Entwicklung eines Kraftfeldes zur Strukturvorhersage von Helixproteinen

DISSERTATION

zur Erlangung des Grades eines
Doktors der Naturwissenschaft
der Abteilung Physik
der Universität Dortmund

vorgelegt von

THOMAS-ALEXANDER HERGES

Oktober 2003

Inhaltsverzeichnis

Kapitel 1

Einleitung

Nach der Aufklärung der Struktur des Genoms des Fadenwurms, der Fruchtfliege, der Maus und neuerdings auch des Menschen, erschließen sich der biomedizinischen Forschung große Potentiale durch die Aufklärung der Struktur und Funktion der im Genom kodierten Proteine. Ein Verständnis über die Funktion der Proteine ist entscheidend für Erkenntnisse über den Lebenszyklus von Zellen, dem Metabolismus und die Frage, wie Zellen Signale an ihre Umgebung senden und Signale von dieser Umgebung empfangen und verarbeiten. Aufgrund des engen Zusammenhangs von Struktur und Funktion wachsen die Bemühungen, Proteine strukturell aufzuklären, und die bedeutendsten Wissenschaftszeitungen der Welt veröffentlichen jede Woche Arbeiten der biomedizinischen Forschung, in denen die Strukturaufklärung von Proteinen wesentlich zum Verständnis wichtiger biologischer Prozesse beiträgt.

Gegenwärtig werden zwei experimentelle Methoden zur Strukturaufklärung der Proteinkonfiguration unter physiologischen Bedingungen, der sogenannten *nativen* Struktur, eingesetzt: die Kernspinresonanzspektroskopie und die Röntgenstrukturanalyse. Diese Methoden gehören im Bereich der Festkörperphysik neben der Neutronenstreuung zu den traditionell wichtigsten Techniken der Strukturaufklärung, doch ist ihre Anwendung auf Proteine ungleich komplizierter oder bei einigen Proteinen gänzlich unmöglich. Experimentelle Verfahren liefern darüber hinaus im wesentlichen statische Informationen, aus denen sich direkte Rückschlüsse über die Dynamik des Mechanismus der Proteinfunktion nur schwer entnehmen lassen.

Theoretische Verfahren zur biomolekularen Strukturvorhersage können hier prinzipiell wichtige Beiträge liefern. Die etablierten Verfahren sind jedoch häufig zu ungenau (homologiebasierte Verfahren), um strukturell unterschiedliche Proteine gut zu beschreiben, oder schlicht zu aufwendig (Molekulardynamik), als daß sie mit den heute zur Verfügung stehenden Rechnerresourcen biologisch relevante Ergebnisse liefern könnten. In den letzten Jahren ist ein alternativer Ansatz zur biomolekularen Strukturaufklärung entwickelt worden, der seinerseits auf die von Anfinsen stammende *thermodynamische Hypothese* zurückzuführen ist, wel-

che besagt, daß sich das Protein im thermodynamischen Gleichgewicht mit seiner physiologischen Umgebung befindet. In der vorliegenden Arbeit gelang die erste erfolgreiche Umsetzung dieses Ansatzes mit atomistischer Auflösung und vertretbaren Rechnerresourcen bei Proteinen mit bis zu 60 Aminosäuren. Hier ist insbesondere die erstmalige reproduzierbare Faltung des aus 40 Aminosäuren bestehenden HIV-accessory Proteins 1F4I zu nennen.

Der besagte Ansatz teilt das Protein-Struktur-Problem in zwei komplementäre Aufgabenstellungen: die Entwicklung atomistisch aufgelöster, generischer Zielfunktionen (Kraftfelder) für die Freie Energie (oder Freie Enthalpie) und die Entwicklung von Methoden (Optimierungsverfahren) zur Identifizierung des globalen Minimums der Freien Energie, welches nach der thermodynamischen Hypothese der nativen Struktur entspricht.

Daher schließt an das folgende, einführende Kapitel, in dem näher auf die Stoffklasse der Proteine und ihre Thermodynamik eingegangen wird, die Beschreibung von Optimierungsverfahren und bereits existierender etablierter Kraftfeldern an. Dabei wird auch auf das am CARB – Centre for Advanced Research in Biotechnology (Rockville, Maryland, USA) – in der Arbeitsgruppe von John Moult entwickelte Kraftfeld Bezug genommen, welches als Ausgangspunkt dieser Arbeit dient. Nach unseren Ergebnissen ist das CARB Kraftfeld jedoch für größere Systeme nicht geeignet und wurde, wie in Kapitel 4 beschrieben wird, während meiner Dissertation fast vollständig neu parametrisiert. Ziel dieser Parametrisierung war die Identifikation eines nativ-ähnlichen Zustandes des Proteins 1VII als globales Minimum der Freien Energie. Für die Parameteranpassung selbst wurden verschiedene experimentelle Daten verwendet, um ein Kraftfeld zu erhalten, welches auf eine Vielzahl von Proteinen anwendbar sein sollte. In Kapitel 5 wird gezeigt, daß dieses Kraftfeld für 6 Helixproteine eine nativ-ähnliche Konfiguration als globales Minimum der Freien Energieoberfläche identifiziert.

Bis dahin haben wir uns auf (α-) Helixproteine konzentriert, da das Strukturmotiv des (β-) Faltblattes bei kleinen Proteinen eher selten ist und kleine Faltblattstrukturen nur eine geringe Stabilität besitzen. In Kapitel 6 werden vorläufige Ergebnisse für Faltblattstrukturen präsentiert und ein Weg aufgezeigt, das vorliegende Kraftfeld systematisch auf Faltblattstrukturen zu erweitern.

Kapitel 2

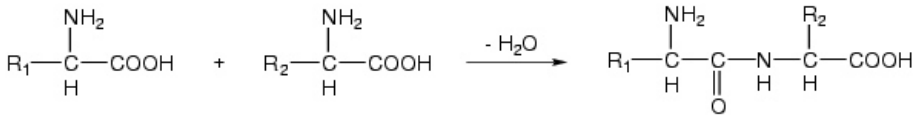
Grundlagen der Proteinfaltung

2.1 Phänomenologie der Proteine

Proteine stellen eine der wichtigsten Substanzklassen für alle lebenden Systeme dar. Die deutsche Bezeichnung "Eiweiß" ist irreführend, da Proteine in allen Organen aller Lebewesen vorkommen und wichtige Funktionen in und außerhalb der Zellen übernehmen. So sind z.B. alle Enzyme und alle Antikörper Proteine. Weitere Funktionen sind Signalübermittlung (z.B. Hormone), Gerüst- und Stützfunktion und viele andere mehr. Sie stellen etwa 17% unserer Körpermasse.

2.1.1 Chemischer Aufbau

Trotz dieser Bandbreite sind Proteine verhältnismäßig einfach aufgebaut. Sie entstehen aus einer lineare Verkettung von Aminosäuren durch Peptidbindungen. Die Peptidbindung entsteht formal durch die Eliminierung von Wasser zwischen der Carboxylgruppe und Aminogruppe zweier aufeinanderfolgenden Aminosäuren.



Praktisch alle biologisch relevanten Proteine enthalten nur 20 verschiedene Aminosäuren, die identische Hauptketten (*engl. backbone* oder *mainchain*) und unterschiedliche Seitengruppen (*engl. sidechain*) haben (Abbildung 2.2). Das Bindeglied zwischen diesen beiden ist seitens der Hauptkette das C_α-Atom. Die Atome der Seitengruppen werden von der Hauptkette fortlaufend mit β, γ, δ, ϵ und ζ indiziert.

Die Abfolge der Aminosäuren, ausgehend vom N-Terminus, bestimmt eindeutig den chemischen Aufbau eines Proteins. Diese Abfolge bezeichnet man als *Primärstruktur* oder schlicht als Sequenz. Die Länge der Sequenzen verschiede-

ner Proteine variiert sehr stark bis hin zu Proteinen mit mehreren hundert Aminosäuren.

Eine Besonderheit der Peptidbindung ist die Delokalisierung des π-Orbitals der Carboxylgruppe bis zum Stickstoff der gebundenen Aminogruppe. Dies zwingt die vier Atome der beiden beteiligten Gruppen in eine Ebene. Verkippungen dieser Peptidbindungsebene sind auf wenige Grad beschränkt und energetisch benachteiligt. Bis auf wenige Ausnahmen nehmen Peptide die *trans*-Konfiguration ein, bei der aufeinanderfolgende C_α-Atome auf gegenüberliegenden Seiten angeordnet sind. Die Bindungen des C_α-Atoms entlang der Hauptkette sind frei drehbar und bilden die Freiheitsgrade der Hauptketten. Zur Illustration kann man sich die Hauptkette als über Scharniere, welche die C_α-Atome repräsentieren, verbundene Rechtecke vorstellen (Abbildung 2.1).

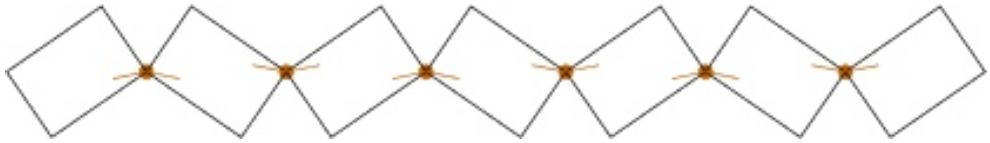


Abbildung 2.1: Reduktion der Hauptkette auf ihre Freiheitsgrade

Die Dihedralwinkel der Aminosäure i werden, wiederum vom N-Terminus ausgehend, mit ϕ_i und ψ_i bezeichnet. So läßt sich der Verlauf der Hauptkette bei fixierten Bindungsabständen anhand des Winkelsatzes $\{(\phi_i, \psi_i)\}$ eindeutig darstellen. In Gegenzug sind Strukturmotive, wie α-Helix und β-Faltblatt, mit bestimmten Winkelbereichen assoziiert. Trägt man für ein Protein die ϕ_i und ψ_i Werte in ein Koordinatensystem ein, so kann an diesem sogenannten *Ramachandran-Diagramm* das Auftreten dieser Strukturmotive abgelesen werden.

An die C_α-Atome sind entlang der Hauptkette ein Kohlenstoff- und ein Stickstoffatom gebunden. Die beiden noch fehlenden Bindungspartner sind bei Glycin 2 Wasserstoffatome. Bei allen anderen Aminosäuren ist eines der beiden H-Atome durch kleinere kohlenstoffreiche Verbindungen ersetzt. Abhängig davon, welches der Wasserstoffatome substituiert ist, spricht man von L- oder R-Aminosäuren; links- oder rechtsdrehend. Die in der Natur auftretenden Proteine sind ausschließlich aus L-Aminosäuren aufgebaut. Zwei Aminosäuren sind exemplarisch in Abb. 2.2 abgebildet. Da die Hauptketten verschiedener Proteine sich nur in der Länge unterscheiden, sind praktisch alle Eigenschaften des Proteins – Struktur, Funktion, Stabilität, Affinität zu anderen Molekülen u.v.m. – allein von den Seitengruppen bestimmt.

[1]In der Abbildung sind Kohlenstoffatome grau, Sauerstoffatome rot und die des Stickstoffs blau eingefärbt. Wasserstoffatome sind nicht eingezeichnet.

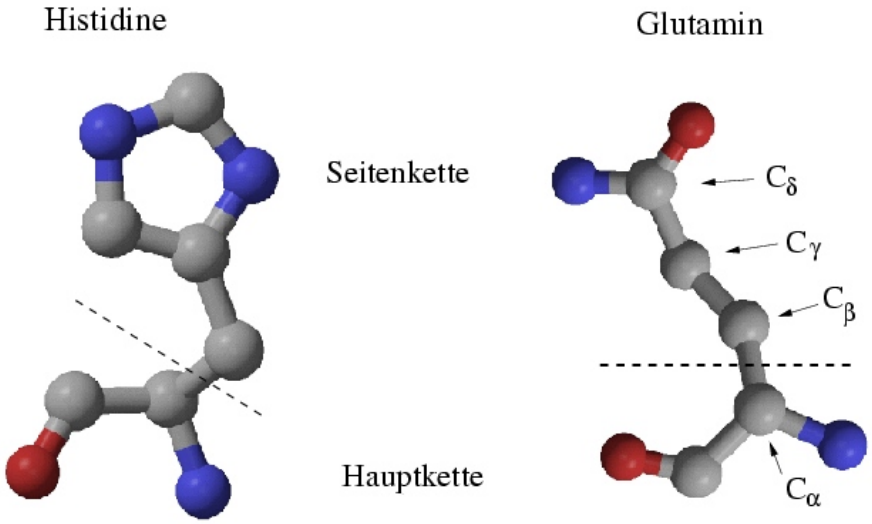


Abbildung 2.2: Zwei der zwanzig Aminosäuren [1]

2.1.2 Sekundär- und Tertiärstruktur

Die dreidimensionale Struktur eines Proteins ist durch die Aminosäuresequenz und damit von der Abfolge der Seitengruppen bestimmt [Anf73, AS75]. Als Sekundärstruktur bezeichnet man die wiederkehrenden Strukturelemente der Peptidhauptkette. Diese werden durch Wasserstoffbrückenbindungen stabilisiert, die sich zwischen den CO- und HN-Gruppen der Hauptkette ausbilden. Die wichtigsten Sekundärstrukturen sind die *α-Helix* und das *β-Faltblatt*.

Legt man zwei gestreckte Aminosäurenketten nebeneinander, so können sich bei bestimmter Anordnung Wasserstoffbrücken zwischen den beiden Strängen ausbilden. Dabei ist es aus sterischen Gründen notwendig, beide Stränge ziehharmonikaähnlich zu falten, weshalb dieses Strukturelement Faltblatt genannt wird. Den griechischen Buchstaben β hat diese Geometrie verliehen bekommen, weil das β-Keratin der Haare diese Struktur besitzt. Die Seitengruppen der Aminosäuren stehen dabei nahezu senkrecht nach oben und unten von der Faltblattebene ab[2]. Durch die Abfaltung der einzelnen Ebenen ist es möglich, daß sich Wasserstoffbrücken nicht nur zwischen gegenläufigen, antiparallelen Ketten ausbilden, sondern auch zwischen gleichläufigen, parallelen Ketten (Abbildung 2.3). Antiparallele Faltblattstrukturen entstehen häufig, indem eine gestreckte Peptidkette an 2 aufeinanderfolgenden C_α-Atomen um 180° geknickt wird. Diese Formation wird *β-Wende* (*engl. turn*) oder *Haarnadelbiegung* genannt.

Wickelt man die Peptidkette schraubenförmig um einen Zylinder, so stehen sich bei passendem Zylinderdurchmesser CO- und HN-Gruppen von Windung zu Windung gegenüber. Je nach Durchmesser ergeben sich unterschiedliche Struk-

[2]Bei einer Mischung von *L*- und *R*-Aminosäuren würden die Seitenketten teilweise in der Faltblattebene liegen und würden die Faltblattstruktur sprengen.

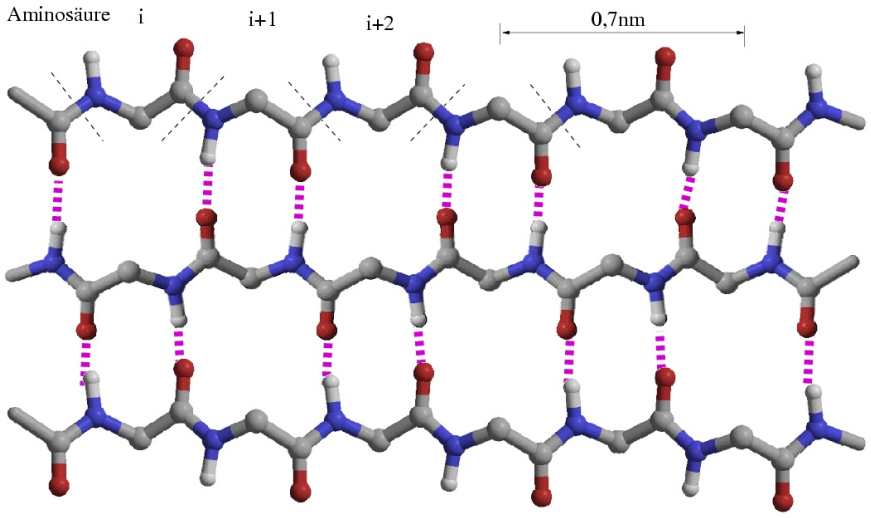


Abbildung 2.3: β Faltblattstrukturen

turen. In der Natur ist die α-Helix mit 3.6 Aminosäurenresten pro Windung sehr verbreitet (Abbildung 2.4). Die Seitengruppen weisen bei allen Helixstrukturen von der Zylinderachse weg.

Die Einteilung in Primär-, Sekundär, Tertiär- und Quartärstruktur gibt den Grad der Organisation der Struktur an. Die Primärstruktur hat keine geometrische Bedeutung, sondern beschreibt allein die Abfolge der Aminosäuren im Peptid. Unter Sekundärstruktur versteht man die erste Stufe der Organisation dieser Aminosäurensequenz. Der Begriff *Tertiärstruktur* bezeichnet die genaue dreidimensionale Struktur des Proteins, also die Position aller Proteinatome im Raum. Diese Struktur ist nicht starr, sondern populiert unter physiologischen Bedingungen ein Ensemble (mit einer RMS-Abweichung von bis zu 2Å) von Strukturen [IKP88]. Über 50% der Struktur aller strukturaufgeklärten Proteine sind in einer Form von Sekundärstruktur integriert [SBO94]. Für immer wiederkehrende Motive, die aus Sekundärstrukturelementen zusammengesetzt sind, wird gelegentlich der Begriff Supersekundärstruktur herangezogen.

Die zu einem globulären Protein zusammengefalteten Peptidketten finden sich zum Teil zu höheren Aggregaten zusammen. Man bezeichnet die räumliche Gestalt dieser Aggregate als Quartärstruktur. Meist bestehen Quartärstrukturen aus wenigen Proteinen, doch übernimmt z.B. die Eisenspeicherung beim Pferd ein Protein namens Apoferittin ($1AEW$), welches aus 20 Untereinheiten besteht.

2.1.3 Funktion und Struktur

Die räumliche Struktur eines Proteins unterliegt lokalen Fluktuationen. Es kommt jedoch nicht zu spontanen und autonomen Umlagerungen einzelner Strukturelemente. Dies garantiert, daß etwa $1AEW$ wiederholt Eisen aufnehmen und spei-

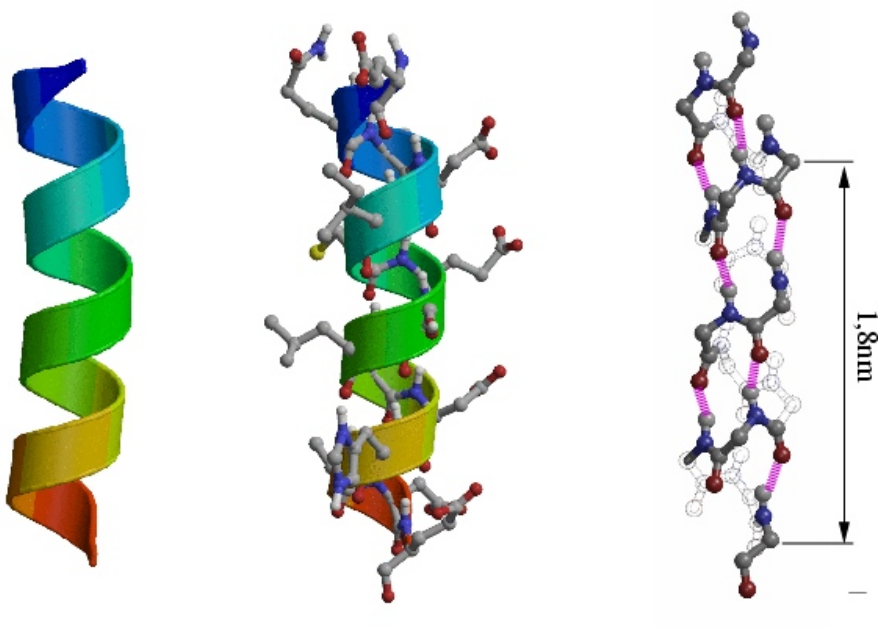


Abbildung 2.4: α Helixstruktur. Links: *Cartoon*-Darstellung. Mitte: *Cartoon* mit allen Atomen. Rechts: Hauptkettenatome mit Wasserstoffbrückenbindungen

chern kann und diese Funktion nicht zwischenzeitlich verlorengeht.

Seitdem 1958 die erste dreidimensionale Struktur eines Proteins (Myoglobin) bestimmt wurde, ist die Zahl der strukturaufgelösten Proteine auf über 20,000 Strukturen, die in einer zentralen Proteindatenbank [BWF$^+$00] ("Protein Data Bank", kurz PDB) eingetragen sind, angewachsen. Diese Datenbank ist unter $http://www.rcsb.org/pdb/$ (oder schlicht $http://www.pdb.org$) frei zugänglich und hält die Koordinaten der Proteinatome sowie einige weitere Informationen über die Proteine bereit. Die Kennung eines Proteins in dieser Datenbank besteht aus einem 4-stelligen alphanumerischen Code und definiert so einen wichtigen Standard für die Proteinnomenklatur.

Darüber hinaus gibt es Bemühungen, die existierenden Datenbankeinträge in Strukturklassen einzuteilen. Die größten Datenbanken, die sich dieser Aufgabe gewidmet haben, sind CATH [P$^+$02] ($http://www.biochem.ucl.ac.uk/bsm/cath/$) und SCOP [MBHC95] ($http://scop.berkeley.edu/$).

Das Interesse an der dreidimensionalen Struktur eines Proteins begründet sich in dem engen Zusammenhang zwischen Struktur und Funktion. Erst wenn die funktionellen Gruppen der Seitenketten richtig zusammenwirken, kann ein Protein kleinere Moleküle binden und transportieren, sich selbst an andere Zellbestandteile heften, oder die elektrostatischen Bedingungen für andere Moleküle/Liganden verändern, und so zum Beispiel die Affinität des Hemmoleküls um eine Größenordnung herabsetzen.

Es ist nicht weiter verwunderlich, daß Proteine mit vergleichbarer Struktur

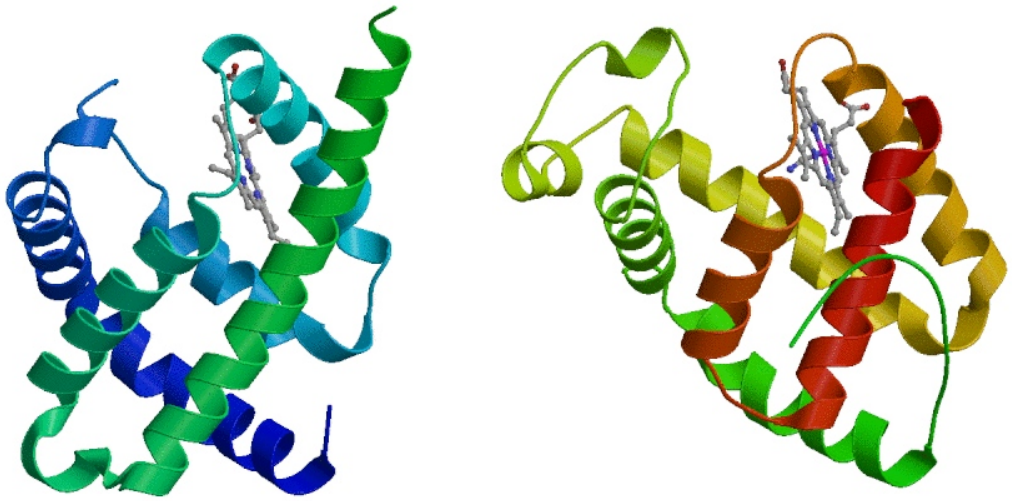


Abbildung 2.5: 1VHB (links) und 2LHB (rechts) haben ähnliche Strukturen und biologische Funktionen, obwohl ihre Sequenz nur zu 8% übereinstimmen. Der abgebildete Ligand ist das sauerstoffbindende Hemmolekül.

oftmals ähnliche Aufgaben erfüllen. Als Beispiel sei hier Vitreoscilla Stercoparia Hemoglobin (1VHB) und Petromyzon Marinus Hemoglobin (2LHB) angeführt (Abbildung 2.5). Wie die Namen schon besagen, gehören beide Proteine zur Familie des Globins, dessen Aufgabe im Sauerstofftransport besteht. Das heißt, sie besitzen beide die gleiche Funktion und wie in der Abbildung zu erkennen, auch ähnliche Strukturen. Interessanterweise ist ihre Sequenz nur zu 8% identisch.

Die Transferierbarkeit der Funktionsweise von einem Protein auf ein anderes, ist jedoch mit Vorsicht zu genießen, wie ein Vergleich von Lysozyme (1RE2) und α-lactalbumin (1A4V), welche in Abb.2.6 wiedergegeben sind, zeigt. Die Sequenz beider Proteine ist zu 40% identisch, und ihre Strukturen sind vergleichbar[3]. Dennoch gehört Lysozyme zur Proteinklasse der Enzyme (die genaue Bezeichnung ist O-glycosyl Hydrolase), wohingegen α-lactalbumin keine katalytische Funktion ausübt.

Der Unterschied zwischen 1RE2 und 1A4V erschließt sich erst, wenn man die Seitengruppen bestimmter Aminosäuren betrachtet. Wie in den meisten Fällen sind auch bei Lysozme nur wenige Aminosäuren an der Wechselwirkung mit anderen Stoffen direkt beteiligt. Bei 1A4V ist eben dieser Proteinbereich des “aktiven” 1RE2 durch einen “inaktiven” ersetzt worden. Folglich ist die Wechselwirkung mit anderen Stoffen bei gleicher räumlicher Struktur der Hauptkette deutlich verschieden.

[3]Proteine, die ein gewissen Maß an Sequenzübereinstimmung haben und vergleichbare Strukturen ausbilden, nennt man *homolog*

Auf ganz andere Weise demonstrieren Subtilisin (1GNS) und Chymotrypsin (1AB9) (Abbildung 2.7), daß die Funktionsweise eines Proteins durch wenige Aminosäuren dominiert wird. Hier ist es der Natur gelungen, für die Bewältigung einer Aufgabe zwei verschiedene Lösungen zu finden. Man spricht in diesem Zusammenhang von *konvergenter Evolution*. Bei beiden Proteinen ist die genaue Lage des katalytischen Dreiecks $Ser - His - Asp$ identisch im Raum positioniert, und sie können so auf gleiche Weise mit anderen Stoffen interagieren. Die Residuennummern des Dreiecks sind 125, 64, 32 für 1GNS und 195, 57, 102 für 1AB9. Die Atome der Seitengruppen dieser Aminosäuren sind in Abbildung 2.7 explizit dargestellt. Es ist also keinesfalls so, daß hier ein kurzes Fragment für die katalytische Wirkung verantwortlich ist, und der Rest nur die Struktur dieses Abschnittes stabilisiert. Vielmehr handelt es sich hier um einen hochgradig kooperativen Effekt.

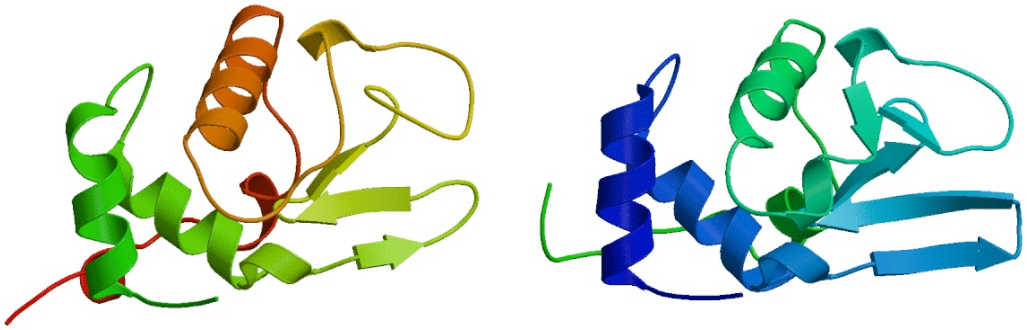


Abbildung 2.6: Lysozyme (1RE2) und α-lactalbumin (1A4V) sind in vereinfachter Hauptkettendarstellung praktisch nicht zu unterscheiden, erfüllen aber nicht die gleichen biologischen Aufgaben

Die Aminosäuren, die nicht direkt an der Funktionsweise des Proteins beteiligt sind, sind nicht zuletzt für die Stabilität des Proteins entscheidend. Ohne stabile Struktur wäre die räumliche Anordnung der aktiven Aminosäuren starken räumlichen Schwankungen unterzogen oder würde zweitweise verlorengehen. Untersuchungen von Staphylococcal Nuclease (1JOQ), einem Enzym mit 149 Aminosäuren, sind hierfür ein deutliches Beispiel. Schneidet man dieses Protein bei Residuum 126 ab, so reduziert sich die Aktivität des Proteins auf 0.12% des ursprünglichen Wertes, obwohl die Residuen des aktiven Zentrums weiterhin vorhanden sind [AS75].

Wenn die Funktionweise eines Proteins von dessen genauer Struktur und diese wiederum von der Proteinsequenz als Ganzes abhängt, wird eine Erweiterung unserer Kenntnisse über Proteinstrukturen sich positiv auf unser Verständnis von biologischen Abläufen auswirken. Dieses Verständnis wäre nicht nur dem Fachgebiet der Biologie zuträglich, sondern wäre auch in vielen Bereichen der

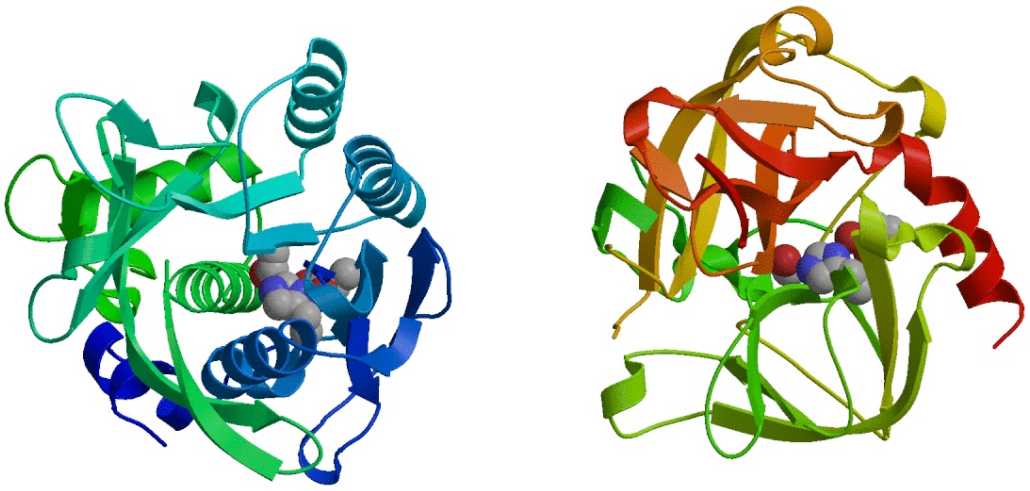


Abbildung 2.7: Ein Beispiel für konvergente Evolution. Das katalytische Dreieck (Kugelkalotten) beider Proteine (1GNS links,1AB9 rechts), ist gleich und so sind beide Proteine biologisch in gleicher Weise aktiv.

Medizin förderlich. Zum Beispiel auf dem Gebiet der Erbkrankheiten, denn Fehler in der DNS gehen in fehlerhafte Sequenzen über und können so zu fehlgefalteten Proteinstrukturen führen. Aufgrund ihrer Struktur können sie die ursprüngliche Aufgabe des Proteins nicht wahrnehmen.

2.1.4 Die thermodynamische Hypothese

Proteine nehmen unter physiologischen Bedingungen eine eindeutige native Struktur ein. Ändern sich diese Bedingungen, wie (sehr) hohe oder niedrige Temperaturen, veränderter pH-Wert oder Salzkonzentrationen, verlieren sie diese. Man spricht dann von Denaturierung. Werden diese Veränderungen zurückgenommen, geht das Protein wieder in seine native Struktur über. Anson und Mirsky konnten schon 1931 zeigen, daß Hämoglobin nach Denaturierung wieder in eine Struktur übergeht, deren Eigenschaften mit der Ausgangssubstanz identisch ist. Das wohl beeindruckendste Beispiel für reversibles Falten stammt von Anfinsen aus dem Jahr 1961 [AS75]. Bei Untersuchungen von Bovine Pancreatic Ribonuclease beobachtete er, daß dieses nach Entfalten, wobei seine natürlichen Disulfidbrücken aufbrachen, aus 105 möglichen Kombinationen der Disulfidbrücken die natürliche Anordnung wiederherstellte. Dies ist insofern beeindruckend, da das Aufbrechen zufällig gebildeter Disulfidbrücken zunächst einmal energetisch ungünstig ist und dieser Energieverlust erst durch erneute Disulfidbrückenbildung kompensiert werden kann. Folglich hängt die vom Protein eingenommene Struktur nicht von der Ausgangsstruktur ab, aus der die Rückfaltung startet, denn ansonsten wäre eine Mischung von Strukturen mit unterschiedlichen Disulfidbrücken zu erwarten. An-

finsen schloß, daß die Struktur eines Proteins durch seine chemische Zusammensetzung, also die Abfolge der Aminosäuren, bestimmt sein muß. In Kombination mit der Erkenntnis, daß die Sequenz eines Proteins in der DNS festgelegt ist und die Struktur eines Proteins seine Funktion bestimmt, hat Anfinsen gezeigt, wie in der DNS die Funktion jedes Proteins gespeichert ist.

Heute zeigt eine Vielzahl von Studien, daß reversible Faltung bei nahezu allen (wasserlöslichen) Proteinen möglich ist. Das Auftreten mehrerer konkurrierender Strukturen, wie z.B. bei Vertretern der Serpin Familie, die in zwei verschiedenen Konfigurationen vorkommen [CES91, CES93], ist eher die Ausnahme als die Regel. Demzufolge ist die native Struktur (bei reversibel faltenden Proteinen) weder von einer Ausgangskonfiguration, wie sie etwa nach der Bildung am Ribosom vorliegt, noch von Wahl (oder Existenz) eines spezifischen Faltungspfades abhängig. Ebensowenig ist für die Faltung die Beihilfe anderer Proteine notwendig. Die native Struktur wird unter physiologischen Bedingungen spontan und autonom eingenommen[4]. Sie ist folglich eine *Zustandsfunktion*, welche dem globalen Minimum der Freien Energie zustrebt. Diese von Anfinsen 1962 ursprünglich als *thermodynamische Hypothese* bezeichnete Eigenschaft der Proteinfaltung ist heute allgemein akzeptiert [Anf73]. Sie bezieht sich auf das Verhalten des Gesamtsystems und steht nicht im Widerspruch zu der Beobachtung, daß Proteine bei hoher Konzentration aggregieren können, wie dies etwa bei der Alzheimer Erkrankung der Fall ist.

Seither sind viele Bemühungen unternommen worden, die Mechanismen zu ergründen, welche es einem Protein ermöglichen, aus einer beliebigen Konfiguration heraus seine native Struktur zu finden. Die bisherigen Experimente vermitteln den Eindruck, daß es für verschiedene Proteine unterschiedliche Faltungsstrategien zu geben scheint. Eine endgültige Klärung des Faltungsmechanismus steht jedoch noch aus. Der nächste Abschnitt versucht, einige der bisherigen Erkenntnisse kurz zusammenzustellen.

2.1.5 Der Beginn der Faltung

Die random-coil Struktur

Ein Modell zur Proteinfaltung muß bei bei einer sogenannten *random-coil* Struktur beginnen und bei einer eindeutigen "gefalteten" Struktur enden, die alle nativen Kontakte aufweist. Der Begriff der nativen Struktur ist eindeutig und wohldefiniert. Hingegen ist die Natur der random-coil Struktur weniger erschlossen, und es existieren unterschiedliche Definitionen. Dies liegt nicht zuletzt daran, daß es sich hier nicht um einen einzelnen Zustand, sondern um ein Ensemble handelt. Der Begriff random-coil Ensemble hat sich jedoch nicht etablieren können. 1970

[4]Die in der Natur auftretenden Chaperone (*engl.* Anstandsdame) sind Proteine, die der Faltung anderer Proteine assistieren. Sie wirken jedoch nur als Katalysator eines Vorgangs, der auch ohne sie stattfinden würde.

definierte Tanford den random-coil Zustand eines Polymeres als eine Konfiguration, in der Drehungen um jeden frei rotierbaren Winkel möglich sind, wie dies in kleinen Molekülen der Fall ist [Tan70]. Shortle definiert den random-coil Zustand als diejenige Struktur, in der keine Seitengruppen-Seitengruppen Wechselwirkungen auftreten [Sho96]. Smith et al. haben angenommen, daß die Winkel jedes einzelnen Residuums in einer random-coil Konfiguration unabhängig von den Winkeln aller anderen Residuen sind [SFSD96].

Wesentlich an allen Definitionen ist, daß im nativen Zustand nur ein Winkelsatz mit kleinen Fluktuationen populiert ist, wohingegen für random-coil Zustände ein breites Spektrum an Winkeln existiert. Dies bedeutet wiederum, daß random-coil Zustände in einem sehr großen Phasenraum existieren und eine entsprechend hohe Entropie besitzen. Demgegenüber ist der Phasenraum der nativen Struktur praktisch singulär. Der hohe Verlust an Konfigurationsentropie während des Faltungsprozesses ist diesem entgegengerichtet und begünstigt die random-coil Konfigurationen. Er ist gleichzeitig der betragsgrößte Beitrag zur Freien Energie.

Nimmt man an, daß für jede Seitengruppe drei gleichpopulierte niederenergetische Konfigurationen existieren, von denen in der nativen Struktur nur eine eingenommen wird, so läßt sich die Änderung der Konfigurationsentropie einer Seitengruppe zu

$$\begin{aligned} \Delta S &= -R\left(\sum_{i=1}^{1} 1\ln(1) - \sum_{i=1}^{3}(1/3)\ln(1/3)\right) \\ &= -R\ln 3 \\ &= -9.1\frac{J}{mol\,K} = -2.2\frac{cal}{mol\,K} \end{aligned}$$

abschätzen (Anhang A). Bei Raumtemperatur trägt die Konfigurationsentropie jedes Residuums somit rund $0.7 kcal\,mol^{-1} = 2.7\,kJ\,mol^{-1}$ zur Freien Energie der Faltung bei.

Der hydrophobe Kollaps und das Levinthalsche Paradoxon

Random-coil Konfigurationen sind im allgemeinen räumlich ausgedehnt, um Drehungen um die Dihedralwinkel zu ermöglichen. In wässeriger Lösung wären die Aminosäuren dieser Proteinstuktur praktisch alle in Kontakt mit dem Lösungsmittel. Nach allgemein akzeptierter Hypothese, die von Gittersimulationen bestätigt wird [SSK94a, Bal94], kollabiert diese sehr rasch zu einer semi-kompakten Struktur, wodurch die Kontaktfläche zum Lösungsmittel deutlich reduziert wird. Dieser Mechanismus wird als *hydrophober Kollaps* bezeichnet.

Der hydrophobe Kollaps ist unter anderem am Protein barnase (1BNR) experimentell beobachtet worden [Wal96, FD96].Die random-coil und die semi-kompakte *ungeordnete Tröpfchenstruktur* (*random globule*) haben einen Gyra-

tionsradius[5] von 31 beziehungsweise 15Å. Letzterer ist dem Gyrationsradius der nativen Struktur mit 13.4Å sehr ähnlich.

Röntgenkleinwinkelstreu-Experimente an Myoglobin haben weitere Einblicke in die Natur dieses Kollapses erlaubt. So tritt bei Myoglobin der Kollaps ein, bevor sich native Kontakte ausbilden [EJW+95]. Letztere bilden sich scheinbar erst in einem späteren Zeitpunkt der Faltung. In dem ungeordneten semi-kompakten Zustand kompensiert die unspezifische Gruppierung hydrophober Seitenketten im Inneren den Verlust an Konfigurationsentropie.

Für Myoglobin wurde die Anzahl der möglichen Hauptkettenkonfigurationen aus einer Abschätzung der Hauptkettenentropie pro Residuum bestimmt [SSK94b]. Hiernach existieren circa 10^{44} (≈ 2.57 Konfigurationen pro Residuum) random-coil Zustände, die zu Beginn des Faltungsprozesses in circa 10^{26} (≈ 1.7 Konfigurationen pro Residuum) zufällige semi-kompakte Strukturen kollabieren. Am Ende der Faltung verbleibt nur eine einzelne Struktur, die native (Abbildung 2.8).

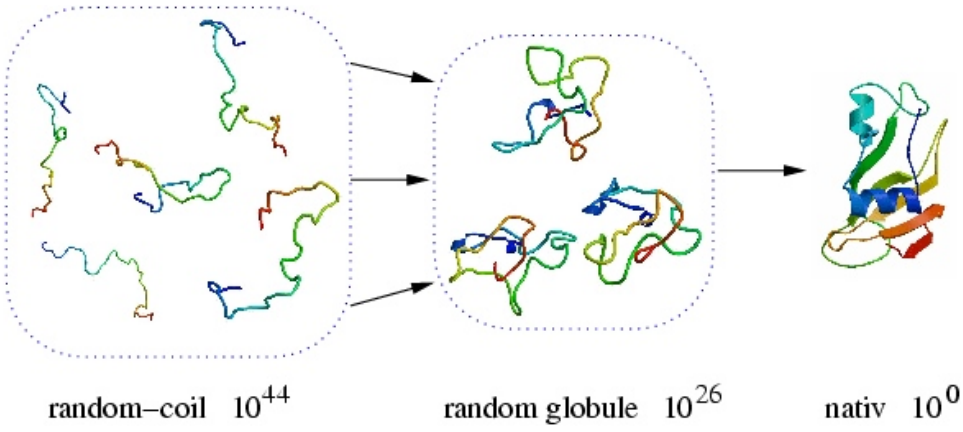


Abbildung 2.8: Hydrophober Kollaps des Proteins barnase ($1BNR$)

Wenn es dem Protein möglich wäre, in nur $10^{-11}s$ zwischen zwei dieser 10^{44} Konfigurationen zu wechseln, so wären $10^{33}s \approx 10^{25}$ Jahre notwendig, jede dieser Konfigurationen einmal einzunehmen. Allein im Raum der semi-kompakten Zustände wären hierzu $10^{26} \cdot 10^{-11}s = 10^{15}s \approx 30$ Millionen Jahre aufzubringen. Somit bleibt auch nach dem hydrophoben Kollaps ein so großer Konfigurationsraum übrig, daß ein Protein von einer semi-kompakten oder gar random-coil

[5]Der Gyrationsradius ist die mittlere Abweichung der Proteinatome von deren "masselosen" Schwerpunkt.

$$R_g^2 = \sum_{\text{Atome } i} \left(\vec{r}_i - \frac{1}{N}\sum_j \vec{r}_j\right)^2$$

Diese Definition weicht von der üblichen Definition mit Masse ab, doch ist der Unterschied, da C, N, O ähnlich schwer sind, nur gering.

Struktur ausgehend, die native Struktur nicht durch zufälliges ausprobieren der Konfigurationen finden kann.

Levinthal schloß, daß Faltungspfade existieren müssen, die eine random-coil Zustand zur nativen Struktur führen[6] [Lev68]. Der Endpunkt des Faltungspfades mußte, nach der Auffassung von Levinthal, nicht notwendigerweise im Minimum der Freien Energie liegen.

Mit dem Konzept des Faltungspfades geht einher, die Proteinfaltung als sequentielle chemische Reaktion aufzufassen, für die eine einzelne Reaktionkoordinate existiert. Für sehr kleine Peptide ist dieses Bild experimentell bestätigt [AS75]. Bei diesen Peptiden sind thermodynamisch nur zwei Zustände bedeutend, der ungefaltete U und der gefaltete/native Zustand N. Der Übergang $U \rightleftharpoons N$ ist einem diskontinuierlichem Phasenübergang ähnlich (*quasi-first-order*), wobei eine moderate Energiebarriere die Zustände U und N trennt.

Die beiden großen Beiträge zur differentiellen Freien Energie sind Konfigurationsentropie und die hydrophobe Wechselwirkung, wobei der Verlust an Konfigurationsentropie der Faltung entgegenwirkt und der hydrophobe Effekt die native Struktur gegenüber seinem unstrukturierten Pendant begünstigt [Dil90, Hon99, Tan70]. Daher weist der hydrophobe Effekt vom random-coil Zustand in Richtung des nativen. Die Änderung der anderen Wechselwirkungen wird in Vergleich hierzu als eher klein eingestuft [Dil90].

Die Zusammensetzung der Freie Energie des Faltungsprozesses läßt sich wie in Abb. 2.9 dargestellt illustrieren. Um dieser Abbildung Zahlen zuzuordnen, sei hier Lysozyme mit 130 Residuen als Beispiel aufgeführt [MP95]. Der Beitrag der unpolaren Gruppen zur Freien Energie der Denaturierung bei $25°C$ wurde zu 450 $kcal\,mol^{-1}$ und der der polaren Gruppen zu 87 $kcal\,mol^{-1}$ bestimmt, wobei erster durch der hydrophoben Effekt dominiert wird. Von dieser Energie ist die Konfigurationsentropie von 523 $kcal\,mol^{-1}$ bei $25°C$ abzuziehen. Die Freie Energie der Entfaltung beträgt somit nur 14 $kcal\,mol^{-1} = 59\,kJ\,mol^{-1}$, etwa 0.1 $kcal\,mol^{-1} = 0.5\,kJ\,mol^{-1}$ pro Residuum.

2.1.6 Stabilität der Proteine

Der Unterschied in der Freien Energie zwischen einer random-coil und der nativen Struktur beträgt bei Raumtemperatur generell nur leicht über 0.1 $kcal\,mol^{-1}$ pro Residuum. Bei starker Erwärmung geht die native Struktur des Proteins verloren. Wie hoch die Denaturierungstemperatur liegt, ist von vielen Faktoren abhängig. Mutationsexperimente, bei denen einzelne Aminosäuren ausgetauscht werden, haben hier viele Einsichten gewährt. Die von Forschern häufigst angeführ-

[6] Levinthal hatte 1968 keine auf experimentellen Daten beruhenden Abschätzung der Konfigurationsraumgröße vorliegen. Seine Schätzung für Barnase (mit 110 Residuen) läge bei $3^{109} \approx 10^{52}$, da er von 3 Konfigurationen pro Residuum, genauer gesagt pro Hauptkettendihedralwinkelpaar, ausging. Dies überschätzt den Zeitaufwand zwar, aber seine Argumentation gilt auch im vermeintlich kleinen 10^{26} Einträge umfassenden Konfigurationenraum.

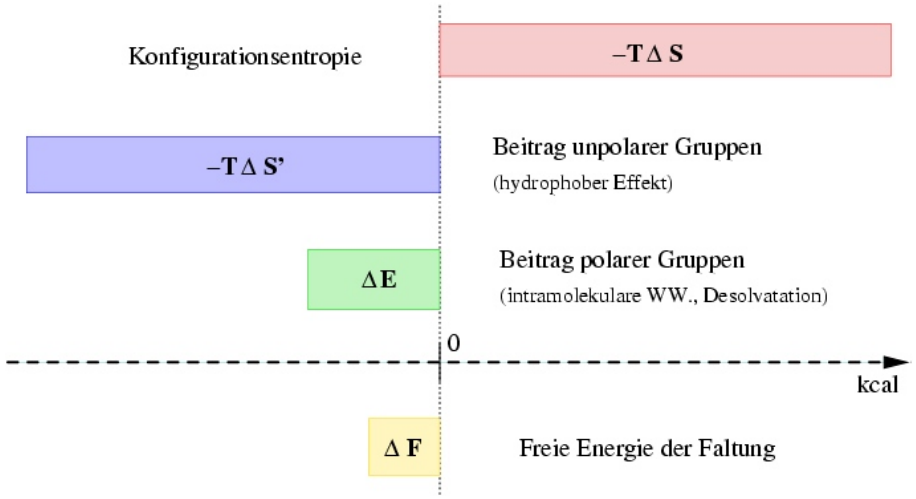


Abbildung 2.9: Beiträge zur Freien Energie des Faltungsprozesses

ten Faktoren, die zu einer größeren Thermostabilität führen, sind zum einen eine erhöhte Zahl von (Hauptketten-) Wasserstoffbrückenbindungen, und zum anderen eine Erhöhung des polaren Anteils der Proteinoberfläche [VWA97]. Doch obwohl der hydrophobe Effekt die treibende Kraft der Faltung ist, werden Proteine thermisch nicht stabiler, wenn sie mehr unpolare Gruppen vor dem Lösungsmittel verbergen können [HY95, Hon99].

Um ein Protein in seiner nativen Struktur nicht nur gegenüber den völlig unstrukturierten random-coil Zuständen, sondern auch gegenüber kompakten Strukturen zu stabilisieren, ist eine ausgeprägt Energielücke (in der inneren Energie) notwendig, da die Konfigurationsentropie des nativen Zustandes sehr gering ist. Einige Gittersimulationen haben gezeigt, daß Proteine mit großer Energielücke rasch falten [SSK94a]. In diesen Simulationen wurde auch die umgekehrte Folgerung aufgezeigt, daß nämlich alle rasch faltenden Protein eine ausgeprägte Energielücke haben. Diese Umkehrung hat sich als Eigenschaft der speziellen Simulation erwiesen, und konnte widerlegt werden [UM96, PO98].

Heute geht man davon aus, daß die Faltungsrate sehr wahrscheinlich von der *Kontaktordnung* (engl. *contact-oder* CO) des Proteins abhängig ist [BRTB02]. Die Kontaktordnung ist der mittlere Sequenzabstand der Residuen, die in der Tertiärstruktur Kontakte ausbilden, geteilt durch die Sequenzlänge. Die Abhängigkeit der Faltungsrate von dieser Größe ist darin begründet, daß Proteine mit geringer Kontaktordnung schon im frühen Faltungsstadium stabilisierende Wechselwirkungen und damit Kontakte ausbilden können, die den Verlust an Konfigurationsentropie aufwiegen, wohingegen Proteine mit vielen nicht-lokalen Kontakten[7]

[7]Im Kontext der Proteinfaltung unterscheidet man lokale und nicht-lokale Wechselwirkungen von kurz- und langreichweitigen. Lokal ist hierbei bezüglich des Anstands in der Sequenz zu verstehen. Die (räumlich) langreichweitige Coulomb-Wechselwirkung hingegen unterwirft Sequenznachbarn in der gleichen Weise wie Aminosäuren, die in der Sequenz weit voneinander

erst eine *Entropiebarriere* überwinden müssen.

Die Notwendigkeit der Existenz einer Energielücke für eine stabile native Struktur bleibt weiterhin bestehen. Da wir nur ungenügend Zugang zur Konfigurationsentropie haben, weil sie nur in linearer Approximation über die Lösungsmittelparameter in unser Potentialoberfläche der Freien Energie enthalten ist, können wir den Unterschied zwischen dem nativen Ensemble und des random-coil Ensemble nicht auflösen. Wie in allen Kraftfelder der Freien Energie, sind in PFF01 random-coil Zustände stets mit äußerst schlechten Werten in der Freien Energie behaftet. Die Proteinstrukturvorhersage kann sich auf den Unterschied in der Freien Energie zwischen kompakten niederenergetischen Strukturen beschränken, weshalb diese Fehleinschätzung der random-coil Zustände nicht weiter von Bedeutung ist.

Die absoluten Zahlenwerte für die Freie Energie einzelner Konfigurationen sind nicht sehr aussagekräftig, vielmehr sollte man sich auf die Unterschiede in der Freien Energie zwischen zwei Strukturen konzentrieren. Für eine stabile native Struktur ist ein Unterschied in der Freien Energie zu anderen kompakten Strukturen von nicht weniger als $2k_BT$, also grob $1\,kcal\,mol^{-1}$ (Anhang A). Die Einheit $kcal\,mol^{-1}$ ist somit die *natürliche* Energieskala biologischer Systeme.

2.1.7 Das Framework- und das Nukleation-Kondensations-Modell

Es ist unstrittig, daß ein random-coil Zustand zu Beginn der Faltung sehr rasch in eine semi-kompakte Struktur übergeht. Unklar ist jedoch, wieviel Struktur er besitzt und wie diffus dieser Zwischenzustand ist. Myoglobin präsentiert sich als unstrukturierter Tropfen. Doch andere Proteine erscheinen in einem anderem Licht. Bei Ribonuclease A (kurz: RNase A) bilden sich frühzeitig (native) Sekundärstrukturmotive aus, die eher auf einen Faltungsmechanismus gemäß eines sogenannten Framework-Modells schließen lassen [UB88].

Im Framework-Modell bilden sich zunächst stabile Sekundärstrukturelemente aus, die die notwendigen Rahmenbedingungen für die nachfolgende Anordnung dieser Elemente zur vollständigen nativen Struktur bilden. Die Bildung der Tertiärstruktur soll durch einen Diffusionsprozeß erfolgen.

Dies steht im Widerspruch zu Experimenten an CI2 (Barey Chymotrypsin Inhibitor 2, PDB-Code: 2CI2), einem Protein mit 83 Residuen. Seine Faltungsdynamik entspricht einem diskontinuierlichem Phasenübergang von einer relativ offenen Struktur aus, wie sowohl NMR als auch Messungen des zirkularen Dichroismus (CD) gezeigt haben [DLI+96, JeF93]. Im Framework Modell sollten Proteinfragmente in wässeriger Lösung stabile Sekundärstrukturanteile ausbilden, die der nativen Struktur zumindest ähnlich sind. Experimente, in denen das Protein beginnend beim N-Terminus sukzessive aufgebaut wurde, haben jedoch

---

entfernt sind.

die Ausbildung der Helix erst beobachten können, nachdem nahezu das gesamte Protein zusammengesetzt war (Residuen 1-60). Es gibt auch keine schlüssigen Hinweise, daß kleinere Fragmente überhaupt eine eindeutige und stabile Struktur aufweisen. Dies bedeutet nicht, daß die Helix oder andere Strukturelemente sich nicht kurzzeitig ausgebildet haben können.

Beeindruckend ist in diesem Kontext die Beobachtung, daß, wenn man aus CI2 Residuum 40 und 41 herausschneidet und damit in 2 Fragmente teilt, keines dieser Fragmente irgendeine Struktur ausbildet. Mischt man jedoch die Fragmente, so assoziieren die beiden Peptidketten unverzüglich zu einer nativähnlichen Struktur (1CIQ) [GRSDF94] (Abbildung 2.10). Ein Resultat, welches mit dem Framework Modell nicht vereinbar ist.

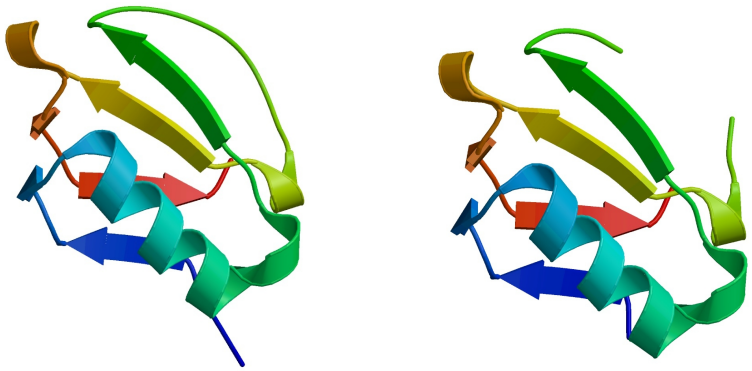


Abbildung 2.10: Struktur des Barey Chymotrypsin Inhibitor 2 (links,2CI2) und die Struktur der beiden Fragmente (rechts,1CIQ)

In Kombination mit (Punkt-)Mutationsexperimenten, bei denen einzelne Aminosäuren (meist durch Aminosäuren ohne funktionelle Gruppen) ersetzt werden, und Molekulardynamik Simulationen, hat sich das Nukleation-Kondensation Modell entwickelt. Demnach bildet sich zunächst die Helix aus (Nukleation), gefolgt von der "Stabilisierung" durch drei spezifische Seitengruppen-Wechselwirkungen zwischen Lysin-2, Glutamin-7 und Asparagin-23[8]. Nach der Vorstellung dieses Modells kondensiert anschließend die Struktur um diesen Keim.

Dieses Modell ist mit den meisten Beobachtungen an anderen Proteinen konsistent. Allerdings kann es unter anderem nicht erklären, warum das CI2 Fragment aus den Aminosäuren 1 bis 40 nicht ohne das fehlende Stück eine stabile Helix ausbildet, obwohl die für die Strukturbildung wichtige Seitengruppenwechselwirkung 2-7-23 innerhalb dieses Fragments liegt.

[8]Die Farbgebung in Abb. 2.10 ist blau für den N-Terminus/Residuum 1 und rot für den C-Terminus/Residuum 64. Die Helix umfaßt die Residuen 12 bis 24.

Es gibt noch eine Reihe offener Fragen bezüglich des Faltungsmechanismus. Das Nukleation-Kondensation Modell an CI2 kombiniert Informationen aus Kernspinresonanzspektroskopie (NMR), Messungen des zirkularen Dichroismus (CD), Punktmutationen und Molekulardynamik Simulationen. Schon für ein so kleines Protein sind viele Methoden notwendig, um einen Eindruck über das Faltungsszenario zu erlangen. Dies zeigt, in welch komplexem Bild sich der Mechanismus der Proteinfaltung präsentiert. Möglicherweise wird in naher Zukunft ein konsistentes Modell der Proteinfaltung existieren, doch das bleibt abzuwarten. Daher möchte ich das Kapitel mit einer Illustration der momentan im Gespräch befindlichen Faltungsmodelle abschließen (Abbildung 2.11).

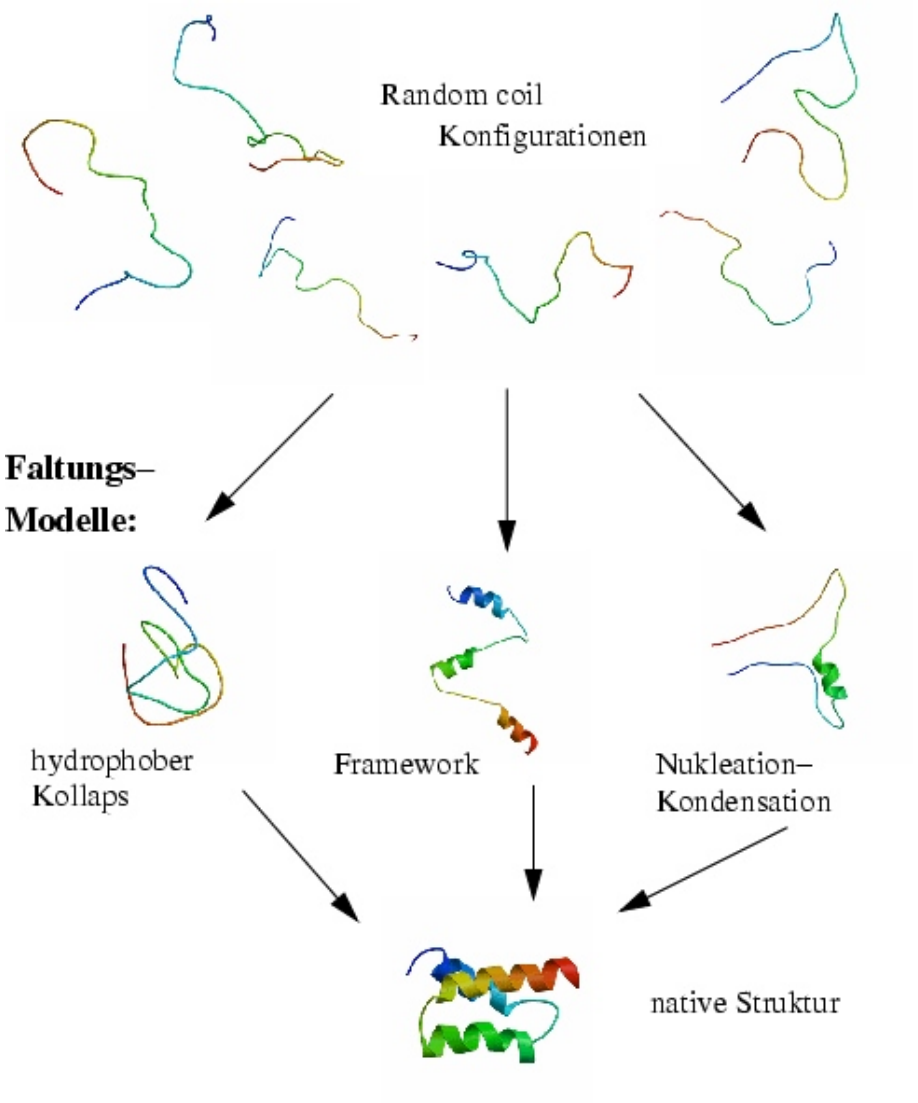


Abbildung 2.11: Die drei gängigen Modelle der Proteinfaltung. a) hydrophobe Kollaps Modell, b) Framework Modell c) Nukleation-Kondensations Modell

2.2 Thermodynamische Grundlagen der Proteinfaltung

In den folgenden Abschnitten soll der Zusammenhang zwischen der Hamiltonfunktion und der Freien Energie durch gewissen Näherungen vereinfacht werden. Anschließend werden dann die im Rahmen dieser Arbeit wichtigsten stochastischen Optimierungsverfahren erläutert. Die verwendeten Notationen sind in Anhang B beschrieben.

Die native Struktur ist gemäß der thermodynamischen Hypothese das Minimum der Freien Energie F. Daher wollen wir uns etwas näher mit diesem thermodynamischen Potential auseinandersetzten, welches nach den Gesetzen der statistischen Physik mit dem Zustandsintegral $\mathcal{Q}$ verknüpft ist

$$F = -\beta^{-1} \ln \mathcal{Q} \, . \tag{2.1}$$

$\mathcal{Q}$ wird durch Integration des Boltzmann Faktors $e^{-\beta H(\vec{p},\vec{q})}$ über den Phasenraum[9] $\Gamma = (h^{3N})^{-1} \iint d\vec{p}\, d\vec{q}$ des Systems berechnet. $H(\vec{p},\vec{q})$ ist der Hamiltonian des Protein/Lösungsmittel-Systems in den verallgemeinerten Koordinaten $\vec{q}$ und Impulsen $\vec{p}$.

Wenn die Wahrscheinlichkeitsverteilung $\rho(\vec{p},\vec{q}) \sim e^{-\beta H(\vec{p},\vec{q})}$, wie bei der nativen Struktur, um eine Konfiguration konzentriert ist, so läßt sich die Verteilung um diese Struktur herum entwickeln. Die Freie Energie kann dann in harmonischer Näherung als Energie dieser Struktur plus einem Vibrationsbeitrag geschrieben werden. Wenn mehrere energetisch gleiche Minima eng zusammenliegen, so ist dieser Summe noch ein weiterer Entropieterm zuzuschlagen, der Konfigurationsentropie S_{konf} genannt wird. Am Ende dieses Abschnittes erhalten wir

$$\Delta F = \Delta E_{protein} - T\Delta S_{wasser} - T\Delta S_{konf} \, . \tag{2.2}$$

Diesem Ergebnis wollen wir uns nun schrittweise nähern.

2.2.1 Das Konfigurationsintegral

Schreibt sich die Hamiltonfunktion als Summe von kinetischer und potentieller Energie $H(\vec{q},\vec{p}) = T(\vec{p}) + V(\vec{q})$, mit den Ortkoordinaten $\vec{q}$ und den Impulskoordinaten $\vec{p}$, so kann der kinetische Anteil im kanonischen Zustandsintegral $\mathcal{Q}$

[9] Der Vorfaktor des Zustands- bzw. Phasenraumintegral, welcher *korrekte Boltzmann Abzählung* genannt wird, ist $(h^{3N})^{-1}$ und nicht $(h^{3N}N!)^{-1}$, da die Konfigurationen der Proteinatome aufgrund ihrer Bindungen zum Rest des Proteins *unterscheidbar* sind. Formal müßte hier zwischen den Wassermolekülen und dem Protein unterschieden und ein Faktor $N_w!$ für die Wassermoleküle notiert werden. Da wir jedoch später auf ein implizites Lösungsmittelmodell wechseln, wird hier auf diese Unterscheidung verzichtet.

abgespalten werden.

$$\begin{aligned} \mathcal{Q} &= \frac{1}{h^{3N}} \int exp(-\beta H(\vec{q},\vec{p}))\, d\vec{p}\, d\vec{q} && (2.3)\\ &= \frac{1}{h^{3N}} \int exp(-\beta\,[T(\vec{p}) + V(\vec{q})])\, d\vec{p}\, d\vec{q} && (2.4)\\ &= \underbrace{\frac{1}{h^{3N}} \int exp(-\beta T(\vec{p}))\, d\vec{p}}_{\lambda^{-3N}} \cdot \underbrace{\int exp(-\beta V(\vec{q}))\, d\vec{q}}_{\mathcal{Z}} && (2.5) \end{aligned}$$

Das Integral der kinetischen Energie läßt sich ausrechnen und führt zur Definition der *thermischen De-Broglie-Wellenlänge* λ

$$\frac{1}{\lambda^3} := \frac{1}{h^3} \int exp\left(-\frac{\beta\vec{p}^2}{2m}\right) d\vec{p} = \left(\frac{2\pi m}{\beta h^2}\right)^{\frac{3}{2}} \quad (2.6)$$

Die Freie Energie ist dann[10]

$$F = -\beta^{-1} \ln \lambda^{-3N} - \beta^{-1} \ln \mathcal{Z} \quad (2.7)$$

Der kinetische Anteil an der Freien Energie ist nur abhängig von der Temperatur des Systems und geht folglich bei fester Temperatur als Konstante in die Freie Energie ein. Für die Proteinstrukturvorhersage spielen konstante Anteile keine Rolle, da wir nur an Differenzen in der Freien Energie interessiert sind und wir können den kinetischen Anteil von vornherein aus der Betrachtung ausschließen. Die für die Strukturvorhersage wichtige Freie Energie F lautet somit

$$\begin{aligned} F &= -\beta^{-1} \ln \mathcal{Z} && (2.8)\\ &= -\beta^{-1} \ln \left(\int exp(-\beta V(\vec{q}))\, d\vec{q} \right) . && (2.9) \end{aligned}$$

2.2.2 Implizite Lösungsmittel

Ausgehend von den ersten beiden Atomen des Proteins, läßt sich die Konfiguration durch Angabe der Dihedralwinkel, Bindungswinkel und -längen eindeutig angeben. Sei $\mathcal{P}$ die Menge aller Proteinkonfigurationen $\vec{q}$. Der Konfigurationsraum des Lösungsmittels sei $\mathcal{W}$, dessen Elemente wir als $\vec{w}$ notieren. In den Simulationen werden wir es nur mit Wasser als Lösungsmittel zu tun haben. Der Hamiltonian des Gesamtsystems $H_{gesamt}(\vec{q},\vec{w})$ teilt sich in drei Komponenten auf. In den Hamiltonian $H_p(\vec{q})$, der intramolekularen Wechselwirkungen des Proteins,

[10] Die Schreibweise λ^{3N} gilt strenggenommen nur, wenn alle Teilchen die gleiche Masse habe. Für die hier behandelten Systeme lautet die korrekte Schreibweise $\lambda_C^{3\#C} \cdot \lambda_H^{3\#H} \cdot \lambda_N^{3\#N} \cdot \lambda_O^{3\#O} \cdot \lambda_S^{3\#S}$, mit den unterschiedlichen Wellenlängen für Kohlenstoff, Wasserstoff, Stickstoff, Sauerstoff und Schwefel.

den Hamiltonian $H_w(\vec{w})$, der entsprechend nur die Beschreibung des Wassers übernimmt, und den Wechselwirkungshamiltonian zwischen Protein und Wasser $H_{pw}(\vec{q},\vec{w})$.

$$H_{gesamt}(\vec{q},\vec{w}) = H_p(\vec{q}) + H_{pw}(\vec{q},\vec{w}) + H_w(\vec{w}) \tag{2.10}$$

Für die Freie Energie gilt dann

$$\begin{aligned} F &= -\beta^{-1}\ln \mathcal{Z} && (2.11)\\ \mathcal{Z} &= \int_{\mathcal{P}\otimes\mathcal{W}} exp(-\beta H_{gesamt}(\vec{q},\vec{w}))\; d\vec{w}\; d\vec{q} && (2.12)\\ &= \int_{\mathcal{P}\otimes\mathcal{W}} exp(-\beta H_p(\vec{q}))\; exp(-\beta\left[H_{pw}(\vec{q},\vec{w}) + H_w(\vec{w})\right])\; d\vec{w}\; d\vec{q} && (2.13)\\ &= \int_{\mathcal{P}} exp(-\beta H_p(\vec{q}))\;\cdot && (2.14)\\ &\quad \cdot\left\{\int_{\mathcal{W}} exp(-\beta\left[H_{pw}(\vec{q},\vec{w}) + H_w(\vec{w})\right])\; d\vec{w}\right\} d\vec{q} && (2.15) \end{aligned}$$

Das Integral über $\mathcal{W}$ können wir zu einen effektiven Lösungsmittelhamiltonian umschreiben:

$$\widehat{H}_w^\beta(\vec{q}) = -\frac{1}{\beta}\ln\left\{\int_{\mathcal{W}} exp(-\beta\left[H_{pw}(\vec{q},\vec{w}) + H_w(\vec{w})\right])\; d\vec{w}\right\}\;. \tag{2.16}$$

Wir erhalten somit das **Konfigurationsintegral mit implizitem Lösungsmittel**

$$\mathcal{Z} = \int_{\mathcal{P}} exp\Big(-\beta\left[H_p(\vec{q}) + \widehat{H}_w^\beta(\vec{q})\right]\Big)\; d\vec{q}\,. \tag{2.17}$$

Wenn wir von lokalen Minima sprechen, so meinen wir damit die Minima von $H_p(\vec{q}) + \widehat{H}_w^\beta(\vec{q})$, und die Energiedifferenzen zwischen zwei Strukturen I und J notieren wir zu $\Delta E = E^I - E^J = [H_p(\vec{x}^{\,I}) - H_p(\vec{x}^{\,J})] + [\widehat{H}_w(\vec{x}^{\,I}) - \widehat{H}_w(\vec{x}^{\,J})]$. Zum jetzigen Zeitpunkt ist der effektive Lösungsmittelhamiltonian ein rein formales Hilfsmittel ohne eigene physikalische Bedeutung. Wir werden uns jedoch später auf die Betrachtung der Proteinkonfigurationen in lokalen Minima von $H_p + \widehat{H}_w^\beta$ beschränken können. Die Zeit für den Übergang der Proteinkonfiguration zwischen zwei Minima ist nach dem Ergebnis von Molekulardynamik Simulationen deutlich größer als die Relaxationszeit des Wassers [CK95]. Die der statistischen Physik zugrundeliegende Äquivalenz zwischen *Ensemblemittel* und *Zeitmittel* können wir somit auch auf den Hamiltonian $\widehat{H}_w$ anwenden. Somit entpuppt sich der effektive Hamiltonian als Freie Energie des Wassers bei fester Proteinkonfiguration.

Die Notation $\widehat{H}_w^\beta$ für den implizieren Lösungsmittelhamiltonian soll andeuten, daß es formal möglich ist, für das Lösungsmittel eine andere Temperatur anzusetzen als für das Protein. Diese Bemerkung ist notwendig, da das Modell für das

Lösungsmittel bei uns auf $300K$ festgelegt ist, die Temperatur des Proteins aber variiert wird.

Im Lösungsmittelterm sind zwei verschiedene Effekte subsumiert. Zum einen die Hydrophobizität, welche an der Kontaktfläche zwischen Protein und dem Wasser auftritt, und die Wirkung auf die Elektrostatik. Für die beiden Anteile werden wir später unterschiedliche Näherungen ansetzten.

2.2.3 Bindungswinkel und -längen

Eine Proteinkonfiguration ist durch die Dihedralwinkel ϕ, ψ und χ der Haupt- und Seitenketten (Anhang B), sowie der Bindungswinkel und -längen α, l bestimmt. Da Dihedralwinkel "frei" drehbar sind, liegen die zugehörigen Eigenfrequenzen mehrere Größenordnungen unterhalb derjenigen der "starren" chemischen Bindungen. Mit hohen Frequenzen der starren chemischen Bindung geht einher, daß deren Amplituden nicht sonderlich groß sind. Die Bewegung der Bindungswinkel und -längen erfolgt daher auf einer anderen Zeitskala als die der Dihedralwinkel. Anders ausgedrückt: werden die Dihedralwinkel von Zeitpunkt t auf $t + \Delta t$ verändert, so relaxieren die Bindungen innerhalb des Zeitfensters Δt. Folglich kann man die Bewegungsgleichungen dieser Freiheitsgrade trennen, und wir können uns auf die Rotationen der Dihedralwinkel konzentrieren.

Dennoch lassen Molekulardynamik Simulationen den Schluß zu, daß die Vibrationen der Bindungen einen signifikanten Beitrag zur Freien Energie liefern [KIP87]. Inwieweit dieser Beitrag bei der Differenzbildung zwischen zwei Konfigurationen herausfällt, wird im folgenden Abschnitt diskutiert.

2.2.4 Konfigurationsentropie

Unter physiologischen Bedingungen ist der native Zustand ein Ensemble aus Strukturen, die untereinander einen RMSD-Wert von bis zu 2Å aufweisen [IKP88]. Diese Fluktuationen sind zum einen kleine Schwingungen um die Gleichgewichtslage und zum anderen Übergänge zwischen mehreren energetisch gleichwertigen Minima, die im Phasenraum eng beieinanderliegen. Bei diesen benachbarten Minima handelt es sich zumeist um Strukturen, deren Seitenkettenanordnungen an der Proteinoberfläche unterschiedlich ist. Die Sekundärstruktur, also die Hauptkettenkonfiguration, bleibt dabei, ebenso wie die Seitenkettenanordnung im Inneren, nahezu unangetastet [SSRD91]. Da Proteine eine sehr hohe Packungsdichte haben bezeichnet man den Proteinkern gelegentlich als *Festkörper-ähnlich* und die Oberfläche als *Liquid-ähnlich* (kurz: surface-molten-solids) [ZVK99].

In harmonischer Näherung kann das Protein durch ein System (mehrdimensionaler) harmonischer Oszillatoren beschrieben werden. Nach dem Gleichverteilungssatz der klassischen Thermodynamik liefert jeder quadratische im Hamiltonian auftretende Term zur inneren Energie den Beitag $\frac{1}{2}\beta^{-1}$, unabhängig von

der Frequenz des Oszillators. Mit den Boltzmannfaktoren ρ_I der Eigenmoden unterschiedlicher Konfigurationen I ergibt sich für die *Konfigurationsentropie* der Ausdruck

$$S = \sum_I \rho_I S_I^v - k_B \sum_I \rho_I \ln \rho_I \ , \tag{2.18}$$

der aus einem Anteil der Vibrationsentropie S_I^v der einzelnen Moden und der "Mischungsentropie" der verschiedenen Moden besteht [KIP87].

Gemäß einer Analyse von Karplus, Ichiye und Pettitt [KIP87] für Bovine Pancreatic Trypsin (BTPI) und Lysozyme ist die verbreitete Annahme, daß die Vibrationsentropie S_I^v für alle Moden als auch für unterschiedliche Proteinkonfigurationen gleich wäre, nur mangelhaft erfüllt. Ebensowenig läßt sich S_I^v zu Null approximieren. Dennoch ist die Summe der Vibrationsentropien $S_{vibr} = \sum_I \rho_I S_I^v$ nahezu unabhängig von der Konfiguration. Für BTPI und Lysozyme beläuft sich die mittlere Vibrationsentropie praktisch jeder Konfiguration auf ca. $34\,cal$ pro Mol, Residuum und Kelvin. Dies gilt für die native Struktur ebenso wie für eine einzelne random-coil Struktur. Für andere Proteine kann man ein ähnliches Verhalten erwarten, wobei die Schwankungen um den Entropiemittelwert mit wachsender Proteingröße sehr wahrscheinlich abnehmen. Bei Proteinfragmenten von weniger als etwa 15 oder 20 Aminosäuren könnte allerdings ein Beitrag der Vibrationsentropie vorhanden sein, der einen signifikanten Einfluß auf die Stabilität der nativen Struktur hat.

Wenn die mittlere Vibrationsentropie zweier Proteinkonfigurationen gleich ist, so fällt dieser Beitrag bei der Differenzbildung heraus und kann, wie allgemein üblich, vernachlässigt werden. Der für die Proteinstrukturvorhersage interessante Teil der **Konfigurationsentropie** ist dann durch

$$S_{konf} = -k_B \sum_I \rho_I \ln \rho_I \tag{2.19}$$

geben. Dabei können die unterschiedlichen Proteinbestandteile unterschiedlich in diese Summe eingehen. Das Proteininnere und die Hauptkette sind in kompakten niederenergetischen Konfigurationen verhältnismäßig starr und tragen kaum zu dieser Summe bei. Die Seitengruppen an der Proteinoberfläche können verschiedene Konfigurationen annehmen, da die Wassermoleküle sich der Proteinstruktur anpassen. Sie sind es, die die Konfigurationsentropie maßgeblich bestimmen [KIP87]. Der genauere Zusammenhang zwischen der Konfigurationsentropie und der Position der Aminosäuren wird uns im Abschnitt zur Lösungsmittelenergie beschäftigen.

2.2.5 Die Freie Energie als Funktion der Proteinstruktur

Obwohl wir wiederholt von "der nativen Struktur" oder "dem random-coil Zustand" sprachen, haben wir diesen Strukturen implizit immer ein Ensemble zuge-

ordnet. Diesen Gedanken wollen wir weiterverfolgen und, indem wir dem Ensemble eine Freie Energie zuordnen, letztendlich einer einzelnen Struktur eine Freie Energie zuweisen.

Durch die Verteilungsfunktion ρ, auch Wahrscheinlichkeitsdichte genannt, wird jedem Zustand eines Ensembles ein statistisches Gewicht zugeordnet. Bei der Faltung eines Proteins ausgehend von einem random-coil Zustand, geht die zugehörige Verteilung nach und nach in die Verteilung des nativen Zustandes über. Die statistische Physik lehrt uns, daß die native Verteilung durch

$$\rho^{nat}(\vec{q}) = \frac{e^{-\beta H(\vec{q})}}{\int e^{-\beta H(\vec{y})}\, d\vec{y}} \tag{2.20}$$

gegeben ist. Dieser Ausdruck gibt die Lage des Minimums der Freien Energie F als Funktion der Verteilung wieder:

$$F(\rho) = \int \Big[\rho(\vec{q})H(\vec{q}) + \beta^{-1}\rho(\vec{q})\ln\rho(\vec{q})\Big] d\vec{q} = U - TS \tag{2.21}$$

Wenn wir von der nativen Struktur stellvertretend für ein Ensemble sprechen, dann bedeutet dies, daß die Verteilung ρ in der Nähe dieser Struktur konzentriert ist. Für die Simulation wählen wir als Vertreter dieses Ensembles das Maximum der Verteilung bzw. das zugehörige lokale Minimum von H. Aus dem vorherigen Abschnitt wissen wir um die Beiträge lokaler Fluktuationen zur inneren Energie und der Vibrationsentropie. Somit können wir der nativen Struktur eine Energie $E^{nat} = H(\vec{q}^{nat}) = H_p(\vec{x}^{nat}) + \widehat{H}_w(\vec{q}^{nat})$, einen vibrationsbedingten Beitrag zur inneren Energie ΔU und zur Entropie S_{vibr} zuordnen. Wenn jetzt noch bekannt wäre, zwischen wievielen verschiedenen Konfigurationen das native Ensemble wechseln kann, so wäre uns auch der Wert der Konfigurationsentropie bekannt und insgesamt gilt

$$F = E^{nat} + \Delta U - TS_{vibr} - TS^{nat}_{konf} \tag{2.22}$$

Von den Beiträge ΔU und S_{vibr} wissen wir, daß sie für alle Strukturen gleich sind. Damit ergibt sich die **Differenz der Freien Energien** zweier Strukturen zu

$$\Delta F = \Delta E - T\Delta S_{konf}\ . \tag{2.23}$$

Wir können diese Gleichung noch etwas umschreiben, denn es ist $\Delta E = E^I - E^J = H(\vec{q}^I) - H(\vec{q}^J)$ mit $H(\vec{q}) = H_p(\vec{q}) + \widehat{H}_w(\vec{q})$ und da wir den effektiven Lösungsmittelhamiltonian als Freie Energie des Wassers verstehen können schreiben wir $\Delta E = [H_p(\vec{q}^I) - H_p(\vec{q}^J)] + [\widehat{H}_w(\vec{q}^I) - \widehat{H}_w(\vec{q}^J)] = \Delta H_p + \Delta F_w$. Setzen wir dies in obige Gleichung ein, erhalten wir

$$\Delta F = \Delta H_p + \Delta F_w - T\Delta S_{konf}\ . \tag{2.24}$$

Wenn wir desweiteren annehmen, daß sich die innere Energie des Wassers sich nicht ändert, kann man dies auch in der Form

$$\Delta F = \Delta H_p - T_w \Delta S_w - T \Delta S_{konf} \tag{2.25}$$

schreiben, wie am Anfang dieses Abschnittes behauptet.

Im Ausdruck $\Delta H_p + \Delta F_w$ werden zwei Proteinkonfigurationen miteinander in Beziehung gesetzt. Die Konfigurationsentropie hingegen vergleicht die Umgebungen der beiden Proteinkonfigurationen und ergibt sich aus der Anzahl der verschiedenen Minima um die jeweils betrachtete Struktur. In kompakten niederenergetischen Strukturen ist der Proteinkern dicht gepackt und dort finden keine strukturellen Veränderungen statt. S_{konf} wird von den Anteilen des Proteins bestimmt, die sich an der Oberfläche aufhalten, also in Kontakt mit dem Wasser sind. All diesen Strukturen ist gemein, daß im Kern hydrophobe Seitenketten gruppiert sind und alle hydrophilen Seitenketten in Kontakt mit dem Wasser stehen. Daher kann man in einfachster Näherung annehmen, daß für niederenergetische Strukturen die Konfigurationsentropie gleich ist und aus der Differenz der Freien Energie herausfällt[11]. Damit ergibt sich der Unterschied in der Freien Energie allein aus den beiden Strukturen I und J, ohne explizite Angabe ihrer Umgebung

$$\Delta F = [H_p(\vec{q}^I) + \widehat{H}_w(\vec{q}^I)] - [H_p(\vec{q}^J) + \widehat{H}_w(\vec{q}^J)] \; . \tag{2.26}$$

[11] Wir werden später sehen, daß unser Ansatz für den effektiven Lösungsmittelhamiltonian eine gewisse Approximation der Konfigurationsentropie enthält. Weshalb wir dann auf obige Näherung nicht mehr angewiesen sind.

Kapitel 3

Theoretische Modellierung

Der thermodynamischen Hypothese von Anfinsen folgend ist die native Struktur eines Proteins das globale Minimum der Freien Energie. Infolgedessen läßt sich die native Struktur eines Proteins vorhersagen, wenn eine hinreichend genaue Approximation der Freie Energie des Proteins in seiner physiologischen Umgebung bekannt ist und eine Methode vorliegt, die das globale Minimum der Freien Energieoberfläche zuverlässig identifiziert.

In diesem Kapitel sollen daher Verfahren zur Bestimmung des globalen Minimums vorgestellt werden und anschließend die Wechselwirkungen besprochen werden, die für ein Protein in wässeriger Lösung von Bedeutung sind.

3.1 Stochastische Optimierungsverfahren

3.1.1 Einleitung

Gehen wir davon aus, daß der Hamiltonian oder die Freie Energie als Funktion der Proteinkonfiguration bekannt ist, so verbleibt das Optimierungsproblem, das globale Minimum der Freien Energie zu bestimmen. Die Hauptschwierigkeit der Optimierungsverfahren liegt im exponentiellen Wachstum des Konfigurationsraumes mit der Systemgröße. Eine Formulierung dieses Problem im Kontext der Proteinfaltung ist das Levinthalsche Paradoxon (Seite 12), demzufolge es unmöglich ist, die native Proteinstruktur durch eine rein zufällige Suche im Konfigurationsraum aufzuspüren.

Da Proteine ihre native Struktur relativ schnell einnehmen, erfolgt die Optimierung in der Natur nicht rein zufällig und das Paradoxon resultiert aus der implizierten Annahme, daß die Potentialenergieoberfläche flach sei. In der Realität spielt die Topologie der Potentialoberfläche eine entscheidende Rolle bei der Konvergenz zum globalen Minimum. Es wird angenommen, daß die Potentialenergieoberfläche eines Proteins eine ausgeprägte Tricherstruktur (engl. *funnel*) besitzt, der die Proteinstruktur im Prinzip an jedem Punkt im Konfigurations-

raum in Richtung native Struktur weist.

Eine Strategie zur Identifizierung des Minimums liegt daher – trotz der Systemgröße – in der Lösung der Bewegungsgleichungen, die sich aus dem Hamiltonian ergeben. In einer solchen Molekulardynamik Simulation (Seite 44) läßt sich der vollständige Faltungsprozeß deterministisch nachverfolgen. Dieses Verfahren ist in seiner Natur seriell, das heißt nicht nur, daß zwei Simulationen von den gleichen Startkonfigurationen aus identisch verlaufen, sondern auch, daß Simulationen von unterschiedlichen Startkonfigurationen unabhängig voneinander sind. Da jede Simulation für sich das globale Minimum finden muß und dies nach einer deterministisch festgelegten Zeit erfolgt, sind Simulationen von verschiedenen Startkonfigurationen aus unnötig.

Die Ergebnisse der bisherigen Simulationen dieser Art zeigen, daß die deterministischen Optimierungsverfahren nicht effizient genug arbeiten, um Proteinstrukturvorhersage in größerem Umfang zu betreiben. Daher werden vermehrt stochastische Verfahren verwendet, die schon auf anderen Themengebieten erfolgreich eingesetzt werden. Sie sind Gegenstand der folgenden Abschnitte. Bei einer stochastische Simulation wird eine nicht-deterministische Abfolge von Konfigurationsraumpunkten konstruiert, deren Energie zumeist abfällt, und so zum Minimum führt, aber gelegentlich bergauf über Energiebarrieren hinwegläuft.

Es werden nun zwei stochastische Verfahren vorgestellt, namens Monte-Carlo (MC) und Simulated Annealing (SA). Erstes dient zur Bestimmung thermodynamischer Erwartungswerte bzw. zur Konstruktion von Zuständen mit einer bestimmten Verteilung. Letzteres lokalisiert Minima auf der Potentialenergieoberfläche. Anschließend wird die Beziehung zwischen Molekulardynamik Simulationen und Monte-Carlo basierenden Verfahren beleuchtet und die Methode der Basin-Hopping-Technique vorgestellt, die es ermöglichte die native Struktur des Proteins 1F4I vorherzusagen. Die von uns eingesetzten Optimierungsverfahren arbeiten ohne Gradienten.

3.1.2 Monte-Carlo Simulation (MC)

Ausgehend von Gleichung 2.20 besteht die Aufgabe der Proteinstrukturvorhersage in der Bestimmung der Verteilung ρ^{nat} aus der Proteinsequenz. Hierzu kann man sich des Monte-Carlo-Verfahrens bedienen [KGV83].

Dieses ist in der Lage, eine Approximation $\tilde{\rho}$ der Wahrscheinlichkeitsdichte

$$\rho(\vec{q}) = \frac{exp(-\beta H(\vec{q}))}{\int exp(-\beta H(\vec{q}))\ d\vec{q}} = \frac{exp(-\beta H(\vec{q}))}{\mathcal{Z}}, \tag{3.1}$$

zu bestimmen, ohne $\tilde{\rho}$ explizit anzugeben. Mit Hilfe der Verteilung $\tilde{\rho}$ können thermodynamische Erwartungswerte einer Observablen A gemäß

$$\langle A \rangle = \int A(\vec{q})\rho(\vec{q})\ d\vec{q} \approx \int A(\vec{q})\tilde{\rho}(\vec{q})\ d\vec{q} \tag{3.2}$$

berechnet werden. Die Freie Energie berechnet sich dann z.B. gemäß Gleichung 3.27 aus

$$F = +\beta^{-1} \ln\langle \exp(+\beta H)\rangle \; .$$

Die Approximation gelingt, indem eine Sequenz von Konfigurationen $\{\vec{q}\}$ erzeugt wird, in der (im Limes unendlich langer Sequenz) zwei Konfigurationen $\vec{q}_1, \vec{q}_2$ mit der relativen Häufigkeit

$$\frac{\rho(\vec{q}_1)}{\rho(\vec{q}_2)} = exp(-\beta\,[H(\vec{q}_1) - H(\vec{q}_2)]) \tag{3.3}$$

auftreten.

Die Konstruktionsvorschrift dieser Konfigurationenfolge $\{\vec{q}\}$ trägt den Namen *Metropolis Algorithmus* und ist relativ simpel. Von einem Zustand $\vec{q}_i$ mit der Energie E_i wird durch eine (kleine) Änderung ein Zustand $\vec{q}_j$ mit Energie E_j erzeugt. Liegt die Energie E_j unterhalb von E_i, so dient $\vec{q}_j$ aus Ausgangspunkt für die nächste Iteration und wird an die Folge $\{\vec{q}\}$ angehängt. Für $E_j > E_i$ geschieht dies mit einer Wahrscheinlichkeit von

$$A_{ij} = exp[-\beta(E_j - E_i)] \, . \tag{3.4}$$

Wird die vorgeschlagene Konfiguration q_j verworfen, so wird ein weiteres Mal q_i in die Sequenzliste aufgenommen. Die Wahrscheinlichkeit, $\vec{q}_j$ zu akzeptieren, läßt sich auch kombiniert in der Form

$$A_{ij} = min\,\{1, exp[-\beta(E_j - E_i)]\} \tag{3.5}$$

angeben. Dieses Akzeptanzwahrscheinlichkeit heißt *Metropolis Kriterium*.

Erwartungswerte einer Observablen A ergeben sich als Mittelwert $\langle A\rangle = (1/|\{\vec{q}\}|)\sum_{\{\vec{q}\}} A(\vec{q})$ entlang der Konfigurationssequenz $\{\vec{q}\}$. Ist im Voraus bekannt, welche Erwartungswerte berechnet werden soll, so kann man auf die Speicherung der Konfigurationen verzichten und stattdessen die Werte $A(\vec{q})$ aufsummieren und am Ende der Simulation durch die Anzahl der Iterationsschritte teilen.

An die Konstruktion dieser Konfigurationenfolge, genauer gesagt an die Generierung einer neuen Konfiguration, sind gewissen Bedingungen geknüpft, auf die hier kurz eingegangen werden soll. Dabei streifen wir die Theorie der *Markov Ketten*, welche sich mit statistisch unabhängigen Übergängen beschäftigt. Diese Unabhängigkeit wird oft mit den Worten *"the future depends on the past only through the present"* definiert und führt uns direkt zu der ersten Bedingung an die Generierung neuer Konfigurationen: sie müssen statistisch unabhängig sein. Das Monte-Carlo Verfahren ist damit prinzipiell unfähig, aus vergangenen Generierungsversuchen zu "lernen", um die Proteinstruktur gezielt zu optimieren.

Betrachten wir den Fortlauf einer Monte-Carlo Simulation etwas näher. Befindet sich das System zu einem beliebigen Iterationsschritt im Zustand $\vec{q}_i$ und

sei G_{ij} die Wahrscheinlichkeit von Zustand $\vec{q}_i$ die Änderung in den Zustand $\vec{q}_j$ zu generieren, so ist die Übergangswahrscheinlichkeit P_{ij} von $\vec{q}_i$ zu $\vec{q}_j$

$$P_{ij} = \begin{cases} G_{ij}A_{ij} & \text{für } i \neq j \\ 1 - \sum_{j' \neq i} P_{j'i} & \text{für } i = j \end{cases} . \tag{3.6}$$

Dies gilt für jede einzelne Iteration. Bei zwei aufeinanderfolgenden Iterationen ergibt sich für die Übergangswahrscheinlichkeit $P_{ij}^{(2)}$ und danach entsprechend für $P_{ij}^{(3)}$, usw. zu

$$P_{ij}^{(2)} = \sum_k P_{ik}P_{kj} , \quad P_{ij}^{(3)} = \sum_{k,l} P_{ik}P_{kl}P_{lj} , \quad \ldots \tag{3.7}$$

Die in der statistischen Physik wichtige Bedingung der *Ergodizität* wird auch an $P_{ij}^{(n)}$ gestellt; d.h. für beliebige i, j muß es ein $n > 0$ mit $P_{ij}^{(n)} > 0$ geben. Dies ist die zweite Bedingung an das Monte-Carlo Verfahren, und ist so zu lesen, daß jede beliebige Konfiguration i von jeder anderen Konfiguration j aus zu erreichen ist. Die dritte Anforderung trägt die Bezeichnung *detailed balance* und ist etwa durch $G_{ij} = G_{ji}$ erfüllt. Sie ist so zu verstehen, daß der Wechsel zwischen zwei Konfigurationen in beide Richtungen (prinzipiell) gleich wahrscheinlich ist. Berücksichtigt man, daß die Zustände $\vec{q}_i$ und $\vec{q}_j$ mit unterschiedlicher Wahrscheinlichkeit vorliegen, so schreibt sich die *detailed balance* Bedingung für eine Wahrscheinlichkeitsverteilung $\tilde{\rho}(\vec{q})$ wie folgt:

$$P_{ij}\ \tilde{\rho}(\vec{q}_i) = P_{ji}\ \tilde{\rho}(\vec{q}_j) \tag{3.8}$$

Eine solche Verteilung ist stationär ("invariant unter P"), d.h.

$$\sum_i \tilde{\rho}(\vec{q}_i)P_{ij} = \tilde{\rho}(\vec{q}_j) \tag{3.9}$$

(wie sich durch Einsetzten obiger Gleichung zeigen läßt), und weiterhin gilt

$$\tilde{\rho}(\vec{q}_j) = \lim_{n\to\infty} P_{ij}^{(n)} \tag{3.10}$$

unabhängig von i. Für $G_{ij} = G_{ji}$ erfüllt sich die detailed balance Bedingung, und es gilt

$$\tilde{\rho}(\vec{q}_j) \to \frac{exp(-\beta E_j)}{\int exp(-\beta E(\vec{q}))d\vec{q}} = \rho(\vec{q}_j). \tag{3.11}$$

Allgemein gilt: für normiertes f und $A_{ij} = \min\{1, f(i)/f(j)\}$ ergibt sich unter den Voraussetzungen *a*) P_{ij} ist ergodisch und *b*) f erfüllt die detailed balance Bedingung: $\lim_{n\to\infty} P_{ij}^{(n)} = f(j)$. Der detailierte Beweis ist in praktisch jedem Vorlesungsskript oder Buch zum Thema Markov-Ketten zu finden, siehe etwa [Chu67]

3.1.3 Simulated Annealing (SA)

Simulated Annealing (SA) ist einer der populärsten Algorithmen im Bereich der stochastischen Optimierung. Der Begriff Annealing beschreibt einen Prozeß, bei dem es durch Erhitzen und anschließendes langsames Abkühlen möglich ist, das System aus einem lokalen Minimum in seinen Grundzustand zu überführen und so z.B. Fehlstellen in einem Festkörper zu beseitigen. Ein schnelles Abkühlen wird auch als *Quenchen* bezeichnet, wobei das System zumeist in einem metastabilen Zustand eingefroren wird.

Der Prozeß des Annealing kann in den Metropolis Algorithmus integriert werden, indem dessen Temperatur langsam erniedrigt wird. Dabei hängt das Konvergenzverhalten stark von der Kühlungsstrategie $\{\beta_n\}_{n\in\mathbb{N}}$ ab. Die Temperaturadjustierung muß so langsam erfolgen, daß das System bei jeder Temperatur equilibriert ist (adiabatische Zustandsänderung), d.h. bis asymptotisch die kanonische Verteilung erreicht wurde. Für diese Verteilung gilt

$$\lim_{\beta\to\infty} \tilde{\rho}(\vec{q}) = \lim_{\beta\to\infty} \frac{exp(-\beta E(\vec{q}))}{\sum_i exp(-\beta E_i)} \tag{3.12}$$

$$= \lim_{\beta\to\infty} \frac{exp(-\beta(E(\vec{q}) - E^\star))}{\sum_i exp(-\beta(E_i - E^\star))} \tag{3.13}$$

mit der Energie des globalen Minimums $E^\star = \min\{E_i\}$. Die Exponenten sind somit entweder Null oder negativ, und für $\beta \to \infty$ folgt $\tilde{\rho}(\vec{q}) \neq 0$ nur für $E(\vec{q}) = E^\star$.

Nach der Theorie der Markov-Ketten existiert ein ϵ, welches die Konvergenz zum Grundzustand für alle Abkühlstrategien $\{\beta_n\}_{n\in\mathbb{N}}$ garantiert, sofern diese folgende Bedingung erfüllen:

$$\beta_n \leq \epsilon \log n \ , n \in \mathbb{N} \ .$$

Ein Abkühlschema proportional $\log n$ wäre jedoch viel zu zeitaufwendig, so daß man schnellere Abkühlraten verwenden und auf mathematische Zusage der Konvergenz verzichten muß. Da wir darauf angewiesen sind, in kurzer Zeit viele unterschiedliche niederenergetische Strukturen zu erzeugen, kühlen wir das System sehr schnell ab. Je nach gewünschter Laufzeit einer Simulation wählen wir ein $\gamma \in \mathbb{R}$ und setzen

$$\beta_n = \beta_0 \cdot \gamma^n \ , n \in \{1, 2, \ldots, n_{max}\} \tag{3.14}$$

Im Vergleich zu $\beta \sim \log n$ zeigt sich, daß wir mehr Rechenzeit im hohen und mittleren Temperaturbereich verbringen. Daher können unsere Simulationen größere Entfernungen auf der Potentialenergieoberfläche zurücklegen. Problematisch ist jedoch stets die abschließende lokale Optimierung bei niedrigen Temperaturen.

3.1.4 Strukturgenerierung

Das Monte-Carlo Verfahren gibt uns die Möglichkeit, die Verteilungsfunktion ρ näherungsweise zu bestimmen. Der wesentliche Unterschied zur Molekulardynamik, die über ihre Trajektorien auch Zugang zu thermodynamischen Erwartungswerten hat, ist darin zu sehen, daß es im Metropolis Algorithmus den Begriff Zeit nicht gibt. Die nächste Iteration einer Monte-Carlo Simulation kann zu einer Konfiguration führen, die in der Molekulardynamik erst nach vielen Zeitschritten erreicht werden könnte, und so sind mittels Monte-Carlo große Sprünge auf der Potentialenergieoberfläche möglich (globale Veränderungen). Ein Umstand, der zur schnellen Relaxation der Verteilungsapproximation $\tilde{\rho}$ beiträgt. Die Schritte einer Molekulardynamik Simulation sind im Gegensatz zur Monte Carlo Simulation gerichtet, sie erfolgen stets in Richtung des Gradienten der Potentialenergieoberfläche. In manchen Fällen kann es günstiger sein, dem Gradienten zu folgen, als zufällige Sprünge im Konfigurationsraum durchzuführen. Bei sehr zackigen Energieoberflächen, wie sie bei einem Protein vorliegt, ist dies jedoch unwahrscheinlich.

Wichtig für das Konvergenzverhalten ist auch die Strategie, nach der neue Konfigurationen generiert werden. In unseren Simulationen entstehen neue Konfigurationen durch Drehungen um die Dihedralwinkel; sie sind in drei *Rotationskategorien* eingeteilt. Allen dreien ist gemein, daß nur Veränderungen an einer einzelnen Aminosäure vorgenommen werden und dort auch entweder nur an der Haupt- oder an der Seitenkette. Bei der ersten Kategorie handelt es sich ursprünglich um zufällige Winkelveränderungen von maximal $\pm 5°$. Bei tiefen Temperaturen muß der Winkelbereich jedoch weiter eingeschränkt werden, um die Akzeptanzrate der Simulation nicht zu sehr abfallen zu lassen. Bei Rotationen der zweiten Kategorie werden die Dihedralwinkel auf zufällige Werte zwischen $-180°$ und $180°$ gesetzt und so größere Sprünge auf der Potentialenergieoberfläche möglich. Desweiteren existiert eine Bibliothek von Dihedralwinkeln, die strukturaufgeklärten Proteinen entnommen wurden [AJ95], und in Kategorie drei dazu verwandt wird, Konfigurationen mit diesen Winkeln zu erzeugen.

Rotationen der dritten Kategorie verletzen die detailed balance Bedingung und werden nicht zur Bestimmung thermodynamischer Erwartungswerte eingesetzt. Für die SA Simulation sind diese Rotationen jedoch sehr hilfreich, da sie den Algorithmus mehr in "natürlichen" Bereichen suchen lassen und so das Konvergenzverhalten positiv beeinflussen. Eine genauere Analyse solcher Winkelbibliotheken findet sich in einer Arbeit von Abagyan und Totrov [AT94].

3.1.5 Der Temperaturbegriff in der Simulation

In unserem Kraftfeld ist das Lösungsmittel implizit enthalten, das heißt, es existiert ein effektiver Hamiltonian, der alle Wechselwirkungen mit und innerhalb des Wassers beschreibt. Dieser ist jedoch nur von der Struktur des Proteins

abhängig (Gleichung 2.16). Die zugehörigen effektiven Lösungsmittelparameter sind an Transferenergien angepaßt, bei denen das Protein aus einer Oktanolumgebung in wässerige Lösung transferiert wurde. Das zugrundeliegende Experiment ist bei einer konstanten Temperatur von $300K$ durchgeführt worden. Folglich beschreibt unserer Lösungsmittelhamiltonian die Wechselwirkung des Proteins mit dem Wasser nur richtig bei einer Temperatur von $300K$. Das Monte-Carlo Verfahren zur Bestimmung thermodynamischer Erwartungswerte ist an diese Temperatur gebunden.

Gemäß Gleichung 2.26 kann die Differenz der Freien Energie für zwei Strukturen ausgerechnet werden. Hierbei handelt es sich stets um die Differenz bei $300K$ unabhängig von der Temperatur der Simulation. Das heißt, während eines SA Laufs bewegen wir uns auf der Potentialoberfläche der Freien Energie bei $300K$. Der Temperatur der Simulation kommt daher keine eigene physikalische Bedeutung zu.

Es gibt allerdings noch eine andere Interpretationsmöglichkeit, und zwar kann man die Temperatur des Proteins und die des Lösungsmittels trennen. Die Temperatur des Wassers wird nun auf $300K$ festgelegt. Die Temperatur der Simulation ist dann die Temperatur des Proteins. Der Unterschied im Formalismus besteht darin, ob man entropische Beiträge mit simuliert und diese dann für $T \to 0$ verschwinden, oder ob man sie von vornherein aus der Differenz der Freie Energie herausnimmt.

3.1.6 Schwellwert Kriterium (TA)

Wenn man nicht an thermodynamischen Erwartungswerten interessiert ist, so ist man nicht an das Metropolis-Kriterium gebunden und kann stattdessen aus einer Vielzahl anderer Kriterien auswählen. Eines der einfachsten wird als *threshold acceptance* (TA) bezeichnet [SM98].

$$A_{ij} = \begin{cases} 1 & \text{für } E_j - E_i < \beta^{-1} \\ 0 & \text{sonst} \end{cases} \tag{3.15}$$

Die Optimierung geschieht ähnlich wie bei SA, indem man bei einer hohen Temperatur beginnend neue Konfigurationen gemäß 3.15 akzeptiert. Sukzessive wird die Temperatur langsam abgesenkt, bis man eine vorgegebene Temperatur erreicht oder eine vorgegebene Anzahl von Konfigurationsvorschlägen generiert hat. Üblicherweise wird die Temperatur analog zu SA gemäß 3.14 abgesenkt.

Eigenständige Simulationen sind von uns mit diesem Kriterium nicht durchgeführt worden. Es ist vielmehr Bestandteil der folgenden beiden Optimierungsstrategien.

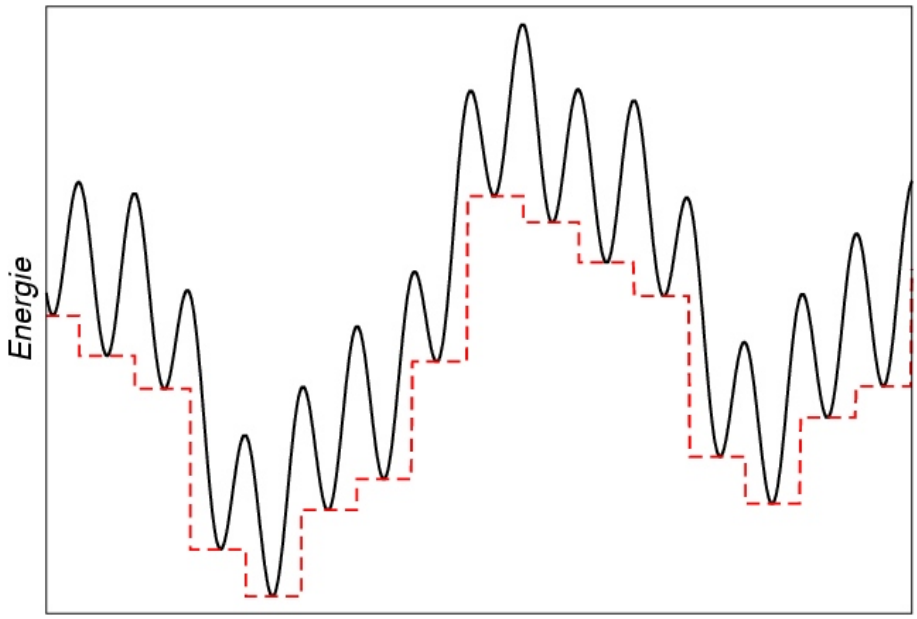


Abbildung 3.1: Schematische Darstellung einer Potentialenergieoberfläche. Die durchgezogene Linie ist das Originalpotential, während die gestrichelte Linie das effektive Potential der Minima wiedergibt.

3.1.7 Basin-Hopping-Technique (BHT)

Dieses Optimierungsverfahren ist die einfache Variante der folgenden EMC-Simulation (evolutionäres Monte-Carlo), welche wir vor diesem Algorithmus entwickelt und verwendet haben. Aus diesem stammt auch die Kombination von SA und TA. Obwohl EMC und damit auch diese Simulationstechnik an eine Arbeit von Schneider und Morgenstern [SM98] angelehnt sind, wurden die hier verwendeten Algorithmen mehrfach überarbeitet und von einigen Parametern befreit.Die Intention dieses Verfahrens war ursprünglich nicht die globale Optimierung, sondern es wurde entworfen, um möglichst viele unterschiedliche niederenergetische Strukturen zu erzeugen. Der Name *Basin-Hopping Technique* steht in Bezug auf eine Arbeit von Wales für Lennard-Jones Cluster [WD97]. Die Potentialenergieoberfläche für Lennard-Jones-Cluster ist ähnlich zerklüftet wie die eines Proteins. Wales Idee bestand darin, ausgeprägte lokale Minima in kleinen Bereichen der Potentialenergieoberfläche zu identifizieren und ein Monte-Carlo Simulation auf diesen Minima durchzuführen (Abbildung 3.1).

Das Grundgerüst des BHT-Algorithmus bilden aufeinanderfolgende SA Simulationen. Jede dieser Simulationen beginnt bei einer (Simulations- oder Protein-) Temperatur von 400 bis $600K$ und kühlt das System auf etwa $1K$ ab. Diese hohe Temperatur stellt sicher, daß es signifikante strukturelle Änderungen des Proteins innerhalb der Simulationszeit geben kann. Auf der anderen Seite ist $1K$ niedrig genug, um ein lokales Minimum zu identifizieren. Wenn wir auf die lokale Mini-

mierung besonderen Wert legen, so verringern wir diese Temperatur noch weiter auf 0.5 oder $0.2K$.

Als Startkonfiguration für die nächste Iteration bzw. SA Simulation wird die Endkonfiguration der aktuellen Iteration genommen, solange deren Energie nach dem TA Kriterium nicht als "zu schlecht" eingestuft wird, d.h. solange die neue Struktur eine ähnlich oder bessere Energie erreicht. In den Simulationen hat sich eine Schranke von $\Delta E = 3\,kcal\,mol^{-1}$ als sinnvoll erwiesen.

Die Laufzeit der einzelnen SA Läufe steigt mit fortlaufender Iterationszahl an, beginnend mit $n_1 = 10^5$ Schritten. Da die Gesamtzeit des Verfahrens nicht allzusehr mit der Anzahl der Iterationen anwachsen soll, führen wir die i-te Iteration mit $n_i = \sqrt{i} \cdot n_1$ SA-Schritten durch. Gleichzeitig ist damit gewährleistet, daß wir in den ersten Iterationen nicht zuviel Rechenzeit in die lokale Optimierung investieren, sondern von energetisch schlechten Startstrukturen weg in Bereiche kommen, die energetisch niedriger liegen. In diesen niederenergetischen Bereichen ist es notwendig, sowohl länger lokal zu optimieren, als auch zwischenzeitlich länger aufzuheizen, um große Barrieren zu überwinden.

Um einen ersten Eindruck zur Arbeitsweise dieses Verfahrens zu erlangen sei auf Abb. 3.2 verwiesen. Hier ist die Energie als Funktion der Schrittzahl am Beispiel einer 1F4I Simulation angegeben. Um die Stärke der Fluktuationen zu verringern, ist die Energie nur nach jeweils 5000 Schritten wiedergegeben. Hierdurch erkennt man nicht, daß – wenn nach einem SA Lauf eine schlechte Struktur gefunden wurde (z.B. nach 5.2×10^6 Schritten) – nicht von dieser Struktur sondern von der vorherigen aus weitersimuliert wird. Die Simulationsbedingungen in diesem Beispiel waren von den aktuell benutzen verschieden. Die Schrittzahl n_0 lag bei $70,000$, und es wurde nur bis $5K$ abgekühlt. Dennoch ist dieses Beispiel hier aufgenommen worden, um auf einige Probleme hinzuweisen. Obwohl wir in der Heizphase Strukturen vorfinden, deren Energie um mehr als $60\,kcal\,mol^{-1}$ oberhalb des Minimums liegen, ist die RMSB Abweichung zweier aufeinanderfolgender Minima im Mittel nur 1.1Å und maximal 2.6Å. Desweiteren ist zu erkennen, daß die lokale Minimierung jeder Iteration wahrscheinlich noch nicht abgeschlossen ist, denn bei vollendeter lokaler Minimierung erwarten wir einen Abschnitt, in dem sich die Energie nicht mehr verändert.

Hieraus können wir zwei Schlüsse ziehen. Ersteinmal benötigen strukturelle Änderungen auch bei sehr hohen Temperaturen eine gewissen Zeit, denn obwohl die Energie instantan der Temperaturerhöhung folgt, ist damit nicht unbedingt eine Strukturänderung verbunden. Zweitens ist bei einer Abkühlstrategie $\beta_n \sim \gamma^n$ eine Endtemperatur von $5K$ noch zu hoch, um eine gute lokale Optimierung zu erhalten.

3.1.8 Eine evolutionäre Erweiterung (EMC)

Evolutionäre Algorithmen (EA) bezeichnen eine Klasse stochastischer Optimierungsverfahren, deren bekanntester Vertreter, der Genetische Algorithmus (GA),

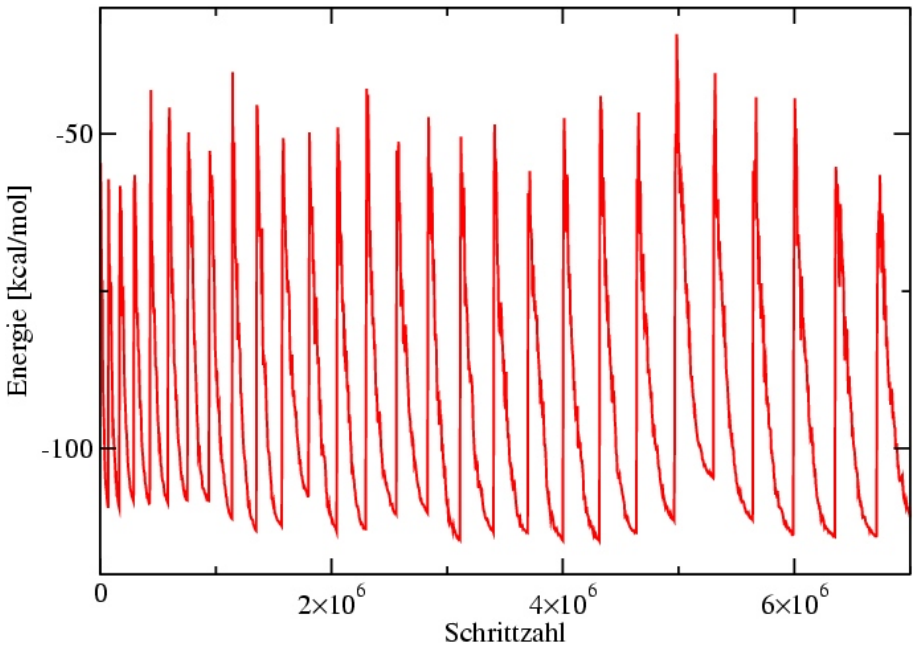


Abbildung 3.2: Energieverlauf einer BHT Simulation

gelegentlich für das Proteinstrukturproblem eingesetzt wird. Wie Kay Hamacher und Wolfgang Wenzel gezeigt haben, skaliert der numerische Aufwand jedoch exponentiell mit der Systemgröße [HW99], weshalb die Anwendung des GA auf Probleme mit verhältnismäßig wenigen Freiheitsgraden beschränkt werden sollte. Auf die Strukturoptimierung bezogen bedeutet dies, man sollte die direkte Nutzung des genetische Algorithmus nur bei Peptiden mit weniger als 20 Aminosäuren in Erwägung ziehen. Als übergeordnete Strategie unterliegen jedoch nicht notwendigerweise alle evolutionären Algorithmen diesem Problem.

Ohne das Thema evolutionärer Algorithmen weiter vertiefen zu wollen, wenden wir uns direkt dem Verfahren zu, das wir als evolutionäres Monte-Carlo (EMC) bezeichnen (Abb. 3.3). Gegeben sei eine Population $P(0)$ von N Konfigurationen $I \in P(0)$ mit Energien E_I. M dieser Konfigurationen I unterwerfen wir einer SA Simulation der Länge n_0, an deren Ende wir neue Strukturen I' mit Energien $E_{I'}$ erhalten. Aus dem Satz $P'(0)$ der $M + N$ neuen und alten Strukturen wählen wir N Kandidaten aus und fassen sie zu einer neuen Population $P(1)$ zusammen. Von dieser Population durchleben wieder M Mitglieder einen SA Lauf und so weiter.

Die M in jeder Iteration ausgewählten Strukturen können parallel auf verschiedenen Rechnern simuliert werden. Bislang haben wir das EMC Verfahren nur mit etwa $N = 20$ Konfigurationen pro Populationen durchgeführt und davon $M = 4 - 6$ parallel in SA Simulationen geschickt. Die Ergebnisse sind durchaus

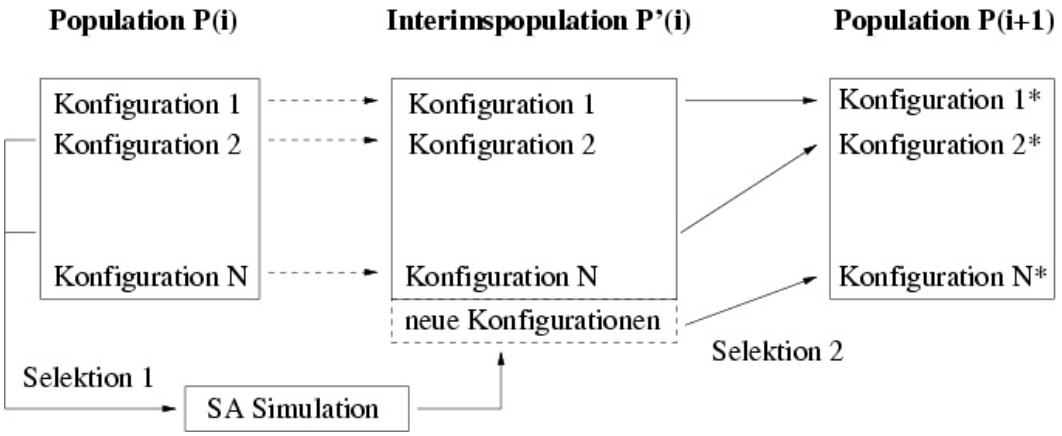


Abbildung 3.3: Ablaufskizze des evolutionären Monte-Carlo

positiv. Allerdings waren weitreichende Anpassungen notwendig, die sich darauf beziehen, welche Strukturen man für die SA Simulationen auswählt und welche Strukturen in die nächste Population aufgenommen werden. Einige Anpassungen begründen sich dadurch, daß auch für das EMC Verfahren in erster Linie unterschiedliche niederenergetische Strukturen generieren soll.

Wenn man bedenkt, daß, wie im letzten Abschnitt gezeigt, eine SA Simulation Strukturen generieren kann, die praktisch mit der Ausgangsstruktur übereinstimmen, d.h. gleiche Energie und eine RMSB Abweichung von unter 1Å haben, so ergibt sich schnell folgendes *Super-Gau-Szenario*. Wenn wir mit hoher Wahrscheinlichkeit die M energetisch niedrigsten Strukturen der Population $P(i)$ auswählen, und die anschließende SA Simulation keine Änderungen mit sich bringt, so sind die ausgewählten M Konfigurationen nun zweifach in dem aus $N + M$ Strukturen bestehendem Strukturensatz $P'(i)$ enthalten. Bildet man nun einfach aus den energetisch besten Strukturen dieses Satzes die nächste Population $P(i + 1)$, so sind die energetisch besseren Strukturen nun doppelt enthalten. In den folgenden Iterationen $P(i + 2)$, $P(i + 3)$, usw., forciert sich diese Situation, bis das Verfahren "konvergiert" ist und nur die ursprünglich beste Struktur (nun N-fach) in der Population vertreten ist.

Um dieser Problematik zu begegnen, sind einige Selektionsmechanismen eingeführt worden. Zunächst einmal werden SA Simulationen von alle Strukturen einer Population mit der gleichen Wahrscheinlichkeit aus gestartet. Hiermit wird auch vermeidlich schlechten Strukturen die Chance zur Verbesserung gegeben. Die von SA gefundene Struktur verdrängt die energetisch schlechteste, wenn sie eine tiefere Energie hat und eine RMSB Abweichung von mehr als 1Å zu allen Strukturen der Population hat. D.h. die SA Struktur wird übernommen, wenn sie sowohl neu als auch energetisch gut ist. Sollte die neue Struktur einer schon vorhandenen Struktur ähnlich sein ($RMSB < 1$Å), so kann sie nur diese aus der Population verdrängen. Die Frage, ob die neue Struktur diese alte ersetzt, wird

durch das TA Kriterium mit einer Schranke von $\Delta E = 3\,kcal\,mol^{-1}$ entschieden.

Die Kriterien, eine neue Struktur aufzunehmen, sind so konzipiert, daß sie für jeden SA Lauf einzeln angewandt werden können und nicht erst die Ergebnisse aller M Simulationen vorliegen müssen. Durch diese asynchrone Verarbeitung ist das evolutionäre Monte-Carlo Verfahren auch für den Einsatz auf heterogenen Computerclustern geeignet.

Darüber hinaus wird die Laufzeit der SA Simulationen angepaßt. Sie wird erhöht, wenn N-mal keine der neu generierten Strukturen akzeptiert wird. Verkürzungen der Laufzeit haben in keiner Formulierung eine sinnvollen Effekt erzielt. Vielmehr kam nach einer Verkürzung praktisch immer ein Zyklus, in dem keine neue Struktur akzeptiert wurde, d.h. die Laufzeit wurde direkt wieder erhöht. Dabei ist zu beachten, daß bei einer asynchronen Verarbeitung Laufzeiterhöhungen sich erst nach einer Verzögerung auf die Simulationen auswirken, da zunächst noch die Ergebnisse der Läufe mit der niedrigeren Schrittzahl abgearbeitet werden.

Es gab eine Reihe von Variationen der Selektionkriterien und der Laufzeitanpassung. Das TA Kriterium wurde eingeführt, da es ein klares Ja-Nein Kriterium darstellt, im Gegensatz zu Kriterien die nur Wahrscheinlichkeiten für ein Ja angeben, wie etwa das Metropolis Kriterium. Dies hatte den praktischen Hintergrund, daß bei Ja-Nein Aussagen die Fehlersuche in der streckenweise recht undurchsichtigen Verwaltung der Strukturen deutlich einfacher ausfällt.

Der Algorithmus des Freien Faltens ergibt sich als Spezialfall des EMC mit $N = 1$ mit spezifischer Laufzeitanpassung.

3.2 Aufbau biomolekularer Kraftfelder

Mit den vorgestellten Verfahren sollte es uns möglich sein, das globale Minimum einer Zielfunktion zu identifizieren. Die folgenden Abschnitte beschäftigen sich nun mit den Konstituenten der zu optimierenden Funktion, welche durch die Freie Energie eines Proteins in Wechselwirkung mit seiner natürlichen Umgebung gegeben ist. Die Wechselwirkungsenergie zwischen den Atomen beruht auf elektromagnetischen Kräften. Wenn die quantenmechanischen Wellenfunktionen der Atome aus der Lösung der Schrödinger Gleichung bekannt sind, so ist es im Prinzip möglich aus den Ladungsverteilungen $\rho_{1,2}$ die Energie aus Integration über die Coulomb-Potentiale, gemäß des *Hellmann-Feyman Theorems*, zu bestimmen:

$$E = \iint \frac{\rho_1(\vec{r}_1)\rho_2(\vec{r}_2)}{|\vec{r}_1 - \vec{r}_2|}\, d\vec{r}_1 d\vec{r}_2$$

Da quantenmechanische Rechnungen aufgrund der Systemgröße (bei mangelnden Symmetrien) unmöglich sind, wird auf Methoden der klassischen Mechanik zurückgegriffen. Die wohl realistischste Beschreibung eines Proteins ist die Molekulardynamik Simulation aller Atome des Proteins und des Lösungsmittels. Für

diese Art von Simulationen sind verschiedene Funktionen entwickelt worden, die einem Punkt im Phasenraum eine Energie bzw. eine Kraft zuordnen. Diese Funktionen werden Kraftfelder genannt, unabhängig davon, ob einer Konfiguration eine Kraft oder eine Energie zugeordnet wird. Werden die Atome des Lösungsmittels mitsimuliert, so spricht man von einem *explizitem Lösungsmittel*. Hängt die Hamiltonfunktion des Systems nicht von der Koordinaten der Lösungsmittelatome ab, sondern enthält einen Anteil der über die Koordinaten des Proteins den Einfluß der Wasseratome beschreibt, bezeichnet man dies als *implizites Lösungsmittelmodell*.

3.2.1 Wechselwirkungen chemisch gebundener Atome

Die analytischen Formen biomolekularer Kraftfelder haben viele Gemeinsamkeiten. Zunächst einmal beschränken sich alle Kraftfelder auf Zweikörperkräfte. Darüber hinaus wird zwischen Wechselwirkungen "gebundener" (*bond*) und "ungebundener" (*non-bond*) Atome unterschieden. Erstere beziehen sich auf Wechselwirkungen zwischen Atompaaren, die durch ein, zwei oder drei konsekutive kovalente Bindungen verbunden sind. Diese werden als 1-2, 1-3 oder 1-4 Wechselwirkungen bezeichnet. Unter ungebundenen Wechselwirkungen werden alle weiteren Beiträge subsumiert.

Die 1-2 Wechselwirkung beschreibt chemische Bindungen. Diese Bindungen besitzen Schwingungsfreiheitsgrade, die sich aus ihrer quantenmechanischen Natur ableiten. Diese Freiheitsgrade können in gewissen Grenzen in ein klassisches Potential integriert werden. Hierbei begnügt man sich zumeist damit, diese Freiheitsgrade in harmonischer Näherung zu beschreiben, d.h. eine Abweichung δl der Bindungslänge vom Gleichgewichtsbindungsabstandes l_0 führt für $\delta l \ll l$ zu einer Energieänderung proportional δl^2. Eine genauere Beschreibung des Potentialverlaufes als Funktion des Bindungsabstandes bietet das Morse-Potential

$$E(l) = E_0(1 - e^{-A(l - l_0)})^2 \ ,$$

welches in guter Näherung über einen großen Wertebereich von l gültig ist. Die Wesensmerkmale des Morse-Potential sind die *Core*-Abstoßung für kleine l und für $l \gg l_0$ der Übergang in ein r^{-1} der Coulomb-Wechselwirkung. Die Wechselwirkungskonstanten lassen sich aus quantenchemischen Rechnungen beziehen.

Das Potential für die Bindungswinkel θ wird der 1-3-Wechselwirkung zugeordnet. In einem klassischen Kraftfeld wird auch hier zumeist in harmonischer Näherung gearbeitet, $E \sim (\theta - \theta_0)^2$. Bei Rotationen um chemische Bindungen, 1-4-Wechselwirkungen, wird sehr oft ein Potential der Form $E \sim 1 + \cos(n\phi + \gamma)$ verwendet. Als Beispiel für die Verwendung dieses Potential sei das Kraftfeld AMBER angeführt, in dem Drehungen um $X - C - C - X$ mit $n = 3$, $\gamma = 0$ beschrieben werden.

3.2.2 Wechselwirkungen ungebundener Atome

Bei der klassischen Beschreibung der Wechselwirkung zwischen zwei Atomen wollen wir der Einfachheit halber damit beginnen, die Atome durch Punkte im Raum mir verschiedenen Eigenschaften wie Masse, Ladung und Polarisierbarkeit zu repräsentieren. Sind die beiden Punkte unendlich weit voneinander entfernt, so wird ihre Wechselwirkungsenergie zu Null angenommen. Führt man die Punkte bis auf eine Entfernung r zusammen, so ist dafür die Arbeit

$$E(r) = \int_{\infty}^{r} \vec{F}(r)dr = -\int_{r}^{\infty} \vec{F}(r)dr$$

aufzubringen, wobei $F(r)$ die Kraft zwischen den Atomen angibt. Eine repulsive Kraft wird dabei stets positiv, eine attraktive stets negativ gewertet. Aus der Arbeit erhält man die Kraft zurück, wenn man den Gradienten der Arbeit bildet $\vec{F}(r) = \nabla E(r)$.

Coulomb-Wechselwirkung

Die Wechselwirkungsenergie zweier Ionen mit den Ladungen q_1 und q_2, welche durch einen Abstand r voneinander getrennt sind, ergibt sich aus dem Coulomb-Gesetz

$$E = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r} \ .$$

Sie ist kleiner Null, also attraktiv, wenn die Ladungen unterschiedliches Vorzeichen haben. Befinden sich die Ionen in einem homogenem Medium, so ist die Dielektrizitätskonstante ϵ_0 mit einem Faktor ϵ zu multiplizieren. In einem Festkörper können frei bewegliche Elektronen zu einer Abschirmung der Ionenladungen führen, wodurch das $1/r$-Verhalten der Wechselwirkungsenergie nicht mehr erfüllt ist. Die Gültigkeit des Coulomb-Gesetzes bleibt nach Einführung einer abstandsabhängigen Dielektrizitätsfunktion $\epsilon \mapsto \epsilon(r)$ erhalten. Einen vergleichbaren Effekt haben Grenzenflächen zwischen Medien mit unterschiedlichem ϵ. Hier erfolgt die Abschirmung durch Polarisation der Grenzschicht. Die Wechselwirkung mit der Polarisation läßt sich für einfache Geometrien durch sogenannte Spiegelladungen beschreiben.

Das bekannteste Beispiel zur Thematik der Spiegelladungen besteht darin, eine Ladung q zu betrachten, die in einem Halbraum mit der Dielektrizitätskonstante ϵ_1 liegt und einen Abstand d von der Grenzfläche zum Halbraum mit der Dielektrizitätskonstante ϵ_2 hat. Die Grenzfläche wird in die xy-Ebene mit $z = 0$ gelegt, sowie die Ladung in den Punkt $(0; 0; d)$. Gesucht ist nach dem Potential Φ des elektrischen Feldes $\vec{E}$. Die Lösung des Randwertproblems

$$\begin{aligned} \epsilon_1 \nabla \vec{E} &= 4\pi\rho \quad \text{für } z > 0 \\ \epsilon_2 \nabla \vec{E} &= 4\pi\rho \quad \text{für } z < 0 \\ \nabla \times \vec{E} &= \vec{0} \end{aligned}$$

$$\lim_{z\to 0^+} \begin{Bmatrix} E_x \\ E_y \\ \epsilon_1 E_z \end{Bmatrix} = \lim_{z\to 0^-} \begin{Bmatrix} E_x \\ E_y \\ \epsilon_2 E_z \end{Bmatrix}$$

ist für $z > 0$ gegeben durch

$$\Phi(\vec{r}) = \frac{1}{4\pi\epsilon_1}\left(\frac{q}{d_1} + \frac{q'}{d_2}\right) .$$

Der Anteil $R = \frac{q'}{\epsilon_1 d_2}$ wird Reaktionsfeld genannt. $d_{1,2}$ sind die Abstände von $\vec{r}$ zu der Ladung $\vec{q}$ bzw. zu der Spiegelladung q' bei $(0; 0; -d)$, die folgende Ladung trägt:

$$q' = -\left(\frac{\epsilon_2 - \epsilon_1}{\epsilon_2 + \epsilon_1}\right) q . \tag{3.16}$$

Setzt man die Werte der Dielektrizitätskonstanten für ein Protein $\epsilon_1 \approx 3\epsilon_0$ und für Wasser $\epsilon_2 \approx 80\epsilon_0$ ein, so zeigt sich, daß eine Ladung q in einem proteinähnlichem Medium eine Grenzschichtpolarisation erzeugt, die einer Spiegelladung im Wasser von $q' \approx -0.93q$ entspricht. Dieser Wert ist unabhängig vom Abstand der Ladung von der Grenzfläche und führt zu einer attraktiven Wechselwirkung der Ladung q mit der Grenzschichtpolarisation. Eine Ladung in einem Protein wird daher immer bestrebt sein, an die Oberfläche des Proteins zu gelangen.

Eine für Proteine realistischere Geometrie ist eine Ladung q in einer Kugel (mit ϵ_1), die von einem Medium (mit ϵ_2) umgeben ist. Sei a der Radius der Kugel und s der Abstand der Ladung vom Kugelmittelpunkt, so kann das Reaktionsfeld

$$R(r, \cos\theta) = \frac{\epsilon_1 - \epsilon_2}{\epsilon_1}\frac{q}{a}\sum_{n=0}^{\infty}\frac{n+1}{n + \frac{\epsilon_2}{\epsilon_1}(n+1)}\left(\frac{rs}{a^2}\right)^2 P_n(\cos\theta) ,$$

wobei P_n die Legendre Polynome bezeichnet, in erster Näherung zu einer Spiegelladung im Abstand a^2/s vom Kugelmittelpunkt mit

$$q' = -\left(\frac{\epsilon_2 - \epsilon_1}{\epsilon_2 + \epsilon_1}\right)\frac{a}{s}q \tag{3.17}$$

zusammengefaßt werden. Eine Herleitung und Diskussion dieses Ergebnisses in Bezug auf Monte-Carlo und Molekulardynamik Simulationen ist in [Fri75] zu finden. Für Ladungen nahe der Proteinoberfläche ($s \approx a$) sind die beiden Spiegelladungen 3.16 und 3.17 nahezu identisch. Mit den Dielektrizitätskonstanten für das Protein und für Wasser, folgt in 3.17 wegen $a \approx s$ rasch $|q'| \approx |q|$. Dies bedeutet, daß bei kompakten (kugelförmigen) Proteinstrukturen Ladungen an oder nahe der Proteinoberfläche eine (praktisch) gleichgroße Spiegelladung in unmittelbarer Nähe erzeugen. Damit ist die Ladung q etwa auf der gegenüberliegenden Seite des Proteins nicht mehr zu sehen, sondern nur noch ein Dipolmoment der Ladungen q und q'.

Wechselwirkungen permanenter und induzierter Dipole

Wenn ein Molekül mit einem permanenten Dipolmoment $\vec{p}$ einem elektrischen Feld $\vec{E}$ ausgesetzt ist, so hat es die Energie

$$E = -\vec{p} \cdot \vec{E} = -|\vec{p}| \cdot |\vec{E}| \cos\theta \; .$$

Wird das elektrische Feld von einer Ladung im Abstand r erzeugt, so ist die resultierende Energie proportional r^{-2}. Wird das Feld durch einen weiteren Dipol ersetzt, so erhält man eine Energie die proportional zu r^{-3} ist.

Werden Ladungen in einem Molekül einem elektrischen Feld ausgesetzt, so führt dies zu einer Verschiebung der Molekülladungen, und das Molekül wird polarisiert. Die Stärke des induzierten Dipols $\vec{p}$ ist

$$\vec{p} = \alpha\epsilon_0\vec{E} \; ,$$

wobei α die Polarisierbarkeit des Moleküls angibt. In erster Näherung ist α ein Skalar mit der Dimension eines Volumens. Die Wechselwirkungsenergie dieses induzierten Dipols mit dem ihn erzeugenden elektrischen Feld ist

$$E = -\int_0^{|\vec{E}|} \vec{p} \cdot d\vec{E} = -\frac{1}{2}\epsilon_0\alpha|\vec{E}|^2 \; . \tag{3.18}$$

Wenn das elektrische Feld von einer Ladung q induziert wird, so ist

$$E = -\epsilon_0\alpha\frac{2q^2}{(4\pi\epsilon_0)^2}\frac{1}{r^4} \; .$$

Diese Energie entspricht stets einer attraktiven Kraft zwischen Ladung und Molekül.

Van-der-Waals Wechselwirkung

Durch Fluktuationen der Ladungsverteilung in einem Atom entstehen kurzzeitig Dipole, die andere Atome polarisieren können. Da das Feld eines Dipols mit r^{-3} variiert, folgt aus Gleichung 3.18, daß die Wechselwirkungsenergie induzierter Dipole mit r^{-6} variiert. Die Frequenz der zugrundeliegenden Fluktuationen ist sehr groß, sodaß die Polarisierbarkeit nicht mehr als statische Größe aufgefaßt werden kann. Vielmehr ist sie mit der Bewegung der Elektronen in einem elektromagnetischen Feld mit Frequenzen des optischen Spektrums verknüpft. Daher wird diese Wechselwirkung auch als Dispersionskraft bezeichnet, und ihre Stärke kann nicht mehr nur aus der klassischen Physik abgeleitet werden. Eine Kombination von klassischer und quantenmechanischer Theorie ergibt eine Wechselwirkungsenergie von

$$E = -\frac{3\hbar\omega_1\omega_2}{2(\omega_1+\omega_2)} \cdot \frac{\alpha_1\alpha_2}{r^6} \; . \tag{3.19}$$

Hierbei werden die Atome als harmonische Oszillatoren mit den Frequenzen $\omega_{1,2}$ und statischen Polarisierbarkeiten $\alpha_{1,2}$ beschrieben.

Vermeidung einer Pauliprinzipverletzung

Das attraktive van-der-Waals-Potential wird kombiniert mit einem starken repulsivem Potential, um einen Überlapp der Atomorbitale für kleine Abstände r zu verhindern. Die gebräuchlichste Form ist das *Lennard-Jones-Potential*, welches sich als Funktion des Abstands r folgendermaßen schreibt:

$$V_{LJ}(r) = V_0 \left[\left(\frac{\tau}{r}\right)^{12} - 2\left(\frac{\tau}{r}\right)^{6} \right] . \tag{3.20}$$

Das Minimum dieses Potentials liegt bei $r = \tau$ mit $V(\tau) = -V_0$. Die Werte für τ und V_0 sind atomspezifisch, wobei für $r \gg \tau$ die Formel 3.20 in 3.19 übergeht.

Wasserstoffbrückenbindungen

Die für die Proteinstruktur zentrale Wechselwirkung der Wasserstoffbrückenbindung wird in einer klassischen Theorie stets als Dipol-Dipol-Wechselwirkung klassifiziert.

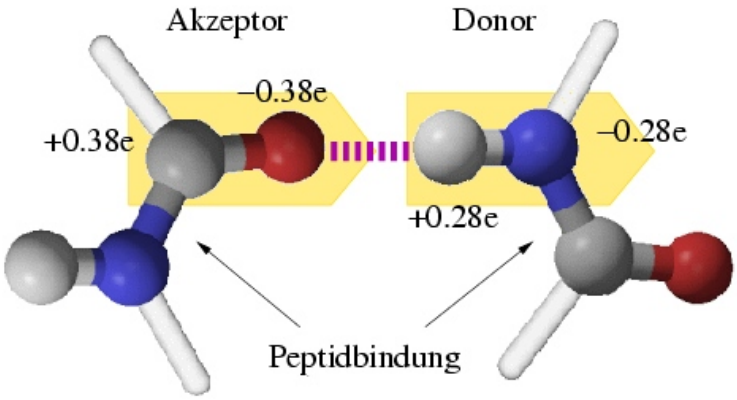


Abbildung 3.4: Wasserstoffbrückenbindung zwischen zwei Peptidhauptkettensegmenten. Angegeben sind die Partialladungen der Atome.

Voraussetzung ist, daß ein Wasserstoffatom (Donor) an ein stark elektronegatives Atom, z.B. Sauerstoff (bei Wasser) oder Stickstoff (Proteinhauptkette), gebunden ist (Abbildung. 3.4). Dieses positiv polarisierte Wasserstoffatom kann in Wechselwirkung mit einem negativ polarisierten Atom (Akzeptor) treten, wie z.B. dem Sauerstoff in der Proteinhauptkette. Bei kurzen Donor-Akzeptor-Abständen kann sich der Bindungsabstand des Wasserstoffatoms zu seinem Donor-Bindungspartner, zugunsten eines kürzeren Donor-Akzeptor-Abstands, vergrößern. Diese Abstandsänderung ist allerdings recht klein; sie wird daher innerhalb einer klassischen Beschreibung vernachlässigt. Die Energie einer Wasserstoffbrückenbindung wird approximativ als proportional zur Dipol-Dipol-Wechselwirkung des donorseitigen HN-Dipols und des akzeptorseitigen CO-Dipols angesetzt.

Für die Computersimulation ist es hilfreich, die Dipole als zwei getrennte Ladungen zu repräsentieren, obwohl die Donor-Akzeptor-Abstände mit dem Abstand zwischen Wasserstoff- und Sauerstoffatom vergleichbar sind, d.h. daß formal die Benutzung einer Fernfeldnäherung nicht möglich ist.

$$E = \frac{0.38e \cdot 0.28e}{4\pi\epsilon\epsilon_0}\left(\frac{1}{|\vec{r}_C - \vec{r}_H|} - \frac{1}{|\vec{r}_C - \vec{r}_N|} - \frac{1}{|\vec{r}_O - \vec{r}_H|} + \frac{1}{|\vec{r}_O - \vec{r}_N|}\right) \quad (3.21)$$

Allerdings weicht die Winkel- und Abstandsabhängigkeit dieses Potentials von der natürlichen, quantenchemisch begründeten, Abhängigkeit ab. Daher wurde in vielen Kraftfeldern ein zusätzlicher Term eingeführt, der bei der Beschreibung von Wasserstoffbrückenbindungen helfen soll.

Das 1-6-12-Potential

Die Wechselwirkung nicht miteinander gebundener Atome wird in einem klassischen Potential durch Terme beschrieben, die Potenzen des inversen Atomabstandes r sind. Dabei sind die sogenannten 1-6-12 Terme in praktisch jedem Kraftfeld integriert. Diese Terme stehen für die Coulomb-Wechselwirkung ($\sim r^{-1}$), die Dispersions- ($\sim r^{-6}$) und repulsive Wechselwirkung ($\sim r^{-12}$), wobei die letzten beiden Summanden zur Lennard-Jones Wechselwirkung zusammengefaßt werden.

3.2.3 Molekulardynamische Simulationen

Sind alle im Protein-Wasser System auftretenden Kräfte bekannt, so kann innerhalb einer klassischen Theorie die Bewegung aller Atome durch Differentialgleichungen beschrieben werden. Zur Vereinfachung der Schreibweise fassen wir die Atomkoordinaten $\vec{x}_i = (x_{i1}; x_{i2}; x_{i3})^T$ aller Atome $i \in \{1, \ldots, N\}$ zu einer Koordinate

$$x = (x_{11}; x_{12}; x_{12}; \ldots; x_{N1}; x_{N2}; x_{N2})^T$$

im $3N$-dimensionalen Raum zusammen. Für das Protein 1VII mit 36 Residuen gehen hier N =596 Atome gemeinsam mit den Wassermolekülen ein, wobei auf letztere verzichtet werden kann, wenn es über ein implizites Lösungsmittel eine zusätzliche Kraft gibt, die die Wassermoleküle implizit beschreibt.

Bei endlichen Temperaturen folgt das Kräftegleichgewicht des Systems der *Langevin Dynamik*, also einer stochastischen Differentialgleichung [kse98]

$$M\ddot{x} + C\dot{x} + \nabla V(x) = D\dot{W} \; . \quad (3.22)$$

M bezeichnet die Matrix, welche die Atommassen auf der Diagonalen enthält, C gibt die Dämpfung des Systems an und V ist das Potential der oben beschriebenen Kräfte. Auf der rechten Seite der Gleichung steht eine zufällige Kraft, die aus Fluktuationen durch Stößen mit der Umgebung resultiert, wodurch Energie dispergiert. $\dot{W}(t)$ beschreibt einen normierten Wiener Prozeß, welcher auch als

normiertes weißes Rauschen bezeichnet wird[1]. Die Diffusionsmatrix D und der Dämpfungsterm C hängen über das Fluktuations-Dissipations-Theorem zusammen

$$DD^T = 2k_BT\,C\ ,$$

wodurch die Verbindung zur Temperatur des Systems gegeben ist. Ein üblicher Ansatz für die Matrix C ist es, eine skalaren Dämpfungskoeffizienten $\gamma > 0$ einzuführen, und $C = \gamma M$ zu setzen. Dann ergibt sich aus obigem Theorem $D = \sqrt{2k_BT\gamma}M^{1/2}$.

Im Tieftemperaturlimes geht die Differentialgleichung 3.22 in

$$M\ddot{x} + C\dot{x} + \nabla V(x) = 0$$

über. Für die Energie $E = \frac{1}{2}\dot{x}^T M\dot{x} + V(x)$ folgt $\dot{E} = \frac{1}{2}\ddot{x}^T M\dot{x} + \nabla V(x)^T\dot{x} = -\dot{x}^T C\dot{x}$. Da C positiv definit ist, verliert das System solange Energie bis alle Energie in der potentiellen Energie enthalten ist.

Für verschwindende Dämpfung $C \to 0$ erhält man die *Newtonschen Bewegungsgleichungen* eines Systems ohne Reibung

$$M\ddot{x} = -\nabla V(x)\ .$$

Diese partielle Differentialgleichung findet Verwendung bei Simulationen, die das Wasser explizit berücksichtigen. Die Temperatur des System steht gemäß des *Gleichverteilungssatzes der klassischen statistischen Physik* in Beziehung zur kinetischen Energie der Atome

$$\langle\frac{1}{2}\dot{x}^T M\dot{x}\rangle = \frac{3}{2}\,N\,k_BT\ .$$

3.2.4 Etablierte Kraftfelder

Um einen ersten Eindruck von der analytische Form eines biomolekularen Kraftfeldes zu erhalten, sei hier exemplarisch das AMBER Potential (in einer älteren

[1]Für Zeitentwicklung eines solchen Prozesses von t zu $t+\Delta t$ gilt: $W(t+\Delta t) = W(t)+z\sqrt{\Delta t}$ mit normalverteiltem z bei Varianz 1.

Version) angegeben.

$$
\begin{array}{rll|l}
E = & \sum_{\text{Bindungen}} & K_1(l-l_0)^2 & l: \text{Bindungsabstand} \\
+ & \sum_{\text{Bindungswinkel}} & K_2(\theta-\theta_0)^2 & \theta: \text{Bindungswinkel} \\
+ & \sum_{\text{Dihedralwinkel}} & \frac{K_3}{2}[1-\cos(n\phi+\gamma)] & \phi: \text{Dihedralwinkel} \\
+ & \sum_{\text{Atompaare}} & \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon\, r} & q: \text{Partialladungen} \\
+ & \sum_{\text{Atompaare}} & \left(\frac{A_{ij}}{r^{12}}-\frac{B_{ij}}{r^{6}}\right) & r: \text{Atomabstand} \\
+ & \sum_{\text{Wasserstoffbrücken}} & \left(\frac{C}{r^{12}}-\frac{D}{r^{10}}\right) &
\end{array}
\tag{3.23}
$$

Mit diesem Kraftfeld ist 1998 die bislang längste Molekulardynamik Simulation eines Proteins mit explizitem Lösungsmittel durchgeführt worden [DK98, DWK98]. Es handelt sich dabei um 1VII, welches eine drei-Helix-Struktur besitzt und mit 36 Residuen für alle Atome einschließende Computersimulationen eine beachtliche Größe hat. Trotz enormen Rechenaufwands von insgesamt 85 CPU-Jahren auf einem der größten existierenden Rechnercluster wurde die native Struktur während der gesamten Simulation nicht ein einziges Mal eingenommen. Die Simulationszeit entspricht einer Mikrosekunde der natürlichen Bewegung des Proteins, dessen Faltungszeit etwa bei 10 Mikrosekunden liegt. Es bleibt offen, ob bei einer längeren Simulationszeit eine Trajektorie entstanden wäre, die durch die native Struktur geführt hätte. Andere Szenarien sind jedoch ebenso wahrscheinlich. Etwa, daß das zugrundegelegte Kraftfeld eine andere Struktur als globales Minimum der Freien Energie ausweist. Damit würden die meisten mit diesem Kraftfeld erzeugten Trajektorien niemals an der nativen Struktur vorbeikommen.

Doch auch wenn die native Struktur auf einer Molekulardynamik Trajektorie liegt, so bedeutet dies nicht, daß diese Struktur auch das globale Minimum der Freien Energie des Kraftfeldes ist. Falls es sich nur um einen metastabilen Zustand handelt, wird die Proteinstruktur eine gewisse Zeit der nativen Struktur ähnlich sein, um dann in Richtung des kraftfeldspezifischen Minimums der Freien Energie weiterzuziehen. Um zu klären, ob das Kraftfeld die Natur richtig wiedergibt, wäre eine Vielzahl sehr langer Simulationen notwendig. Dies war zu dem damaligen Zeitpunkt technisch unmöglich durchzuführen. Seither gibt es nur einen großangelegten Versuch, die Freie Energie auf systematische Weise aus Molekulardynamik Simulationen zu gewinnen. Dieses Projekt bündelt Rechenzeit leerstehender Computer in Form eines speziellen Bildschirmschoners. Dieser Bildschirmschoner ist im Internet unter der Bezeichnung “folding@home” beziehbar[2].

[2] $http://folding.stanford.edu$

Die bisherigen Resultate von folding@home zeigen, daß es mittels Molekulardynamik Simulationen auch mit implizitem Lösungsmittel und moderner Computertechnik nicht möglich ist, (für mehr als die kleinsten Peptide) prediktiv Proteinstrukturvorhersage zu betreiben. Die erfolgreiche Faltung von 1L2Y, einem Protein mit 20 Residuen, umfaßte einen Rechenaufwand von 250 CPU-Jahren [SZP02]. Ein Faltungsversuch mit unserem Kraftfeld und Einsatz der Monte-Carlo Simulationstechnik war nach einem halben CPU-Jahr vollständig konvergiert [SHW03].

Die Verwendung der Molekulardynamik zur Proteinstrukturvorhersage zeigt hier eine weitere Schwäche. Die native Struktur entspricht einer "mittleren" Struktur der Trajektorie, deren Fluktuationen bei $300K$ sehr stark sind, so daß es nur bedingt möglich ist, den Begriff mittlere Struktur sauber zu definieren [SZP02].

Wie wir in einem späteren Kapitel sehen werden, ist durch die Verwendung impliziter Lösungsmittel die Freie Energie unmittelbar zugänglich geworden und muß nicht erst im nachhinein ermittelt werden. Damit ist es nicht mehr notwendig, entropische Beiträge zu simulieren, sondern man kann das Proteinstrukturproblem als reines Optimierungsproblem auf der Potentialoberfläche der Freien Energie verstehen. Es ist allgemein üblich, die Terminologie der Molekulardynamik teilweise zu übernehmen. So werden wir auch weiterhin von einem Kraftfeld sprechen, obwohl es treffender wäre, auf den Begriff "Potentialenergieoberfläche der Freien Energie" zu wechseln. Wie die Molekulardynamik Simulationen gezeigt haben, sind deterministische Verfahren mit der bei Proteinen vorliegenden Konfigurationsraumgröße überfordert, weshalb ausschließlich stochastische Optimierungsverfahren zum Einsatz kommen.

Mit dem Übergang zur Freien Energie ist es auch möglich, die Auswirkungen kleiner Kraftfeldveränderungen effizienter zu untersuchen und aufgrund dieser Analysen weitere Kraftfeldanpassungen durchzuführen. Bei der Molekulardynamik ist ein solcher Optimierungszyklus mit immensem Zeitaufwand verbunden und mit den existierenden Computerressourcen nicht durchführbar. Momentan ist ein neuer Supercomputer namens *Blue Gene*[3] in Planung, der diese Lücke teilweise füllen und Molekulardynamik Simulationen für Proteine in großem Maßstab ermöglichen soll. Ein geeignetes Kraftfeld auf diesem Computer sollte die Beantwortung vieler grundlegenden Fragen ermöglichen und zur Aufklärung des Faltungsmechanismus beitragen.

Biologische Fragestellungen, für deren Beantwortung die Kenntnis der Freien Energie ausreicht – also jede Frage zur Gleichgewichts-Thermodynamik – bedürfen keines solchen Supercomputers. Die Simulationen dieser Arbeit beruhen großteils auf einem am Institut für Nanotechnologie (INT) des Forschungszentrum

[3]Dieser Rechner soll 2006 in Betrieb gehen und mit 1 PetaFLOPS den momentan stärksten Rechner namens *Earth Simulator* mit 41 TeraFLOPS von der Spitzenposition der stärksten Rechner der Welt verdrängen; siehe $http://www.top500.org$.

Karlsruhe existierendem Computercluster zur Proteinstrukturvorhersage, welches in der ersten Stufe aus 36 Computern mit 1GHz PentiumIII Prozessoren bestand und derzeit aus 24 1GHz PentiumIII(Coppermine), 18 1.2GHz AMD Athlon und 28 2.4GHz Intel Xeon Prozessoren zusammengesetzt ist.

Wir wollen nun zunächst unser Augenmerk auf andere Kraftfelder richten und einige der bekanntesten Kraftfelder namentlich kurz erwähnen (s.u.). Insbesondere die ersten drei Kraftfelder sind untrennbar mit den Namen ihrer Entwickler Scheraga, Karplus und Kollman verknüpft sind. Den angegebenen Literaturstellen können genauere Informationen zu den Kraftfeldern entnommen werden. Die meisten Kraftfelder sind für Molekulardynamik Simulationen entwickelt worden und in ein Programmpaket integriert. Die vorherrschende Programmiersprache ist FORTRAN.

ECEPP	Empirical Conformational Energy Program for Peptides	[MMBS75]
CHARMM	Chemistry at Harvard Macromolecular Mechanics	[BBO$^+$83]
AMBER	Assisted Model Building with Energy Refinement	[PCC$^+$95]
OPLS	Opimized Potentials for Liquid Simulations	[Jor81]
MM3	Molecular Mechanics	[AYL89]
GROMOS	Groningen Molecular Simulations	[SHT$^+$99]
CFF	Consistent Forcefield	[HE94]
ESFF	Extensible Systematic Forcefield	[ESF]

Die letzten beiden Kraftfelder CFF und ESFF sind unter anderem im Programmpaket InsightII integriert, welches in unserer Arbeitgruppe in Zusammenhang mit dem *Receptor Ligand Docking* Problem benutzt wird. Das *Receptor Ligand Docking* Problem beschäftigt sich, vereinfacht gesagt, damit, Affinitäten zwischen Molekülen (Liganden) und Proteinen zu bestimmen. ESFF ist besonders für die Behandlung verschiedenster Liganden geeignet, da hier auch selten auftretende Atome parametrisiert sind; Atome, die für andere Kraftfelder nicht existieren.

AMBER und CHARMM sind sich in der analytischen Form sehr ähnlich und verhältnismäßig einfach aufgebaut und nicht zuletzt deshalb weit verbreitet. In Tabellen 3.1 und 3.2 sind die Bestandteile verschiedener Kraftfelder zusammengetragen, wobei die Vorfaktoren, sofern möglich, weggelassen wurden. Die Kraftfeldkürzel sind: *A* AMBER, *C* CHARMM, *E* ESFF, *F* CFF und *M* MM3. Das ESFF Kraftfeld ist nur als Beispiel für die Verwendung des Morse Potentials

Beschreibung von	Potential	Kraftfeld
Bindungslängen	$(l-l_0)^2$	CA
	$k_2(l-l_0)^2+k_3(l-l_0)^3+k_4(l-l_0)^4$	F
	$(l-l_0)^2[1-2(l-l_0)]$	M
	$(1-e^{-A(l-l_0)})^2$ "Morse Potential"	E
Bindungswinkel	$(\theta-\theta_0)^2$	CA
	$k_2(\theta-\theta_0)^2+k_3(\theta-\theta_0)^3+k_4(\theta-\theta_0)^4$	F
	$(\theta-\theta_0)^2[1-7\times10^{-8}(\theta-\theta_0)^4]$	M
θl-Kreuzterme	$(\theta_{AB}-\theta_{AB,0})[(l_A-l_{A,0})+(l_B-l_{B,0})]$	M
	$k_{\theta\theta'}(\theta-\theta_0)(\theta'-\theta'_0)+k_{l\theta}(l-l_0)(\theta-\theta_0)+$ $+k_{ll'}(l-l_0)(l-l'_0)$	F
Dihedralwinkel	$\frac{1}{2}(1+\cos(n\phi+\gamma))$	A
	$1\pm\cos(n\phi)$	C
	$\frac{k_1}{2}(1+\cos\phi)+\frac{k_2}{2}(1+\cos2\phi)+$ $+\frac{k_3}{2}(1+\cos3\phi)$	MF
$\phi\theta l$-Kreuzterme	$(l-l_0)(k_1\cos\phi+k_2\cos2\phi+k_3\cos3\phi)$	F
	$(\theta-\theta_0)(k_1\cos\phi+k_2\cos2\phi+k_3\cos3\phi)$	F
	$\cos\phi(\theta-\theta_0)(\theta'-\theta'_0)$	F

Tabelle 3.1: Wechselwirkungen des 1-2, 1-3 und 1-4 Typs

aufgenommen worden. Aber auch ohne dieses ist das Spektrum der verwendeten Potentialformen beunruhigend vielfältig. Es gibt offensichtlich noch immer keinen Konsens über die funktionale Form der Bestandteile eines biomolekularen Kraftfeldes, auch wenn die zugrundegelegten Kräfte prinzipiell verstanden sind.

In den Einträgen der Tabelle 3.2 zur Elektrostatik und den Wasserstoffbrücken sind keine Zuordnungen zu den Kraftfeldern vorgenommen worden, da diese Bestandteile der Kraftfelder von Version zu Version variieren. So wurde z.B. in AMBER das zur Beschreibung von Wasserstoffbrückenbindungen dienende 10-12-Potential durch Parameteranpassung in das Lennard-Jones Potential integriert. In CHARMM sind verschiedene Potentiale integriert, die jedoch alle geringere Potenzen des inversen Abstands haben: $r^{-m}-r^{-n}$ mit $m=0,2,4$ und $n=0,2$. Um die Winkelabhängigkeit der Wasserstoffbrücken zu berücksichtigen, sind diese teilweise noch mit Kosinusfunktionen der Donor- und Akzeptorwinkel multipliziert. Der Eintrag $q_iq_j/(4\pi\epsilon r)$ im Abschnitt der Wasserstoffbrücken ist als Summe über die Donor und Akzeptoratome zu lesen.

Zu diesen intramolekularen Wechselwirkungen kommen diejenigen hinzu, die zwischen dem Protein und dem Lösungsmittel existieren. Diese zerfallen in zwei Beträge. Zum Einen ändert sich die Elektrostatik des Proteins und wird in den

Beschreibung von	Potential	Kraftfeld
van-der-Waals	$Ar^{-12} - Br^{-6}$	CA
	$Ar^{-9} - Br^{-6}$	F
	$e^{-Ar/r_0} - B\left(\frac{r_0}{r}\right)^6$	M
Elektrostatik	$q_i q_j/(4\pi\epsilon\, r)$ wahlweise	
	$\epsilon = \epsilon_0\,,\ \epsilon = \epsilon_0 r\,,\ \epsilon = \epsilon_0(r + \ldots)$	
Wasserstoffbrücken	$q_i q_j/(4\pi\epsilon\, r)$	
	$Ar^{-12} - Br^{-10}$	
	$\cos\theta(Ar^{-12} - Br^{-6})+$	
	$+(1 - \cos\theta)(Cr^{-12} - Dr^{-6})$	
	u.v.m.	

Tabelle 3.2: Wechselwirkungen oberhalb von 1-4

entsprechenden Bestandteil der intramolekularen Wechselwirkungen integriert, was zu unterschiedlichsten Dielektrizitätsfunktionen führt. Der Andere Beitrag des Lösungsmittel wird als hydrophober Effekt bezeichnet. Bei dessen Beschreibung wird zumeist auf den Ansatz von Eisenberg und McLachlan zurückgegriffen, dem gemäß der Energiebeitrag proportional zur Oberfläche A_i der Atome ist [EM86]

$$E = \sum_i \sigma_i A_i\ .$$

Teils aus Ermangelung einer Alternative herrscht hier ein gewisser Konsens unter den Kraftfeldnutzern und -entwicklern. Es hat sich in vielen vergleichenden Molekulardynamik Simulationen gezeigt, daß dieses Modell recht gute Resultate liefert und dabei mehrere Größenordnungen weniger Rechenzeit benötigt als eine Simulation mit explizitem Lösungsmittel [SSH94].

Die hier beschriebene Arbeit baut auf einem in der Arbeitsgruppe von John Moult am CARB (Centre for Advanced Research in Biotechnology) entwickeltem Kraftfeld auf. In seiner funktionalen Form ist dieses Kraftfeld von bestechender Schlichtheit [Avb92, AM95, AJ95]. Unter Vernachlässigung der Bindungswinkel und -längen sind Dihedralwinkel die einzigen Freiheitsgrade des Proteins. Ein spezielles Dihedralwinkelpotential der Form $\frac{1}{2}(1 + \cos(n\phi + \gamma))$ ist nicht explizit vorhanden, sondern kann als in die Elektrostatik integriert betrachtet werden. Ebenso werden Wasserstoffbrückenbindungen in einfacher Weise durch die elektrostatische Wechselwirkung der 2 Donor- und der 2 Akzeptoratome beschrieben, die man als Approximation der Dipol-Dipol-Wechselwirkung auffassen darf. Das Kraftfeld besteht also nur aus van-der-Waals Wechselwirkung, (Haupt- und Seitenketten-) Elektrostatik und Lösungsmittelbeiträgen. Allerdings verbergen

sich, wie in anderen Kraftfeldern auch, noch einige Details in den Wechselwirkungskoeffizienten, und so spaltet z.B. der elektrostatische Term dreifach auf.

Fügt man den Bezeichnungen von Gleichung 3.23 die Oberfläche A_i des i-ten Atoms hinzu, dann schreibt sich das am CARB entwickelte Kraftfeld wie folgt:

$$
\begin{array}{lll|l}
E = & \displaystyle\sum_{i,j} & \left(\frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^6}\right) & \text{Lennard-Jones} \\
+ & \displaystyle\sum_{\substack{i,j \\ i \in \text{Seitenkette}}} & \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon_{i,j}\, r} & \text{Seitenketten-ES} \\
+ & \displaystyle\sum_{\substack{i,j \in \text{Hauptkette} \\ res(i)=res(j)\pm 1}} & \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon_i\, r} & \text{lokale Hauptketten-ES} \\
+ & \displaystyle\sum_{\substack{i,j \in \text{Hauptkette} \\ |res(i)-res(j)|\geq 2}} & \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon_{hb}\, r} & H\text{-Brücken} \\
+ & \displaystyle\sum_{i \in \text{“Typ I”}} & \sigma_i A_i & \text{Lösungsmittel I} \\
+ & \displaystyle\sum_{i \in \text{“Typ II”}} & \sigma_i A_i^3 & \text{Lösungsmittel II}
\end{array}
\tag{3.24}
$$

Die Spezifikation der Lösungsmittelbeiträge Typ I und Typ II sind hier nicht weiter von Bedeutung, ebenso wie die Dielektrizitätskonstanten. Beides wird uns später genauer beschäftigen.

3.2.5 Zur Transferierbarkeit der Kraftfeldterme

Auch wenn wir nach und nach unser eigenes Kraftfeld entwickelt haben, so inkooperieren wir teilweise Ergebnisse anderer Arbeitsgruppen, die mit anderen Kraftfeldern arbeiten. Daher soll an dieser Stelle darauf hingewiesen werden, daß die Adaption einzelner Kraftfeldterme für die Freie Energie problematischer ist als etwa für die innere Energie [MvG94]. Dies wird uns im Abschnitt über die elektrostatische Wechselwirkung nochmals beschäftigen. Die Problematik ist jedoch auch bei der Umsetzung experimenteller Daten in theoretische Modelle wichtig und spielt z.B. für die Transferenergien des Proteins von Oktanol zu Wasser eine Rolle, die uns in einem späteren Abschnitt begegnen werden.

Wenn wir zu unserem bestehenden Kraftfeldhamiltonian H_1 einen neuen Bestandteil H_2 hinzufügen wollen, so setzt sich die innere Energie von $H = H_1 + H_2$ zusammen aus den inneren Energien der einzelnen Hamiltonians:

$$U = \langle H \rangle = \langle H_1 \rangle + \langle H_2 \rangle = U_1 + U_2 \; . \tag{3.25}$$

Hierbei sind die thermischen Erwartungswerte bezüglich H zu verstehen. Für eine Observablen $A(\vec{p},\vec{q})$ lautet dieser thermische Erwartungswert

$$\langle A\rangle = \frac{\iint A(\vec{p},\vec{q})\, e^{-\beta H(\vec{p},\vec{q})} d\vec{p}\, d\vec{q}}{\iint e^{-\beta H(\vec{p},\vec{q})} d\vec{p}\, d\vec{q}} \; .$$

Die Freie Energie verhält sich etwas anders als die innere Energie. Zunächst läßt sich die Freie Energie

$$F = -\beta^{-1} \ln\left[\frac{1}{h^{3N}} \iint e^{-\beta H(\vec{p},\vec{q})} d\vec{p}\, d\vec{q}\right] \tag{3.26}$$

als Erwartungswert reformulieren (wenn man von einer Konstanten absieht, die in der Differenz der Freien Energie nicht interessiert)

$$\begin{aligned} F &= -\beta^{-1} \ln\left[\frac{\iint e^{-\beta H(\vec{p},\vec{q})} d\vec{p}\, d\vec{q}}{\iint e^{+\beta H(\vec{p},\vec{q})} e^{-\beta H(\vec{p},\vec{q})} d\vec{p}\, d\vec{q}}\right] \\ &= +\beta^{-1} \ln\langle e^{+\beta H}\rangle \; . \end{aligned} \tag{3.27}$$

Die Entropie läßt sich ebenfalls als Ensemble-Mittel angeben:

$$S = -\frac{\partial F}{\partial T} = \langle H\rangle/T - k_B \ln\langle e^{+\beta H}\rangle \; . \tag{3.28}$$

Nach einer Entwicklung der Exponential- und Logarithmusfunktion folgt[4]

$$\begin{aligned} F &= \beta^{-1} \ln\left(1 + \sum_{n>0} \frac{\beta^n \langle H^n\rangle}{n!}\right) \\ &= \beta^{-1} \sum_{i>0} \frac{(-1)^{i+1}}{i} \left(\sum_{n>0} \frac{\beta^n \langle H^n\rangle}{n!}\right)^i \\ &= \beta^{-1} \left[\left(\beta\langle H\rangle + \frac{1}{2}\beta^2\langle H^2\rangle + \sum_{n>2} \frac{\beta^n \langle H^n\rangle}{n!}\right)\right. \\ &\quad -\frac{1}{2}\left(\beta\langle H\rangle + \sum_{n>1} \frac{\beta^n \langle H^n\rangle}{n!}\right)^2 \\ &\quad \left. + \sum_{i>2} \frac{(-1)^{i+1}}{i} \left(\sum_{n>0} \frac{\beta^n \langle H^n\rangle}{n!}\right)^i\right] \\ &= \langle H\rangle + \frac{1}{2}\beta\left[\langle H^2\rangle - \langle H\rangle^2\right] + O(\beta^2) = U - TS \end{aligned}$$

[4]Die Schreibweise $O(\beta^2)$ steht stellvertretend für $\sum_{i=0,1,2,3} O(\beta^2\langle H\rangle^i\langle H^{3-i}\rangle)$, denn β^2 besitzt im Gegensatz zu $\beta^2\langle H\rangle^i\langle H^{3-i}\rangle$ nicht die Einheit einer Energie.

Man kann nun jedem Hamiltonian eine Freie Energie $F_{1,2} = \beta^{-1} \ln\langle e^{\beta H_{1,2}}\rangle = U_{1,2} - TS_{1,2}$ zuordnen, aber wegen

$$\begin{aligned}\langle H^2\rangle - \langle H\rangle^2 &= \langle H_1^2\rangle - \langle H_1\rangle^2 + \langle H_2^2\rangle - \langle H_2\rangle^2 \\ &\quad + \underbrace{\langle H_1 H_2\rangle + \langle H_2 H_1\rangle}_{=\,2\langle H_1 H_2\rangle} - 2\langle H_1\rangle\langle H_2\rangle\end{aligned}$$

folgt im allgemeinen nicht $F = F_1 + F_2$ sondern

$$\begin{aligned}F &= F_1 + F_2 + \beta\left[\langle H_1 H_2\rangle - \langle H_1\rangle\langle H_2\rangle\right] + O(\beta^2) \\ &= F_1 + F_2 - TS_{12} \\ &= U_1 + U_2 - TS \qquad \text{mit } S = S_1 + S_2 + S_{12}\,.\end{aligned} \tag{3.29}$$

Die Freie Energie ist somit eine Summe über die einzelnen Energien und der Entropie, die neben den Anteilen der beiden separaten Hamiltonians, auch die Korrelationen zwischen den Hamiltonians enthält. Inwieweit dieser entropische Anteil weiter separiert werden kann, ist davon abhängig, ob die Freiheitsgrade des Systems entkoppelt werden können. Z.B. kann man in guter Approximation zwischen bond und non-bond Wechselwirkungen unterscheiden [KIP87].

Für einen gegebenen Hamiltonian kann das Hinzufügen oder Variieren einer Wechselwirkung die Natur der erreichbaren Zustände ebenso wie deren Population verändern. Der neue Hamiltonian hat eine entsprechende Freie Energie. Um diesen Sachverhalt etwas konkreter zu fassen, stelle man sich ein Protein in Vakuum oder in Oktanol (mit zugehörigem Hamiltonian H_1) und in Wasser (Hamiltonian H) vor. Damit ist die Wechselwirkung des Proteins mit dem Wasser durch $H_2 = H - H_1$ gegeben. Die mit dieser Wechselwirkung assoziierte Freie Energie F_2 ist im allgemeinen nicht gleich der Differenz $F - F_1$, sondern hängt in bestimmter Weise von allen anderen auftretenden Wechselwirkungen ab.

Folglich kann der Einfluß des Lösungsmittels nur bedingt in elektrostatische Anteile und hydrophoben Effekt aufgeteilt werden. Ebenso problematisch ist die Aufteilung der Wasserstoffbrückenbindung in Lennard-Jones-, elektrostatische- und Lösungsmittelanteile sowie Zusätze etwa in Form eines 10-12-Potentials. Und wenn man sich vor Augen hält, daß in AMBER das 10-12-Potential für die Wasserstoffbrückenbindungen durch ein neu parametrisiertes 6-12-Lennard-Jones Potential ersetzt wurde, so stellt sich die Frage, durch welche Wechselwirkung Wasserstoffbrücken stabilisiert werden, und gleichzeitig, welchen Wert eine solche eindeutige Zuordnung hat?

Nach einem Ergebnis von Yang und Honig [YH95] mit einer älteren Version des CHARMM Kraftfeldes, welche wie AMBER neben dem Lösungsmittel nur Lennard-Jones und elektrostatische Beiträge hat, tragen Wasserstoffbrückenbindungen wenig zur Stabilität von α-Helices bei. Yang und Honig folgern, daß die treibende Kraft der Helixbildung die verstärkte Lennard-Jones-Wechselwirkung in der kompakten Helixkonfiguration zusammen mit dem hydrophoben Effekt sei,

wohingegen Wasserstoffbrücken nur den Entropieverlust der Hauptkette kompensieren. Allerdings führt schon der hydrophobe Effekt zu kompakten Strukturen, also zu einer Reduktion der Konfigurationsentropie, und dieser Anteil müßte in der Analyse genauer berücksichtigt werden. Wir halten die Folgerung von Yang und Honig für kraftfeldspezifisch und für kaum übertragbar.

Die obigen Betrachtungen haben zu Folge, daß es praktisch ausgeschlossen ist, einzelne Parametersätze, wie z.B Lennard-Jones Radien und Potentialtiefen, zwischen verschiedenen Kraftfeldern auszutauschen. Die Parameter jedes Kraftfeldes müssen separat bestimmt werden und dürfen sich nur qualitativ an anderen Kraftfeldern orientieren.

Kapitel 4

Das Kraftfeld PFF01

4.1 Einleitung

Klassische Kraftfelder sind für unterschiedliche Aufgaben entwickelt worden und haben eine lange Tradition in der theoretischen Chemie. Sie werden verwendet, um bestimmte Experimente theoretisch zu erklären und sind entsprechend auf die Reproduktion einzelner Aspekte, wie Vibrationsspektren oder spezifische Wärme, optimiert. In Zuge dieser Optimierung haben verschiedenste Wechselwirkungspotentiale ihren Platz innerhalb spezifischer Kraftfelder eingenommen. Einige Kraftfelder dienen zur Beschreibung kleiner Moleküle, die in Lösung oder in der Gasphase miteinander (und mit dem Lösungsmittelmolekülen) interagieren.

Bei der Proteinstrukturvorhersage befassen wir uns dagegen mit einem einzelnen Molekül aus mehreren hundert Atomen, welches sich stets in ein und demselben Lösungsmittel (Wasser) befindet. Die Beschreibung der Wechselwirkungen kann sich daher ausschließlich an der Stoffklasse der Proteine orientieren und so genauer auf die spezifischen Bedingungen eingehen. Einer der wesentlichen Unterschied zu kleinen Molekülen ist darin zu sehen, daß Proteine innere Strukturbereiche vor dem Lösungsmittel abschirmen können (hydrophober Kern), wohingegen dies kleinen Molekülen unmöglich ist.

In der Arbeitsgruppe von John Moult am CARB ist ein Kraftfeld zur Proteinstrukturvorhersage entwickelt worden, welches bei Proteinfragmenten die native Konfiguration nachbilden konnte. In Simulationen mit Proteinen, die aus mehr als 30 Aminosäuren konstituiert sind, hat sich herausgestellt, daß die einzelnen Kraftfeldterme überarbeitet werden mußten. In mehreren Schritten ist so ein eigenständiges Kraftfeld entstanden, dessen aktuelle Konstituenten sowie die Bestimmung ihre Parameter in diesem Kapitel näher erläutert werden. Momentan trägt das Kraftfeld den schlichten Namen *Protein Force Field 1*, kurz PFF01, und

besitzt folgende Zusammensetzung:

$$\begin{array}{lcl} E & = & E_{lj} + E_{side} + E_{main} + E_{hb} + E_{pse} \\ E_{lj} & : & \text{Lennard-Jones-Wechselwirkung} \\ E_{side} & : & \text{Seitengruppen-Elektrostatik} \\ E_{main} & : & \text{Hauptketten-Elektrostatik} \\ E_{hb} & : & \text{Wasserstoffbrückenpotential} \\ E_{pse} & : & \text{Lösungsmittelbeitrag} \end{array} \tag{4.1}$$

Die Indizierungen sind aus der englischen Bezeichnung (side-/mainchain, hydrogenbond(hb) und protein stabilization energy(pse)) abgeleitet.

PFF01 gehört zu den sogenannten *all-atom* Kraftfeldern, d.h. im Gegensatz zu *united-atom* Kraftfeldern, bei denen Seitengruppen zu einer Kugel oder Ellipsoid zusammengefaßt sind, werden alle schweren Atome (C, N, O und S) des Proteins explizit einbezogen. Wasserstoffatome, die an Kohlenstoff gebunden sind und keine Partialladung tragen, werden in das Kohlenstoffatom integriert. Dieses Kohlenstoffatom entspricht somit einem Methyl. Nur wenn das Wasserstoffatom einen polaren Bindungspartner hat und eine eigene Partialladung besitzt, wird es explizit berücksichtigt.

Jedem Atom i wird ein Potentialtyp $pt(i)$ zugeordnet. Der Potentialtyp soll eine eindeutige Zuordnung der Parameter des Atoms i, wie Ladung, van-der-Waals-Radius und Lösungsmittelparameter, ermöglichen. An dieser Zuordnung sind im Laufe der Kraftfeldoptimierung mehrfach kleinere Veränderungen vorgenommen worden. In PFF01 unterteilen wir die 5 Elemente – Kohlenstoff, Sauerstoff, Stickstoff, Schwefel und Wasserstoff – in elf Potentialtypen. In Anhang C sind die Zuordnungen der Potentialtypen zu den Atomen der 20 Aminosäuren und eine Liste aller atomspezifischen Parameter zu finden. Dort sind auch Abbildungen der 20 verschiedenen Seitengruppen vorhanden. An den Potentialtypbezeichnungen $pt(i)$ ist an dieser Stelle nur wichtig, daß der erste Buchstabe das jeweilige Element wiedergibt. Ein *cme* ist daher ein Kohlenstoffatom und ein $n2$ ein Stickstoffatom.

Im Folgendem werden die einzelnen Kraftfeldanteile und die Bestimmung der zugehörigen Parameter besprochen.

4.2 Das Lennard-Jones Potential

Das Lennard-Jones Potential besteht aus einem attraktiven Term ($1/r^6$), der durch wechselseitig induzierte Dipol-Dipol-Wechselwirkung entsteht, und einem stark repulsiven Potential, welches aus dem Pauli-Prinzip hergeleitet wird und näherungsweise durch ein $1/r^{12}$ beschrieben wird. Die Implementierung innerhalb des CARB Kraftfeldes lautet

$$V_{ij}(r) = A_{ij} r^{-12} - B_{ij} r^{-6} \ ,$$

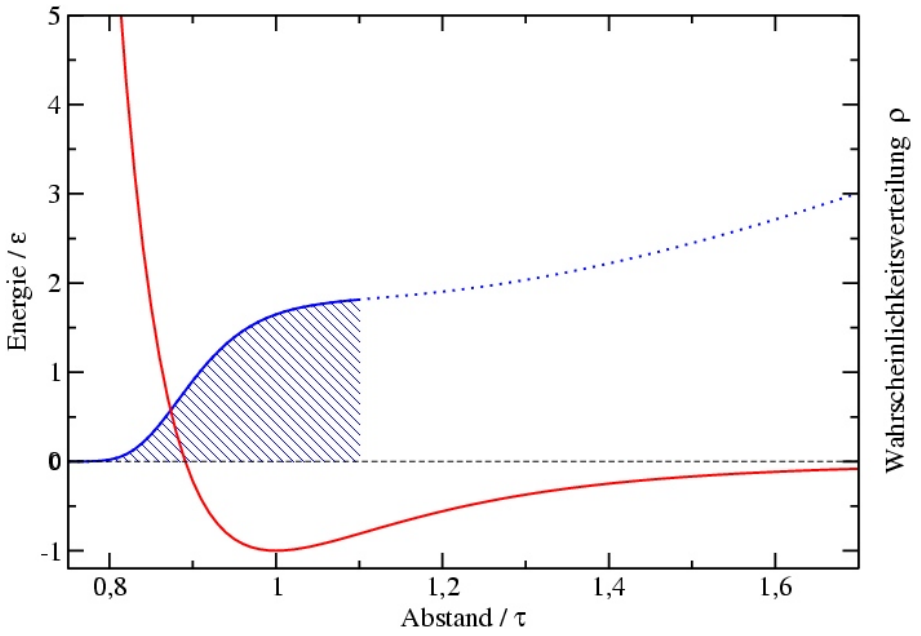


Abbildung 4.1: Verlauf des Lennard-Jones Potential (Gleichung 4.2, rot) und die Verteilungsfunktion $\rho(r)$ (Gleichung 4.10, blau) mit $\epsilon = 0.3\,kcal\,mol^{-1}$ und $T = 300K$.

welche äquivalent ist zu

$$V(r) = \epsilon \left[\left(\frac{\tau}{r} \right)^{12} - 2 \left(\frac{\tau}{r} \right)^{6} \right] . \tag{4.2}$$

Das Minimum dieses Potentials liegt bei $r = \tau$ mit $V(\tau) = -\epsilon$. Die Werte für τ und ϵ sind atomspezifisch ($\epsilon \equiv \epsilon_{ii}$, $\tau \equiv \tau_{ii}$) und $\frac{1}{2}\tau$ wird van-der-Waals- oder Lennard-Jones Radius des jeweiligen Atoms bezeichnet. Für die Wechselwirkung unterschiedlicher Elemente errechnet man die Parameter aus

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \tag{4.3}$$

$$\tau_{ij} = \sqrt{\tau_{ii}\tau_{jj}} \tag{4.4}$$

Oft wird auch eine modifizierte Formulierung benutzt (z.B. in Kraftfeldern wie AMBER und CHARMM):

$$V(r_{ij}) = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{4.5}$$

$$\sigma_{ij} = \frac{1}{2} \left(\sigma_{ii} + \sigma_{jj} \right) \tag{4.6}$$

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \, . \tag{4.7}$$

Hier gibt σ die Position der Nullstelle des Lennard-Jones Potentials und nicht die des Minimums an, welches (unabhängig von den beteiligten Atomen) bei $\sqrt[6]{2}\sigma$ liegt und durch $V(2^{-1/6}\sigma) = -\epsilon$ gegeben ist.

Sei $\tau_{jj} = (1+\eta)\tau_{ii}$, dann gilt

$$\begin{aligned}\sqrt{\tau_{ii}\tau_{jj}} &= \sqrt{1+\eta}\sqrt{\tau_{ii}} \\ &\leq (1+\frac{1}{2}\eta)\sqrt{\tau_{ii}} \\ &= \frac{1}{2}(1+(1+\eta))\sqrt{\tau_{ii}} = \frac{1}{2}(\sqrt{\tau_{ii}}+\sqrt{\tau_{jj}}) \ .\end{aligned}$$

Das heißt, im CARB Kraftfeld und infolgedessen auch in PFF01 sind die Gleichgewichtsabstände verschiedener Atomtypen geringer als etwa in $CHARMM$ oder $AMBER$. Dieser Unterschied ist typischerweise sehr gering. Für den Kohlenstoff-Sauerstoff-Abstand folgt z.B. mit den Gleichgewichtsabständen von 4.1Å für Kohlenstoff und 3.1Å für Sauerstoff: $\sigma_{CO} = \frac{1}{2}(4.1\text{Å}+3.1\text{Å}) = 3.6\text{Å}$ und $\tau_{CO} \approx 3.57\text{Å}$.

Die Formel 4.2 ist für die Simulation am Computer gut geeignet, da der Term proportional r^{-12} das Quadrat des r^{-6}-Termes ist. Man bezeichnet diesen Ansatz auch als "Soft-Sphere-Potential", da ein geringfügiger Überlapp der Orbitale erlaubt ist. Das "Hard-Sphere-Potential" hat dagegen die Form

$$V_{HS}(r) = \begin{cases} 0 & \text{wenn } r > \tau \\ \infty & \text{sonst} \end{cases} \ . \tag{4.8}$$

Beiden Potentialen ist gemein, daß sie den vollständigen Kollaps der Struktur verhindern. Der wesentliche Unterschied beider Potentiale ist in der Existenz des Lennard-Jones-Minimums bei $r = \tau$ zu sehen. Für ungebundene Atome bzw. Moleküle führt dies zur Bildung von kompakten Strukturen, sogenannten Lennard-Jones-Clustern.

In den Kraftfeldern CHARMM und AMBER, welche für Molekulardynamik Simulationen entwickelt wurden, wird das Lennard-Jones Potential verwandt, wohingegen die Arbeitsgruppe von J. Moult dem Hard-Sphere-Potential den Vorzug gibt. Das Hard-Sphere-Potential hat für die Computersimulation den Vorteil, daß es nur bezüglich der nächsten Nachbarn ausgewertet werden muß, wodurch der numerische Aufwand sinkt. Auf der anderen Seite kann ein nicht-differenzierbares Potential nicht für Molekulardynamik Simulationen herangezogen werden. Dies gilt auch für anderen Simulationsverfahren, welche die Existenz eines Gradienten voraussetzen, oder wie das stochastische Tunneln (STUN [WH99]) kurzzeitig in Bereiche unphysikalisch großer Energien eindringen.

Die Entscheidung, welches der beiden Potentiale in einem Kraftfeld verwendet wird, ist solange rein technischer Natur, wie die Potentialtiefe des Lennard-Jones-Potentials so gering gewählt wird, daß sie keinen wesentlichen Beitrag zur Faltungsenergie liefert. Inwieweit diese Annahme gerechtfertigt ist, ist unklar.

Wie erwähnt, wurde in AMBER ein Wasserstoffbrücken stabilisierendes 10-12-Potential durch neue Lennard-Jones Parameter ersetzt, wodurch das Potentialminimum zur Stabilisierung der Sekundärstrukturelemente beiträgt. Diese These wird von Molekulardynamik Simulationen des CHARMM Kraftfeldes unterstützt, die ebenfalls der Lennard-Jones Wechselwirkung eine starke stabilisierende Bedeutung zuordnen [YH95]. Nach unserer Auffassung trägt die zugrundeliegende zeitliche Inhomogenität der Dipolmomente wenig zur Stabilisierung von Wasserstoffbrückenbindungen bei. In PFF01 ist die zentrale Größe des Lennard-Jones Potentials der Gleichgewichtsabstand τ, deren Werte für die einzelnen Atome im folgenden bestimmt wird.

Der Gleichgewichtsabstand soll jedoch nicht aus experimentell bestimmten Energien abgeleitet werden, sondern aus den nativen Strukturen der Proteindatenbank extrahiert werden. Aus den dort eingetragenen Koordinaten der Atome läßt sich die Verteilung der Atomabstände bestimmen, wobei keine 1-2, 1-3 und 1-4 gebundenen Atome berücksichtigt werden. Für die Verteilung des Abstands r zweier Atome i und j, deren Koordinaten mit $\vec{r}_i$ und $\vec{r}_j$ bezeichnet seien, gilt

$$\rho(r) = \mathcal{Z}^{-1} \int exp\left[-\beta H(\vec{q})\right] \delta\left(|\vec{r}_i - \vec{r}_j| - r\right) \, d\vec{q} \tag{4.9}$$

Für kurze Abstände wird der r^{-12} Anteil des Lennard-Jones-Potentials alle anderen auftretenden Kräfte dominieren. Die Wahrscheinlichkeit, in der nativen Struktur zwei Atome im Abstand r zueinander aufzufinden, ist dann durch

$$\rho(r) \sim r^2 \; e^{-\beta\epsilon \left[\left(\frac{\tau}{r}\right)^{12} - 2\left(\frac{\tau}{r}\right)^{6}\right]} \tag{4.10}$$

gegeben[1]. Da strukturaufgeklärte Proteine zugrundegelegt werden, ist β die Temperatur der Experimente, also circa $300K$. Die Gültigkeit dieser Formel beschränkt sich auf den abstoßenden Bereich $r \lesssim \tau$. Für $r = \frac{3}{4}\tau$ ist $\rho(r)$ praktisch auf Null abgefallen. Dies führt zu einem sehr kleinen Abstandsbereich, dessen Daten zur Bestimmung der Lennard-Jones-Parameter herangezogen werden können. Diese Datenmenge erscheint nicht ausreichend, um für jedes Element 2 Parameter (ϵ, τ) hinreichend genau zu bestimmen. Unter der Annahme, daß der Potentialtiefe des Lennard-Jones-Potentials in der Proteinfaltung keine große Bedeutung zukommt, haben wir die Potentialtiefe aus einer anderen Quelle übernommen und die Verteilung 4.10 nur bezüglich des Gleichgewichtsabstand an die Strukturdaten angepaßt.

In die Kohlenstoffatome der hydrophoben Seitengruppen (ALA, LEU, ILE, VAL) werden in vorliegenden Kraftfeld die gebundenen Wasserstoffatome mit

[1]Der Faktor r^2 ergibt sich nach Übergang zu Kugelkoordinaten und Integration über den Winkelanteil. Damit ist die Verteilung im $\mathbb{R}^3$ nicht normiert, sondern nur noch in einem endlichen Volumen $V \subset \mathbb{R}^3$.

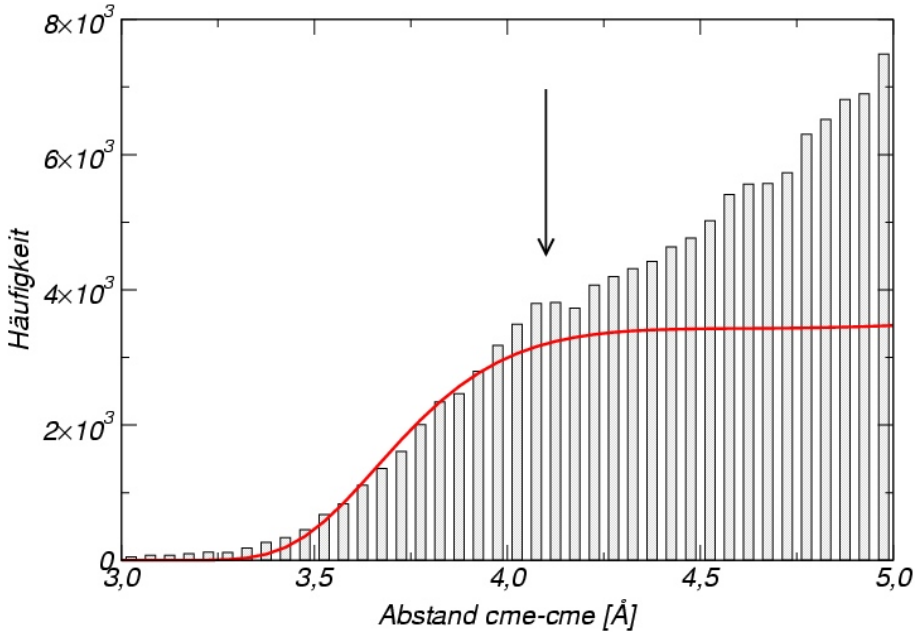


Abbildung 4.2: Vergleich der natürlichen Abstandsverteilung mit dem Ergebnis der Parameteroptimierung für 4.10 ($T = 300K$, $\epsilon = 0.3\,kcal\,mol^{-1}$). Die Markierung zeigt die Lage des Gleichgewichtsabstandes $\tau = 4.1$Å an.

eingerechnet. Somit entspricht jedes (sp^3-hybridisierte) Kohlenstoffatom des Kraftfeldes näherungsweise einem Methanmolekül. Für dieses Molekül existieren experimentelle Daten für die Lennard-Jones Potentialtiefe von $\epsilon \approx 0.3\,kcal\,mol^{-1}$ [MG97], welche wir für alle Atomtypen übernommen haben.

Ausgehend von sp^3-hybridisiertem Kohlenstoff (Potentialtyp *cme*), dem bei weitem häufigsten Potentialtyp, haben wir die Gleichgewichtsabstände der unterschiedlichen Elemente bestimmt. Als Grundlage dienen die 138 Proteinstrukturen M^{138}, die in Anhang B.1 aufgelistet sind. Die Abstandsverteilung für Kohlenstoff innerhalb der Proteinstrukturen zusammen mit der Verteilung 4.10 für den optimalen Wert $\tau = 4.1$Å ist in Abb. 4.2 wiedergegeben. Der experimentelle Wert der Gleichgewichtslage für Methan liegt bei 3.81Å. Die (sp^3-hybridisierten) Kohlenstoffatome dieses Kraftfeldes, welche die gebundenen Wasserstoffatome integriert haben, sind somit etwas größer als freies Methan. Auf der anderen Seite wissen wir aus Simulationen mit dem ursprünglichen CARB Kraftfeld, daß ein Gleichgewichtsabstand von 4.6Å zu groß ist. Das Ergebnis der Anpassung ist somit durchaus konsistent mit dem Experiment und unseren Erfahrungen mit den Parametern des CARB Kraftfeldes.

Für die Sauerstoff- und Stickstoffatome sind die natürlichen Verteilungen der $C-N$ bzw. $C-O$ Abstände und nicht die der $N-N$ bzw. $O-O$ Atome als Grundlage der Parameterbestimmung genommen worden, um elektrostatische Einflüsse

auf die Lennard-Jones Parameter zu vermeiden. Der Gleichgewichtsabstand für N und O ergibt sich nach 4.4 gemäß $\tau_{XX} = \tau_{CX}^2/\tau_{CC}$ mit $X = N, O$. Ähnlich ist es bei den Parametern für Schwefel und Wasserstoff, die ebenfalls an die Abstände zu Kohlenstoffatomen angepaßt wurden. Hier liegt der Grund nicht in der elektrostatischen Wechselwirkung, sondern an der statistisch nicht ausreichenden Datenmenge an Wasserstoff-Wasserstoff und Schwefel-Schwefel Kontakten. In Tabelle 4.1 sind die Resultate der Parameteroptimierung der einzelnen Atomtypen angegeben. Die Verteilungen der nativen Strukturen und der theoretischen Verteilung 4.10 sind in Abb. 4.3 dargestellt. Der Lennard-Jones Gleichgewichtsabstand für das Wasserstoffatom der Hauptkette wurde nachträglich verändert, um die Geometrie der Helices genauer wiedergeben zu können.

Element	Potentialtyp	τ [$\AA$]
Kohlenstoff[2](sp^3-hybridisiert)	cme, cp	4.10
Kohlenstoff (in Ringstruktur)	cr	3.28
Stickstoff	n1, n2, n3	3.55
Sauerstoff	o1, o2	3.10
Schwefel	s	3.80
Wasserstoff (Seitenkette)	h	1.95
Wasserstoff (Hauptkette)	hn	2.25

Tabelle 4.1: Lennard-Jones Gleichgewichtsabstand der einzelnen Atomtypen

4.3 Elektrostatik

Eine der schwierigsten Aspekte der Energiebestimmung einer Proteinkonfiguration ist der Einfluß der dielektrischen Umgebung auf die elektrostatische Wechselwirkung. In biomolekularen Kraftfeldern wird die elektrostatische Wechselwirkung durch einen statischen Satz von Punktladungen approximiert. Ziel dieses Abschnittes ist die Berechnung der elektrostatischen Energie dieser Punktladungen. Polarisationseffekte werden dabei nicht auf mikroskopischer Ebene beschrieben, sondern als über ein Gebiet "verschmierte" Reaktion des Mediums, oft in Form einer ortsabhängigen Dielektrizitätsfunktion.

Die auf dieser Abstraktionsebene exakte elektrostatische Energie läßt sich aus der Lösung der Poisson-Boltzmann-Gleichung gewinnen. Diese wird im nächsten Abschnitt behandelt. Darauf folgt die Betrachtung einiger Näherungen und Einzelaspekte, die für die Proteinstrukturvorhersage wichtig sind, sowie die Beschreibung eines speziellen Potentials der Wasserstoffbrückenbindungen.

[2]In die Kohlenstoffatome des Kraftfeldes sind die gebundenen Wasserstoffatome integriert.

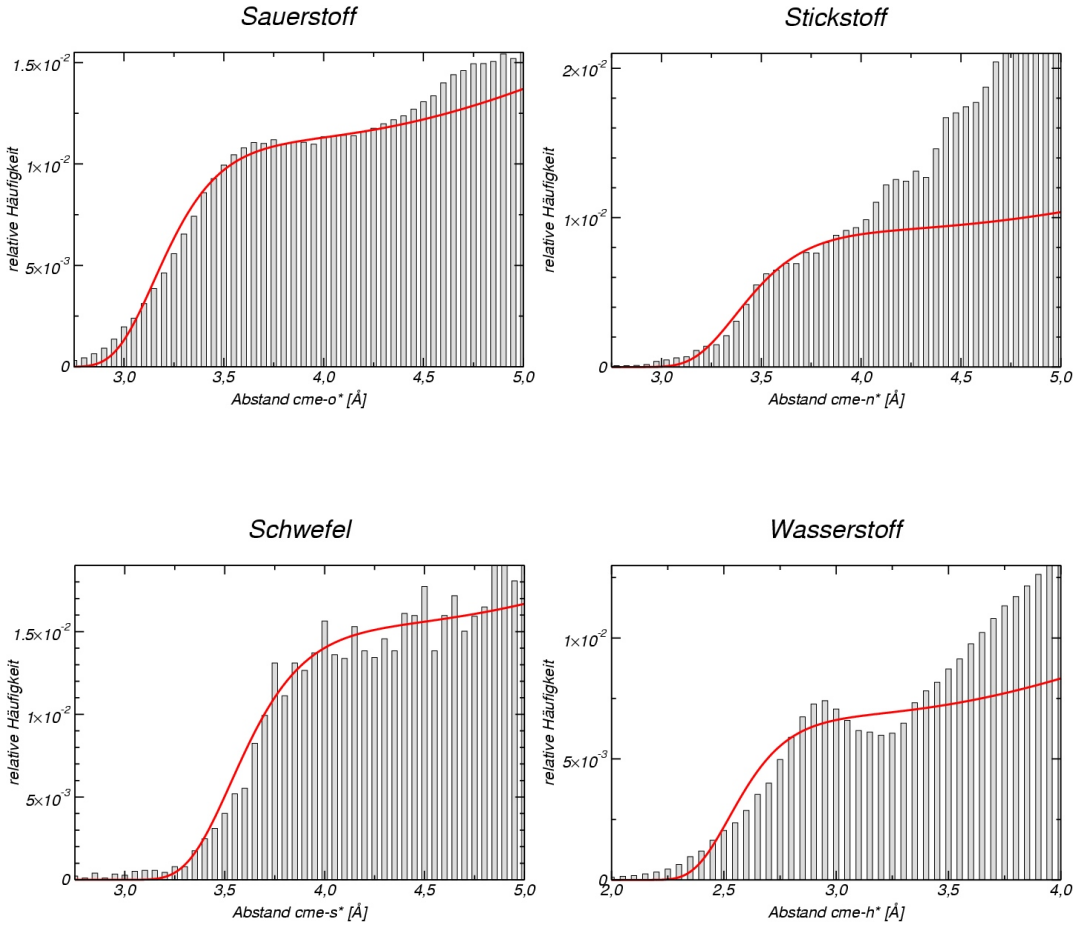


Abbildung 4.3: Natürliche Abstandverteilung zusammen mit dem Ergebnis der Parameteranpassung für 4.10 ($T = 300K$, $\epsilon = 0.3\,kcal\,mol^{-1}$).

4.3.1 Die Poisson-Boltzmann-Gleichung

Die fundamentale Relation der Elektrostatik ist die *Poisson*-Gleichung

$$\epsilon_0 \nabla^2 \phi(\vec{r}) = -\rho(\vec{r}) \ , \tag{4.11}$$

welche die räumliche Variation des Potentials ϕ an der Position $\vec{r}$ mit der Ladungsdichte ρ in Beziehung setzt. Wenn die Ladungsverteilung durch einen Satz von Punktladungen gegeben ist, so ist die Lösung der Poisson-Gleichung durch das Coulomb-Gesetz gegeben:

$$\phi(\vec{r}) = \frac{1}{4\pi\epsilon_0} \sum_i \frac{q_i}{|\vec{r} - \vec{r}_i|} \ . \tag{4.12}$$

Hierbei notieren $\vec{r}_i$ die Position und q_i die Stärke der i-ten Punktladung.

Sind alle Ladungen eines System explizit gegeben, so lassen sich alle elektrostatischen Wechselwirkungen durch das Coulomb-Gesetz beschreiben. In vielen Fällen sind die Details der Wechselwirkungen nicht von Bedeutung, und die Poisson-Gleichung kann umgeschrieben werden, so daß die unbedeutenden Details herausfallen. Wenn eine Region des Systems "gleichmäßig" auf die Anwesenheit eines elektrischen Feldes $\vec{E}$ reagiert, ist diesem Bereich eine Suszeptibilität χ zugeordnet, womit die Polarisationsdichte durch $\vec{P} = \chi\vec{E}$ gegeben ist. Wenn das Gesamtsystem eine homogene Suszeptibilität besitzt, kann nach Einführung der relativen Dielektrizitätskonstanten $\epsilon = 4\pi\chi + 1$ die Poisson-Gleichung und das Coulomb-Gesetz zu

$$\epsilon_0\epsilon \nabla^2 \phi(\vec{r}) = -\rho(\vec{r}) \ , \quad \phi(\vec{r}) = \frac{1}{4\pi\epsilon_0\epsilon} \sum_i \frac{q_i}{|\vec{r} - \vec{r}_i|} \tag{4.13}$$

umgeschrieben werden. Wenn die Dielektrizitätskonstante räumlich variiert, so ist das Coulomb-Gesetz nicht länger gültig, und die Poisson-Gleichung geht über in

$$\epsilon_0 \nabla \cdot \epsilon(\vec{r}) \nabla \phi(\vec{r}) = -\rho(\vec{r}) \ . \tag{4.14}$$

In dem für uns interessanten Szenario eines Proteins in wässeriger Lösung liegt ϵ zwischen 3 und 4 für Punkte innerhalb des Proteins, und im Wasser gilt $\epsilon = 80$. Darüber hinaus muß die Anwesenheit beweglicher Ionen (im Wasser gelöster Salze) berücksichtigt werden. Sei n_- die Ladungsdichte der Kationen und n_+ der Anionen, dann lautet die Poisson-Gleichung

$$\epsilon_0 \nabla \cdot \epsilon(\vec{r}) \nabla \phi(\vec{r}) = -4\pi \Big[\rho(\vec{r}) + n_+(\vec{r}) - n_-(\vec{r}) \Big] \ . \tag{4.15}$$

In einem *Mean-field* Ansatz kann die Ionenverteilung als proportional zum Boltzmann-Faktor der Energie der Ionen mit Ladung $\pm q$ im *Mean-field* $\phi(\vec{r})$ angesetzt werden:

$$n_\pm \sim \exp[-\beta(\pm q)\phi(\vec{r})] \ . \tag{4.16}$$

Damit geht Gleichung 4.15 in die *Poisson-Boltzmann-Gleichung* über:

$$\epsilon_0 \nabla \cdot \epsilon(\vec{r}) \nabla \phi(\vec{r}) - \kappa^2(\vec{r}) \sinh \phi(\vec{r}) = -\rho(\vec{r}) \; , \tag{4.17}$$

mit dem Debye-Hückel Abschirmparameter κ, der mit der Ionenstärke I verknüpft ist. Dieser Abschirmparameter ist gleich (N_A : Avogadro Konstante)

$$\sqrt{\frac{8\pi N_A e^2 I}{1000\, \epsilon k_B T}} \sim \sqrt{\beta I} \; .$$

Für niedrige lokale Potentiale $\beta q \phi(\vec{r}) \ll 1$ (Gleichung 4.16) kann man die Poisson-Boltzmann-Gleichung entwickeln und erhält so die *linearisierte Poisson-Boltzmann-Gleichung*

$$\epsilon_0 \nabla \cdot \epsilon(\vec{r}) \nabla \phi(\vec{r}) - \kappa^2(\vec{r}) \phi(\vec{r}) = -\rho(\vec{r}) \; , \tag{4.18}$$

Diese Gleichung inkooperiert elektrische und Dipol-Polarisation in ϵ und Abschirmung durch Ionen in κ.

Setzen wir für $\epsilon = 1$ eine Ladung in den Ursprung des Koordinatensystems, $\rho = q\delta(\vec{r})$, so ist Lösung der linearisierten Poisson-Boltzmann-Gleichung gegeben durch

$$\phi(\vec{r}) = \frac{1}{4\pi\epsilon_0} \frac{q\, e^{-\kappa|\vec{r}|}}{|\vec{r}|} \; . \tag{4.19}$$

Das Coulomb-Potential der Ladung q wird unter Anwesenheit mobiler Ionen durch eine exponentielle Abschirmung abgeschwächt und geht im Limes verschwindender Salzkonzentration $\kappa \to 0$ in das reine Coulomb-Potential über. Damit ist die elektrostatische Wechselwirkung zwischen Ladungen für Entfernungen größer als die Abschirmlänge κ^{-1} irrelevant. Unter physiologischen Bedingungen, also Salzkonzentrationen von ca. $100mM$, ist die Abschirmlänge von der Größenordnung $\kappa^{-1} \approx 1\, nm$.

Die numerische Lösung der Poisson-Boltzmann Gleichung kann über finite Differenzen Methoden erfolgen. Betrachten wir ein kubisches Gitter mit einem Gitterabstand h. Unter Anwendung des *Mittelwertsatzes der Funktionentheorie* gilt für die Lösung der PB Gleichung am Gitterpunkt i

$$\phi_i = \frac{\sum\limits_{j \in N(i)} \epsilon_{ij} \phi_j + q_i/h}{\sum\limits_{j \in N(i)} \epsilon_{ij} + (h\kappa_i)^2} \tag{4.20}$$

Dabei bezeichnet $N(i)$ die nächsten Nachbarn des Gitterplatzes i, ϵ_{ij} die Dielektrizitätskonstante am Mittelpunkt zwischen i und j und q_i die Ladung in i. Der

Wert für κ_i ist dann von Null verschieden, wenn i innerhalb des Bereiches liegt, in dem sich Ionen aufhalten können. Dies ist ähnlich der Definition der SAS-Oberfläche des Wassers (Seite 80) zu verstehen, nur mit einem Radius von 2Å. Das bedeutet, wenn sowohl das Lösungsmittel als auch die gelösten Satze als Kontinuum behandelt werden, existiert um das Protein ein dünner Bereich "reinen" Wassers, der sogenannten *Stern Layer*, der insbesondere dann interessant wird, wenn das Protein von einer sphärischen Geometrie abweicht und kleine Taschen bildet.

Obige Formel wird im Zuge einer Fixpunktiteration zur Lösung der Poisson-Boltzmann-Gleichung herangezogen [NH91]. Eine sehr effiziente Lösungsmethode ist in das Programm APBS (Adaptive Poisson-Boltzmann Solver [BSJ+01]) integriert, welches im Rahmen wissenschaftlicher Arbeiten frei erhältlich ist[3] und ein Interface zur Verarbeitung von Proteinstrukturen im Format der Proteindatenbank PDB und zur Darstellung des Ergebnisses mit den Visualisierungsprogrammen VMD und MOLMOL hat. Die Geschwindigkeit des APBS-Programmes ist dennoch zu gering, als daß es für Simulationen eingesetzt werden könnte. Für Proteinsimulationen ist man daher weiterhin auf Näherungen angewiesen.

4.3.2 Das generalisierte Born-Modell

Ausgangspunkt des Born-Modells ist ein einzelnes Ion in Lösung. Das Ion wird als Kugel mit Radius a und einer Punktladung q im Mittelpunkt modelliert, welches von einem Medium mit einer Dielektrizitätskonstanten ϵ_w umgeben ist. Das Kugelinnere wird später dem Protein entsprechen und erhält daher die Dielektrizitätskonstante ϵ_p. Aufgrund der sphärischen Symmetrie reduziert sind dieses Modell auf ein eindimensionales Problem und das elektrostatische Potential ergibt sich sehr schnell aus der ersten Maxwell-Gleichung in Integralform und der Stetigkeitbedingung bei $r = a$ zu

$$\phi(r) = \begin{cases} \frac{q}{4\pi\epsilon_w\, r} & , r \geq a \\ \frac{q}{4\pi\epsilon_p\, r} - \frac{q}{4\pi\, a}\left(\frac{1}{\epsilon_p} - \frac{1}{\epsilon_w}\right) & , r < a \end{cases}$$

Der zweite Term für $r < a$ ist das sogenannte Reaktionspotential. Die Wechselwirkungsenergie der Punktladung mit diesem Potential ist die Hälfte des Produktes der Ladung q und dem Potential, wobei der Faktor 1/2 auftritt, da es sich hier nicht um eine Ladung in einem äußeren Feld handelt, sondern das Reaktionspotential wird durch die Ladung selbst induziert. Die Polarisationenergie der Solvatation eines sphärischen Ions ist somit gegeben durch

$$E_{Born} = -\frac{1}{2}\, q\, \phi(r) = -\frac{q^2}{8\pi a}\left(\frac{1}{\epsilon_p} - \frac{1}{\epsilon_w}\right) . \tag{4.21}$$

[3] APBS-homepage: *http : //agave.wustl.edu/apbs/*

Diese Energie wird auch *Born Energie* genannt.

Für die Wechselwirkung zwischen zwei gelösten Ionen betrachtet man zunächst die beiden Sonderfälle, bei denen die Ionen unendlich weit entfernt sind (reiner Coulomb-Fall) bzw. sich am gleichen Ort befinden (reiner Born-Fall). Für die Polarisationenergie einer beliebigen Verteilung von N Ladungen wurde als Interpolation zwischen den beiden Extrema das sogenannte *generalisierte Born-Modell* vorgeschlagen [STHH91]

$$\begin{aligned} E_{GB} &= -\frac{1}{8\pi}\left(\frac{1}{\epsilon_p}-\frac{1}{\epsilon_w}\right)\sum_{i,j=1}^{N}\frac{q_i\,q_j}{f_{GB}(r_{ij})} \\ f_{GB}(r_{ij}) &= \sqrt{r_{ij}^2 + a_i a_j\, e^{-r_{ij}^2/(4a_i a_j)}} \\ &\doteq \begin{cases} r_{ij} & , r_{ij} \gg \sqrt{a_i a_j} \quad \text{Coulomb-Fall} \\ \sqrt{a_i a_j} & , r_{ij} \ll \sqrt{a_i a_j} \quad \text{Born-Fall} \end{cases} \end{aligned}$$

Der *effektive Born Radius* a_i des Atoms i ist definiert als derjenige Radius, der eingesetzt in die Born Gleichung 4.21 die elektrostatische Energie des Moleküls angibt, wenn alle bis auf die i-te Ladung abgeschaltet sind. Dieser Wert ist folglich abhängig von der Struktur des Proteins und der Position der Ladung im Protein. Es existieren verschiedene Methoden, den Radius näherungsweise zu bestimmen, ohne die Poisson-Boltzmann-Gleichung explizit lösen zu müssen. Auf diese soll hier jedoch nicht weiter eingegangen werden.

Wesentlich sind hier zwei Punkte. Zum einen reproduziert das generalisierte Born Modell, bei optimaler Wahl der effektiven Radien a_i, die Resultate der Poisson-Boltzmann Gleichung recht gut [OCB02] und zum anderen existiert in Bezug auf das Proteinstrukturproblem zumindest ein Vergleich zwischen den Resultaten des generalisierten Born Modells und der Verwendung des Coulomb-Gesetzes mit angepaßten Dielektrizitätskonstanten [Fri02]. Demnach ist der Gesamtenergiebeitrag des generalisierte Born-Modells (mit recht genauer Bestimmung der effektiven Radien a_i bei 653 niederenergetischen Strukturen von 3ICB) proportional zur Gesamt-Coulomb-Energie, bei einem Korrelationkoeffizienten von 0.94. Diese Korrelation war für eine abstandsabhängige Dielektrizitätskonstante der Form $\epsilon \sim r$ deutlich geringer. Wir gehen daher davon aus, das die Verwendung des Coulomb-Gesetzes mit angepaßten Dielektrizitätskonstanten im Bereich der Proteinstrukturvorhersage eine verhältnismäßig gute Approximation des generalisierten Born-Modells ist.

Es ist bekannt, daß die effektiven Born Radien a_i recht genau bestimmt werden müssen, um eine gute Approximation der Lösung der Poisson-Boltzmann zu bieten [OCB02]. Allerdings ist die genaue Ermittlung der effektiven Born Radien mit einem hohen Zeitaufwand verbunden, weil die genaue Geometrie des Proteins berücksichtigt werden muß. Wir haben daher bislang der alten Form des CARB-Kraftfeldes folgend auf die Verwendung des Coulomb-Gesetzes mit gruppenspezifischen Dielektrizitätskonstanten zurückgegriffen.

4.3.3 Ionisation einzelner Seitengruppen

Einige Seitengruppen können Säure-Base-Reaktionen eingehen

$$HA + H_2O \rightleftharpoons H_3O^+ + A^- \ .$$

Die abkürzende Schreibweise $HA \rightleftharpoons H^+ + A^-$ ist irreführend, denn die Beteiligung des Wassers ist eine notwendige Bedingung für die Reaktion. Da man die Konzentration des Wassers (bei verdünnten Lösungen) als konstant annimmt, muß ihr Wert nicht in die Dissoziationskonstanten

$$K_a = \frac{[H_3O^+][A^-]}{[HA]} \ .$$

aufgenommen werden. In isolierter Form hat nur Histidin eine pK_a-Wert ($= -\log K_a$) von 6.04, der dem Wert der Dissoziationskonstanten des Wassers[4] ähnlich ist. Somit koexistiert Histidin, sofern es im Kontakt mit Wasser steht, in protonierten und deprotonierten Konfigurationen. Alle anderen Aminosäuren haben (isoliert) eine eindeutige Struktur. Asparagin- und Glutaminsäure liegen stets als COO^- vor und sowohl Lysin als auch Arganin sind protoniert. Innerhalb eines Kraftfeldes müssen letztere vier Aminosäuren nur durch eine chemische Konfiguration wiedergegeben werden. Bei Histidin beschränkt sich das CARB Kraftfeld auf die protonierte Form, wohingegen einige andere Kraftfelder die deprotonierte Form bevorzugen.

Es ist nicht zu erwarten, daß die geladenen Seitengruppen in der (de-)protonierten Konfiguration regelmäßig im Inneren des Proteins anzutreffen sind. Viel wahrscheinlicher ist, daß im Proteinkern die Seitengruppen zumeist in einem neutralen Zustand vorkommen. Diese Annahme wird für Asparaginsäure durch eine SCF Rechnung unterstützt [KM01].

Das Säure-Base-Gleichgewicht der Seitengruppen wird in keinem der in einem vorigen Kapitel vorgestellten Kraftfelder berücksichtigt. Somit werden elektrostatische Wechselwirkungen verborgenerer Seitenketten möglicherweise falsch wiedergegeben.

4.3.4 Ein einfaches Bild des Proteins in wässeriger Lösung

Wie im 3.2.2 Kapitel beschrieben, kann die Poisson-Gleichung für einfache Geometrien analytisch gelöst werden. Approximiert man die Proteinkonfiguration durch eine Kugel, so induziert eine Ladung an der Proteinoberfläche eine Spiegelladung umgekehrten Vorzeichens (Gleichung 3.17). Diese Spiegelladung schirmt die sie induzierende Ladung weitestgehend ab. Diese Abschirmung geht für im Proteininneren liegende Ladungen verloren, da der Abstand zur Spiegelladung q'

[4] Der pH-Wert destillierten Wassers ist 7.0. In biologischen Systemen liegt der pH-Wert oft im leicht sauren (< 7). Eine wichtige Ausnahme ist hierbei das Blut mit $pH \approx 7.4$.

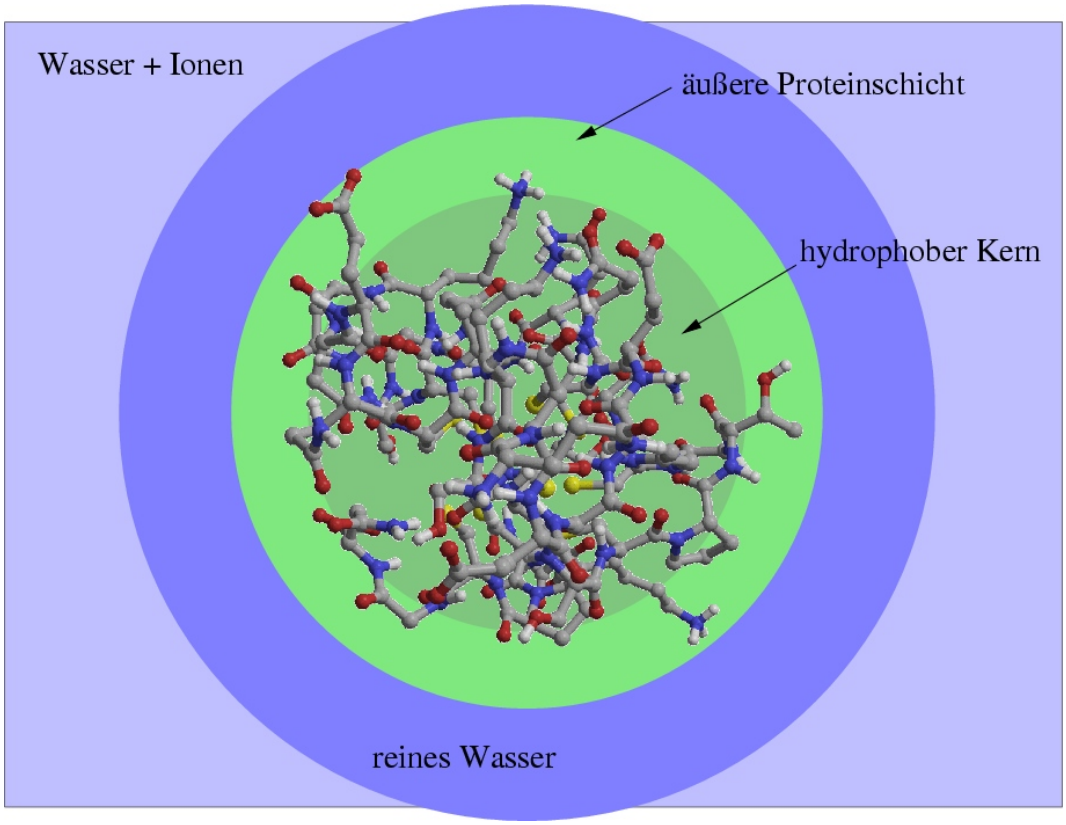


Abbildung 4.4: Die unterschiedlichen Schichten eines Proteins in wässeriger Lösung

mit $1/s$ wächst, wenn s den Abstand der Ladung q zum Kugelmittelpunkt mißt. Dennoch gibt es eine geringe attraktive Wechselwirkung zwischen den beiden Ladungen, die in Richtung der Proteinoberfläche weist. Das heißt, daß die elektrostatische Energie zweier Ladungen an der Proteinoberfläche niedriger liegt, als die von Ladungen im Protein[5].

In einem groben Modell für die elektrostatische Wechselwirkung kann man sich das Protein als eine hydrophobe Kugel umgeben von einer Schicht vorstellen, in welcher sich die geladenen Seitengruppen sammeln. Das Protein ist von einem Mantel aus Wassermolekülen umgeben, und schwimmt in einer Lösung verschiedener Salze und anderer Stoffe, mit denen das Protein nur gelegentlich in Berührung kommt[6]. Diese Modellvorstellung ist in Abb. 4.4 illustriert.

Diesem Bild folgend, unterscheiden wir elektrostatische Wechselwirkungen an

[5]Die "Ausschmierung" der Wassermoleküle zu einem kontinuierlichem Medium mit einer homogenen Dielektrizitätskonstanten vernachlässigt die Freie Energie des Wassers, welche von sich aus im Rahmen des hydrophoben Effektes die Ladungen des Proteins an die Oberfläche zieht.

[6]Einige Proteine sind für die Speicherung von Ionen oder Molekülen zuständig, und das angeführte Modell ist dann nicht mehr anwendbar.

der Proteinoberfläche von denen im Kern des Proteins. Die vorherrschende elektrostatische Wechselwirkung im Proteininneren sind Wasserstoffbrückenbindungen. Die Oberflächenelektrostatik ist dominiert durch die Wechselwirkungen der Seitengruppen.

4.3.5 Elektrostatik des CARB-Kraftfeldes

Die elektrostatische Wechselwirkung wird im CARB Kraftfeld durch das Coulomb-Gesetz mit gruppenspezifischen Dielektrizitätskonstanten beschrieben, wobei Wasserstoffbrücken gänzlich durch die Elektrostatik der Hauptkette beschrieben werden (Gleichung 3.21).

$$E_{side} + E_{main} = \frac{1}{4\pi\epsilon_0} \sum_{i,j} \frac{q_i\, q_j}{r_{ij}} \cdot \epsilon^{-1}_{\kappa(i),\kappa(j)} \tag{4.22}$$

Für die folgenden Betrachtungen ist es hilfreich, sich diese Energie nicht auf atomarer Auflösung, sondern als Wechselwirkung von Gruppen von Atomen vorzustellen, also z.B. als Interaktion der NH_3^+-Gruppe des Lysin mit der OH-Gruppe des Serin, statt einer Wechselwirkung des Serin-O mit einem Lysin-H. Die Partialladungen q_i sind eine modifizierte Version der Partialladungen des CVFF Kraftfelds (*consistent valence forcefield*) von Dauber-Osguthrope [DORO$^+$88]. Das CVFF Kraftfeld ist traditionell das Standardkraftfeld des Discover-Programmpaketes zur Proteinsimulation und wurde an Proteinen und Protein/Liganden-Komplexen kalibriert.

Die Zuordnung der Atome i auf die Werte von κ ist in den Tabellen des Anhangs C wiederzufinden und die zugehörigen Dielektrizitätskonstanten $\epsilon_{\kappa(i),\kappa(j)}$ in Tabelle 4.2. Darüber hinaus gilt $\epsilon_{\kappa(i),\kappa(j)} = \epsilon_{\kappa(j),\kappa(i)}$ und für $\kappa(i) = 0$ gilt $\epsilon_{0,\kappa(j)} = \epsilon_{\kappa(j),0} = \infty$.

$\kappa = 1$ bzw. $\kappa = 2$ sind für die Amino- bzw. Carboxylgruppe der Hauptkette reserviert und dienen im CARB-Kraftfeld der Beschreibung von Wasserstoffbrückenbindungen. Sie stellen mit $\epsilon = 1/0.375731 \approx 2.66$ die stärkste elektrostatische Wechselwirkung dar. Etwa einen Faktor 2.5 kleiner ist die Wechselwirkung der Gruppen mit $\kappa = 3, 4$ und 5. Dies sind die partiell geladenen OH, CO und NH_2 Gruppen der (Asn, Gln, Ser, Thr, Tyr) Seitenketten. Nochmals einen Faktor 3 geringer ist die Wechselwirkung zu den geladenen COO^- und $NH_x^{(+)}$ Gruppen (in Asp, Glu, Arg, Lys, His, Trp) mit $\kappa = 6$.

Die Tabelle 4.2 des CARB-Kraftfeldes wurde über einen *Potential-of-mean-force* Ansatz gewonnen [AJ95]. Sie ist in die Elektrostatik des Kraftfeldes PFF01 übernommen worden.

4.3.6 Elektrostatik der Hauptkette

In Abbildung 4.5 ist der atomare Aufbau der Hauptkette wiedergegeben. Die verschiedenen Aminosäuren unterscheiden sich nur in der Struktur der Seitengrup-

κ	1	2	3	4	5	6
1	ϵ_{11}^{-1} =0.375731	0.375731	0	0.143396	0.143396	0.043222
2		0.375731	0.161852	0.143396	0.143396	0.031012
3			0	0	0.161852	0.045452
4				0.143396	0.143396	0.043222
5					0.143396	0.031012
6						0.013097

Tabelle 4.2: inverse gruppenspezifische Dielektrizitätskonstanten $\epsilon^{-1}_{\kappa(i),\kappa(j)}$

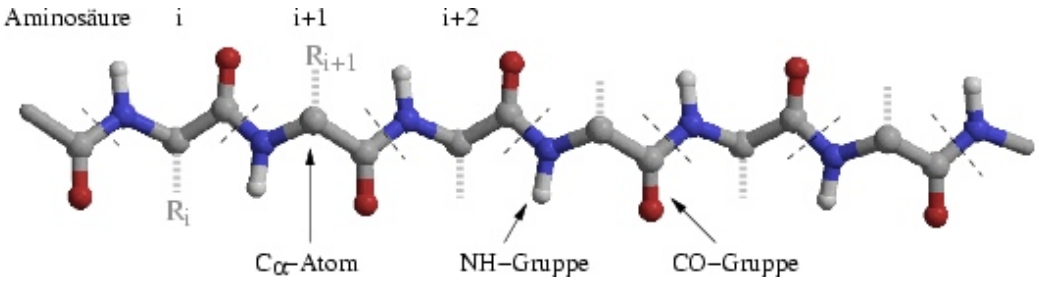


Abbildung 4.5: Darstellung der Hauptkette einer allgemeinen Aminosäuresequenz. Verschiedene Aminosäuren unterscheiden sich im Substituenten R

pen, hier mit R bezeichnet, welche an das jeweilige C_α-Atom der Hauptkette gebunden sind. Das C_α-Atom ist von zwei Dipolen (der NH- und der CO-Gruppen) flankiert. Die Wechselwirkungsenergie der Dipole wird über die Coulombenergie der beteiligten Punktladungen (Abbildung 3.4) berechnet:

$$E_{main} = \frac{1}{4\pi\epsilon} \sum_{i,j\in\{C,O,N,H\}} \frac{q_i\, q_j}{r_{ij}} \tag{4.23}$$

mit $\epsilon = \epsilon_0/0.375731$ (Tabelle 4.2). In vielen Kraftfeldern wird die Hauptkettenelektrostatik für die Beschreibung der Wasserstoffbrückenbindungen herangezogen (Seite 3.2.2).

4.3.7 Elektrostatik der Seitengruppen

Die Wechselwirkung der Seitengruppen haben wir bislang nur geringfügig gegenüber dem CARB-Kraftfeld modifiziert. Dies liegt wesentlich daran, daß der Beitrag der Seitengruppen-Elektrostatik zum Unterschied der Freien Energie zweier Strukturen als untergeordnet eingestuft wird. Die dominanten Beiträge ergeben sich aus dem Lösungsmittel und den Wasserstoffbrückenbindungen innerhalb des Proteins. Die vorgenommenen Änderungen an der Seitengruppenelektrostatik sind folgendermaßen motiviert:

Nach Simulation mit den (alten) Lösungsmittelparametern σ_{FS-TK} (s.u.) hat sich gezeigt, daß (partiell) geladene Seitenketten, die sich im Proteininneren grup-

pieren, in einigen Fällen zu metastabilen Zuständen geführt haben, deren Freie Energie nicht deutlich oberhalb der Energie nativen Struktur liegt. Nach unserer Vorstellung ist dies (zumindest für kleine Proteine) nicht korrekt. Wir gehen davon aus, daß wegen des Säure-Base-Gleichgewichtes die betreffenden Seitengruppen im Proteinkern nur in neutraler Form vorliegen. Daher sind wir dazu übergegangen, die Ladungen mit der dem Lösungmittel zugänglichen Oberfläche zu skalieren. Diese Oberfläche bezieht sich nicht auf die einzelnen Atome, sondern auf die Oberfläche der (Seiten-)Gruppe, zu der das Atom gehört. Ebenso werden alle Ladungen einer Gruppe mit demselben Faktor skaliert.

Würde man diese Skalierung proportional zur Oberfläche ansetzen, so wäre es denkbar, daß zwei gegensätzliche Ladungen auf der Proteinoberfläche sich nur bis auf einen Mindestabstand anziehen, da eine weiter Annäherung zu einem Verlust an freier Oberfläche führen würde. In einer ersten Abschätzung gehen wir davon aus, daß, wenn zwei Gruppen der elektrostatischen Wechselwirkung an der Proteinoberfläche in Kontakt sind, etwa noch ein Viertel ihrer Oberflächen Zugang zum Wasser hat. Wenn also für eine Gruppe mehr als ein Viertel der maximal möglichen SAS-Oberfläche in der vorliegenden Konfiguration in Kontakt mit dem Lösungsmittel steht, behält diese Gruppe vollständig ihre Ladungen. Liegt die SAS-Oberfläche unterhalb dieser Grenze werden die Ladungen linear herunterskaliert, bis sie auf Null abgefallen sind, wenn die Atomgruppe keine freie Oberfläche mehr hat.

4.4 Wasserstoffbrückenbindungen

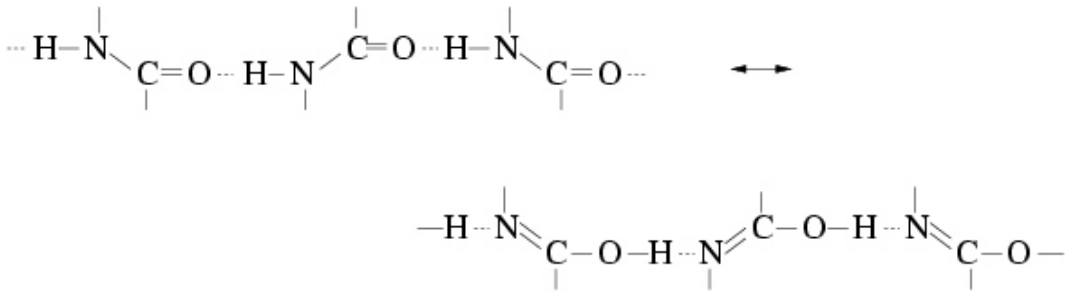
Einen wichtigen Anteil an der Freien Energie und der Struktur des nativen Zustandes haben Wasserstoffbrückenbindungen, die sich im Protein zwischen den NH- und CO-Gruppen der Hauptkette ausbilden. In Kraftfeldern zur Simulation von Proteinen kommt der Elektrostatik der Hauptkette eine wichtige, teilweise sogar zentrale Rolle bei der Ausbildung dieser Wasserstoffbrücken zu. So werden in einigen Versionen des CHARMM und AMBER Kraftfeldes H-Brücken durch eine Kombination aus Hauptkettenelektrostatik und Lennard-Jones-Wechselwirkung beschrieben. Es ist jedoch bekannt, daß quantenmechanische Effekte, die nicht in der Elektrostatik enthalten sind, wesentlich zur Konfiguration der Wasserstoffbrücken beitragen. Da die Ausbildung von Sekundärstrukturmerkmalen mit einer großen Zahl an Wassersoffbrücken einhergeht, muß die Stärke und Geometrie der Wasserstoffbrückenbindung richtig wiedergegeben werden.

4.4.1 Grenzen der Elektrostatik

Der Beitrag einer Wasserstoffbrücke zu Freien Energie wurde an vielen keinen Molekülen experimentell bestimmt. Die Ergebnisse variieren zwischen $-2.8\,kcal\,mol^{-1}$ und $+1.9\,kcal\,mol^{-1}$ je nachdem, welches Molekül betrachtet wurde [Avb92, MT94].

Die Stabilität der Wasserstoffbrückenbindungen hängt darüber hinaus stark vom eingesetzten Lösungsmittel ab. So liegen die Werte für N-methylacetamid in Wasser bei $+3.1\,kcal\,mol^{-1}$, in Dioxan bei $+0.39\,kcal\,mol^{-1}$ und bei $-0.92\,kcal\,mol^{-1}$ in Tetra-Chlor-Kohlenwasserstoff [Avb92].

Für eine $CO\cdots HN$ Wasserstoffbrückenbindung sagen quantenchemische Berechnungen in der Gasphase eine Freie Energie von $-6.5kcal\,mol^{-1}$ voraus (Dieser Wert gilt für die optimale Geometrie; in der α-Helix-Konfiguration sind es nur $-4.9\,kcal\,mol^{-1}$ und -5.6 bis $-6.3\,kcal\,mol^{-1}$ in der β-Faltblatt-Anordnung) [HY95]. Rechnet man hiervon die Transferenergien von Gas-zu-Wasser und Wasser-zu-Öl ab, so ist eine $CO\cdots HN$ Wasserstoffbrückenbindung in Wasser nur marginal stabil mit $-0.5\,kcal\,mol^{-1}$ und in einem organischen Lösungsmittel liegt der Beitrag zur Freien Energie bei $-3.8\,kcal\,mol^{-1}$ [HY95]. Jüngste DFT Studien an Polyalanin sagen für eine isolierte Wasserstoffbrücke eine Stärke von $-3.5\,kcal\,mol^{-1}$ voraus [INS+03]. Für eine unendliche Polyalaninkette in α-Helix Konfiguration wurde ein Wert von $-8.6\,kcal\,mol^{-1}$ prognostiziert. Somit würde eine α-Helix wesentlich durch einen kooperativen Effekt der Wasserstoffbrücken stabilisiert. Die Verstärkung einzelner Wasserstoffbrücken erfolgt dabei unter anderem durch Delokalisierung der Wasserstoffbrücken über die Peptidbindung [Kau59]:



Andere Studien sagen allerdings eine kooperative Verstärkung einzelner Wasserstoffbrücken von weniger als 50%, statt mehr als 200% voraus [Smi94].

In biomolekularen Kraftfeldern ist es nicht möglich, alle Einflüsse auf die Stärke einer Wasserstoffbrücke zu berücksichtigen. In der Regel werden sie durch die elektrostatische Wechselwirkung der vier Atome der Amino- und der Carboxylgruppe beschrieben (Abbildung 3.4).

$$E_{ij} = E_{main}(CO \in res(i), HN \in res(j)) \tag{4.24}$$

$$= \frac{0.38e \cdot 0.28e}{4\pi\epsilon\epsilon_0}\left(\frac{1}{|\vec{r}_C - \vec{r}_H|} - \frac{1}{|\vec{r}_C - \vec{r}_N|}\right. \tag{4.25}$$

$$\left. - \frac{1}{|\vec{r}_O - \vec{r}_H|} + \frac{1}{|\vec{r}_O - \vec{r}_N|}\right) \tag{4.26}$$

Zusammen mit der Lennard-Jones Wechselwirkung ist es möglich, den natürlichen $H-O$-Abstand und die Winkelabhängigkeit der Wasserstoffbrückenbindungsstärke näherungsweise zu reproduzieren.

In Gleichung 4.26 ist eine "kooperative" Verstärkung durch die langreichweitigen Coulomb-Wechselwirkung enthalten. Berücksichtigt man die über die Peptidbindung nächstgelegenen Dipole CO bzw. HN mit, so liegt die Energie der Konfiguration $NH-CO\cdots HN-CO$ rund 50% über der reinen $CO\cdots HN$ Energie.

In einer Helixkonfiguration läßt sich die von Gleichung 4.26 vorhergesagte Stärke des kooperativen Effektes wie in Abbildung 4.6 dargestellt zusammenfassen. Dazu wurde ein mit dem Programmpaket TINKER [RP03] generiertes Polyalaninpeptid mit 40 Residuen im Vakuum simuliert. Die Helixkonfiguration ist das globale Minimum der Potentialenergieoberfläche. Für ein Residuum i lassen sich die elektrostatischen Beiträge gemäß Gleichung 4.26 der Residuen $i+1$ bis einschließlich Residuum $i+d$ aufaddieren. Die Beiträge der Residuen mit kleineren Nummern $(i-1,\, i-2,\, \ldots)$ wurden nicht berücksichtigt[7].

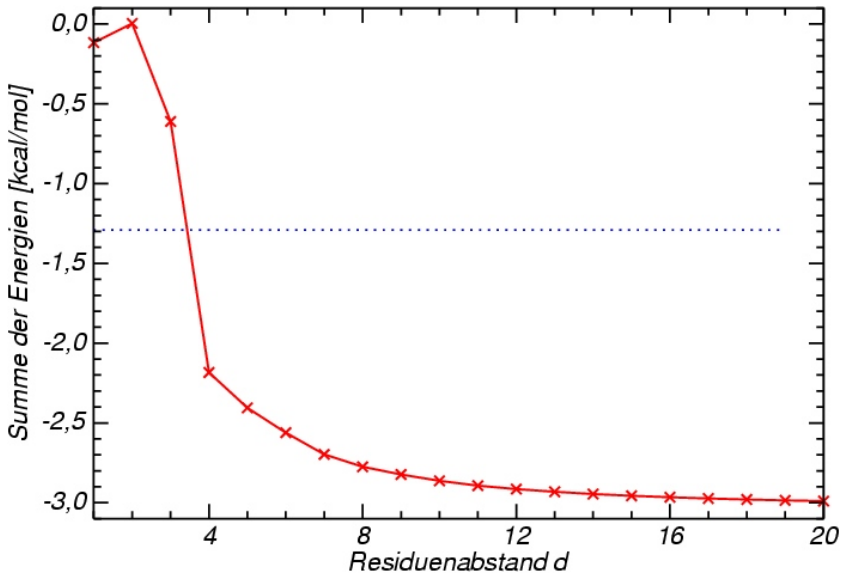


Abbildung 4.6: Wasserstoffbrückenbindungsenergie $\sum_{j>i}^{|i-j|\leq d} E_{ij}$

In Abbildung 4.6 sind die Summen der mittleren Energiebeiträge als Funktion des Residuenabstands d dargestellt. Für $d > 6$ gehen die Energiezuwächse in das $1/r^2$ der Dipol-Dipol-Wechselwirkung über, d.h. nach 1.5 Helixwindungen. Die Energie der lokalen Wasserstoffbrückenbindung beträgt $-1.29\, kcal\, mol^{-1}$ und

[7]Die Gesamtenergie der Hauptketten-Elektrostatik ist die Summe der Einzelbeiträge über alle Wechselwirkungspaare. Diese Summe läßt sich entweder in Vorwärtsrichtung als $\sum_i \sum_{j>i} E_{ij}$ oder in beide Richtungen als $1/2 \sum_i \sum_{j\neq i} E_{ij}$ berechnen. Für das Diagramm ist der Teil $\sum_{j>i}^{d} E_{ij}$ der Doppelsumme in Vorwärtsrichtung berechnet worden.

ist als punktierte Linie eingezeichnet. Dieser Energiebeitrag wird in die Summe $\sum_{j>i}^{|i-j|\leq d} E_{ij}$ bei $d = 4$ aufgenommen, da eine α-Helix Wasserstoffbrücken zwischen der i-ten und $(i+4)$-ten Aminosäure ausbildet. Die kooperative Verstärkung der Wasserstoffbrückenbindung in eine Helixrichtung beträgt im CARB-Kraftfeld etwa 130%.

Folglich ist eine Helix im CARB-Kraftfeld faktisch nur wegen des langreichweitigen Charakters der zugrundeliegenden Wechselwirkung thermodynamisch stabil. Abschirmeffekte durch das Lösungsmittel und dort gelöste Ionen verhindern mit großer Sicherheit kooperative Effekte dieser Größe. Es verwundert nicht, daß im Zusammenspiel mit dem CARB-eigenen Lösungsmittelmodell erste Simulationen an größeren Proteinen stets zu Strukturen mit einzelnen sehr langen Helices geführt haben. Wir gehen davon aus, daß die reine elektrostatische Behandlung die Stärke der lokalen Wasserstoffbrücke unterbewertet und die des kooperativen Anteils überbewertet.

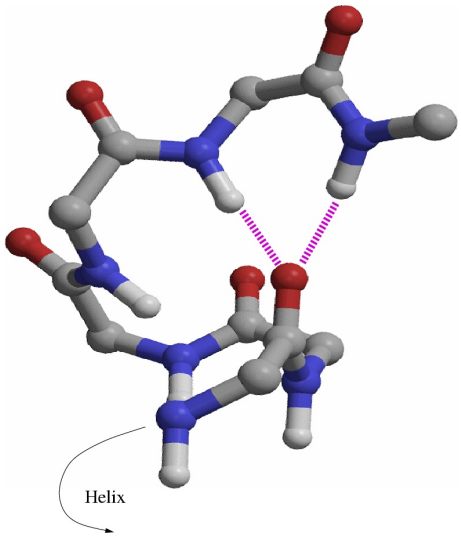


Abbildung 4.7: Ein wiederholt auftretendes Strukturelement des Kraftfeldes mit reinem elektrostatischen Wasserstoffbrückenpotential

Abgesehen von der Überbetonung langreichweitiger Effekte hat der gewählte Ansatz, Wasserstoffbrücken über Dipol-Dipol-Wechselwirkungen bzw. Gleichung 4.26 mit einer festen Dielektrizitätskonstanten zu beschreiben, wiederholt zu bestimmten lokalen Hauptketten-Konfigurationen geführt. Eine dieser Konfigurationen ist in Abbildung 4.7 dargestellt und zeigt eine doppelte Wasserstoffbrücke, die über untypische (aber nicht Lennard-Jones-verbotene) Dihedralwinkel zustandegekommen ist. Diese Konfiguration war in den Simulationen sowohl am Ende von Helices als auch als stabilisierendes Element unstrukturierter Bereiche anzutreffen. Allerdings wird in den 138 Proteinen (M^{138}, Anhang B.1), die uns schon im Abschnitt zur van-der-Waals Wechselwirkung begegnet sind, keines der Sekundärstrukturelemente durch eine doppelte Wasserstoffbrücke terminiert. Auch

wenn das Sauerstoffatom 2 einfachbesetzte $2p$-Orbitale hat und daher 2 Wasserstoffbrücken ausbilden kann, ist dieses Motiv nur äußerst selten in Proteinen realisiert. Bei diesem Strukturelement handelt es sich wahrscheinlich um ein Artefakt der Gleichung 4.26, welche die Natur nicht richtig wiederzugeben scheint. Hier muß festgehalten werden, daß sich dies Argument auf die ausgebildeten Winkel bezieht, wohingegen der Abstand zwischen den H- und O-Atomen recht gut mit den natürlichen Abständen übereinstimmt.

Das Strukturmotiv aus Abbildung 4.7 läßt sich durch Einführung eines Potentials der Dihedralwinkel, etwa in der Form $\cos(n\phi + \gamma)$ (vgl. AMBER und CHARMM), energetisch benachteiligen. Dabei würden die langreichweitigen Eigenschaften der elektrostatischen Wechselwirkung jedoch unangetastet bleiben, weshalb wir einer andere Lösung den Vorzug gegeben haben.

4.4.2 Bestimmung eines Korrekturpotentials

Für isolierte Aminosäuren lassen sich die Ladungsverteilung der Elektronen berechnen und den Atomen so Partialladungen zuordnen. Die elektrostatische Wechselwirkung zweier Aminosäuren im Vakuum, die durch eine große Distanz voneinander getrennt sind, läßt sich durch das Coulomb-Gesetz beschreiben, welches auch Grundlage für Gleichung 4.7 ist. Bringt man die beiden Aminosäuren zusammen, so daß sie eine Wasserstoffbrückenbindung ausbildet, werden Polarisationseffekte und Delokalisierung wichtig. Die Dielektrizitätskonstante in Gleichung 4.7 ist durch ein *potential-of-mean-force*-Ansatz aus strukturaufgeklärten Proteinen abgeleitet worden [AM95]. Da in diesen Strukturen sehr viele Wasserstoffbrücken ausgebildet sind, sind in der Dielektrizitätskonstante sowohl die Elektrostatik als auch alle weiteren quantenmechanischen Einflüsse subsumiert. Wenn Geometrie und Stärke der Wasserstoffbrücken in den Simulationen mit dieser Dielektrizitätskonstanten richtig reproduziert worden wären, so könnte man auf die Trennung der klassischen elektrostatischen Beschreibung und der quantenmechanischen Behandlung verzichten. Wir haben daher die Verteilung einiger wichtiger Charakteristika der Wasserstoffbrückenbindungen in strukturaufgeklärten Proteinen untersucht und – in Übereinstimmung mit den Resultaten des vorangegangenen Abschnitts – gefunden, daß eine Korrektur zur klassischen Beschreibung des Coulomb-Gesetzes nötig ist.

Ziel dieses Abschnittes ist die Beschreibung eines (klassischen) Kontaktpotentials, welches die quantenmechanischen Korrekturen enthalten soll. Die Wasserstoffbrückenbindungsenergie E_{hb} läßt sich dann durch eine Kombination der Hauptkettenelektrostatik E_{main} und des Kontaktpotentials E_{pmf} mit $\lambda \in [0;1]$ beschreiben:

$$E_{hb} = \lambda E_{main} + (1 - \lambda) E_{pmf} \tag{4.27}$$

Hierbei bietet λ einen kontinuierlichen Übergang zwischen der reinen elektrostatischen Beschreibung und dem neuen Potential. Wie noch gezeigt wird, liegt der

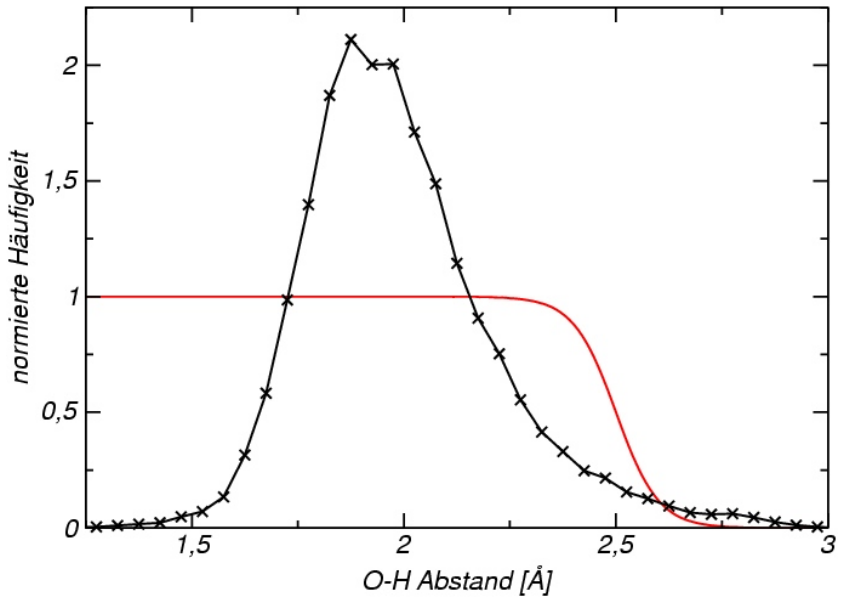


Abbildung 4.8: Normierte $O-H$-Abstandsverteilung (schwarz) und der abstandsabhängige Anteil des Wasserstoffbrückenpotentials (rot)

optimale Wert bei $\lambda = 0.25$.

In den Simulationen mit einer rein elektrostatischen Beschreibung konnte beobachtet werden, daß die Verteilung der H-O-Atomabstände denen der natürlichen entspricht. Die Aufgabe des Potential E_{pmf} für die Proteinstrukturvorhersage besteht darin, die natürliche Winkelverteilung in den Simulationen nachzubilden. Daher wurde für die funktionale Form des neuen Wasserstoffbrückenpotentials E_{pmf} ein Produkt aus einer abstandsabhängigen Funktion, die das Potential auf den interessanten Abstandsbereich beschränkt, und einer winkelabhängigen Funktion angesetzt.

Die natürliche Abstandsverteilung der Wasserstoffbrücken in den 138 strukturaufgeklärten Proteinen M^{138} (Anhang B.1) ist in Abbildung 4.8 wiedergegeben. Der abstandsabhängige Faktor von E_{pmf} wird durch eine Funktion beschrieben, die für kleine Abstände gleich 1 und dann bei einem bestimmten Wert kontinuierlich und differenzierbar auf Null abfällt. Als Prototyp einer solchen Funktion haben wir

$$t(r; r_0, \sigma_r) = \frac{1}{2}\left[1 - \tanh\left(\frac{r - r_0}{\sigma_r}\right)\right] \tag{4.28}$$

gewählt. Für den Abstandsanteil von E_{pmf} speziell $r_0 = 2.5$Å und $\sigma_r = 0.1$Å. Der Verlauf dieser Funktion ist in die Abbildung 4.8 aufgenommen worden.

Die Geometrie der NH- und CO-Gruppen erlaubt gemäß Abbildung 4.9 die Definition dreier Winkel. Mit dem Winkel ν wird der Winkel zwischen den Atomen N, H und O bezeichnet und der Winkel der NH- und CO-Dipole trägt den

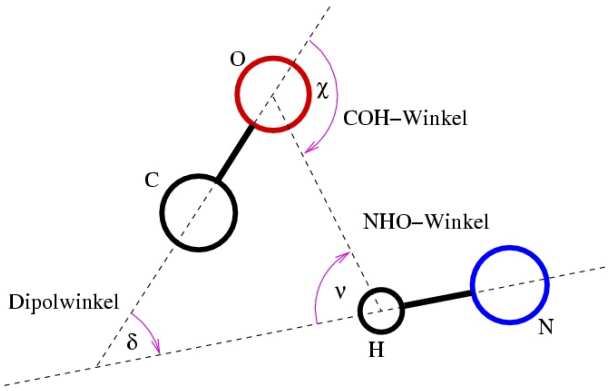


Abbildung 4.9: Definition der Wasserstoffbrückenwinkel

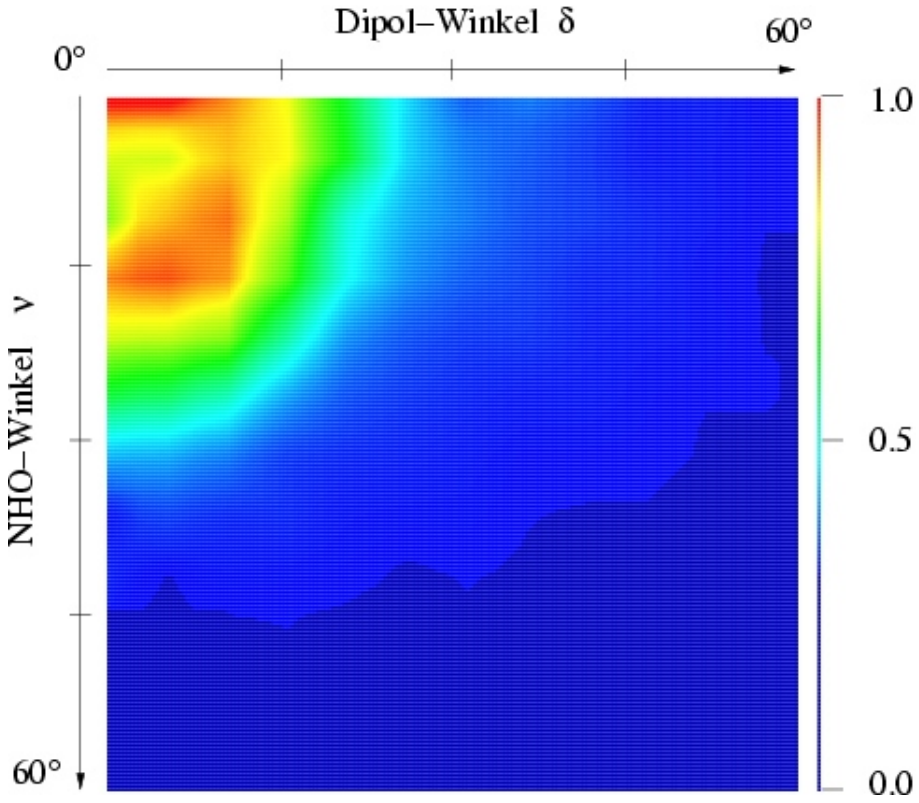


Abbildung 4.10: Konturgraph der Winkelverteilung in Wasserstoffbrückenbindungen.

Namen δ. Die Winkel seien dabei so definiert, daß 0° der gestreckten NH-OC Konfiguration entspricht. Für den dritten Winkel χ, den Winkel zwischen C, O und H, gilt $\chi = \nu + \delta$; er muß nicht weiter berücksichtigt werden.

Bei der Betrachtung der Winkelverteilung ist zu berücksichtigen, daß die Fläche für einen Winkelbereich α bis $\alpha+\Delta\alpha$ (mit $\alpha = \nu, \delta, \chi$) von α abhängig ist, da die Verteilung im dreidimensionalen Raum nicht nur über den Abstand, sondern noch über den verbleibenden Winkel zu integrieren ist. Wie man sich leicht klarmachen kann, ist die Fläche des Winkelbereichs α bis $\alpha + \Delta\alpha$ proportional $\sin\alpha$. Daher werden natürliche Winkelverteilung durch $\sin\alpha$ dividiert und somit "flächennormiert" betrachtet.

Im Konturgraphen 4.10 ist die flächennormiert Häufigkeit natürlicher Wasserstoffbrücken gegen die beiden Winkel ν und δ aufgetragen, wobei im dunkelblauen Bereich keine Wasserstoffbrücke mit den entsprechenden Winkeln in den 138 Strukturen aus Anhang B.1 existiert, und die Farbe rot signalisiert, daß Wasserstoffbrücken sehr oft mit diesen Winkeln gebildet werden. Der rote rechteckige Bereich für Winkel kleiner $\nu < 20°$ und $\delta < 10°$ legt einen Produktansatzes des winkelabhängigen Anteils des Wasserstoffbrückenpotentials E_{pmf} in ν und δ nahe. Für größere Winkel muß dieses Potential jedoch um Korrelationen zwischen ν und δ korrigiert werden. Der Abfall des Verteilung unter den Wert 0.5 ist am türkisfarbenen Bereich zu erkennen. Der Verlauf dieser Übergangslinie läßt sich durch

$$\zeta^2(\nu, \delta; \nu_0, \delta_0) := \left(\frac{\nu}{\nu_0}\right)^2 + \left(\frac{\delta}{\delta_0}\right)^2 = konst. \tag{4.29}$$

mit $\nu_0 = 30°$ und $\delta_0 = 24°$ beschreiben.

Der winkelabhängige Teil des neuen Wasserstoffbrückenpotential wird durch eine Funktion des Typs 4.28 mit dem Argument ζ (und ζ_0, σ_ζ) beschrieben. Das Potential hat seine volle Stärke für kleine Winkel ν und δ und fällt für größere Winkel ($\zeta \gtrsim \zeta_0$) kontinuierlich auf Null ab. Die Werte $\zeta_0 = 1.5$ und $\sigma_\zeta = 0.05$ wurden so gewählt, daß die natürlichen Wasserstoffbrücken in den 138 strukturaufgeklärten Proteinen M^{138} durch das neue Potential erkannt wurden.

Fassen wir zusammen:
In PFF01 werden Wasserstoffbrücken gemäß

$$E_{hb} = \lambda E_{main} + (1 - \lambda) E_{pmf} \ , \tag{4.30}$$

durch zwei Anteile beschrieben, ein langreichweitiges elektrostatisches Potential und ein Kontaktpotential. Das langreichweitige Potential hat die Form

$$\begin{aligned} E_{main} \quad = \quad & \frac{0.38e \cdot 0.28e}{4\pi\epsilon_{11}\epsilon_0} \left(\frac{1}{|\vec{r}_C - \vec{r}_H|} - \frac{1}{|\vec{r}_C - \vec{r}_N|} \right. \\ & \left. - \frac{1}{|\vec{r}_O - \vec{r}_H|} + \frac{1}{|\vec{r}_O - \vec{r}_N|} \right) \end{aligned}$$

mit $\epsilon_{11} = 1/0.375731$ (Tabelle 4.2). Das Kontaktpotential lautet (Abbildung 4.9)

$$E_{pmf}(r,\nu,\delta) = E_0 \cdot \underbrace{t(r;r_0,\sigma_r)}_{\text{abstandsabhängig}} \cdot \underbrace{t(\zeta(\nu,\delta;\nu_0,\delta_0);\zeta_0,\sigma_\zeta)}_{\text{winkelabhängig}} \tag{4.31}$$

mit $t(x;x_0,\sigma_x) = \frac{1}{2}[1-\tanh(\frac{x-x_0}{\sigma_x})]$ und

$$\begin{aligned} E_0 &= -2.31\,kcal\,mol^{-1} \\ r_0 &= 2.5\,\text{Å} \\ \sigma_r &= 0.1\,\text{Å} \end{aligned} \qquad \begin{aligned} \nu_0 &= 30^\circ \\ \delta_0 &= 24^\circ \\ \zeta_0 &= 1.5 \\ \sigma_\zeta &= 0.05 \end{aligned}$$

Der Parameter λ bietet die Möglichkeit, die Beschreibung der Wasserstoffbrücken kontinuierlich von der langreichweitigen elektrostatischen Beschreibung durch das Coulomb-Gesetz in eine Beschreibung zu überführen, die sich an experimentell aufgelösten Strukturen orientiert, um quantenmechanische Einflüsse miteinzubeziehen.

Um den Wert für λ zu bestimmen, sind mehrere nativen Strukturen mit verschiedenen λ-Werten relaxiert worden. Die mittlere RMSB-Abweichung der relaxierten Struktur nur nativen war bei $\lambda = 1/4$ minimal. Dieser Wert führt dazu, daß effektiv die Dielektrizitätskonstanten für die Seitengruppen- und die Hauptketten-Elektrostatik annähernd gleich sind (In Tabelle 4.2 führt $\lambda = 1/4$ zu $\epsilon_{11}^{-1} = 0.375731 \mapsto 0.375731\lambda \approx 0.094$). Folglich hat die Einführung des neuen Potentials zu einer Vereinfachung der Parameterstruktur des kompliziertesten Kraftfeldbestandteiles, nämlich der Elektrostatik der Proteinatome, geführt.

4.5 Wechselwirkung mit dem Lösungsmittel

Die Proteinstrukturvorhersage behandelt – zum derzeitigen Stand der Forschung – ausschließlich Proteine in wässeriger Lösung. Die Interaktion der Wassermoleküle untereinander und mit dem Protein wirken sich auf verschiedene Aspekte aus. Zum einen führt die Polarisierbarkeit des Wassers zu einer Abschirmung der elektrostatischen Wechselwirkung innerhalb des Proteins, wie es im vorherigen Abschnitt besprochen wurde. Ebenso wird die Desolvatationsenergie partial geladener Atome in der Elektrostatik berücksichtigt. Dynamische Aspekte sind dabei weitestgehend unberücksichtigt geblieben. Diese dynamischen Beiträge spiegeln sich in der Entropie des Protein/Wasser-Systems wider, welche sich ihrerseits in zwei Bestandteile zerlegen läßt. Die Bewegungen des Proteins führen zur Konfigurationsentropie und der Verlust dynamischer Wasserstoffbrückenbindungen des Wasser wird als *hydrophober Effekt* bezeichnet.

In Kraftfeldern der Freien Energie wird die Entropie nicht simuliert, sondern im Rahmen eines *impliziten Lösungsmittelmodells* behandelt. So gelingt es, einer einzelnen Proteinkonfiguration effektiv eine Entropie zuzuordnen. Bei der

Betrachtung impliziter Lösungsmittelmodelle werden wir uns zunächst auf die Interaktion des Proteins mit dem Wasser konzentrieren und nachfolgend die Konfigurationsentropie des Proteins besprechen.

4.5.1 Implizite Lösungsmittelmodelle

Seit den 80er-Jahren folgen implizite Lösungsmittelmodelle dem Ansatz, daß der Lösungsmittelbeitrag zur Freien Energie mit der Proteinoberfläche in Beziehung gesetzt werden kann. Der Begriff der Proteinoberfläche ist als die dem Wasser durch van-der-Waals Kontakt zugängliche Oberfläche zu verstehen. Diese Oberfläche wird im Englischen auch mit *Solvent Accessible Surface Area* (SASA) bezeichnet. Die folgende Definition geht zurück auf Lee und Richards [LR71]: Die Atome des Proteins seien durch Kugeln repräsentiert, deren Durchmesser dem Lennard-Jones Gleichgewichtsabstand τ_{ii} gleichgesetzt wird. Die Oberfläche ergibt sich nun durch "Abrollen" einer fiktiven Wasserkugel mit einem Radius von 1.4Å über die Proteinatomkugeln (Abbildung 4.11). Die SAS-Oberfläche ist nicht die Oberfläche des Proteins, die von einer Kugel mit Radius 1.4Å berührt werden kann, sondern die Menge aller Punkte, an die der Kugelmittelpunkt der Wasserkugel plaziert werden kann[8]. In Anhang D sind verschiedene numerische Methoden zur Berechnung der SAS-Oberfläche beschrieben.

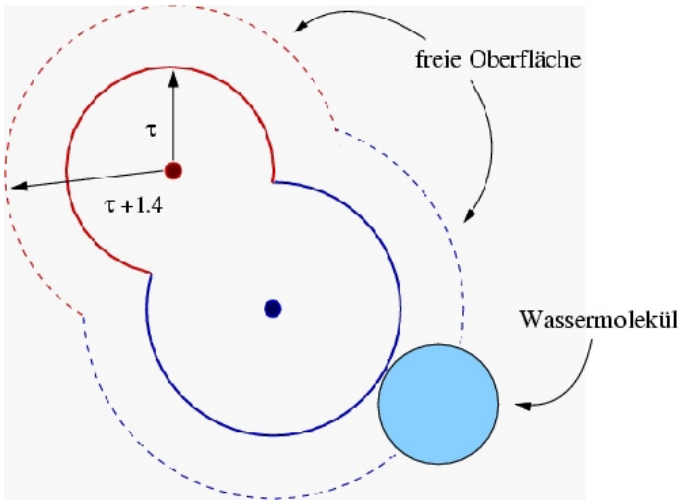


Abbildung 4.11: Definition der Proteinoberfläche (SASA) am Beispiel zweier Atome

Die Proteinoberfläche ist eine Funktion der Konfiguration. Während der Faltung von einer gestreckten Struktur hin zur nativen, ändert sich die mittlere

[8] An dieser Stelle wird der Begriff des van-der-Waals Radius anders interpretiert als andernorts im CARB Kraftfeld üblich. Hier ergibt sich der Gleichgewichtsabstand durch Addition ($\tau_{ij} = \frac{1}{2}(\tau_{ii} + \tau_{jj})$) und nicht durch Multiplikation ($\tau_{ij} = \sqrt{\tau_{ii}\tau_{jj}}$).

Freie Oberfläche einer Aminosäure. Das Verhältnis dieser Oberflächenänderung zur Oberfläche einer gestreckten Referenzstruktur wurde von Rose et al. bestimmt [R+85] (Abbildung 4.12). Es zeigte sich, daß die Aminosäuren entlang dreier Linien arrangiert sind und sich diesbezüglich in drei Familien einteilen lassen; in hydrophob, leicht polar und stark polar. Die Zuordnung der Aminosäuren lautet:

hydrophob (rot) Ala, Cys, Val, Ile, Leu, Met, Phe, Trp

leicht polar (grün) Gly, Ser, Thr, His, Tyr

stark polar (blau) Pro, Asp, Asn, Glu, Gln, Lys, Arg

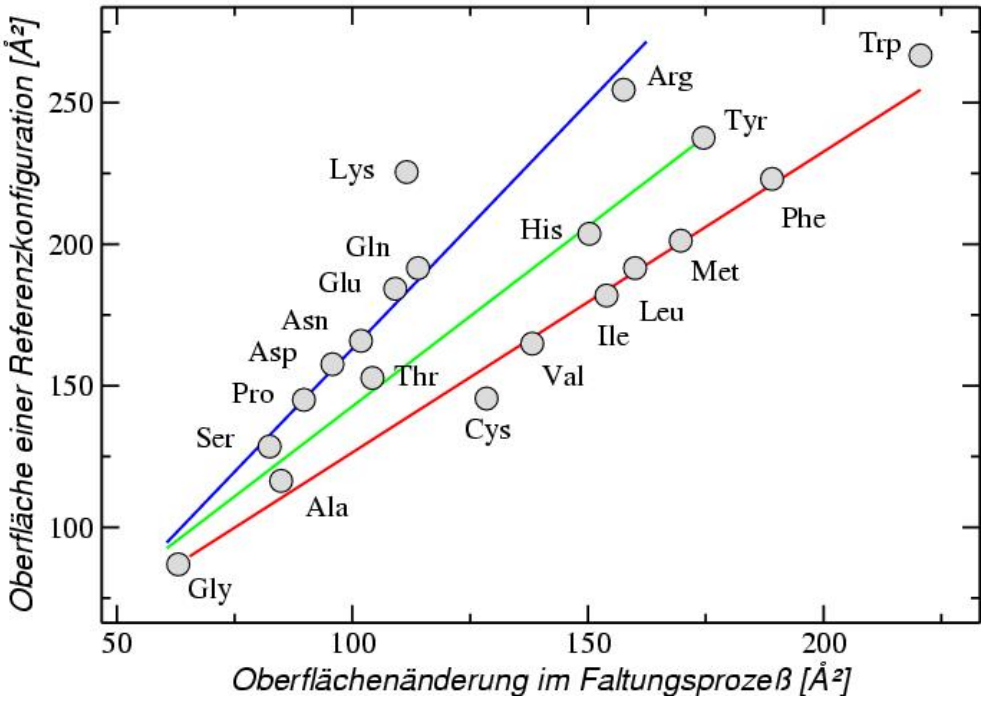


Abbildung 4.12: Reproduktion des Graphen von Rose et al. [R+85]

Diese Einteilung ergibt sich allein aus den Oberflächenveränderungen während der Faltung, wenngleich die Namensgebung durch die chemische Zusammensetzung der Aminosäuren motiviert ist. Bemerkenswert an dieser Verteilung ist die Einstufung von Tryptophan als hydrophob und Prolin als stark polar, welche im Widerspruch zu vielen Kategorisierungen steht, welche auf Transferenergien (s.u.) der Aminosäuren zwischen Vakuum-und-Wasser oder Öl-und-Wasser beruhen.

In oberflächenbasierten Lösungsmittelmodellen wird die freie Oberfläche in Relation zur Freien Energie des Proteins und des Wassers gesetzt. Dabei wird versucht, den Unterschied zwischen dem Energiebeitrag einer Aminosäure im Inneren des Proteins und dem einer Aminosäure an der Proteinoberfläche zu erfassen. Unter der Annahme, daß Oktanol als Lösungsmittel eine Umgebung schafft,

die dem Inneren eines Proteins ähnlich ist, kann aus dem Transfer eines Peptids von Oktanol zu Wasser der Lösungsmittelbeitrag zur Freien Energie ermittelt werden. Um dies zu tun, bedienen wir uns eines weit verbreiteten Ansatzes, der Eisenberg und McLachlan zugeordnet wird [EM86]. Ihr Ansatz beruht im Wesentlichen auf zwei Annahmen:

- Die Freie Energie des Transfers der Aminosäuren von Wasser ins Innere des Proteins ist die Summe über die Transferenergien der einzelnen Atome.
- Die Transferenergie jedes einzelnen Atoms i ist eine lineare Funktion seiner Oberfläche A_i.

Ordnet man jedem Atom i über seinen Potentialtyp $pt(i)$ einen Lösungsmittelparameter (ASP:Atomic Solvation Parameter) $\sigma_{pt(i)}$ zu, so kann die Beziehung zwischen der Transferenergie und den Atomoberflächen zu

$$\Delta F = \sum_i \sigma_{pt(i)} A(i) \tag{4.32}$$

zusammengefaßt werden.

Aus den Transferenergien und den Atomoberflächen lassen sich die Lösungsmittelparameter nach Gleichung 4.32 bestimmen. Die 20 Aminosäuren $X = Ala, Val, \ldots$ lassen sich, um möglichst gleiche Bedingungen zu gewährleisten, mit Glycin flankiert und als Tripeptid in der Form $Gly - X - Gly$ untersuchen. Fauchere und Pliska haben die Transferenergien der Tripeptide von Wasser zu n-Oktanol experimentell bestimmt [FP83].

Arbeiten auf dem Gebiet der Theorie der Polymere haben ergeben, daß bei Transferenergien die unterschiedlichen Volumina der Lösungsmittel sowie der Tripeptide berücksichtigt werden müssen und diese im allgemeinen zu einer Erhöhung der Transferenergien führen. Die experimentellen Daten von Fauchere und Pliska sind in einer Arbeit von Sharp entsprechend korrigiert worden [SNFH91] und sind den Orginaldaten in Tabelle 4.3 gegenübergestellt. Desweiteren ist in die Tabelle eine Hydrophobizitätsskala aufgenommen, die aus einem *potential-of-mean-force* (PMF) aus 88 strukturaufgeklärten Proteinen ermittelt wurde [CS92]. Sie ist stellvertretend für ein Reihe von Hydrophobizitätsskalen aufgenommen worden, die auf unterschiedliche Weise ermittelt wurden [CS92]. Der Vergleich der Transferenergien und der PMF Daten zeigt die Schwierigkeit, den Begriff Hydrophobizität zu quantifizieren. So liegt die Korrelation[9] zwischen den korrigierten Transferenergien und der Hydrophobizitätsskala nur bei 0.77. Die Abweichung

[9]Der empirische Korrelationskoeffizient ist definiert durch

$$R(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \text{ mit } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

der Korrelation von 1 wird insbesondere von Cystin, Prolin und Lysin hervorgerufen, gefolgt von Glycin und Alanin. Im Vergleich zur Arbeit von Rose et al. zeigt sich, daß schon dort das Verhalten der Aminosäuren Lysin, Cystin und Alanin von dem der anderen Aminosäuren abwich (Abbildung 4.12).

Aminosäure	Transferenergie nach Fauchere und Pliska $[kcal\,mol^{-1}]$	Volumen-korrigierte Energien $[kcal\,mol^{-1}]$	Hydrophobizität nach [CS92]
Ala	0.42	1.02	0.2
Arg	-1.38	0.77	-0.6
Asn	-0.79	0.39	-0.4
Asp	-1.05	0.08	-1.3
Cys	2.10	3.18	2.0
Gln	-0.30	1.47	-1.0
Glu	-0.87	0.77	-1.2
Gly	0.00	0.00	0.0
His	0.18	1.94	0.5
Ile	2.46	4.86	1.5
Leu	2.32	4.88	0.6
Lys	-1.35	1.00	-1.5
Met	1.68	3.79	0.5
Phe	2.45	5.02	1.0
Pro	0.67	2.50	-0.9
Ser	-0.05	0.58	-0.7
Thr	0.36	1.52	-0.4
Trp	3.07	6.13	1.6
Tyr	1.31	3.89	0.6
Val	1.67	3.50	0.7

Tabelle 4.3: Transferenergien aller Aminosäuren relativ zu Glycin im Vergleich zu einer Hydrophobizitätsskala, welche aus Proteinstrukturen abgeleitet wurde.

Für die Bestimmung der Lösungsmittelparameter nach Gleichung 4.32 wurden die Transferenergien der Tripeptide $Gly - X - Gly$ sowohl in der unkorrigierten als auch in der Volumen-korrigierten Fassung verwendet. Diese Transferenergien sollen der Energie entsprechen, die notwendig ist, einen Aminosäurerest von der Oberfläche ins Innere des Proteins zu bringen[10]. Demzufolge haben die Tripeptide in Oktanol, welches dem Inneren des Proteins entsprechen soll, keine freie Oberfläche, d.h. $A_{Oktanol}(i) \equiv 0\text{Å}^2$ für alle i. Folglich beziehen sich die Oberflächen

[10]Die Lennard-Jones-Wechselwirkung zwischen dem Protein und dem Lösungsmittel ist in der Transferenergie enthalten, insbesondere die zwischen dem Protein und dem Lösungsmittel Oktanol. In Inneren des Proteins fließt der Lennard-Jones-Beitrag somit doppelt in die Freie Energie ein. Dies ist in unserem Kraftfeld jedoch nicht weiter von Bedeutung, da das Lennard-Jones-Minimum nicht sehr ausgeprägt ist.

auf der rechten Seite von Gleichung 4.32 auf die Peptidstruktur in Wasser und sind durch die Wahl einer Seitengruppenkonfigurationen festgelegt. Diese Wahl bestimmt, wie sich herausstellen wird, maßgeblich die Lösungsmittelparameter und die Qualität der Anpassung.

Die zweite Größe, die in die Lösungsmittelparameter eingeht, ist die Oberfläche ($\sim 4\pi(r_{vdw}+1.4)^2$) der Atome. Nachdem die Lennard-Jones-Radien neu bestimmt wurden, war es daher notwendig die Anpassung der Parameter neu durchzuführen. Im Zug dieser Parameteroptimierung ist die Zahl der Potentialtypen von 34 auf 11 reduziert worden und gleichzeitig das CARB-eigene Lösungsmittelmodell, welches neben den linearen Ausdrücken $\sigma_{pt(i)}A(i)$ auch kubische Terme $\xi_{pt(i)}A(i)^3$ integriert hat, auf das Eisenberg-McLachlan Modell vereinfacht worden.

Für die Generierung der Peptidkonfigurationen wurde das Programmpaket TINKER [RP03] eingesetzt, welches die Möglichkeit bietet, eine Konfiguration aus der Angabe der Sequenz zu erzeugen. Die mit TINKER generierten Strukturen der Tripeptide wurden teilweise nachbehandelt, um gestreckte Konfigurationen zu formen, in denen gleiche Atome in unterschiedlichen Aminosäuren (z.B. C_β in Phe, Tyr, His und Trp) vergleichbar große Oberflächen haben. Denn wenn diese Atome den gleichen Lösungsmittelparameter erhalten, so ist zu erwarten, daß die Tendenz, dem Wasser Kontaktflächen zu bieten, die gleiche ist, und dies sollte sich in der Struktur grob wiederfinden lassen.

Die zu den TINKER-Strukturen gehörigen Oberflächen sind zusammen mit den von Fauchere und Pliska gemessenen Transferenergien in die erste Lösungsmittelparameteranpassung eingeflossen. In Tabelle 4.4 sind die Resultate der Parameteroptimierung unter der Bezeichnung σ_{FP-TK} aufgelistet, wobei das Subskript $FP-TK$ so zu verstehen ist, daß hier die Transferenergien von Fauchere und Pliska (FP) gegen die Oberflächen der mit TINKER (TK) erzeugten Strukturen angepaßt wurden.

Für die Parameteranpassung steht jedoch nur ein experimenteller Wert pro Aminosäure zur Verfügung; d.h. (abzüglich Glycin) nur 19 Bestimmungsgleichungen. Um eine für die Proteinstrukturvorhersage ausreichende Qualität der Lösungsmittelparameter zu gewährleisten, muß die Zahl der Freiheitsgrade möglichst niedrig gewählt werden. Wasserstoffatome erhalten daher aufgrund ihrer geringen Größe keinen eigenen Parameter. In ersten Näherung kann man davon ausgehen, daß jedem der vier Elemente (C, N, O, S) unabhängig von dessen Hybridisierung und Partialladung nur ein Lösungsmittelparameter zugeordnet werden muß. Von den 4 optimierten Werten aus sind die Parameter der verschiedenen Potentialtypen sukzessive nachoptimiert worden, um das Fehlerquadrat

$$\chi^2 = \sum_{X\in\{Ala,Arg,\ldots\}} \left(\Delta F_X - \sum_{i\in X} \sigma_{pt(i)} A_i \right)^2 \tag{4.33}$$

zu minimieren. Auf diese Weise konnte eine systematische Differenzierung der

Elemente nach ihrer Hybridisierung und Partialladung stattfinden. Dabei ist zu beachten, daß etwa für $n3$, welches nur in Lysin vorkommt, nur eine Bestimmungsgleichung vorliegt und diese einen utopischen Wert für σ_{n3} prognostiziert. Hier ist etwas Feingefühl notwendig, um die Lösungsmittelparameter in einem "natürlichen" Rahmen zu halten[11].

Schon in dieser Formulierung wurde eine deutlich Verbesserung des Kraftfeldes erzielt. Zum ersten Mal konnte eine Konkurrenz zwischen Wasserstoffbrückenbindungen, die zum damaligen Zeitpunkt rein elektrostatischer Natur waren, und dem hydrophoben Effekt beobachtet werden. In den Simulationen bildeten sich kompakte Strukturen mit hohem Sekundärstukturanteil aus, wohingegen das Lösungsmittelmodell des CARB-Kraftfeldes stets zu langen Helices und den damit verbundenen gestreckten Konfigurationen führte. Wie im nächsten Kapitel gezeigt wird, war die native Struktur von 1VII jedoch nicht das globale Minimum der durch das Kraftfeld beschriebenen Potentialenergieoberfläche.

Zur Stabilisierung der nativen Struktur sind einige der Lösungsmittelparameter variiert worden. Diese Veränderungen von σ_{FS-TK} waren zeitweise so groß, daß überprüft werden mußte, ob mit den jeweiligen Parameter die Transferenergien mittels Gleichung 4.32 noch reproduziert werden konnten. Bei den Aminosäuren, die hierbei die größten Abweichungen zwischen Experiment und theoretischer Vorhersage aufwiesen, zeigt sich, daß es prinzipiell möglich gewesen wäre, diesen Fehler durch Änderungen der Seitenkettenkonfiguration auf nahezu Null zu reduzieren! Dieser signifikante Einfluß der Peptidstruktur auf die Lösungsmittelparameter weist darauf hin, daß man bei der Auswahl der Konfiguration mit großer Sorgfalt vorgehen muß. Offenbar ist es notwendig, die Transferenergien genau den Strukturen gegenüberzustellen, die in wässeriger Lösung vorliegen, um die Lösungsmittelparameter möglichst frei von störenden Einflüssen zu halten.

Sharke und Rupley haben die Atomoberflächen strukturaufgeklärter Proteine untersucht und so die Oberflächen der Aminosäuren, so wie sie wahrscheinlich in wässeriger Lösung vorliegen, bestimmt [SR73]. Nach allgemein akzeptierter Auffassung bilden die Konfigurationen von Sharke und Rupley eine verhältnismäßig gute Grundlage für die Lösungsmittelparameteranpassung. Es bleibt offen, ob und in wie weit die elektrostatische Wechselwirkung zwischen Seitenketten und Hauptkette die Struktur eines Tripeptids beeinflußt. Es sollte in diesem Zusam-

[11] Eisenberg und McLachlan [EM86] haben sich auf 5 Potentialtypen beschränkt. Dadurch ist unter anderem die Transferenergie für Lysin nur sehr schlecht reproduziert worden. Ihre Beschränkung hat den Vorteil, daß den optimierten Parametern eine Standardabweichung zugeordnet werden kann. Für Kohlenstoff haben sie einen Wert von $\sigma_C = 16 \pm 2\, cal/(mol\text{Å}^2)$ bestimmt, also eine moderate Standardabweichung. Für das Stickstoff-Sauerstoff-Paar hingegen liegt der Wert bei $\sigma_{N/O} = -6 \pm 4\, cal/(mol\text{Å}^2)$. In der Arbeit von Pickett und Sternberg [PS93] (ebenfalls mit 5 Potentialtypen) liegen die Fehler sogar noch etwas höher: $\sigma_{N/O} = -7 \pm 7\, cal/(mol\text{Å}^2)$ und $\sigma_{O^-} = -8 \pm 16\, cal/(mol\text{Å}^2)$, also 200%. In unserer Rechnung mit 4 Potentialtypen fanden wir die größte Standardabweichung bei Stickstoff mit 180%, was unter anderem darin begründet liegt, daß das geladene $n3$ zunächst nicht von den polaren $n1$, $n2$ Potentialtypen unterschieden werden kann.

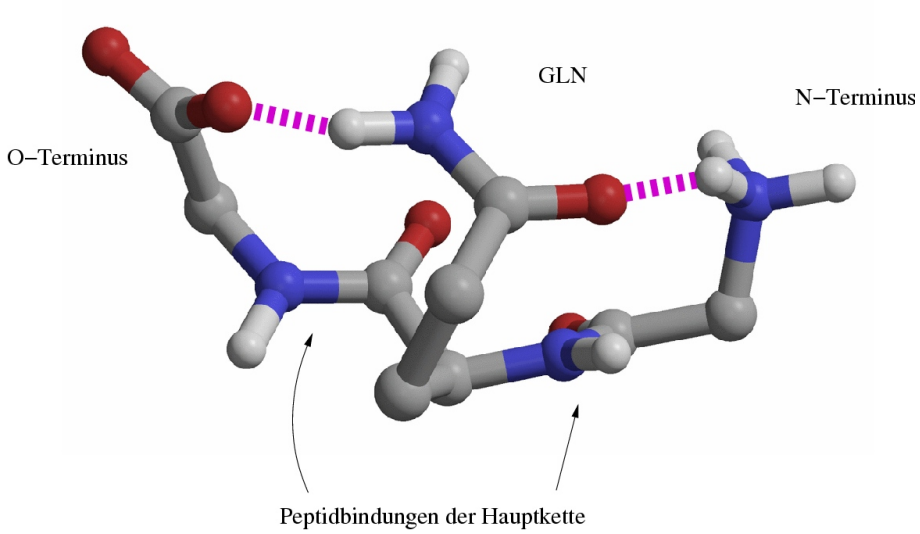


Abbildung 4.13: Eine Konfiguration des Tripeptids Gly-Gln-Gly mit starker elektrostatischer Seitenkette-Hauptketten Wechselwirkung

menhang nicht vergessen werden, daß die C- und N-Termini der Tripeptide (de-)protoniert sind und in unmittelbarer Nähe der Seitengruppe liegen. So ist etwa für Glutamin die Konfiguration in Abbildung 4.13 möglich.

Zusätzlich zum Übergang auf die neuen Konfigurationen haben wir auch berücksichtigt, daß die beiden Lösungsmittel n-Oktanol und Wasser unterschiedliche Volumina besitzen. Der Lösungsmittelparametersatz in Tabelle 4.4, welcher aus den Volumen-korrigierten Transferenergien und den Sharke und Rupley Oberflächen bestimmt wurde, trägt die Bezeichnung σ_{SH-SR}. Abbildung 4.14 zeigt den Vergleich zwischen den experimentellen Energien und den theoretischen Werten nach Gleichung 4.32. Dieser Parametersatz ist derjenige, der in der Kraftfeldversion PFF01 verwendet wird und zur Stabilisierung mehrerer Proteine beigetragen hat.

4.5.2 Konfigurationsentropie

Mit dem Ansatz von Eisenberg und McLachlan wird der Anteil zur Freien Energie des Protein/Wasser-Systems beschrieben, der als hydrophober Effekt bezeichnet wird und sich primär an der Entropie des Wassers orientiert. Dabei geht es um die Bewegungen der Wassermoleküle und die Ausbildung dynamischer Wasserstoffbrückenbindungen. In diesem Abschnitt behandeln wir die Mobilität der Proteins.

In einer random-coil Struktur in wässeriger Lösung unterliegen Seitenketten nicht nur kleineren Vibrationen, sondern wechseln zwischen energetisch ähnlichen Konfigurationen, deren Dihedralwinkel stark voneinander abweichen. Dies führt zu einem Beitrag in der Freien Energie, der als Konfigurationsentropie bezeichnet wird. Im Verlauf der Faltung werden einige Seitenketten vor dem Wasser abge-

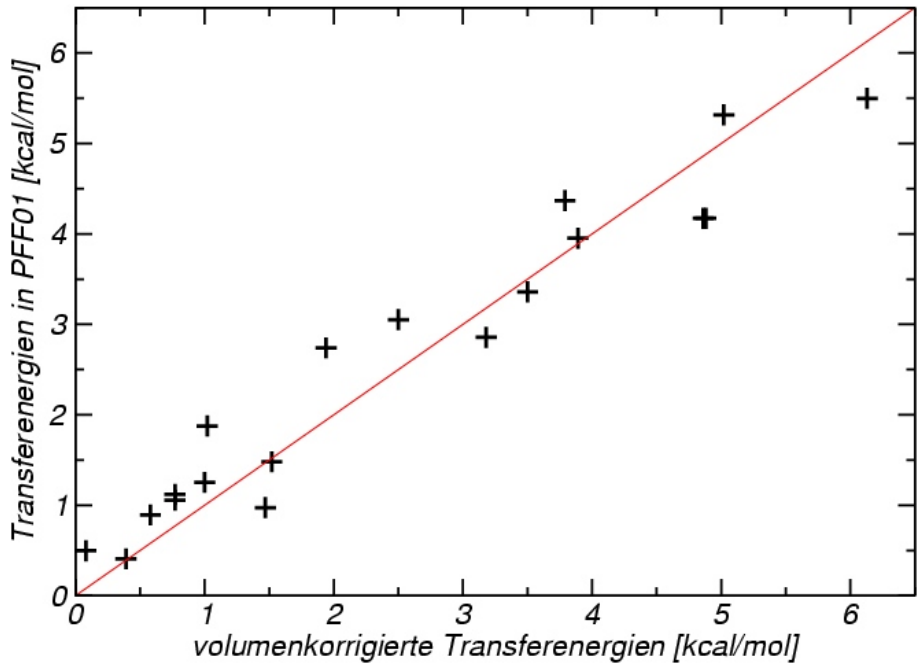


Abbildung 4.14: Vergleich der Volumen-korrigierten experimentellen Transferenergien und den nach Gleichung 4.32 berechneten Werten.

Potentialtyp	σ_{FS-TK}	σ_{SH-SR}
cme	45	84
cp	39	-6
cr	63	93
n1	-60	-30
n2	-60	-15
n3	-120	-45
o1	-30	-30
o2	-60	-15
s	30	84
h	*	*
hn	*	*

Tabelle 4.4: Liste der Lösungsmittelparameter in $cal\,mol^{-1}\,\text{Å}^{-2}$. Wasserstoffatome erhalten den Wert ihres Bindungspartners. In PFF01 werden die σ_{SH-SR} Parameter verwendet.

schirmt und verlieren aufgrund der Packungsdichte des Proteins die Möglichkeit, verschiedene Minima einzunehmen. Dieses Verhalten konnte an mehreren Proteinen beobachtet werden [SSRD91]. Die gleiche Argumentation kann auch auf die Hauptkette übertragen werden, deren Beweglichkeit so zu Null approximiert werden kann. Die Fähigkeit einer Seitengruppe, verschiedene Konfigurationen einzunehmen, scheint dabei verhältnismäßig sprunghaft einzusetzen, wenn die freie Oberfläche der Seitenkette circa 50% bis 60% der Oberfläche einer gestreckten Konfiguration hat [PS93]. Eine Seitengruppe benötigt demnach ein kritisches Mindestwasservolumen in seiner Umgebung, bevor es mehrere Konfigurationen einnimmt.

Da n-Oktanol ($C_8H_{17}O$) deutlich mehr Volumen und Masse hat als ein Wassermolekül (H_2O), ist die Beweglichkeit einer gelösten Peptidseitengruppe eingeschränkt. Auch wenn die Reduktion des Bewegungsspielraums durch n-Oktanol nicht die gleiche Stärke hat, wie sie der Innenbereich eines Proteins bietet, so ist dennoch anzunehmen, daß der gelösten Seitengruppe effektiv nicht ihr kritisches Mindestwasservolumen zur Verfügung gestellt wird. Die Konfigurationsentropie der Seitengruppe in n-Oktanol ist folglich gleich Null und in den Transferenergien von Fauchere und Pliska ist der Beitrag der Konfigurationsentropie zur Freien Energie der Faltung enthalten. Somit ist in unserem Lösungsmittelmodell die Konfigurationsentropie integriert, und zwar in linearer Approximation.

In einer Arbeit von Pickett [PS93] ist versucht worden, den Beitrag der Konfigurationsentropie aus den Transferenergien herauszurechnen. Dies ist jedoch eher für Molekulardynamik Simulationen von Interesse, die sich über ihre Trajektorien die Konfigurationsentropie erarbeiten und ansonsten diesen entropischen Beitrag doppelt zählen würden. Für Kraftfelder der Freien Energie sehen wir bislang keine Notwendigkeit die Konfigurationsentropie aus dem Lösungsmittelmodell herauszutrennen, um diese Größe dann durch ein anderes Modell wieder einzuführen, sondern begnügen uns stattdessen mit der integrierten linearen Näherung.

Kapitel 5

Faltungssimulationen

5.1 Einleitung

Proteinstrukturvorhersage geschieht derzeit ausschließlich durch sehr aufwendige experimentelle Verfahren. Von Computersimulationen erhofft man sich für die Zukunft eine Unterstützung, wenn es z.B. um Strukturen von Proteinen geht, deren Sequenz der mehrerer experimentell strukturaufgeklärter Proteins ähnelt. Biomolekulare Simulationen könnten so zwischen existierenden Daten interpolieren. Darüber hinaus wären Simulationen dieser Art auch bei anderen Fragestellungen nützlich, etwa zu Themen wie biologische Regelmechanismen, Transport, Molekülaffinität oder Aggregation von Proteinen.

Die Zielsetzung der Proteinstrukturvorhersage ist zum gegenwärtigen Stand der Forschung die Nachbildung wesentlicher struktureller Merkmale der nativen Konfiguration. Bei Helixproteinen sind dies in ersten Näherung Anzahl und Anordnung der Helices. Es wird in diesem Kapitel gezeigt, inwieweit es möglich war, diese Merkmale zu reproduzieren. Dabei wird stets auf die Aspekte der nativen Struktur aufmerksam gemacht, die in den Simulationen nicht vorhergesagt werden konnten.

Neben der Geometrie des Proteins ist auch die Freien Energie zu beachten. Die Beiträge der einzelnen Energieterme in den beschriebenen Approximationen des vorangegangenen Kapitels liegen mindestens eine Größenordnung höher als ihre Summe (Abbildung 2.9). Daher ist es schwierig, die richtige Balance zwischen den Energietermen zu finden. Die native Struktur zu stabilisieren bedeutet, daß die Differenz der Freien Energien zwischen der (relaxierten) nativen und jeder anderen Konfiguration mindestens 1 $kcal\,mol^{-1}$ beträgt. Ein generisches Kraftfeld muß diese Balance bei einer großen Zahl von Proteinen herstellen. Dies ist bislang niemandem zufriedenstellend gelungen. Die Gründe liegen zum Teil auf der Seite der Kraftfelder, aber auch auf Seite der eingesetzten Optimierungsverfahren.

Das ursprüngliche Kraftfeld, von dem aus PFF01 entwickelt wurde, ist am CARB (Centre for Advanced Research in Biotechnology) in der Gruppe um

John Moult entwickelt worden [Avb92, AM95, AJ95] und konnte dort mit gewissem Erfolg auf Proteinfragmente mit bis zu 18 Residuen angewandt werden [PM97a, PM97b]. Als Optimierungverfahren für Proteinfragmente wurde am CARB vorwiegend der genetische Algorithmus eingesetzt. Simulationen haben allerdings gezeigt, daß der genetische Algorithmus bei größeren Systemen das globale Minimum nicht zuverlässig identifiziert. Unser Ansatz war, durch Einsatz anderer/neuer Optimierungsverfahren die Proteingröße von den damals 18 Residuen deutlich zu erhöhen.

Wir wählten daher zunächst ein Protein mit mehr als 18 Residuen aus, welches schon von anderen Gruppen untersucht wurde. Das Protein 1VII mit 36 Residuen ist ein autonom faltendes Teilstück des Proteins 1QQV (*Chicken Villin Headpiece*) und ist nach einem gescheiterten Faltungsversuch von Duan und Kollman von besonderem Interesse [DK98, DWK98]. Bei unseren Simulationen im CARB-Kraftfeld mit verschiedenen Optimierungsverfahren konnten Konfigurationen mit einer Freien Energie unter der der relaxierten nativen Struktur generiert werden. Diesen Strukturen war gemein, daß sie aus ausgedehnten Helixkonfigurationen bestanden, die keinen hydrophoben Kern ausbildeten (Abbildung 5.1). Mit den Simulationen konnte gezeigt werden, daß das damalige CARB-Kraftfeld nicht für die Proteinstrukturvorhersage von (Helix-)Proteinen geeignet ist. Dieses Ergebnis steht nicht im Widerspruch zu den Resultaten der Fragmente, da diese wegen ihrer geringen Größe nicht in der Lage sind, Seitengruppen in einem hydrophoben Kern gänzlich vor dem Kontakt mit dem Wasser zu verbergen.

Die Schwachstellen des CARB-Kraftfeldes bestanden im Lösungsmittelmodell und in der Beschreibung der Wasserstoffbrückenbindungen. Diese Bestandteile sind von uns in mehreren Schritten überarbeitet worden, indem die strukturellen und energetischen Unterschiede der fehlgefalteten Konfigurationen zur nativen Struktur identifiziert und ursächlich beseitigt wurden. Im nächsten Abschnitt soll daher kurz erläutert werden, wie die Unterschiede zwischen zwei Strukturen herausgearbeitet werden können, um dann darauffolgend die Schritte der Kraftfeldoptimierung näher zu betrachten, die letzten Endes zu PFF01 geführt haben.

Es schließen zwei Versuche an, PFF01 auf weitere Proteine anzuwenden: 1F4I [Abschnitt 5.5] und 1L2Y [Abschnitt 5.6] (Abbildung 5.2). In beiden Fällen ist eine reproduzierbare Faltung gelungen. Bei der Simulation von 1F4I handelt es sich um die erstmalige reproduzierbare Faltung eines Proteins mit mehr als 20 Residuen in einem *all-atom* Kraftfeld mit physikalischen Wechselwirkungen. Nach diesem Erfolg verbleibt noch zu demonstrieren, daß das Kraftfeld auch andere Proteine stabilisieren kann. Mit dem in Abschnitt 5.7 vorgestellten Verfahren werden daher gezielt unterschiedlichste niederenergetische Strukturen dreier weiterer Proteine erzeugt: 1BDD (52 Residuen), 1ENH (54 Residuen) und 1GYZ (60 Residuen) [Abschnitte 5.8-5.10]. Es wird gezeigt, daß eine Stabilisierung dieser Proteine möglich war. Diese Ergebnisse zeigen, daß das Kraftfeld PFF01 für die Anwendung auf die gesamte Proteinfamilie der Helixproteine geeignet zu sein scheint.

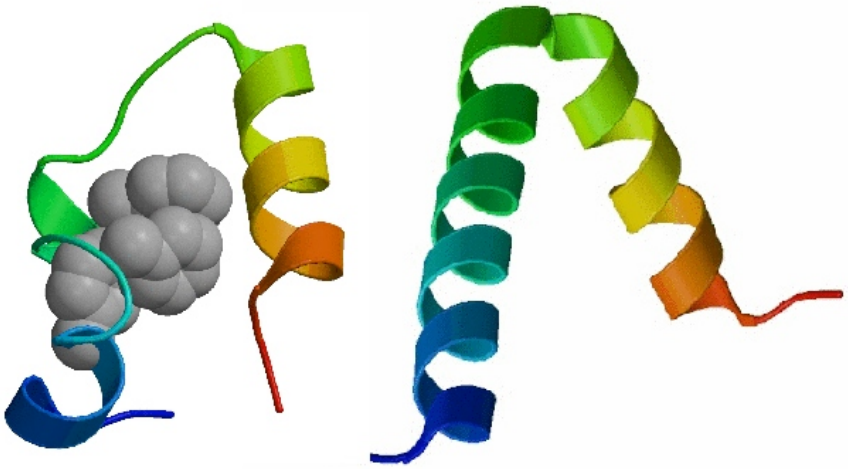


Abbildung 5.1: Tertiärstruktur und hydrophober Kern (Kugelkalotten) des Proteins 1VII (36 Residuen, links) und eine Konfiguration mit gemäß des CARB-Kraftfeldes niedrigerer Freien Energie

5.2 Analysetechniken

Es soll nun kurz skizziert werden, welche Verfahren zur Analyse der simulierten Strukturen eingesetzt wurden. Eine genaue Beschreibung dieser Verfahren findet sich in Anhang B. Die Methoden betrachten dabei entweder einzelne Strukturen, oder vergleichen zwei Konfigurationen miteinander und arbeiten komplementär zur dreidimensionalen Darstellung der Struktur.

Bei der separaten Betrachtung der Konfigurationen steht die Frage im Vordergrund, welche Aminosäuren an der Bildung von Helices beteiligt sind und welche in einer abweichenden Anordnung vorliegen. Man spricht hier von *Sekundärstrukturanalyse*. Das von uns diesbezüglich eingesetzte Programm trägt die Bezeichnung DSSP. Jeder Struktur wird so eine Buchstabenfolge zuordnen, bei der ein "H" an Position i besagt, daß Residuum i in eine Helix eingebunden ist. Im Kontext dieser Arbeit kann jeder andere Buchstabe als unstrukturierter Bereich interpretiert werden und wird zur besseren Unterscheidung mit einem Kleinbuchstaben notiert.

Die Sekundärstruktur gibt keine Auskunft über die räumliche Anordnung der Atome und auch bei identischer Sekundärstruktur können sich zwei Strukturen im dreidimensionalen Raum deutlich unterscheiden. Ein Maß für den Abstand zweier (Tertiär-) Strukturen ist die mittlere quadratische Distanz der Proteinatome, die *Root Mean Square Derivation*. Da wir primär an der räumlichen Ge-

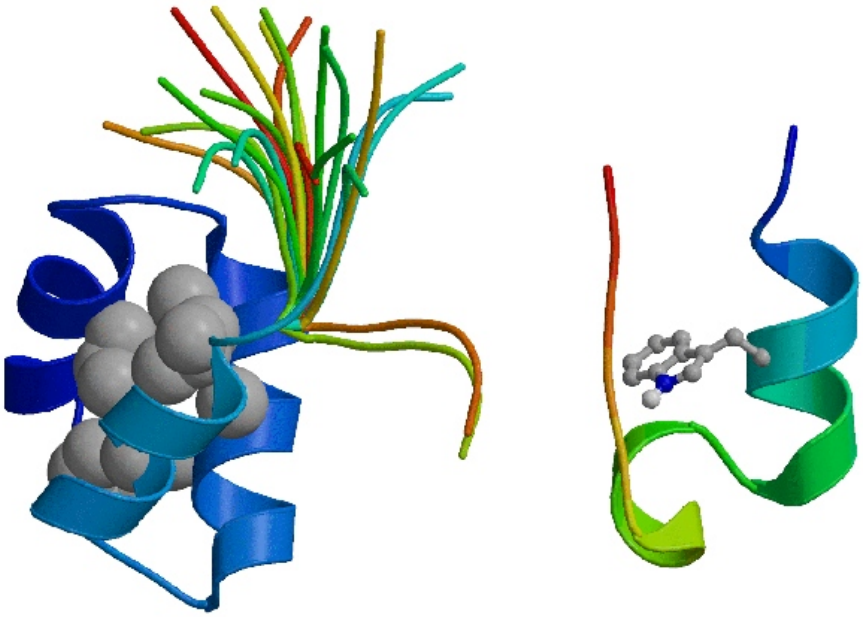


Abbildung 5.2: Tertiärstruktur und hydrophober Kern der Proteine 1F4I (40(+5) Residuen, links) und 1L2Y (20 Residuen, rechts). Für 1F4I sind 21 Strukturen in der Proteindatenbank PDB enthalten, die sich strukturell nur in den letzten 5 türkis bis rot dargestellten Aminosäuren unterscheiden. Diese 5 Aminosäuren werden in den Simulationen nicht berücksichtigt.

stalt der Hauptkette interessiert sind, wird der RMSD-Wert meist nur aus den Abständen der Hauptkettenatomen berechnet. Wir schreiben kurz RMSB statt $\text{RMSD}_{backbbone}$. Beide Abstandsbegriffe sind modulo Translationen und Rotationen der Proteine als Ganzes zu verstehen, d.h. vor der Berechnung der Zahlenwerte, werden die Proteine bestmöglich überlagert. Die natürlichen Fluktuationen der nativen Proteinstruktur bei 300K führen zu Konfigurationen mit einer RMSD-Abweichung von etwa 2.0Å [IKP88].

In den RMSB-Wert gehen die Positionen der Seitengruppen nicht ein, über deren Lage man sich anhand eines C_β-Mosaik einen Überblick verschaffen kann. In diesem Mosaik werden Unterschiede in den Abständen der C_β-Atomen farbig wiedergegeben. Sind in den beiden Strukturen die Abstände der C_β-Atome nahezu gleich, so wird die Mosaikposition schwarz eingefärbt; grau steht für einen mittleren Abstandsunterschied, weiß repräsentiert eine hohe Abweichung. Das C_β-Mosaik ist recht empfindlich gegen strukturelle Unterschiede und kann in Ergänzung zur grafischen Darstellung der Hauptkette bei der Identifikation der Strukturunterschiede beitragen.

In Anhang B findet sich neben den genauen Definitionen dieser Techniken auch ein erläuterndes Beispiel.

5.3 Verbreiterung des Faltungspfades

Schon die ersten Simulationen zeigten, daß eine reproduzierbare Strukturvorhersage mit den derzeit verfügbaren Rechnerresourcen an der Stärke der Lennard-Jones-Wechselwirkung scheitert. Eine erfolgreiche Faltung ist nur möglich, wenn die Simulation einen Pfad mittlerer und tiefer Energien von einer random-coil Struktur in die native findet. Im Niederenergiebereich, in dem nur kompakte Strukturen existieren, führt allerdings nahezu jede Konfigurationsänderung zu einem Überlapp der Atome und wird daher aufgrund einer extrem hohen Lennard-Jones Energie unterbunden. Dies gilt insbesondere für große Proteine und bei Simulationen im Dihedralwinkelraum, da eine Winkeländerung, den Proteinbereich bis zu einem der Termini im dreidimensionalen Raum dreht. Somit sind Faltungspfade bei großen Proteinen extrem schmal und die Wahrscheinlichkeit, daß ein Optimierungsverfahren diesem durch einen rein stochastischen Prozeß (d.h. ohne Gradienten) folgen kann, liegt nahe Null. Es erscheint notwendig, den Weg zum globalen Minimum der Potentialenergieoberfläche künstlich zu verbreitern, indem man den abstoßenden Teil des Lennard-Jones Potentials abschwächt. Da die Lennard-Jones Radien nicht verändert werden können, muß man daher die Potentialtiefe insgesamt reduzieren.

Damit die Energieunterschiede verschiedener niederenergetischer Proteinstrukturen nicht durch die Lennard-Jones-Wechselwirkung dominiert wird, war in den Simulationen eine Abschwächung des Lennard-Jones-Potentials bis auf $\epsilon = 0.01 kcal\, mol^{-1}$ notwendig. Mit diesen Werten war es möglich, mehrere Proteine

in ihrer nativen Struktur zu stabilisieren und teilweise zu falten. Die Gyrationsradien aller niederenergetischen Konfigurationen der bislang untersuchten Proteine stimmen mit den Gyrationsradien der nativen Strukturen gut überein.

5.4 Die Optimierung des Kraftfeldes an 1VII

Der auf dem Gebiet der Proteinfaltung berühmteste Faltungsversuch eines Proteins stammt von Duan und Kollman [DK98, DWK98]; er war nicht erfolgreich. Das betrachtete Protein war 1VII mit 36 Residuen. Die Simulationszeit belief sich auf 85 CPU-Jahre und entspricht 1 Mikrosekunde auf der natürlichen Zeitskala. Unter physiologischen Bedingungen liegt die experimentell beobachtete Faltungszeit allerdings ein bis zwei Größenordnungen darüber. So bleibt die Frage unbeantwortet, ob bei entsprechend langer Simulationszeit, eine nativ-ähnliche Konfiguration als globales Minimum der Freien Energie des AMBER Kraftfeldes ausgewiesen worden wäre.

Bislang ist es nur dem Folding@home-Projekt gelungen, eine Molekulardynamik Simulation auf einer Zeitskala durchzuführen, die der natürlichen Faltungszeit entspricht. Nach den bisher veröffentlichten Daten werden in den Simulationen die drei Helices der nativen Struktur nicht stabil ausgebildet, und das zugrundegelegte CHARMM Kraftfeld scheint nicht in der Lage, die native Struktur richtig vorherzusagen. Dabei sollte beachtet werden, daß die Fluktuationen der Struktur in einer Molekulardynamik Simulation bei 300K so groß sind, daß sich die "mittlere" vorhergesagte Struktur nur grob identifizieren läßt [SZP02]. Die Strukturvorhersagen von Kraftfeldern der Freien Energie liegt für 1VII bei über 5Å RMSB-Abweichung [LHH03]. Unser Kraftfeld ist an 1VII optimiert worden, und so liegt die RMSB-Abweichung deutlich unter dem Wert anderer Kraftfelder.

Die Optimierung des Kraftfeldes erfolgte durch den sogenannten *Decoy*-Ansatz. Hierbei werden in Simulationen niederenergetische Strukturen erzeugt und, soweit möglich, das globale Minimum der Freien Energieoberfläche der jeweils aktuellen Kraftfeldversion bestimmt. Findet das Optimierungsverfahren Strukturen, die unterhalb der Freien Energie der relaxierten nativen Struktur liegen, so ist gezeigt, daß das Kraftfeld fehlerhaft ist. Die wesentliche Aufgabe der Kraftfeldoptimierung besteht darin, den Anteil des Kraftfeldes zu identifizieren, der für die Fehlfaltung verantwortlich ist, und diesen zu korrigieren.

Um sicherzustellen, daß die am Kraftfeld vorgenommenen Änderungen nicht im Widerspruch zu bis dato gewonnenen Ergebnissen stehen, müssen nicht nur neue Simulationen durchgeführt werden, sondern auch die Freie Energie der Konfigurationen vergangener Simulationen ausgewertet werden. Die Gesamtzahl der Strukturen für 1VII liegt mittlerweile bei über $76,000$, und eine Energiebestimmung all dieser Strukturen würde eine zu große Zeitspanne in Anspruch nehmen. In diesem Zusammenhang hat es sich bewährt, einen *Decoy*-Satz anzulegen, in dem keine nahezu identischen Strukturen enthalten sind, d.h. die paarweise

RMSB-Abweichung liegt bei mindestens 1.0Å.

Die 14,000 Konfigurationen dieses Decoy-Satzes lassen sich nach ihren Freien Energien sortieren, und Strukturen unterhalb ein gewählten Energie lassen sich in Gruppen einteilen, die untereinander maximal 3.0Å voneinander abweichen. Trägt man diese Gruppierung gegen die Energie auf, so entsteht eine baumartige Struktur, wie sie in Abbildung 5.3 wiedergegeben ist. Die Entstehung eines Decoy-Baumes aus einer bzw. mehreren Faltungssimulationen läßt sich folgendermaßen nachvollziehen. Die Hochtemperaturkonfigurationen sind energetisch sehr schlecht, decken allerdings durch ihre große Anzahl einen großen Bereich des Konfigurationsraumes ab und bilden so den Stamm des Decoy-Baumes. Werden die Strukturen während der Optimierungsphase abgekühlt und erreichen so niedrigere Energien, pflanzt sich der Baum nach unten hin fort. An gewissen Stellen müssen sich die Strukturen entscheiden, in welches Energietal sie hineinlaufen wollen. Die Energietäler werden durch die Äste des Baumes repräsentiert. Die Astspitzen sind die energetisch tiefsten Punkte der Energietäler.

Den Verzweigungsstellen der Äste kommt dabei im Prinzip eine wichtige Bedeutung zu. Wenn eine Simulation ein Energietal identifiziert hat (z.B. Struktur C in Abbildung 5.3), muß die Energie dieser Konfiguration erst auf den Wert der Übergangsenergie angehoben werden, bevor die Struktur in einer nachfolgenden Abkühlung in ein neues Minimum (A, B, M oder N) übergehen kann. Die Energien, bei denen die Äste zusammenwachsen, geben allerdings nur eine obere Schranke für die Übergangsenergien an, da dieser Baum nur aus den Decoy-Strukturen gewonnen wurde und keine separate Bestimmung der Übergangsenergien stattgefunden hat. Es sei angemerkt, daß die Anzahl der Strukturen in den verschiedenen Decoy-Gruppen keine Aussagekraft hat und nicht in der Wahrscheinlichkeit der thermodynamischen Verteilung $\rho \sim \exp(-\beta H)$ auftreten, da hier Strukturen aus sehr vielen Simulationen, darunter auch Rechnungen mit anderen Kraftfeldvarianten, zusammengetragen wurden.

Bei der Kraftfeldoptimierung kann man in erster Näherung davon ausgehen, daß die energetisch niedrigsten Mitglieder einer Decoy-Gruppe (Astspitzen) die wesentlichen Merkmale der gesamten Gruppe aufweisen, und sich daher zunächst auf die Betrachtung der Minima beschränken.

Der Baum in Abbildung 5.3 entstand nach Simulationen ohne das in Abschnitt 4.4 diskutierte Korrekturpotential – das heißt, mit einer rein elektrostatischen Beschreibung der Wasserstoffbrückenbindungen (Seite 72) – und mit den Lösungsmittelparametern σ_{FS-TK} (Seite 87). Unter dem Baum sind die Strukturen der Minima, sowie die native Struktur, unter der Bezeichnung der experimentellen Methode mit der sie bestimmt wurde (NMR), dargestellt. Die Struktur N ist die relaxierte native Struktur. Sie ist nicht das globale Minimum der Freien Energie. Dieses ist (wahrscheinlich) durch die Struktur M gegeben, deren Helices anders angeordnet sind. Dieser Baum und die zugehörigen Strukturen stellen ein wichtiges Zwischenergebnis dar. Hier war es zum ersten Mal gelungen, einen Wettbewerb zwischen hydrophoben Effekt und Sekundärstrukturbildung durch

Wasserstoffbrückenbindungen zu beobachten. So ist durchaus beachtlich, daß die niederenergetischen Strukturen allesamt drei Helices ausbilden und längere unstrukturierte Abschnitte besitzen, wie es auch in der nativen Struktur zu sehen ist. In diesem Stadium der Kraftfeldentwicklung war die relaxierte native Struktur ein ausgeprägtes Minimum der Potentialenergieoberfläche und gleichzeitig eine der energetisch niedrigsten Konfigurationen.

Es ist mit der damaligen Kraftfeldparametrisierung nicht gelungen, die nativen Struktur von 1VII als globales Minimum der Potentialenergieoberfläche auszuweisen. In den Simulationen wurden Konfigurationen gefunden, deren Freie Energie unterhalb der relaxierten nativen Struktur lagen. Die Energieterme, die diese Strukturen gegenüber der nativen begünstigten, waren der Lösungsmittelbeitrag und die Wasserstoffbrückenbindungsenergie. Beide Kraftfeldbestandteile sind überarbeitet worden und haben zu einer deutlichen Verbesserung geführt, in dem Sinne, daß nur noch sehr wenige Strukturen energetisch unterhalb der relaxierten Struktur lagen.

Die beiden Parametersätze für den Lösungsmittelbeitrag σ_{FS-TK} und σ_{SH-SR} weichen signifikant voneinander ab und sind jeweils mit großen Fehlern versehen. Wie im Abschnitt zur Parameteranpassung beschrieben, existiert eine starke Abhängigkeit der Lösungsmittelparameter von den zugrundegelegten Referenzkonfigurationen und den experimentell bestimmten Transferenergien. Es wurde gezeigt, daß die Wahl der Referenzkonfigurationen problematisch ist, und die Transferenergien nicht konsistent mit anderen Hydrophobizitätsskalen sind. Wir sind daher davon ausgegangen, daß in den Lösungsmittelparametern noch ein systematischer Fehler enthalten sein könnte.

Um diesem Umstand Rechnung zu tragen, wurde in das Lösungsmittelmodell für jede Aminosäure $res \in \{Ala, Arg, \ldots\}$ ein Multiplikator π_{res} eingeführt. Mit diesem Faktor werden die Parameter der jeweiligen Aminosäure multipliziert. Der Ansatz von Eisenberg und McLachlan (Gleichung 4.32) geht dann über in

$$\Delta F = \sum_i \sigma_{pt(i)} \pi_{res(i)} A(i) \; . \tag{5.1}$$

Mit diesem Ansatz kann das spezifische Verhalten einzelner Aminosäuren besser berücksichtigt werden. Die Faktoren, die zu einer Stabilisierung der relaxierten nativen Struktur von 1VII geführt haben, sind in Tabelle 5.1 wiedergegeben. Den Faktor der hydrophoben Seitengruppen von $\pi = 0.8$ verstehen wir so, daß die elektrostatische Wechselwirkung auf die partialladungsfreien hydrophoben Seitengruppen keinen Einfluß nimmt, welche im Gegensatz dazu für partial geladene Seitengruppen ein Beitrag zur Transferenergie bzw. Referenzkonfiguration liefert (Abbildung 4.13) und so in einem systematischen Unterschied zwischen den geladenen und ungeladenen/hydrophoben Seitengruppen mündet. Bei Tryptophan liegt der π-Faktor 25% unter dem Wert für die anderen hydrophoben Seitengruppen. Diese Abweichung ist sehr stark, war jedoch notwendig, um die

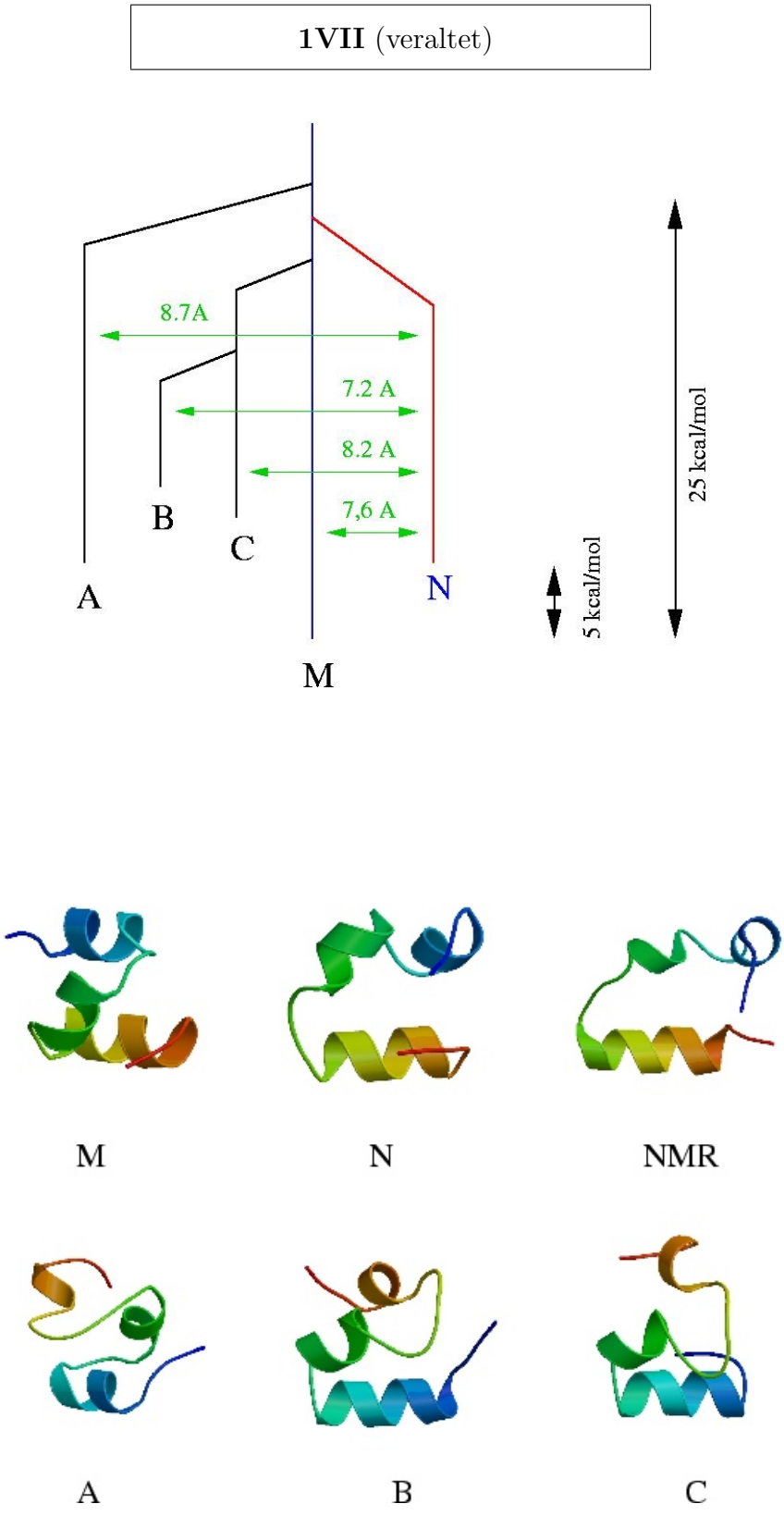


Abbildung 5.3: Der 1VII-Baum vor PFF01

π	Aminosäure
0.8	Ala, Ile, Leu, Met, Phe, Val
0.6	Trp
1.0	sonst

Tabelle 5.1: Faktoren der einzelnen Aminosäuren in Gleichung 5.1.

native Struktur von 1VII zu stabilisieren. Tryptophan ist nach den Volumenkorrigierten Transferenergien diejenige Aminosäure mit dem größten Unterschied zwischen Proteininnerem und Proteinoberfläche. Es ist erwähnt worden, daß diese Aminosäure auf verschiedenen Hydrophobizitätsskalen sehr unterschiedlich eingestuft wird. Wir vermuten daher, daß der experimentelle Wert der Transferenergie dieser Aminosäure als Tripeptid ihren hydrophoben Charakter in einem Protein nicht richtig wiedergibt.

Mit diesem letzten Endes sehr begrenzten Eingriff in das Lösungsmittelmodell war es möglich, eine relaxierte native Struktur als globales Minimum der Potentialenergieoberfläche zu identifizieren. An dieser Stelle ist die Kraftfeldoptimierung beendet worden, und das Kraftfeld PFF01 war damit vollständig festgelegt.

Trägt man alle Simulationsdaten für 1VII zusammen, so ergibt sich der in Abbildung 5.4 wiedergegeben Decoy-Baum. In diese Abbildung ist die Energieskala und die Anzahl der Decoys als Funktion der Energie hinzugefügt worden. In der darunterstehenden Tabelle sind die Energien zusammen mit den RMSB-Abweichungen zur nativen Konfiguration und der Sekundärstruktur aufgelistet. Die oberste Sekundärstruktur ist die der nativen Konfiguration. Der Buchstabe zwischen der RMSB-Abweichung und der Sekundärstruktur gibt die Bezeichnung im Baumdiagramm wieder. Auf der nachfolgenden Seite sind die Strukturen und das zugehörige C_β-Mosaik (Anhang B.3.2) der relaxierten Struktur N gegen die native abgebildet.

Die niedrigste in Simulationen gefundene Freie Energie ist die einer nativähnlichen Struktur mit einer RMSB-Abweichung von 3.56Å. Vergleicht man die native und die Struktur N etwas genauer, so erkennt man am C_β-Mosaik, daß es einen Bereich gibt, in dem die Konfiguration N keine nativen Kontakte ausbildet. Der Grund hierfür liegt möglicherweise in der Fehleinschätzung der Seitenketten-Elektrostatik in PFF01. Die beiden kurzen Helices sind in der nativen Struktur durch eine Wasserstoffbrücke zwischen der Seitengruppe des Arganin 15 und des Hauptketten-CO-Gruppe des Leucin 2 miteinander verbunden, wohingegen das Arganin 15 in der Struktur N ins Lösungsmittel ragt. Dieses Detail wird in PFF01, wie in faktisch allen Simulationen mit anderen Kraftfeldern, nicht richtig wiedergegeben. Die Charakteristika der nativen Struktur in Bezug auf die Hauptkette sind jedoch sehr gut in der relaxierten Konfiguration wiedergegeben, ebenso wie der hydrophobe Kern von 1VII, der im Wesentlichen aus drei Phenylalanin besteht (Abbildung 5.1). Der Energieunterschied zwischen der nativ-ähnlichen

Struktur N und der energetisch zweitplazierten Struktur B liegt bei $1\,kcal\,mol^{-1}$ und ist knapp ausreichend, die nativ-ähnliche Struktur zu stabilisieren.

Es ist uns bislang nur einmal gelungen, eine native-ähnliche Struktur in einer Faltungssimulation, also von einer random-coil Struktur aus, zu finden. Obwohl der Decoy-Baum eine gewisse Trichterstruktur (engl. *funnel*) aufweist, ist er im Niederenergiebereich sehr breit, und obwohl die Verbindungspunkte der Äste nur eine obere Schranke für die Energiebarriere zwischen den Ästen ist, so ist aufgrund der hohen Zahl an verfügbaren Strukturen anzunehmen, daß die Übergangsenergie nur wenige $kcal\,mol^{-1}$ unter den angegebenen Energien liegt. Somit ist es in den Simulationen äußerst schwer, aus einem der Minima in ein anderes zu wechseln.

Alle Astspitzen besitzen einen mit der nativen Struktur vergleichbaren Sekundärstrukturanteil. A, C und E bilden zwei Helices und N, B, D und F, in Übereinstimmung mit der nativen Struktur, drei Helices. Dies ein ein signifikanter Unterschied zu früheren Simulationen mit AMBER [DK98, DWK98] oder der aktuellen Simulation mit ECEPP [LHH03], in denen nur selten drei Helixkonfiguration gefunden wurde.

Eine weitere Optimierung des Kraftfeldes allein an 1VII – etwa um den Abstand der nativen Struktur zu der im Kraftfeld relaxierten Konfiguration zu verringern, erscheint nicht sinnvoll, da wahrscheinlich eher 1VII-spezifische Details in das Kraftfeld integriert würden, die bei der Simulation an anderen Proteinen wieder entfernt werden müßten.

5.5 Die Faltung des HIV-accessory Proteins 1F4I

Nachdem das Kraftfeld für ein Protein optimiert wurde, ist es notwendig, dieses auf mindestens ein weiteres System anzuwenden, um den generischen Charakter des Kraftfeldes zu testen. Ohne weitere Parameterveränderungen wurde mit PFF01 das Protein 1F4I simuliert. In der nativen Struktur von 1F4I bilden die Residuen 41 bis 45 keine stabile Struktur aus, sondern stehen in beliebiger Richtung von den anderen Residuen ab (Abbildung 5.2). Daher wurden diese Residuen entfernt und nur der strukturierte Bereich der ersten 40 Aminosäuren betrachtet. Ein Kontakt einer Seitengruppe mit einer anderen oder mit der Hauptkette, wie in 1VII, existiert nicht.

Zur Faltung von 1F4I wurde erstmals eine spezielle Strategie verwendet, die eine Simulation von einer random-coil Struktur zum globalen Minimum in mehrere Phasen aufteilt. Auf den konzeptionelle Hintergrund dieses Verfahrens wird in einem späteren Abschnitt gesondert eingegangen. An dieser Stelle sei nur kurz der Simulationsablauf skizziert. Ausgangspunkt der Faltung bilden 20 random-coil Strukturen, die in BHT-Simulationen (Basin-Hopping-Technique) bei 20%-iger Lösungsmittelreduktion von 800 auf $300K$ abgekühlt wurden. Die TA-Akzeptanzschwelle wurde auf $15\,kcal\,mol^{-1}$ gesetzt. In einer zweiten Phase erhielt das

Lösungsmittel seine volle Stärke zurück. Die 20 Endstrukturen dieser Simulationen wurden in einer dritten Phase der Optimierung durch die BHT Methode unterzogen, d.h. die Endtemperatur eines SA Laufen lag bei 1K. Jede Simulation durchlief insgesamt 10^7 Monte-Carlo Schritte.

Damit wurden 20 verschiedene random-coil Strukturen unabhängig voneinander simuliert. Am Ende der Simulation wurden die 20 Strukturen gemäß ihrer Energie sortiert, und die fünf energetisch niedrigsten Strukturen hatten eine RMSB-Abweichung zur nativen Struktur von weniger als 3.00Å. Die mit $-119.54\,kcal\,mol^{-1}$ beste Struktur hatte eine RMSB-Abweichung von nur 2.34Å. Dies ist die erste Simulation eines *all-atom* Kraftfeldes, basierend auf physikalischen Wechselwirkungen, in der reproduzierbar (in 5 von 20 Simulationen) ein Protein mit mehr als 20 Residuen gefaltet wurde.

Der zugehörige Decoy-Baum hat sich auch nach weiteren Simulationen kaum verändert und ist in Abbildung 5.6 zusammen mit der Liste der Energien, RMSB-Abweichungen und den Sekundärstrukturen zu sehen. Es fällt auf, daß die erste Struktur, die mehr als 3Å von der Konfiguration N abweicht und daher im Baum einen eigenen Eintrag erhält, mit $F = -114.06\,kcal\,mol^{-1}$ schon $5\,kcal\,mol^{-1}$ über der Energie von N liegt[1]. Erst weitere $5\,kcal\,mol^{-1}$ später kommt eine neue Decoy-Gruppe hinzu.

Die ausgeprägte Trichterstruktur in der Potentialenergieoberfläche von 1F4I hat sehr wahrscheinlich zur raschen und reproduzierbaren Faltung von 1F4I beigetragen. Dennoch scheinen die Energietäler A, B von C, N, D und E, F durch hohe Energiebarrieren getrennt. Dies könnte erklären, warum nicht nahezu alle 20 Simulationen die Struktur N gefunden haben.

5.6 Das Trp-Cage Protein 1L2Y

Diese Protein ist im Labor synthetisiert worden und erfüllt daher keine biologisch Funktion. Mit 20 Aminosäuren ist dies das kleinste Protein, welches einem Zwei-Zustands-Faltungsmechanismus folgt, daß heißt, es ist entweder vollständig gefaltet oder entfaltet [SSR02]. Wie in Abbildung 5.8 zu sehen, bildet es zwei kurze Helices und ein unstrukturiertes Endstück aus. Eine Besonderheit in der nativen Struktur ist, daß der hydrophobe Kern faktisch nur aus Tryptophan 6 besteht, dessen Seitenketten-NH-Gruppe eine Wasserstoffbrücke zur Hauptketten-CO-Gruppe des Arganin 16 bildet. Diesem Umstand verdankt das Protein 1L2Y den Namen "Trp-Cage".

1L2Y konnte im vergangenen Jahr in Molekulardynamik Simulationen mit

[1]Es wäre falsch anzunehmen, daß es unter den mittlerweile 60,000 Strukturen für 1F4I keine Strukturen mit Energien zwischen $-119.54\,kcal\,mol^{-1}$ und $-116.25\,kcal\,mol^{-1}$ gebe. Es ist vielmehr so, daß diese Strukturen von der Struktur N nicht mehr als 1Å abweichen, und daher nicht in den Decoy-Satz aufgenommen sind. Für höhere Energien gilt das entsprechende Argument, so daß sich die 60,000 Strukturen nur in 4,500 Decoys gruppieren.

dem CHARMM [SZP02] und AMBER [SSR02] Kraftfeld gefalten werden. Wobei zu erwähnen ist, daß das Strukturelement, des hydrophoben Kerns und der Wasserstoffbrücke des Tryptophan 6 nicht (CHARMM) oder nur in wenigen Simulationen (AMBER) aufgetreten ist. Unsere Simulationen in PFF01 sind mit dem Molekulardynamik Simulationen qualitativ vergleichbar, wobei auch wir den Trp-Käfig nicht in allen Simulationen vorhersagen konnten, insbesondere die Konfiguration des globalen Minimums bildet diesen Käfig nicht aus (Abbildung 5.8). Dies steht wahrscheinlich im Zusammenhang mit den schon früher zu beobachtenden Problemen, welche bei dieser Aminosäure aufgetreten sind, und im Lösungsmittelmodel zu einem π-Faktor von 0.6 geführt haben (Gleichung 5.1 und Tabelle 5.1).

Den experimentell nachgewiesenen Zwei-Zustands-Faltungsmechanismus konnten wir insofern beobachten, als daß es keine niederenergetische Struktur von 1L2Y mit größerer RMSB-Abweichung zur relaxierten nativen Struktur gibt. Die Simulationszeit im CHARMM Kraftfeld beläuft sich auf 250 CPU-Jahre. Mittels stochastischer Optimierung im PFF01 Kraftfeld mit 25 Simulationen war insgesamt nur ein halbes CPU-Jahr bis zur vollständigen Konvergenz aufzubringen [SHW03]. Diese Zahlen sind ein starkes Indiz dafür, daß im Bereich der Proteinstrukturvorhersage stochastische Optimierungsverfahren den Molekulardynamik Simulation überlegen sind.

Der Zwei-Zustands-Faltungsmechanismus hat dazu geführt, daß wir dieses Protein praktisch immer und mit sehr geringem Rechenaufwand falten können. Das Fehlen einer niederenergetischen fehlgefalteten Struktur macht dieses Protein jedoch eher unattraktiv für die Untersuchung von Optimierungsstrategien.

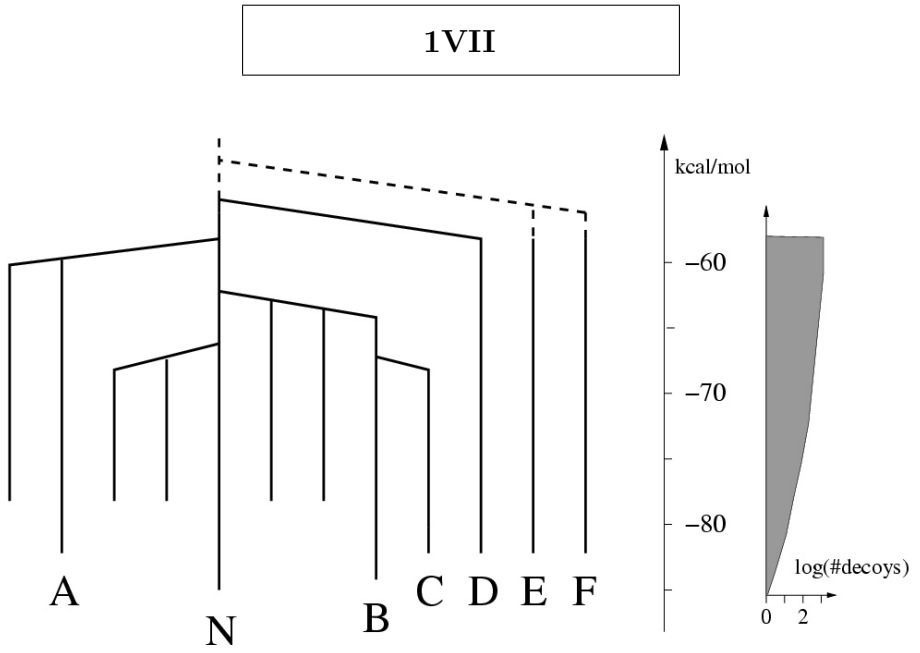


Abbildung 5.4: Der 1VII-Baum für PFF01

F	RMSB		Sekundärstruktur
$\frac{kcal}{mol}$	$[\AA]$		cccHHHHHttsssсHHHHttscHHHHHHHHHHttcc
-85.14	3.56	N	cccHHHHHHHHtscHHHHHHscHHHHHHHHHHHtcc
-84.11	6.36	B	cccHHHHHHtHHHHHHHHHHssscctttHHHHHHHc
-83.54	7.27	C	cHHHHHHHHHssscсsscscHHHHHHHHHHHHHtcc
-83.17	5.96	E	cHHHHHHHHHtssscsccssHHHHHHHHHHHHHtcc
-83.10	6.29		cccHHHHHHcHHHHHHHHHHssscсHHHHHHHHHHc
-82.59	6.40		cccHHHHHHcHHHHHHHHHHssscсHHHHHHHHtcc
-82.43	6.14	D	cHHHHHHHHtHHHHHHHHHHssscttтcHHHHHttc
-82.28	5.80	F	cHHHHHHcHHHHHHHHHHcHHHHHHHHHHHHHHHc
-82.17	6.44		cccHHHHHHcHHHHHHHHHHssscсHHHHHHHHHHc
-82.03	7.85	A	cccHHHHHHHHtscHHHHHHscttttsssssscctttc
-82.01	4.02		cccHHHHHttttccHHHHHHscHHHHHHHHHHcccc
-81.73	8.21		cccssccscttтHHHHtcccccsscHHHHHHHHtcc
-81.72	6.89		cccsHHHHHHHHHHHHHHHHssctтssccHHHHtcc
-81.49	4.85		cccHHHHHHHcHHHHHHHHsHHHHHHHHHHHHHtcc
-81.46	6.46		ccccHHHHHHHHHHHHHHcHHHHHHHHHHHHHHtcc
-81.35	7.62		cccHHHHHHHtssccsscscHHHHHHHHHHHHHtcc
-81.16	5.49		cccHHHHHHHHHHHHHHHHHscccscHHHHHHHHHc
-81.06	6.42		cccHHHHHHcHHHHHHHHHHssscсHHHHHHHHHHc
-80.98	7.74		cccHHHHHHHtssscsccssHHHHHHHHHHHHHtcc
-80.84	6.62		cccHHHHHHcHHHHHHHHHHssscсHHHHHHHHHHc
-80.79	4.46		cccHHHHHHHcHHHHHHHHsHHHHHHHHHHHHHtcc
-80.75	7.51		cHHHHHHHHHHtscHHHHsscsssctttccscsttc

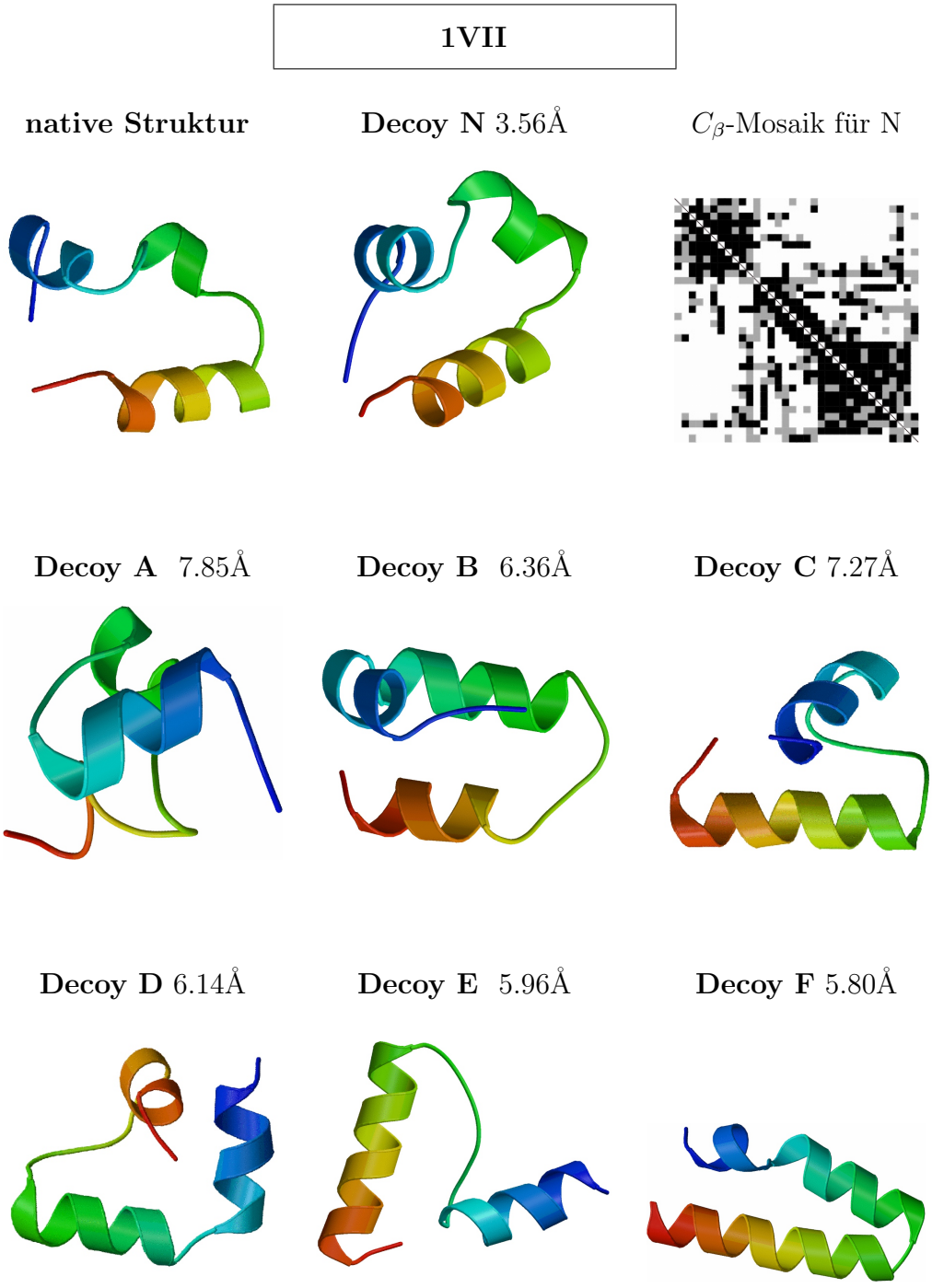


Abbildung 5.5: Die niederenergetischen Strukturen des 1VII-Baumes für PFF01

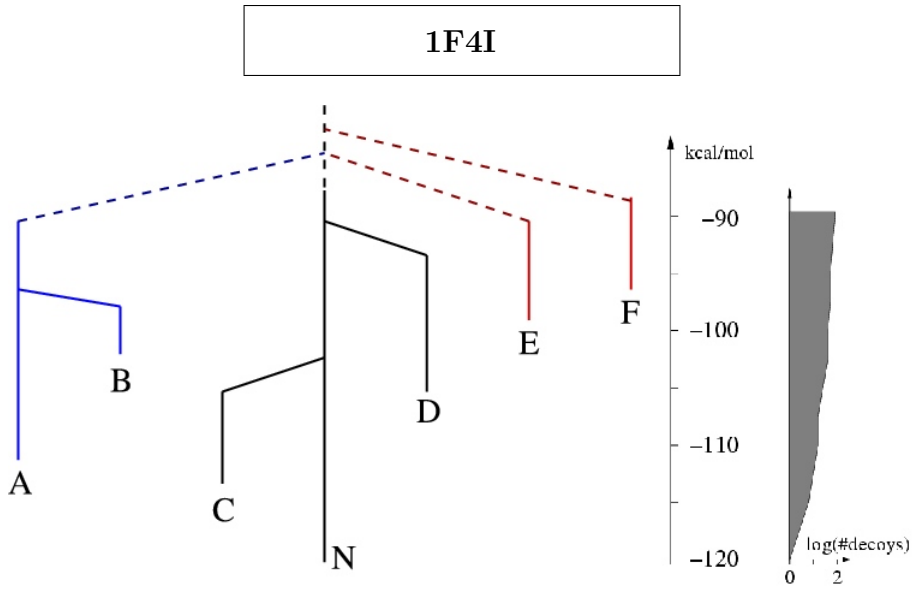


Abbildung 5.6: Der 1F4I-Baum für PFF01

F $\frac{kcal}{mol}$	RMSB $[\AA]$		Sekundärstruktur ccHHHHHHHHHHttccHHHHHHHHHHtttscsHHHHHHHHHHc
-119.54	2.34	N	cHHHHHHHHHHHHtccHHHHHHHHHHHHHHcHHHHHHHHHHc
-117.52	2.41		cHHHHHHHHHHHsscccHHHHHHHHHHHHHHcHHHHHHHHHHc
-116.25	2.76		cHHHHHHHHHHHHsccHHHHHHHHHHHHHHcHHHHHHHHHHc
-116.12	1.92		cHHHHHHHHHHHHsccHHHHHHHHHHHHHHcHHHHHHHHHHc
-115.63	2.30		cHHHHHHHHHHHHHccHHHHHHHHHHHHHHcHHHHHHHHHHc
-114.67	2.43		cHHHHHHHHHHHHtccHHHHHHHHHHHHcsHHHHHHHHHHc
-114.14	1.53		cHHHHHHHHHHHHHscHHHHHHHHHHHHtscHHHHHHHHHHc
-114.06	6.48	C	cHHHHHHHHHHHHssssHHHHHHHHHHHHHcHHHHHHHHHHc
-113.06	1.50		cHHHHHHHHHHHHcccHHHHHHHHHHHHtscHHHHHHHHHHc
-112.90	2.26		cHHHHHHHHHHHHtccHHHHHHHHHHHHssHHHHHHHHHHc
-112.89	6.66		cHHHHHHHHHHHHcsssHHHHHHHHHHHHHcHHHHHHHHHHc
-112.40	2.16		cHHHHHHHHHHHHHccHHHHHHHHHHHHHHcHHHHHHHHHHc
-112.02	3.70		cHHHHHHHHHHHHHccHHHHHHHHHHHHHHcHHHHHHHHHHc
-111.20	2.85		cccHHHHHHHHHHcccHHHHHHHHHHHHHHcHHHHHHHHHHc
-110.86	4.61	A	cHHHHHHHHHHHHtccsscHHHHHHHHHHcHHHHHHHHHHc
⋮	⋮		⋮
-103.88	5.92	D	cHHHHHHHHHHHHccHHHHHHHHHcsscccccHHHHHHHHHc
-102.86	7.55	B	cHHHHHHHHHHHHtscHHHHHHcttsscscHHHHHHHHHHc
-97.99	5.77	E	cHHHHHHHHHHHHtcssscgggttscscscHHHHHHHHHHc
-93.17	6.44	F	cHHHHHHHHHHHHcHHHHscccsscsssssccHHHHHHHHHc

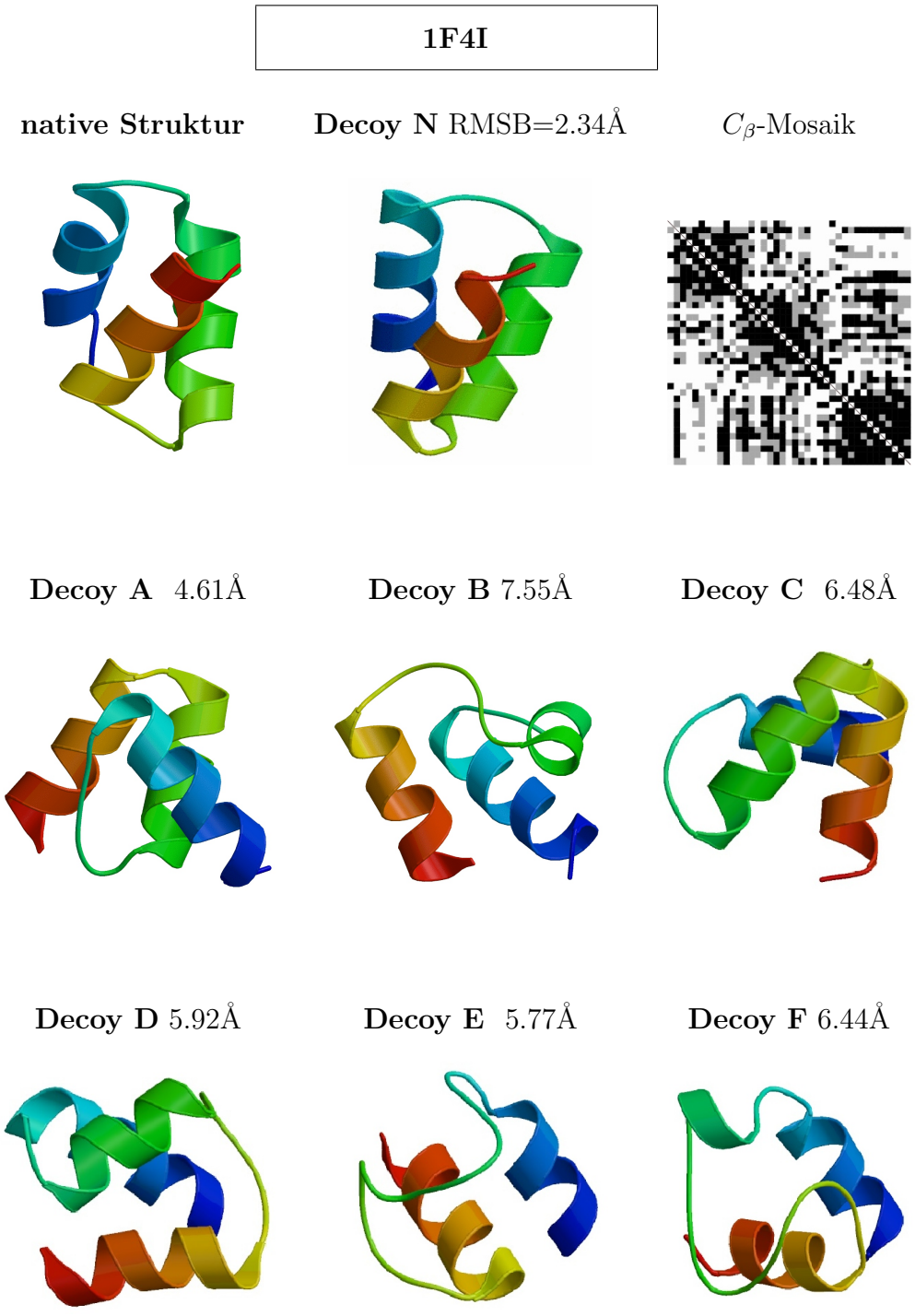


Abbildung 5.7: Die niederenergetischen Strukturen des 1F4I-Baumes für PFF01

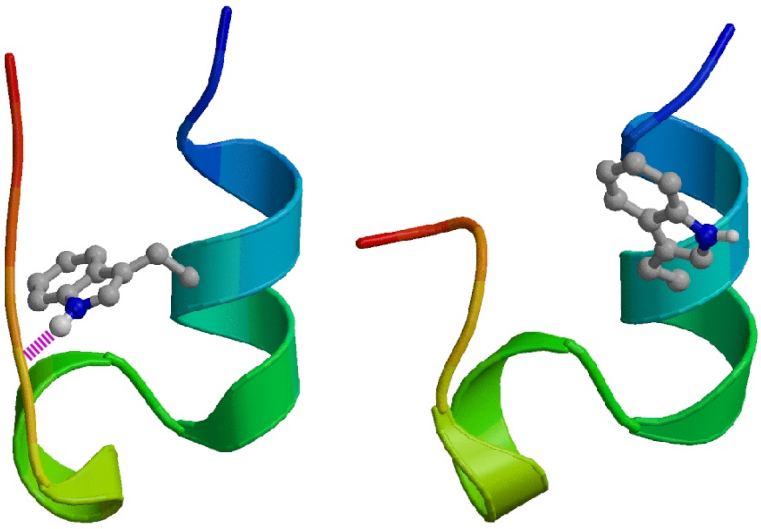


Abbildung 5.8: Vergleich der nativen Struktur von 1L2Y (links) und der in PFF01 gefalteten Konfiguration(rechts)

5.7 Analyse der Hochtemperatur-Simulationen

Die bei 1F4I erstmals eingesetzte Strategie, random-coil Strukturen einige Zeit bei hohen Temperaturen zu simulieren und erst anschließend lokal zu optimieren, soll in diesem Abschnitt näher betrachtet werden, da diese auch bei anderen Proteinen angewandt wurde. In seiner derzeit eingesetzte Formulierung weicht das Verfahren leicht von der 1F4I-Variante ab und enthält weniger freie Parameter.

Jeder Faltungsversuch, beginnend von einem Satz mit circa 20 bis 50 unterschiedlichen random-coil Strukturen, besteht aus zwei Phasen. In der Ersten wird bei einer konstanten Temperatur simuliert, da bei einer Abkühlung auf tiefe Temperaturen die random-coil Strukturen direkt in das nächstgelegenen lokalen Minimum hineinlaufen und dort verbleiben würden. Die Simulationstemperatur ist mit 500 bis $700K$ recht hoch gewählt, um große strukturelle Veränderungen zu ermöglichen. Es sei daran erinnert, daß die Simulationstemperatur nicht die physikalische Temperatur des Protein/Wasser-Systems ist [Abschnitt 3.1.5]. Jede dieser Simulationen umfaßt im allgemeinen ein bis zwei Millionen Monte-Carlo Schritte und dauert für die Protein 1VII und 1F4I ein bis zwei Tage. Während dieser Zeit werden in regelmäßigen Abständen die jeweils aktuellen Konfigurationen gespeichert. In der zweiten Phase erfolgt die Identifikation lokaler Minima durch die BHT.

Es gelangen jedoch nicht die Endkonfigurationen des ersten Simulationsabschnittes in die zweite Phase, sondern ein Teil der aufgezeichneten Zwischenstrukturen. Der energetisch niedrigstliegende dieser Zustände wird stets in die nächste Runde übernommen. In energetisch aufsteigender Reihenfolge werden (von den insgesamt circa 1000 gespeicherten Konfigurationen) Strukturen hinzugenommen, die zu den bislang ausgewählten einen RMSB-Abstand von mindestens 5.0Å haben. Damit ist sichergestellt, daß sehr unterschiedliche Strukturen in die nächste Optimierungsphase gelangen und daß in der zweiten Phase in verschiedenen Bereichen des Konfigurationsraumes optimiert wird. Die Dauer der zweiten Phase liegt bei etwa ein bis zwei CPU-Monaten pro Struktur.

Obwohl die Hochtemperaturphase vergleichsweise kurz ist, kommt ihr eine große Bedeutung in den Faltungssimulationen zu, die zum Teil durch eine spezielle Modifikation der Kraftfeldparameter bedingt ist. In Anlehnung an das Framework-Modell der Proteinfaltung (Abbildung 2.11), wonach sich zunächst die Sekundärstruktur und dann die dreidimensionale Anordnung der Atome in der Tertiärstruktur bildet, wird in der Hochtemperaturphase die Lösungsmittelenergie knapp 20% reduziert. Auf diese Weise erlangt die elektrostatische Wechselwirkung und insbesondere die Wasserstoffbrückenbindungsenergie ein größeres Gewicht. Die Ausbildung von Sekundärstrukturelementen – nicht nur das der α-Helix, sondern auch das β-Faltblatt – ist in dieser Simulationsphase begünstigt. Würde der Lösungsmittelbeitrag auf Null reduziert, so wäre die energetisch günstigste Konfiguration die einer einzelnen Helix mit maximaler Länge.

In der Tat finden sich unter den energetisch besten 50% der Hochtempera-

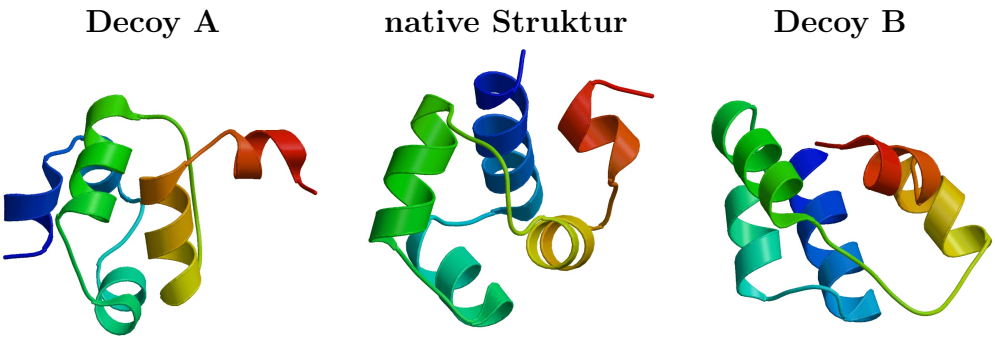


Abbildung 5.9: Die native Struktur von 1R63 (Mitte) im Vergleich mit dem Hochtemperaturfavoriten (links) und der Struktur mit nahezu identischer Sekundärstruktur (rechts)

turkonfigurationen fast ausschließlich Zustände mit hohem Sekundärstrukturanteil. Die Reduktion des Lösungsmittelhamiltonian ist so moderat, daß nicht etwa Strukturen mit einzelnen sehr langen Helices, also relativ gestreckte Konfigurationen, den Wettstreit um die niedrigste Energie gewinnen, sondern kompakte Strukturen, wenngleich der Sekundärstrukturanteil im Durchschnitt höher liegt als in der nativen Struktur. Es hat sich gezeigt, daß Konfigurationen mit einer der nativen ähnlichen Sekundärstruktur verhältnismäßig häufig unter den energetisch Besten der Hochtemperatursimulation anzutreffen sind.

Auch wenn zu Beginn der zweiten Phase schon Strukturen existieren, die die native Sekundärstruktur haben, so ist die Anordnung dieser Strukturelemente im dreidimensionalen Raum meist von der nativen verschieden. Hierfür sei 1R63 als Beispiel aufgeführt. Mit 63 Residuen ist es das größte Protein, dem wir uns bislang zugewandt haben. Es handelt sich, wie in Abbildung 5.9 zu sehen, um ein fünf-Helix Protein. (Der N-Terminus ist in blau, der C-Terminus in rot wiedergegeben. Die Helices werden vom N-Terminus ausgehend durchnumeriert.) In diese Abbildung ist auch die beste Struktur eines Hochtemperaturlaufs aufgenommen (Decoy B). Die Sekundärstruktur dieser beiden Konfigurationen ist praktisch identisch. In Tabelle 5.2 sind die Sekundärstrukturen der zehn energetisch besten Hochtemperaturkonfigurationen aufgelistet. Die RMSB-Abweichungen aller Hochtemperaturstrukturen von der nativen Konfiguration liegt bei über 8Å, das heißt, obwohl die Hochtemperaturstrukturen die Helices der nativen Konfiguration ausgebildet haben, sind dieses völlig verschieden im Raum angeordnet.

Sekundärstruktur	
cHHHHHHHHHHHHtccHHHHHHHHtscHHHHHHHHHtccssctt**H**HHHHHHHtccHHHHHtcc	
cHHHHHccHHHHHtsscHHHHHHHcccHHHHHHHHHcssccsscc**H**HHHHHHHHcsHHHHHtc	**A**
cHHHHHHHHHHHHtsccHHHHHHHtHHHHHHHHHHttcsccsssc**H**HHHHHHcccHHHHHHHc	
cHHHHHHHHHHHHssccHHHHHHHtHHHHHHHHHHttcsccsssctt**H**HHHHtccHHHHHHcc	
cHHHHHHHHHHHHssscHHHHHHttsHHHHHHHHHHcccccsscc**H**HHHHHHcccHHHHHHcc	**B**
cHHHHHHHHHHHHtsccHHHHHHHtHHHHHHHHHHttccccsssc**H**HHHHHHtsscsHHHHtc	
cHHHHHHHHHHHHtsscHHHHHHHcccHHHHHHHHHtcttcscscsccscHHHHtcsscsscc	
cHHHHcHHHHHssscsHHHHHHHHcccHHHHHHHHHcssccsscc**H**HHHHHHHtcHHHHHHtc	
cctHHHHHHHHHHtccHHHHHHHHssHHHHHHHHHtssccsscc**H**HHHHHHHHccHHHHHtc	
cHHHHHHHHHHHHcccttccttsccHHHHHHHHHHHHHssscscsttHHHHHHHHHHHtc	
ccHHHHHHHHHHHHscHHHHHHscccHHHHHHHHHHHHHtsttccSHHHHctttscHHHHHcc	

Tabelle 5.2: Die Sekundärstruktur der zehn energetisch niedrigsten Strukturen einer Hochtemperatursimulation für 1R63. In der obersten Zeile ist die Sekundärstruktur der nativen Konfiguration angegeben und die Sekundärstrukturen zu den Decoys A und B sind mit A und B indiziert.

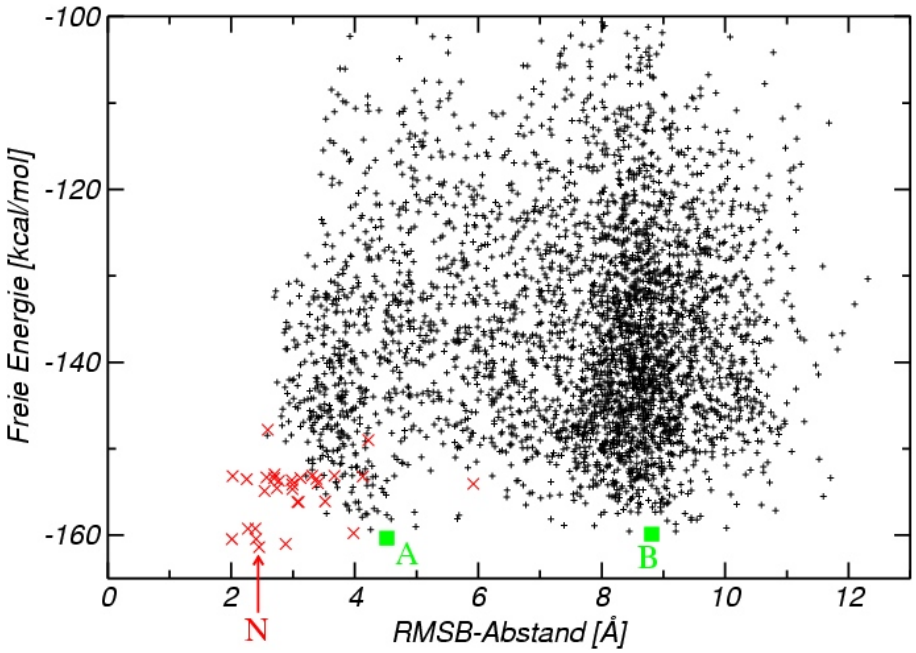
5.8 1BDD

Die erfolgreiche Faltung eines Proteins mit 40 Residuen (1F4I) ist einer der wesentlichen Punkte dieser Dissertation. Ein weiterer Aspekt betrifft die Frage, inwieweit es gelungen ist, ein Kraftfeld zu parametrisieren, welches auf eine ganze Proteinklasse, die der Helixproteine, angewendet werden kann. Wir suchten daher nach weiteren kleinen Helixproteinen und haben uns für drei Proteine entschieden, die in diesem und den folgenden zwei Abschnitten behandelt werden: 1BDD, 1ENH und 1GYZ. Für diese drei Proteine wird gezeigt, daß eine nativ-ähnliche Struktur als das globale Minimum der Freien Energie (im bislang untersuchen Teil des Konfigurationsraumes) identifiziert werden konnte. Zu diesem Zweck wird die native Struktur jedes der drei Proteine in einem BHT-Lauf (Basin-Hopping-Technique) relaxiert und das der nativen Struktur nächstliegende ausgeprägte lokale Minimum identifiziert. Damit ist das Ziel eines bei einer random-coil Struktur beginnenden Faltungsversuches festgelegt, und zwar sowohl in Bezug auf die Struktur, als auch auf den Wert der Freie Energie.

Betrachten wir nun die Ergebnisse für 1BDD [ZK97] mit 52 Residuen[2], welches wie 1F4I aus drei Helices besteht. Die von random-coil Zuständen beginnenden Simulation gliederte sich in die weiter oben beschriebenen zwei Abschnitte, eine Hochtemperatur- und eine Optimierungsphase. Die Zahl der optimierten Strukturen ist zu gering, um einen Decoy-Baum sinnvoll anfertigen zu können. Einen Überblick über die Energien und RMSB-Abstände der erfolgten Simulationen ist in Abbildung 5.10 wiedergegeben. Im RMSB-Abstands/Energie-Diagramm sind die Relaxationsläufe der nativen Struktur in rot und die Ergebnisse der freien Faltungssimulationen in schwarz abgebildet. Die energetisch niedrigste Struktur der Relaxationsläufe wird mit N bezeichnet. Die beiden grünen Punkte gehören zu den beiden energetisch niedrigsten Konfigurationen der Faltungssimulationen, und sind als Decoy A und B bezeichnet.

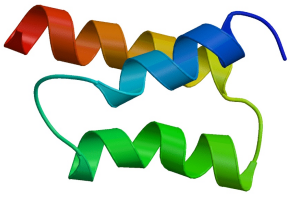
Die in den Simulationen gefundene Konfiguration mit der niedrigsten Freien Energie (Decoy A) liegt $1\,kcal\,mol^{-1}$ über der relaxierten nativen Struktur und ist 4.52Å entfernt. Im Vergleich der Strukturen ist zu erkennen, daß in Decoy A die erste Helix (blau) aufgebrochen ist. Das eingesetzte Optimierungsverfahren, welches nur Rotationen eines einzelnen Residuums kennt, ist nicht in der Lage, Konfigurationsänderungen an Decoy A vorzunehmen, um das fehlende $kcal\,mol^{-1}$ in der Freien Energie zu gewinnen. Decoy B ist eine fünf-Helix Struktur und besitzt eine ausgeprägte sphärische Geometrie. Hier zeigt sich zum wiederholten Male, daß im Kraftfeld PFF01 im Gegensatz zu der ursprünglichen Parametrisierung des CARB-Kraftfeldes ein Wettstreit zwischen hydrophoben Effekt und Sekundärstrukturbildung durch Wasserstoffbrücken beobachtet werden kann.

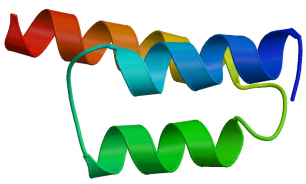
[2]Die 1BDD-Struktur der Proteindatenbank PDB weist 60 Aminosäuren auf, von denen die ersten fünf und die letzten drei keine stabile Struktur ausbilden. Diese acht Residuen werden in den Simulationen nicht berücksichtigt.



native Struktur **relaxierte Struktur N** C_β-Mosaik für N

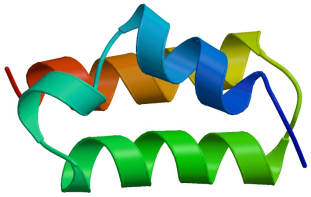
$F = -161.39\frac{kcal}{mol}$ RMSB=2.45Å





Decoy A **Decoy B**

$F = -160.35\frac{kcal}{mol}$ RMSB=4.52Å $F = -159.88\frac{kcal}{mol}$ RMSB=8.81Å



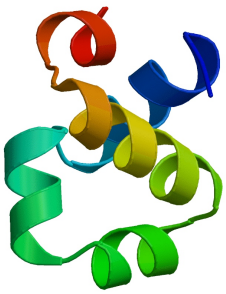


Abbildung 5.10: 1BDD-Simulationen mit PFF01

5.9 1ENH

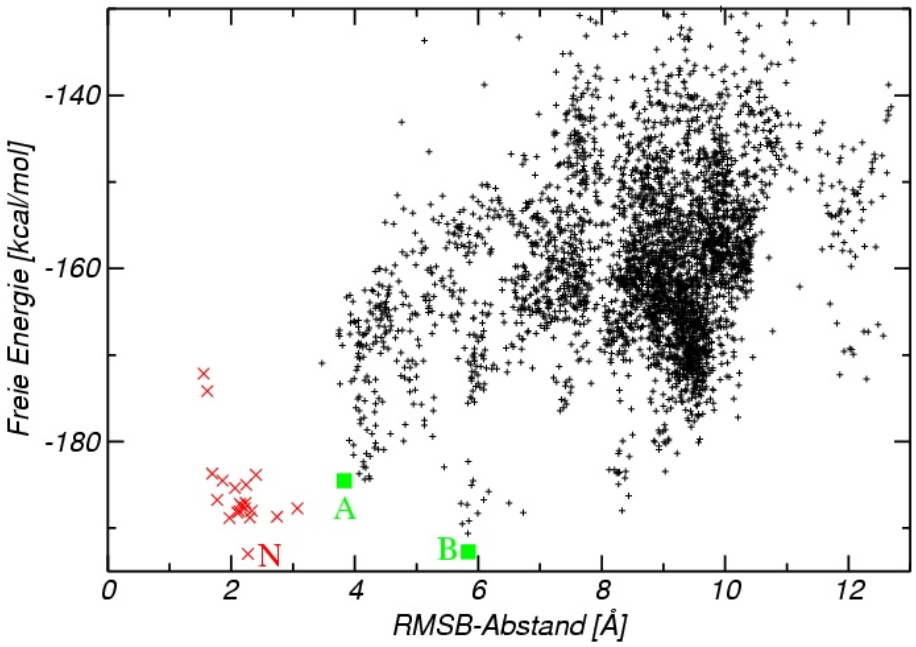
In einer aktuellen Arbeit wurde 1ENH zusammen mit weiteren drei Helix-Proteinen in einem stark vereinfachten Modell behandelt, um Hypothesen zum Faltungsmechanismus durch theoretische Berechnungen zu unterstützen [IKW02]. Mit 54 Residuen entzieht sich die Faltungsdynamik dieses Proteins jedoch einer vollständigen Betrachtung durch die Molekulardynamik Simulation. Auch wenn unsere Simulationen den Faltungsmechanismus nicht enträtseln können, so könnte ein Decoy-Baum mit den richtigen Übergangsenergien einiges über die Thermodynamik des Systems aussagen. Doch bevor ein solcher Baum angefertigt werden kann, müssen wir klären, ob 1ENH mit PFF01 und dem vorliegenden Simulationsprogramm gefaltet werden kann.

Bei der Relaxation in PFF01 bildet 1ENH eine vierte Helix aus (Decoy N in Abbildung 5.11), wohingegen die ersten sieben Residuen in der nativen Struktur eine gestreckte Konfiguration einnehmen, die zu einer verbesserten elektrostatischen Wechselwirkung der Seitenketten führt. Dabei wird die hydrophile Seitengruppe des Glutamin 42 weitestgehend vor dem Lösungsmittel abgeschirmt; eine Konfiguration, die in PFF01 energetisch benachteiligt ist. Im C_β-Mosaik sind die schwarz eingefärbten Helixanteile zu erkennen. Berücksichtigt man, daß auch die grauen Bereiche im Rahmen der natürlichen Fluktuationen liegen, so weicht die relaxierte Struktur nur bei der zusätzliche Helix und beim Übergang der zweiten zur dritten Helix von der nativen Struktur ab.

In Abbildung 5.11 sind die Ergebnisse der Relaxationen (rot) und der Freien Faltung (schwarz) wiedergegeben. Die Konfiguration, die bei mittelmäßiger Energie der nativen Struktur ähnelt, ist als Decoy A abgebildet. Allerdings ist die in den Simulationen vorhergesagte Struktur (Decoy B) 5.84Å von der nativen Struktur entfernt und besitzt die gleiche Energie wie die relaxierte Struktur N. Auch wenn die mittlere Helix bei Decoy B länger ist als in der nativen Struktur, so sind die Anordnungen der Helices in den Strukturen recht ähnlich.

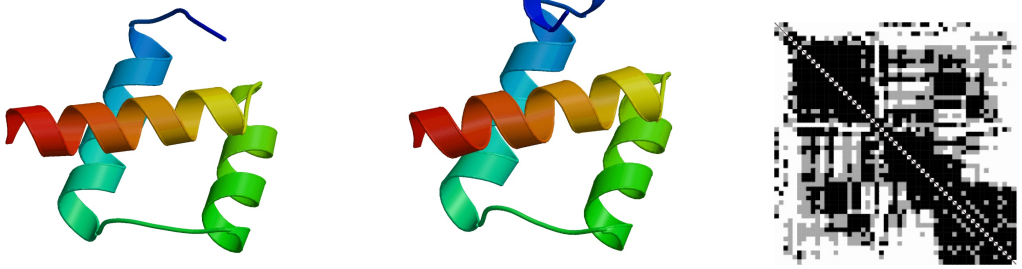
Wie aus dem Energie/Abstands-Diagramm zu entnehmen, fand das Optimierungsverfahren für 1ENH mit 54 Residuen keine Struktur mit weniger als 3.5Å RMSB-Abweichung zur nativen Struktur, und der Bereich bis 4.0Å ist sehr dünn besetzt. Hinzu kommt, daß nur wenige Strukturen eine Freie Energie unterhalb von $-180\,kcal\,mol^{-1}$ haben. Im übernächsten Abschnitt werden wir uns, nachdem wir die Ergebnisse für 1GYZ kennengelernt haben, etwas ausführlicher mit den Optimierungsverfahren und den 1ENH Simulationen beschäftigen.

Es sei hier noch an Tabelle 5.3 auf Seite 116 verwiesen, in der die Energien, RMSB-Abweichungen und Sekundärstrukturen der besten 15 Konfigurationen aufgelistet sind. Wie zu erkennen, ist die Sekundärstruktur aller niederenergetischen Konfigurationen der nativen recht ähnlich.



native Struktur **relaxierte Struktur N** C_β-Mosaik für N

$F = -192.99\frac{kcal}{mol}$ RMSB=2.27Å



Decoy A **Decoy B**

$F = -184.54\frac{kcal}{mol}$ RMSB=3.83 $F = -192.75\frac{kcal}{mol}$ RMSB=5.84Å

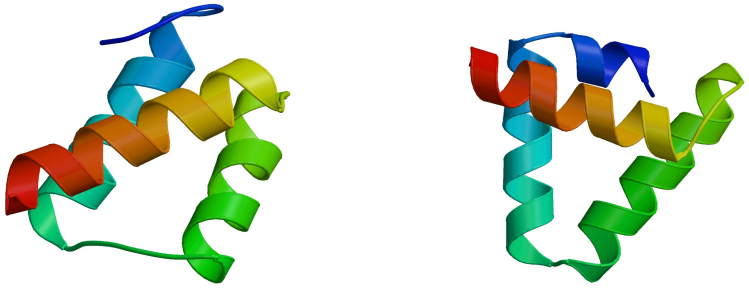


Abbildung 5.11: 1ENH-Simulationen mit PFF01

5.10 1GYZ

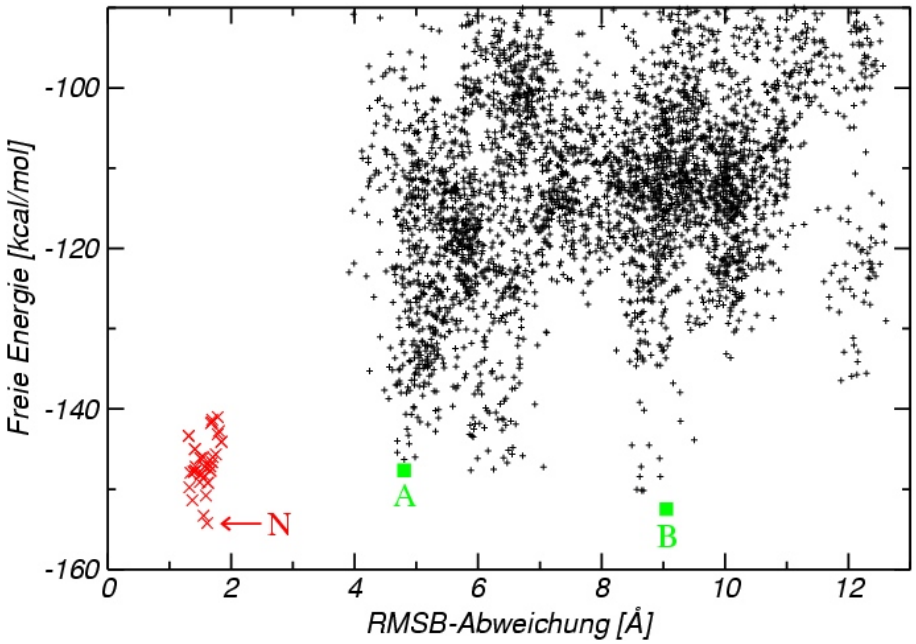
Die in den letzten Abschnitten vorgestellten Proteine bilden drei Helices aus. Die Betrachtung eines Proteins mit vier Helices in der nativen Struktur macht es erforderlich, die Systemgröße weiter zu erhöhen. Mit 60 Residuen ist 1GYZ eines der kleinsten Peptide, die in diesem Kontext zur Verfügung stehen.

Die Relaxation der nativen Struktur ist mit einer RMSB-Abweichung von 1.61Å sehr dicht an der natürlichen Konfiguration geblieben und liegt nahezu $2\,kcal\,mol^{-1}$ unterhalb aller Strukturen, die wir in den Faltungssimulationen erreicht haben (Abbilding 5.12). Die relaxierte Struktur kann in Bezug auf die Hauptkette als identisch zur nativen Struktur angesehen werden. Der RMSD-Abstand, also der mittlere quadratische Abstand aller Atome, liegt bei 2.72Å. Daran, daß das C_β-Mosaik nicht vollständig schwarz eingefärbt ist, obwohl die Strukturen sehr ähnlich sind, ist die hohe Empfindlichkeit des C_β-Mosaiks gegenüber strukturellen Unterschieden zu erkennen.

Im oberen Diagramm der Abbildung ist zu erkennen, daß das Optimierungsverfahren zum wiederholten Male Schwierigkeiten hat, niederenergetische Konfigurationen ausfindig zu machen. Die Strukturvorhersage (Decoy B) der Simulation ist 9.05Å von der nativen Struktur entfernt. Positiv ist jedoch zu vermerken, daß Decoy B vier Helices enthält, von denen die letzten zwei in der Länge und die letzten drei auch in ihrer relativen Ausrichtung denen der nativen Struktur ähnlich sind.

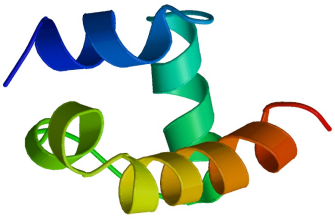
Das Optimierungsverfahren war außerstande, Strukturen mit einem RMSB-Abstand unter 3.8Å zu generieren. Selbst unter 4.5Å sind nur sehr wenige niederenergetische Strukturen zu sehen. Ähnlich wie bei 1ENH ist auch der Energiebereich unterhalb von etwa $20\,kcal\,mol^{-1}$ über der relaxierten nativen Struktur nur dünn besiedelt. Dabei muß jedoch berücksichtigt werden, daß für 1ENH und 1GYZ momentan jeweils nur 10,000 Strukturen vorliegen. Decoy A ist eine der wenigen Strukturen der Faltungssimulation, die, bei mäßiger Energie, relativ dicht an die native Struktur gelangt ist. Die Sekundärstruktur stimmt, wie der Tabelle 5.3 entnommen werden kann, sehr gut mit der nativen Struktur überein. In der Abbildung 5.12 und der Tabelle 5.3 ist zu erkennen, daß der wesentliche Unterschied zwischen der nativen Struktur und Decoy A in der Länge und Orientierung der zweiten Helix besteht.

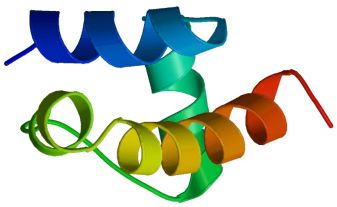
Tabelle 5.3 zeigt, daß die drei äußeren Helices bei allen niederenergetischen 1GYZ-Strukturen ausgebildet sind. Im Zwischenbereich werden eine oder zwei Helices gebildet, deren Position variiert. Die Simulationen bilden jedoch tendenziell weniger Sekundärstruktur aus als in der nativen Struktur vorhanden. Hingegen bilden die Simulationen zu 1ENH offenbar eher mehr Sekundärstruktur als die Natur.



native Struktur **relaxierte Struktur N** C_β-Mosaik für N

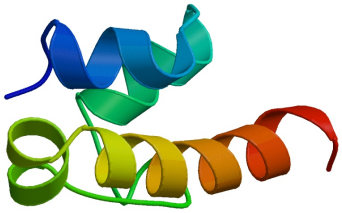
$F = -154.23\frac{kcal}{mol}$ RMSB=1.61Å





Decoy A **Decoy B**

$F = -147.65\frac{kcal}{mol}$ RMSB=4.80Å $F = -152.48\frac{kcal}{mol}$ RMSB=9.05Å



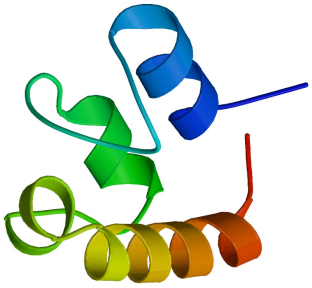


Abbildung 5.12: 1GYZ-Simulationen mit PFF01

1ENH

$F[\frac{kcal}{mol}]$	RMSB[$\AA$]	Sekundärstruktur	
nativ		ccccccccHHHHHHHHHHHHHHcsssccHHHHHHHHHHHHtccHHHHHHHHHHHHHHHc	
-192.99	2.27	cHHHHHscHHHHHHHHHHHHHHssssccHHHHHHHHHHHHtccHHHHHHHHHHHHHHHc	N
-192.75	5.80	cHHHHHHsHHHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHtscHHHHHHHHHHHHHHHc	B
-190.63	5.79	cHHHHHHsHHHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHtccHHHHHHHHHHHHHHHc	
-189.51	5.71	cHHHHHHsHHHHHHHHHHHHHHsHHHHHHHHHHHHHHHHHtscHHHHHHHHHHHHHHHc	
-189.19	5.80	cHHHHHHsHHHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHtscHHHHHHHHHHHHHHHc	
-188.31	5.96	cHHHHHtsHHHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHtscHHHHHHHHHHHHHHHc	
-188.23	6.68	cHHHHHHsHHHHHHHHHHHHHHsHHHHHHHHHHHHHHHHHtscHHHHHHHHHHHHHHHc	
-187.99	8.31	cccHHHHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHHc	
-187.30	5.78	cHHHHHtsHHHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHtccHHHHHHHHHHHHHHHc	
-187.16	5.69	cHHHHHHssHHHHHHHHHHHHHsHHHHHHHHHHHHHHHHHtscHHHHHHHHHHHHHHHc	
-187.11	6.45	cHHHHHHsHHHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHtccHHHHHHHHHHHHHHHc	
-186.75	6.05	cHHHHHtsHHHHHHHHHHHHHHsHHHHHHHHHHHHHHHHHtccHHHHHHHHHHHHHHHc	
-186.29	8.43	cccHHHHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHHc	
-185.78	6.13	cHHHHHtsHHHHHHHHHHHHHHcHHHHHHHHHHHHHHHHHtccHHHHHHHHHHHHHHHc	
-184.54	3.83	ccccsscHHHHHHHHHHHHHHscccsHHHHHHHHHHHHscHHHHHHHHHHHHHHHHHc	A

1GYZ

$F[\frac{kcal}{mol}]$	RMSB[$\AA$]	Sekundärstruktur	
nativ		ccHHHHHHHttttccsHHHHHHHHHHHtccccssssHHHHHHcHHHHHHHHHHHHHHHttcc	
-154.23	1.61	cHHHHHHHHHHHtccsHHHHHHHHHHHtcccssscsHHHHHHcHHHHHHHHHHHHHHHHttc	N
-152.48	9.05	cccHHHHHHHHtccsssscccttttcHHHHHHssscccHHHHHcHHHHHHHHHHHHHHHtccc	B
-150.18	8.68	cHHHHHHHHHHtsssccttcHHHHHHHHHHssscccHHHHHsHHHHHHHHHHHHHHHttc	
-150.16	8.65	ccccHHHHHHHttcsssscccttttcHHHHHHssscccHHHHHcHHHHHHHHHHHHHHHsccc	
-150.05	8.57	ccHHHHHHHHHtscttccscHHHHHHHHHHcsscccHHHHHsHHHHHHHHHHHHHHHtscc	
-148.43	8.62	ccHHHHHHHHHtsssccttcHHHHHHHHHHssscccHHHHHsHHHHHHHHHHHHHHHttc	
-148.43	8.95	cccHHHHHHHHttcsssscccttttcHHHHHHssscccHHHHHcHHHHHHHHHHHHHHHsccc	
-147.65	4.80	ctttcHHHHHHHtcccHHHHHHHHtsBtttBsscccHHHHHHcHHHHHHHHHHHHHHHHHtc	A
-147.65	5.88	ccsscHHHHHHHtHHHHHHHHHtcssccHHHHHttccHHHHtsHHHHHHHHHHHHHHHttcc	
-147.52	6.25	ccsttHHHHHHHtHHHHHHHHHtcssccHHHHHttccHHHHtsHHHHHHHHHHHHHHHttcc	
-147.24	7.06	ccttcHHHHHHHtHHHHHHHHHccssccHHHHHssccHHHHHcHHHHHHHHHHHHHHHsccc	
-147.19	8.64	cHHHHHHHHHHtsssccttcHHHHHHHHHHssscccHHHHHsHHHHHHHHHHHHHHHttc	
-146.69	6.47	ccttcHHHHHHHtHHHHHHHHHccssccHHHHHsscsHHHHHcHHHHHHHHHHHHHHHttcc	
-146.30	4.80	ctttcHHHHHHHtcccHHHHHHHHttBtttBsscccHHHHHHcHHHHHHHHHHHHHHHHHtc	
-146.19	8.94	cccHHHHHHHHttcsssscsctttcsHHHHHssscccsHHHHHcHHHHHHHHHHHHHHHsccc	

Tabelle 5.3: Sekundärstruktur der energetisch besten Konfigurationen für 1ENH und 1GYZ. Die rechtsstehenden Buchstaben markieren die Namen der Decoys.

Kapitel 6

Diskussion

Eine erfolgreiche de-novo Proteinstrukturvorhersage, also die Bestimmung der Tertiärstruktur allein aus der Kenntnis der Proteinsequenz, ist eines der größten ungelösten Probleme der biophysikalischen Chemie. Das wachsende Interesse an der Strukturaufklärung von Proteinen erklärt sich durch die enge Beziehung zwischen Struktur und biologischer Funktion. Experimentelle Strukturaufklärung ist jedoch recht aufwendig und schwierig; in einigen Fällen sogar gänzlich unmöglich. Desweiteren liefern diese Experimente primär statische Informationen. Eine Theorie, die eine genaue Modellierung biophysikalischer Wechselwirkungen in kurzer Zeit erlaubt, würde die experimentelle Strukturaufklärung komplementieren. Darüber hinaus würde diese Theorie eine Basis für ein breites Spektrum an Simulationen von biologischen Systemen auf mikroskopischer Ebene schaffen. Diese Simulationen würden die Möglichkeit bieten, Hypothesen über Regelmechanismen und Protein-Protein Aggregation schnell und einfach zu testen. Computerunterstützte Proteinstrukturvorhersage geschieht derzeit mit homologiebasierten Verfahren, deren Resultate wichtige Hinweise für die Proteinstruktur liefern, jedoch nicht mit der Genauigkeit, die für eine Analyse biologischer Prozesse notwendig ist. Die homologiebasierten Methoden besitzen kein deutliches Verbesserungspotential. Der davon grundsätzlich verschiedene Ansatz der Molekulardynamik Simulation arbeitet – ein geeignetes Kraftfeld vorausgesetzt – sehr genau. Der Rechenaufwand ist jedoch viel zu hoch, als daß dieses Verfahren im großen Maßstand einsetzbar wäre. Dies ist auch der Grund dafür, daß existierende Kraftfelder nicht an Proteinen, sondern nur an kleinen Molekülen optimiert worden sind.

Nach der *thermodynamischen Hypothese* von Anfinsen, für die er 1972 den Nobelpreis in Chemie erhielt, ist die Proteinstruktur das globale Minimum der Freien Energie. Damit ist das Protein-Struktur-Problem den Methoden der statistischen Physik zugänglich. Die thermodynamische Hypothese wird mittlerweile von vielen Experimenten bestätigt. Unterstützungen dieser Hypothese durch theoretische Modelle liegen jedoch nur in groben Approximationen der Proteine vor.

Das vorliegende Kraftfeld ist das erste realistische Proteinmodell, welches für

mehrere Proteine die jeweilige native Struktur als globales Minimum auszeichnet. Dies ist das Ergebnis mehrerer Arbeitsschritte, die sich zunächst recht schwierig gestalteten. Wir konzentrierten uns dabei auf ein einzelnes Protein (1VII) und folgten dem Ansatz, daß ein fehlerhaftes Kraftfeld für mindestens eine nicht-native Struktur eine Freie Energie prognostiziert, die unterhalb der Freien Energie der nativen Struktur liegt. Wenn ein solches nicht-natives Decoy gefunden wurde, ist das Kraftfeld entsprechend modifiziert worden. Es wurde darauf Wert gelegt, daß die notwendigen Kraftfeldveränderungen physikalisch motiviert sind. Auch aus diesem Grund wurden zur Parameteranpassung zumeist experimentelle Ergebnisse berücksichtigt, die sich aus aufgeklärten Proteinstrukturen und Lösungsmittel-Transferenergien ergaben. Simulationen an 1VII haben gezeigt, daß in den experimentellen Daten der Transferenergie verschiedene Aspekte integriert sind und zu systematischen Parametrisierungsfehlern geführt haben. Für 1VII konnte der Einfluß dieser Fehler auf die Freie Energie identifiziert und behoben werden. Es war anfänglich nicht zu erwarten, daß die gewählte Parametrisierung für mehr als dieses eine Protein geeignet ist.

Wie in dieser Arbeit gezeigt wurde, weist das vorliegende Kraftfeld für 6 Proteine unterschiedlicher Größe eine nativ-ähnliche Struktur als globales Minimum der Freien Energie aus. Dies deutet darauf hin, daß auch die scheinbar 1VII-spezifischen Parameteranpassungen auf die ganze Proteinfamilie der Helixproteine übertragbar sind. In Bezug auf Molekulardynamik Simulationen muß darauf hingewiesen werden, daß diese Parameteranpassung erst nach umfangreichen Simulationen eines relativ großen Proteins (1VII mit 36 Residuen) als notwendig erkannt und in mehreren Schritten umgesetzt werden konnte. Beide Aspekte, die Einsicht in die Notwendigkeit der Anpassung und deren Umsetzung, scheitern bei Molekulardynamik Simulationen daran, daß die verfügbaren Rechnerkapazitäten mehrere Größenordnungen zu gering sind.

Bei stochastischen Optimierungsverfahren ist noch eine signifikante Effektivitätssteigerung zu erwarten. Diese kann sich aus der Entwicklung gänzlich neuer Optimierungsstrategien und aus der Anpassung existierender Verfahren an die spezifischen Bedingungen der Proteinfaltung ergeben. Hierbei ließen sich auch experimentell gewonnene Erkenntnisse über Faltungsmechanismen berücksichtigen, wie wir es bei den Hochtemperatursimulationen (Abschnitt 5.7) getan haben, indem wir dem Framework-Modell (Abbildung 2.11) folgend die Ausbildung von Sekundärstrukturelementen begünstigt haben.

Nach unserer Einschätzung sollte sich eine nachfolgende Arbeit zunächst auf die Weiterentwicklung der Optimierungsverfahren konzentrieren. Im folgenden Abschnitt soll daher der Aspekt der Optimierung noch etwas ausführlicher behandelt werden. Die vorliegende Arbeit hat sich der Kraftfeldentwicklung für die Familie der Helixproteine gewidmet. Um zu klären, inwieweit dieses Kraftfeld auf β-Faltblattstrukturen erweiterbar ist, sollen anschließend vorläufige Ergebnisse für eine β-Faltblattstruktur präsentiert werden.

6.1 Lokale versus globale Optimierung

Bei 1L2Y und 1F4I hat die Optimierung reproduzierbar die nativen Struktur gefunden und diese als globales Minimum der Protentialoberfläche ausgewiesen. Für 1VII, 1BDD, 1ENH, 1GYZ und 1R63 war es den Optimierungsverfahren nicht möglich, Strukturen ausfindig zu machen, die energetisch unter der relaxierten nativen Struktur lagen. Dies spricht für die Qualität des Kraftfeldes PFF01. Diese Aussage ist jedoch solange mit Vorsicht zu genießen, wie kein Optimierungsverfahren vorliegt, welches die native Struktur dieser Proteine findet. Es ist denkbar, daß ein solches Verfahren gleichzeitig Strukturen aufspüren würde, die energetisch unterhalb der relaxierten nativen Struktur liegen.

Daher hängt der Erfolg weiterer Simulationen nicht nur vom Fortschritt in der Approximation der Freien Energie als Funktion der Proteinstruktur, sondern auch von der Qualität der Optimierungsverfahren ab. Die Ergebnisse zu 1ENH und 1GYZ zeigen, daß die verwendeten Verfahren nicht hinreichend in der Lage sind, den Niederenergiebereich der Potentialenergieoberfläche stark und in der Verteilung breit zu populieren. Dies gilt wahrscheinlich nicht nur für große Proteine, sondern auch für Proteine mit einer sehr breiten Trichterstruktur in der Potentialenergieoberfläche, wie sie bei 1VII vorliegt.

Das Problem der Optimierungsverfahren besteht hierbei in der lokalen Optimierung. Das heißt, daß Simulated Annealing mit den eingesetzten Rotationskategorien für die Identifizierung eines tiefen lokalen Minimums auf der Potentialenergieoberfläche eines Proteins eher ungeeignet ist. Als übergeordnete Optimierungsmethode hat sich die Basin-Hopping-Technique hingegen innerhalb gewisser Grenzen bewährt. Um dies zu illustrieren, wurde der Fortlauf der BHT Simulation während der Optimierungsphase für 1ENH in Abbildung 6.1 sichtbar gemacht. Dazu wurden (für jede Simulation getrennt) die nach den SA-Iterationen gefundenen Strukturen durchnumeriert und die RMSB- und RMSD-Abstände[1] zweier Strukturen i und $j(> i)$ in eine Matrix R eingetragen. Das Matrixelement R_{ij} (oberhalb der Diagonalen) gibt den RMSB- und das Element R_{ji} (unterhalb der Diagonalen) den RMSD-Abstand an. Beide Zahlenwerte sind bei 4.0Å abgeschnitten worden.

In Abbildung 6.1 sind die RMSB-RMSD-Matrizen für 9 ausgewählte Simulationen farbig dargestellt. Die drei Matrizen der ersten Zeile geben Simulationen wieder, in denen die Struktur während der Simulation viele unterschiedliche Bereiche aufgesucht hat, wie die dominierende Farbe rot (RMSB, RMSD$\geq$ 4Å) zeigt. Insbesondere in der rechten Matrix ist an ihrer Blockform zu sehen, daß das Optimierungsverfahren eine gewisse Zeit einen kleinen Bereich des Konfigurationsraumes intensiv durchsucht (RMSB-Werte blau bis türkis, d.h. < 2Å; RMSD-Werte grün, d.h. < 2.5Å), bevor es die Struktur in andere Regionen schickt

[1]Der RMSD-Abstand ist die mittlere quadratische Abweichung aller Atome nach optimaler Überlagerung der beiden Strukturen. Der RMSB-Wert ist analog für die Atome der Hauptkette definiert [Abschnitt B.3.1].

und dort eine lokale Suche erfolgt. Diese drei Simulationen sind stellvertretend für etwa 20% aller 1ENH-Simulationen aufgenommen worden. Sehr viel häufiger finden sich RMSB-RMSD-Matrizen der Art, wie sie in der zweiten Zeile abgebildet sind. Hier ist die Zeitspanne, die eine Struktur innerhalb eines kleinen Bereiches des Konfigurationsraumes optimiert wird, deutlich höher. Einige wenige Simulationen finden jedoch während der gesamten Laufzeit nicht aus ihrer Umgebung heraus. Drei solche Beispiele sind in der untersten Zeile wiederzufinden. Die vorherrschende Farbe der RMSB-Werte ist türkis bis blau (< 1Å), woraus sich ergibt, daß die Hauptkette des Peptids ihre Konfiguration faktisch nicht verändert hat. Zusammen mit dem grünen RMSD-Bereich läßt sich hieraus folgern, daß in den Simulationen nie akzeptable Konfigurationsänderungen der Hauptkette vorgeschlagen wurden. Man beachte, daß die RMSD-Werte nur in sehr seltenen Fällen türkis eingefärbt sind, also die RMSD-Abweichungen fast ausnahmslos über 1.5Å liegen.

In einer typische BHT Simulation reihen sich Abschnitte mit geringen RMSB-RMSD-Werten (lokale Suche) aneinander. Der Übergang zwischen diesen Abschnitten ist durch strukturelle Veränderungen gekennzeichnet, die zu einer RMSB-Abweichung über 3Å führen (globale Schritte). Somit zeigt die BHT genau das gewünschte Verhalten und arbeitet zumeist recht zuverlässig. Dennoch reichen die globalen Schritte bei 1ENH und 1GYZ nicht aus, um in der aufgewandten Rechenzeit Strukturen nahe der nativen Konfiguration zu generieren. An 1ENH (54 Residuen) wurden zwei Faltungsversuche unternommen, die jeweils 20 unabhängige Simulationen umfaßten. Die Hochtemperaturphase dauerte jeweils einen Tag und die Rechenzeit der Optimierungsphase beläuft sich auf etwa einen Monat pro Struktur (10^7 Monte-Carlo Schritte[2]) auf einem 1.4GHz Rechner, d.h. für 20 1ENH-Simulationen auf insgesamt knapp 2 CPU-Jahre.

Die gleichen Computerressourcen waren für die Faltung von 1F4I mit 40 Residuen (in der Optimierungsphase) notwendig. Es ist daher nicht zu überraschend, daß wir nur einen kleinen Teil des Konfigurationsraumes während der Simulationen von 1ENH und 1GYZ einsehen konnten und keine Strukturen nahe der nativen fanden. Wie vereinzelte SA Läufe zeigen, existieren jedoch auch in diesen Region energetisch sehr tiefliegende Strukturen, die nur mit einer geringen Erfolgsquote aufgespürt werden konnten. Das heißt, zunächst bedarf die lokale Optimierung einer Qualitätssteigerung und muß überarbeitet werden.

6.2 Grenzen und Ausbaupotential von PFF01

Neben den Optimierungsverfahren besitzt auch das Kraftfeld selbst noch Potential zu Verbesserung. Der Schwerpunkt dieser Arbeit liegt auf Helixproteinen, und bei 1F4I konnte erstmals die native Struktur eines Peptids mit 40 Residuen

[2]Im Vergleich mit den Simulationsprogrammen zum CARB- und ECEPP-Kraftfeld ist das Simulationsprogramm zu PFF01 somit einen Faktor 10 schneller [Han].

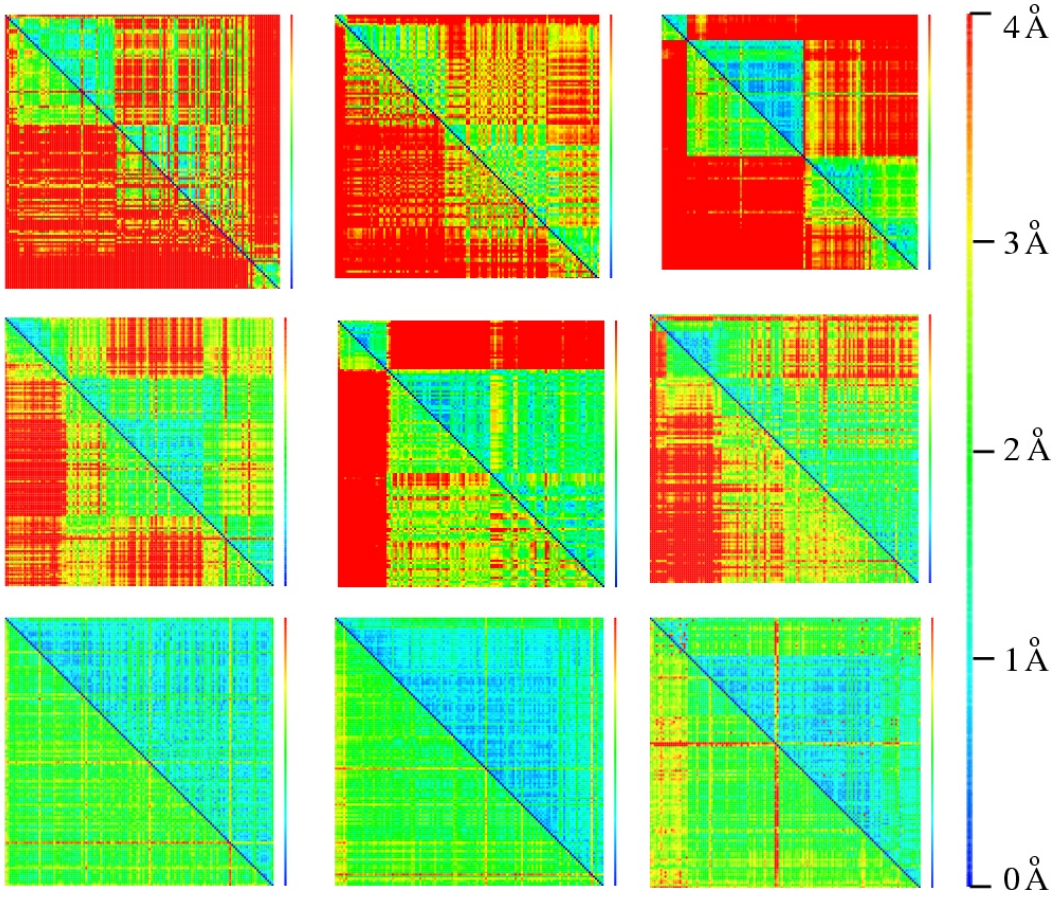


Abbildung 6.1: Strukturelle Veränderungen in den BHT Simulationen für 1ENH. Das obere rechte Dreieck der Matrizen repräsentiert jeweils die RMSB-Abstände, das untere linke die RMSD-Werte.

allein durch Simulation in einem *all-atom* Kraftfeld mit physikalischen Wechselwirkungen vorhergesagt werden. Andere Helixproteine betreffend scheitert eine reproduzierbare Vorhersage bislang an mangelnden CPU-Ressourcen. Das Kraftfeld selbst weist jedoch eine nativähnliche Konfiguration als globales Minimum des bislang zugänglichen Bereiches der Freien Energieoberfläche aus. Wir wollen uns nun dem zweiten wichtigen Sekundärstrukturelement vieler Proteine zuwenden, dem β-Faltblatt.

Faltblattstrukturen stellen bislang für alle Kraftfelder eine große Herausforderung dar. Unsere Versuche, PFF01 auf β-Faltblätter anzuwenden, konzentrieren sich auf das Protein 1BHI, welches aus einer α-Helix und einem β-Faltblatt besteht (Abbildung 6.2). 1BHI ist eines der kleinsten Peptide, welches in einer stabilen Konfiguration vorliegt und ein β-Faltblatt ausbildet. Das β-Faltblatt ist in der nativen Struktur jedoch nicht sehr gut ausgebildet und wird nur durch zwei Wasserstoffbrücken stabilisiert. Neben 1BHI ist noch ein Fragment von 2BG1 zu nennen, welches in wässeriger Lösung ein einzelnes β-Faltblatt ausbildet. Diese Struktur ist jedoch recht instabil und geht in Lösung immer wieder für kurze Zeit verloren [DLK99]. Es wird allgemein angenommen, daß die Bildung des ersten anti-parallelen β-Faltblattes nicht sehr stabil ist und eine Faltblatt-Struktur erst nach der Anlagerung des dritten Strangs thermodynamisch stabil ist [HY95]. Diese Auffassung ist jedoch nicht unumstritten.

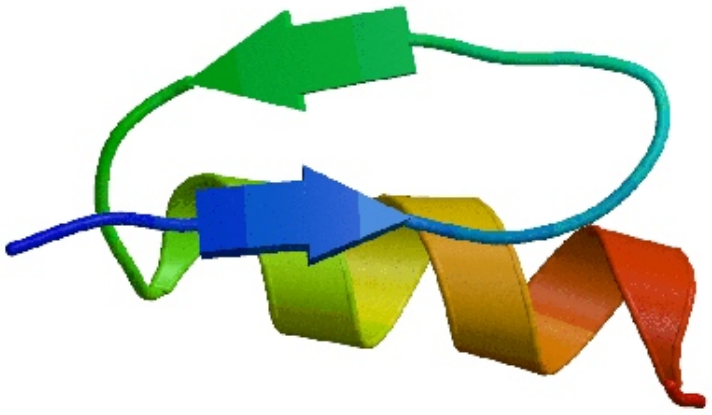


Abbildung 6.2: Die native Struktur des Proteins 1BHI

Die ersten Faltungsversuche von 1BHI waren nicht erfolgreich und haben zu Strukturen geführt, die unterhalb der relaxierten nativen Struktur liegen. Insbesondere eine zwei Helix Struktur hat eine sehr niedrige Freie Energie. Obwohl zum Teil Faltblattstrukturen in der Simulation auftreten, liegt ihre Energie mehrere $kcal\,mol^{-1}$ oberhalb einer zwei Helix Struktur. Daraufhin wurde eine 1BHI Konfiguration generiert, die in der Region des nativen β-Faltblattes alle zugehörigen

Wasserstoffbrücken ausgebildet. Die Energie diese Struktur ist mit derjenigen der zwei Helix Konfiguration vergleichbar. Somit scheint das Kraftfeld PFF01 nicht grundsätzlich für β-Faltblattstrukturen ungeeignet, da ausgeprägte Faltblattstrukturen offenbar eine sehr niedrige Energie besitzen. Es hat sich herausgestellt, daß der lokalen Hauptkettenwechselwirkung[3] für die Stabilisierung des β-Faltblattes eine tragende Rolle zukommt. Die richtige Parametrisierung dieser speziellen Wechselwirkung sollte uns in die Lage versetzen, die Ausbildung von β-Faltblätter in ausreichendem Maße zu unterstützt, ohne die bisherigen Ergebnisse der Helixproteine negativ zu beeinflussen.

Nach unseren bisherigen Erfahrungen werden für die Bildung ausgedehnter Faltblattstrukturen die in Abschnitt 3.1.4 beschriebenen Rotationskategorien, die zur Generierung unterschiedlicher Konfigurationen eingesetzt werden, entsprechend zu erweitern sein. Weniger ausgeprägte Faltblattstrukturen sind dagegen auch mit den derzeitigen Rotationskategorien nicht selten. Sogar in den Simulationen der Helixproteine konnten vereinzelt Strukturen mit geringem Faltblattanteil beobachtet werden. In Abbildung 6.3 ist eine dieser Strukturen exemplarisch für 1F4I wiedergegeben und soll abschließend demonstrieren, daß eine Erweiterung von PFF01 auf β-Faltblätter zwar schwierig ist aber wahrscheinlich dennoch in naher Zukunft möglich sein wird.

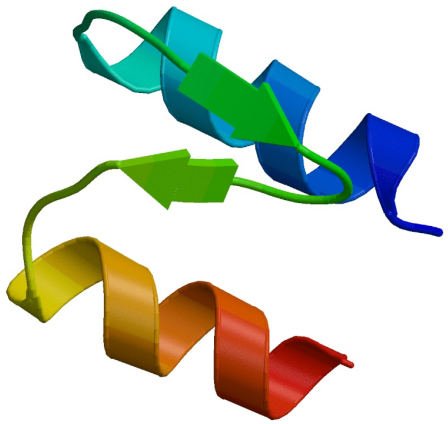


Abbildung 6.3: Faltblattstruktur eines 1F4I Decoys

[3]Dies sind (elektrostatische) Wechselwirkung der Peptidbindungsdipole in der Sequenz benachbarter Aminosäuren. Sie lassen sich entweder durch Dipol-Dipol-Wechselwirkungen oder durch ein Dihedralwinkelpotential beschreiben [AM95].

Anhang A

Einheitentabelle und Umrechnungsfaktoren

Konstanten	Avogadro	N_A	6.0221367×10^{23}	Moleküle mol^{-1}
	Elementarladung	e	$1.6021773 \times 10^{-19}$	C
	Dielektrizitätskonstante	ϵ_0	$8.8541878 \times 10^{-12}$	$AsV^{-1}m^{-1}[= C^2\,J^{-1}\,m^{-1}]$
	Allgemeine Gaskonstante	R	8.3145	$J\,K^{-1}\,mol^{-1}$
	Boltzmann-Konstante	k_B	1.3807×10^{-23}	$J\,K^{-1}\ = R/N_A$

Umrechnungen:

Länge	1 Angström	Å =	$10^{-10}m\ =\ 0.1nm$
Energie	1 Kalorie	$cal\ =$	$4.184J$

Damit gilt bei $T = 25°C = 298.15K$

$$N_A \cdot k_B T \ = \ RT = 0.593\ \frac{kcal}{mol}$$

$$\frac{1}{4\pi\epsilon_0} \ = \ 332.151\ \frac{kcal\,\text{Å}}{mol\,e^2}$$

Anhang B

Geometrie der Proteine

In diesem Abschnitt wird der Formalismus zur Darstellung und Verarbeitung der Proteinkonfigurationen erläutert. Dabei wird der Zusammenhang zwischen Sekundärstrukturelementen und Dihedralwinkelverteilung ebenso beleuchtet werden, wie die von uns eingesetzten Vergleichsmöglichkeiten zwischen zwei Strukturen, welche ohne Einsatz eines Kraftfeldes auskommen.

B.1 Der Proteinsatz M^{138} nicht-homologer Strukturen

Die Parametrisierung des Kraftfeldes beruft sich nicht nur auf experimentell bestimmte Energien, sondern orientiert sich auch an nativen Strukturen. Dazu bedarf es eines Satzes strukturaufgeklärter Proteine, die möglichst genau gemessen und wenig Gemeinsamkeiten in ihrer Sequenz haben. Von diesen Strukturen sind die Lennard-Jones Parameter und das Korrekturpotential der Wasserstoffbrückenbindungen abgeleitet worden.

Aus einer Arbeit von Abagyan und Totrov[AT94] haben wir eine Liste experimentell aufgeklärter Proteine übernommen, deren Auslösung unter 2.0Åliegt und deren Sequenz paarweise zu maximal 50% übereinstimmt. Von der ursprünglichen Liste haben wir diejenigen Elemente entfernt, deren Einträge in der Proteindatenbank *PDB* nur Koordinaten für die C_α-Atome oder der Hauptkettenatome enthalten. In wenigen Fällen waren die in der Arbeit von Abagyan und Totrov benannten PDB-Kürzel nicht in der Datenbank vorhanden und mußten daher auch aus der Liste gestrichen werden. Die resultierende Liste umfaßt 138 Strukturen. Die PDB-Codes der Strukturdaten, aus denen sowohl die Lennard-Jones Parameter wie auch die Parameter des Wasserstoffbrückenbindungspotentials abgeleitet wurden, lauten:

1AAP-A, 1ACX, 1AKE-A, 1ALC, 1APT, 1BBH-B, 1BBP-B, 1C53, 1CRN, 1CSE-I, 1CTF, 1DFN-B, 1DRB-B, 1ECN, 1ER8-E, 1FIA-A, 1FKB, 1FXD, 1GD1-

R, 1GKY, 1GOX, 1GP1-B, 1GPB, 1HIP, 1HMO-D, 1HNE-E, 1HOE, 1IFB, 1L06, 1LMB-B, 1LTE, 1LTS-A, 1LTS-C, 1LTS-H, 1MEE-A, 1OMD, 1OVA-D, 1PAZ, 1PGX, 1PII, 1PK4, 1PPD, 1PPT, 1R69, 1RBP, 1RNB, 1SAR-B, 1SGC, 1SGT, 1TGL, 1TGS-I, 1THB-C, 1TMN-E, 1TRB, 1UBW, 1YCC, 1YPI-B, 256B-B, 2ACT, 2ALP, 2APR, 2AZA-B, 2CBC, 2CCY-B, 2CDV, 2CI2-I, 2CNA, 2CSC, 2CYP, 2FBJ-H, 2FBJ-L, 2FCR, 2FX2, 2GBP, 2HHB-B, 2LH7, 2LHB, 2LTN-B, 2LTN-C, 2MCM-A, 2MCG-2, 2MHR, 2OVO, 2PAB-A, 2PCY, 2PKA-A, 2PKA-Y, 2POR, 2PRK, 2RN2, 2RSP-A, 2SCP-A, 2SNM, 2SOD-Y, 2TEC-E, 2TRX-B, 2TSC-B, 2WRP-R, 2ZTA-A, 31BI, 351C, 3BLM, 3C2C, 3CBH, 3CHY, 3CLA, 3DFR, 3GRS, 3RP2-B, 4BP2, 4CPV, 4ENL, 4FAB-H, 4ICB, 4INS-D, 4LYZ, 4MBA, 4PEP, 4PTI, 5ABP, 5EBX, 5EST-E, 5HVP-A, 5P21, 5PAL, 5RUB-B, 5RXN, 5TNC, 6CHA-A, 6CPA, 6CPP, 6FAB-H, 6LDH, 6RNT, 6RXN, 7AAT-A, 7ACN, 8DFR

Die kleinsten Proteine in der Liste sind $1DFN - B$ und $4INS - D$ mit jeweils 30 Residuen. $7ACN$ mit 753 Aminosäuren ist der größte Eintrag. Insgesamt umfaßt diese Liste 24893 Aminosäuren, also im Mittel 183 Residuen pro Protein.

B.2 Strukturwiedergabe

B.2.1 Lokale Geometrie

Die Struktur eines Proteins wird durch die Koordinaten jedes Atoms i im dreidimensionalen Raum repräsentiert, also durch Angabe aller Ortsvektoren $\vec{x}_i = (x_{i1}, x_{i2}, x_{i3})^T \in \mathbb{R}^3$. Die Anzahl der Atome werden wir im allgemeinen mit dem Buchstaben N oder M angeben und die Gesamtheit der Atomkoordinaten einer Struktur I mit $\{\vec{x}_i^I\}_{i=1}^N$, oder kurz $\vec{x}$

$$\vec{x} \equiv \{\vec{x}_i^I\}_{i=1}^N \,. \tag{B.1}$$

Für die in dieser Arbeit auftretenden Phasenraumintegrale setzten wir

$$\int \cdot \, d\vec{x} \equiv \int \cdots \int_{-\infty}^{\infty} \cdot \, dx_{1,1} dx_{1,2} dx_{1,3} dx_{2,1} \cdots dx_{N,3} \tag{B.2}$$

Sind zwei Atome i und j kovalent miteinander verbunden, so ist der Bindungsvektor durch $\vec{r}_{ij} = \vec{x}_j - \vec{x}_i$ gegeben und der Bindungsabstand errechnet sich zu $l = |\vec{r}| = \sqrt{\vec{r} \cdot \vec{r}}$. Den Mittelpunkt des Proteins notieren wir mit $\vec{r}_c$ und berechnen

$$\vec{r}_c = \frac{1}{N} \sum_{i=1}^{N} \vec{x}_i \tag{B.3}$$

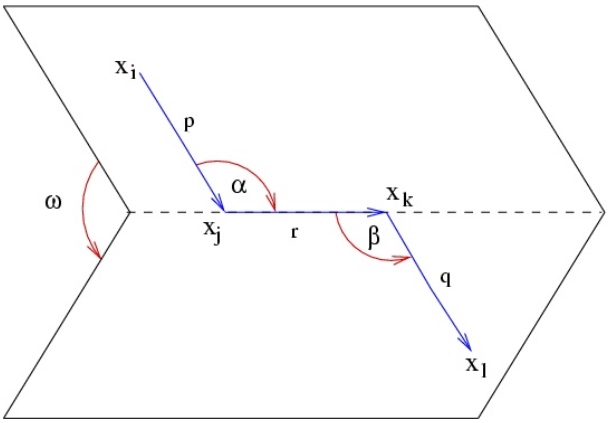


Abbildung B.1: Definition der Bindungs- und Dihedralwinkel

Als Maß für die Größe eines Proteins dient die mittlere quadratische Abweichung von dessen Mittelpunkt, der sogenannte Gyrationsradius R_g

$$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |\vec{x}_i - \vec{r}_c|^2} = \sqrt{\frac{1}{N^2}\sum_{\substack{i,j=1 \\ i<j}}^{N} |\vec{x}_i - \vec{x}_j|^2} \tag{B.4}$$

In anderen Kontexten berücksichtigt der Gyrationsradius die unterschiedlichen Massen der Atome. Für Proteine wird diese Unterscheidung zumeist ignoriert, da die Massen von Kohlenstoff, Sauerstoff und Stickstoff recht ähnlich sind.

Entlang einer Kette kovalenter Bindungen lassen sich gemäß der Notation in Abb. B.1 Bindungs- und Dihedralwinkel definieren. Den Bindungswinkel $\alpha = \sphericalangle(i{-}j{-}k) \in [0°; 180°)$ erhält man aus dem Skalarprodukt und der Torsionswinkel $\omega = \sphericalangle(i{-}j{-}k{-}l) \in (-180°; 180°]$ ist der Winkel zwischen den Normalenvektoren der Ebenen $i-j-k$ und $j-k-l$.

$$\cos\alpha = \frac{\vec{p}\cdot\vec{r}}{|\vec{p}||\vec{r}|} \tag{B.5}$$

$$\cos\beta = \frac{\vec{r}\cdot\vec{q}}{|\vec{r}||\vec{q}|} \tag{B.6}$$

$$\cos\omega = \frac{(\vec{p}\times\vec{r})\cdot(\vec{r}\times\vec{q})}{|\vec{p}\times\vec{r}||\vec{r}\times\vec{q}|} \tag{B.7}$$

Das Vorzeichen von ω ist das des Spatproduktes $(\vec{p}\times\vec{r})\cdot\vec{q}$.

Üblicherweise werden Bindungswinkel in einem Protein mit dem griechischen Buchstaben θ beschriftet, sowie die Dihedralwinkel, die die Drehung um die Hauptkettenbindungen $N - C_\alpha$, $C_\alpha - C$ und $C - N$ beschreiben, mit ϕ, ψ und ω. Die Dihedralwinkel der Seitenkette tragen die Bezeichnung χ_i $(i = 1, 2, \ldots)$ und werden von der Hauptkette nach außen durchnumeriert. Wie in der Einleitung

erwähnt (Seite 4), liegt es in der Natur der Peptidbindung begründet, daß ω nie deutlich von 180° abweicht. Im vorliegendem Kraftfeld wird ω beim dem Werte fixiert, den das Protein gemäß der Koordinaten in der Proteindatabank PDB hat.

Gemäß der Homogenität und Isotropie des Raumes definiert der vollständige Satz aus Bindungslängen, -winkel und Dihedralwinkeln eindeutig die Struktur des Proteins. Einige Aminosäuren enthalten Mehrfachbindungen und Ringstrukturen, sodaß auf die Angabe einiger Winkel verzichtet werden kann, ohne die Eindeutigkeit zu verlieren.

B.2.2 Das Ramachandran-Diagramm

Alle Sekundärstrukturelemente sind eindeutig gewissen Dihedralwinkelwerten zuzuordnen. Das häufigste Sekundärstrukturelement ist das der (rechtshändigen) α-Helix. Der axiale Abstand zwischen zwei aufeinanderfolgenden Monomeren ist 1.5Åund das Verhältnis zwischen der Steigung und dem axialen Abstand ist 3,6. Nach 5 Windungen und 18 Aminosäuren erhält man eine exakte Wiederkehr der Atomabfolge. Der Neigungswinkel zwischen zwei aufeinanderfolgenden Aminosäuren beträgt 100° und die Dihedralwinkelpaare ϕ, ψ sind für alle Aminosäuren gleich. Die Helix wird durch Wasserstoffbrückenbindungen zwischen der CO-Gruppe der i-ten und der HN-Gruppe der $i+4$-ten Aminosäure stabilisiert.

Eine andere möglich Helixkonfiguration ist die sogenannte 3_{10}-Helix. Hierbei bilden sich Wasserstoffbrücken zwischen der i-ten und $(i+3)$-ten Aminosäure aus. Die Stabilität der 3_{10}-Helix ist deutlich geringer als die der α-Helix, da die CO-Gruppe nicht entlang der Helix weist, sondern leicht nach außen geneigt ist. Die stabilisierenden Wasserstoffbrückenbindungen sind somit verhältnismäßig schlecht ausgebildet.

Ein anderes Motiv, bei dem sich Wasserstoffbrücken zwischen der i-ten und $i+3$-ten Aminosäure ausbilden, ist der β-Turn. Auf dieses Strukturelement soll hier, ebenso wir auf der β-Faltblatt, jedoch nicht weiter eingegangen werden. Die durchschnittlichen Dihedralwinkel der häufigsten Sekundärstrukturelemente sind in Tabelle B.1 angeben.

Struktur	ϕ [deg]	ψ [deg]	Residuen pro Windung	Abstand zweier Monomere [Å]
α-Helix	-57	-47	3.6	1.5
3_{10}-Helix	-49	-26	3.0	2.0
$\uparrow\uparrow$ β-Faltblatt	-119	113	-	-
$\uparrow\downarrow$ β-Faltblatt	-139	135	-	-

Tabelle B.1: Durchschnittswerte der Dihedralwinkel für die vier häufigsten Sekundärstrukturelemente[Dau98]

Trägt man die Dihedralwinkel von strukturaufgeklärten Proteinen in einem

Diagramm ein, so ergeben sich Häufungspunkte an den in der Tabelle aufgelisteten Stellen. In Abb. B.2 ist dies für den oben beschriebenen Proteinsatz M^{138} ohne die Aminosäuren Glycin und Prolin geschehen. Für die noch zu beschreibende Bestimmung der Lennard-Jones Parameter ist dieses Diagramm nicht ohne Relevanz, da genau dort die Proteine betrachtet werden, die auch in Abb. B.2 berücksichtigt wurden.

Wie in der Abbildung zu erkennen, kommen große Winkelbereiche bei nativen Strukturen nicht vor. Dies liegt laut allgemeiner Lehrmeinung in der van-der-Waals Wechselwirkung begründet, denn bei diesen Winkeln kommt es zum Überlapp verschiedener Atomorbitale. In der Hard-Sphere Approximation, die den Überlapp zweier Orbitale gänzlich verbietet, lassen sich die "verbotenen" Bereiche besonders deutlich erkennen (Abbildung B.3). Da Glycin kein C_β Atom vorweisen kann, ist ihm auch der Bereich frei zugänglich, der für andere Aminosäuren wegen eines Überlapps mit dem C_β-Atomorbital verboten ist. Entsprechend ist die Verteilung der Dihedralwinkel signifikant von denen der anderen Aminosäuren verschieden (Abbildung B.4). Die Bezeichnung "verbotener" Bereich ist, zumindest bei einem Überlapp mit einem Wasserstofforbital, stark überzogen, denn wie man im Vergleich der beiden Diagramme B.2 und B.3 im Punkt $(\phi;\psi) = (-90°;0°)$ erkennt, ist dieser Bereich sehr wohl populiert.

Prolin ist die einzige Aminosäure, die zwei kovalente Bindungen zwischen Haupt- und Seitenkette besitzt (Abbildung B.5). Aus diesem Grund sind die Dihedralwinkel stärker eingeschränkt als bei allen anderen Aminosäuren, und so geht Prolin im allgemeinen nicht in das Ramachandran-Diagramm ein.

B.2.3 Sekundärstrukturanalyse

Eine Helix wird durch Wasserstoffbrückenbindungen stabilisiert, wobei sich die Wasserstoffbrücken zwischen der i-ten und $(i+4)$-ten Aminosäure ausbilden. Folglich können 3 aufeinanderfolgende Aminosäuren keine (stabile) Helix bilden, selbst wenn ihre Dihedralwinkel die entsprechenden Werte einnehmen. Diesem Umstand muß man Rechnung tragen, will man aus der Winkelverteilung Rückschlüsse auf die Sekundärstruktur ziehen. Ähnlich ist die Situation bei β-Faltblättern. So wird aus einem einzelnen Aminosäurenstrang erst dann ein β-Faltblatt, wenn dieser über Wasserstoffbrücken mit einem zweiten Strang verbunden ist. Eine zuverlässige Zuordnung von Sekundärstruktur muß sich folglich an den Dihedralwinkeln und der Elektrostatik der Hauptkette orientieren.

Für die Sekundärstrukturanalyse verwenden wir ein Programm von Kabsch und Sanders, welches unter dem Namen DSSP (*Database of Secondary Structure in Proteins*) frei erhältlich ist[KS83]. Es ist eines der weitverbreitesten Sekundärstrukturanalyseprogramme. Die Proteinabbildungen dieser Arbeit sind weitestgehend mit *Molscript*[Kra91] und *Raster3D*[LB00] erstellt worden, wobei das Sekundärstrukturanalyseprogramm von *Molscript* durch DSSP ersetzt wurde.

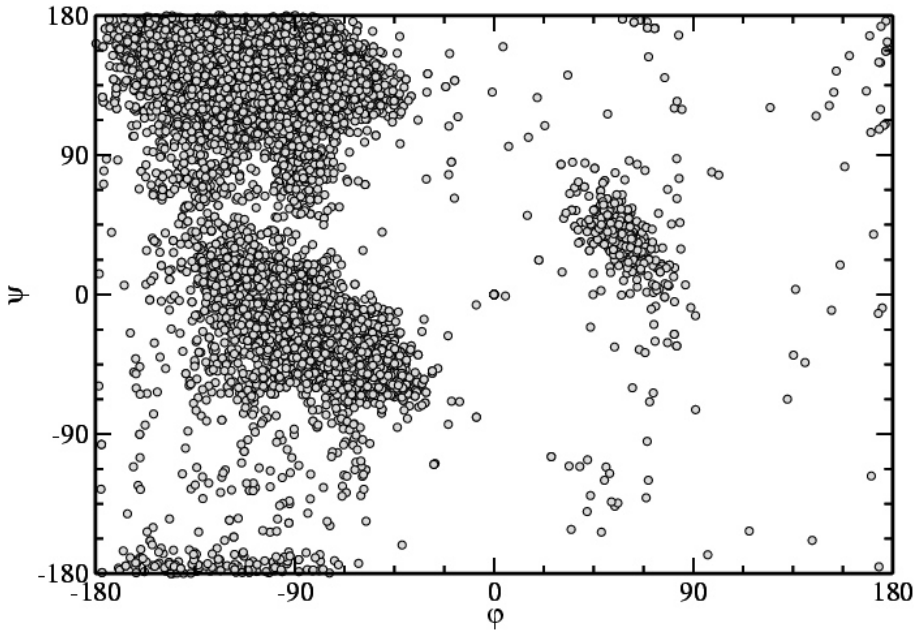


Abbildung B.2: Ramachandran-Diagramm strukturaufgeklärter Proteine (ohne die Aminosäuren Glycin und Prolin)

Bei DSSP werden die Sekundärstrukturen auf 8 verschiedene Zustände abgebildet:

- H = α-Helix
- G = 3_{10}-Helix
- I = π-Helix
- B = Residuum in isolierter β-Brücke
- E = ausgedehnter β-Strang
- T = Wasserstoffbrückengebundene Drehung (*turn*)
- S = Kurve (*bend*)
- C = Knäuel

Das Strukturelement *C* für Knäuel wird eingefügt, wenn kein Kriterium für die anderen sieben Strukturelemente greift. Eine Folge von mehreren aufeinanderfolgenden *C*s in der Sekundärstruktur steht für einen "unstrukturierten" Bereich. Um diese Bereiche optisch besser sichtbar zu machen, wird für ein Knäuel auch ein Kleinbuchstabe ("c") niedergeschrieben.

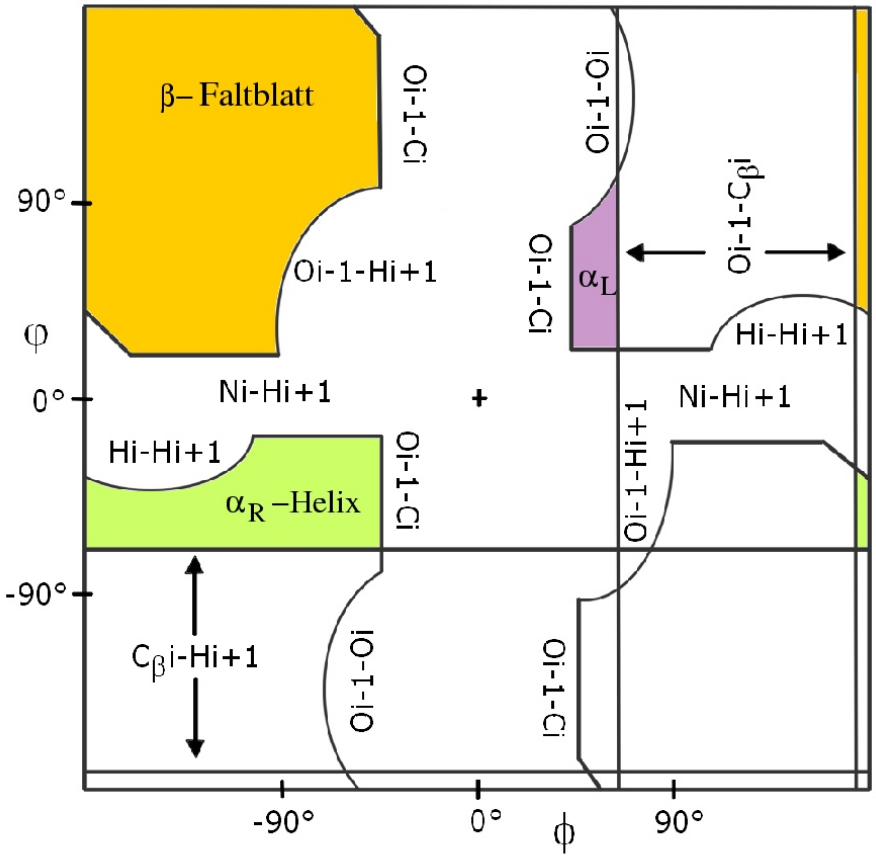


Abbildung B.3: Ramachandran-Diagramm in der Hard-Sphere Approximation. Die erlaubten Bereiche sind farbig gekennzeichnet und die verbotenen Zonen sind den überlappenden Atomen zugeordnet, z.B. kennzeichnet $C_{\beta i} - H_{i+1}$ einen Überschneidung des C_β-Atoms der i-ten Aminosäure mit dem H-Atom der Hauptketten NH-Gruppe in Aminosäure $i + 1$.

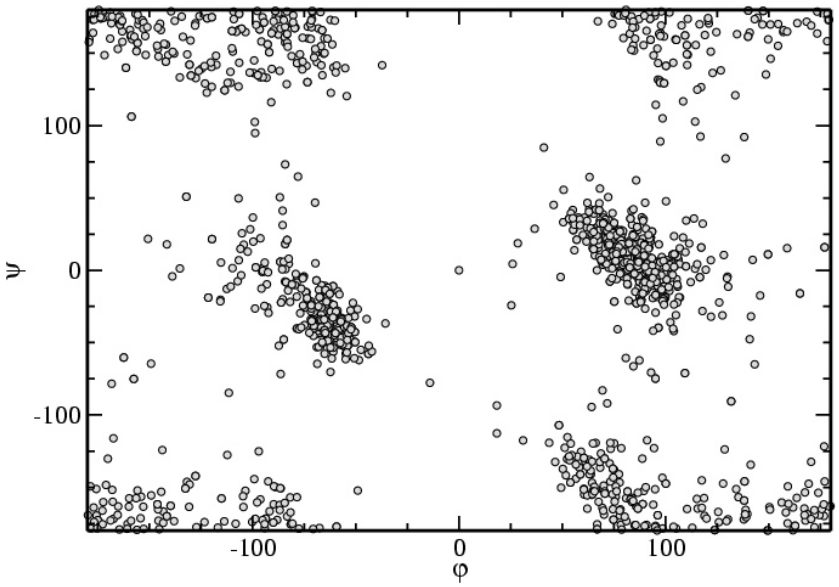


Abbildung B.4: Ramachandran-Diagramm für Glycin

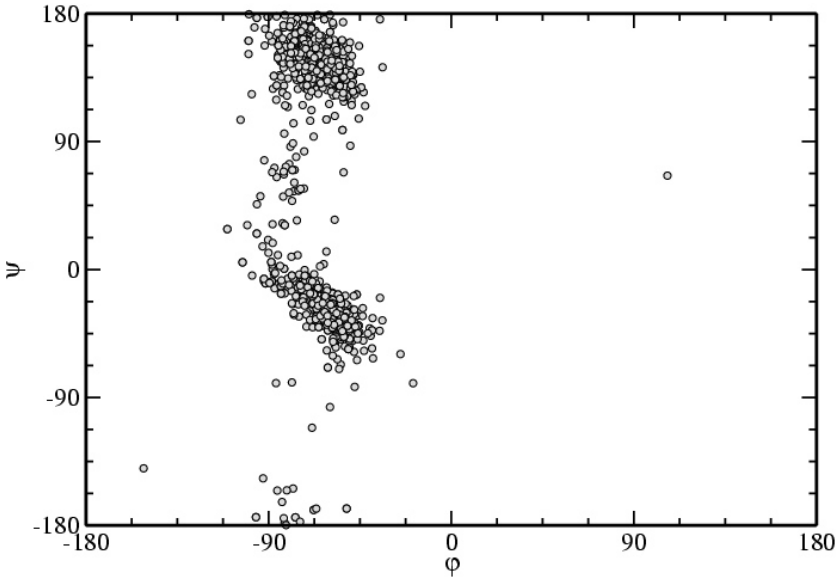


Abbildung B.5: $\phi - \psi$-Diagramm für Prolin

B.3 Vergleich zweier Strukturen

Zwei Möglichkeiten zur Gegenüberstellung mehrerer Konfigurationen sind praktisch schon genannt worden. So kann man die von DSSP erhaltene Sekundärstrukturanalyse der zwei Strukturen gegenüberstellen, oder die Ramachandran Diagramme miteinander vergleichen. Dabei ist es hilfreich, sich nicht das Ramachandran Diagramm des gesamten Proteins anzusehen, sondern jeweils nur auszugsweise von Residuum i bis Residuum j.

B.3.1 Root Mean Square Derivation

Für die Simulation ist es wichtig, zwei Strukturen energetisch vergleichen zu können. Um zu klären, ob eine Simulation die native Struktur eingenommen hat, benötigen wir ein Maß für den Unterschied bzw. Abstand zwischen zwei Strukturen.

Einen Abstandsbegriff für zwei Strukturen I und J mit den Koordination $\{\vec{r}_i^I\}_{i=1}^N$ und $\{\vec{r}_j^J\}_{j=1}^N$ kann man definieren durch

$$\sqrt{\frac{1}{N}\sum_{i,j=1}^{N}|\vec{r}_i - \vec{r}_j|^2} \quad . \tag{B.8}$$

Allerdings muß man diesen Begriff modulo Translationen $\vec{t}$ und Rotationen R des Proteins als Ganzes verstehen. Daher betrachten wir als Abstand zweier Strukturen folgende *Root Mean Square Derivation* [1]

$$RMSD(I,J) := \min_{\vec{t},R} \sqrt{\frac{1}{N}\sum_{i,j=1}^{N}|\vec{r}_i^I - R\vec{r}_j^J + \vec{t}|^2} \quad . \tag{B.9}$$

Diese mißt den Abstand nach bestmöglicher Überlagerung der beiden Strukturen. Sie weicht allerdings rasch von Null ab, auch wenn sich die beiden Strukturen für das menschliche Auge als sehr ähnlich präsentieren. Sehr oft wird daher der RMSD-Wert nur aus den Hauptkettenatomen berechnet. Wir schreiben kurz RMSB statt $\text{RMSD}_{backbbone}$. Will man die Anzahl der Atome noch weiter reduzieren, so beschränkt man sich auf die C_α-Atome. Diese geben im Gegensatz etwa zu den Stickstoffatomen der Hauptkette, rudimentär die Lage der Seitenketten wieder und sind gleichzeitig diejenigen Atome der Hauptkette, deren Dihedralwinkel drehbar sind. Dieser Wert wird mit $RMSD_\alpha$ bezeichnet.

Einige Strukturmerkmale gehen bei den Werten für RMSB und RMSD_α jedoch verloren. So macht es keinen Unterschied, ob eine Seitenkette nach innen

[1] Wie man leicht zeigen kann ist $\vec{t} = 0$ für $\langle\vec{r}^I\rangle = \langle\vec{r}^J\rangle$ und es verbleibt die Aufgabe die Spur einer 3×3-Matrix zu maximieren. D.h. man kommt bei $R = R_x(\alpha)R_y(\beta)R_z(\gamma)$ und $\frac{\partial}{\partial\alpha,\beta,\gamma}\text{RMSD} = 0$ ohne komplizierte Ausdrücke trigonometrischer Funktionen aus.

oder nach außen weist. Folglich geben diese Werte nur an, inwieweit vorhandene Sekundärstrukturelemente richtig angeordnet sind. Positiv zu vermerken ist, daß die richtige Anordnung der Sekundärstruktur jedoch oftmals mit der richtigen Lage der Seitengruppen einhergeht.

B.3.2 Das C_β-Mosaik

Um auch diejenigen Fälle, für die obige Bedingung nicht gilt, zu erfassen, wird gelegentlich der relative Abstand der C_β-Atome innerhalb einer Struktur betrachtet und die $C_\beta - C_\beta$-Abstandsmatrizen zweier Strukturen voneinander subtrahiert. Liegt der Abstandsunterschied unterhalb von 0.75Å so wird eine schwarze, zwischen 0.75 und 1.5Å ein graue Markierung in ein Mosaik eingezeichnet. Das so erhaltene Mosaik gibt uns einen guten Eindruck über die Kontakte und somit über die Lage der Seitenketten. Da bei α-Helices und β-Faltblättern die C_β-Abstände fixiert sind, kann man diese Sekundärstrukturelemente ebenfalls in diesen Abbildungen wiederfinden, sofern sie in beiden Strukturen vorhanden sind. Die Abstandsintervalle $[0; 0.75]$ und $(0.75; 1.5]$ liegen unterhalb der experimentellen Auflösung und sind daher ihrerseits mit großen Fehlern behaftet. Wir haben uns für diese Intervalle entschieden, um auch zwei Strukturen aus den Simulationen miteinander vergleichen zu können. Das Werkzeug des C_β-Mosaiks ist sehr empfindlich gegenüber strukturellen Unterschieden.

Für die Beschränkung auf die C_β-Atome gibt es mehrere Gründe. So ist die Ausrichtung der Seitenkette maßgeblich durch dieses Atom bestimmt, welches seinerseits eine gewisse Unempfindlichkeit gegenüber der exakten Position des Seitengruppenrestes hat. Hiermit haben wir das Problem, welches auch beim RMSD-Wert vorliegt, vermeiden können. Auch ist die geringe Größe des C_β-Mosaiks von Vorteil. Schließlich handelt es sich hier nur um eine von mehreren möglichen Analysetechniken, sodaß zuviele Details nur hinderlich wären.

B.4 Ein Beispiel: 1BHI

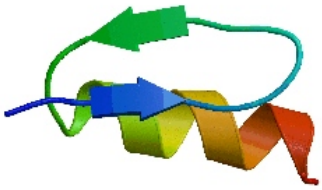
In Abb. B.6 sind die native Struktur von 1BHI, einem Protein mit 29 Residuen, sowie drei aus verschiedenen Monte-Carlo Simulationen gewonnenen Strukturen (Decoys), abgebildet. Die native Struktur und das Decoy A sind in der vereinfachten Hauptkettendarstellung sehr ähnlich. Wäre diese Struktur durch einen Faltungsversuch und nicht durch lokale Relaxation entstanden, so würden wir dies als erfolgreiche Faltung bezeichnen. Dennoch gibt es kleine Abweichungen. Ein Unterschied ist aus der Sekundärstrukturanalyse zu erkennen. In der nativen Struktur ist das β-Faltblatt zwischen den Residuen 3, 4 und 13, 14 ausgebildet, wohingegen bei Decoy A die Residuen 5 und 6 beteiligt sind. Im C_β-Mosaik ist der Helixbereich von Residuum 21 aufwärts schwarz eingefärbt. Wenn im Verhältnis zur nativen Struktur nur der Bereich vom N-Terminus bis zum Turn um 2 Re-

siduen verschoben wäre, wie die Sekundärstrukturanalyse suggeriert, ohne den Rest der Struktur zu verändern, wäre im C_β-Mosaik auch der Bereich etwa ab Residuum 10 dunkler, d.h. das β-Faltblatt ist nicht nur verschoben sondern auch geringfügig um die Helix herumgedreht.

In Decoy B bilden die ersten 6 Residuen eine Helix. Im C_β-Mosaik ist die Nebendiagonale der ersten Residuen trotz des strukturellen Unterschiedes schwarz. Den schwarzen Mosaiksteinen der ersten Nebendiagonalen kommt somit kaum eine Aussagekraft zu. Erst wenn die schwarzen Mosaiksteine die Diagonale verlassen, stimmt die Geometrie der beiden Strukturen lokal überein. Dies zeigt sich, an der ersten Residuen des Decoys C, welche eine β-Faltblatt ausbilden und im C_β-Mosaik eine deutliche Färbung des oberen linken Bereiches bewirken.

Die Tabelle der RMS-Abweichungen kann als Anhaltspunkt dafür genommen werden, daß eine RMSD-Abweichung von unter 4.0Å notwendig ist, um die Struktur verhältnismäßig gut zu approximieren. Im Vergleich dazu liegt die experimentelle Auflösung im Bereich von 2Å, dem Bereich natürlichen Fluktuationen in wässriger Lösung.

Das wichtigste Instrument ist die interaktive dreidimensionale Darstellung am Computer. Hier ist es möglich, zwischen verschiedenen Darstellungsweise zu wechseln; etwa von der Darstellung der Sekundärstruktur zur Darstellung aller Atome und deren kovalenten Bindungen. Im Zusammenspiel mit den Energiebeiträgen der einzelnen Aminosäuren ist es möglich, diejenigen Proteinabschnitte zu identifizieren, die bei mehreren Strukturen gleich bzw. unterschiedlich sind. Somit lassen sich die Charakteristika der jeweiligen Struktur recht schnell herausarbeiten. Diese können ihrerseits dann mit anderen Methoden weiter analysiert werden.

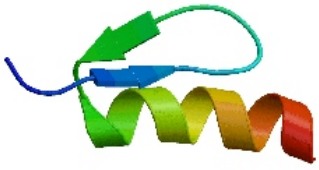


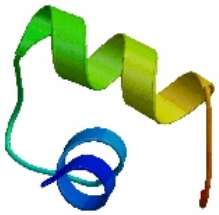
native Struktur von 1BHI

Decoy C

Decoy B

Decoy A





C_β–Mosaik:

Sekundärstruktur:

nativ	ccEEcccTTTccEESSHHHHHHHHHHHHc
A	ccccEEcSScccEEScHHHHHHHHHHHHc
B	cHHHHHScTTTTcccHHHHHHHHHTcScc
C	cccEEEcccTTccccHHHHHHScccEEEc

RMS–Abweichungen zur nativen Struktur:

	RMSD	RMSB	$RMSD_\alpha$
A	3.83	2.91	2.99
B	8.98	6.43	6.67
C	9.18	8.32	8.10

Abbildung B.6: Ein einfaches Analyseschema am Beispiel 1BHI

Anhang C

Datentabellen der Kraftfeldparameter

Die analytische Form des Kraftfeldes PFF01 lautet:

$$
\begin{array}{rcll}
E = & \displaystyle\sum_{i,j} & \left(\frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^{6}}\right) & (=: E_{lj}) \text{ Lennard-Jones} \\
+ & \displaystyle\sum_{i,j} & \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon_{\kappa(i),\kappa(j)}\, r} & (=: E_{side}) \text{ Seitenketten-ES} \\
 & i \in \text{Seitenkette} & & \\
+ & \displaystyle\sum_{i,j \in \text{Hauptkette}} & \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon_{hb}\, r} & (=: E_{main}) \text{ Hauptketten-ES} \\
+ & \displaystyle\sum_{CO\cdots HN} & f(CO\cdots HN) & (=: E_{hb}) \;\; H\text{-Brücken} \\
+ & \displaystyle\sum_{i} & \sigma_{pt(i)}\pi_{res(i)}A_i & (=: E_{pse}) \text{ Lösungsmittel}
\end{array}
\tag{C.1}
$$

Die Parametrisierung der Funktion $f(CO\cdots HN)$ ist in Gleichung 4.31 wiedergegeben. Es bezeichnet r den Abstand zweier Atome i und j. A_i ist die freie Oberfläche des Atoms i. Die Summen erstrecken sich nur über diejenigen Atompaare, die nicht über 1, 2 oder 3 konsekutive kovalente Bindungen miteinander verbunden sind. Die Werte für r und A_i sind von der Proteinkonfiguration abhängig. Alle anderen auftretenden Werte sind durch die Chemie des Proteins festgelegt und stimmen größtenteils überein. So sind etwa die Lennard-Jones Parameter A_{ii} und B_{ii} für alle sp^3-hybridisierten Kohlenstoffatome gleich, ebenso wie die (relative) Dielektrizitätskonstante bei gleichen funktionellen Gruppen identisch ist. Die κ-Werte liefern dabei die Zuordnung der Atome zu ihren funktionellen Gruppen. Dennoch werden im folgenden die Parameter jedes Atoms einzeln aufgelistet, um eine direkte Gegenüberstellung der Aminosäuren und der Parameter zu ermöglichen.

Die Hauptkette

Ein Protein ist eine lineare Kette von Aminosäuren, wobei die C_α-Atome von zwei Peptidbindungen flankiert sind. Lediglich die beiden Termini, also die erste und letzte Aminosäure, nehmen an nur einer Peptidbindung teil. Die NH_2-Gruppe des N-Terminus ist stets protoniert und die $COOH$-Gruppe des C-Terminus deprotoniert. Wenn wir die atomare Zusammensetzung der Hauptkette einer Aminosäure angeben wollen, benötigen wir drei verschiedene Varianten, N-Terminus, "normale" Hauptkette und den C-Terminus. Triglycin entspricht (unter Vernachlässigung der nichtpolarisierten Wasserstoffatome) genau dieser Anordnung.

Das Triglycin ist in Abb. C.1 dargestellt. In der darunterstehenden Tabelle sind die zugehörigen Identifikationen und Parameter aufgelistet. Diese Tabelle ist folgendermaßen aufgebaut. In der ersten Spalte ist der Name der Aminosäure angegeben, wobei der dreibuchstabigen Abkürzung noch ein N für N-Terminus oder ein C für C-Terminus zugefügt wird. In der Tabelle folgt in der 2. Spalte der Name des Atoms, wobei der erste Buchstabe den Typ des Atoms angibt: H für Wasserstoff, C für Kohlenstoff u.s.w.. Der zweite Buchstabe entspricht im Normalfall der Durchnumerierung der Seiteketten in griechischen Buchstaben (wobei Programm intern C_α mit CA, C_β mit CB bezeichnet wird). Die dritte und vierte Spalte gibt den Potentialtyp an, dessen erster Buchstabe wieder dem Typen des Atoms entspricht. Im CARB Kraftfeld sind 34 verschiedene Potentialtypen enthalten, die zunächst auf 26 (3.Spalte) und anschließend auf elf reduziert wurden (4.Spalte). Da diese Zuordnung jedoch teilweise verändert worden ist und in zukünftigen Kraftfeldverbesserungen eventuell wieder abgeändert wird, sind beide Notationen aufgeführt. Zur besseren Unterscheidung werden wir die Bezeichnungen der 3. Spalte als CARB Notation und die der 4. Spalte als PFF01 Notation referieren.

Die PFF01 Notation hat folgenden Aufbau: Der Großteil der Kohlenstoffatome besitzt den Potentialtyp *cme*. Der Buchstabe *c* steht für Kohlenstoff und das Suffix *me* ist die Abkürzung für Methan, womit angedeutet werden soll, daß dieses Kohlenstoffatom die gebundenen Wasserstoffatome mit einschließt. Davon separiert sind Kohlenstoffatome, die an stark polare Sauerstoff oder Stickstoffatome gebunden sind. Diese werden mit *cp* bezeichnet. Die dritte Kohlenstoffvariante liegt in Ringstrukturen, wie in Phenylalanin, Tyrosin, Histidin und Tryptophan, und trägt die Bezeichnung *cr*. Alle einfach an Kohlenstoff gebundenen Sauerstoffatome sind vom Typ *o*1; z.B. in Serin, Threonin und Tyrosin. Sauerstoffatome mit Mehrfachbindungen, darunter auch die mesomeriestabilisierten Bindungen in COO^-, werden mit *o*2 bezeichnet; letztere finden sich in Asparagin- und Glutaminsäure, erstere in Asparagin und Glutamin. Das Stickstoffatom der NH Gruppe in Arganin, Histidin und Tryptophan ist $n1$; das der NH_2 Gruppen in Asparagin und Glutamin ist $n2$. Der Potentialtyp $n3$ ist speziell für das Stickstoffatom der NH_3^+ Gruppe des Lysin reserviert. Die Schwefelatome in Cystin und Methionin sind beide vom Typ *s*. Bei den Wasserstoffatomen unterscheiden wir

das der Hauptkette (hb) von allen weiteren (h).

Die in den folgenden Tabellen rechtsstehenden Parameter sind, wie der Gleichung C.1 zu entnehmen, die Partialladung q, der Index $\kappa \in [0;6]$ für die Dielektrizitätskonstante $\epsilon_{\kappa(i),\kappa(j)}$, die beiden Lennard-Jones Parameter A, B und der Lösungsmittelparameter σ. Es sei vorweg erwähnt, daß $\epsilon = \infty$ für $\kappa = 0$ gilt, d.h. daß Atome mit $\kappa = 0$ nicht in die elektrostatische Wechselwirkung eingehen. Zumeist tragen Atome mit $\kappa = 0$ keine Partialladungen und $\kappa = 0$ sorgt Programm intern dafür, daß nicht unnötigerweise Nullen aufsummiert werden. Die Klammerung der κ-Werte dient der Einteilung der Atome in "elektrostatische Gruppen", kurz ES-Gruppen.

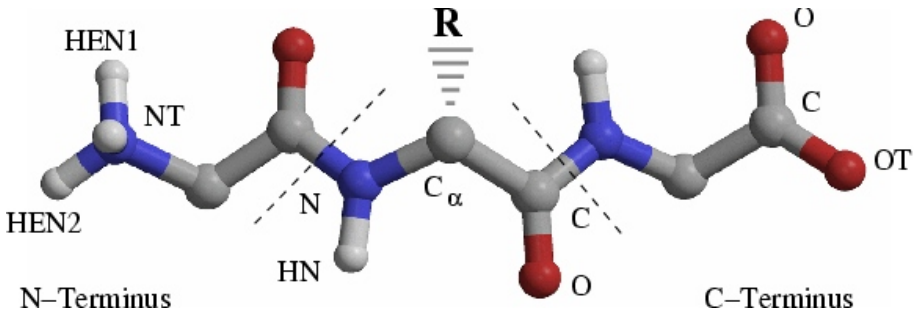
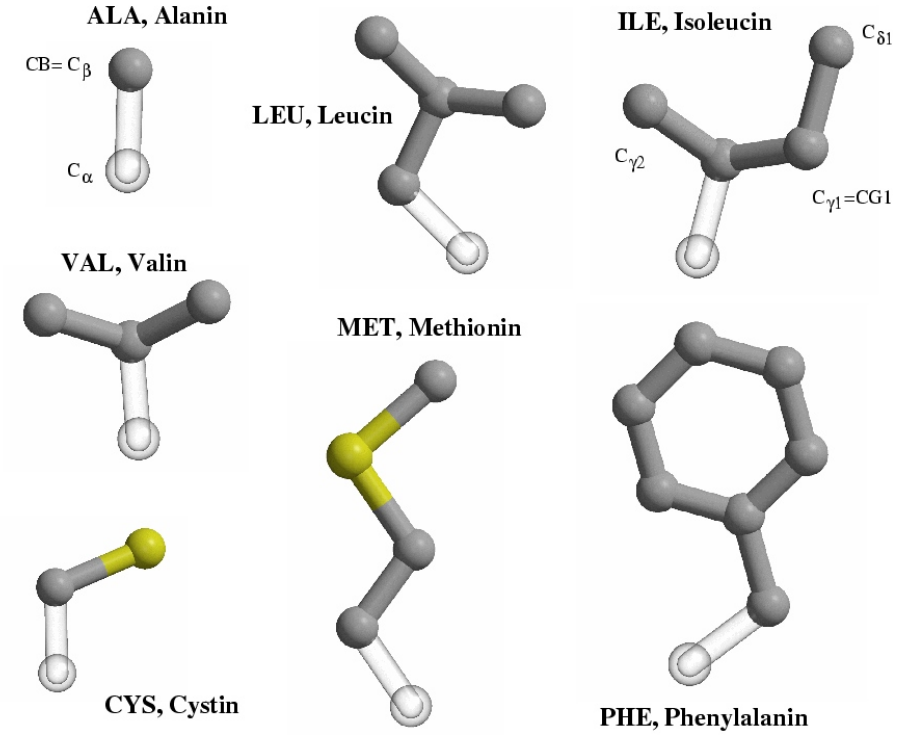


Abbildung C.1: Aufbau des Triglycin. Unter Variation des Restes R entstehen die anderen 19 Tripeptide der Form $Gly - X - Gly$; $X = Ala, Asn, \ldots$

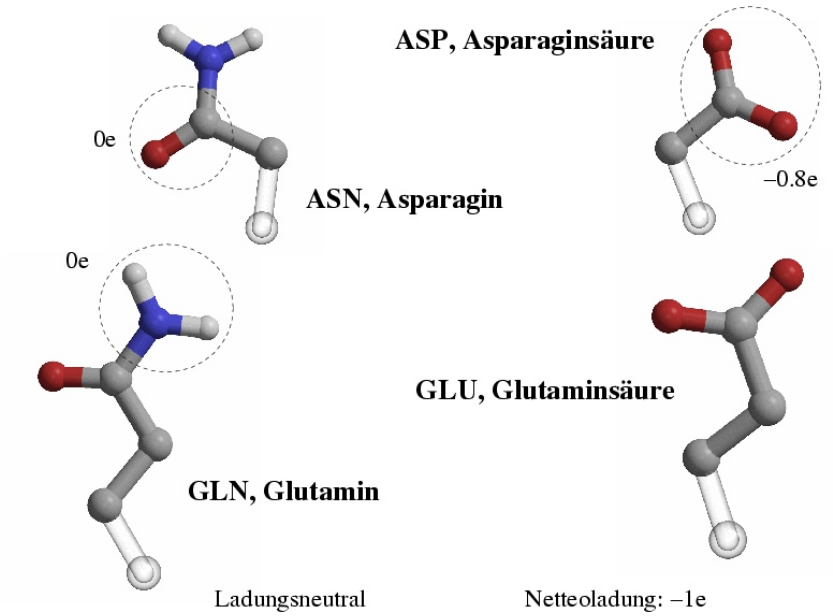
Aminosäuren				q_i	$\kappa(i)$	A_{ii}	B_{ii}	σ_i
GLYN	NT	n3	n3	-0.600	0	40062.70	40.03	-15
GLYN	HEN1	h3	h	0.450	0	30.23	1.10	-15
GLYN	HEN2	h3	h	0.450	0	30.23	1.10	-15
GLYN	HEN3	h3	h	0.450	0	30.23	1.10	-15
GLYN	CA	cb	cme	0.250	0	225634.90	95.00	28
GLYN	C	cd	co	0.380	⌜2	225634.90	95.00	-2
GLYN	O	odd	o1	-0.380	⌞2	7876.63	17.75	-10
GLY	N	n	n1	-0.280	⌜1	40062.70	40.03	-10
GLY	HN	hn	hn	0.280	⌞1	97.56	1.98	-10
GLY	CA	caa	cme	0.000	0	225634.90	95.00	28
GLY	C	cd	cp	0.380	⌜2	225634.90	95.00	-2
GLY	O	odd	o1	-0.380	⌞2	7876.63	17.75	-10
GLYC	N	n	n1	-0.280	⌜1	40062.70	40.03	-10
GLYC	HN	hn	hn	0.280	⌞1	97.56	1.98	-10
GLYC	CA	caa	cme	-0.250	0	225634.90	95.00	28
GLYC	C	cd	cp	0.650	0	225634.90	95.00	-2
GLYC	O	odd	o1	-0.700	0	7876.63	17.75	-10
GLYC	OT	o-	o1	-0.700	0	7876.63	17.75	-10

Die Seitenketten

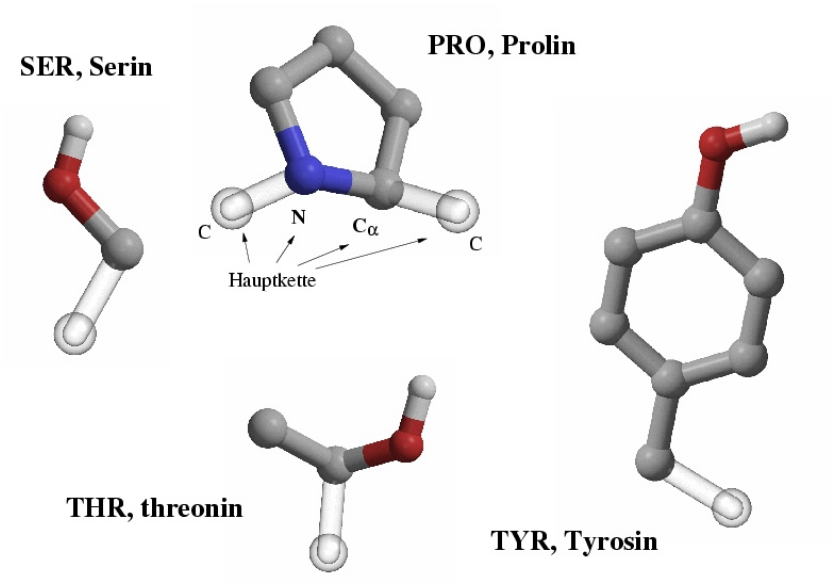
In den folgenden Abbildungen und Tabellen finden sich die Parameter der 19 Aminosäurenreste. Die Bindung zur Hauptkette ist jeweils angedeutet.



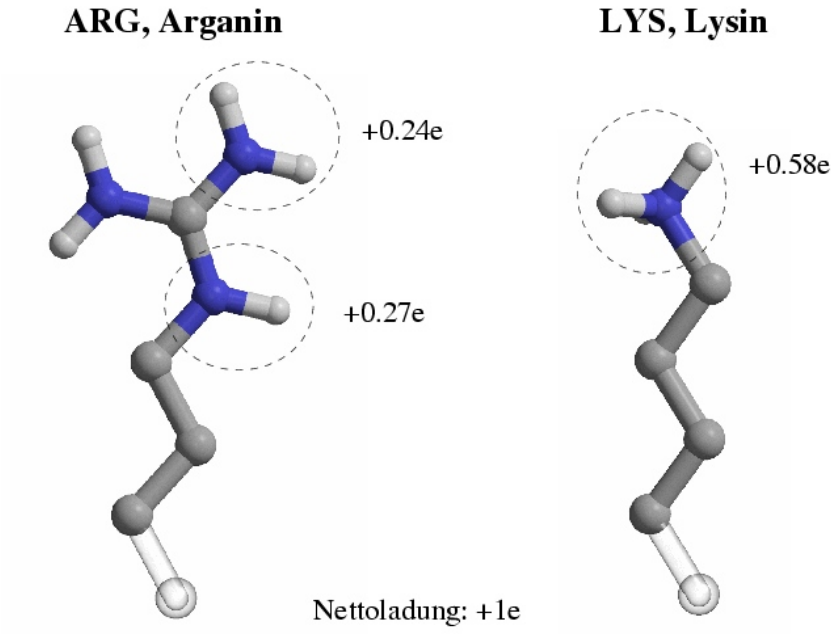
				q_i	$\kappa(i)$	A_{ii}	B_{ii}	σ_i
ALA	CB	cme	cme	0.000	0	225634.90	95.00	28
ILE	*	cme	cme	0.000	0	225634.90	95.00	28
LEU	*	cme	cme	0.000	0	225634.90	95.00	28
VAL	*	cme	cme	0.000	0	225634.90	95.00	28
PHE	CB	cme	cme	0.000	0	225634.90	95.00	28
PHE	CG -Z	cr	cr	0.000	0	32094.71	35.83	31
CYS	CB	cme	cme	0.000	0	225634.90	95.00	28
CYS	SG	s	s	0.000	0	90657.38	60.22	28
MET	CB	cme	cme	0.000	0	225634.90	95.00	28
MET	CG	cme	cme	0.000	0	225634.90	95.00	28
MET	SD	s	s	0.000	0	90657.38	60.22	28
MET	CE	cme	cme	0.000	0	225634.90	95.00	28



				q_i	$\kappa(i)$	A_{ii}	B_{ii}	σ_i
ASN	CB	cme	cme	0.000	0	225634.90	95.00	28
ASN	CG	c-	cp	0.380	⌜5	225634.90	95.00	-2
ASN	OD1	od	o2	-0.380	⌞5	7876.63	17.75	-5
ASN	ND2	n2	n2	-0.560	⌜4	40062.70	40.03	-5
ASN	HNA	h2	h	0.280	4	30.23	1.10	-5
ASN	HNB	h2	h	0.280	⌞4	30.23	1.10	-5
ASP	CB	cme	cme	-0.200	⌜6	225634.90	95.00	28
ASP	CG	c-	cp	0.340	6	225634.90	95.00	-2
ASP	OD1	o-	o2	-0.570	6	7876.63	17.75	-5
ASP	OD2	o-	o2	-0.570	⌞6	7876.63	17.75	-5
GLN	CB	cme	cme	0.000	0	225634.90	95.00	28
GLN	CG	cme	cme	0.000	0	225634.90	95.00	28
GLN	CD	c-	cp	0.380	⌜5	225634.90	95.00	-2
GLN	OE1	od	o2	-0.380	⌞5	7876.63	17.75	-5
GLN	NE2	n2	n2	-0.560	⌜4	40062.70	40.03	-5
GLN	HNA	h2	h	0.280	4	30.23	1.10	-5
GLN	HNB	h2	h	0.280	⌞4	30.23	1.10	-5
GLU	CB	cme	cme	0.000	0	225634.90	95.00	28
GLU	CG	cme	cme	-0.200	⌜6	225634.90	95.00	28
GLU	CD	c-	cp	0.340	6	225634.90	95.00	-2
GLU	OE1	o-	o2	-0.570	6	7876.63	17.75	-5
GLU	OE2	o-	o2	-0.570	⌞6	7876.63	17.75	-5



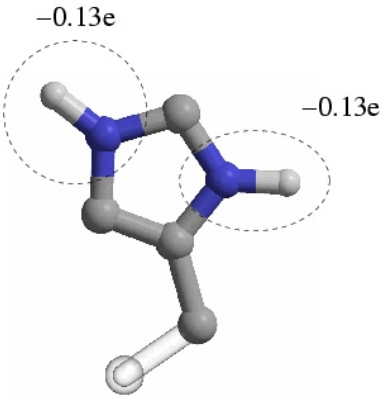
				q_i	$\kappa(i)$	A_{ii}	B_{ii}	σ_i
PRO	N	n	n1	-0.420	0	40062.70	40.03	-10
PRO	CA	caa	cme	0.210	0	225634.90	95.00	28
PRO	CB	cme	cme	0.000	0	225634.90	95.00	28
PRO	CG	cme	cme	0.000	0	225634.90	95.00	28
PRO	CD	cme	cme	0.210	0	225634.90	95.00	28
SER	CB	cme	cme	0.030	⌜3	225634.90	95.00	28
SER	OG	os	o1	-0.380	3	7876.63	17.75	-10
SER	HOG	hd	h	0.350	⌞3	30.23	1.10	-10
THR	CB	cme	cme	0.030	⌜3	225634.90	95.00	28
THR	OG1	os	o1	-0.380	3	7876.63	17.75	-10
THR	HOG	hd	h	0.350	⌞3	30.23	1.10	-10
THR	CG2	cme	cme	0.000	0	225634.90	95.00	28
TYR	CB	cme	cme	0.000	0	225634.90	95.00	28
TYR	CG	cr	cr	0.000	0	32094.71	35.83	31
TYR	CD1	cr	cr	0.000	0	32094.71	35.83	31
TYR	CD2	cr	cr	0.000	0	32094.71	35.83	31
TYR	CE1	cr	cr	0.000	0	32094.71	35.83	31
TYR	CE2	cr	cr	0.000	0	32094.71	35.83	31
TYR	CZ	cr	cr	0.030	⌜3	32094.71	35.83	31
TYR	OH	oy	o1	-0.380	3	7876.63	17.75	-10
TYR	HOH	hd	h	0.350	⌞3	30.23	1.10	-10



				q_i	$\kappa(i)$	A_{ii}	B_{ii}	σ_i
ARG	CB	cme	cme	0.000	0	225634.90	95.00	28
ARG	CG	cme	cme	0.000	0	225634.90	95.00	28
ARG	CD	cme	cme	0.190	⌜6	225634.90	95.00	28
ARG	NE	na	n1	-0.500	6	40062.70	40.03	-10
ARG	HNE	h1	h	0.370	6	30.23	1.10	-10
ARG	CZ	cb [1]	cp	0.460	6	225634.90	95.00	-2
ARG	NH1	n+	n1	-0.500	6	40062.70	40.03	-10
ARG	NH2	n+	n1	-0.500	6	40062.70	40.03	-10
ARG	HHA	h+	h	0.370	6	30.23	1.10	-10
ARG	HHB	h+	h	0.370	6	30.23	1.10	-10
ARG	HHC	h+	h	0.370	6	30.23	1.10	-10
ARG	HHD	h+	h	0.370	⌟6	30.23	1.10	-10
LYS	CB	cme	cme	0.000	0	225634.90	95.00	28
LYS	CG	cme	cme	0.000	0	225634.90	95.00	28
LYS	CD	cme	cme	0.120	⌜6	225634.90	95.00	28
LYS	CE	cb	cp	0.300	6	225634.90	95.00	-2
LYS	NZ	n3	n3	-0.500	6	40062.70	40.03	-15
LYS	HZA	h3	h	0.360	6	30.23	1.10	-15
LYS	HZB	h3	h	0.360	6	30.23	1.10	-15
LYS	HZC	h3	h	0.360	⌟6	30.23	1.10	-15

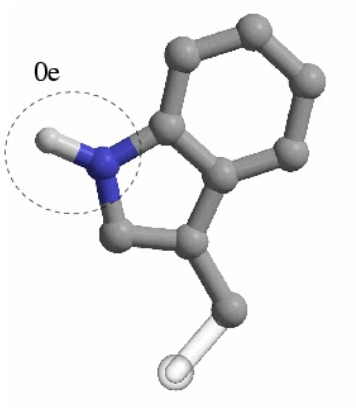
[1] Im CARB Kraftfeld ist das C_ζ Atom vom Typ *cme*. Allerdings ist es dem C_ζ aufgrund seiner starken polaren Ladung möglich, mit dem Sauerstoff des Wassers ähnlich stark zu interagieren, wie es den polaren Wasserstoffatomen des Arganin möglich ist[Smi94] und wird in gewissen Kontexten als Wasserstoffbrücken-Donator bezeichnet. Somit ist dieses Kohlenstoff

HIS, Histidin



Nettoladung: +1e

TRP, Tryptophan



				q_i	$\kappa(i)$	A_{ii}	B_{ii}	σ_i
HIS	CB	cme	cme	0.110	⌜6	225634.90	95.00	28
HIS	CG	cr	cr	0.170	6	32094.71	35.83	31
HIS	ND1	n1	n1	-0.500	6	40062.70	40.03	-10
HIS	HD1	hp	h	0.370	6	30.23	1.10	-10
HIS	CD2	cr	cr	0.330	6	32094.71	35.83	31
HIS	CE1	cr	cr	0.650	6	32094.71	35.83	31
HIS	NE2	n1	n1	-0.500	6	40062.70	40.03	-10
HIS	HE2	hp	h	0.370	⌞6	30.23	1.10	-10
TRP	CB	cme	cme	0.000	0	225634.90	95.00	28
TRP	CG	cr	cr	0.000	0	32094.71	35.83	31
TRP	CD1	cr	cr	0.000	0	32094.71	35.83	31
TRP	CD2	cr	cr	0.000	0	32094.71	35.83	31
TRP	NE1	nw	n1	-0.280	⌜6	40062.70	40.03	-10
TRP	HNE	h1	h	0.280	⌞6	30.23	1.10	-10
TRP	CE2	cr	cr	0.000	0	32094.71	35.83	31
TRP	CE3	cr	cr	0.000	0	32094.71	35.83	31
TRP	CZ2	cr	cr	0.000	0	32094.71	35.83	31
TRP	CZ3	cr	cr	0.000	0	32094.71	35.83	31
TRP	CH2	cr	cr	0.000	0	32094.71	35.83	31

als hydrophil zu bezeichnen und wir weisen ihm den Potentialtyp *cp* zu.

Anhang D

Integration auf Kugeloberflächen

Die Berechnung der dem Lösungsmittel zugänglichen Proteinoberfläche, genauer gesagt die SAS-Oberfläche, ist derjenige Bestandteil einer Energieberechnung für eine Struktur, der den höchsten Rechenaufwand darstellt. Hier ist ein Kompromiß zwischen Rechenzeit und Genauigkeit zu finden.

Zur Berechnung der SAS-Oberfläche wird der Radius r_i jedes Atoms i um den eines fiktiven Wassermoleküls von $1.4\mathring{A}$ erhöht. Die Freie Oberfläche eines Atoms i ist dann diejenige Fläche, die nicht innerhalb eines Nachbaratoms $j \in N(i)$ liegt. Als Nachbaratome $j \in N(i)$ werden die Atome bezeichnet, deren Kugeloberfläche sich mit der des i-ten Atoms schneidet, d.h., wenn die Summe der Radien kleiner als der Abstand d_{ij} der Atome ist. Für die numerische Bestimmung dieser Oberfläche gibt es "exakte" und "approximative" Verfahren.

Eine Beispiel für ein approximatives Verfahren ist die LCPO-Methode (Linear Combination of Pairwise Overlaps [WSS99]). Bei dieser Methode wird das analytische Ergebnis

$$A_{ij} = 2\pi r_i \left(r_i - \frac{d_{ij}}{2} - \frac{r_i^2 - r_j^2}{2d_{ij}} \right)$$

welches den Teil der Oberfläche des Atoms i angibt, der von Atom j abgedeckt wird, benutzt, um die Oberfläche des Atoms i unter Anwesenheit aller Nachbaratome $N(i)$ als Funktion des Atompaareoberflächen näherungsweise zu bestimmen:

$$\begin{aligned} A_i^{LCPO} &= P_1 \cdot 4\pi r_i^2 + P_2 \cdot \sum_{j \in N(i)} A_{ij} + P_3 \cdot \sum_{\substack{j,k \in N(i) \\ j \neq k; k \in N(j)}} A_{jk} + \\ &\quad + P_4 \cdot \sum_{k \in N(i)} A_{ij} \left(\sum_{k \in N(i) \cap N(j)} A_{jk} \right) \end{aligned}$$

Die Parameter P_1 bis P_4 sind Atomtyp-spezifisch, wobei LCPO 18 verschiedene Atomtypen unterscheidet. Wir konnten im Gegensatz zu [WSS99] keine gute

Übereinstimmung der LCPO mit den exakten Oberflächen erzielen. Eine andere Methode, die gern in Kombination mit dem CHARMM Kraftfeld verwendet wird, lautet

$$\begin{aligned} A_i &= 4\pi r_i^2 \prod_{j \in N(i)} \left[1 - p_i p_{ij} b_{ij}/(4\pi r_i^2)\right] \\ \text{mit} \quad b_{ij} &= \pi\, r_i (r_i + r_j - d_{ij})(1 + (r_j - r_i) d_{ij}^{-1}) \end{aligned}$$

wobei $p_{ij} = 0.8875$ für kovalent gebundene Atome und ansonsten $p_{ij} = 0.3516$ gilt. p_i sind atomspezifische Parameter[FAC02]. Erste Ergebnisse zeigen auch hier keine gute Übereinstimmung mit den exakten Oberflächen. Ein Vorteil der "approximativen" Verfahren liegt darin, daß sie eine einfache Berechnung des Gradienten zulassen, wodurch sie insbesondere für Molekulardynamik Simulationen interessant sind.

Unter exakten Verfahren sind hier die Algorithmen zu verstehen, die eine Kenngröße – Punktzahl N oder Punkt-/Ebenenabstand Δz – haben, die im Limes $N \to \infty$, $\Delta z \to 0$ das analytisch exakte Ergebnis liefern. Darunter fällt z.B. das Verfahren von Lee und Richards[LR71], welches das Protein in Scheiben schneidet und auf den zweidimensionalen Ebenen die analytisch bestimmbaren Sehnenlängen, die im Kontakt mit dem Lösungsmittel sind, aufaddiert. Die Gesamtlänge wird dann mit der Schichtdicke Δz multipliziert, um die freie Oberfläche zu erhalten. Im Limes unendlich feiner Schnitte erhält man den exakten Wert der freien Oberfläche. In das CARB Simulationspaket ist dieser Algorithmus integriert, wobei der Schnittebenenabstand auf $0.39\mathring{A}$ festgelegt wurde.

Eine andere Möglichkeit der Oberflächenintegration besteht darin, Punkte auf der Kugeloberfläche zu verteilen und die Anzahl der Punkte, die nicht innerhalb anderer Nachbarn liegen zu zählen. Wenn allen Punkten ein gleichgroßes Flächenelement zugeordnet ist, so verhält sich die freien Oberfläche zur Gesamtoberfläche wie die frei liegenden Punkte zur Gesamtpunktzahl. Für $N \to \infty$ erhält man das exakte Ergebnis. Bei fester Gesamtpunktzahl N ist die Genauigkeit des Verfahrens von der Verteilung der Punkte auf der Oberfläche abhängig. Eine mögliche Punktverteilung mit gleichen Flächenelementen entsteht, wenn man die Kugel in Streifen gleicher Dicke schneidet ($\Delta z = const.$) und die verbleibenden Scheiben in gleichgroße Stücke unterteilt:

$$x = \cos\phi \, \sin\theta \; , \; y = \sin\phi \, \sin\theta \; , \; z = \cos\theta$$

$$\begin{aligned} \text{mit} \qquad \cos\theta &= 1 - \frac{j + 1/2}{M + 1} \qquad ; \; j = 1, 2, \ldots, M \\ \phi &= i \cdot \frac{2\pi}{N/M} \qquad ; \; i = 1, 2, \ldots, N/M \end{aligned}$$

Wenn man die Scheiben an den Polen nochmals halbiert und in diesen Scheiben die Anzahl der ϕ Werte halbiert, um die Gesamtpunktzahl konstant zu halten,

wird die Genauigkeit etwas erhöht. Nach dieser Vorschrift werden Punktmengen mit $N > 1000$ generiert.

Eine weitere Methode, Punkte möglichst gleichmäßig zu verteilen, ist es, die Punkte als Elektronen auf einer Kugeloberfläche zu interpretieren und die Verteilung mit der niedrigsten elektrostatischen Energie zu bestimmen. Dies ist das sogenannte *Thompson*-Problem, welches für Punktzahlen $N > 100$ ein numerisch sehr rechenintensive Aufgabe darstellt, die nur äußerst langsam konvergiert.

Die Oberflächenberechnung der Tripeptide von Sharke und Rupley [SR73] ist mit 92 gleichverteilten Punkten durchgeführt worden. Hier sei angemerkt, daß es geometrisch möglich ist, 92 Punkte gleichmäßig auf einer Kugeloberfläche zu verteilen, wohingegen dies etwa für 93 Punkte nur näherungsweise möglich ist. Die Zahlen, für die eine völlig gleichmäßige Verteilung möglich ist, werden *magic numbers* genannt. Einige Simulationsprogramme arbeiten mit mehreren Hundert Punkten auf der Kugeloberfläche. In unserem Simulationsprogramm verwenden wir 3072 Punkte für die Integration auf der Kugeloberfläche, da eine kleinere Punktmenge nicht die gewünschte Genauigkeit besitzt. Hierbei handelt es sich nicht um den Versuch, die freie Proteinoberfläche möglichst genau auszurechnen, sondern diese große Punktemenge ist notwendig, um der Isotropie des Raumes zu genügen.

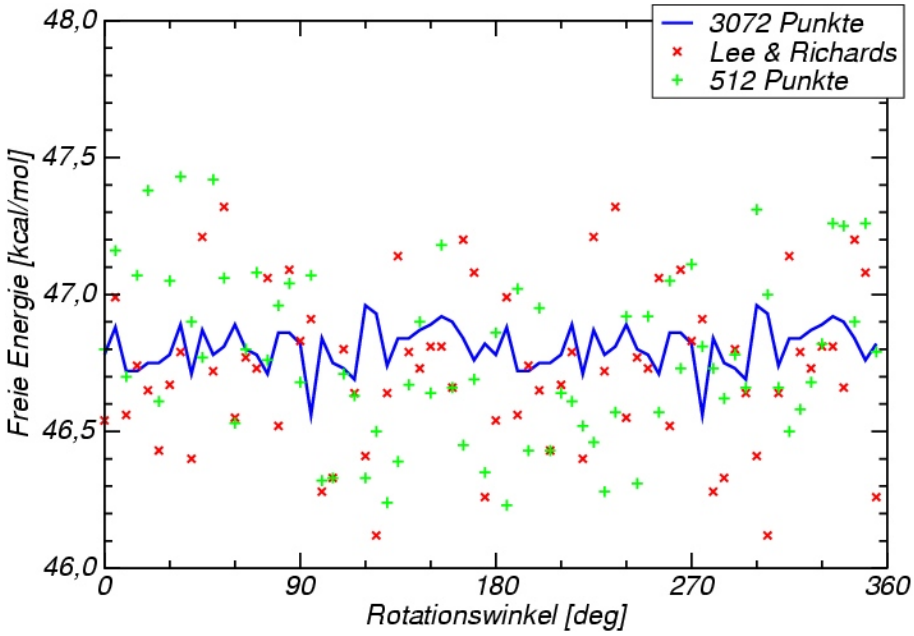


Abbildung D.1: Vergleich verschiedener Integrationsmethoden

Dies soll am Beispiel der nativen Struktur von 1VII verdeutlicht werden. Dreht man die Proteinstruktur um die x-Achse so ändert sich die (exakte) freie Ober-

fläche des Proteins nicht. Entsprechend ist auch der Beitrag des Lösungsmittels zur Freie Energie konstant. In Abb. D.1 sind die Lösungsmittelenergien der nativen Struktur von 1VII jeweils nach einer Rotation um 5° um die x-Achse für verschiedene Integrationsmethoden aufgetragen. Sowohl die Integrationsroutine des CARB Kraftfeldpaketes als auch die Integration mit $N = 512$ nahezu gleichverteilten Punkten liefern je nach Orientierung der Struktur Energien, die um bis zu $1.2\,kcal\,mol^{-1}$ voneinander abweichen. Bei $N = 92$ Punkten liegt der maximale Unterschied bei $3.5\,kcal\,mol^{-1}$, und ein Drittel der Energien liegen nicht innerhalb des in Abb. D.1 dargestellten Bereiches von $46 - 48\,kcal\,mol^{-1}$.

Für die Proteinstrukturvorhersage muß das Kraftfeld in der Lage sein, einen Energieunterschied der nativen Struktur zu einer mißgefalteten von wenigen $kcal\,mol^{-1}$ zu identifizieren. Die Genauigkeit der Integration sollte daher so gewählt sein, daß der Beitrag zur Freien Energie, der aus der Proteinoberfläche stammt, nicht mehr als etwa ein halbes $kcal\,mol^{-1}$ schwankt. Wir haben uns daher für die Punktintegration mit $N = 3072$ (teilweise sogar $N = 5000$) entschieden.

Literaturverzeichnis

[AJ95] F. Avbelj and J.Moult. Determination of the conformation of folding initiation sites by computer simulations. *Proteins: Structure, Function and Genetics*, 23:129–141, 1995.

[AM95] F. Avbelj and J. Moult. Role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry*, 34:755–764, 1995.

[Anf73] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[AS75] C. B. Anfinsen and H. A. Scheraga. Experimental and theoretical aspects of protein folding. *Advances in Protein Chemistry*, 29:205–300, 1975.

[AT94] R. Abagyan and M. Totrov. Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. *Journal of Molecular Biology*, 235:983–1002, 1994.

[Avb92] F. Avbelj. Use of a potential of mean force to analyze free energy contributions in protein folding. *Biochemistry*, 31:6290–6297, 1992.

[AYL89] N. L. Allinger, Y. H. Yuh, and J.-H. Lii. The MM3 force field for hydrocarbons. *Journal of the American Chemical Society*, 111:8551–8566, 1989.

[Bal94] R. L. Baldwin. Matching speed and stability. *Nature*, 369:183–184, 1994.

[BBO+83] 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. Comput. Chem.*, 4:187–217, 1983.

[BRTB02] R. Bonneau, I. Ruczinski, J. Tsai, and D. Baker. Contact order and ab initio protein structure prediction. *Protein Science*, 11:1937–1944, 2002.

[BSJ+01] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA*, 98:10037–10041, 2001.

[BWF+00] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourn. The protein data bank. *Nucleic Acids Research*, 28:245–242, 2000.

[CES91] R. W. Carrol, O. Evans, and P. E. Stein. Mobile reactive centre of serpins and the control of thrombosis. *Nature*, 353:576–578, 1991.

[CES93] R. W. Carrol, O. Evans, and P. E. Stein. Corrections: Mobile reactive centre of serpins and the control of thrombosis. *Nature*, 364:737, 1993.

[Chu67] K. L. Chung. *Markov Chains with Stationary Transition Probabilities, second ed.* Springer, New York, 1967.

[CK95] A. Calflisch and M. Karplus. *Journal of Molecular Biology*, 252:672–708, 1995.

[CS92] G. Casari and M. J. Sippl. Structure-derived hydrophobic potential. *Journal of Molecular Biology*, 224:725–732, 1992.

[Dau98] M. Daune. *Molecular Biophysics. Structures in motion.* Oxford University Press, 1998.

[Dil90] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.

[DK98] Y. Duan and P. A. Kollman. Pathway to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740–744, 1998.

[DLI+96] V. Daggett, A. Li, L. S. Itzhaki, D. E. Otzen, and A. R. Fersht. Structure of the transition state for folding of a protein derived from experiment and simulation. *Journal of Molecular Biology*, 257:430–440, 96.

[DLK99] A. R. Dinner, T. Lazaridis, and M. Karplus. Understanding β-hairpin formation. *Proc. Natl. Acad. Sci. USA*, 96:9068–9073, 1999.

[DORO+88] P. Dauber-Osguthorpe, V. A. Roberts, D. J. Osguthorpe, J. Wolff, M. Genest, and A. T. Hagler. Structure and energetics of ligand binding to proteins: E. coli dihydrofolate reductase- trimethoprim, a drug-receptor system. *Proteins: Structure, Function and Genetics*, 4:31–47, 1988.

[DWK98] Y. Duan, L. Wang, and P.A. Kollman. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. *Proc. Natl. Acad. Sci. USA*, 95:9897–9902, 1998.

[EJW+95] D. Eliezer, P. A. Jennings, P. E. Wright, S. Doniach, K. O. Hodgson, and H. Tsuruta. *Sience*, 270:487–488, 1995.

[EM86] D. Eisenberg and A.D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986.

[ESF] *Discover 2.9.7 / 95.0 / 3.0.0 User Guide, Oktober 1995. San Diego: Biosym/MSI, 1995.*

[FAC02] P. Ferrara, J. Apostolakis, and A. Caflisch. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins: Structure, Function and Genetics*, 46:24–33, 2002.

[FD96] D. L. Freeman and J. D. Doll. *Annu. Rev. Phys. Chem.*, 47:42–80, 1996.

[FP83] J. L. Fauchere and V. Pliska. *Eur. J. med. Chem.-Chim. ther.*, 18:369–375, 1983.

[Fri75] H. L. Friedman. Image approximation to the reaction field. *Molecular Physics*, 29(5):1533–1543, 1975.

[Fri02] R. A. Friesner, editor. *Computational Methods for Protein Folding: A Special Volume of Advances in Chemical Physics*, volume 120. John Wiley & Sons, Inc., 2002.

[GRSDF94] G. de Part Gay, J. Ruiz-Sanz, B. Davis, and A. R. Fersht. The structure of the transition state for the association of two fragments of the barley chymotrypsin inhibitor-2 to generate native-like protein: implications for the mechanisms of protein folding. *Proc. Natl. Acad. Sci. USA*, 91:10943, 1994.

[Han] U. H. E. Hansmann. private Kommunikation.

[HE94] A. T. Hagler and C. S. Ewig. On the use of quantum energy surfaces in the derivation of molecular force fields. *Comp. Phys. Comm.*, 84:131–155, 1994.

[Hon99] B. Honig. Protein folding: From the levinthal paradox to structure prediction. *Journal of Molecular Biology*, 293:283–293, 1999.

[HW99] K. Hamacher and W. Wenzel. Scaling behavior of stochastic minimization algorithms in a perfect funnel landscape. *Phys. Rev. E*, 59(1):938–941, 1999.

[HY95] B. Honig and A. Yang. Free energy balance in protein folding. *Advances in Protein Chemistry*, 46:27–57, 1995.

[IKP88] C. L. Brooks III., M. Karplus, and B. M. Pettitt. Proteins: a theoretical perpective of dynamics, structure and thermodynamics. *Advances in Chemical Physics*, 71:1–259, 1988.

[IKW02] S. A. Islam, M. Karplus, and D. L. Weaver. Application of the diffusion-collision model to the folding of three-helix bundle proteins. *Journal of Molecular Biology*, 318:199–215, 2002.

[INS+03] J. Ireta, J. Neugebauer, M. Scheffler, A. Rojo, and M. Galván. Density functional theory study of the cooperativity of hydrogen bonds in finite and infinite α-helices. *J. Phys. Chem. B*, 107:1432–1437, 2003.

[JeF93] S. E. Jackson, N. elMasry, and A. R. Fersht. Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: a critical test of the protein engineering method of analysis. *Biochemistry*, 32(42):11270–11278, 1993.

[Jor81] W. L. Jorgeson. Quantum and statistical mechanical studies of liquids. 11. transferable intermolecular potential functions. application to liquid methanol including internal rotation. *Journal of the American Chemical Society*, 103:341, 1981.

[Kau59] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry*, 14:1–63, 1959.

[KGV83] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[KIP87] M. Karplus, T. Ichiye, and B.M. Pettitt. Configurational entropy of native proteins. *Biophys. J.*, 52:1083–1085, 1987.

[KM01] P. S. Kushwaha and P. C. Mishra. Stability of the normal, zwitterionic neutral and anionic forms od spatic scid in gas phase and aqueous media. *Journal of Molecular Structure (Theochem)*, 549:229–242, 2001.

[Kra91] P. J. Kraulis. MOLSCRIPT; a program to produce both detailed and schematic plots of protien structures. *J. Appl. Cryst.*, 24:946–950, 1991.

[KS83] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[kse98] B. Øksendal. *Stochastic Differential Equations: an introduction with applications. 5th edition*. Berlin: Springer, 1998.

[LB00] M. C. Lawrence and P. Bourke. A program for generating electron density isosurfaces. *J. Appl. Cryst.*, 33:990–991, 2000.

[Lev68] Cyrus Levinthal. Are there pathways for protein folding? *J. Chim. Phys.*, 65:44–45, 1968.

[LHH03] C. Lin, C. Hu, and U. H. E. Hansmann. Parallel tempering simulations of HP-36. *Proteins: Structure, Function and Genetics*, 52:436–445, 2003.

[LR71] B. Lee and F. M. Richards. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, 55:379–400, 1971.

[MBHC95] A. G. Murzin, S. E. Brenner, T. J. P. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[MG97] M. Miyahara and K. E. Gubbins. Freezing/melting phenomena for lennard-jones methane in slit pores: A monte carlo study. *J. Chem Phys.*, 106(7):2865–2880, 1997.

[MMBS75] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga. Energy parameters in polypeptides. vii. geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *J. Phys. Chem.*, 79:2361, 1975.

[MP95] G. I. Makhatadze and P. L. Privalov. *Advances in Protein Chemistry*, 47:308–417, 1995.

[MT94] I. K. McDonald and J. M. Thornton. Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology*, 238:777–793, 1994.

[MvG94] A. E. Mark and W. F. van Gunsteren. Decomposition of the free energy of a system in terms of specific interactions. *Journal of Molecular Biology*, 240:167–176, 1994.

[NH91] A. Nicholls and B. Honig. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the poisson-boltzmann equation. *J. Comp. Chem.*, 12(4):435–445, 1991.

[OCB02] A. Onufriev, D. A. Case, and D. Bashford. Effective born radii in the generalized born approximation: The importance of being perfect. *J. Comp. Chem.*, 23:1297–1304, 2002.

[P+02] F. M. Pearl et al. The CATH extended protein-family database: providing structural annotations for genome sequences. *Protein Science*, 11:233–244, 2002.

[PCC+95] D. A Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Commun.*, 91:1–41, 1995.

[PM97a] J. T. Pedersen and J. Moult. Ab initio protein folding simulations with genetic algorithms: Simulations on the complete sequence of small proteins. *Proteins: Structure, Function and Genetics*, 1:179–184, 1997.

[PM97b] J. T. Pedersen and J. Moult. Protein folding simulations with genetic algorithms and a detailed molecular description. *Journal of Molecular Biology*, 269:240–259, 1997.

[PO98] E. Pitard and H. Orland. The role of the energy gap in protein folding dynamics. *cond-mat/9811252*, 1998.

[PS93] S. D. Pickett and M. J. E. Sternberg. Empirical scale of side-chain conformational entropy in protein folding. *Journal of Molecular Biology*, 231:825–839, 1993.

[R+85] Rose et al. *Science*, 229:834–838, 1985.

[RP03] P. Ren and J. W. Ponder. *J. Phys. Chem. B*, 107:5933–5947, 2003.

[SBO94] N. D. Socci, W. Bialek, and J.N. Onuchic. Properties and origin of protein secondary structure. *Physical Review E*, 49(4):3440–3443, 1994.

[SFSD96] L. J. Smith, K. M. Fiebig, H. Schwalbe, and C. M. Dobson. The concept of the random coil - residual structure in peptides and denatured proteins. *Fold. Des.*, 1(5):95–106, 1996.

[Sho96] D. Shortle. The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J.*, 10(1):27–34, 1996.

[SHT+99] W. R. P. Scott, P. H. Hunenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Kruger, and W. F. van Gunsteren. The gromos biomolecular simulation program package. *J. Phys. Chem.*, 103:3596–3607, 1999.

[SHW03] A. Schug, T.-A. Herges, and W. Wenzel. Reproducible protein folding with the stochastic tunneling method. *Physical Review Letters*, 91:in press, 2003.

[SM98] J. Schneider and I. Morgenstern. Bouncing toward the optimum: Improving the results of monte carlo optimization algorithms. *Phys. Rev. E*, 58(4):5085–5095, 1998.

[Smi94] D. A. Smith. *Modeling the Hydrogen Bond*. ACS Symposium Series 569, 1994.

[SNFH91] K. A. Sharp, A. Nicholls, R. Friedman, and B. Honig. Extracting hydrophobic free energies fram experimental data: Relationship to protein folding and theoretical models. *Biochemistry*, 30:9686–9697, 1991.

[SR73] A. Sharke and J. A. Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of Molecular Biology*, 79:351–371, 1973.

[SSH94] D. Sirkoff, K. A. Sharp, and B. Honig. Accurate calculation of hydration free energy using macroscopic solvent models. *J. Phys. Chem.*, 98:1978–1988, 1994.

[SSK94a] A. Sali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369:248–251, 1994.

[SSK94b] A. Sali, E. Shakhnovich, and M. Karplus. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. *Journal of Molecular Biology*, 235:1614–1636, 1994.

[SSR02] C. Simmerling, B. Sockbine, and A. E. Roitberg. All-atom structure prediction and folding simulations of a stable protein. *Journal of the American Chemical Society*, 124:11258–11259, 2002.

[SSRD91] L. J. Smith, M. J. Sutcliffe, C. Redfield, and C. M. Dobson. Analysis of ϕ and χ_1 torsion angles for hen lysozyme in solution from ^1H NMR spin-spin cpouling constants. *Biochemistry*, 30, 1991.

[STHH91] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, 112:6127–6129, 1991.

[SZP02] C. D. Snow, B. Zagrovic, and V.A. Pande. The trp cage: Folding kinetics and unfolded state topology via molecular dynamics simulations. *Journal of the American Chemical Society*, 124:14548–14549, 2002.

[Tan70] C. Tanford. Protein denaturation. *Advances in Protein Chemistry*, 24:1–95, 1970.

[UB88] J. B. Udgaonkar and R. L. Baldwin. NMR evidence for an early framework intermediate on the folding pathway of ribonuclease a. *Nature*, 335:694–699, 1988.

[UM96] R. Unger and J. Moult. Local interactions dominate folding in a simple protein model. *Journal of Molecular Biology*, 259:988–994, 1996.

[VWA97] G. Vogt, S. Woell, and P. Argos. Protein thermal stability, hydrogen bonds and ion pairs. *Journal of Molecular Biology*, 269:631–643, 1997.

[Wal96] D.J. Wales. Structure, dynamics, and thermodynamics of clusters: Tales from topographic potential surfaces. *Sience*, 271:925–929, 1996.

[WD97] D. J. Wales and J. P. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *J. Phys. Chem.*, 101:5111–5116, 1997.

[WH99] W. Wenzel and K. Hamacher. A stochastic tunnelling approch for global minimization. *Physical Review Letters*, 82:3003, 1999.

[WSS99] J. Weiser, P. S. Shenkein, and W. C. Still. Approximate atomic surfaces from liner combinations of pairwise overlaps (lcpo). *J. Comp. Chem.*, 20(2):217–230, 1999.

[YH95] A. Yang and B. Honig. Free energy determinats of secondary structure formation: I. α-helices. *Journal of Molecular Biology*, 252:351–365, 1995.

[ZK97] Y. Zhou and M. Karplus. Folding thermodynamics of a model three-helix-bundle protein. *Proc. Natl. Acad. Sci. USA*, 94:14429–14432, 1997.

[ZVK99] Y. Zhou, D. Vitkup, and M. Karplus. Native proteins are surface-molten solids: Application of the lindemann criterion for the solid versus liquid state. *Journal of Molecular Biology*, 285:1371–1375, 1999.

An dieser Stelle möchte ich allen danken, die an der Entstehung und Durchführung dieser Arbeit teilgenommen haben.

Insbesondere danke ich Herrn Priv.-Doz. Dr. Wolfgang Wenzel für die Themenstellung und für die vielen Diskussionen, denen wichtige Impulse für meine Arbeit entsprungen sind. Für die Bereitstellung des Programmpaketes, welches das ursprüngliche CARB-Kraftfeld enthält, danke ich Prof. Dr. John Moult, sowie Dr. Susan Gregurick für die Einführung in dessen Benutzung und den Erläuterungen der umfangreichen Programmarchitektur.

Weiterhin bin ich den Mitgliedern des Lehrstuhles "Theoretische Physik I" der Universität Dortmund und den Mitarbeitern des Instituts für Nanotechnologie des Forschungszentrums Karlsruhe für die kollegiale Atmosphäre und Hilfsbereitschaft zu Dank verpflichtet.

Holger Merlitz, Alexander Schug und Philipp Stampfuß haben zudem die mühevolle Aufgabe übernommen, Teile dieser Arbeit korrekturzulesen.