}^3J|\varphi, D, I)$ abgebildet, welche sich aus den ${}^3J$(C-C)-Kopplungen und der Kristallstruktur ergibt.

3.8.2 Selbstkonsistente Parametrisierung

Bei einer unbekannten Struktur ist die direkte Bestimmung der Karplus-Koffizienten nicht möglich. Doch selbst wenn eine Kristallstruktur vorliegt, ist der Zugang problematisch: Erstens werden skalare Kopplungen an Molekülen in Lösung nicht im Kristall gemessen; wenn sich die Konformationen systematisch unterscheiden, wird die Karplus-Kurve verzerrt; es ist schwer abzuschätzen, ob Abweichungen von Meßfehlern oder strukturellen Unterschieden herrühren. Zweitens sind die Dihedralwinkel während der Messung beweglich; man mißt nur zeitlich gemittelte Kopplungen (dies haben beispielsweise Brüschweiler und Case [59] versucht zu berücksichtigen).

Wegen dieser Schwierigkeiten haben Schmidt et al. [56] vorgeschlagen, auf die Verwendung einer Kristallstruktur zu verzichten und Messungen skalarer Kopplungskonstanten „selbstkonsistent“ zu analysieren. Dadurch soll vermieden werden, die Eigenschaften des Moleküls in kristalliner Form auf seine Eigenschaften in Lösung zu übertragen.

Die Methode geht von m skalaren Kopplungen unterschiedlichen Typs aus, die denselben Dihedralwinkel θ betreffen; meist ist dies der φ Winkel, der durch die Atome C_{-1}, N, C_α und C des Proteinrückgrats definiert wird und an sechs Kopplungen beteiligt sein kann. Jeder Kopplungstyp ${}^3J_j$ wird durch eine eigene Karplus-Kurve beschrieben:

$${}^3J_j(\theta) = A_j\,\cos^2(\theta + \delta_j) + B_j\,\cos(\theta + \delta_j) + C_j, \quad j = 1, \ldots, m.$$

Die Phasen δ_j, welche die Kopplungen durch die verschiedenen Bindungen auf denselben Dihedralwinkel zurückführen, sind bei starrer kovalenter Struktur bekannt.

Der Winkel θ wird gemeinsam durch die verschiedenen Kopplungen festgelegt und wäre bei bekannter Parametrisierung der Karplus-Kurven sogar überbestimmt. Andererseits können natürlich bei bekannten Dihedralwinkeln θ_i entlang des Proteinrückgrats die Parametrisierungen der Karplus-Kurven bestimmt werden, denn es muß im Idealfall

$$^3J_{i,j} = A_j \cos^2(\theta_i + \delta_j) + B_j \cos(\theta_i + \delta_j) + C_j$$

gelten. Wenn genügend Meßwerte $^3J_{i,j}$ vorliegen, können alle Unbekannten, sowohl die Parametrisierungen $\{A_j, B_j, C_j\}$ als auch die Dihedralwinkel $\{\theta_1, \ldots, \theta_n\}$, bestimmt werden.

Schmidt et al. [56] haben einen iterativen Algorithmus entwickelt, der die mittlere quadratische Abweichung zwischen gemessenen und theoretischen Kopplungen minimiert. Das Optimierungsproblem ist linear in den Karplus-Koeffizienten und nicht-linear in den Dihedralwinkeln. Während der Minimierung können die Dihedralwinkel alle Werte annehmen; Vorwissen über die molekulare Struktur wird nicht integriert. Das Modell kann dadurch erweitert werden, daß man den Dihedralwinkeln eine Fluktuation unbekannter Größe zugesteht, die ebenfalls optimiert wird.

In der Wahrscheinlichkeitsrechnung garantieren die Cox'schen Theoreme Konsistenz; somit zielt der Ansatz von Schmidt et al. auf etwas ab, das erst in der inferentiellen Strukturbestimmung volle Gültigkeit erlangt.

Das Schmidt'sche Verfahren ist ein Spezialfall der Bayes'schen Formulierung: Die Parametrisierungen $\{A_j, B_j, C_j\}$ der Karplus-Kurven $^3J_j(\theta)$ sind unbekannt und werden neben den eigentlichen Hypothesenparametern, den Dihedralwinkeln θ_i, durch die A posteriori-Verteilung bestimmt. Simulation der gemeinsamen A posteriori-Verteilung $p(\theta_1, \ldots, \theta_n, \{A_j, B_j, C_j\})$ erlaubt, sowohl die Koeffizienten der Karplus-Kurven als auch die Dihedralwinkel zu schätzen. Weil die Abweichungen der skalaren Kopplungen mit einem Gauß'schen Fehlermodell beschrieben wurden, hat die Schmidt'sche

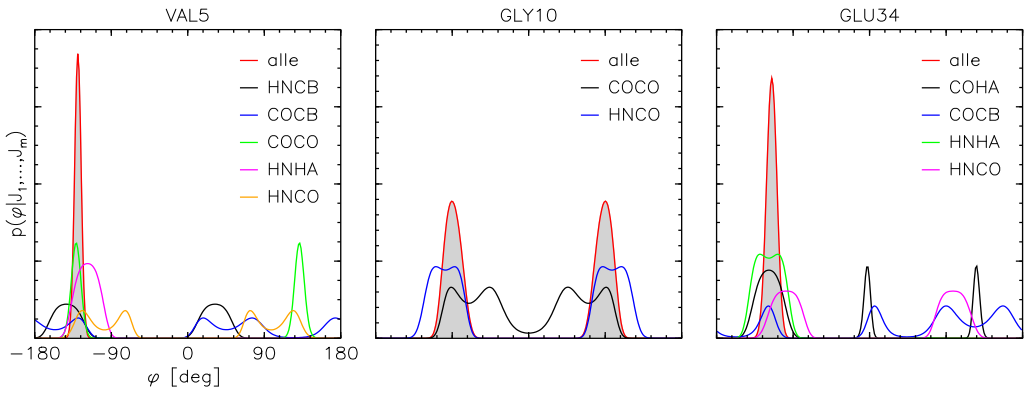


Abbildung 3.39: Bedingte A posteriori-Verteilung eines Dihedralwinkels bei gemeinsamer Analyse von Messungen skalarer Kopplungonstanten unterschiedlichen Typs. Je nach Position in der Aminosäurekette sind die Messungen für φ_i mehr oder weniger vollständig.

Zielfunktion dieselbe funktionelle Form wie der negative Logarithmus der A posteriori-Verteilung.

Beim Gibbs sampling werden Stichproben der Dihedralwinkel von bedingten A posteriori-Verteilungen der Form

$$p(\theta_i|{}^3J_{i,1},\ldots,{}^3J_{i,m})$$

gezogen, wobei ${}^3J_{i,j}$ hier Messungen der Kopplungen sind, die den Dihedralwinkel θ_i betreffen. Die Anzahl der Messungen für die jeweiligen Dihedralwinkel variiert von Winkel zu Winkel. Für drei Beispiele sind die bedingten A posteriori-Verteilungen abgebildet, die sich aus den sechs Datensätzen von Ubiquitin ergeben (Abb. 3.39).

Abbildung 3.40 illustriert, daß bei $\beta = 0$ die Bayes'sche Analyse äquivalent zur Methode von Schmidt et al. ist: Durch gemeinsame Schätzung der Dihedralwinkel und der unbekannten Parametrisierungen der Karplus-Kurven, werden die skalaren Kopplungen selbstkonsistent analysiert. Die A posteriori-Verteilungen streuen um die Schätzwerte, die sich aus der Analyse der Kristallstruktur ergeben.

Zusätzlich zu den Parametern werden auch die Dihedralwinkel bestimmt,

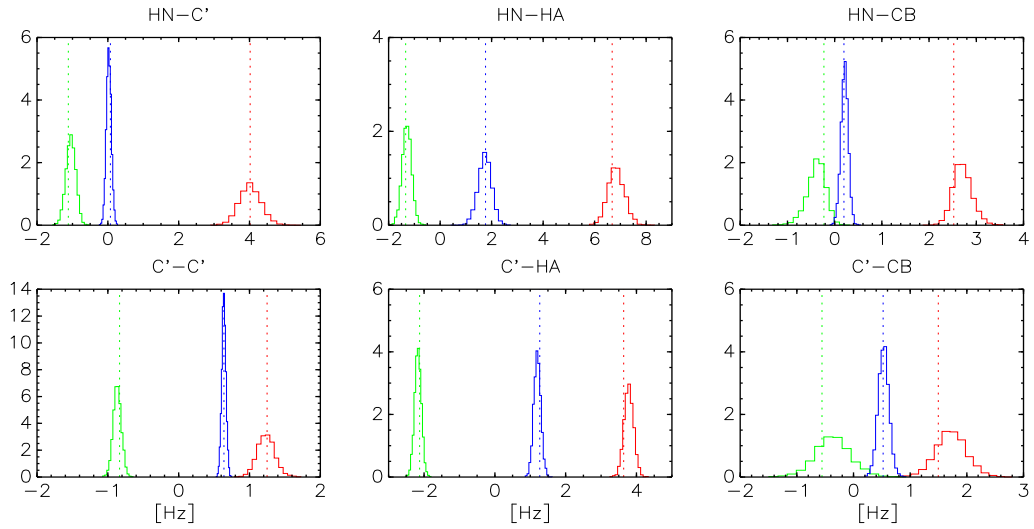


Abbildung 3.40: Marginale A posteriori-Verteilungen der Karplus-Parameter bei gleichzeitiger Schätzung der Dihedralwinkel und der zusätzlichen Parameter (rot: $p(A|D,I)$, grün: $p(B|D,I)$, blau: $p(C|D,I)$). Die A priori-Verteilung der Dihedralwinkel ist eine Gleichverteilung, d.h. im Boltzmann-Ensemble ist $\beta = 0$. Die Schätzwerte aus der Kristallstruktur wurden als gestrichelte Linien eingetragen.

die durch Bindungen definiert sind, für die skalare Kopplungen gemessen wurden. Abbildung 3.41 zeigt die Konfidenzintervalle der φ-Winkel. Die Intervalle wurden aufgrund der marginalen A posteriori-Verteilungen $p(\varphi_i|D,I)$ bei einer Konfidenz von 68 % berechnet. In den meisten Fällen ist der Dihedralwinkel der Kristallstruktur in diesem Intervall enthalten oder liegt in seiner Nähe. Obwohl die Paramterisierungen der Karplus-Kurven unbekannt sind, können allein aus den skalaren Kopplungen die φ-Winkel gut rekonstruiert werden. In manchen Fällen ist die marginale A posteriori-Verteilung $p(\varphi_i|D,I)$ multimodal: die Konfidenzbereiche hängen nicht zusammen. Dies betrifft Winkel, für die nicht alle der sechs möglichen Kopplungen gemessen wurden, und deren A posteriori-Verteilung deshalb mehrdeutig bleibt.

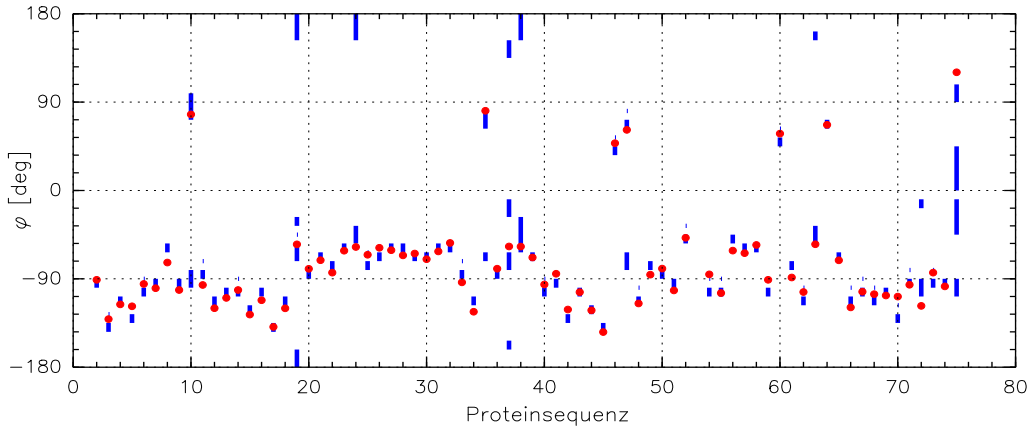


Abbildung 3.41: Konfidenzintervalle aller φ-Dihedralwinkel, aufgetragen gegen die Nummer des zugehörigen Aminosäurerests (blau). Die Intervalle wurden bei einer Konfidenz von 68 % aus den marginalen A posteriori-Verteilungen $p(\varphi_i|D,I)$ der Dihedralwinkel φ_i abgeleitet. Die Dihedralwinkel der Kristallstruktur sind als rote Kreise eingetragen.

3.8.3 Marginalisierung der Karplus-Koeffizienten

Vom Standpunkt der Strukturbestimmung interessieren in erster Linie die Dihedralwinkel und nicht die Parametrisierungen der Karplus-Kurven. Durch Integration lassen sich die Karplus-Koeffizienten aus der A posteriori-Verteilung entfernen, so daß man während der Strukturberechnung nicht mehr auf die Koeffizienten zurückgreifen muß (siehe Abschnitt 3.2.1). Die Daten werden durch eine reduzierte Likelihood-Funktion

$$L(\theta) \;\propto\; \prod_j \left|\mathbf{A}_j^T\mathbf{A}_j\right|^{-1/2} \left[\mathbf{j}_j^T(\mathbf{I}-\mathbf{A}_j\mathbf{A}_j^\dagger)\,\mathbf{j}_j\right]^{-(n_j-3)/2} \tag{3.32}$$

beschrieben. Sie hängt nur noch von den Dihedralwinkeln über die Matrizen $(\mathbf{A}_j)_{kl} = [\cos(\theta_l-\delta_j)]^{2-k}$, $l=1,\ldots,n_j$, $k=0,1,2$ ab.

Zur Veranschaulichung sind in Abbildung 3.42 die Konturlinien der reduzierten Likelihood-Funktion (3.32) für ein Paar von Dihedralwinkeln gezeigt. Die Messungen der einzelnen Kopplungstypen resultieren in mehrdeutigen Verteilungen. Erst die Gesamtheit aller Messungen vermag die Mehrdeutigkeit aufzulösen. Die marginale Likelihood-Funktion könnte dazu verwendet

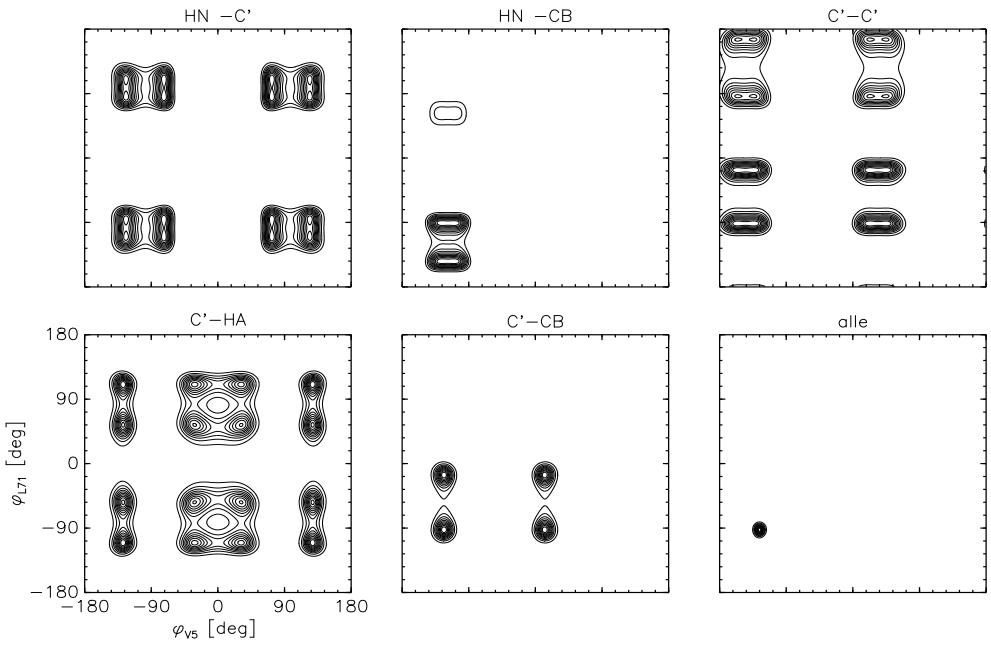


Abbildung 3.42: Marginale A posteriori-Verteilungen zweier Dihedralwinkel. Für beide Dihedralwinkel wurden fünf der sechs möglichen Kopplungen gemessen. Die Konturlinien der marginalen Likelihood-Funktionen, die sich aus den einzelnen Datensätzen ergeben, bleiben mehrdeutig. Erst die gemeinsame Likelihood-Funktion aller Daten (rechts unten) vermag, die Mehrdeutigkeit aufzulösen.

werden, die Dihedralwinkel ohne Kenntnis der Karplus-Koeffizienten zu bestimmen.

3.9 Bestimmung der mittleren Orientierung

In anisotroper Umgebung kann sich ein Makromolekül im Mittel ausrichten. Die Elemente des Orientierungstensors sind unbekannt, so daß bei Verwendung dipolarer Kopplungen zur Strukturbestimmung ein ähnliches Problem wie bei den Koeffizienten der Karplus-Kurve besteht.

Mit diesem Problem wird unterschiedlich umgegangen: (i) Losonczi et al. [60] berechnen die Saupe-Matrix aus einer bekannten Struktur mittels einer

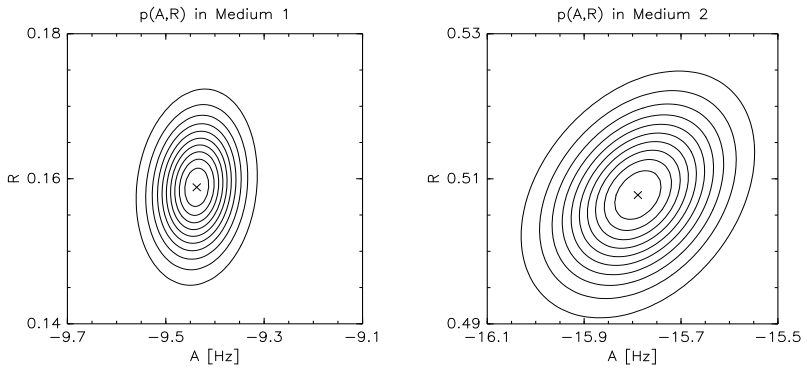


Abbildung 3.43: Bedingte A posteriori-Verteilungen $p(A, R|\theta, \hat{\mathbf{U}})$ der axialen und der rhombischen Komponente in beiden Medien, in denen die dipolaren Kopplungen von Ubiquitin gemessen wurden. Es wurde die NMR-Struktur von Cornilescu et al. [37] (PDB-code 1d3z) verwendet. Die Rotation $\hat{\mathbf{U}}$ diagonalisiert die geschätzte Saupe-Matrix $\hat{\mathbf{S}}$. Die Maxima der A posteriori-Verteilung sind als Kreuze eingezeichnet. Der Fehler wurde ausintegriert.

Singulärwertzerlegung, (ii) Clore et al. [61] haben vorgeschlagen, den axialen und den rhombischen Anteil der Ausrichtung anhand des Histogramms der gemessenen dipolaren Kopplungen abzuschätzen, (iii) Hus et al. [62] bestimmen die Orientierungen der Peptidebenen und die Saupe-Matrizen gemeinsam und verwenden diese Schätzwerte anschließend in einer Strukturverfeinerung, (iv) Moltke und Grzesiek [63] entfernen die Saupe-Matrix analytisch aus der Hybridenergie.

3.9.1 Methode der kleinsten Quadrate

Modelliert man Abweichungen zwischen gemessenen und berechneten dipolaren Kopplungen mit der Methode der kleinsten Quadrate, so werden die fünf Elemente der Saupe-Matrix durch das lineare Gleichungssystem

$$\mathbf{A}_D\,\mathbf{s} = \mathbf{d}$$

bestimmt (siehe Abschnitt 3.1.2 für Details und Definitionen). Die Matrix $\mathbf{A}_D$ hängt von der Konformation ab und ist bei bekannter Struktur gege-

ben. Multiplikation mit der Pseudo-Inversen ergibt die Lösung: $\mathbf{s} = \mathbf{A}_D^{\dagger}\mathbf{d}$. Die Pseudo-Inverse kann über eine Singulärwertzerlegung berechnet werden. Losonczi et al. [60] haben vorgeschlagen, die Saupe-Matrix auf diese Weise aus einer bekannten Struktur zu bestimmen.

Die Methode von Losonczi et al. entspricht einer Bestimmung der Saupe-Matrix über die bedingte A posteriori-Verteilung $p(\mathbf{S}|\theta, D, I)$ und wird beim Gibbs sampling aller Hypothesenparameter implizit angewendet. Die bedingte A posteriori-Verteilung der Elemente der Saupe-Matrix ist eine fünfdimensionale Gauß-Glocke (siehe Abschnitt 3.2.2). Die Verteilung der Elemente s_i läßt sich auf die axiale und die rhombische Komponente A und R umtransformieren. Abbildung 3.43 zeigt die A posteriori-Verteilung bei bekannter Struktur. Wie bei der Analyse der skalaren Kopplungskonstanten hat die probabilistische Behandlung den Vorteil, daß die Genauigkeit der Schätzwerte direkt angeben werden kann.

3.9.2 Histogrammethode

Zur Bestimung einer unbekannten Struktur ist dieser Zugang freilich nicht brauchbar. Die Histogrammethode [61] geht von einer Gleichverteilung der Bindungsvektoren aus, für die dipolare Kopplungen gemessen wurden. Bei gegebener axialer und rhombischer Komponente und Vernachlässigung experimenteller Fehler sind die dipolaren Kopplungen gemäß

$$n(D|A,R) = \frac{1}{4\pi}\int_{-1}^{1} \mathrm{d}\cos\theta \int_{0}^{2\pi} \mathrm{d}\varphi \;\; \delta\big(D - D(\theta,\varphi)\big)$$

verteilt. Der Definitionsbereich und das Maximum der Verteilung sind bekannt: es werden Werte zwischen $-A(1+3R/2)$ und $2A$ angenommen, am wahrscheinlichsten ist $-A(1-3R/2)$. Durch Vergleich des Histogramms der gemessenen Werte mit der theoretischen Kurve kann man den axialen und den rhombischen Anteil bestimmen. Im einfachsten Fall liest man den Definitionsbereich und das Maximum der Verteilung ab und berechnet daraus Werte von A und R. Abbildung 3.44 zeigt Histogramme bei verschiedenen Rhombizitäten.

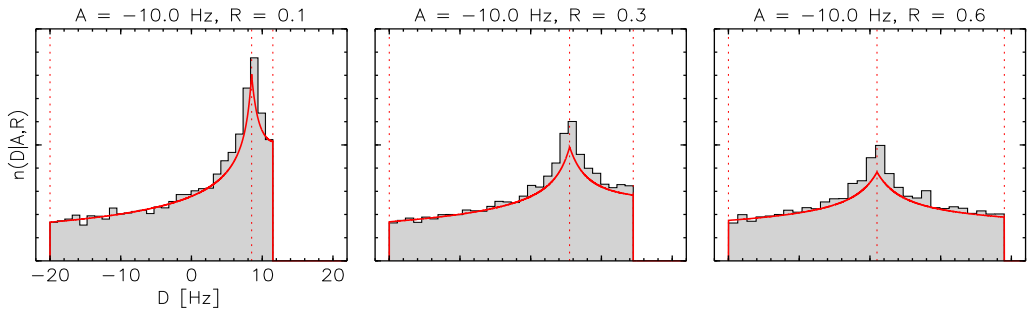


Abbildung 3.44: Analytische Entwicklung der Verteilung der dipolaren Kopplungen (rote Kurve) mit ihren charakteristischen Werten (gestrichelte Linien); aus diesen lassen sich Schätzwerte der axialen und der rhombischen Komponente berechnen. Zusätzlich sind die Histogramme eingezeichnet, die eine Monte-Carlo-Integration bei gleichverteilten Dihedralwinkeln ergibt.

Analytische Entwicklung von $n(D|A,R)$ in eine Reihe erlaubt, die Histogrammethode weiter zu verfeinern. Warren und Moore [64] bestimmen die axiale und rhombische Komponente durch Maximierung der Likelihood-Funktion

$$L(A,R) = \prod_i n(D_i|A,R).$$

Durch Faltung des idealen Histogramms mit einer Gauß-Verteilung können zudem experimentelle Ungenauigkeiten berücksichtigt werden.

Die Histogrammethode ergibt sich aus dem Modell aus Abschnitt 3.1.2, wenn man das Vorwissen über die Konformationen vernachlässigt ($\beta = 0$). In diesem Fall sind die Bindungsvektoren a priori gleichverteilt. Für $\sigma = 0$ ist die reduzierte Datenverteilung, welche man durch Integration über die Einstellungen der Dihedralwinkel erhält, das Histogramm $n(D|A,R)$. A Posteriori-Schätzung der Parameter A und R ist analog zur Maximum-Likelihood-Methode. Der Fehler σ berücksichtigt Ungenauigkeiten der Messungen; er wird bei Simulation der A posteriori-Verteilung $p(A,R,\sigma|D,I)$ an die Daten angepaßt. Damit ist die verfeinerte Histogrammethode von Warren und Moore [64] ein Spezialfall des Bayes'schen Zugangs. Die Annahme, daß die Bindungsvektoren isotrop verteilt sind, trifft bei gefalteten Proteinen nicht

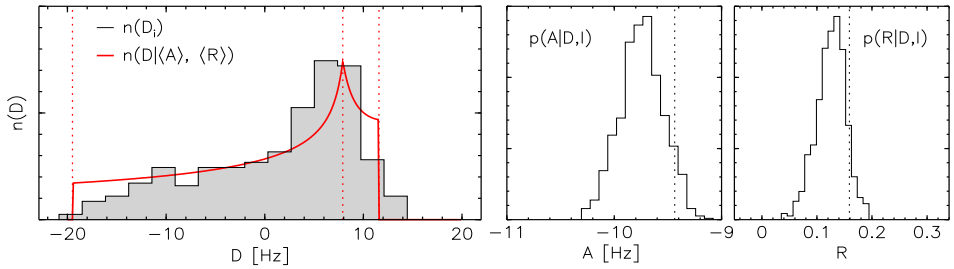


Abbildung 3.45: Analyse dipolarer Kopplungen in Ubiquitin. Links: Das Histogramm der für eine Orientierung gemessenen und normierten Kopplungen (schwarze Kurve) stimmt mit dem rekonstruierten Histogramm (rote Kurve) überein, welches sich aus den mittleren Werten der axialen und rhombischen Komponente ergibt. Marginale A posteriori-Verteilungen der axialen und rhombischen Komponente sind rechts abgebildet; gestrichelte Linien liegen bei den aus der NMR-Struktur abgeleiteten Schätzwerten (s. Abb. 3.43).

zu. Sie ist in der Bayes'schen Analyse nicht notwendig: die Bindungsvektoren werden mit dem Orientierungstensor zusammen geschätzt.

Abbildung 3.45 stellt die Ergebnisse einer solchen Analyse zusammen. Die dipolaren Kopplungen, gemessen für eine Ausrichtung, wurden analysiert ($n = 348$). Das Histogramm der normierten dipolaren Kopplungen stimmt mit dem Histogramm überein, das sich als Vorhersage für die Verteilung der dipolaren Kopplungen aus den Monte-Carlo-Stichproben ergibt. Die marginalen A posteriori-Verteilungen der axialen und rhombischen Komponente konzentrieren sich in der Nähe der Schätzwerte, die aus der NMR-Struktur 1d3z folgen.

Selbst bei wenigen Daten sind noch Aussagen über die Orientierung möglich. Für die 63 dipolaren Kopplungen des N-H_N-Bindungsvektors wurde dieselbe Analyse durchgeführt. Das vorhergesagte Histogramm (Abb. 3.46) ähnelt stark dem, das eine Analyse aller Daten ergibt. Die marginalen A posteriori-Verteilungen der axialen Komponente des Orientierungstensors streuen zwar stärker als in der Analyse aller Kopplungen, sind aber ungefähr um dieselben Werte konzentiert.

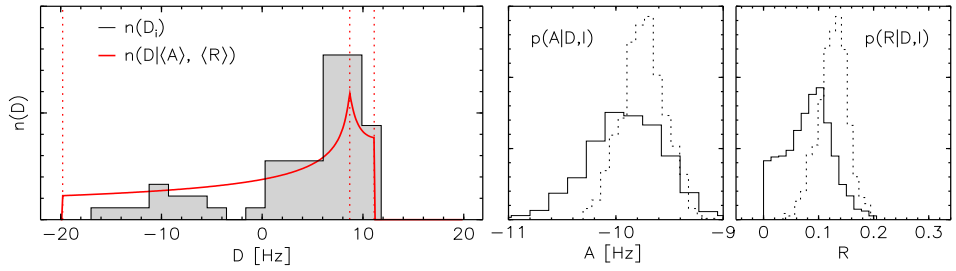


Abbildung 3.46: Ergebnisse der gleichen Analyse wie in Abb. 3.45, wobei als Daten lediglich die 63 Kopplungen der N-H_N-Bindung verwendet wurden. In die marginalen A posteriori-Verteilungen auf der rechten Seite wurden gestrichelt die Histogramme eingezeichnet, die aus einer Analyse des vollen Datensatzes resultieren.

3.9.3 Berechnung der Orientierung und der Struktur

Hus et al. [62] haben einen Algorithmus entwickelt, um nicht nur die Saupe-Matrizen, sondern auch die Orientierungen der Peptidebenen aus Messungen dipolarer Kopplungen in verschiedenen Lösungsmitteln zu bestimmen. Eine anschließende Strukturberechnung benutzt die ermittelten Schätzwerte. Auf diese Weise konnte die Faltung eines Proteins allein aus Messungen dipolarer Kopplungen berechnet werden.

Diese Strategie ist in der Bayes'schen Formulierung enthalten. Hier werden die Peptidebenen indirekt über die Dihedralwinkel parametrisiert. Simulation der vollen A posteriori-Verteilung

$$p(\theta, \mathbf{S}_1, \ldots, \mathbf{S}_m | D, I)$$

mit jeweils einer Saupe-Matrix $\mathbf{S}_j$ für jede Orientierung liefert Schätztwerte der mittleren Orientierungen und der Dihedralwinkel.

Abbildung 3.47 zeigt die Ergebnisse für die Ubiquitin-Daten. Die marginalen A posteriori-Verteilungen der axialen und rhombischen Anteile liegen in der Nähe der Schätzwerte, die aus NMR-Struktur abgeleitet wurden.

Neben den Saupe-Matrizen werden die Dihedralwinkel geschätzt. Die A posteriori-Verteilungen der Dihedralwinkel können multimodal bleiben,

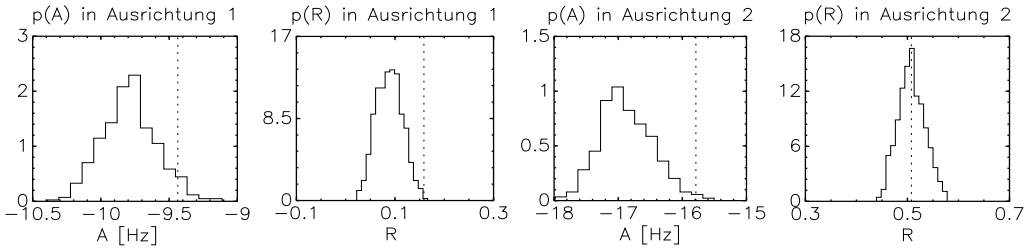


Abbildung 3.47: Marginale A posteriori-Verteilungen der axialen und der rhombischen Anteile für die dipolaren Kopplungen, gemessen an Ubiquitin in zwei verschiedenen Medien. Die A priori-Verteilung der Dihedralwinkel war eine Gleichverteilung ($\beta = 0$). Die gestrichelten Linien zeigen die Schätzwerte an, die eine Analyse der NMR-Struktur liefert.

weil sich die dipolaren Kopplungen nicht eindeutig umkehren lassen. Verwendung dipolarer Kopplungen allein reicht nicht aus, um kompakte Konformationen zu bestimmen. Durch zusätzliche Berücksichtigung von Vorwissen werden die Dihedralwinkel genauer festgelegt. In Abbildung 3.48 sind Stichproben der Dihedralwinkel für eine Simulation bei $\beta^{-1} = 0\,\mathrm{K}$ und $\beta^{-1} = 300\,\mathrm{K}$ zu sehen. Bei Vernachlässigung von Vorwissen über die erlaubten Konformationen weist die Verteilung der Dihedralwinkel eine Symmetrie auf, die Mehrdeutigkeiten der dipolaren Kopplungen widerspiegelt. Diese Symmetrie kann durch Verwendung zusätzlichen Vorwissens zu einem gewissen Grad gebrochen werden.

3.9.4 Analytische Eliminierung der Saupe-Matrix

Moltke und Grzesiek [63] entledigen sich der Notwendigkeit, zur Analyse der dipolaren Kopplungen die mittlere Orientierung des Moleküls kennen zu müssen, indem sie die unbekannten Parameter aus der Zielfunktion für Dihedralwinkel und Orientierungen analytisch entfernen. Dies ist möglich, weil das Optimierungsproblem linear in den Elementen der Saupe-Matrix ist: Die Zielfunktion ist die χ^2-Abweichung (3.7); Ableitung nach den Elementen s_i der Saupe-Matrix ergibt ein lineares Gleichungssystem für den Orientierungs-

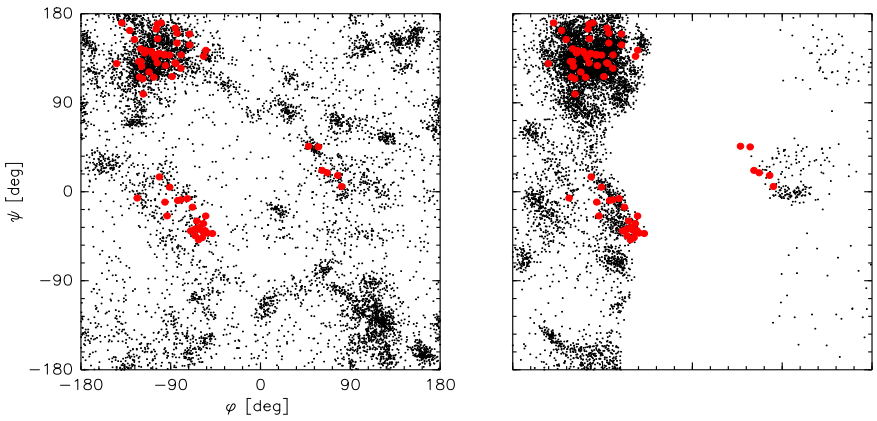


Abbildung 3.48: Stichproben der φ-, ψ-Winkel, die sich aus einer Analyse der dipolaren Kopplungen von Ubiquitin ergeben. Links: Dihedralwinkel sind a priori gleichverteilt $\beta^{-1} = 0\,\mathrm{K}$, rechts: Simulation bei $\beta^{-1} = 300\,\mathrm{K}$. Als rote Kreise wurden die Dihedralwinkel der Kristallstruktur eingetragen.

tensor:

$$\mathbf{A}_D\,\mathbf{s} = \mathbf{d},$$

wobei die $n \times 5$-Matrix $\mathbf{A}_D$ von den Dihedralwinkeln abhängt.

Das Gleichungssystem kann durch Multiplikation mit der Pseudo-Inversen nach $\mathbf{s}$ aufgelöst werden und liefert als besten Wert der Saupe-Matrix das Maximum der Likelihood-Funktion (siehe auch Abschnitt 3.1.2). Setzt man diese Lösung in $\chi^2(\theta, \mathbf{s})$ ein, so erhält man eine Zielfunktion, die nur noch auf dem Konfigurationsraum definiert ist:

$$F(\theta) = \mathbf{d}^T \left(\mathbf{I} - \mathbf{A}_D\,\mathbf{A}_D^{\dagger}\right) \mathbf{d}.$$

Minimierung von F erlaubt, die Struktur zu berechnen, ohne die Saupe-Matrix überhaupt zu kennen.

Die wahrscheinlichkeitstheoretische Entsprechung der Zielfunktion von Moltke und Grzesiek ist die marginale A posteriori-Verteilung (3.19)

$$p(\theta) = \int \mathrm{d}s_1 \cdots \mathrm{d}s_5 \;\; p(\theta, s_1, \ldots, s_5).$$

Der negative Logarithmus von $p(\theta)$ weist jedoch noch zusätzliche Terme auf:

$$\begin{aligned} -\log p(\theta) &= \beta\, E(\theta) + \frac{1}{2} \log \left| \mathbf{A}_D^T \mathbf{A}_D \right| \\ &\quad + \frac{n-5}{2} \log \left[\mathbf{d}^T \left(\mathbf{I} - \mathbf{A}_D \mathbf{A}_D^\dagger \right) \mathbf{d} \right]. \end{aligned}$$

Integration über den unbekannten Fehler σ schwächt die Abhängigkeit von den Daten weiter ab: der Anteil $F(\theta)$ aus dem Formalismus von Moltke und Grzesiek geht nur logarithmisch in die Bayes'sche Zielfunktion ein.

Wie im Falle der skalaren Kopplungskonstanten könnte $-\log p(\theta)$ direkt zur Berechnung der Struktur verwendet werden.

Kapitel 4

Diskussion

Makromolekulare Strukturbestimmung ist ein Induktionsproblem und daher mit den Mitteln der Wahrscheinlichkeitsrechnung zu lösen.

Ausgehend von dieser Sichtweise wurden in dieser Arbeit probabilistische Modelle zur Analyse kernspektroskopischer Daten entwickelt und auf das Problem zusätzlicher Parameter angewendet.

Die Wahrscheinlichkeitstheorie ist ein konsistenter, in sich geschlossener Formalismus. In einer wahrscheinlichkeitstheoretischen Formulierung ist ein Strukturbestimmungsproblem vollständig durch die A Posteriori-Wahrscheinlichkeit der Konformationen gelöst; Heuristiken sind überflüssig.

Der Formalismus ist allgemein anwendbar. Als Daten können nicht bloß Kernresonanzmessungen dienen, sondern jede Information, die Aufschluß über die molekulare Struktur gibt.

4.1 Modellierung kernspektroskopischer Meßgrößen

Die A Posteriori-Wahrscheinlichkeit bewertet jede Konformation des Makromoleküls eindeutig im Licht der Daten und allgemeiner Vorkenntnisse. Alle Informationen, die in die A posteriori-Verteilung einfließen, sind als bedingende Sachverhalte aufgeführt. Die Strukturberechnung wird so objekti-

ver: alle Annahmen sind explizit genannt und nicht mehr in algorithmischen Abläufen verborgen. Ein Vergleich mit traditionellen Zugängen legt offen, welche impliziten Annahmen dort gemacht werden, und wo deren Schwächen und Unzulänglichkeiten liegen.

Das Prinzip der Inferentiellen Strukturbestimmung wurde ursprünglich nur auf Distanzen angewendet, die aus einer Analyse von NOESY-Spektren stammen [2]. Es ist aber auch direkt auf andere Datentypen anwendbar. In dieser Arbeit wurden zusätzlich zu dem publizierten Modell für NOESY-Volumina (Abschnitt 3.1.3) wahrscheinlichkeitstheoretische Modelle für die Beschreibung skalarer und dipolarer Kopplungen entwickelt. Als Theorien zur Beschreibung dieser Meßgrößen wurden empirische Zusammenhänge bzw. Näherungen verwendet: die Karplus-Kurve quantifiziert den Zusammenhang zwischen skalaren Kopplungskonstanten und den Dihedralwinkeln, die Saupe-Matrix beschreibt die mittlere Ausrichtung des Moleküls während der Messung dipolarer Kopplungen. Beide Theorien vernachlässigen die Flexibilität der Polypeptidkette, aufgrund derer bloß ein dynamisches Mittel gemessen wird. Die Karplus-Koeffizienten und die Elemente der Saupe-Matrix sind zusätzliche Parameter, die ebenfalls aus den Daten erschlossen werden müssen; in einer wahrscheinlichkeitstheoretischen Formulierung ist dies direkt möglich.

Die Objektivität der Methode zeigt sich auch darin, daß sehr allgemeine Invarianzforderungen die Fehlergesetze und die A priori-Verteilungen festlegen. Ein Gauß'sches Fehlermodell beschreibt die Abweichungen zwischen den gemessenen und den theoretischen Werten für die skalaren und dipolaren Kopplungen. Ein solches Fehlergesetz läßt sich durch das Maximum-Entropie-Prinzip motivieren; im Falle beschränkter Größen, wie den positiven dipolaren Relaxationsraten, ergibt sich als natürliches Pendant, eine Gauß-Verteilung der logarithmischen Werte. Die Forderung der Skaleninvarianz des Fehlers und des Kalibrationsfaktors liefert Jeffreys' prior als deren A priori-Verteilung. Bei Vernachlässigung des Lösungsmittels liefert das Maximum-Entropie-Prinzip das Boltzmann-Ensemble der molekularen Konfigurationen

als konformationelle A priori-Verteilung. Eine Parametrisierung der Struktur in Dihedralwinkeln reduziert die Dimensionalität des Konformationsraums um eine Größenordnung.

4.2 Das Problem zusätzlicher Parameter

Die wahrscheinlichkeitstheoretische Sicht der Strukturbestimmung hat nicht nur formale, sondern auch praktische Vorteile. Ein Strukturbestimmungsproblem ist stets wohl definiert. Alle Unbekannten, auch zusätzlich zu den Koordinaten eingeführte Parameter, werden gemeinsam bestimmt.

Jede Observable, die Aufschluß über die Struktur gibt, kann direkt analysiert werden, ohne daß die Parameter ihrer Theorie bekannt sein müßten. Es ist nicht mehr notwendig, die Parameter vor der Strukturberechnung abzuschätzen. Dadurch ist eine flexiblere Modellierung der Daten möglich. Die Parametrisierungen komplizierterer Modelle können auf dieselbe Weise geschätzt werden wie bei den hier vorgestellten Modellen.

Jegliche experimentelle Information wird gemeinsam analysiert. Dadurch, daß alle Daten mit eigenen Fehlern in die Analyse eingehen, können mehrere Datensätze gemeinsam konsistent genutzt werden. Der Einfluß eines Datensatzes paßt sich verhältnismäßig zu den anderen Datensätzen und zu den Vorkenntnissen an. Dadurch werden die Koordinaten genauer festgelegt.

In den hier entwickelten Modellen ist es möglich, die Theorieparameter und Fehler durch analytische Integration zu entfernen. Dies illustriert, daß Strukturbestimmung in der wahrscheinlichkeitstheoretischen Formulierung stets vollständig bestimmt ist. Über die marginale A posteriori-Verteilung könnte die Struktur berechnet werden, ohne die zusätzlichen Parameter überhaupt zu schätzen.

4.2.1 Gewichtung der Daten

Wird die Struktur durch Minimierung einer Zielfunktion der Form (1.4) bestimmt, stellt sich die Frage nach der Wahl des Gewichtungsfaktors λ. Schon

Jack und Levitt [65] haben festgestellt: „ the best choice of this parameter, so as to produce the most nearly 'correct' structure, is something of a problem." Sie schlagen vor, λ so zu wählen, daß während der Minimierung E und λF von derselben Größenordnung sind. Brünger und Karplus [66] schließen sich diesem Vorschlag an: sie bestimmen λ, indem sie eine kurze Moleküldynamik ohne Datenterm F berechnen und die Werte von E und F vergleichen. Ebenfalls von Brünger und Mitarbeitern [15, 16] wurde die Kreuzvalidierung in der makromolekularen Strukturberechnung als Methode eingeführt, um das Gewicht λ zu bestimmen.

Kreuzvalidierung hat mehrere Nachteile: Es gibt kein übergeordnetes Prinzip, aus dem sie sich ableiten ließe. Die Wahl der Testdaten ist unklar – idealerweise sollte das Ergebnis unabhängig von dieser Wahl sein. Gerade NOESY-Daten sind jedoch nicht gleich wichtig: NOEs zwischen Protonen, die in der Sequenz weit entfernt sind, gestatten erst, kompakte Konformationen zu berechnen; sequentielle NOEs bestimmen eher die Kalibrierung. Werden zuviele langreichweitige NOEs den Testdaten zugeteilt, können keine kompakten Strukturen berechnet werden. Die Einteilung in Arbeits- und Testdaten kann somit stark die Stabilität der Kreuzvalidierung beeinflussen. Außerdem legt die Anzahl der Testdaten die Genauigkeit von λ fest: je mehr Testdaten verwendet werden, umso schärfer ist λ bestimmt. Mit abnehmender Anzahl von Messungen in der Hybridenergie konvergiert die Strukturberechnung jedoch schlechter. Dagegen liefert die Bayes'sche Analyse auch bei dünnen Datensätzen stabile Ergebnisse, wie die Analyse der SH3-Daten in Abschnitt 3.5.1 zeigt. Auch die Wahl der Bewertungsfunktion ist nicht klar; unterschiedliche R- oder Qualitäts-Faktoren wurden vorgeschlagen. Sie haben allesamt den Nachteil, bloß Abweichungen zu den Testdaten zu messen, Vorwissen, welches in der A priori-Verteilung ausgedrückt wird, jedoch zu vernachlässigen. Bei Anwendung der Kreuzvalidierung steht man somit wieder vor dem Problem, Entscheidungen (Unterteilung der Daten, Wahl der Bewertungsfunktion) ad hoc treffen zu müssen.

In der Inferentiellen Strukturbestimmung stellt die Wahl eines oder meh-

rerer Gewichte kein Problem dar. Sie stehen im direkten Zusammenhang mit den Fehlern der Datensätze: $\lambda_j \propto \sigma_j^{-2}$. Im allgemeinsten Fall werden die Gewichte als zusätzliche Hypothesenparameter zur den Atompositionen behandelt und aus den Daten geschätzt. Dazu äquivalent ist die analytische Elimierung der Gewichte durch Marginalisierung (siehe Abschnitte 2.2.3 und 3.2), welche aber nur bei einfachen Fehlermodellen möglich ist. In der Praxis zeigt sich, daß die Monte-Carlo-Schätzung des Gewichts dieselben Ergebnisse wie eine Kreuzvalidierung liefert. Es entfällt jedoch die Einteilung in Arbeits- und Testdaten, welche bei dünnen Datensätzen immer problematischer wird.

4.2.2 Parametrisierung der Karplus-Kurve

Die Entwicklungskoeffizienten der Karplus-Kurve sind vor einer Strukturbestimmung nicht bekannt; im Prinzip könnten sie quantenmechanisch berechnet werden, dies setzt aber die Kenntnis der Struktur voraus. Deshalb sind die Karplus-Koeffizienten Unbekannte, die zusätzlich zu den Koordinaten und Fehlern aus den Daten geschätzt werden müssen. Dies ist in der herkömmlichen, optimierungsbasierten Strukturbestimmung nicht möglich, so daß man auf bekannte Parametrisierungen zurückgreifen muß. Lediglich die „selbstkonsistente“ Methode von Schmidt et al. [56] ist in der Lage, die Karplus-Koeffizienten zusammen mit den Dihedralwinkeln zu bestimmen. Die Einschränkungen dieser Methode sind jedoch: (i) Vorwissen über die erlaubten Dihedralwinkeleinstellungen wird vernachlässigt, (ii) unterschiedliche Fehler für Messungen der jeweiligen Kopplungstypen werden nicht berücksichtigt und können nicht aus den Daten bestimmt werden.

Daß die Dihedralwinkel alle Werte annehmen dürfen, ist eine Vernachlässigung von Vorwissen über die molekularen Konformationen und gleichbedeutend mit $\beta = 0$ im Boltzmann-Ensemble. In der inferentiellen Strukturberechnung muß man nicht auf dieses Vorwissen verzichten, sondern kann es durch die A priori-Verteilung in die Parameterschätzung einfließen lassen. Berücksichtigung von Vorwissen über die Konformationen hilft, Mehrdeutigkeiten der A posteriori-Verteilung der Dihedralwinkel aufzulösen.

Skalare Kopplungen sind je nach Typ unterschiedlich genau meßbar [55]. Unterschiedliche Genauigkeiten sollten in der Parameterschätzung berücksichtigt werden. Im Schmidt'schen Algorithmus werden jedoch die Fehler aller Messungen, auch unterschiedlichen Kopplungstyps, gleich gewählt und auf 0.25 Hz gesetzt. Eine Bayes'sche Analyse paßt dagegen die Fehler an die Daten an (siehe Abschnitt 3.7).

4.2.3 Bestimmung der Saupe-Matrix

Auf dieselbe Problematik trifft man bei der Analyse dipolarer Kopplungen. Die Art des Experiments legt die Saupe-Matrix fest, ohne eine Kenntnis der Struktur läßt sich aber die mittlere Orientierung der Probenmoleküle vorab nicht angeben. Die Inferentielle Strukturbestimmung kann die unbekannten Elemente der Saupe-Matrix genauso behandeln wie die Karplus-Koeffizienten, d.h. gemeinsam mit den Koordinaten aus den Daten schätzen. Je nachdem, welches Vorwissen man annimmt, ergeben sich bekannte Algorithmen zur Bestimmung der Saupe-Matrix als Spezialfälle der wahrscheinlichkeitstheoretischen Behandlung: Ist die Struktur bekannt, können die Elemente der Saupe-Matrix direkt aus der bedingten A posteriori-Verteilung $p(s_1, \ldots, s_5|X)$ bestimmt werden; das Maximum dieser Verteilung liegt bei der Lösung, die das Verfahren von Losonczi et al. [60] liefert. Die Histogrammethode von Clore et al. [61] sowie deren Verfeinerung als Maximum-Likelihood-Methode [64] erhält man ebenfalls direkt aus der Bayes'schen Behandlung, wenn man das Vorwissen über die erlaubten Konformationen vernachlässigt ($\beta = 0$).

Allgemeiner sind die Ansätze von Hus et al. [62] sowie von Moltke und Grzesiek [63]. Sie setzen keine bekannte Struktur voraus. Hus et al. bestimmen die Saupe-Matrix zusammen mit den Orientierungen der Peptid-Ebenen; ihr Algorithmus basiert auf der gemeinsamen Analyse mehrerer Datensätze, denen unterschiedliche mittlere Orientierungen zugrundeliegen. Hus et al. konnten allein aus den dipolaren Kopplungen Strukturen berechnen, die sehr nahe zur NMR-Struktur (1d3z) von Ubiquitin sind. Dazu mußten sie wei-

teres Vorwissen in Betracht ziehen: Die Mehrdeutigkeit der Orientierungen der Peptidebenen wurde durch Chiralitätsbedingungen aufgelöst. Zusätzlich wurden in Sequenzabschnitten, in denen nur wenige Kopplungen gemessen werden (bspw. bei Prolinen), Bedingungen eingeführt, um die Orientierungen der Peptidebenen möglichst getreu fortzusetzen. Diese Informationen müßten in die A priori-Verteilungen des Bayes'schen Analyse eingehen, um ähnlich gute Resultate zu erzielen.

Der Ansatz von Moltke und Grzesiek ist nur scheinbar verschieden von der Methode von Hus et al.: Er entspricht einer analytischen Marginalisierung der Saupe-Matrix; nach den Ausführungen in Abschnitt 2.2.3 ist aber eine Integration über unbekannte Parameter wie die Elemente der Saupe-Matrix gleichwertig mit deren Schätzung durch Monte-Carlo-Simulation oder Optimierung einer geeigneten Zielfunktion. Moltke und Grzesiek haben die optimale Saupe-Matrix analytisch in Abhängigkeit von der gesuchten Struktur dargestellt; in einer Optimierung aller Parameter, der Struktur und der Saupe-Matrix, würden sich dieselben Werte ergeben.

4.2.4 Äquivalenz von analytischer und numerischer Marginalisierung

Alle zusätzlichen Parameter, sowohl die Gewichte der Datensätze sowie unbekannte Parameter der Theorie, können entweder mit den Koordinaten geschätzt oder durch analytische Integration entfernt werden. Daß die beiden Zugänge in Hinblick auf die berechneten Strukturen gleichwertig sind, wurde in Abschnitt 3.5.1 für den Fall des Kalibrationsfaktors der NOESY-Modells nachgewiesen. Genauso ließen sich die marginalen A posteriori-Verteilungen der skalaren Kopplungskonstanten (Gl. (3.15)) sowie der dipolaren Kopplungen (Gl. (3.20)) implementieren und zur ausschließlichen Bestimmung der Struktur nutzen. Eine direkte Schätzung der zusätzlichen Parameter stellt jedoch, verglichen mit der Simulation der bedingten konformationellen A posteriori-Verteilung, einen vernachlässigbaren Mehraufwand dar, so daß sich die Implementierung der Modelle, die durch analytische Marginalisierung

gewonnen werden, aus praktischer Sicht nicht lohnt.

4.2.5 Gemeinsame Verwendung aller Datensätze

In der Inferentiellen Strukturbestimmung werden alle zur Verfügung stehenden Daten gemeinsam analysiert; es ist nicht nötig, die Analyse eines bestimmten Datentyps abzusondern, wie dies beispielsweise bei Verwendung der Histogrammethode zur Bestimmung der Saupe-Matrix aus gemessenen dipolaren Kopplungen der Fall ist. Die unbekannten Fehler der Datensätze (also ihre Gewichtung) sowie unbekannte Theorieparameter werden optimal angepaßt. Die allgemeinste Form der A posteriori-Verteilung ist in Abschnitt 3.2.4 angeben. Wieder kann durch Marginalisierung eine Verteilung abgeleitet werden, die nur noch auf dem Konformationsraum definiert ist. Die Ergebnisse aus Abschnitt 3.5.2 zeigen, daß eine solche gemeinsame Analyse mit Hilfe des Replica-Algorithmus durchgeführt werden kann. Eine gemeinsame Verarbeitung aller zur Verfügung stehenden Information hat den Vorteil, daß durch die Vermeidung von Zwischenschritten keine Information verlorengeht und die berechneten Strukturen genauer bestimmt und näher zur Kristallstruktur sind.

4.3 Strukturberechnung

In der Praxis liefert eine Bayes'sche Analyse Strukturen von vergleichbarer Qualität wie traditionelle Methoden (siehe Abschnitt 3.5). Als Formalismus zur Beschreibung unvollständigen Wissens macht die Wahrscheinlichkeitstheorie zusätzlich Aussagen über die Genauigkeit der geschätzten Größen. Man erhält die „Fehlerbalken“ der Struktur, aber auch sonstiger Unbekannter; und damit nicht nur ein Ergebnis, sondern auch dessen Glaubwürdigkeit.

Aufgrund der Komplexität des Problems sind nur wenige analytische Rechnungen möglich; Monte-Carlo-Methoden berechnen numerische Näherungen. Ein Nachteil ist, daß das verwendete Replica-Schema sehr rechen-

aufwendig ist. Dem kann durch mehrere Entwicklungen entgegengewirkt werden: Erstens sind Monte-Carlo-Methoden denkbar, die nicht mehrere Kopien des Systems gleichzeitig simulieren und so den Rechenaufwand vervielfachen. Zweitens können Implementationsdetails optimiert werden. Drittens wird durch die Beschleunigung der Prozessoren eine Anwendung zukünftig praktikabler.

Die Bayes'sche Formulierung des Strukturbestimmungsproblems stellt auch einen einfachen Formalismus bereit, um Zielfunktionen für Optimierungsverfahren abzuleiten. Statt einer Monte-Carlo-Simulation der A posteriori-Verteilung kann deren negativer Logarithmus minimiert werden, um die wahrscheinlichsten Parametereinstellungen zu bestimmen. Hier kann entweder der volle Satz an Hypothesenparametern (d.h. Strukturparameter und zusätzliche Parameter) oder ein reduzierter Parametersatz optimiert werden. In einfachen Modellen lassen sich die zusätzlichen Parameter analytisch entfernen und aus der resultierenden marginalen A posteriori-Verteilung eine Zielfunktion ableiten, die auf einem reduzierten Parameterraum definiert ist.

Man mag argumentieren, eine Bayes'sche Strukturbestimmung durch Monte-Carlo-Simulation erfordere soviel Mehraufwand, daß sich die Anwendung nicht lohne. Dem läßt sich außer den erwähnten Entwicklungen entgegenhalten, daß der Mehraufwand auch ein reichhaltigeres Ergebnis liefert, und daß Optimierungsverfahren bloß den Eindruck erwecken, mit weniger aufwendigen Mitteln dieselben Aussagen machen zu können. Eine Simulation der A posteriori-Verteilung mittels Monte-Carlo-Verfahren wird dann unumgänglich, wenn man neben den optimalen Parameterwerten auch deren Genauigkeit angeben will. Hierfür sind Optimierungsverfahren ungeeignet. Die Varianzen, die aus einer Variation der Startbedingungen resultieren, entbehren einer statistischen Grundlage. Ein sauberes, von Implementationsdetails unabhängiges Strukturensemble läßt sich nur mit Monte-Carlo-Methoden berechnen; ohne eine wahrscheinlichkeitstheoretische Formulierung der Strukturbestimmung wäre darüberhinaus das Strukturensemble strenggenommen garnicht eindeutig definiert, weil es von der Wahl der

zusätzlichen Parameter abhinge. Versucht man mittels Optimierung eine vergleichbar vollständige Analyse durchzuführen, wie dies eine Simulation der A posteriori-Verteilung darstellt, würde sich auch dort der Aufwand schnell potenzieren; dies wird beispielsweise bei der Bestimmung mehrerer Datengewichte durch Kreuzvalidierung deutlich.

4.4 Ausblick

Neben algorithmischen Verbesserungen sind Erweiterungen der Datenmodelle denkbar. Außer den behandelten Observablen könnten weitere experimentell zugängliche Größen verarbeitet werden: Beispiele wären Messungen anisotroper chemischer Verschiebungen oder die Beobachtung von Quadrupolwechselwirkungen, die analog zu dipolaren Kopplungen beschrieben werden.

Auch die Beschreibung der Observablen ließe sich verfeinern. Ein Relaxationsmatrixansatz [4] könnte beispielsweise die Spindiffusion modellieren. Darüberhinaus müßten die Messungen als Zeit- und Ensemblemittel beschrieben werden und nicht als instantane Größen.

Weitaus wichtiger wäre jedoch, die Verwendung von Spektren zu ermöglichen, deren Resonanzen noch nicht zugeordnet wurden. Ebenso wichtig wäre es, Vorwissen über biomolekulare Strukturen besser durch die konformationelle A priori-Verteilung zu repräsentieren.

Eine möglichst vollständige Beschreibung der Meßgrößen wurde in dieser Arbeit nicht angestrebt und viele Teilprobleme blieben unbehandelt. Insofern sind die vorgestellten Ergebnisse vorläufig und können leicht erweitert und verbessert werden.

Unverändert bleiben wird jedoch die prinzipielle Vorgehensweise: Daten aus Experimenten zur makromolekularen Strukturbestimmung mit den Mitteln der Wahrscheinlichkeitstheorie zu analysieren.

Anhang A

Wahrscheinlichkeitsverteilungen

A.1 Gauß-Verteilung

Die Gauß- oder Normalverteilung ist definiert als:

$$\mathrm{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{A.1}$$

mit Mittelwert μ und Standardabweichung σ. Sie ist die Verteilung mit maximaler Informationsentropie, die die Bedingungen $\langle x\rangle = \mu$ und $\langle (x-\langle x\rangle)^2\rangle = \sigma^2$ erfüllt.

Die ungeraden Momente der Gaußverteilung verschwinden, die geraden Momente sind Potenzen der Varianz σ^2:

$$\left\langle (x-\mu)^{2n}\right\rangle = \frac{(2n)!}{2\,n!}\,\sigma^{2n}. \tag{A.2}$$

Für Messungen $D = \{x_1,\ldots,x_n\}$ läßt sich die Likelihood-Funktion schreiben als

$$p(D|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}}\,\exp\left\{-\frac{n}{2\sigma^2}\left[(\bar{x}-\mu)^2 + s^2\right]\right\} \tag{A.3}$$

mit den hinreichenden Statistiken $\bar{x} = \frac{1}{n}\sum_i x_i$ (arithmetisches Mittel) und $s^2 = \frac{1}{n}\sum_i (x_i-\bar{x})^2$ (Varianz).

Um μ und σ aus den Messungen D schätzen zu können, muß ihnen noch eine

A priori-Wahrscheinlichkeit zugewiesen werden. Eine Verteilung, die minimale Vorkenntnisse annimmt, folgt aus den Invarianzgesetzen [25, 20] (Skaleninvarianz für σ und Translationsinvarianz für μ):

$$p(\mu, \sigma | I) = \sigma^{-1}, \tag{A.4}$$

wobei $\mu \in [-\infty, \infty]$ und $\sigma \in [0, \infty]$.

Damit ist die A posteriori-Verteilung

$$p(\mu, \sigma | D, I) \propto \sigma^{-(n+1)} \exp\left\{-\frac{n}{2\sigma^2} \left[(\mu - \overline{x})^2 + s^2 \right]\right\}. \tag{A.5}$$

Die marginalen A posteriori-Verteilungen sind eine t-Verteilung im Falle des Mittelwerts

$$p(\mu | D, I) = \frac{1}{\sqrt{\pi(n-1)s^2}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \left[\frac{(\mu - \overline{x})^2}{(n-1)\, s^2} + 1 \right]^{-n/2} \tag{A.6}$$

und eine Gamma-Verteilung für σ^{-2} (siehe A.3)

$$p(\sigma | D, I) = \frac{(s^2/2)^{n/2}}{\Gamma(n/2)} \sigma^{-(n+1)} \exp\left\{-s^2/(2\sigma^2)\right\} \tag{A.7}$$

A.2 Lognormal-Verteilung

Durch die Transformation $x \to \log x$ läßt sich die Gauß-Verteilung auf die positive Achse beschränken. Die Logarithmen sind gaußverteilt:

$$\mathrm{LN}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}\, x} \exp\left\{-\frac{1}{2\sigma^2} \log^2(x/\mu)\right\}, \tag{A.8}$$

wobei μ durch $\log \mu$ ersetzt wurde. Die Variablentransformation ergibt als Grundmaß für x: $\mathrm{d}\log x = x^{-1}\, \mathrm{d}x$.

Die Lognormal-Verteilung kann durch ein Maximum-Entropie-Argument motiviert werden: Kennt man für eine positive Observable mit Grundmaß $\mathrm{d}\mu(x) = \mathrm{d}\log x$ sowie die ersten beiden Momente des Logarithmus, also

$$\log \mu = \langle \log x \rangle, \quad \sigma^2 = \left\langle (\log x - \langle \log x \rangle)^2 \right\rangle,$$

so ergibt sich die Lognormal-Verteilung (A.8) als die Verteilung maximaler Entropie, die die beiden Bedingungen erfüllt. Außer den beiden Forderungen gehen in sie keine weiteren Annahmen ein.

Aus n Messungen $D = \{x_1, \ldots, x_n\}$, $x_i > 0$, ergibt sich die A posteriori-Verteilung

$$p(\mu, \sigma | D, I) \propto \frac{1}{\mu\, \sigma^{(n+1)}} \exp\left\{ -\frac{n}{2\sigma^2} \left[\log^2(\mu/\overline{x}) + s^2 \right] \right\}, \qquad \text{(A.9)}$$

wobei hier $\overline{x} = (\prod_i x_i)^{1/n}$ das geometrische Mittel der Messungen ist und $s^2 = \frac{1}{n} \sum_i \log^2(x_i/\overline{x})$ die Varianz der logarithmischen Werte.

A.3 Gamma-Verteilung

Die Gamma-Verteilung ist für positive Variablen $x > 0$ definiert als

$$\mathrm{G}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{\,\alpha-1}\, \exp\{-\beta x\}, \qquad \text{(A.10)}$$

wobei $\Gamma(\alpha)$ die Γ-Funktion ist. Die Verteilung ist durch zwei positive Parameter $\alpha > 0$ und $\beta > 0$ parametrisiert.

Die Momente der Gamma-Verteilung sind

$$\langle x^n \rangle = \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)}\, \beta^{-n} = \frac{(\alpha + n - 1)\,(\alpha + n - 2) \cdots (\alpha + 1)\,\alpha}{\beta^n}, \qquad \text{(A.11)}$$

und man erhält

$$\langle x \rangle = \frac{\alpha}{\beta}, \quad \left\langle\, (x - \langle x \rangle)^2 \,\right\rangle = \frac{\alpha}{\beta^2}. \qquad \text{(A.12)}$$

Anhang B

Interne Koordinaten

B.1 Externe und interne Koordinaten

Ein Kettenmolekül besteht aus N Atomen, die durch kovalente Bindungen verbunden sind und so eine verzweigte Kette bilden. Eine Möglichkeit, die Konfigurationen des Moleküls darzustellen, ist über die kartesischen Koordinaten $\mathbf{x}_i, i = 1, \ldots, N$ der Atome. Diese Parametrisierung ist jedoch nicht an die kovalente Geometrie des Moleküls angepaßt. Besser ist es, Konformationen durch externe und interne Koordinaten zu parametrisieren [67].

Kovalente Kräfte schränken die Abstände kovalent gebundener Atome ein, so daß die Länge des Bindungsvektors

$$\mathbf{l}_i = \mathbf{x}_i - \mathbf{x}_{i-1} \tag{B.1}$$

nur geringen Schwankungen unterliegt. Wir setzen $\mathbf{x}_0 = \mathbf{0}$, d.h. $\mathbf{l}_1 = \mathbf{x}_1$. Die neuen Koordinaten $\{\mathbf{l}_i\}$ haben $3N$ Freiheitsgrade und es gilt

$$\mathbf{x}_i = \sum_{j=1}^{i} \mathbf{l}_j.$$

Die Bindungsvektoren lassen sich in Kugelkoordinaten darstellen. Die Länge des Bindungsvektors ist

$$l_i = \|\mathbf{l}_i\|. \tag{B.2}$$

Der Polarwinkel ist der Bindungswinkel zwischen aufeinanderfolgenden Bindungen:

$$\cos\kappa_i = (\mathbf{l}_i^T \mathbf{l}_{i-1}) \,/\, (l_i l_{i-1}). \tag{B.3}$$

Wir setzen $\mathbf{l}_0 = \mathbf{e}_z$, d.h. κ_1 ist der Polarwinkel von $\mathbf{x}_1$ in einem äußeren Koordinatensystem $\left(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\right)$. Der Azimuthalwinkel

$$\cos\theta_i = \mathbf{n}_i^T \mathbf{n}_{i-1}, \quad \mathbf{n}_j = (\mathbf{l}_j \times \mathbf{l}_{j-1})/\|\mathbf{l}_j \times \mathbf{l}_{j-1}\| \tag{B.4}$$

ist der Winkel zwischen den beiden Ebenen mit Normalenvektoren $\mathbf{n}_i$ und $\mathbf{n}_{i-1}$. Wir setzen $\mathbf{l}_{-1} = -\mathbf{e}_x$, dann ist θ_1 der Azimuthalwinkel von $\mathbf{x}_1$ im äußeren Koordinatensystem. θ_i beschreibt Drehungen um den Bindungsvektor $\mathbf{l}_{i-1}$, d.h. die Winkel θ_i sind die Dihedralwinkel des Kettenmoleküls.

Da die Wechselwirkungen innerhalb des Moleküls nicht von der äußeren Orientierung abhängen, können die sechs äußeren Freiheitsgrade von den inneren abgesondert werden: der Translationsvektor $\mathbf{t}$ und die Rotationsmatrix $\mathbf{R}$ beschreiben die Position und die Orientierung des Moleküls. Ist der Translationsvektor die Position des ersten Atoms,

$$\mathbf{t} = \mathbf{l}_1 = \mathbf{x}_1,$$

so werden die Freiheitsgrade l_1, κ_1 und θ_1 in $\mathbf{t}$ absorbiert. Wir wählen die Orientierung der verbleibenden internen Koordinaten so, daß $\mathbf{l}_2 = l_2 \mathbf{R}\, \mathbf{e}_1$ und $\mathbf{l}_3$ im lokalen Koordinatensystem in der x-y-Ebene liegt ($\mathbf{l}^T \mathbf{R}\, \mathbf{e}_3 = 0$). Die Freiheitsgrade κ_2, θ_2 und θ_3 gehen in $\mathbf{R}$ auf.

Die verbleibenden inneren Freiheitsgrade sind $l = \{l_i, i = 2, \ldots, N\}$, $\kappa = \{\kappa_i, i = 3, \ldots, N\}$ und $\theta = \{\theta_i, i = 4, \ldots, N\}$. Damit gibt es $3N - 6$ innere Freihheitsgrade. Die äußeren Parameter $\mathbf{t}$ und $\mathbf{R}$ stellen sechs weitere Freiheitsgrade.

Die Fluktuationen der inneren Freiheitsgrade l_i, κ_i und θ_i sind sehr unterschiedlich in ihrer Stärke, weil verschiedene physikalische Kräfte sie einschränken. Die Bindungslängen und -winkel werden in erster Linie durch kovalente Kräfte zwischen benachbarten Atomen festgelegt. Die Dihedralwinkel werden dagegen neben direkten Einschränkungen durch die chemische Bindung durch nicht-kovalente Kräfte, d.h. van der Waals- und elektrostatische

Kräfte, eingeschränkt. Abhängig von der Natur der chemischen Bindung kann der Dihedralwinkel nur bestimmte Einstellungen einnehmen (so zum Beispiel der χ_1-Winkel der Methylgruppe von Alanin).

In der Strukturberechnung halten wir die kovalenten Parameter l_i und κ_i fest, weil sie bei gewöhnlichen Versuchsbedingungen nur wenig schwanken. Die verbleibenden Strukturparameter sind die Drehwinkel um die kovalenten Bindungen. Dadurch verringern wir die Anzahl der inneren Freiheitsgrade um eine Größenordnung. In Polypeptiden sind zusätzlich die Planaritäten von Ringen in den Seitenketten fixiert; die Phasenwinkel sind fest und die Peptidebene wird als starr angesehen ($\omega = 180^o$).

B.2 Transformation auf kartesische Koordinaten

Für jedes der N Atome kann ein lokales Koordinatensystem definiert werden. Die Achsen des Koordinatensystems des i-ten Atoms im Koordinatensystem des j-ten Atoms seien $\mathbf{e}_x^{(i,j)}$, $\mathbf{e}_y^{(i,j)}$, $\mathbf{e}_z^{(i,j)}$.

Es gilt $\mathbf{e}_x^{(i,i)} = \mathbf{l}_i / l_i$, d.h. die x-Achse im eigenen lokalen Koordinatensystem ist durch den ersten Bindungsvektor definiert. $\mathbf{e}_y^{(i,i)}$ liegt in der Ebene, die durch die Bindungsvektoren $\mathbf{l}_i$ und $\mathbf{l}_{i-1}$ aufgespannt wird; die Projektion auf $\mathbf{e}_x^{(i-1,i)}$ soll dabei positiv sein. $\mathbf{e}_z^{(i,i)}$ ist der Normalenvektor der Ebene, die $\mathbf{e}_x^{(i,i)}$ und $\mathbf{e}_y^{(i,i)}$ bzw. $\mathbf{l}_i$ und $\mathbf{l}_{i-1}$ aufspannen. Das heißt:

$$\mathbf{e}_z^{(i,i)} = \frac{\mathbf{l}_i \times \mathbf{l}_{i-1}}{\|\mathbf{l}_i \times \mathbf{l}_{i-1}\|}.$$

Die fehlende y-Achse berechnet sich zu

$$\mathbf{e}_y^{(i,i)} = \mathbf{e}_z^{(i,i)} \times \mathbf{e}_x^{(i,i)}.$$

Wenn wir nun die Achsen des Koordinatensystems des $(i+1)$-ten Atoms in dem des i-ten darstellen wollen, gilt

$$\begin{aligned} \mathbf{e}_x^{(i+1,i)} &= \mathbf{R}_x(\theta_i)\,(\cos\kappa_i, \sin\kappa_i, 0)^T \\ &= \mathbf{R}_x(\theta_i + \pi)\,\mathbf{R}_z(-\kappa_i)\,\mathbf{e}_x \end{aligned}$$

und

$$\mathbf{e}_z^{(i+1,i)} = -\mathbf{R}_x(\theta_i)\,\mathbf{e}_z = \mathbf{R}_x(\theta_i+\pi)\,\mathbf{e}_z = \mathbf{R}_x(\theta_i+\pi)\,\mathbf{R}_z(-\kappa_i)\,\mathbf{e}_z.$$

Es folgt:

$$\mathbf{e}_y^{(i+1,i)} = \mathbf{e}_z^{(i+1,i)} \times \mathbf{e}_x^{(i+1,i)} = \mathbf{R}_x(\theta_i+\pi)\,\mathbf{R}_z(-\kappa_i)\,\mathbf{e}_y.$$

Nun kann jeder Vektor $\mathbf{v}$, der in dem $(i+1)$-ten Koordinatensystem durch $\mathbf{v}^{(i+1)}$ dargestellt wird, in einem vorherigen Koordinatensystem dargestellt werden [67]:

$$\mathbf{v}^{(i)} = \mathbf{T}^{(i)}\,\mathbf{v}^{(i+1)}$$

mit der Transformationsmatrix $\mathbf{T}^{(i)} \equiv \mathbf{T}(\theta_i, \kappa_i)$:

$$\mathbf{T}(\theta,\kappa) = \mathbf{R}_x(\theta+\pi)\,\mathbf{R}_z(-\kappa) = \begin{pmatrix} \cos\kappa & \sin\kappa & 0 \\ \cos\theta\sin\kappa & -\cos\theta\cos\kappa & \sin\theta \\ \sin\theta\sin\kappa & -\sin\theta\cos\kappa & -\cos\theta \end{pmatrix}. \tag{B.5}$$

Die Dihedralwinkel θ liegen in $[0, 2\pi]$, die Bindungswinkel κ in $[0, \pi]$.

Eine Darstellung im Koordinatensystem des Moleküls, d.h. im Koordinatensystem des ersten Atoms, erhält man durch wiederholte Anwendung der Transformationsregel

$$\mathbf{v}^{(1)} = \mathbf{R}^{(i)}\,\mathbf{v}^{(i+1)}$$

mit

$$\mathbf{R}^{(i)} = \prod_{k=2}^{i} \mathbf{T}^{(k)} = \prod_{k=2}^{i} \mathbf{T}(\theta_k, \kappa_k).$$

Die kartesischen Koordinaten lassen sich nun aus internen und externen Koordinaten berechnen:

$$\mathbf{x}_i = \sum_{k=1}^{i} \mathbf{l}_k = \sum_{k=2}^{i} l_k\,\mathbf{R}^{(k)}\,\mathbf{e}_x \;+\; \mathbf{t}.$$

Wegen $\mathbf{R}^{(i)} = \mathbf{R}^{(i)}(\{(\theta_k, \kappa_k) \mid k \le i\})$ hängt $\mathbf{x}_i$ nur von den ersten i Dihedral- und Bindungswinkeln ab; $\mathbf{t}$ und $\kappa_2, \theta_2, \theta_3$ sind hier die äußeren Freiheitsgrade.

B.3 Jacobi-Determinante der Transformation

Wenn man Wahrscheinlichkeiten, die in kartesischen Koordinaten parametrisiert sind, auf externe und interne Koordinaten übertragen will, muß die Jacobi-Determinante der Abbildung bekannt sein,

Zuerst wird der Übergang von kartesischen Koordinaten zur Darstellung durch Bindungsvektoren betrachtet. Es gilt

$$\mathbf{l}_1 = \mathbf{x}_1, \quad \mathbf{l}_i = \mathbf{x}_i - \mathbf{x}_{i-1},\ i \geq 2.$$

Dieser Koordinatenwechsel ist linear:

$$\begin{pmatrix} \mathbf{l}_1 \\ \mathbf{l}_2 \\ \mathbf{l}_3 \\ \vdots \\ \mathbf{l}_N \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots \\ -\mathbf{I} & \mathbf{I} & \mathbf{0} & \\ \mathbf{0} & -\mathbf{I} & \mathbf{I} & \ddots \\ \vdots & \vdots & \vdots & \ddots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \vdots \\ \mathbf{x}_N \end{pmatrix}$$

mit Jacobi-Determinante

$$\left| \frac{\partial(\mathbf{l}_1, \ldots, \mathbf{l}_N)}{\partial(\mathbf{x}_1, \ldots, \mathbf{x}_N)} \right| = 1. \tag{B.6}$$

In externen und internen Koordinaten $\mathbf{t}, l, \kappa, \theta$ werden die Bindungsvektoren dargestellt durch

$$\mathbf{l}_1 = \mathbf{t}, \quad \mathbf{l}_i = \mathbf{R}\ l_i\, \mathbf{R}^{(i)} \mathbf{e}_x,\ i \geq 2$$

mit

$$\mathbf{R}^{(i)} = \prod_{k=1}^{i-1} \mathbf{T}(\theta_k, \kappa_k),\ i \geq 2,$$

wobei $\kappa_2, \theta_2, \theta_3$ die äußeren Rotationsfreiheitsgrade des Moleküls darstellen.

Um die Jacobi-Determinante für den Koordinatenwechsel $\{\mathbf{l}_1, \ldots, \mathbf{l}_N\} \to \{\mathbf{t}, l_2, \ldots, l_N, \kappa_2, \ldots, \kappa_N, \theta_2, \ldots, \theta_N\}$ zu berechnen, brauchen wir die Ableitungen der Bindungsvektoren nach den neuen Koordinaten. Es gilt:

$$\frac{\partial\, \mathbf{l}_i}{\partial\, \mathbf{t}} = \delta_{i1}\, \mathbf{I}, \quad \frac{\partial\, (l_{ix}, l_{iy}, l_{iz})}{\partial\, (l_j, \kappa_j, \theta_j)} = \begin{cases} \mathbf{J}_{ij}, & i \geq j \\ \mathbf{0}, & i < j \end{cases}.$$

Damit ist die Jacobi-Matrix der Koordinatentransformation:

$$\frac{\partial\left(\mathbf{l}_1,\ldots,\mathbf{l}_N\right)}{\partial\left(\mathbf{t},l,\kappa,\theta\right)}=\begin{pmatrix}\mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{J}_{22} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{J}_{32} & \mathbf{J}_{33} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{J}_{42} & \mathbf{J}_{43} & \mathbf{J}_{44} & \cdots & \mathbf{0}\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\end{pmatrix}.$$

Die Jacobi-Determinante ist:

$$\left|\frac{\partial\left(\mathbf{l}_1,\ldots,\mathbf{l}_N\right)}{\partial\left(\mathbf{t},l,\kappa,\theta\right)}\right|=\prod_i\ \left|\mathbf{J}_{ii}\right|.$$

Es gilt:

$$\mathbf{J}_{ii}=\left(\frac{\partial\,\mathbf{l}_i}{\partial\,l_i},\frac{\partial\,\mathbf{l}_i}{\partial\,\kappa_i},\frac{\partial\,\mathbf{l}_i}{\partial\,\theta_i}\right)=\mathbf{R}^{(i-1)}\ \left(\mathbf{T}^{(i)}\mathbf{e}_x,\ l_i\,\frac{\partial\,\mathbf{T}^{(i)}}{\partial\,\kappa_i}\,\mathbf{e}_x,\ l_i\,\frac{\partial\,\mathbf{T}^{(i)}}{\partial\,\theta_i}\,\mathbf{e}_x\right)$$

und damit

$$\left|\mathbf{J}_{ii}\right|=l_i^2\sin\kappa_i.$$

So daß insgesamt gilt:

$$\left|\frac{\partial\left(\mathbf{l}_1,\ldots,\mathbf{l}_N\right)}{\partial\left(\mathbf{t},l,\kappa,\theta_)\right.}\right|=\prod_i\ l_i^2\sin\kappa_i.$$

Indem man die Abbildungen $\{\mathbf{x}_i\}\rightarrow\{\mathbf{l}_i\}$ und $\{\mathbf{l}_i\}\rightarrow\{\mathbf{t},l,\kappa,\theta\}$ hintereinander ausführt, erhält man die Jacobi-Determinante für die gesamte Transformation:

$$J(l,\kappa,\theta)\equiv\left|\frac{\partial\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)}{\partial\left(\mathbf{t},l,\kappa,\theta_)\right.}\right|=\prod_i\ l_i^2\sin\kappa_i. \tag{B.7}$$

Anhang C

Das ISD-Softwarepacket

Die Grundidee der inferentiellen Strukturbestimmung wurde zusammen mit Wolfgang Rieping entwickelt und in einem neuen Software-Packet implementiert. Die Programmiersprache ist Python und C: zeitintensive Routinen wurden in C verfaßt und in den Python-Code eingebunden (wrapping). Die Bibliothek umfaßt ungefähr 40000 Zeilen Python- und C-Code. Die Software wurde auf dem Betriebssystem Linux entwickelt und ausschließlich hierfür getestet; sie sollte sich aber auch leicht auf anderen Betriebssystemen installieren lassen.

Die Konformationen eines Makromoleküls werden durch Dihedralwinkel dargestellt (siehe Anhang B). Bisher ist nur die Analyse von Proteinen, die sich aus den zwanzig Standardaminosäuren zusammensetzen, möglich. Die kovalenten Parameter wurden dem ECEPP/2 Kraftfeld [68, 69] entnommen und entsprechen damit der Bausteinbibliothek, die in DYANA [70] verwendet wird.

Als A priori-Verteilung dient das Boltzmann-Ensemble des Moleküls im Vakuum. Van der Waals-Kräfte werden durch das approximative Potential (2.15) mit PROLSQ-Parametern [24] beschrieben. Dieses Kraftfeld wird auch in CNS benutzt [53].

Als Daten können Messungen der Abstände, der Dihedralwinkel und der Orientierungen von Abstandsvektoren dienen. Spezialfälle dieser allgemeinen

Meßgrößen sind NOE-Volumina, skalare Kopplungskonstanten und dipolare Kopplungen. Andere Datentypen können auch verwendet werden.

Die Simulation der A posteriori-Verteilung erfolgt mit den in Kapitel 2.3 beschriebenen Monte-Carlo-Verfahren. Die Dihedralwinkel werden mit Hybrid-Monte-Carlo simuliert. Das Schema zur Integration der Pseudo-Hamilton-Gleichungen ist der leapfrog-Algorithmus.

Die zusätzlichen Parameter, also die Theorieparameter (Karplus-Koeffizienten, Elemente der Saupe-Matrix, Kalibrationsfaktor) und die Fehler der Datensätze, werden mit Zufallszahlengeneratoren gezogen, die aus der R-Bibliothek stammen.

Beim Replica-Monte-Carlo dient die parallel virtual machine (PVM) dazu, die Simulation der verschiedenen Replicas auf einen Cluster von Linux-Maschinen zu verteilen.

Literaturverzeichnis

[1] Laplace, S. *Théorie analytique des probabilités.* Gauthier-Villars, Paris, (1820).

[2] Habeck, M., Rieping, W., and Nilges, M. A new principle for macromolecular structure determination. In 23rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Erickson, G. and Zhai, Y., editors, 157–166. American Institute of Physics, (2004).

[3] Ernst, R. R., Bodenhausen, G., and Wokaun, A. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions.* Oxford University Press, New York, (1990).

[4] Macura, S. and Ernst, R. R. Elucidation of cross relaxation in liquids by two-dimensional NMR spectroscopy. *Molecular Physics* **41**, 95–117 (1980).

[5] Karplus, M. Vicinal proton coupling in nuclear magnetic resonance. *J. Am. Chem. Soc.* **85**, 2870–2871 (1963).

[6] Solomon, I. Relaxation Processes in a System of Two Spins. *Phys. Rev.* **99**, 559–565 (1955).

[7] Neuhaus, D. and Williamson, M. P. *The nuclear Overhauser effect in structural and conformational analysis.* VCH Publishers Inc., New York, (1989).

[8] Brünger, A. T. and Nilges, M. Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR spectroscopy. *Q. Reviews of Biophys.* **26**, 49–125 (1993).

[9] Wüthrich, K. *NMR of Proteins and Nucleic Acids.* John Wiley, New York, (1986).

[10] Havel, T., Kuntz, I. D., and Crippen, G. M. The theory and practice of distance geometry. *Bull. Math. Biol.* **45**, 665–720 (1983).

[11] Kaptein, R., Zuiderweg, E. R., Scheek, R. M., Boelens, R., and van Gunsteren, W. F. A protein structure from nuclear magnetic resonance data: lac repressor headpiece. *J. Mol. Biol.* **182**, 179–182 (1985).

[12] Jaynes, E. Prior Information and Ambiguity in Inverse Problems. In *SIAM -AMS Proceedings*, volume 14, 151–166. American Mathematical Society, (1984).

[13] Güntert, P., Braun, W., and Wüthrich, K. Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *J. Mol. Biol.* **217**, 517–530 (1991).

[14] Nilges, M. and O'Donoghue, S. I. Ambiguous NOEs and automated NOESY assignment. *Prog. NMR Spec.* **32**, 107–139 (1998).

[15] Brünger, A. T. The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–474 (1992).

[16] Brünger, A. T., Clore, G. M., Gronenborn, A. M., Saffrich, R., and Nilges, M. Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. *Science* **261**, 328–331 (1993).

[17] Güntert, P. Structure calculation of biological macromolecules from NMR data. *Q. Reviews of Biophys.* **31**, 145–237 (1998).

[18] Wittgenstein, L. *Tractatus logico-philosophicus.* Suhrkamp Verlag, Frankfurt a. M., (1919).

[19] Cox, R. T. *The Algebra of Probable Inference.* John Hopkins University Press, Baltimore, (1961).

[20] Jaynes, E. T. *Probability Theory: The Logic of Science.* Cambridge University Press, Cambridge, (2003).

[21] Bayes, T. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.* **53**, 370–418 (1763).

[22] Rieping, W., Habeck, M., and Nilges, M. Inferential Structure Determination. submitted, (2004).

[23] Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev. Lett.* **106**, 620–630 (1957).

[24] Hendrickson, W. A. Stereochemically restrained refinement of macromolecular structures. *Methods in Enzymology* **115**, 252–270 (1985).

[25] Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A* **186**, 453–461 (1946).

[26] Neal, R. M. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, (1993).

[27] Geman, S. and Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. PAMI* **6**, 721–741 (1984).

[28] Duane, S., Kennedy, A. D., Pendleton, B., and Roweth, D. Hybrid Monte Carlo. *Phys. Lett. B* **195**, 216–222 (1987).

[29] Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, A., and Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1957).

[30] Allen, M. P. and Tildesley, D. J. *Computer Simulation of Liquids.* Clarendon Press, Oxford, (1987).

[31] Jain, A., Vaidehi, N., and Rodrigues, G. A fast recursive algorithm for molecular dynamics simulation. *J. Comp. Phys.* **106**, 258–268 (1993).

[32] Swendsen, R. H. and Wang, J.-S. Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.* **57**, 2607–2609 (1986).

[33] Tsallis, C. Possible Generalization of Boltzmann-Gibbs Statistics. *J. Stat. Phys.* **52**, 479–487 (1988).

[34] Hansmann, U. H. E. and Okamoto, Y. New Generalized-Ensemble Monte Carlo Method for Systems with Rough Energy Landscape. *Phys. Rev. E* **56**, 2228–2233 (1997).

[35] Mal, T. K., Matthews, S. J., Kovacs, H., Campbell, I. D., and Boyd, J. Some NMR experiments and a structure determination employing a $\{^{15}\mathrm{N}, ^{2}\mathrm{H}\}$ enriched protein. *J. Biomol. NMR* **12**, 259–276 (1998).

[36] Noble, M., Musacchio, A., Saraste, M., and Wierenga, R. Crystal structure of the SH3 domain in human Fyn; comparison of the three-dimensional structures of SH3 domains in tyrosine kinases and spectrin. *EMBO J.* **12**, 2617–2624 (1993).

[37] Cornilescu, G., Marquardt, J. L., Ottiger, M., and Bax, A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.* **120**, 6836–6837 (1998).

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

[39] Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291 (1993).

[40] Vriend, G. and Sander, C. Quality control of protein models: Directional atomic contact analysis. *J. Appl. Cryst.* **26**, 47–60 (1993).

[41] Helgaker, T., Jaszuński, M., and Ruud, K. Ab Initio Methods for the Calculation of NMR Shielding and Indirect Spin-Spin Coupling Constants. *Chem. Rev.* **99**, 293–352 (1999).

[42] Case, D. A. Interpretation of chemical shifts and coupling constants in macromolecules. *Curr. Opin. Struct. Biol.* **10**, 197–203 (2000).

[43] Saupe, A. and Englert, G. High-Resolution Nuclear Magnetic Resonance Spectra of Orientated Molecules. *Phys. Rev. Lett.* **11**, 462–464 (1963).

[44] Bax, A., Kontaxis, G., and Tjandra, N. Dipolar Couplings in Macromolecular Structure Determination. *Methods in Enzymology* **339**, 127–174 (2001).

[45] Tjandra, N. and Bax, A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **278**, 1111–1114 (1997).

[46] Zweckstetter, M. and Bax, A. J. Prediction of Sterically Induced Alignment in a Dilute Liquid Crystalline Phase: Aid to Protein Structure Determination by NMR. *J. Am. Chem. Soc.* **122**, 3791–3792 (2000).

[47] Nilges, M. Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J. Mol. Biol.* **245**, 645–660 (1995).

[48] Spronk, A. E. M., Nabuurs, S. B., Bovin, M. J. J., Krieger, E., Vuister, W., and Vriend, G. The precision of NMR structure ensembles revisited. *J. Biomol. NMR* **25**, 225–234 (2003).

[49] Rosenblatt, M. Remarks on a Multivariate Transformation. *Ann. Math. Stat.* **23**, 470–472 (1952).

[50] Yip, P. F. and Case, D. A. Incorporation of internal motions in NMR refinements based on NOESY data. In Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy, Hoch, J. C., Poulsen, F. M., and Redfield, C., editors, volume 225 of *NATO ASI Series*, 317–330 (Plenum, New York, 1991).

[51] Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. Roy. Stat. Soc. B* **36**, 111–147 (1974).

[52] Nilges, M., Macias, M. J., O'Donoghue, S. I., and Oschkinat, H. Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β-spectrin. *J. Mol. Biol.* **269**, 408–422 (1997).

[53] Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. Crystallography and NMR system (CNS): A new software suite for macromolecular structure determination. *Acta Cryst. sect. D* **54**, 905–921 (1998).

[54] Pardi, A., Billeter, M., and Wüthrich, K. Calibration of the angular dependence of the amide proton-C_α proton coupling constants, $^3J_{HN\alpha}$, in a globular protein. *J. Mol. Biol.* **180**, 741–751 (1984).

[55] Wang, A. and Bax, A. Determination of the Backbone Dihedral Angles ϕ in Human Ubiquitin from Reparametrized Empirical Karplus Equations. *J. Am. Chem. Soc.* **118**, 2483–2494 (1996).

[56] Schmidt, J. M., Blümel, M., Löhr, F., and Rüterjans, H. Self-consistent ^3J coupling analysis for the joint calibration of Karplus coefficients and evaluation of torsion angles. *J. Biomol. NMR* **14**, 1–12 (1999).

[57] Case, D. A., Dyson, H. J., and Wright, P. E. Use of chemical shifts and coupling constants in NMR structural studies on peptides and proteins. *Methods in Enzymology* **239**, 392–416 (1994).

[58] Vuister, G. W. and Bax, A. Quantitative J Correlation: A New Approach for Measuring Homonuclear Three-Bond $J(\mathrm{H^N}\text{-}\mathrm{H}^\alpha)$ Coupling Constants in ^{15}N-Enriched Proteins. *J. Am. Chem. Soc.* **115**, 7772–7777 (1993).

[59] Brüschweiler, R. and Case, D. A. Adding Harmonic Motion to the Karplus Equation for Spin-Spin Coupling. *J. Am. Chem. Soc.* **116**, 11199–11200 (1994).

[60] Losonczi, J. A., Andrec, M., Fischer, M. W. F., and Prestegard, J. H. Oder Matrix Analysis of Residual Dipolar Couplings Using Singular Value Decomposition. *J. Magn. Reson.* **138**, 334–342 (1999).

[61] Clore, G. M., Gronenborn, A. M., and Bax, A. A Robust Method for Determining the Magnitude of the Fully Asymmetric Alignment Tensor of Oriented Macromolecules in the Absence of Structural Information. *J. Magn. Reson.* **133**, 216–221 (1998).

[62] Hus, J.-C., Mario, D., and Blackledge, M. Determination of Protein Backbone Structure Using Only Residual Dipolar Couplings. *J. Am. Chem. Soc.* **123**, 1541–11542 (2001).

[63] Moltke, S. and Grzesiek, S. Structural constraints from residual tensorial couplings in high resolution NMR without an explicit term for the alignment tensor. *J. Biomol. NMR* **15**, 77–82 (1999).

[64] Warren, J. J. and Moore, P. B. A Maximum Likelihood Method for Determining D_a^{PQ} and R for Sets of Dipolar Coupling Data. *J. Magn. Reson.* **149**, 271–275 (2001).

[65] Jack, A. and Levitt, M. Refinement of Large Structures by Simultaneous Minimization of Energy and R Factor. *Acta Cryst. sect. A* **34**, 931–935 (1978).

[66] Brünger, A. T. and Karplus, M. Molecular Dynamics Simulations with Experimental Restraints. *Acc. Chem. Res.* **24**, 54–61 (1991).

[67] Flory, P. J. *Statistical Mechanics of Chain Molecules.* Carl Hanser Verlag, Munich, (1969).

[68] Momany, F. A., McGuire, R. F., Burgess, A. W., and Scheraga, H. A. Energy Parameters in Polypeptides. VII. Geometric Parameters, Partial Atomic Charges, Nonbonded Interactions, Hydrogen Bond Interactions, and Intrinsic Torsional Potentials for the Naturally-Occurring Amino Acids. *J. Phys. Chem.* **79**, 2361–2381 (1975).

[69] Némethy, G., Pottle, M. S., and Scheraga, H. A. Energy Parameters in Polypeptides. 9. Updating of Geometrical Parameters, Nonbonded Interactions, and Hydrogen Bond Interactions for the Naturally-Occurring Amino Acids. *J. Phys. Chem.* **87**, 1883–1887 (1983).

[70] Güntert, P., Mumenthaler, C., and Wüthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–98 (1997).

Danksagung

Ich danke Prof. Dr. Dr. Hans-Robert Kalbitzer für die unkomplizierte Zusammenarbeit und seine Bereitwilligkeit, meine Arbeit als externe Dissertation gegenüber der Fakultät für Biologie der Universität Regensburg zu vertreten.

Ich danke Dr. Michael Nilges dafür, daß er mir die Möglichkeit gab, diese Arbeit in seiner Forschungsgruppe am Institut Pasteur, Paris durchzuführen, und finanzielle Unterstützung im Rahmen des EU-Projekts NMRQUAL zur Verfügung stellte. Seine stete Ansprechbarkeit und Offenheit sowie sein Interesse an der Arbeit ermöglichten ein sehr angenehmes Arbeiten.

Ich danke Wolfgang Rieping für die gute Zusammenarbeit und zahlreiche Diskussionen, in denen das Konzept der Inferentiellen Strukturbestimmung entwickelt wurde. Erst durch das zusammen mit ihm geschriebene Simulationspacket ISD konnten die in dieser Arbeit entwickelten Ideen auch praktisch umgesetzt werden.

Lebenslauf

Persönliche Daten

Name:	Habeck
Vorname:	Michael
Titel:	Diplom-Physiker
Staatsangehörigkeit:	deutsch
Geburtsdatum:	27. Januar 1973
Wohnort:	Paris
Familienstand:	ledig

Ausbildung

Hochschulreife:	Abitur am 23. Juni 1992, Ahrweiler
Zivildienst:	von August 1992 bis September 1993 in der Ehrenwall'schen Klinik, Ahrweiler
Studium:	Physikstudium von Oktober 1993 bis Mai 1999 in Siegen und Heidelberg; Diplom am 31. Mai 1999 in Heidelberg
Gastwissenschaftler:	EMBL, Heidelberg in der Gruppe von Dr. M. Nilges (Juli 1999 bis Oktober 2000)
Software-Entwickler:	LION Bioscience AG, Heidelberg (Oktober 1999 bis Oktober 2000)

Doktorarbeit:	Nigles-Gruppe EMBL, Heidelberg von Oktober 2000 bis März 2001; seit März 2001 am Institut Pasteur, Paris in der Gruppe Bioinformatique Structurale geleitet von M. Nilges