Qualitätskriterien für Proteinstrukturen aus 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

Wolfgang Rieping aus Mannheim

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

Seit den Anfängen von makromolekularer Strukturbestimmung durch hochaufgelöste Kernspinresonanz-Spektroskopie ist die Frage nach der Qualität der erzeugten Strukturen wiederholt gestellt worden: Experimentelle Daten sind verrauscht und unvollständig, NMR-Parameter hängen von einer Vielzahl physikalischer Effekte ab, die nicht oder nur näherungsweise beschrieben werden können. Die vorliegende Arbeit befaßt sich mit der Fragestellung, wie sich experimentelle Ungenauigkeiten und Näherungen in den theoretischen Modellen quantitativ auf die Verläßlichkeit auswirken, mit der die Positionen der einzelnen Atome einer NMR-Struktur berechnet werden können.
Es zeigt sich, daß die Bestimmung dieser Verläßlichkeit als Teil des Strukturberechnungsprozesses selbst aufgefaßt werden muß, konventionelle Methoden für eine objektive Beantwortung dieser Fragestellung jedoch grundsätzlich ungeeignet sind. Die in dieser Arbeit vorgestellten Methoden basieren auf dem Prinzip der induktiven Strukturbestimmung, einem neuen, wahrscheinlichkeitstheoretischen Zugang zur Strukturbestimmung.
Es wurde eine Methode entwickelt, welche die objektive Berechnung der strukturellen Unsicherheitsbehaftung einer NMR-Struktur im Sinne eines atomweisen Fehlerbalkens gestattet. Die berechneten Koordinatenunsicherheiten werden eindeutig von den experimentellen Daten sowie von Zusatzannahmen bestimmt, die für die Interpretation der Daten vonnöten sind. Die Methode basiert auf einem probabilistischen Verteilungsmodell, das mit Hilfe eines Markov-Ketten-Monte-Carlo-Algorithmus aus den Daten geschätzt wird. Beispielrechnungen mit NOESY-Daten demonstrieren die Methode sowie die Effizienz des Algorithmus.
Die Qualität eines Datensatzes nimmt unmittelbaren Einfluß auf die Koordinatenunsicherheiten. Ein intuitiv zu interpretierendes Maß für die Qualität eines Datensatzes folgt aus der statistischen Beschreibung der experimentellen Messungen. Verglichen mit herkömmlichen, externen Qualitätsmaßen ist

für seine Bestimmung kein zusätzlicher Rechenaufwand nötig. Eigenschaften des Konsistenzmaßes sowie die Auswirkungen von Inkonsistenzen in den Daten auf die Verläßlichkeit einer NMR-Struktur werden anhand zahlreicher Simulationen untersucht.

Um systematische Fehler und Unsicherheiten in den Koordinaten einer NMR-Struktur zu reduzieren, wurde ein Datenmodell für dipolare Kreuzrelaxationsraten entwickelt, welches dynamikinduzierte Inkonsistenzen in den Messungen explizit berücksichtigt. In dem verfolgten Ansatz wird der Grad der Inkonsistenz für jede Messung individuell modelliert. Die Zahl der zu bestimmenden Parameter übersteigt daher grundsätzlich die Anzahl der Messungen. Beispielrechnungen zeigen, daß in einem wahrscheinlichkeitstheoretischen Zugang auch komplexe Modelle verläßlich aus den Daten geschätzt werden können. Dies bedeutet ein hohes Maß an Flexibilität bei der Beschreibung experimenteller Meßgrößen: Mit Hilfe des vorgestellten Datenmodells wird die Qualität einer Struktur deutlich verbessert und Unsicherheiten in den dreidimensionalen Koordinaten zugleich reduziert.

Inhalt

Kapitel 1

Einleitung

Detailliertes Wissen über die dreidimensionale Struktur biologischer Makromoleküle ist für ein Verständnis ihrer intrazellulären Funktion von großer Wichtigkeit und gewinnt bei der Modellierung mechanistischer Eigenschaften von Biomolekülen zunehmend an Bedeutung. In der Strukturbiologie haben sich zwei experimentelle Techniken zur Aufklärung der atomaren Struktur von Proteinen und Nukleinsäuren etabliert: Röntgenkristallographie und Kernspinresonanz-Spektroskopie. In der Röntgenkristallographie werden Beugungsmuster von Röntgenstrahlung an geordneten Mikrokristallen für die Strukturbestimmung genutzt. Typische Implementierungen dieser Technik erzielen eine Auflösung von unter 2 Å und erlauben die Rekonstruktion der Positionen der schweren Atome eines Moleküls im Raum.

1.1 Kernspinresonanz-Spektroskopie

Die Kernspinresonanz-Spektroskopie (*Nuclear Magnetic Resonance*, NMR) hat sich seit ihren ersten Anwendungen für die experimentelle Strukturaufklärung vor etwa 20 Jahren [1, 2] als zweite Standardmethode etabliert. NMR-Experimente gestatten die Bestimmung der dreidimensionalen Struktur eines Makromoleküls in wässriger Lösung. Die Züchtung von Kristallen entfällt daher. Viele NMR-Observable sind von der molekularen Konforma-

tion abhängig und gestatten somit detaillierte Rückschlüsse über mikroskopische Strukturgrößen: NMR-Parameter enthalten Information über interatomare Abstände, über Werte von Dihedralwinkeln oder Winkel zwischen Bindungsvektoren. Diese Information kann mit Hilfe theoretischer Modelle analysiert und für die Berechnung der dreidimensionalen Struktur eines Makromoleküls genutzt werden. NMR-Observable sind darüber hinaus sensitiv gegenüber der Bewegung eines Moleküls. Neben der Strukturaufklärung erlauben NMR-Experimente daher auch die Charakterisierung dynamischer Eigenschaften biologischer Makromoleküle. So läßt sich beispielsweise die Beweglichkeit des Proteinrückgrats auf der Piko- bis Nanosekunden-Zeitskala anhand von ^{15}N Relaxations Experimenten studieren [3, 4].

1.1.1 Kerne im äußeren Magnetfeld

Die physikalische Größe, die einem NMR-Experiment zugrunde liegt, ist der Eigendrehimpuls (*Spin*) eines Atomkerns. Atomkerne mit nicht verschwindendem Spin $\mathbf{I}$ besitzten ein magnetisches Dipolmoment $\mu = \gamma\mathbf{I}$. γ bezeichnet das gyromagnetische Verhältnis des Atomkerns. In der Strukturbiologie relevante Elemente, im wesentlichen sind dies ^{1}H, ^{13}C, ^{15}N, ^{19}F und ^{31}P, tragen Kernspin 1/2. Die Kopplung eines Spin-1/2 Kerns an ein äußeres, in z-Richtung orientiertes, statisches Magnetfeld der Stärke B_0, führt zu einer Aufspaltung der Energie in zwei Niveaus $E_m = -\gamma I_z B_0 = -m\gamma\hbar B_0$. Die korrespondierenden Spinzustände werden durch die magnetische Quantenzahl $m = \pm 1/2$ charakterisiert. Im thermodynamischen Gleichgewicht gehorchen die Besetzungszahlen der Spinzustände der Boltzmann-Statistik:

$$N_{-1/2}/N_{+1/2} = \exp(-\beta\Delta E) \quad \text{mit} \quad \Delta E = \gamma\hbar B_0. \tag{1.1}$$

$N_{-1/2}$ ($N_{+1/2}$) bezeichnet die Besetzungszahl des Spinzustands mit höherer (niedrigerer) Energie. $\beta = 1/k_B T$, k_B bezeichnet die Boltzmann-Konstante und T die Temperatur der Probe. Die Asymmetrie in den Besetzungszahlen führt zu einer makroskopischen Nettomagnetisierung in z-Richtung. Selbst bei hohen Feldstärken unterscheiden sich die Besetzungszahlen nur schwach,

was für die geringe Intensität eines NMR-Signals verantwortlich ist. Durch Einstrahlen eines Hochfrequenz (HF) Signals bei der *Larmorfrequenz* eines Atomkerns,

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

werden Übergange zwischen den Spinzuständen induziert, wodurch sich die Besetzungszahlen der Zustände gezielt manipulieren lassen. Nach Gl. (1.2) hängt die Resonanzfrequenz eines Atomkerns von der Stärke des äußeren Magnetfelds ab und wird zusätzlich von seiner chemischen Umgebung beeinflußt: Die Bewegung eines Moleküls induziert lokale Magnetfelder, welche sich dem äußeren Feld überlagern. Dadurch ändert sich das vom Kern wahrgenommene effektive Magnetfeld, was eine Verschiebung seiner Resonanzfrequenz zur Folge hat („*chemische Verschiebung*").

1.1.2 Relaxation

Nach dem Einstrahlen eines HF-Signals und der damit einhergehenden Änderung der relativen Besetzungszahlen kehrt das System in sein thermodynamisches Gleichgewicht zurück. Dieses Phänomen wird als *Relaxation* bezeichnet. Die Ursache von Relaxationsprozessen ist die Kopplung der Kernspins an ihre Umgebung. Die Kopplung erfolgt über lokale Magnetfelder, welche durch die thermische Bewegung des Moleküls erzeugt werden. Die Messung von Relaxationsparametern durch NMR-Experimente gestattet somit Rückschlüsse über die physikalischen Prozesse, welche für die Relaxation verantwortlich sind. Phänomenologisch unterscheidet man zwischen Spin-Gitter- und Spin-Spin Relaxation. Spin-Spin Relaxation zerstört die Phasenkohärenz der Magnetisierungskomponenten und bedingt dadurch den Zerfall der makroskopischen Transversalmagnetisierung. Spin-Spin Relaxation erfolgt durch Übergänge zwischen Spinzuständen, deren relative Besetzungszahlen davon jedoch unbeeinflußt bleiben. Bei der Spin-Gitter Relaxation führt eine Änderung der relativen Populationen von Spinzuständen zu der Wiederherstellung der z-Magnetisierung. Als Mediator für diesen Energieaustausch fungiert das *Git-*

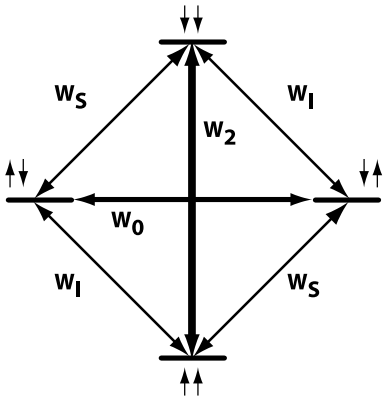

Abbildung 1.1: Energieniveaus, Übergangskanäle und Ratenkonstanten für ein System von zwei gekoppelten Spin-1/2 Kernen. Für den NOE wichtige Übergänge sind der Doppeltquanten- und Nullquantenübergang mit den Ratenkonstanten W_2 bzw. W_0.

ter, d.h. die zahlreichen Translations-, Rotations- und Schwingungsfreiheitsgrade der Probe, welche ein Quasikontinuum von Energiezuständen bilden.

1.1.3 Der nukleare Overhauser Effekt

Ein wichtiger Relaxationsparameter ist der nukleare Overhauser Effekt (NOE) [5, 6]. Der NOE basiert auf dem Energieübertrag zwischen Spinzuständen, welcher über die dipolare Wechselwirkung realisiert wird. Der NOE ist für die NMR-Strukturaufklärung von großer Bedeutung, da seine Stärke von der räumlichen Nähe der wechselwirkenden Kerne abhängt.

Dipolare Relaxation in zwei-Spin Systemen

Ein System, bestehend aus zwei gekoppelten, homonuklearen Kernspins I und S, besitzt vier Spinzustände. Bei abwesender skalarer Kopplung induziert die dipolare Wechselwirkung Übergänge zwischen diesen Zuständen über vier Kanäle mit den Ratenkonstanten W_S, W_I, W_0 und W_2 (vgl. Abb. 1.1). Die Übergänge W_S und W_I ändern den Zustand eines Spin; der Nullquantenübergang W_0 ändert beide Spins, läßt den Gesamtspin jedoch unverändert; der

Doppelquantenübergang W_2 ändert den Gesamtspin des Systems um 1. Befinden sich beide Spins in räumlicher Nähe, so beeinflußt die Manipulation von Spin S den Zustand von Spin I: Während einer Anregung von S führen die Übergänge W_2 und W_0 zu einer Erhöhung bzw. Verminderung der Besetzungszahlen von Spin I; es wird also Magnetisierung von S auf I übertragen. Dies ist der nukleare Overhauser Effekt. Der Magnetisierungsübertrag findet durch den Raum statt und ist aufgrund der dipolaren Wechselwirkung von der Distanz beider Kerne abhängig.

Die Solomon Gleichungen

Die Solomon Gleichungen [7] bilden die Grundlage der NOE-Theorie und beschreiben die Relaxation von zwei isolierten Spins im Falle einer reinen Dipol-Dipol Wechselwirkung. Die zeitliche Entwicklung der Magnetisierung von Spin I in z-Richtung gehorcht der Differentialgleichung

$$\frac{dI_z(t)}{dt} = -\left[I_z(t) - I_z^0\right]\rho_I - \left[S_z - S_z^0\right]\sigma_{IS}. \tag{1.3}$$

$\rho_I = W_0 + 2W_I + W_2$ und $\sigma_{IS} = W_2 - W_0$ bezeichnen die Auto- bzw. Kreuzrelaxations-Ratenkonstanten, mit

$$\rho_I = \frac{2\pi}{5}\gamma_H^4\hbar^2\left[J(0) + 3J(\omega_0) + 6J(2\omega_0)\right], \tag{1.4}$$

$$\sigma_{IS} = \frac{2\pi}{5}\gamma_H^4\hbar^2\left[6J(2\omega_0) - J(0)\right]. \tag{1.5}$$

ω_0 und γ_H bezeichnen die Larmorfrequenz bzw. das gyromagnetische Verhältnis von Spin I und S. Die *spektrale Dichte* $J(\omega)$ beschreibt die Modulation der dipolaren Kopplung, welche durch Fluktuationen des interatomaren Verbindungsvektors zwischen beiden Spins, $\mathbf{r}(t)$, relativ zum äußeren Magnetfeld, verursacht wird:

$$J(\omega) = \int_{-\infty}^{+\infty} dt\, C(t)\, cos(\omega t). \tag{1.6}$$

$C(t)$ bezeichnet die Autokorrelationsfunktion von $\mathbf{r}(t)$. Im Falle einer isotopen, mit der internen Bewegung unkorrelierten Rotation des Moleküls, faktorisiert die Korrelationsfunktion in einen äußeren Anteil $C_O(t)$ und einen

inneren Anteil $C_I(t)$ [8, 9]. $C_O(t)$ fällt exponentiell mit Korrelationszeit τ_c ab; für die vollständige Korrelationsfunktion folgt [10]:

$$C(t) = \frac{1}{4\pi} e^{-t/\tau_c} \left\langle \frac{P_2(\hat{\mathbf{r}}(0) \cdot \hat{\mathbf{r}}(t))}{d^3(0) d^3(t)} \right\rangle . \tag{1.7}$$

$d = \|\mathbf{r}\|$, $\hat{\mathbf{r}}$ bezeichnet den normierten interatomaren Verbindungsvektor im lokalen Koordinatensystem des Moleküls, $P_2(\cdot)$ das Legende Polynom 2. Ordnung. $\langle \cdot \rangle$ steht für den Ensemble-Mittelwert.

Bei Annahme eines starren Moleküls entfällt der Ensemble-Mittelwert in Gl. (1.7) und die Kreuzrelaxationsrate ist eine einfache Funktion des interatomaren Abstands beider Kerne:

$$\sigma_{IS} = \frac{1}{10} \gamma_H^4 \hbar^2 \frac{1}{d^6} \left(\frac{6\tau_c}{1 + 4\omega^2 \tau_c^2} - \tau_c \right) . \tag{1.8}$$

Kreuzrelaxationsraten lassen sich im Prinzip durch eine Analyse sogenannter *Buildup*-Kurven aus NOE-Messungen extrahieren. *Buildup*-Kurven geben zugleich Aufschluß über die Stärke der Spindiffusionsbehaftung einer Messung. Für kurze Mischzeiten verläuft der zeitliche Aufbau des NOE linear mit einer Anstiegsrate proportional zu σ_{IS} [11]. Kreuzrelaxationsraten werden daher üblicherweise durch die integrierte Intensität („Volumen“) eines NOE bei kurzer Mischzeit angenähert. Für das theoretische Volumen eines NOE folgt in der *Isolated Spin Pair Approximation* (ISPA):

$$V = \gamma d^{-6}. \tag{1.9}$$

γ bezeichnet einen Skalenfaktor, der bei bekanntem τ_c im Prinzip aus Gl. (1.8) berechnet werden kann, in der Praxis jedoch unbekannt ist [6].

Experimentelle Bestimmung von Kreuzrelaxationsraten

Für die Bestimmung eines NMR-Spektrums hat sich die Methode der Fourier-Transformations NMR [12, 13] bewährt. Durch Einstrahlen eines HF-Impluses werden alle Kernspins gleichzeitig in Resonanz versetzt. Die Antwort des Systems, d.h. der zeitliche Abfall der Magnetisierung auf ihren Gleichgewichtswert, wird in Form des freien Induktionsabfalls (*Free Induction Decay*, FID)

aufgenommen und mittels Fourier Transformation in ein Frequenzspektrum überführt. Die einzelnen Resonanzen des Spektrums korrespondieren zu den Magnetisierungskomponenten der bei unterschiedlichen Frequenzen resonierenden Kernspins. Eine systematische Bestimmung von Kreuzrelaxationsraten durch eindimensionale Experimente ist zeitaufwendig. Das zweidimensionale NOESY (*Nuclear Overhauser Enhancement Spectroscopy*) Experiment [12] bildet die Grundlage mehrdimensionaler NMR-Techniken und gestattet die simultane Beobachtung aller dipolaren Kreuzrelaxationsprozesse: Mit Hilfe einer Folge von HF-Impulsen werden die Magnetisierungskomponenten räumlich benachbarter Kernspins durch Ausnutzen des NOE miteinander korreliert. Nach einer zweidimensionalen Fourier Transformation sind räumlich benachbarte Spins in einem NOESY-Spektrum in Form von *Kreuzresonanzen* sichtbar. Aufgrund der Abstandsabhängigkeit des NOE enthält die integrierte Intensität („Volumen“) einer Kreuzresonanz Information über den interatomaren Abstand der involvierten Spinsysteme und kann somit für die Strukturrechnung genutzt werden.

1.2 Strukturbestimmung

Theoretische Methoden zur makromolekularen Strukturbestimmung verfolgen das Ziel, aus experimentellen Daten und *a-priori*-Wissen Information über die Positionen der einzelnen Atome eines Moleküls im Raum zu gewinnen. Dies ist das *Strukturbestimmungsproblem*. Zentraler Bestandteil von Strukturberechnungsmethoden ist der theoretische Zusammenhang zwischen den Koordinaten der makromolekularen Konformation und den experimentellen Meßgrößen. Diese Theorie ist üblicherweise in Form eines *Vorwärtsmodells* implementiert, welches die Berechnung einer Meßgröße V_i als Funktion der Strukturkoordinaten $\mathbf{x}$ gestattet. Eine informative Meßgröße ist beispielsweise der NOE. Das Vorwärtsmodell, die ISPA in Gl. (1.9), beschreibt den Zusammenhang zwischen der theoretischen Größe einer dipolaren Kreuzrelaxationsrate und dem korrespondierenden interatomaren Abstand. Für die

Herstellung einer exakten Relation greift die Theorie fast immer auf zusätzliche Hilfsgrößen zurück: Die Skala observierter Kreuzrelaxationsraten ist in der Regel unbekannt und muß in der ISPA durch einen freien Parameter berücksichtigt werden. Allgemein ist das Vorwärtsmodell daher eine Funktion der Koordinaten und der Hilfsparameter $\alpha = \{\alpha_1, \ldots, \alpha_H\}$:

$$V_i = V_i(\mathbf{x}, \alpha). \tag{1.10}$$

Für die Berechnung der dreidimensionalen Struktur eines Moleküls seien experimentelle Daten D in Form von N observierten Größen $D = \{\tilde{V}_1, \ldots, \tilde{V}_N\}$ gegeben.

1.2.1 Konventionelle Methoden

Konventionelle Methoden formulieren Strukturbestimmung als Inversionsproblem: Dieser Ansatz hat das Ziel, die gesuchte Konformation durch Auswerten des inversen Vorwärtsmodells an den observierten Daten auf direkte Weise zu berechnen.
Die mathematische Umsetzung des Inversionsproblems erfolgt durch die Formulierung eines Optimierungsproblems: Dazu werden mit Hilfe einer *Verlustfunktion* $f(\tilde{V}_i, V_i(\mathbf{x}, \alpha))$ Zwangsbedingungen formuliert, um die observierten Strukturgrößen mit den theoretischen Werten in Beziehung zu setzen. Die Verlustfunktion ist per Definition minimal, wenn die berechnete mit der observierten Strukturgröße übereinstimmt, d.h. wenn gilt $\tilde{V}_i \equiv V_i(\mathbf{x}, \alpha)$. In der Praxis werden Zwangsbedingungen häufig über quadratische Verlustfunktionen („harmonische Potentiale“) realisiert. Die Gesamtheit aller Zwangsbedingungen definiert die Verlustfunktion für den Gesamtdatensatz,

$$E_{\text{Daten}}(\mathbf{x}; \alpha) = \sum_{i=1}^{N} f(\tilde{V}_i, V_i(\mathbf{x}, \alpha)), \tag{1.11}$$

wobei alle Meßwerte als voneinander unabhängig angenommen wurden. Die „Energie“-Funktion der Daten in Gl. (1.11) quantifiziert die Übereinstimmung der molekularen Konformation mit den experimentellen Daten und ist

im Falle vollkommener Übereinstimmung minimal.
Für NMR-Daten führt die Minimierung von Gl. (1.11) zu physikalisch unrealisitischen Konformationen, weshalb das Inversionsproblem regularisiert werden muß. NMR-Daten werden daher generell ergänzt durch physikalisches *a-priori*-Wissen über kovalente und nichtkovalente Wechselwirkungen des Moleküls. Terme zur Beschreibung dieser Wechselwirkungen sind in Molekulardynamik (MD) Kraftfeldern in Form einer physikalischen Energiefunktion E_{Phys} implementiert [14]. Die Zielfunktion des Optimierungsproblems ist die Hybridenergiefunktion

$$E_{\text{Hybrid}}(\mathbf{x}) = E_{\text{Phys}}(\mathbf{x}) + wE_{\text{Daten}}(\mathbf{x};\alpha). \tag{1.12}$$

Das Datengewicht („Kraftkonstante“) w ist ein Skalierungsfaktor, der empirisch bestimmt werden muß. Das Inversionsproblem wird durch Minimierung der Hybridenergiefunktion bezüglich $\mathbf{x}$ gelöst. Praktisch geschieht die Lokalisierung des Energieminimums mit Hilfe von nichtlinearen Minimierungsstrategien [15], d.h. die Inverse des Vorwärtsmodells wird in Form eines Algorithmus implementiert. Die Konformation mit minimaler Gesamtenergie wird als native Struktur des Moleküls interpretiert und erfüllt im Idealfall die Daten sowie die physikalischen Randbedingungen. Die Hilfsgrößen α und das Datengewicht w sind freie Parameter des Minimierungsproblems, die nicht durch Gl. (1.12) festgelegt sind und über externe Kriterien bestimmt werden müssen. Der Skalenfaktor γ in der ISPA wird beispielsweise durch die „Kalibrierung“ der NOE-Daten bestimmt. Die Vorschrift,

1. Formuliere Hybridenergiefunktion,
2. Berechne native Konformation durch Minimierung der Gesamtenergie,

bildet die Grundlage existierender Strukturbestimmungsmethoden und wird als (prozedurale) Formulierung des Strukturbestimmungsproblems angesehen [15, 16, 17]. Die globale Konvergenz der verwendeten Minimierungsstrategien kann in der Praxis nicht garantiert werden. Es hat sich daher als günstig erwiesen, durch die wiederholte Anwendung der Berechnungsvorschrift einen

Satz von Strukturen zu generieren, das *NMR-Strukturensemble*. Die Streuung der Strukturen ist ein Maß für die Eindeutigkeit der Lösung und die numerische Stabilität des verwendeten Minimierungsprotokolls.

1.2.2 Qualität von NMR-Strukturen

Die Frage nach der Qualität von NMR-Strukturen wird seit den Anfängen von NMR-Strukturbestimmung wiederholt gestellt: Experimentelle Daten sind verrauscht, NMR-Parameter hängen von einer Vielzahl physikalischer Effekte ab, die nicht oder nur näherungsweise beschrieben werden können. NMR-Daten sind zudem von indirekter Natur und bedürfen stets der Vorprozessierung, welche eine zusätzliche Fehlerquelle darstellen kann. Die Analyse des NOE beispielsweise ist von großer Schwierigkeit in mehrerer Hinsicht: Neben experimentellen Unsicherheiten führen insbesondere Näherungen in der Theorie zu Abweichungen der beobachteten von den berechneten Kreuzrelaxationsraten. Dies führt zu inkonsistenten Datensätzen in dem Sinne, daß nicht alle Messungen simultan anhand einer einzelnen Struktur erklärt werden können.

Näherungen in der Theorie

Systematische Beiträge zu NOE-Intenstitäten werden primär durch Spindiffusion [18, 13] und die interne Dynamik eines Moleküls verursacht. Spindiffusion entsteht durch den indirekten Übertrag von Magnetisierung durch ein Netzwerk räumlich benachbarter Spins. Multispineffekte werden durch die Relaxationsmatrixtheorie beschrieben und können somit bei der Berechnung von theoretischen Kreuzrelaxationsraten näherungsweise berücksichtigt werden [19, 20, 21].
Die Berechnung von Dynamikkorrekturen ist jedoch von großer Schwierigkeit: Kreuzrelaxationsraten sind Zeit- und Ensemble-Mittelwerte und somit sensitiv gegenüber den mikroskopischen Details der internen Dynamik eines Moleküls. Abstandsfluktuationen eines Protonpaars können die Intensität eines NOE signifikant erhöhen [22, 23]. Die atomare Bewegung ist komplex und

findet auf mehreren Zeitskalen statt. Eine verläßliche Simulation aller relevanten Effekte ist schwierig, weshalb Dynamikbeiträge bei der Berechnung von theoretischen Kreuzrelaxationsraten meist unberücksichtigt bleiben.

Unvollständigkeit der Daten

Ein zusätzliches Problem ist die Unvollständigkeit von NMR-Datensätzen: Aufgrund seiner inversen Abstandsabhängigkeit ist der NOE nur für Wasserstoffe mit einem interatomaren Abstand von weniger als etwa 5 Å meßbar. Selbst unter idealen experimentellen Voraussetzungen ist die Zahl der Messungen daher unzureichend, um die Struktur eines Makromoleküls eindeutig festzulegen. Dieses Problem wird durch die Berücksichtigung von physikalischem *a-priori*-Wissen reduziert, bleibt prinzipiell jedoch bestehen. In der Praxis kann sich die Unvollständigkeit und Fehlerhaftigkeit der Ausgangsdaten durch Probleme bei der Interpretation der Spektren (Fehler bei der Auswahl von Kreuzresonanzen („*peak picking*“) in stark populierten Regionen eines Spektrums, Fehlzuordnungen von Resonanzen oder Kreuzresonanzen, Resonanzverdoppelung durch starke Kopplung, Rausch-Artefakte in den Spektren oder konformationelle Mittelungseffekte) weiter erhöhen.

Bewertung der Qualität einer Struktur

In der Röntgenkristallographie haben sich verschiedene R-Faktoren zur Bestimmung der Qualität einer Struktur bewährt [24, 25]. Ein R-Faktor bewertet die Übereinstimmung der experimentellen Daten mit den theoretischen Werten, die anhand einer Struktur berechnet wurden. Ein zusätzliches Qualitätsmaß ist durch den Debye-Waller-Faktor („B-Faktor“) gegeben, der eigentlich ein Maß für die Dynamik einzelner Atome ist, in der Praxis jedoch häufig zur Bewertung der Verläßlichkeit einer Kristallstruktur dient.
Die Bestimmung der Verläßlichkeit einer NMR-Struktur ist konzeptionell involvierter: Im Gegensatz zur Röntgenkristallographie liefern NMR-Experimente lediglich indirekte Strukturinformation mit einem vergleichsweise geringem Informationsgehalt. Eine direkte Bestimmung der strukturellen Ver-

läßlichkeit rein auf Basis der experimentellen Daten ist aus diesem Grunde bisher nicht möglich. Die Qualität von NMR-Strukturen wird daher mittels externer Kriterien bewertet. Existierende Ansätze [26, 27, 28] definieren Qualitäts-Indikatoren für Strukturgrößen wie Haupt- und Seitenketten-Dihedralwinkel, die Atompackung oder Bindungswinkel. Dazu werden berechnete Strukturgrößen mit typischen Werten verglichen, wie sie in der Strukturdatenbank PDB (*Protein Data Bank*) [29] vorgefunden werden. Andere Verfahren verwenden wissensbasierte Bewertungsfunktionen, um die Schlüssigkeit der Aminosäuresequenz mit der Struktur zu testen und Fehler in der Faltung eines Proteins zu identifizieren [30]. Die Streuung eines NMR-Strukturensembles (beispielsweise definiert über den RMSD[1] zur mittleren Struktur oder die zirkuläre Varianz von Dihedralwinkeln) wird als Maß für die Eindeutigkeit der erzeugten Struktur verwendet; die Richtigkeit der Struktur läßt sich damit jedoch nicht abschätzen. Für die Bewertung der Konsistenz einer berechneten Struktur mit den experimentellen Daten wurden verschiedene, dem R-Faktor der Röntgenkristallographie verwandte Qualitätsmaße vorgeschlagen [31, 32, 33].

Fehler und Unvollständigkeiten in den experimentellen Daten sowie Näherungen in den theoretischen Modellen sind unvermeidbar und führen zu entsprechenden Unsicherheiten in den dreidimensionalen Koordinaten einer Struktur. Für die Bestimmung der Qualität einer NMR-Struktur – im Sinne eines Fehlerbalkens – ist somit entscheidend, wie sich experimentelle Ungenauigkeiten und Näherungen in der Theorie quantitativ auf die Verläßlichkeit auswirken, mit der die Positionen der einzelnen Atome bestimmt werden können. Die Bestimmung dieser Unsicherheitsbehaftung ist somit Teil des Strukturberechnungsprozesses selbst. Für eine quantitative Behandlung dieser Fragestellung muß daher auch geklärt werden, wie der Begriff der „Unsicherheit“ mathematisch konkretisiert und in der Strukturrechnung geeignet berücksichtigt werden kann.

[1] *Root mean square deviation.* Der RMSD wird bezüglich zwei optimal überlagerter Koordinatensätze berechnet.

Schwierigkeiten des konventionellen Zugangs

Eine Diskussion dieser Problemstellungen innerhalb der konventionellen Formulierung von Strukturbestimmung ist problematisch: Im Inversionszugang bleiben Unsicherheiten in den Ausgangsdaten per Definition unberücksichtigt. Zwar werden Fehler in den Zieldistanzen mittels *ad hoc* Konzepten wie *Abstandsschranken* in der Rechnung berücksichtigt. Die Größe der Schranken wird jedoch nicht durch die Daten festgelegt, sondern muß empirisch bestimmt werden.

Ein weiteres Problem stellen Hilfsgrößen, wie der bereits angesprochene NOE-Skalenparameter dar: Sie sind für die Interpretation der Daten notwendig, können jedoch nicht experimentell bestimmt werden. Die Hybridenergiefunktion in Gl. (1.12) ist lediglich Zielfunktion für die Koordinaten der Struktur – Hilfsgrößen gehen hingegen als freie Parameter ein und sind somit nicht Teil der Lösung des Inversionsproblems. Die Hybridenergiefunktion ist daher unvollständig in dem Sinne, daß aus ihr keine definitiven Regeln für die Bestimmung aller Unbekannten des Problems, also Strukturkoordinaten *und* Hilfsgrößen, abgeleitet werden können. Ein analoges Problem tritt bei Parametern auf, die für die Formulierung der Hybridenergie eingeführt werden müssen: Das Datengewicht w beispielsweise ist von Natur aus keine experimentelle Observable; die Bestimmung eines geeignenten Wertes auf Basis der Hybridenergie ist auch in diesem Fall nicht möglich.

In der Praxis werden Zusatzparameter daher entweder fixiert oder empirisch bestimmt. Konventionelle Strukturbestimmungsmethoden greifen folglich stets auf Heuristiken zurück, was in Hinsicht auf die objektive Interpretierbarkeit der berechneten Strukturen problematisch ist: Heuristiken enthalten *ad hoc* Elemente und sind demzufolge subjektiv. Strukturbestimmung wird instabil in dem Sinne, daß die generierten Koordinaten empfindlich von der Wahl der Heuristik bzw. von der Wahl spezieller Parametereinstellungen abhängen.

Objektive Aussagen über die Unsicherheitsbehaftung von Atomkoordinaten oder über die Schlüssigkeit der verwendeten Datensätze sind innerhalb des Inversionszugangs aus diesen Gründen grundsätzlich unmöglich.

1.2.3 Strukturbestimmung als Induktionsproblem

Die genannten Schwierigkeiten entstehen durch eine unangemessene Formulierung des Strukturbestimmungsproblems, wodurch charakteristische Eigenschaften des Problems nicht geeignet berücksichtigt werden können. Das Strukturbestimmungsproblem ist in seiner Formulierung als Inversionsproblem mathematisch schlecht gestellt und dadurch strenggenommen nicht lösbar: Unvollständige Datensätze können durch mehrere Konformationen erklärt werden. Selbst ein eindeutiges Vorwärtsmodell ist daher degeneriert und nicht invertiertbar. Durch die Berücksichtigung physikalischen *a-priori*-Wissens wird die Degeneriertheit reduziert, bleibt prinzipiell aber bestehen. Experimentelle Fehler und Näherungen in der Theorie führen zu Inkonsistenzen in den Daten. Die Messungen sind somit nicht durch einen einzelnen Koordinatensatz simultan erklärbar. Dadurch können unterschiedliche „nichtnative“ Konformationen mit identischer „Energie“ existieren, wodurch ein formal invertierbares Vorwärtsmodell praktisch degeneriert ist.
Experimentelle Datensätze sind unvollständig und fehlerbehaftet; theoretische Modelle basieren auf Näherungen und freien Parametern, die nicht experimentell bestimmt werden können. Die Ausgangsinformation ist unvollständig, weshalb der *deduktive* Schluß von den Daten auf die Struktur eines Moleküls, wie er im Inversionszugang angestrebt wird, prinzipiell unmöglich ist. Die dreidimensionalen Koordinaten einer Struktur können niemals eindeutig rekonstruiert werden. Die Frage nach der „wahren“ Konformation eines Moleküls ist daher bedeutungslos.

Unvollständige Ausgangsinformation gestattet generell nur den schwächeren, *induktiven* Schluß: Strukturbestimmung ist ein *Induktionsproblem* und verfolgt die allgemeine Fragestellung, *zu welchem Grad* eine Konformation mit der experimentellen Evidenz und relevanter Hintergrundinformation verträglich ist [34, 35].
Anstelle nach der „wahren“ Konformation eines Moleküls zu fragen, steht die Bewertung *aller möglichen* Konformationen eines Moleküls im Vordergrund.

Die Bewertung erfolgt dabei auf Basis der verfügbaren Ausgangsinformation, d.h. experimenteller Evidenz und *a-priori*-Wissen. Die Verläßlichkeit, mit der eine NMR-Struktur berechnet werden kann, wird unmittelbar von der Vollständigkeit der Ausgangsinformation bestimmt: Stehen nur wenige Konformationen mit den Daten in Einklang, so konnte die Struktur verläßlich bestimmt werden. Im Falle unschlüssiger Ausgangsdaten sind hingegen viele Konformationen mit den Daten verträglich – die Konformation des Moleküls kann nicht verläßlich bestimmt werden.

1.3 Überblick über die vorliegende Arbeit

Ich verfolge in meiner Dissertation zwei Ziele: Erstens, die Entwicklung neuer Methoden mit welchen die Verläßlichkeit von NMR-Strukturen sowie die Qualität von NOE-Datensätzen in objektiver Weise bestimmt werden kann. Zweitens, die Berücksichtigung von Inkonsistenzen in NOE-Daten während der Strukturrechnung, um systematische Fehler und Unsicherheiten in den erzeugten Koordinaten zu reduzieren. Ich konzentriere mich dabei auf die Fragestellung, wie Fehler in den Daten, Näherungen in den theoretischen Modellen sowie die resultierenden Unsicherheiten in den Atompositionen mathematisch einheitlich und objektiv beschrieben werden können. Die entwickelten Methoden basieren auf dem Prinzip der *induktiven Strukturbestimmung* [34, 35], einer wahrscheinlichkeitstheoretischen Formulierung des Strukturbestimmungsproblems.

Die Arbeit gliedert sich in vier Teile. Kapitel 2 behandelt theoretische Grundlagen: Kapitel 2.1 führt den Bayes'schen Wahrscheinlichkeitsbegriff ein und bespricht die Gesetze zur konsistenten Manipulation von Wahrscheinlichkeiten. Die Bayes'sche Wahrscheinlichkeitstheorie bildet die Grundlage der in dieser Arbeit vorgestellten Methoden. Kapitel 2.2 handelt von der induktiven Strukturbestimmung und erläutert, wie ein induktives Strukturbestimmungsproblem mit Hilfe der Wahrscheinlichkeitstheorie allgemein formuliert und gelöst wird. Minimierungsalgorithmen, wie sie in der konventionellen

NMR-Strukturbestimmung verwendet werden, sind für die Berechnung eines induktiven Strukturbestimmungsproblems ungeeignet. Kapitel 2.3 behandelt die algorithmischen Aspekte des statistischen Zugangs. Es werden die Grundlagen von Markov-Ketten-Monte-Carlo (MCMC) Techniken sowie existierende MCMC-Algorithmen zur Simulation von Wahrscheinlichkeitsverteilungen vorgestellt. Die bei der Strukturbestimmung auftretenden Wahrscheinlichkeitsverteilungen sind komplex, weshalb eine Simulation mit Standardtechniken ineffizient ist. Die numerische Behandlung eines induktiven Strukturbestimmungsproblems erfolgt mit Hilfe einer hierarchischen Simulationsstrategie, welche mehrere MCMC-Techniken kombiniert und verallgemeinert. Diese Simulationsstrategie bildet den Kernalgorithmus, welcher für alle Strukturrechnungen in dieser Arbeit verwendet wurde. Verwendete Softwarekomponenten, Testsysteme und Testdatensätze sind in den Kapiteln 2.4 und 2.5 aufgeführt.

Die Unsicherheitsbehaftung einer NMR-Struktur ist implizit im Ergebnis einer induktiven Strukturrechnung kodiert. Der Ergebnisteil der Arbeit beginnt mit Kapitel 3.1. Es wird eine allgemeine Methode beschrieben, welche die Berechnung der Koordinatenunsicherheiten einer Struktur, im Sinne eines atomweisen Fehlerbalkens gestattet. Die Methode basiert auf einem wahrscheinlichkeitstheoretischen Modell für dessen Berechnung ein MCMC-Algorithmus entwickelt wurde. Ich stelle das Modell sowie den Algorithmus im Detail vor und demonstriere die Methode anhand von mehreren Testrechnungen. Die Rechnungen basieren auf NOESY-Datensätzen für das Protein BPTI und die Tudor Domäne des humanen SMN Proteins. Beide Proteine dienen in dieser Arbeit als Testsysteme.

Kapitel 3.2 widmet sich der Bestimmung der Qualität eines NOE-Datensatzes. Ich stelle zwei Bewertungsmaße vor: Das eine Maß quantifiziert die Qualität des Gesamtdatensatzes, das zweite Maß die Schlüssigkeit jeder Einzelmessung in Bezug auf die Gesamtheit aller Daten. Beide Größen folgen in natürlicher Weise aus der statistischen Beschreibung experimenteller Daten und werden während der Strukturrechnung bestimmt. Ihre Bestimmung ist Teil der Ge-

samtstrategie für die Berechnung eines induktiven Strukturbestimmungsproblems und erfordert daher keinen zusätzlichen Aufwand. Anhand von zahlreichen Testsimulationen diskutiere ich die Eigenschaften beider Maße sowie den Einfluß von Inkonsistenzen in den Daten auf die Qualität einer NMR-Struktur.

Um systematische Fehler und Unsicherheiten in den berechneten Koordinaten zu reduzieren, wurde ein Datenmodell für dipolare Kreuzrelaxationen entwickelt, welches Inkonsistenzen in den Messungen explizit berücksichtigt (Kapitel 3.3). Das Datenmodell bewertet die Konsistenz der Messungen auf einer individuellen Basis. Dies erfordert die Einführung einer Vielzahl unbekannter Parameter, deren Anzahl die Zahl der Messungen übersteigt. Ich demonstriere, daß in der induktiven Strukturbestimmung auch komplexe Modelle verläßlich aus den Daten geschätzt werden können. Die modellseitige Berücksichtigung systematischer Fehler in den Daten gestattet ferner die Bewertung jeder Einzelmessung in Hinsicht auf ihre strukturelle Erfüllbarkeit.

In Kapitel 4 werden die Ergebnisse der Arbeit diskutiert, Kapitel 5 gibt einen Ausblick auf mögliche Erweiterungen des statistischen Zugangs.

Kapitel 2

Materialien und Methoden

2.1 Bayes'sche Wahrscheinlichkeitstheorie

In der traditionellen, „frequentistischen" Definition des Wahrscheinlichkeitsbegriffs quantifiziert eine Wahrscheinlichkeit die relative Häufigkeit des Auftretens eines Ereignisses in einer Folge mehrfach wiederholter Experimente oder in einem Ensemble identisch präparierter Systeme.
Der *Bayes'schen Wahrscheinlichkeitstheorie* [36] liegt eine allgemeinere Definition des Wahrscheinlichkeitsbegriffs zugrunde, welche Wahrscheinlichkeit als Maß für die Plausibilität einer Hypothese auffaßt. Der Bayes'sche Wahrscheinlichkeitsbegriff kann daher in Situationen angewendet werden, in welchen ein sicherer Schluß auf das Zutreffen oder Nichtzutreffen einer Hypothese aufgrund der Unvollständigkeit der zur Verfügung stehenden Information nicht möglich ist.

Die Cox'schen Axiome

Um den Begriff der „Plausibilität" zu formalisieren, wird jeder Hypothese ein *Plausibiliätswert* zugeordnet. Ein Plausibilitätswert ist eine abstrakte Größe, die unser *persönliches Vertrauen* in den Wahrheitsgehalt einer Hypothese repräsentiert. R. Cox formulierte mit den nach ihm benannten Axiomen grundlegende Forderungen nach logischer Konsistenz, die ein Plausibilitäts-

maß erfüllen muß, um in sich schlüssig zu sein. Er konnte zeigen, daß jedes Plausibilitätsmaß auf eine Wahrscheinlichkeit abgebildet werden kann, und daß die Regeln der Wahrscheinlichkeitstheorie den einzigen Formalismus für konsistentes Schließen aus unvollständiger Information bilden [37, 38, 36]. Damit wurde der Bayes'schen Wahrscheinlichkeitstheorie eine axiomatische Grundlage gegeben.

2.1.1 Die Regeln der Wahrscheinlichkeitstheorie

Der Cox'sche Beweis gibt somit die Rechtfertigung, daß die gewöhnlichen Regeln der Wahrscheinlichkeitsrechnung auch auf beliebige Plausibilitätsmaße angewendet werden können. Die Regeln der Wahrscheinlichkeitstheorie sind die Summen- und die Produktregel:

$$P(A|B) + P(\bar{A}|B) = 1, \tag{2.1}$$

$$P(AB|I) = P(A|BI)P(B|I). \tag{2.2}$$

A, B und I bezeichnen logische Aussagen (bzw. Hypothesen), die entweder *wahr* oder *nicht wahr* sind. AB ist das logische Produkt, d.h. die Aussage „A und B sind wahr“ und $\bar{A}$ die logische Verneinung.
Die *bedingte Wahrscheinlichkeit* $P(A|B)$ ist eine reelle Zahl zwischen Null und Eins und repräsentiert den Grad unserer persönlichen Überzeugung in Hinsicht auf den Wahrheitsgehalt von Aussage A, unter der Annahme, daß B wahr ist. $P(A|B)$ bewertet somit die Plausibilität des logischen Schlusses „aus dem Zutreffen von B folgt die Richtigkeit von A.“ Bei der „subjektiven“ Definition des Wahrscheinlichkeitsbegriffs werden Hypothesen immer relativ in Bezug auf vorhandenes Wissen I bewertet. Persönliches Vertrauen, ob eine Aussage wahr ist oder nicht, läßt sich ohne diesen Bezug nicht ausdrücken. Wahrscheinlichkeiten sind daher niemals absolut, sondern stets bedingt. Die Bedingung einer Wahrscheinlichkeit drückt dabei eine logische, keine kausale Abhängigkeit aus: Die logische Implikation $B \rightarrow A$ ist also nicht notwendigerweise die Folge eines kausalen Zusammenhangs.

Aus der Summen- und Produktregel folgen zwei Regeln, die für die Lösung von Induktionsproblemen von zentraler Bedeutung sind: Der *Bayes'sche Satz* und die *Marginalisierungsregel* [36]:

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

$$P(A|I) = P(AB|I) + P(A\bar{B}|I). \tag{2.4}$$

Der Bayes'sche Satz in Gleichung (2.3) folgt aus der Produktregel und der Kommutativität des logischen Produkts. Mit Hilfe des Bayes'schen Satzes können bedingte und bedingende Aussagen miteinander vertauscht werden: Auf diese Weise läßt sich die Plausibilität des logischen Schlusses $B \rightarrow A$ über die Plausibilität des Schlusses $A \rightarrow B$ ausdrücken. Der Bayes'sche Satz verkörpert somit den *induktiven Schluß* und ist in Situationen gültig, in welchen der deduktive Schluß aufgrund logischer Uneindeutigkeiten nicht anwendbar ist.
Der Bayes'schen Sichtweise von Wahrscheinlichkeiten wird häufig eine fehlende Objektivität vorgeworfen: Da eine „persönliche Überzeugung" subjektiv sei, müßten Wahrscheinlichkeiten folglich selbst subjektiv sein. Wie Jaynes hervorgehoben hat [36], ist eine Wahrscheinlichkeit tatsächlich ein Maß für die *persönliche* Überzeugung, ob eine Aussage wahr ist. Diese Überzeugung sollte jedoch auf der gesamten Hintergrundinformation beruhen, die für eine Bewertung relevant ist. Objektivität besagt nur, daß Personen, die über identische Hintergrundinformation verfügen, gleiche Wahrscheinlichkeiten vergeben müssen, und daß diese Hintergrundinformation in einer bedingten Wahrscheinlichkeit explizit angegeben werden muß.

2.1.2 Induktionsprobleme

Ich erläutere einige Begrifflichkeiten der Wahrscheinlichkeitstheorie sowie die generelle Vorgehensweise bei der Formulierung und Lösung von Induktionsproblemen am Beispiel der experimentellen Datenanalyse.
Die Problemstellung lautet wie folgt: Aus experimentellen Daten D und Hin-

tergrundinformation I soll auf den unbekannten Zustand eines Systems geschlossen werden. Eine Theorie gestattet dabei die Berechnung der Meßgrößen aus einem gegebenen Zustand. Dies ist ein Problem der Induktion: Experimentelle Daten sind prinzipiell fehlerbehaftet, weshalb der Zustand, in welchem sich das System zum Zeitpunkt der Messung befand, nicht exakt rekonstruiert werden kann.
Um das Induktionsproblem quantitativ zu formulieren, werden die möglichen Zustände des Systems durch einen Satz von alternativen Hypothesen $\varphi = \{\varphi_1, \ldots, \varphi_N\}$ numeriert. Die Hypothese φ_i formalisiert die Aussage „Das System befindet sich in Zustand φ_i." Die Gesamtheit aller Hypothesen spannt den *Hypothesenraum* des Induktionsproblems auf. Im allgemeinen Fall besitzt das System ein Kontinuum von Zusänden, die durch einen kontinuierlichen *Hypothesenparameter* φ numeriert werden.

Likelihood-Funktion, *a-priori*- und *a-posteriori*-Verteilung

Die Lösung des Induktionsproblems bedeutet die Bewertung jeden Punktes im Hypothesenraum anhand der gegebenen Ausgangsinformation, d.h. anhand der Daten und allgemeiner Hintergrundinformation. Die zu bestimmende Größe ist also die bedingte *Wahrscheinlichkeitsdichte* $p(\varphi|D, I)$. $P_R = \int_R d\varphi\, p(\varphi|D, I)$ ist dabei die Wahrscheinlichkeit, daß sich der Zustand des Systems in der Region R des Zustandsraums befindet.
Die Verteilung der Wahrscheinlichkeiten drückt unser Unwissen über den tatsächlichen Zustand des Systems aus: Eine vollständige Ausgangsinformation erlaubt die eindeutige Rekonstruktion des Zustands. In diesem Fall besäße genau eine Hypothese eine endliche Wahrscheinlichkeit, die Wahrscheinlichkeit aller übrigen Hypothesen wäre Null. Unvollständige Information läßt hingegen keinen dezisiven Schluß zu: Der Zustand des Systems ist unsicher – die Wahrscheinlichkeit ist verteilt. Die Wahrscheinlichkeitsverteilung $p(\varphi|D, I)$ repräsentiert demzufolge unser *Wissen* über den Zustand des Systems, jedoch keine realen Fluktuationen des Zustands [36]. Verteilt ist also nicht der Zustand, sondern die Wahrscheinlichkeit.

Durch Anwenden des Bayes'schen Satzes wird das Induktionsproblem formal gelöst:

$$p(\varphi|D,I) = p(\varphi|I)\frac{p(D|\varphi,I)}{p(D|I)}. \tag{2.5}$$

Die Normierungskonstante $p(D|I)$ ist von φ unabhängig und für den Schluß auf den Zustand des Systems nicht weiter von Interesse. Die gesuchte Verteilung ist die *a-posteriori*-Verteilung $p(\varphi|D,I)$. Sie ist das Produkt aus der *a-priori*-Verteilung $p(\varphi|I)$ und der *Likelihood* $p(D|\varphi,I)$.

Die *a-priori*-Verteilung repräsentiert Hintergrundwissen über den Zustand des Systems, welches auch ohne Kenntnis der Daten vorhanden ist. Die *Likelihood* ist eine Funktion des Hypothesenparameters φ und bewertet, wie wahrscheinlich die Daten D beobachten werden, wenn sich das System in Zustand φ befindet. Der Bayes'sche Satz zerlegt die *a-posteriori*-Verteilung somit in zwei natürliche Teile: In *a-priori*-Wissen und experimentelle Information. Die theoretischen Werte der Meßgrößen werden mit Hilfe der Theorie aus einem gegebenen Zustand berechnet, die *Likelihood*-Funktion beschreibt die Diskrepanz zwischen observierten und berechneten Werten. Der logische Schluß vom Zustand des Systems auf die Daten wird durch die Theorie implementiert, seine Plausibilität durch die *Likelihood*-Funktion bewertet. Die Lösung des Induktionsproblems, d.h. die Bewertung der Plausibilität des umgekehrten Schlusses – von den Daten auf den Zustand des Systems – erfolgt durch Anwenden des Bayes'schen Satzes.

Nuisance-Parameter

Für die Berechnung der experimentellen Meßgrößen sind oftmals zusätzliche Parameter α vonnöten, die typischerweise unbekannt und nicht von eigentlichem Interesse sind. Gößen dieser Art werden in der Wahrscheinlichkeitstheorie als *Nuisance* (Ärgernis) Parameter bezeichnet. *Nuisance*-Parameter unterscheiden sich nicht von „echten" Hypothesenparametern und werden in analoger Weise mit Hilfe des Bayes'schen Satzes bestimmt. Dazu wird der Hypothesenraum, der anfangs nur durch die Hypothesen φ aufgespannt wurde, um die Hypothesen α erweitert, welche die möglichen Werte aller

Nuisance-Parameter numerieren. Die *Likelihood*-Funktion besitzt dann die allgemeine Form $p(D|\varphi, \alpha, I)$. Die *a-posteriori-Verbundverteilung* für φ und α folgt unmittelbar aus dem Bayes'schen Satz:

$$p(\varphi, \alpha|D, I) \propto p(\varphi, \alpha|I)p(D|\varphi, \alpha, I). \tag{2.6}$$

Die *a-priori*-Verbundverteilung $p(\varphi, \alpha|I)$ drückt Hintergrundwissen über die möglichen Werte von φ und α aus. Für den Schluß auf den Zustand des Systems sind die Werte der *Nuisance*-Parameter nicht von Interesse. Mit Hilfe der Marginalisierungsregel aus Gleichung (2.4), verallgemeinert auf kontinuierliche Hypothesenparameter, lassen sich *Nuisance*-Parameter eliminieren und man erhält die gesuchte *marginale a-posteriori*-Verteilung für φ:

$$\begin{aligned} p(\varphi|D, I) &= \int d\alpha\, p(\varphi, \alpha|D, I) \\ &\propto \int d\alpha\, p(D|\varphi, \alpha, I)p(\varphi, \alpha|I), \end{aligned} \tag{2.7}$$

wobei Gleichung (2.6) verwendet wurde. Die Eliminierung nicht interessierender Parameter durch Marginalisierung garantiert, daß die Unsicherheitsbehaftung der *Nuisance*-Parameter vollständig in der marginalen *a-posteriori*-Verteilung berücksichtigt wird: Durch Integration über α in Gl. (2.7) wird die *Likelihood*-Funktion über *alle möglichen* Werte der *Nuisance*-Parameter gemittelt; die *a-priori*-Verteilung fungiert dabei als Gewichtungsfunktion. Marginalisierte Dichten weisen daher stets eine höhere Unsicherheitsbehaftung auf als bedingte Dichten [39]. Gleichung (2.7) ist die vollständige Lösung des Induktionsproblems und gestattet die direkte Berechnung der relativen Wahrscheinlichkeiten aller Zustände φ.

2.2 Induktive Strukturbestimmung

Theoretische Methoden zur makromolekularen Strukturbestimmung verfolgen das Ziel, aus experimentellen Daten und *a-priori*-Wissen, Information über die Positionen der einzelnen Atome eines Moleküls im Raum zu gewinnen. Die grundsätzliche Schwierigkeit bei der Lösung dieses Problems liegt in

der Unvollständigkeit der Ausgangsinformation, wodurch die dreidimensionale Struktur eines Makromoleküls niemals eindeutig rekonstruiert werden kann.
Die induktive Strukturbestimmung [34, 35] faßt Strukturbestimmung als Induktionsproblem auf und verfolgt die allgemeine Fragestellung nach der Plausibilität einer Konformation im Lichte der verfügbaren Ausgangsinformation. Die Bayes'sche Wahrscheinlichkeitstheorie bildet den konsistenten Rahmen, um dieses Induktionsproblem eindeutig zu lösen. Ungenauigkeiten in den Messungen und Näherungen in den theoretischen Modellen werden in Form einer *Likelihood*-Funktion explizit berücksichtigt; eine Wahrscheinlichkeitsverteilung repräsentiert die maximale Information über die dreidimensionale Struktur des Moleküls, welche aus der gegebenen Ausgangsinformation abgeleitet werden kann.

Formulierung des Induktionsproblems

Das Zielmolekül bestehe aus M Atomen. Die Konformation des Moleküls sei durch die Angabe von M kartesischen Koordinaten $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ eindeutig bestimmt. Für die Bewertung jeder Konformation stehen experimentelle Messungen D sowie Hintergrundinformation I zur Verfügung. Das Induktionsproblem besteht darin, von der unvollständigen Ausgangsinformation D,I auf die unbekannte Konformation des Moleküls, $\mathbf{x}$, zu schließen. Der Hypothesenraum des Induktionsproblems wird durch einen vollständigen Satz alternativer Hypothesen aufgespannt. Die Konformation des Moleküls ist eindeutig durch Angabe seiner kartesischen Koordinaten $\mathbf{x}$ definiert; $\mathbf{x}$ fungiert daher als kontinuierlicher Hypothesenparameter. Die zu bewertenden Hypothesen haben somit die Form

$$\mathbf{x} = \text{„Die Konformation des Moleküls ist } \mathbf{x}.\text{“}$$

Das Induktionsproblem wird gelöst, indem jeder Hypothese, d.h. jeder Konformation, eine Wahrscheinlichkeit zugeordnet wird. Die Lösung soll dabei objektiv in dem Sinne sein, daß die Wahrscheinlichkeiten ausschließlich

von der verfügbaren Ausgangsinformation bestimmt werden. Die Bewertung der Hypothesen erfolgt also anhand der bedingten Wahrscheinlichkeitsdichte $p(\mathbf{x}|D,I)$. $p(\mathbf{x}|D,I)$ drückt unser gesamtes Wissen über die möglichen Konformationen des Moleküls aus, welches aus der gegebenen Ausgangsinformation prinzipiell ableitbar ist und repräsentiert somit die vollständige Lösung des Strukturbestimmungsproblems.

2.2.1 Die *Likelihood*-Funktion der Daten

Die Berücksichtigung der experimentellen Daten erfolgt durch die Angabe einer *Likelihood*-Funktion $p(D|\mathbf{x},I)$. Die *Likelihood*-Funktion entspricht der Verlustfunktion für die Daten in konventionellen Strukturbestimmungsmethoden (vgl. Kap. 1.2.1). Sie bewertet, wie wahrscheinlich die experimentellen Daten D beobachten werden, wenn die Konformation des Moleküls $\mathbf{x}$ ist. Die *Likelihood*-Funktion wird stets aus zwei Objekten gebildet: Einem Vorwärtsmodell und einer Fehlerverteilung. Das Vorwärtsmodell berechnet die theoretischen Werte der Meßgrößen aus den Strukturkoordinaten. Die Fehlerverteilung beschreibt die Diskrepanz zwischen observierten und berechneten Daten.
Im allgemeinen greift das Vorwärtsmodell für die Berechnung einer Meßgröße auf zusätzliche Hilfsgrößen $\alpha = \{\alpha_1, \ldots, \alpha_H\}$ zurück. Ein Beispiel ist der NOE-Skalenfaktor γ in der ISPA. Unbekannte Größen werden als *Nuisance*-Parameter aufgefaßt und über den Hypothesenparameter α in der *Likelihood*-Funktion berücksichtigt. Die *Likelihood*-Funktion besitzt somit die allgemeine Form $p(D|\mathbf{x},\alpha,I)$. I faßt alle Annahmen zusammen, die für die Interpretation der Daten vonnöten sind. Dazu zählt insbesondere der analytische Ausdruck des Vorwärtsmodells sowie die Form der Fehlerverteilung.

2.2.2 Die *a-priori*-Verteilung für die Struktur

Hintergrundwissen über Werte von Hilfsgrößen und über mögliche Konformation eines Moleküls, welches auch ohne Kenntnis der experimentellen Daten vorhanden ist, wird durch die *a-priori*-Verbundverteilung $p(\mathbf{x},\alpha|I)$ repräsentiert. Strukturelles *a-priori*-Wissen betrifft die Primärstruktur des Moleküls,

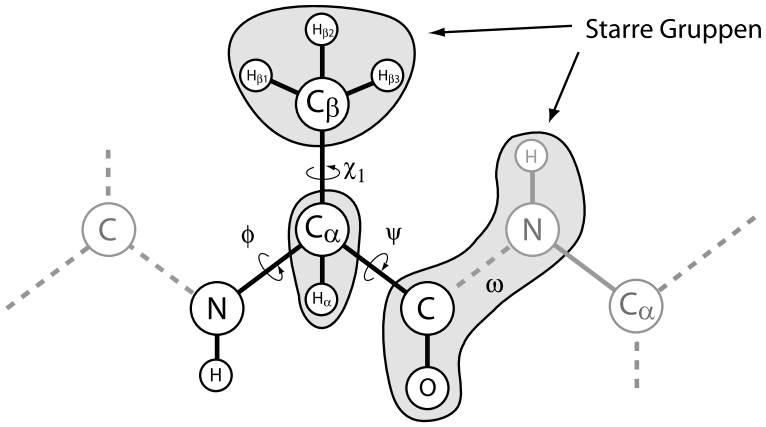

Abbildung 2.1: Parametrisierung der Polypeptidkette am Beispiel der Aminosäure Alanin. Atome, die sich in einer räumlich konstanten Anordnung zueinander befinden, werden zu starren Gruppen zusammengefaßt, die über drehbare Bindungen miteinander verbunden sind. Die Parametrisierung erfolgt über Dihedralwinkel. Dargestellt sind die Hauptketten-Dihedralwinkel ϕ, ψ und ω sowie der Seitenketten-Dihedralwinkel χ_1.

aus welcher der atome Aufbau des Systems abgeleitet werden kann. Zusätzliches *a-priori*-Wissen umfaßt kovalente und nicht-kovalente Wechselwirkungen zwischen den Atomen sowie Wechselwirkungen des Moleküls mit seiner Umgebung. Physikalische Interaktionen werden durch eine physikalische Energiefunktion beschrieben.

Parametrisierung der Polypeptidkette

Die in Proteinsystemen beobachtete Varianz von Bindungslängen und Bindungswinkeln ist gering und kann in guter Näherung als konstant angenommen werden. Dies erlaubt die Zerlegung der Polypeptidkette in starre Gruppen, welche über drehbare Bindungen miteinander verbunden sind (vgl. Abb. 2.1). Jede dieser Gruppen wird aus Atomen gebildet, die sich in einer räumlich konstanten Anordnung zueinander befinden. Die Einführung von starren Gruppen hat mehrere Vorteile: Zum einen kann die Polypeptidkette vollständig durch Dihedralwinkel parametrisiert werden, wodurch die

Zahl der Freiheitsgrade um eine Größenordnung reduziert wird. Zum anderen müssen in der Energiefunktion weder Potentiale für Bindungslängen und Bindungswinkel noch für planare und chirale Gruppen eingeführt werden. Dies reduziert den für die Auswertung der Energiefunktion benötigten Aufwand und wirkt sich zudem günstig auf die simulatorischen Eigenschaften des Systems aus. Die kartesischen Koordinaten aller Atome sind eindeutig durch einen Satz von Dihedralwinkeln und die Definition der starren Gruppen bestimmt. Die Parameter für die kovalente Geometrie der starren Gruppen wurden der Definition des ECEPP/2 Kraftfelds [40, 41] entnommen. Alle Dihedralwinkel um die Peptidbindungen, $\{\omega_i\}$ (vgl. Abb. 2.1), wurden auf 180° fixiert.

Nichtkovalente Wechselwirkungen

Für die Beschreibung nichtkovalenter Wechselwirkungen zwischen den Atomen wird ein rein repulsiver van der Waals-Term berücksichtigt, wie er im PROLSQ Kraftfeld [42] angegeben ist:

$$E_{ij}^{\mathrm{vdW}} = \frac{k}{2}\left(\|\mathbf{x}_i - \mathbf{x}_j\| - d_0\right)^4 \quad \text{für } \|\mathbf{x}_i - \mathbf{x}_j\| \leq d_0; \quad 0 \text{ sonst.} \tag{2.8}$$

i, j bezeichnet den Index des wechselwirkenden Atompaars und $\mathbf{x}_i$ bzw. $\mathbf{x}_j$ die kartesischen Koordinaten der beiden Atome. Die Parameter k und d_0 bezeichnen die Kraftkonstante bzw. den Mindestabstand der Wechselwirkung. Die Werte für d_0 sind von den Atomtypen der Wechselwirkungspartner abhängig und wurden aus den Atomradien berechnet, wie sie im PROLSQ Kraftfeld angegeben sind. Die Kraftkonstante ist atomtypunabhängig und beträgt $50\,\mathrm{kcal\,mol^{-1}\AA^{-4}}$. Elektrostatische Wechselwirkungen und Lösungsmitteleffekte werden in der gegenwärtigen Implementierung aus Effizienzgründen vernachlässigt. Damit ergibt sich die physikalische Energiefunktion zu

$$E(\mathbf{x}) = \frac{1}{2}\sum_{ij} E_{ij}^{\mathrm{vdW}}(\mathbf{x}_i, \mathbf{x}_j), \tag{2.9}$$

wobei die Summation über alle Atompaare läuft. Ferner wird angenommen, daß sich das System auf der Temperatur T befindet. Die Verteilung, wel-

che dieses Wissen ohne Zusatzannahmen repräsentiert, ist die Boltzmann-Verteilung [43]:

$$p(\mathbf{x}|I) = Z^{-1}(\beta)\, e^{-\beta E(\mathbf{x})}. \tag{2.10}$$

$\beta = 1/k_B T$, k_B bezeichnet die Boltzmann-Konstante und $Z(\beta)$ die kanonische Zustandssumme. Das Hintergrundwissen I faßt alle für die Herleitung der *a-priori*-Verteilung getroffenen Annahmen zusammen:

> $I =$ "Das System befindet sich auf der Temperatur T und besitzt konstante Bindungslängen, Bindungswinkel sowie konstante Dihedralwinkel für alle Peptidbindungen; als physikalische Wechselwirkung wird ein repulsiver van der Waals-Term berücksichtigt. Die Kraftfelder ECEPP/2 und PROLSQ liefern für die Beschreibung der kovalenten Geometrie und der physikalischen Wechselwirkungen in Proteinsystemen geeignete Parameterwerte."

Das kanonische Ensemble für die Energiefunktion $E(\mathbf{x})$ repräsentiert somit das Wissen über die Positionen der einzelnen Atome der Struktur welches sich bei Abwesenheit von experimentellen Daten rein aus den in I zusammengefaßten Annahmen ergibt.

2.2.3 Die *a-posteriori*-Verbundverteilung

Bei bekannter *Likelihood*-Funktion und *a-priori*-Verbundverteilung folgt die *a-posteriori*-Verbundverteilung für die Koordinaten der Struktur und alle Hilfsgrößen aus dem Bayes'schen Satz:

$$\begin{aligned} p(\mathbf{x}, \alpha|D, I) &\propto p(\mathbf{x}, \alpha|I) p(D|\mathbf{x}, \alpha, I) \\ &\propto e^{-\beta E(\mathbf{x})}\, p(\alpha|I) p(D|\mathbf{x}, \alpha, I), \end{aligned} \tag{2.11}$$

wobei die logische Unabhängigkeit von Koordinaten und Hilfsgrößen angenommen wurde. Strukturbestimmung verfolgt die Fragestellung, welche Implikationen sich aus der Ausgangsinformation für die Koordinaten der Struktur ergeben. Die Werte von Hilfsgrößen sind in diesem Zusammenhang nicht

von Interesse. Die marginale *a-posteriori*-Verteilung für die Koordinaten der Struktur, die *Strukturverteilung*, folgt durch Integration über alle *Nuisance*-Parameter:

$$\begin{aligned} p(\mathbf{x}|D,I) &= \int d\alpha\, p(\mathbf{x},\alpha|D,I) \\ &\propto e^{-\beta E(\mathbf{x})} \int d\alpha\, p(D|\mathbf{x},\alpha,I)\, p(\alpha|I). \end{aligned} \tag{2.12}$$

Eine Inversion des Vorwärtsmodells oder die empirische Bestimmung von Hilfsgößen durch Heuristiken ist in der induktiven Strukturbestimmung hinfällig: Gleichung (2.12) gestattet die direkte Berechnung der relativen Wahrscheinlichkeit einer gegebenen Konformation. Hilfsgrößen werden formal aus den Daten geschätzt und durch Marginalisierung eliminiert. Auf diese Weise wird das Unwissen über die tatsächlichen Werte der Hilfsgrößen vollständig in der Lösung berücksichtigt. Die Verteilung der Wahrscheinlichkeiten repräsentiert das Unwissen über die Positionen der Atome der Struktur im Raum: Gestattet die Ausgangsinformation einen hinreichend eindeutigen Schluß, so konzentriert sich die Wahrscheinlichkeitsmasse in einer oder wenigen Regionen des Konformationsraums. Unvollständige Daten lassen hingegen nur ungenaue Aussagen zu, was sich in über weite Bereiche des Konformationsraums verteilten Wahrscheinlichkeiten äußerte. Im Grenzfall vollständiger Information existiert genau eine Konformation $\hat{\mathbf{x}}$ mit $p(\hat{\mathbf{x}}|D,I) \neq 0$, d.h. das Induktionsproblem hat eine eindeutige Lösung, die durch Inversion bestimmt werden kann. Die induktive Strukturbestimmung enthält den konventionellen Zugang somit als Spezialfall.

Verallgemeinerung auf mehrere Datensätze

Die Berücksichtigung mehrerer Datensätze erfolgt durch Verallgmeinerung von Gleichung (2.12). Jeder der K Datensätze $D^{(k)}$ werde durch eine individuelle *Likelihood*-Funktion $p(D^{(k)}|\mathbf{x},\alpha^{(k)},I)$ mit einem separatem Satz von Hilfsparametern $\alpha^{(k)}$ beschrieben. Im Falle unabhängiger Datensätze faktorisieren die Gesamt-*Likelihood* sowie die *a-priori*-Verteilungen für die Sätze

von Hilfsparametern und die *a-posteriori*-Verbundverteilung besitzt die allgemeine Form

$$p(\mathbf{x}, \{\alpha^{(k)}\}|\{D^{(k)}\}, I) \propto e^{-\beta E(\mathbf{x})} \prod_{k=1}^{K} p(D^{(k)}|\mathbf{x}, \alpha^{(k)}, I)\, p(\alpha^{(k)}|I). \qquad (2.13)$$

Die Strukturverteilung folgt wiederum durch Marginalisierung aller *Nuisance*-Parameter:

$$p(\mathbf{x}|\{D^{(k)}\}, I) \propto e^{-\beta E(\mathbf{x})} \prod_{k=1}^{K} \int d\alpha^{(k)}\, p(D^{(k)}|\mathbf{x}, \alpha^{(k)}, I)\, p(\alpha^{(k)}|I). \qquad (2.14)$$

2.3 Simulation der Strukturverteilung

Die Lösung eines induktiven Strukturbestimmungsproblems bedeutet die Berechnung der korrespondierenden Strukturverteilung. Techniken für die Berechnung von Wahrscheinlichkeitsverteilungen unterscheiden sich dabei grundlegend von Optimierungsstrategien, wie sie in der konventionellen NMR-Strukturrechnung verwendet werden: Während Optimierungsalgorithmen den Konformationsraum auf Basis eines Optimalitätskriteriums gezielt nach Strukturen minimaler Energie durchsuchen können, ist die Berechnung der Strukturverteilung von größerer Komplexität: Die Strukturverteilung ordnet *jedem* Punkt im Konformationsraum eine Wahrscheinlichkeit zu – alle Konformationen einer Struktur sind daher Teil der Lösung. Eine vorurteilsfreie Berechnungsstrategie muß somit in der Lage sein, den gesamten Konformationsraum geeignet abbilden zu können. Optimierungsalgorithmen sind für die Berechnung von Wahrscheinlichkeitsverteilungen aus diesem Grunde ungeeignet.

Eine direkte Berechnung der Wahrscheinlichkeiten aller Konformationen ist aufgrund der Größe des Konformationsraums unmöglich. Für eine effiziente Berechnung der Strukturverteilung sind daher Verfahren von Interesse, welche den Konformationsraum gezielt nach „relevanten“ Regionen, also Bereichen signifikanter Wahrscheinlichkeitsmasse absuchen und auf diese Weise ein repräsentatives Abbild der Strukturverteilung erzeugen. Zu diesen Verfahren zählt die Klasse der *Sampling*-Algorithmen, eine verallgemeinerte Form

von Zufallszahlengeneratoren. *Sampling*-Methoden erzeugen ein repräsentatives Abbild einer Wahrscheinlichkeitsverteilung, indem sie Stichproben $\varphi^{(i)}$ gemäß einer gegebenen Wahrscheinlichkeitsdichtefunktion $p(\varphi)$ generieren. Die Verteilung wird durch die Dichte der Proben angenähert; sie wird in diesem Sinne also nicht berechnet, sondern *simuliert*.

Die Effizienz eines *Sampling*-Algorithmus wird von den Eigenschaften der Zielverteilung bestimmt, die im Falle von Proteinsystemen vier Charakteristika aufweist: (1) eine hohe Dimensionalität; (2) Multimodalität; (3) die Parameter sind stark korreliert; (4) Moden sind durch Bereiche geringer Wahrscheinlichkeit voneinander getrennt. Die effiziente Behandlung dieser Probleme erfolgt mit einer hierarchischen *Sampling*-Strategie [34, 44], welche als *Markov-Ketten-Monte-Carlo* (MCMC) Methode konzipiert ist. In der Strategie kommen drei MCMC-Methoden zum Einsatz: Gibbs-*Sampling* [45], Hybrid-Monte-Carlo [46] und Replika-Austausch-Monte-Carlo [47]. Die Strategie ist zugleich die erste Anwendung von Replika-Austausch-Monte-Carlo in der NMR-Strukturrechnung. Abschnitt 2.3.1 bespricht die Grundlagen von MCMC-Methoden. In den Abschnitten 2.3.2 bis 2.3.4 werden die genannten MCMC-Methoden vorgestellt, welche in Abschnitt 2.3.5 in Form eines verallgemeinerten Replika-Algorithmus kombiniert werden.

2.3.1 Markov-Ketten-Monte-Carlo-Methoden

Monte-Carlo (MC) Methoden simulieren eine Wahrscheinlichkeitsverteilung, indem mittels einer stochastischen Vorschrift aus bisher gezogenen Proben $\varphi^{(i-1)}, \ldots, \varphi^{(0)}$ eine neue Stichprobe $\varphi^{(i)}$ erzeugt wird. MCMC-Methoden realisieren den Ziehungsprozess über die Konstruktion einer Markov-Kette, d.h. einer Folge von Stichproben, bei der die Wahrscheinlichkeit für die Ziehung einer neuen Probe nur von der unmittelbar zuvor erzeugten Probe abhängt. Mathematisch läßt sich zeigen, daß *homogene* und *ergodische* Markov-Ketten konstruiert werden können, welche eine *invariante* Verteilung besitzen und unabhängig von der Anfangsbedingung gegen diese konvergieren. Die Simulation einer Verteilung p durch MCMC erfolgt über die Konstruktion einer

homogenen und ergodischen Markov-Kette, deren invariante Verteilung p ist. Homogene Markov-Ketten sind vollständig charakterisiert durch die Angabe einer Verteilung $p^{(0)}$ für den Wert des ersten Zustands $\varphi^{(0)}$ der Markov-Kette und der *Übergangswahrscheinlichkeit* $T(\varphi', \varphi)$. Bei homogenen Markov-Ketten ist die Übergangswahrscheinlichkeit von der Position i innerhalb der Kette unabhängig. Falls die letzte Konfiguration gemäß $p^{(i-1)}$ erzeugt wurde, ist der neue Zustand der Kette verteilt nach

$$p^{(i)}(\varphi') = \int d\varphi\, T(\varphi', \varphi) p^{(i-1)}(\varphi). \tag{2.15}$$

Die invariante Verteilung einer Markov-Kette $\pi(\varphi)$ ist diejenige Verteilung, die bei einem Markov-Übergang reproduziert wird, d.h. falls gilt

$$\pi(\varphi') = \int d\varphi\, T(\varphi', \varphi) \pi(\varphi). \tag{2.16}$$

Wurden also Zustände $\varphi^{(i)}, \varphi^{(i-1)}, \ldots$ gemäß der invarianten Verteilung erzeugt, so gehorchen die nachfolgend erzeugten Proben ebenfalls der invarianten Verteilung. Die Ziehung von Stichproben von der Zielverteilung p erfolgt nun durch die Realisierung einer Markov-Kette, deren Übergangswahrscheinlichkeit p invariant läßt. Gängige MCMC-Verfahren nutzen *reversible* Markov-Ketten, welche die stärkere Bedingung des *detaillierten Gleichgewichts* erfüllen. Das detaillierte Gleichgewicht ist ein Spezialfall der allgemeinen Invarianzbedingung aus Gleichung 2.16 und besagt, daß die Wahrscheinlichkeit für den Übergang $\varphi \rightleftharpoons \varphi'$ richtungsunabhängig sein muß, sofern der Ausgangszustand von der Zielverteilung gezogen wurde, also

$$T(\varphi', \varphi)\pi(\varphi) = T(\varphi, \varphi')\pi(\varphi'). \tag{2.17}$$

Um genau eine invariante Verteilung zu besitzen, muß eine Markov-Kette zudem ergodisch sein: Unabhängig von der Verteilung der Anfangskonfiguration $p^{(0)}$ müssen im Grenzfall $i \to \infty$ die Verteilungen $p^{(i)}$ gegen die invariante Verteilung konvergieren. Ein mathematischer Beweis zeigt, daß die Eigenschaft der Ergodizität bei einer großen Klasse von Markov-Ketten gegeben ist [48].

Monte-Carlo-Integration

Bei der Berechnung von Erwartungswerten, marginalisierten Dichten und anderen statistischen Größen treten Integrale der Form

$$E_p[f] = \int d\varphi f(\varphi) p(\varphi) \tag{2.18}$$

auf. $E_p[f]$ bezeichnet den Erwartungswert der Funktion f bezüglich der Verteilung p. Die numerische Berechnung von Erwartungswerten durch Monte-Carlo-Integration erfolgt auf Basis von Stichproben $\varphi^{(i)} \sim p(\varphi)$, indem das Integral in Gleichung 2.18 durch eine Summe angenähert wird:

$$E_p[f] \approx \frac{1}{N} \sum_{i=1}^{N} f(\varphi^{(i)}). \tag{2.19}$$

Bei einer endlichen Zahl von Stichproben ist der *MC-Schätzer* in Gleichung (2.19) unsicherheitsbehaftet und weist eine endliche Varianz auf, die von der Zahl der Proben abhängt:

$$\mathrm{Var}[E_p] = \frac{1}{N} \mathrm{Var}_p[f]. \tag{2.20}$$

Var_p bezeichnet den MC-Schätzer für die Varianz von f bezüglich p und $\mathrm{Var}[E_p]$ die Varianz, d.h. die Unsicherheitsbehaftung, des MC-Schätzers in Gleichung (2.19).

2.3.2 Der Gibbs-Algorithmus

Der Gibbs-Algorithmus [45] definiert eine allgemeine Vorschrift, nach welcher Proben $\varphi^{(i)}$ von einer Wahrscheinlichkeitsverteilung $p(\varphi)$ für die d-dimensionale Zufallsvariable $\varphi = \{\varphi_1, \ldots, \varphi_d\}$ durch die Simulation der bedingten Verteilungen für die einzelnen Komponenten $p(\varphi_k | \{\varphi_{j \neq k}\})$ generiert werden können. Für die Erzeugung einer neuen Probe werden alle Komponenten $\{\varphi_k\}$ sukzessive von ihren bedingten Verteilungen gezogen, wobei die zuvor gezogenenen Komponenten auf der bedingenden Seite der jeweiligen Verteilung eingesetzt werden. Der Gibbs-Algorithmus ist daher besonders für die Simulation von Wahrscheinlichkeitsverteilungen geeignet, deren bedingte

Dichten direkt mittels Zufallszahlengeneratoren simuliert werden können. Die Erzeugung einer neuen Stichprobe $\varphi^{(i+1)}$ bei gegebener Stichprobe $\varphi^{(i)}$ erfolgt nach dem folgenden Schema:

$$\begin{aligned}
\varphi_1^{(i+1)} &\sim p(\varphi_1|\varphi_2^{(i)},\ldots,\varphi_d^{(i)}), \qquad (2.21)\\
&\ldots\\
\varphi_k^{(i+1)} &\sim p(\varphi_k|\varphi_1^{(i+1)},\ldots,\varphi_{k-1}^{(i+1)},\varphi_{k+1}^{(i)},\ldots,\varphi_d^{(i)}),\\
&\ldots\\
\varphi_d^{(i+1)} &\sim p(\varphi_d|\varphi_1^{(i+1)},\ldots,\varphi_{d-1}^{(i+1)}).
\end{aligned}$$

„$\sim$“ ist eine Kurznotation für „gezogen von“. Durch Iteration der obigen Vorschrift wird eine Markov-Kette konstruiert, deren invariante Verteilung $p(\varphi)$ ist, d.h. deren Zusände $\varphi^{(i)}$ gemäß $p(\varphi)$ verteilt sind.

2.3.3 Hybrid-Monte-Carlo

Hybrid-Monte-Carlo (HMC) [46] wurde für die Simulation hochdimensionaler Verteilungen stark korrelierter Variablen entwickelt und kombiniert die Techniken von Metropolis-Monte-Carlo [49] und Molekulardynamik [50].
Die generelle Vorgehensweise in HMC ist analog zu Metropolis-Monte-Carlo: Um eine Stichprobe („Zustand“) von der Verteilung $p(\varphi)$ für die d-dimensionale Zufallsvariable φ zu ziehen, wird auf Basis des aktuellen Zustands ein Kandidat für den nächsten Zustand erzeugt. Der Kandidat wird anschließend gemäß dem Metropolis-Kriterium akzeptiert oder verworfen. Anders als bei Metropolis-MC wird der Kandidat jedoch nicht durch eine stochastische Störung des aktuellen Zustands erzeugt, sondern anhand einer Dynamiktrajektorie bestimmt: Korrelierte Variablen ändern sich in konzertierter Weise, wodurch der Zustandsraum effizient durchlaufen und nichtlokale Kanditaten erzeugt werden.
Um das Konzept einer Dynamiktrajektorie für die Simulation einer Wahrscheinlichkeitsverteilung nutzbar zu machen, wird in Analogie zu einem phy-

sikalischen System eine potentielle „Energie“ definiert:

$$U(\varphi) = -\log p(\varphi). \tag{2.22}$$

Die Energiefunktion $U(\varphi)$ dient ausschließlich algorithmischen Zwecken und besitzt keinerlei physikalische Realität. Eine Erweiterung des Parameterraums um kanonisch konjugierte Impulse $\pi = \{\pi_1, \ldots, \pi_d\}$ gestattet die Definition eines dynamischen Systems mit der Hamilton Funktion

$$H(\varphi, \pi) = \frac{\pi^2}{2} + U(\varphi). \tag{2.23}$$

Hybrid-Monte-Carlo realisiert eine Markov-Kette im Phasenraum und erzeugt auf diese Weise Zustände $(\varphi, \pi)^{(i)}$ von der kanonischen Verteilung,

$$p(\varphi, \pi) \propto \exp\{-H(\varphi, \pi)\} = \exp\{-\pi^2/2\}\, p(\varphi). \tag{2.24}$$

Jeder Markov-Übergang besteht aus einem dynamischen und einem stochastischen Anteil. Bei gegebenem Zustand $(\varphi, \pi)^{(i)}$ wird der nächste Zustand gemäß folgender Vorschrift erzeugt:

1. Ziehe Anfangsimpulse π_0 der Dynamiktrajektorie von einer Normalverteilung mit Erwartungswert 0 und Varianz 1;

2. Berechne Dynamiktrajektorie der Länge τ mit $q = (\varphi^{(i)}, \pi_0)$ als Startpunkt;

3. Betrachte den Endpunkt der Trajektorie $q^* = (\varphi(\tau), \pi(\tau))$ als Kandidaten für den nächsten Zustand. Akzeptiere den Kandidaten gemäß dem Metropolis-Kriterium mit der Wahrscheinlichkeit $\min(1, \exp(-\Delta H))$ und setzte $(\varphi, \pi)^{(i+1)} = q^*$; anderenfalls setze $(\varphi, \pi)^{(i+1)} = (\varphi, \pi)^{(i)}$.

$\Delta H = H(q^*) - H(q)$ bezeichnet die Differenz der Gesamtenergie von End- und Anfangspunkt der Trajektorie. Die Simulation der Markov-Kette erfolgt durch die wiederholte Anwendung der Schritte (1) bis (3). Da die kanonische Verteilung faktorisiert, ergeben sich die gesuchten Stichproben $\{\varphi^{(1)}, \varphi^{(2)}, \ldots\}$ durch Verwerfen der jeweiligen Impulse. Die Invarianz der Gesamtenergie H

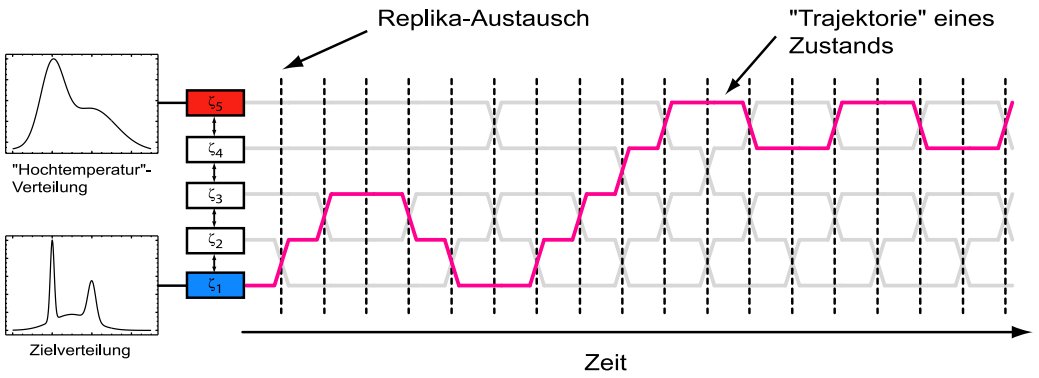

Abbildung 2.2: Illustration von Replika-Austausch-Monte-Carlo. Die Zielverteilung (blau) wird anhand einer Kette von Hilfsverteilungen („Kopien“) sukzessive in eine effizient zu simulierende „Hochtemperatur“-Verteilung (rot) deformiert. Alle Kopien werden parallel simuliert; in festen Abständen (gestrichelte vertikale Linien) werden Zustände benachbarter Kopien nach dem Metropolis-Kriterium ausgetauscht und propagieren dadurch stochastisch durch die Kette (hellrot).

unter Hamilton'scher Dynamik bedeutet eine theoretische Akzeptanzrate von 1, die lediglich durch Ungenauigkeiten bei der numerischen Berechnung der Trajektorie reduziert wird und somit von der Länge der Dynamiktrajektorie abhängt.

2.3.4 Replika-Austausch-Monte-Carlo

Wahrscheinlichkeitsverteilungen besitzen häufig eine Vielzahl lokaler Moden, die durch Bereiche geringer Wahrscheinlichkeit („Energiebarrieren“) voneinander getrennt sind. Damit die Zustände einer Markov-Kette von ihrer invarianten Verteilung gezogen werden, müssen alle Moden mit den korrekten Populationsgewichten besucht werden. Methoden wie Hybrid-Monte-Carlo werden typischerweise von einer der Moden „eingefangen“ und sind für die Simulation multimodaler Wahrscheinlichkeitsverteilungen aus diesem Grunde ungeeignet.

Die Technik des Replika-Austausch-Monte-Carlo [47] umgeht diese Probleme, indem mehrere, nicht miteinander wechselwirkende Kopien der Zielver-

teilung $p(\varphi)$ parallel simuliert werden. Jede Kopie wird dabei geeignet transformiert, so daß etwaige Moden der Zielverteilung nach der Transformation weniger stark pronounciert ausfallen (vgl. Abb. 2.2); im physikalischen Bild korrespondiert dies zu einer „Aufheizung" des Systems. Die Form der einzelnen Verteilungen wird durch einen temperaturartigen Parameter ζ („*Replika-Parameter*") gesteuert. Die Familie transformierter Verteilungen $f(\varphi;\zeta)$ wird so gewählt, daß die „Hochtemperatur"-Verteilung effizient simuliert werden kann. Um die Hochtemperatur-Verteilung sukzessive in die Zielverteilung, d.h. in die untransformierte Verteilung $f(\varphi;\zeta_1) \equiv p(\varphi)$ zu überführen, werden K Kopien in Form einer Kette organisiert. Replika-Austausch-Monte-Carlo realisiert eine Markov-Kette für die Verbundverteilung,

$$p_{\text{replika}}(\varphi_1,\ldots,\varphi_K) \propto \prod_{i=1}^{K} f(\varphi_i;\zeta_i), \tag{2.25}$$

welche aufgrund der Unabhängigkeit aller Kopien faktorisiert. φ_i und ζ_i bezeichnet den Variablensatz bzw. den Replika-Parameter von Kopie i. Für einen Markov-Übergang wird jede Kopie unabhängig von den übrigen Kopien simuliert. Für die Simulation der einzelnen Kopien können beliebige MCMC-Verfahren (z.B. Hybrid-Monte-Carlo) verwendet werden. Nach einer festen Anzahl von Schritten werden die jeweils zuletzt generierten Stichproben $\hat{\varphi}_i$ und $\hat{\varphi}_j$ benachbarter Kopien i bzw. j gemäß dem Metropoliskriterium mit der Wahrscheinlichkeit

$$P_{\text{acc}} = \min\left\{1, \frac{f(\hat{\varphi}_j;\zeta_i) f(\hat{\varphi}_i;\zeta_j)}{f(\hat{\varphi}_i;\zeta_i) f(\hat{\varphi}_j;\zeta_j)}\right\} \tag{2.26}$$

ausgetauscht. Normierungskonstanten der transformierten Verteilungen müssen nicht bekannt sein, da sich konstante Faktoren in Gleichung (2.26) gegeneinander wegheben. Durch den stochastischen Austausch von Zuständen werden nichtlokale Sprünge innerhalb der individuellen Markov-Ketten der einzelnen Kopien erzeugt. Bei geeigneter Wahl der Hochtemperaturverteilung (idealerweise einer Gleichverteilung) ist die Ergodizität der Gesamtsimulation und damit die Ergodizität der Markov-Kette für die Zielverteilung garantiert.

2.3.5 Ein verallgemeinerter Replika-Algorithmus

Nuisance-Parameter lassen sich häufig nicht analytisch marginalisieren. Im allgemeinen Fall ist die Strukturverteilung daher nicht in geschlossener Form, sondern in Integraldarstellung gegeben. Die Berechnung von Marginalisierungsintegralen auf simulatorischem Wege erfolgt durch Simulation der zugrundeliegenden *a-posteriori*-Verbundverteilung [51]. Um Konformationen $\mathbf{x}^{(i)}$ von der Strukturverteilung in Gleichung (2.14) zu ziehen, wird demzufolge ein Verfahren benötigt, mit welchem *a-posteriori-Stichproben* $(\mathbf{x}, \{\alpha^{(k)}\})^{(i)}$ von der *a-posteriori*-Verbundverteilung aus Gleichung (2.13) gezogen werden können. Die Simulation der *a-posteriori*-Verbundverteilung hat darüber hinaus den Vorteil, daß außer der Strukturverteilung, auch die marginalen *a-posteriori*-Verteilungen der übrigen Hypothesenparameter bestimmt werden können.

Zerlegung des Simulationsproblems

Das Simulationsproblem wird mit Hilfe des Gibbs-Algorithmus in zwei Teilprobleme zerlegt: Für die Erzeugung einer *a-posteriori*-Stichprobe $(\mathbf{x}, \{\alpha^{(k)}\})^{(i)}$ werden zuerst die *Nuisance*-Parameter $\alpha^{(k)}$ der einzelnen Datenmodelle von ihren bedingten Verteilungen gezogen, gefolgt von der Generierung eines neuen Koordinatensatzes $\mathbf{x}$. Das allgemeine Gibbs-Schema besitzt somit folgende Form:

$$\begin{aligned}
[\alpha^{(1)}]^{(i+1)} &\sim p(\alpha^{(1)}|\mathbf{x}^{(i)}, D^{(1)}, I), \\
&\dots \\
[\alpha^{(K)}]^{(i+1)} &\sim p(\alpha^{(K)}|\mathbf{x}^{(i)}, D^{(K)}, I), \\
\mathbf{x}^{(i+1)} &\sim p(\mathbf{x}|\{\alpha^{(k)}\}^{(i+1)}, \{D^{(k)}\}, I). \qquad (2.27)
\end{aligned}$$

Die Erzeugung einer neuen Konformation gemäß Gl. (2.27) ist aufwendig, was auf die genannten Eigenschaften der Strukturverteilung zurückzuführen ist.

Erzeugung konformationeller Stichproben

Hybrid-Monte-Carlo gestattet die effiziente Simulation hochdimensionaler und stark korrelierter Wahrscheinlichkeitsverteilungen und ist daher für die Erzeugung konformationeller Stichproben $\mathbf{x}^{(i)}$ geeignet. Die für die Berechnung der Dynamiktrajektorie erforderliche Energiefunktion besitzt gemäß Gleichung (2.13) die allgemeine Form

$$U(\theta) = \beta E(\mathbf{x}(\theta)) - \sum_{k=1}^{K} \log p(D^{(k)}|\mathbf{x}(\theta), \cdot). \tag{2.28}$$

Die Notation $\mathbf{x}(\theta)$ verdeutlicht, daß die kartesischen Koordinaten der Struktur intern durch einen Satz von Dihedralwinkeln $\theta = \{\theta_1, \ldots, \theta_M\}$ parametrisiert werden (vgl. Kap. 2.2.2). Die Dynamiktrajektorie wird durch die numerische Integration der Hamilton'schen Bewegungsgleichungen [52] direkt im Dihedralwinkelraum berechnet:

$$\begin{aligned} \frac{d\theta_i}{dt} &= +\frac{\partial H}{\partial \pi_i} = \pi_i, \\ \frac{d\pi_i}{dt} &= -\frac{\partial H}{\partial \theta_i} = -\frac{\partial U}{\partial \theta_i}, \end{aligned}$$

wobei π_i den zu θ_i kanonisch konjungierten Impuls bezeichnet. Eine Konformation $\mathbf{x}$ wird auf indirekte Weise gezogen, indem ein neuer Satz von Dihedralwinkeln gemäß $\theta \sim p(\mathbf{x}(\theta)|\cdot)$ generiert wird, welcher anschließend in kartesische Koordinaten überführt wird. Die Diskretisierung der Bewegungsgleichungen erfolgt durch das *Leapfrog*-Schema [53]. Die *Leapfrog*-Diskretisierung ist von erster Ordnung und zeitreversibel. Die Zeitreversibilität des Integrators ist eine notwendige Voraussetztung, damit die durch Hybrid-Monte-Carlo konstruierte Markov-Kette die Bedingung des detaillierten Gleichgewichts erfüllt [48]. Die Lösung der Hamilton'schen Bewegungsgleichungen im Dihedralwinkelraum führt zu einer unphysikalischen Trajektorie im kartesischen Raum. Die Berechnung der Dynamiktrajektorie dient jedoch ausschließlich algorithmischen Zwecken. Ihr unphysikalischer Charakter stellt daher kein Problem dar, solange die Erhaltung der Gesamtenergie unabhängig von der verwendeten Parametrisierung ist.

Einbettung des Gibbs-Algorithmus in ein Replika-Schema

Die Strukturverteilung besitzt eine Vielzahl von Moden, wodurch sich die Effizienz des Gibbs/HMC-Schemas signifikant reduzieren kann. Um die Konvergenzgeschwindigkeit der Methode zu erhöhen, wurde der bisher besprochene Gibbs-Algorithmus in ein Replika-Austausch-Monte-Carlo-Schema eingebettet. Multimodalitäten der *a-posteriori*-Verteilung gründen sich sowohl auf Eigenschaften des Boltzmann-Ensembles, als auch auf Eigenschaften der *Likelihood*-Funktion: Van der Waals-Wechselwirkungen schränken die Beweglichkeit der Atome aufgrund von Energiebarrieren stark ein, die *Likelihood*-Funktion führt zu einer Kompaktierung der Strukturen, wodurch die Beweglichkeit des Systems weiter verringert wird. Um diese Effekte wirksam zu reduzieren, werden die *Likelihood*-Funktion und das Boltzmann-Ensemble unabhängig voneinander transformiert: Zwei Replika-Parameter $\zeta = (\lambda, q)$ parametrisieren die Familie transformierter *a-posteriori*-Verteilungen:

$$f(\mathbf{x}, \{\alpha^{(k)}\}; \lambda, q) = e^{-\beta E(\mathbf{x};q)} \prod_{k=1}^{K} \left[p(D^{(k)}|\mathbf{x}, \alpha^{(k)}, I)\right]^{\lambda} p(\alpha^{(k)}|I). \qquad (2.29)$$

Die beiden ersten Terme der rechten Seite von Gleichung (2.29) bezeichnen das transformierte Boltzmann-Ensemble bzw. die transformierte *Likelihood*-Funktion. Die *a-priori*-Verteilungen der *Nuisance*-Parameter, $p(\alpha^{(k)}|I)$, sind von der Transformation nicht betroffen.
Die Transformation der *Likelihood*-Funktion bewirkt, daß die Daten für $\lambda = 0$ vollständig vernachlässigt werden; $\lambda = 1$ entspricht dem untransformierten Fall. Das Boltzmann-Ensemble wird durch eine nichtlineare Transformation der physikalischen Energiefunktion $E(\mathbf{x})$ in ein Tsallis-Ensemble [54, 55] überführt:

$$E(\mathbf{x}; q) = \frac{q}{\beta(q-1)} \log\left\{1 + \beta(q-1)(E(\mathbf{x}) - E_{\min})\right\}.$$

$E_{\min}$ bezeichnet die minimale erreichbare Energie; für die Energiefunktion in Gleichung (2.9) gilt $E_{\min} = 0$. Der Parameter q steuert die Stärke der Deformation: Das Tsallis-Ensemble ist für $q = 1$ identisch mit dem Boltzmann-

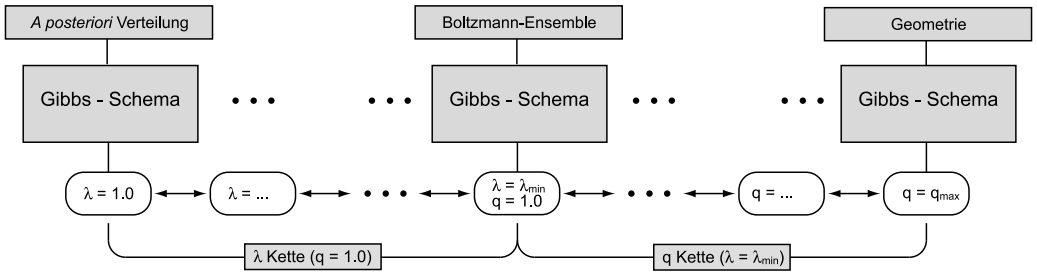

Abbildung 2.3: Standardkonfiguration der Replika-Kette. Die transformierten Hilfsverteilungen („Kopien“) sind in zwei Teilketten organisiert: In der λ-Kette (linker Teil) wird der Einfluß der Daten, in der q-Kette (rechter Teil) die Stärke der van der Waals-Wechselwirkung schrittweise reduziert, wodurch sich die einzelnen Systeme sukzessive freier bewegen können. Die untransformierte *a-posteriori*-Verteilung befindet sich ganz links ($\lambda = q = 1$).

Ensemble; für $q > 1$ wird die strukturelle *a-priori*-Verteilung sukzessive abgeflacht und geht für $E(\mathbf{x}; q) > E_{\min}$ im Grenzfall $q \to \infty$ in eine Gleichverteilung über. Die untransformierte *a-posteriori*-Verteilung entspricht $q = \lambda = 1$.

Standardkonfiguration

Abbildung 2.3 zeigt die Konfiguration der Replika-Kette, wie sie in allen Simulationen in dieser Arbeit verwendet wurde. Die Replika-Kette besteht aus zwei Teilen, der λ-Kette und der q-Kette. Die *a-posteriori*-Verteilung bildet den Anfang der Kette (ganz links). Die Verteilungen aller Kopien werden parallel und unabhängig voneinander mit Hilfe des Gibbs-Schemas simuliert. In der λ-Kette wird die Gewichtung der Daten durch schrittweises Absenken von λ von 1 auf $\lambda_{\min}$ langsam reduziert, so daß sich das System frei von kompaktierenden Einflüssen bewegen kann. Die in der letzten Kopie der λ-Kette erzeugten Konformationen bilden näherungsweise ein Boltzmann-Ensemble. In der q-Kette wird die Stärke der van der Waals-Wechselwirkung sukzessive reduziert, indem q graduell von 1 auf $q_{\max}$ erhöht wird. Auf diese Weise können sich Atome während der Dynamiktrajektrie des Hybrid-Monte-Carlo-Algorithmus gegenseitig durchdringen, womit die Effizient des Algorithmus

gesteigert wird. In der „Hochtemperatur"-Kopie (ganz rechts) sind Daten und physikalische Wechselwirkungen nahezu abgeschaltet. Bei geeigneter Wahl von ($\lambda_{\min}$, $q_{\max}$) ist die *a-posteriori*-Verteilung hinreichend flach, wodurch die Ergodizität der Markov-Kette garantiert ist.
Das Konvergenzverhalten der Markov-Kette wird von den einzelnen Akzeptanzwahrscheinlichkeiten und somit vom Überlapp der Verteilungen benachbarter Kopien bestimmt. Bei der Simulation einer Replika-Kette muß somit ein Kompromiß getroffen werden zwischen der Effizienz der Simulation (d.h. ihrer Konvergenzgeschwindigkeit) und dem dafür notwendigen numerischen Aufwand, d.h. der Zahl der Kopien.

2.4 Software

Die meisten Abbildungen wurden mit der Software-Bibliothek Biggles [56] und dem Programm MOLMOL [57] erstellt. Die Bewertung berechneter Konformationen hinsichtlich allgemeiner Qualitätsmerkmale erfolgte mit den Programmen PROCHECK [26], WHATIF [58] und PROSA [30].

ISD-Simulationspaket

Alle Strukturrechnungen in dieser Arbeit wurden mit dem ISD (*Inferential Structure Determination*) Simulationspaket durchgeführt. Das Softwarepaket stellt die vollständige Infrastruktur für die Simulation und Analyse eines induktiven Strukturbestimmungsproblems aus NMR-Daten zur Verfügung. Neben zahlreichen Datenmodellen für die Integration experimenteller Strukturinformation enthält das Paket eine parallelisierte Version des beschriebenen Replika-Austausch-Monte-Carlo-Algorithmus. Das ISD-Simulationspaket wurde in Kollaboration mit M. Habeck (Institut Pasteur, Paris) entwickelt. Eine detailliertere Beschreibung des Pakets habe ich in Anhang B zusammengestellt.

2.5 Testsysteme und Datensätze

Für die Strukturrechnungen standen simulierte und experimentelle NOESY-Datensätze für das Protein BPTI [59, 60] bzw. die Tudor Domäne des humanen SMN Proteins [61, 62] zur Verfügung.

BPTI

BPTI hat eine Länge von 58 Aminosäuren und bildet drei Disulfidbrücken aus (Reste 5-55, 14-38, 30-51). Die Faltung besteht aus einer N-terminalen 3_{10}-Helix (Reste 2-7), einem verdrillten, antiparallelen *β-hairpin* Motiv (Reste 18-24 und 29-35) sowie einer C-terminalen α-Helix (Reste 48-55).

Der verwendete Datensatz wurde aus einer 6.6ns langen Molekular Dynamik Trajektorie *in vacuo* generiert und besteht aus 1543 simulierten dipolaren Kreuzrelaxationsraten [23, 63]. Die Berechnung der MD Trajektorie erfolgte mit dem Programm X-PLOR [64] auf Basis der CHARMM Energiefunktion PARAM19 [65] mit impliziter Lösungsmittelbeschreibung. Für alle Wasserstoffpaare mit einem mittleren effektiven Abstand von weniger als 4.5 Å wurden Vektorautokorrelationsfunktionen berechnet. Konvergierte Korrelationsfunktionen wurden in spektrale Dichten überführt, aus welchen dipolare Kreuzkorrelationsraten berechnet wurden. Der Modelldatensatz enthält somit Dynamikbeiträge in realistischer Weise; Multispineffekte wurden bei der Berechnung der Kreuzrelaxationsraten nicht berücksichtigt (für weitere Details siehe [23]).
Aus der Trajektorie wurden die flexiblen Regionen von BPTI bestimmt und die mittlere Struktur berechnet [63]. Die flexiblen Regionen umfassen die Reste 9-10, 13-15 und 39-40. Ich verwende die mittlere Struktur als Referenzstruktur (`Bref`).

SMN Tudor Domäne

Die SMN Tudor Domäne hat einer Länge von 56 Aminosäuren und besitzt eine β-Faß Faltung. Der strukturierte Teil der Domäne umfaßt die Reste 92-144 des humanen SMN Proteins [61].
Die beiden experimentellen Datensätze[1] wurden aus ^{13}C und ^{15}N markierten NOESY-Spektren abgeleitet und bestehen aus 1444 bzw. 431 Kreuzrelaxationsraten. Beide Datensätze sind vollständig zugeordnet. Die Kristallstruktur der Tudor Domäne [62] (PDB-Zugriffsnummer 1mhn) dient als Referenzstruktur (`Tref`).

[1] Beide Datensätze wurden freundlicherweise von Dr. M. Sattler, EMBL Heidelberg, zur Verfügung gestellt.

Kapitel 3

Ergebnisse

3.1 Strukturelle Unsicherheitsbehaftung

Die praktische Lösung eines induktiven Strukturbestimmungsproblems erfolgt über die Simulation der Strukturverteilung mit Hilfe der in Kapitel 2.3.5 besprochenen Replika-Austausch-Monte-Carlo-Strategie. Die dabei erzeugten konformationellen Stichproben repäsentieren die maximale Information über die möglichen Konformationen der Zielstruktur. Unvollständigkeiten in der Ausgangsinformation führen zu strukturellen Unsicherheiten, die sich in einer nichtverschwindenden Varianz der Strukturverteilung und somit in der Streuung der Proben äußern. Die Unsicherheitsbehaftung von Atompositionen kann dem Simulationsergebnis daher nicht direkt entnommen werden, sondern ist implizit in der Dichte der Proben kodiert.
Ich stelle in den folgenden Abschnitten eine Bayes'sche Methode vor, welche die Darstellung und Interpretation der Strukturverteilung in natürlicher Weise gestattet. Aus einem Satz konformationeller Stichproben wird ein analytisches Modell geschätzt, welches die Strukturverteilung approximiert. Die Parameter des Modells gestatten die Berechnung der Unsicherheitsbehaftung aller Atompositionen sowie die Angabe einer mittleren Struktur.
Ich demonstriere die Methode an zwei Systemen: Dem Protein BPTI und der Tudor Domäne des humanen SMN Proteins. Für beide Proteine standen

zugeordnete NOESY-Datensätze zur Verfügung (vgl. Kap. 2.5). Die Berücksichtiung der Daten in der Strukturrechnung erfolgt mit Hilfe eines Datenmodells für dipolare Kreuzrelaxationsraten, welches in Abschnitt 3.1.1 vorgestellt wird. Abschnitt 3.1.2 beschreibt das Verteilungsmodell zur Approximation der Strukturverteilung. Die Schätzung des Modells erfolgt durch einen MCMC-Algorithmus, der in Abschnitt 3.1.3 hergeleitet wird. Abschnitt 3.1.4 widmet sich der Simulation der Strukturverteilungen beider Testsysteme sowie der Berechnung der atomweisen Koordinatenunsicherheiten.

3.1.1 Modellierung von NOESY-Daten

In der induktiven Strukturbestimmung erfolgt die Berücksichtigung eines experimentellen NOESY-Datensatzes durch die Angabe einer entsprechenden *Likelihood*-Funktion. Das experimentelle NOESY-Spektrum sei in Form von N zugeordneten NOE-Intensitäten $D = \{\tilde{V}_i, \ldots, \tilde{V}_N\}$ gegeben.

3.1.1.1 Datenmodell für dipolare Kreuzrelaxationsraten

Als physikalisches Modell für die Berechnung von NOE-Intensitäten aus den Koordinaten einer gegebenen Struktur verwende ich die ISPA (Gl. 1.9). Das Vorwärtsmodell lautet also:

$$V_i(\mathbf{x}, \gamma) = \gamma d_i^{-6}(\mathbf{x}). \tag{3.1}$$

$V_i(\mathbf{x})$ bezeichnet die berechnete Intensität und $d_i(\mathbf{x})$ die Distanz des involvierten Atompaars in der Struktur; der Wert des Skalenfaktors γ ist unbekannt. Wurde der NOE durch die dipolare Wechselwirkung zwischen zwei Gruppen magnetisch equivalenter Spins hervorgerufen (wie beispielsweise Methylgruppen oder aromatische Ringe), berechne ich die Intensität des NOE als Summe über alle paarweisen Partialintensitäten:

$$V_i(\mathbf{x}, \gamma) = \gamma \sum_{jk \in J \times K} d_{i,jk}^{-6}(\mathbf{x}). \tag{3.2}$$

J und K bezeichnen die Indexmengen beider Gruppen und $d_{i,jk}(\mathbf{x})$ die Distanz zwischen den Atomen j und k. Gleichung (3.2) folgt aus einem dis-

kreten Sprungmodell zur Beschreibung langsamer dynamischer Mittelungseffekte [66]. Bedingt durch Meßfehler und systematische Effekte, welche in der ISPA nicht berücksichtigt werden, können observierte und berechnete Kreuzrelaxationsraten voneinander abweichen. Die physikalische Natur dieser Beiträge sowie ihr Einfluß auf die Größe der beobachteten Kreuzrelaxationsraten sei unbekannt. Dieses Unwissen wird in Form einer Wahrscheinlichkeitsverteilung für die observierte Intensität, dem *Datenmodell*, beschrieben. Ich leite das Datenmodell aus drei Annahmen her:

1. NOE Intensitäten sind positive Größen;
2. Die Abweichung zwischen observierter und berechneter Intensität läßt sich durch ein mulitplikatives Fehlergesetz beschreiben. Der Fehler ist folglich additiv für die Logarithmen der Intensitäten:

$$\log(\tilde{V}_i) = \log(V_i(\mathbf{x}, \gamma)) + \epsilon.$$

   $\tilde{V}_i$ bezeichnet die gemessene Intensität und ϵ die Abweichung der logarithmierten Intensitäten.
3. ϵ streut um 0 mit der Varianz σ^2, deren Wert unbekannt ist.

Die Verteilung, welche die in (3) aufgeführten Eigenschaften besitzt und darüber hinaus keine weiteren Annahmen enthält, wird durch das Prinzip der maximalen Entropie eindeutig bestimmt und ist die *Normalverteilung* [36]:

$$p(\epsilon|I) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\epsilon^2\right\}.$$

Durch die Variablentransformation $\epsilon \rightarrow \log \tilde{V}_i/V_i(\mathbf{x}, \gamma)$ und unter Beachtung des Vorwärtsmodells ergibt sich als Datenmodell für die observierte Kreuzrelaxationsrate die *Lognormalverteilung*

$$p_{\mathrm{NOE}}(\tilde{V}_i|\mathbf{x}, \sigma^2, \gamma, I) = \frac{1}{\sqrt{2\pi\sigma^2}} \tilde{V}_i^{-1} \exp\left\{-\frac{1}{2\sigma^2}\log^2\left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})}\right)\right\} \qquad (3.3)$$

mit Ortsparameter $\gamma d_i^{-6}(\mathbf{x})$ und Formparameter σ (vgl. Anhang A.1). I repräsentiert die Hintergrundinformation, welche für die Formulierung des Datenmodells relevant ist; in diesem Fall sind dies die Annahmen (1) bis (3).

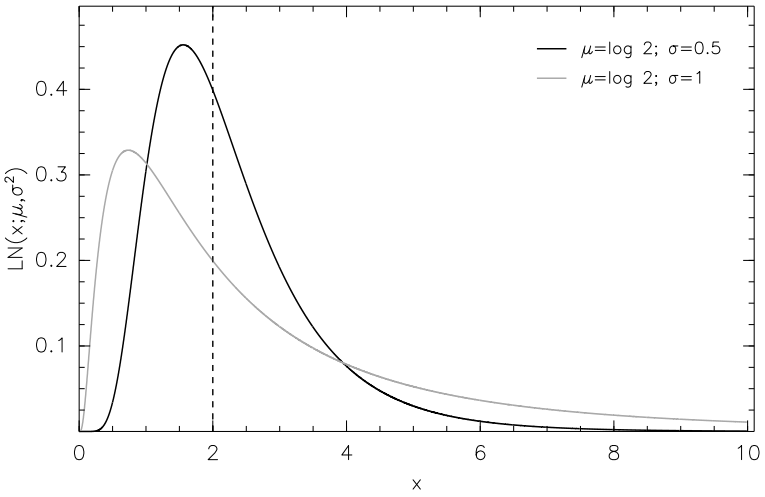

Abbildung 3.1: Datenmodell für NOE-Intensitäten. Graph der Lognormalverteilung für verschiedene Formparameter bei gleichem Ortsparameter.

Hilfsgrößen

Für die Formulierung des Datenmodells und die Interpretation der experimentellen Daten wurden die Hilfsgrößen σ und γ eingeführt. σ beschreibt die Größe der Abweichung von observierter und berechneter Kreuzrelaxationsrate, γ die Skala der observierten Werte. Keine der beiden Größen kann auf experimentellem Wege bestimmt werden: γ ließe sich im Prinzip zwar aus dem Vorfaktor der ISPA in Gleichung (1.8) ableiten, jedoch wird die Skala von NOE-Intensitäten durch die Prozessierung der Rohdaten beeinflußt und muß daher als unbekannt angesehen werden. Der Wert von σ wird von der Realitätstreue des gewählten Vorwärtsmodells bestimmt und ist von Natur aus keine experimentelle Observable. Beide Parameter gehen daher als *Nuisance*-Parameter in das Datenmodell ein und werden später durch Marginalisierung eliminiert.

Eigenschaften des Datenmodells

Die Lognormalverteilung (vgl. Abb. 3.1) aus Gleichung (3.3) ist für positive Intensitäten $\tilde{V}$ definiert und besitzt die folgenden Eigenschaften:

$$\int_0^\infty d\tilde{V}\, p(\tilde{V}|\mathbf{x}, \sigma^2, \gamma, I) \;=\; 1, \tag{3.4}$$

$$\begin{aligned} \text{Median}[\tilde{V}] &= V(\mathbf{x}), && (3.5)\\ \langle \tilde{V} \rangle &= V(\mathbf{x})\, e^{\sigma^2/2}, && (3.6)\\ \text{Var}[\tilde{V}] &= [V(\mathbf{x})]^2\ e^{\sigma^2} \left(e^{\sigma^2} - 1\right). && (3.7) \end{aligned}$$

Var bezeichnet die Varianz. Gemäß Gleichung (3.5) ist die Wahrscheinlichkeit, daß die beobachtete Intensität kleiner bzw. größer ist als der berechnete Wert, jeweils 1/2. Trotz der Asymmetrie der Dichtefunktion der Lognormalverteilung (bedingt durch die Positivität von Kreuzrelaxationsraten) erfolgt die Integration der Messungen somit symmetrisch und ist in diesem Sinne vorurteilsfrei.

3.1.1.2 Die *a-posteriori*-Verbundverteilung

Die Quantifizierung der Übereinstimmung der Struktur mit den experimentellen Daten erfolgt anhand einer *Likelihood*-Funktion. Im Falle logisch voneinander unabhängiger Messungen faktorisiert diese vollständig in die Anteile für die einzelnen Intensitäten:

$$p(D|\mathbf{x}, \sigma^2, \gamma, I) = \prod_{i=1}^{N} p_{\text{NOE}}(\tilde{V}_i|\mathbf{x}, \sigma^2, \gamma, I).$$

Durch Einsetzen des Datenmodells aus Gl. (3.3) und Ausnutzen der Eigenschaften der Exponentialfunktion ergibt sich die *Likelihood*-Funktion zu:

$$p(D|\mathbf{x}, \sigma^2, \gamma, I) \propto \sigma^{-N} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \log^2\left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})}\right)\right\}, \qquad (3.8)$$

wobei alle konstanten Vorfaktoren aus Gl. (3.3) weggelassen wurden. Für die Angabe der *a-posteriori*-Verteilung für alle unbekannten Hypothesenparameter, $(\mathbf{x}, \sigma^2, \gamma)$, muß die *a-priori*-Verteilung, die bisher lediglich für die Koordinaten der Struktur konkret formuliert wurde (vgl. Kap. 2.2.2), auf den vollständigen Hypothesenraum erweitert werden. Dazu nehme ich an, daß der Skalenfaktor γ und die Datenvarianz σ^2 logisch voneinander unabhängig sind. Unter dieser Annahme faktorisiert die *a-priori*-Verteilung:

$$p(\mathbf{x}, \sigma^2, \gamma|I) = Z^{-1}(\beta) e^{-\beta E(\mathbf{x})} p(\sigma^2|I) p(\gamma|I).$$

Als *a-priori*-Verteilungen für σ^2 und γ wähle ich den jeweiligen *Jeffreys-prior* [67, 68, 36]. Jeffreys-*prior* repräsentieren völliges „Nichtwissen" über den Wert eines Hypothesenparameters und werden mit Hilfe der Methode der Transformationsgruppen aus Symmetrien eines Datenmodells oder, in allgemeiner Form, aus der Fisher-Informations-Matrix hergeleitet [36, 69]. σ und γ zählen zur Klasse der Skalenvariablen, mit einem Jeffreys-*prior* von $p(\sigma|D) = \sigma^{-1}$ bzw. $p(\gamma|I) = \gamma^{-1}$ [67, 36]. Aus dem Jeffreys-*prior* für σ folgt die *a-priori*-Verteilung für σ^2 durch die Variablentransformation $\sigma \rightarrow \sigma^2$. Damit ergibt sich die vollständige *a-priori*-Verteilung für $(\mathbf{x}, \sigma^2, \gamma)$ zu:

$$p(\mathbf{x}, \sigma^2, \gamma|I) = Z^{-1}(\beta)\, e^{-\beta E(\mathbf{x})}\, (\sigma^2\gamma)^{-1}.$$

Die *a-posteriori*-Verbundverteilung für die Konformation $\mathbf{x}$ sowie für alle *Nuisance*-Parameter $\alpha = (\gamma, \sigma^2)$ folgt aus dem Bayes'schen Satz und besitzt gemäß Gleichung (2.11) die Form

$$p(\mathbf{x}, \sigma^2, \gamma|D, I) \propto \gamma^{-1}\sigma^{-(N+2)} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\log^2\left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})}\right) - \beta E(\mathbf{x})\right\}, \tag{3.9}$$

wobei alle konstanten Faktoren der Übersicht halber weggelassen wurden. Die *a-posteriori*-Verteilung in Gleichung (3.9) repräsentiert die maximale Information hinsichtlich der möglichen Werte der Hypothesenparameter $(\mathbf{x}, \sigma^2, \gamma)$, welche aus dem Datensatz D und der Hintergrundinformation I extrahiert werden kann. Die Ableitung dieses Wissens erfolgte in objektiver Weise: Die analytische Form von Gleichung (3.9) folgt mit Hilfe des Bayes'schen Satzes eindeutig aus dem Datenmodell und der *a-priori*-Verteilung. Mathematische Ausdrücke für das Datenmodell und die *a-priori*-Verteilung wurden aus explizit formulierter Hintergrundinformation hergeleitet und enthalten darüberhinaus keine weiteren Annahmen.

3.1.1.3 Die Strukturverteilung

Nach Gleichung (2.12) ergibt sich die Strukturverteilung durch Marginalisierung aller *Nuisance*-Parameter:

$$p(\mathbf{x}|D,I) \propto e^{-\beta E(\mathbf{x})} \int_0^\infty d\sigma^2\, d\gamma\, \gamma^{-1} \sigma^{-(N+2)} \times \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \log^2 \left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})} \right) \right\}. \tag{3.10}$$

Gleichung (3.10) ist die formale Lösung des hier formulierten induktiven Strukturbestimmungsproblems für zugeordnete NOESY-Daten. Die Verteilung der konformationellen Wahrscheinlichkeiten spiegelt die Aussagekraft der experimentellen Daten in Verbindung mit der berücksichtigten Hintergrundinformation wider und ist ein Maß für das Unwissen über die wahre Konformation des Zielmoleküls. Bildlich läßt sich die Ausdehnung der Strukturverteilung daher im Sinne eines statistischen Fehlerbalkens interpretieren. Der analytische Ausdruck der Strukturverteilung folgt mit Hilfe der Regeln der Wahrscheinlichkeitstheorie *eindeutig* aus dem Datenmodell und der *a-priori*-Verteilung und hängt darüber hinaus von keinem freien Parameter ab: Unbekannte Hilfsgrößen werden durch Marginalisierung eliminiert, wodurch ihre Unsicherheitsbehaftung vollständig in der Strukturverteilung berücksichtigt wird. Die möglichen Konformationen des Moleküls werden somit ausschließlich von den experimentellen Daten und von Zusatzwissen determiniert, welches für die Interpretation der Daten vonnöten ist; die Strukturverteilung ist in diesem Sinne eine *objektive* Darstellung von struktureller Unsicherheit.

3.1.2 Approximation der Strukturverteilung

Unsicherheiten in den Koordinaten der Atome sind implizit in der Geometrie der Strukturverteilung kodiert. Um die atomare Unsicherheitsbehaftung zu extrahieren, ist es zweckmäßig, die Strukturverteilung mit Hilfe eines analytischen Modells zu approximieren. Die Approximation der Strukturverteilung erfolgt dabei vor dem Hintergrund, anhand des Modells Aussagen über die

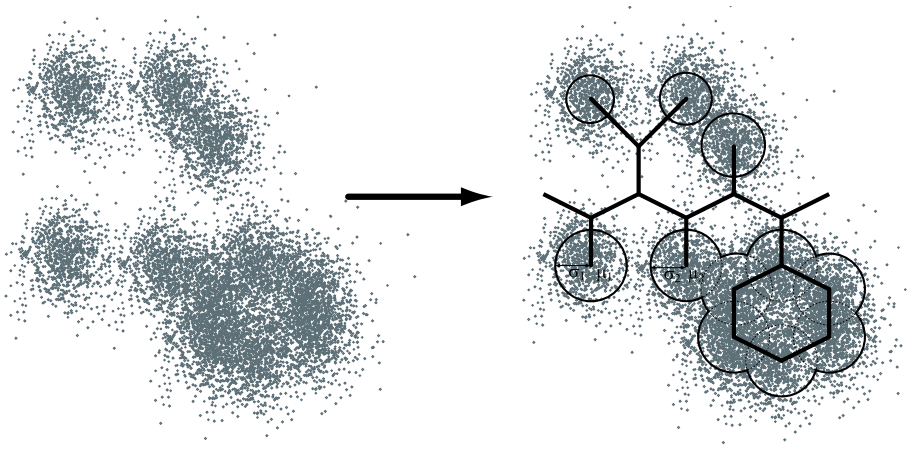

Abbildung 3.2: Parametrisierung des Verteilungsmodells. Die Strukturverteilung ist implizit über konformationelle Stichproben gegeben (links). Das Verteilungsmodell (rechts) beschreibt die Form der Strukturverteilung anhand einer mittleren Struktur mit den Koordinaten $\mu_1, \mu_2, \ldots$; atomare Unsicherheiten werden durch individuelle Varianzen $\sigma_1^2, \sigma_2^2, \ldots$ modelliert.

geometrische Form der Strukturverteilung treffen zu können.
Das Problem besteht somit darin, die Strukturverteilung, von der nur ein Satz konformationeller Stichproben $\mathbf{x}^{(i)} \sim p(\mathbf{x}|D, I)$ bekannt ist, durch eine Wahrscheinlichkeitsdichtefunktion $p_{\text{par}}(\mathbf{x}|\varphi)$ anzunähern. $p_{\text{par}}(\mathbf{x}|\varphi)$ wird als *Verteilungsmodell* bezeichnet, dessen geometrische Form durch den Parametersatz φ bestimmt wird. Die Anpassung des Verteilungsmodells an die Strukturverteilung erfolgt durch die Wahl geeigneter Parameterwerte φ. Im Wahrscheinlichkeitskalkül wird dieses Problem durch die Bestimmung der *a-posteriori*-Verteilung für die Hypothesenparameter φ gelöst; die konformationellen Stichproben fungieren dabei als „Daten“, aus denen das Verteilungsmodell *geschätzt* wird.

3.1.2.1 Ein analytisches Modell

Ich wählte die Parametrisierung des Verteilungsmodells daher so, daß alle Parameter eine geometrische Interpretation besitzen. Die Parametrisierung des

Modells ist in Abbildung 3.2 illustriert. Ich beschreibe die prinzipielle Form der Strukturverteilung anhand einer mittleren Struktur. Unsicherheiten in den Atompositionen werden durch einen Satz von individuellen Varianzen modelliert, welche die Ausdehnung der Strukturverteilung um die mittlere Konformation beschreiben. Bei der Definition der mittleren Struktur kommt erschwerend hinzu, daß die gezogenen Konformationen auf unterschiedlichen Koordinatensystemen definiert sind: Theoretische NOE-Intensitäten und van der Waals-Wechselwirkungen hängen nur von interatomaren Abständen ab. Die *a-posteriori*-Verteilung in Gleichung (3.9) ist deshalb invariant unter Rotation und Translation des kartesischen Koordinatensystems. Unterschiede in den Koordinatensystemen müssen daher im Verteilungsmodell berücksichtigt werden.

Eine von der Strukturverteilung gezogene Konformation bestehe aus M Atomen mit den kartesischen Koordinaten $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ und sei aus der mittleren Struktur mit den kartesischen Koordinaten $\mu = \{\mu_1, \ldots, \mu_M\}$ durch Rotation und anschließende Translation hervorgegangen. Die Position des j-ten Atoms der Konformation möge von der Position des j-ten Atoms der mittleren Struktur um $\mathbf{e}_j$ abweichen. Damit lautet der Zusammenhang zwischen der i-ten Konformation und der mittleren Struktur:

$$\mathbf{x}_j^{(i)} = \mathbf{R}^{(i)}\mu_j + \mathbf{t}^{(i)} + \mathbf{e}_j. \tag{3.11}$$

$\mathbf{x}_j^{(i)}$ bezeichnet die kartesischen Koordinaten des j-ten Atoms, $\mathbf{R}^{(i)}$ die Rotationsmatrix und $\mathbf{t}^{(i)}$ den Translationvektor der i-ten Konformation. Die Koordinaten der mittleren Struktur, die Abweichungen in den Koordinaten $\{\mathbf{e}_1, \ldots, \mathbf{e}_M\}$ sowie die individuellen Transformationen $(\mathbf{R}, \mathbf{t})^{(i)}$ sind unbekannt.

Für die Formulierung des Verteilungsmodells nehme ich die logische Unabhängigkeit der $\{\mathbf{e}_j\}$ an und ferner, daß die Verteilung von $\mathbf{e}_j$ konformationsunabhängig durch eine isotrope und um Null zentrierte Normalverteilung

mit Varianz σ_j^2 beschrieben werden kann:

$$\begin{aligned} p(\{\mathbf{e}_j\}|\{\sigma_j^2\}, I) &= \prod_{j=1}^{M} \mathrm{N}(\mathbf{e}_j; 0, \sigma_j^2) \\ &= \prod_{j=1}^{M} \left(2\pi\sigma_j^2\right)^{-3/2} \exp\left\{-\frac{1}{2\sigma_j^2}\mathbf{e}_j^{\mathsf{T}}\mathbf{e}_j\right\}. \end{aligned} \tag{3.12}$$

Aus Gln. (3.11) und (3.12) folgt als parametrisches Verteilungsmodell für die kartesischen Koordinaten der i-ten Konformation eine im Konformationsraum um die mittlere Struktur μ zentrierte, $3M$-dimensionale Normalverteilung mit diagonaler Kovarianzmatrix:

$$\begin{aligned} &p_{\mathrm{par}}\left(\mathbf{x}^{(i)}|\{\mu_j\}, \{\sigma_j^2\}, \mathbf{R}^{(i)}, \mathbf{t}^{(i)}, I\right) = \\ &\prod_{j=1}^{M} \left(2\pi\sigma_j^2\right)^{-3/2} \exp\left\{-\frac{1}{2\sigma_j^2}\|\mathbf{x}_j^{(i)} - \mathbf{R}^{(i)}\mu_j - \mathbf{t}_j^{(i)}\|^2\right\}. \end{aligned} \tag{3.13}$$

3.1.2.2 Definition der Koordinatenunsicherheiten

Eine konkrete Definition der Unsicherheit einer Atomposition orientiert sich an der Parametrisierung des Verteilungsmodells. Ein vor dem Hintergrund der gewählten isotropen Fehlerverteilung natürliches Maß für Unsicherheit ist die Größe der „Unsicherheitssphäre“, welche jedem Atom anhand der Fehlerverteilung zugeordnet wird. Ich definiere die Unsicherheit in der Position des j-ten Atoms, δ_j, als Radius der ihm zugeordnenten Unsicherheitssphäre. Das j-te Atom ist demzufolge mit einer Wahrscheinlichkeit von 68% innerhalb einer Kugel mit Radius δ_j um die mittlere Position μ_j lokalisiert; bei einer 2σ-Unsicherheitsbehaftung entsprechend mit Wahrscheinlichkeit 96% in einer Kugel mit Radius $2\delta_j$. Aus Gleichung (3.12) folgt:

$$\delta_j = \langle\|\mathbf{e}_j\|\rangle = \sqrt{3}\sigma_j. \tag{3.14}$$

Die Unsicherheiten $\{\delta_j\}$ werden im Verteilungsmodell somit bis auf einen konstanten Faktor durch die Hypothesenparameter $\{\sigma_j\}$, die kartesischen Koordinaten der mittleren Struktur explizit durch $\{\mu_j\}$ modelliert. Die konformationsspezifischen Transformationen $\{(\mathbf{R}, \mathbf{t})^{(i)}\}$ wurden lediglich für die

Herstellung des Zusammenhangs zwischen den Koordinaten der gezogenen Strukturen und der mittleren Struktur eingeführt und enthalten keine Information über die Form der Strukturverteilung. Ich fasse Rotationsmatrizen und Translationsvektoren daher als *Nuisance*-Parameter auf und werde beide Größen später durch Marginalisierung eliminieren.
Die Schätzung der Koordinatenunsicherheiten und der mittleren Struktur aus einer Menge von Konformationen erfolgt, wie im Wahrscheinlichkeitskalkül üblich, auf Basis der marginalen *a-posteriori*-Verteilung für $\{\sigma_j^2\}$ und $\{\mu_j\}$, welche es im Folgenden abzuleiten gilt.

3.1.2.3 Die *a-posteriori*-Verteilung

Aus Gleichung (3.13) folgt als *Likelihood*-Funktion für N unabhängig von der Strukturverteilung gezogene Konformationen $D = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$:

$$p\left(D|\{\mu_j\},\{\sigma_j^2\},\{\mathbf{R}^{(i)}\},\{\mathbf{t}^{(i)}\},I\right) = \prod_{j=1}^{M}\left(2\pi\sigma_j^2\right)^{-3N/2}\exp\left\{-\frac{1}{2\sigma_j^{-2}}\sum_{i=1}^{N}\|\mathbf{x}_j^{(i)}-\mathbf{R}^{(i)}\mu_j-\mathbf{t}^{(i)}\|^2\right\}. \tag{3.15}$$

Die *a-posteriori*-Verbundverteilung ergibt sich nach dem Bayes'schen Satz aus der *Likelihood*-Funktion und der *a-priori*-Verteilung für alle unbekannten Hypothesenparameter. Ich formuliere die *a-priori*-Verteilung unter der Annahme der logischen Unabhängigkeit aller Hypothesenparameter. Als *a-priori*-Verteilung für $\{\mu_j\}$, $\{\sigma_j^2\}$ und $\{\mathbf{t}^{(i)}\}$ wähle ich den jeweiligen Jeffreys-*prior*. Der Jeffreys-*prior* für die Ortsparameter $\{\mu_j\}$ und $\{\mathbf{t}^{(i)}\}$ ist gleichverteilt [36]. Aus dem Jeffreys-*prior* für die Skalenvariable σ_j folgt durch die Variablentransformation $\sigma_j \to \sigma_j^2$ die *a-priori*-Verteilung für die Varianzen $\{\sigma_j^2\}$:

$$p(\{\sigma_j^2\}|I) = \prod_{j=1}^{M}\sigma_j^{-2}.$$

Die relativen Orientierungen der Konformationen in Bezug auf die mittlere Struktur ist *a-priori* unbekannt, was in der *a-priori*-Verteilung für die Rotationsmatrizen zu berücksichtigen ist. Ferner ist zu beachten, daß Rotationsmatrizen die Orthogonalitätsrelation $\mathbf{R}^{\mathsf{T}}\mathbf{R} = \mathbf{1}$ erfüllen, was durch ihre interne

Parametrisierung in Eulerwinkeln gewährleistet wird: Jede Rotationsmatrix wird dabei zerlegt in sukzessive Rotationen um die z-, y- und z-Achse, also

$$\mathbf{R}^{(i)} \equiv \mathbf{R}(\alpha^{(i)}, \beta^{(i)}, \gamma^{(i)}) = \mathbf{R}_z(\gamma^{(i)})\, \mathbf{R}_y(\beta^{(i)})\, \mathbf{R}_z(\alpha^{(i)}), \tag{3.16}$$

mit $\alpha^{(i)}, \gamma^{(i)} \in [0, 2\pi]$ und $\beta^{(i)} \in [0, \pi]$. Ich wähle gleichverteilte *a-priori*-Verteilungen für alle Eulerwinkel, wodurch die Hauptachsen der Rotationsmatrizen isotrop im Raum verteilt sind. Damit erhält man für die *a-priori*-Verbundverteilung:

$$p(\{\mu_j\}, \{\sigma_j^2\}, \{\alpha^{(i)}, \beta^{(i)}, \gamma^{(i)}\}, \{\mathbf{t}^{(i)}\}|I) = 4^{-N}\pi^{-3N}\prod_{j=1}^{M}\sigma_j^{-2}. \tag{3.17}$$

Aus der *Likelihood*-Funktion in Gl. (3.15) und der *a-priori*-Verbundverteilung in Gl. (3.17) folgt die *a-posteriori*-Verbundverteilung für alle Hypothesenparameter:

$$\begin{aligned} &p(\{\mu_j\}, \{\sigma_j^2\}, \{\alpha^{(i)}\}, \{\beta^{(i)}\}, \{\gamma^{(i)}\}, \{\mathbf{t}^{(i)}\}|D, I) \propto \\ &\prod_{j=1}^{M}\left(\sigma_j^2\right)^{-(3N/2+1)} \exp\left\{-\frac{1}{2\sigma_j^{-2}}\sum_{i=1}^{N}\|\mathbf{x}_j^{(i)} - \mathbf{R}^{(i)}\mu_j - \mathbf{t}^{(i)}\|^2\right\}, \end{aligned} \tag{3.18}$$

wobei alle konstanten Faktoren vernachlässigt wurden und die Definition der Rotationsmatrizen in Gl. (3.16) zu beachten ist. Die gesuchte marginale *a-posteriori*-Verteilung für die Varianzen und die mittlere Struktur folgt durch Integration der *a-posteriori*-Verbundverteilung über alle Eulerwinkel und Translationsvektoren:

$$\begin{aligned} p(\{\mu_j\}, \{\sigma_j^2\}|D, I) = &\int \prod_{i=1}^{N} d\alpha^{(i)}\, d\beta^{(i)}\, d\gamma^{(i)}\, d^3\mathbf{t}^{(i)} \\ \times \quad &p(\{\mu_j\}, \{\sigma_j^2\}, \{\alpha^{(i)}\}, \{\beta^{(i)}\}, \{\gamma^{(i)}\}, \{\mathbf{t}^{(i)}\}|D, I). \end{aligned} \tag{3.19}$$

Die marginale *a-posteriori*-Verteilung für $\{\sigma_j^2\}$ und $\{\mu_j\}$ hängt von keinen freien Parametern ab und wird ausschließlich von dem gegebenen Satz von Konformationen und der Hintergrundinformation bedingt. Sie repräsentiert die vollständige Information über die mittlere Struktur und die atomweisen Varianzen in objektiver Weise. Das Marginalisierungsintegral in Gl. (3.19)

ist nicht analytisch lösbar. Ich leite im folgenden Abschnitt einen Markov-Ketten-Monte-Carlo-Algorithmus ab, der die numerische Simulation der *a-posteriori*-Verteilung in Gleichung (3.18) gestattet. Das Marginalisierungsintegral wird auf diese Weise auf simulatorischem Wege gelöst.

3.1.3 Berechnung des Verteilungsmodells

Ein geeigneter Algorithmus für die Bestimmung aller unbekannten Größen folgt aus der wahrscheinlichkeitstheoretischen Formulierung des Problems: Die analytische Form der *a-posteriori*-Verbundverteilung erlaubt die Simulation des Modells durch Gibbs-*Sampling*. Die hierfür benötigten bedingten *a-posteriori*-Verteilungen ergeben sich aus Gleichung (3.18) durch Zusammensammeln der entsprechenden Terme:

$$p(\sigma_j^2|\cdot) = \mathrm{IG}\left(\sigma_j^2; \frac{3N}{2}, \frac{N}{2}\left\langle \|\mathbf{x}_j^{(i)} - \mathbf{R}^{(i)}\mu_j - \mathbf{t}^{(i)}\|^2 \right\rangle^{(i)}\right), \tag{3.20}$$

$$p(\mu_j|\cdot) = \mathrm{N}\left(\mu_j; \left\langle \mathbf{R}^{\mathsf{T}}(\mathbf{x}_j^{(i)} - \mathbf{t}^{(i)})\right\rangle^{(i)}, \frac{\sigma_j^2}{N}\right), \tag{3.21}$$

$$p(\mathbf{t}^{(i)}|\cdot) = \mathrm{N}\left(\mathbf{t}^{(i)}; \frac{\left\langle \sigma_j^{-2}\left(\mathbf{x}_j^{(i)} - \mathbf{R}^{(i)}\mu_j\right)\right\rangle_j}{\left\langle \sigma^{-2}\right\rangle}, \left\langle \sigma^2 \right\rangle\right), \tag{3.22}$$

$$p(\mathbf{R}^{(i)}|\cdot) \propto \exp\left\{ M \left\langle \sigma_j^{-2}\left(\mathbf{t}_j^{(i)} - \mathbf{x}_j^{(i)}\right)^{\mathsf{T}} \mathbf{R}^{(i)}\, \mu_j \right\rangle_j \right\}. \tag{3.23}$$

$\langle\cdot\rangle^{(i)}$ und $\langle\cdot\rangle_j$ bezeichnet den Mittelwert bezüglich des oberen bzw. unteren Index; in eindeutigen Fällen wurde der Index weggelassen. Die Ziehung von $\{\sigma_j^2\}$, $\{\mu_j\}$ und $\{\mathbf{t}^{(i)}\}$ in den Schritten (3.20) bis (3.22) des Gibbs-Algorithmus realisiere ich direkt mittels Zufallszahlengeneratoren für die inverse Gammaverteilung bzw. die Normalverteilung.

Ziehung der Eulerwinkel

Um die Rotationsmatrizen, respektive den Satz von Eulerwinkeln, ebenfalls mittels Zufallszahlengeneratoren ziehen zu können, bringe ich die bedingte *a-posteriori*-Verteilung in Gl. (3.23) mit Hilfe der Identität $\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{y} \equiv$

$\mathrm{Sp}\left\{\mathbf{A}\,\mathbf{y}\mathbf{x}^{\mathsf{T}}\right\}$ auf folgende Form:

$$p(\alpha^{(i)},\beta^{(i)},\gamma^{(i)}|\cdot) \;\propto\; \exp\left\{M\,\mathrm{Sp}\left(\underbrace{\mathbf{R}_z(\gamma^{(i)})\mathbf{R}_y(\beta^{(i)})\mathbf{R}_z(\alpha^{(i)})}_{\mathbf{R}^{(i)}}\mathbf{P}^{(i)}\right)\right\}. \quad (3.24)$$

$\mathrm{Sp}(\cdot)$ bezeichnet die Spur und $\mathbf{P}^{(i)}$ die Projektionsmatrix

$$\mathbf{P}^{(i)} = \left\langle \sigma_j^{-2}\,\mu_j\left(\mathbf{x}_j^{(i)} - \mathbf{t}^{(i)}\right)^{\mathsf{T}}\right\rangle_j. \quad (3.25)$$

Die bedingte *a-posteriori*-Verteilung für einen Eulerwinkel $\zeta \in \{\alpha^{(i)},\beta^{(i)},\gamma^{(i)}\}$ ergibt sich durch Ausführen der Spur in Gl. (3.24) und besitzt aufgrund der speziellen Form der Euler-Matrizen stets die Form

$$p(\zeta|\cdot) = C\exp\left\{M\left(A\cos\zeta + B\sin\zeta\right)\right\}. \quad (3.26)$$

C bezeichnet eine Normierungskonstante. Die Koeffizienten A und B ergeben sich nach Ausführen der Spur durch Zusammensammeln der von ζ unabhängigen Terme. Durch Einführen eines *Phasen-* und *Formparameters* ϕ bzw. κ mit

$$A = \kappa_\zeta\cos\phi_\zeta \quad \text{und} \quad B = \kappa_\zeta\sin\phi_\zeta$$

geht Gleichung (3.26) mit Hilfe der Additionstheoreme für trigonometrische Funktionen in eine von Mises-Verteilung (s. Anhang A.4) über, für welche ebenfalls Zufallszahlengeneratoren existieren [70]:

$$\begin{aligned} p(\zeta|\cdot) &= C\exp\left\{\kappa_\zeta\cos\left(\zeta-\phi_\zeta\right)\right\} \\ &\equiv \mathrm{M}(\zeta;\kappa_\zeta,\phi_\zeta). \end{aligned}$$

$\mathrm{M}(\zeta;\kappa,\phi)$ bezeichnet eine von Mises-Verteilung für die zyklische Variable ζ mit Formparameter κ und Phasenparameter ϕ. Die bedingten *a-posteriori*-Verteilungen für die Eulerwinkel $\{\alpha^{(i)},\beta^{(i)},\gamma^{(i)}\}$ ergeben sich damit zu:

$$\begin{aligned} p(\alpha^{(i)}|\cdot) &= \mathrm{M}(\alpha^{(i)};\kappa_{\alpha^{(i)}},\phi_{\alpha^{(i)}}), \\ p(\beta^{(i)}|\cdot) &= \mathrm{M}(\beta^{(i)},\kappa_{\beta^{(i)}},\phi_{\beta^{(i)}}), \\ p(\gamma^{(i)}|\cdot) &= \mathrm{M}(\gamma^{(i)},\kappa_{\gamma^{(i)}},\phi_{\gamma^{(i)}}). \end{aligned} \quad (3.27)$$

Für alle Atome j**:** Ziehe Varianz und mittlere Position.

- $[\sigma_j^2]^{(k+1)} \sim p(\sigma_j^2|\cdot)$,
- $\mu_j^{(k+1)} \sim p(\mu_j|\cdot)$.

Für alle Konformationen i**:** Ziehe Transformation.

- $[\mathbf{t}^{(i)}]^{(k+1)} \sim p(\mathbf{t}^{(i)}|\cdot)$,
- Berechne Projektionsmatrix $\mathbf{P}^{(i)}$ entsprechend Gl. (3.25).

Für die Eulerwinkel $\zeta \in \{\alpha^{(i)}, \beta^{(i)}, \gamma^{(i)}\}$**:**

- Berechne Phasen- und Formparameter ϕ_ζ bzw. κ_ζ aus $\mathbf{P}^{(i)}$ und den Euler-Matrizen der beiden nicht betrachteten Winkel;
- ziehe Eulerwinkel gemäß $\zeta \sim p(\zeta|\cdot)$.

- Berechne Rotationsmatrix $[\mathbf{R}^{(i)}]^{(k+1)}$ gemäß Gl. (3.16).

Abbildung 3.3: Gibbs-Algorithmus für die Simulation der *a posteriori*-Verbundverteilung des Verteilungsmodells.

Der Gibbs-Algorithmus

Der Gibbs-Algorithmus für die Simulation des Verteilungsmodells basiert auf den bedingten *a-posteriori*-Verteilungen in Gln. (3.20) bis (3.22) und (3.27) und ist in Abbildung (3.3) dargestellt: Für die Erzeugung einer neuen Probe in Schritt k+1 werden zu Beginn alle Varianzen sowie die Koordinaten der mittleren Struktur von Gln. (3.20) bzw. (3.21) gezogen, gefolgt von der Generierung der individuellen Transformationen: Für jede Konformation wird dazu ein neuer Translationsvektor von Gl. (3.22) gezogen und anschließend jeder der drei Eulerwinkel sequenziell gemäß Gln. (3.27) generiert.

Durch die wiederholte Anwendung dieser Vorschrift wird eine Markov Kette realisiert, deren Zustände $(\{\mu_j\}, \{\sigma_j^2\})^{(k)}$ gemäß der marginalen *a-posteriori*-Verteilung in Gl. (3.19) verteilt sind. Die Stichproben der *Nuisance*-Parameter

$(\mathbf{R}, \mathbf{t})^{(k)}$ sind nicht von weiterem Interesse und werden verworfen.
Der beschriebene Gibbs-Algorithmus wurde als Computerprogramm realisiert und in das ISD-Simulatonspaket integriert. Zeitkritische Routinen sind in C, die gesamte Infrastruktur des Programms in der Skriptsprache Python implementiert.

3.1.4 Testrechnungen

Ich diskutiere die Methode anhand von BPTI und der Tudor Domäne (vgl. Kap. 2.5). Abschnitt 3.1.4.1 zeigt die Realisierung des Replika-Austausch-Monte-Carlo-Algorithmus für die Simulation der Strukturverteilungen auf Basis zugeordneter NOESY-Daten. Abschnitte 3.1.4.2 und 3.1.4.3 befassen sich mit der Simulation beider Strukturverteilungen und der Berechnung der atomaren Unsicherheitsbehaftung.

3.1.4.1 Realisierung des Replika-Algorithmus

Die konkrete Form des Gibbs-Algorithmus zur Simulation der Strukturverteilung folgt aus dem in Kapitel 2.3.5 angegebenen allgemeinen Gibbs-Schema. Die Erzeugung einer Stichprobe von der *a-posteriori*-Verbundverteilung in Gleichung (3.9) erfolgt in drei Schritten, in welchen die Datenvarianz σ^2, der Skalenparameter γ sowie die Koordinaten der Struktur $\mathbf{x}$ nacheinander von ihren bedingten *a-posteriori*-Verteilungen gezogen werden. Die bedingten *a-posteriori*-Verteilungen für σ^2 und γ können unmittelbar aus Gl. (3.9) abgelesen werden:

$$p(\sigma^2|\mathbf{x}, \gamma, D, I) = \mathrm{IG}\left(\sigma^2; N/2, \frac{1}{2}\sum_{i=1}^{N} \log^2\left(\tilde{V}_i/\gamma d_i^{-6}(\mathbf{x})\right)\right), \quad (3.28)$$

$$p(\gamma|\mathbf{x}, \sigma^2, D, I) = \mathrm{LN}\left(\gamma; \frac{1}{N}\sum_{i=1}^{N} \log\left(\tilde{V}_i/d_i^{-6}(\mathbf{x})\right), \sigma^2/N\right), \quad (3.29)$$

$\mathrm{IG}(\cdot)$ und $\mathrm{LN}(\cdot)$ bezeichnet die inverse Gammaverteilung (s. Anhang A.2) bzw. die Lognormalverteilung. Ich verwende Zufallszahlengeneratoren für die inverse Gammaverteilung bzw. die Lognormalverteilung, um σ^2 und γ direkt von ihren bedingten *a-posteriori*-Verteilungen in Gln. (3.28) bzw. (3.29)

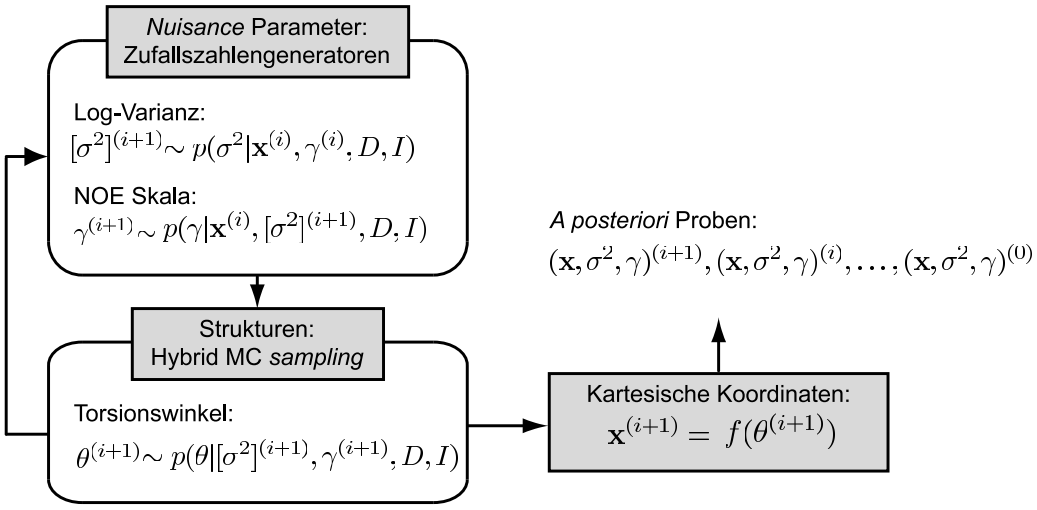

Abbildung 3.4: Gibbs-Algorithmus zur Simulation der Strukturverteilung. Für einen MC Schritt werden die *Nuisance*-Parameter von ihren bedingten *a-posteriori*-Verteilungen gezogen. Anschließend werden neue Dihedralwinkel *via* Hybrid-Monte-Carlo von der konformationellen *a-posteriori*-Verteilung gezogen und in kartesische Koordinaten überführt.

zu ziehen. Die Erzeugung einer neuen Konformation erfolgt durch Hybrid-Monte-Carlo. Die für die Berechnung der Dynamiktrajektorie erforderliche Energiefunktion folgt ebenfalls aus Gleichung (3.9):

$$\begin{aligned} U(\theta) &= -\log p(\mathbf{x}(\theta)|\sigma^2, \gamma, D, I) \\ &= \beta E(\mathbf{x}(\theta)) + \frac{1}{2\sigma^2} \sum_{i=1}^{N} \log^2 \left(\tilde{V}_i / \gamma d_i^{-6}(\mathbf{x}(\theta)) \right). \end{aligned} \tag{3.30}$$

Abbildung 3.4 illustriert das vollständige Gibbs-Schema: Für jeden Monte-Carlo-Schritt werden zu Beginn die *Nuisance*-Parameter σ^2 und γ mittels Zufallszahlengeneratoren von ihren bedingten *a-posteriori*-Verteilungen aus Gln. (3.28) bzw. (3.29) gezogen. Die Erzeugung einer neuen Struktur erfolgt mittels Hybrid-Monte-Carlo: Ein Satz von Dihedralwinkeln θ wird von der konformationellen *a-posteriori*-Verteilung gezogen, woraus die neuen kartesischen Koordinaten $\mathbf{x}$ der Struktur berechnet werden. Die Verallgemeinerung des Gibbs-Schemas auf mehrere NOESY-Datensätze erfolgt nach der in Kapitel 2.2.3 angegebenen Art und Weise.

λ_{01}-λ_{07}:	1.0000	0.9688	0.9375	0.9062	0.8750	0.8438	0.8125
λ_{08}-λ_{14}:	0.7812	0.7500	0.7188	0.6875	0.6562	0.6250	0.5938
λ_{15}-λ_{21}:	0.5625	0.5312	0.5000	0.4688	0.4375	0.4062	0.3750
λ_{22}-λ_{25}:	0.3438	0.3125	0.2812	0.2500			
q_{01}-q_{07}:	1.0010	1.0020	1.0032	1.0045	1.0059	1.0074	1.0091
q_{08}-q_{14}:	1.0110	1.0131	1.0154	1.0180	1.0208	1.0239	1.0274
q_{15}-q_{21}:	1.0312	1.0354	1.0400	1.0452	1.0509	1.0572	1.0641
q_{22}-q_{25}:	1.0718	1.0803	1.0896	1.1000			

Tabelle 3.1: (λ, q) Einstellungen des Replika-Algorithmus. Werte der Replika-Parameter für die λ-Kette (obere Hälfte, $q = 1$) und die q-Kette (untere Hälfte, $\lambda = \lambda_{\min} = 0.25$).

3.1.4.2 Simulation der Strukturverteilungen

Die Simulation der Strukturverteilungen von BPTI und der Tudor Domäne erfolgte durch Replika-Austausch-Monte-Carlo unter Verwendeung des oben beschriebenen Gibbs-Schemas. Die Varianz der Daten σ^2 sowie der NOE-Skalenfaktor γ sind in beiden Fällen *a priori* unbekannt und wurden während der Rechnung aus den Daten geschätzt. Die thermodynamischen Eigenschaften beider Systeme wurden jeweils über ein Boltzmann-Ensemble bei Temperatur $T = 300$K berücksichtigt.

Simulation von BPTI

Die drei von BPTI ausgebildeten Disulfidbrücken wurden anhand einer normalverteilten Distanz zwischen den jeweiligen SG-Schwefelatomen mit einem Gleichgewichtswert von 2.02 Å modelliert. Die λ- und q-Kette bestanden aus jeweils 25 Kopien. Die Replika-Parameter (λ, q), welche die Transformation der individuellen *a-posteriori*-Verteilungen parametrisieren, wurden so gewählt, daß sich während der Simulation homogene Austauschraten ergaben (s. Tab. 3.1). Für jeden Replika-Schritt wurden 25 Gibbs/HMC Schritte berechnet; die Länge der Dynamiktrajektorie des Hybrid-Monte-Carlo-

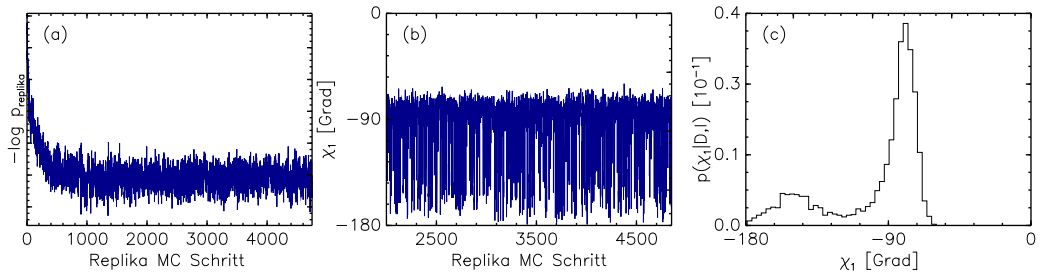

Abbildung 3.5: Zeitliche Entwicklung typischer Kenngrößen. (a) Negativer Logarithmus („Gesamtenergie“) der Replika-Verbundverteilung. In der Anfangsphase der Simulation konvergiert die Markov-Kette zu ihrer Gleichgewichtsverteilung, was sich in dem Abfall der Gesamtenergie auf einen Plateauwert äußert. (b) Mischverhalten eines Dihedralwinkels (Phe33-χ_1) im konvergierten Teil der Simulation. (c) Marginale *a-posteriori*-Verteilung eines Dihedralwinkels am Beispiel von χ_1.

Algorithmus betrug 250 Schritte. Alle 50 Kopien wurden parallel für 4850 Schritte simuliert, wodurch 4850 *a-posteriori*-Stichproben der Hypothesenparameter $(\mathbf{x}, \sigma^2, \gamma)$ erzeugt wurden. Als Anfangsbedingungen wählte ich jeweils eine vollständig ausgedehnte Konformation sowie die Werte $\sigma = \gamma = 1$. Mit diesen Einstellungen ergaben sich homogene mittlere Akzeptanzraten für einen Replika-Austausch von etwa 60%.

Konvergenz der Markov-Kette

Die Markov-Kette startet aus dem Nichtgleichgewicht und konvergiert während der *Konvergenzphase* der Simulation zu ihrer stationären Verteilung (die *a-posteriori*-Verteilung). Um systematischen Fehlern bei der Berechnung von statistischen Schätzern vorzubeugen, ist es wichtig zu entscheiden, ab welchem Schritt die Zustände von der Gleichgewichtsverteilung gezogen werden. Bislang existieren keine hinreichenden Kriterien für die Feststellung des globalen Konvergenzpunktes von Markov-Ketten. Jedoch hat sich die Analyse der zeitlichen Entwicklung ausgewählter Größen als Funktion der Modellparameter in dieser Hinsicht als ausreichend genau herausgestellt [71].
Abbildung 3.5(a) zeigt den Verlauf der „Gesamtenergie“ der Replika-Simula-

tion, definiert als negativer Logarithmus der Replika-Verbundverteilung. Anhand des Verlaufs der Gesamtenergie setzte ich das Ende der Konvergenzphase konservativ auf Schritt 2000 fest: Innerhalb dieser Spanne fällt die Energie auf ihren Gleichgewichtwert ab und verweilt auf dem Plateau für die restliche Dauer der Simulation. Für die nachfolgenden Analysen wurden alle Stichproben aus der Konvergenzphase verworfen; die Datenbasis bestand somit aus 2850 Konformationen. Für den konvergierten Teil der Simulation ist in Abbildung 3.5(b,c) exemplarisch der zeitliche Verlauf des Seitenketten-Dihedralwinkels χ_1 von Phe33 sowie die korrespondierende marginale *a-posteriori*-Verteilung dargestellt. Für diese Strukturgröße ist die Markov-Kette ebenfalls zu ihrer Gleichgewichtsverteilung konvergiert und zeigt ein gutes Mischverhalten. Die Analyse der übrigen Dihedralwinkel ergab ein vergleichbares Resultat.

Verhalten der Replika-Simulation

Abbildung 3.6 zeigt die Strukturverteilungen für Kopien mit unterschiedlichen (λ, q)-Parameterwerten, dargestellt als Ramachandran Diagramme. Kopien der λ-Kette befinden sich in der oberen, Kopien der q-Kette in der unteren Hälfte der Abbildung. Die *a-posteriori*-Verbundverteilung ist die Zielverteilung und befindet sich in der oberen Reihe ganz links ($q = \lambda = 1$). Die korrespondierende ϕ/ψ Verteilung ist in den für gefaltete Proteine typischen Regionen des Ramachandran-Diagramms populiert.
Durch Absenken von λ in der ersten Hälfte der Replika-Kette wird der kompaktierende Einfluß der Daten reduziert, was sich in Strukturverteilungen mit wachsender Breite äußert (obere Reihe, mitte und rechts). Dateninduzierte Moden werden dadurch abgesenkt, wodurch sich die Beweglichkeit des Systems erhöht (vgl. Abb. 3.7(a,c)). Bei $\lambda = 0.25$ (oben rechts) ist der Einfluß der *Likelihood*-Funktion stark reduziert, und die Strukturverteilung entspricht näherungsweise der eines Boltzmann-Ensembles.
Die Erhöhung des Tsallis-Parameters q in der zweiten Hälfte der Replika-Kette führt zu einer graduellen „Aufheizung“ des Systems (untere Reihe, von

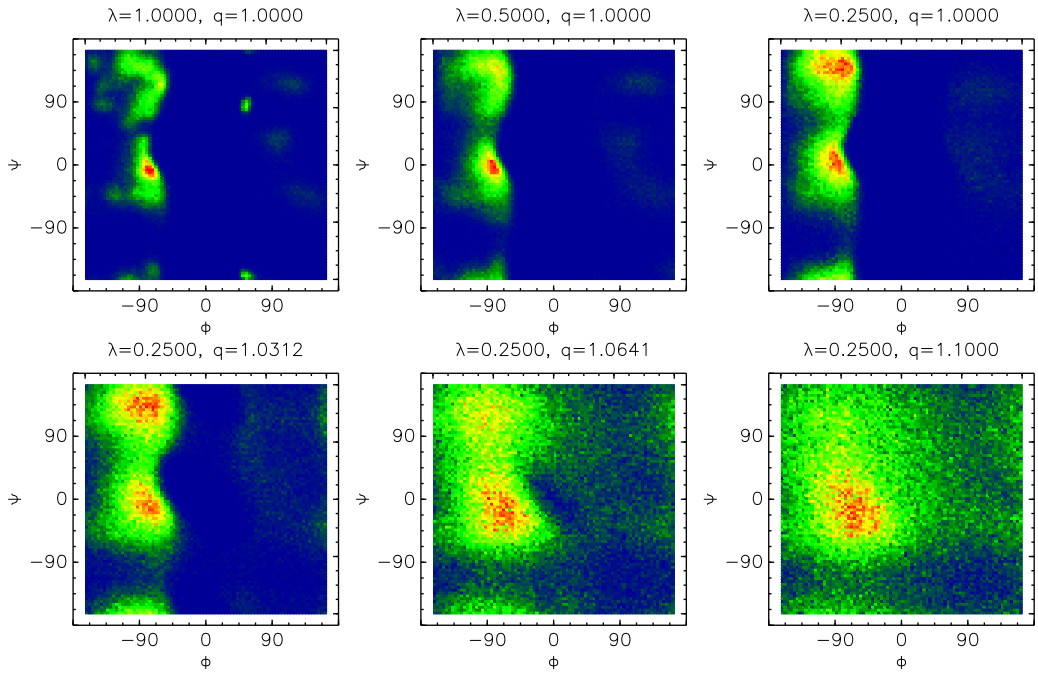

Abbildung 3.6: Strukturverteilungen bei unterschiedlichen (λ, q)-Einstellungen. Hohe (niedrige) Wahrscheinlichkeiten sind rot (blau) eingefärbt. Obere Reihe (λ-Kette): Zielverteilung ist die *a-posteriori*-Verteilung (links). Absenken von λ bedeutet eine schwächere Gewichung der Daten; die Kompaktierung der Strukturen ist schwächer, was sich in einer breiteren Strukturverteilung äußert. Untere Reihe (q-Kette): Erhöhung von q reduziert die Stärke physikalischer Wechselwirkungen. Die „Hochtemperatur"-Verteilung (rechts) zeigt wenig Struktur.

links nach rechts). Moden, die durch van der Waals-Wechselwirkungen der Atome entstehen, werden reduziert (vgl. Abb. 3.7(b)), was sich in einer weiteren Verbreiterung der Verteilungen widerspiegelt. Die Strukturverteilung der „Hochtemperatur"-Kopie ($q = 1.1$, unten rechts) zeigt keine nennenswerten Multimodalitäten, wodurch die Ergodizität der Markov-Kette gesichert wird.

Bedingt durch die Unvollständigkeit von Daten und Hintergrundwissen werden σ^2 und γ nicht exakt festgelegt, sondern weisen eine endliche Unsicherheitsbehaftung auf (vgl. Abb. 3.8). Die relativen Unsicherheiten beider

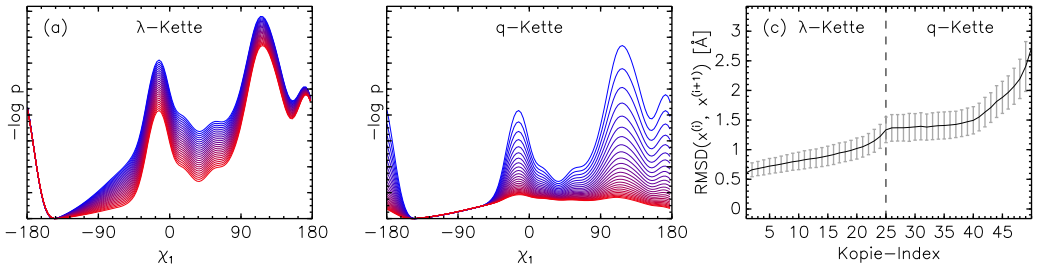

Abbildung 3.7: Energieprofile und Simulationseffizienz. (a,b) Energiebarrieren werden durch Reduktion von λ bzw. Erhöhung von q schrittweise reduziert (gezeigt am Beispiel von Asn24-χ_1). Der Wert von λ (q) nimmt von blau nach rot ab (zu). (c) CA-RMSD aufeinanderfolgender konformationeller Stichproben $\mathbf{x}^{(i)}$ bzw. $\mathbf{x}^{(i+1)}$. Die Effizienz der Simulation, d.h. die Beweglichkeit der einzelnen Systeme, wächst mit steigender Stärke der Deformation.

Größen sind jedoch gering und betragen 4.1% bzw. 3.6%. Die wahrscheinlichste der 2850 erzeugten Konformationen besitzt einen CA-RMSD von 0.85 Å zur Referenzstruktur `Bref` (vgl. Abb. 3.9(a)).

Simulation der SMN Tudor Domäne

Die Simulation der Strukturverteilung erfolgte auf Basis der experimentellen ^{13}C und ^{15}N Datensätze, welche simultan in der Rechnung verwendet wurden. Die Anfangsbedingung der Simulation sowie die Einstellungen des Replika-Algorithmus waren identisch mit der Werten der BPTI Rechnung. Um Austauschraten befriedigender Größe zu erzielen, mußten die (λ, q)-Parameterwerte angepaßt werden. Alle 50 Kopien wurden für 4000 Schritte parallel simuliert. Die Festsetzung des Konvergenzpunktes der Simulation erfolgte nach den oben beschriebenen Kriterien. Die Länge der Konvergenzphase betrug 1500 Schritte. Die wahrscheinlichste unter den 2500 gezogenen Konformationen besitzt einen CA-RMSD von 0.81 Å zur Referenzstruktur `Tref` (vgl. Abb. 3.9(b)).

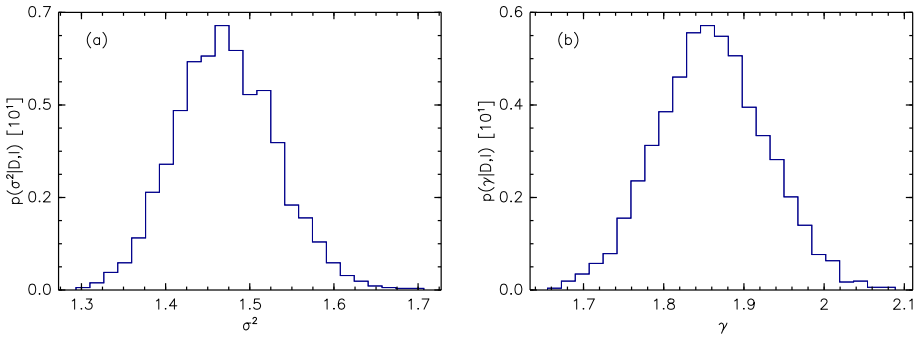

Abbildung 3.8: *Nuisance*-Parameter der BPTI-Simulation. (a) Datenvarianz σ^2. (b) NOE-Skalenparameter γ. Unsicherheiten sind auf die endliche Größe des Datensatzes zurückzuführen.

3.1.4.3 Berechnung der Koordinatenunsicherheiten

Die Berechnung der Unsicherheiten der kartesischen Koordinaten beider Systeme erfolgte durch Approximation der jeweiligen Strukturverteilungen durch das Verteilungsmodell. Als Datenbasis für die Schätzung des Modells verwendete ich jeweils $N = 500$ Konformationen, die zufällig aus den konformationellen Stichproben ausgewählt wurden. Die Zahl der zu schätzenden Parameter beträgt $4M + 6N$ ($3M$ kartesische Koordinaten für die mittlere Struktur, M Varianzen und $6N$ Parameter für die individuellen Transformationen der Koordinatensysteme). Pro Monte-Carlo-Schritt sind dies 6532 zu schätzende Parameter für BPTI ($M = 883$ Atome) bzw. 6468 Parameter für die Tudor Domäne ($M = 867$ Atome).

Simulation des Verteilungsmodells durch MCMC

Zur Bestimmung der Parameter beider Verteilunsmodelle simulierte ich die *a-posteriori*-Verteilung aus Gl. (3.18) mit Hilfe des Gibbs-Algorithmus aus Abschnitt 3.1.3 für jeweils 400 Monte-Carlo-Schritte. Die benötige Rechenzeit betrug in beiden Fällen weniger als 3 Minuten (Linux PC, 1.2 GHz Athlon Prozessor). Aus dem zeitlichen Verlauf der Gesamtenergie (s. Abb. 3.10) ist ersichtlich, daß beide Markov-Ketten trotz der hohen Anzahl unbe-

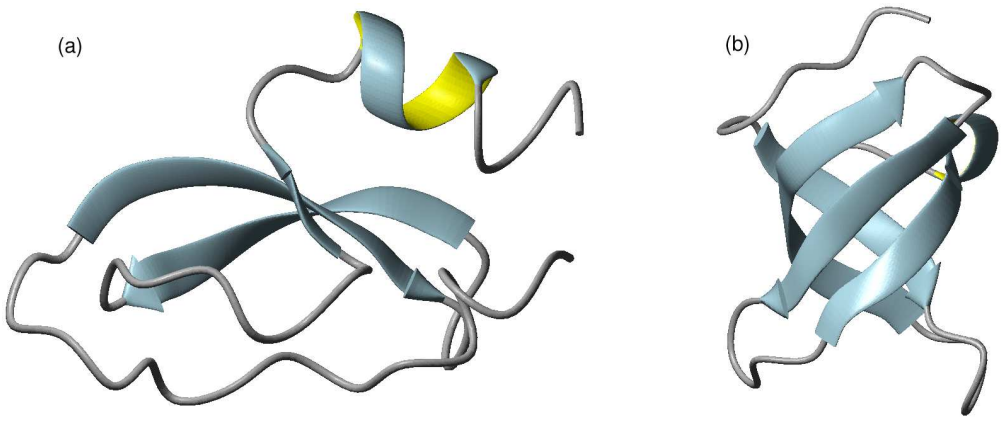

Abbildung 3.9: Wahrscheinlichste Konformationen. (a) BPTI (CA-RMSD zur Referenzstruktur `Bref`: 0.85 Å). (b) SMN Tudor Domäne (CA-RMSD zur Kristallstruktur `Tref`: 0.81 Å).

kannter Parameter rasch zu ihren Gleichgewichtsverteilungen konvergierten: Die *a-posteriori*-Verteilung ist somit sehr gut bestimmt. Im Falle der Tudor Domäne konvergierte die Simulation nahezu instantan nach etwa 10 Schritten (vgl. Abb. 3.10(b)). Ich legte das Ende der Konvergenzphase für beide Rechnungen konservativ auf Schritt 100 fest. Nichtkonvergierte Zustände beider Simulationen wurden von der Berechnung statistischer Größen ausgenommen.

Aus den $K = 300$ konvergierten *a-posterori*-Proben wurden Erwartungswerte und statistische Unsicherheitsbehaftungen aller Hypothesenparameter berechnet. Erwartungswerte und statistische Ungenauigkeiten der Koordinatenunsicherheiten $\{\delta_j\}$ berechnete ich gemäß Gleichung (3.14).

Alle Koordinatenunsicherheiten $\{\delta_j\}$ konnten verläßlich aus den verwendeten Datenbasen geschätzt werden: Die statistische Unsicherheitsbehaftung, welche aus der endlichen Größe der Datenbasis folgt, beträgt maximal 2%. In den folgenden Abbildungen habe ich auf die Darstellung von Fehlerbalken daher verzichtet.

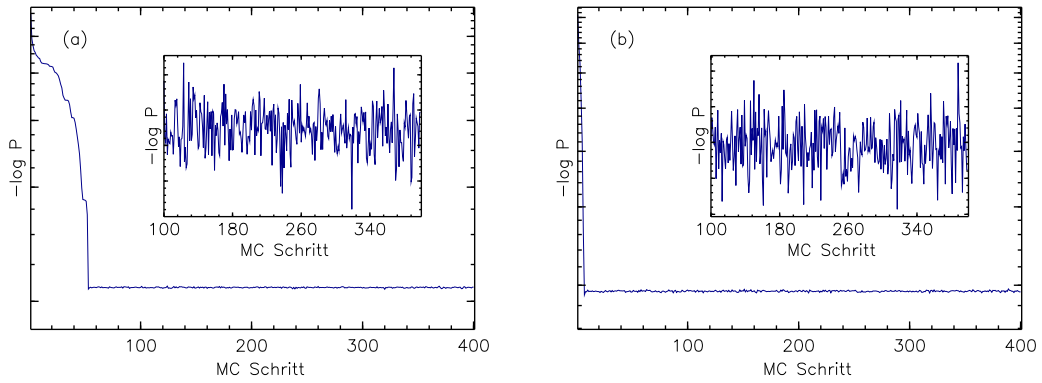

Abbildung 3.10: Verlauf der Gesamtenergie bei der Simulation des Verteilungsmodells. (a) BPTI. (b) Tudor Domäne. Die *Insets* zeigen den Verlauf der Gesamtenergie für den konvergierten Teil (ab Schritt 100) der Simulation.

Abbildungen 3.11(a) und 3.12 zeigen die approximierten Strukturverteilungen von BPTI und der Tudor Domäne in graphischer Darstellung. Die Unsicherheitssphären der Hauptkettenatome C,CA,N werden durch Kreise um die mittlere Position μ_j repräsentiert. Die Größe der Radien entspricht der Unbestimmtheit δ_j in der Position des jeweiligen Atoms. Die übrigen Atome wurden der Übersicht halber nicht dargestellt. Punkte korrespondieren zu Atomen konformationeller Stichproben, auf Basis derer das Verteilungsmodell geschätzt wurde. Die kartesischen Koordinaten der Hauptkettenatome streuen um ihre mittleren Positionen. Die Stärke der Streuung ist positionsabhängig und wird durch die individuellen Unsicherheitssphären mit sehr guter Übereinstimmung beschrieben. Die unstrukturierten Termini der Tudor Domäne sind in Form von Unsicherheitssphären mit signifikanter Größe zu erkennen (vgl. Abb. 3.12).
Abbildung 3.11(b,c) zeigt exemplarisch die Verteilung für alle Atome der Reste Phe45 bzw. Ala48 von BPTI. Die Verteilung von Haupt- und Seitenkettenatomen von Phe45 wird durch das isotrope Modell mit guter Genauigkeit angenähert. Abweichungen des Modells von der Strukturverteilung sind für die HB# Methylgruppe von Ala48 zu erkennen: Die Rotation strukturell uneingeschränkter Methylgruppen führt zu einer torusartigen Verteilung der

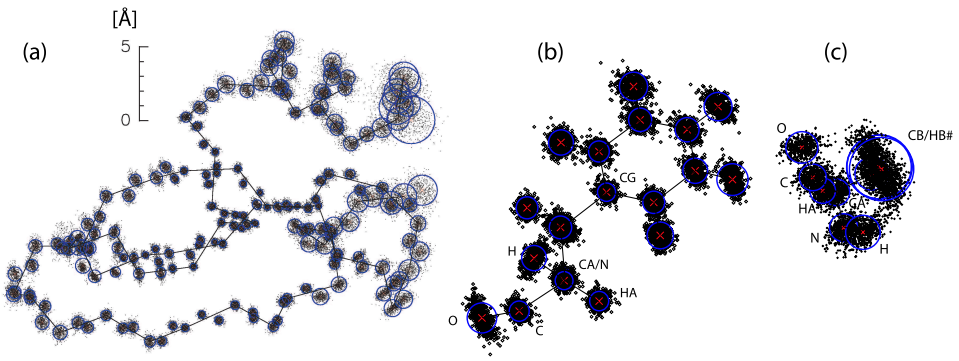

Abbildung 3.11: Graphische Darstellung des Verteilungsmodells für BPTI. Hauptkettenatome C,CA,N werden durch Kreise um ihre mittlere Position repräsentiert. Die Größe der Radien entspricht der Unbestimmtheit in der Position des jeweiligen Atoms. Punkte korrespondieren zu Atomen konformationeller Stichproben. Mittlere Positionen der CA-Atome sind miteinander verbunden.

betroffenen Wasserstoffatome. Die Ausdehnung der Strukturverteilung wird in diesem Fall in zwei Raumrichtungen korrekt wiedergegeben, in der dritten Richtung jedoch überschätzt.

Koordinatenunsicherheiten

Abbildung 3.13(a) zeigt die berechneten 1σ-Koordinatenunsicherheiten aller Atome von BPTI. Hauptkettenatome sind in der linken, Seitenkettenatome in der rechten Hälfe der Abbildung dargestellt. Die 1543 Messungen des Datensatzes legen die Positionen der Atome bis auf eine mittlere Unsicherheit von $\langle\delta\rangle = 0.78 \pm 0.58$ Å fest. Unterschiede bestehen zwischen Hauptketten- und Seitenkettenatomen. In Abbildung 3.14(a) ist die Unsicherheitsbehaftung der CA-Positionen in graphischer Form dargestellt.
Ein ähnliches Bild zeigte sich bei der Tudor Domäne (s. Abb. 3.13(b)). Die mittlere 1σ-Unsicherheitsbehaftung der Atompositionen beträgt $\langle\delta\rangle = 0.80 \pm 0.64$ Å. Auch in diesem Fall sind deutliche Unterschiede in der Unsicherheitsbehaftung von Haupt- und Seitenketten-Atomen zu beobachten.

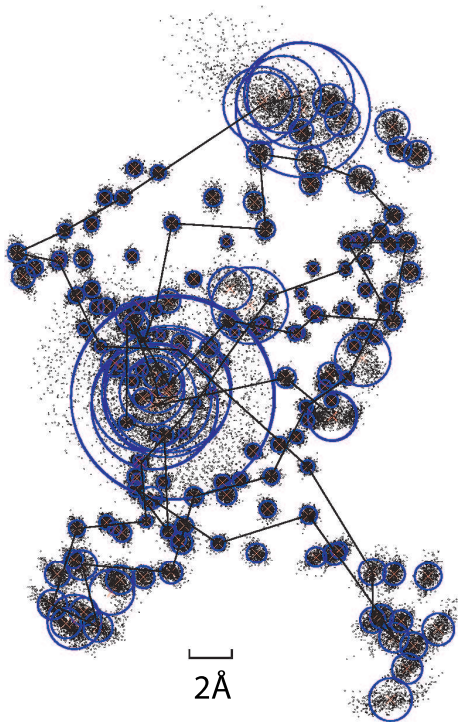

Abbildung 3.12: Graphische Darstellung des Verteilungsmodells für die SMN Tudor Domäne. Hauptkettenatome C,CA,N werden durch Kreise um ihre mittlere Position repräsentiert. Die Größe der Radien entspricht der Unbestimmtheit in der Position des jeweiligen Atoms. Punkte korrespondieren zu Atomen konformationeller Stichproben. Mittlere Positionen der CA-Atome sind miteinander verbunden.

Für beide Proteine sind die mittlere Unsicherheitsbehaftung der kartesischen Koordinaten und der Abstand zu der jeweiligen Referenzstruktur (gemessen anhand des RMSD) somit von vergleichbarer Größe. Die hohe Unsicherheitsbehaftung der unstrukturierten Termini ist deutlich sichtbar (vgl. Abb. 3.14).

Die Abweichung der berechneten Koordinaten von einem Referenzkoordinatensatz läßt sich über verschiedene Maße definieren, beispielsweise über den RMSD. Aufgrund der Unsicherheitsbehaftung von Atompositionen ist diese Abweichung jedoch nicht exakt bestimmt, sondern weist selbst eine gewisse Unsicherheit auf. In der induktiven Strukturbestimmung erfolgt der Vergleich

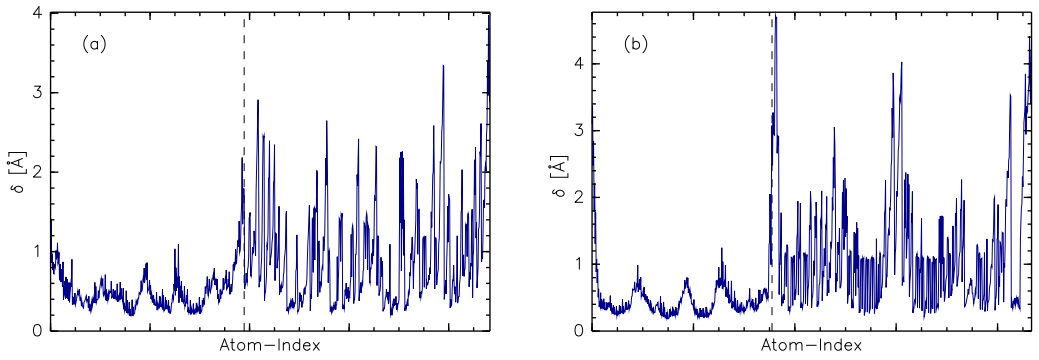

Abbildung 3.13: 1σ-Unsicherheitsbehaftung von Atompositionen. Hauptkettenatome befinden sich in der jeweils linken, Seitenkettenatome in der rechten Hälfte einer Abbildung. (a) BPTI. Hauptkettenatome sind bis auf $\langle\delta\rangle = 0.51 \pm 0.27$ Å, Seitenkettenatome bis auf $\langle\delta\rangle = 1.0 \pm 0.66$ Å bestimmt. Für 90% der Seitenkettenatome ist die Unsicherheitsbehaftung kleiner als 2.0 Å. (b) Tudor Domäne. Hauptkettenatom sind im Mittel bis auf 0.50 ± 0.41 Å, Seitenkettenatome auf $\langle\delta\rangle = 1.07 \pm 0.71$ Å festgelegt.

der Strukturverteilung mit einer Referenzstruktur daher grundsätzlich auf Basis einer Wahrscheinlichkeitsverteilung für das jeweilige Ähnlichkeitsmaß. Abbildung 3.15 zeigt die Wahrscheinlichkeitsverteilungen für den CA-RMSD bezüglich den Referenzstrukturen `Bref` bzw. `Tref`, die aus den Strukturverteilungen von BPTI und der SMN Tudor Domäne berechnet wurden. Die Abweichung der wahrscheinlichsten Konformation von der Referenzstruktur ist in beiden Fällen gering (0.85 Å bzw. 0.81 Å). Die Strukturverteilung suggeriert jedoch eine größere Abweichung: Im Falle von BPTI beträgt die Wahrscheinlichkeit, daß die Struktur von der Referenzstruktur stärker abweicht als die wahrscheinlichste Konformation:

$$P(\mathrm{RMSD} > 0.85|D,I) = \int_0^\infty dr\, p(r|D,I)\, \theta(r-0.85) \approx 0.89.$$

$p(r|D,I)$ steht symbolisch für die Wahrscheinlichkeitsverteilung für den RMSD, $\theta(\cdot)$ bezeichnet die Heaviside-Funktion. Mit anderen Worten wird die Hypothese, daß die Abweichung der berechneten Struktur von der Referenzstruktur weniger als 0.85 Å beträgt, nur mit einer Wahrscheinlichkeit

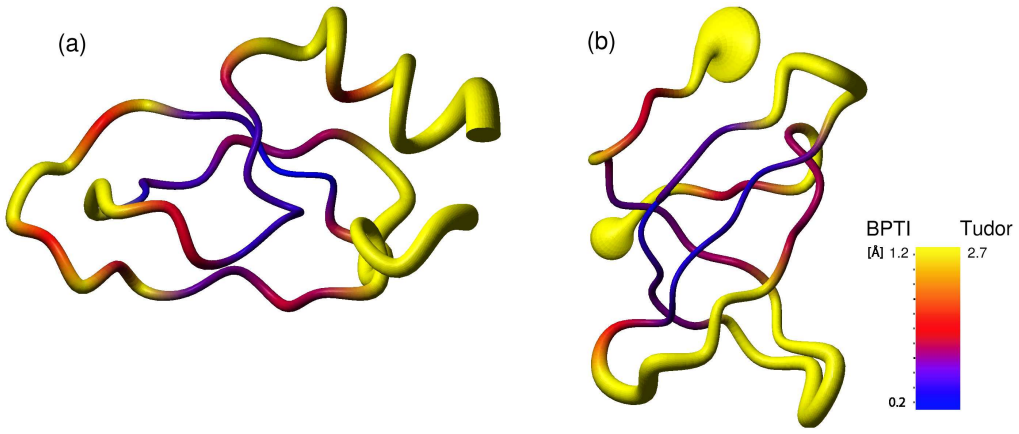

Abbildung 3.14: Unsicherheitsbehaftung in den CA-Positionen. (a) BPTI. (b) Tudor Domäne. Stärke der Unsicherheitsbehaftung ist in Farbe und Strichstärke kodiert.

von 0.11 unterstützt. Für die Tudor Domäne beträgt die Wahrscheinlichkeit 0.21. Die Möglickeit, daß die berechnete Struktur tatsächlich nah an der Referenzstruktur liegt, wird somit nicht ausgeschlossen; die Wahrscheinlichkeit dafür ist jedoch gering.

Die Bewertung struktureller Ähnlichkeit auf Basis einer wahrscheinlichsten Konformation (oder einer Struktur minimaler Energie) ist daher nicht sehr aussagekräftig und kann eine Unterschätzung der strukturellen Abweichung bedeuten. Der wahrscheinlichkeitstheoretische Zugang ist vorurteilsfreier, da die Unsicherheitsbehaftung der dreidimensionalen Koordinaten einer Struktur bei der Angabe eines Ähnlichkeitsmaßes berücksichtigt wird: Im Falle von BPTI beträgt das 68%-Konfidenzintervall der RMSD-Verteilung $[0.86, 1.0]$ Å. Die Abweichung der berechneten BPTI Struktur zur Referenzstruktur liegt demzufolge mit einer Wahrscheinlichkeit von 0.68 zwischen 0.86 Å und 1.0 Å (vgl. Abb. 3.15(a)). Strukturelle Unsicherheiten führen somit zu einer Unsicherheit im Wert des Ähnlichkeitsmaßes von etwa 15% Å. Im Falle der Tudor Domäne beträgt das 68%-Konfidenzintervall $[0.79, 0.95]$ Å (vgl. Abb. 3.15(b)).

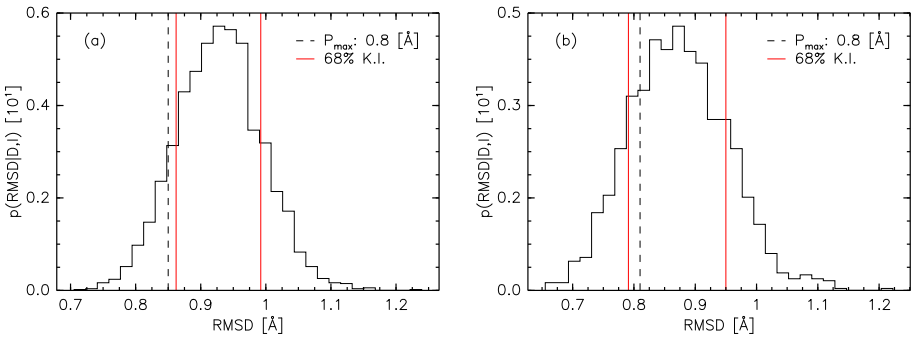

Abbildung 3.15: Wahrscheinlichkeitsverteilungen des CA-RMSD zu Referenzstrukturen. (a) BPTI; Referenzstruktur: `Bref`. (b) Tudor Domäne; Referenzstruktur: `Tref`. Wahrscheinlichste Konformationen ($P_{\max}$) überschätzen strukturelle Ähnlichkeit. Die Angabe eines Konfidenzintervalls (rot) berücksichtigt atomare Unsicherheiten.

3.2 Qualität von NOE-Datensätzen

Die Qualität eines NOE-Datensatzes ist neben der Vollständigkeit der Messungen ein weiterer Faktor, welcher zu Unsicherheiten in den berechneten Strukturen führen kann. Neben experimentellen Ungenauigkeiten und Fehlern bei der Interpretation der Daten wird die Qualität eines Datensatzes maßgeblich durch Näherungen in den theoretischen Modellen zur Beschreibung der experimentellen Strukturgrößen begrenzt: Beiträge zu NOE-Intensitäten durch Spindiffusion oder interne Dynamik bleiben in der ISPA beispielsweise unberücksichtigt. Dies kann zu systematischen Abweichungen von berechneten und observierten Kreuzrelaxationsraten führen, wodurch realistische Datensätze mit der Hypothese eines starren Moleküls in der Regel unverträglich und in diesem Sinne *inkonsistent* sind.

Standardverfahren bewerten die Konsistenz von NOE Datensätzen typischerweise anhand des Vergleichs von berechneten und observierten Strukturgrößen („R-Faktoren“) und basieren oft auf der Methode der Kreuzvalidierung. Für die Berechnung eines freien R-Faktors durch Kreuzvalidierung wird der Datensatz in einen Arbeits- und einen Testdatensatz zerlegt. Der

Arbeitsdatensatz wird für die Strukturrechnung, der Testdatensatz für die Validierung der vorhergesagten Daten verwendet. Arbeits- und Testdatensatz werden dabei so erzeugt, daß jede Messung genau einmal im Testdatensatz vorkommt. Der freie R-Faktor berechnet sich als Mittelwert über die individuellen R-Faktoren. In der Praxis gestaltet sich die Bestimmung von kreuzvalidierten Maßen aufwendig, da die Strukturrechnung mehrfach durchgeführt werden muß. Darüber hinaus kann die Zerlegung in Arbeits- und Testdatensatz bei kleinen Datensätzen zu Konvergenzproblemen führen, wodurch die Methode instabil und nicht universell einsetzbar ist.

Ich stelle in den folgenden Abschnitten zwei statistische Maße für die Bewertung der Konsistenz experimenteller NOE-Datensätze vor. Beide Größen folgen unmittelbar aus der wahrscheinlichkeitstheoretischen Beschreibung dipolarer Kreuzrelaxationsraten durch ein Datenmodell. Das in Abschnitt 3.2.1 diskutierte Maß dient der Beurteilung der Konsistenz des Gesamtdatensatzes, die in Abschnitt 3.2.2 hergeleitete Größe bewertet die Konsistenz jeder Einzelmessung in Hinblick auf die Gesamtheit aller Messungen. In Abschnitt 3.2.3 diskutiere ich die Eigenschaften der beiden Maße anhand der Datensätze für BPTI und die SMN Tudor Domäne.

3.2.1 A-posteriori-Bewertung der Daten

Das Datenmodell für dipolare Kreuzrelaxationsraten aus Gleichung (3.3) bewertet die Größe der Abweichung der observierten von der berechneten Kreuzrelaxationsraten einer Einzelmessung anhand des Hypothesenparameters σ. σ repräsentiert die Breite der Fehlerverteilung und wurde *a priori* als unbekannt angenommen. Die Schätzung unbekannter Größen erfolgt durch Bestimmung der *a-posteriori*-Verteilung: Die marginale *a-posteriori*-Verteilung für σ, $p(\sigma|D, I)$, drückt unser Wissen über die Diskrepanz von beobachteter und berechneter Intensität einer Einzelmessung aus, wenn der Gesamtdatensatz und das Hintergrundwissen gemeinsam berücksichtigt werden. σ stellt daher ein Maß für die Verträglichkeit der Messungen untereinan-

der und die Schlüssigkeit von Daten und Hintergrundwissen dar. Die Bewertung der Konsistenz eines Datensatzes erfolgt demzufolge niemals absolut, sondern stets in Bezug auf relevantes Hintergrundwissen – ausgedrückt durch die Bedingung von σ durch I: Die Schlüssigkeit der Meßwerte wird beispielsweise davon bestimmt, wie die experimentellen Daten interpretiert werden; σ hängt demnach von der verwendeten Theorie ab. Einen weiteren Faktor bildet physikalisches *a-priori*-Wissen: In die Bewertung fließt stets Wissen ein, ob eine Messung *a priori* zu erwarten ist, z.B. ob sie strukturell erfüllbar ist oder nicht. Mit anderen Worten läßt sich grundsätzlich keine „intrinsische" oder absolute Konsistenz eines Datensatzes definieren. Eine Bewertung basiert stets auf relevantem Hintergrundwissen.
Aus der bedingten *a-posteriori*-Verteilung für σ^2 in Gleichung (3.28) folgt für den bedingten Erwartungswert von σ im Fall $N \gg 1$:

$$\langle\sigma\rangle|\cdot \approx \left\{\frac{1}{N}\sum_{i=1}^{N}\log^2\left(\tilde{V}_i/\gamma d_i^{-6}(\mathbf{x})\right)\right\}^{1/2}.$$

Die relative Unsicherheitsbehaftung von σ,

$$\frac{\sqrt{\mathrm{Var}(\sigma)}}{\langle\sigma\rangle}|\cdot \approx \frac{1}{N}, \tag{3.31}$$

ist von den Messungen hingegen unabhängig und wird durch die Größe des Datensatzes bestimmt. Der Grad der Erfüllbarkeit eines Datensatzes spiegelt sich somit unmittelbar in der Größe von σ wider: Steht der Gesamtdatensatz mit der Hypothese einer starren Struktur in Einklang, so ist die mittlere erwartete Abweichung von beobachteten und berechneten Intensitäten gering, was sich in einem kleinen Wert von σ äußerte. Umgekehrt führen Inkonsistenzen in den Daten zu nichtverschwindenden Diskrepanzen der Intensitäten, was einen endlichen Wert von σ verursacht.

Transformation des Datenmodells

σ quantifiziert die Standardabweichung der Streuung der logarithmierten Intensitäten – Zahlenwerte sind daher nur schlecht intuitiv interpretierbar. Es

ist zweckmäßig, den Formparameter von der Skala für Intensitäten auf die natürlichere Distanzskala zu überführen. Dazu transformiere ich das Datenmodell für Intensitäten mit Hilfe der Variablentransformation $\tilde{V} \to \gamma \tilde{d}^{-6}$ auf das korrespondierende Datenmodell für Distanzen. Aus Gleichung (3.3) folgt unter Beachtung des Vorwärtsmodells aus Gleichung (3.1) als Datenmodell für die „observierte“ Zieldistanz $\tilde{d}$:

$$p_{\mathrm{NOE}}(\tilde{d}|\mathbf{x}, \sigma_d^2, D, I) \quad = \quad \frac{1}{\sqrt{2\pi\sigma_d^2}}\, \tilde{d}^{-1} \exp\left\{ -\frac{1}{2\sigma_d^2} \log^2\left(\frac{\tilde{d}}{d(\mathbf{x})}\right)\right\},$$

wobei

$$\sigma_d = \sigma/6. \tag{3.32}$$

Nach Gleichung (3.32) ist die Zieldistanz ebenfalls lognormalverteilt mit Ortsparameter $d(\mathbf{x})$ und Formparameter σ_d. Für $\sigma_{\mathrm{d}} \ll 1$ folgt aus den Eigenschaften der Lognormalverteilung für die relative quadratische Abweichung der observierten von der berechneten Distanz:

$$\left\langle \left(\frac{\tilde{d} - d(\mathbf{x})}{d(\mathbf{x})}\right)^2 \right\rangle \approx \sigma_d^2.$$

Für $\sigma_d \leq 0.6$ beträgt der Fehler dieser Näherung weniger als 10%. Für kleine Werte quantifiziert σ_d somit die relative Abweichung der Zieldistanz von der korrespondierenden Distanz in der Struktur und ist damit intuitiv interpretierbar. Wie die Testrechnungen demonstrieren werden, ist die Bedingung $\sigma_d \ll 1$ in praktischen Fällen immer erfüllt.

Im Gegensatz zu kreuzvalidierten Konsistenzmaßen erfordert die Bestimmung von σ_d keinen zusätzlichen Rechenaufwand: Die Schätzung des Hypothesenparameters σ ist notwendiger Teil der Strategie zur Simulation der Strukturverteilung (s. Gibbs-Schema in Kapitel 3.1.4.1). Aus den *a-posteriori*-Stichproben für σ folgt σ_d samt Fehlerabschätzung umittelbar aus Gl. (3.32).

3.2.2 Konsistenz von Einzelmessungen

σ_d quantifiziert die Konsistenz eines Datensatzes in Hinsicht auf das Modell zur Berechnung der theoretischen Intensitäten sowie die berücksichtigte

Hintergrundinformation, gestattet jedoch keine Aussage über die Konsistenz einer Einzelmessung mit der Gesamtheit aller Daten. Ich diskutiere die Konsistenz einer Einzelmessung anhand der Vorhersage derselben Messung durch das Datenmodell, welche unter Berücksichtigung der Gesamtheit aller Daten und dem gegebenem Hintergrundwissen erfolgt.
Im konkreten Fall bedeutet dies: Welche Intensität eines NOE würde man bei Berücksichtigung des gesamten Datensatzes und der gegebenen Hintergrundinformation nach Durchführung eines neuen Experiments erwarten? Ein statistischer Vergleich der vorhergesagten mit der bereits gemessenen Intensität erlaubt die *a-posteriori*-Bewertung von Einzelmessungen und damit die Identifikation von Meßwerten, die nicht dem Trend des Datensatzes folgen und in diesem Sinne mit der Gesamtheit aller Messungen inkonsistent sind.

Vorhersage experimenteller Daten

Aus dem Datenmodell für dipolare Kreuzrelaxationsraten folgt durch Marginalisierung aller Hypothesenparameter die *a-posteriori*-Verteilung für eine Intensität $\tilde{V}_{N+1}$:

$$p_{\mathrm{pred}}(\tilde{V}_{N+1}|D,I) = \int d\sigma^2\, d\gamma\, d^{3M}\mathbf{x}\, p_{\mathrm{NOE}}(\tilde{V}_{N+1}|\mathbf{x},\sigma^2,\gamma) p(\mathbf{x},\sigma^2,\gamma|D,I). \tag{3.33}$$

Die Verteilung in Gleichung (3.33) drückt aus, was man für die Intensität $\tilde{V}_{N+1}$ bei erneuter Durchführung des Experiments erwartet, wenn ein zuvor gemessener Datensatz $D = \{\tilde{V}_1, \ldots, \tilde{V}_N\}$ und die Hintergrundinformation I als bekannt vorausgesetzt werden. Die obige *a-posteriori*-Verteilung repräsentiert also die *Vorhersage* einer Einzelmessung und wird daher allgemein als *Vorhersageverteilung*, hier für die Intensität $\tilde{V}_{N+1}$, bezeichnet. Die Bedingung der Vorhersageverteilung durch D und I bringt dabei zum Ausdruck, daß das Unwissen über den möglichen Versuchsausgang durch die Gesamtheit aller Daten und die Hintergrundinformation bestimmt wird. Vorhersageverteilungen werden stets durch Integration über alle Hypothesenparameter gebildet, wobei die *a-posteriori*-Verbundverteilung als Gewichtungsfunktion fungiert.

In Gl. (3.33) wird auf diese Weise die Unsicherheitsbehaftung der Koordinaten, des NOE-Skalenfaktors sowie der Varianz der Daten vollständig bei der Vorhersage berücksichtigt: Strukturelle Mehrdeutigkeiten würden sich in multimodalen Verteilungen äußern, Unsicherheiten in den Hypothesenparametern erhöhen die Breite einer Vorhersageverteilung und führen somit zu konservativeren Vorhersagen.

Bewertung von Einzelmessungen

Die Frage *"was ist eine wahrscheinliche Messung?"* läßt sich naturgemäß nur in Bezug auf den erwarteten Wert der Messung beantworten. Ist der erwartete Wert, bzw. allgemein die Verteilung der erwarteten Werte der Messung bekannt, läßt sich die Frage präziser stellen: In welchem Intervall B wird der Meßwert bei einer vorgegebenen Sicherheit K erwartet? Liegt der Meßwert innerhalb dieses Intervalls, so entspricht er mit der Sicherheit (*Konfidenz*) K der Erwartung und ist in diesem Sinne korrekt. Liegt er jedoch außerhalb des Intervalls, so ist die Messung mit Konfidenz K falsch.
Die eingangs formulierte Frage nach der Richtigkeit einer Messung wird durch die Berechnung des *Konfidenzintervalls* B_K der Vorhersageverteilung statistisch sauber beantwortet. Das Konfidenzintervall ist definiert als das kleinste aller Intervalle B, welche den Anteil K der Wahrscheinlichkeitsmaße enthalten, für die also gilt:

$$K = \int_B d\tilde{V}\, p_{\text{pred}}(\tilde{V}|D, I). \tag{3.34}$$

Ich definiere einen Meßwert als *korrekt mit Konfidenz* K, wenn er innerhalb des Konfidenzintervalls B_K seiner Vorhersageverteilung liegt. Liegt beispielsweise eine gemessene Intensität $\tilde{V}_i$ außerhalb des 80% Konfidenzintervalls, so ist die Messung mit 80% Wahrscheinlichkeit falsch. Diese Wahrscheinlichkeit ist dabei nicht absolut, sondern immer in Bezug auf die übrigen Daten zu interpretieren. Die Größe von Konfidenzintervallen mit gleicher Konfidenz wird nach Gl. (3.34) durch die Breite der Vorhersageverteilung bestimmt und hängt somit unter anderem von der Unsicherheitsbehaftung der Strukturkoordinaten ab. Ein großes Konfidenzintervall impliziert daher nicht, daß eine

Messung schlecht war, sondern besagt, daß der Meßwert anhand der gegebenen Information nicht präziser festgelegt werden kann. Unsicherheiten in den Hypothesenparametern werden somit in konsistenter Weise berücksichtigt, indem bei schlecht bestimmten Meßwerten entsprechend höhere Abweichungen toleriert werden. Der Versuch, inkorrekte Messungen über den Vergleich mit ihrem Soll-Wert zu identifizieren, ist hingegen vorurteilsbehaftet, da hier eine identische Unsicherheitsbehaftung aller Messungen angenommen wird.

Für die Konfidenz K_i, daß die Messung $\tilde{V}_i$ mit den übrigen Messungen in Einklang steht, folgt aus der obigen Definition:

$$K_i = \min\{K|\tilde{V}_i \in B_K\}. \tag{3.35}$$

K_i ist also der Konfidenzwert des kleinsten Konfidenzintervalls, welches den Meßwert gerade enthält. Ein Konfidenzwert von $K_i = 1$ besagt, daß die Messung mit der Vorhersage perfekt in Einklang steht und in diesem Sinne mit dem Gesamtdatensatz konsistent ist; Messungen mit $K_i = 0$ folgen nicht dem Trend der Daten und sind in Bezug auf das Datenmodell, die Hintergrundinformation sowie die übrigen Messungen inkorrekt. Die Konfidenzwerte $\{K_i\}$ gestatten damit die Bewertung jeder Einzelmessung eines Datensatzes auf einer absoluten Skala.

3.2.3 Eigenschaften der Qualitätsmaße

Ich demonstriere die Eigenschaften des Konsistenzmaßes anhand des Modelldatensatzes für BPTI sowie der beiden experimentellen Datensätze für die Tudor Domäne. Dazu wurden mehrere Testdatensätze unterschiedlicher Größe und Konsistenz erzeugt.

Datensätze unterschiedlicher Größe Aus dem Datensatz für BPTI wurden 5 Datensätze B_X generiert, indem $X = \{20, 40, 50, 70, 100\}$ % der Daten auf zufälliger Basis ausgewählt wurden. Analog wurden aus den ^{13}C und

^{15}N Datensätzen der Tudor Domäne jeweils 6 Datensätze T_X^{13} bzw. T_X^{15} mit $X = \{20, 40, 60, 70, 80, 100\}$ % der Daten erzeugt.

Datensätze unterschiedlicher Konsistenz Ausgangspunkt bildete jeweils ein vollständig konsistenter Datensatz. Diesen generierte ich aus den Originaldaten, indem ich die Intensitäten aller Messungen durch Werte ersetzte, welche anhand einer Referenzstruktur berechnet wurden. Die Messungen des Ausgangsdatensatzes sind somit simultan anhand einer einzelnen Struktur erklärbar. Um dem Datensatz zu stören, wurde der Anteil f aller Intensitäten gegen die korrespondierenden Werte des Originaldatensatzes ausgetauscht; die Auswahl der Messungen erfolgte auf zufälliger Basis. Der Originaldatensatz entspricht folglich $f = 1$, der konsistente Datensatz $f = 0$. Um die Intensitätsskalen anzugleichen, rekalibrierte ich den Originaldatensatz bezüglich des konsistenten Datensatzes.
Im Falle von BPTI verwendete ich die mittlere Struktur `Bref` (vgl. Kap. 2.5), für die Tudor Domäne die publizierte NMR-Struktur (PDB-Zugriffsnummer 1g5v[1]) als Referenzstruktur. Aus dem Datensatz für BPTI und den experimentellen Datensätzen für die Tudor Domäne wurden jeweils 6 Datensätze BF_f bzw. TF_f^{13} und TF_f^{15} mit $f = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ erzeugt.

Ich simulierte die *a-posteriori*-Verteilungen beider Testsysteme auf Basis der Datensätze B_N, T_N, BF_f und TF_f durch Replika-Austausch-MC; 19 Simulationen insgesamt. Im Falle der Tudor Domäne wurden die Testdatensätze T_N^{13} und T_N^{15} bzw. TF_f^{13} und TF_f^{15} jeweils simultan in der Rechnung verwendet.

Konsistenz und Konfidenzen

Ich bestimmte die Konsistenz der (ungestörten) Originaldatensätze für BPTI und die Tudor Domäne, B_{100} bzw. T_{100}^{13},T_{100}^{15} (s. Abb. 3.16). Im Falle von BPTI beträgt die mittlere relative Abweichung von observierten und berechneten

[1]Aus den 10 angegeben Modellen wählte ich die Konformation mit geringstem CA-RMSD zur Kristallstruktur 1mhn.

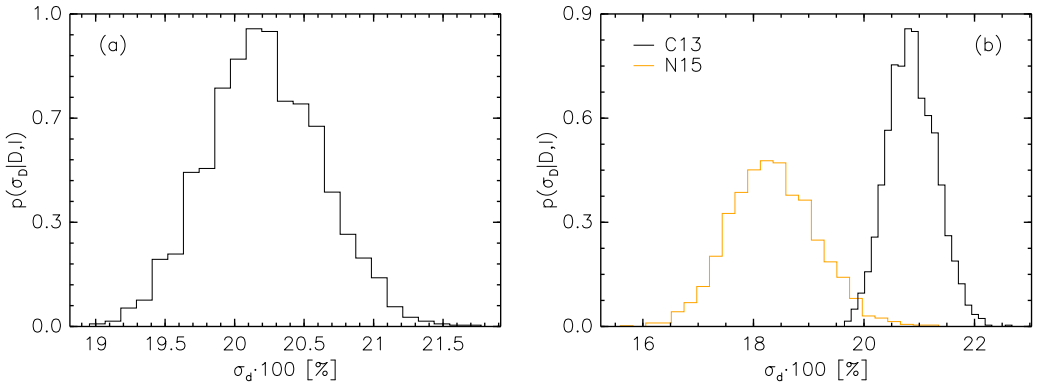

Abbildung 3.16: Konsistenz der Datensätze für BPTI und die SMN Tudor Domäne. (a) Marginale *a-posteriori*-Verteilung für σ_d des Datensatzes B_{100}. (b) Dito für die Datensätze T^{13}_{100} und T^{15}_{100}.

Distanzen $\langle\sigma_d\rangle = 20.2 \pm 0.4\%$ (Abb. 3.16(a)). Für die Tudor Domäne ergaben sich vergleichbare Werte: Die mittlere Abweichung liegt für den ^{13}C Datensatz bei $20.9 \pm 0.4\%$, der ^{15}N Datensatz läßt sich anhand einer starren Struktur mit einer Abweichung von $18.4 \pm 0.8\%$ besser erklären (Abb. 3.16(b)). Unterschiede in der Unsicherheit der geschätzten Größen sind nach Gleichung (3.31) auf die differierende Größe der Datensätze zurückzuführen.

Um die Konsistenz jeder einzelnen der 1543 simulierten Kreuzrelaxationsraten des BPTI Datensatzes zu bestimmen, berechnete ich die Vorhersageverteilungen aller Messungen. Die Berechnung des Marginalisierungsintegrals in Gl. (3.33) erfolgte auf numerischem Wege durch Monte-Carlo-Integration. Abbildung 3.17 zeigt typische Vorhersageverteilungen in Rot, korrespondierende Häufigkeitsverteilungen wurden anhand der konformationellen Stichproben berechnet und sind grau dargestellt. Inkonsistenzen in den Daten führen stets zu endlichen Werten von σ^2, was sich unmittelbar auf die Form der Vorhersageverteilungen auswirkt: Verglichen mit Häufigkeitsverteilungen sind Vorhersageverteilungen unspezifischer und somit konservativer (vgl. Abb. 3.17(a)). Eine Vorhersage von NOE-Intensitäten auf Basis der konformationellen Stichproben ist daher grundsätzlich vorurteilsbehaftet: Die

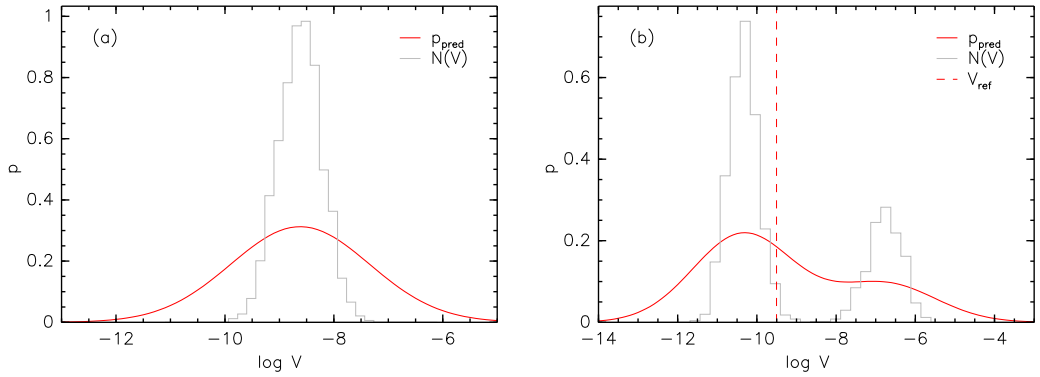

Abbildung 3.17: Vorhersageverteilungen für observierte NOE-Intensitäten des BPTI-Datensatzes. Vorhersageverteilungen sind rot, korrespondierende Häufigkeitsverteilungen grau dargestellt. (a) Vorhersageverteilungen berücksichtigen die endliche Konsistenz eines Datensatzes und sind generell unspezifischer. (b) Der wahre Meßwert V_{ref} wird von der Vorhersageverteilung unterstützt.

Größe des Bereichs, über den sich die Meßwerte erstrecken können, wird systematisch unterschätzt. Im Beispiel der multimodalen Häufigkeitsverteilung (Abb. 3.17(b)) liegt der wahre Intensitätswert (berechnet anhand der BPTI Referenzstruktur) in einer Region mit verschwindender Wahrscheinlichkeit und würde auf Basis der Struktur daher als inkorrekt klassifiziert werden. Die Aussage der Vorhersageverteilung ist hingegen konservativer: Die beiden alternativen Intensitätswerte werden nicht aufgelöst; dies bringt zum Ausdruck, daß Details dieser Art aus den Daten nicht geschlussfolgert werden können. Der wahre Wert ist mit der Vorhersage jedoch kompatibel.

Aus den 1543 Vorhersageverteilungen bestimmte ich die Konfidenzen $\{K_i\}$ aller Messungen gemäß Vorschrift (3.35). Abbildung 3.18 zeigt zwei typische Vorhersageverteilungen (umgerechnet auf die Distanzskala) für eine konsistente und eine inkonsistente Messung: Im Falle der konsistenten Messung (Abb. 3.18(a), $K = 0.92$) stimmen Ziel- und Referenzdistanz nahezu miteinander überein. Die Häufigkeitsverteilung ist zu größeren Abständen hin verschoben, d.h. der Meßwert wird von der Struktur nur mit geringer Wahr-

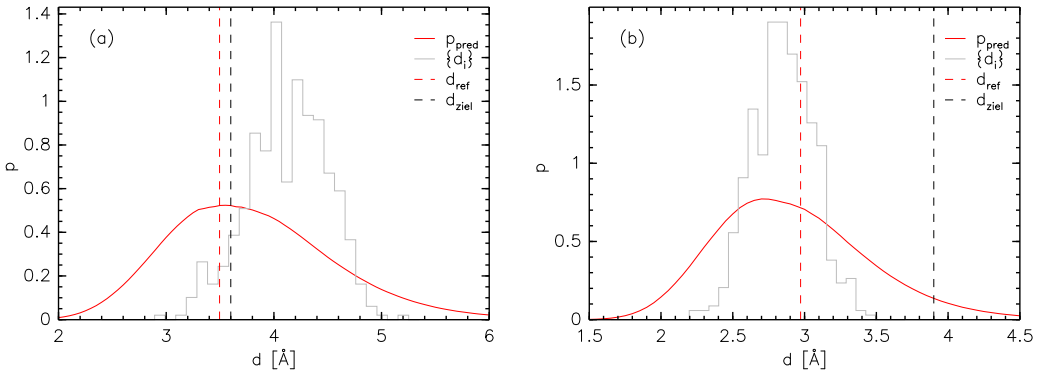

Abbildung 3.18: Konsistente und inkonsistente Messungen. Vorhersageverteilungen sind rot, Häufigkeitsverteilungen grau dargestellt. (a) Konsistente Messung. Die Zieldistanz d_{ziel} wird von den Daten unterstützt und stimmt mit dem wahren Wert d_{ref} und der Vorhersageverteilung gut überein. (b) Inkonsistente Messung. d_{ziel} steht mit den übrigen Daten nicht in Einklang. Die Vorhersageverteilung ist um den wahren Wert konzentriert.

scheinlichkeit unterstützt. Meßwert und Vorhersageverteilung sind jedoch schlüssig, d.h. die Messung wird im Lichte der übrigen Daten als korrekt eingestuft. Die inkonsistente Messung (Abb. 3.18(b), $K = 0.03$) liegt im rechten Schwanz der Vorhersageverteilung und wird von den übrigen Daten somit nicht unterstützt. Die Vorhersageverteilung ist um den korrekten Wert konzentriert.

Betrachtet man die Gesamtheit aller Daten, so werden 70% der Messungen mit einer Konfidenz von mehr als 0.5 als korrekt eingestuft (vgl. Abb. 3.19(a)). Messungen mit kleiner Konfidenz zeigen zum Teil signifikante Abweichungen von bis zu 80% zur Referenzdistanz (Abb. 3.19(b)). Schlechte Messungen können auf diese Weise anhand ihrer Konfidenzwerte *a posteriori* identifiziert werden.

3.2.3.1 Stabilität

Um eine Überinterpretation der Daten zu vermeiden, ist die relative Gewichtung von Datenwissen und *a-priori*-Wissen von großer Wichtigkeit. Ei-

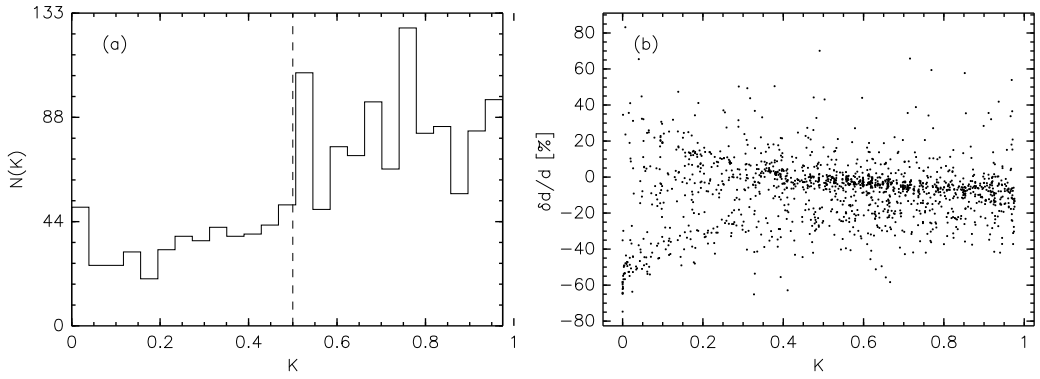

Abbildung 3.19: Konfidenzwerte für den BPTI Datensatz. (a) Histogramm für alle K_i. 70% der Messungen sind mit Konfidenz $K_i > 0.5$ korrekt. (b) Prozentuale Abweichung der Ziel- zur Referenzdistanz in Abhängigkeit der Konfidenz.

ne unsachgemäße Gewichtung kann inbesondere bei Datensätzen geringer Größe dazu führen, daß datenseitige Inkonsistenzen von den Strukturen reproduziert werden. Dies ist im Hinblick auf eine vorurteilsfreie Bestimmung der Datenkonsistenz von Bedeutung, da „Konsistenz“ eine intrinsische Eigenschaft eines Datensatzes ist (in dem in Abschnitt 3.2.1 formulierten Sinne) und daher nicht von seiner Größe abhängen sollte. In der induktiven Strukturbestimmung wird diese Problematik *per constructum* vermieden, da die Datenkonsistenz als unbekannter Hypothesenparameter im Datenmodell berücksichtigt wird und während der Rechnung aus der Ausgangsinformation geschätzt wird. Um die Stabilität von σ_d unter Variation der Größe eines Datensatzes zu demonstrieren, wurden die Rechnungen für BPTI und die Tudor Domäne für Datensätze unterschiedlicher Größe (BF_f bzw. TF_f^{13} und TF_f^{15}) wiederholt.

In Abbildung 3.20(a) ist σ_d gegen die Größe der verschiedenen BPTI Datensätze aufgetragen. Die individuellen σ_d sind im Rahmen ihrer Unsicherheitsbehaftung von der Größe des jeweiligen Datensatzes unabhängig. Selbst für den Datensatz B_{20}, mit lediglich 5 NOEs pro Aminosäure, wurde die Datenkonsistenz korrekt wiedergegeben. Ein analoges Bild ergibt sich bezüglich der experimentellen Datensätze der Tudur Domäne (Abb. 3.20(b)): Für kei-

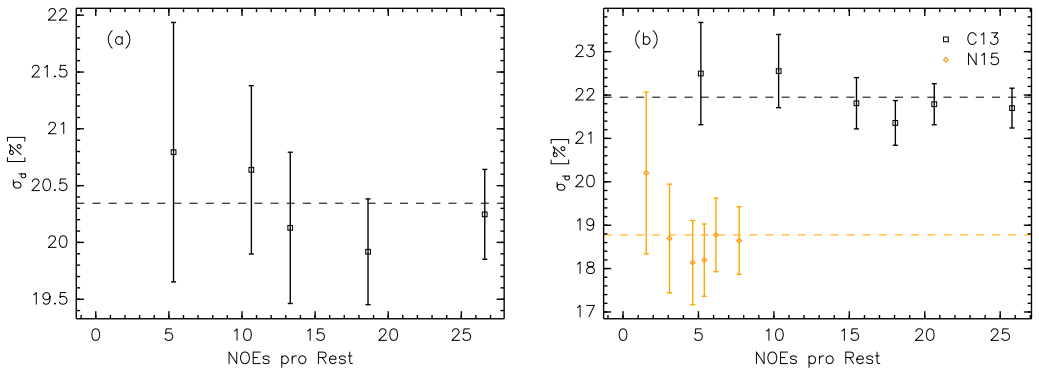

Abbildung 3.20: Stabilität von σ_d unter Variation der Datensatzgröße. (a) Werte für BPTI. Datenpunkte korrespondieren zu den Datensätzen B_X. (b) Werte für die Tudor Domäne, getrennt für die Datensätze T_X^{13} und T_X^{15}. Horizontale Linien korrespondieren zu pro-Datensatz-gemittelten Werten.

nen der beiden Datensätze ist eine nennenswerte Abhängigkeit von σ_d von der Datendichte zu erkennen. Die Größe eines Datensatzes führt demzufolge zu keinen systematischen Fehlern bei der Bestimmung der Konsistenz, sondern spiegelt sich lediglich in der Stärke der Unsicherheitsbehaftung von σ_d wider.

3.2.3.2 Datensätze unterschiedlicher Konsistenz

Um eine quantitative Abschätzung dafür zu erhalten, wie sich die strukturelle Erfüllbarkeit eines Datensatzes auf den Zahlenwert von σ_d auswirkt, leite ich im Folgenden eine theoretische Beziehung zwischen beiden Größen ab. Anhand dieser Beziehung werde ich nachfolgend zeigen, daß die Konsistenz eines Datensatzes während der Strukturrechnung zurückgerechnet werden kann.
Ich nehme vereinfachend an, daß der gegebene Datensatz in zwei Mengen Ω_K und Ω_I zerlegt werden kann, die aus konsistenten bzw. inkonsistenten Messungen gebildet werden. Die Intensitäten konsistenter Messungen seien über das NOE-Datenmodell simultan erklärbar. Die Gesamtzahl der Messungen sei N, wovon der Anteil $f \in [0, 1]$ auf die Menge Ω_K, der Anteil $1 - f$ auf Ω_I

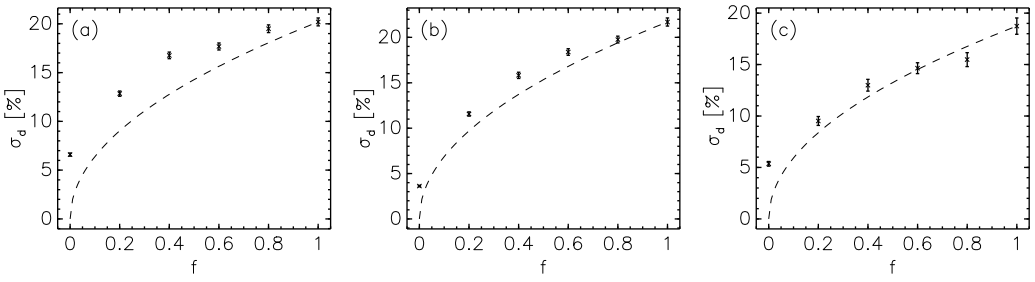

Abbildung 3.21: Zurückgerechnete Datenkonsistenz. f bezeichnet den Anteil inkonsistenter Messungen. (a) BPTI Datensätze BF_f. (b,c) Datensätze der Tudor Domäne TF_f^{13} bzw. TF_f^{15}. Gestrichelt: Verlauf der theoretischen Beziehung.

entfalle. Unter dieser Annahme folgt aus Gleichung (3.28) für die marginale *a-posteriori*-Verteilung für σ^2:

$$\begin{aligned} p(\sigma^2|\cdot) &\propto \sigma^{-(N+2)} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i\in\Omega_K} R_i + \sum_{i\in\Omega_I} R_i\right]\right\} \\ &\propto \sigma^{-(N+2)} \exp\left\{-\frac{N}{2\sigma^2}\left[(1-f)R_K + fR_I\right]\right\}, \end{aligned} \tag{3.36}$$

wobei $R_i = \log^2\left(\tilde{V}_i/\gamma d_i^{-6}(\mathbf{x})\right)$. R_K und R_I bezeichnen die mittleren Diskrepanzen bezüglich aller konsistenten bzw. inkonsistenten Messungen. Im Grenzfall unabhängig voneinander erfüllbarer Messungen verschwindet nach Voraussetzung die mittlere Diskrepanz für den konsistenten Anteil der Daten, d.h. es gilt $R_K \equiv 0$. In diesem Fall folgt aus Gleichung (3.36) und den Eigenschaften der inversen Gammaverteilung

$$\langle\sigma^2\rangle|\cdot = \frac{N}{N-1} f R_I.$$

Für $N \gg 1$ ergibt sich damit die Abhängigkeit von σ_d vom Anteil inkonsistenter Messungen, f:

$$\langle\sigma_d(f)\rangle \approx \sigma_d(1)\sqrt{f}, \tag{3.37}$$

d.h. die erwartete relative Abweichung der Zieldistanzen von den Distanzen in der Struktur ist proportional zur Wurzel des Anteils inkonsistenter Messungen.

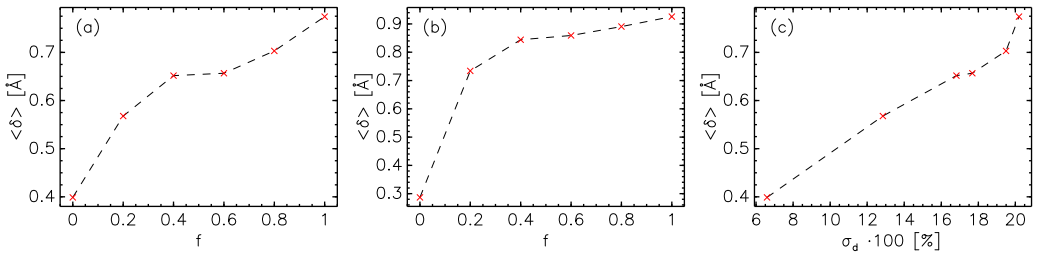

Abbildung 3.22: Abhängigkeit der mittleren Koordinatenunsicherheit von der Datenkonsistenz. (a,b) Unsicherheitsbehaftung als Funktion des Anteils inkonsistenter Messungen für BPTI bzw. die Tudor Domäne. (c) Unsicherheitsbehaftung als Funktion der Konsistenz der BPTI-Datensätze BF_f.

Abbildung 3.21 zeigt die zurückgerechnete Datenkonsistenz für die Datensätze BF_f, TF_f^{13} und TF_f^{15}. Die theoretische Beziehung aus Gleichung (3.37) ist getrichelt dargestellt. Der Verlauf von σ_d ist in allen drei Testfällen in guter Übereinstimmung mit der theoretischen Vorhersage; die streng monotone Abhängigkeit wird reproduziert. Relativ zu den theoretischen Werten ist die zurückgerechnete Datenkonsistenz systematisch zu größeren Werten verschoben. Dieses Verhalten ist auf Inkompatiblitäten der Kraftfelder zurückzuführen, die für die Simulation der Strukturverteilungen (Kraftfeld des ISD-Simulationspakets) und die Berechnung der Referenzstrukturen verwendet wurden. Zudem verwendet das ISD-Simulationspaket eine interne Parametrisierung durch Dihedralwinkel und fixiert den Hauptketten-Dihedralwinkel ω, wodurch zusätzliche Inkompatibilitäten entstehen können. Auf Basis der Referenzstrukturen berechnete Intensitäten sind daher nicht simultan erfüllbar, was sich in einem nichtverschwindenden Wert von σ_d bei $f = 0$ äußert: Sowohl für BPTI als auch für die Tudor Domäne führen Inkompatibilitäten in den Kraftfeldern zu relativen Abweichungen in den Distanzen von $\sigma_d(0) \approx 5\%$.

3.2.3.3 Konsistenz und strukturelle Unsicherheit

Um ein detailliertes Bild über den Einfluß von Inkonsistenzen in den Daten auf die atomare Unsicherheitsbehaftung einer Struktur zu erhalten, berech-

nete ich die Koordinatenunsicherheiten bezüglich aller Strukturverteilungen, die auf Basis der Datensätze BF_f und TF_f^{13}, TF_f^{15} simuliert wurden. Für die Berechnung verwendete ich die in Kapitel 3.1 entwickelte Methode. In Abbildung 3.22(a,b) sind die mittleren Koordinatenunsicherheiten von BPTI und der Tudor Domäne gegen den Anteil inkonsistenter Messungen in den jeweiligen Datensätze aufgetragen. Die Unsicherheit in den Atompositionen nimmt in beiden Fällen monoton mit wachsender Inkonsistenz der Daten zu. Die genannten Inkompatibilitäten der Kraftfelder führen zu einer unteren Schranke der Koordinatenunsicherheiten (0.4 Å für BPTI bzw. 0.3 Å für die Tudor Domäne). Abbildung 3.22(c) zeigt die atomare Unsicherheitsbehaftung als Funktion der geschätzten Datenkonsistenz σ_d für BPTI.
Die Abhängigkeit der atomaren Unsicherheitsbehaftung von der Konsistenz eines Datensatzes läßt sich theoretisch motivieren: Für Datensätze typischer Größe sind die *Nuisance*-Parameter σ^2 und γ gut bestimmt (die Varianz beider Größen ist invers proportional zur Größe des Datensatzes). Die marginale *a-posteriori*-Verbundverteilung $p(\sigma^2, \gamma | D, I)$ kann daher in guter Näherung durch Delta-Funktionen an ihrem Maximum ($\hat{\sigma}^2$,$\hat{\gamma}$) approximiert werden. Die Strukturverteilung ist in diesem Fall gleich der bedingten Strukturverteilung ausgewertet an den Maxima der *Nuisance*-Parameter:

$$\begin{aligned} p(\mathbf{x}|D,I) &= \int d\sigma^2 d\gamma \, p(\mathbf{x}|\sigma^2,\gamma,D,I) p(\sigma^2,\gamma|D,I) \\ &\approx p(\mathbf{x}|\hat{\sigma}^2,\hat{\gamma},D,I). \end{aligned} \tag{3.38}$$

Um einen analytischen Ausdruck für die Skala zu erhalten, auf der sich die Stärke der Streuung der kartesischen Koordinaten bewegt, schreibe ich Gleichung (3.38) in Gauß'scher Näherung:

$$p(\mathbf{x}|\hat{\sigma}^2,\hat{\gamma},D,I) \approx \frac{\det(H_L)^{1/2}}{(2\pi)^{3M/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^{\mathsf{T}} H_L(\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}})\right\}. \tag{3.39}$$

Gleichung (3.39) folgt aus der Taylor-Entwicklung um das Minimum $\hat{\mathbf{x}}$ des negativen Logarithmus von Gl. (3.38), $L(\mathbf{x}) = -\log p(\mathbf{x}|\hat{\sigma}^2,\hat{\gamma},D,I)$, bis zur 2. Ordnung. $H_L(\mathbf{x})$ bezeichnet die Hesse-Matrix von L. In Gauß'scher Näherung ist die Strukturverteilung demzufolge um die wahrscheinlichste Struktur

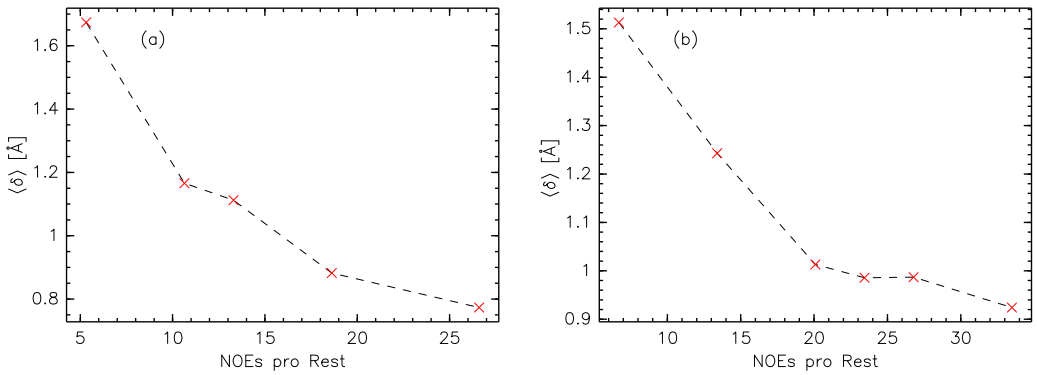

Abbildung 3.23: Abhängigkeit der atomaren Unsicherheitsbehaftung von der Größe eines Datensatzes. (a) BPTI-Datensätze B_X. (b) Datensätze der Tudor Domäne T_X^{13} und T_X^{15}.

konzentriert und nach Gleichung (2.11) von der Form:

$$p(\mathbf{x}|D,I) \approx \frac{\det(H_{L'})^{1/2}}{(2\pi\hat{\sigma}^2/N)^{3M/2}} \exp\left\{-\frac{N}{2\hat{\sigma}^2}(\mathbf{x}-\hat{\mathbf{x}})^{\mathsf{T}} H_{L'}(\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}})\right\}, \qquad (3.40)$$

mit

$$L'(\mathbf{x}) \;=\; \frac{\hat{\sigma}^2}{N}L(\mathbf{x}) = \frac{\hat{\sigma}^2\beta}{N}E(\mathbf{x}) + \frac{1}{2}\left\langle \sum_{i=1}^{N} \log^2\left(\tilde{V}_i/\hat{\gamma}d_i^{-6}(\mathbf{x})\right)\right\rangle,$$

wobei $H_{L'}$ die Hesse-Matrix von L' bezeichnet. Für hinreichend große Datensätze wird L' vom Datenterm dominiert, so daß die $\hat{\sigma}^2$-Abhängigkeit von $H_{L'}$ vernachlässigbar ist. Die Ausdehnung der Strukturverteilung bewegt sich nach Gl. (3.40) auf der Skala $\hat{\sigma}^2/N$. Für die atomare Unsicherheitsbehaftung ergibt sich entsprechend $\delta \propto \sigma_d/\sqrt{N}$.

In dieser Näherung erwartet man daher einen linearen und monoton steigenden Zusammenhang zwischen der Inkonsistenz eines Datensatzes und der strukturellen Unsicherheitsbehaftung. Dieser Zusammenhang ist im Falle des BPTI-Datensatzes in guter Näherung erfüllt (vgl. Abb. 3.22(c)). Die Unsicherheitsbehaftung wird außer von der Datenkonsistenz auch von der Größe eines Datensatzes bestimmt: Die Koordinatenunsicherheiten fallen mit der

Wurzel der Zahl der Messungen, was in den Simulationen in guter Näherung reproduziert wird (vgl. Abb. 3.23). Somit äußern sich beide Arten der Unvollständigkeit experimenteller Daten direkt in der Unsicherheitsbehaftung der berechneten Konformationen. Die σ^2-Abhängigkeit der Ausdehnung der Strukturverteilung unterstreicht noch einmal die Notwendigkeit, diesen Parameter aus den Daten zu schätzen, und nicht zu fixieren.
Die Linearität der abgeleiteten Beziehung ist nur bei scharf bestimmten *Nuisance*-Parametern, großen Datensätzen und normalverteilten Strukturkoordinaten erfüllt. Bei typischen Strukturbestimmungsproblemen sind die beiden ersten Voraussetzungen näherungsweise erfüllt. Abweichungen der Strukturverteilung von einer Gauß'schen Form führen in Gl. (3.40) zu entsprechenden Korrekturtermen höherer Ordnung.

3.3 Modellierung inkonsistenter NOE-Daten

Wie in den vorigen Abschnitten gezeigt wurde, erhöhen Inkonsistenzen in den Daten die Unschärfe der Strukturverteilung und nehmen folglich Einfluß auf die lokale Unsicherheitsbehaftung einer Struktur. Systematische Abweichungen von observierten und berechneten Kreuzrelaxationsraten führen außerdem zu strukturellen Verzerrungen, wodurch lokale Unsicherheiten in den Atompositionen zusätzlich verfälscht werden. Voraussetztung für eine vorurteilsfreie Interpretation der Strukturverteilung ist daher die Minimierung von möglichen Inkonsistenzen durch die Wahl eines geeigneten Datenmodells.
Neben der Spindiffusion kann die Bewegung eines Moleküls signifikante Inkonsistenzen in NOESY-Datensätzen verursachen [72]. Spindiffusionskorrekturen zu Kreuzrelaxationsraten können über mehrere Verfahren berechnet werden [73, 74] und ließen sich somit im Datenmodell für dipolare Kreuzrelaxationsraten im Prinzip berücksichtigen. Die Berechnung von Dynamikkorrekturen zu NOE-Intensitäten ist hingegen von großer Schwierigkeit: Dynamikeffekte lassen sich bislang nicht in allgemeiner und analytisch geschlos-

sener Form beschreiben, weshalb ihre Berücksichtigung in der Strukturrechnung aufwendig ist.

Die folgenden Abschnitte behandeln ein Datenmodell für dipolare Kreuzrelaxationsraten, welches Inkonsistenzen in den Messungen explizit berücksichtigt. In dem gewählten Zugang werden Dynamikbeiträge nicht anhand einer physikalischen Theorie berechnet, sondern ihr Einfluß auf die beobachtete Fehlerverteilung durch ein realistisches Fehlermodell geeignet modelliert. Die Berücksichtigung von Inkonsistenzen erfolgt dabei individuell für jeden Meßwert: Dies erfordert die Einführung einer Vielzahl von *Nuisance*-Parametern, deren Zahl die Anzahl der Messungen übersteigt. Es wird sich zeigen, daß in der induktiven Strukturbestimmung auch komplexe Modelle mit einer Vielzahl unbekannter Parameter verläßlich aus den Daten geschätzt werden können. Die modellseitige Berücksichtigung systematischer Fehler in den Daten gestattet ferner die Bewertung jeder Einzelmessung in Hinsicht auf ihre strukturelle Erfüllbarkeit. Abschnitte 3.3.1 und 3.3.2 behandeln die Herleitung des Datenmodells sowie die Realisierung der Replika-Austausch-Monte-Carlo-Strategie. Abschnitte 3.3.3 und 3.3.4 zeigen die Anwendung des Modells auf BPTI und die SMN Tudor Domäne.

3.3.1 Datenmodellierung

Bei Kenntnis der wahren Konformation eines Moleküls lassen sich inkonsistente Meßwerte von konsistenten Messungen anhand der Abweichung der oberservierten von den theoretischen Intensitäten isolieren. Betrachtet man die Häufigkeitsverteilung dieser Abweichungen, d.h. die „experimentelle“ Fehlerverteilung der Daten, so sind konsistente Messungen bei kleinen Abweichungen, inkonsistente Messungen hingegen tendenziell im Schwanz der Verteilung lokalisiert. Es ist die Aufgabe des Datenmodells, diese Abweichungen vorurteilsfrei zu beschreiben.
Für die Formulierung des Datenmodells nehme ich an, daß die Gesamtheit aller Messungen als Mischung von Einzelmessungen unterschiedlichen Typs

beschrieben werden kann. Jede Einzelmessung gehöre genau einer von zwei Klassen an: Die eine Klasse definiere konsistente, die andere Klasse inkonsistente Messungen. Die relativen Populationen beider Klassen seien jedoch unbekannt. Datenmodelle dieser Art werden allgemein als *Mischmodelle* bezeichnet; bei dem im Folgenden beschriebenen Datenmodell handelt es sich also um ein zweikomponentiges Mischmodell.
Die Modellierung des Datensatzes als Mischung gestattet die individuelle Beschreibung jeder Messung durch klassenspezifische Datenmodelle: Konsistente Messungen sind per Definition mit der ISPA verträglich und werden mit guter Genauigkeit durch das bisherige Datenmodell für Kreuzrelaxationsraten aus Kapitel 3.1.1.1 beschrieben. Große Abweichungen in den Intensitäten, wie sie für inkonsistente Messungen zu erwarten sind, werden durch ein separates Datenmodell berücksichtigt, welches den Schwanz der experimentellen Fehlerverteilung modelliert: Ich beschreibe inkonsistente Messungen anhand einer Fehlerverteilung mit signifikanter Breite.
Die Zuordnung der Messungen zu einer der beiden Klassen geschieht während der Strukturrechnung: Der Datensatz wird dabei zerlegt in Intensitäten, welche simultan anhand einer Konformation erklärt werden können, und Messungen, welche mit der Struktur nicht in Einklang stehen. Inkonsistente Meßwerte werden aufgrund der Breite der ihnen zugeordneten Fehlerverteilung automatisch schwächer gewichtet, wodurch strukturelle Verzerrungen wirksam reduziert werden.

3.3.1.1 Ein Klassifikationsmodell

Das Datenmodell bestehe aus den Komponenten K und I. K bezeichne die Komponente zur Modellierung konsistenter, I die Komponente zur Beschreibung inkonsistenter Messungen. Den Komponenten seien individuelle Datenmodelle $p_K(\tilde{V}_i|\cdot)$ bzw. $p_I(\tilde{V}_i|\cdot)$ zugeordnet, welche die Wahrscheinlichkeit beschreiben, daß ein konsistenter bzw. inkonsistenter NOE mit Intensität $\tilde{V}_i$ observiert wird. Der *Mischparameter* ω quantifiziert das Mischungsverhältnis beider Komponenten und bezeichnet die Wahrscheinlichkeit mit der eine

Messung von der K-Komponente generiert wurde. Das Mischmodell ist die gewichtete Summe der individuellen Datenmodelle:

$$p_{\mathrm{MIX}}(\tilde{V}_i|\omega,\cdot) = \omega\, p_K(\tilde{V}_i|\cdot) + (1-\omega)\, p_I(\tilde{V}_i|\cdot). \tag{3.41}$$

Um einen expliziten Ausdruck für das Mischmodell zu erhalten, spezifiziere ich die individuellen Datenmodelle der K- und I-Komponente:

K-Komponente

Für die Beschreibung konsistenter Messungen wähle ich das Lognormal-Datenmodell für Kreuzrelaxationsraten aus Gleichung (3.3) mit Formparameter σ_K:

$$p_K(\tilde{V}_i|\mathbf{x},\sigma_K^2,\gamma,I) = \frac{1}{\sqrt{2\pi\sigma_K^2}}\tilde{V}_i^{-1}\exp\left\{-\frac{1}{2\sigma_K^2}\log^2\left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})}\right)\right\}. \tag{3.42}$$

I-Komponente

Werden dynamikinduzierte Beiträge zu Kreuzkorrelationsraten in einem Korrekturfaktor Γ_i^{dyn} absorbiert, so lautet das Vorwärtsmodell für die Berechnung der Intensität eines NOE allgemein:

$$V_i^{\mathrm{dyn}}(\mathbf{x},\gamma) = \Gamma_i^{\mathrm{dyn}} V_i(\mathbf{x},\gamma).$$

$V_i(\mathbf{x},\gamma)$ bezeichnet die ISPA. Die Abwesenheit jedweder Dynamikbehaftung entspricht demnach $\Gamma_i^{\mathrm{dyn}} = 1$. Beschreibt man die Diskrepanz zwischen beobachteter und berechneter Intensität – analog zum statischen Fall – durch eine Lognormalverteilung mit log-Varianz σ^2, so folgt als Datenmodell für die observierte dynamikbehaftete Intensität:

$$p(\tilde{V}_i|\mathbf{x},\sigma^2,\gamma,\Gamma_i^{\mathrm{dyn}},I) = \frac{1}{\sqrt{2\pi\sigma^2}}\tilde{V}_i^{-1}\exp\left\{-\frac{1}{2\sigma^2}\log^2\left(\frac{\tilde{V}_i}{\gamma\Gamma_i^{\mathrm{dyn}} d_i^{-6}(\mathbf{x})}\right)\right\}. \tag{3.43}$$

Details der molekularen Dynamik, und damit der Wert von Γ_i^{dyn} seien unbekannt. Dieses Unwissen wird durch eine Wahrscheinlichkeitsverteilung für

Γ_i^{dyn} repräsentiert: Γ_i^{dyn} ist positiv und streue um 1. Die Varianz der Streuung sei für jeden NOE identisch, jedoch unbekannt. Die natürliche Wahrscheinlichkeitsverteilung, welche dieses Wissen repräsentiert, ist eine Lognormalverteilung mit Ortsparameter 1 und Formparameter σ_Γ:

$$p(\Gamma_i^{\mathrm{dyn}}|\sigma_\Gamma^2, I) = \frac{1}{\sqrt{2\pi\sigma_\Gamma^2}}(\Gamma_i^{\mathrm{dyn}})^{-1} \exp\left\{-\frac{1}{2\sigma_\Gamma^2}\log^2\Gamma_i^{\mathrm{dyn}}\right\}. \tag{3.44}$$

Das Datenmodell der I-Komponente des Mischmodells geht aus Gln. (3.43) und (3.44) unmittelbar durch Marginalisierung von Γ_i^{dyn} hervor:

$$\begin{aligned} p_I(\tilde{V}_i|\mathbf{x}, \sigma_I^2, \gamma, I) &= \int_0^\infty d\Gamma_i^{\mathrm{dyn}} p(\tilde{V}_i|\mathbf{x}, \sigma^2, \gamma, \Gamma_i^{\mathrm{dyn}}, I)\, p(\Gamma_i^{\mathrm{dyn}}|\sigma_\Gamma^2, I) \\ &= \frac{1}{\sqrt{2\pi\sigma_I^2}}\tilde{V}_i^{-1} \exp\left\{-\frac{1}{2\sigma_I^2}\log^2\left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})}\right)\right\}, \end{aligned} \tag{3.45}$$

wobei $\sigma_I^2 = \sigma^2 + \sigma_\Gamma^2$. Nach Gleichung (3.45) ist die Diskrepanz zwischen observierter und berechneter Intensität im Falle einer inkonsistenten Messung ebenfalls lognormalverteilt. Die Breite der Fehlerverteilung wird über den Formparameter σ_I bestimmt.

Die explizite Form des Mischmodells (vgl. Abb. 3.24) ergibt sich durch Einsetzen von Gln. (3.42) und (3.45) in Gl. (3.41) zu:

$$\begin{aligned} p_{\mathrm{MIX}}\left(\tilde{V}_i|\mathbf{x}, \omega, \sigma_K^2, \sigma_I^2, \gamma, I\right) &= \frac{1}{\sqrt{2\pi}}\tilde{V}_i^{-1}\left[\frac{\omega}{\sigma_K}\exp\left\{-\frac{1}{2\sigma_K^2}\log^2\left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})}\right)\right\}\right. \\ &+ \left.\frac{1-\omega}{\sigma_I}\exp\left\{-\frac{1}{2\sigma_I^2}\log^2\left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})}\right)\right\}\right]. \end{aligned} \tag{3.46}$$

Als *Likelihood*-Funktion für N unabhängig voneinander observierte Kreuzrelaxationsraten ergibt sich gemäß Gl. (3.46) ein Produkt von Summen, was der expliziten Berücksichtigung aller 2^N möglichen Aufteilungen von N Intensitäten auf 2 Klassen entspricht. Diese *Likelihood*-Funktion läßt sich auf keine analytisch geschlossene Form bringen und ist für eine effiziente Simulation daher ungeeignet.

Mit Hilfe der *missing data*-Formulierung [75] läßt sich das Mischmodell aus einem erweiterten Datenmodell ableiten, in welchem die Klassenzugehörigkeit einer Messung explizit berücksichtigt wird. Die *Likelihood*-Funktion des

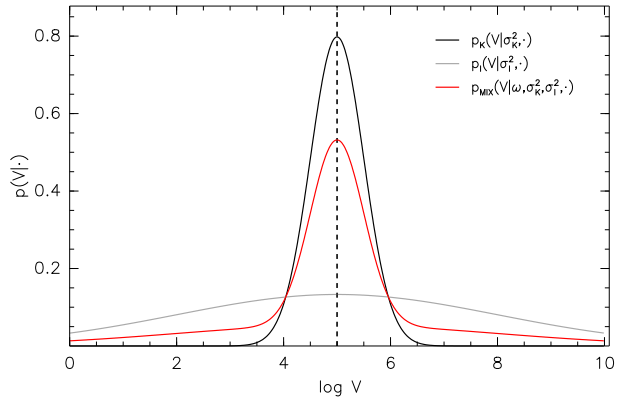

Abbildung 3.24: Illustration des Mischmodells für NOE-Intensitäten. Das Mischmodell (rot) ist die gewichtete Summe der Datenmodelle zur Modellierung konsistenter (K-Komponente, schwarz) und inkonsistenter Messungen (I-Komponente, grau); der Mischparameter ω spezifiziert die Stärke der Mischung. Die Intensitäten inkonsistenter NOEs können signifikant von den theoretischen Werten abweichen, was im Datenmodell durch die unspezifische I-Komponente berücksichtigt wird.

erweiterten Datenmodells läßt sich in analytisch geschlossener Form angeben. Dazu wird jeder Messung ein *Klassenzugehörigkeitsparameter* $z_i \in \{0, 1\}$ zugeordnet, der angibt, von welcher Komponente die betreffende Messung generiert wurde:

$$z_i = \begin{cases} 1 & : \quad i\text{-te Messung wurde von } K\text{-Komponente erzeugt,} \\ 0 & : \quad i\text{-te Messung wurde von } I\text{-Komponente erzeugt.} \end{cases} \tag{3.47}$$

Bei bekannter Klassenzugehörigkeit einer Messung lassen sich die individuellen Datenmodelle der K- und I-Komponente formal durch das erweiterte Datenmodell

$$p(\tilde{V}_i|z_i, \cdot) = \Big[p_K(\tilde{V}_i|\cdot)\Big]^{z_i} \cdot \Big[p_I(\tilde{V}_i|\cdot)\Big]^{1-z_i} \tag{3.48}$$

darstellen. Wurde die Messung von der K-Komponente generiert, so gilt $p_K(\tilde{V}_i|\cdot) = p(\tilde{V}_i|z_i = 1, \cdot)$ und analog $p_I(\tilde{V}_i|\cdot) = p(\tilde{V}_i|z_i = 0, \cdot)$. Der Klassenzugehörigkeitsparameter ist unbekannt und wird durch Marginalisierung eliminiert. Die Berücksichtigung des *a-priori*-Wissens, daß eine Messung mit der Wahrscheinlichkeit ω von der K-Komponente generiert wurde, wird durch

die *a-priori*-Verteilung für z_i berücksichtigt:

$$p(z_i|\omega, I) = \omega^{z_i}(1-\omega)^{1-z_i}. \tag{3.49}$$

Die allgemeine Form des Mischmodells aus Gl. (3.41) folgt unmittelbar aus dem erweiterten Datenmodell durch Marginalisierung von z_i:

$$\begin{aligned} p_{\mathrm{MIX}}(\tilde{V}_i|\omega,\cdot) &= \sum_{z_i\in\{0,1\}} p(\tilde{V}_i|z_i,\cdot)p(z_i|\omega,I) \\ &= \sum_{z_i\in\{0,1\}} \Big[\omega\, p_K(\tilde{V}_i|\cdot)\Big]^{z_i} \cdot \Big[(1-\omega)\, p_I(\tilde{V}_i|\cdot)\Big]^{1-z_i} \\ &= \omega p_K(\tilde{V}_i|\cdot) + (1-\omega)p_I(\tilde{V}_i|\cdot). \end{aligned} \tag{3.50}$$

Durch Einsetzen der Datenmodelle der K- und I-Komponente aus Gln. (3.42) und (3.45) in Gleichung (3.48) ergibt sich für das erweiterte Datenmodell der kompakte Ausdruck:

$$\begin{aligned} p(\tilde{V}_i|\mathbf{x},\sigma_K^2,\sigma_I^2,\gamma,z_i,I) \;\propto\; & \left[\frac{1}{\sigma_K}\right]^{z_i}\left[\frac{1}{\sigma_I}\right]^{1-z_i} \\ & \times \exp\left\{-\frac{1}{2\sigma_{\mathrm{eff},i}^2}\log^2\left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})}\right)\right\}, \end{aligned} \tag{3.51}$$

wobei

$$\sigma_{\mathrm{eff},i}^2 = \frac{\sigma_K^2\sigma_I^2}{z_i\,\sigma_I^2 + (1-z_i)\,\sigma_K^2}. \tag{3.52}$$

Die *Likelihood*-Funktion für N unabhängige Intensitäten $D = \{\tilde{V}_1,\ldots,\tilde{V}_N\}$ folgt unmittelbar aus Gleichung (3.51):

$$\begin{aligned} p(D|\mathbf{x},\sigma_K^2,\sigma_I^2,\gamma,\{z_i\},I) &= \prod_{i=1}^{N} p(\tilde{V}_i|\mathbf{x},\sigma_K^2,\sigma_I^2,\gamma,z_i,I) \\ &\propto \sigma_K^{-N_K}\,\sigma_I^{-(N-N_K)} \\ &\quad \times \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\frac{1}{\sigma_{\mathrm{eff},i}^2}\log^2\left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})}\right)\right\}. \end{aligned} \tag{3.53}$$

$N_K = \sum_{i=1}^{N} z_i$ bezeichnet die Zahl der konsistenten Messungen. Im Mischmodell wird somit jeder Messung eine effektive Varianz $\sigma_{\mathrm{eff},i}^2$ zugeordnet,

welche die strukturelle Erfüllbarkeit der Messung quantifiziert. Die inversen effektiven Varianzen fungieren in der *Likelihood*-Funktion als Gewichtungsfaktoren. Das Gewicht jeder Messung ist von der jeweiligen Klassenzugehörigkeit abhängig und setzt sich anteilig aus den Varianzen der K- bzw. I-Komponente zusammen.

3.3.1.2 Die *a-posteriori*-Verteilung

Die Klassifikationsparameter des Datenmodells, $\{z_i\}$ und ω, sowie die Parameter der K- und I-Komponente, σ_K^2 und γ, sind unbekannt und werden wie gewohnt zusammen mit den Koordinaten der Struktur aus dem Datensatz geschätzt. Aufgrund der NOE-weisen Zuordnung der Klassenzugehörigkeitsvariablen übersteigt die Zahl der Modellparameter generell die Zahl der Daten.

Den Formparameter der I-Komponente σ_I werde ich nicht aus den Daten schätzen, sondern als gegeben annehmen: Gln. (3.48) und (3.49) sind invariant unter Vertauschung beider Klassen, d.h. unter der Transformation

$$\begin{aligned} \sigma'_K &= \sigma_I, & \sigma'_I &= \sigma_K, \\ \omega' &= 1-\omega, & z' &= 1-z. \end{aligned}$$

Aufgrund dieser Invarianz ist die Definition der Klassenzugehörigkeitsparameter in Gl. (3.47), d.h. die Zuordnung einer Messung zur konsistenten Klasse bei $z_i = 1$, strenggenommen willkürlich. Technisch gesehen ist die simultane Schätzung von σ_K^2 und σ_I^2 unproblematisch. ω und $\{z_i\}$ ließen sich in diesem Fall jedoch nicht eindeutig und im Sinne ihrer Definition interpretieren. Die Symmetrie kann gebrochen werden, indem σ_I auf einen sinnvollen Wert fixiert wird[2].

Die Koordinaten der Struktur seien von den Parametern des Mischmodells logisch unabhängig. Dann besitzt die *a-priori*-Verbundverteilung die folgende

[2] Alternativ ließe sich im Datenmodell das Vorwissen $\sigma_I \gg \sigma_K$ berücksichtigen. Dies zöge jedoch ein aufwendigeres Datenmodell mit entsprechend höherem Rechenaufwand nach sich und wurde daher nicht weiter verfolgt.

Struktur:

$$p(\mathbf{x}, \sigma_K^2, \gamma, \omega, \{z_i\}|I) \;=\; p(\mathbf{x}|I)\, p(\sigma_K^2, \gamma, \omega|I) \prod_{i=1}^{N} p(z_i|\omega, I). \qquad (3.54)$$

Als *a-priori*-Verteilung für σ_K^2 und γ wähle ich den jeweiligen Jeffreys-*prior*. Unter Beachtung der expliziten Form der *a priori* Verteilung für $\{z_i\}$ in Gl. (3.49) sowie der Boltzmann-Verteilung als konformationelle *a-priori*-Verteilung ergibt sich die *a-priori*-Verbundverteilung zu:

$$p(\mathbf{x}, \sigma_K^2, \gamma, \omega, \{z_i\}|I) \propto (\sigma_K^2\gamma)^{-1}\, \omega^{N_K+\frac{1}{2}} (1-\omega)^{N-N_K+\frac{1}{2}} e^{-\beta E(\mathbf{x})}, \qquad (3.55)$$

wobei als *a-priori*-Verteilung für ω eine Betaverteilung (s. Anhang A.3) mit Formparametern $\alpha = \beta = 1/2$ verwendet wurde.
Die *a-posteriori*-Verbundverteilung folgt aus Gln. (3.53) und (3.55):

$$\begin{aligned} p(\mathbf{x}, \sigma_K^2, \omega, \gamma, \{z_i\}|\sigma_I^2, D, I) \;&\propto\; \omega^{N_K+\frac{1}{2}}\, (1-\omega)^{N-N_K+\frac{1}{2}}\, \sigma_K^{-(N_K+2)} \gamma^{-1} \qquad (3.56)\\ &\times\; \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} \frac{1}{\sigma_{\mathrm{eff},i}^2} \log^2 \left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})} \right) - \beta E(\mathbf{x}) \right\}. \end{aligned}$$

3.3.2 Realisierung des Replika-Algorithmus

Die Berechnung der Strukturverteilung durch Integration über alle *Nuisance*-Parameter erfolgt auf numerischem Wege durch Simulation der *a-posteriori*-Verbundverteilung aus Gl. (3.56). Die benötigten bedingten *a-posteriori*-Verteilungen für die Hypothesenparameter $\sigma_K^2, \omega, \gamma$ folgen unmittelbar aus Gl. (3.56) unter Beachtung von Gl. (3.52):

$$p(\sigma_K^2|\cdot) \;=\; \mathrm{IG}\left(\sigma_K^2; N_K/2, \frac{1}{2} \sum_{i=1}^{N} z_i \log \left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})} \right) \right), \qquad (3.57)$$

$$p(\gamma|\cdot) \;=\; \mathrm{LN}\left(\gamma; \frac{\sum_{i=1}^{N} \sigma_{\mathrm{eff},i}^{-2} \log\left(\tilde{V}_i / d_i^{-6}(\mathbf{x}) \right)}{\sum_{i=1}^{N} \sigma_{\mathrm{eff},i}^{-2}}, 1/\sum_{i=1}^{N} \sigma_{\mathrm{eff},i}^{-2} \right), \qquad (3.58)$$

$$p(\omega|\cdot) \;=\; \mathrm{B}\left(\omega; N_K + 1/2, N - N_K + 1/2 \right). \qquad (3.59)$$

Bedingt durch die Größe realistischer Datensätze ist die Erzeugung der NOE-weisen Klassenzugehörigkeitsparameter $\{z_i\}$ durch Ziehen von ihren bedingten *a-posteriori*-Verteilungen numerisch teuer. Um den Rechenaufwand zu

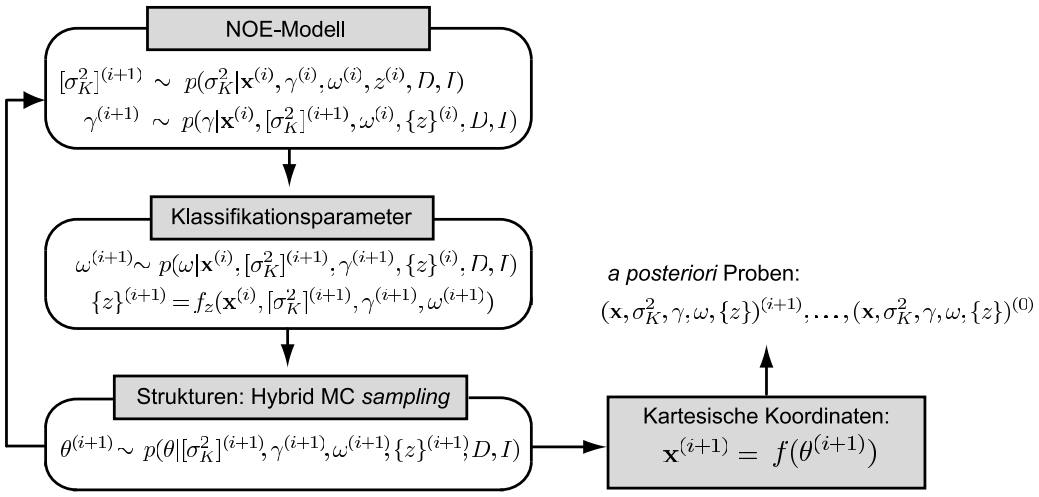

Abbildung 3.25: Gibbs-Algorithmus zur Simulation der Strukturverteilung: Für jeden MC Schritt werden erst die Parameter der individuellen Datenmodelle gezogen, gefolgt von den Klassifikationsparametern des Mischmodells. Die Erzeugung einer neuen Konformation erfolgt durch Hybrid-Monte-Carlo.

reduzieren, generiere ich den Satz der $\{z_i\}$ daher nicht durch Ziehen von den bedingten Verteilungen, sondern anhand ihres bedingten Erwartungswertes; dies entspricht der Approximation der bedingten *a-posteriori*-Verteilungen durch Delta-Funktionen, also $p(z_i|\cdot) = \delta(z_i - \langle z_i \rangle)$. Der bedingte Erwartungswert für z_i folgt aus Gln. (3.48) und (3.49):

$$\begin{aligned} \langle z_i \rangle|\cdot &= \frac{\omega p_K(\cdot)}{\omega p_K(\cdot) + (1-\omega) p_I(\cdot)} \\ &= \left[1 + \frac{1-\omega}{\omega} \frac{\sigma_K}{\sigma_I} \exp\left\{ \frac{\sigma_K^2 - \sigma_I^2}{2\sigma_K^2 \sigma_I^2} \log^2\left(\frac{\tilde{V}_i}{\gamma d_i^{-6}(\mathbf{x})} \right) \right\} \right]^{-1} . \quad (3.60) \end{aligned}$$

Das vollständige Gibbs-Schema ist in Abbildung 3.25 dargestellt: Für die Erzeugung einer neuen Stichprobe in Schritt $i{+}1$ werden σ_K^2, γ und ω nacheinander von Gln. (3.57), (3.58) bzw. (3.59) gezogen, gefolgt von der Berechnung des neuen Satzes von Klassenzugehörigkeitsparameter $\{z_i\}$ gemäß Gl. (3.60). Die Erzeugung einer neuen Konformation erfolgt wie bisher durch Hybrid-Monte-Carlo.

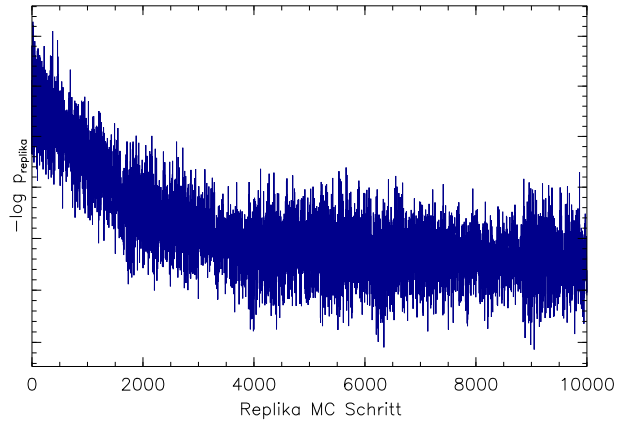

Abbildung 3.26: Konvergenz der Replika-Simulation. Die Konvergenzphase erstreckt sich über die ersten 4000 Stichproben.

3.3.3 Testrechnung I: BPTI

Um die Arbeitsweise des Mischmodells sowie die Auswirkungen der realistischeren Beschreibung der Daten auf die Qualität einer Struktur zu demonstrieren, berechnete ich die Strukturverteilung von BPTI auf Basis des Modelldatensatzes aus Kapitel 2.5. Da Multispineffekte weder im Mischmodell noch bei der Berechnung der simulierten Kreuzrelaxationsraten berücksichtigt wurden, stellt der gewählte Datensatz eine geeignete Testumgebung für das neue Datenmodell dar: Abweichungen von beobachteten und berechneten Kreuzrelaxationsraten sind rein auf die interne Dynamik von BPTI zurückzuführen, so daß eine Analyse der Klassifikationsparameter zusätzlich Rückschlüsse über die Dynamikbehaftung jeder Einzelmessung gestattet. Ich vergleiche die Ergebnisse mit der Simulation von BPTI aus Kapitel 3.1.4.2, bei der die Beschreibung der observierten Kreuzrelaxationsraten auf Basis des nicht-klassifizierenden (Lognormal)-Datenmodells erfolgte.

Simulation der Strukturverteilung

Um das Mischmodell zu vervollständigen, mußte ein geeigneter Wert für den Formparameter der I-Komponente spezifiziert werden. Als konservative

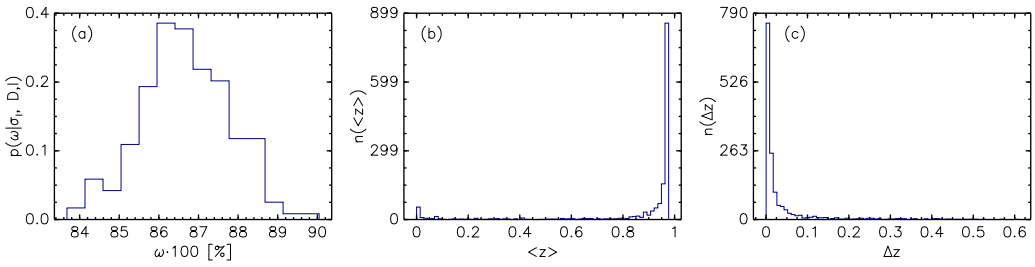

Abbildung 3.27: Klassifikationsparameter des Mischmodells: (a) Marginale *a-posteriori*-Verteilung für den Mischparameter ω. (b) Histogramm für die Mittelwerte aller 1543 Klassenzugehörigkeitsparameter. Alle Messungen wurden hinreichend eindeutig als konsistent oder inkonsistent klassifiziert. (c) Unsicherheitsbehaftung der Klassenzugehörigkeitsparameter. Δz bezeichnet die Größe des 68%-Konfidenzintervalls.

Abschätzung nahm ich an, daß interne Dynamik zu mittleren relativen Fehlern von bis zu 75% in den Zieldistanzen führen kann, was einem Wert von $\sigma_I = 4.5$ entspricht.
Die Zahl der unbekannten Größen übersteigt die Zahl der Daten signifikant: Neben den 256 Dihedralwinkelfreiheitsgraden mußten die Klassenzugehörigkeitsparameter aller Messungen aus den Daten bestimmt werden: Für den Modelldatensatz sind dies 1543 zusätzliche Parameter. Die zu simulierende Wahrscheinlichkeitsverteilung besitzt insgesamt 1802 Dimensionen, was 0.85 Messungen pro Parameter entspricht.
Ich simulierte die *a-posteriori*-Verbundverteilung für 10000 MC-Schritte. Trotz der großen Zahl der zu schätzenden Hypothesenparameter konvergierte die Markov-Kette rasch zu ihrer Gleichgewichtsverteilung (vgl. Abb. 3.26). Die Konvergenzphase erstreckte sich über die ersten 4000 Stichproben.

3.3.3.1 Klassifikation und Dynamikbehaftung

Das Klassifikationsverhalten des Mischmodells wurde durch Analyse der Klassenzugehörigkeitsparameter sowie des Mischparameters ω untersucht. ω quantifiziert den Bruchteil der Messungen, die nicht anhand einer starren Struk-

tur erklärt werden können. Die Klassenzugehörigkeitsparameter $\{z_i\}$ geben ferner Auskunft darüber, aus welchen Messungen sich dieser Bruchteil zusammensetzt. $\langle\omega\rangle = (86.5 \pm 1.0)\%$ der Daten wurden der K-Komponente zugeordnet (vgl. Abb. 3.27(a)); die verbleibenden Messungen wurden als inkonsistent klassifiziert und durch die I-Komponente des Modells beschrieben. Die Mehrheit der Daten wird durch das Mischmodell eindeutig als „konsistent“ ($\langle z\rangle \approx 1$) oder „inkonsistent“ ($\langle z\rangle \approx 0$) klassifiziert (Abb. 3.27(b)).

Die Mehrheit der Klassenzugehörigkeitsparameter wird durch die Daten mit einer hohen Genauigkeit festgelegt: Für mehr als 90% der Parameter beträgt die Größe des 68%-Konfidenzintervalls weniger als 0.1 (s. Abb. 3.27(c)).
Um zu untersuchen, ob die Klassenzugehörigkeit der Messungen Rückschlüsse auf die Abweichung von Ziel- und Referenzdistanzen gestattet, zerlegte ich den Datensatz in drei Gruppen:

1. Konsistente Messungen: $z_i \geq 0.9$ (80% der Daten);
2. Inkonsistente Messungen: $z_i \leq 0.1$ (6% der Daten);
3. Nicht-klassifizierte Messungen: $0.1 < z_i < 0.9$ (14% der Daten).

In Abbildung 3.28 sind die Referenzdistanzen d_{ref} gegen die Zieldistanzen d_{exp} aufgetragen und entsprechend ihrer Gruppenzugehörigkeit eingefärbt. Zieldistanzen wurden gemäß $d_{\mathrm{exp}} = \tilde{V}^{-1/6}$ berechnet und anschließend bezüglich der Referenzdistanzen kalibriert. Die Dynamikbehaftung einer Messung äußert sich in einer Abweichung von Ziel- und Referenzdistanz, wobei die Referenzdistanz tendenziell unterschätzt wird (vgl. Abb. 3.28). Die Abbildung zeigt deutlich, daß dynamikbehaftete Meßwerte in der Mehrzahl als „inkonsistent“ klassifiziert wurden. Konsistente Messungen streuen hingegen um die korrekten Distanzen. Nicht-klassifizierte Messungen sind in Grau dargestellt. Auf exakte numerische Werte für die Abweichung von Ziel- und Referenzdistanzen konnte aus der Größe der Klassenzugehörigkeitsparameter nicht geschlossen werden; eine eindeutige Korrelation zwischen beiden Größen wurde nicht beobachtet.

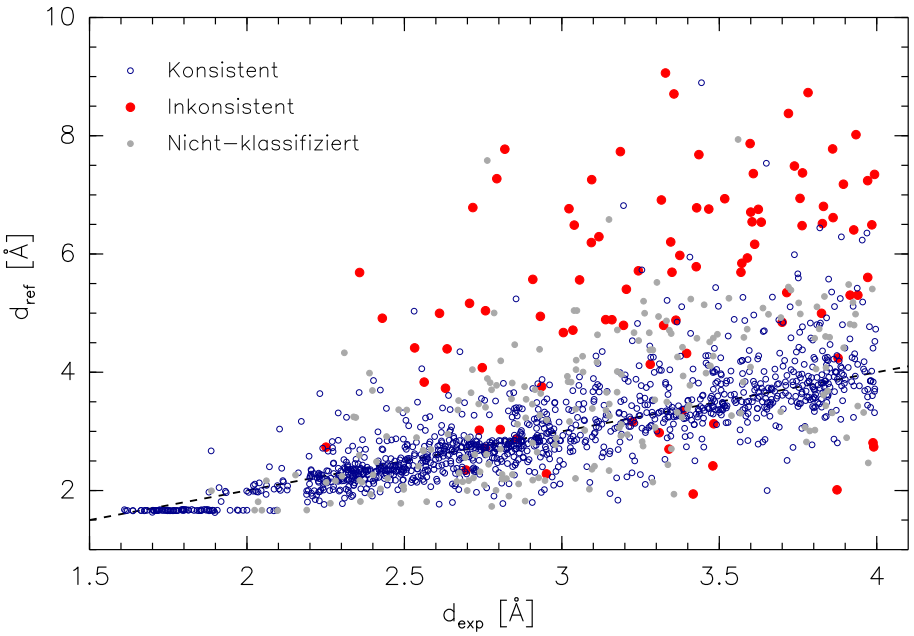

Abbildung 3.28: Identifikation dynamikbehafteter Messungen. Distanzen aus der Referenzstruktur sind gegen die Zieldistanzen aufgetragen. Inkonsistene Messungen sind rot, konsistente Messungen blau und nicht-klassifizierte Messungen grau dargestellt. Dynamikbehaftete Messungen zeigen systematische Abweichungen von den Referenzdistanzen und wurden in der Mehrzahl identifiziert.

3.3.3.2 Strukturelle Qualität

Die *Likelihood*-Funktion des Mischmodells gewichtet Abweichungen von observierten und theoretischen Kreuzrelaxationsraten für jede Messung individuell. Die Stärke der Gewichtung (σ_{eff}^{-2} aus Gl. (3.52)) wird von der Klassenzugehörigkeit der jeweiligen Messung bestimmt. Inkonsistente Meßwerte tragen dabei überproportional wenig zur *Likelihoood*-Funktion bei (vgl. Abb. 3.29(a)) – die Schlüssigkeit einer Konformation mit den Daten wird somit primär auf Basis der konsistenten Messungen bewertet. Verglichen mit dem nicht-klassifizierenden Datenmodell werden lokale Verzerrungen in der Struktur dadurch vermindert und systematische Abweichungen von der Referenzstruktur deutlich reduziert: Die Abweichung der wahrscheinlichsten Konformation zur Referenzstruktur verbessert sich auf 0.69 Å (erwartete Abwei-

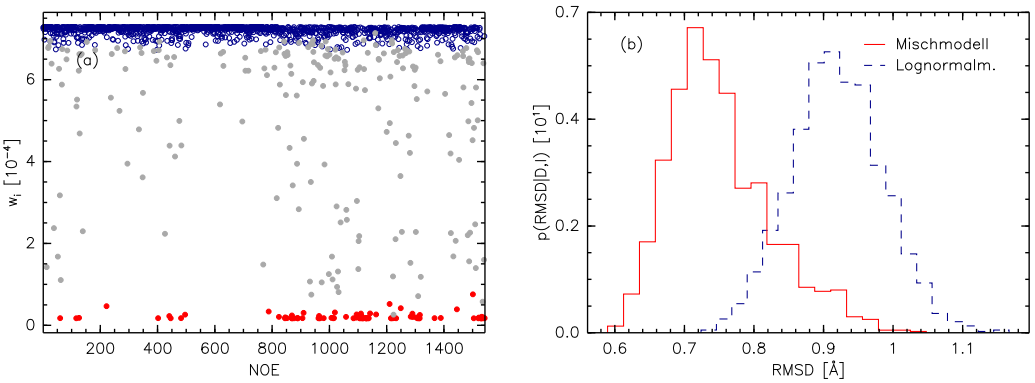

Abbildung 3.29: Einfluß der Datengewichtung auf die Strukturqualität. (a) Individuelle Gewichte $w_i = \sigma_{\mathrm{eff},i}^{-2} / \sum_i \sigma_{\mathrm{eff},i}^{-2}$ für konsistente (blau), inkonsistente (rot) und nicht-klassifizierte Messungen (grau). (b) Marginale *a-posteriori*-Verteilung für den CA-RMSD zur Referenzstruktur.

chung 0.75 ± 0.08 Å) (nicht-klassifizierendes Datenmodell: 0.85 Å, erwartete Abweichung 0.92 ± 0.07 Å) (vgl. Abb. 3.29(b)).

Verbesserungen ergaben sich auch in Hinblick auf charakteristische Eigenschaften gefalteter Proteine: Für die 100 wahrscheinlichsten Konformationen berechnete ich mit Hilfe der Programme PROCHECK [26], WHATIF [58] und PROSA [30] eine Reihe von Qualitäts-Indikatoren (s. Tab. 3.2). Das Mischmodell führt zu einer deutlich realistischeren ϕ/ψ-Verteilung (RAMCHK und PROCHECK Ramachandran-Statistiken) und verbessert die Pakkung der Struktur (QUACHK, NQACHK), was auf eine Verminderung lokaler Verzerrungen hindeutet. Für die Rückgratkonformation (BBCCHK) ergibt sich keine Verbesserung. Die wissensbasierte PROSA-Energie bewertet die Schlüssigkeit der Aminosäuresequenz mit der Struktur. Auch für diesen Index erzeugt das Mischmodell einen deutlich besseren Wert.

Qualitäts-Indikator [**z-***score*]	**Mischmodell**	**Lognormalmodell**
QUACHK[a]	-3.73 ± 0.23	-4.35 ± 0.26
NQACHK[a]	-2.87 ± 0.66	-3.10 ± 0.53
RAMCHK[b]	-4.41 ± 0.53	-4.77 ± 0.55
BBCCHK[c]	-2.37 ± 0.77	-2.08 ± 0.78
Ramachandran ϕ/ψ Statistik[d] **[%]:**		
Bevorzugte Region	76.1 ± 4.4	71.7 ± 4.7
Zusätzlich erlaubte Region	23.9 ± 4.5	26.9 ± 4.8
Wissensbasierte Energiefunktion:		
Paarweises Potential[e]	-1.43 ± 0.31	-0.97 ± 0.18
RMSD[f] **[Å]**		
Wahrscheinlichste Struktur	0.69	0.85
Erwartet	0.75 ± 0.08	0.92 ± 0.07

WHATIF: [a] Packungsqualität, [b] ϕ/ψ Verteilung, [c] Rückgratkonformation.
[d] Berechnet mit dem Programm PROCHECK, ohne Glycin und Prolin.
[e] Berechnet mit dem Programm PROSA.
[f] Relativ zur Referenzstruktur `Bref`

Tabelle 3.2: Strukturelle Qualitäts-Indikatoren für BPTI. Vergleich von Misch- und Lognormalmodell. Für die Indikatoren (^{a,b,c,d,e}) sind jeweils die Mittelwerte bezüglich der 100 wahrscheinlichsten Konformationen inklusive Standardabweichung angegeben.

3.3.3.3 Datenkonsistenz

Um zu demonstrieren, daß die als konsistent klassifizierten Messungen mit der Hypothese einer starren Struktur deutlich besser in Einklang stehen als die Gesamtheit aller Messungen, berechnete ich das Datenkonsistenzmaß aus Abschnitt 3.2, σ_d, bezüglich der Meßwerte in der K-Komponente: Dabei ergab sich eine mittlere relative Abweichung der Zieldistanzen von den korrespondierenden Distanzen in der Struktur von $\sigma_{K,d} = 11.3 \pm 0.5$ % (s. Abb. 3.30(a)). Für den Gesamtdatensatz ist die Abweichung mit $\sigma_d = 20 \pm 0.4$ % beinahe doppelt so hoch. Dies zeigt, daß bereits ein geringer Anteil inkon-

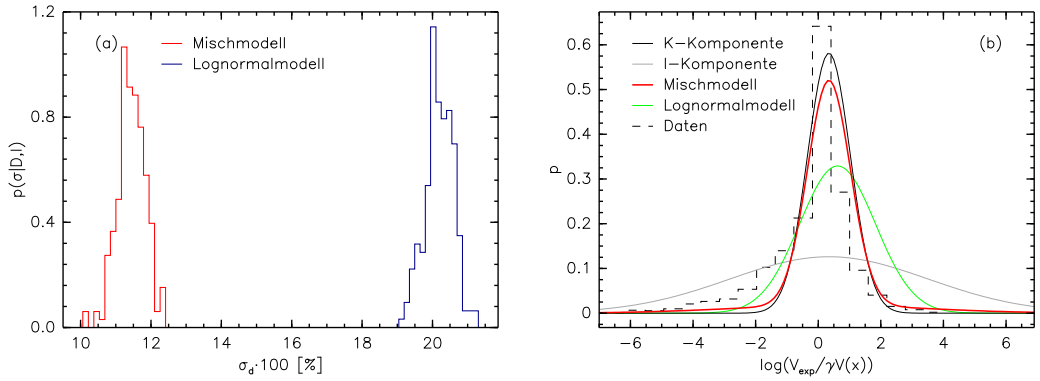

Abbildung 3.30: Auswirkung der Datenklassifikation auf Konsistenz und Fehlerverteilung. (a) Datenkonsistenz σ_d für konsistente Messungen (rot, Mischmodell) im Vergleich zum nicht-klassifizierenden Lognormalmodell (blau). (b) Vergleich der Fehlerverteilungen des Mischmodells, des Lognormalmodells und der experimentellen Fehlerverteilung.

sistenter Messungen (im Falle des Modelldatensatzes lediglich 6%) zu einer signifikant niedrigeren Konsistenz eines Datensatzes führen kann, obwohl die überwiegende Mehrheit der Daten keine signifikanten Unschlüssigkeiten aufweist. Vor diesem Hintergrund ist zu erwarten, daß sich das Datenkonsistenzmaß bei der Strukturrechnung mit nicht-zugeordneten NOE-Datensätzen als nützlich erweisen kann: Bei der praktischen Arbeit mit Verfahren für die automatische Zuordnung von NOESY-Spektren zeigt sich, daß Fehlzuordnungen oder Rauschartefakte auch in den späten Phasen einer Strukturrechnung häufig unvermeidbar sind. Die Sensitivität von σ_d gegenüber Inkonsistenzen kann dabei helfen, Probleme dieser Art über einen Vergleich mit erwarteten Werten von σ_d zu identifizieren, die sich für typische, „saubere“ Datensätze ergeben: Rechnungen mit anderen Testsystemen deuten darauf hin, daß der Wert von $\sigma_d \approx 20\%$ für saubere Datensätze und Proteine mittlerer Größe (bis 100 Aminosäuren) universell zu sein scheint.

Die separate Beschreibung von konsistenten und inkonsistenten Messungen spiegelt sich auch in der Form der Fehlerverteilung der beiden Datenmodelle

wider. Abbildung 3.30(b) zeigt die Fehlerverteilung des Mischmodells und des Lognormalmodells im Vergleich zur experimentellen („wahren“) Verteilung, die aus dem Datensatz mit Hilfe der Referenzstruktur bestimmt wurde. Die Verteilungen beziehen sich jeweils auf die Diskrepanz von observierter und berechneter Kreuzrelaxationsrate. Konsistente Messungen zeigen kleine Abweichungen und sind folglich um 0 konzentriert; ihre Verteilung wird durch die K-Komponente des Mischmodells mit sehr guter Genauigkeit angenähert. Inkonsistente Messungen sind im Schwanz der experimentellen Verteilung zu beobachten, der durch die I-Komponente modelliert wird. Deutlich ist zu erkennen, daß das Lognormalmodell (grün) die Breite der experimentellen Verteilung signifikant überschätzt und zudem systematisch verschoben ist.

Das Mischmodell ließe sich leicht auf mehrere Komponenten verallgemeinern, wodurch zusätzliche Details der experimentellen Fehlerverteilung im Datenmodell berücksichtigen werden könnten: Bei einem Mischmodell mit K Komponenten würde die Klassenzugehörigkeit des i-ten Meßwerts durch einen Klassenzugehörigkeitsvektor z_{ik}, $k = 1 \ldots K$ beschrieben; $z_{ik} = 1$ ordnet den i-ten Meßwert der k-ten Komponente zu. Die funktionale Form des Mischmodells in Gleichung (3.51) wäre in diesem Falle ensprechend zu verallgemeinern:

$$p(\tilde{V}_i|\{z_{ik}\},\cdot) = \prod_{k=1}^{K} \left[p_k(\tilde{V}_i|\cdot)\right]^{z_{ik}},$$

wobei $p_k(\tilde{V}|\cdot)$ das Datenmodell bezeichnet, welches der k-ten Komponente des Mischmodells zugeordnet ist. Die Zahl der zu schätzenden Parameter ist demnach von der Ordnung $O(KN)$, wobei N die Anzahl der Meßwerte bezeichnet. Ein entsprechender Gibbs-Algorithmus für die Simulation des erweiterten Mischmodells folgte unmittelbar aus der entsprechend modifizierten *a posteriori*-Verbundverteilung. Wie die Rechnungen gezeigt haben, ist die Unsicherheitsbehaftung der Klassenzugehörigkeitsparameter im Falle des zweikomponentigen Mischmodells gering. Es ist daher zu erwarten, daß auch das verallgemeinerte Mischmodell verläßlich aus den Daten geschätzt werden kann.

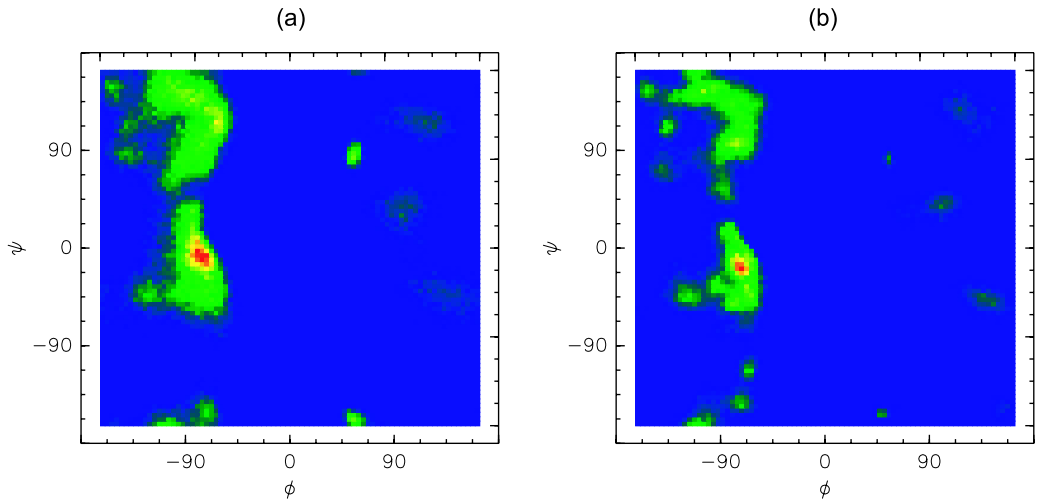

Abbildung 3.31: Effekt der Modellierung inkonsistenter Messungen auf die Strukturverteilung, dargestellt als ϕ/ψ-Diagramm. (a) Lognormalmodell. (b) Mischmodell. Im Mischmodell werden Inkonsistenzen geeignet behandelt, wodurch sich die strukturelle Unsicherheitsbehaftung deutlich verringert.

3.3.3.4 Strukturelle Unsicherheitsbehaftung

Aufgrund der Abhängigkeit der strukturellen Unsicherheitsbehaftung von der Konsistenz der Daten ist zu erwarten, daß die individuelle Gewichtung der Messungen, außer zu einer Reduktion von systematischen Fehlern, auch zu einer Verminderung der atomaren Unsicherheitsbehaftung führt. In Abbildung 3.31 ist der Einfluß der realistischeren Fehlerverteilung des Mischmodells auf die Ausdehnung der Strukturverteilung deutlich zu erkennen: Verglichen mit der Referenzsimulation ist die Wahrscheinlichkeitsmasse im Falle des Mischmodells besser lokalisiert.

Um ein detaillierteres Bild über lokale Unterschiede zwischen den beiden Strukturverteilungen zu erhalten, berechnete ich die Koordinatenunsicherheiten aller Atome mit Hilfe der in Abschnitt 3.1.2 beschriebenen Methode. Die mittlere 1σ-Unsicherheitsbehaftung fällt um mehr als 20% von $\langle\delta\rangle = 0.78 \pm 0.58$ Å (Lognormalmodell) auf $\langle\delta\rangle = 0.60 \pm 0.53$ Å für das Mischmodell. Der CA-RMSD der mittleren Struktur zur Referenzstruktur `Bref` beträgt 0.67 Å. Die realistischere Beschreibung der experimentellen Daten

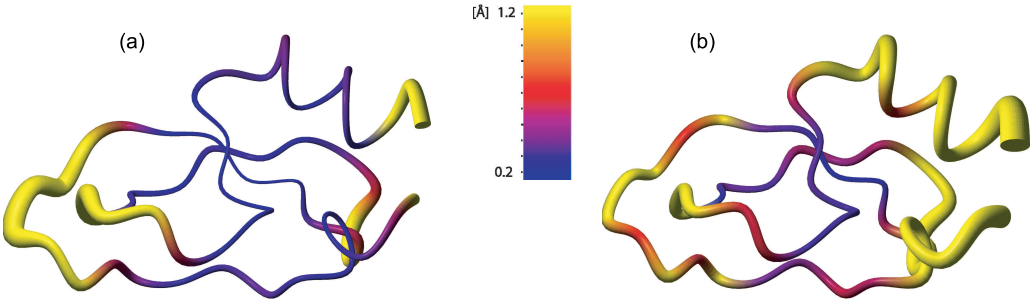

Abbildung 3.32: CA-Unsicherheitsbehaftung mit und ohne Modellierung inkonsistenter Messungen. Dargestellt sind mittlere Strukturen, Unsicherheitsbehaftung ist in Strichstärke und Farbe korrespondiert. (a) Das Mischmodell führt zu einer schärferen Strukturverteilung mit realistischer Streuung. (b) Das Lognormalmodell produziert unspezifischere Unsicherheiten.

verbessert damit sowohl die Richtigkeit als auch die Genauigkeit der generierten Koordinaten. Die hohe Zahl der zu bestimmenden Parameter des Datenmodells zieht also keine erhöhte Unsicherheitsbehaftung der Struktur nach sich. Die auf Basis des Mischmodells berechnete Struktur (Abb. 3.32(a)) ist sichtbar genauer definiert. Im Falle des nicht-klassifizierenden Datenmodells (Abb. 3.32(b)) ist die Unsicherheitsbehaftung unspezifisch über die Struktur verteilt. Im Detail unterscheidet sich die Unbestimmtheit der CA-Positionen um bis zu 50% (vgl. Abb. 3.33(a)). Signifikante Abweichungen sind in den mobilen Regionen (Reste 13-15 und 39-40) und in den starren Regionen (Reste 1-8 und 41-58) zu beobachten.

Verglichen mit dem Mischmodell wird die atomare Unsicherheitsbehaftung durch das Lognormalmodell in mobilen Regionen des Moleküls unterschätzt, in starren Regionen dagegen überschätzt. Dies demonstriert den Einfluß inkonsistenter Messungen auf die räumliche Verteilung der Koordinatenunsicherheiten, wenn das Datenmodell nicht zwischen konsistenten und inkonsistenten Messungen differenziert: Im Falle des nicht-klassifizierenden Modells werden alle Messung durch dieselbe Fehlerverteilung modelliert, d.h. Abweichungen von berechneten und observierten Intensitäten werden für alle Mes-

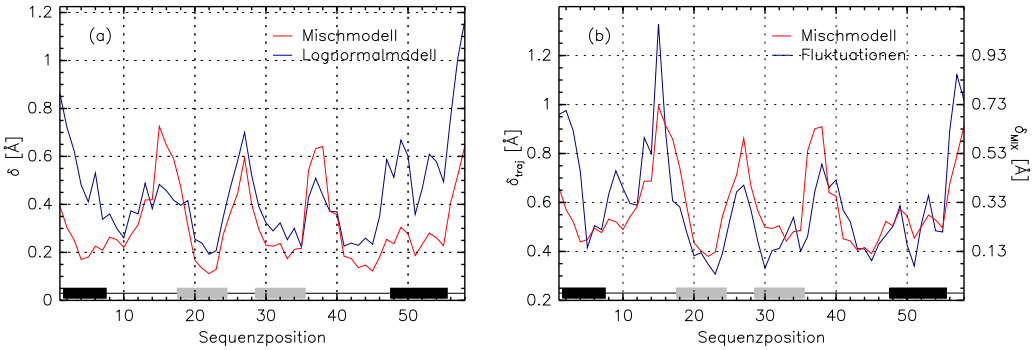

Abbildung 3.33: CA-Unsicherheiten und CA-Fluktuationen. (a) Unsicherheiten für Lognormalmodell (blau) und Mischmodell (rot). Das Lognormalmodell unterschätzt (überschätzt) Unsicherheiten in mobilen (starren) Regionen. (b) Fluktuationen in der Trajektorie (blau, linke Ordinate) und Unsicherheiten für Mischmodell (rot, rechte Ordinate). Mittelwerte beider Kurven wurden überlagert. *β-strands* sind grau, $\alpha/3_{10}$-Helices schwarz dargestellt.

sungen gleich gewichtet. Die Fehler in den korrespondierenden Abständen verteilen sich dabei so über die Struktur, daß die *mittlere* Abweichung minimal ist.

Das Mischmodell hingegen modelliert jede Messung durch eine individuelle Fehlerverteilung und besitzt somit die notwendige Flexibilität, um Abweichungen in den Intensitäten nach dem Grad der Erfüllbarkeit einer Messung zu gewichten. Für konsistente Messungen werden dabei geringere Abweichungen toleriert, als für inkonsistente Meßwerte. Die lokale Unsicherheitsbehaftung der Struktur hängt demzufolge davon ab, ob aus der betreffenden Region des Moleküls vermehrt konsistente- oder inkonsistente Messungen hervorgegangen sind. Im Falle des Modelldatensatzes „verstärken“ sich konsistente Messungen aus starren Regionen, was zu einer entsprechend niedrigeren Unsicherheitsbehaftung dieser Bereiche führt; inkonsistente Meßwerte aus mobilen Regionen werden dagegen schwächer gewichtet, was die erhöhte Koordinatenunsicherheit erklärt.

Die 1σ-Unsicherheiten der CA-Positionen zeigen eine gute positionsweise

Übereinstimmung mit der Stärke der CA-Fluktuationen (Korrelationskoeffizient 0.68), die aus der MD Trajektorie berechnet wurden (vgl. Abb. 3.33(b)). Insbesondere werden die signifikanten Fluktuationen in den flexiblen *loop*-Regionen sowie die geringe Mobilität der C-terminalen α-Helix reproduziert. Die Größe der Fluktuationen wird im Mittel um 0.26 Å unterschätzt. Aufgrund der Abhängigkeit struktureller Unsicherheiten von der Größe eines Datensatzes (vgl. Kap. 3.2.3.3) ist eine Skalengleichheit jedoch nicht zu erwarten.

3.3.4 Testrechnung II: SMN Tudor Domäne

^{15}N Relaxationsexperimente und kristallographische B-Faktoren zeigen, daß die SMN Tudor Domäne eine sehr geringe Hauptkettenflexibilität aufweist [62]. Ich wählte die Tudor Domäne als Testsystem, um das Verhalten des Mischmodells für NOESY-Daten zu studieren, welche eine schwache Dynamikbehaftung, jedoch nichtverschwindende Spindiffusionsanteile und Inkonsistenzen aufgrund von heteronuklearer Relaxation aufweisen. Die Berechnung der Strukturverteilung erfolgte auf Basis der in Kapitel 2.5 beschriebenen Datensätze. Beide Datensätze wurden durch individuelle Mischmodelle beschrieben und simultan in der Strukturrechnung verwendet. Die Simulation der Tudor Domäne aus Kapitel 3.1.4.2, bei der die Beschreibung der observierten Kreuzrelaxationsraten auf Basis des nicht-klassifizierenden (Lognormal)-Datenmodells erfolgte, dient als Referenzsimulation.

Simulation der Strukturverteilung

Verglichen mit dem BPTI-Datensatz konnte die erwartete Diskrepanz in den Intensitäten inkonsistenter Messungen verringert werden: Bei Rechnungen mit Werten für σ_I, welche einem relativen Fehler in den Zieldistanzen von 75% und 50% entsprechen, wurden jeweils alle Messungen der K-Komponente zugeordnet, d.h. als konsistent klassifiziert. Das Mischmodell reduzierte sich in diesem Fall auf das Lognormalmodell. Die Stärke der Inkonsistenzen ist im

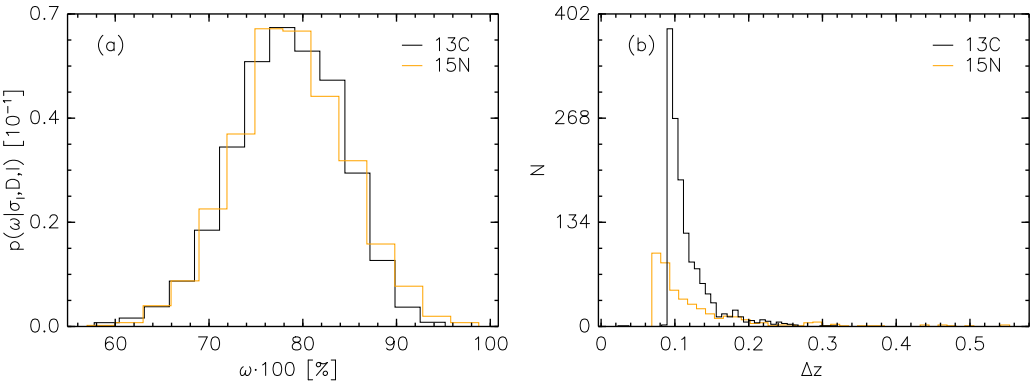

Abbildung 3.34: Klassifikationsparameter des Mischmodells für die Datensätze ^{13}C und ^{15}N. (a) Marginale *a-posteriori*-Verteilungen für die Mischparameter. (b) Unsicherheitsbehaftungen (68% Konfidenzintervall) Δz der Klassenzugehörigkeitsparameter; Histogramme beziehen sich auf die Gesamtheit aller Messungen.

Falle der experimentellen Datensätze somit deutlich geringer. Für den Produktionslauf wählte ich den Wert $\sigma_I = 0.45$, was der Annahme entspricht, daß inkonsistente Messungen relative Distanzfehler in der Größenordnung von 25% aufweisen können. Die Zahl der zu schätzenden Hypothesenparametern betrug 2145, was etwa 0.87 Messungen pro Parameter entspricht. Alle 50 Kopien der Replika-Kette wurden für 8000 Schritte parallel simuliert. Die Länge der Konvergenzphase betrug 3000 Schritte.

3.3.4.1 Klassifikationsverhalten

Aus den *a-posteriori*-Stichproben wurden marginale *a-posteriori*-Verteilungen für alle Hypothesenparameter bestimmt, mittels derer das Klassifikationsverhalten des Mischmodells analysiert wurde.
Die Konsistenz beider Datensätze ist von vergleichbarer Größe: Jeweils 78% der Messungen wurden als konsistent klassifiziert; 22% der Daten zeigten sich mit der Hypothese einer starren Struktur hingegen unvereinbar und wurden der I-Komponente des Modells zugeordnet (vgl. Abb.3.34(a)).
Abb. 3.34(b) zeigt die Häufigkeitsverteilungen für die Unsicherheitsbehaftung

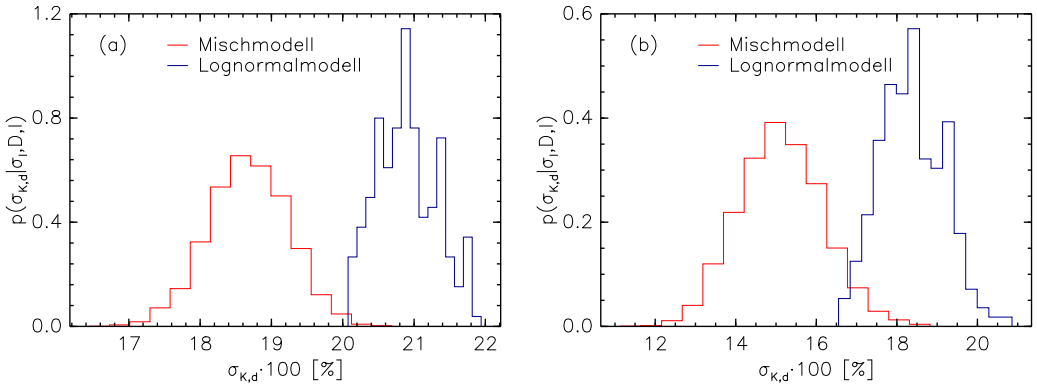

Abbildung 3.35: Genauigkeit der Datenmodellierung. (a) Verteilungen für das Datenkonsistenzmaß σ_d für den ^{13}C Datensatz. (b) dito für den ^{15}N Datensatz. Das Mischmodell ist rot, das Lognormalmodell blau dargestellt.

Δz der Klassenzugehörigkeitsparameter, definiert als Größe des 68% Konfidenzintervalls der jeweiligen marginalen *a-posteriori*-Verteilung. Wie schon bei der Rechnung für BPTI ist der Gibbs-Algorithmus stabil und das Mischmodell kann trotz der hohen Zahl unbekannter Parameter verläßlich aus den Daten geschätzt werden. Für 90% der z_i liegt die Unsicherheitsbehaftung unter 0.15 (^{13}C Datensatz) bzw. unter 0.2 (^{15}N Datensatz).

Verglichen mit dem nicht-klassifizierenden Modell führt die individuelle Gewichtung der Daten zu einer besseren Übereinstimmung konsistente Messungen mit den aus der Struktur berechneten Werten; die mittlere relative Abweichung in den Distanzen reduziert sich für beide Datensätze (vgl. Abb. 3.35). Verglichen mit der Rechnung für BPTI aus Abschnitt 3.3.3 fällt der Gewinn an Genauigkeit jedoch geringer aus. Dies ist auf die Vernachlässigung von Multispineffekten bei der Berechnung des BPTI-Datensatzes zurückzuführen: Der BPTI-Datensatz besitzt daher eine hohe Konsistenz, die ausschließlich durch das Auftreten stark dynamikbehafteter Messungen reduziert wird. Mit anderen Worten lassen sich inkonsistente Messungen von den konsistenten Messungen hinreichend eindeutig trennen. Die hohe Konsistenz des

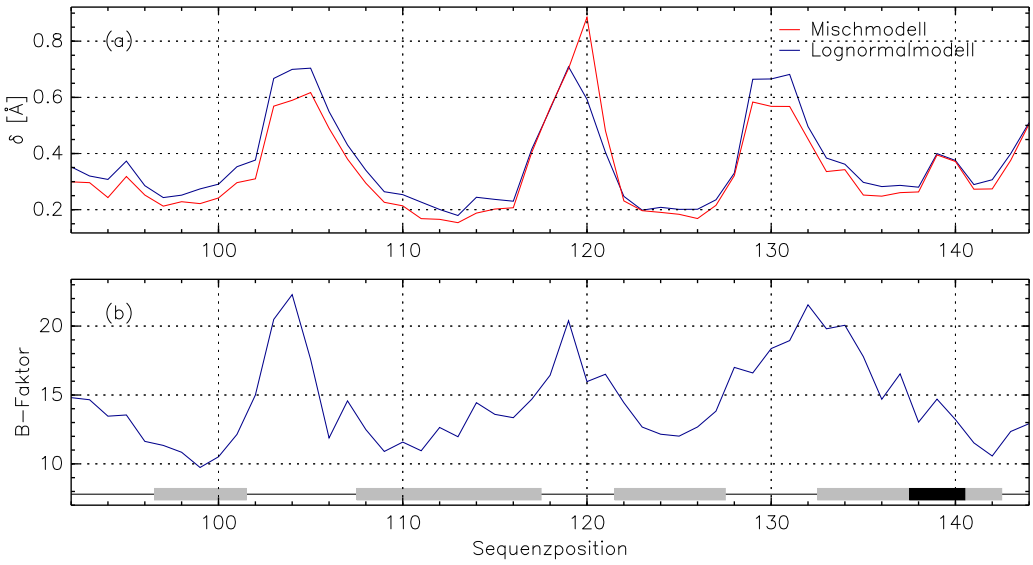

Abbildung 3.36: Unsicherheitsbehaftung der berechneten NMR Struktur im Vergleich zur Kristallstruktur. (a) 1σ Unsicherheiten in den CA-Koordinaten der NMR-Struktur berechnet aus den Simulationen auf Basis des Mischmodells (rot) und des Lognormalmodells (blau). (b) B-Faktoren der CA-Positionen der Kristallstruktur. *β-strands* sind grau, α-Helices schwarz dargestellt.

Datensatzes spiegelt sich daher in den Messungen der K-Komponente wider. Im Falle der Tudor Domäne besitzen die experimentell bestimmten Kreuzrelaxationsraten hingegen nichtverschwindende Spindiffusionsanteile. Die Berechnung der theoretischen Intensitäten auf Basis der ISPA führt somit zu einer grundauf höheren Inkonsistenz der beiden experimentellen Datensätze. Wie die Rechnungen in Abschnitt 3.2.3 gezeigt haben, ist die effektive Konsistenz (bezüglich aller Messungen) des BPTI-Datensatzes und der Tudor Datensätze von etwa gleicher Größe. Im Falle der Tudor Domäne ist die effektive Konsistenz jedoch nicht auf wenige Messungen, sondern auf die Gesamtheit aller Messungen zurückzuführen. Inkonsistente Meßwerte lassen sich daher nicht so eindeutig identifizieren wie im Falle des BPTI-Datensatzes.

3.3.4.2 Strukturelle Qualität

Aus den konformationellen Stichproben wurden die Unsicherheitsbehaftungen aller Atome extrahiert. Die Positionen der CA-Atome sind im Falle des

Qualitäts-Indikator [**z-***score*]	**Mischmodell**	**Lognormalmodell**
QUACHK[a]	-3.10 ± 0.31	-3.35 ± 0.28
NQACHK[a]	-2.15 ± 0.52	-2.73 ± 0.40
RAMCHK[b]	-2.29 ± 0.71	-2.32 ± 0.57
BBCCHK[c]	0.41 ± 0.67	-0.66 ± 0.70
Ramachandran ϕ/ψ Statistik[d] **[%]:**		
Bevorzugte Region	75.0 ± 4.5	77.1 ± 4.3
Zusätzlich erlaubte Region	22.9 ± 4.8	20.8 ± 4.9
Wissensbasierte Energiefunktion:		
Paarweises Potential[e]	-1.71 ± 0.17	-1.67 ± 0.16
RMSD[f] **[Å]**		
Mittlere Struktur (CA,N,C',O)	0.78	0.86

WHATIF: [a] Packungsqualität, [b] ϕ/ψ Verteilung, [c] Rückgratkonformation.

[d] Berechnet mit dem Programm PROCHECK, ohne Glycin und Prolin.

[e] Berechnet mit dem Programm PROSA.

[f] Relativ zur Kristallstruktur 1mhn für die Reste 2-54.

Tabelle 3.3: Strukturelle Qualitäts-Indikatoren für die Tudor Domäne. Vergleich von Misch- und Lognormalmodell. Für die Indikatoren (^{a,b,c,d,e}) sind jeweils die Mittelwerte bezüglich der 100 wahrscheinlichsten Konformationen inklusive Standardabweichung angegeben.

Mischmodells besser bestimmt (vgl. Abb. 3.36(a)); Unterschiede in der Unsicherheitsbehaftung sind jedoch gering. Die CA-Unsicherheiten zeigen eine sehr gute positionsweise Übereinstimmung mit den B-Faktoren der Kristallstruktur (vgl. Abb. 3.36(b)) (Korrelationskoeffizient 0.69).

Auch im Falle der Tudor Domäne werden systematische Fehler in der Strukturverteilung durch die modellseitige Berücksichtigung von Inkonsistenzen in den Daten reduziert. Der CA-RMSD der wahrscheinlichsten Struktur zur Kristallstruktur sinkt auf 0.75 Å (0.81 Å für das Lognormalmodell). Für die mittlere Struktur ergaben sich vergleichbare Werte (vgl. Tab. 3.3).

Eine Verbesserung zeigte sich auch bei allgemeinen Qualitäts-Indikatoren, die

mit Hilfe der Programme PROCHECK, WHATIF und PROSA für die 100 wahrscheinlichsten Konformationen berechnet wurden. Tabelle 3.3 faßt die Ergebnisse zusammen: Das Mischmodell führt zu einer deutlich verbesserten Packungsqualität (QUACHK, NQACHK), zu einer realisitischeren Rückgratkonformation (BBCCHK) und verbessert die Schlüssigkeit der Aminosäuresequenz mit der Struktur (PROSA-Energie). Die ϕ/ψ-Verteilung (RAMCHK) sowie PROCHECK Ramachadran-Statistiken weisen in beiden Fällen für NMR-Strukturen sehr gute Werte auf.

Kapitel 4

Diskussion

In der Strukturbiologie hat sich die Kernspinresonanz-Spektroskopie neben der Röntgenkristallographie als zweite Standardmethode zur experimentellen Strukturaufklärung etabliert. Die Frage nach der Qualität einer NMR-Struktur ist seit den Anfängen von Strukturbestimmung durch hochaufgelöste NMR wiederholt gestellt worden: NMR-Daten sind unvollständig und schwierig zu interpretieren; insbesondere die Analyse des NOE gestaltet sich komplex, da seine Intensität von zahlreichen Effekten abhängt, die theoretisch nicht oder nur näherungsweise beschrieben werden können.
Für die Bestimmung der Qualität einer NMR-Struktur ist daher quantitativ zu klären, wie sich Unsicherheiten in den experimentellen Daten und Approximationen in den theoretischen Modellen auf die Genauigkeit auswirken, mit der die Positionen der einzelnen Atome einer Struktur berechnet werden können. Diese Fragestellung ist mit konventionellen Methoden zur Strukturbestimmung nicht in befriedigender Weise beantwortbar: Standardmethoden versuchen die „wahre“ Konformation eines Moleküls durch Inversion der Daten zu berechnen, d.h. Strukturbestimmung wird mathematisch als Inversionsproblem formuliert. Unschlüssigkeiten zwischen Theorie und Experiment sowie resultierende Unsicherheiten in den dreidimensionalen Koordinaten einer Struktur bleiben in diesem Zugang per Definition unberücksichtigt bzw. können nicht repräsentiert werden. Ein zusätzliches Problem stellen freie Pa-

rameter dar, die für die Berechnung des theoretischen Wertes einer Meßgröße oder die Formulierung der Hybridenergiefunktion eingeführt werden müssen: Konventionelle Methoden bestimmen diese Parameter mittels Heuristiken und enthalten daher stets subjektive Elemente. Dies führt dazu, daß die generierten Koordinaten empfindlich von der Wahl der Heuristiken oder speziellen Parametereinstellungen (wie beispielsweise dem Wert von „Kraftkonstanten“ oder dem Protokoll zur Minimierung der Hybridenergie) abhängen. Eindeutige und objektive Aussagen im Hinblick auf die Unsicherheitsbehaftung der generierten Koordinaten sind mit konventionellen Strukturberechnungsmethoden daher nicht möglich.
Die induktive Strukturbestimmung vermeidet diese Problematik, indem die Unvollständigkeit der Ausgangsinformation bei der mathematischen Formulierung des Strukturbestimmungsproblems berücksichtigt wird: Strukturbestimmung wird als Induktionsproblem aufgefaßt und formal mit Hilfe der Bayes'schen Wahrscheinlichkeitstheorie gelöst. Anstelle nach der „wahren“ Konformation eines Moleküls zu fragen, steht die Bewertung aller möglichen Konformationen eines Moleküls im Vordergrund. Die Bewertung erfolgt dabei auf Basis der verfügbaren Ausgangsinformation, d.h. experimenteller Evidenz und *a-priori*-Wissen. Ein Datenmodell beschreibt Unschlüssigkeiten zwischen Theorie und Experiment, wodurch die Qualität eines Datensatzes explizit in der Strukturrechnung berücksichtigt wird. Das Prinzip der induktiven Strukturbestimmung diente in der vorliegenden Arbeit als Grundlage für die Behandlung der folgenden Problemstellungen:

1. Die Entwicklung einer Methode zur Berechnung der atomweisen Unsicherheitsbehaftung der dreidimensionalen Koordinaten einer NMR-Struktur;
2. Die Bestimmung der Qualität des für die Berechnung einer Struktur verwendeten NOE-Datensatzes;
3. Die Entwicklung eines Datenmodells für inkonsistente NOE-Datensätze zur Verminderung systematischer Fehler und Unsicherheiten in den berechneten Koordinaten.

4.1 Verläßlichkeit einer NMR-Struktur

4.1.1 Darstellung von struktureller Unsicherheit

Ergebnis einer induktiven Strukturrechnung ist eine Wahrscheinlichkeitsverteilung für die dreidimensionalen Koordinaten aller Atome des Zielmoleküls. Der Bayes'sche Wahrscheinlichkeitsbegriff interpretiert Wahrscheinlichkeiten als Grad der persönlichen Überzeugung in Hinsicht auf den Wahrheitsgehalt einer Hypothese. Die Verteilung der konformationellen Wahrscheinlichkeiten ist daher ein Maß für das Unwissen über die wahre Konformation des Makromoleküls: Die Ausdehnung der Strukturverteilung spiegelt die Aussagekraft der experimentellen Daten in Verbindung mit der berücksichtigten Hintergrundinformation wider und läßt sich bildlich im Sinne eines statistischen Fehlerbalkens interpretieren. Über die Richtigkeit der dreidimensionalen Koordinaten, d.h. über systematische Abweichungen von einer externen Referenzstruktur, läßt sich auf Basis der Strukturverteilung naturgemäß keine Aussage treffen.
Bedingte Wahrscheinlichkeiten bewerten logische Implikationen, denen nicht notwendigerweise ein kausaler Zusammenhang zugrundeliegen muß. Aus der Verteiltheit der dreidimensionalen Koordinaten kann daher nicht zwangsläufig auf die Existenz realer Fluktuationen des Moleküls geschlossen werden. Strukturellen Unsicherheiten können durchaus dynamische Phänomene zugrundeliegen: So kann die Bewegung eines Moleküls die Vollständigkeit eines NOESY-Datensatzes reduzieren und dadurch unmittelbaren Einfluß auf Schärfe der Strukturverteilung nehmen. Jedoch gehen in die beobachtete Varianz *a posteriori* neben dynamischen auch andere, in Kapitel 1.2.2 besprochene Unsicherheitsfaktoren ein, welche nur bei expliziter Modellierung voneinander trennbar sind.

4.1.2 Objektivität

Ein expliziter Ausdruck der Strukturverteilung wurde für den Fall zugeordneter NOESY-Daten hergeleitet. Für eine objektive Interpretierbarkeit einer Strukturverteilung hat sich das wahrscheinlichkeitstheoretische Konzept

der Marginalisierung von großer Wichtigkeit erwiesen: Die Beschreibung der Fehlerverteilung und die Interpretation der experimentellen Messungen erforderte die Einführung von zwei Hilfsparametern: σ quantifiziert die Größe der Abweichung der observierten von den berechneten Kreuzrelaxationsraten, γ beschreibt die Skala der observierten Werte. Beide Größen sind *a priori* unbekannt und wurden als *Nuisance*-Parameter aufgefaßt, die durch Marginalisierung eliminiert wurden. Dies garantiert, daß die Unsicherheitsbehaftung beider Parameter vollständig und in konsistenter Weise in der Strukturverteilung berücksichtigt wird. Der analytische Ausdruck der Strukturverteilung folgte mit Hilfe der Regeln der Wahrscheinlichkeitstheorie eindeutig aus dem Datenmodell und der *a-priori*-Verbundverteilung für alle Hypothesenparameter und hängt darüber hinaus von keinem freien Parameter ab. Subjektive Elemente, die in konventionellen Methoden durch die Verwendung von Heuristiken zur Bestimmung von Hilfsgrößen unvermeidbar sind, werden per Ansatz vermieden. Die möglichen Konformationen eines Makromoleküls werden ausschließlich von den experimentellen Daten und von Hintergrundinformation festgelegt, welche für die Interpretation der Daten vonnöten ist; die Strukturverteilung ist in diesem Sinne eine objektive Darstellung von struktureller Unsicherheit.

Vergleich mit NMR-Strukturensemblen

In der konventionellen Strukturbestimmung wird strukturelle Variabilität durch NMR-Strukturensembles repräsentiert, deren Mitglieder durch wiederholtes Ausführen des jeweiligen Minimierungsprotokolls von unterschiedlichen Anfangsbedingungen erzeugt werden. NMR-Strukturensembles liegt keine quantitative Definition zugrunde: Nach der gängigen Interpretation enthält jede Konformation eine „Teilwahrheit“ über die wahre Struktur des Moleküls, weshalb alle Mitglieder des Ensembles als gleichwertig angesehen werden. Strukturelle Variabilität wird daher häufig als Maß für die Genauigkeit angesehen, mit der die Koordinaten einer Struktur aus den Daten berechnet werden können.

Die in Strukturensemblen beobachtete Varianz spiegelt zum einen Charakteristiken der Energiefläche wider; sie wird jedoch zusätzlich von den Eigenschaften des verwendeten Minimierungsprotokolls bestimmt: Ein perfekter Minimierungsalgorithmus konvergierte im Falle einer nicht-degenerierten Energiefläche, unabhängig von der gewählten Startbedingung, immer zu derselben Lösung. Die Breite eines Ensembles läßt sich durch eine gezielte Wahl von Protokollparametern daher zu weiten Teilen steuern. Die Beschaffenheit von Strukturensemblen wird darüber hinaus von der Wahl spezieller Heuristiken für die Bestimmung von Hilfsgrößen und der Wahl freier Parameter, wie beispielsweise der Größe von Kraftkonstanten oder von Abstandsschranken, beeinflußt. Die strukturellen Eigenschaften eines NMR-Ensembles sind deshalb manuell „justierbar“ und gehorchen folglich keiner analytischen Verteilung. NMR-Strukturensembles besitzten daher keine statistische Grundlage und sind nicht objektiv interpretierbar.

Die Strukturverteilung hingegen repräsentiert Unwissen über die wahre Konformation eines Makromoleküls in expliziter Form und ist somit die statistisch saubere Definition eines NMR-Strukturensembles. Unsicherheiten in den Koordinaten einer Struktur sind Teil der Lösung eines induktiven Strukturbestimmungsproblems. Sie können direkt berechnet werden und sind von der Wahl des verwendeten Algorithmus *per constructum* unabhängig: Die Theorie der Markov-Ketten formuliert eindeutige Bedingungen, wie beispielsweise das detaillierte Gleichgewicht, die von *Sampling*-Algorithmen erfüllt werden müssen. Dies garantiert, daß bei der numerischen Simulation einer Wahrscheinlichkeitsverteilung keine artifiziellen Verzerrungen im Berechnungsergebnis entstehen. Die in dieser Arbeit verwendete Monte-Carlo-Strategie zur Simulation der Strukturverteilung ließe sich somit im Prinzip durch andere *Sampling*-Algorithmen ersetzten. Das Resultat der Rechnung wäre davon nicht betroffen.

4.1.3 Unsicherheitsbehaftung von Atompositionen

Unsicherheiten in den Koordinaten der Atome können der Strukturverteilung nicht direkt entnommen werden, sondern sind implizit in ihrer Form kodiert. Es wurde eine Bayes'sche Methode vorgestellt, mit welcher die Unsicherheiten in den dreidimensionalen Koordinaten einer NMR-Struktur aus einer Strukturverteilung extrahiert werden können.
Die Methode basiert auf einem analytischen Verteilungsmodell, welches die Strukturverteilung auf Basis konformationeller Stichproben approximiert. Es erwies sich als günstig, das Modell durch eine mittlere Struktur zu parametrisieren. Der verwendete Jeffreys-*prior* kann zu einer unphysikalischen Konfiguration der mittleren Atompositionen führen. Die mittlere Struktur dient jedoch lediglich als Referenz, welche die prinzipielle Form der Strukturverteilung beschreibt, jedoch nicht von eigentlichem Interesse ist. Ihr unphysikalischer Charakter ist aus diesem Grunde unproblematisch. Die Abweichung der Atomkoordinaten von den mittleren Positionen wurde als isotrop und normalverteilt angenommen. Als natürliches Maß für die Unsicherheit einer Atomposition diente der Radius der Unsicherheitssphäre, welche jedem Atom anhand der Fehlerverteilung zugeordnet wird. Das Verteilungsmodell gestattet somit die Bestimmung der Unsicherheiten aller Atompositionen ähnlich einem kristallographischen B-Faktor.

Berechnung des Verteilungsmodells

Durch den wahrscheinlichkeitstheoretischen Ansatz zur Formulierung des Verteilungsschätzungsproblems wurden zwei Schwierigkeiten per Ansatz gelöst: Erstens, die konsistente Behandlung von Hilfsparametern, die für die Problemformulierung vonnöten, jedoch nicht von weiterem Interesse sind. Zweitens, die Angabe eines geeigneten Algorithmus zur Berechnung des Modells. Die *a-posteriori*-Verteilung eines induktiven Strukturbestimmungsproblems aus NOESY-Daten ist invariant unter Rotation und Translation des kartesischen Koordinatensystems. Unterschiede in den individuellen Koordinatensystemen konformationeller Stichproben werden im Verteilungsmodell

daher in Form von Hilfsparametern berücksichtigt. Die einzelnen Transformationen enthalten keine Information über die Form der Strukturverteilung. Rotationsmatrizen und Translationsvektoren wurden als *Nuisance*-Parameter aufgefaßt und mittels Marginalisierung eliminiert. Die beschriebene Methode hängt dadurch von keinem freien Parameter ab: Die berechneten Koordinatenunsicherheiten werden eindeutig von den konformationellen Stichproben (den „Daten“) und notwendigen Zusatzannahmen über die Form der Strukturverteilung bestimmt und sind in diesem Sinne objektiv interpretierbar.
Ein Gibbs-Algorithmus für die Bestimmung aller Hypothesenparameter folgte unmittelbar aus der analytischen Form der *a-posteriori*-Verbundverteilung. Die individuellen Rotationsmatrizen konnten in der gewählten Darstellung durch Euler-Matrizen effizient mittels Zufallszahlengeneratoren behandelt werden; eine explizite Superposition konformationeller Stichproben auf die mittlere Struktur war dadurch überflüssig.
Die Methode wurde an den Strukturverteilungen des Proteins BPTI und der Tudor Domäne des humanen SMN Proteins getestet. Die topologische Form beider Strukturverteilungen ist komplex: Bedingt durch ihre hohe Dimensionalität, multiple Moden und starke Korrelationen in den Parametern ist eine Simulation mit Standardtechniken ineffizient. Für die Simulation kam ein Replika-Austausch-Monte-Carlo-Algorithmus zur Anwendung. Die Kombination mehrerer Markov-Ketten-Monte-Carlo-Verfahren erwies sich dabei als allgemeines und mächtiges Konzept, mit dem die genannten Eigenschaften der Verteilung zuverlässig behandelt werden konnten.
Die Simulation des Verteilungsmodells demonstrierte die Effizienz und Stabilität des Gibbs-Algorithmus: Trotz der großen Anzahl unbekannter Hypothesenparameter (jeweils mehrere Tausend) konvergierten beide Markov-Ketten innerhalb kurzer Zeit zu ihren Gleichgewichtsverteilungen. Abweichungen der Strukturverteilung von der modellseitig angenommenen isotropen Form ergaben sich für Methylgruppen und unstrukturierte Abschnitte. Anisotopien und Korrelationen in den Atomkoordinaten ließen sich leicht durch eine Verallgemeinerung des Verteilungsmodells berücksichtigen: Für die Modellierung

lokaler Anisotropien müssten im Fehlermodell in Gleichung (3.12) die diagonalen Kovarianzmatrizen lediglich durch atomweise 3-dimensionale Kovarianzmatrizen $\mathbf{\Sigma}_j$ ersetzt werden. Korrelationen zwischen Atomen ließen sich über eine vollständige $3M$-dimensionale Kovarianzmatrix $\mathbf{\Sigma}$ erfassen. Das Verteilungsmodell besäße dann die allgemeine Form

$$p_{\mathrm{par}}\left(\mathbf{x}^{(i)}|\mu,\mathbf{\Sigma},\mathbf{R}^{(i)},\mathbf{t}^{(i)},I\right)=\frac{1}{(2\pi)^{3M/2}\det^{1/2}\mathbf{\Sigma}}\exp\left\{-\frac{1}{2}\mathrm{Sp}\left(\mathbf{a}\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}^{-1}\right)\right\},$$

mit $\mathbf{a}=\mathbf{x}^{(i)}-\hat{\mathbf{R}}^{(i)}\mu-\hat{\mathbf{t}}^{(i)}$. $\mathbf{x}^{(i)}$ und μ bezeichnen den $3M$-dimensionalen Vektor der kartesischen Koordinaten der i-ten Konformation bzw. der mittleren Struktur. Die $3M$-dimensionale Matrix $\hat{\mathbf{R}}^{(i)}$ und der $3M$-dimensionale Vektor $\hat{\mathbf{t}}^{(i)}$ besitzen die Form

$$\hat{\mathbf{R}}^{(i)}=\begin{pmatrix}\mathbf{R}^{(i)} & & 0\\ & \ddots & \\ 0 & & \mathbf{R}^{(i)}\end{pmatrix},\quad \hat{\mathbf{t}}^{(i)}=\begin{pmatrix}\mathbf{t}^{(i)}\\ \vdots\\ \mathbf{t}^{(i)}\end{pmatrix}. \tag{4.1}$$

Ein entsprechender Gibbs-Algorithmus für die Simulation des verallgemeinerten Modells folgte unmittelbar aus der korrespondierenden *a-posteriori*-Verteilung. Der numerische Aufwand für die Schätzung dieses Modells skalierte allerdings quadratisch mit der Zahl der Atome, wohingegen der Aufwand für die Simulation des gezeigten isotropen Modells lediglich linear mit der Zahl der Atome steigt.

4.2 Qualität eines NOE-Datensatzes

Um den Einfluß von Inkonsistenzen in den Daten auf die Unsicherheitsbehaftung einer NMR-Struktur zu untersuchen, wurde ein Konsistenzmaß für zugeordnete NOE-Datensätze definiert. Das Maß σ_d quantifiziert die Konsistenz eines Datensatzes und läßt sich interpretieren als mittlere relative Abweichung der Zieldistanzen von den korrespondierenden Distanzen in der Struktur. Das Konsistenzmaß folgt direkt aus dem Datenmodell zur Beschreibung dipolarer Kreuzrelaxationsraten: Der Hypothesenparameter σ quantifiziert

die Abweichung von observierten und berechneten Kreuzrelaxationsraten einer Einzelmessung und ist *a priori* unbekannt. Das Konsistenzmaß wird auf Basis der *a-posteriori*-Verteilung für σ bestimmt und quantifiziert demzufolge die Verträglichkeit der Messungen untereinander sowie die Schlüssigkeit von Daten und Hintergrundwissen. In der induktiven Strukturbestimmung folgt ein Maß für die Erfüllbarkeit eines Datensatzes somit in natürlicher Weise aus der statistischen Beschreibung der experimentellen Daten. Die Bewertung der Konsistenz eines Datensatzes erfolgt dabei niemals absolut, sondern stets in Bezug auf relevantes Hintergrundwissen: Die Schlüssigkeit der Meßwerte wird beispielsweise davon bestimmt, wie die experimentellen Daten interpretiert werden. Einen weiteren Faktor bildet physikalisches *a-priori*-Wissen: In die Bewertung fließt stets Wissen ein, ob eine Messung *a priori* zu erwarten ist, z.B. ob sie strukturell erfüllbar ist. Eine „intrinsische“ Konsistenz eines Datensatzes ist in diesem Sinne bedeutungslos.
Im Gegensatz zu kreuzvalidierten Konsistenzmaßen erfordert die Berechnung von σ_d keinen zusätzlichen Rechenaufwand: Die Bestimmung der marginalen *a posteriori*-Verteilung $p(\sigma|D, I)$ und damit die Bestimmung des Konsistenzmaßes, ist integraler Bestandteil der Strategie zur Simulation der Strukturverteilung. Eine mehrfache Durchführung der Strukturrechnung, wie sie für die Berechnung kreuzvalidierter Qualitätsmaße vonnöten ist, entfällt daher.

4.2.1 Konsistenz eines Datensatzes

Testrechnungen für die Tudor Domäne und BPTI zeigten, daß die Konsistenz eines Datensatzes während der Strukturrechnung aus den Daten zurückgerechnet werden kann: Der Verlauf von σ_d als Funktion des Inkonsistenzgrades eines Datensatzes zeigte in allen betrachteten Fällen eine gute Übereinstimmung mit der theoretischen Vorhersage, die auf Basis einer analytischen Relation berechnet wurde. Inkompatibilitäten der Kraftfelder, welche für die Simulation der Strukturverteilungen und die Berechnung der Referenzstrukturen (diese dienten als Basis für die Erzeugung der einzelnen Testdatensätze) verwendet wurden, führten zu Abweichungen von theoretischen und geschätz-

ten Werten. Dies verdeutlicht, daß das Konsistenzmaß von der Verträglichkeit der Messungen untereinander *und* der Schlüssigkeit von Daten und Hintergrundwissen bestimmt wird: In den gezeigten Beispielen äußerten sich die genannten Inkompatibilitäten auch im Falle simultan erfüllbarer Messungen in einer nichtverschwindenden Diskrepanz der observierten und berechneten Distanzen. Eine Abhängigkeit des Maßes von der Zahl der Messungen konnte dagegen nicht festgestellt werden: σ_d verhielt sich stabil unter Variation der Größe eines Datensatzes; diese Invarianzeigenschaft ist eine notwendige Voraussetzung für eine allgemeingültige Interpretation des Maßes.
Die Erfüllbarkeit einer Messung wird im Datenmodell für Kreuzrelaxationsraten durch den Hypothesenparameter σ explizit berücksichtigt. Inkonsistenzen in den Daten nehmen daher unmittelbaren Einfluß auf die Breite der Strukturverteilung: In Gauß'scher Näherung wird ein linearer Zusammenhang zwischen der Inkonsistenz eines Datensatzes und der mittleren Unsicherheitsbehaftung der Atompositionen erwartet. Zusätzlich skaliert die atomare Unsicherheitsbehaftung invers proportional mit der Wurzel der Größe eines Datensatzes. Beide Abhängigkeiten wurden in den Testsimulationen in guter Näherung reproduziert. Unvollständige Ausgangsdaten äußern sich somit in einer erhöhten strukturellen Unsicherheitsbehaftung der Atompositionen, wodurch eine Überinterpretation der generierten Koordinaten vermieden wird.

4.2.2 Konsistenz von Einzelmessungen

Die statistische Interpretierbarkeit des Datenmodells erlaubte ferner eine objektive Definition der Konsistenz einer Einzelmessung hinsichtlich der Gesamtheit aller Messungen eines Datensatzes. Das Konzept einer Vorhersageverteilung hat sich dabei als wertvoll erwiesen: Vorhersageverteilungen basieren auf dem Datenmodell und werden durch Marginalisierung aller Hypothesenparameter gebildet. Auf diese Weise wird garantiert, daß die Unsicherheitsbehaftung der dreidimensionalen Koordinaten sowie die Qualität des gesamten Datensatzes vollständig bei der Vorhersage berücksichtigt wer-

den. Die Richtigkeit einer Messung konnte mit Hilfe von Konfidenzintervallen in Verbindung mit einer Vorhersageverteilung in Form eines Konfidenzwertes auf einer absoluten Skala definiert werden. Am Beispiel des BPTI-Datensatzes wurde gezeigt, daß Vorhersageverteilungen stets konservativere Aussagen bezüglich der erwarteten Verteilung eines Meßwerts treffen als Häufigkeitsverteilungen, die aus konformationellen Stichproben abgeleitet wurden. Eine Überschätzung der Genauigkeit einer Messung wird auf diese Weise vermieden. Die Bewertung der Richtigkeit einer Messung auf Basis einer Häufigkeitsverteilung, wie sie beispielsweise bei Verfahren wie der *Violation Analysis* [76] zur Identifikation inkonsistenter Distanzeinschränkungen erfolgt, ist in diesem Sinne vorurteilsbehaftet.
Das Verfahren ist nicht auf den vorgeführten Fall dipolarer Kreuzrelaxationsraten beschränkt: In der induktiven Strukturbestimmung erfolgt die Beschreibung experimenteller Messungen stets durch ein entsprechendes Datenmodell. Die Abweichung zwischen observiertem und berechnetem Meßwert, welche *a priori* unbekannt ist, geht dabei in Form eines Hypothesenparameters ein, der während der Strukturrechnung bestimmt werden kann. Aufgrund der statistischen Interpretierbarkeit des Datenmodells folgt ein Maß für die Konsistenz der Messungen unmittelbar aus der gewählten Parametrisierung des Datenmodells. Vorhersageverteilungen und Konfidenzintervalle sind allgemeine statistische Konzepte. Die beschriebene Bewertung von Einzelmessungen ist somit auf beliebige experimentelle Observable übertragbar, für die ein Datenmodell verfügbar ist.

4.3 Inkonsistente NOE-Datensätze

Es zeigte sich, daß die Unsicherheitsbehaftung einer Struktur durch Inkonsistenzen in den Daten erhöht wird. Eine zusätzliche Verfälschung der Koordinatenunsicherheiten wird durch strukturelle Verzerrungen hervorgerufen, die aufgrund systematischer Abweichungen von observierten und berechneten Kreuzrelaxationsraten entstehen können. Inkonsistente Messungen lassen

sich mit Hilfe von Vorhersageverteilungen identifizieren. Die Identifikation geschieht jedoch nach der Strukturrechnung. Das zweikomponentige Mischmodell für dipolare Kreuzrelaxationsraten erlaubt die Identifizierung von Inkonsistenzen in den Daten während der Strukturrechnung. Das Modell behandelt einen Datensatz als Mischung von konsistenten und inkonsistenten Messungen. Beide Klassen werden durch separate Fehlerverteilungen beschrieben. Auf diese Weise wird die experimentelle („wahre") Fehlerverteilung, in der sich inkonsistente Messungen in Form von Ausreißern äußern, realistisch modelliert.

Verglichen mit dem nicht-klassifizierenden Datenmodell hat das Mischmodell zwei entscheidende Vorteile: Erstens, werden lokale Verzerrungen und Unsicherheiten in den dreidimensionalen Koordinaten einer Struktur deutlich reduziert. Die Klassifikation der Messungen erfolgt dabei während der Strukturrechnung durch die Schätzung aller Klassifikationsparameter. Eine mehrfache Durchführung der Strukturrechnung, wie sie bei verwandten, iterativen Verfahren [23] vonnöten ist, enfällt daher. Zweitens, sind alle Hypothesenparameter des Modells intuitiv interpretierbar: So gestatten die Klassenzugehörigkeitsparameter Aussagen über die strukturelle Erfüllbarkeit jeder Einzelmessung.

Anhand des Modelldatensatzes für BPTI wurde gezeigt, daß dynamikbehaftete Messungen mit Hilfe des Mischmodells von konsistenten Messungen getrennt und auf diese Weise identifiziert werden können. Aufgrund der individuellen Gewichtung der Meßwerte bewertet das Mischmodell die Schlüssigkeit der Struktur mit den Daten primär auf Basis der konsistenten Messungen. Systematische Abweichungen der observierten von den berechneten Kreuzrelaxationsraten, wie sie bei inkonsistenten Messungen auftreten, werden dadurch automatisch schwächer gewichtet. Auf diese Weise konnten die Richtigkeit und die Genauigkeit der Struktur sowie externe Qualitäts-Indikatoren deutlich verbessert werden. Die Reduzierung lokaler Verzerrungen führte darüber hinaus zu einer qualitativen Übereinstimmung der atomaren Unsicherheitsbehaftung mit physikalischen Fluktuationen, welche aus der

BPTI-Molekulartrajektorie berechnet wurden. Die Unsicherheitsbehaftung der NMR-Struktur unterschätzt dabei die Stärke der Fluktuationen. Eine Skalengleichheit ist aufgrund der Abhängigkeit der Koordinatenunsicherheiten von der Größe eines Datensatzes nicht zu erwarten. Die Gleichsetzung von Inkonsistenz und Dynamikbehaftung ist im Falle des BPTI-Datensatzes gerechtfertigt: Bei der Ableitung der simulierten Kreuzrelaxationsraten wurden Multispineffekte nicht berücksichtigt, so daß Inkonsistenzen ausschließlich auf die interne Dynamik von BPTI zurückzuführen sind. Ferner ließen sich Spindiffusionskorrekturen im Prinzip mittels mehrerer Verfahren im Mischmodell berücksichtigen, so daß eine Trennung dynamischer Effekte auch für spindiffusionsbehaftete Messungen möglich wäre.

Aufgrund der individuellen Klassifikation jeder Einzelmessung übersteigt die Zahl der Hypothesenparameter grundsätzlich die Zahl der Messungen: Sowohl für BPTI als auch für die Tudor Domäne standen weniger als 0.9 NOE-Messungen für die Schätzung eines Hypothesenparameters zur Verfügung. Dennoch erwies sich die Simulation der Strukturverteilung durch Replika-Austausch-Monte-Carlo und damit die Bestimmung der Klassifikationsparameter als unproblematisch: Dies zeigt, daß die Wahl eines geeigneten Datenmodells nicht von technischen Überlegungen im Hinblick auf die Zahl der unbekannten Parameter geleitet sein muß: In der induktiven Strukturbestimmung wird der Wert einer unbekannten Größe nicht „bestimmt“ – vielmehr wird die Plausibilität aller möglichen Werte des Parameters im Lichte der Daten und vorhandener Hintergrundinformation in Form einer *a-posteriori*-Verteilung bewertet. Der Gibbs-Algorithmus zur Simulation der *a-posteriori*-Verbundverteilung des Mischmodells folgte unmittelbar aus der Problemformulierung und ist daher garantiert konsistent. Die große Zahl der zu schätzenden Parameter führt somit zu keinen algorithmischen Instabilitäten. Die Hybridenergiefunktion, welche konventionellen Methoden zugrundeliegt, ist hingegen nur Zielfunktion für die dreidimensionalen Koordinaten einer Struktur: Sie gibt keine Regeln für die Bestimmung von Zusatzparametern. Die Konsi-

stenz externer Heuristiken zur Bestimmung von Zusatzparametern ist nicht garantiert, was die Verwendung komplexer Datenmodelle aufgrund algorithmischer Instabilitäten erschwert, wenn nicht gar unmöglich macht.

Die Verwendung des Mischmodells führt auch bei schwach dynamikbehafteten Daten mit Beiträgen durch Spindiffusion oder heteronukleare Relaxation zu einer verbesserten Qualität der erzeugten Strukturen. Wie die Testrechnungen mit den experimentellen Datensätzen der Tudor Domäne zeigen, lassen sich Inkonsistenzen, welche auf andere Effekte als auf interne Dynamik zurückzuführen sind, jedoch nicht eindeutig isolieren: In dem gezeigten Beispiel reduziert sich das Mischmodell im Wesentlichen auf das nichtklassifizierende Datenmodell. Um dynamikinduzierte Inkonsistenzen von Spindiffusionseffekten in befriedigender Weise unterscheiden zu können, wird es daher nötig sein, die ISPA, welche momentan für die Berechnung von Kreuzrelaxationsraten verwendet wird, durch einen vollständigen Relaxationsmatrix-Ansatz zu ersetzen. Dazu ist lediglich das verwendete Vorwärtsmodell zu modifizieren; der beschriebene Replika-Austausch-Algorithmus für die Simulation der Strukturverteilung wäre von diesem Schritt nicht betroffen und kann direkt übernommen werden.

Kapitel 5

Ausblick

Die induktive Strukturbestimmung hat sich als allgemeines Prinzip erwiesen, welches existierende Konzepte aus der konventionellen Strukturbestimmung klärt und in einem kohärenten Rahmen integriert: Die statistisch fundierte Definition eines NMR-Strukturensembles ist die Strukturverteilung, aus welcher die Unsicherheitsbehaftung von Atompositionen in objektiver Weise extrahiert werden kann. Die Qualität eines Datensatzes geht direkt in die Beschreibung experimenteller Daten ein und wird über Regeln, die eindeutig aus der Problemformulierung hervorgehen, während der Strukturrechnung bestimmt. Die Möglichkeit, auch komplexe Modelle verläßlich aus den Daten bestimmen zu können, bedeutet ein hohes Maß an Flexibilität bei der Modellierung experimenteller Daten: Der für NOESY-Daten hergeleitete Ausdruck der Strukturverteilung repräsentiert die Unsicherheitsbehaftung einer NMR-Struktur in objektiver Weise – unter der Annahme, daß die Zuordnung von Resonanzen und Kreuzresonanzen bekannt ist (diese Annahme ist Teil der Hintergrundinformation I). Um Artefakten aufgrund von fehlerhaften oder unvollständigen Zuordnungen vorzubeugen, wäre es wünschenswert, daß die Berurteilung struktureller Unsicherheit auf Basis eines nicht-zugeordneten NOESY-Spektrums erfolgt. Entsprechendes gilt für die Qualität eines Datensatzes. Dies wurde bereits von anderen Autoren im Zusammenhang mit der Bewertung von NMR-Strukturen mittels R-Faktoren angemerkt [33]. In

der induktiven Strukturbestimmung wäre es denkbar, einen analytischen Ausdruck für die Strukturverteilung bezüglich roher NMR-Daten herzuleiten: Die unbekannten Resonanzzuordnungen würden in diesem Fall als Satz von *Nuisance*-Parametern in das Datenmodell eingehen. Konsistente Regeln für die Bestimmung aller Unbekannten, also Strukturkoordinaten und Resonanzzuordnungen, folgten in gewohnter Weise in Form eines Gibbs-Schemas aus der korrespondierenden *a-posteriori*-Verbundverteilung. Die Angabe eines Datenmodells für nicht-zugeordnete Kreuzrelaxationsraten ist unproblematisch. Inwieweit eine numerische Lösung dieses Problems realisierbar ist, bliebe einem Versuch überlassen.

Neben der strukturellen Unsicherheitsbehaftung würde die routinemäßige Angabe der Datenqualität das Bild der Gesamtqualität und der Verläßlichkeit einer NMR-Struktur ergänzen. Aufgrund von fehlenden Konzepten, dem Mehraufwand für die Berechnung externer Qualitätsmaße sowie der mangelnden Integration bestehender Verfahren, wird auf die Validierung von Proteinstrukturen in der Praxis oftmals verzichtet. Validierung ist ein integraler Bestandteil des Strukturbestimmungsprozesses, weshalb ich hoffe, daß die in dieser Arbeit entwickelten Methoden der Validierung von NMR-Strukturen zu einer breiteren Anwendung verhelfen können.

Anhang A

Wahrscheinlichkeitsverteilungen

Die folgenden Abschnitte fassen die wichtigsten Eigenschaften aller in dieser Arbeit verwendeten Wahrscheinlichkeitsverteilungen zusammen. Besprochen werden die Lognormalverteilung, die inverse Gammaverteilung, die Betaverteilung sowie die von Mises-Verteilung.

A.1 Lognormalverteilung

Die Lognormalverteilung ist die „Normalverteilung für positive Größen“. Ihre Dichtefunktion (vgl. Abb. A.1) folgt unmittelbar aus der Dichtefunktion der Normalverteilung,

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \tag{A.1}$$

durch die Variablentransformation $x \to \log x$:

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} x^{-1} \exp\left\{-\frac{1}{2\sigma^2}(\log^2(x/\mu)\right\}, \tag{A.2}$$

wobei $x > 0$. $\mu > 0$ und $\sigma > 0$ bezeichnen den Orts- bzw. Formparameter der Lognormalverteilung. Die Lognormalverteilung in Gl. (A.2), in Kurznotation $\mathrm{LN}(x;\mu,\sigma^2)$, ist bezüglich x normiert und besitzt die folgenden Eigenschaften:

$$\mathrm{Median}[x] \;=\; \mu, \tag{A.3}$$

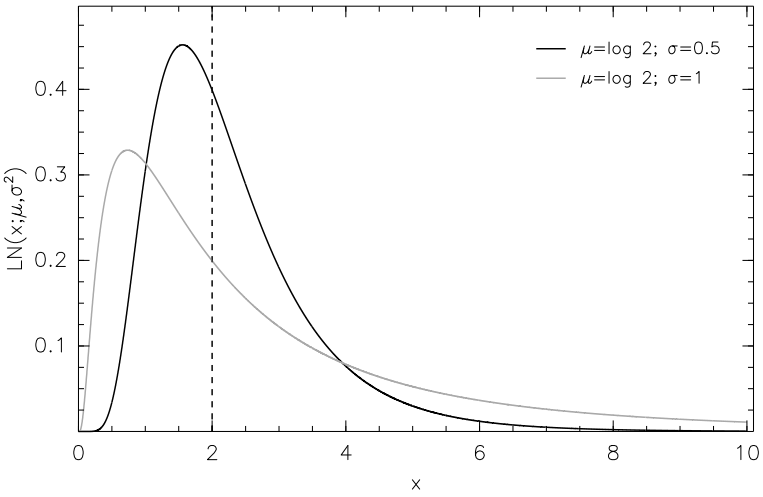

Abbildung A.1: Lognormalverteilung: Graph der Dichtefunktion für verschiedene Formparameter σ bei gleichem Ortsparameter μ. Der Median der Lognormalverteilung entspricht dem Ortsparameter μ und ist gestrichelt dargestellt.

$$\mathrm{E}[x] = \mu e^{\sigma^2/2}, \tag{A.4}$$

$$\mathrm{Var}[x] = \mu^2 e^{\sigma^2}\left(e^{\sigma^2}-1\right). \tag{A.5}$$

A.2 Inverse Gammaverteilung

Die inverse Gammaverteilung (vgl. Abb. A.2) ist für positive x definiert und folgt aus der Gammaverteilung mit der Dichtefunktion,

$$p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x), \tag{A.6}$$

durch die Variablentransformation $x \to x^{-1}$:

$$p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)}\exp(-\beta x^{-1}). \tag{A.7}$$

$\Gamma(\cdot)$ ist die Gamma-Funktion, $\alpha > 0$ und $\beta > 0$ bezeichnen die Formparameter der inversen Gammaverteilung. Die inverse Gammaverteilung in Gl.

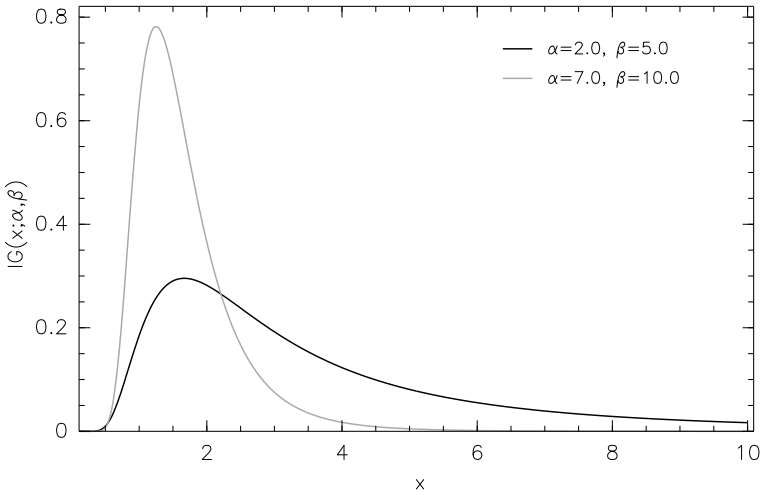

Abbildung A.2: Inverse Gammaverteilung: Graph der Dichtefunktion für verschiedene Formparameter α und β.

(A.7), in Kurznotation $\mathrm{IG}(x;\alpha,\beta)$, ist bezüglich x normiert und bestizt die folgenden Eigenschaften:

$$\mathrm{E}[x] = \frac{\beta}{\alpha-1}, \tag{A.8}$$

$$\mathrm{Var}[x] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \qquad \text{für } \alpha > 2. \tag{A.9}$$

A.3 Betaverteilung

Die Betaverteilung mit den Formparametern $\alpha > 0$ und $\beta > 0$ ist für $x \in [0,1]$ definiert und bestitzt die Dichtefunktion

$$p(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1-x)^{\beta-1}x^{\alpha-1}. \tag{A.10}$$

Die Dichtefunktion der Betaverteilung in Gl. (A.10), Kurznotaton $\mathrm{B}(x;\alpha,\beta)$, ist in Abbildung A.3 dargestellt. Für den Erwartungswert und die Varianz von x folgt:

$$\mathrm{E}[x] = \frac{\alpha}{\alpha+\beta}, \tag{A.11}$$

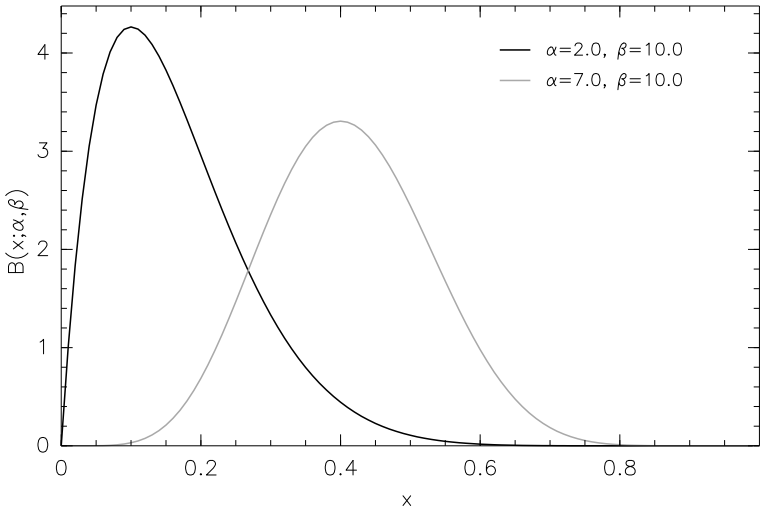

Abbildung A.3: Betaverteilung: Graph der Dichtefunktion für verschiedene Formparameter α und β.

$$\mathrm{Var}[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{A.12}$$

A.4 Von Mises-Verteilung

Die von Mises-Verteilung ist die „Normalverteilung für zyklische Variable“ und ist für $x \in [0, 2\pi)$ definiert. Die Dichtefunktion (vgl. Abb. A.4) einer von Mises-Verteilung für x mit Ortsparameter $\phi \in [0, 2\pi)$ und Formparameter $\kappa > 0$ ist definiert als

$$p(x|\phi,\kappa) = \frac{\exp(\kappa\cos(x-\phi))}{2\pi I_0(\kappa)}. \tag{A.13}$$

$I_0(\cdot)$ bezeichnet die modifizierte Besselfunktion der ersten Art und Ordnung 0. Für den Erwartungswert und die zirkuläre Varianz folgt [77]:

$$\mathrm{E}[x] = \phi, \tag{A.14}$$

$$\mathrm{Var}[x] = 1 - \frac{I_\phi(\kappa)}{I_0(\kappa)}. \tag{A.15}$$

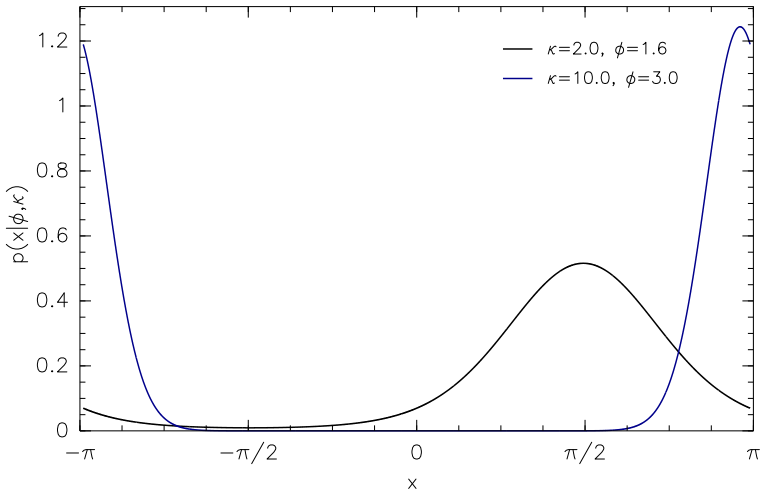

Abbildung A.4: Von Mises-Verteilung: Graph der Dichtefunktion für verschiedene Werte des Orts- und Formparameters ϕ bzw. κ.

Anhang B

ISD-Simulationspaket

Das ISD (*Inferential Structure Determination*) Simulationspaket stellt die vollständige Infrastruktur für die Simulation und Analyse eines induktiven Strukturbestimmungsproblems aus NMR-Daten zur Verfügung. Teile der Infrastruktur, wie beispielsweise Daten-Ein- und Ausgabe sowie die Implementierung einer parallelisierten Version des Replika-Austausch-Monte-Carlo-Algorithmus, entstanden in Kollaboration mit M. Habeck, Unite de Bioinformatique Structurale, Institut Pasteur.

Implementierung

Das Softwarepaket ist in Form einer Softwarebibliothek organisiert und wurde in den Programmiersprachen C und Python [78] implementiert. Die Quellen umfassen etwa 40000 Zeilen Programmcode. Zeitkritische Teile wurden als C-Routinen implementiert, die mittels spezieller Schnittstellen (*„Wrapper“*) von Python aus angesprochen werden. Die gesamte Infrastruktur des Pakets ist in Python geschrieben. Der C-Teil der Bibliothek umfaßt Routinen für die Berechnung von Energien und Gradienten aller Datenmodelle sowie des physikalischen Kraftfelds. Ferner sind der Hybrid-Monte-Carlo-Algorithmus sowie die Gibbs-Schritte zur Schätzung von *Nuisance*-Parametern numerisch aufwendig und wurden ebenfalls in C implementiert.

Die Realisierung des Simulationspakets als Softwarebibliothek bedeutet ein hohes Maß an Flexibilität und Erweiterbarkeit. Die hybride Softwarearchitektur erlaubt zudem eine zeiteffiziente Programmentwicklung: Die zahlreichen *high level*-Routinen der Skriptsprache Python führen zu einer signifikanten Reduktion des Aufwands für Implementierung[1] und Fehlersuche; die Effizienz der numerischen Routinen wird durch ihre Implementierung in C garantiert.

Kraftfeld und Datenmodelle

Die Polypeptidkette wird intern durch Dihedralwinkel parametrisiert. Die Definition sowie die kovalente Geometrie der starren Gruppen wurden dem ECEPP/2 Kraftfeld [40, 41] entnommen. Alle Dihedralwinkel um die Peptidbindungen sind auf 180° fixiert; Wasserstoffatome werden explizit modelliert. Potentiale für die Hauptketten-Dihedralwinkel ϕ und ψ sowie für den Seitenketten-Dihedralwinkel χ_1 werden nicht berücksichtigt. Die gegenwärtige Implementierung unterstützt 5 Atomtypen: C, H, N, O, S. Die Nomenklatur für Atomnamen ist konform mit dem von IUPAC [79] vorgeschlagenen Standard. Nichtkovalente Wechselwirkungen werden in Form eines rein repulsiven van der Waals-Terms berücksichtigt, wie er im PROLSQ Kraftfeld [42] angegeben ist; die Werte aller Atomradien stammen ebenfalls aus dem PROLSQ Kraftfeld. Die Kraftkonstante des van der Waals-Terms ist atomtypunabhängig und beträgt $50\,\mathrm{kcal\,mol^{-1}\AA^{-4}}$. Elektrostatische Wechselwirkungen und Lösungsmitteleffekte werden in der gegenwärtigen Implementierung aus Effizienzgründen vernachlässigt.

Neben den Modellen zur Behandlung inkonsistenter NOESY-Messungen und zugeordneter NOESY-Spektren, welche in dieser Arbeit vorgestellt wurden, werden 2 weitere experimentelle NMR-Datentypen unterstützt: RDCs (*Residual Dipolar Couplings*) und J-Kopplungen. Zusätzliche Strukturinformation kann mittels allgemeiner Modelle für Distanz- und Dihedralwinkelfreiheitsgrade für die Strukturrechnung genutzt werden. Die Kombination dieser Mo-

[1] Im Vergleich zu C etwa um einen Faktor 5-10 in der Länge des Quellcodes.

delle erlaubt die Berechnung einer Struktur aus einer unbeschränkten Zahl experimenteller Datensätze.

Parallelisierter Replika-Algorithmus

Leistungsstarke Cluster-Architekturen oder Mehrprozessormaschinen werden durch eine parallelisierte Version des Replika-Austausch-Algorithmus unterstützt. Die Parallelisierung wurde in Form einer *Master-Slave*-Architektur realisiert: Jede Kopie der Replika-Kette wird als separater *Slave*-Prozeß simuliert. Die Steuerung der gesamten Simulation erfolgt zentral über den *Master*-Prozeß, der alle benötigten Parametereinstellungen verwaltet und für die Durchführung des Replika-Austausch-Schritts zuständig ist. Technisch erfolgt die Parallelisierung mit Hilfe der portablen Softwarebibliothek PVM (*Parallel Virtual Machine*) [80]. PVM stellt allgmeine Routinen für die nachrichtenbasierte Kommunikation in heterogenen Rechnernetzwerken zur Verfügung und unterstützt eine Vielzahl von Rechnerarchitekturen.

Infrastruktur

Die Daten-Ein- und Ausgabe erfolgt über die Dokument-Beschreibungssprache XML (*eXtended Markup Language*) [81]. Das ISD-XML-Format umfaßt Definitionen für Sequenzinformation sowie für zugeordnete NOE-, RDC- und J-Kopplungs-Daten. Das XML-Format für NOE-Daten kann mittels Konversionsroutinen unmittelbar in das XML-Format von ARIA Version 2.0 [82], einer Software zur automatischen Zuordnung von NOESY-Daten, übersetzt werden. Die Anbindung von ARIA an das CCPN-Datenmodell (*Collaborative Computing Project for NMR*) [83] erlaubt daher den beidseitigen Datenaustausch mit Fremdsoftware, die ebenfalls das CCPN-Datenformat unterstützt.

Literatur

[1] M. P. Williamson, T. F. Havel, and K. Wüthrich. Solution conformation of proteinase inhibitor from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry. *J. Mol. Biol.*, 182(2):295–315, Mar 20 1985.

[2] R. Kaptein, E. R. Zuiderweg, R. M. Scheek, R. Boelens, and W. F. van Gunsteren. A protein structure from nuclear magnetic resonance data: lac repressor headpiece. *J. Mol. Biol.*, 182(1):179–182, Mar 5 1985.

[3] G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.*, 104:4546–4558, 1982.

[4] L. E. Kay, D. A. Torchia, and A. Bax. Backbone dynamics of proteins as studied by ^{15}N inverse detected heteronuclear spectroscopy: application to staphylococcal nuclease. *Biochemistry*, 28(23):8972–8979, Nov 14 1989.

[5] K. Wüthrich. *NMR of Proteins and Nucleic Acids.* John Wiley, New York, 1986.

[6] D. Neuhaus and M. P. Williamson. *The nuclear Overhauser effect in structural and conformational analysis, 2nd ed.* Wiley-VCH Inc., New York, 2000.

[7] I. Solomon. Relaxation processes in a system of two spins. *Phys. Rev.*, 99(2):559–565, July 1955.

[8] P. S. Hubbard. Nonexponential Relaxation of Rotating Three-Spin Systems in Molecules of a Liquid. *J.Chem.Phys*, 52:563–568, 1970.

[9] J. Tropp. Dipolar relaxation and nuclear Overhauser effects in nonrigid molecules: The effect of fluctuating internuclear distances. *J.Chem.Phys.*, 72:6035–6043, 1980.

[10] D. Wallach. Effect of Internal Rotation on Angular Correlation Functions. *J.Chem.Phys.*, 47:5258–5268, 1967.

[11] A. Kumar, G. Wagner, R. R. Ernst, and K. Wüthrich. Buildup rates of the nuclear Overhauser effect measured by two-dimensional proton magnetic resonance spectroscopy: implications for studies of protein conformation. *J. Am. Chem. Soc.*, 103:3654–3658, 1981.

[12] J. Jeener, B. H. Meier, P. Bachmann, and R. R. Ernst. Investigation of exchange processes by two-dimensional NMR spectroscopy. *J. Chem. Phys.*, 71:4546–4553, 1979.

[13] S. Macura and R. R. Ernst. Elucidation of cross relaxation in liquids by two-dimensional NMR spectroscopy. *Molecular Physics*, 41:95–117, 1980.

[14] R. M. Scheek, W. F. van Gunsteren, and R. Kaptein. Molecular dynamics simulations techniques for determination of molcular structures from nuclear magnetic resonance data. *Methods in Enzymology*, 177:204–218, 1989.

[15] A. T. Brünger and M. Nilges. Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Q. Reviews of Biophys.*, 26(1):49–125, Feb 1993.

[16] P. Güntert. Structure calculation of biological macromolecules from NMR data. *Q. Reviews of Biophys.*, 31(2):145–137, 1998.

[17] G. M. Clore and C. D. Schwieters. Theoretical and computational advances in biomolecular NMR spectroscopy. *Curr. Opin. Struct. Biol.*, 12:146–153, 2002.

[18] A. Kalk and H. J. C. Berendsen. Proton magnetic relaxation and spin diffusion in proteins. *J. Magn. Reson.*, 24:275–268, 1976.

[19] R. Boelens, T. M. G. Koning, and R. Kaptein. Determination of biomolecular structures from proton–proton NOEs using a relaxation matrix approach. *J. Mol. Struct.*, 173:299–311, 1989.

[20] B. A. Borgias and T. L. James. MARDIGRAS: a procedure for matrix analysis of relaxation for discerning geometry of an aqueous structure. *J. Magn. Reson.*, 87:475–487, 1990.

[21] T. L. James. Relaxation matrix analysis of two-dimensional nuclear Overhauser effect spectra. *Curr. Opin. Struct. Biol.*, 1:1042–1053, 1991.

[22] D. M. LeMaster, L. E. Kay, A. T. Brünger, and J. H. Prestegard. Protein dynamics and distance determinations by NOE measurement. *FEBS Lett.*, 236:71–76, 1988.

[23] T. Schneider, A. T. Brünger, and M. Nilges. Influence of internal dynamics on accuracy of protein NMR structures: derivation of realistic model distance data from a long molecular dynamics trajectory. *J. Mol. Biol.*, 285:727–740, 1999.

[24] A.T. Brunger, R.L. Campbell, G.M. Clore, A.M. Gronenborn, M. Karplus, G. Petsko, and M. M. Teeter. Solution of a Protein Crystal Structure with a Model Obtained from NMR Interproton Distance Restraints. *Science*, 235:1049–1053, 1987.

[25] A. T. Brünger. The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355:472–474, 1992.

[26] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, 26:283–291, 1993.

[27] R. Laskowski, J. Rullman, M. MacArthur, R. Kaptein, and J. Thornton. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*, 8:477–486, 1996.

[28] G. Vriend and C. Sander. Quality control of protein models: Directional atomic contact analysis. *J. Appl. Cryst.*, 26:47–60, 1993.

[29] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.*, 112:535–542, 1977.

[30] M. J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins Struct. Funct. Genet.*, 17:355–362, 1993.

[31] C. Gonzales, J. A. C. Rullmann, A. M. J. J. Bonvin, R. Boelens, and R. Kaptein. Toward an NMR R factor. *J. Magn. Reson.*, 91:659–664, 1991.

[32] A. T. Brünger, G. M. Clore, A. M. Gronenborn, R. Saffrich, and M. Nilges. Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. *Science*, 261:328–331, 1993.

[33] W. Gronwald, R. Kirchhöfer, A. Görler, W. Kremer, B. Ganslmeier, K. P. Neidig, and H. R. Kalbitzer. Rfac, a program for automated NMR R-factor estimation. *J. Biomol. NMR*, 17:137–151, 2000.

[34] M. Habeck, W. Rieping, and M. Nilges. A new principle for macromolecular structure determination. In G. Erickson and Y. Zhai, editors, *23rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, pages 157–166. American Institute of Physics, 2004.

[35] W. Rieping, M. Habeck, and M. Nilges. Inferential Structure Determination. 2004. Submitted.

[36] E. T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press, Cambridge UK, 2003.

[37] R. T. Cox. Probability, frequency and reasonable expectation. *Am. J. Phys.*, 14:1–13, 1946.

[38] R. T. Cox. *The Algebra of Probable Inference.* John Hopkins University Press, 1961.

[39] D. S. Sivia. *Data Analysis. A Bayesian Tutorial.* Oxford University Press Inc., New York, 1996.

[40] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga. Energy Parameters in Polypeptides VII, Geometric Parameters, Partial Charges, Non-bonded Interactions, Hydrogen Bond Interactions and Intrinsic Torsional Potentials for Naturally Occuring Amino Acids. *J. Phys. Chem.*, 79:2361–2381, 1975.

[41] G. Nemethy, M. A. Pottle, and H. A. Scheraga. Energy Parameters in Polypeptides, 9. Updating of Geometrical Parameters, Non-bonding Interactions and Hydrogen Bonding Interactions for Naturally Occuring Amino Acids. *J. Phys. Chem.*, 87:1883–1887, 1983.

[42] W. A. Hendrickson. Stereochemically restrained refinement of macromolecular structures. *Methods in Enzymology*, 115:252–270, 1985.

[43] E. T. Jaynes. Information Theory and Statistical Mechanics. *Phys. Rev. Lett.*, 106:620–630, 1957.

[44] W. Rieping, M. Habeck, and M. Nilges. Structure calculation from NMR data – a Bayesian view. In M. Sattler, M. Nilges, and H. Oschkinat, editors, *NMR analysis of protein structure.* Springer-Verlag, Heidelberg, 2004. To appear.

[45] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. PAMI*, 6(6):721–741, 1984.

[46] S. Duane, A. D. Kennedy, B. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195:216–222, 1987.

[47] R. H. Swendsen and J.-S. Wang. Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.*, 57:2607–2609, 1986.

[48] R. M. Neal. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.

[49] N. Metropolis, M. Rosenbluth, A. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing. *J. Chem. Phys.*, 21:1087–1092, 1957.

[50] B.J. Alder and T.E. Wainwright. Studies in molecular dynamics. I. General method. *J. Chem. Phys.*, 31:459–466, 1959.

[51] M. H. Chen, Q. M. Shao, and J. G. Ibrahim. *Monte Carlo Methods in Bayesian Computation.* Springer Verlag, Inc., New York, 2002.

[52] Herbert Goldstein. *Classical Mechanics.* Addison-Wesley Publishing Company, Reading, MA, 2nd edition, 1980.

[53] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids.* Clarendon Press, Oxford, 1987.

[54] C. Tsallis. Possible Generalization of Boltzmann-Gibbs Statistics. *J. Stat. Phys.*, 52:479–487, 1988.

[55] U. H. E. Hansmann and Y. Okamoto. New Generalized-Ensemble Monte Carlo Method for Systems with Rough Energy Landscape. *Phys. Rev. E*, 56:2228–2233, 1997.

[56] M. Nolta. Biggles. A Plotting Module for Python. `http://biggles.sourceforge.net`, 2004.

[57] R. Koradi, M. Billeter, and K Wüthrich. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.*, 14:51–55, 1996.

[58] G. Vriend. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.*, 8:52–56, 1990.

[59] M. Marquart, J. Walter, J. Deisenhofer, W. Bode, and R. Huber. The geometry of the reactive site of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Cryst.*, B39:480–487, 1983.

[60] K. Berndt, P. Güntert, L. Orbons, and K. Wütrich. Determination of a high-quality nuclear magnetic resonance solution structure of the bovine pancreatic trypsin inhibitor and comparison with thee crystal structures. *J. Mol. Biol.*, 227:757–775, 1993.

[61] P. Selenko, R. Sprangers, G. Stier, D. Buehler, U. Fischer, and M. Sattler. SMN Tudor domain structure and its interaction with the Sm proteins. *Nature Struct. Biol.*, 8(1):27–31, 2001.

[62] R. Sprangers, M. Groves, I. Sinning, and M. Sattler. High-resolution X-ray and NMR Structures of the SMN Tudor Domain: Conformational Varation in the Binding Site for Symmetrically Dimethylated Arginine Residues. *J. Mol. Biol.*, 327:507–520, 2003.

[63] F.-R. Chalaoux, S. I. O'Donoghue, and M. Nilges. Molecular dynamics and accuracy of NMR structures: effects of error bounds and data removal. *Proteins Struct. Funct. Genet.*, 34:453–463, 1999.

[64] A.T. Brünger. *X–PLOR. A System for X-ray Crystallography and NMR.* Yale University Press, 1992.

[65] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

[66] A. Görler and H. R. Kalbitzer. Relax, a flexible program for the back calculation of NOESY spectra based on complete relaxation matrix formalism. *J. Magn. Reson.*, 124(1):177–188, 1997.

[67] H. Jeffreys. *Theory of Probability.* Oxford University Press, 1939.

[68] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A*, 186:453–461, 1946.

[69] H. L. Harney. *Bayesian Inference. Parameter Eestimation and Decisions.* Springer Verlag, Berlin, Heidelberg, 2003.

[70] L. Devroye. *Non-uniform Random Variate Generation.* Springer Verlag, New York, 1986.

[71] R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal. Markov Chain Monte Carlo in Practice: A Roundtable Discussion. *The American Statistcan*, 52(2):93–100, 1998.

[72] W. F. van Gunsteren, R. M. Brunne, P. Gros, R. C. van Schaik, C. A. Schiffer, and A. E. Torda. Accounting for molecular mobility in structure determination based on nuclear magnetic resonance spectroscopic and X-ray diffraction data. *Methods in Enzymology*, 261:619–654, 1994.

[73] P. F. Yip and D. A. Case. A new method for refinement of macromolecular structures based on nuclear Overhauser effect spectra. *J. Magn. Reson.*, 83:643–648, 1989.

[74] P. L. Linge, M. Habeck, W. Rieping, and M. Nilges. Correction of spin diffusion during iterated automatic NOE assignment. *J. Magn. Reson.*, 167:334–342, 2004.

[75] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, 39:1–38, 1977.

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

[77] E. W. Weisstein. Von Mises Distribution. MathWorld - A Wolfram Web Resource. http://mathworld.wolfram.com/vonMisesDistribution.html, 17.04.2004.

[78] G. van Rossum and J. de Boer. Linking a stub generator (AIL) to a prototyping language (Python). In EurOpen, editor, *EurOpen. UNIX Distributed Open Systems in Perspective. Proceedings of the Spring 1991 EurOpen Conference, Tromsø, Norway, May 20–24, 1991*, pages 229–247, 1991.

[79] J. L. Markley, A. Bax, Y. Arata, C. W. Hilbers, R. Kaptein, B. D. Sykes, P. E. Wright, and K. Wüthrich. Recommendations for the presentation of NMR structures of proteins and nucleic acids. *J. Mol. Biol.*, 280(5):933–952, 1998.

[80] V.S. Sunderam. PVM: A Framework for Parallel Distributed Computing. *Journal of Concurrency: Practice and Experience*, 2(4):315–339, 1990.

[81] The World Wide Web Consortium. Extensible Markup Language (XML) 1.0, W3C recommendation. `http://www.w3.org/TR/REC-xml`, 1999.

[82] M. Habeck, W. Rieping, and M. Nilges. NOE assignment with ARIA 2.0 - the nuts and bolts. In A. K. Downing, editor, *Protein NMR Techniques*. Humana Press, Totowa, NJ, 2004. To appear.

[83] R. H. Fogh, J. Ionides, E. Ulrich, W. Boucher, W. Vranken, J. P. Linge, M. Habeck, W. Rieping, T. N. Bhat, J. Westbrook, K. Henrick, G. Gilliland, H. Berman, J. Thornton, M. Nilges, J. Markley, and E. Laue. The CCPN project: an interim report on a data model for the NMR community. *Nature Struct. Biol.*, 9(6):416–418, 2002.

Danksagung

In den vergangenen drei Jahren haben mehrere Personen zum Gelingen dieser Arbeit beigetragen. Ich möchte mich daher herzlich bedanken bei

Herrn Prof. Dr. Dr. Hans Robert Kalbitzer für die offizielle Betreuung und die bereitwillige Vertretung meiner Arbeit als externe Dissertation gegenüber der Fakultät für Biologie der Universität Regensburg;

Herrn Dr. Michael Nilges für die Finanzierung und Betreuung der Arbeit am Institut Pasteur, für inspirierende, auch über wissenschaftliche Themen hinausgehende Diskussionen sowie für die phantastische Atmosphäre in seiner Arbeitsgruppe;

Michael Habeck für unzählige Diskussionen, für die Zusammenarbeit bei der Entwicklung der induktiven Strukturbestimmung und bei der Implementierung des ISD-Simulationspakets.

Weiterer Dank gilt Dr. Tru Huynh für die professionelle Betreuung der Rechnerumgebung in unserer Arbeitsgruppe am Institut Pasteur und Dr. Michael Sattler, EMBL Heidelberg, für die Bereitstellung der experimentellen NMR-Datensätze der Tudor Domäne.

Mein besonderer Dank gilt meinen Eltern, die mich während meiner Ausbildung in allen Lebenslagen tatkräftig unterstützten und diese Arbeit dadurch erst ermöglicht haben.

Lebenslauf

Persönliche Daten

Name:	Wolfgang Rieping
Geburtsdatum:	3. April 1974
Geburtsort:	Mannheim
Nationalität:	Deutsch
Familienstand:	Ledig

Ausbildung

1984 - 1993	Albertus-Magnus-Gymnasium Viernheim. Abitur, 15. Juni, 1993.
1993 - 1999	Physikstudium an der Ruprecht-Karls-Universität Heidelberg. Diplom, 31. Juni, 1999.
1999 - 2000	Gastwissenschaftler am EMBL Heidelberg, Structural Biology Programme (Juli 1999 - Oktober 2000).
2000 -	Promotion. Aufenthalt am EMBL (Oktober 2000 - Februar 2001), seit März 2001 am Institut Pasteur Paris, Unite de Bioinformatique Structurale.

Berufliche Tätigkeiten

1999 - 2000	Software-Entwickler bei der LION Bioscience AG, Heidelberg. Abteilung *Scientific Development* (Oktober 1999 - Oktober 2000).