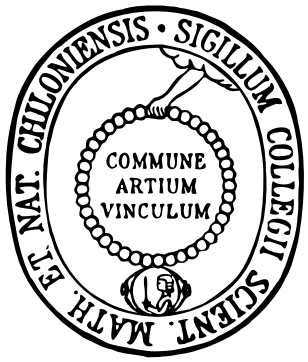
Entwicklung und Optimierung einer neuartigen Potentialfunktion mit Anwendung in der globalen Geometrie-Optimierung zur Vorhersage der Proteinfaltung

Dissertation

zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel



vorgelegt von

Florian Koskowski

Kiel 2009

Entwicklung und Optimierung einer neuartigen Potentialfunktion mit Anwendung in der globalen Geometrie-Optimierung zur Vorhersage der Proteinfaltung

Dissertation

zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Florian Koskowski

Kiel 2009

Referent:	Prof. Dr. Bernd Hartke
Koreferent:	Prof. Dr. Rainer Herges
Tag der mündlichen Prüfung:	04.11.2009
Zum Druck genehmigt:	04.11.2009

Der Dekan:

Abstract

Proteins in biological systems play a major role for metabolic processes, for example as catalysts in chemical reactions or as transportation units in membranes. Their function is determined by their three-dimensional structure in the native state, which itself is encoded in the amino acid sequence. Understanding this connection is a very important scientific goal as it is the key for the prediction of the protein's biological activity. There are many different approaches to solve this problem. Methods relying on the use of broad statistical database information perform the best, but are limited to already known structure-sequence-patterns and do not explain the important basic mechanisms and complex interactions in the folding process. To understand this, methods describing the interactions between single amino acids are needed, but those are working well just for few protein classes and are not able to predict the native state structure for all proteins. This drawback was the starting point for this work, which aimed at developing a program for generating native state candidates without the use of database information, taking as information just the amino acid sequence. A novel coarse-grained and smooth force field was established to describe the internal interactions in a protein by combining statistical potentials with optimized linear combinations of different basis functions. The preliminary work for setting up the protein model and the definition of the potential functions included statistical analysis of known database structures, development of side chain descriptors and the generation of new decoy proteins. The force field parameters were fitted to the resulting experimental data whereas the weighting factors of the potential functions were determined through a linear programming procedure which made use of direct comparison of the decoys with native structures. The linear programming problem was solved by an interior-point method.

The resulting potential was used in a ranking test against different other potentials from the literature and showed here an average performance of 60 % correctly identified proteins. The force field was also implemented in a newly written global optimization program utilizing a genetic algorithm for the energy surface exploration and for the identification of possible native state geometries. The results showed several problems which originated from the high dimensionality of the energy surfaces in connection with the force field parameterization scheme, leading to a weak energy-structure-correlation and to failed identifications of the native state.

To improve the parameterization of the force field systematically, the global optimization program and the linear programming protocol were merged into an iterative cyclic procedure, but the results did not show a distinct progress in the prediction ability of the potential.

In summary, the encountered difficulties of the developed program could be traced back to problems constructing near-native structures with the genetic algorithm and to the limited number of decoy structures in the parameter optimization process resulting in non-optimal force field parameters. Especially these points should play a central role in future developments of the program.

Kurzzusammenfassung

Proteine spielen in biologischen Systemen für viele Stoffwechselvorgänge eine zentrale Rolle, beispielsweise als Katalysatoren in chemischen Reaktionen oder als Transporteinheiten durch Membranen. Für ihre Funktion ist ihre dreidimensionale Struktur im nativen Zustand entscheidend, welche in der Sequenz der Aminosäuren codiert ist. Die Kenntnis über diesen Zusammenhang ist der Schlüssel zur Vorhersage der biologischen Aktivität eines Proteins, wodurch dieser Fragestellung eine große Bedeutung zukommt. Zur Beantwortung existieren bereits viele verschiedene Ansätze, wobei Methoden, die auf dem breiten Einsatz statistischer Daten bekannter Proteine beruhen, heutzutage die größten Erfolge erzielen. Diese funktionieren jedoch lediglich bei bereits bekannten Sequenz-Geometrie-Mustern und beantworten zudem nicht die sehr wichtigen Fragen nach den zugrundeliegenden Mechanismen und komplexen Wechselwirkungen der Strukturbildung bzw. Faltung. Hierfür werden Verfahren benötigt, die die Wechselwirkungen zwischen den einzelnen Aminosäuren erfassen, wobei diese Algorithmen heutzutage lediglich für einige ausgewählte Proteine in der Lage sind, den nativen Zustand vorherzusagen. An dieser Stelle setzt auch diese Arbeit ein, ein Programm zu entwickeln, das auf Basis der alleinigen Kenntnis der Sequenz ohne Abgleich mit Datenbankstrukturen Vorschläge für den nativen Zustand generiert. Hierzu wurde ein neuartiges vergröbertes, glattes Kraftfeld der internen Proteinwechselwirkungen unter Verwendung verschiedener etablierter Techniken erstellt, wobei statistische Potentiale mit optimierten Linearkombinationen unterschiedlicher Basisfunktionen kombiniert wurden. Als Grundlage hierzu wurden zur Erstellung des Proteinmodells und für die Definition des Potentials bekannte Datenbankstrukturen statistisch analysiert, Seitenkettenmodelle entwickelt und neue falsche Proteinstrukturen generiert. Die Parameter des Kraftfeldes wurden an die erhaltenen experimentellen Daten angepasst und die Gewichtungskoeffizienten der Potentialfunktionen über ein lineares Optimierungsproblem bestimmt, welches Informationen aus einem direkten Vergleich zwischen falschen und nativen Strukturen verwendet. Die Lösung dieses linearen Problems wurde mittels eines Innere-Punkte-Algorithmus' berechnet.

Das resultierende Potential wurde zum einen gegen verschiedene Literaturpotentiale über einen Erkennungstest verglichen, wobei eine Identifikationsleistung von ca. 60 % erreicht wurde, was im oberen Mittelfeld anzuordnen ist. Zum anderen wurde es in einen neu implementierten globalen Geometrieoptimierungsalgorithmus integriert, das einen genetischen Algorithmus verwendet, um die Energiehyperflächen zu erkunden und Vorschläge für die Geometrie des nativen Zustandes zu entwickeln. Hierbei zeigten sich Probleme, die durch die hohe Dimensionalität der Energieflächen in Zusammenhang mit der Kraftfeld-Parametrisierung bedingt waren, wodurch nur eine geringe Struktur-Energie-Korrelation erreicht und die Bestimmung des nativen Zustandes mit dieser Methode verhindert wurde.

Um die Parametrisierung des Kraftfeldes systematisch zu verbessern, wurde die globale Geometrieoptimierung mit dem linearen Problem zur Parameter-Optimierung zu einem iterativen zyklischen Algorithmus zusammengefasst. Die Iterationen zeigten jedoch keinen eindeutigen Trend für die Entwicklung der Vorhersageleistung des Potentials.

Insgesamt konnten die Schwierigkeiten des in dieser Arbeit entwickelten Programmes vor allem auf Probleme in der Erzeugung nah-nativer Strukturen im genetischen Algorithmus und in der beschränkten Anzahl an falschen Proteinen in der Parameter-Optimierung zurückgeführt werden, wodurch keine optimalen Kraftfeldparameter erhalten werden konnten, weshalb in zukünftigen Entwicklungen diese Punkte besonders im Mittelpunkt stehen sollten.

Inhaltsverzeichnis

1 Begriffe und Abkürzungen

Übersicht über Begriffe, Abkürzungen und Symbole

Backbone	Proteinrückgrat
$\mathbf{b}_i$	Bindungsvektor zwischen zwei C^α-Atomen, $\mathbf{b}_i = \mathbf{x}_i - \mathbf{x}_{i-1}$
C_i^α	Das α-ständige Kohlenstoff der i-ten Aminosäure
C_i^β	Das β-ständige Kohlenstoff der i-ten Aminosäure
DNA / DNS	*Deoxyribonucleic acid* (engl.) / Desoxyribonukleinsäure (deutsch)
GA	Genetischer Algorithmus
κ	Der Bindungswinkel zwischen drei sequentiellen C^α-Atom
Lit.	Literaturreferenzen in Übersichtsdarstellungen
N	Anzahl der Aminosäuren in einem Protein
Random Coil	Zufallsgeometrie der Peptidkette, statistisches Knäuel
SAS	*Solvent Accessible Surface* - Lösungsmittelzugängliche Oberfläche
$\mathbf{q}_i$	Vektor vom C_i^α-Atom zu dessen zugehörigen Seitenketten-Pseudoatom
$\boldsymbol{\rho}_i$	Ortsvektor des Seitenketten-Pseudoatoms. $\boldsymbol{\rho}_i = \mathbf{x}_i + \mathbf{q}_i$
$\mathbf{S}$	Sequenzvektor des Proteins, mit den Elementen (Aminosäuren) s_i, so dass $\mathbf{S} = (s_1, s_2, ..., s_N)$
s_i	Element des Sequenzvektors $\mathbf{S}$. Entspricht einer der zwanzig Aminosäuren: $s_i \in \{1, 2, ..., 20\}$ (siehe Tab. 1.1).
σ	Die Standardabweichung einer Verteilung
τ	Der Torsionswinkel, der durch vier sequentielle C^α-Atome gegeben ist
$\mathbf{X}$	Gesamtgeometrie eines Proteins
$\mathbf{x}_i$	Ortsvektor des i-ten C^α-Atoms bzw. von C_i^α
$\mathbf{z}_i$	Ortsvektor zum Mittelpunkt (Zentrum) der Strecke zwischen zwei sequentiellen C^α-Atomen: $\mathbf{z}_i = \frac{1}{2}(\mathbf{x}_i + \mathbf{x}_{i+1})$.
$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$	Inneres Produkt zweier m-dimensionaler Vektoren $\mathbf{v}_1$ und $\mathbf{v}_2$. $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = \sum_{i=1}^{m} v_{1,i} \cdot v_{2,i}$
$\mathbf{v} \times \mathbf{v}$	Äußeres Produkt (Kreuzprodukt) zweier Vektoren $\mathbf{v}_1$ und $\mathbf{v}_2$ des $\mathbb{R}^3$. $\mathbf{v} \times \mathbf{v} = \begin{pmatrix} v_{1,2}v_{2,3} - v_{1,3}v_{2,2} \\ v_{1,3}v_{2,1} - v_{1,2}v_{2,3} \\ v_{1,1}v_{2,2} - v_{1,1}v_{2,1} \end{pmatrix}$

Zur expliziten Kennzeichnung eines nativen Proteins im Gegensatz zu einem anderen wird in dieser Arbeit ein Stern verwendet, wie z. B. $\mathbf{S}^*$ (Sequenzvektor eines nativen Proteins) oder $\mathbf{x}_i^*$ (Position des i-ten C^α-Atoms in einem nativen Protein).

Kanonische Aminosäuren

Nr.	Name	3er-Code[1]	1er-Code[2]
1	Alanin	Ala	A
2	Cystein	Cys	C
3	Asparaginsäure	Asp	D
4	Glutaminsäure	Glu	E
5	Phenylalanin	Phe	F
6	Glycin	Gly	G
7	Histidin	His	H
8	Isoleucin	Ile	I
9	Lysin	Lys	K
10	Leucin	Leu	L
11	Methionin	Met	M
12	Asparagin	Asn	N
13	Prolin	Pro	P
14	Glutamin	Gln	Q
15	Arginin	Arg	R
16	Serin	Ser	S
17	Threonin	Thr	T
18	Valin	Val	V
19	Tryptophan	Trp	W
20	Tyrosin	Tyr	Y

Tabelle 1.1: Die natürlichen Aminosäuren, sortiert in Reihenfolge der Ein-Buchstaben-Abkürzung.
[1]Drei-Buchstaben-Abkürzung, [2]Ein-Buchstaben-Abkürzung

2 Einleitung - Warum Proteine?

Lebende Organismen, in ihrer gängigen Definition, müssen einen beständigen Stoffwechsel aufrechterhalten, um die Kontinuität ihrer Funktionsweise zu gewährleisten ("Leben"). Dieser Stoffwechsel beinhaltet die Aufnahme, Umsetzung und Abgabe von chemischen Verbindungen und Energie in unterschiedlichen Formen mit der Umgebung. Die Entwicklung der hierzu nötigen Stoffkreisläufe in den Organismen basiert auf Optimierungsprozessen, die an menschlichen Maßstäben gemessen über lange Zeiträume hinweg abliefen, welche gemeinhin als Evolution bezeichnet werden. Hierbei erfolgt eine beständige Selektion auf Grundlage der Anpassungsfähigkeit des Individuums an seine Umgebung.
Die Grundlage für den biologischen Erfolg eines Individuums bildet ein funktionierender und effizienter Metabolismus. Dieser ist auf der molekularen Ebene ein komplexes Zusammenspiel einer großen Anzahl unterschiedlichster Substanzklassen. Da viele der an den Stoffwechselvorgängen beteiligten Verbindungen häufig speziellen Anforderungen bezüglich ihrer Funktionalität genügen müssen und dies über einen komplexeren Aufbau realisiert wird, müssen diese im Organismus selbst synthetisiert werden und können häufig nicht aus der Umgebung aufgenommen werden. Die *In-vivo*-Synthesen erfordern einen nicht geringen Anteil der dem Organismus zur Verfügung stehenden Energie, wodurch im Hinblick auf die Evolution zu erwarten ist, dass dieses Merkmal ebenfalls einer Optimierung unterlag. Betrachtet man die unterschiedlichen Verbindungsklassen der Naturstoffe wie beispielsweise Lipide, Kohlenhydrate, Proteine oder Nucleotide sowie deren Bausteine bzw. Monomere, die häufig in biologischen Systemen synthetisiert werden, und vergleicht die benötigten biochemischen Energien zur Durchführung der an der Synthese beteiligten Prozesse, so zeigt sich, dass die Proteinbiosynthese häufig weit energieintensiver als die anderen Biosynthesen ist. Als Beispiel hierzu sei das Bakterium *E.Coli* aufgeführt, welches ca. 90 % der gesamten zur Verfügung stehenden Biosyntheseenergie für die Proteinbiosynthese aufwendet [1]. Dies umfasst unter anderem sowohl den Transport der einzelnen Aminosäuren zum Syntheseort und dort die Bildung der notwendigen chemischen Bindungen zur Erzeugung des vollständigen Proteins wie auch die Aufrechterhaltung vieler notwendiger Kontrollfunktionen, die die richtige Faltung und die Funktionsfähigkeit der Proteine vom Synthese- bis zum Zielort überprüfen, um Schädigungen des Organismus durch mutierte Sequenzen oder falsche Proteinstrukturen zu verhindern.

Hiermit in Verbindung steht wiederum eine Vielzahl an Funktionseinheiten im Organismus, die fehlerhafte Proteine reparieren oder wieder in die Monomerbestandteile abbauen [2]. Diese strengen Kontrollmechanismen haben zur Folge, dass ca. 30 % aller neu synthetisierten Proteine sofort wieder abgebaut werden, ohne dass diese den Syntheseapparat in der Zelle überhaupt verlassen [3]. Aber selbst die Proteine, die ihren Zielort erreichen, besitzen insgesamt nur eine Halbwertszeit von ca. 2 bis 20 Minuten [4], wodurch im Durchschnitt pro Stunde ein bis zwei Prozent der gesamten Proteinmenge wieder abgebaut wird [5]. Aus diesen Gründen müssen im Organismus Proteine beständig neu synthetisiert werden.
Diese beiden Gesichtspunkte, der große Energiebedarf der Proteinbiosynthese und die relativ kurzen Lebensdauern der synthetisierten Proteine, erscheinen im Licht einer "langen" Evolution und Optimierung der biochemischen Prozesse besonders widersprüchlich, wenn hierzu weiterhin beachtet wird, dass beispielsweise Enzyme zumeist nur ein einziges reaktives Zentrum besitzen, welches im Verhältnis zur Gesamtgröße des Proteins bzw. im Vergleich zum Rest des Proteins, der an der Reaktion nahezu unbeteiligt bleibt, meist sehr klein ist. Diese vermeintlichen Gegensätze lösen sich aber unter der Betrachtung der vielfältigen Anwendungen und Reaktionen, die dem Organismus durch den Einsatz von speziell konstruierten Proteinen ermöglicht werden. Dies reicht von strukturgebenden Proteinen, die zum makroskopischen Aufbau eines Organismus beitragen können, über die Katalyse spezifischer Reaktionen zur Erzeugung anderer Naturstoffe, die dem Organismus ohne Einbeziehung des Proteinkatalysators nur schwer zugänglich wären, bis hin zu Schutzfunktionen im Organismus, die schädigende Fremdstoffe neutralisieren und defekte Bereiche des Organismus' reparieren. Hierzu wird in Tabelle 2.1 ein allgemeiner beispielhafter Überblick darüber gegeben, in welchen Bereichen im Organismus Proteine verwendet werden [1]. Aus dieser kurzen Tabelle ist bereits ersichtlich, dass Proteine an sehr unterschiedlichen Prozessen teilnehmen und in vielen wichtigen Bereichen vorkommen.
Insbesondere die Effektivität der Proteine als biologische Katalysatoren im Detail zu verstehen, ist von großem Interesse, da deren Wirkungsgrad meist weit über industriell verwendeten Katalysatoren liegt. Als Beispiel hierfür sei die Umwandlung des chemisch inerten Stickstoffs (N_2) in leichter umsetzbare Verbindungen genannt. Dieser Prozess ist für die heutige Industrie von eminenter Bedeutung, da dieser die Rohstoffe für die Herstellung sehr vieler weiterer Produkte zur Verfügung stellt. Das heutige Standardverfahren hierzu ist der Haber-Bosch-Prozess, welcher aufgrund der stabilen Stickstoffbindung zur Umsetzung des N_2 ca. 450°C und 200 bis 300 bar benötigt [6]. In biologischen Systemen dagegen, in denen der analoge Prozess als Stickstoff-Fixierung bezeichnet wird, wird durch die Einbeziehung von Proteinen in die Reaktionskette unteranderem eine starke Absenkung der Aktivierungsenergie der N_2-Spaltungsreaktion erreicht, wodurch diese Reaktion in den Organismen bei Raumtemperatur und Normaldruck ablaufen kann.

Art der Proteins	Beispiel Anwendung und Funktion
Strukturproteine	Zusammenhalt von Zellgruppen, Bildung von Fasern und Gewebe wie Haut, Nägel, Klauen, Exoskelette
Enzyme	Katalyse chemischer Reaktionen, z. B. Dehydrierung, Phosphorylierung, Elektronentransfer
Transportproteine	Bindung und Transport verschiedener Stoffe, z. B.: Ionen durch Membranen, Sauerstoff, Fettsäuren
Speicherproteine	Zur Bereithaltung verschiedener Stoffe bei Bedarf, z. B.: Aminosäuren, Metallionen
Hormone	Regulierung des Stoffwechsels und des Wachstums
Schutzproteine	Abwehr beim Eindringen fremder Stoffe (Antikörper) und Reparatur verletzten Gewebes, z. B. bei der Blutgerinnung
Toxine	Angriff auf andere Organismen

Tabelle 2.1: Beispiele für Funktionen von Proteinen in Organismen [1].

Zum Verständnis dieser wie auch der anderen Reaktionen der Proteine und deren Vielzahl an biologischen Anwendungen ist es notwendig, den molekularen Aufbau der Proteine aufzuklären, da deren Funktionsprinzipien über die dreidimensionale Anordnung der Aminosäuren bestimmt ist. Die durch diese spezifische Anordnung erzeugte Geometrie ist auf die auszuführende Reaktion und die Eigenschaften der Umgebung abgestimmt, in der das Protein aktiv ist. Ändern sich die Eigenschaften dieser Umgebung, verliert das Protein zumeist seine Struktur und damit auch seine aktiven Eigenschaften. Transmembran-Proteine beispielsweise, die in den Stofftransport durch die Zellwand involviert sind, sind in einer polaren Umgebung anders gefaltet als in der apolaren Lipid-Membran, wodurch sie ihre Funktion verlieren. Diese Strukturänderung bzw. Entfaltung ist für viele Proteine ein reversibler Prozess, so dass bei einer Wiederherstellung der natürlichen Umgebungsparameter auch das Protein wieder in seinen aktiven Zustand zurückkehrt. Diese aktive Struktur der Proteine wird als nativer Zustand bezeichnet und ist häufig der Schlüssel zum Verständnis der Funktionsweise eines Proteins. Er besitzt meist eine gut definierte Geometrie bzw. besteht aus einem Ensemble dicht beieinander liegender Geometrien. Durch die Kenntnis der dreidimensionalen Struktur des nativen Zustandes lassen sich häufig die Reaktionsmechanismen der Proteine erklären und verstehen und ermöglichen Ansätze, diese ggf. zu beeinflussen oder auf andere Systeme zu übertragen.

Die dreidimensionale Struktur natürlich vorkommender Proteine ist wiederum für eine Aminosäuresequenz spezifisch, so dass im Idealfall aus der reinen Kenntnis der Abfolge der Ami-

nosäuren auf die Funktion des Proteins geschlossen werden kann. Diesen Zusammenhang zwischen der eindimensionalen Sequenz und der dreidimensionalen räumlichen Anordnung der Aminosäuren aufzuklären und herzustellen, ist ein breites Feld aktiver Forschung mit den unterschiedlichsten Methoden und Motivationen. Es ist eine Grundlagenfrage, die für viele wichtige Bereiche besondere Relevanz besitzt, angefangen bei der biomedizinischen Forschung, die Funktionsweise von Stoffwechselkreisläufen zu verstehen, bis hin zu einer industriellen Anwendung, beispielsweise zur Entwicklung neuer Medikamente mit gezielten Effekten.
Heutzutage sind für sehr viele Proteine die dreidimensionale Struktur wie auch die zugehörige Sequenz durch verschiedene experimentelle Methoden detailliert bekannt und über Datenbanken allgemein zugänglich. Ebenso existieren viele verschiedene Ansätze, um aus der Sequenz eines Proteins auf dessen native Struktur zu schließen. Doch trotz dieser breiten Informationsbasis und den großen geleisteten Anstrengungen ist bis heute das Proteinfaltungsproblem noch nicht vollständig gelöst, da noch keine allgemeingültige Methode existiert, in vernünftiger Zeit rein aus der Kenntnis der Sequenz ohne intensiven Einsatz von Datenbank-Information und Vergleiche mit anderen Sequenzen auf die Geometrie zu schließen. Für bestimmte Proteine und Proteinklassen existieren zwar Methoden, die in der Lage sind, bis zu einer gewissen Genauigkeit eine native Struktur zu bestimmen, dennoch fehlt immer noch eine universelle Lösung.
Diesem Problem widmet sich auch diese Arbeit, welche das Ziel verfolgt, einen allgemeinen Algorithmus zu entwickeln, einer Aminosäuresequenz den nativen Zustand zuzuordnen. Die Nahziele sind hierbei zunächst, bekannte Strukturen zu reproduzieren und die native Struktur in einer Menge andersartig gefalteter Proteine erkennen zu können. Ein Fernziel dieser Arbeit wäre beispielsweise, diesen Algorithmus, wenn die Erkennung bekannter Proteine erfolgreich war, auf Sequenzen mit bisher unbekannter Struktur zu erweitern, welche experimentell nicht oder nur schwer zugänglich sind, oder um zum Beispiel die Struktur mutierter oder künstlich generierter Sequenzen vorhersagen zu können.

Der zentrale Aspekt des in dieser Arbeit verfolgten Ansatzes zu diesem Problem war die Entwicklung eines neuartigen, reduziert-dimensionalen Proteinkraftfeldes durch eine bisher noch nicht verwendete Verknüpfung von statistischen Potentialen mit optimierten Linearkombinationen von Potentialfunktionen über Basisfunktionen. Hierzu wurde ein vereinfachtes Proteinmodell konstruiert, auf dessen Basis bekannte Datenbank-Proteine analysiert wurden, um die daraus erhaltenen Informationen für die Entwicklung verschiedener Seitenkettenmodelle und für die Definitionen der Potentialfunktionen zu verwenden. Zusätzlich wurden neue falsche Proteinstrukturen generiert, um durch diese Informationen über die Unterschiede zwischen der nativen Geometrie im Vergleich zu den nicht-nativen zu erhalten, indem sich die Aminosäuren in den falschen Strukturen anders anordnen als in der nativen Struktur. Mit Hilfe dieser Informationen wurden Potentialfunktionen konstruiert und erzeugt, die den nati-

ven Zustand gegenüber nicht-nativen Strukturen stabilisieren. Hierzu wurden die Parameter der Basisfunktionen an die Datenbank-Informationen angepasst, während die Gewichtungskoeffizienten der Potentialfunktionen über ein lineares Optimierungsproblem mit dem Ziel bestimmt wurden, der nativen Struktur den kleinsten Wert der Energiefunktion für alle Geometrien zu einer Sequenz zuzuordnen. Um Proteinstrukturen zu einer bestimmten Sequenz vorhersagen zu können, wurde auf der Basis des parametrisierten Kraftfeldes ein globaler Geometrie-Optimierungsalgorithmus implementiert, der einen genetischen Algorithmus verwendet, um den nativen Zustand im Konformationsraum zu identifizieren, welcher wie gefordert im Idealfall dem globalen Minimum der Energiefläche entsprechen sollte. Ebenso wurde das erhaltene Potential durch einen Erkennungstest, in welchem der native Zustand in einer vorgegebenen Menge an festen Geometrien erkannt werden muss, mit verschiedenen anderen Literaturpotentialen verglichen, die auf ähnlichen Proteinmodellen beruhten.

Die hierzu durchgeführten Arbeiten werden in den folgenden Kapiteln näher beschrieben. Zunächst werden im dritten Kapitel die allgemeinen Grundlagen des Aufbaus der Proteine sowie die wichtigsten Aspekte der dreidimensionalen Struktur der Proteine und des nativen Zustandes beschrieben. Darauf folgend werden im vierten Kapitel die durchgeführten Arbeiten, angewendeten Methoden und die Ergebnisse dieser Arbeit präsentiert, wobei auf die Punkte zur Auswahl der nativen Proteine, die statistische Auswertung der Datenbank-Strukturen, der Generierung der falschen Strukturen, die Entwicklung der Seitenkettenmodelle, der Kraftfeldansatz, die Parameteroptimierung und die Anwendung im Erkennungstest und in der globalen Geometrie-Optimierung eingegangen wird. Das fünfte Kapitel fasst die erhaltenen Ergebnisse zusammen und bietet eine Diskussion dieser sowie einen Ausblick auf zukünftige Arbeiten. Der Anhang im sechsten Kapitel enthält einige detaillierte Daten zu verwendeten Methoden, die nicht im vierten Kapitel aufgeführt wurden.

3 Theoretischer Hintergrund

3.1 Der Aufbau von Proteinen

Einleitend für diese Arbeit wird mit dem generellen Aufbau eines Proteins begonnen, um zu erläutern, wie die Aminosäuren auf Basis ihrer atomaren Struktur zu der dreidimensionalen Gesamtstruktur eine Proteins beitragen. Die Beschreibung der Proteinstrukturen erfolgt grundsätzlich auf vier verschiedenen hierarchisch konzipierten Ebenen, deren Anfang die Primärstruktur darstellt, welche die Zusammensetzung des Proteins angibt, über die Sekundär- und Tertiärstruktur, die die dreidimensionale Geometrie des Proteins an sich beschreiben, bis hin zu Quartärstruktur, die Protein-Protein- oder andere Molekül-Protein-Komplexe charakterisieren kann. Zusätzlich zu diesen vier Ebenen werden häufig zwei weitere Ebenen verwendet, deren deskriptive Inhalte zwischen denen der Sekundär- und der Tertiärstruktur einzuordnen sind. Es handelt sich hierbei um die Über- bzw. Supersekundärstruktur (*super secondary structure*) und die Domänen (*domains*) [7].

3.1.1 Primärstruktur

Proteine gehören zu den Biopolymeren, deren Monomerbausteine die Aminosäuren sind. Im Gegensatz zu künstlich hergestellten Proteinen, die prinzipiell aus jeder chemischen Modifikation einer Aminosäure aufgebaut sein können, werden die in der Natur vorkommenden Proteine nur aus einem Satz von zwanzig verschiedenen Standardaminosäuren, den sog. natürlichen oder kanonischen Aminosäuren (siehe Tab. 1.1), gebildet.[1]
Bei den natürlichen Aminosäuren handelt es sich ausschließlich um Carbonsäuren mit einer zur Carboxylfunktion α-ständigen Aminogruppe. Mit Ausnahme von Glycin besitzen diese am α-Kohlenstoffatom (C^{α}) ein Chiralitätszentrum. Alle natürlichen Aminosäuren sind L-Aminosäuren.

[1] Es sind weitere natürliche, aber selten vorkommende Aminosäuren bekannt, wie beispielsweise Selenocystein, Pyrrolysin oder Hydroxyprolin. Diese entstehen durch chemische Modifikation der 20 kanonischen Aminosäuren. Da diese in der Gesamtheit der Proteine eine untergeordnete Rolle spielen, werden sie in dieser Arbeit nicht näher behandelt.

Charakteristisch für jede Aminosäure sind die Seitenketten R, welche größtenteils die physikochemischen Eigenschaften der Aminosäure bestimmen. Diese sind lineare oder verzweigte aliphatische Kohlenwasserstoff-Ketten, die auch aromatische Ringe oder funktionelle Gruppen wie Amin-, Amid-, Carboxyl- oder Sulfidgruppen enthalten können. Prolin ist strukturell eine Ausnahme, da dies die einzige Aminosäure ist, bei der die Seitenkette zusätzlich kovalent mit der Aminogruppe verknüpft ist, wodurch sich ein Fünfring ergibt (siehe b) in Abb. 3.1), welcher besondere Auswirkungen auf die Proteinstruktur hat (s. u.).

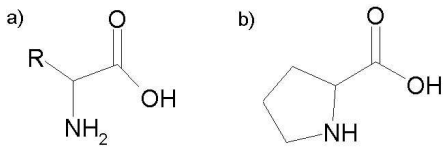


Abbildung 3.1: Struktur der Aminosäuren. a) Allgemeine Aminosäure mit Seitenkette R. b) Prolin.

Im Protein sind die Aminosäuren kovalent über eine Amidbindung miteinander verknüpft, die nur unter Einfluss starker Säuren, hoher Temperatur oder durch enzymatische Reaktion hydrolisiert. Die Verknüpfung mehrerer Aminosäuren wird als Peptid bezeichnet. Allgemein lässt sich der Aufbau eines Proteins angeben als

$$\text{H-(NH-CHR}_i\text{-CO)}_n\text{-OH} \tag{3.1}$$

wobei R_i die Seitenkette der i-ten Aminosäure ist. Aufgrund dieser Art der Verknüpfung besitzt ein Protein somit ein Ende mit einer freien Aminofunktion, welches als N-Terminus bezeichnet wird, und ein Ende mit einer freien Carboxylfunktion, welches als C-Terminus bezeichnet wird. Unter physiologischen Bedingungen[2] liegt der N-Terminus normalerweise protoniert als NH_3^+-Gruppe und der C-Terminus als deprotonierte Carboxylatgruppe COO^- vor.

Die Reihenfolge der Aminosäuren wird als Primärstruktur bezeichnet, und sie bestimmt die strukturellen und biologischen Eigenschaften eines Proteins [8]. Weil die Biosynthese der Proteine beim N-Terminus beginnt [9], werden die Aminosäuren im Protein zur Identifizierung entsprechend beginnend beim N-Terminus durchnummeriert.

Diejenigen Atome, die nicht zur Seitenkette zugehörig sind, werden auch als Proteinrückgrat bezeichnet. Dies sind die Amidgruppe, das C^α-Atom und die Carbonylgruppe.

Neben der Verknüpfung der Aminosäuren durch die Amidbindung können in Proteinen intramolekular auch Quervernetzungen auftreten, indem zwischen zwei Cystein-Seitenketten durch eine Reduktionsreaktion eine kovalente Bindung gebildet wird, was zu einer Disulfidbrücke führt. Diese Brücken geben der Struktur einerseits eine große Festigkeit, während sie andererseits auch sehr stark die geometrische Anordnung der Proteinteile bestimmen, da die

[2] Als "physiologische Bedingungen" sollen in dieser Arbeit die Werte der Parameter wie Temperatur, Druck, Ionenstärke etc. des Systems Protein und Umgebung verstanden werden, unter denen das Protein in seiner nativen Form vorliegt. Diese Werte können von Protein zu Protein unterschiedlich sein.

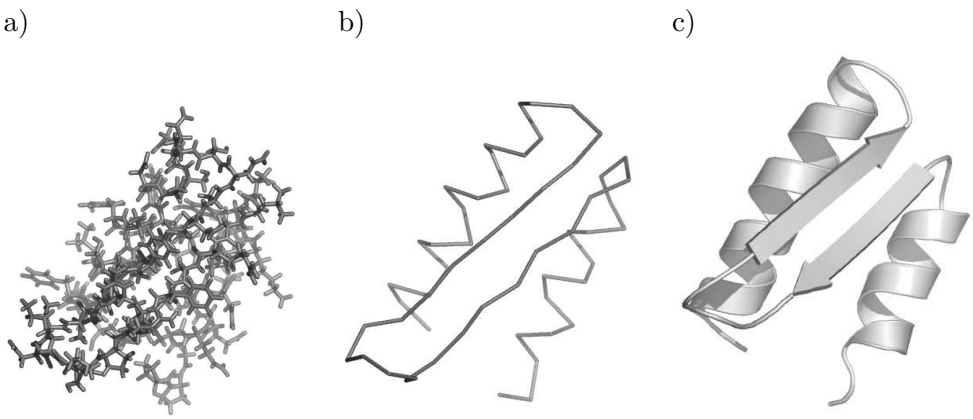


Abbildung 3.2: Die Reste 40 bis 96 des Proteins mit der PDB-Kennung 1wjw [10] in verschiedenen Darstellungen: a) Mit allen Atomen, b) nur C^{α}-Atome und c) in der *Cartoon*-Form.

Disulfidbrücken nicht zufällig, sondern nur zwischen bestimmten Sequenzpositionen gebildet werden. Allgemein lässt sich dies darstellen durch:

$$\text{-}H_2C^{\beta}\text{-SH} + \text{HS-}C^{\beta}H_2\text{-} \rightarrow \text{-}H_2C^{\beta}\text{-S-S-}C^{\beta}H_2\text{-} \tag{3.2}$$

Die Ausbildung von Disulfidbrücken findet während der Proteinsynthese oder Faltung nicht spontan statt, sondern wird über Hilfsmoleküle durchgeführt.

Da die dreidimensionale Struktur in einer Darstellung mit allen Atomen kompliziert ist, werden häufig, sofern bestimmte Details der Geometrie nicht von entscheidender Bedeutung sind, Proteine zur besseren visuellen Anschauung stark vereinfacht. Dies kann von der Auslassung lediglich der Wasserstoffatome bis hin zu einer Darstellung einer Aminosäure durch eine einzigen Punkt geschehen, wobei in der Regel das C^{α}-Atom gewählt wird. Beispiele für verschiedene Protein-Repräsentationen sind in Abb. 3.2 gezeigt.

Da die Struktur einer einzelnen Aminosäure und die Peptidbindung weitreichende Folgen für die gesamte dreidimensionale Geometrie eines Proteins hat, werden im folgenden hierfür die wichtigsten Aspekte aufgeführt.
Die Peptidbindung besitzt zwei Resonanzstrukturen, was sich bereits in der grundlegenden Strukturaufklärung der Peptide durch Pauling *et al.* 1951 zeigte [11], da die Länge der Bindung in der Peptideinheit nicht mit der einer Einfachbindung übereinstimmt, sondern kürzer ist. Die Peptidbindung hat ca. 40 % Doppelbindungscharakter und eine Resonanzenergie von ca. 20 kcal/mol [12–14]. Dies führt dazu, dass diese Gruppe von Atomen planar und die Rotation um diese Bindung stark eingeschränkt ist. Hieraus resultiert ein permanenter Dipol mit positiver Partialladung an der Amidgruppe und negativer Partialladung an der Carbonylgruppe. Zudem zeigt die Amidgruppe im gesamten pH-Bereich von 0 bis 14 nur eine sehr eingeschränkte Tendenz zur De-/Protonierung. Die Dipol-Dipol-Wechselwirkungen die-

ser Gruppe sind ein entscheidender Faktor in der dreidimensionalen Proteinstruktur, da sie zur Bildung von Wasserstoffbrücken zwischen Peptideinheiten führen. In gefalteten Proteinen bilden nur rund 2 % der Carbonyl- und 6 % der Amidgruppen keine Wasserstoffbrücken aus [15].

Da die meisten Proteine bei Raumtemperatur in ihrer aktiven Form vorliegen, sind die kovalenten Bindungen, sowohl im Rückgrat als auch in den Seitenketten, aufgrund der zur Verfügung stehenden thermischen Energie sehr nahe ihrer Gleichgewichtslage, ebenso wie die Bindungswinkel, wobei diese von der Gleichgewichtslage um ca. zwei bis drei Grad abweichen können [16]. Einzige wesentliche Ausnahme ist hier der N-C$^{\alpha}$-C-Winkel, der unter sterischen Zwängen eine größere Flexibilität als die anderen Winkel zeigt [17]. Die größten Freiheiten bieten daher die Torsionswinkel, deren energetische Rotationsbarrieren, mit Ausnahme des Torsionswinkel der Peptideinheit, vergleichsweise klein sind. Das Rückgrat besitzt pro Peptideinheit drei Torsionswinkel, die mit ϕ, ψ und ω bezeichnet werden (siehe Abb. 3.4 für deren Zuordnung). Hinzu kommen die Torsionswinkel der Seitenketten, die mit χ_k mit k = 1,2,... bezeichnet werden[3].

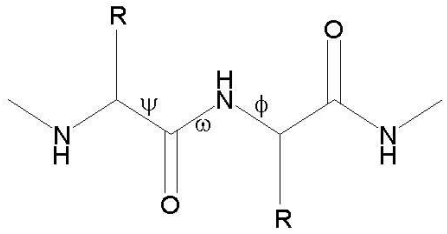


Abbildung 3.4: Torsionswinkel im Rückgrat. Markiert sind die zentralen Bindungen des jeweiligen Torsionswinkels.

Aufgrund des Doppelbindungscharakters der Peptidbindung ist die Energiebarriere für eine Rotation um den Torsionswinkel ω hoch. Sie beträgt ungefähr 20 kcal/mol, wodurch dieser Torsionswinkel nahezu nur in den zwei Konformationen *cis* ($\omega = 0°$) und *trans* ($\omega = 180°$) vorkommt, wobei aus sterischen Gründen die *trans*-Konfiguration in der überwiegenden Mehrheit der Fälle auftritt, was dadurch verursacht wird, dass sich bei zwei sequentiellen Aminosäuren in der *cis*-Konformation die Seitenketten sehr nahe kommen. Prolin ist aufgrund der besonderen Ring-Gestalt der Seitenkette eine Ausnahme. Bei dieser Aminosäure beträgt die Rotationsbarriere um den Winkel ω nur 13 kcal/mol und der Energieunterschied zwischen der *cis*- und *trans*-Konformation nur 2 kcal/mol. Aus diesem Grund treten *cis*-Bindungen häufig in Verbindung mit Prolinresten auf [19].

Die größten Freiheiten zur Realisierung verschiedener Geometrien bieten somit die Torsionswinkel ϕ und ψ, obwohl eine Analyse von Röntgenstrukturen zeigt, dass der Torsionswinkel ω trotz der hohen Energiebarriere von ca. 20 kcal/mol von den oben genannten Idealwerten um bis zu $\pm 20°$ abweichen kann. Dennoch sind in erster Näherung meist ϕ und ψ zur Definitio-

[3]Sowohl die Benennung der Torsionswinkel als auch die Zuordnung der Atome zu diesen ist durch die IUPAC in einer Richtlinie festgelegt [18].

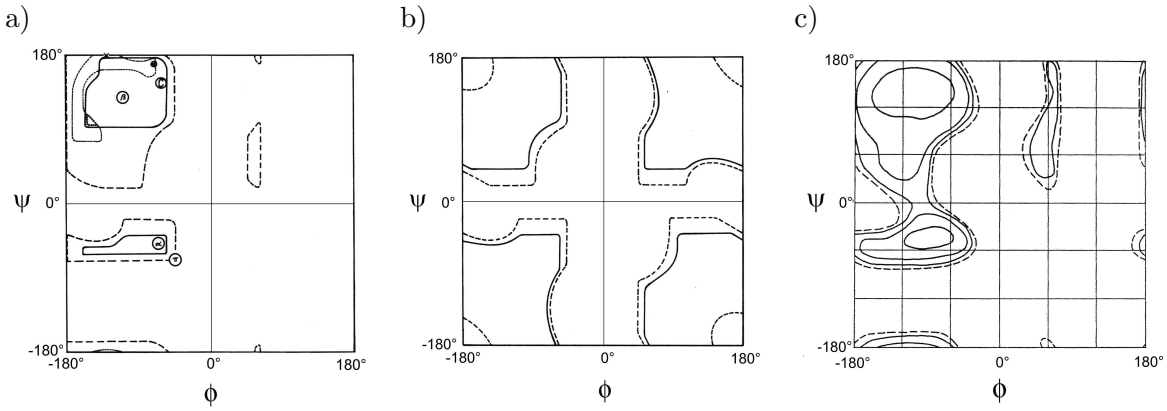


Abbildung 3.5: Zugängliche Bereiche für unterschiedliche Aminosäuren aus [19]. Dargestellt sind in a) die Bereiche im Harte-Kugel-Modell für eine allgemeine Aminosäure, außer Glycin und Prolin, in b) der Bereich für Glycin und in c) eine Auftragung der potentiellen Energie für ein Alanin-Dipeptid ohne das Harte-Kugel-Modell. Die gestrichelte Linie entspricht in diesem 0 kcal/mol, jede weitere Linie einem Energieunterschied von -1 kcal/mol. Wesentlicher Unterschied zum Harte-Kugel-Modell ist bspw. der erlaubte Brückenbereich bei $(\phi,\psi) = (-90°, 0°)$ und der etwas ausgedehntere Bereich bei $(\phi,\psi) = (60°, 120°)$.

nen und Beschreibung bestimmter Rückgratstrukturen ausreichend, so dass in den folgenden Abschnitten die erlaubten Werte für diese Torsionswinkel allein und die daraus resultierenden Folgen für die anderen hierarchischen Strukturebenen dargelegt werden.

3.1.2 Sekundärstruktur

Die Sekundärstruktur eines Proteins beschreibt die Geometrie bzw. Konformation des Proteinrückgrates. Diese kann in guter Näherung durch die Abfolge der Torsionswinkel ϕ und ψ wiedergegeben werden, da die anderen Freiheitsgrade wie oben beschrieben nahe ihrer Gleichgewichtswerte sind und deren energetisch höhere Zustände kaum populiert sind.
Die Ausrichtungen der Seitenketten fließen in die Angabe der Sekundärstruktur gewöhnlich nicht mit ein.

Während der Torsionswinkel ω durch den partiellen Doppelbindungscharakter der Peptidbindung keine freie Drehung ausführen kann, rotieren ϕ und ψ um einfache kovalente Bindungen ohne diese Einschränkung. Dennoch sind auch bei diesen nicht alle Drehwinkel bzw. alle Paarungen (ϕ,ψ) für diese Torsionswinkel gleichermaßen zugänglich. Analysen experimentell bestimmter Strukturen zeigen, dass 90 % der in Proteinen auftretenden (ϕ,ψ)-Torsionswinkelpaare nur ca. 14 % des insgesamt möglichen ϕ-ψ-Bereiches abdecken [20].
Diese Einschränkung der Torsionswinkel kann bereits gut im Rahmen eines vereinfachten

Harte-Kugel-Modells erklärt werden [21, 22], in welchem sich aufgrund der festen Atomradien für bestimmte (ϕ,ψ)-Paare erlaubte, scharf begrenzte Bereiche ergeben, in denen es zu keiner Kollision der Atome kommt (Für eine Analyse, welche Atomkollisionen die zugänglichen Bereiche einschränken siehe [23]). Es zeigte sich, dass die erlaubten Bereiche vor allem von der Größe und Form der Seitenketten abhängen, aber dennoch für alle Aminosäuren, bis auf zwei Ausnahmen, sehr ähnlich sind. In Abb. 3.5 a) sind diese zugänglichen Bereiche dargestellt. Von diesem weicht im wesentlichen zum einen nur Glycin ab, da aufgrund der nur aus einem Wasserstoffatom bestehenden Seitenkette die sterischen Ansprüche sehr viel geringer sind, was zur Folge hat, dass der zugängliche Bereich größer wird. Dies ist in Abb. 3.5 b) dargestellt. Zum Anderen ist dagegen bei Prolin aufgrund der Ringbildung der Seitenkette mit dem Rückgrat der ϕ-Torsionswinkel auf Werte des Bereiches bei ungefähr ϕ = -60°±20° begrenzt[4].
Das vereinfachte Harte-Kugel-Modell wurde auch durch detailliertere Berechnungen der potentiellen Energie bestätigt [24–26]. Geringe Unterschiede zeigten sich dabei in der Ausdehnung der erlaubten Gebiete, die im Modell ohne harte Kugeln größer sind. In Abb. 3.5 c) ist eine solche berechnete Potentialenergiekarte für ein Alanin-Dipeptid dargestellt.

Auf der Ebene der Sekundärstruktur sind Proteine durch einen Wechsel von (geometrisch) linearen Einheiten und Biegungen charakterisiert, wobei linear hier auf die Sekundärstruktur als ganzes bezogen ist, während die atomaren Bindungen in diesen ihre natürlichen Winkel einnehmen. Die linearen, regelmäßigen Struktureinheiten entstehen durch eine sequentielle Wiederholung von bestimmten (ϕ,ψ)-Paarungen, die vornehmlich in Bereichen mit niedriger potentieller Energie liegen. Es existieren in bekannten nativen Strukturen zwar auch abschnittsweise Geometrien, in denen Paarungen aus Bereichen mit hoher potentieller Energie vorkommen, diese treten aber nur selten unter bestimmten Bedingungen auf [27].
Die wichtigsten Strukturklassen, die durch Wiederholung von Paaren entstehen, sind die α-Helices und die β-Faltblätter. Die dritte wichtige Klasse, die dazu im Gegensatz aus einer eher unregelmäßigen Reihenfolge an Torsionswinkel-Paaren besteht, sind die sog. Schlaufen und Windungen (*loops and turns*), die keine lineare Geometrie besitzen, sondern für Richtungswechsel im Rückgrat sorgen und als Bindeglieder zwischen linearen Struktureinheiten dienen. Bezogen auf die Gesamtheit der bekannten Proteine lassen sich 89 % aller Aminosäuren im gefalteten Zustand einer dieser drei Klassen zuordnen, wobei wiederum bezogen auf die Gesamtheit der Proteine auf alle drei Klassen ungefähr die gleichen Anteile entfallen [28].
Die Zuordnung einer Aminosäure zu einer bestimmten Strukturklasse erfolgt meist auf Grundlage bestimmter geometrischer Kriterien, wobei hier die Schwierigkeit besteht, für diese Kriterien entsprechende Grenzwerte festzulegen, nach denen dann die Zuordnung durchgeführt wird, da es für jede Struktur gewisse Übergangsbereiche gibt. Dies betrifft vor allem die

[4]Die unterschiedlichen Werte für ϕ bei Prolin entsprechen einem bestimmten *ring-puckering*.

Reste an den Randbereichen einer Sekundärstruktureinheit, deren Geometrie unregelmäßiger ist, wodurch die Zuordnung nicht immer eindeutig ist. In der Literatur existieren viele unterschiedliche Methoden, Sekundärstrukturen zu bestimmen und zu klassifizieren. Zur einheitlichen Zuordnung wurde in dieser Arbeit der verbreitete Ansatz von W. Kabsch und C. Sander verwendet [29], welcher im entsprechenden frei verfügbaren Programm DSSP zur automatischen Erkennung von Sekundärstrukturen [30] enthalten ist.
Im folgenden werden die wichtigsten Sekundärstrukturen detaillierter dargestellt.

Helices

Helices gehören zu den mit am häufigsten vorkommenden Strukturen, so dass in der Natur viele Proteine existieren, die nur oder nahezu vollständig aus Helices aufgebaut sind (siehe z. B. 1et1 [31] oder Hämoglobin, PDB-Kennung 1o1i). Obwohl theoretisch verschiedene Formen der Helix möglich sind, werden die Strukturen mit großem Abstand durch die α-Helix dominiert. Andere Helixtypen sind zwar ebenfalls experimentell nachgewiesen, treten aber nur in seltenen Fällen auf.
Die Helices in den natürlich vorkommenden Proteinen sind stets rechtsgängige Helices, was darin begründet ist, dass natürliche Proteine nur L-Aminosäuren enthalten, wodurch die rechtsgängige Helix für diese Aminosäuren eine sterisch günstigere Anordnung als die linksgängige Helix ist, in welcher es zu einem engen Kontakt zwischen dem Carbonylsauerstoffatom des Rückgrats und dem C^{β}-Atom der Seitenkette kommt.

Die Bildung eine Helix wird durch verschiedene Faktoren begünstigt und stabilisiert: Die Torsionswinkel des Rückgrates liegen in Bereichen mit niedriger potentieller Energie bzw. für die selteneren Helices am Rande von diesen. Die Durchschnittswerte der Torsionswinkel für die häufigsten Helices sind in Tab. 3.1 aufgelistet (siehe dazu auch Abb. 3.5). Die Seitenketten haben eine sterisch günstige Anordnung, da sie zur Helixachse radial nach außen gerichtet sind. Die Seitenketten sequentieller Aminosäuren nehmen so einen Winkel von ca. 100° zueinander ein. Der permanente Dipol der Peptidbindung ermöglicht die Ausbildung von Wasserstoffbrücken, welche in einer Helix gewöhnlich zu sequenznahen Aminosäuren ausgebildet werden und so wesentlich zur Stabilisierung der Helix beitragen. Es gibt aber auch Helices, in die Wasserstoffbrücken intermolekular zu anderen Proteinen ausgebildet werden (s. u. Kollagenhelix). Aus dem Wasserstoffbrückenbindungsmuster, das sich für die verschiedenen Helices unterscheidet, resultieren die anderen geometrischen Eigenschaften wie bspw. der Schlaufenabstand oder der Schlaufendurchmesser.

Intramolekular stabilisierte Helices können durch Angabe der Anzahl an Aminosäuren pro Schlaufe n und der Anzahl der Atome m, die entlang des Rückgrates den Ring vom Donor zum Akzeptor der Wasserstoffbrücke bilden, in der Kurzform n_m angegeben werden. Zur besseren Veranschaulichung sind die intramolekular stabilisierten Helices mit den auftretenden Wasserstoffbrücken in Abb. 3.6 dargestellt.

In einer α-Helix sind die Dipolvektoren der Peptidbindung, die die Wasserstoffbrücken bilden, nahezu optimal zueinander orientiert, wodurch diese Helix eine starke Stabilisierung erfährt und sich für die Gesamthelix ein Dipolmoment ergibt [32]. Daher dienen in manchen Proteinen die Enden einer Helix als Bindungsstellen z. B. für Ionen.

Helixtyp	$\phi/^\circ$	$\psi/^\circ$	n_m
3_{10}-Helix	-49	-26	3_{10}
α-Helix	-57	-47	3.6_{13}
π-Helix	-55	-70	4.3_{16}

Tabelle 3.1: Übersicht der wichtigsten Helices. Werte aus [19, 28].

Die α-Helices zeigen sehr unterschiedliche Längen, wobei sie insgesamt eine durchschnittliche Länge von 10 Resten besitzen, aber auch bis zu 40 Resten lang werden können [33]. Experimentelle α-Helix-Strukturen haben häufig eine leichte Krümmung der Helix als ganzes bis zu 30° [34]. Dies kann dazu führen, dass an bestimmten Stellen in der Helix die Wasserstoffbrücken unterbrochen werden.

Wie aus den Auftragungen der potentiellen Energie deutlich wird (siehe Abb. 3.5), können prinzipiell alle natürlich vorkommenden Aminosäuren Helices formen, lediglich die individuellen Tendenzen der Aminosäure hierzu unterscheiden sich stark. Zusätzlich beeinflusst die lokale Sequenz und die Proteinumgebung die Tendenz zur Helixbildung [35]. Obwohl auch für Prolin der zur Bildung einer α-Helix nötige (ϕ, ψ)-Bereich zugänglich ist, findet man es gewöhnlich nicht innerhalb einer α-Helix, sondern nur vor oder nach dieser, da diese Aminosäure keine Wasserstoffbrücken mit einer Aminogruppe ausbilden kann. Weiterhin sind auch Sequenzen mit vielen lange und/oder geladene Seitenketten ungeeignet, Helices auszubilden, da sich die Seitenketten gegenseitig abstoßen. So bildet beispielsweise Polylysin im neutralen pH-Bereich keine Helices, da die Seitenketten protoniert und damit gleichsinnig positiv geladen vorliegen. Dagegen wird aber im pH-Bereich um 12 eine spontane Helixbildung von Polylysin beobachtet, weil die Seitenketten hier neutral geladen sind. Ähnliche Eigenschaften weisen auch die anderen Seitenketten wie z. B. Glutamin- oder Asparaginsäure auf. Im Allgemeinen gilt, dass die Aminosäuren Ser, Ile, Thr, Glu, Asp, Asn, Lys, Arg, Gly, Tyr und Cys eher eine Helix destabilisieren, während Ala, Leu, Phe, Trp, Met, His, Gln und Val Helices stabilisieren [1].

Betrachtet man die Verteilung der Aminosäuren entlang einer Helix, so zeigt sich, dass diese nicht uniform ist, sondern dass die statistische Wahrscheinlichkeit eine Aminosäure an einer bestimmten Position entlang der Helix anzutreffen, abhängig von den physiko-chemischen

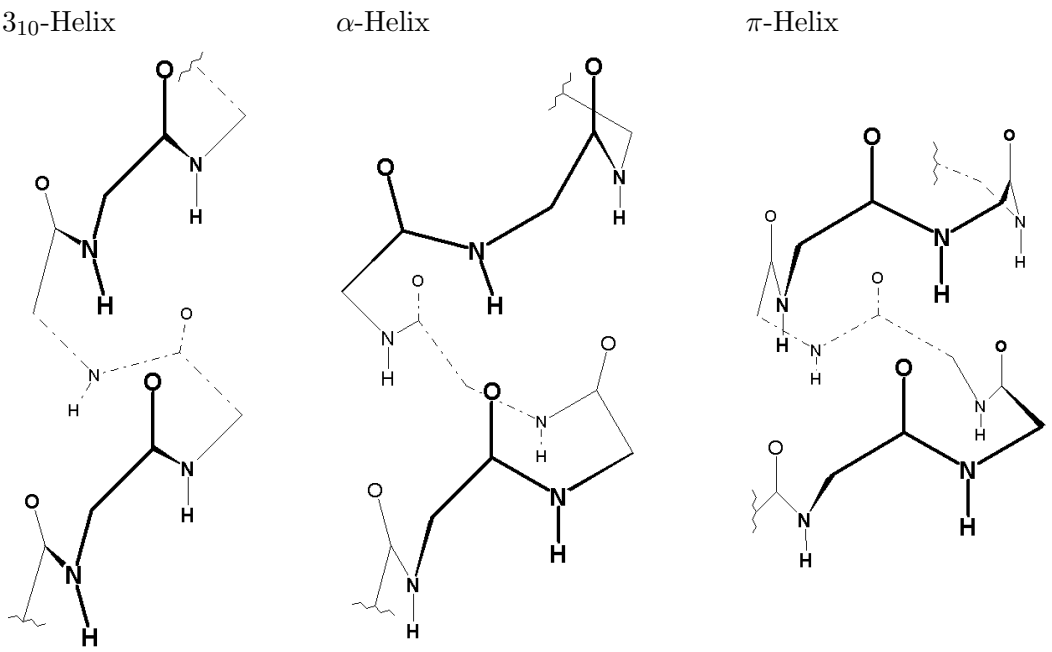


Abbildung 3.6: Darstellung verschiedener Helixtypen. Gezeigt sind nur die Schweratome des Rückgrates zur Verdeutlichung der Wasserstoffbrückenbindungen. Dicke, einfache und gestrichelte Bindungen dienen der dreidimensionalen Darstellung.

Eigenschaften der Aminosäure bzw. von deren Seitenkette ist. So weist das Besetzungsmuster der Aminosäuren diesbezüglich eine Periodizität auf, welche analog zur Periodizität Helix-Geometrie verläuft. Durchschnittlich alternieren alle vier Reste hydrophobe Aminosäuren, die kleine Seitenketten besitzen, mit eher hydrophilen Aminosäuren mit größeren Seitenketten. (Eine ausführliche Analyse dieser Verteilung von Aminosäuren in Helices findet man beispielsweise in [33]) Dies führt schließlich dazu, dass die gesamte Helix entlang ihrer Achse eine eher hydrophile und eine eher hydrophobe Seite besitzt [36]. Dies hat Auswirkungen darauf, wie α-Helices in gefalteten Proteinen mit den anderen Strukturen arrangiert sind.

Am Rand des α-Helix-Bereiches in einer Ramachandranauftragung (siehe Abb. 3.5 a und b) liegen zwei weitere Typen von Helices, die zwar in Proteinen beobachtet werden, aber äußerst selten auftreten. Dies sind die 3_{10}- und die π-Helix. Diese Helices bilden keine ausgedehnten Strukturen, sondern treten nur in sehr kurzer Form mit einer Windung auf. Beobachtet werden sie z. B. an den Enden einer regulären α-Helix im Übergangsbereich zu anderen Sekundärstrukturen oder als isolierte Einheit um dem Proteinrückgrat eine Richtungsänderung zu geben.

Im Unterschied zur α-Helix sind die Dipolvektoren der Wasserstoffbrückenbindungen in einer 3_{10}-Helix nicht mehr ideal zueinander ausgerichtet. Zudem sind die Seitenketten sterisch ungünstig angeordnet, da die Seitenketten der Reste i und $i+2$ in die gleiche Raumrichtung orientiert sind.

Bei der π-Helix sind die Seitenketten zwar günstiger als bei der 3_{10}-Helix ausgerichtet, aber dennoch immer noch sterisch ungünstiger als in der α-Helix. Auch die Dipolvektoren der Wasserstoffbrücken sind besser zueinander orientiert als in der 3_{10}-Helix. Aber aufgrund des größeren Radius der π-Helix sind die Atome des Proteinrückgrates weiter voneinander entfernt, wodurch die Dispersionswechselwirkungen schwächer sind.
Durch diese Faktoren sind die Bildung der 3_{10}- und der π-Helix verglichen mit der α-Helix ungünstiger, so dass diese beispielsweise als Faltungsintermediate auftreten, die sich dann weiter zur α-Helix umformen.

Hier sei noch am Rande vermerkt, dass neben diesen drei wichtigen Helices weitere Typen existieren, die meist Übergangsformen der anderen Helices darstellen und nur vereinzelt auftreten wie beispielsweise die Typ II α-Helix [37] oder die ϵ-Helix [38].

Ein sich von den bisher genannten Helices unterscheidender Typus kommt in dem für Wirbeltiere sehr wichtigen Faserprotein Kollagen vor, das zur Bildung des Bindegewebes dient und bei Tieren bis zu 1/3 des Körpergewichtes ausmachen kann.
Kollagen ist nicht aus einem einzigen Protein aufgebaut, sondern ist ein Aggregat aus drei Proteinen. Jeder Proteinstrang ist allgemein aus dem Aminosäuretriplet $(\mathrm{Gly,A,B})_n$ zusammengesetzt, wobei A und B verschiedene Aminosäuren sein können, nach denen sich die Klasse des Kollagens richtet. Sehr häufig treten für A oder B Prolin bzw. dessen modifizierte Form Hydroxyprolin auf. Eine solche chemische Modifikation der Seitenkette, oder beispielsweise auch durch Verknüpfung mit Kohlenhydratmolekülen, ist in Kollagen ein häufig auftretendes Merkmal, und betrifft neben Prolin auch die anderen Aminosäuren. (Für ein ausführliches *Review* zu Kollagenen siehe z. B. [39].)
Im Gegensatz zu den anderen Helixtypen bilden die Einzelstränge des Kollagens linksgängige Helices, wobei die Torsionswinkel des Proteinrückgrates ungefähr die Werte $(\phi, \psi) = (-60°, +140°)$ einnehmen [28]. Eine Windung enthält durchschnittlich 3.3 Aminosäuren. Die Bildung der linksgängigen Helix wird durch den sehr hohen Anteil an Glycin begünstigt, da die sterischen Anforderungen dieser Aminosäure wesentlich geringer sind, weil der oben erwähnte ungünstige Kontakt zwischen dem Carbonylsauerstoffatom und dem C^{β}-Atom hier einem Kontakt zu einem Wasserstoffatom gewichen ist, welches wesentlich kleiner ist. Der eingeschränkte Konformationsraum der vermehrt auftretenden Prolinreste im Kollagen verstärkt ebenfalls die Tendenz zur Formung einer linksgängigen Helix.

Neben dieser vergleichsweise ungewöhnlichen Zusammensetzung und der Bildung von linksgängigen Helices ist ebenso auch das Wasserstoffbrückenbindungsmuster konträr zu den vorher beschriebenen Helices, da die Wasserstoffbrücken nicht intra-, sondern intermolekular zwischen den drei Peptidsträngen ausgebildet werden, was zu einer Umwicklung der Ein-

zelstränge führt. Hierbei entsteht eine weitere, übergeordnete helicale Struktur, die aber im Gegensatz zu den einzelnen beitragenden Proteinsträngen rechtsgängig ist. Eine Windung dieser übergeordneten Helix enthält durchschnittlich 10 Aminosäuren und hat einen Länge von 86 Å.
Wie zuvor können bei der Bildung der Wasserstoffbrücken sowohl die Carbonyl- als auch Amidgruppen des Rückgrates beteiligt sein. Zusätzlich zeigt sich aber auch, dass in diesem Fall auch häufig die Hydroxygruppe der Hydroxyprolinseitenkette involviert sein kann. Während diese Ausbildung einer Wasserstoffbrücke von einer Seitenkette zum Rückgrat im Kollagen eine wichtige Rolle spielt, ist dieses Motiv dagegen in anderen Proteinen bezogen auf dessen Häufigkeit nur von untergeordnete Bedeutung. Da die Stabilisierung intermolekular stattfindet, ist ebenfalls wichtig, dass die Aminosäuren im Kollagenstränge eher kleine Seitenkette besitzen, damit sich die einzelnen Stränge nahe genug kommen können, um die entsprechenden Wasserstoffbrücken ausbilden zu können. Sterisch anspruchsvolle Seitenketten sind hier eher hinderlich. Aus diesem Grund bilden Glycin und Prolin einen großen Anteil der vorkommenden Aminosäuren [40].
Aufgrund dieser speziellen intermolekularen Stabilisierung ist ein einzelner Proteinstrang mit Kollagenkonformation in Isolation nicht stabil.

Faltblätter

In der zweiten wichtigen Sekundärstruktur, der β-Konformation, nimmt das Proteinrückgrat eine nahezu gestreckte Geometrie ein, so dass der Gesamtverlauf des Rückgrats annähernd einer geraden Linie folgt. Die Torsionswinkel des Rückgrates nehmen hierbei Werte ein, die im Bereich bei $(\phi, \psi) = (-120, +130°)$ der Ramachandranauftragung liegen. Aufgrund der gesteckten Konformationen und den damit verbundenen geringeren sterischen Zwängen entlang des Stranges ist in diesem (ϕ, ψ)-Gebiet der erlaubte zugängliche Bereich größer als beispielsweise der für die α-Helix. Damit geht einher, dass die Potentialenergiefläche in diesem Bereich relativ flach ist und somit kleine Variationen in den Torsionswinkeln vergleichsweise auch nur mit kleinen Änderungen der potentiellen Energie verbunden sind. Aus diesem Grund sind die experimentell gefundenen Torsionswinkel in der β-Konformation weniger scharf definiert als beispielsweise die Torsionswinkel für eine helicale Struktur.

Aufgrund der gestreckten Geometrie der β-Konformation ist, wie auch schon bei der Kollagenhelix beschrieben, deren Torsionswinkel am Rand des β-Gebietes liegen, keine Ausbildung von Wasserstoffbrücken zu dicht benachbarten Aminosäuren möglich. Dies führt dazu, dass isolierte β-Stränge nicht stabil bzw. in Isolation weniger stabil als andere mögliche Sekundärstrukturen sind, und damit nicht beobachtet werden. Daher liegen Sequenzabschnitte mit

β-Konformation in Proteinen mindestens paarweise vor, um die nötigen Wasserstoffbrücken bilden zu können. Hierbei spricht man dann von einem β-Faltblatt. Bei der Zusammenlagerung von mehreren Teilen eines Proteinstrangs kann dies parallel oder antiparallel geschehen, in dem Sinne, dass man dem Proteinrückgrat wie oben erwähnt vom N- zum C-Terminus eine Richtung zuordnet. In beiden Fällen unterscheiden sich das Muster der Wasserstoffbrückenbindungen, welches in Abb. 3.7 dargestellt wird. Im antiparallelen Faltblatt sind die Wasserstoffbrücken paarweise angeordnet und ungefähr rechtwinklig zum Rückgrat orientiert, während sie im parallelen Faltblatt einen gewissen Winkel zum Rückgrat einnehmen. Auch die gemittelten Werte der Torsionswinkel im Proteinrückgrat sind jeweils leicht unterschiedlich: Für den parallelen Fall sind die Mittelwerte $(\phi, \psi) = (-119, +113°)$ und für den antiparallelen Fall bei $(\phi, \psi) = (-139, +135°)$ [28], allerdings streuen diese Werte im antiparallelen Fall wesentlich größer über den erlaubten Bereich als für die parallele Anordnung [41]. Bilden mehr als zwei Abschnitte ein Faltblatt, so können sie sich auch in einer Mischung aus parallel und antiparallel zusammenlagern, wobei bei größeren Proteinen ein starker Trend zu gemischten Faltblätter zu erkennen ist [42]. Ebenfalls beobachtbar ist, dass sich bei paralleler Anordnung häufig wesentlich mehr Stränge zu einem Faltblatt zusammenfinden, während bei antiparalleler Anordnung häufig nur zwei Abschnitte zusammenkommen. Generell ist das β-Faltblatt als eine gewellte Ebene beschreibbar, bei der die C^{α}-Atome die Knickpunkte bzw. Spitzen darstellen. Die Zusammenlagerung zweier Abschnitte erfolgt stets so, dass die C^{α}-Atome der beiden Abschnitte benachbart sind, während dies für die anderen funktionellen Gruppen des Rückgrats nur in der parallelen Anordnung erfüllt ist. Die Wasserstoffbrücken, die die Sequenzabschnitte miteinander verbinden, liegen ebenfalls alle in einer Ebene, deren Dipole sind aber nicht wie einer α-Helix räumlich identisch orientiert, sondern zwischen zwei β-Strängen stets antiparallel.

Die Seitenketten entlang eines Proteinstranges in der β-Konformation zeigen jeweils abwechselnd aus den zwei Seiten dieser Ebene heraus. Dabei zeigen aber Seitenketten, die zu den benachbarten C^{α}-Atomen gehören, die in unterschiedlichen Sequenzabschnitten sind, jeweils in die gleiche Richtung. Es zeigt sich, dass die Paarung der Seitenketten, der zu unterschiedlichen Abschnitten gehörenden aber im β-Faltblatt benachbarten C^{α}-Atome nicht vollständig willkürlich ist, sondern dass es leichte Präferenzen gibt, welche Seitenketten zusammenkommen [43]. Wie auch in Helices werden Paarungen mit geladenen Seitenketten vermieden und Paarungen mit zwei hydrophoben Seitenketten eher bevorzugt. Auch findet man häufiger Paarungen, in denen eine verzweigte Seitenketten neben einer unverzweigten vorliegt. Es besteht eine Tendenz dazu, dass die Start- und Endreste eines Faltblattes eher hydrophil sind, während die inneren Aminosäure eher hydrophob sind, und dass eine Seite des Faltblattes insgesamt hydrophil und die andere hydrophob ist.

Untersuchungen zeigten, dass die Formung eines Stranges in eine β-Struktur von der lokalen nicht-bindenden Wechselwirkung der Seitenketten mit dem Rückgrat abhängt und dass

a) b)

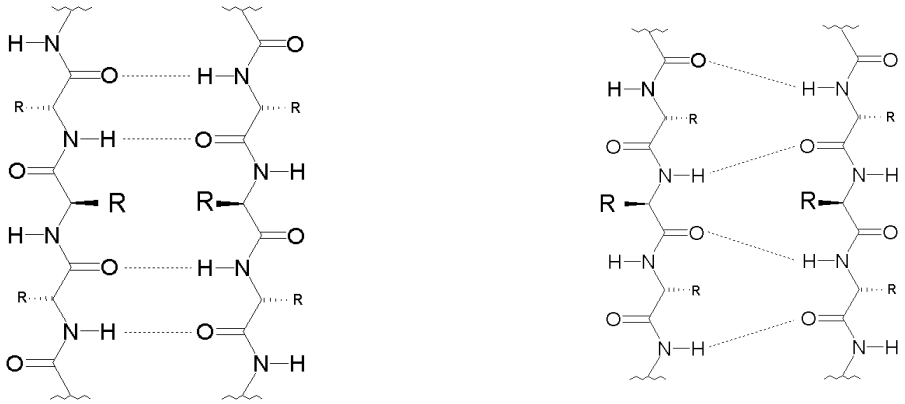


Abbildung 3.7: Wasserstoffbrückenmuster der β-Faltblattstrukturen, in a) anti-paralleler oder b) in paralleler Ausrichtung der Rückgratabschnitte.

dies die Tendenzen, diese Struktur auszubilden, direkt mit der Sequenz der Aminosäuren zusammenhängt [44, 45]. Zudem scheint die Ausbildung einer hydrophoben Oberfläche durch die Seitenketten eine der entscheidenden Triebkräfte zur Zusammenlagerung verschiedener Rückgratabschnitte in β-Konformation zu einer Faltblattstruktur zu sein [46].

Schlaufen und Windungen

Die bisher behandelten Sekundärstrukturen bilden aufgrund der Repetition der Torsionswinkel räumlich lineare Einheiten. Native Proteine besitzen aber insgesamt eine annähernd sphärische oder ellipsoidale Gestalt. Um diese Form zu erreichen, benötigt man Strukturelemente, die dem Rückgrat eine Richtungsänderung ermöglichen. Diese Einheiten fallen unter die Kategorie Schlaufen und Windungen (*loops and turns*). Wie auch die anderen Sekundärstrukturen können diese in vielen unterschiedlichen Längen auftreten. Windungen, die viele Aminosäuren enthalten, sind sehr flexibel und kommen in unterschiedlichsten Konformationen vor, so dass hier bis jetzt noch kein Ordnungsschema existiert. Dagegen lassen sich aber die sehr kurzen Windungen, die sog. *tight turns* oder *reverse turns*, in bestimmte Klassen einteilen. Sie umfassen im Regelfall (mindestens) vier Aminosäuren. Ähnlich wie die Vorhersage der α-Helix auf Basis von Wasserstoffbrückenbindungsmöglichkeiten durch Pauling und Corey wurden auch die engen Windungen früh durch theoretische Überlegungen durch C. M. Venkatachalam vorhergesagt [47]. Die Vorhersage der Windungen basierte auf einer günstigen H-Brückenwechselwirkung zwischen dem i-ten Carbonylsauerstoffatom und dem $(i + 3)$-ten Stickstoffatom. Dieses Motiv wird in ungefähr der Hälfte aller engen Win-

Klasse	$\phi_{i+1}/^\circ$	$\psi_{i+1}/^\circ$	$\phi_{i+2}/^\circ$	$\psi_{i+2}/^\circ$	Kommentar
I	-60	-30	-90	0	Deformierte 3_{10}-Helix
I'	60	30	90	0	
II	-60	120	80	0	
II'	60	-120	-80	0	
III	-60	-30	-60	-30	Identisch mit 3_{10}-Helix
III'	60	30	60	30	
IV	-	-	-	-	Verschiedene Typen. Zu Typ IV gehörend, wenn mindestens zwei Torsionswinkel um mehr als 40° von Typ I bis III' abweichen.
V	-80	80	80	-80	
V'	80	-80	-80	80	
VI	-	-	-	-	Wenn ein *cis*-Prolinrest an der $(i+2)$-Position vorliegt.
VII a	-	<60	180	-	
VII b	-	180	<60	-	

Tabelle 3.2: Übersicht über die Definitionen enger Windungen (*tight turns*) basierend auf [48, 53].

dungen beobachtet [48]. Die andere Hälfte besitzt keine derartige Stabilisierung durch eine Wasserstoffbrückenbindung, weil die hierzu benötigten Atome häufig in anderen Sekundärstrukturen involviert sind oder Wasserstoffbrücken zum Lösungsmittel ausbilden.
Die Klassifizierung der kurzen Windungen erfolgt ebenfalls anhand der Rückgrattorsionswinkel. Tabelle 3.2 gibt eine Übersicht über die Windungsklassen. Die aufgeführten Torsionswinkelwerte sind idealisierte Werte. Neben dieser Möglichkeit der Einteilung über die Torsionswinkel existieren weitere Methoden zur automatisierten Erkennung von Windungen wie zum Beispiel die Analyse des Rückgratverlaufes anhand der C^α-Atomkoordinaten. Hierbei werden beispielsweise die relativen Orientierungen der Vektoren zwischen den sequentiellen C^α-Atomen verglichen [49, 50].
Neben diesen kurzen Windungen mit einer Länge von vier Resten existiert noch die γ-Windung (γ-*turn*), welche kürzer ist und nur durch drei Aminosäuren geformt wird [51, 52].
Enge Windungen dienen häufig als Bindeglied zwischen den anderen Sekundärstrukturen wie zum Beispiel zwischen zwei α-Helices oder einer α-Helix und einem Faltblatt. Ein sehr häufig auftretendes Motiv ist die Verbindung zweier sequenznaher β-Faltblattkonformationen (β-*hairpin*), bei dem eine kurze Windung zwischen diesen liegt und diese antiparallel verbindet, so dass eine $\beta T \beta$-Struktur entsteht (mit T als *turn*).
Eine allgemeine Charakteristik von Windungen ist neben der typischen Position zwischen den linearen Struktureinheiten zum einen ihre Zusammensetzung und zum anderen ihre Lage im gefalteten Zustand.

So werden enge Windungen bevorzugt von polaren Aminosäuren gebildet [54]. Der Grund kann wie auch schon bei β-Faltblättern darin liegen, dass bestimmte Sequenzen in Verbindung mit lokalen Wechselwirkungen der polaren Aminosäuren die Bildung von Windungen begünstigen. Zusätzlich sind häufig die Aminosäuren Glycin und Prolin an Windungen beteiligt, was an den geringeren sterischen Anforderungen des Glycins und an der Ausbildung von *cis*-Bindungen über Prolin begründet ist [48, 55]. Aufgrund der Polarität der beteiligten Aminosäuren zeigen Windungen eine starke Tendenz dazu, im gefalteten Zustand eher an der Oberfläche des Proteins an der Grenzfläche zum Lösungsmittel aufzutreten [50], so dass beispielsweise eine hydrophobe lineare Struktur zur Oberfläche verläuft, der Oberflächen-Kontakt dann über eine Windung vermittelt wird, der dann wieder eine andere hydrophobe Struktur folgt, die wieder ins innere des Proteins orientiert ist.

3.1.3 Supersekundärstruktur, Domänen und Tertiärstruktur

Die in den vorhergehenden Abschnitten beschriebenen Geometrien des Proteinrückgrates sind die Grundlage zur Bildung der Tertiärstuktur, welche die Anordnung der Sekundärstrukturelemente räumlich zueinander beschreibt. Sie gibt somit die Gesamtfaltung eines Proteins an, auf deren Basis sich Proteine in Klassen einteilen lassen. In der Literatur sind inzwischen sehr viele Tertiärstrukturen und Proteinklassen bekannt, so dass im Rahmen dieser Einleitung nur eine kurze Übersicht über diesen Bereich gegeben werden kann.

Die Entstehung der dreidimensionalen Struktur eines Proteins basiert auf einer Reihe von verschiedenen Wechselwirkungen und muss bestimmte geometrischen Einschränkungen berücksichtigen, so dass die Formierung des nativen oder eines anderen Zustandes auf einer Balancierung der verschiedensten Beiträge beruht. Im folgenden sollen kurz einige wichtige Aspekte hiervon aufgelistet werden:

- Planare Peptidgruppen, eingeschränkte Torsionswinkel, relativ fixierte Bindungslängen und -winkel im Rückgrat und Seitenketten
- Unterschiedliche Eigenschaften der Seitenketten: Polare, unpolare, geladene Seitenketten, hydrophobe Wechselwirkungen, sterische Ansprüche, bevorzugte Rotationszustände
- Wasserstoffbrücken zwischen Peptidgruppen, zum Lösungsmittel, zwischen Seitengruppen und Rückgrat und zwischen Seitenketten
- Einbau von Fremdionen oder -molekülen
- Entropieabnahme im Protein
- Umgebungseffekte: pH-Wert, Salzkonzentration, Polarität der Umgebung

Diese und weitere Aspekte können jeweils eine Rolle bei der Definition der Tertiärstruktur zu einer bestimmten Sequenz spielen (siehe hierzu auch Abschnitt 3.2).
Deren Beschreibung wird in der Literatur allgemein ebenfalls in hierarchische Ebenen aufgeteilt, welche bestimmte Aspekte bzw. Abschnitte der Proteinstruktur erfassen. Die nach der Sekundärstruktur nächst höhere Beschreibungsebene ist die Supersekundärstruktur, die lokal die Zusammenlagerung der Sekundärstrukturen wiedergibt. Hier gibt es, aufgrund der geringen Anzahl an verschiedenen Sekundärstrukturen ebenfalls häufig wiederkehrende Strukturen, mit denen sich sehr viele Proteine beschreiben lassen. Häufig beobachtete Motive sind beispielsweise parallel oder antiparallel orientierte Bündel aus α-Helices, die sich, wenn die Helices ausgedehnter sind, umeinanderwickeln. Des weiteren die oben beschriebenen *β-Hairpins*, also zwei antiparallele β-Strukturen verbunden über eine enge Windung, oder auch gemischte Zusammenlagerungen von mehreren parallelen oder antiparallelen β-Strukturen (beispeisweise ein sog. *Greek Key* oder *Jellyroll*), welche z. B. Schichtstrukturen oder Hohlräume in Proteinen formen können (*β-Barrel*). Daneben treten auch gemischte Strukturen auf, in denen α-Helices die β-Faltblätter miteinander verbinden und z. B. β-α-β- oder β-α-β-α-β-Einheiten bilden (Rossmann-Motiv).
Wie mit diesen Beispielen bereits angedeutet wurde, besteht in Proteinen bei Ausbildung der Tertiär- bzw. der Supersekundärstruktur eine Tendenz dahingehend, dass Aminosäuren und Sekundärstrukturen, die sich in der Sequenz nahe sind, auch räumlich nah beieinander liegen und dass deren Orientierung zueinander bevorzugt antiparallel ist [56]. Obwohl es noch kein einheitliches Faltungsmodell zur Bestimmung der Faltungswege der Proteine existiert, kann angenommen, dass diese Nahordnung einen wichtigen Einfluss auf die Faltung hat, indem diese dafür sorgt, dass der bei der Faltung abzusuchende Konformationsraum kleiner wird und dass durch die Ausbildung der Sekundärstruktur in kleineren Abschnitten, die nah beieinander liegen, die Entropiebarriere für diesen Vorgang geringer ist [19, 57].
Eine Beschreibung der Supersekundärstruktur ist neben der rein geometrischen Einteilung auch dahingehend von Vorteil, da diese Strukturmotive häufig auch mit bestimmten Funktionen in Verbindung stehen. Beispielsweise dient eine α-β-α-Struktur in vielen Fällen als Bindungsstelle für Calciumionen oder für die DNA [58].

Auf der nächst höheren Strukturebene, die durch Zusammenlagerung von Supersekundärstrukturen gebildet wird, sind die sog. Domänen. Dies sind große kompakte Abschnitte in Proteinstrukturen, die zumeist eine in sich abgeschlossene vom Rest des Proteins größtenteils unabhängige Struktur mit eigenem Faltungsmuster bilden. Viele der sehr großen Proteine bestehen aus Domänen, die durch kurze Sequenzabschnitte miteinander verbunden sind. Solche Domänen weisen zumeist die gleichen Charakteristika wie einzelne ganze globuläre Proteine auf, indem sie einen hydrophoben Kern mit gut definierter Sekundärstruktur ausbilden, welcher wenige oder keine Wassermoleküle enthält, während an der Oberfläche eher hydrophile

Aminosäuren zu finden sind [59].
Wie auch der Supersekundärstruktur zuvor, können ganze Domänen häufig mit einer bestimmten Funktion in Verbindung gebracht werden, so das aus einer ähnlichen Faltung zweier Domänen auf eine ähnliche Funktion geschlossen werden kann [60]. Dies betrifft vor allem Domänen, deren Sequenzen eine Ähnlichkeit von größer als 30 % aufweisen. Domänen können sich auch gegeneinander verschieben, wodurch dem Protein eine gewisse Flexibilität gegeben wird, was beispielsweise zur Ausführung der Funktion von Bedeutung sein kann.
Eine Unterscheidung zwischen einer Supersekundärstruktur und einer Domäne ist bisweilen nicht eindeutig. Auch existieren unterschiedliche Ansätze und Methoden, Domänen in Proteinstrukturen zu identifizieren, so dass diese nicht immer kongruente Resultate liefern.
Grundsätzlich werden Domänen nach den in ihnen dominierenden Sekundärstrukturen klassifiziert, so dass sie als α- oder β- oder als gemischte (α+β)-Domänen bezeichnet werden. Eine weiteres Unterscheidungskriterium ist, ob die Zusammenlagerung der Sekundärstrukturen in ihnen eher parallel oder antiparallel ist. Neben diesen Kategorien werden Domänen weiter danach eingeteilt, ob sie viele Disulfidbrücken oder viele Metallionen enthalten.

3.1.4 Quartärstruktur

Die Quartärstruktur der Proteine beschreibt das Zusammenspiel der Proteine mit anderen Molekülen, z. B. Aggregate aus mehreren Proteinen. Ein sehr bekanntes Beispiel hierfür ist Hämoglobin, das für die Bindung und den Transport von Sauerstoff im menschlichen Körper zuständig. Hämoglobin ist ein Komplex aus vier voneinander unabhängigen Proteinen, der über nicht-bindende Wechselwirkungen stabilisiert wird.
Da sich diese Arbeit auf die Beschreibung der proteininternen Wechselwirkungen beschränkt, wird hier auf eine ausführlichere Darstellung der Quartärstruktur verzichtet.

3.2 Der native Zustand

In den vorangegangenen Abschnitten wurden die wichtigsten Strukturmerkmale beschrieben, die in unterschiedlichen Kombinationen und Verhältnissen wesentliche Anteile der nativen Strukturen der Proteine bilden. Diese wurden über sehr viele Jahre der Evolution über Mutationen in der Sequenz angepasst, wobei die unterschiedlichsten Aspekte wie beispielsweise Faltungsstabilität, Reaktivität oder Reaktionsspezifität eine Rolle spielen. Einige dieser zu berücksichtigen Konzepte sind in nativen Strukturen auf geradezu grenzwertige Balancen hin optimiert. So müssen zum Beispiel Sequenzen erzeugt werden bzw. entstehen, deren nativer bzw. aktiver Zustand kinetisch erreichbar und gleichzeitig thermodynamisch stabil ist, so dass

• Fehler in der DNA-Information
• Fehler beim Auslesen der DNA-Information
• Mutation oder Fehler während des Synthesevorgangs
• Unvollständiger oder falscher Faltungsweg
• Unvollständige oder falsche Aggregation bei Multiproteinkomplexen
• Zu hohe oder zu niedrige Temperatur
• Oxidation, Reduktion, Amidierung, Glycosylierung oder Nitrosylierung von Aminosäuren
• Enzymatische Reaktion mit anderen Proteinen
• Zu hohe Salzkonzentration, falscher pH-Wert
• Reaktion mit Detergentien wie z. B. Fettsäuren

Tabelle 3.3: Faktoren, die zur Denaturierung oder zu einem falsch gefalteten Protein führen können.

die Faltung ohne große Energiebarrieren überwinden zu müssen dorthin laufen kann und dass das Protein auch in diesem Zustand verbleibt, um die Funktion auszuführen, und nicht zu einem anderen inaktiven Zustand weiterfaltet. Hierbei darf aber die native Proteinstruktur energetisch nicht so niedrig sein (thermodynamisch stabil), dass die Reaktionen an bzw. in dem Protein nicht mehr ausgeführt werden können, beispielsweise dadurch, dass die Energiebarriere für ein Substrat, um Zugang zum reaktiven Zentrum eines Enzyms zu erhalten, zu groß ist [61].
Aufgrund dieser labilen Balance bestimmter Eigenschaften vieler nativer Proteinen, in denen der native Zustand verglichen mit anderen inaktiven Zuständen energetisch nur minimal stabiler ist, reagieren diese empfindlich auf Änderungen der Umgebungseigenschaften, wodurch es bei einer zu starken Veränderung der Systemvariablen zu einer Denaturierung des Proteins kommt. Aber ebenso können andere Faktoren, wie eine mutierte Sequenz oder ein falscher Faltungsweg dafür sorgen, dass der native Zustand nicht erreicht oder entfaltet wird. Einige Beispiele für denaturierende Faktoren sind hierzu in Tab. 3.3 aufgeführt. Solche induzierten Denaturierungen sind für gewöhnlich reversibel, sofern sie dabei nicht die chemische Struktur bzw. die Zusammensetzung des Proteins verändern. Nach Aufhebung der Störung und Wiederherstellung des natürlichen Systems kehren die meisten Proteine wieder in ihren nativen Zustand zurück.

Trotz dieser gewissen Labilität und der Abhängigkeit von den Eigenschaften der Umgebung der biologisch relevanten Proteinstrukturen zeigen experimentelle Ergebnisse, dass Proteine in ihrer natürlichen Umgebung meist eine gut definierte Struktur besitzen, die, von thermischen Fluktuationen abgesehen, zeitlich sehr stabil ist und nur wenige Änderungen erfährt. Ausnah-

men hiervon können Vorgänge sein, die notwendigerweise eine Konformationsänderung des Proteins voraussetzen, wie beispielsweise die Dislokation des Proteins vom Syntheseort zum Einsatzort im Organismus, wobei es beim Transport durch eine Membran teilweise entfaltet wird, um den Kanal passieren zu können, oder z. B. bei der Ausführung der Funktion selber. So ist das membranüberspannende kanalerzeugende Molekül selbst ein Protein, dass sich beim Transportvorgang weiten kann, um die Überführung von großen Proteinen zu ermöglichen [62, 63]. Dies kann beispielsweise durch die oben erwähnte Verschiebung von Domänen geschehen, wodurch die Sekundärstruktur im wesentlichen erhalten bleibt und sich nur die Tertiärstruktur ändert. Nach Beendigung der Reaktion kehrt das Protein im Regelfall in den Ausgangszustand zurück.

Sowohl Experimente aus den Anfängen der Proteinforschung wie auch aktuelle Ergebnisse zeigen, dass der Faltungsprozess zum nativen Zustand für viele Proteine völlig spontan abläuft [64–66], sofern das System die Eigenschaften der nativen Umgebung hat und der Faltungsprozess keine weiteren Hilfsmoleküle (Chaperone) benötigt, um z. B. Disulfidbrücken zu bilden. Die physikochemischen Eigenschaften und Interpretationen dieses Prozesse sind seit Beginn der intensiven Forschung an Proteinen, als gut aufgelöste Röntgenstrukturen und die ersten theoretischen Modelle verfügbar waren, Gegenstand intensiver Diskussion in der Literatur. Auch wenn bereits viele Fragen der Proteinfaltung beantwortet werden konnten, so sind bis heute aufgrund der Komplexität dieses Problems noch nicht alle Details verstanden bzw. entziehen sich (noch) einer eindeutigen Interpretation und Erklärung.
Eine entscheidende Frage hier ist, welche Triebkräfte für den spontanen Faltungsprozess wichtig sind und wie der native Zustand stabilisiert ist. Eine grundlegende Antwort hierzu lieferte Anfinson 1973 mit der Aufstellung der thermodynamischen Hypothese [8], welche im wesentlichen zwei Kernpunkte beinhaltet:

- Der native Zustand ist das globale Minimum der freien Energie,
- Er steht im Gleichgewicht mit den anderen Faltungszuständen.

Nach diesem Ansatz ist der native Zustand somit unter thermodynamischer Kontrolle, wobei der Faltungsprozess durch die Minimierung der freien Energie getrieben wird. Der Beweis dieser Hypothese ist experimentell wie auch theoretisch schwer zu führen. Aufgrund der Größe und der Komplexität des Proteinfaltungsproblems können hierzu derzeit keine hochgenauen Berechnungen auf Basis quantenmechanischer Methoden durchgeführt werden. Stattdessen werden Modellsysteme untersucht, deren Auflösung von vereinfacht bis hin zu stark vergröbert reicht. Die einzelnen Ansätze unterscheiden sich teilweise auch stark in Detailfragen, z. B. wie die Potentialenergiefläche aussieht oder wie die Umgebung bzw. das Lösungsmittel behandelt wird. Aufgrund dieser Tatsache sind die erhaltenen Ergebnisse meist sehr unterschiedlich und schwer zu vergleichen. Es muss besonders im Hinblick auf die stark vereinfachten Modelle

stets die Frage gestellt werden, ob die daraus resultierenden Energieflächen, auf der die Proteine berechnet werden, die wesentlichen bzw. die wichtigen Eigenschaften der realen Fläche erfassen, und so das wirkliche Verhalten nähern. Da aber die vielen unterschiedlichen theoretischen Methoden wie auch beispielsweise Denaturierung-Faltungsexperimente zu ähnlichen Ergebnissen bezüglich des Faltungsverhaltens und der Interpretation der dominierenden Kräfte kommen, wird allgemein der Schluss gezogen, dass die thermodynamische Hypothese für die Mehrheit der Proteine korrekt ist, und der native Zustand das globale Minimum der freien Energie ist (siehe hierzu z. B. das ausführliche *Review* von K. A. Dill [67] oder auch [68–73]). Die hierzu entwickelten Modelle und Untersuchungen zeigen, dass die dem Faltungsprozess zugrundeliegende frei Energiefläche, dargestellt in der Abhängigkeit von der geometrischen Differenz der Proteinstruktur zum nativen Zustand, trichterförmig ist, mit dem Minimum an der Position des nativen Zustandes, was eine starke Tendenz zur Bildung dieses Zustandes zur Folge hat. Des weiteren ist der native Zustand durch eine gewisse Energielücke vom nächsten nicht-nativen lokalen Minimum bzw. meta-stabilem Zustand getrennt und besitzt einen gewissen Einzugsbereich, der vielen eng benachbarte lokale Minima mit niedrigen Energiebarrieren entspricht [74–76]. Diese entsprechen thermischen Bewegungen der Proteinstruktur, beispielsweise der Rotation der Seitenketten oder Bewegungen von Sekundärstrukturteilen. Normalerweise sind hiermit keine größeren Veränderungen der Struktur verbunden.
Zu diesem Modell der thermodynamischen Erklärung existieren allerdings einige Ausnahmen, bei denen der native Zustand ein lokales und nicht das globale Minimum der freien Energie ist. Diese Ergebnisse sprechen eher für das auf dem Levinthal-Theorem basierenden Modell der kinetischen Stabilisierung des nativen Zustandes, in welchem dieser einen meta-stabilen Zustand auf der Energiehyperfläche darstellt, der durch hohe Energiebarrieren von den anderen Zuständen und dem globalen Minimum getrennt ist [77]. Hierbei kann es ein Ensemble dieser kinetisch stabilisierten Minima geben kann, die alle eine ähnliche Geometrie besitzen, aber auf unterschiedlichen Faltungswegen aus verschiedenen denaturierten Anfangszuständen erreicht werden können [78]. In solchen Proteinen ist das globale Minimum häufig mit einer biologisch inaktiven Geometrie verbunden [79–81].
Welches der beiden Modelle die korrekte Beschreibung liefert, ist noch nicht eindeutig geklärt. Es besteht auch die Möglichkeit, dass beide Modelle richtig sind, wobei dann die Art der Stabilisierung des nativen Zustandes protein- bzw. sequenzspezifisch wäre. Dies kann darauf begründet sein, dass die Evolution der Proteine parallel und unabhängig voneinander ablief, wodurch verschiedene Systeme bzw. Organismen unterschiedlich stabilisierte Proteine hervorbrachten, angepasst an die jeweiligen Bedingungen, in denen sie existieren.

Neben der Diskussion, ob der native Zustand thermodynamisch oder kinetisch stabilisiert wird, herrscht in der Literatur ebenfalls Uneinigkeit darüber, welche Wechselwirkungen tatsächlich essentiell für dessen Stabilisierung sind und wie darüber detailiert die Energiefläche

des Proteins aussieht. Hier führen ebenfalls verschiedene Ansätze zu teils widersprüchlichen Aussagen und Meinungen. Inzwischen ist anerkannt, dass der native Zustand wie oben bereits erwähnt eine gewisse energetische Distanz zu den anderen Zuständen besitzt, dass aber die Stabilisierungsenergie mit ca. 20 bis 60 kJ/mol nicht sehr groß ist [82]. Diese entspricht lediglich der Energie einiger weniger Wasserstoffbrückenbindungen. Kompliziert wird die Bestimmung der Stabilisierungsenergie, da sie sich aus einer Summe wesentlich großerer Energiebeiträge ergibt, die sich teilweise gegenseitig aufheben. Im folgenden sollen nun einige Eigenschaften des nativen Zustandes sowie die treibenden Kräfte diesen zu bilden kurz beschrieben werden.

Obwohl ein Protein aus flexiblen Einheiten, wie verschiebbaren Domänen oder Seitenketten, die unterschiedliche Rotamerzustände einnehmen können, besteht, ist die Packungsdichte des nativen Zustandes dennoch sehr hoch. Sie beträgt ca. 75 % Raumfüllung, was der durchschnittlichen Raumerfüllung der Kristalle entspricht [83]. Diese Raumerfüllung geht mit einer wohl geordneten Struktur der Proteine einher, deren Ausbildung während der Faltung aber mit einem großen Verlust an Konfigurationsentropie verbunden ist, da sowohl das Rückgrat wie teilweise auch die Seitenketten auf einige wenige Zustände beschränkt werden [19]. Betrachtet man ein Protein im Vakuum, so wird dieser Verlust der Konfigurationsentropie durch die Bildung der Wasserstoffbrücken bzw. der anderen proteininternen Wechselwirkungen überkompensiert, so dass der native Zustand dort sehr stark stabilisiert wird. Betrachtet man dagegen jedoch Proteine in ihrer natürlichen Umgebung im gelösten Zustand in einem wässrigen Medium, so wird die Thermodynamik des Faltungsprozesses komplizierter, da in diesem weitere Effekte hinzukommen [67, 84]. Dies ist zunächst der hydrophobe Effekt, welcher darauf beruht, die Grenzfläche zwischen apolaren Seitenketten wie Valin, Cystein oder Leucin zum Wasser zu minimieren. Dieser Effekt wurde lange Zeit als die dominierende Triebkraft zur Bildung des nativen Zustandes angesehen [67, 85]. Dieser ist sicherlich sehr wichtig, aber er muss differenzierter betrachtet werden. Zunächst muss beachtet werden, dass es aufgrund der unregelmäßigen Verteilung der Aminosäuren entlang der Sequenz nicht möglich ist, lediglich die hydrophoben Seitenketten im Kern des Proteins einzulagern, so dass auch polare oder geladene Seitenketten im inneren zu finden sind. Daher sind im gefalteten Zustand durchschnittlich 83 % der hydrophoben, 63 % der polaren und 54 % der geladenen Seitenketten eingelagert, ebenso wie auch 82 % der polaren Rückgrat-Peptidgruppen [86]. Durch die Einlagerung der polaren und geladenen Seitenketten sowie der Peptidgruppen entsteht ein großer destabilisierender Desolvatationsenergiebeitrag, da diese Gruppen beim Faltungsprozess, bei welchem das Wasser bis auf sehr wenige Moleküle aus dem Protein gedrängt wird, vom Lösungsmittel getrennt werden. Es ist unklar, ob die Einlagerung der geladenen bzw. polaren Seitenketten im inneren stabilisierend oder destabilisierend wirkt [87, 88]. Wie sich zeigt, ist diese Desolvatationsenergie gleich oder größer dem Energiegewinn durch die Bildung der proteininternen Wechselwirkungen wie z. B. Wasserstoffbrücken [64, 89], wodurch

sich diese Beiträge nahezu aufheben, so dass der native Zustand durch diese Energiebeiträge weit weniger stabilisiert wird, als früher angenommen. Daher ist die Betrachtung der reinen proteininternen Wechselwirkungen nicht ausreichend, die spontane Faltung zu erklären. Wie sich in neueren theoretischen Untersuchungen zeigte, ist der entscheidende Energiebeitrag zur Stabilisierung des nativen Zustandes die Änderung der Entropie der umgebenden Wassermoleküle [90], wobei jedoch der detaillierte Beitrag sich von der ursprünglichen Annahme unterscheidet. Das ursprüngliche Bild des hydrophoben Effektes bestand in der Annahme, dass Wassermoleküle in der Nähe zu apolaren Gruppen des denaturierten Proteins in ihrer Rotation eingeschränkt sind. Nach der Einlagerung dieser Gruppen im inneren des Proteins durch die Faltung werden diese Wassermoleküle frei gesetzt und besitzen mehr Rotationsfreiheitsgrade, was einer Zunahme der Entropie der Wassermoleküle entspricht. Wie sich aber in den aktuellen Arbeiten zeigten, ist der Hauptanteil des Energiegewinnes nicht im Rotations-, sondern im Translationsanteil der Entropie zu finden. Dieser entsteht dadurch, dass im ungefalteten Zustand die Seitenketten ein großes ausgeschlossenes Volumen besitzen, das für die Wassermoleküle nicht zugänglich. Dieses Volumen verringert sich während des Faltungsprozess, wodurch die Translationsentropie der Wassermoleküle zunimmt.
Fasst man diese Ergebnisse zusammen, so zeigt sich, dass der native Zustand im Vakuum durch die internen Wechselwirkungen stark überstabilisiert wird, während dagegen die Thermodynamik im wässrigen Medium durch die Solvatationseffekte komplizierter ist und man die spontane Faltung nur unter Berücksichtigung des Wassers erklären kann.

3.3 Experimentelle Methoden zur Strukturaufklärung

Die genaue Strukturaufklärung stellt die Grundlage für das Verständnis jeden molekularen Vorgangs da, auch wenn sich teilweise bestimmte Abschnitte von Reaktionen aufgrund ihrer Geschwindigkeit oder anderer Faktoren nur schwer oder nicht messen lassen. Nicht nur in diesen Fällen, sondern ganz allgemein, kann die Kombination aus Experimenten und Theorie Impulse für den jeweils anderen Forschungsbereich geben. Während die Theorie bestimmte Beobachtungen erklären oder Ideen für neue Experimente bereitstellen kann, so ist das Experiment stets der Gradmesser für die Theorie, da nur diejenigen Modelle vernünftig sein können, die einen Versuchsaufbau und die aus diesem erhaltene Resultate richtig beschreiben. Für diese wie auch andere theoretische Arbeiten, werden experimentelle Ergebnisse benötigt, um Parameter für das gewählte Modell zu erhalten. In diesem speziellen Fall werden die dreidimensionalen Strukturen von Proteinen benötigt, um über diese eine Potentialfunktion anzupassen, die diese Strukturen reproduzieren soll. Eine kurze Übersicht, mit welchen experimentellen Methoden Informationen über Proteinstrukturen sich erhalten lassen, ist in [91] zu finden.
Experimentell existieren heutzutage verschiedene Methoden, die Struktur der Proteine aufzuklären. Im folgenden werden die wichtigsten Methoden hierbei kurz angesprochen. Die grundsätzliche Schwierigkeit der Protein-Strukturbestimmung liegt zunächst darin, ähnlich wie ebenfalls bei vielen anderen Biomolekülen auch, das zu untersuchende Protein aus seiner Umgebung zu isolieren, um es gezielt analysieren zu können. Dieser Prozess kann sehr aufwendig sein. Einige Proteine verhalten sich hierbei besonders problematisch, da sie ihre Struktur verändern, wenn sie aus ihrer nativen Umgebung entfernt werden. Dies ist z. B. ein Effekt, den man bei Membran-Proteinen beobachten kann.
Aufgrund der großen Anzahl an Atomen bei bereits mittelgroßen Proteinen, werden zur Strukturbestimmung Verfahren benötigt, die viele detaillierte Informationen über das Molekül liefern. Abhängig davon, in welcher Umgebung das Protein untersucht werden soll bzw. untersucht werden kann, stehen heutzutage hauptsächlich zwei Methoden zur Verfügung. Dies sind die Röntgenstrukturanalyse und die NMR-Spektroskopie, wobei beide Methoden unterschiedliche Struktureigenschaften messen und zueinander komplementäre Methoden sind [92].
Die Röntgenstrukturmethode basiert auf der Beugung von Röntgenstrahlung an den Netzebenen von Kristallen, wobei die Strahlung von den Elektronen gestreut wird. Für diese Methode ist somit ein Proteinkristall notwendig, was für einige Proteine problematisch sein kann, da der Extraktionsprozess und/oder die Kristallisation sehr schwierig sein können. Aus den erhaltenen Beugungsbildern lässt sich die Elektronendichteverteilung berechnen, mit deren Hilfe die dreidimensionale Struktur bestimmt werden kann. Qualitativ höherwertige Strukturen sollten eine Auflösung von besser als 1.5 bis 2 Å besitzen. In Großforschungsanlagen können sehr

gute Auflösungen erzielt werden, indem Röntgenstrahlung mit einer hohen Strahlungsdichte aus einem Synchrotron verwendet wird
Eine ähnliche Methode ist die Sreuung von Neutronen. Diese erfordert ebenfalls einen großen apparativen Aufwand, da die Neutronen in einem Kern-Spaltungsreaktor erzeugt werden müssen. Im Gegensatz zu Röntgenstrahlen werden die Neutronen an den Atomkernen gestreut. Hierbei ist der besondere Vorteil, dass die Neutronenbeugung ein bessere Auflösung liefert, und dass auch die Wasserstoffatome bestimmt werden können, die in der Röntgenbeugung normalerweise nicht detektiert werden.

Mit der Röntgenstrukturmethode können nur Kristalle untersucht werden, wodurch es mit dieser Methode nicht möglich ist, Proteine in ihrer natürlichen wässrigen Umgebung bei dynamischen Prozessen zu beobachten. Für diese Fälle und für Proteine, die sich nicht oder nur schwer kristallisieren lassen, kann die zweite wichtige Strukturaufklärungsmethode, die NMR-Spektroskopie, hilfreich sein. Mit dieser werden Übergänge von unterschiedlichen Spinzuständen von Atomkernen beobachtet, wobei der Kern im Grundzustand einen nicht verschwindenden Spin besitzen muss, damit diese Methode angewendet werden kann. Solche in Proteinen vorkommende und hierfür verwendete Atomkerne sind gewöhnlich ^{1}H, ^{13}C und ^{15}N. Die Struktur eines Moleküls kann aus verschiedenen Messdaten ermittelt werden. Häufig werden die Daten des Kern-Overhauser-Effektes verwendet, es können aber z. B. auch die chemische Verschiebung, dipolare Kopplungen oder Relaxationszeiten herangezogen werden [93, 94]. Die hieraus gewonnen Informationen über die Struktur werden über verschiedene Algorithmen in Geometrien umgerechnet, wobei aber abhängig von der Qualität und der Menge an Daten sowie der Dynamik der Struktur die erhaltene Geometrie nicht immer eindeutig ist, sondern es vorkommen kann, dass verschiedene oder ähnliche Geometrien zu den gleichen NMR-Daten passen, wodurch beispielsweise Datenbank-Strukturen diverse Alternativmodelle für eine Proteinstruktur enthalten können. Eine weiterer Nachteil der NMR-Spektroskopie ist, dass nicht beliebig große Moleküle untersucht werden können.

Neben diesen beiden sehr wichtigen Methoden zur Strukturbestimmung existieren weitere, die Informationen über die Geometrie liefern, aber nicht direkt in eine vollständige Struktur umgerechnet werden können. Eine weitere Methode ist beispielsweise die Verwendung von polarisierter ultravioletter Strahlung im Bereich um 200 nm, mit welcher sich Informationen über den Gehalt an den unterschiedlichen Sekundärstrukturen gewinnen lassen, da diese bestimmte Elektronenübergänge der Amidgruppe anregt. Die Übergangsenergien sind spezifisch für bestimmte Konformationen des Rückgrates, woraus auf die Sekundärstruktur geschlossen werden kann [95].

Ebenfalls wichtig ist die Massenspektrometrie bei der unter Verwendung verschiedener Methoden Molekülionen hergestellt werden, deren Masse-zu-Ladung-Verhältnis bestimmt werden kann. Aufgrund der Überschussenergie im Ionisationsprozess defragmentieren im Regelfall die untersuchten Moleküle, so dass nach der Auftrennung dieser Fragmente nach ihrem Masse-zu-Ladung-Verhältnis ein Reihe von Signalen gemessen werden kann, die charakteristisch für das jeweilige Molekül sind. Auf diese Weise lassen sich beispielsweise durch Vergleich mit Datenbanken Proteine oder Fragmente bestimmen. Sequenzinformationen über das Protein werden auch durch einen gezielten Abbau des Proteins beispielsweise durch Enzymreaktionen und anschließender Analyse der Reaktionsprodukte gewonnen [96].

4 Methoden und Ergebnisse

In diesem Abschnitt werden die für die vorliegende Arbeit angewendeten Methoden und die erhaltenen Ergebnisse dargestellt. Die zwei Hauptpunkte dieser Arbeit waren die Erstellung eines vergröberten Kraftfeldes zur Beschreibung der proteininternen Wechselwirkungen und dessen Anwendung in der globalen Geometrieoptimierung zur Vorhersage des nativen Zustandes. Um diese Ziele zu erreichen, mussten zunächst einige notwendigen Vorarbeiten geleistet werden. Im Folgenden wird hierzu ein kurzer Überblick über die Reihenfolge der wichtigsten Punkte und den Ablauf der Arbeiten gegeben, gefolgt von detaillierten Beschreibungen der entsprechend in einzelne Themenblöcke gegliederten Arbeitsschritte.

Die Grundlage zur Beschreibung einer Proteinstruktur ist die Wahl eines Modells. Hierzu wurde, um die Anzahl an Freiheitsgraden zu reduzieren, ein vereinfachtes Proteinmodell angesetzt, das drei Wechselwirkungszentren pro Aminosäure enthielt.
Die Beschreibung der Wechselwirkungen zwischen diesen Punkten wurde als empirisches Potential angesetzt, das sich aus der Linearkombination verschiedener Basisfunktionen zusammensetzte. Da es nicht das Ziel dieses Ansatzes war, thermodynamische Daten oder *ab-initio*-Potentiale zu reproduzieren, lässt sich das gewählte Kraftfeld auch als eine reine Bewertungsfunktion für Proteinstrukturen auffassen. Zur Bestimmung der im Potential enthaltenen Parameter wie beispielsweise Gleichgewichtsabstände usw. musste zunächst ein Referenz-Datensatz ausgewählt werden, an welchen die Parameter angepasst wurden. Hierzu wurden die verfügbaren experimentell bekannten Protein-Strukturdatensätze herangezogen und aus diesen ein qualitativ hochwertiger Datensatz ausgewählt. Die in diesem Datensatz enthaltenen Geometrien wurden im Hinblick auf die angestrebte Form des Potentials nach bestimmten statistischen Mustern analysiert und die erhaltenen Daten zur Gewinnung der Parameterwerte für die Kraftfeld-Funktionen genutzt, indem diese entweder direkt in die Basisfunktionen als Konstanten eingingen oder indem bestimmte Potentialfunktionen an die statistischen Daten angepasst wurden. Des weiteren wurden die Datenbank-Informationen zur Realisierung drei unterschiedlicher Seitenkettenmodelle genutzt.
Nach der Festlegung der Basisfunktionen mussten die Gewichtungskoeffizienten der Funktionen untereinander bestimmt werden. Die Zielgröße für die Bestimmung der Gewichtungskoeffizienten war der Energieunterschied zwischen den nativen Proteinstrukturen und Strukturen,

die zwar die gleiche Aminosäuresequenz wie das native Protein besaßen, deren Geometrien sich aber von diesen unterschied. Hierbei sollten die Koeffizienten so bestimmt werden, dass die Energie der nativen Struktur für alle Strukturen einer Sequenz am energetisch niedrigsten ist. Zur Durchführung wurden daher zusätzlich zu den nativen Datenbank-Strukturen die Geometrien falsch gefalteter Proteine benötigt. Die hierzu erhältlichen Datensätze wurden zunächst einer Analyse unterzogen, wobei sich bei diesen viele Probleme zeigten, die beispielsweise darin bestanden, dass viele Datensätze zu wenige falsche Strukturen für eine effektive Koeffizienten-Optimierung enthielten, diese unpassend für das gewählte Proteinmodell waren oder dass die Strukturen qualitative Mängel enthielten, indem sie unrealistische interatomare Abstände enthielten. Aus diesen Gründen wurden eigene Programme zur Erzeugung von falschen Strukturen implementiert.
Diese hiermit neu generierten falschen Proteine wurden, zusammen mit den bekannten nativen Datenbank-Strukturen, anschließend zur Bestimmung der Gewichtungskoeffizienten der Potentialfunktionen über die Formulierung eines linearen Ungleichungssystems verwendet, zu dessen Lösung ein bereits existierendes Programmpaket angewendet wurde.
Das resultierende Potential wurde in einen neu programmierten genetischen Algorithmus implementiert, um mittels diesem die Energieflächen global nach energetisch sehr niedrigen Minima absuchen zu können. Da aufgrund der Größe des Konformationsraumes und der begrenzten Anzahl an Proteinen zur Parametrisierung des Potentials nicht zu erwarten war, in einem Schritt eine optimale Energiefunktion zu erhalten, in welcher alle nativen Zustände das globale Minimum einnehmen, wurde die globale Geometrie-Optimierung und die Optimierung der Gewichtungskoeffizienten zu einem iterativen Zyklus zusammengefasst, in welchem Proteinstrukturen, die in der Geometrie-Optimierung erzeugt wurden und eine niedrigere Energie als die native Struktur besaßen, zu den falschen Strukturen für die Koeffizienten-Optimierung hinzugefügt wurden. Dieser Zyklus sollte solange fortgesetzt werden, bis in der Geometrie-Optimierung keine Strukturen mehr auftraten, die energetisch günstiger als das native Protein waren.
Neben der globalen Geometrie-Optimierung wurde als Vergleich, um die Vorhersageleistung des Potentials bestimmen zu können, ein Erkennungstest mit publizierten falschen Strukturen gegen andere Protein-Potentiale durchgeführt. Die für diese Tests bewerteten Protein-Strukturen waren nicht im Optimierungssatz enthalten.
Dieser Ablauf der einzelnen Punkte der Arbeit ist in Abb. 4.1 zur Übersicht zusammengefasst. Die einzelnen Punkte werden in den folgenden Abschnitten ausführlich beschrieben.

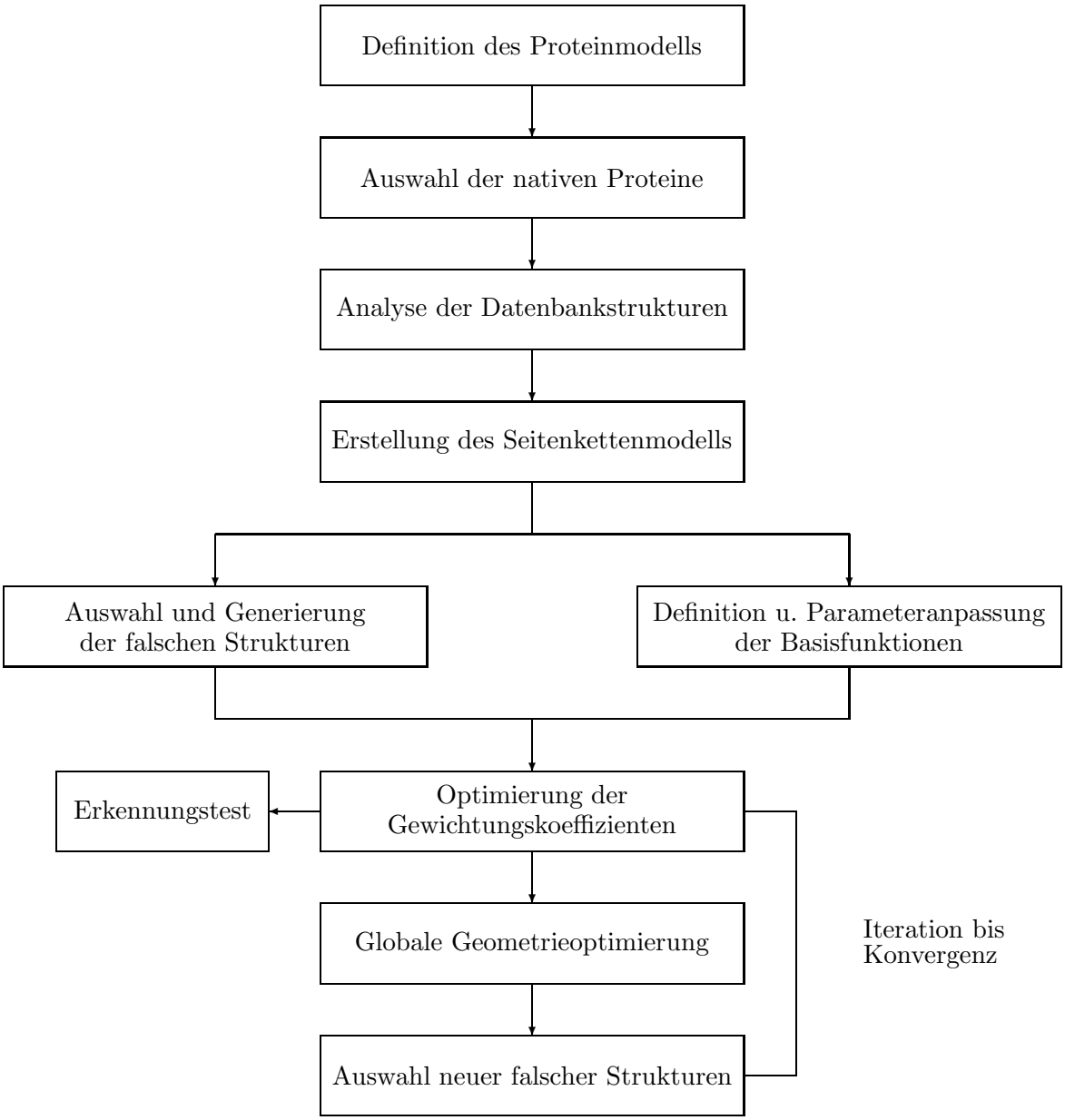


Abbildung 4.1: Verlaufsübersicht der durchgeführten Arbeiten zur Erstellung des Kraftfeldes und dessen Anwendungen.

4.1 Das Proteinmodell

Obwohl nach dem heutigen Wissensstand die (relativistische) Quantenmechanik die exakteste mathematisch-physikalische Beschreibung atomarer oder molekularer Systeme erlaubt, ist die Anwendung ihrer Gleichungen für größere Systeme numerisch nicht durchführbar, da die computertechnischen Anforderungen wie auch die Rechenzeit in Abhängigkeit von der Systemgröße stark zunehmen. Daher ist man für die Beschreibung solcher Systeme auf die Verwendung vergröberter Modelle und dahingehend auch vereinfachter Potentialfunktionen zur Beschreibung der Wechselwirkungen zwischen den Körpern angewiesen. Die Wahl des Modells ist somit ein Kompromiss zwischen der angestrebten Genauigkeit der Berechnung und den zu erzielenden Ergebnissen auf der einen Seite und den verfügbaren Computerressourcen und der erforderlichen oder realisierbaren Rechenzeit auf der anderen Seite.
Für die Simulationen von Proteinen, welche besonders die Bereiche der Vorhersage der Faltung, die Analyse der (klassischen) Dynamik und die Wechselwirkungen zwischen Proteinen umfassen, werden sehr viele unterschiedliche Modelle verwendet. Eine Übersicht mit einer großen Anzahl an Literaturverweisen über vereinfachte Proteinmodelle ist in [97] zu finden. Die unterschiedlichen Darstellungen einer Proteingeometrie lassen sich zunächst grundsätzlich darin einteilen, welche Interaktionspunkte verwendet werden bzw. welche Atome zu einem Punkt zusammengefasst werden oder welche direkt repräsentiert sind. Die einfachsten Darstellungen verwenden hierbei ganze Aminosäuren als einen einzigen Punkt. Komplexere Modelle haben zumeist die Gemeinsamkeit, dass zunächst der Verlauf des Rückgrates beschrieben wird, wobei dies zumeist über die Koordinaten der C^{α}-Atome erfolgt. Dies wird in Abhängigkeit von der gewählten Genauigkeit um zusätzliche Punkte erweitert, wobei es sich beispielsweise zur Beschreibung der Seitenkette um das C^{β}-Atom oder um den Seitenkettenschwerpunkt handeln kann, oder Punkte zur Beschreibung der Wasserstoffbrückenbindungen oder weiterer Atome des Rückgrates, um z. B. die Torsionswinkel zu erfassen. Grundsätzlich werden aber in der Regel nur die Schweratome explizit einbezogen und die Wasserstoffatome ausgelassen.
Des weiteren können sich die Proteinmodelle darin voneinander unterscheiden, ob die Positionsangabe der Interaktionspunkte auf einem Gitter mit diskreten Zuständen erfolgt oder ob dies kontinuierlich bzw. gitterfrei geschieht. Ebenso sind zwischen diesen beiden Darstellungen Mischformen möglich, indem ein Teil der Interaktionspunkte diskrete und der andere Teil kontinuierliche Zustände einnehmen kann.
In direktem Zusammenhang mit der Wahl der unterschiedlichen Möglichkeiten der Darstellung des molekularen Aufbaus erfolgt auch die Festlegung zur Beschreibung der Wechselwirkungen. Hierauf wird im Abschnitt 4.4 näher eingegangen.

Das Potential, das in dieser Arbeit entwickelt wurde, basierte ebenfalls auf einer vergröberten Darstellung eines Proteins. Dieses verwendete drei Wechselwirkungszentren pro Aminosäure, welche durch das C^{α}-Atom, den Schwerpunkt der Seitenkette und durch den Punkt Z zur Beschreibung der Wasserstoffbrückenbindungen gegeben waren. Der Punkt Z (für Zentrum) liegt auf der halben Strecke zwischen zwei sequentiellen C^{α}-Atomen und repräsentiert die O=C-N-H-Einheit bzw. die Amidbindung zwischen zwei Aminosäuren. Die Beschreibung der Wasserstoffbrücken über diesen Punkt wird in Abschnitt 4.4.6 dargelegt.
Dieses hier gewählte Proteinmodell ist insgesamt in Abb. 4.2 veranschaulicht. Um ein glattes Potential definieren zu können, wurden kontinuierliche Koordinaten verwendet bzw. die Koordinaten wurden im Rahmen der Maschinenpräzision angegeben und diese in Einheiten von Ångström verwendet[1].

Die Rückgratgeometrie wurde über die Positionen der C^{α}-Atome beschrieben. Ein Ziel des gewählten Modells war es, eine Darstellung der Geometrie zu erhalten, die nach Möglichkeit über eine einfache Vektormathematik behandelbar ist, um so eine umständlichere Umrechnung in interne Koordinaten und Schwierigkeiten in der lokalen Strukturoptimierung mit Gradienteninformationen zu vermeiden. So bilden die dreidimensionalen Vektoren $\mathbf{b}_i$ zwischen zwei sequentiellen C^{α}-Atomen i und $i-1$ die Grundlage zur Beschreibung eines Proteins. Diese werden auch als Bindungsvektoren bezeichnet. Sei der Ortsvektor des C_i^{α}-Atoms gegeben mit $\mathbf{x}_i$, so ist der Bindungsvektor $\mathbf{b}_i = \mathbf{x}_i - \mathbf{x}_{i-1}$. Setzt man für das N-terminale C^{α}-Atom ($\mathbf{x}_1$) eine beliebige Position an, ist das gesamte Rückgrat des Proteins über die Vektoren $\mathbf{b}$ bestimmt. Die internen Koordinaten können über diese vektoriellen Größen folgendermaßen ermittelt werden [76, 98]: Der Abstand r_{ik} zweier C^{α}-Atome i und k ist gegeben durch

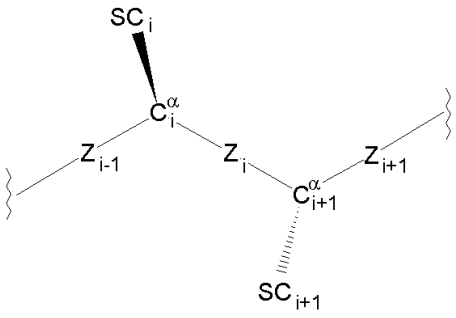


Abbildung 4.2: Proteinmodell mit drei Wechselwirkungszentren: C^{α}-Atom, Z-Punkt für die Wasserstoffbrücken und die Seitenkette SC.

$$r_{ik} = \sqrt{\langle \mathbf{x}_i - \mathbf{x}_k, \mathbf{x}_i - \mathbf{x}_k \rangle} \tag{4.1}$$

mit dem inneren Produkt $\langle \mathbf{x}, \mathbf{x}' \rangle$ zweier Vektoren (siehe Übersicht Seite 1). Im Programm wird zumeist der quadratische Abstand r_{ik}^2 verwendet, um das Wurzelziehen zu vermeiden. Alternativ zu Gl. 4.1 lässt sich der quadratische Abstand auch über

$$r_{ik}^2 = \langle \mathbf{x}_i, \mathbf{x}_i \rangle + \langle \mathbf{x}_k, \mathbf{x}_k \rangle - 2 \langle \mathbf{x}_i, \mathbf{x}_k \rangle \tag{4.2}$$

[1]Programmintern erfolgte eine Koordinatendarstellung in Einheiten von 0.1 Ångström (= 1 Å'), so dass beispielsweise zwei sequentielle C^{α}-Atome bei einer *trans*-Stellung des Rückgrat-Torsionswinkels ω einen Abstand von ca. 38 Å' besitzen.

ermitteln. Der Bindungswinkel κ_i zwischen drei sequentiellen C^α-Atomen $i, i+1$ und $i+2$ ist

$$\kappa_i = \frac{\langle \mathbf{b}_{i+1}, \mathbf{b}_{i+2} \rangle}{\|\mathbf{b}_{i+1}\| \|\mathbf{b}_{i+2}\|} \tag{4.3}$$

und der Torsionswinkel τ_i zwischen vier sequentiellen C^α-Atomen $i, i+1, i+2$ und $i+3$ ergibt sich aus

$$\tau_i = \mathrm{sgn}\Big(\langle \mathbf{b}_{i+3} \times \mathbf{b}_{i+2}, \mathbf{b}_{i+1} \rangle \Big) \left| \frac{\langle \mathbf{b}_{i+1} \times \mathbf{b}_{i+2}, \mathbf{b}_{i+2} \times \mathbf{b}_{i+3} \rangle}{\|\mathbf{b}_{i+1} \times \mathbf{b}_{i+2}\| \|\mathbf{b}_{i+2} \times \mathbf{b}_{i+3}\|} \right| \tag{4.4}$$

wobei $\mathbf{b} \times \mathbf{b}'$ das äußere Produkt (siehe Übersicht Seite 1) und "sgn" die Signumfunktion ist. Das Vorzeichen des Torsionswinkels τ_i wird somit durch das Vorzeichen des inneren Produktes

$$\langle \mathbf{b}_{i+3} \times \mathbf{b}_{i+2}, \mathbf{b}_{i+1} \rangle \tag{4.5}$$

festgelegt. Wie besonders aus den Gl. 4.3 bis 4.5 hervorgeht, sind für eine Umrechnung der kartesischen Ortsvektoren $\mathbf{x}_i$ in interne Koordinaten eine Anzahl an Rechenschritten erforderlich. Um diese Anzahl möglichst klein zu halten und damit die Berechnung des Potentials möglichst effizient zu gestalten, war ein Ziel der Definition des Kraftfeldes, dieses im wesentlichen nur von Vektoren direkt bzw. von inneren Produkten der Vektoren abhängig zu machen. Für die Berechnung des Potentials konnte dies realisiert werden, so dass dieses auf einfachen Abständen bzw. inneren Produkten beruht. In der Implementation des genetischen Algorithmus' zeigten sich jedoch Schwierigkeiten, wenn explizite Rückgratgeometrien bzw. bestimmte Sekundärstrukturen vektoriell erzeugt werden sollten. Besonders problematisch war hierbei die Erzeugung von α-Helices, da hierzu eine Unterscheidung zwischen der rechtsgängigen und der linksgängigen α-Helix unabdingbar ist, da praktisch nur die rechtsgängige Helix von Bedeutung ist. Die Unterscheidung beider auf Basis von Vektoren war in Bezug auf die Anzahl an notwendigen Rechenschritten so aufwendig, dass dies äquivalent mit der Berechnung der internen Koordinaten war, so dass im genetischen Algorithmus dennoch auf diese zurückgegriffen wurde.

Die weiteren Punkte, die der Beschreibung einer Proteingeometrie dienten, waren das bereits oben erwähnte Zentrum Z bzw. der zugehörige Ortsvektor $\mathbf{z}_i$, der sich aus $\mathbf{z}_i = \frac{1}{2}(\mathbf{x}_i + \mathbf{x}_{i+1})$ ergibt, sowie der Seitenkettenschwerpunkt, dessen Ortsvektor mit $\boldsymbol{\rho}_i$ bezeichnet wird. Der Verbindungs- bzw. Richtungsvektor $\mathbf{q}_i$ zwischen dem C_i^α-Atom und der zugehörigen Seitenkette am Ort $\boldsymbol{\rho}_i$ ist gegeben durch $\mathbf{q}_i = \boldsymbol{\rho}_i - \mathbf{x}_i$. Im folgenden werden die entwickelten Modelle zur Positionierung der Seitenketten-Schwerpunkte zu einer gegebenen Rückgratgeometrie beschrieben.

4.2 Seitenkettenapproximation

Sowohl auf den Faltungsprozess wie auch auf die Organisation des nativen Zustandes haben die Seitenketten einen sehr wichtigen Einfluss. Da das Rückgrat in jeder Aminosäure stets die gleiche chemische Zusammensetzung hat, unterscheiden sich die Aminosäuren anhand der physikochemischen Eigenschaften der Seitenketten, welche unterschiedliche funktionelle Gruppen enthalten können. Diese bestimmen die von einer Aminosäure präferierte Umgebung, und wie verschiedene Aminosäuren bzw. Seitenketten miteinander wechselwirken können.
Zunächst einmal wird durch die Seitenkette der Zugang zum Rückgrat beispielsweise zur Ausbildung von Wasserstoffbrücken durch sterische Hinderung erschwert bzw. auf bestimmte Raumbereiche außerhalb des ausgeschlossenen Volumens der Seitenkette beschränkt. Andererseits treten Seitenketten über polare bzw. apolare Gruppen miteinander oder auch mit dem Rückgrat in Wechselwirkungen, woraus unter anderem der hydrophobe Effekt entsteht, der dafür sorgt, dass hydrophobe Seitenketten eine kompakte Packung im inneren des Proteins anstreben, wozu sich im Gegensatz die geladenen Aminosäuren eher an der Oberfläche befinden und vom Lösungsmittel solvatisiert werden. Wichtig für die Struktur von vielen Proteinen ist auch die Ausbildung von Disulfidbrücken zwischen zwei Cystein-Seitenketten. Daneben sind an molekularen Erkennungsprozessen, Protein-Protein-Wechselwirkungen oder an den katalytischen Reaktionen von Enzymen ebenfalls im Regelfall bestimmte Seitenketten beteiligt, hier beispielsweise durch Koordination an Metallionen oder durch direkte Beteiligung von Heteroatomen an Reaktionen.
Ein anderer Aspekt, auf den hier nicht näher eingegangen wird, ist die thermodynamische Bilanz des Faltungsprozesses, in die die Seitenketten neben dem Enthalpiebeitrag auch durch den Entropiebeitrag entscheidend mit einfließen, da diese im Faltungsprozess, wie auch das Rückgrat, auf sehr wenige zugängliche Zustände beschränkt werden, was einer Abnahme der Entropie entspricht.

Aus diesen verschiedenen Gründen ist es sinnvoll, wenn nicht unerlässlich, für ein akkurates Proteinmodell die Seitenketten mit einzubeziehen. Im folgenden werden kurz zwei Modelle vorgestellt, wie Seitenketten in der Literatur in vergröberten Darstellungen behandelt werden, die einen zu dieser Arbeit ähnlichen Ansatz zu Grunde gelegt haben.
Zunächst sei der Ansatz von H. A. Scheraga *et al.* des UNRES-Kraftfeldes erwähnt, in welchem der Seitenkettenschwerpunkt zur Repräsentation verwendet wird [99–101]. Dieser wird als Ellipse mit aminosäureabhängigen Dimensionen beschrieben. Für die Position des Schwerpunktes werden insgesamt ein Abstand, der für jede Seitenkette fest ist, und zwei Winkel benötigt, wobei dies zum einen der Winkel des Seitenkettenvektors zur Winkelhalbierenden

des Dreiecks C^{α}_{i-1},C^{α}_{i},C^{α}_{i+1} und zum anderen der Rotationswinkel um diese Winkelhalbierende ist. Hierauf basierend wurden mehrere Gauss-Funktionen definiert, die direkt abhängig von diesen drei Winkeln waren. Die zugehörigen Parameter wie Gewichtungsfaktor, Breite und Höhe der Gauss-Funktionen wurden an statistische Verteilungen der Seitenketten aus Datenbankproteinen angepasst, mit denen schließlich die potentielle Energie der Seitenkette bezüglich des Rückgrates genähert wurde. D. h., dass die i-te Seitenkettenposition in diesem Modell zum einen von der Rückgratgeometrie am C^{α}_{i}-Atom abhängt, zum anderen aber auch, dass die Seitenkette im Rahmen der Gauss-Funktionen beweglich ist und sich der Umgebung anpassen kann.
Das zweite als Kontrast hier angeführte Beispielmodell vereinfacht die Darstellung der Seitenkette im Gegensatz zur flexiblen Positionierung des UNRES-Ansatzes sehr stark. In diesem, von M. Levitt *et al.* verwendeten Modell, wurde ebenfalls lediglich ein Punkt für jede Seitenkette angesetzt [102]. Dies war das C^{β}-Atom oder der Seitenkettenschwerpunkt, abhängig vom Typ der Aminosäure. Eine Seitenkette wird in diesem Modell so positioniert, dass zunächst zwei Einheitsvektoren durch die Vektoren zwischen den C^{α}-Atomen $i-1, i$ und $i, i+1$ berechnet werden, die ein lokales Koordinatensystem für die Seitenkette bilden. Über eine feste Linearkombination dieser Vektoren für jede Aminosäure wurden entweder bei kurzen Seitenketten die Koordinaten des C^{β}-Atoms oder für längere die des Seitenkettenschwerpunktes berechnet. Die Linearkombination für das C^{β}-Atom wurde aber so gewählt, dass dessen Position nicht mit der des realen C^{β}-Atoms übereinstimmte, sondern so, dass möglichst gute Resultate für das Kraftfeld erhalten wurden. Dieses Modell unterscheidet sich vom ersteren dahingehend, dass die Position der Seitenkette nur von der Rückgratgeometrie abhängt und nicht beweglich ist, was eine sehr starke Vereinfachung darstellt. Außerdem werden die Seitenketten als kugelförmig und nicht ellipsoidal angenommen.
Gemeinsam an diesen Ansätzen ist, dass die Eigenschaft der Seitenketten verwendet wird, dass sie sich relativ zum Rückgrat in bestimmten Raumbereichen statistisch häufiger auftritt, was bestimmten bevorzugten Rotamerzuständen der Seitenkette entspricht, wie bspw. der *all-trans*-Konformation einer Lysinseitenkette. Diese Eigenschaft wurde auch für diese Arbeit verwendet, wobei mehrere Seitenkettenmodelle entwickelt wurden, die im folgenden beschrieben werden. In diesem Abschnitt werden die Methoden dargestellt, mit welchen die Seitenkettenpositionen in Bezug zum Rückgrat bestimmt wurden. Die Wechselwirkungsterme der Seitenketten für das Potential werden dagegen in Abschnitt 4.4.4 beschrieben.

Vor der Beschreibung der Seitenkettenmodelle seien an dieser Stelle vorweg einige Bezeichnungshinweise bzw. -definitionen zum besseren Verständnis des Textes erläutert:
Im folgenden sei $\mathbf{S}$ der Sequenzvektor eines Proteins mit N Aminosäuren. Die i-te Komponente dieses Vektors sei mit $s_i \in \{1, 2, ..., 20\}$ bezeichnet und gibt gemäß Spalte 1 in Tab. 1.1 die Aminosäuren durch eine Zahlencodierung an, so dass $\mathbf{S} = (s_1, s_2, ..., s_N)$. Der Sequenzvektor

$\mathbf{S}$ ist spezifisch für ein Protein P, so dass $\mathbf{S} = \mathbf{S}_P$ gilt bzw. $\mathbf{S}_P = (s_{1,P}, s_{2,P}, ..., s_{N(P),P})$. Zur einfacheren Darstellung wird auf die P-abhängige Schreibweise verzichtet.
Des weiteren wird im folgenden allgemeiner Bezug auf die zwanzig Aminosäuren genommen. Der Typ einer Aminosäure, unabhängig von einem Protein und Sequenzposition, sei hier mit $a \in \{1, 2, ..., 20\}$ bezeichnet.

4.2.1 Potentialmethode

Das erste verwendete Modell für die Positionierung der Seitenkette wurde so konzipiert, dass die Lage der Seitenkette relativ zum zugehörigen C^{α}-Atom bzw. zum lokalen Rückgrat als Terme in das Gesamtpotential mit eingehen. Das Ziel war es, anhand statistischer Daten charakteristische Orientierungen der Seitenketten in nativen Proteinen bezüglich eines Koordinatensystems zu bestimmen, um mit Hilfe dieser ein Potential zu formulieren, so dass Seitenkettenpositionen in Nähe zu nativen Positionen einen niedrigeren Energiebeitrag liefern als Seitenketten auf nicht-nativen Positionen, und so, dass im Schritt einer lokalen Optimierung, die Seitenketten zu diesen nativen Orientierungen getrieben werden.
Das zugrundeliegende gewählte Koordinatensystem basierte auf den Bindungsvektoren $\mathbf{b}_i$ und $\mathbf{b}_{i+1}$ für die jeweilige Seitenkette an der Sequenzposition i, ähnlich zum Ansatz von [102]. Unter Verwendung dieser beiden Vektoren wurde ein nahezu orthogonales (kartesisches) Koordinatensystem definiert, in dessen Raum die Positionen der Seitenketten als Linearkombination der Koordinatenachsenvektoren beschrieben wurden. Die Vektoren, die dieses Koordinatensystem aufspannten, waren zum einen die beiden Vektorsummen $\mathbf{b}_i + \mathbf{b}_{i+1}$ und $\mathbf{b}_i$ - $\mathbf{b}_{i+1}$ der beiden Bindungsvektoren. Diese beiden Vektoren sowie die Bindungsvektoren sind in Abb. 4.3 in ihrer gemeinsamen Ebene dargestellt. $\mathbf{b}_i + \mathbf{b}_{i+1}$ und $\mathbf{b}_i$ - $\mathbf{b}_{i+1}$ sind nahezu orthogonal zueinander, da die beiden Bindungsvektoren $\mathbf{b}_i$ und $\mathbf{b}_{i+1}$ in nativen Proteinen fast ausschließlich die einheitliche Länge von ca. 3.8 Å besitzen. Ausnahmen sind hierbei *cis*-Bindungen, die aber nur in sehr selten Fällen auftreten (siehe hierzu auch Abb. 4.19). Daher ist in Abb. Abb. 4.3 b) die Verteilung des Winkels α der Vektoren $\mathbf{b}_i + \mathbf{b}_{i+1}$ und $\mathbf{b}_i$ - $\mathbf{b}_{i+1}$ dargestellt, die aus der Auswertung der PDB-Strukturen des TOP500H-Proteinsatzes erhalten wurde, welche verdeutlicht, dass beide Vektoren nahezu orthogonal zueinander sind.
Da der Seitenkettenschwerpunkt nur in seltenen Fällen gerade in der durch diese beiden Vektoren aufgespannten Ebene liegt, wurde zusätzlich als dritter Vektor für das Koordinatensystem der Vektor senkrecht zu der durch die Atome C^{α}_{i-1},C^{α}_{i} und C^{α}_{i+1} aufgespannten Ebene verwendet, der sich aus dem äußeren Produkt $\mathbf{b}_i \times \mathbf{b}_{i+1}$ ergibt.

Basierend auf diesen Koordinatenvektoren wurde nun untersucht, ob die Projektionen der Seitenkettenvektoren $\mathbf{q}_i$, welche vom C^{α}_{i}-Atom zu dessen Seitenkettenschwerpunkt zeigen,

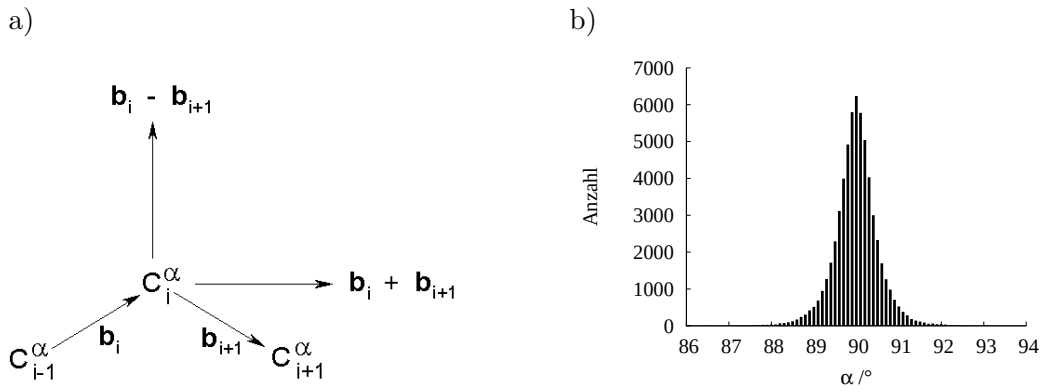


Abbildung 4.3: a) Basisvektoren zur Beschreibung der Seitenketten. b) Statistik des Winkels α zwischen $\mathbf{b}_i + \mathbf{b}_{i+1}$ und $\mathbf{b}_i$ - $\mathbf{b}_{i+1}$.

bestimmte charakteristische Häufungen enthalten. Angestrebt wurde, im Falle von eindeutigen Mustern in den Verteilungen, die Maxima derselben als Minima für Potentialfunktionen zur Positionierung der Seitenkette zu verwenden. Hierzu wurden die 273 Proteine des reduzierten TOP500H-Proteindatensatzes verwendet und für alle Seitenketten die folgenden inneren Produkte $\chi_{i,j}$ ausgewertet, wobei i die Position entlang der Sequenz und j die Ordnungsnummer des inneren Produktes angibt: $\chi_{i,1} = \langle \mathbf{q}_i, \mathbf{b}_i \rangle$, $\chi_{i,2} = \langle \mathbf{q}_i, \mathbf{b}_{i+1} \rangle$, $\chi_{i,3} = \langle \mathbf{q}_i, \mathbf{b}_i \times \mathbf{b}_{i+1} \rangle$, $\chi_{i,4} = \langle \mathbf{q}_i, \mathbf{b}_{i+1} - \mathbf{b}_i \rangle$ und $\chi_{i,5} = \langle \mathbf{q}_i, \mathbf{b}_i + \mathbf{b}_{i+1} \rangle$. Die inneren Produkte der beiden Bindungsvektoren $\mathbf{b}_i$ und $\mathbf{b}_{i+1}$ wurden zum Vergleich mit ausgewertet, aber nicht zur Definition der Seitenkettenposition herangezogen. Die anhand dieser Funktionen in Intervallen bestimmten Werte wurden jeweils den zwanzig Aminosäuretypen a zugeordnet. Die resultierenden Statistiken sind in den Abb. 4.6 bis 4.10 dargestellt. Diese Statistiken wurden dahingehend ausgewertet, die Abszissenwerte der Maxima zu jeder Aminosäure und jedem inneren Produkt zu bestimmen und diese als Minima für die Potentialfunktion zu verwenden. Für ein so erhaltenes Minimum wurde damit die Funktion

$$V_{i,j}^{(scp)}(\chi_{i,j}) = \left[\chi_{i,j} - \gamma_{j,s_i}\right]^2 \tag{4.6}$$

angesetzt, wobei γ_{j,s_i} das Minimum dieser Funktion ist, das jeweils von der Funktionsnummer j und von einem der zwanzig Aminosäuretypen s_i an der Sequenzposition i abhängt. (scp) steht für *side chain potential*. Diese Funktion beschreibt eine quadratische Abhängigkeit von der Differenz zum statistisch erwarteten Wert des inneren Produktes. Diese Formulierung enthält sowohl Winkel- wie auch Längeninformationen der Vektoren. Da aber, wie aus den den Statistiken in den Abb. 4.6 bis 4.10 ersichtlich ist, für ein inneres Produkt zu einer Aminosäure häufig mehrere Maxima auftreten, wurde aus diesem Grund Gl. 4.6 in ein Produkt über alle bestimmten Minima überführt:

$$V_{i,j}^{(scp)}(\chi_{i,j}) = \prod_{k \in \mathbf{M}_{j,s_i}} \left[\chi_{i,j} - \gamma_{j,s_i,k}\right]^2 \tag{4.7}$$

Die Menge $\mathbf{M}_{j,s_i}$ enthält die Ordnungsnummern der Minima der j-ten Funktion und des Aminosäuretyps s_i und besitzt so viele Elemente wie der Statistik der Funktion $\chi_{i,j}$ zu dem Aminosäuretyp $a = s_i$ Maxima zugeordnet wurden. Das Produkt wird also über alle Minima berechnet. Dementsprechend hängt nun auch das Minimum $\gamma_{j,s_i,k}$ von der Ordnungsnummer des Faktors k ab. Für eine native Seitenkettengeometrie nahe eines Minimums γ geht der entsprechende Term des Produktes gegen Null und damit das gesamte Produkt.
Um nun das Gesamtpotential zur Positionierung einer Seitenkette i zu berechnen, muss zusätzlich über alle aktiven (ausgewählten) inneren Produkte summiert werden. Gegeben sei hier die Menge $\mathbf{A}$, die die Ordnungsnummern dieser aktiven Funktionen enthält, so dass $j \in \mathbf{A}$ gilt. Die Summation führt zu folgender Gleichung für die Seitenkette i mit dem Positionierungspotential $E_i^{(scp)}$:

$$E_i^{(scp)} = \sum_{j \in \mathbf{A}} V_{i,j}^{(scp)}(\chi_{i,j}) = \sum_{j \in \mathbf{A}} \prod_{k \in \mathbf{M}_{j,s_i}} \left[\chi_{i,j} - \gamma_{j,s_i,k}\right]^2 \tag{4.8}$$

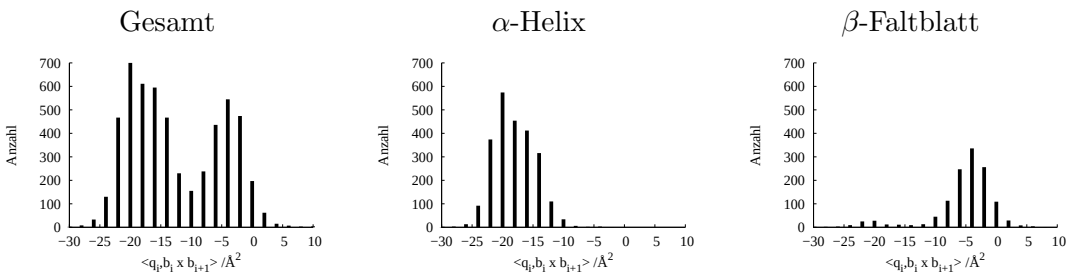
Die gesamte Seitenkettenpositionierungsenergie eines Proteins ergibt sich aus der Summe über alle N Aminosäuren $\sum_{i=1}^{N} E_i^{(scp)}$.
Die erhaltenen Statistiken zur Bestimmung der Minima $\gamma_{j,s_i,k}$ in Gl. 4.8, welche im folgenden näher beschrieben werden, fallen je nach Aminosäure und funktionaler Form des inneren Produktes $\chi_{i,j}$ stark unterschiedlich aus.
Zunächst lassen sich aus den Statistiken allgemeine geometrische Angaben zu den Seitenketten ableiten.
Die Projektion $\langle \mathbf{q}_i, \mathbf{b}_i - \mathbf{b}_{i+1} \rangle$ zeigt fast durchweg nur positive Vorzeichen. Die Seitenketten liegen also vom C_i^α aus betrachtet in Richtung dieses Vektors. Bezieht man dies direkt auf Abb. 4.3 a), so liegen die Seitenketten oberhalb des dort gezeigten $\mathbf{b}_i + \mathbf{b}_{i+1}$-Vektors, mit dem C_i^α-Atom als Startpunkt.
Die Häufigkeiten zu $\langle \mathbf{q}_i, \mathbf{b}_i + \mathbf{b}_{i+1} \rangle$ zeigen unterschiedliche Charakteristika. Für die kurzen Seitenketten wie Alanin, Isoleucin, Threonin und Valin sind die Verteilungen um 0 Å^2 zentriert. Deren Seitenkettenvektoren $\mathbf{q}_i$ liegen also nahezu orthogonal zu $\mathbf{b}_i + \mathbf{b}_{i+1}$, da deren Schwerpunkte nahe dem C^β-Atom liegen. Viele andere Aminosäuren wie z. B. Phenylalanin, Histidin, Lysin, Leucin, Tryptophan oder Tyrosin zeigen Verteilungen, die eine gewisse Symmetrie bezüglich des Nullpunktes zeigen, wobei am Nullpunkt eher ein Minimum erreicht wird. Dies bedeutet, dass diese Seitenkettenvektoren entweder eine eher parallele oder antiparallele Ausrichtung in Bezug auf den Vektor $\mathbf{b}_i + \mathbf{b}_{i+1}$ haben. Diese Häufungen beruhen auf bevorzugten Rotamerzuständen der Seitenketten, was besonders bei Seitenketten mit Ringsystemen ausgeprägt ist. Aufgrund sterischer Wechselwirkungen beispielsweise bei der Rotation des Phenylringes um die C^α-C^β-Bindung gibt es bestimmte Vorzugsorientierungen (Dieses wird im Kapitel 4.2.3 näher erläutert). Diese nicht kontinuierliche Verteilung von Rotameren spiegelt sich in den statistischen Daten wieder.

Wie zu erwarten, streuen die Verteilungen für kurze Seitenketten wie z. B. Alanin, Prolin, Leucin oder Valin weniger als für die längeren Seitenketten, aufgrund geringerer Freiheitsgrade.
Häufig treten in allen Statistiken Verteilungen mit ein bis drei differenzierbaren Maxima auf. Zur Formulierung des Potentials basierend auf Funktionen wie in Gl. 4.6 kritische Verteilungen sind beispielsweise die für Phenylalanin, Histidin oder Tryptophan in Abb. 4.6, da hier die beiden Maxima stark unterschiedliche Formen besitzen. Die Verteilungsmaxima zwischen 10 und 15 Å sind relativ eng, während dagegen der Teil der Verteilung um bei 0 Å breit ist. Die Funktion aus Gl. 4.6 bzw. 4.7 für beliebig viele Minima beschreibt allerdings eine gleichmäßige Verteilung, so dass die Minima alle eine ähnliche Breite und die gleiche Tiefe besitzen. Dies kann nun in der Potentialfunktion dazu führen, dass entweder das breite Minimum zu schmal beschrieben wird, wenn der Gewichtungskoeffizient der Funktion groß ist oder dass andersherum das scharfe Maximum zu breit beschrieben wird, wenn der Gewichtungskoeffizient klein ist. Geometrisch kann dies dazu führen, dass Bereiche zur Seitenkettenpositionierung erlaubt werden, die nicht in nativen Proteinen vorkommen, oder dass Bereiche ausgeschlossen werden, die in nativen Protein zugänglich sind. Beide Fälle sind in einer Geometrieoptimierung nicht erwünschenswert, wobei für den Fall des zu großen zugänglichen Konformationsraumes noch die Möglichkeit besteht, dass dieser durch Wechselwirkungen mit anderen Atomen beschränkt und dadurch korrigiert wird.

Weiterhin treten bei einigen Aminosäuren sehr breite und asymmetrische Verteilungen auf. Besonders Arginin besitzt für alle inneren Produktfunktionen relativ zu den anderen Aminosäuren unspezifische Muster ohne markante Minima oder Maxima. Hier ist es besonders schwierig mittels quadratischer Funktionen eine angemessene Potentialform zu finden. Dies ist darin begründet, das Arginin die längste Seitenketten aller zwanzig natürlichen Aminosäuren besitzt und zudem mehrere polare funktionale Gruppen enthält, wodurch sie bevorzugt an der Oberfläche eines Proteins zu finden ist, wo die Rotationen der Seitenkette um die Bindungen weniger eingeschränkt ist als im inneren. Aber auch andere Aminosäuren wie z. B. Lysin, Tryptophan oder Tyrosin können je nach Funktion diffuse, unspezifische Verteilungen zeigen.

Eine Analyse der Verteilungen der inneren Produkte aufgeschlüsselt nach den Sekundärstrukturen, in denen die Seitenketten vorlagen, zeigte, dass eine Aufspaltung der Verteilung in mehrere Maxima nur in bestimmten Fällen auf unterschiedliche Sekundärstrukturen zurückgeführt werden kann. Dies soll am Beispiel der beiden Aminosäuren Leucin und Serin und zwei unterschiedlichen Produkt-Funktionen veranschaulicht werden. Verglichen wurden die Verteilungen jeweils zu den Funktionen $\langle \mathbf{q}_i, \mathbf{b}_i \times \mathbf{b}_{i+1} \rangle$ und $\langle \mathbf{q}_i, \mathbf{b}_i - \mathbf{b}_{i+1} \rangle$. Die entsprechenden Gesamtverteilungen und die nach den Sekundärstrukturen aufgeschlüsselten Vertei-

Leucin; Verteilung für $\langle \mathbf{q}_i, \mathbf{b}_i \times \mathbf{b}_{i+1} \rangle$.



Serin; Verteilung für $\langle \mathbf{q}_i, \mathbf{b}_i - \mathbf{b}_{i+1} \rangle$.

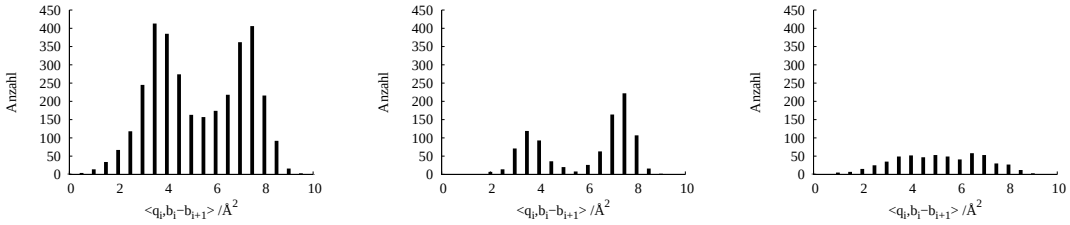


Abbildung 4.4: Unterschiedliche Verteilungen der Sekundärstrukturen für zwei Aminosäuren und zwei innere Produkte.

lungen, die durch das DSSP-Programm [29] zugeordnet wurden, sind in Abb. 4.4 dargestellt. In dieser zeigt sich für Leucin und das Produkt $\langle \mathbf{q}_i, \mathbf{b}_i \times \mathbf{b}_{i+1} \rangle$, dass zum Maximum bei -20 Å^2 Seitenketten mit helicaler Umgebung beitragen, aber kaum welche aus einer β-Struktur. Diese tragen dagegen zum Maximum bei ca. -5 Å^2 bei. Dieses Muster ist somit eine direkte Folge der unterschiedlichen Geometrie der α-Helix und des β-Faltblattes: In der Helix befinden sich die Seitenketten aufgrund sterischer Effekte größtenteils oberhalb der $\mathbf{b}_i$,$\mathbf{b}_{i+1}$-Ebene, also antiparallel zu $\mathbf{b}_i \times \mathbf{b}_{i+1}$, während sie in der gestreckten β-Konformation in dieser Ebene liegen, also annähernd orthogonal zum Ebenenvektor.

Betrachtet man dagegen Serin und die Projektion des Seitenkettenvektors auf $\mathbf{b}_i - \mathbf{b}_{i+1}$, so tragen hier beide Sekundärstrukturen zu jeweils beiden Maxima bei. Hier kann also für das Doppelmaximum nicht die gleiche Begründung wie zuvor für Leucin gegeben werden. Eine Begründung für diese Aufspaltung sind hier unterschiedliche Rotamerzustände der Seitenkette. Hierfür wurde der Torsionswinkel der Atome untersucht, der durch die Atome C^{α}_{i-1},C^{α}_{i}, C^{β}_{i} und durch das Sauerstoffatom der Hydroxygruppe der Serinseitenkette gegeben ist. Es zeigte sich hier, dass zu diesen Zuständen hauptsächlich zwei Torsionswinkel beitragen: Eine Orientierung der Seitenkette mit einem Torsionswinkel von ca. 0° und eine von ca. 90°. Eine andere Erklärung für die Entstehung unterschiedlicher Maxima kann hier dagegen ebenfalls durch einen intrinsischen Modelleffekt gegeben werden: Die Basisvektoren hängen unmittelbar von den beteiligten C^{α}-Koordinaten und damit vom C^{α}_{i-1},C^{α}_{i},C^{α}_{i+1}-Winkel ab. Ändert man diesen Winkel, führt dies zu einer Rotation des Koordinatensystem, bzw. zu einer Rotation der Basisvektoren $\mathbf{b}_i + \mathbf{b}_{i+1}$ und $\mathbf{b}_i - \mathbf{b}_{i+1}$ mit dem Vektor $\mathbf{b}_i \times \mathbf{b}_{i+1}$ als Drehachse. Hält man die Seitenkettenkoordinaten fest und ändert nur diesen Winkel von einem Wert, der

einer α-Konformation entspricht (ca. 90°) in eine β-Konformation (ca. 130°) und berechnet den Unterschied der Projektion des festen Seitenkettenvektors $\mathbf{q}_i$ auf $\mathbf{b}_i - \mathbf{b}_{i+1}$ vor und nach der Rotation, so beträgt der Unterschied ca. 4 $\mathrm{\AA}^2$, was der Differenz der beiden Maxima entspricht. Es ist aber unklar, inwieweit dieser Effekt tatsächlich zu der Statistik beiträgt, da mit einer Veränderung der Winkels sich auch die Position der Seitenkette verschieben kann. Problematisch ist die Interpretation hier, da durch das vergröberte Modell keine Informationen über die Rückgrattorsionswinkel vorliegen, da gänzlich verschiedene ϕ,ψ-Kombinationen den gleichen C^{α}_{i-1},C^{α}_{i},C^{α}_{i+1}-Winkel erzeugen können. Hinzukommt, dass die Seitenkettenposition aber ebenfalls von den expliziten Torsionswinkeln innerhalb der Seitenkette selber abhängt. Hierdurch ist schwierig zu unterscheiden, ob sich lediglich nur die Positionen der C^{α}-Atome geändert hat und mit diesen der Winkel, oder ob auch die Seitenkettenposition abweicht bzw. ob ein anderer Rotamerzustand zu Grunde liegt.

Trotz dieser vereinzelten Probleme zeigten jedoch viele Verteilungen der inneren Produkte ausgeprägte Muster, so dass dieser Ansatz implementiert wurde. Die benötigten Minima der Funktionen wurden entweder direkt aus den statistischen Daten abgelesen oder in weniger eindeutigen Fällen mittels eines Fits über normalverteilte Gauss-Funktionen $g^{(fit)}$ der Form

$$g^{(fit)}(\chi) = \exp\left(-A\Big(\frac{\chi - B}{C}\Big)^2\right) \tag{4.9}$$

bestimmt, wobei A, B und C die frei bestimmbaren Parameter sind und χ das innere Produkt. Der Parameter B enthält die Positionsangabe (Abszissenwert) für das Maximum. Der Fit wurde über den Marquardt-Levenberg-Algorithmus mit Hilfe des Programmes GNUPLOT durchgeführt [103].
Zusätzlich zu den bisher beschriebenen inneren Produkten wurde noch eine Abstandsfunktion in das Positionierungspotential aufgenommen, um den Seitenkettenschwerpunkt im richtigen Abstand zum C^{α}-Atom zu halten. Hierbei wurde wie bereits bei den anderen inneren Produkten verfahren. Die erhaltenen Verteilungen der Abstandsfunktion $\chi_{i,6} = \|\mathbf{q}_i\|$ ist in Abb. 4.5 dargestellt. Diese zeigten, dass für viele Aminosäuren der Abstand des Seitenkettenschwerpunktes zum C^{α}-Atom sehr konstant ist. Längere Seitenketten wie Glutamin, Lysin oder Arginin besaßen dagegen wiederum breitere Verteilungen, was unterschiedlichen Rotameren entspricht. Für diese Verteilungen wurde für jede Aminosäure ebenfalls eine Funktion Gl. 4.7 verwendet.

Zusammenfassend wurde mittels der Funktionen $\chi_{i,3} = \langle \mathbf{q}_i, \mathbf{b}_i \times \mathbf{b}_{i+1} \rangle$, $\chi_{i,4} = \langle \mathbf{q}_i, \mathbf{b}_{i+1} - \mathbf{b}_i \rangle$, $\chi_{i,5} = \langle \mathbf{q}_i, \mathbf{b}_i + \mathbf{b}_{i+1} \rangle$ und $\chi_{i,6} = \|\mathbf{q}_i\|$ die Seitenkettenpositionierung definiert, und die Gewichtungskoeffizienten dieser Funktionen zusammen mit den restlichen Potentialen optimiert. Nach der Optimierung gegen die falschen Proteinstrukturen zeigten sich bei der anschließenden Verwendung dieses Seitenkettenansatzes im genetischen Algorithmus allerdings Probleme, die darauf beruhten, dass die Gewichtungskoeffizienten der Seitenkettenpositionierungsfunktionen relativ zu den Koeffizienten der anderen Funktionen mit kleinen Werten optimiert wurden. Dies führte dazu, dass im Schritt der lokalen Kraftfeldoptimierung die anderen Potentiale, hierbei besonders die nicht-bindenden Wechselwirkungen, über die Seitenkettenpositionierung dominierten und die Seitenketten dadurch zu physikalisch unrealistischen Positionen verschoben werden konnte, beispielsweise zu einem zu großen Abstand vom zugehörigen C^α-Atom oder oder in falscher Relation zur C^α_{i-1},C^α_i,C^α_{i+1}-Ebene.
An dieser Stelle war unklar, ob dieser Effekt auf einer schlechten Wahl der Positionierungsfunktionen oder auf der Methode der Parameteroptimierung bzw. auf der Auswahl der falschen Strukturen beruhte. Da sehr viele unterschiedliche falsche Proteine, von sehr gering verzerrten über erfolgreich verwendete Literatur-Datensätze bis zu vom nativen Zustand sehr weit entfernte Strukturen verwendet worden waren, wurde diesem Aspekt weniger Relevanz zugemessen, ebenso wie der Parameteroptimierungsmethode, die ebenfalls von verschiedenen Autoren bereits erfolgreich angewendet wurde. Problematisch bei der Verwendung von solchen Potentialfunktionen kann die Kopplung aller Parameter im Kraftfeld sein, da alle gleichzeitig abhängig voneinander bestimmt werden. Um Parameter für die Seitenkettenpositionierung zu erhalten, müssen die Positionen der nativen Seitenketten bzw. der restlichen Umgebung verändert werden. Dies beeinflusst wiederum die anderen Parameter, beispielsweise die der nicht-bindenden Wechselwirkungen der Seitenketten untereinander, da sich durch eine Verschiebung der Seitenketten deren Anordnung ebenfalls verändert. Im Hinblick hierauf wurde dann in der Kraftfeldentwicklung mehr Gewicht auf eine korrekte Beschreibung der nichtbindenden Wechselwirkungen gelegt und weniger auf die Positionierung der Seitenketten. Aus diesen Gründen wurde dieses Modell nicht verwendet und auf einen expliziten Term zur Beschreibung der Seitenketten im Potential verzichtet. Stattdessen wurden, wie im folgenden beschrieben wird, andere Ansätze verfolgt, die Seitenketten zu positionieren, ohne dass diese Parameter in die Optimierung der Gewichtungskoeffizienten der Potentialterme mit einfließen.

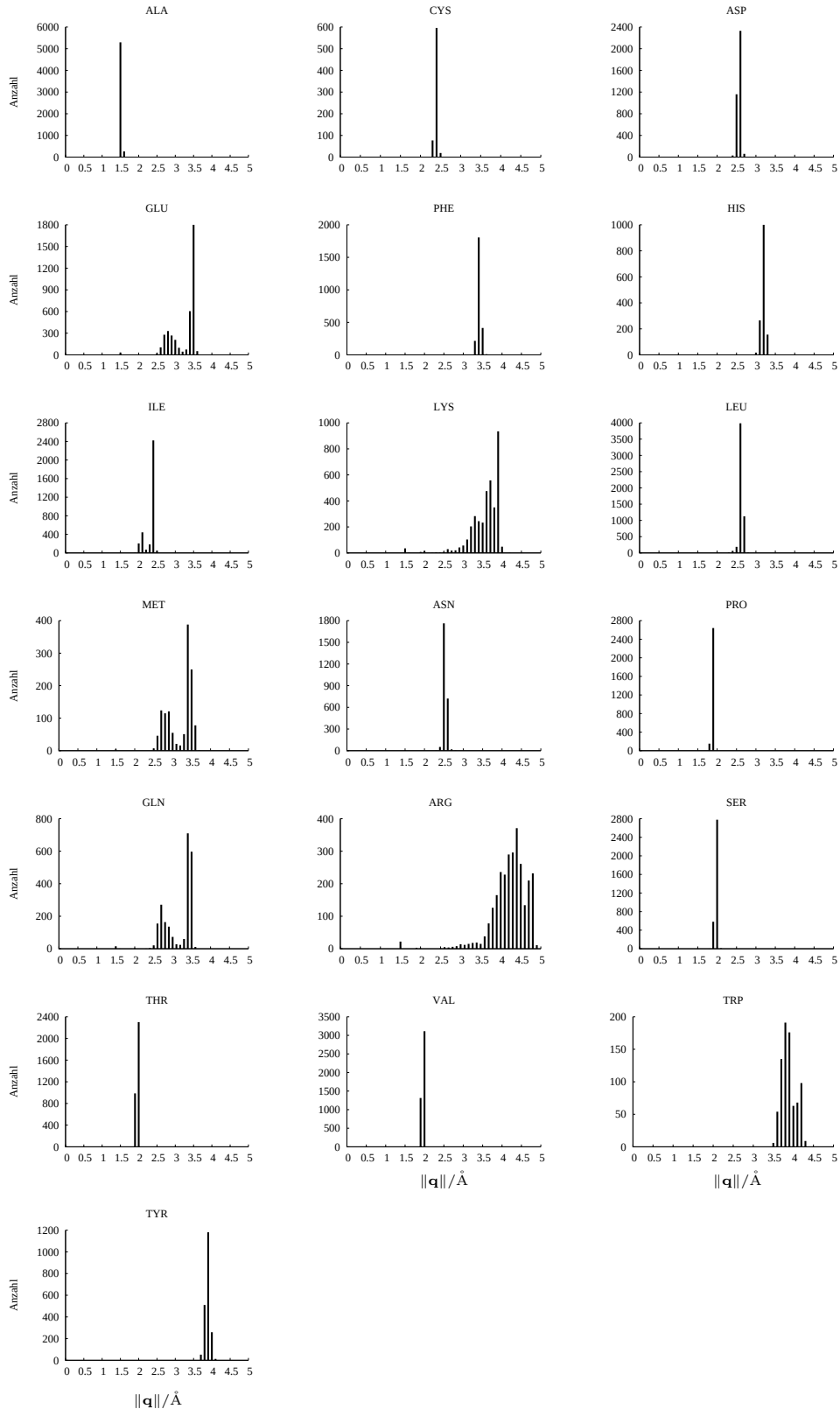


Abbildung 4.5: Länge $\|\mathbf{q}\|$ des Seitenkettenvektors. (Intervallbreite: 0.1 Å)

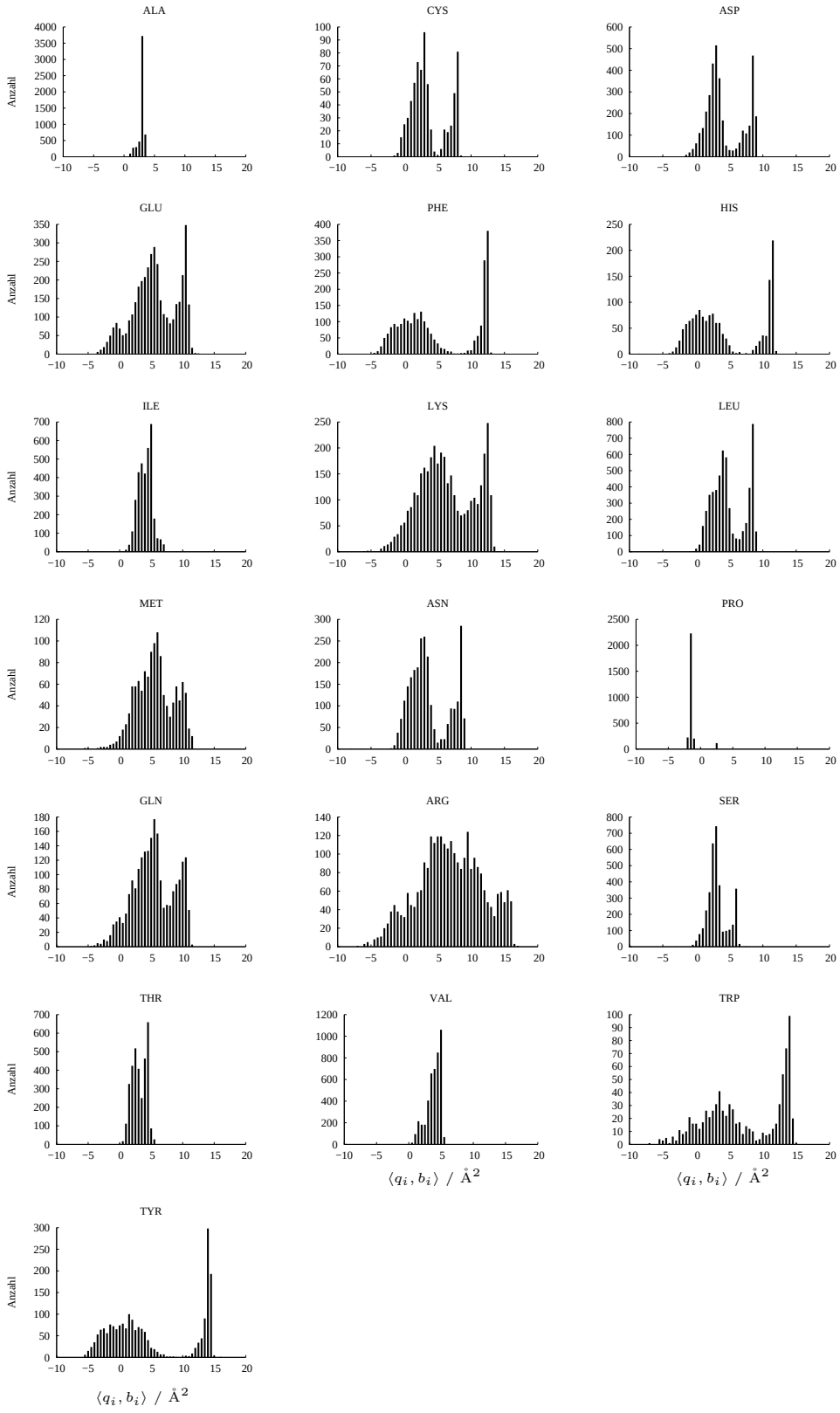


Abbildung 4.6: Projektion des Seitenkettenvektors $\mathbf{q}_i$ auf $\mathbf{b}_i$. (Intervallbreite: 0.5 Å^2)

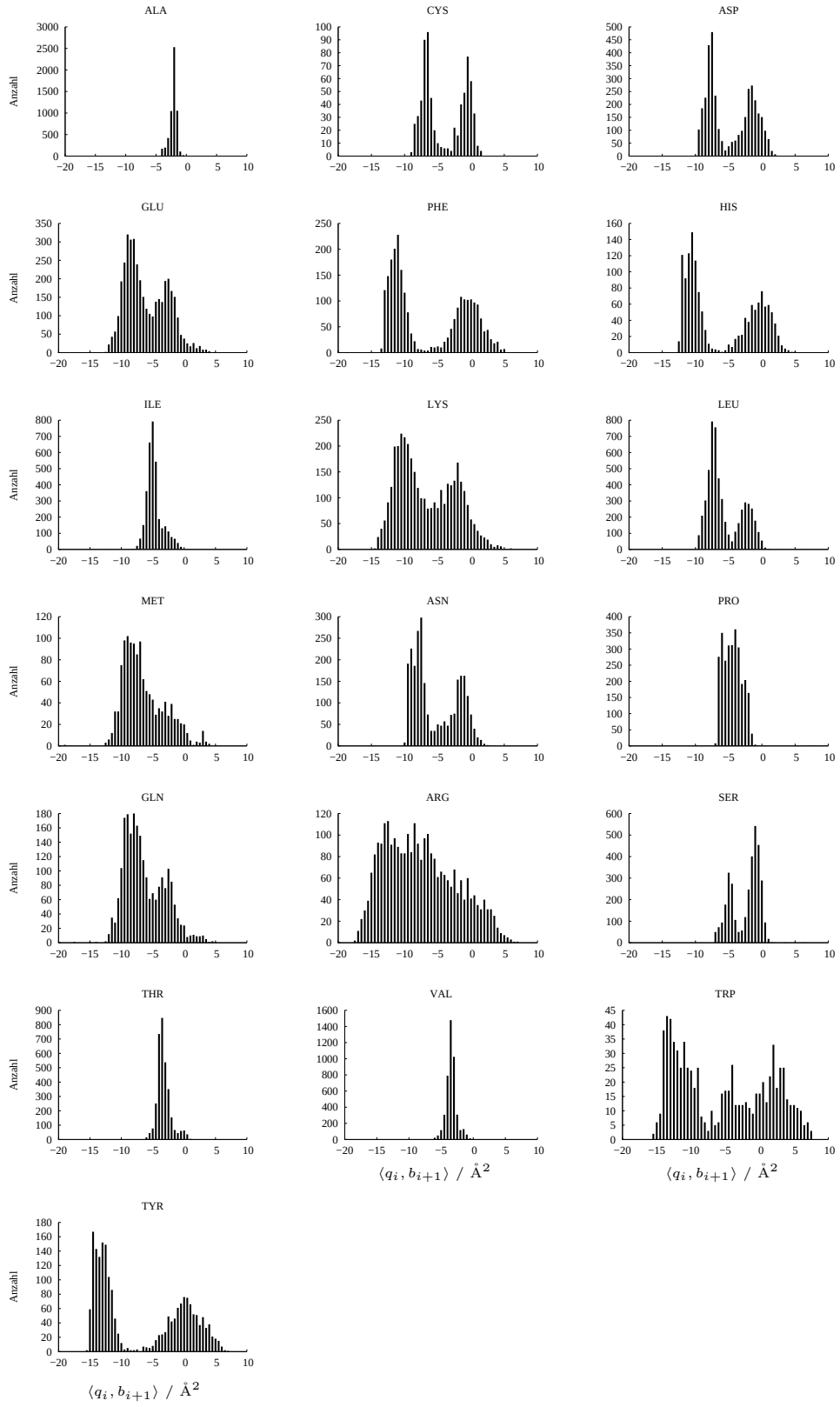


Abbildung 4.7: Projektion des Seitenkettenvektors $\mathbf{q}_i$ auf $\mathbf{b}_{i+1}$. (Intervallbreite: 0.5 Å^2)

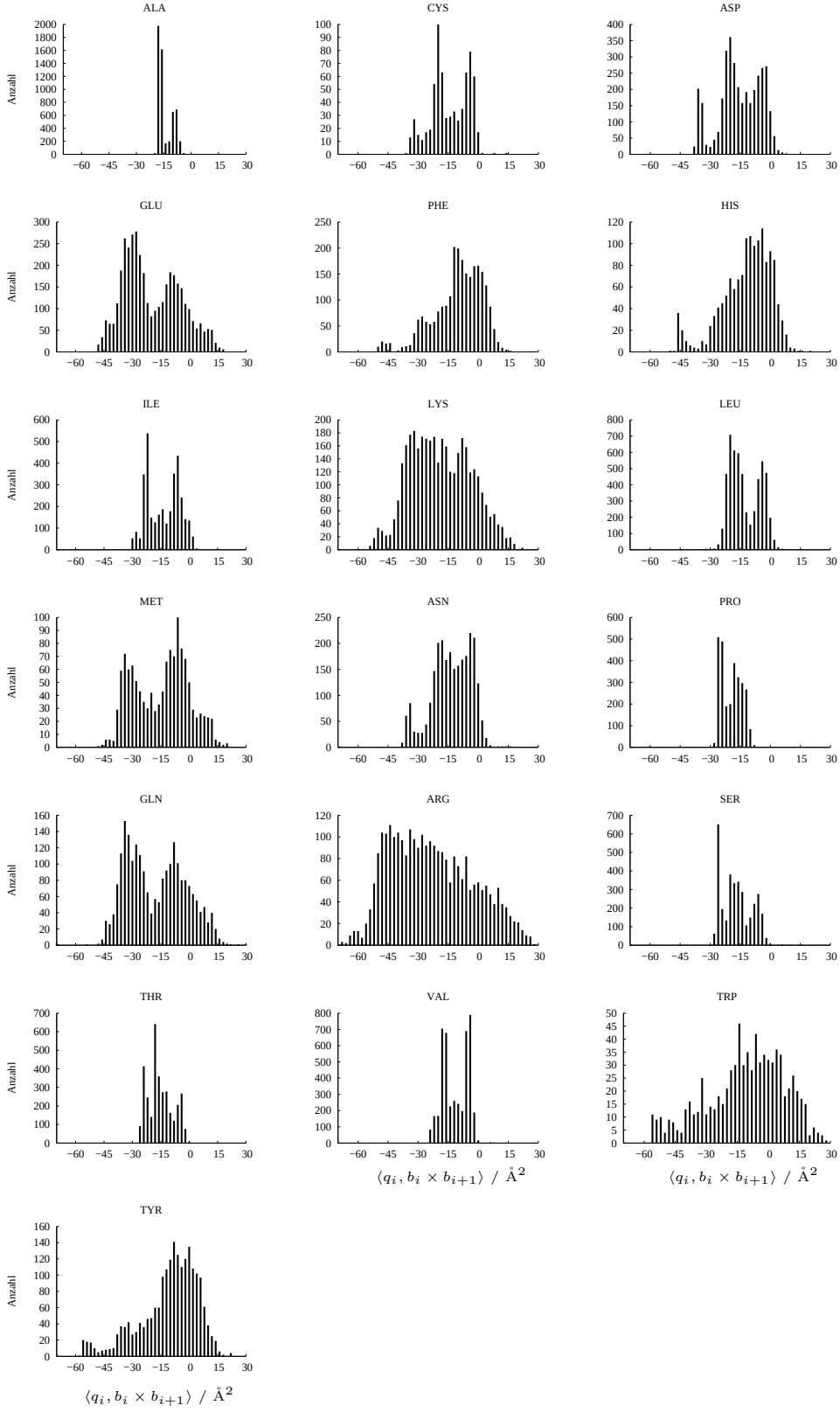


Abbildung 4.8: Projektion des Seitenkettenvektors $\mathbf{q}_i$ auf $\mathbf{b}_i \times \mathbf{b}_{i+1}$. (Intervallbreite: 2 Å^2)

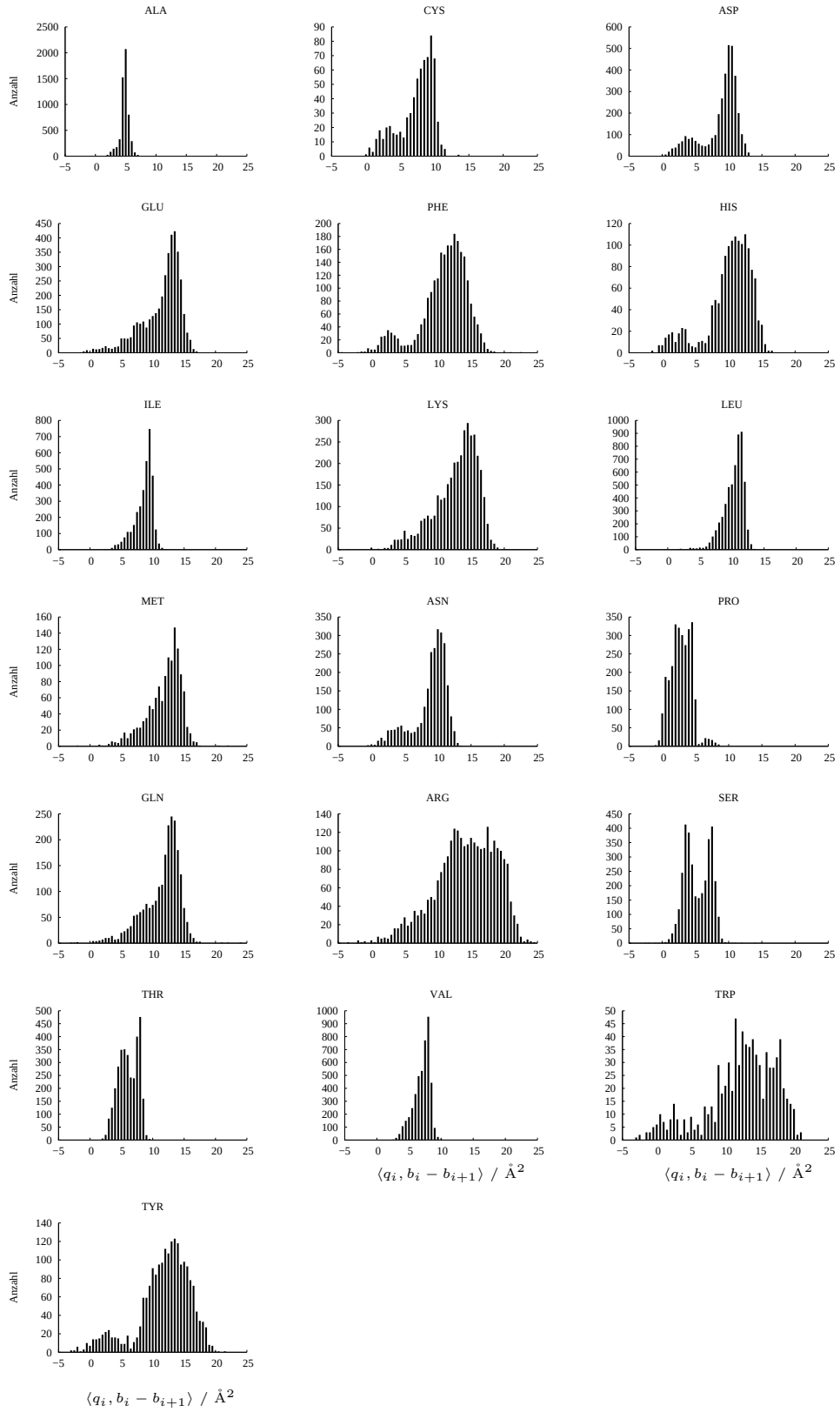


Abbildung 4.9: Projektion des Seitenkettenvektors $\mathbf{q}_i$ auf $\mathbf{b}_i$ - $\mathbf{b}_{i+1}$. (Intervallbreite: 0.5 Å^2)

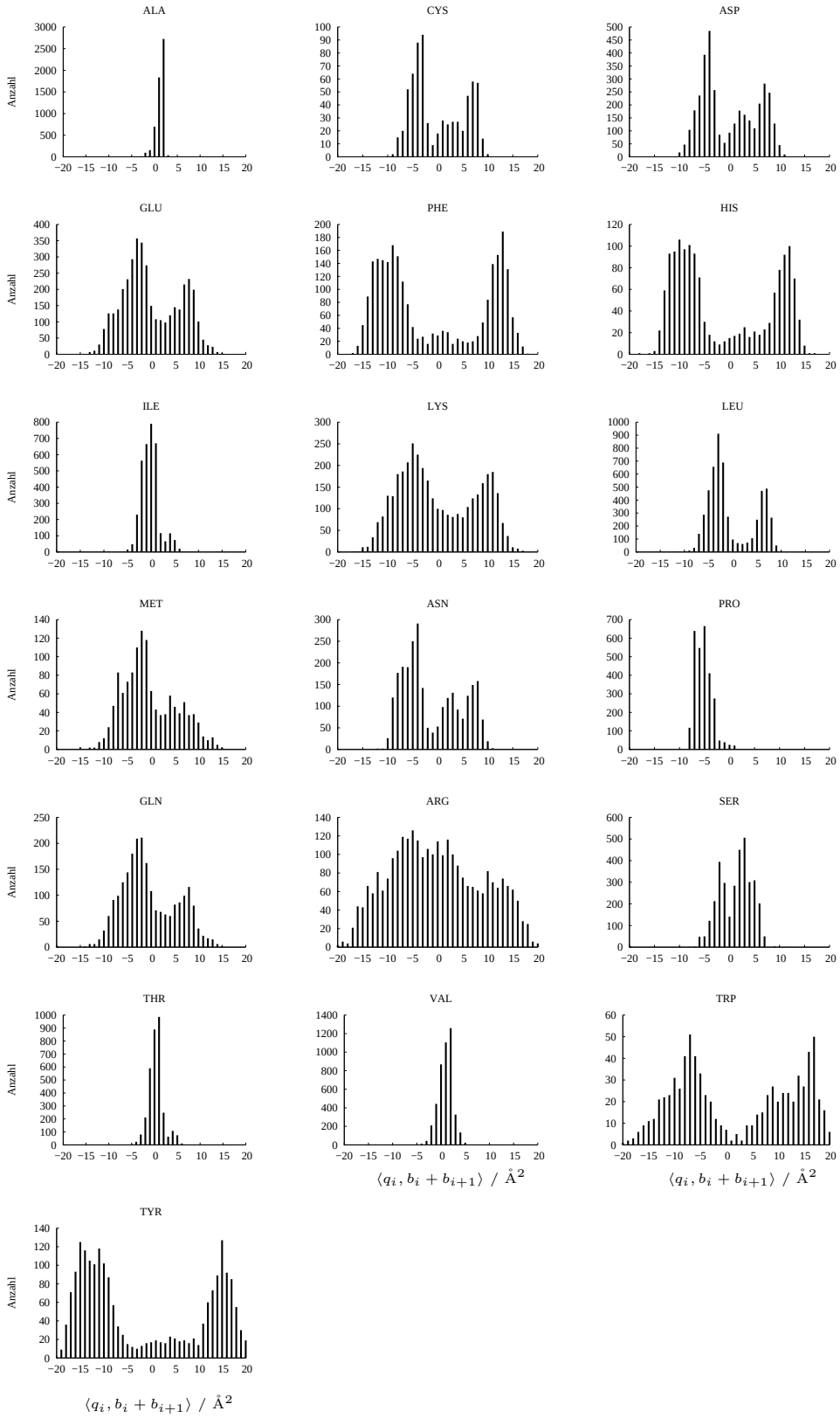


Abbildung 4.10: Projektion des Seitenkettenvektors $\mathbf{q}_i$ auf $\mathbf{b}_{i+1} + \mathbf{b}_i$. (Intervallbreite: 1 Å^2)

4.2.2 Fixierte Seitenkette

Nach dem ersten Ansatz zur Positionierung der Seitenkettenschwerpunkte mittels einer Potentialfunktion wurde untersucht, ob sich die Seitenketten hinreichend approximieren lassen, wenn die Rückgratkoordinaten in einer einzigen festen Linearkombination für jede Aminosäure verwendet werden. Hierzu wurde zunächst angenommen, dass sich die Position des Seitenkettenschwerpunktes aus den Rückgratkoordinaten ableiten lässt, indem bestimmte Rückgratgeometrien eine mehr oder weniger bestimmte Seitenkettenposition bedingen. Beispielsweise sind die Seitenketten in einer α-Helix radial zur Helixachse angeordnet, während sie in einer Faltblattstruktur ungefähr senkrecht zu dessen Ebene orientiert sind.
Dieser Ansatz basierte prinzipiell auf der Seitenkettenbeschreibung aus [102], in dem das C^{β}-Atom oder der Schwerpunkt der Seitenkette durch eine Linearkombination von Rückgratvektoren berechnet wird. In dieser Arbeit wird dieser Ansatz durch Hinzunahme weiterer Funktionen verallgemeinert und erweitert.
Das mathematische Modell zur Näherung der Seitenkettenposition ist gegeben durch die Gleichung:

$$\mathbf{q}_i = \sum_{k=1}^{n_f} c_{k,s_i}\, \mathbf{f}_k(\mathbf{X}) \tag{4.10}$$

Hier ist $\mathbf{q}_i$ der Vektor, der vom C_i^{α}-Atom der Aminosäure des Typs $s_i \in \{1, 2, ..., 20\}$ der i-ten Komponente des Sequenzvektors $\mathbf{S}$ zu der zugehörigen Seitenkette zeigt. Die c_{k,s_i} sind feste Koeffizienten, die nur von der Art der Aminosäure s_i und dem entsprechenden Summanden k abhängen. Die $\mathbf{f}_k$ sind verschiedene vektorwertige Funktionen, die von der Geometrie $\mathbf{X}$ des Rückgrates abhängen, und n_f ist die Gesamtanzahl an verwendeten Funktionen. Die Funktionen $\mathbf{f}_k$ wurden so gewählt, dass sie auf der lokalen bzw. sequenznahen Rückgratgeometrie der i-ten Seitenkette basierten: Hierzu wurden die Bindungsvektoren $\mathbf{b}_{i+j}$ mit $-1 \leq j \leq 2$ angesetzt, da die reinen Koordinaten der C_{i+j}^{α}-Atome mit $-2 \leq j \leq 2$ nicht invariant gegenüber Translation und Rotation sind. Weiter entfernte Bindungsvektoren wurden nicht verwendet, da sie außer in ausgedehnten einheitlichen Sekundärstrukturen nicht mehr charakteristisch für die lokale Rückgratgeometrie um eine Seitenkette sind, wie beispielsweise im Übergangsbereich zwischen verschiedenen Strukturtypen. Wie sich in vergleichenden Rechnungen allerdings zeigte, verschlechterten sich die Ergebnisse von Gl. 4.10, wenn Funktionen abhängig von $\mathbf{b}_{i-1}$ und $\mathbf{b}_{i+2}$ miteinbezogen wurden. Dies zeigte, dass diese Vektoren sehr wenige oder sogar falsche Informationen zur Seitenkettenposition beitragen, so dass die Funktionen $\mathbf{f}_k$ schließlich nur von $\mathbf{b}_{i+j}$ mit $0 \leq j \leq 1$ abhängig gewählt wurden.
Der gesamte zur Verfügung gestellte Satz an Funktionen $\mathbf{f}_k$ enthielt somit die Bindungsvektoren $\mathbf{b}$ an sich, ihre normierten Richtungsvektoren sowie deren Kombinationen mittels Addition und Vektorprodukt, so dass Gl. 4.10 ausformuliert zu

$$\begin{aligned}\mathbf{q}_i \;=\; & c_{1,s_i}\mathbf{b}_i \;+\; c_{2,s_i}\frac{\mathbf{b}_i}{\|\mathbf{b}_i\|} \;+\; c_{3,s_i}\mathbf{b}_{i+1} \;+\; c_{4,s_i}\frac{\mathbf{b}_{i+1}}{\|\mathbf{b}_{i+1}\|} \;+\; c_{5,s_i}(\mathbf{b}_i \times \mathbf{b}_{i+1}) \\ & + c_{6,s_i}\frac{\mathbf{b}_i \times \mathbf{b}_{i+1}}{\|\mathbf{b}_i \times \mathbf{b}_{i+1}\|} \;+\; c_{7,s_i}(\mathbf{b}_i + \mathbf{b}_{i+1}) \;+\; c_{8,s_i}\frac{\mathbf{b}_i + \mathbf{b}_{i+1}}{\|\mathbf{b}_i + \mathbf{b}_{i+1}\|} \\ & + c_{9,s_i}(\mathbf{b}_i - \mathbf{b}_{i+1}) \;+\; c_{10,s_i}\frac{\mathbf{b}_i - \mathbf{b}_{i+1}}{\|\mathbf{b}_i - \mathbf{b}_{i+1}\|}\end{aligned} \tag{4.11}$$

wird, wobei $(\cdot \times \cdot)$ das äußere bzw. Kreuzprodukt ist. Diese Gleichung enthält alle der zur Approximation zur Verfügung gestellten Funktionen, wobei bei der Optimierung jedoch die Anzahl n_f der effektiv verwendeten Funktionen $\mathbf{f}_k$ so variiert wurde, dass alle Werte $n_f = 2, 3, ..., 10$ durchlaufen wurden. Hierbei wurden zu jeder vorgegebenen Gesamtanzahl alle Permutationen der oben angegebenen Funktionen optimiert, mit dem Ziel, die optimale Kombination dieser Funktionen zu bestimmen. Diese Permutationen waren reihenfolgeunabhängig und jede Funktionen durfte in einer Permutationen nur einmal verwendet werden. Zudem wurden die redundanten Permutationen ausgeschlossen, so dass beispielsweise keine Permutation zugelassen wurden, die gleichzeitig die Funktionen zu den Koeffizienten c_1, c_3 und c_7 enthielten. Tatsächlich zeigte sich aber, dass das Ergebnis unabhängig vom Ausschluss der redundanten Kombinationen war, da sich ihre Koeffizienten im Falle ihrer Einbeziehung nach der Minimierung gegeneinander aufhoben.

Zur Bestimmung der Koeffizienten wurden die 273 Proteine des TOP500H-Proteinsatzes verwendet. Es wurde ein Gleichungssystem $\mathbf{Fc} = \mathbf{Q}$ aufgestellt, in dem die ausgewerteten Funktionen $\mathbf{f}_k$ aus Gl. 4.11 in der Koeffizientenmatrix $\mathbf{F}$ zusammengefasst wurden und die Vektoren $\mathbf{q}_i$, die von den C^α-Atomen zu den Seitenketten führen, zusammengefasst als Rechte-Seite-Vektor $\mathbf{Q}$ auftauchen. Der Vektor $\mathbf{c}$ ist der Lösungsvektor dieses Gleichungssystems. Jede der drei kartesischen Vektorkomponenten der verwendeten Seitenkette bildet eine Zeile des Gleichungssystems, so dass aus der Gesamtanzahl $n_{sc}(a)$ an verwendeten Seitenketten vom Typ $a \in \{1, 2, ..., 20\}$ ein Gleichungssystem mit $3n_{sc}(a)$ Zeilen und zwei bis zehn Spalten entstand, deren Anzahl durch die Menge n_f der verwendeten Funktionen $\mathbf{f}_k$ gegeben ist. Dieses Gleichungssystem ist somit stark überbestimmt. Für jeden Aminosäuretyp a wurde ein separates Gleichungssystem aufgestellt und gelöst.

Die Kleinste-Quadrate-Optimierung der Koeffizienten wurde über eine Singulärwertzerlegung (*singular value decomposition*) der Koeffizientenmatrix berechnet, wofür Standardroutinen aus den *Numerical Recipes* implementiert wurden [104]. Zur Überprüfung der Singulärwertzerlegung wurde die Koeffizientenmatrix $\mathbf{F}$ aus der Matrixzerlegung $\mathbf{F}'$ durch Matrizenmultiplikation wieder hergestellt und die Differenz dieser beiden Matrizen elementweise über $\sum_{ij} |f_{ij} - f'_{ij}|$ ermittelt. Das numerische Ergebnis dieses Vergleichs ist in Tab. 4.1 aufgelistet. Da die verwendeten Koordinaten in Å in der Größenordnung um 10^0 gegeben sind, die Differenzen aber im Bereich um 10^{-11} Å liegen, ist die Zerlegung für diesen Fall exakt genug.

Nach der Bestimmung des Koeffizientenvektors **c** wurde dieser an den TOP500H-Proteinen getestet. Dazu wurde der erhaltene Vektor **c** in Gl. 4.11 eingesetzt, die approximierten Seitenkettenpositionen $\mathbf{q}_i$ berechnet und diese mit den Positionen der Seitenketten $\mathbf{q}_i^*$ in den nativen PDB-Strukturen über $d_i^{(fix)} = \|\mathbf{q}_i - \mathbf{q}_i^*\|$ verglichen. Der Wert $d_i^{(fix)}$, welcher die Abweichung der approximierten Position von der erwarteten PDB-Struktur angibt, wurde separat für jeden Aminosäuretyp a zusammengefasst und der Mittelwert der Verteilung $\mu(a)$ sowie die Standardabweichung berechnet. Zur Charakterisierung der Güte einer Permutation wurde der Erwartungswert $E(\mu)$ des Mittelwertes bezogen auf alle zwanzig Aminosäuretypen gemäß der Gleichung

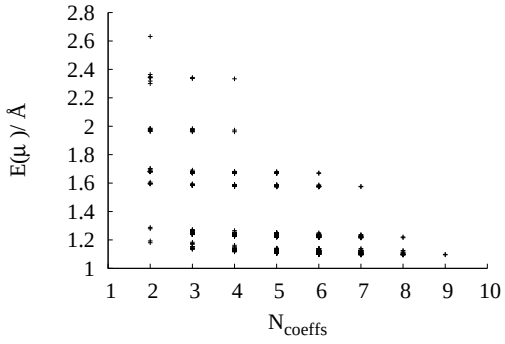


Abbildung 4.11: Abhängigkeit des Erwartungswertes $E(\mu)$ von der Anzahl der Koeffizienten.

$$E(\mu) = \sum_{a=1}^{20} \frac{\mu(a)}{n_{sc}(a)} \tag{4.12}$$

verwendet. Ausgeführt werden im folgenden nur die Ergebnisse der Permutation, die zum niedrigsten Erwartungswert $E(\mu)$ führte, da diese im Schnitt die beste Näherung darstellt.

Die erhaltenen Erwartungswerte $E(\mu)$ der optimierten Permutationen lagen im Intervall zwischen 1.095 Å und 2.632 Å. Sie korrelierten nur sehr gering mit der Anzahl der Funktionen n_f. Zur Anschauung sind in Abb. 4.11 die Erwartungswerte gegen n_f aufgetragen. Charakteristisch für das Ergebnis war zum einen, dass lediglich die Obergrenze des Erwartungswertes von n_f abhing, zum anderen, dass die Erwartungswerte Gruppen und keine gleichmäßige Verteilung bildeten sowie dass viele Erwartungswerte sehr ähnliche Werte annahmen, wodurch (scheinbar) nur wenige Datenpunkte in der Grafik vorhanden sind. Die erste Erwartungswertgruppe lag bei ca. $E(\mu) = 1.2$ Å. Die nächsten Gruppen hatten alle ungefähr den gleichen Abstand untereinander von ca. 0.3 bis 0.4 Å, so dass die nächsten Gruppen bei ca. $E(\mu) =$ 1.6, 2.0, 2.4 und 2.7 Å lagen.
Die Permutation mit dem niedrigsten Erwartungswert von 1.095 Å enthielt acht Funktionen. Nicht in ihr enthalten waren die Funktionen zu den Koeffizienten c_1 und c_3, den unnormierten Bindungsvektoren. Die Daten zu dieser Permutation sind in Tab. 4.1 aufgelistet, in der die Anzahl an verwendeten Aminosäuren aus den PDB-Dateien, die einzelnen Mittelwerte der Aminosäuren zusammen mit deren Standardabweichung enthalten sind, sowie in Tab. 4.2, in der die Zahlenwerte der minimierten Koeffizienten gezeigt werden.

Bemerkenswert ist, dass es neben der Permutationen mit dem niedrigsten $E(\mu)$ weitere gab, die ähnlich gute Ergebnisse lieferten, obwohl diese wesentlich weniger Funktionen enthielten. Beispielsweise besaß die Permutation, die nur die zwei Funktionen zu den Koeffizienten c_5 und c_{10} enthielt, einen Erwartungswert von 1.181 Å, welcher sich nur um 0.086 Å von der Permutation mit dem niedrigsten Erwartungswert, in der acht Funktionen enthalten waren, unterschied. Dies lässt sich so interpretieren, dass einige Kombinationen weniger Funktionen ausreichend sind, wesentliche geometrische Eigenschaften der Seitenketten zu erfassen und dass das Hinzufügen von weiteren Funktionen zu diesen nicht unmittelbar zu einer deutlich besseren Vorhersage führt, da der zusätzliche Informationsgehalt klein oder redundant ist. Dies wird auch durch einen Vergleich der Mittelwerte $\mu(a)$ der einzelnen Aminosäuren zu diesen beiden Permutationen gestützt, da die Unterschiede der $\mu(a)$ zwischen beiden Permutationen für alle Aminosäuren nur im Bereich von 0.02 Å bis 0.15 Å lagen (mit Ausnahme von Prolin, s. u.), und dies auch für die längeren, flexibleren Seitenketten wie z. B. Lysin galt. Dies bedeutet, dass der niedrigere Erwartungswert für die Permutation mit wenigen Funktionen auf allen Aminosäuren basiert und nicht, wie theoretisch möglich, auf einigen wenigen Seitenketten, die besser beschrieben werden als andere. Dementsprechend ist diese Erfassung der essentiellen Eigenschaften der Seitenkettenposition ein systematischer Effekt und kein zufälliger. Dennoch lieferte die Permutation mit acht Koeffizienten für alle Aminosäure stets die niedrigeren Erwartungswerte. Einzig für Prolin fiel der Unterschied zwischen beiden Permutationen größer aus als oben beschrieben: Für acht Koeffizienten betrug der Mittelwert $\mu(a) = 0.338$ Å, während dieser für zwei Koeffizienten $\mu(a) = 1.006$ Å betrug. Hier konnte mit diesen zwei Funktionen die besondere Positionierung des Prolinringes nur ungenügend reproduziert werden.

Die schlechtesten Vorhersagen lieferte dieses Modell für Seitenketten mit Ringsystemen wie Phenylalanin, Tyrosin oder Tryptophan (siehe Tab. 4.1). Die Mittelwerte dieser Aminosäuren lagen alle im Bereich um zwei Å oder mehr. Dies kann damit begründet werden, dass dieses Modell die Seitenkettenschwerpunkte approximiert, welche für diese Aminosäuren ungefähr in der Ebene bzw. nahe des Mittelpunktes der Ringsysteme liegen. Diese Ringsysteme können aber um die C^{α}-C^{β}-Bindung rotieren, wobei es aufgrund der Sterik drei verschiedene Vorzugsorientierungen gibt, deren Schwerpunkte relativ weit auseinander liegen. Im Minimierungsprozess der Koeffizienten wird aber über alle Seitenkettenrotamere gemittelt, was dazu führt, dass der resultierende approximative Seitenkettenvektor hier nicht zu einer bestimmten Seitenkettenposition eines bestimmten Rotamers zeigt, sondern zum mittleren Schwerpunkt aller Rotamere. Dieser befindet sich ungefähr in der Verlängerung der C^{α}-C^{β}-Bindung und fällt nicht mit einem realen Atom zusammen. Dieser Effekt tritt selbstverständlich auch bei den anderen Aminosäuren auf, aber aufgrund des großen geometrischen Unterschiedes der verschiedenen Rotamerzustände der Seitenketten mit Ringsystemen, ist er bei diesen beson-

Aminosäure a	$n_{sc}(a)$	$\mu(a)$/Å	Std.-Abw./Å	**F**-**F'**-Diff./Å
Ala	5481	0.202280	0.154075	$1.861 \cdot 10^{-11}$
Cys	689	1.113443	0.408252	$3.190 \cdot 10^{-12}$
Asp	3521	1.191222	0.480843	$1.030 \cdot 10^{-11}$
Glu	3814	1.483005	0.536114	$2.110 \cdot 10^{-11}$
Phe	2408	2.052221	0.526981	$6.955 \cdot 10^{-12}$
Gly	-	-	-	
His	1433	1.913942	0.510178	$5.558 \cdot 10^{-12}$
Ile	3319	0.611411	0.302099	$1.307 \cdot 10^{-11}$
Lys	3495	1.673741	0.603760	$2.634 \cdot 10^{-11}$
Leu	5312	0.874867	0.375665	$2.131 \cdot 10^{-11}$
Met	1254	1.386779	0.514696	$5.426 \cdot 10^{-12}$
Asn	2531	1.199021	0.476949	$8.407 \cdot 10^{-12}$
Pro	2738	0.338357	0.216021	$2.024 \cdot 10^{-11}$
Gln	2204	1.461318	0.526982	$1.048 \cdot 10^{-11}$
Arg	2732	2.122466	0.840714	$1.797 \cdot 10^{-11}$
Ser	3323	0.769788	0.262031	$1.090 \cdot 10^{-11}$
Thr	3248	0.505159	0.259025	$1.043 \cdot 10^{-11}$
Val	4347	0.453578	0.228821	$1.749 \cdot 10^{-11}$
Trp	790	2.483827	0.696586	$3.044 \cdot 10^{-12}$
Tyr	1982	2.526374	0.627611	$9.645 \cdot 10^{-12}$

Tabelle 4.1: Statistische Werte zur Optimierung. n_{sc} ist die Anzahl an verwendeten Seitenketten. $\mu(a)$ ist der Mittelwert der Abweichung von der PDB-Seitenkettenposition. Std.-Abw. ist die Standardabweichung dieses Mittelwertes und **F**-**F'**-Diff. zeigt den numerischen Test der Singulärwertzerlegung.

ders hervortretend. Dies Ungenauigkeit hat wichtige Auswirkungen auf die Verwendung dieses Modells im Rahmen des Potentials. Für Tryptophan beispielsweise beträgt der Mittelwert der Abweichung ca. $\mu(a) = 2.5$ Å mit einer Standardabweichung von ca. 0.7 Å. Dies führt dazu, dass ein auf dieser Grundlage berechnetes Paarpotential zwischen zwei Tryptophanseitenkettenschwerpunkten basierend auf deren Abstand, eine große durchschnittliche Ungenauigkeit von ca. 5 Å hat.

Zur weiteren Übersicht über die Genauigkeit diese Ansatzes wurden für alle Aminosäure die statistische Häufigkeit der Werte $d^{(fix)}$ in Intervallen zu 0.1 Å berechnet. Die Ergebnisse hierzu sind in Abb. 4.12 gezeigt. Die meisten dieser Statistiken zeigten eine gleichmäßige Verteilung um den Mittelwert und keine stärkere Gewichtung zu niedrigeren oder höheren $d^{(fix)}$-Werten. Lediglich für die sehr kurzen Seitenketten wie Alanin, Valin oder Prolin liefert das Modell geringe Abweichungen.

Aminosäure	Koeffizientenindex i									
a	1	2	3	4	5	6	7	8	9	10
Ala	0	0.587	0	-1.315	-0.089	0.223	-0.011	0.840	0.399	-1.799
Cys	0	15.669	0	-6.054	-0.012	-0.898	-0.980	-1.744	2.844	-1.729
Asp	0	1.426	0	-2.634	-0.078	-0.005	0.714	-3.290	0.585	-2.154
Glu	0	12.371	0	-3.885	-0.190	1.233	-1.085	-0.177	2.375	-3.623
Phe	0	11.208	0	-0.481	0.054	-1.461	-1.976	3.377	1.489	-2.339
Gly	-	-	-	-	-	-	-	-	-	-
His	0	22.197	0	-12.282	0.010	-0.991	-1.348	0.100	4.537	-2.368
Ile	0	3.655	0	-0.956	-0.139	0.905	-0.514	0.943	0.773	-2.746
Lys	0	10.564	0	-4.055	-0.197	1.353	-0.996	0.859	2.079	-3.584
Leu	0	8.086	0	-3.280	-0.205	1.942	-0.631	-0.049	1.623	-2.878
Met	0	15.062	0	-4.111	-0.049	-0.387	-1.404	-0.262	2.633	-3.190
Asn	0	6.548	0	-2.429	-0.016	-0.722	-0.155	-2.495	1.139	-1.760
Pro	0	4.366	0	-2.399	-0.134	0.428	-0.416	0.066	1.274	-2.472
Gln	0	15.226	0	-2.690	-0.165	1.036	-1.614	-0.273	2.532	-3.353
Arg	0	16.002	0	-1.785	-0.158	0.447	-2.131	1.543	2.499	-3.933
Ser	0	3.675	0	-2.905	-0.062	-0.416	-0.043	-0.146	0.990	-1.837
Thr	0	3.168	0	-1.488	-0.062	-0.309	-0.225	0.086	0.716	-1.949
Val	0	2.440	0	-1.740	-0.097	0.459	-0.275	1.287	0.681	-2.263
Trp	0	10.209	0	2.641	-0.022	-0.483	-2.208	3.503	1.040	-2.971
Tyr	0	17.295	0	1.134	0.056	-1.517	-2.969	3.170	2.059	-2.469

Tabelle 4.2: Optimierte Koeffizienten $c_{i,a}$ für die Vektorlinearkombination mit dem niedrigsten Erwartungswert.

Diese Ergebnisse zeigten insgesamt, dass mit diesem Modell eine Platzierung der Seitenkette mit einer damit vom Typ der Aminosäuren abhängigen Ungenauigkeit möglich ist. Es werden wesentliche grundlegende geometrische Eigenschaften erfasst. Jedoch zeigten sich auch die Grenzen dieses Modells, die hauptsächlich darin begründet sind, dass die Positionen der Seitenketten nicht alleine und hinreichend durch die Geometrie des Proteinrückgrates bestimmt ist, sondern dass die Ausrichtung der Seitenketten von der gesamten lokalen Umgebung und somit auch von Atomen, die in der Sequenz entfernt, aber räumlich nah sind, welches im wesentlichen die anderen Seitenketten betrifft. Hinzu kommt, dass die räumliche Orientierung der Seitenketten an der Oberfläche eines Proteins sehr flexibel ist. Aufgrund der großen mittleren Fehler dieses Modells wurde es nicht zur Seitenkettenpositionierung im Kraftfeldansatz verwendet. Stattdessen wurde dieses Modell zur Clustermethode erweitert, die im folgenden Abschnitt beschrieben wird.

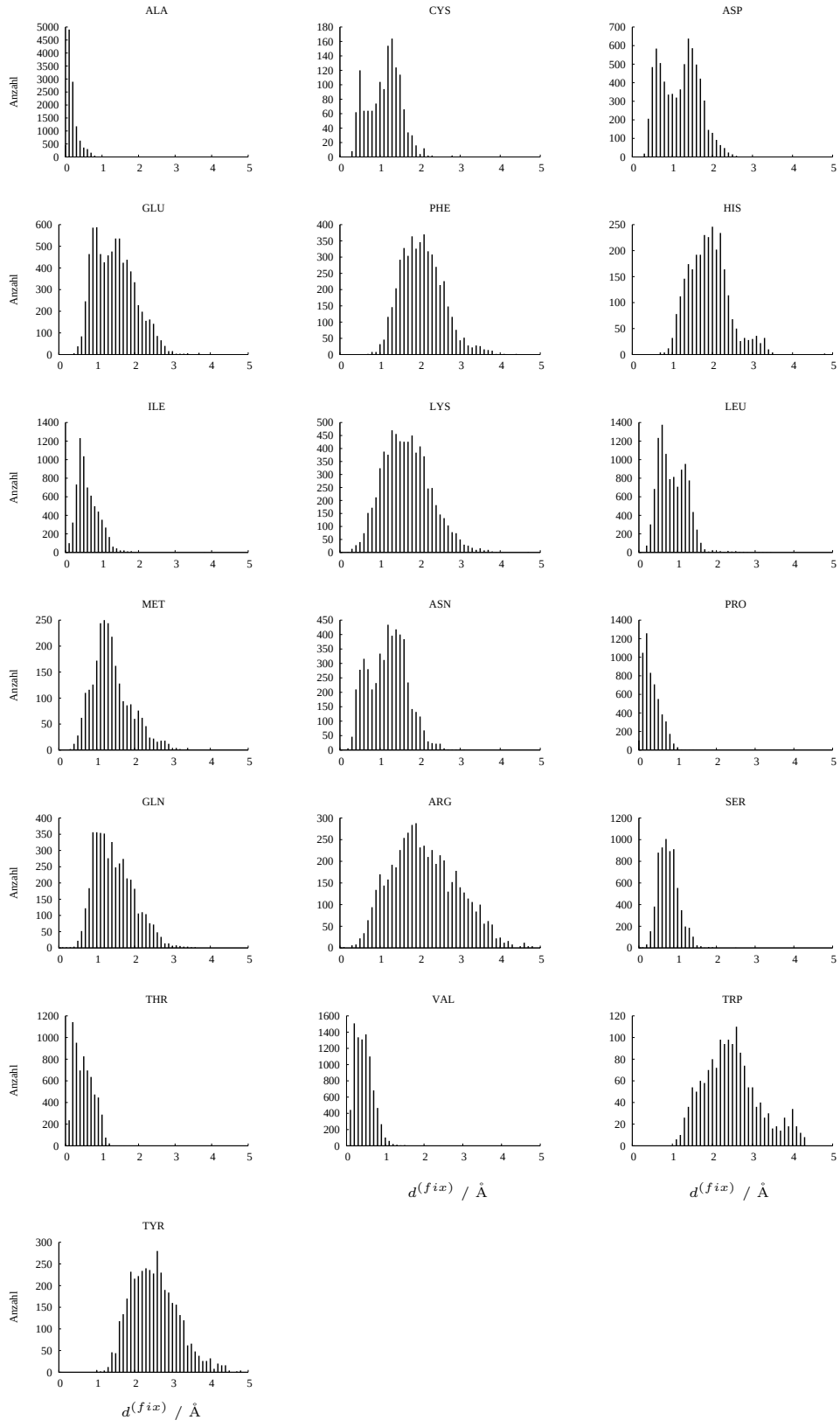


Abbildung 4.12: Abstand $d^{(fix)}$ zwischen der approximierten Seitenkettenposition mit festen Koeffizienten (siehe Gl. 4.11) und der PDB-Position (Intervallbreite: 0.1 Å.)

4.2.3 Clustermethode

Ein wesentlicher Schwachpunkt der fixierten Seitenkette aus dem vorherigen Abschnitt war, dass bei der Optimierung der Koeffizienten über alle Seitenkettenschwerpunkte gemittelt wurde, so dass am Ende eine mittlere Seitenkettenposition für alle Rückgratgeometrien erhalten wurde, die nur abhängig von den C^{α}-Koordinaten war, sich aber nicht der Umgebung der Seitenkette anpassen konnte. Dies ist zur Beschreibung einer nativen Struktur ungenügend, da die dichte, umgebungsabhängige Packung der Seitenketten wesentlich zum nativen Zustand beiträgt. Daher wurde der Ansatz der fixierten Seitenkette dahingehend erweitert, mehrere mögliche Positionen einer Seitenkette zu beschreiben, wobei diese Beschreibung wiederum auf einer Linearkombination der Rückgratvektoren beruht.
Im folgenden wird zunächst das generelle Vorgehen beschrieben und im Anschluss daran dazu Details und die Ergebnisse beschrieben.

Wie weiter oben Text beschrieben wurde, war ein Schluss, der aus den Häufigkeitsverteilungen der Seitenketten, die bei der Potentialmethode gewonnen wurde (siehe Abschn. 4.2.1), dass bestimmte Positionen von Seitenketten bevorzugt werden. Dies war besonders bei Seitenketten mit einem Ringsystem wie Tryptophan oder Histidin auffällig und für diese besonders bei der Verteilung zum Vektor $\mathbf{b}_i + \mathbf{b}_{i+1}$ (siehe Abb. 4.10). Eine dreidimensionale Darstellung der Verteilung der Schwerpunkte beispielsweise von Phenylalanin zeigte, dass relativ zur C^{α}_{i-1},C^{α}_{i},C^{α}_{i+1}-Ebene bestimmte Bereiche gibt, in denen der Seitenkettenschwerpunkt mit einer erhöhten Wahrscheinlichkeit auftritt. Für das dritte Modell wurde wie bereits auch für die Potentialmethode ein orthogonales Koordinatensystem am i-ten C^{α}-Atom definiert, das aus den drei normierten Vektoren

$$\mathbf{e}_{1,i} = \frac{\mathbf{b}_{i+1} - \mathbf{b}_i}{\|\mathbf{b}_{i+1} - \mathbf{b}_i\|}, \qquad \mathbf{e}_{2,i} = \frac{\mathbf{b}_{i+1} + \mathbf{b}_i}{\|\mathbf{b}_{i+1} + \mathbf{b}_i\|} \quad \text{und} \quad \mathbf{e}_{3,i} = \frac{\mathbf{b}_i \times \mathbf{b}_{i+1}}{\|\mathbf{b}_i \times \mathbf{b}_{i+1}\|} \tag{4.13}$$

bestand. Hierbei waren wieder die $\mathbf{b}_k = \mathbf{x}_k$ - $\mathbf{x}_{k-1}$ die Bindungsvektoren der C^{α}-Atome. Mit Hilfe dieser Basisvektoren lässt sich jeder Seitenkettenvektor $\mathbf{q}_i$ als Linearkombination $\mathbf{q}_i = \sum_{n=1}^{3} \lambda_n \, \mathbf{e}_{n,i}$ mit $\lambda \in \mathbb{R}$ angeben, was eine analoge Darstellung zu Gl. 4.11 mit den entsprechenden Koeffizienten ist.
Allerdings wurde in diesem Ansatz nicht angestrebt, gemittelte Werte für die Koeffizienten zu erhalten, sondern es wurde bezüglich dieses Koordinatensystems untersucht, ob sich für jede Aminosäure bestimmte unterschiedliche Häufungspunkte in den $\lambda_{i,n}$ finden ließen. Hierzu wurden wiederum die Proteine des TOP500H-Satzes analysiert, und mit den Daten der PDB-Strukturen zu jeder Seitenkette i ein Gleichungssystem der Form

$$\begin{pmatrix} e_{1,i,1} & e_{1,i,2} & e_{1,i,3} \\ e_{2,i,1} & e_{2,i,2} & e_{2,i,3} \\ e_{3,i,1} & e_{3,i,2} & e_{3,i,3} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} q_{i,1} \\ q_{i,2} \\ q_{i,3} \end{pmatrix} \tag{4.14}$$

aufgestellt und die Lösungen $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^T$ bestimmt. Die Lösungen wurden dann derjenigen Lösungsmenge $\boldsymbol{\Lambda}_a$ mit $a \in \{1, 2, ..., 20\}$ zugeordnet, für die $s_i = a$ gilt, wobei s_i wieder die i-te Komponente des Sequenzvektors $\mathbf{S}$ ist.
Diese Lösungsmengen $\boldsymbol{\Lambda}_a$ für alle Aminosäuren sind in den Abb. 4.15 bis 4.18 dargestellt. In den Abbildungen sind die zweidimensionalen Projektionen der dreidimensionalen Lösungsvektoren auf die jeweiligen λ-Achsen gezeigt.
Die Punktwolken liegen für die meisten Aminosäuren auf einer dünner Kugeloberfläche, deren Zentrum das zur Seitenkette gehörende C^{α}-Atom ist. Der Radius dieser Kugel entspricht der Länge des Seitenkettenvektors (siehe hierzu auch Abb. 4.5) und damit gilt auch für die Lösungsvektoren $\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2} = \|\boldsymbol{\lambda}\| = \|\mathbf{q}\|$. Diese Kugeloberflächen sind für viele Aminosäuren in den Abb. 4.15 bis 4.18 meist gut in der mittleren Spalte erkennbar, welche die Projektion auf λ_1 und λ_3 enthält. So z. B. für Alanin, Asparaginsäure, Threonin oder Valin. Ebenso aber auch in der Projektion auf λ_1 und λ_2 für für Phenylalanin, Tryptophan oder Tyrosin. Aufgrund der Projektion und der damit reduzierten Dimensionalität wirken diese Punktwolken leicht verschmiert. Dabei handelt es sich um Punkte aus verschiedenen Schnittebenen parallel zur Projektionsebene, die durch das jeweilige λ_j und λ_k aufgespannt wird. Aminosäuren, die eine breitere Verteilung der Seitenkettenvektorlänge aufzeigten, wie beispielsweise Lysin oder Arginin, enthielten mehrere dicht beieinanderliegende Kugeloberflächen in geringem Abstand entsprechend der Längenverteilung.
Wie aus den graphischen Übersichten deutlich wird, fanden sich für viele Aminosäuren Bereiche mit hoher Punktdichte, dies besonders, wie zu erwarten war, bei den kurzen Seitenketten. Die sehr flexiblen Seitenketten wie Lysin und Arginin zeigen diffuse Verteilungen und kaum Bereiche mit erhöhter Punktdichte. Zwischen diesen beiden Extremen befinden sich Aminosäuren wir Glutamin, Asparagin oder Methionin, welche eine gewisse Ausbildung an Häufigkeitszentren zeigten, die aber uneinheitlich verteilt waren. Eine besondere Stellung nahmen hier die Seitenketten mit Ringsystemen wie Phenylalanin, Histidin, Tyrosin oder Tryptophan ein. Diese zeigten in den Verteilungen jeweils drei gut voneinander abgegrenzte Gebiete, welche jeweils bestimmten Rotamerzuständen der Seitenkette entspricht.

Im nächsten Schritt wurden die Verteilungsdaten einer Clusteranalyse unterzogen. Die erhaltenen Lösungen $\boldsymbol{\lambda}$ sollten so für jede Aminosäure in Zentren eingeteilt werden, die für bestimmte Seitenkettenpositionen charakteristisch bzw. repräsentativ sind. Für eine Clusteranalyse existieren mehrere verschiedene Verfahren. Hier wurde das sog. *k-means*-Verfahren verwendet [105–107] und dazu die FORTRAN90-Routinen von J. Burkardt [108] implementiert, welche mehrere Varianten des *k-means*-Verfahrens bereitstellen, unter anderem unterschiedliche Clusterinitialisierungs- und -partitionierungsalgorithmen für Daten im $\mathbb{R}^m$ mit $m \geq 1$. Ganz allgemein muss vor dem Beginn des Algorithmus die Gesamtanzahl n_c der Zentren (Cluster) angegeben werden, in die die Punkte aufgeteilt werden. Darauf erzeugt das Pro-

gramm aus den Eingabepunkten die Koordinaten der Clusterzentren $\bar{\boldsymbol{\xi}}_k$ mit $k \in \{1, 2, ..., n_c\}$, worauf die n_d Datenpunkte $\boldsymbol{\xi}_i$ mit $i \in \{1, 2, ..., n_d\}$ ihrem nächstgelegenen Clusterzentrum zugeordnet werden, und eine Bewertung G über eine vorgegebene Funktion berechnet wird, um die Güte der Clustereinteilung zu bestimmen. Diese Funktion ist definiert durch

$$G = \sum_{i=1}^{n_d} \left[\min_{k=1}^{n_c} \|\boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}}_k\| \right]^2 \tag{4.15}$$

Summiert wird jeweils über den kürzesten quadratischen Abstand eines Datenpunktes zu dem entsprechenden Clusterzentrum. Im Anschluss werden die Clusterkonfigurationen variiert, d. h. es werden die Koordinaten der Zentren verändert wie auch die Zugehörigkeit der Punkte zu den Zentren. Für die neue Konfiguration wird dann wieder der Wert der Funktion 4.15 bestimmt. Dieser Vorgang wird wiederholt bis eine Konvergenz eintritt. Als Resultat erhält man schließlich die Koordinaten der Clusterzentren und die Zuordnung der Datenpunkte zu diesen, die den kleinsten Wert G besitzen.
Im vorliegenden Fall waren die Lösungen $\mathbf{\Lambda}_a$ für jedes $a \in \{1, 2, ..., 20\}$ die Datenpunkte, die in Cluster eingeteilt wurden. Die resultierenden Clusterzentren repräsentierten dann Häufungspunkte der λ_k und damit bestimmte Seitenkettenpositionen, relativ zu einer bestimmten Rückgratgeometrie.
Das Clusterproblem ist *NP*-hart. Es kann nicht gewährleistet werden, dass ein bestimmter Lauf das bestmögliche Resultat liefert. Aus diesen Gründen wurden zu jeder Aminosäure a und zu jeder vorgegebenen Gesamtanzahl an Clustern n_c insgesamt 1000 separate Läufe durchgeführt und von diesen das Ergebnis mit dem niedrigsten G verwendet. Es zeigte sich, dass sehr viele der 1000 Läufe das gleiche Resultat lieferten bzw. sehr dicht beieinander lagen. Durch eine Erhöhung des Wertes von n_c lässt sich G stets erniedrigen, bis jeder Punkt seine eigenes Clusterzentrum ist. Aus diesem Grund wurden zwei Kriterien zur Bestimmung der verwendeten Clusterpartitionierung angewendet: Zum einen wurde zunächst die maximale Anzahl an Clusterzentren n_c auf 20 beschränkt. Zum anderen wurde ein zweites Kriterium angewendet, dass die Abdeckung der Datenpunkte durch die Clusterzentren beschreibt. Hierzu wurde zu jedem Clusterzentrum die Varianz der ihm zugeordneten Datenpunkte berechnet. Aus diesen n_c Varianzen wurde dann unter Verwendung der Anzahl der Datenpunkte (Population) eines Clusterzentrums der Erwartungswert $E(\mathrm{Var})$ der Varianz für die gesamte Partitionierung berechnet. Dies erfolgte über folgende Gleichungen: Gegeben sei die Population m_k eines Clusters k. Die dem k-ten Cluster zugeordneten Datenpunkte seien $\boldsymbol{\xi}_{i,k}$ mit $i \in \{1, 2, ..., n_d\}$. Der mittlere Abstand μ_k dieser Punkte von Clusterzentrum beträgt

$$\mu_k = \frac{1}{m_k} \sum_{i=1}^{m_k} \|\boldsymbol{\xi}_{i,k} - \bar{\boldsymbol{\xi}}_k\| \tag{4.16}$$

Die Varianz $\mathrm{Var}(k)$ des k-ten Clusters ist gegeben durch

$$\mathrm{Var}(k) = \frac{1}{m_k - 1} \sum_{i=1}^{m_k} \left(\boldsymbol{\xi}_{i,k} - \mu_k\right)^2 \tag{4.17}$$

und der Erwartungswert $E(\mathrm{Var}(k))$ der Varianz für alle n_c Cluster ist

$$E(\mathrm{Var}(k)) = \sum_{k=1}^{n_c} \frac{m_k}{\sum_{j=1}^{n_c} m_j} \mathrm{Var}(k) = \sum_{k=1}^{n_c} w_k \cdot \mathrm{Var}(k) \tag{4.18}$$

mit der Wahrscheinlichkeit w_k bzw. der anteiligen Population des Clusters k an der Gesamtpopulation $\sum_{j=1}^{n_c} m_j$.

Betrachtet man den Verlauf von $E(\mathrm{Var})$ in Abhängigkeit von der vorgegebenen Anzahl an Clustern n_c (siehe Abb. 4.14), so zeichnen sich die Graphen der meisten Aminosäuren qualitativ durch 3 Bereiche aus. Bei wenigen Clustern (2 bis ca. 4) fällt die Erniedrigung der Varianz durch Hinzufügen eines weiteren Clusters relativ groß aus. Dem folgt ein Übergangsbereich mit abnehmendem Gefälle, welcher je nach Aminosäure im Bereich zwischen ca. 5 bis 10 Clustern liegt. Für mehr als 10 Cluster ist der Verlauf nahezu asymptotisch, so dass sich hier durch Hinzufügen weiterer Zentren die Varianz kaum weiter erniedrigt. Für viele längere Aminosäuren ist dieser asymptotische Bereich bei einem Varianzerwartungswert von ca. 0.05 Å^2 erreicht (siehe z. B. Cys, Asp, Glu, Phe, His, Met, Gln, Ser in Abb. 4.14). Aus diesem Grund wurde für $E(\mathrm{Var})$ ein Wert von 0.05 Å^2 als Grenze festgelegt, da ab diesem der weitere Informationsgewinn durch Hinzufügen eines weiteren Clusters sehr gering ausfällt. Die Clusteranzahl mit einem maximalen Wert von 20, deren Varianzerwartungswert diesen Wert unterschreitet, wurde somit zur Approximation der Seitenkette verwendet.

Aminosäure	$E(\|\mathbf{q}\|)$/Å	n_c
Ala	1.50	2
Cys	2.39	4
Asp	2.57	5
Glu	3.23	14
Phe	3.41	9
Gly	-	-
His	3.19	8
Ile	2.33	3
Lys	3.57	15
Leu	2.61	5
Met	3.19	13
Asn	2.53	6
Pro	1.89	3
Gln	3.17	13
Arg	4.22	20
Ser	1.98	2
Thr	1.97	2
Val	1.97	2
Trp	3.89	20
Tyr	3.88	16

Tabelle 4.3: Distanzerwartungswert $E(\|\mathbf{q}\|)$ und Anzahl n_c der bestimmten Cluster.

Die hieraus resultierende Anzahl an Clustern n_c ist in Tab. 4.3 dargestellt. Zum Vergleich, ob ein direkter Zusammenhang zwischen Anzahl der Cluster und der Länge der Seitenkette besteht, wurde der Erwartungswert der Länge des Seitenkettenvektors $E(\|\mathbf{q}\|)$ berechnet. Hierzu wurden die Daten zu Abb. 4.5 verwendet, in der die Abstände in Intervallen zu je 0.1 Å bestimmt wurden. Analog zu den Gleichungen

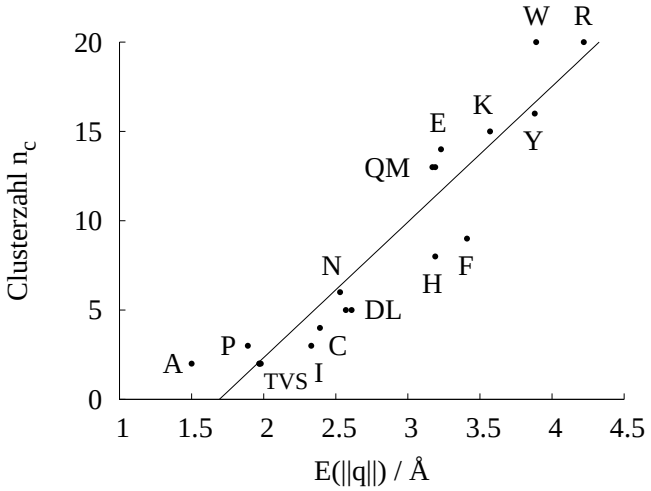


Abbildung 4.13: Anzahl der Clusterzentren n_c in Abhängigkeit vom Erwartungswert der Seitenkettenlänge $E(\|\mathbf{q}\|)$ mit Regressionsgrade. (Die Aminosäuren sind mit ihrem Ein-Buchstaben-Code gekennzeichnet.)

4.16 bis 4.18 wurde unter Verwendung der Population dieser Intervalle der Erwartungswert $E(\|\mathbf{q}\|)$ bestimmt. Diese Werte sind ebenfalls in Tab. 4.3 aufgelistet. Eine Auftragung der Anzahl der verwendeten Cluster gegen diesen Erwartungswert (siehe Abb. 4.13) zeigt eine lineare Abhängigkeit. Der Korrelationskoeffizient der Regressionsgeraden beträgt 0.955. Die zugehörige Gerade hat die Form n_c = -12.85±1.87 + (7.59±0.64 $Å^{-1}$)·$E(\|\mathbf{q}\|)$.
Die weiteren Daten der Clusterzentren, wie beispielsweise die Koordinaten der Clusterzentren, die Varianzen, kürzeste, mittlere und größte Abstände der Punkte zu ihren Zentren sind im Anhang in der Tabelle 6.4 aufgeführt.

Mit diesem Ansatz war es möglich, grundsätzlich so viele unterschiedliche Seitenkettenpositionen wie Clusterzentren vorhanden waren zu verwenden. Wie im vorhergehenden Abschnitt zur fixierten Seitenketten ausgeführt wurde, können die verschiedenen Clusterzentren zu verschiedenen Rotamerzuständen oder aber auch zum gleichen Rotamer für verschiedene Rückgratgeometrien gehören. Diese Unterscheidung wird in der vergröberten Darstellung nicht aufgelöst, ebenso wie beispielsweise die unterschiedlichen Kombinationen der Rückgrattorsionswinkel. Die Darstellung der Seitenkette basierte auf einer einfachen vektoriellen Darstellung, wodurch eine aufwendigere Umrechnung zwischen kartesischen und internen Koordinaten entfiel.
Dieses Modell wurde so für die verschiedenen Anwendungen verwendet, beispielsweise zur Positionierung der Seitenketten bei der Erzeugung der falschen Proteinstrukturen (siehe Abschnitt 4.3.2) und im genetischen Algorithmus.

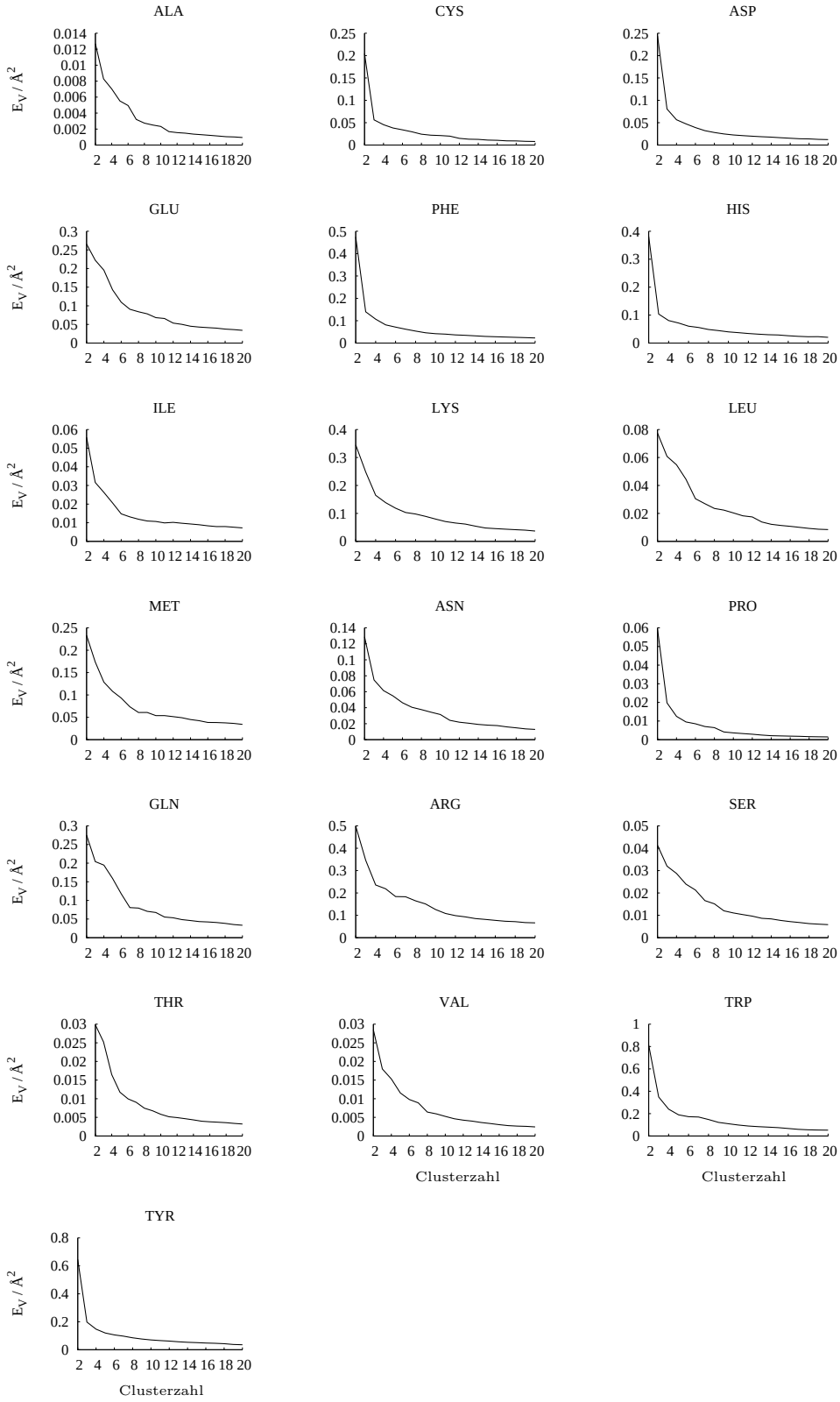


Abbildung 4.14: Erwartungswert der Varianz $E(\mathrm{Var}(k))$ in Abhängigkeit von der Clusterzahl n_c (siehe Gl. 4.18).

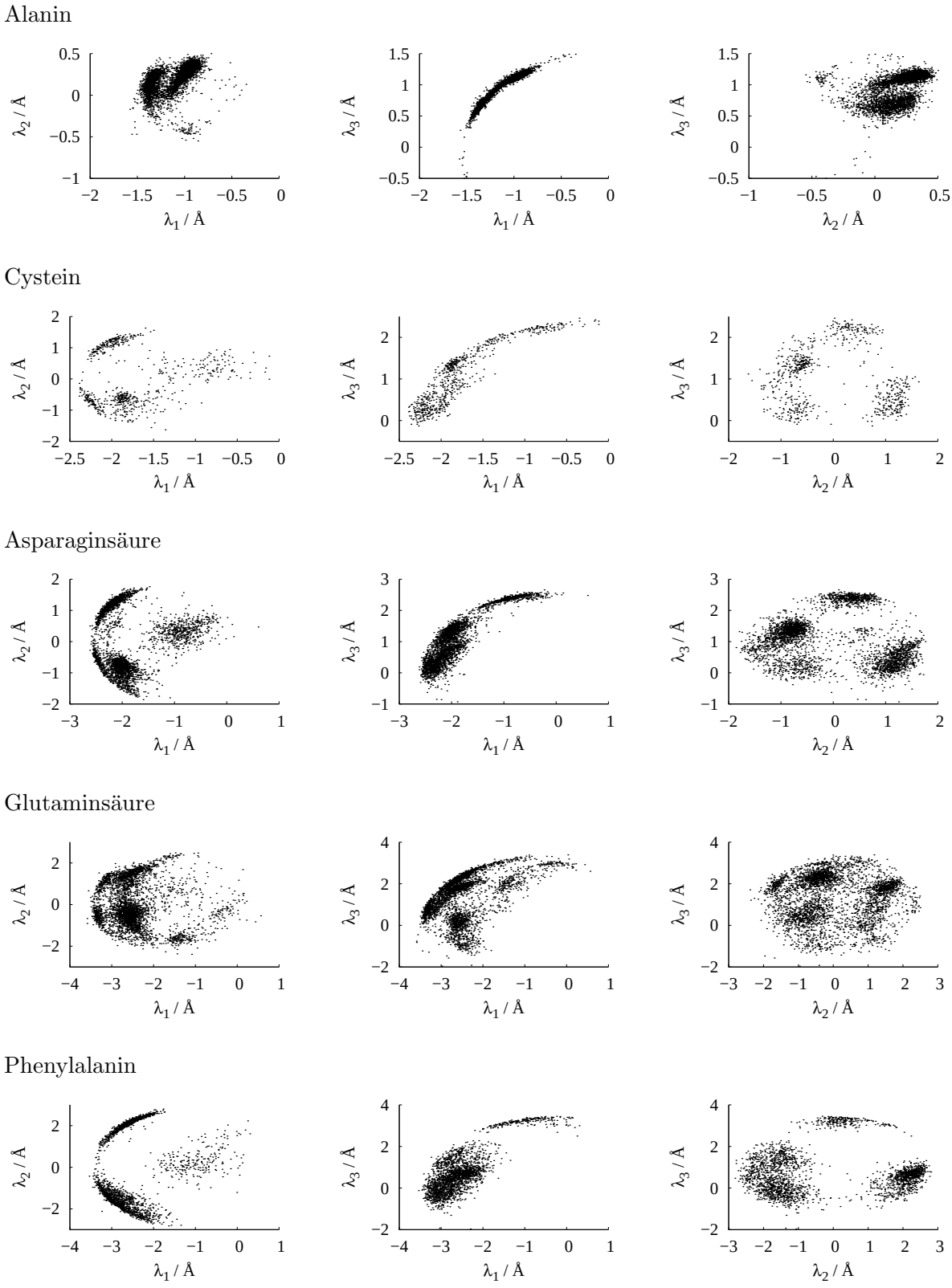


Abbildung 4.15: Lösungsmenge $\mathbf{\Lambda}_a$ (siehe Gl. 4.14) für die Aminosäuren Ala, Cys, Asp, Glu und Phe.

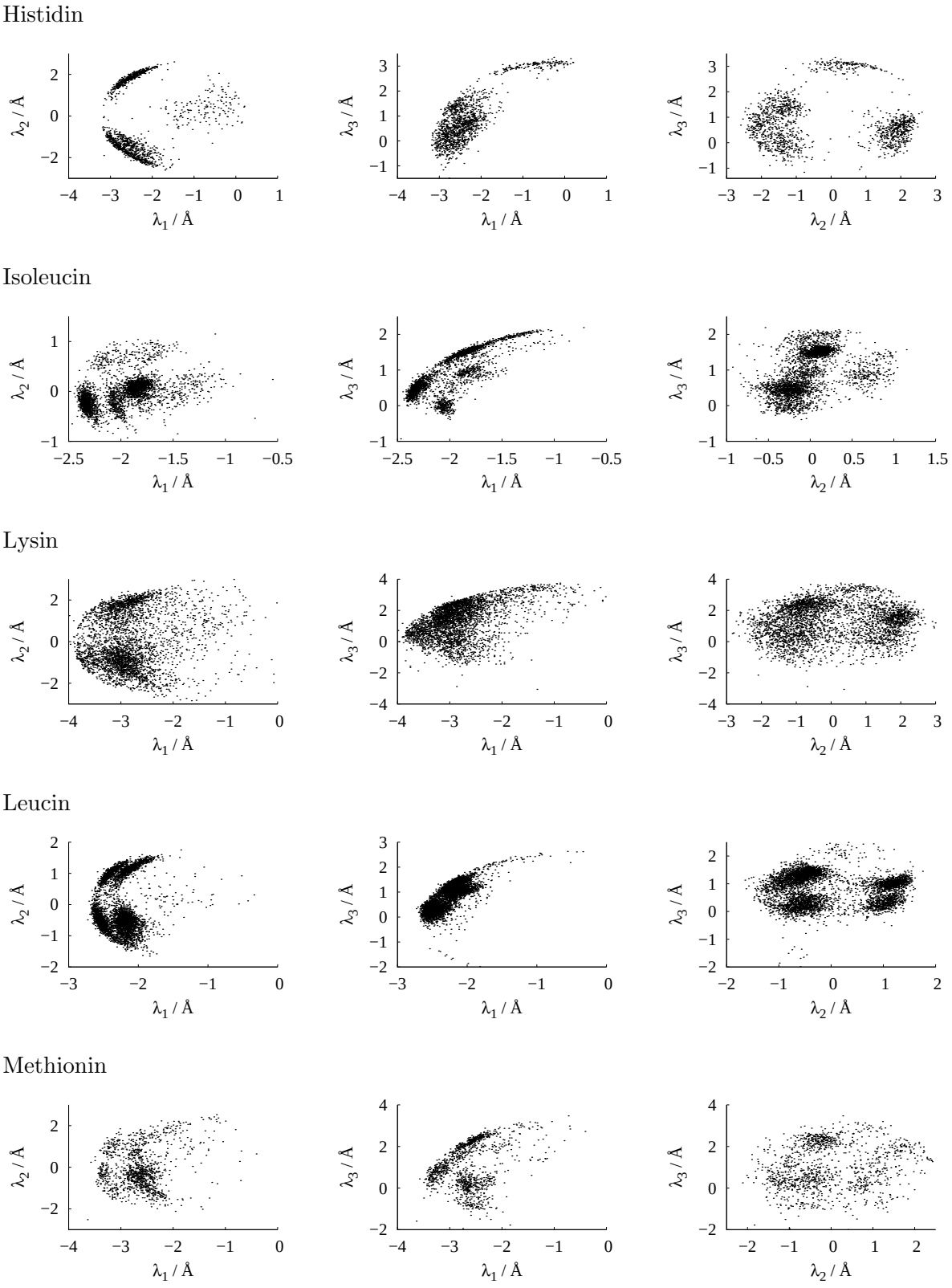


Abbildung 4.16: Lösungsmenge $\mathbf{\Lambda}_a$ (siehe Gl. 4.14) für die Aminosäuren His, Ile, Lys, Leu und Met.

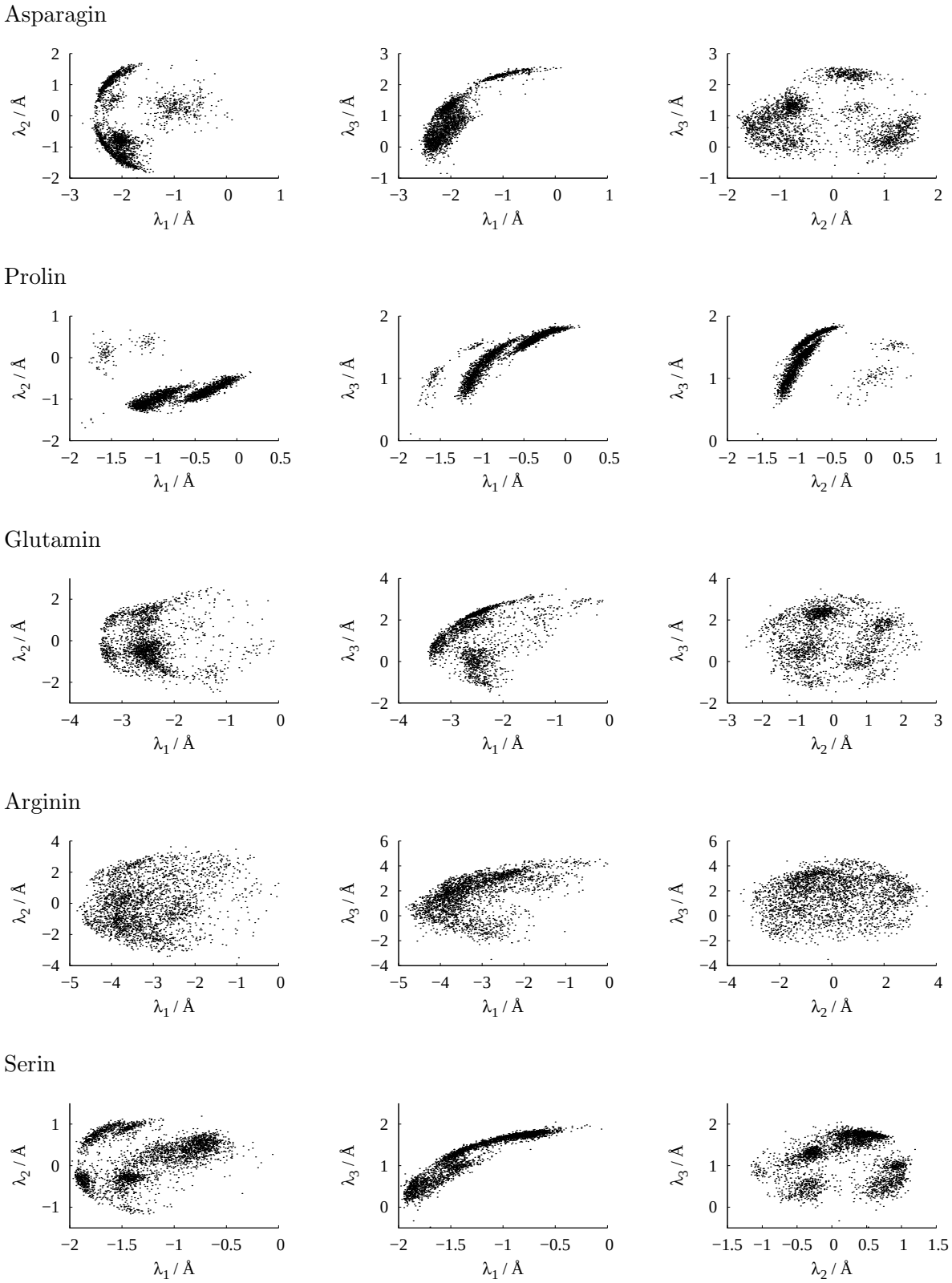


Abbildung 4.17: Lösungsmenge $\mathbf{\Lambda}_a$ (siehe Gl. 4.14) für die Aminosäuren Asn, Pro, Gln, Arg und Ser.

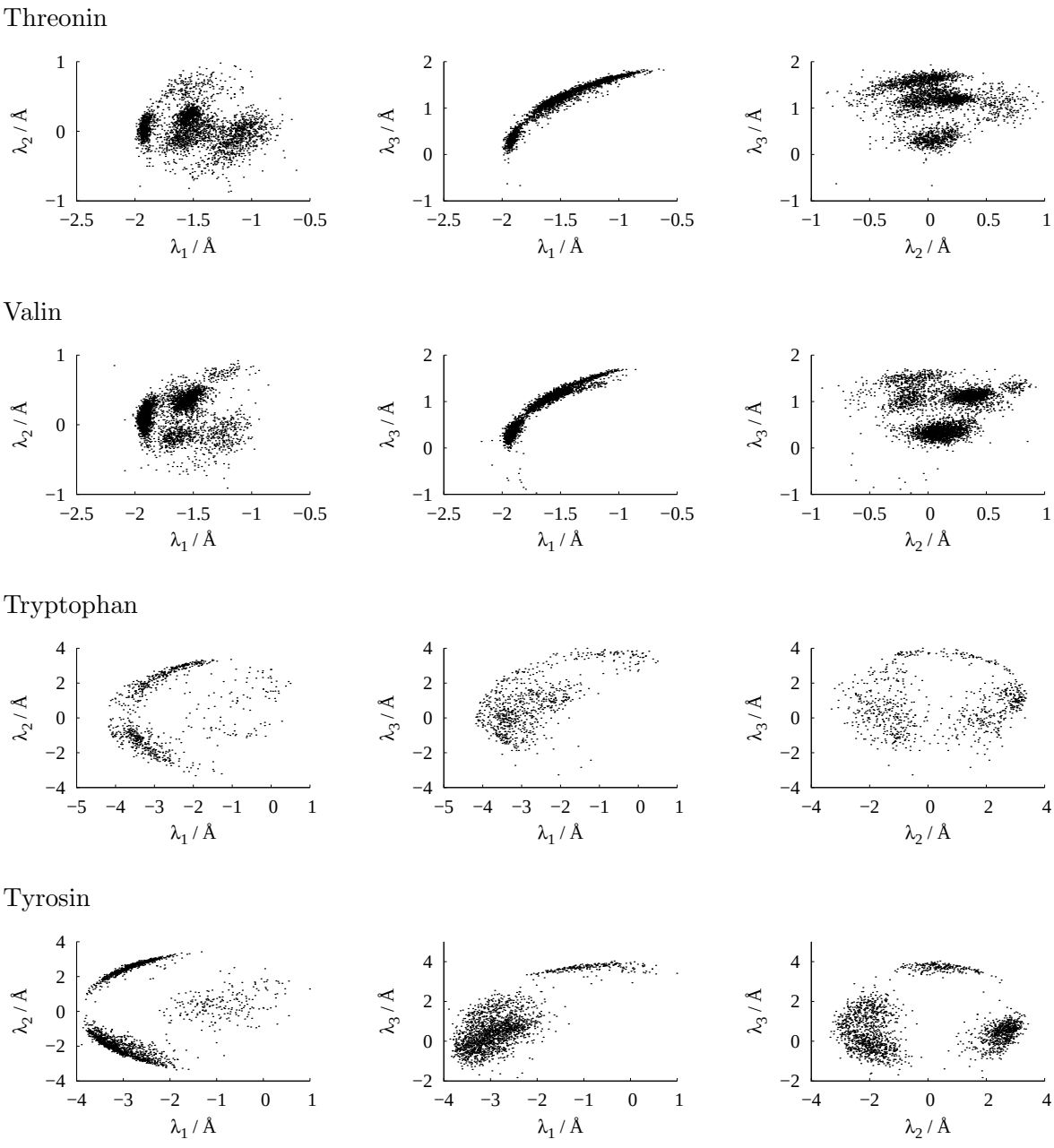


Abbildung 4.18: Lösungsmenge $\mathbf{\Lambda}_a$ (siehe Gl. 4.14) für die Aminosäuren Thr, Val, Trp und Tyr.

4.3 Auswahl der Proteine und Aminosäureklassen

Der in dieser Arbeit verwendete Ansatz zur Bestimmung der optimalen Kraftfeldparameter basierte darauf, diese Parameter für eine gewählte Bewertungsfunktion so einzustellen, dass die bekannten nativen Strukturen stets niedrigere Werte bezüglich dieser Bewertungsfunktion liefern als alle andere Strukturen. Dies erforderte zum einen einen Satz an nativen Strukturen, welche hier durch Einkristall-Röntgenstrukturen gegeben waren, die der PDB-Datenbank entnommen wurden. Andererseits benötigte man falsche bzw. nicht-native Strukturen (engl.: *decoys*) zur Parameteroptimierung. Um mit dieser Methode ein robustes Kraftfeld zu erhalten, müssen die nativen und falschen Proteinstrukturen gewisse Bedingungen erfüllen. Hierauf und auf die konkrete Auswahl der Strukturen wird im folgenden näher eingegangen.

4.3.1 Native Strukturen

Die Qualität empirischer Paarpotentiale hängt unmittelbar mit der Qualität der Strukturen zusammen, anhand derer die Parameter bestimmt wurden. Paarpotentialterme können beispielsweise an ausgewählte Molekülkoordinaten angepasst werden. Je größer die Unsicherheiten bei der Bestimmung dieser Koordinaten war, umso ungenauer werden im Regelfall auch die Ergebnisse sein, die mit dem resultierenden Paarpotential erhalten werden. Daher sind die Referenzstrukturen neben den notwendigen akkuraten Termen der Potentialfunktion das zweite wichtige Standbein eines Kraftfeldes. Für ein vergröbertes Proteinkraftfeld werden dementsprechend gut aufgelöste Proteinstrukturen benötigt. Hier ist die Röntgenstrukturanalyse die Methode der Wahl, da sie in der Regel die verlässlichsten und genauesten Strukturdaten liefert. Viele der mit dieser Methode bestimmten Strukturen werden zur Verwendung über Datenbanken bereitgestellt, wie beispielsweise die PDB-Datenbank [109, 110]. Die zweite wichtige Methode zur Strukturaufklärung neben der Röntgenstrukturanalyse ist die NMR-Spektroskopie. Für diese Arbeit wurden jedoch Proteine ausgeschlossen, die mittels NMR-Daten bestimmt wurden, um nach Möglichkeit lösungsmittelbedingte Effekte in den Geometrien auszuschließen. Zusätzlich liefert die NMR-Methode häufig mehrere Alternativstrukturen, die alle die aus den NMR-Daten erhaltenen Geometriebedingungen erfüllen. Damit ist die Eindeutigkeit des nativen Zustands nicht immer zweifelsfrei gegeben. Dies beruht zum einen auf der Methode, mit der die NMR-Daten in Geometrien konvertiert werden, und zum anderen auf eventuellen Konformationsänderungen der Proteine in der Lösung. Drittens ist nicht immer gewährleistet, dass die mit NMR-Daten erhaltenen Strukturen auch realistische bzw. physikalisch sinnvolle interatomare Abstände beinhalten. So können NMR-Protein-Modelle beispielsweise zu kurze nicht-bindende Abständen enthalten, welche sich wiederum auf die Parameteroptimierung auswirken.

Prinzipiell liefert dagegen die Röntgenstrukturanalyse sehr genaue Daten über Kristallstrukturen. Bei großen Molekülen steigt jedoch der Aufwand, um eine gute Auflösung zu erhalten. Aus diesem Grund sind in der PDB-Datenbank auch viele Proteine mit großen Massen mit einer schlechteren Auflösung enthalten (> 3 Å). Um "gute" Kraftfeldparameter zu erhalten, sollten jedoch Röntgenstrukturen mit einer möglichst guten Auflösung verwendet werden. Diese sollte idealerweiser besser als 2 Å sein. Des weiteren werden Proteinsequenzen mit möglichst niedriger Ähnlichkeit benötigt, um einen großen Sequenzraum abzudecken. Basierend auf verschiedenen Kriterien (bspw. Auflösung, R-Faktor, Kettenlänge usw.) existieren bereits fertige Listen, die eine gefilterte Auswahl der PDB-Proteine enthalten, so dass diese im günstigsten Fall nicht selber zusammengestellt werden müssen. (Siehe z. B. den PISCES-Internet-Server, der viele bereits existierende Listen zur freien Verfügung bereitstellt oder die PDB-Proteindatenbank nach den Benutzereingaben filtert und eine entsprechende Liste liefert [111, 112].) Für diese Arbeit wurde der bereits etablierte TOP500H-Satz an Proteinen verwendet [113], welcher aus einem kleineren Proteinsatz von Hobohm und Sander hervorgegangen ist [114]. Der TOP500H-Satz enthält insgesamt 500 Proteine aus der PDB-Datenbank, die mittels Röntgenstrukturanalyse bestimmt wurden. Diese Zusammenstellung wurde benutzt, da bei der Auswahl dieser Proteine mehrere Qualitätskriterien herangezogen wurden: Beispielsweise wurde bei der Auswahl beachtet, dass die Proteinsequenzen eine nur sehr geringe Sequenzähnlichkeit (Identität) besitzen, dass die Auflösung für alle Strukturen besser als 1.8 Å ist und dass nur sehr wenige nicht-bindende Überlappungen (Van-der-Waals-Kollisionen) vorkommen. Außerdem wurde der kristallographische B-Faktor einbezogen, Bindungswinkel- und Seitenkettenanalysen zur Vermeidung außergewöhnlich verzerrter Strukturen durchgeführt und weitere Kriterien angewendet. Der resultierende Proteinsatz ist im Internet frei verfügbar [115]. Diese Zusammenstellung bietet somit einen großen Satz an gut aufgelösten Proteinstrukturen. Um mit diesem Proteinsatz dem in dieser Arbeit verwendeten Kraftfeldansatz weiter gerecht zu werden, wurde dieser nach den folgenden Kriterien weiter differenziert. Ein Protein wurde dabei ausgeschlossen, wenn es aus mehreren Strängen bestand bzw. Kettenunterbrechungen enthielt (C^{α}-Abstand der Atome i und $i+1$ größer als 4.5 Å), wenn es Disulfidbrücken, Fremdatome, -moleküle oder -ionen mit Ausnahme von eingelagertem Wasser beinhaltete oder wenn Seitenketten chemisch modifiziert vorlagen. Weiterhin wurden Strukturen nicht verwendet, in denen Schweratome der Seitenkette fehlten, da diese zur Berechnung des Seitenkettenschwerpunktes notwendig sind. Fehlende Schweratome im Rückgrat mit Ausnahme von C^{α}-Atomen führten nicht zum Ausschluss, da diese im Kraftfeldansatz nicht benötigt wurden. Nach Anwendung dieser Kriterien blieben insgesamt 48 Proteine zur Parameteroptimierung übrig. Die verwendeten Proteine sind in Tabelle 4.5 mit zusätzlichen Daten aufgelistet. Die Proteingrößen lagen zwischen 50 und 402 Aminosäuren. Die Gesamtsekundärstrukturanteile der α-Helix mit 30.8 % und des β-Faltblattes mit 24.9 % sind sehr ähnlich, wodurch eine ausgeglichene Gewichtung beider Strukturtypen gewährleistet wird.

Name	Auflösung/Å	N	n_α	w_α/%	n_β	w_β/%	n_t	w_t/%	SAS/Å^2
1a12(A)	1.7	401	5	1.2	170	42.4	71	17.7	16265
1a1y(I)	1.0	63	11	17.5	14	22.2	13	20.6	4394
1a3a(D)	1.8	144	54	37.5	25	17.4	22	15.3	7686
1a92(A)	1.8	50	43	86.0	0	0.0	4	8.0	5104
1agj(A)	1.7	242	33	13.6	81	33.5	59	24.4	11019
1ako	1.7	268	79	29.5	73	27.2	40	14.9	12794
1amm	1.2	174	5	2.9	80	46.0	25	14.4	8772
1atz(A)	1.8	184	67	36.4	40	21.7	21	11.4	8396
1auo(A)	1.8	218	70	32.1	50	22.9	33	15.1	9833
1ay7(B)	1.7	89	42	47.2	16	18.0	14	15.7	4986
1bf4(A)	1.6	63	9	14.3	33	52.4	13	20.6	4478
1bfg	1.6	126	0	0.0	49	38.9	36	28.6	6561
1bgf	1.4	124	92	74.2	0	0.0	12	9.7	8072
1bkr	1.1	108	58	53.7	0	0.0	19	17.6	6178
1bm8	1.7	99	37	37.4	27	27.3	14	14.1	5525
1byi	0.9	224	79	35.3	45	20.1	31	13.8	10887
1c02(A)	1.8	166	98	59.0	0	0.0	27	16.3	8857
1cem	1.6	363	168	46.3	8	2.2	64	17.6	13036
1chd	1.7	198	57	28.8	47	23.7	39	19.7	8861
1dhn	1.6	121	39	32.2	40	33.1	6	5.0	7564
1dpt(A)	1.5	117	32	27.4	31	26.5	22	18.8	6838
1edg	1.6	380	132	34.7	44	11.6	73	19.2	15618
1elk(A)	1.5	153	101	66.0	0	0.0	18	11.8	8217
1erv	1.6	105	43	41.0	28	26.7	14	13.3	5806
1es5	1.4	260	95	36.5	56	21.5	35	13.5	10647
1fna	1.8	91	0	0.0	42	46.2	8	8.8	5464
1ftr(A)	1.7	296	63	21.3	105	35.5	55	18.6	14002
1gvp	1.6	87	0	0.0	40	46.0	12	13.8	6550
1hcr(A)	1.8	52	25	48.1	0	0.0	4	7.7	4541
1hka	1.5	158	45	28.5	40	25.3	23	14.6	8743
1ifc	1.1	131	15	11.5	77	58.8	18	13.7	7144
1iib(A)	1.8	103	50	48.5	21	20.4	17	16.5	6077
1lmb(4)	1.8	92	59	64.1	0	0.0	12	13.0	6281
1mla	1.5	305	154	50.5	45	14.8	34	11.1	12114
1mml	1.8	251	70	27.9	51	20.3	46	18.3	12896
1mol(A)	1.7	94	17	18.1	50	53.2	10	10.6	6036

1msi	1.2	66	4	6.1	8	12.1	17	25.8	3671
1nar	1.8	289	105	36.3	64	22.1	33	11.4	13206
1nkd	1.0	59	51	86.4	0	0.0	5	8.5	4616
1npk	1.8	150	56	37.3	24	16.0	28	18.7	8080
1pcf(A)	1.7	66	20	30.3	34	51.5	6	9.1	5591
1qs1(A)	1.5	402	117	29.1	113	28.1	63	15.7	19320
1qts	1.4	247	37	15.0	103	41.7	32	13.0	12941
1stn	1.7	136	33	24.3	41	30.1	27	19.9	7950
1ten	1.8	89	0	0.0	48	53.9	10	11.2	5213
1tfe	1.7	142	69	48.6	33	23.2	15	10.6	8920
1ttb(A)	1.7	127	7	5.5	58	45.7	17	13.4	6913
1tud	1.7	60	0	0.0	23	38.3	7	11.7	4336
Gesamt	1.6	165.2	51.0	30.8	41.2	24.9	25.5	15.4	8479

Tabelle 4.5: Verwendete native Strukturen. Name: PDB-Code mit der Ketten-ID in Klammern, Auflösung: Auflösung der Röntgenkristallstruktur, N: Gesamtanzahl der Aminosäuren, n_α: Reste in einer α-Helix, w_α: proz. Anteil aller Reste in einer Helix, n_β: Reste in einem β-Faltblatt, w_β: proz. Anteil aller Reste in einem Faltblatt, n_t: Reste in einer Windung/Schlaufe, w_t: proz. Anteil der Reste in einer Windung/Schlaufe, SAS: Lösungsmittelzugängliche Oberfläche. In der "Gesamt"-Zeile sind die über alle Proteine gemittelten Werte aufgeführt. Die Auflösung wurde den entsprechenden PDB-Dateien entnommen, während die anderen Daten aus den DSSP-Dateien ermittelt wurden.

4.3.2 Falsche Strukturen

Die Optimierung von Kraftfeldparametern über einen Vergleich von nativen gegen falsche Strukturen durchzuführen bzw. Ergebnisse auf diese Weise zu testen, ist ein Konzept, das seit einiger Zeit Verbreitung gefunden hat [102, 116–132]. Hierbei muss zunächst einmal der Begriff "falsch" definiert werden, da er sich auf mehrere Aspekte beziehen kann. In dieser Arbeit wird der Begriff "falsch", bezogen auf ein Protein, rein auf geometrischen Kriterien der dreidimensionalen Proteinstruktur begründet. Andere Merkmale, die die Abweichung eines Proteins zu einem anderen beschreiben, wie bspw. die Sequenz, eingelagerte Moleküle oder die Umgebung, werden hierbei nicht berücksichtigt bzw. als identisch angenommen. Somit wird eine falsche Struktur als eine Geometrie definiert, die sich von der Geometrie des nativen Zustandes unterscheidet, die aber sonst die gleiche (chemische) Zusammensetzung und Umgebung besitzt. Um diesen Unterschied zwischen zwei Proteinstrukturen zu quantifizieren, werden in der Literatur häufig zwei Merkmale benutzt: Zum einen der RMSD-Wert (*root mean squared deviation*), der die Differenz zweier N-dimensionaler Koordinatenvektoren $\mathbf{x}$ und $\mathbf{x}'$ beschreibt:

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i' - x_i)^2} \tag{4.19}$$

Dieser Wert erfordert eine direkte (bekannte) Zuordnung der Vektorelemente beider Vektoren zueinander. Diese Zuordnung ist bei Polypeptiden dadurch gewährleistet, dass nur Proteine mit gleiche Länge und gleicher Sequenz verglichen werden und dass die Numerierung der Elemente bzw. der Aminosäuren stets am N-Terminus beginnt. Weiterhin ist zu beachten, dass der RMSD-Wert weder rotations- noch translationsinvariant ist. Daher müssen, um vergleichbare Ergebnisse zu erhalten, die beiden Koordinatenvektoren vor der Berechnung des RMSD-Wertes optimal übereinandergelegt werden. Prinzipiell können zur Berechnung des RMSD-Wertes alle Atome in einem Protein herangezogen werden. Oftmals werden aber nur die Koordinaten der C^{α}-Atome verwendet, wobei ein rein auf C^{α}-Atomen basierender RMSD-Wert kurz auch als cRMSD-Wert bezeichnet wird. Die zur Berechnung des RMSD-Wertes nötige optimale Übereinanderlagerung wurde in dieser Arbeit mit dem Unterprogramm SUPERPOSE des TINKER-Programmpaketes durchgeführt [133–139].
Der zweite häufig verwendete Merkmal zur Beschreibung des geometrischen Unterschiedes ist die sogenannte Kontaktordnung. Hierbei werden üblicherweise bei Proteinen ganze Aminosäuren durch Punkte repräsentiert, für die zumeist entweder die Koordinaten der C^{α}-Atome, der C^{β}-Atome oder des Seitenkettenschwerpunkte gewählt werden. Basierend auf diesen Koordinaten wird folglich analysiert, welche und wie viele Aminosäuren in Kontakt miteinander stehen. Hierzu wird ein maximaler Radius r_{max} definiert, welcher meist im Bereich zwischen

8 und 15 Å liegt. Zwei Punkte i, j im Protein sind dann in Kontakt, wenn ihr Abstand r_{ij} kleiner als r_{max} ist. Die Kontaktordnung beschreibt somit in vereinfachter Form die Distanzen innerhalb eines Proteins und enthält damit Informationen über die dreidimensionale Gestalt des Proteins. Dieser Wert wird häufig bei Faltungsstudien als ein Maß für die Nativität eines Zustandes verwendet. Hierbei wird beispielsweise angegeben, wieviele Kontakte des nativen Zustandes zu einem bestimmten Zeitpunkt des Experiments gebildet worden sind.

Im Normalfall sind die RMSD- und die Kontaktordnungsdaten nicht miteinander korreliert, wodurch sie gleichzeitig verwendet können. Beide Werte haben zur Beschreibung der Ähnlichkeit zweier Strukturen ihre Vor- und Nachteile. Für eine Diskussion und Beispiele hierzu siehe [140] und die dort angegebenen Referenzen.

Unabhängig von der Wahl des Kriteriums, verbleibt das Problem der Definition für "falsch", d. h. die Festlegung eines Zahlenwertes, ab welchem eine Proteinstruktur als nicht mehr nativ angesehen wird. Diese Grenze ist selbstverständlich fließend und nicht eindeutig. In der Literatur haben sich für den cRMSD-Wert bestimmte Werte etabliert: So gilt ein cRMSD von Null bis drei Å als sehr gute Übereinstimmung, Werte zwischen 3 und 6 Å als strukturell ähnlich und größere Werte als strukturell entfernt [141].

Neben der Schwierigkeit der reinen Definition einer falschen Struktur besteht bei Proteinen das Problem, dass experimentell nur sehr wenige nicht-nativen Strukturen bekannt sind. Dieses beruht weniger auf der Erzeugbarkeit missgefalteter Proteine, da eine Denaturierung bereits durch kleine Veränderungen der Systemvariablen leicht zu erreichen ist, sondern viel mehr auf einem Mangel an experimentellen Methoden, die in der Lage sind, in ausreichend hoher Auflösung und gleichzeitig in kurzer Zeit Proteinstrukturen zu analysieren, da unter Normalbedingungen denaturierte Proteine schnellen Strukturfluktuaktionen unterworfen sind. Aus diesem Grund muss zur Generierung von falschen Strukturen auf computergestützte Methoden zurückgegriffen werden.
Hierzu existieren bereits einige Arbeiten in der Literatur, die die Konstruktion von falschen Strukturen im Allgemeinen oder für eine spezielle Anwendung behandeln. Viele der resultierenden Struktursätze werden nach der Fertigstellung im Internet zum freien Herunterladen angeboten. Die wichtigsten und bekanntesten Sätze (ohne Anspruch auf Vollständigkeit) sind in Tab. 4.6 zusammengestellt. Die zusätzlich bekannten Datensätze mit den Bezeichnungen "asilomar", "ifu" und "sgpa" [155] wurden nicht aufgeführt, da sie mit teilweise lediglich drei enthaltenen Strukturen extrem klein sind.
Die Methoden, die bei der Erzeugung der Struktursätze zum Einsatz kommen, decken ein sehr breites Spektrum ab. Beginnend beispielsweise bei Modellen mit diskreten Zuständen (Lattice_ssfit, 4state_reduced), über Methoden zur Rekombination von Datenbank-Struktur-

Satz-Name	Proteine	Geometrien pro Protein	Lit.	Download
4state_reduced	7	665	[102]	[142]
Abm_database	4	200	[143]	[144]
Fisa	6	1432	[145]	[142]
Fisa-Casp3	6	1432	[145]	[142]
Hg_structural	29	29	[146]	[142]
Ig_structural	61	60	[146]	[142]
Ig_structural_hires	20	19	[146]	[142]
Lattice_ssfit	8	2000	[147]	[142]
Lee	12	30	[148]	[149]
LKF	183	200	[130]	[150]
LKF2	1400	1000	[131]	[151]
Lmds	11	439	[152]	[142]
Misfold	23	26	[153]	[142]
Pdb_error	3	3	[154]	[142]
ProtG	1	7000	?	[155]
Rosetta	56	1000	[145]	[149]
Semfold	6	12900	[156]	[142]
Tsai	41	1800	[157]	[149]
Vhp_mcmd	1	6255	[158]	[142]
Wang	6	200	[159]	?

Tabelle 4.6: Übersicht über die bekanntesten Sätze mit falschen Strukturen. Die angegebene Anzahl an falschen Geometrien pro Protein ist ein Richtwert und kann innerhalb des gleichen Datensatzes für die darin enthaltenen Proteine verschieden sein.

fragmenten (z. B. Rosetta, Fisa), über Sequenzübertragung (z. B. Misfold) bis hin zu Molekülдynamik- (z. B. Wang), Monte-Carlo-Simulationen und genetischen Algorithmen (z. B. ProtG). Auch die Anzahl an verwendeten nativen Proteinen und die Anzahl an erzeugten Strukturen ist zwischen den Sätzen sehr unterschiedlich, ebenso wie deren Qualität. Hierbei es dem Benutzer selbst überlassen zu entscheiden, ob der Struktursatz für ein Problem bzw. eine Anwendung geeignet ist. Viele der veröffentlichten Strukturen enthalten Fehler oder basieren auf experimentell weniger gut aufgelösten nativen Strukturen, so dass es praktisch unerlässlich ist, vor der Verwendung eines dieser Sätze eine eigenständige Analyse der Strukturen durchzuführen.

Hierzu ist eine von D. Gilis publizierte Arbeit hilfreich [140], in der insgesamt 12 der in Tab. 4.6 aufgelisteten Struktursätze untersucht wurden. Ziel war es, aus den verwendeten Sätzen die qualitativ minderwertigen Strukturen herauszufiltern und einen neuen Gesamtsatz herzu-

stellen, wobei die folgenden Kriterien zur Analyse herangezogen wurden: a) die Qualität der Strukturbestimmungsmethode der nativen Struktur, wodurch bspw. mittels NMR bestimmte Proteine ausgeschlossen wurden, b) die Vollständigkeit der Atome und c) die Übereinstimmung der Sequenz bzw. der Sequenzlänge der erzeugten falschen Struktur im Vergleich mit der nativen. Interessanterweise fallen durch das letzte Kriterium bereits sehr viele Strukturen weg, da in den untersuchten Sätzen viele falsche Strukturen eine vom nativen Protein abweichende Sequenzlänge besitzen, was nach deren Erzeugung nicht korrigiert wurde. Dies betrifft vor allem die falschen Strukturen, die mittels Rekombinationen von Datenbankfragmenten erzeugt wurden, weil die Fragmente häufig auf bestimmte Längen beschränkt wurden.

Es resultierte ein Satz mit 45 Proteinen und 47301 falschen Strukturen. Dieser wurde in dieser Arbeit anhand weiterer Kriterien, die in Abschnitt 4.3.1 für die nativen Proteine beschrieben sind, analysiert, beispielsweise in Bezug auf die Auflösung der Röntgenkristallstruktur, auf Kettenunterbrechungen oder Disulfidbrücken. Nach dieser blieben weniger als 10000 falsche Strukturen aus dem von D. Gilis publizierten Satz übrig. Diese Zahl war in Übereinstimmung mit der Literatur [124, 125] zu klein für eine effektive Parameteroptimierung. Außerdem war die Schnittmenge mit dem verkleinerten TOP500H-Datensatz an hoch aufgelösten nativen Proteinen sehr klein. Aus diesen Gründen wurden die publizierten falschen Strukturen bis auf eine Ausnahme (s. u.) nicht verwendet und stattdessen eigene Programme zu Erzeugung falscher Strukturen programmiert, die im folgenden eingehender beschrieben werden. Folgende Ziele standen bei der Erzeugung im Vordergrund:

- Generierung von Strukturen sehr nah an der nativen Struktur. Dieser Faktor ist wichtig, um Parameter zur Bestimmung der richtigen Struktur in einem Satz sehr ähnlicher Proteine dicht am nativen Zustand zu erhalten.

- Erzeugung von sehr kompakten aber falschen Strukturen mit nativ-artigen Strukturelementen. Also Geometrien, die Baueinheiten aus bekannten nativen Proteinen enthalten, die aber falsch zueinander angeordnet sind bzw. nicht zu der Zielsequenz passen.

- Unabhängigkeit von einem bestimmten Kraftfeld, um hierdurch intrinsische Präferenzen des Kraftfeldes zu vermeiden. Beispielsweise tendieren einige Kraftfelder zur Übergewichtung von α-Helices [160].

Durch die Generierung eigener falscher Strukturen konnte die maximale Anzahl an Proteinen ausgenutzt werden, die das Programm zur Bestimmung der Parameter zulässt. Diese Zahl ist zwar von der Anzahl der zu optimierenden Parameter abhängig, dennoch war die Menge der verwendeten falschen Strukturen in dieser Arbeit trotz relativ vieler Parameter sehr viel größer als in vielen anderen publizierten Arbeiten und sehr viel größer als die ca. 10000 übriggebliebenen Strukturen nach der Gilis-Analyse. Zudem konnte sicher gestellt werden,

dass die Strukturen keine unerwünschten Fehler enthielten und dass sie auf hoch aufgelösten Röntgenstrukturen basierten.

Die Erzeugung der falschen Strukturen wurde mit unterschiedlichen Techniken realisiert: Grundlegend sei hier zunächst angemerkt, dass die Geometrie der falschen Geometrien ebenfalls nur über die C^{α}-Atome und die Seitenkettenschwerpunkte modelliert wurde. Die Seitenkettendarstellung beruhte auf dem Cluster-Ansatz, der in Abschnitt 4.2.3 dargelegt wird. Generell wurde bei der Generierung eines falschen Proteins zunächst die C^{α}-Rückgratstruktur verändert und berechnet, da sie bzw. die C^{α}-Bindungsvektoren die notwendige Basis zur Positionierung der Seitenketten ist. Im zweiten Schritt wurden die Seitenkettenschwerpunkte auf zufällige Clusterpositionen gesetzt. Wenn es zu Kollisionen mit anderen Seitenketten oder C^{α}-Atomen kam, wurden die anderen nicht verwendeten Clusterpositionen getestet. Wenn dennoch keine kollisionsfreie Anordnung gefunden werden konnte, wurde die Struktur verworfen. Für die Kollisionstests wurden die aus den Datenbankdaten ermittelten minimalen Seitenkette-zu-Seitenkette- und Seitenkette-zu-Rückgrat-Abstände verwendet (siehe im Anhang Tab. 6.1 und 6.2).
In allen im folgenden beschriebenen Routinen war der Schritt der Kollisionstests zusammen mit der Platzierung der Seitenketten jeweils der mit Abstand zeitaufwendigste Programmteil bei der Ausführung.

Die ersten erstellten Routinen wählten zunächst eine zufällige Anzahl sequentieller C^{α}-Atome aus, wobei die für die Auswahl maximale Anzahl proportional zur Gesamtlänge des Proteins war, aber höchstens 50 Aminosäuren umfasste. Anschließend wurde aus einer Liste eine Veränderungsoperation für diesen Abschnitt zufällig gewählt. Hierbei waren folgende Operationen möglich: Translation des Abschnittes entlang eines Vektors, Rotation des Abschnittes um einen Winkel mit einer beliebig orientierten Rotationsachse, Rotation des Abschnittes um den Vektor, der das erste und das letzte ausgewählte C^{α}-Atom verbindet, oder schließlich Neufaltung des Abschnittes. Für die in internen Koordinaten durchgeführte Neufaltung wurden unterschiedliche Methoden verwendet, wobei aber stets nur die C^{α}-Bindungswinkel und -Torsionswinkel verändert wurden, während die Abstände sequentieller C^{α}-Atome unverändert blieben. Durch die erste Neufaltungsmethode wurden die neuen Werte für die Winkel entweder vollständig zufällig oder aus einer Liste mit bekannten Sekundärstrukturen gewählt, welche im ersten Ansatz nur die α-Helix, das β-Faltblatt und einige idealisierte Windungen enthielt. In der zweiten Methode zur Neufaltung wurde dagegen zunächst der Abstand des ersten und letzten C^{α}-Atoms des gewählten Abschnittes bestimmt. Mit dieser wurde danach aus einem Satz an Datenbank-Proteinstrukturen ein anderer Abschnitt zufällig gewählt, der die gleiche Anzahl an C^{α}-Atomen und den gleichen Abstand vom ersten zum letzten C^{α}-Atom hat, wobei der Abstand zusätzlich um einen Toleranzbereich erweitert wurde, um eine grö-

ßere Menge andersartig gefalteter Abschnitte zuzulassen. Nach der Auswahl eines passenden Fragmentes wurde versucht, dieses konfliktfrei in die Struktur des falschen Proteins anstelle des gewählten Abschnittes zu übernehmen, wobei nur die Koordinaten der C^{α}-Atome übernommen wurden. Dazu wurden die beiden Endpunkte des neuen Fragmentes zunächst mit dem SUPERPOSE-Programm mit den ursprünglichen Positionen der Endpunkte überlagert. Eventuelle geometrische Diskrepanzen, die häufig nach der Überlagerung zwischen den alten und neuen Endpunkten auftreten können, da eine andere Faltung im Normalfall einen anderen Endpunkt-zu-Endpunkt-Abstand mit sich bringt, wurden unter Verwendung des FCCD-Algorithmus geschlossen [161, 162]. Nach den C^{α}-Atomen wurden die Seitenketten gesetzt und die Kollisionstests durchgeführt.
Diese Methoden, Segmente relativ zum Rest des Proteins zu verschieben, umzuorientieren und/oder umzufalten, um kompakte, falsche, aber native Motive enthaltende Strukturen zu erhalten, zeigten insgesamt betrachtet nicht die erwünschten Resultate und waren zudem sehr langsam, weil nahezu alle neuen Strukturen aufgrund von Kollisionen verworfen werden mussten. Problematisch war hier vorallem die kompakte Packung der Aminosäuren und Seitenketten im inneren des Proteins. Bei diesen besteht nur ein sehr geringer Spielraum, längere Segmente geometrisch neu anzuordnen, wobei vor allem die Seitenketten die Veränderungen einschränken. Geometrien ohne Seitenketten ließen sich auf diese Weise relativ gut verändern, passten aber im Rahmen dieser Arbeit nicht in das gewünschte Modell.
Aufgrund dieser Ergebnisse wurden andere, einfachere Strukturerzeugungsmethoden entwickelt. Erstens wurde die Methode der Sequenzübertragung verwendet. Hierbei wird die Sequenz des nativen Zielproteins über die Sequenz eines anderen Datenbank-Proteins geschrieben. Dabei musste das zu überschreibende Datenbank-Protein verschiedenen Anforderungen genügen: Es durfte nicht Teil des Optimierungssatzes sein, die Sequenzen durften eine maximale Identität von 30 % aufweisen, der CRMSD-Wert zum Zielprotein musste ≥ 6 Å sein, es durfte keine Kettenunterbrechungen beinhalten und es musste selbstverständlich gleich viele oder mehr Aminosäuren besitzen wie das Zielprotein. Der Vorgang der Sequenzübertragung wird in Gl. 4.20 dargestellt: Die Sequenz $\mathbf{S}^*$ des nativen Proteins mit N Aminosäuren wird auf die Sequenz $\mathbf{S}$ eines anderen Proteins übertragen ($\rightarrow$), welches die Länge M hat ($N \leq M$). Ist $M > N$ so wird die Sequenz des nativen Proteins nur auf einen ausgewählten Teil der Länge N des anderen Sequenzvektors übertragen und nur dieser verwendet. Bestimmt wird dieser Abschnitt durch die Wahl von i in Gl. 4.20.

$$\mathbf{S}^* = (s_1^*, s_2^*, ..., s_N^*) \rightarrow (s_{1+i}, s_{2+i}, ..., s_{N+i}) \quad \text{mit} \quad 0 \leq i \leq M - N \tag{4.20}$$

Dies bedeutet, dass aus einem nativen und einem Datenbank-Protein insgesamt $M - N$ neue falsche Proteine erhalten werden können. Allerdings sind die Strukturen, die man für i und $i+1$ (aus Gl. 4.20) oder allgemein für i und $i+j$ mit einem kleinen $j \in \mathbb{N}$ erhält, sehr ähnlich, da sich nur die letzten und die ersten Aminosäuren unterscheiden. Aus diesem Grund wurde

bei der Implementierung, weil eine große Anzahl an Datenbankproteinen zur Sequenzübertragung zur Verfügung stand, und um sehr ähnliche Strukturen mit geringem Informationsunterschied zu vermeiden, i als Vielfaches von fünf gewählt: $i \in \{0, 5, 10, 15, ...\}$.
Während die Sequenzübertragung für die Rückgratatome unproblematisch ist, können dagegen Komplikationen bei der anschließenden Positionierung der Seitenketten auftreten. Diese müssen in der Regel neu orientiert werden, da sich nach der Sequenzübetragung die Umgebungen der einzelnen Seitenketten vollständig ändern. Problematisch ist hierbei, dass die überschriebenen Datenbankstrukturen auf ihre ursprünglichen Aminosäuresequenzen optimiert waren. Daher kann der Fall eintreten, dass nach der Sequenzübertragung die neue Seitenkettensequenz mit der Geometrie, die durch das Rückgrat vorgegeben wird, unvereinbar ist, beispielsweise wenn kurze Seitenkette wie Glycin oder Alanin in kompakten Strukturen durch Seitenketten mit großer Raumerfüllung wie Lysin oder Arginin ersetzt werden. Um aus diesem Grund bei der Positionierung der Seitenketten größere Möglichkeiten zu haben, wurden sie zunächst auf eine zufällig gewählte Clusterposition gesetzt und danach parallel zu einem beliebig orientierten Vektor verschoben, unter der Einschränkung, dass der Abstand zum zugehörigen C^{α}-Atom erhalten bleiben musste. Die Länge dieses Vektors wurde mittels einer um Null zentrierten Gaussverteilung bestimmt, deren Breite durch die Standardabweichung der Distanzen der zum gewählten Clusterzentrum gehörenden Punkte gegeben war, die durch den Clusterzentren-Bestimmungsprozess erhalten wurde (siehe Abschnitt 4.2.3 und Anhang 6.2). Die maximale Länge des Vektors wurde zusätzlich noch auf den Mittelwert dieser Distanzen begrenzt. Nach dem Überschreiben der Sequenz und dem Setzen der Seitenkettenschwerpunkte wurden die Kollisionstests durchgeführt.
Mit dieser Methode konnten sehr viele falsche Strukturen mit kompakter, nativer Faltung erhalten werden, da die überschriebenen Strukturen ebenfalls native Proteine waren. Diese waren in der Regel strukturell sehr weit von der Geometrie des Zielproteins entfernt. Nachteilig war bei diesem Verfahren, neben einer nicht-optimalen Anordnung der Seitenketten, dass die dreidimensionale Struktur des überschriebenen Proteins Lücken bzw. Hohlräume enthalten konnte. Diese konnten dadurch entstehen, dass häufig nicht das vollständige überschriebene Protein verwendet werden konnte, wenn es eine längere Sequenz als das Zielprotein besaß, so dass bestimmte Sequenzabschnitte ausgelassenen werden mussten. Solche Proteine könnten in einem Folgeschritt mit einem etablierten Kraftfeld lokal optimiert werden. Dieser Schritt wurde hier nicht durchgeführt, da per Konstruktion Kollisionen vermieden wurden und diese Proteine insgesamt sehr weit vom nativen Zustand entfernt sind, wodurch sie in der Regel selbst nach einer Relaxierung unkritisch für die Parameteroptimierung sind [122].
Mit der Methode der Sequenzübertragung ließen sich weite Bereiche des Konformationsraumes abdecken, die nicht in der Nähe das gewünschten nativen Zustandes lagen. Daher wurde eine weitere Methode implementiert, um Strukturen zu erhalten, die dem nativen Zustand des Zielproteins sehr ähnlich waren.

Hierzu wurde direkt von der Geometrie des Zielproteins ausgegangen und diese in kleinen Schritten verändert. Zunächst wurde eine beliebige Position entlang der Sequenz ausgewählt und für diese im Anschluss entschieden, ob entweder nur die Position der Seitenkette oder ob sowohl die Position des C^{α}-Atoms wie auch die zugehörige Seitenkettenposition verändert wird. Bei einer Verschiebung der Seitenkette alleine wurde zunächst bestimmt, welchem Clusterzentrum der Schwerpunkt in der unveränderten Geometrie am nächsten liegt. Dann wurde die Seitenkette zu einem anderen Clusterzentrum positioniert und parallel zu einem zufällig orientierten Vektor verschoben. Wenn das C^{α}-Atom und die Seitenkette gewählt wurden, wurde zunächst die Position des C^{α}-Atoms um die Achse $\mathbf{v}_{rot}$ rotiert, die durch das in der Sequenz vorausgehende und durch das folgende C^{α}-Atom geht: Sei i das zu rotierende C^{α}-Atom, so ist die Rotationsachse $\mathbf{v}_{rot}$ definiert als $\mathbf{v}_{rot} = \mathbf{x}_{i+1} - \mathbf{x}_{i-1}$ mit den C^{α}-Koordinaten $\mathbf{x}_k$. Diese Rotationsachse wurde gewählt, da hierbei die Abstände zu C^{α}_{i-1} und C^{α}_{i+1} erhalten bleiben. Bei einer beliebigen Verschiebung des C^{α}_i-Atoms entstehen ansonsten viele Strukturen, die zu kurze Abstände zu den direkt benachbarten Atomen besitzen, wodurch diese Strukturen wieder verworfen würden. Nach der kollisionsfreien Drehung des C^{α}_i-Atoms mussten die Seitenketten der Aminosäuren $i-1$, i und $i+1$ neu berechnet werden, da das Seitenkettenmodell auf den beiden Bindungsvektoren zwischen den Atomen C^{α}_{i-1}, C^{α}_i und C^{α}_{i+1} beruhte. Ohne eine Neubestimmung der Seitenketten am C^{α}_{i-1}- und C^{α}_{i+1}-Atom können dort physikalisch unmögliche Geometrien entstehen. Hierbei wurde die i-te Seitenkette wieder an eine zufällige Clusterposition gesetzt und entlang eines Vektors verschoben. Die anderen beiden Seitenketten wurde so gesetzt, dass ihre Position relativ zu den Bindungsvektoren vor der Verschiebung des C^{α}_i-Atoms erhalten blieb. Dazu wurden vor der Verschiebung die Koeffizienten der drei Basisvektoren mittels eines linearen Gleichungssystem bestimmt, die zusammen die Seitenkettenposition erzeugen. Nach der Verschiebung des C^{α}_i-Atoms wurden diese Koeffizienten mit den neuen Basisvektoren verwendet, um die neue Seitenkettenposition zu erhalten. Wenn es hierbei aber zu Kollisionen kam, wurde wie bei der i-ten Seitenkette verfahren und andere Clusterpositionen getestet.

Diese Änderungen der C^{α}- und Seitenkettenkoordinaten wurden so lange durchgeführt bis die Differenz zwischen der Ausgangsstruktur und der veränderten Struktur größer als ein zu Anfang des Programmes festgelegter Wert war. Für vergleichbare RMSD-Werte müssen die Strukturen zunächst ideal übereinandergelegt werden. Das Übereinanderlegen kann jedoch dazu führen, dass schon kleine Änderungen in der Struktur zu größeren RMSD-Werten führen, da beispielsweise das gesamte Protein rotiert werden muss. Dies kann zu der Fehlbewertung führen, dass sich die Strukturen scheinbar in größerem Maße unterscheiden, während es in Wahrheit lediglich ein Effekt der Überlagerung ist. Daher wurde für das Abbruchkriterium auf die Überlagerung der Strukturen verzichtet und die Differenz in den Koordinaten direkt ausgewertet.

Der Wert D, der das Abbruchkriterium repräsentierte und die mittlere Differenz jeder Aminosäure zur originalen Struktur beschrieb, hatte die Form:

$$D = \frac{1}{N-2} \sum_{j=2}^{N-1} \left[\|\mathbf{x}_j^* - \mathbf{x}_j'\| + \|\boldsymbol{\rho}_j^* - \boldsymbol{\rho}_j'\| \right] \tag{4.21}$$

wobei N die Anzahl der Aminosäuren ist, $\mathbf{x}_j^*$ bzw. $\boldsymbol{\rho}_j^*$ die dreidimensionalen Ortsvektoren der C^α- bzw. der Seitenkettenschwerpunkte der nativen Struktur und die $\mathbf{x}_j'$ und $\boldsymbol{\rho}_j'$ die Koordinaten der veränderten Struktur. Wenn keine Kollisionen auftraten und D größer als der vorgegebene Wert war, wurde die veränderte Struktur akzeptiert. Je kleiner D gewählt wird, umso ähnlicher ist die falsche Struktur zur nativen. Das Problem hierbei ist, dass, wenn D zu klein gewählt wird, das Parameteroptimierungsproblem nicht mehr lösbar sein kann, da die Strukturen zu ähnlich sind, um im gewählten Kraftfeldansatz unterscheidbar zu sein. Daher wurden mit dieser Methode unterschiedliche Sätze von falschen Strukturen erzeugt und getestet, bis zu welchem D das Problem noch lösbar war. Als kleinster Wert für D wurde 0.1 Å gewählt und in Schritten von 0.1 Å vergrößert.

Mit diesen beiden Sätzen selbst erzeugter falscher Strukturen konnten sowohl nah-native wie auch weit entfernte Bereiche des Konformationsraumes der Proteine erfasst werden, wobei trotzdem aufgrund der sehr großen konformatorischen Freiheit eines durchschnittlich langen Proteins (100 bis 300 Aminosäuren) der gesamte Konformationsraum mit den aktuell begrenzten Computerkapazitäten nicht abgedeckt werden kann. Insgesamt wurden zu jedem der 48 Proteine aus dem reduzierten TOP500H-Satz bis zu 10000 falsche Strukturen mittels Sequenzübertragung und 10000 mittels Strukturverzerrung erzeugt.
Zu diesen eigenerzeugten Proteinstrukturen wurde noch ein Satz falscher Strukturen aus der Literatur hinzugefügt. Hierbei handelte es sich um den LKF2-Satz (siehe Tab. 4.6), der in der Analyse von D. Gilis nicht miteinbezogen war. Dieser bietet pro Protein 1000 falsche Strukturen, die durch Reorganisation von Strukturfragmenten erzeugt wurden und wenige Fehler enthalten, wie sich nach der selbst durchgeführten Analyse zeigte. Diese Strukturen können, bezogen auf die Differenz zum nativen Zustand, zwischen die durch Verzerrung und die durch Sequenzübertragung erzeugten Strukturen eingeordnet werden. Verwendet wurde aber nicht der gesamte LKF2-Satz, sondern nur die mit dem reduzierten TOP500H-Satz gemeinsamen Proteine.
Auf der Basis dieser drei Sätze mit falschen Strukturen wurden die ersten Parameter des Kraftfeldes optimiert. Mit diesen Parametern wurden dann ausgewählte Sequenzen global optimiert und dabei erhaltene Proteine, die eine niedrigere Energie als das native Protein besaßen, als neue falsche Strukturen für eine weitere Parameteroptimierung verwendet, so dass schließlich insgesamt vier verschiedene Sätze mit sehr unterschiedlichen Geometrien zur Verfügung standen. Die Parameteroptimierung konnte folglich an insgesamt 48 nativen Proteinen und mehr als 700000 falschen Strukturen durchgeführt werden.

4.3.3 Aminosäureklassen

Die in der Natur vorkommenden Proteine enthalten zwanzig verschiedene Aminosäuren (und einige weitere seltenere), deren Sequenzen die verschiedenen Faltungen und Funktionen der Proteine bestimmen. Dieser Satz an Molekülbausteinen ist im Vergleich zu der zugänglichen Menge an möglichen organischen Verbindungen zum Aufbau eines Proteins geradezu klein. Dennoch ist man sowohl von der experimentellen [163] wie auch von der theoretischen [164] Betrachtungsseite bestrebt, diesen Satz an Aminosäuren durch eine Einteilung in bestimmte Klassen weiter zu vereinfachen. Die Gründe hierfür sind vielfältig und reichen von der Vergrößerung des Grundlagenwissens über die Zusammenhänge der Eigenschaften der Aminosäuren untereinander, die beispielsweise auf ähnliche physikalisch-chemische Eigenschaften zurückgeführt werden können, über ein besseres Verständnis der Faltung und der Organisation innerhalb eines Proteins bis hin zu möglichen Erkenntnissen über die Evolution [165]. Auf der anderen Seite besitzt die Klassifizierung einen praktischen computertechnischen Aspekt. Numerische Simulationen von Proteinen oder anderen Verbindungen basieren zumeist auf einer mathematischen Grundlage, die Funktionen verwendet, die in irgendeiner Form spezifisch für die enthaltenen Atome oder Bausteine des simulierten Objektes sind. Die Komplexität des zugrundeliegenden Beschreibungsmodells hängt somit auch von der Anzahl der unterschiedenen Bausteine ab. Im Falle eines vergröberten Proteinmodells bzw. -potentials entspricht dies der Größe des verwendeten Aminosäurealphabets. Daher lässt sich durch eine Reduktion der Anzahl an unterschiedenen Aminosäuren eine numerische Simulationen effizienter gestalten. Hierbei muss selbstverständlich berücksichtigt werden, dass eine Reduktion der Anzahl an Aminosäuren zwar rechentechnische Vorteile bieten kann, dies aber gleichzeitig auch mit einem Verlust an Genauigkeit und Detailinformation verbunden ist.

Die Anzahl an Basisfunktionen hängt für das in dieser Arbeit angesetzte Potential ebenfalls von der Größe des Aminosäurealphabets ab. Wie weiter unten in den folgenden Abschnitten ausführlich gezeigt wird, bedingt aufgrund der technischen Beschränkungen eine größere Zahl an Basisfunktionen eine kleinere Zahl an falschen Strukturen, die für die Optimierung der Parameter benötigt werden. Die Anzahl an Basisfunktionen skaliert hierbei teilweise quadratisch mit der Größe des Aminosäurealphabets. Aus diesem Grund wurde das Ziel angestrebt, die Anzahl an unterschiedenen Aminosäuren so klein wie möglich zu wählen, um eine möglichst große Anzahl an falschen Strukturen verwenden zu können, gleichzeitig aber auch eine genügende Anzahl an Aminosäureklassen zur Verfügung zu stellen, um eine akkurate Beschreibung und Diversifizierung der Wechselwirkungen zu gewährleisten.
Da in der Literatur bereits verschiedene Einteilungen bekannt sind, wurde auf diese zurückgegriffen. Sowohl die Methoden, die zur Klassifizierung der Aminosäuren verwendet wurden, wie auch die resultierenden Gruppen von Aminosäuren sind hierbei sehr unterschiedlich. Die

Größe der Alphabete reicht von einem vollständigen 20-Buchstaben-Satz bis zu einer Einteilung, die nur zwei Buchstaben enthält und beispielsweise lediglich hydrophobe und hydrophile Aminosäuren unterscheidet. Zwischen diesen existieren viele verschiedene Gruppierungen, die z. B. fünf, sieben oder zehn unterschiedliche Klassen enthalten. In Tab. 4.7 ist eine Übersicht über bekannte Arbeiten zur Aminosäureklassifikation gegeben. Aus Anwendersicht ist eine Auswahl eines bestimmten Alphabets schwierig, weil zum einen sehr unterschiedliche Techniken bei der Einteilung angewendet wurden, was die Vergleichbarkeit beeinträchtigt, und weil zumeist mit einer Technik nicht nur ein einziges Alphabet erstellt wurde, sondern mehrere verschiedene mit einer unterschiedlichen Anzahl an Klassen. Ebenso kommen verschiedene Autoren bei einer gleichen Anzahl an Klassen zu einer unterschiedlichen Besetzung dieser durch die Aminosäuren. Zusätzlich werden in der Literatur von verschiedenen Autoren abweichende Zahlen dafür genannt, wieviele Aminosäureklassen für ein Proteinmodell ausreichend sind bzw. benötigt werden. Diese Zahlen reichen von zwei [164] bis zu mehr als zehn [166] Aminosäureklassen und können je nach gewähltem Modell unterschiedlich erfolgreich sein.
Aus diesen Gründen wurden im Programm mehrere unterschiedliche Aminosäurealphabete implementiert, die vom Benutzer ausgewählt werden können. Implementiert wurden die folgenden Alphabete: a) je ein Alphabet mit zwei bzw. drei Klassen, eingeteilt nach physikochemischen Eigenschaften der Seitenkette [82], b) die Einteilungen von Cieplak *et al.* [165], Wang *et al.* [167] und von Chang *et al.* [168], die alle jeweils fünf Klassen umfassen, sowie c) die acht Klassen umfassenden Alphabete ebenfalls von Wang *et al.* und Chang *et al.*
In dieser Arbeit wurden bei der Erstellung des Kraftfeldes nur die Klassifizierung nach Wang *et al.* verwendet, die insgesamt acht Klassen enthält.
Während der Arbeit wurde mit den unterschiedlichen Gruppierungen experimentiert, wobei sich zeigte, dass, wenn falsche Strukturen verwendet werden, die sehr ähnlich dem nativen Zustand sind, eine größere Anzahl an Klassen notwendig ist, um diese Strukturen von der nativen unterscheiden zu können. Bei einem zu kleinen Alphabet konnte die Parameter-Optimierung nicht durchgeführt werden bzw. sie endete ohne die für die Parameter-Optimierung gestellten Bedingungen erfüllen zu können (siehe hierzu Abschnitt 4.5). Das schließlich verwendete Acht-Buchstaben-Alphabet ist in Tab. 4.8 dargestellt. Zum Vergleich ist ebenfalls das Fünf-Buchstaben-Alphabet, welches mit der gleichen Methode erstellt wurde, in dieser Tabelle enthalten.

Klassifizierungstechnik	Jahr	Lit.
Sekundärstrukturpräferrenzen	1976	[169]
Mutationswahrscheinlichkeiten	1978	[170]
Mutationswahrscheinlichkeiten und physiko-chem. Eigenschaften	1986	[171]
Differenzen (*mismatch*) zur MJ-Matrix	1999	[172]
Eigenwertanalyse der MJ-Matrix	2000	[165]
Korellationskoeffizient von Ähnlichkeitsmatrizen	2000	[173]
Branch-and-bound-Algorithmus	2002	[174]
Kontaktenergien	2002	[167]
Substitutionsmatrizen	2003	[175]
Markov-Prozess	2004	[176]
Selbstorganisierende Karte	2004/5	[168, 177]
Multi-dimensionale Skalierung der MJ-Matrix	2007	[178]

Tabelle 4.7: Einteilung der Aminosäuren in Klassen. (Ergänzt aus [178]). MJ-Matrix ist die Miyazawa-Jernigan-Matrix, die Kontaktenergien zwischen Aminosäuren enthält [179].

Klassen	Nr.	Zugehörige Aminosäuren
5	1	Phe, Ile, Leu
	2	Cys, Met, Val, Trp, Tyr
	3	Ala, His, Thr
	4	Gly, Pro
	5	Asp, Glu, Lys, Asn, Gln, Arg, Ser
8	1	Phe, Ile, Leu
	2	Cys, Tyr
	3	Met, Val, Trp
	4	His
	5	Ala, Thr
	6	Gly, Pro
	7	Gln, Arg, Ser
	8	Asp, Glu, Lys, Asn

Tabelle 4.8: Verwendete Aminosäuren-Klassifizierungen mit fünf und acht Klassen nach [168], mit der Ordnungsnummer $Nr.$ der einzelnen Aminosäuregruppen.

4.4 Kraftfeldansatz

Die Faltung eines Proteins zum nativen Zustand wird durch die Minimierung der freien Energie des Gesamtsystems Protein und Umgebung getrieben. Aufgrund der Komplexität dieses Problems beruhend auf der Vielzahl an Atomen und Freiheitsgraden und dem damit einhergehenden großen Konformationsraum sowie einer notwendigen langen Simulationsdauer bis zur Vollendung der Faltung, ist eine solche Berechnung für viele Proteine heutzutage außerhalb des technisch realisierbaren. Daher vereinfacht man vielfach das Problem zunächst dahingehend, das Protein ohne explizite Solvenshülle isoliert im Vakuum bzw. mit einem impliziten Solvatationsmodell zu beschreiben. Damit diese Simplifizierung erfolgreich ist, muss die grundlegende Voraussetzung erfüllt sein, dass die Geometrie des nativen Zustandes im wesentlichen nicht durch die Eigenschaften der Umgebung, sondern durch die Wechselwirkungen innerhalb des Proteins bestimmt wird. Hinweise für die Richtigkeit dieser Annahme erhielt man beispielsweise durch den Vergleich von Proteinstrukturen, die im Kristall mit Röntgenbeugung bestimmt wurden, mit in Lösung NMR-spektroskopisch bestimmten Strukturen (siehe z. B. [180–183]). Hier zeigte sich, dass sich die nativen Proteinstrukturen in diesen beiden Umgebungen kaum unterschieden. Im Besonderen zeigte die Kernregion im inneren eines Proteins, die eine dicht gepackte, gut definierte Sekundärstruktur besitzt, nahezu keine Abhängigkeit von der Umgebung. Geometrische Unterschiede konnten dagegen an Windungen und in den Randbereichen beobachtet werden, die eine weniger ausgeprägte Sekundärstruktur und damit größere Flexibilität besitzen. Insgesamt wurden für die untersuchten Proteine aber nur cRMSD-Gesamtdifferenzen von 1 bis 2 Å gefunden.
Unter dieser Annahme, dass die Faltung zwar durch die Minimierung der freien Energie des Gesamtsystems getrieben wird, die Umgebung aber nicht maßgeblich bestimmend für die Struktur des nativen Zustandes ist, kann das Problem dahingehend genähert bzw. reduziert werden, das globale Minimum der proteininternen Wechselwirkungen anstelle der des Gesamtsystems zu bestimmen.
Aufgrund der "kombinatorischen Explosion" [184] der zugänglichen Proteingeometrien, ist der zu durchsuchende Konformationsraum trotz dieser Vereinfachung des Problems dennoch sehr groß. Daher wird häufig die Darstellung des Proteins an sich auch noch einmal vereinfacht, um weniger Wechselwirkungen zwischen unterschiedlichen Interaktionspunkten berechnen zu müssen, um in vernünftiger Zeit die potentielle Energie für viele Geometrien berechnen zu können. Hierzu werden in der Literatur vergröberte Potentiale entwickelt und angewendet, die auf sehr wenigen Wechselwirkungszentren pro Aminosäure basieren, da selbst Kraftfelder mit vielen vereinigten Atomen kaum rechentechnische Vorteile gegenüber Kraftfeldern, die alle Atome enthalten, bieten [97]. Dagegen haben stark vergröberte Darstellungen eine sehr viel geringer Zahl an Freiheitsgraden und eine glattere Potentialenergiefläche, wodurch eine globale Optimierung und Strukturvorhersage des Proteins ohne Vorkenntnisse technisch

realisierbar wird. Bei einer solchen Vergröberung der Energiefläche eines Proteins muss natürlich immer beachtet werden, dass dadurch auch Detailinformationen über die Struktur und die Wechselwirkungen verloren gehen, wodurch die Struktur abhängig vom gewählten Modell und Potential nur bis zu einer bestimmten Auflösung beschrieben werden kann, da die Vergröberung eines Potentials immer eine Mittelung über verschiedene Kräfte beinhaltet.

In der Literatur werden viele unterschiedliche Proteinmodelle und Kraftfeldansätze verwendet, so dass an dieser Stelle keine umfassende Übersicht über diese gegeben werden kann und hier nur auf beispielhafte Arbeiten verwiesen wird. Für ein *Review* über klassische Kraftfelder, die auf einer vollständigen atomaren Darstellung eines Proteins beruhen siehe z. B. den Artikel von J. W. Ponder und D. A. Case [185]. Die Publikation von A. Kolinski und J. Skolnick [97] dagegen enthält beispielsweise eine ausführliche Zusammenfassung zu vergröberten Potentialen und reduzierten Modellen. Stark vereinfachte Kontakt-Energie-Ansätze werden beispielsweise in der frühen wichtigen Arbeit von S. Miyazawa und R. L. Jernigan [179] oder z. B. in [186] dargestellt. Viele der publizierten Kraftfelder basieren auf Paarpotentialen. Deren funktionale Formen können beispielsweise auf mathematischen Ausdrücken beruhen, die an quantenmechanische Potentiale angepasst wurden wie z. B. das Lennard-Jones-Potential (siehe z. B. [99]) oder sie können beispielsweise durch Einteilung Abstandsintervalle auf Basis einer statistischen Analyse erstellt definiert worden sein (siehe z. B. [128]). Mehrkörperpotentiale finden hierbei weniger Verbreitung. Für Dreikörperpotentiale siehe z. B. [187, 188], und für Vierkörperpotentiale, die häufig auf einem Ansatz mit Delaunay- oder Voronoi-Mosaiken beruhen z. B. [189, 190]. Vielkörperpotentiale sind in der Literatur sehr rar vertreten, siehe [191] als Beispiel.

4.4.1 Die Basisfunktionen

Das in dieser Arbeit entwickelte Potential basierte auf einer vergröberten Darstellung des Proteins. Es wurden drei Wechselwirkungszentren pro Aminosäure verwendet. Dies waren das C^{α}-Atom, der Punkt Z zur Beschreibung der Wasserstoffbrückenbindung, welcher auf der halben Strecke zwischen zwei C^{α}-Atomen liegt, und der Schwerpunkt der Seitenkette. Hierbei entspricht lediglich das C^{α}-Atom einem realen Atom im Protein, während die anderen beiden Punkte virtuelle Zentren sind (vergleiche auch Abb. 4.2). Die Berechnung der Wechselwirkungen anhand dieser Punkte beruhte ausschließlich auf Paarpotentialen. Es wurden in diesem Ansatz keine Mehrkörper-Potentiale verwendet.
Generell wurde die Gesamtenergie E eines Proteins mit der Geometrie $\mathbf{X}$ und der Sequenz $\mathbf{S}$ als Linearkombination von Basisfunktionen ϕ_n angesetzt:

$$E(\mathbf{X}, \mathbf{S}) = \sum_{n=1} c_n \cdot \phi_n(\mathbf{X}, \mathbf{S}) \tag{4.22}$$

Funktionenmenge	Beschreibung
$\{\phi^{(B)}\}$	Wechselwirkung von Aminosäuren, die in der Sequenz dicht beieinander liegen (B: *Bonded*)
$\{\phi^{(NB)}\}$	Nicht-bindende Wechselwirkungen (NB: *Non Bonded*)
$\{\phi^{(SC)}\}$	Seitenkettenzentrierte Potentiale (SC: *Side Chains*)
$\{\phi^{(SU)}\}$	Oberflächenpotential (SU: *Surface*)
$\{\phi^{(HB)}\}$	Wasserstoffbrückenpotential (HB: *Hydrogen Bonds*)

Tabelle 4.9: Übersicht über die verwendeten Basisfunktionsklassen.

wobei die c_n die Gewichtungskoeffizienten sind. Zur einfacheren Darstellungen des Ansatzes sind in Gl. 4.22 die weiteren Abhängigkeiten der Gewichtungskoeffizienten und Basisfunktionen, welche spezifisch für bestimmte Wechselwirkungen sind, weggelassen. Diese werden in den folgenden Kapiteln näher erläutert. In einem klassischen Kraftfeldansatz entsprechen die ϕ_n den bekannten Termen wie beispielsweise Coloumb- oder Lennard-Jones-Energie. Da in dem hier gewählten Ansatz die Basisfunktionen nur auf wenigen Punkten zentriert waren und sie dadurch gemittelte effektive Wechselwirkungen mehrerer realer Zentren und verschiedener Potentiale repräsentierten, war die funktionale Form des vergröberten Potentials a-priori nicht bekannt. Daher konnten für Basisfunktionen zunächst keine bekannten atomaren Potentialfunktionen angesetzt werden. Stattdessen musste ein Ansatz erfolgen, der das Potential möglichst allgemeingültig definiert, so dass sich aus der Kombination mehrerer dieser Funktionen die effektive Energiefläche des Proteins ergab. Hierbei ist die Wahl der mathematischen Funktionen in einem vergröberten Potential ein wichtiger Punkt, da schlecht gewählte Funktionen zur Folge haben können, dass der Verlauf der echten Potentialfläche nur unzureichend oder sogar falsch wiedergegeben wird oder dass die wichtigen Merkmale nicht erfasst werden können.

In dieser Arbeit wurde daher bei der Wahl der Potentialbasisfunktionen ein kombinierter Ansatz verfolgt: Einerseits wurden bestimmte Protein-Geometrie-Charakteristika wie z. B. die häufigsten Sekundärstrukturen direkt für Potentialfunktionen verwendet, um den möglichen Konformationsraum von vornherein zu verkleinern. Andererseits wurden für bestimmte Wechselwirkungen, deren effektive Form nicht bekannt ist, einfache mathematische Funktionen angesetzt, um mit diesen eine optimierte Linearkombination zu bilden, die in Summe das vergröberte Potential beschreibt.

Insgesamt wurden fünf unterschiedliche Basisfunktionsklassen zur Gestaltung des vergröberten Potentials angesetzt. Diese sind in Tab. 4.9 aufgelistet und werden in den folgenden Abschnitten eingehender beschrieben.

4.4.2 Nahwechselwirkungsterme

Die Potentialfunktionen $\{\phi^{(B)}\}$ beschreiben die Wechselwirkungen von Aminosäuren, die in der Sequenz und somit auch räumlich dicht benachbart sind. Diese Funktionen sind auf den C^{α}-Atomen zentriert. Da die Sekundärstruktur der Proteine auf wenige wiederkehrende Muster beschränkt ist, welche auf den zu Anfang beschriebenen Einschränkungen der atomaren Struktur der Aminosäuren beruhen, wurde für die Nahwechselwirkungsterme untersucht, ob sich entsprechende Muster ebenfalls in der reinen C^{α}-Geometrie für sequenznahe Aminosäuren bestimmen lassen, um über eine Darstellung basierend auf Abständen und Vektoren Potentialfunktionen zu definieren. Hierzu wurden 278 Proteine aus dem TOP500H-Satz an hoch aufgelösten Proteinstrukturen analysiert (siehe 4.3.1 zur Beschreibung dieser Proteinmenge) und für diese die statistische Häufigkeiten der Funktionswerte der euklidischen Abstände der C^{α}-Atome und der inneren Produkte ihrer Bindungsvektoren in Intervallen bestimmt. Der Abstand $r_{i,j}$ zweier C^{α}-Atome i und j mit den Ortsvektoren $\mathbf{x}_i$ und $\mathbf{x}_j$ ist gegeben durch $r_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|$. Ein Bindungsvektor $\mathbf{b}_i$ zwischen zwei sequentiellen C^{α}-Atomen ist gegeben durch $\mathbf{b}_i = \mathbf{x}_i - \mathbf{x}_{i-1}$, wobei das innere Produkt zweier Bindungsvektoren $\mathbf{b}_i$ und $\mathbf{b}_j$ gegeben ist durch $\langle \mathbf{b}_i, \mathbf{b}_j \rangle$. Unter Verwendung der Datenbank-Proteine wurden somit die Abstandsfunktionen

$$r_{i,i+1},\ r_{i,i+2},\ r_{i,i+3} \text{ und } r_{i,i+4} \tag{4.23}$$

sowie die inneren Produkte

$$\langle \mathbf{b}_i, \mathbf{b}_{i+1} \rangle, \langle \mathbf{b}_i, \mathbf{b}_{i+2} \rangle \text{ und } \langle \mathbf{b}_i, \mathbf{b}_{i+3} \rangle \tag{4.24}$$

ausgewertet. Die resultierenden Statistiken der Abstandsfunktionen sind in Abb. 4.19 und die der inneren Produkte in Abb. 4.20 dargestellt. Die Häufigkeiten der wichtigsten Sekundärstrukturen werden zusätzlich in den Grafiken separat gezeigt. Das Ziel war es, signifikante Häufungen in diesen Statistiken in Potentiale umzusetzen. Dabei wurde so verfahren, dass nicht, wie in vielen aus der Literatur bekannten statistischen Potentialen, die Verteilungen über die Boltzmann-Gleichung in Relation zu einem bestimmten Referenzzustand direkt in eine Wechselwirkungsenergie bestimmt wurde, sondern dass die Abszissenwerte der Maxima als Minima einfacher glatter Potentialfunktionen verwendet wurden, deren Gewichtungskoeffizienten in der Gesamtentwicklung des Potentials bestimmt wurden.

Die erste Funktion $r_{i,i+1}$ beschreibt den Abstand zweier direkt benachbarter C^{α}-Atome. Wie aus der Abbildung ersichtlich ist, enthält diese Verteilung zwei Maxima, die jeweils einer *cis*- und einer *trans*-Bindung entsprechen, wobei die *trans*-Bindung weitaus häufiger auftritt und dessen Verteilung sehr eng auf einen Bereich von ca. 3.8±0.05 Å verteilt ist.
Betrachtet man die anderen Verteilungen in Abb. 4.19 allgemein, so zeigt der Trend, dass zu

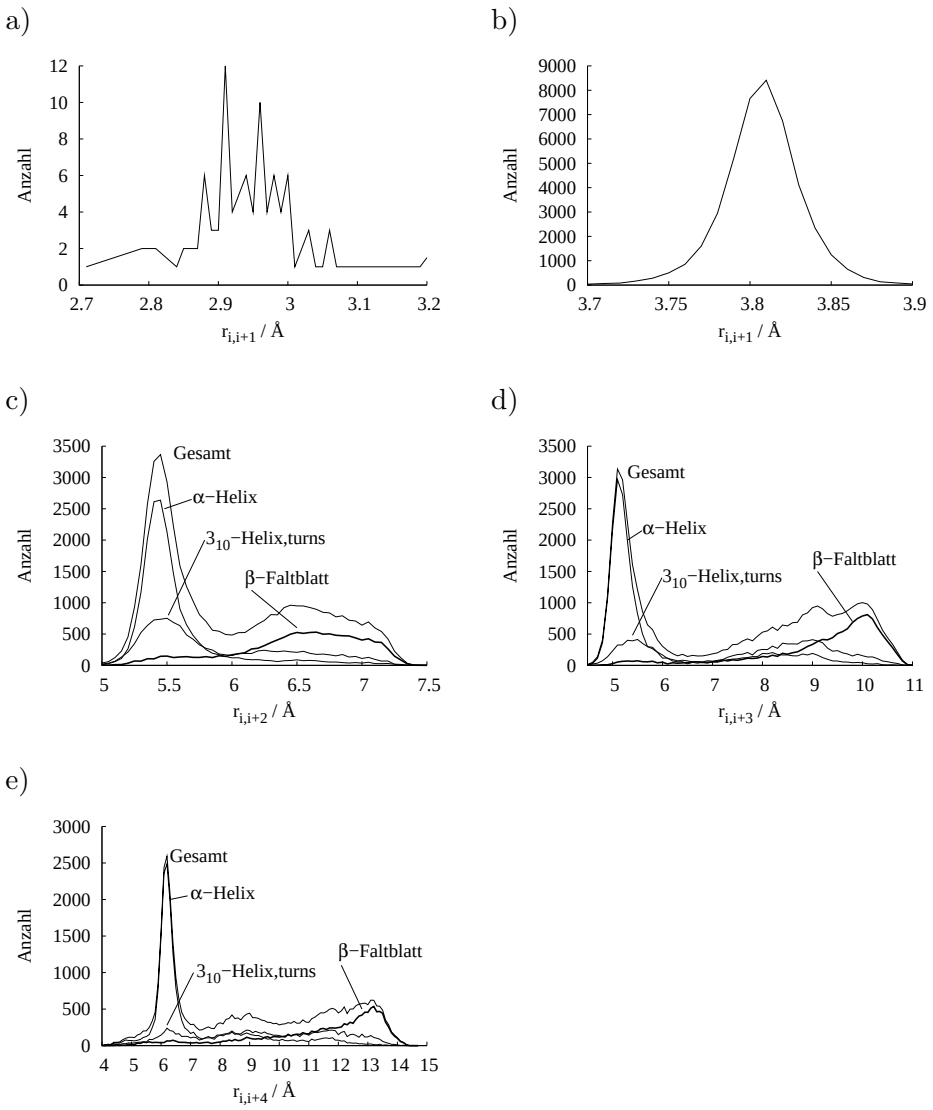


Abbildung 4.19: Statistiken der Abstandsfunktionen. Die Funktion $r_{i,i+1}$ (erste Zeile bzw. a und b) ist auf zwei Grafiken aufgeteilt. Die linke zeigt die Verteilung für cis-Bindungen, die rechte für $trans$-Bindungen. Die Sekundärstrukturen wurde mit Hilfe des DSSP-Programms zugeordnet.

größeren Sequenzabständen hin die Verteilungen diffuser werden, was darauf beruht, dass eine Sekundärstruktur nur lokal die Geometrie definiert aber nicht über weite Strecken korreliert. Wie aber weiterhin ersichtlich ist, enthält jede Verteilung zu den Abstandsfunktionen einen dominierenden α-Helix-Peak bei kleinen Abständen, der mit dem Maximum der 3_{10}-Helix zusammenfällt, während dagegen die anderen Strukturen breitere Verteilungen zeigen. Dies beruht darauf, dass die α-Helix in Proteinen die längsten Strukturen ausbildet, die zudem auch durch eine große Regelmäßigkeit ausgezeichnet sind, woraus die schmale Verteilung des Peaks resultiert. Dagegen sind β-Faltblatt-Strukturen allgemein kürzer und durch Verdrehungen in der Struktur weniger regelmäßig, woraus die breitere und unspezifischere Verteilung resultiert. Wie aufgrund der geometrischen Unterschiede der α-Helix und des β-Faltblattes

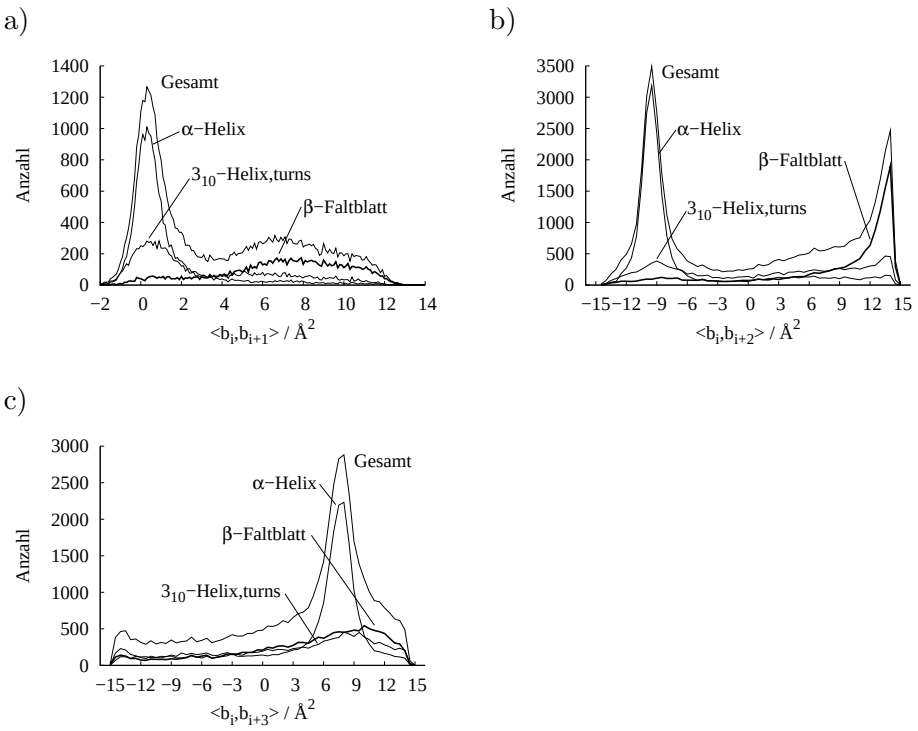


Abbildung 4.20: Statistiken der inneren Produkte der Bindungsvektoren. Die Sekundärstrukturen wurde mit Hilfe des DSSP-Programms zugeordnet.

zu erwarten war, sind in allen Abstandsfunktionen die Maxima ihrer Verteilungen getrennt. Aus der Sichtweise zur Definition eines Potentials basierend auf diesen Verteilungen eignen sich die Verteilungen α- bzw. der 3_{10}-Helix gut, da sie schmal und von den anderen Strukturen getrennt vorliege. Die wichtige Verteilung des β-Faltblattes dagegen ist weniger gut geeignet, weil die Definition des Maximums teilweise weniger eindeutig ist und weil die Verteilung breiter und asymmetrischer ist. Es zeigte sich jedoch in durchgeführten Faltungstests mit dem genetischen Algorithmus, dass Abstandsfunktionen notwendig sind, um realistische Rückgratgeometrien zu erzeugen. Eine indirekte Mitberücksichtigung der Abstandsinformationen durch andere Funktionen wie beispielsweise durch die weiter unten beschriebenen inneren Produktfunktionen war nicht ausreichend, so dass hierfür eigene Funktionen definiert wurden. Für das Potential wurden daher die Abstandsverteilungen zu $r_{i,i+1}$, $r_{i,i+2}$ und $r_{i,i+3}$ in Potentiale umgesetzt und für diese zunächst jeweils ein Doppelminimumpotential angesetzt, um α-Helices und β-Faltblätter zu erzeugen, deren Maxima in allen Statistiken getrennt vorlagen. Hierzu wurden die Abszissenwerte $\alpha_{n,1}$ und $\alpha_{n,2}$ der beiden Maxima 1 und 2 aus den Verteilungen bestimmt, wobei n die Ordnungsnummer der Basisfunktion ist. Mit diesen wurden Potentialfunktionen folgender Form definiert: $\phi_n^{(B)}(r_{i,i+n}) = \Big((r_{i,i+n} - \alpha_{n,1})(r_{i,i+n} - \alpha_{n,2})\Big)^2$. Die Minima der Funktionen $\phi_2^{(B)}$ und $\phi_3^{(B)}$ entsprechen somit bestimmten Sekundärstrukturen, während dagegen die Minima der Funktion $\phi_1^{(B)}$ eine *cis*- oder *trans*-Bindung beschreibt. Somit wurden mit den bestimmten Maxima der Verteilungen folgende Basisfunktionen ba-

sierend auf C^{α}-Abständen verwendet:

$$\begin{aligned}
\phi_1^{(B)}(r_{i,i+1}) &= \Big((r_{i,i+1} - 2.96\text{Å})(r_{i,i+1} - 3.81\text{Å})\Big)^2 \\
\phi_2^{(B)}(r_{i,i+2}) &= \Big((r_{i,i+2} - 5.45\text{Å})(r_{i,i+2} - 6.65\text{Å})\Big)^2 \\
\phi_3^{(B)}(r_{i,i+3}) &= \Big((r_{i,i+3} - 5.1\text{Å})(r_{i,i+3} - 10.1\text{Å})\Big)^2
\end{aligned} \tag{4.25}$$

Die Potentiale erhalten damit eine generelle Doppelminimum-Form wie in Abb. 4.21 dargestellt, wobei sich zwischen den Funktionen jeweils lediglich die Positionen der Minima unterscheiden. Die Gewichtungskoeffizienten dieser Funktionen $\phi^{(B)}(r_{i,j})$ wurden abhängig von den Aminosäuretypen der Positionen i und j gewählt, um es dem Optimierungsprozess zu ermöglichen, die spezifischen Wechselwirkungen zwischen unterschiedlichen Aminosäuren separat zu gewichten. Dabei wurde weiterhin die Sequenzreihenfolge der Aminosäuren unterschieden, so dass das Aminosäure-Paar (m, n) mit $m, n \in \{1, \ldots, 20\}$ unabhängig vom Paar (n, m) behandelt wird, wobei n und m die Aminosäuretypen der Sequenzpositionen i und j sind. Bei der Verwendung von allen 20 Aminosäuren würde man so 400 zu optimierende Koeffizienten für jede einzelne Funktion aus den Gleichungen 4.25 erhalten, was insgesamt für alle drei Potentiale zu 1200 Parametern führt. Dies ist eine sehr große Anzahl an Parametern, insbesondere wenn zusätzlich noch die Koeffizienten der anderen Basisfunktionen hinzukommen. Um die Zahl der Koeffizienten zu verkleinern wurde, daher ein reduziertes Aminosäure-Alphabet verwendet.

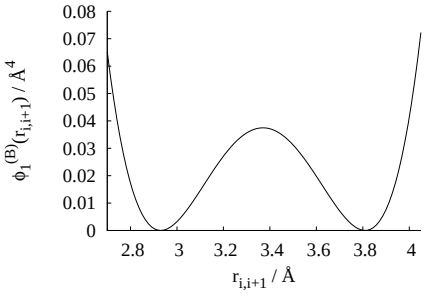


Abbildung 4.21: Potentialfunktion für $r_{i,i+1}$.

Diese sequenzreihenfolgeabhängige Wahl der Koeffizienten wurde so getroffen, da große Teile der Sekundärstrukturen durch eine lokale Reihenfolge an bestimmten Aminosäuren codiert wird. Durch entsprechende Kenntnis dieser Codierung können 70 bis 75 % der Sekundärstruktur vorhergesagt werden [192–194]. Die langreichweitigen Wechselwirkungen spielen bei der Bildung der Sekundärstrukturen zumeist nur eine unterstützende Rolle. Diese können aber in bestimmten Geometrien über die lokale Sequenzbevorzugung dominieren [195].

Wie sich in anfänglichen Parameteroptimierungen zeigte, erhielten die hier verwendeten Koeffizienten der Doppelminimumpotentiale im Vergleich zu den anderen Koeffizienten sehr kleine Werte. Dies hatte in Faltungstests mit dem genetischen Algorithmus zur Folge, dass diese Potentiale auf die Gesamtenergie und den Gradienten praktisch keinen Einfluss hatten und dass dadurch unrealistische Geometrien in der globalen Optimierung erzeugt wurden, weil die anderen Kräfte über diese Nahwechselwirkungsterme dominierten. Als weiteres Ergebnis wurde festgestellt, dass die Funktion $\phi_3^{(B)}(r_{i,i+3})$ nicht notwendigerweise mitverwendet werden mus-

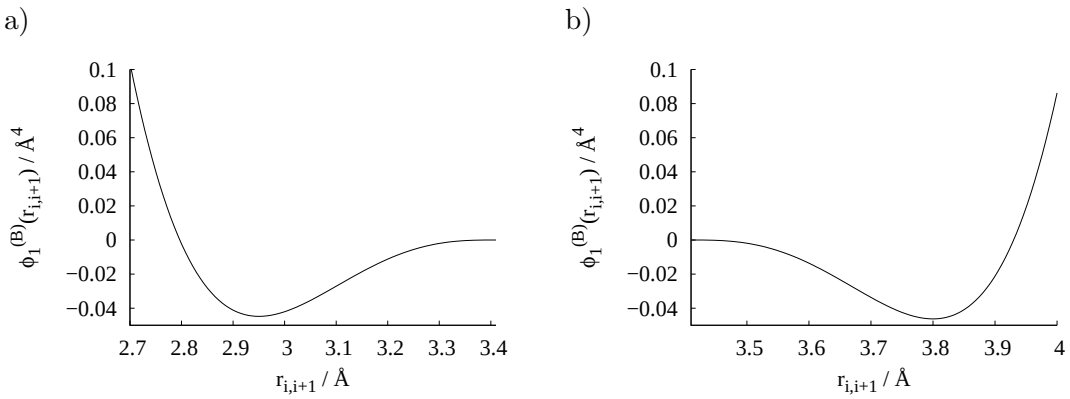


Abbildung 4.22: Geteilte Abstandsbasisfunktionen für $\phi_1^{(B)}(r_{i,i+1})$. In a) ist der linke Teil der Funktion für $r_{i,i+1} < \beta_n$ und in b) der rechte Teil für $r_{i,i+1} \geq \beta_n$ dargestellt.

ste, sondern dass die anderen Funktionen ausreichend waren, korrekte Rückgratgeometrien zu erzeugen, da diese die Informationen der $\phi_3^{(B)}$-Funktion mit enthielten (wie z. B. die Basisfunktionen basierend auf den inneren Produkten).

Aus diesen Gründen wurde eine andere Formulierung für die Nahwechselwirkungsterme abhängig den Abstand der C^α-Atome verwendet, so dass sie zwar weiterhin den Doppelminimumcharakter behielten, dass aber statt dessen beide Minima unterschiedlich zueinander gewichtet werden konnten. Hierzu wurde jedes Potential zunächst zwischen den beiden Minima getrennt und das Potential links und rechts dieser Trennstelle separat durch eine eigenständige Funktion dargestellt. Für die neuen Potentiale wurde folgende allgemeine Form angesetzt:

$$\phi_n^{(B)}(r_{i,i+n}) = \begin{cases} (\beta_n - r_{i,i+n})^3 \cdot (3(\beta_n - r_{i,i+n}) - \gamma_{n,1}) & \text{mit } r_{i,i+n} < \beta_n \\ (r_{i,i+n} - \beta_n)^3 \cdot (3(r_{i,i+n} - \beta_n) - \gamma_{n,2}) & \text{mit } r_{i,i+n} \geq \beta_n \end{cases} \tag{4.26}$$

Hierbei sind β_n, $\gamma_{n,1}$ und $\gamma_{n,2}$ Konstanten. Die Trennposition zwischen beiden neuen Funktionen ist gegeben durch β_n, an welcher beide Funktionen den Wert Null annehmen. Der Wert β_n wurde so gewählt, dass er mit einem Minimum in der Häufigkeitsstatistik (siehe Abb. 4.19) zusammenfällt. Explizit wurde für β_1 in $\phi_1^{(B)}$ der Mittelpunkt zwischen den beiden Verteilungsmaxima gewählt, so dass $\beta_1 = 0.5(\alpha_{1,1} + \alpha_{1,2}) = 3.37\,\text{Å}$. Für $\phi_2^{(B)}$ wurde der Wert $\beta_2 = 6.0$ Å verwendet. Die Konstanten $\gamma_{n,1}$ und $\gamma_{n,2}$ sind die neuen Minima der Funktionen, die aus den ursprünglichen Minima $\alpha_{n,1}$ und $\alpha_{n,2}$ hervorgehen und mit Hilfe der ersten Ableitungen zu Gl. 4.26 bestimmt wurden. Es ergab sich als Bestimmungsgleichung für die neuen Minima: $\gamma_{n,k} = 4\,|\alpha_{n,k} - \beta_n|$ mit $k \in [1, 2]$. Die mit diesen Werten resultierenden Funktionen haben die allgemeine Form wie in Abb. 4.22 dargestellt. Die Koeffizienten wurden wie vorher abhängig von den Aminosäuretypen an Position i und $i + n$ und ebenfalls abhängig von ihrer Reihenfolge gewählt. Mit diesen Funktionen können die beiden Minima unterschiedlich zueinander gewichtet werden. Durch unterschiedliche Koeffizienten ändern sich die Tiefe des

Minimums und die Steigungen. Unveränderlich bleibt dagegen der Übergangs- bzw. Trennpunkt β_n zwischen den zusammengehörigen Potentialen, welcher stets den Funktionswert Null annimmt, wodurch kein Sprung zwischen den Potentialen entsteht und auch die Ableitungen ineinander übergehen. Bei der Optimierung wurden die Koeffizienten dieser Funktionen auf positive Werte beschränkt, um eine Umkehrung des Potentials zu vermeiden.

Neben diesen Abstandsfunktionen wurden auch Basisfunktionen basierend auf den inneren Produkten der Bindungsvektoren $\mathbf{b}_i$ als Nahwechselwirkungsfunktionen verwendet, wobei wie bei den Abstandsverteilungen beschrieben hierbei ebenfalls die Maxima der Verteilungen als Minima für eine Doppelminimum-Potentialfunktion verwendet wurden. Die Funktionen hierfür wurden, identisch zu den erweiterten Abstandsfunktionen, auf zwei Potentiale mit Trennstelle aufgeteilt.
Im ursprünglichen Potentialansatz war vorgesehen gewesen, nur diese und keine zusätzlichen abstandsabhängigen Funktionen zu verwenden, da die inneren Produkte sowohl Distanz- wie auch Winkelinformationen zwischen den Bindungsvektoren enthalten. Wie aber oben bereits erwähnt, waren sie alleine nicht ausreichend, physikalisch richtige Rückgratgeometrien zu erzeugen.
Die zu den inneren Produkten (siehe Gl. 4.24) ausgewerteten Statistiken der Datenbank-Proteine sind in Abb. 4.20 dargestellt. Der Verlauf der Verteilung für $\langle \mathbf{b}_i, \mathbf{b}_{i+1} \rangle$ ist sehr ähnlich zu der entsprechenden Verteilung für $r_{i,i+2}$, da der Abstand zwischen i und $i+2$ neben dem Bindungswinkel der drei beteiligten C^α-Atome auch von den beiden Bindunsglängen $r_{i,i+1}$ und $r_{i+1,i+2}$ abhängt. Da aber, wie aus Abb. 4.19 ersichtlich, der Abstand zweier sequentieller C^α-Atome i und $i+1$ in der großen Mehrheit der Fälle bei ca. 3.8 Å liegt, enthält das innere Produkt $\langle \mathbf{b}_i, \mathbf{b}_{i+1} \rangle$ nur unwesentlich andere Informationen als die reine Abstandsfunktion $r_{i,i+2}$. Daher wurde zur Vermeidung von redundanten Informationen der Funktionen zu $r_{i,i+1}$ und $r_{i,i+2}$ mit $\langle \mathbf{b}_i, \mathbf{b}_{i+1} \rangle$ die letztere Funktion nicht für das Potential verwendet.
Einen interessanteren Verlauf zeigte die Verteilung zum inneren Produkt $\langle \mathbf{b}_i, \mathbf{b}_{i+2} \rangle$,welche zwei gut separierte Maxima bei ca. -10 und 13 $Å^2$ enthält, die durch unterschiedliche Sekundärstruktur-Beiträge dominiert werden. Der Peak bei negativen Abszissenwerte gehört hauptsächlich zur α-Helix, der bei positiven Abszissenwerte zum β-Faltblatt. Vergleicht man diese Statistik mit derjenigen zum Abstand $r_{i,i+3}$ so ist in dieser die Verteilung der β-Struktur schmaler und ähnlich eng begrenzt wie die α-Struktur, so dass sich diese Verteilung als Grundlage für ein Doppelminimumpotential sehr gut eignet. Da das innere Produkt auf den Koordinaten von vier C^α-Atomen beruht, müsste der Koeffizient der zugehörigen Funktion in Abhängigkeit von gleichzeitig vier Aminosäuretypen gewählt werden. Bei zwanzig Aminosäure entspräche dies 160000 Koeffizienten, was bei den für die Optimierung zur Verfügung stehenden Ressourcen außerhalb des technisch realisierbaren lag. Daher wurde der Koeffizient

so aufgeteilt, dass er jeweils nur von den beiden äußeren ($i-1$ und $i+2$) bzw. nur von den inneren beiden (i und $i+1$) Aminosäuren abhing, wobei ebenfalls auch wieder die Reihenfolge der Aminosäuren entlang der Sequenz beachtet wurde.
Die Verteilung für $\langle \mathbf{b}_i, \mathbf{b}_{i+3} \rangle$ zeigt insgesamt nur einen ausgeprägten Peak bei ca 8. $\mathring{\text{A}}^2$ und ansonsten in Richtung kleinerer Abszissenwerte eine recht gleichmäßige Verteilung. Im Unterschied zu anderen Funktionen wird der Peak bei 8. $\mathring{\text{A}}^2$ zwar durch die α-Helix dominiert, enthält aber auch die Häufungsmaxima der anderen Sekundärstrukturen, wobei die maximale Verteilung für die β-Struktur relativ breit und flach und das Maximum leicht zu größeren Abszissenwerten verschoben ist. Für diese Verteilung wurde zunächst ein rein quadratisches Potential angesetzt, welche ihr Minimum bei 8 $\mathring{\text{A}}^2$ besaß. In Faltungsversuchen mit dem genetischen Algorithmus zeigte sich jedoch, dass durch Einbeziehung dieser Funktion in Kombination mit den anderen die α-Helix zu stark gewichtet wurde, was zu langen durchgängigen Helices führte. Daher wurde diese Funktion für das Potential schließlich nicht verwendet.
Die weiteren hier nicht mehr abgebildeten Statistiken der Funktionen $\langle \mathbf{b}_i, \mathbf{b}_j \rangle$ mit $j \geq 4$ zeigen eine ähnliche Verteilung wie die zur Funktion $\langle \mathbf{b}_i, \mathbf{b}_{i+3} \rangle$, wobei der α-Helix-Peak zu größeren Werten von j stetig abnimmt und eine gleichmäßige Verteilung entsteht.

Zusammengefasst wurden für die Nahwechselwirkungen folgende Funktionen und Koeffizienten $c^{(B)}_{n,s_i,s_j}$ verwendet. Hierbei sei $\chi_{i,j} = \langle \mathbf{b}_i, \mathbf{b}_j \rangle$ das innere Produkt.

$$\phi_1^{(B)}(r_{i,i+1}) = c^{(B)}_{1,s_i,s_{i+1}} \cdot (3.41\,\text{Å} - r_{i,i+1})^3 \cdot (3(3.41\,\text{Å} - r_{i,i+1}) - 3.39\,\text{Å}) \quad (4.27)$$
$$\text{mit } r_{i,i+1} < 3.41\,\text{Å}$$

$$\phi_2^{(B)}(r_{i,i+1}) = c^{(B)}_{2,s_i,s_{i+1}}(r_{i,i+n} - 3.41\,\text{Å})^3 \cdot (3(r_{i,i+n} - 3.41\,\text{Å}) - 3.39\,\text{Å}) \quad (4.28)$$
$$\text{mit } r_{i,i+1} \geq 3.41\,\text{Å}$$

$$\phi_3^{(B)}(r_{i,i+2}) = c^{(B)}_{3,s_i,s_{i+2}}(6.09\,\text{Å} - r_{i,i+2})^3 \cdot (3(6.09\,\text{Å} - r_{i,i+2}) - 2.58\,\text{Å}) \quad (4.29)$$
$$\text{mit } r_{i,i+2} < 6.09\,\text{Å}$$

$$\phi_4^{(B)}(r_{i,i+2}) = c^{(B)}_{4,s_i,s_{i+2}}(r_{i,i+2} - 6.09\,\text{Å})^3 \cdot (3(r_{i,i+2} - 6.09\,\text{Å}) - 2.58\,\text{Å}) \quad (4.30)$$
$$\text{mit } r_{i,i+2} \geq 6.09\,\text{Å}$$

$$\phi_5^{(B)}(\chi_{i,i+2}) = c^{(B)}_{5,s_{i-1},s_{i+2}}(-\chi_{i,i+2})^3 \cdot (-3\chi_{i,i+2} - 38\,\text{Å}) \quad (4.31)$$
$$\text{mit } \chi_{i,i+2} < 0$$

$$\phi_6^{(B)}(\chi_{i,i+2}) = c^{(B)}_{6,s_{i-1},s_{i+2}}(\chi_{i,i+2})^3 \cdot (3\chi_{i,i+2} - 38\,\text{Å}) \quad (4.32)$$
$$\text{mit } \chi_{i,i+2} \geq 0$$

$$\phi_7^{(B)}(\chi_{i,i+2}) = c^{(B)}_{7,s_i,s_{i+1}}(-\chi_{i,i+2})^3 \cdot (-3\chi_{i,i+2} - 38\,\text{Å}) \quad (4.33)$$
$$\text{mit } \chi_{i,i+2} < 0$$

$$\phi_8^{(B)}(\chi_{i,i+2}) = c^{(B)}_{8,s_i,s_{i+1}}(\chi_{i,i+2})^3 \cdot (3\chi_{i,i+2} - 38\,\text{Å}) \quad (4.34)$$
$$\text{mit } \chi_{i,i+2} \geq 0$$

4.4.3 Nicht-bindende Wechselwirkungen

Die nicht-bindenden Wechselwirkungen beschreiben das Potential zwischen Aminosäuren, die in der Sequenz entfernt voneinander sind, die aber in der Geometrie nahe beieinander sein können. Die Wechselwirkungspunkte sind hierfür auf den C^{α}-Atomen zentriert und werden nur zwischen diesen berechnet. Hier treten also keine Wechselwirkungsterme mit den anderen definierten Punkten einer Aminosäure auf. Die nicht-bindenden Wechselwirkungen werden für zwei C^{α}-Atome i und j berechnet, für die $|i-j| \geq 5$ gilt. Dieses Potential erfasst in gemittelter Weise die Wechselwirkungen zwischen den Rückgratteilen der Aminosäuren, die sonst in atomaren Kraftfeldern beispielsweise mittels Lennard-Jones- oder Coloumb-Termen beschrieben werden.
Eine statistische Analyse der Abstandsverteilungen sowie der inneren Produkte der Bindungsvektoren von in der Sequenz entfernten Aminosäuren, wie zuvor für die Nahwechselwirkungsterme beschrieben, zeigt für diese keine signifikanten Häufungen mehr zeigen. Einzige Ausnahme hiervon ist wiederum die α-Helix, die aufgrund ihrer Eigenschaft, sehr lange Helices bilden zu können, selbst noch für Werte $|i-j| > 20$ bestimmte Häufungsmuster zeigt. Für die anderen Sekundärstrukturen trifft dies aber nicht mehr zu, so dass diese Information ohne eine Übergewichtung der α-Helix nicht verwendet werden kann. Daher wurden für die nicht-bindenden Wechselwirkungen keine Annahmen über eine Form des Potentials gemacht. Stattdessen wurde die Gesamtwechselwirkung zwischen zwei Aminosäuren in einfachen mathematischen Funktionen entwickelt, die in Linearkombination das Potential beschreiben. Dieser Ansatz ist beispielsweise dem Lennard-Jones-Potential ähnlich, in welchem ein repulsiver und ein attraktiver Funktionsanteil zusammen das Gesamtpotential bilden. Der generelle Ansatz für die nicht-bindenden Wechselwirkungen war gegeben durch

$$\phi^{(NB)}(r_{i,j}) = \sum_{k=1} c_k^{(NB)} \cdot \psi_k^{(NB)}(r_{i,j}) \tag{4.35}$$

mit den Entwicklungskoeffizienten $c_k^{(NB)}$ und den Basisfunktionen $\psi_k^{(NB)}$. An dieser Stelle sei angemerkt, dass die Funktionen $\psi_k^{(NB)}$ eine Teilmenge der Basisfunktionen ϕ_n aus Gl. 4.22 sind. Es sind also weitere Basisfunktionen aus dem Gesamtansatz des Potentials, deren Koeffizienten im Optimierungsprozess gleichzeitig zusammen mit allen anderen bestimmt wurden. Charakteristisch für nicht-bindende Wechselwirkungen im allgemeinen ist, dass sie ein $1/r_{i,j}^n$-Verhalten (mit $n > 0$) zeigen, so dass für größere Abstände die Wechselwirkungsenergie schnell abnimmt und gegen Null strebt. Weiterhin können sich zwei Moleküle bzw. Atome aufgrund der Abstoßung der Elektronenwolken und der Atomkerne nicht beliebig nahe kommen, wodurch das Potential einen repulsiven Anteil für $r_{i,j}$ gegen Null beinhalten muss. Auf diesen Annahmen basierend wurden die Basisfunktionen $\psi_k^{(NB)}$ so gewählt, dass sie zum einen ein unteres Limit für das Potential enthalten, bei welchem der Funktionswert gegen positiv un-

endlich strebt, um dadurch das ausgeschlossene Volumen zu berücksichtigen, und dass sie zum anderen ein oberes Limit (*cut-off*) enthalten, für welches das Potential gegen Null geht und am Limit exakt Null ist. Die Funktionswerte, die zu noch größeren Abständen als dem oberen Limit gehören, wurden gleich Null gesetzt. Die verwendete funktionale Form für die Basisfunktionen $\psi_k^{(NB)}$ der nicht-bindenden Wechselwirkungen waren somit:

$$\psi_k^{(NB)}(r_{i,j}) = \frac{(l_2 - r_{i,j})^n}{(r_{i,j} - l_1)^k} \tag{4.36}$$

Hierbei sind l_1 und l_2 das untere und obere Limit (in Å) und $n, k \geq 1$. Der Zähler dieser Funktion bewirkt, dass die Funktion bei $r_{i,j} = l_2$ exakt Null wird. Der Exponent n bestimmt die Steilheit des Funktionsverlaufes und wurde zu $n = 5$ gewählt. Der Nenner hat den Effekt, dass die Funktion für $r_{i,j} \to l_1$ mit $r_{i,j} > l_1$ gegen positiv unendlich strebt und somit repulsiv wird. Kleinere Abstandswerte ($r_{i,j} < l_1$) wurden nicht mit diesen Funktionen berechnet, sondern statt dessen durch ein rein repulsives Potential erfasst (siehe hierzu Abschnitt 4.7.3).
Die zu diesen Funktionen zugehörigen Koeffizienten wurden so gewählt, dass sie von den Aminosäuretypen s_i und s_j der beteiligten C^α-Atome abhingen, so dass $c_k^{(NB)} = c_{k,s_i,s_j}^{(NB)}$ war. Bei den nicht-bindenden Wechselwirkungen wurde die Sequenzreihenfolge der Aminosäuren im Gegensatz zu den Nahwechselwirkungstermen nicht mit einbezogen, wodurch die Paare (s_i, s_j) und (s_j, s_i) gleich behandelt wurden. Insgesamt wurden für jedes Aminosäurepaar vier Basisfunktionen $\psi_k^{(NB)}$ in der Entwicklung angesetzt, wodurch über die optimierten Linearkombinationskoeffizienten dieser Funktionen unterschiedliche Minima und Maxima erzeugt werden können.
Um zu entscheiden, welche Werte für die Limit-Parameter l_1 und l_2 zu wählen seien, wurde getestet, für welche l_1 und l_2 die Parameteroptimierung noch erfolgreich war. Da sich generell in bekannten Proteinstrukturen die C^α-Atome nicht weiter als ca. 3 Å einander nähern [28], wurde l_1 zwischen 0 und 3 Å und l_2 zwischen 8 und 14 Å variiert, während alle anderen Variablen des Optimierungsproblems konstant gehalten wurden. (Die Prozedur der Parameteroptimierung wird in Abschnitt 4.5 näher beschrieben.) Es zeigte sich, dass der Optimierungserfolg in der Mehrzahl der Fälle von der Wahl von l_1 abhing, während die Wahl von l_2 hierbei weniger Bedeutung besaß. Als Ergebnis konnte l_1 maximal zu 2.2 Å gesetzt werden, bevor das Optimierungsproblem nicht mehr lösbar war. Der Grund hierfür war ein numerisches Problem, welches zu einer Instabilität und somit zu einem Abbruch des Parameteroptimierungsprozesses führen konnte. Dieses entstand dadurch, dass die repulsive Wechselwirkung für C^α-Atome, die einen Abstand nahe 3 Å besitzen, sehr von der Wahl von l_1 abhängt. Für $l_1 \to 3$Å nimmt diese stark zu, was dazu führt, dass in der Parameteroptimierung sehr große gegenüber sehr kleinen Funktionswerten berücksichtigt werden mussten.
Für das obere Limit l_2 zeigten sich keine numerischen Problemen bei unterschiedlich angesetzten Werten, da die Beiträge für große Abstände sehr klein sind, da die Funktionen dort gegen Null streben, so dass l_2 zu 8 Å gewählt wurde.

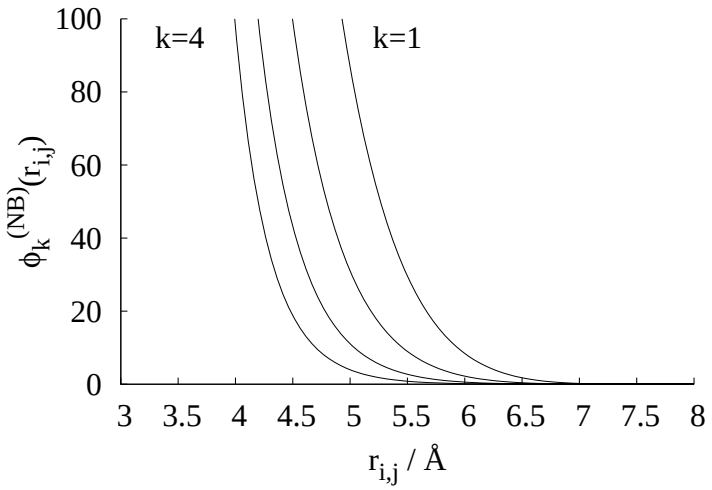


Abbildung 4.23: Verlauf der nicht-bindenden Basisfunktionen (siehe Gl. 4.37). Von links nach rechts sind die Funktionen mit jeweils kleinerem k dargestellt.

Mit diesen Werten für die Limits ergaben sich die Basisfunktionen der nicht-bindenden Wechselwirkungen zu:

$$\phi_k^{(NB)}(r_{i,j}) = c_{k,s_i,s_j}^{(NB)} \frac{(8\,\text{Å} - r_{i,j})^5}{(r_{i,j} - 2.2\,\text{Å})^k} \quad \text{mit} \quad 1 \leq k \leq 4 \tag{4.37}$$

Der Verlauf dieser Funktionen ist in Abbildung 4.23 dargestellt. Die Werte für die Koeffizienten $c_{k,s_i,s_j}^{(NB)}$ wurden während der Optimierung nicht eingeschränkt, so dass auch negative Werte zugelassen wurden, was einer attraktiven Wechselwirkung entspricht. In Summe können somit Linearkombinationen mit mehreren Minima und Maxima entstehen. Lediglich der Koeffizient mit dem größten k wurde auf positive Werte beschränkt, um die Gesamtfunktionen gegen l_1 repulsiv zu halten.

4.4.4 Seitenkettenpotential

Die Seitenketten in Proteinen können auf unterschiedliche Art mit der Umgebung wechselwirken. Neben der einfachen, aber für die Faltung wichtigen Tatsache, dass Seitenketten ein großes ausgeschlossenes Volumen besitzen, können sie aliphatische, aromatische, polare oder geladene Gruppen enthalten. Bei der Modellierung der Wechselwirkungen muss also wiederum ein breites Spektrum an unterschiedlichen Wechselwirkungen berücksichtigt werden. Hierzu wurden Paarpotentiale zentriert auf den Seitenkettenschwerpunkten definiert, die nur zwischen zwei Seitenketten wirken. Wechselwirkungen zwischen dem Rückgrat und Seitenketten wurden hierbei nicht mit einbezogen. Der Ansatz für die funktionale Form dieser Potentiale

war identisch mit dem für die nicht-bindenden Wechselwirkungen, die auf den C^{α}-Atomen zentriert waren:

$$\phi_k^{(SC)}(r_{i,j}) = c_{k,s_i,s_j}^{(SC)} \frac{(8\,\text{Å} - r_{i,j})^5}{(r_{i,j} - 2.2\,\text{Å})^k} \quad \text{mit} \quad 1 \leq k \leq 4 \tag{4.38}$$

Hierbei ist $r_{i,j}$ der Abstand zweier Seitenketten i und j mit den Ortsvektoren $\boldsymbol{\rho}_i$ und $\boldsymbol{\rho}_j$, so dass $r_{i,j} = \|\boldsymbol{\rho}_i - \boldsymbol{\rho}_j\|$. Es wurden auch hier insgesamt vier Funktionen für jedes Aminosäurepaar angesetzt. Die Koeffizienten $c_{k,s_i,s_j}^{(NB)}$ der Seitenketten-Funktionen sind ebenfalls von den Aminosäuretypen des betrachteten Seitenkettenpaares abhängig, ohne Beachtung der Sequenzreihenfolge.

Wie weiter oben bereits erklärt, wurden Cystein-Seitenketten, die eine Disulfidbrücke gebildet hatten, nicht mitberücksichtigt. Dies geschah im Hinblick auf die spätere Awendung des Potentials im genetischen Algorithmus zur globalen Geometrieoptimierung. Die Problematik hierbei ist weniger die Formulierung einer angemessenen Potential-Funktion, sondern die programmtechnische Realisation zur Erzeugung der Disulfidbrücken, da es schwierig und zeitaufwendig ist, Proteinstrukturen so zu generieren oder umzufalten, dass die entsprechenden Cystein-Seitenketten zur Bildung einer Disulfidbrücke zusammenkommen. Zudem ist ohne Kenntnis der experimentellen Struktur von vornherein unklar, wieviele und welche Cystein-Seitenketten tatsächlich die Disulfidbrücken bilden.

Weiterhin sind in diesen Potentialfunktionen (Gl. 4.38) keine Wechselwirkungen der Seitenkette mit dem Rückgrat enthalten. Hierfür wurden separat nicht-bindende Funktionen implementiert, die eine analoge funktionale Form besitzen:

$$\phi_k^{(SB)}(r_{i,j}) = c_{k,s_i,s_j}^{(SB)} \frac{(8\,\text{Å} - r_{i,j})^5}{(r_{i,j} - 2.2\,\text{Å})^k} \quad \text{mit} \quad 1 \leq k \leq 4 \tag{4.39}$$

Hierbei ist $r_{i,j}$ der Abstand der Seitenkette zu einem C^{α}-Atom mit $r_{i,j} = \|\boldsymbol{\rho}_i - \mathbf{x}_j\|$. Die Koeffizienten $c_{k,s_i,s_j}^{(SB)}$ sind spezifisch für das wechselwirkende Aminosäurepaar. Diese Funktionen wurden für den Kraftfeldansatz zwar implementiert, aber in die Parameteroptimierung nicht miteinbezogen, so dass diese Wechselwirkungen im aktuellen Potential noch nicht erfasst werden. Sie werden im genetischen Algorithmus allerdings durch ein repulsives Potential berücksichtigt, um einen zu kurzen Abstand zwischen einer Seitenkette und einem C^{α}-Atom zu vermeiden (siehe Abschnitt 4.7.3). Sie wurden zunächst nicht berücksichtigt, um die Anzahl an zu optimierenden Paramtern möglichst klein zu halten (siehe Abschnitt 4.5.1) und um zu prüfen, inwieweit die anderen Wechselwirkungen ausreichend sind, Proteinstrukturen zu beschreiben.
Sowohl die Behandlung von Disulfidbrücken als auch die explizite Einbeziehung der Wechselwirkungen zwischen dem Rückgrat und den Seitenkettenschwerpunkte sind mögliche Punkte für eine zukünftige Entwicklung des Potentials.

4.4.5 Oberflächenpotential

Viele sowohl theoretische wie auch experimentelle Arbeiten zeigen, dass die Solvatation der Aminosäuren einen wichtigen und treibenden Effekt auf die Faltung und die Definition der Geometrie des nativen Zustandes hat. Dies geschieht zum einen durch den hydrophoben Effekt, durch den bei der Faltung die Grenzfläche zwischen hydrophoben Aminosäuren und dem Lösungsmittel minimiert wird, (siehe hierzu beispielsweise [67, 84, 196, 197]), oder durch die Solvatation und damit Exposition von geladenen Seitenketten gegenüber dem Lösungsmittel [198, 199]. Daher ist die Einbeziehung der Solvatationseffekte in einem Kraftfeldansatz wichtig. Beispielsweise konnte in [200] oder [201] gezeigt werden, dass ein auf hydrophoben Wechselwirkungen beruhendes Potential in der Lage ist, sehr viele falsche von nativen Strukturen zu unterscheiden.
Für eine Beschreibung dieser Wechselwirkung gibt es unterschiedliche Ansätze: Zum einen die explizite Einbeziehung von Wassermolekülen, z. B. in Molekulardynamik-Simulationen. Dieser Ansatz ist sehr rechenintensiv und zeitaufwendig und für eine globale Optimierung ungünstig. Geeigneter sind daher Ansätze, die die Solvatation implizit ohne Wasser behandeln. Eine ausführliche Übersicht über solche Methoden findet sich beispielsweise in [202]. Neben diesen Methoden gibt es noch den Ansatz, die Solvatationsenergie durch Berechnung der lösungsmittelzugänglichen Oberfläche zu nähern. Dieser Ansatz liefert bei atomarer Auflösung keine verlässlichen Daten, sondern ist auf Molekülmodelle mit vergröberter Darstellung beschränkt, in der die behandelten Objekte größer als der Durchmesser eines Wassermoleküls sind [203, 204]. Hierbei wird davon Gebrauch gemacht, dass die freie Solvatationsenergie eines Moleküls proportional zu dessen Oberfläche ist [205–207]. Für Proteine konnte sogar gezeigt werden, dass die Oberfläche einer Aminosäure bzw. Seitenkette proportional zur Anzahl der benachbarten Aminosäuren im maximalen Abstand von 10 Å ist [208, 209].

Für diese Arbeit wurde die Solvatationsenergie auf Basis der Abhängigkeit von der lösungsmittelzugänglichen Oberfläche der Aminosäuren gewählt, welche sich aus den Abständen der Aminosäuren voneinander berechnen lässt, um so ein stetiges Potential zu erhalten. Eine Potential dagegen, welches rein auf der Anzahl der Nachbarn beruht, ist per Definition diskret, sofern keine Interpolation durchgeführt wurde, wodurch ein solches Potential Sprünge enthält.
Analysiert wurde hierzu zunächst der Erwartungswert der zugänglichen Oberflächen aller Aminosäuren im reduzierten TOP500H-Proteinsatz, um zu bestimmen, welche Aminosäuren für das hydrophobe Potential verwendet werden sollten. Die Oberflächen A wurden mit dem DSSP-Programm bestimmt und zu Intervallen in 1 $Å^2$ zusammengefasst. Hierauf wurde der Erwartungswert der Oberfläche $E(A)$ bestimmt. Zusätzlich wurden dem DSSP-Programm die Gesamtoberflächen der freien Aminosäuren berechnet und mit diesen der relative Erwartungs-

Aminosäure	$E(A)$/Å	$E(A/A_0)$/%
Ala	27.6	12.94
Cys	16.2	6.64
Asp	71.3	25.75
Glu	90.1	30.09
Phe	30.7	9.55
Gly	27.9	14.98
His	55.0	18.78
Ile	24.3	8.58
Lys	104.3	33.87
Leu	25.2	8.78
Met	39.4	13.13
Asn	62.1	22.23
Pro	52.3	21.67
Gln	79.5	26.85
Arg	94.9	26.11
Ser	42.9	18.38
Thr	43.3	17.06
Val	22.2	8.55
Trp	41.5	11.38
Tyr	47.2	13.51

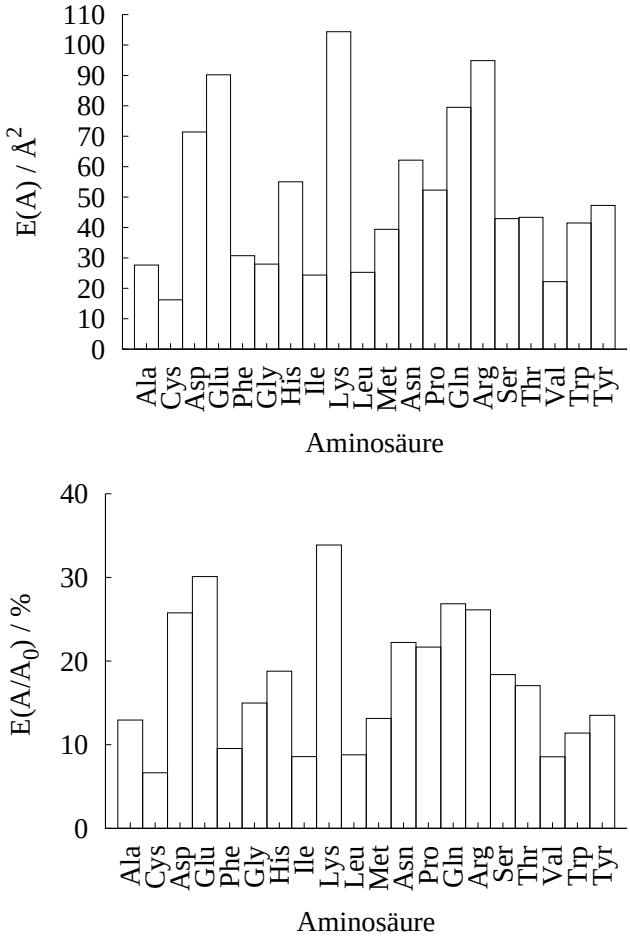


Abbildung 4.24: Erwartungswerte der lösungsmittelzugänglichen Oberfläche absolut $E(A)$ (Tabelle mittlere Spalte und obere Grafik) und relativ bezogen auf die Gesamtoberfläche A_0 einer isolierten Aminosäure (Tabelle rechte Spalte und untere Grafik).

wert $E(A/A_0)$ für jede Aminosäure bestimmt, welcher angibt, wieviel Prozent der gesamten Oberflächen dem Lösungsmittel zugänglich ist. Die zugehörigen Gesamtstatistiken der Oberflächenverteilungen sind in Abb. 4.25 und die Ergebnisse der Erwartungswerte in Abb. 4.24 dargestellt. Die Gesamtoberfläche einer Aminosäure dividiert durch zehn ergibt annähernd die Anzahl der Wassermoleküle, die Zugang zur Aminosäure haben [29].
Markant in Abb. 4.25 sind zunächst die polaren und geladenen Aminosäuren wie Asp, Glu, Lys, Gln und Arg, die breite Verteilungen bis hin zu 250 Å^2 zeigen. Die Erwartungswerte der Oberflächen liegen bei diesen Aminosäuren im Bereich zwischen 70 und 100 Å^2. Zusätzlich besitzen die Aminosäuren Asp, Glu, Gln und Arg auch große Häufigkeiten bei kleinen Oberflächen. Dies zeigt, dass auch polare bzw. geladene Aminosäuren nicht immer an der Oberfläche eines Proteins zu finden sind.

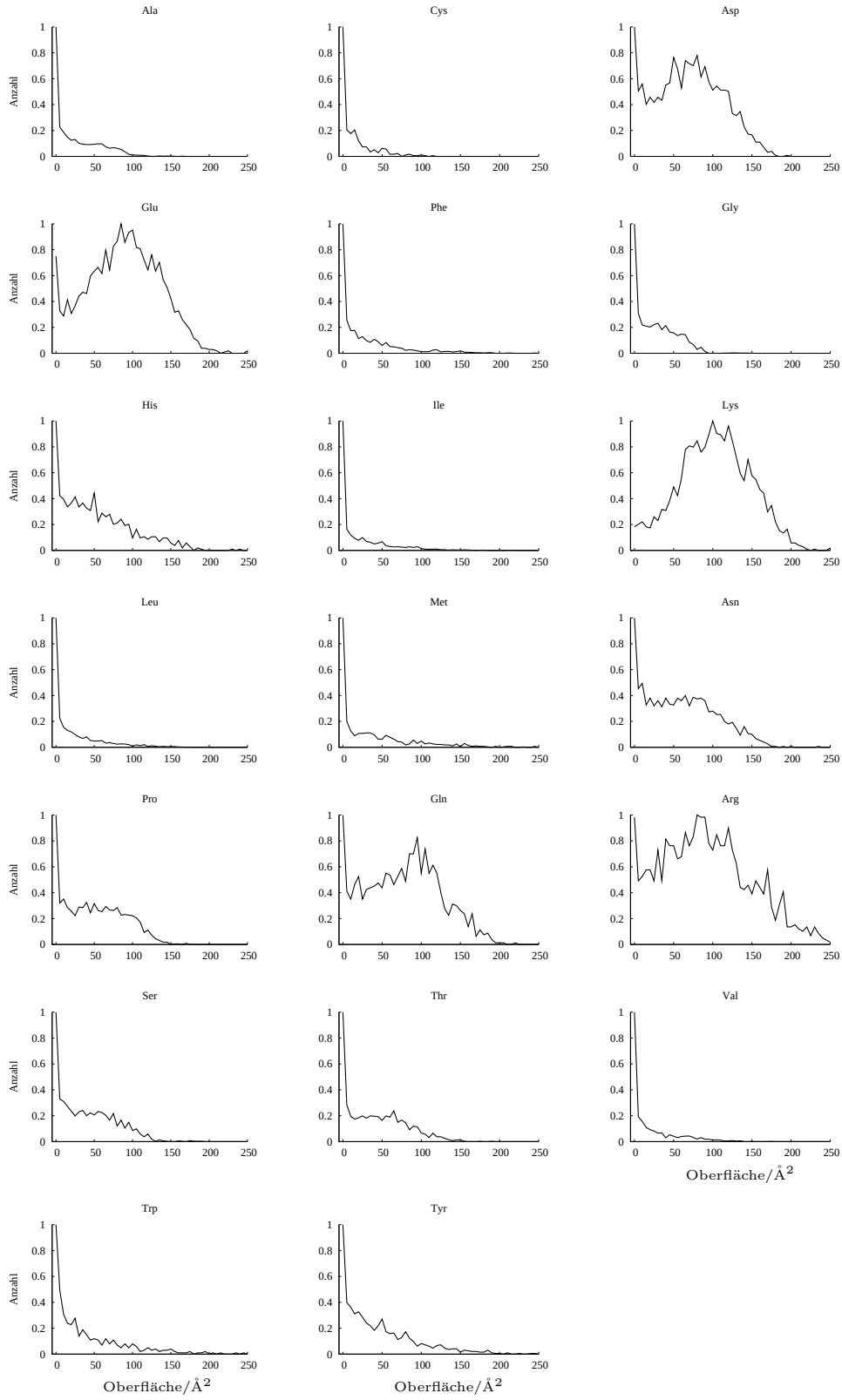


Abbildung 4.25: Normalisierte lösungsmittelzugängliche Oberflächenverteilung im TOP500H-Satz (in Intervallen zu 5 Å).

Die Seitenketten dieser Aminosäuren können sich bspw. zusammenlagern oder mit dem Rückgrat wechselwirken. Wie stark die Stabilisierung durch eine solche Wechselwirkung ist, wird in der Literatur diskutiert [88, 210]. Auf der anderen Seite stehen die mehr hydrophoben Reste, deren Verteilung nahe 0 Å^2 konzentriert ist. Mit Hilfe dieser Analyse sollten die Aminosäuren identifiziert werden, deren Oberflächen in der Regel sehr klein ist, um damit eine Potentialfunktion zu definieren, die diese Aminosäuren zur Minimierung ihrer Oberfläche treibt, um den hydrophoben Effekt zu simulieren. Hierzu wurden die Erwartungswerte $E(A)$ und $E(A/A_0)$ berechnet (siehe Abb. 4.24). Diese Erwartungswerte zeigen, angefangen bei der sehr hydrophoben Aminosäure Cystein, einen kontinuierlichen Übergang ohne größere Sprünge bis hin zur hydrophilsten Aminosäure Lysin. Da anhand der Erwartungswerte keine eindeutige Grenze gezogen werden konnte, welche der Aminosäuren als hydrophob verwendet werden sollte, wurden andere Literaturergebnisse hinzugezogen. Die Definition der Hydrophobizität ist allerdings schwierig und in der Literatur umstritten, so dass hierzu bisher mehr als 40 unterschiedliche Methoden und Skalen existieren. Eine ausführliche Übersicht über diese unterschiedlichen Methoden und Skalen findet man beispielsweise in [211]. Neumaier *et. al* haben die Daten der meisten dieser Skalen einer Hauptkomponentenanalyse (*principal component analysis*) unterzogen, um die wesentlichen gemeinsamen Informationen zu extrahieren, die die Hydrophobizität beschreiben [212]. Die durch diese Analyse neu erstellte Skala diente dann als Basis zur Auswahl der hydrophoben Aminosäuren Cys, Ile, Leu, Met, Phe, Trp, Tyr und Val. Diese Auswahl korrespondiert gut mit dem Erwartungswert $E(A/A_0)$, für welchen diese Aminosäuren alle einen Wert < 15 % besitzen. Lediglich Alanin und Glycin, welche die kürzesten Seitenketten besitzen, fallen heraus, da für diese die Polarität des Rückgrates mehr Bedeutung gewinnt. Für die ausgewählten hydrophoben Aminosäuren wurde im Kraftfeld eine "Straffunktion" eingebunden, welche deren zugängliche Oberfläche minimieren sollte. Hierzu wurde die Funktion $\phi^{(SU)}$ in Abhängigkeit von der Oberfläche A_i einer Aminosäure an der Sequenzposition i definiert zu:

$$\phi^{(SU)}(A_i) = \sum_{k=1}^{4} c_{k,s_i}^{(SU)} A_i^k \tag{4.40}$$

Die $c_{k,s_i}^{(SU)}$ sind die Gewichtungskoeffizienten für die Aminosäure s_i, die in der Optimierungsprozedur auf positive Werte beschränkt wurden. Weil A_i stets positiv ist, ist auch der Gesamtwert dieses Potentials stets ≥ 0. Das gesamte Oberflächenpotential ergibt sich aus der Summation über alle $\phi^{(SU)}(A_i)$ für alle hydrophoben Aminosäuren.

Um Gl. 4.40 zu verwenden, muss für jede Geometrie die zugängliche Oberfläche der hydrophoben Aminosäuren berechnet werden. Hierfür existieren in der Literatur bereits verschiedene Methoden. Häufige Anwendung finden vor allem die sehr genauen Methoden von Lee und Richards [213] und von Richmond [214], die aber wiederum längere Zeiten für die Evaluation der Oberfläche brauchen. Ein andere wesentlich schnellere Methode, die aber etwas weniger

exakte Resultate liefert, basiert auf den Arbeiten von Wodak und Janin [215] und verwendet zur Berechnung der Oberfläche einen probabilistischen Ansatz. Hierbei wird die durch andere Atome besetzte Oberfläche, deren exakte analytische Bestimmung bei der Anwesenheit vieler Atome zeit- und rechenintensiv ist, durch eine Wahrscheinlichkeitsberechnung ersetzt, die nur von den interatomaren Abständen abhängt. Die grundlegende Annahme hierbei ist, dass die zugängliche Oberfläche eines Atom oder allgemein eines kugelförmigen Körpers mittels folgender Formel genähert werden kann:

$$A_i = S_i \prod_{\substack{j=1 \\ j \neq i}}^{N} \left(1 - \frac{b_{ij}(r_{i,j})}{S_i}\right) \tag{4.41}$$

Hierbei ist A_i die lösungsmittelzugängliche Oberfläche für das Atom i, S_i die Gesamtoberfläche des isolierten Atoms, N die Gesamtanzahl an Atomen und b_{ij} die ausgeschnittene Oberfläche durch Überlappung der Atome i und j, welche vom Abstand $r_{i,j}$ abhängig ist. Der Produktterm $1 - \frac{b_{ij}(r_{i,j})}{S_i}$ ist die Wahrscheinlichkeit dafür, dass ein Punkt auf der Oberfläche von i außerhalb eines Schnittbereiches mit Atom j liegt. Bei großen Abständen $r_{i,j}$ geht b_{ij} gegen Null, so dass der entsprechende Produktterm gegen eins geht und die Oberfläche S_i nicht reduziert. Diese Formel beschreibt den ersten Ansatz dieses Oberflächenmodells. Für die genaue Ausformulierung, die bspw. die funktionale Form für $b_{ij}(r_{i,j})$ und Korrekturterme enthält, siehe [215]. Dieser Ansatz wurde bereits erfolgreich auf die Berechnung von Oberflächen für verschiedene organische Moleküle [217] und auch für Proteine in einer atomaren sowie in einer C^{α}-Darstellung erweitert und angewendet [216]. Hierbei zeigte sich, dass die erhaltenen Gesamtoberflächen für ein Molekül sehr gut mit den oben erwähnten genauen Methoden übereinstimmen, wobei die Abweichung nur einige wenige Prozent beträgt, während die Berechnung lediglich einen Bruchteil der Zeit in Anspruch nimmt, die die exakten Methoden benötigen.

j	p_{ij}
$i+1$	1.2431
$i+2$	-0.15956
$i+3$	0.65562
$\geq i+4$	0.5872

Tabelle 4.10: Optimierter Parameter p_{ij} aus [216].

Eine wichtige Veränderung des Ansatzes aus Gl. 4.41 für Proteine in einer reinen C^{α}-Darstellung war die Einführung von Korrekturparametern, die zum einen den Aminosäuretyp (p_i) und zum anderen den Abstand der Aminosäuren in der Sequenz (p_{ij}) spezifizieren, da die direkt benachbarten Aminosäuren in der Regel eine größeren Einfluss auf die zugängliche Oberfläche als die in der Sequenz weiter entfernten Aminosäuren haben. Die resultierende Formel lautete:

$$A_i = S_i \prod_{\substack{j=1 \\ j \neq i}}^{N} \left(1 - \frac{p_i\, p_{ij}\, b_{ij}(r_{i,j})}{S_i}\right) \tag{4.42}$$

Die Optimierung der Parameter p_i und p_{ij} für ein Protein in einer C^{α}-Darstellung wurde von Fraternali *et al.* durchgeführt [216]. Der Parameter p_i wurde für alle zwanzig Aminosäuren

bestimmt. Zusätzlich zu diesem Wert musste zur Berechnung von b_{ij} (siehe [215]) der Kugelradius R_i einer Aminosäure mitbestimmt werden. Beide Werte sind in Tab. 4.11 aufgelistet. Im Gegensatz zu diesen beiden Parametern wurde der Sequenzabstandsparameter p_{ij} auf vier mögliche Werte beschränkt und unabhängig von den beteiligten Aminosäuren angesetzt. Dieser wurde für $j = i + 1, i + 2, i + 3$ und $j \geq i + 4$ optimiert. Die erhaltenen Werte sind in Tab. 4.10 aufgeführt.

Diese Methode der probabilistischen Näherung der lösungsmittelzugänglichen Oberfläche wurde bei der Parameteroptimierung für die falschen Strukturen sowie später in der globalen Optimierung mittels des genetischen Algorithmus' eingesetzt. Die Oberflächen der nativen Strukturen wurden dagegen den entsprechenden DSSP-Dateien entnommen, welche die exakten Werte enthalten.

Aminosäure	$R_i/\text{Å}^2$	p_i
Ala	4.01256	0.99042
Arg	4.62059	0.71552
Asn	4.28348	0.87896
Asp	4.27733	0.88912
Cys	3.74826	0.95123
Gln	4.35527	0.79978
Glu	4.29564	0.77849
Gly	3.86518	1.05203
His	4.32923	0.79982
Ile	4.57110	1.04589
Leu	4.55731	1.02160
Lys	4.16101	0.65241
Met	4.49251	0.97725
Phe	4.69671	1.02315
Pro	4.51078	0.98890
Ser	4.12788	0.98230
Thr	4.15639	0.89542
Trp	5.05033	1.01999
Tyr	4.51586	0.86185
Val	4.28142	0.98227

Tabelle 4.11: Optimierte Parameter zur Berechnung der Oberfläche aus [216].

4.4.6 Wasserstoffbrückenpotential

Die Amideinheit zwischen den C^{α}-Atomen, die die Aminosäuren chemisch miteinander verknüpft, hat in der dreidimensionalen Struktur der Proteine eine weitere wichtige Funktion, indem sie vom Rückgrat ausgehende Wasserstoffbrückenbindungen ermöglicht. In Proteinen in ihrer natürlichen Umgebung sind, wie zu Anfang beschrieben, bis zu 90 % aller Amideinheiten an einer Wasserstoffbrückenbindung beteiligt. Der Hauptanteil der Bindungspartner für die Amideinheit kann drei Klassen zugeordnet werden: Zum einen zu Bindungen zur Umgebung, wie z. B. zu Wassermolekülen, Bindungen zu Seitenketten anderer Aminosäuren, wie Asparaginsäure oder zu anderen Rückgrat-Amideinheiten. Für die Sekundär- und Tertiärstruktur ist besonders die dritte Möglichkeit von entscheidender Bedeutung, da diese Wasserstoffbrückenbindungen maßgeblich zu der Stabilisierung dieser Strukturen beitragen. Aus diesen Gründen ist es sehr wichtig, diese Wechselwirkung mit einzubeziehen.
Die Amideinheit besteht aus einer polaren NHCO-Gruppe, der man einen Dipolvektor vom Wasserstoffatom zum Sauerstoffatom zuordnen kann. Dieser Vektor steht ungefähr senkrecht zur Verbindungslinie der angrenzenden C^{α}-Atomen (bzw. zum Bindungsvektor $\mathbf{b}_i$) und kann aufgrund der großen Freiheiten durch die ϕ- und ψ-Torsionswinkel nahezu frei um diese Verbindungslinie rotieren.
Um die Anzahl an Wechselwirkungszentren klein zu halten, und damit auch die Anzahl an zu optimierenden Parametern, wurde ein virtuelles Zentrum Z eingeführt, das geometrisch auf der Hälfte der Strecke zwischen zwei sequentiellen C^{α}-Atomen liegt und gleichzeitig im Zentrum der Amideinheit. Diese Ansatz wurde auch schon der Literatur beschrieben, siehe beispielsweise [99]. Der Ortsvektor $\mathbf{z}_i$ dieses Zentrums ergibt sich aus $\mathbf{z}_i = \frac{1}{2}(\mathbf{x}_i + \mathbf{x}_{i+1})$ mit den Ortsvektoren der C^{α}_k-Atome $\mathbf{x}_k$. In Abb. 4.26 ist die Lage dieses Zentrums und der Amideinheit veranschaulicht.

Betrachtet man die Anordnung der Amideinheiten in Proteinen, die in regelmäßigen Sekundärstrukturen an Wasserstoffbrücken-Bindungen beteiligt sind, so finden sich hier wiederkehrende Muster. Dies beruht auf dem sterisch eingeschränkten Zugangsbereich zum Rückgrat einer Aminosäure beispielsweise durch die Seitenkette, auf der relativen Fixierung des Torsionswinkels ω durch den partiellen Doppelbindungscharakter der Amidbindung, auf der Elektronenstruktur der an den Wasserstoffbrücken beteiligten Sauerstoff- und Wasserstoffatome, durch die bestimmte Bindungsrichtungen und -abstände bevorzugt werden und letztlich darauf, dass eine Amideinheit im Protein maximal zwei Wasserstoffbrücken ausbildet.

In dieser Arbeit wurde untersucht, inwieweit sich diese Eigenschaften durch ein einfaches Vektormodell beschreiben ließen. Hierzu wurden wieder die 278 Proteine des TOP500H-Satzes untersucht, und für eine Wasserstoffbrückenbindung drei Funktionen definiert: a) eine Abstands-

funktion zweier Zentren zueinander, b) der Winkel zweier Bindungsvektoren zur Richtung der Wasserstoffbrücke und c) der Winkel zweier Wasserstoffbrücken untereinander, die von der gleichen Peptideinheit ausgehen. Für die Erstellung der zugehörigen Statistiken wurde das DSSP-Programm verwendet und nur die Aminosäuren ausgewertet, denen entsprechend diesem Programm eine Wasserstoffbrückenbindung zugeordnet wurde. Die untersuchten Funktionen hatten folgende mathematische Form:

a) Seien $\mathbf{z}_i$ und $\mathbf{z}_k$ zwei Ortsvektoren zu Zentren Z, dann ist der Verbindungsvektor $\mathbf{z}_{ik}$ dieser beiden gegeben zu $\mathbf{z}_{ik} = \mathbf{z}_k$ - $\mathbf{z}_i$. Die Länge dieses Vektors ist $\|\mathbf{z}_{ik}\|$. Dies wurde als Abstandsfunktion ausgewertet.

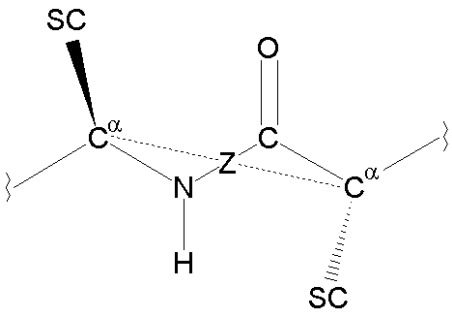


Abbildung 4.26: Position des virtuellen Zentrums Z für die Wasserstoffbrückenbindung in der Amidgruppe. Die gestrichelte Linie zeigt den Bindungsvektor $\mathbf{b}$ der C^{α}-Atome.

b) Da die Wasserstoffbrücken aufgrund der oben erwähnten Eigenschaften nicht in beliebigen Winkeln zueinander auftreten können, wurden die Winkel zwischen den Bindungsvektoren $\mathbf{b}$ zu dem Wasserstoffbrückenvektor $\mathbf{z}_{ik}$ untersucht: Gegeben sei der Vektor $\mathbf{z}_{ik}$ und die zugehörigen Bindungsvektoren $\mathbf{b}_i$ und $\mathbf{b}_k$ der C^{α}-Atome (siehe Abb. 4.27). Ausgewertet wurde das Produkt der Projektionen von $\mathbf{b}_i$ und $\mathbf{b}_k$ auf $\mathbf{z}_{ik}$ durch $\langle \mathbf{z}_{ik}, \mathbf{b}_i \rangle^2 + \langle \mathbf{z}_{ik}, \mathbf{b}_k \rangle^2$, was im Falle sehr ähnlicher Bindungslängen für $\mathbf{b}_i$ und $\mathbf{b}_k$ proportional dem Winkel ist. Die inneren Produkte $\langle \cdot, \cdot \rangle$ wurden quadriert, um die parallelen und antiparallelen Orientierungen der $\mathbf{b}$ zu $\mathbf{z}_{ik}$ gleich zu behandeln, da deren Richtung rein auf deren Definition und nicht auf physikalischen Gegebenheiten beruht und weil die hierzu alternative Betragsfunktion bei der Bildung der Ableitung aufgrund von Unstetigkeitsstellen problematischer ist, während die quadratische Funktion überall stetig ist. Zusätzlich werden die Terme addiert und nicht multipliziert, um Situationen zu vermeiden, in denen eines der beiden inneren Produkte sehr klein ist (nahe Null), wodurch auch das Gesamtprodukt sehr klein wird, während das andere innere Produkt einen großen Wert besitzt, was einer verzerrten Wasserstoffbrücken-Geometrie entsprechen würde. Diese Funktion über die inneren Produkte enthält keine Informationen über den Torsionswinkel, den die beiden beteiligten Bindungsvektoren zueinander haben, welcher über die Atome C_i^{α},C_{i+1}^{α},C_k^{α} und C_{k+1}^{α} gebildet wird.

c) Unter nativen Bedingungen findet zwar, da die thermische Energie nicht ausreichend ist, keine wesentliche Rotation um den Torsionswinkel ω statt, dennoch enthalten Proteine häufig Strukturen, deren ω-Werte vom Idealwinkel abweichen. Dies kann z. B. durch strukturelle

Anforderungen der angrenzenden Proteinbereiche verursacht werden, indem eine solche lokale Verzerrung eine energetisch günstigere Anordnung der anderen Teile ermöglicht. Außerdem sind durch die elektronische Struktur die C=O- und die N-H-Bindung in Wasserstoffbrücken nicht ideal kollinear, sondern in einem bestimmten Winkel zueinander angeordnet. Aus diesen Gründen wurde in der Situation, in der eine Amideinheit zwei Wasserstoffbrücken ausbildet, ausgewertet, welchen Winkel die zwei Wasserstoffbrückenvektoren $\mathbf{z}_{ik}$ und $\mathbf{z}_{ij}$ zueinander einnehmen, die ein gemeinsames Zentrum $\mathbf{z}_i$ besitzen, die aber jeweils zu unterschiedlichen anderen Zentren $\mathbf{z}_k$ und $\mathbf{z}_j$ ausgerichtet sind. Hierfür wurde die Projektion $\langle \mathbf{z}_{ik}, \mathbf{z}_{ij} \rangle$ von $\mathbf{z}_{ik}$ auf $\mathbf{z}_{ij}$ verwendet.
Die mit diesen drei Funktionen erhaltenen Statistiken sind in Abb. 4.28 dargestellt, wobei zusätzlich noch das innere Produkt $\langle \mathbf{z}_{ik}, \mathbf{z}_{ij} \rangle$ in den entsprechenden Winkel $\alpha(\mathbf{z}_{ik}, \mathbf{z}_{ij})$ umgewandelt zur Veranschaulichung mit enthalten ist.

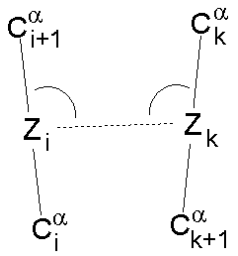


Abbildung 4.27: Wasserstoffbrückenbindungsvektor zwischen zwei Zentren $\mathbf{z}_i$ und $\mathbf{z}_k$.

Aus den Grafiken ist ersichtlich, dass die Länge $\|\mathbf{z}_{ik}\|$ des Wasserstoffbrückenvektors (siehe Abb. 4.28a) auf einen engen Bereich um ca. 4.8 ± 0.2 Å beschränkt ist und dass diese Längenverteilung in den separat aufgeführten Sekundärstrukturen gleich ist. Der zusätzliche Beitrag bei ca. 4.3 Å ist hauptsächlich auf verkürzte, geometrisch ungünstige Wasserstoffbrücken in Windungen zwischen zwei Resten i und $i+3$ zurückzuführen (siehe Einleitung Abschnitt 3.1.2). An diese Verteilung um 4.8 Å wurde eine Funktion der Form

$$f(x) = \frac{A}{1 + B(x + C)^2} \tag{4.43}$$

mit den Parametern A, B und C so angepasst, dass die Lage des Maximums der Funktion und deren Breite mit der Verteilung übereinstimmten. Die Anpassung erfolgte über den Marquardt-Levenberg-Algorithmus. Hieraus resultierte die Basisfunktion für die Länge einer Wasserstoffbrücke:

$$\phi_1^{(HB)}(\mathbf{z}_{ik}) = \frac{c_1^{(HB)}}{1 + 84.02 \left(\|\mathbf{z}_{ik}\| - 4.78\,\text{Å} \right)^2} \tag{4.44}$$

Hierbei wurde für den Koeffizienten $c_1^{(HB)}$ nicht der Wert des angepassten Parameters A übernommen, da $c_1^{(HB)}$ bzw. A die Höhe dieser Funktion und damit deren Wechselwirkungsenergie bestimmt. Statt dessen wurde $c_1^{(HB)}$ im Gesamtoptimierungsprozess des Kraftfeldes neu bestimmt. Damit diese Funktion attraktiv wirkt, wurde $c_1^{(HB)}$ auf negative Werte beschränkt und zunächst nicht abhängig von bestimmten Aminosäuretypen gewählt, da die meisten Aminosäuren gleichmäßig an Wasserstoffbrücken beteiligt sind.

Diese Funktion hat einen ähnlichen Verlauf wie eine Gauss-Funktion mit dem Maximum bei 4.78 Å. Damit wird der repulsive Teil für kleine $\mathbf{z}_{ik}$ nicht richtig beschrieben. Wie sich aber in Faltungstests zeigte, stellte dies kein Problem dar, weil die repulsiven Wechselwirkungen durch andere Funktionen übernommen wurden, wie z. B. die nicht-bindenden Potentiale. Diese einfache funktionale Form wurde zunächst so gewählt, um zu prüfen, ob sich die Wasserstoffbrücken wie in den nativen Strukturen bilden ließen. Für eine spätere Erweiterung war eine Umformulierung in eine Reihenentwicklung analog zu den nicht-bindenden Wechselwirkungen vorgesehen.

Die Summe der Projektionen der Bindungsvektoren $\mathbf{b}_i$ und $\mathbf{b}_k$ auf $\mathbf{z}_{ik}$ (siehe Abb. 4.28b) zeigte einen Trend zu kleinen Werten, wobei für die α-Helix ein Maximum bei ca. 10 Å^4 erreicht wird, während dieses für die Faltblattstruktur bei ungefähr 1 Å^4 liegt. Dies bedeutet, dass in diesen Strukturen beide Bindungsvektoren nahezu orthogonal zum Wasserstoffbrückenvektor $\mathbf{z}_{ik}$ orientiert sind.

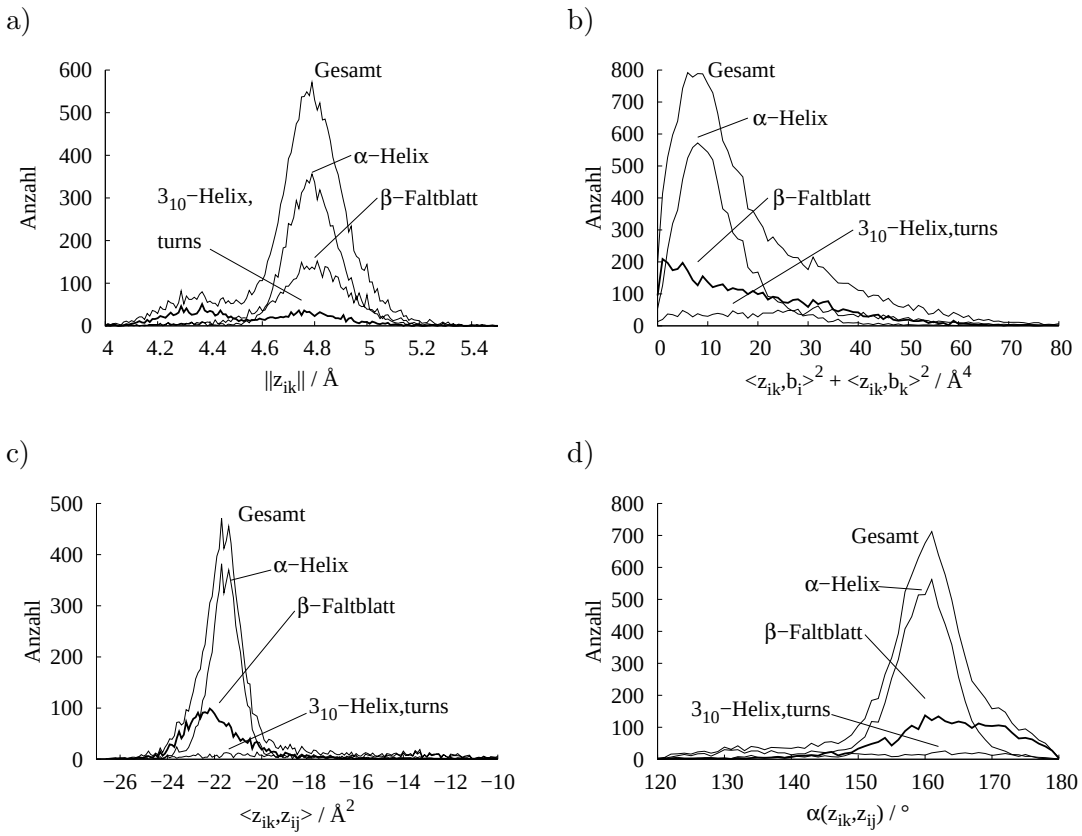


Abbildung 4.28: Darstellung der Statistiken zu den Wasserstoffbrückenfunktionen. a) Länge des Vektors $\mathbf{z}_{ik}$, b) Projektion von $\mathbf{z}_{ik}$ auf die Bindungsvektoren $\mathbf{b}_i$ und $\mathbf{b}_k$, c) inneres Produkt zweier Vektoren $\mathbf{z}_{ik}$ und $\mathbf{z}_{ij}$ und d) Winkel α zwischen $\mathbf{z}_{ik}$ und $\mathbf{z}_{ij}$.

An diese Werte wurde ebenfalls wieder die Funktion aus Gl. 4.43 angepasst. Man erhielt als Basisfunktion $\phi_2^{(HB)}(\mathbf{z}_{ik}, \mathbf{b}_i, \mathbf{b}_k)$, welche den Winkel zwischen den Bindungsvektoren $\mathbf{b}_i$ und $\mathbf{b}_k$ und $\mathbf{z}_{ik}$ beschreibt:

$$\phi_2^{(HB)}(\mathbf{z}_{ik}, \mathbf{b}_i, \mathbf{b}_k) = \frac{c_2^{(HB)}}{1 + 9.85 \cdot 10^{-3} \left(\langle \mathbf{z}_{ik}, \mathbf{b}_i \rangle^2 + \langle \mathbf{z}_{ik}, \mathbf{b}_k \rangle^2 - 8.61 \, \text{Å}^4 \right)^2} \tag{4.45}$$

In dieser Form ist diese Funktion unabhängig vom Abstand der beiden Zentren $\mathbf{z}_i$ und $\mathbf{z}_k$. Dies würde eine unphysikalische Wechselwirkung aller Zentren miteinander implizieren, da sich lediglich nur Bindungsvektoren zueinander orientieren, die eine Wasserstoffbrücke teilen. Wie hierzu verfahren wurde, siehe weiter unten im Abschnitt "Wahl der Wasserstoffbrücken". Der Koeffizient $c_2^{(HB)}$ wurde bei der Optimierung auf negative Werte beschränkt, um das Potential attraktiv zu machen.

Schließlich zeigte die Statistik über den Winkel zweier Wasserstoffbrücken, die an die gleichen Amidgruppe binden (siehe Abb. 4.28c,d), dass es auch hier einen Vorzugswinkel gibt. Das Maximum dieser Verteilung liegt hier bei ungefähr 160°, wobei dieses vor allem durch die α-Helix bestimmt wird, während dagegen β-Faltblatt-Strukturen eine breitere Verteilung hauptsächlich im Bereich zwischen 155° und 175° ohne ein eindeutiges Maximum zeigen. Dies bedeutet, dass für eine α-Helix die beiden Vektoren $\mathbf{z}_{ik}$ und $\mathbf{z}_{ij}$ nicht exakt antiparallel ausgerichtet sind, sondern stets einen Winkel kleiner als ca. 170° zueinander einnehmen, während für das β-Faltblatt auch nahezu antiparallele Orientierungen möglich sind.
Auch hier wurde die Funktion aus Gl. 4.43 an die Verteilung für das innere Produkt $\langle \mathbf{z}_{ik}, \mathbf{z}_{ij} \rangle$, aber nicht an die Winkelfunktion angepasst, um die zusätzlichen numerischen Schritte bei der Umrechnung des inneren Produktes einen Winkel zu vermeiden:

$$\phi_3^{(HB)}(\mathbf{z}_{ik}, \mathbf{z}_{ij}) = \frac{c_3^{(HB)}}{1 + 1.85 \left(\langle \mathbf{z}_{ik}, \mathbf{z}_{ij} \rangle + 21.62 \, \text{Å}^2 \right)^2} \tag{4.46}$$

Der Koeffizient dieser Funktion $c_3^{(HB)}$ wurden ebenfalls bei der Optimierung auf negative Werte beschränkt.

Wahl der Wasserstoffbrücken

Die Bestimmung der Sekundärstrukturen und die Zuordnung der Wasserstoffbrückenbindungen, die für die oben dargestellten Statistiken verwendet wurden, wurden mit Hilfe des DSSP-Programmes vorgenommen. Dieses Programm extrahiert aus PDB-Dateien sehr viele Informationen, um sie in kondensierter Form darzustellen. Das Ziel der Parameteroptimierung war, einen globalen Optimierer zu konstruieren, um mit diesem basierend auf dem Kraftfeld den nativen Zustand vorherzusagen. Aufgrund der Größe des Konformationsraumes müssen sehr

viele Geometrien erstellt und deren Energie berechnet werden. Aus Effizienzgründen sollten dementsprechend die Energie- und Gradientenberechnung möglichst wenig Zeit in Anspruch nehmen. Aus diesen Gründen wurden darauf verzichtet, das DSSP-Programm in der globalen Optimierung zur Bestimmung der Wasserstoffbrücken zu verwenden, da hierzu zunächst die Proteingeometrien in PDB-Dateien hätten umformatiert und dann dem DSSP-Programm übergeben werden müssen, welches dann zusätzlich viele nicht benötigte Daten liefert. Statt dessen wurden eigene Routinen programmiert, die entscheiden, ob zwischen zwei Bindungsvektoren eine Wasserstoffbrücke existiert oder nicht. Dabei wurde von folgender Grundannahme ausgegangen: Der Dipolvektor kann aufgrund der Freiheiten der ϕ- und ψ-Torsionswinkel frei um den Bindungsvektor $\mathbf{b}$ rotieren und sich instantan (in jedem Schritt der Optimierung) auf eine neue Geometrie einstellen. Diese instantane Umorientierung ist darauf begründet, dass eine reine C^{α}-Geometrie im Normalfall nicht mit einer einzigartigen Darstellung des Proteins, in der alle Atomen enthalten sind, korrespondiert, sondern dass verschiedene vollständige Rückgratgeometrien gleichen oder ähnlichen reinen C^{α}-Geometrien entsprechen. Daher lässt sich diese spontane Umorientierung so auffassen, dass die "echte" Orientierung der Amidgruppe im Rückgrat basierend auf einer C^{α}-Geometrie eigentlich nicht bekannt ist, dass sie aber am Ende der Faltung die energetisch günstigste Orientierung einnehmen wird.
Hierauf aufbauend wurden Programme implementiert, die die Umgebung eines Bindungsvektors $\mathbf{b}_i$ nach Wasserstoffbrücken-Bindungspartnern absuchen. Basierend auf den oben beschriebenen Statistiken und Potentialfunktionen wurde mehrere Kriterien verwendet, um zu entscheiden, ob am Zentrum $\mathbf{z}_i$ eine Wasserstoffbrücke vorliegt: Zunächst wurden alle Zentren $\mathbf{z}_j$ gesammelt, die im Abstand von 4.6 bis 5.0 Å um $\mathbf{z}_i$ lagen. Dann wurde untersucht, ob die zugehörigen Bindungsvektoren $\mathbf{b}_j$ das Projektionskriterium $\langle \mathbf{z}_{ik}, \mathbf{b}_i \rangle^2 + \langle \mathbf{z}_{ik}, \mathbf{b}_j \rangle^2 \leq 30\,\text{Å}^4$ erfüllen, welches bestimmt, ob die Bindungsvektoren so zueinander orientiert sind, dass eine Wasserstoffbrücke möglich ist. Nach Anwendung dieser beiden Kriterien konnten mehrere Fälle eintreten: 1. Es gab ein Paar von Bindungsvektoren $\mathbf{b}_i$ und $\mathbf{b}_j$, die die Kriterien erfüllten. In diesem Fall wurde dieses als gebildete Wasserstoffbrücke aufgefasst; 2. Es gab zwei Paare. Dann wurde zusätzlich überprüft, ob beide Zentren-Verbindungsvektoren $\mathbf{z}_{ik}$ und $\mathbf{z}_{ij}$ antiparallel "genug" sind, so dass es physikalisch möglich ist, dass sie zum gleichen Zentrum gehören. Hierzu wurde überprüft, ob $\langle \mathbf{z}_{ik}, \mathbf{z}_{ij} \rangle \leq -20$ galt. Wenn dieses erfüllt war, wurden beide Wasserstoffbrücken akzeptiert. Hierbei wurde zusätzlich geprüft, ob diese Antiparallelität auch für die bereits vorher festgelegten Wasserstoffbrücken-Bindungen galt, wenn diese eine gemeinsame Wasserstoffbrücke mit einem dieser Zentren besaßen. Wenn die Bindungsvektoren nicht antiparallel waren, wurde nur die Bindung verwendet, die nach Gl. 4.44 die niedrigste Abstandsenergie lieferte. Die andere Bindung wurde als physikalisch nicht möglich verworfen. 3. Es gab mehrere mögliche Paare. Trat dieser Fall ein, wurden mit den Bindungsvektoren alle kombinatorisch möglichen Paarungen gebildet, die Paarungen nach Antiparallelität wie zuvor beschrieben selektiert und zum Schluss das antiparallele Paar verwendet, dass nach Gl.

4.44 den niedrigsten Gesamtwert für die Summe beider H-Brücken lieferte. Erfüllte kein Paar die Antiparallelität, so wurde wieder nur eine Wasserstoffbrücke angenommen.

Die Zuordnung der Wasserstoffbrücken gemäß der oben beschriebenen Kriterien erfolgte schrittweise für jede Aminosäure einzeln entlang der Sequenz beginnend am N-Terminus. Diese Art der Festlegung wurde so gewählt, um in der Situation, in der mehrere mögliche konkurrierende Gesamtanordnungen des H-Brücken-Netzwerkes vorlagen, eine separate globale Optimierungssuche nach der besten Anordnung der Wasserstoffbrücken zu vermeiden, um den Algorithmus an dieser Stelle nicht weiter zu verlangsamen. Hierauf basierend wurden die möglichen Wasserstoffbrücken weiter selektiert, indem die Wasserstoffbrücken-Bindungen des Zentrums $\mathbf{z}_j$ zum Zentrum $\mathbf{z}_i$ mit $j > i$ ausgeschlossen wurden, wenn die bereits vorher festgelegten Wasserstoffbrücken an $\mathbf{z}_i$ nicht auch $\mathbf{z}_j$ mit eingeschlossen hatten.

Diese Kriterienanalyse wurde anstelle der einfachen Berechnung der Gesamtenergie im Hinblick darauf betrieben, dass jede Aminosäure maximal zwei Wasserstoffbrücken ausbilden kann, die zusätzlich auch noch eine Vorzugsorientierung besitzen. Bei unkritischer Energieberechnung über alle Aminosäuren mittels der oben beschriebenen Potentialfunktionen können ansonsten unphysikalische Überstabilisierungen auftreten, da es beispielsweise in der Umgebung eines Zentrums viele andere Zentren geben kann, die alleine durch die Abstandsfunktion $\phi_1^{(HB)}$ Energie beitragen. Um dies zu vermeiden, wurden die Menge der möglichen Wasserstoffbrücken auf die physikalisch sinnvollen reduziert.
Diese Prozedur wurde im genetischen Algorithmus bei der globalen Optimierung angewendet. Hierbei wurden die Wasserstoffbrücken-Potentialfunktionen nur auf diejenigen Aminosäuren angewendet, die nach dieser Analyse als Paare "markiert" waren. In der Parameteroptimierung dagegen wurde zur Bestimmung der der Wasserstoffbrücken das DSSP-Programm verwendet.

4.5 Parameteroptimierung

4.5.1 Methoden der Kraftfeldparametrisierung

Die Parametrisierung eines Kraftfeldes setzt neben der Wahl der Potential- bzw. Basisfunktionen des weiteren die Definition der zu reproduzierenden Zielgrößen und eine Methode voraus, mit welcher die optimalen Parameter im Raum der gewählten Basisfunktion bestimmt werden.
In der Literatur existieren viele verschiedene empirische Potentiale, deren Parameter mit ebenso unterschiedlichen Methoden und Zielvorgaben bestimmt wurden, zu denen in der Tabelle 4.12 ein paar Beispiele gegeben werden. Diese Übersicht ist nicht vollständig, weder in den erwähnten Kraftfeldern, noch in den Methoden der Parametergewinnung. Tab. 4.12 soll lediglich verdeutlichen, dass die Kraftfeld-Parametrisierung ein breites Spektrum an Techniken abdeckt.

In der modernen theoretischen Forschung an der Proteinfaltung zur Vorhersage des nativen Zustandes wird heutzutage meist entweder rein mit vergröberten Darstellungen oder mit einem Wechsel zwischen unterschiedlichen Darstellungen des Proteins und somit ebenfalls mit einem Wechsel zwischen unterschiedlichen Potentialen gearbeitet [218]. Hierbei wird zunächst zu Beginn eine vergröberte Darstellung und entsprechend ein vergröbertes Potential verwendet, um größere Bereiche des Konformationsraumes abzudecken, während anschließend, wenn der (vermeintliche) native Zustand in der vergröberten Darstellung identifiziert wurde, zu einem besser aufgelösten bis hin zu einem atomaren Kraftfeld gewechselt wird, in welchen die Strukturdetails optimiert werden. Diese Technik wird angewendet, da es zu Anfang einer Strukturvorhersage ohne Verwendung externer Informationen unklar ist, wo sich der native Zustand auf der Energiefläche befindet, so dass zunächst eine große Anzahl an Punkten auf der Fläche berechnet werden muss, um deren Form und somit die Position des nativen Zustandes bestimmen zu können. In einer vollständigen atomaren Auflösung ist eine solche Berechnung sehr aufwendig, so dass eine vergröberte Darstellung verwendet wird, um die Suche in vernünftiger Zeit zu realisieren.
Eine Reduktion der Anzahl an Interaktionspunkten in einer vergröberten Darstellung führt dazu, dass diese mehreren realen Punkten bzw. Atomen in einem Protein entsprechen, so dass die hierzu angesetzten Potentialfunktionen effektiven Wechselwirkungen entsprechen, also einer Summe atomarer Wechselwirkungen. Die Bestimmung der Parameter und Gewichtungskoeffizienten muss für diese Kräfte daher einen Prozess der Mittelung bzw. Gewichtung der zugrundeliegenden Wechselwirkungen beinhalten. Ein Beispiel hierfür aus den in der Tabelle 4.12 aufgeführten Potentialen ist das UNRES-Kraftfeld, in welchem die Parameter zum

Kraftfeld	Parametrisierung
AMBER	Reproduktion von Dichte und Enthalpien flüssiger Phasen und Anpassung an quantenmechanische Rechnungen [219]
CHARMM	Informationen aus Schwingungsdaten und kristallographischen Strukturen [220]
ECEPP	Anpassung an semi-empirische Potentiale [221–224]
MM2	Optimiert zur Reproduktion von Bildungsenthalpien [225]
UNRES	Bildung von Durchschnittswerten der Parameter des ECEPP-Kraftfeldes und Z-Wert-Optimierung [99–101]
YAMBER	Reparametrisierung des AMBER-Kraftfeldes mit einem *Simulated-annealing*-Monte-Carlo-Verfahren zur Minimierung des RMSD-Wertes zu Kristallstrukturen [226]

Tabelle 4.12: Beispiele für die Parameteroptimierung verschiedener Kraftfelder. Zur Erklärung des Z-Wertes siehe Text Seite 122 bzw. Gl. 4.48.

Teil dadurch bestimmt wurden, dass eine Mittelwertbildung über Funktionen einer atomaren Darstellung, welche in diesem Fall durch das ECEPP-Potentials gegeben war, durchgeführt wurde.

Wie aus der Tabelle 4.12 ersichtlich ist, werden viele Kraftfelder an allgemeine experimentelle thermodynamische Daten angepasst. Dies geschieht vornehmlich mit dem Ziel, eine große Anzahl an Systemen mit unterschiedlicher Zusammensetzung simulieren zu können. Zusätzlich zu diesen Kraftfeldern existieren stärker spezialisierte Kraftfelder, mit eng begrenztem Aufgaben- bzw. Einsatzgebiet wie beispielsweise reine Wasserkraftfelder. Solche Kraftfelder sind auch auf dem Gebiet der Proteinfaltung häufig vertreten. Diese sind darauf spezialisiert, den nativen Zustand einer Aminosäuresequenz vorherzusagen. Für die Parametrisierung dieser "Protein-Kraftfelder" haben sich in der Literatur drei Verfahren etabliert, bei welchen die Parameter bzw. Gewichtungskoeffizienten der Potentialfunktionen nicht anhand beispielsweise thermodynamischer Größen bestimmt werden, sondern diese zielgerichtet direkt an der Problematik der Strukturvorhersage orientiert optimiert werden. Diese Verfahren sind die Optimierung über den quasichemischen Ansatz, die Minimierung des Z-Wertes und die lineare Optimierung.
Gemein zwischen den drei Methoden ist hierbei, dass die Parameter über einen direkten Vergleich zwischen aus Datenbanken stammenden nativen Proteinen und falschen Strukturen bestimmt werden, deren Sequenz zwar mit dem nativen Protein identisch ist, die aber eine andere dreidimensionale Struktur besitzen (siehe hierzu auch Abschnitt 4.3). Die Parameter der gewählten Potentialfunktion (oder allgemeiner Bewertungsfunktion) werden dann so bestimmt,

dass die nativen Strukturen einen besseren, normalerweise kleineren Wert, mit der gewählten Potentialfunktion liefern als die zugehörigen falschen Strukturen, wodurch diese voneinander unterschieden werden können. Das Ziel ist, im Rahmen einer *Ab-initio*-Strukturvorhersage die native Geometrie so in einem großen Satz an Proteinstrukturen identifizieren zu können.

Durch die Methodik der Potential- bzw. Modellvergröberung unterscheidet sich im Normalfall die reale Energiefläche von derjenigen Fläche, auf welcher die Strukturvorhersage durchgeführt, da die Vergröberung einer Glättung der Energiefläche entspricht, indem bestimmte Details ausgelassen werden. Weiterhin unterscheiden sich die Flächen dadurch, dass bei der Parametrisierung durch die hier angesprochenen Methoden nicht die exakte Reproduktion der Fläche das Ziel ist, sondern die Erzeugung einer Fläche, in welcher der native Zustand das globale energetische Minimum ist, während alle anderen Geometrien eine höhere Energie besitzen. Dementsprechend sind auch die energetischen Abstände zwischen zwei Zuständen modell- und methodenabhängig und korrelieren nicht zwangsläufig mit experimentellen Daten. Dadurch sind faltungsdynamische Ergebnisse oder thermodynamische Eigenschaften, die auf Basis solcher Funktionen berechnet wurden, stets zu hinterfragen und mit Vorsicht zu behandeln.
Im folgenden werden zunächst die beiden beiden Methoden des quasichemischen Ansatzes und der Z-Wert-Optimierung kurz vorgestellt und die in dieser Arbeit verwendete Methode der linearen Optimierung anschließend ausführlicher dargestellt.

Quasichemischer Ansatz

Im Rahmen dieses Ansatzes, der große Verbreitung in der Literatur gefunden hat, wird davon ausgegangen, dass sich in einem Protein im nativen Zustand die Anzahl der Kontakte zwischen den Interaktionspunkten, welche beispielsweise durch die Seitenketten oder die C^{α}-Atome gegeben sein können, entsprechend der Boltzmann-Verteilung verhalten. Auf Basis dieser Annahme, die sich in verschiedenen Untersuchungen als zutreffend erwiesen hat [227, 228], wird die Wechselwirkungsenergie e zwischen zwei Aminosäuren a und b allgemein angesetzt als [229]:

$$e = -k_B T \ln\left(\frac{p_{a,b}}{\bar{p}_{a,b}}\right) \tag{4.47}$$

wobei k_B die Boltzmann-Konstante und T die Temperatur sind. Die entscheidenden Größen hierbei sind $p_{a,b}$ und $\bar{p}_{a,b}$, welche in der Literatur verschieden definiert bzw. benutzt werden. Generell geben diese Werte die Häufigkeit wieder, mit welcher das Paar der Aminosäuren a und b in einem Satz von Proteinstrukturen auftaucht. Der Wert $p_{a,b}$ gibt hierbei die Häufigkeit der Kontakte in den nativen Strukturen wieder, welche sich beispielsweise durch Auswertung

bekannter Datenbanken relativ einfach ermitteln lassen. Diese werden in Relation zu der Häufigkeit $\bar{p}_{a,b}$ des selben Paares in einem Referenzdatensatz an Strukturen gesetzt. Die Wahl des Referenzdatensatzes ist eine wichtige und schwierigere Aufgabe, die in der Literatur unterschiedlich gelöst wird (siehe z. B. [230]).
Mit Hilfe dieses Ansatzes werden Energien in diskreten Abstandsintervallen bestimmt, deren Breite generell 1 Å nicht unterschreitet, um eine verlässliche Statistik als Grundlage zu haben. Bei zu schmalen Intervallen können viele Intervalle zu spärlich oder gar nicht besetzt sein.
Potentiale, die aus dieser Methode resultieren, enthalten im Allgemeinen, da sie auf diskreten Intervallen bestimmt wurden, keine glatten Potentialfunktionen, weshalb diese Methode im Rahmen dieser Arbeit ungeeignet war und nicht verwendet wurde.

Z-Wert-Optimierung

Eine weitere Methode, die Gewichtungskoeffizienten zur Erzeugung von Flächen zur Proteinstrukturvorhersage zu bestimmen, ist die Z-Wert-Optimierung (*z-score optimization*) [129, 231].
Der Z-Wert ist generell ein allgemeines Merkmal einer statistischen Verteilung. In der Literatur der Parameteroptimierung für Proteinkraftfelder gegen falsche Strukturen wird dieser gewöhnlich in folgender Form verwendet, wobei verschiedene Autoren unterschiedliche Definitionen verwenden:

$$Z = \frac{E^* - \mu_{E,dec}}{\sigma_{E,dec}} \tag{4.48}$$

Hierbei sind E^* die Energie des nativen Zustandes, $\mu_{E,dec}$ der Mittelwert der Energien der falschen Strukturen und $\sigma_{E,dec}$ dessen Standardabweichung. Der Z-Wert enthält Informationen darüber, ob die Zuordnung des nativen Zustandes statistisch signifikant ist oder ob mit einer Zufallszuordnung das gleiche Resultat hätte erzielt werden können. Je kleiner bzw. negativer dieser Wert ist, desto signifikanter ist die Zuordnung.
Bei der Verwendung des Z-Wertes zur Bestimmung der Gewichtungskoeffizienten werden zwei Merkmale optimiert: Zum einen der Abstand der Energie der nativen Struktur zum Energiemittelwert der falschen Strukturen. Bei einer breiten Streuung der Energien ist es jedoch nicht gewährleistet, dass der native Zustand stets energetisch niedriger als die falschen Strukturen ist. Zum anderen wird die Streuung der Energien der falschen Strukturen über die Standardabweichung mit berücksichtigt.
Diese Methode zur Optimierung der Gewichtungskoeffizienten war eine mögliche Alternative für diese Arbeit, da aber die dritte Methode, die lineare Optimierung, weitere Vorteile bietet, die weiter unten beschrieben werden, wurde die lineare Optimierung der Z-Wert-Optimierung vorgezogen. Zudem liefern beide Methode Ergebnisse, die sich nicht in wesentlichen Punkten unterscheiden, so dass beide Methoden als äquivalent betrachtet werden können [232].

4.5.2 Einführung lineare Optimierung

In diesem Abschnitt wird eine kurze allgemeine Einführung in die Methodik der linearen Optimierungstheorie (*linear programming*) bzw. deren Anwendungen gegeben. Da die lineare Optimierung in dieser Arbeit aber nur als Hilfsmittel verwendet wurde, die in Form bereits vorhandener Programme zur Anwendung kam und dementsprechend keine Methodenentwicklung stattfand, wird dieses Thema an dieser Stelle folglich nur sehr oberflächlich beschrieben. Für eine tiefer gehende und präzisere Darstellung der zugrundeliegenden Mathematik sei daher hier auf die entsprechende Literatur verwiesen. Siehe hierzu beispielsweise [233–235].

Die lineare Optimierung ist ein Verfahren, das vielfältige Anwendungen besonders in ökonomisch orientierten Bereichen findet. Die allgemeine Problemstellung besteht darin, den Funktionswert einer sogenannten Zielfunktion f (*object function*) zu maximieren oder zu minimieren. Diese Zielfunktion hängt von einem Satz Variablen $\{x_i \,|\, i = 1, \ldots, n\}$ ab, die in der Funktion f stets nur als lineare Faktoren $\alpha_i x_i$ mit Koeffizienten α_i auftreten, so dass die Zielfunktion die generelle Form $f(x_1, \ldots, x_n) = \sum_{i=1}^{n} \alpha_i x_i = \boldsymbol{\alpha}^T \mathbf{x}$ besitzt. Für andere Funktionstypen wie beispielsweise Funktionen, die quadratisch von ihren Variablen abhängen, existieren separate Lösungsmethoden, auf die hier nicht näher eingegangen wird. Zu dieser Zielfunktion existiert weiterhin ein endlicher Satz an linearen Nebenbedingungen, welche den Raum beschränken, innerhalb dessen der maximale oder minimale Funktionswert zu suchen ist. Diese definieren die Bedingungen, die die Lösungen des Problems erfüllen müssen. In der Standardform eines linearen Optimierungsproblems treten die m Nebenbedingungen als Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ auf, wobei $\mathbf{A}$ die m, n-Koeffizientenmatrix, $\mathbf{x}$ den n-dimensionalen Lösungsvektor und $\mathbf{b}$ den m-dimensionalen Rechte-Seite-Vektor darstellt. Zusätzlich werden in der Standardform sämtliche Variablen $x_i \geq 0$ angesetzt. Formal ergibt sich somit folgende Standardform des Problems:

$$\text{minimiere} \quad f(x_1, \ldots, x_n) = \boldsymbol{\alpha}^T \mathbf{x} \tag{4.49}$$

mit den Nebenbedingungen

$$\mathbf{Ax} = \mathbf{b} \tag{4.50}$$

$$\mathbf{x} \geq \mathbf{0} \tag{4.51}$$

oder kürzer formuliert

$$\min \boldsymbol{\alpha}^T \mathbf{x},\ \mathbf{Ax} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \tag{4.52}$$

Diese Form ist in der Regel die Grundlage, auf welcher Lösungsalgorithmen arbeiten. Die mathematische Formulierung auftretender realer bzw. anwendungsbezogener Problemstellungen

führt jedoch meist zu einer hiervon abweichenden Form, indem beispielsweise nicht nur Gleichungen, sondern auch Ungleichungen als Nebenbedingungen auftreten, oder indem Variablen eine andere Definitionsmenge besitzen. Für diese Fälle existieren aber mathematische Methoden, ein allgemeines Problem auf die oben angegebene Standardform zu überführen. Hierzu sollen zur Veranschaulichung kurz ein paar Beispiele gegeben werden:

- Gewöhnlich ist in einem Optimierungsalgorithmus entweder nur eine Maximierung oder nur eine Minimierung einer Zielfunktion realisiert, da beide Optimierungsrichtungen ineinander überführbar sind. Die Richtung *minimiere* $f(x_1, \ldots, x_n)$ ist äquivalent zu *maximiere* $-f(x_1, \ldots, x_n)$.
- Ungleichungen u in den Nebenbedingungen vom Typ $u(x_1, \ldots, x_n) \leq \beta$ mit $\beta \in \mathbb{R}$ können durch Einführung einer Schlupfvariablen $s \geq 0$ (*slack variable*) in eine Gleichung der Form $u(x_1, \ldots, x_n) = \beta + s$ überführt werden. Die Schlupfvariable wird dann als zusätzliche Variable in die Optimierung miteinbezogen, was durch eine Identitätstransformation verdeutlicht werden kann, indem man $s \equiv x_{n+1}$ setzt, und $u(x_1, \ldots, x_n)$ zu $u'(x_1, \ldots, x_n, x_{n+1})$ überführt.
- Beschränkte Variablen vom Typ $x_i \geq \gamma$ mit $\gamma \in \mathbb{R}$ lassen sich durch Einführung einer neuen Variablen x_i' substituieren, indem $x_i = x_i' + \gamma$ in den Gleichungen ersetzt wird.
- Unbeschränkte (freie) Variablen $x_i \in [-\infty, \infty]$ können als Differenz $x_i = x_i^{(+)} - x_i^{(-)}$ zweier nicht-negativer Variablen $x_i^{(+)} \geq 0$ und $x_i^{(-)} \geq 0$ formuliert werden.

Neben diesen Beispielen existieren weitere Verfahren, um Probleme in die Standardform zu überführen. Diese Methoden werden in der entsprechenden Literatur ausführlich behandelt.

Im mehrdimensionalen Raum lassen sich die Nebenbedingungen als Ebenen darstellen, deren Gesamtheit den Lösungsraum einschließt, welcher dadurch anschaulich die Gestalt eines Polyeders erhält. Alle Punkte im Inneren dieses Polyeders stellen mögliche Lösungen des Problems dar. Diejenigen Punkte, die den Funktionswert der Zielfunktion optimieren, liegen auf dem Rand (auf der Oberfläche) des Polyeders. Dies kann in Abhängigkeit von der Art der Zielfunktion eine Seite, eine Kante oder ein Eckpunkt sein. Können nicht alle Nebenbedingungen erfüllt werden, so ist das Problem unlösbar.

Die Entstehung der Lösungsalgorithmen linearer Optimierungsprobleme ist eng mit der Entwicklung der Computertechnik verbunden. Wie auch in anderen Gebieten der Mathematik fand die Entwicklung und Verbesserung der Algorithmen in diesem Bereich auf unterschiedlichen Wegen statt, repräsentiert durch verschiedene Personen, die hinter diesen standen, bis sich ca. in der Mitte des 20. Jahrhunderts das umfassende Forschungs- und Anwendungsgebiet der Unternehmensforschung (*operations research*) als Teilgebiet der angewandten Mathema-

tik bildete und etablierte. Zur Lösung von Praxisproblemen brachte dieser Forschungszweig vor allem zwei wichtige Methoden hervor, die hier besonders erwähnt seien, weil sie heute weite Verbreitung gefunden haben und durch Implementation in vielen unterschiedlichen Programmpaketen vielfältig angewendet werden. Es handelt sich zum einem um das Simplex-Verfahren und zum anderen um die Innere-Punkte-Methoden (*interior points methods*).
Während das Simplex-Verfahren zur Auffindung einer optimalen Lösung von Eckpunkt-zu-Eckpunkt auf der Oberfläche des Polyeders des Lösungsraumes springt, bewegen sich die Innere-Punkte-Methoden, wie bereits im Namen deutlich wird, entlang einer Serie von Punkten, die sich im Inneren des Lösungsraum-Polyeders befinden zur Oberfläche.
Die zeitlich später entwickelten Innere-Punkte-Methoden besitzen im Vergleich zum Simplex-Algorithmus den Vorteil, dass bei diesen der rechentechnische Aufstand mit der Größe des zu lösenden Problems weit weniger stark ansteigt als beim Simplex-Verfahren. Die Rechenzeit des Simplex-Algorithmus skaliert im schlechtesten Fall exponentiell, während die Innere-Punkte-Methoden polynomial skalieren. Da diese Arbeit die Lösung eines sehr großen linearen Optimierungsproblems beinhaltete, und die Innere-Punkte-Methoden sehr robust sind und zudem gegenüber dem Simplex-Verfahren den Rechenzeitvorteil besitzen, wurden sie in dieser Arbeit zur Bestimmung der Parameter verwendet.

Da der Innere-Punkte-Lösungsalgorithmus an sich nicht selber programmiert oder verändert wurde, sondern hierzu ein fertiges Programm verwendet wurde, wird im folgenden der generelle Ablauf der Methode nur im groben skizziert, ohne auf die mathematischen Details oder die explizite Implementation einzugehen. An dieser Stelle sei auf die entsprechende mathematischer Literatur verwiesen, in der die Methoden ausführlich und exakt behandelt werden.

Zur Lösung des linearen Optimierungsproblems wurde das Programm BPMPD von C. Mészáros verwendet [236–238], welches zum Beispiel in [239] erhältlich ist. Dieses verwendet den primalen-dualen Innere-Punkte-Algorithmus. Hierbei wird von dem oben formulierten Standardproblem

$$\min \boldsymbol{\alpha}^T \mathbf{x},\ \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \tag{4.53}$$

ausgegangen und zu diesem das zugehörige duale Problem

$$\max \mathbf{b}^T \mathbf{y},\ \mathbf{A}^T \mathbf{y} \leq \boldsymbol{\alpha},\ \mathbf{y} \in \mathbb{R}^m \tag{4.54}$$

definiert, wobei m die Anzahl an Nebenbedingungen, $\mathbf{A}$ die m, n-Koeffizienten-Matrix, $\mathbf{y}$ der m-dimensionale Variablenvektor und $\mathbf{b}$ der m-dimensionale Koeffizientenvektor des dualen Problems ist. Die Überführung in das duale Problem bietet häufig die Vorteile, dass dieses einfacher als das primale zu lösen ist, und dass dieses zusätzliche Informationen bereitstellen kann. Weiterhin ist die duale Form des dualen Problems wiederum das primale. Der Variablenvektor $\mathbf{x}$ des primalen Problems hängt hierbei mit dem Variablenvektor $\mathbf{y}$ des dualen

Problems über die Beziehung $\mathbf{c}^T\mathbf{x} \leq \mathbf{b}^T\mathbf{y}$ zusammen. Eine Lösung des dualen Problems ist somit eine obere Schranke für das primale Problem. Die optimale Lösung des dualen Problems $\bar{\mathbf{y}}$ ist aber gleichzeitig auch die optimale Lösung des primalen Problems $\bar{\mathbf{x}}$, so dass in diesem Fall $\mathbf{c}^T\bar{\mathbf{x}} = \mathbf{b}^T\bar{\mathbf{y}}$ ist.

Zur Durchführung der Lösung werden im ersten Schritt in Gl. 4.54 die Ungleichungen $\mathbf{A}^T\mathbf{y} \leq \boldsymbol{\alpha}$ durch Einführung von Schlupfvariablen $\mathbf{s}$ in Gleichungen überführt

$$\mathbf{A}^T\mathbf{y} \leq \boldsymbol{\alpha} \tag{4.55}$$

$$\Rightarrow \mathbf{A}^T\mathbf{y} + \mathbf{s} = \boldsymbol{\alpha} \tag{4.56}$$

mit $\mathbf{s} = (s_1, \ldots, s_m)^T$ und $s_i \geq 0$. Zur praktischen Umsetzung der Voraussetzung, dass $s_i \geq 0$ ist, werden sogenannte logarithmische Grenzen eingeführt (*logarithmic barrier*), über welche die Schlupfvariablen zur Zielfunktion hinzugefügt werden. Dazu wird zunächst eine Konstante $\mu > 0$ gewählt und mit dieser die Zielfunktion formuliert als

$$\max \mathbf{b}^T\mathbf{y} + \mu \sum_{i=1}^{m} \ln(s_i) \tag{4.57}$$

Da $\lim_{s_i \to 0} \ln(s_i) = -\infty$ erzeugt dies eine Tendenz, dass die Schlupfvariablen nicht gegen Null streben und somit innerhalb des erlaubten Bereiches bleiben, wobei dieser Term um den Faktor μ gewichtet wird. Tatsächlich hängen die Ergebnisse der Optimierung von der Wahl von μ ab, so dieses im Optimierungsprozess mitverändert wird, wie weiter unten beschrieben wird. Damit ergibt sich das zu optimierende Problem aus der Zielfunktion Gl. 4.57 und den Nebenbedingungen Gl. 4.56. Diese können über Lagrange-Multiplikatoren in einer gemeinsamen Gleichung zusammengefasst werden. Die Optimalitätsbedingungen ergeben sich dann aus dem Nullsetzen der ersten Ableitung dieser Lagrange-Funktion bezüglich der Variablen $\mathbf{x}, \mathbf{y}$ und $\mathbf{s}$. Dies führt zu den Gleichungen

$$\mathbf{A}^T\mathbf{y} + \mathbf{s} = \boldsymbol{\alpha} \tag{4.58}$$

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{4.59}$$

$$\mathbf{X}\mathbf{S}\mathbf{e} = \mu\mathbf{e} \tag{4.60}$$

wobei $\mathbf{X} = \mathrm{diag}(x_1, \ldots, x_n)$, $\mathbf{S} = \mathrm{diag}(s_1, \ldots, s_m)$ Diagonalmatrizen sind und $\mathbf{e}$ die Einheitsmatrix. Zur Lösung werden diese Gleichungen nach Null umgestellt, so dass sich die Gleichungen

$$\mathbf{A}^T\mathbf{y} + \mathbf{s} - \boldsymbol{\alpha} = \mathbf{0} \tag{4.61}$$

$$\mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0} \tag{4.62}$$

$$\mathbf{X}\mathbf{S}\mathbf{e} - \mu\mathbf{e} = \mathbf{0} \tag{4.63}$$

ergeben. Zur numerischen Bestimmung der Nullstellen dieser Funktionen wird das Newton-Raphson-Verfahren angewendet, bei welchem ausgehend von einem Startpunkt, der hier einer anfänglichen erlaubten Lösung entspricht, sich iterativ den Nullstellen genähert wird, was einem Pfad innerhalb des oben beschriebenen erlaubten Lösungsraumpolyeders entspricht und wobei die Nullstelle der optimalen Lösung entspricht (bzw. ggf. mehreren Nullstellen und Lösungen). Für die Details zur iterativen Bestimmung der Richtungen und Wahl der Schrittlängen dieses Verfahrens sei auf die entsprechende Literatur verwiesen.
Der numerisch bestimmte optimale Lösungspunkt hängt von der Wahl des Parameters μ ab (siehe Gl. 4.57). Dieser Parameter wird während des Optimierungsprozesses ebenfalls minimiert, wodurch sich der optimale Punkt dem "echten" optimalen Punkt, welcher sich für $\mu = 0$ ergibt, nähert. Eine Methode μ zu wählen ist beispielsweise $\mu = 1$ in der ersten Iteration zu setzen und danach $\mu = 10^{1-k}$ für die k-te Iteration zu verwenden.
Ein analytischer Optimalitätspunkt $\bar{\mathbf{x}}, \bar{\mathbf{y}}$ und $\bar{\mathbf{s}}$ ist erreicht, wenn der Wert $\boldsymbol{\alpha}^T\bar{\mathbf{x}} - \bar{\mathbf{b}}\bar{\mathbf{y}}$ (*duality gap*) nahe Null ist und wenn diese Variablen die Gleichungen 4.58 und 4.59 erfüllen. Gl. 4.60 wird im Programm nicht als Optimalitätskriterium verwendet, was weiter unten erläutert wird. Numerisch wird hier die Optimalität erreicht, wenn die linken Seiten der Gleichungen 4.61 und 4.62 sowie $\boldsymbol{\alpha}^T\bar{\mathbf{x}} - \bar{\mathbf{b}}\bar{\mathbf{y}}$ kleiner als ein zuvor vorgegebener Wert ϵ sind. Dieser wird im Programm zu $\epsilon = 10^{-p}$ festgelegt, wobei $p \in \mathbb{N}$ die Maschinengenauigkeit angibt, mit der Fließkomma-Berechnungen durchgeführt werden können. Hierfür wird in der Literatur meist ein Wert von $p = 8$ angesetzt, dieser kann aber vom Benutzer eingestellt werden. Dementsprechend liegt ein (numerischer) Optimalpunkt vor, wenn die Bedingungen

$$\mathbf{A}^T\mathbf{y} + \mathbf{s} - \boldsymbol{\alpha} \leq 10^{-8} \tag{4.64}$$

$$\mathbf{A}\mathbf{x} - \mathbf{b} \leq 10^{-8} \tag{4.65}$$

$$\boldsymbol{\alpha}^T\mathbf{x} - \mathbf{b}\mathbf{y} \leq 10^{-8} \tag{4.66}$$

erfüllt sind. Die Bedingung Gl. 4.60 bzw. 4.63 wird im Programm nicht als Kriterium der Optimalität herangezogen, da diese Gleichungen nicht zwangsweise erfüllt werden, weil der Wert des vom Benutzer vorgegebenen Parameters μ von der Anzahl an Iterationen abhängt und dadurch in der Nähe eines Optimums einen großen Wert besitzen kann, wenn bspw. der Startpunkt der Optimierung bereits nahe des Optimums lag.
(Als Hinweis sei an dieser Stelle noch erwähnt, dass die oben beschriebenen Optimalitätsbedingungen im Programm in einer skalierten Form realisiert sind. Für die Details siehe [238].)

4.5.3 Das MPS-Format

Zusammen mit der Entwicklung der kommerziellen bzw. industriellen Programmpakete zur linearen Optimierung ist auch ein Format-Standard entstanden, mit welchem die Daten des Problems den Programmen zur Verfügung gestellt werden können, welches auch das hier implementierte BPMPD-Programm verwendet. Hierbei handelt es sich um das MPS-Format (*Mathematical Programming System*), welches das grundlegende Prinzip verwendet, das zu lösende (Un-)Gleichungssystem in spaltenweiser Form (Spaltenvektoren) darzustellen. Ebenfalls im Rahmen einer kurzen Einführung wird dieses Format im folgenden in einer einfachen Form beschrieben, die nicht alle Details wiedergibt. Zusätzlich wird hierzu zur Veranschaulichung ein kurzes Beispielproblem gegeben. Da es sich bei dem MPS-Format um ein relativ einfach strukturiertes und weit verbreitetes Anwendungsformat handelt, erhält man heutzutage besonders im Internet schnell Informationen zum Aufbau und zur Verwendung (siehe z. B. [240, 241]).

Das MPS-Format entstammt der Anfangszeit der Computertechnik, in welcher Lochkarten zur Informationseingabe benutzt wurden. Daher handelt es sich bei diesem um ein Format mit einer strikten äußeren Form, die sich darin widerspiegelt, dass beispielsweise nur bestimmte Spalten der Eingabedatei verwendet werden dürfen bzw. ausgelesen werden. Allerdings akzeptieren heutzutage viele Programme Abweichungen vom originalen strikten MPS-Format, wobei diese programmspezifisch sind, so dass vor der Verwendung eines bestimmten Programmes geprüft werden muss, welche Form die Eingabedatei besitzen darf.

Das Grundprinzip des MPS-Formates besteht darin, dass Problem, welches die zu optimierende Zielfunktion und die Nebenbedingungen beinhaltet, in spaltenweiser Form bzw. in Reihenfolge der Variablen anzugeben und nicht in Zeilenform beispielsweise in Reihenfolge der Nebenbedingungen. Hierzu wird jeder Variablen, der Zielfunktion und jeder Nebenbedingung vom Benutzer ein Name zugeordnet, wodurch das entsprechende Element des (Un-)Gleichungssystems identifiziert wird, was analog zu Matrizenindizes aufgefasst werden kann. Hierbei können nicht nur Zahlen, sondern auch Zeichenkombinationen verwendet werden, wodurch beispielsweise die Lesbarkeit sowohl der Eingabe- als auch Ausgabedaten verbessert werden kann.

Eine MPS-Datei enthält als wichtigste Teile vier Abschnitte: der erste (`ROWS`) umfasst die vom Benutzer angegebene Benennung der Zielfunktion sowie der Nebenbedingungen. Hier wird zusätzlich angegeben, ob eine Nebenbedingung eine Ungleichung oder Gleichung ist. Im zweiten Abschnitt (`COLUMNS`) werden die Elemente der Koeffizientenmatrix $\mathbf{A}$ spaltenweise aufgeführt. Der dritte Abschnitt enthält die rechte Seite des Gleichungssystems (`RHS`) und der vierte die Definitionsbereiche der Variablen (`BOUNDS`) und mögliche Bereichsangaben für die Werte der rechten Seite (`RANGES`). Die Tabelle 4.15 gibt eine Übersicht über diese Abschnitte.

Diese enthält die elementaren Inhalte zur Verwendung des MPS-Formates als Präparationsdatei für ein lineares Optimierungsprogramm. Obwohl es inzwischen viele Entwicklungen gegeben hat, ist dieses Format in der oben beschriebenen Form immer noch weit verbreitet, wenn auch bestimmte Programme Abweichungen oder Erweiterungen unterstützen wie beispielsweise das Ignorieren von Groß- und Kleinschreibung, verlängerte Datenfelder oder Erweiterung auf andere Probleme wie die quadratische Optimierung.
Zur Verdeutlichung dieses Formates soll im folgenden ein kleines Beispielproblem gegeben werden, welches zunächst in der allgemeinen mathematischen Form und darauffolgend als Eingabe-Datei im MPS-Format dargestellt wird.

Beispielproblem

Sei eine zu minimierende Zielfunktion gegeben mit

$$\text{minimiere} \quad 5x_1 - 7x_2 + 0.5x_3 \tag{4.67}$$

mit den Nebenbedingungen

$$x_1 + 0.5x_3 \geq 2 \tag{4.68}$$

$$5x_2 - x_3 \leq -1 \tag{4.69}$$

$$2.1x_1 - 2x_2 = 1 \tag{4.70}$$

mit $x_k \in \mathbb{R}$. als Zwischenschritt zur verbesserten Übersichtlichkeit werden die zu formulierenden Größen in einer tabellarischen Übersicht dargestellt (siehe Tabelle 4.14). Dieses im MPS-Format als Eingabedatei formuliert ist in Abb. 4.29 gezeigt.

	Koeffizienten			Relation	Rechte Seite	Kommentar
Name	x_1	x_2	x_3		b	
ZF	5	-7	0.5			Zielfunktion
NB1	1	0	0.5	$\geq$	2	erste Nebenbedingung
NB2	0	5	-1	$\leq$	-1	zweite Nebenbedingung
NB3	2.1	-2	0	$=$	1	dritte Nebenbedingung

Tabelle 4.14: Tabellarische Darstellung des linearen Beispielproblems.

Datei-Abschnitt	Erklärung
`NAME`	Name bzw. Titel des Problems (ab Spalte 15).

`ROWS`	Enthält die Namen der Zeilen, welche die Zielfunktion und die Nebenbedingungen darstellen. In Spalte 2-3 wird die Art der Gleichung bzw. Ungleichung angegeben und in Spalte 5-12 der Name der Zeile definiert. Als Zeilentypen können N für Zielfunktion, G für $\geq$-Ungleichung, L für $\leq$-Ungleichung und E für eine Gleichung angegeben werden.
	Beispiel:
	`ROWS`
	` N  Kosten`
	` L  Nebenbed`

`COLUMNS`	In diesem Abschnitt werden die Elemente der Koeffizientenmatrix spaltenweise aufgeführt. Die Reihenfolge der Zeilen ist hierbei nicht relevant, lediglich die Reihenfolge der Spalten muss der Reihenfolge der Variablen entsprechen. Matrixelemente, die Null sind, müssen nicht explizit aufgeführt werden. In den Spalten 5-12 der Eingabe-Datei wird der Name der Spalte bzw. der Variablen definiert, in Spalte 15-22 wird der Zeilenname angegeben, die in `ROWS` definiert wurde, und in Spalte 25-36 der Wert des Matrixelementes bzw. des Koeffizienten.
	Beispiel:
	`COLUMNS`
	`    x1        Kosten     3.0`
	`    x1        Nebended  -1.0`

`RHS`	Hier werden die Elemente der Rechte-Seite-Vektoren aufgeführt. Die Zeilenreihenfolge kann beliebig gewählt werden. In Spalte 5-12 der Eingabe-Datei wird der Name des Rechte-Seite-Vektors definiert, in Spalte 15-22 wird der Zeilenname angegeben und in Spalte 25-36 der Wert des Vektorelements.
	Beispiel:
	`RHS`
	`    b1        Nebenbed  -0.333`

`RANGES`	Dieser Abschnitt ist optional und gibt mögliche Intervalle für die Werte der Rechte-Seite-Vektorelemente an. In Spalte 5-12 wird der Name des Bereiches definiert, in Spalte 15-22 wird der Zeilenname angegeben und in Spalte 25-36 der Wert des Bereiches. Die Interpretation eines `RANGES`-Wertes r hängt von der Art der Zeile bzw. Nebenbedingung ab. So wird zum Beispiel in einer $\geq$-Ungleichung (siehe `ROWS`), für die eine rechte Seite mit einem Wert b_i vorgegeben wurde, der Bereich der rechten Seite definiert zu: $b_i + \lvert r \rvert$.
	Beispiel:
	`RANGES`
	`    RNG       Nebenbed  1.2`

`BOUNDS`	In diesem Abschnitt werden die Definitionsbereiche der Variablen $\mathbf{x}$ angegeben. Die Angabe ist optional. Beim Auslassen wird für alle Variablen der Definitionsbereich $x_i \in \mathbb{R}_{\geq 0}$ angenommen. Hier können zum Beispiel die Definitionen `FR` für eine freie Variable ($x_i \in \mathbb{R}$), `UP` für eine Variable, die durch u nach oben beschränkt ist ($x_i \leq u$), oder `LO` für eine durch eine untere Schranke l begrenzte Variable ($x_i \geq l$) verwendet werden.
	Beispiel:
	`BOUNDS`
	` UP BND       x1         19.0`

`ENDATA`	Dieser Eintrag beendet die Eingabe-Datei.

Tabelle 4.15: Generelle Beschreibung der Abschnitte einer MPS-Datei. Die Titelzeilen der einzelnen Abschnitte (`ROWS`, `BOUNDS` etc.) müssen stets in der ersten Spalte der Eingabe-Datei beginnen.

NAME	Test		*Problemname*
ROWS			*Abschnitt Zeilendefinition*
N ZF			*Zielfunktion*
G NB1			*1. Nebenbed. ist eine $\geq$-Ungleichung*
L NB2			*2. Nebenbed. ist eine $\leq$-Ungleichung*
E NB3			*3. Nebenbed. ist eine Gleichung*
COLUMNS			*Abschnitt der Spaltendefinition*
x1	ZF	5.0	
x1	NB1	1.0	
x1	NB2	0.0	#
x1	NB3	2.1	
x2	ZF	-7.0	
x2	NB1	0.0	#
x2	NB2	5.0	
x2	NB3	-2.0	
x3	ZF	0.5	
x3	NB1	0.5	
x3	NB2	-1.0	
x3	NB3	0.0	#
RHS			*Abschnitt der rechten Seite*
b	NB1	2.0	
b	NB2	-1.0	
b	NB3	1.0	
BOUNDS			*Abschnitt der Variablendefinitionsbereiche*
FR BND	x1		*Alle Variablen sind unbeschränkt*
FR BND	x2		
FR BND	x3		
ENDATA			*Ende der Eingabedatei*

Abbildung 4.29: Beispiel einer Eingabedatei im MPS-Format für einen lineares Optimierungsprogramm. Kommentare zur Erklärung sind in kursiver Schriftart dargestellt. Die mit # markierten Zeilen können ausgelassen werden, da sie Null-Elemente der Koeffizienten-Matrix enthalten.

4.5.4 Optimierung der Kraftfeldparameter

Das in dieser Arbeit auftretende Problem der Bestimmung der optimalen Gewichtungskoeffizienten zu den Basisfunktionen lässt sich im Rahmen eines linearen Optimierungsproblems wie folgt formulieren: Als Zielfunktion wird die Minimierung der Summe der Gewichtungskoeffizienten angesetzt, wobei in diesem Fall die Summation über die Absolutbeträge der Koeffizienten erfolgt, damit die Koeffizienten nicht in der Weise optimiert werden, dass sie sich gegenseitig aufheben. Dies führt zur Zielfunktion

$$\text{minimiere} \quad \sum_{i=1}^{n_c} |c_i| \tag{4.71}$$

wobei die c_i die Gewichtungskoeffizienten der n_c Basisfunktionen sind, die in Abschnitt 4.4 beschrieben wurden.
Die Nebenbedingungen zu dieser Zielfunktion, im Sinne der thermodynamischen Hypothese

(siehe hierzu Abschnitt 3.2), wurden so angesetzt, dass die Energien aller falschen Strukturen höher als die der zugehörigen nativen Proteine sein sollte. Die Energie E eines Proteins mit der Geometrie $\mathbf{X}$ und den Basisfunktionen ϕ ergibt sich aus

$$E = \sum_{i=1}^{n_c} c_i \cdot \phi_i(\mathbf{X}) \tag{4.72}$$

Hierauf lässt sich eine Energiedifferenz ΔE zwischen der Energie des nativen Zustandes E^* und der Energie einer falschen Struktur E folgendermaßen schreiben:

$$\Delta E = E - E^* = \sum_{i=1}^{n_c} c_i \cdot \phi_i(\mathbf{X}) - \sum_{i=1}^{n_c} c_i \cdot \phi_i(\mathbf{X}^*) = \sum_{i=1}^{n_c} c_i \Big[\phi_i(\mathbf{X}) - \phi_i(\mathbf{X}^*)\Big] \tag{4.73}$$

Bezogen auf die thermodynamische Hypothese wird somit gefordert, dass

$$\Delta E = \sum_{i=1}^{n_c} c_i \Big[\phi_i(\mathbf{X}) - \phi_i(\mathbf{X}^*)\Big] > 0 \tag{4.74}$$

Erweitert man dies auf mehrere native Proteine p mit jeweils zugehörigen falschen Strukturen d, so lauten die die Nebenbedingungen zur Minimierung der Koeffizientensumme Gl. 4.71:

$$E_{p,d} - E_p^* > 0 \quad \forall\ p \in \{1, 2, \ldots, n_p\} \text{ und } \forall\ d \in \{1, 2, \ldots, n_d(p)\} \tag{4.75}$$

Hier ist n_p die Gesamtanzahl an verwendeten nativen Proteinen und $n_d(p)$ die Anzahl an falschen Strukturen zu jedem nativen Protein p. Die Gesamtzahl an falschen Strukturen und somit auch die Gesamtanzahl an Nebenbedingungen ist gegeben durch

$$n_{d,ges} = \sum_{p=1}^{n_p} n_d(p) \tag{4.76}$$

Die Nebenbedingungen Gl. 4.75 lassen sich als ein Ungleichungssystem $\mathbf{Ac} \geq \mathbf{b}$ formulieren, in welchem die Zeilen der Koeffizientenmatrix $\mathbf{A}$ den Proteinen bzw. der Kombinationen einer nativen mit einer falschen Struktur entsprechen. Die Matrixelemente der Koeffizientenmatrix $\mathbf{A}$ sind somit durch die Differenz der ausgewerteten Basisfunktionen (siehe Gl. 4.73) gegeben

$$\phi_k(\mathbf{X}_{p,d}) - \phi_k(\mathbf{X}_p^*) \tag{4.77}$$

für den k-ten Koeffizienten bzw. die k-te Spalte. Der Vektor $\mathbf{c} = (c_1\, c_2\, \ldots\, c_n)^T$ enthält die Koeffizienten der Basisfunktionen. Der rechte Seite Vektor $\mathbf{b}$ enthält die geforderte minimale energetische Differenz zwischen nativen und nicht-nativen Strukturen, so dass gilt $\mathbf{b} = \mathbf{0}$.
Eine zulässige Lösung des Problems ist dementsprechend nur gegeben, wenn ein Koeffizientenvektor $\mathbf{c}$ existiert, der $\mathbf{Ac} \geq \mathbf{b}$ erfüllt, wodurch alle nativen Strukturen im Vergleich mit den falschen Strukturen eine niedrigere Energie besitzen. Die resultierende Koeffizientenmatrix $\mathbf{A}$ ist in Abb. 4.30 dargestellt.

$$\mathbf{A} = \begin{pmatrix}
\phi_1(\mathbf{X}_{1,1}) - \phi_1(\mathbf{X}_1^*) & \phi_2(\mathbf{X}_{1,1}) - \phi_2(\mathbf{X}_1^*) & \dots & \phi_{n_c}(\mathbf{X}_{1,1}) - \phi_{n_c}(\mathbf{X}_1^*) \\
\phi_1(\mathbf{X}_{1,2}) - \phi_1(\mathbf{X}_1^*) & \phi_2(\mathbf{X}_{1,2}) - \phi_2(\mathbf{X}_1^*) & \dots & \phi_{n_c}(\mathbf{X}_{1,2}) - \phi_{n_c}(\mathbf{X}_1^*) \\
\phi_1(\mathbf{X}_{1,3}) - \phi_1(\mathbf{X}_1^*) & \phi_2(\mathbf{X}_{1,3}) - \phi_2(\mathbf{X}_1^*) & \dots & \phi_{n_c}(\mathbf{X}_{1,3}) - \phi_{n_c}(\mathbf{X}_1^*) \\
\dots & \dots & & \dots \\
\phi_1(\mathbf{X}_{1,n_d(1)}) - \phi_1(\mathbf{X}_1^*) & \phi_2(\mathbf{X}_{1,n_d(1)}) - \phi_2(\mathbf{X}_1^*) & \dots & \phi_{n_c}(\mathbf{X}_{1,n_d(1)}) - \phi_{n_c}(\mathbf{X}_1^*) \\
\phi_1(\mathbf{X}_{2,1}) - \phi_1(\mathbf{X}_2^*) & \phi_2(\mathbf{X}_{2,1}) - \phi_2(\mathbf{X}_2^*) & \dots & \phi_{n_c}(\mathbf{X}_{2,1}) - \phi_{n_c}(\mathbf{X}_2^*) \\
\phi_1(\mathbf{X}_{2,2}) - \phi_1(\mathbf{X}_2^*) & \phi_2(\mathbf{X}_{2,2}) - \phi_2(\mathbf{X}_2^*) & \dots & \phi_{n_c}(\mathbf{X}_{2,2}) - \phi_{n_c}(\mathbf{X}_2^*) \\
\dots & \dots & & \dots \\
\phi_1(\mathbf{X}_{n_p,n_d(n_p)-1}) - \phi_1(\mathbf{X}_{n_p}^*) & \phi_2(\mathbf{X}_{n_p,n_d(n_p)-1}) - \phi_2(\mathbf{X}_{n_p}^*) & \dots & \phi_{n_c}(\mathbf{X}_{n_p,n_d(n_p)-1}) - \phi_{n_c}(\mathbf{X}_{n_p}^*) \\
\phi_1(\mathbf{X}_{n_p,n_d(n_p)}) - \phi_1(\mathbf{X}_{n_p}^*) & \phi_2(\mathbf{X}_{n_p,n_d(n_p)}) - \phi_2(\mathbf{X}_{n_p}^*) & \dots & \phi_{n_c}(\mathbf{X}_{n_p,n_d(n_p)}) - \phi_{n_c}(\mathbf{X}_{n_p}^*)
\end{pmatrix}$$

Abbildung 4.30: Darstellung der Koeffizientenmatrix $\mathbf{A}$, welche die Nebenbedingungen des linearen Optimierungsproblems beinhaltet (siehe Gl. 4.74 und 4.75). $p \in \{1, \dots, n_p\}$ ist die Nummer des Proteins, $d \in \{1, \dots, n_d(p)\}$ die Nummer der falschen Struktur und n_c die Anzahl der Koeffizienten.

Als Hinweis zur Nummerierung der Koeffizienten c_i sei an dieser Stelle vermerkt, dass die Koeffizienten wie in Abschnitt 4.4 beschrieben, aminosäurepaar- und sequenzpositionsabhängig sein können. Zur Verdeutlichung der wichtigen Punkte an dieser Stelle wurde die Darstellung aber vereinfacht. Der Index i kann an dieser Stelle als Ordnungsnummer des Koeffizienten in der Gesamtmenge aller Koeffizienten aufgefasst werden. Ein Koeffizient $c^{(NB)}_{k,s_n,s_m}$ beispielsweise einer nicht-bindenden Basisfunktion $\phi^{(NB)}$ ist abhängig von der Differenz k der Indizes der beteiligten Aminosäuren im Sequenzvektor und von der Art der Aminosäuren s_n und s_m, so dass der Summationsindex i einer bestimmten Kombination dieser drei Zahlen entspricht.

Die Forderung bzw. Voraussetzung einer Lösung des linearen Problems, dass alle Nebenbedingungen gleichzeitig erfüllt sein müssen, bietet gegenüber der Optimierung über den Z-Wert Vorteile, da es bei dieser Methode nicht gewährleistet ist, dass alle nativen Proteine einen niedrigeren Funktionswert gegenüber den falschen Strukturen besitzen. Zudem kann in einer linearen Optimierung, falls für das Problem keine Lösung existieren sollte, direkt bestimmt werden, welche Nebenbedingungen nicht erfüllt werden konnten, wodurch Informationen für eine Verbesserung der Potentialfunktion gewonnen werden können.

Je mehr Nebenbedingungen verwendet werden können, um so mehr Informationen enthalten die resultierenden Parameter über den Konformationsraum der Proteine bzw. über die Wechselwirkungen. Deren Anzahl ist zum einen abhängig von der Programmarchitektur des Lösungsalgorithmus und zum anderen von den zur Verfügung stehenden Computerressourcen. Hierbei ist der am stärksten limitierende Faktor der Hauptarbeitsspeicher der Rechners. Die Lösungsalgorithmen, die für diese Arbeit zur Auswahl standen, sowohl kommerzielle wie auch nicht-kommerzielle Produkte, behandeln das Ungleichungssystem nicht parallel, sondern laden es zur Lösung insgesamt in den Hauptspeicher. Bestimmte Programmteile können

parallelisiert sein, doch diese umfassen beispielsweise nur die Präparation des Ungleichungssystems wie das Einlesen der Daten. Die Lösung des linearen Problems erfolgt seriell.
Zur Bestimmung wie viele Nebenbedingungen bzw. falsche Strukturen $n_{d,ges}$ insgesamt zur Lösung des Ungleichungssystems verwendet werden konnten, wurde aufgrund der Komplexität des Lösungsprogrammes BPMPD empirisch der Hauptspeicherbedarf M in Gigabyte (GB) in Abhängigkeit von der Anzahl an Nebenbedingungen $n_{d,ges}$ und von der Anzahl an Spalten des Gleichungssystems n_c bestimmt, die durch die Anzahl an Koeffizienten der Basisfunktionen gegeben sind. Dies führte zur folgender Näherung:

$$n_{d,ges} \approx \frac{M}{4.28 \cdot 10^{-8}\mathrm{GB} \cdot n_c} \tag{4.78}$$

Aufgrund der programmtechnischen Realisation des verwendeten Lösungsalgorithmus ist der verfügbare Hauptspeicher auf $M = 20$ GB beschränkt. Daher bleibt in Gl. 4.78 lediglich die Anzahl an Koeffizienten n_c als freie Variable übrig. Diese Anzahl hängt zum einen von der Anzahl an gewählten Basisfunktionen, die für jede Art der Wechselwirkung unterschiedlich sein kann, und zum anderen von der Anzahl an Aminosäureklassen ab, da der Großteil der Koeffizienten spezifisch für die Wechselwirkung bestimmter Aminosäurepaare sind. Im folgenden werden für die einzelnen Basisfunktionsklassen die Anzahl an notwendigen Koeffizienten dargestellt und daraus die Skalierung und die Größe und des zu lösenden linearen Problems erläutert.
Da die Nahwechselwirkungsterme von der Reihenfolge der Aminosäuren in der Sequenz abhängen, werden die Aminosäurepaare (s_a, s_b) und (s_b, s_a) unterschiedlich behandelt, sofern $s_a \neq s_b$. Durch diese Unterscheidung skaliert die Anzahl der Koeffizienten der Nahwechselwirkungsterme quadratisch mit der Anzahl an Aminosäureklassen zu einer vorgegebenen Anzahl an Basisfunktionen. Sei $n_\phi^{(B)}$ die Anzahl an Basisfunktionen der Nahwechselwirkungsterme und $y^{(B)}$ die Anzahl an Aminosäureklassen zu diesen, so beträgt die Anzahl an Koeffizienten $\left(y^{(B)}\right)^2 \cdot n_\phi^{(B)}$. Da diese sequenzpositionsabhängige Unterscheidung bei den nicht-bindenden Wechselwirkungen sowohl für die C^α-Atome als auch für die Seitenketten nicht vorgenommen wird, skalieren deren Koeffizientenanzahlen dagegen mit $0.5\ y^{(NB)}\left(y^{(NB)}+1\right) \cdot n_\phi^{(NB)}$.
Die Oberflächenfunktionen wirken nur auf die als hydrophob klassifizierten Aminosäuren. Für diese Funktionen werden die Aminosäuren nicht weiter in Klassen eingeteilt, sondern die sieben hydrophoben Aminosäuren direkt verwendet. Für jede Aminosäure werden vier Basisfunktionen angesetzt. Insgesamt ergibt sich so eine lineare Skalierung $y^{(SU)} \cdot n_\phi^{(SU)}$.
Die Wasserstoffbrücken-Wechselwirkung wurde mit Funktionen genähert, deren Koeffizienten nicht von der Art der beteiligten Aminosäuren abhingen, da angenommen wurde, dass alle Aminosäuren eine ähnliche Tendenz zur Bildung von Wasserstoffbrücken besitzen. Daher skaliert hierfür die Koeffizientenanzahl nur mit der Anzahl an Basisfunktionen $n_\phi^{(HB)}$. Diese Skalierungen und die zur Parameteroptimierung verwendete Anzahl an Funktionen und Aminosäureklassen sind in Tab. 4.16 zusammengefasst.

Funktionsklasse	n_ϕ	y_ϕ	Skalierung	K_ϕ
Nahwechselwirkung	8	5	$(y_\phi^{(B)})^2 \cdot n_\phi^{(B)}$	200
Nicht-bindende Wechselwirkung (C^α)	6	8	$0.5\, y_\phi^{(NB)} \left(y_\phi^{(NB)} + 1\right) \cdot n_\phi^{(NB)}$	216
Nicht-bindende Wechselwirkung (Seitenketten)	4	8	$0.5\, y_\phi^{(SC)} \left(y_\phi^{(SC)} + 1\right) \cdot n_\phi^{(SC)}$	144
Oberflächenpotential	4	7	$y_\phi^{(SU)} \cdot n_\phi^{(SU)}$	28
Wasserstoffbrücken	3	-	$n_\phi^{(HB)}$	3
Gesamt (n_c)				591

Tabelle 4.16: Übersicht über die Funktionsklassen, Anzahl an verwendeten Basisfunktionen n_ϕ, an Aminosäureklassen y_ϕ, Skalierung der Koeffizientenanzahl und die Anzahl an resultierenden Koeffizienten K_ϕ.

Bei der Festlegung der Anzahl an Funktionen und Aminosäureklassen müssen mehrere Punkte beachtet werden: Zu einer akkuraten Beschreibung des Potentials sind ein Minimum an Funktionen nötig, wobei es vom Design des Potentials abhängt, wieviele Funktionen tatsächlich notwendig sind. Die benötigte Minimalanzahl ist *a priori* nicht bekannt und muss getestet werden. Als Folge zu weniger und/oder unpassend gewählter Funktionen können beispielsweise wichtige Elemente der realen Energiefläche nicht oder falsch modelliert werden. Eine zu große Anzahl an Funktionen dagegen kann ebenso kontraproduktiv wirken und die Qualität des Potentials beeinträchtigen, da es zu einer redundanten Beschreibung von Wechselwirkungen kommen kann, wodurch eine falsche Balance der einzelnen Energiebeiträge resultieren kann. Diese Maximalanzahl an Funktionen ist ebenfalls abhängig von der Formulierung des Potentialansatzes und nicht bekannt.
Zusätzlich hängen aufgrund der gewählten Methode zur Bestimmung der Koeffizienten die Anzahl an Potentialfunktionen und die Anzahl an falschen Strukturen zum Vergleich mit den nativen Strukturen direkt zusammen, da sie zusammen die Größe des zu lösenden Problems bestimmen, so dass ein vergrößerter Funktionensatz einen kleineren Satz an Proteinstrukturen während der Optimierung zur Folge hat, da die zur Verfügung stehen Computer-Ressourcen begrenzt sind.

Neben diesen Faktoren ist weiterhin zu beachten, dass in dem Kraftfeldansatz sehr viele weitere Parameter einfließen, die die Form der resultierenden Fläche mitbestimmen. Als Beispiele seien hierfür aufgeführt:

- Auswahl der nativen und falschen Proteine, welche die Informationen über den Konformationsraum, Faltungspräferenzen, Kontakte, Abstände usw. enthalten.
- Wahl der Konstanten in den Basisfunktionen, die z. B. die Positionen der Minima in den Potentialen festlegen.

- Generell die Wahl der Basisfunktion bzw. deren Funktionsform, die für die modellierte Wechselwirkung passend oder im ungünstigen Fall auch unpassend sein kann.
- Definition des Proteinmodells, wodurch die Anzahl und die Position der Interaktionszentren bestimmt wird.

Dies führt zu einem sehr großen Raum an möglichen Variablen, die nicht alle gleichzeitig optimiert werden können. Aufgrund dieser Komplexität wurde in dieser Arbeit der in den vorangegangen Abschnitten beschriebene Kraftfeldansatz als ein Anfangspotential angesetzt und in einer iterativen Optimierungsprozedur getestet, in wie weit sich dieses Potential zur Strukturvorhersage eignet.
Dieses iterative Vorgehen bestand aus einer schrittweisen Anpassung der Gewichtungskoeffizienten **c**, während alle anderen Parameter konstant gehalten wurden. Hierbei wurde folgendermaßen vorgegangen:

1. Definition des Proteinmodells und des Potentials sowie Festlegung der nativen und falschen Strukturen bzw. deren Erzeugung. Berechnung der Basisfunktionen anhand der gewählten Proteinstrukturen, woraus sich die Nebenbedingungen Gl. 4.75 ergeben.
2. Bestimmung der Koeffizienten **c** als lineares Optimierungsproblem mit der Zielfunktion Gl. 4.71 und den Nebenbedingungen Gl. 4.75 unter Verwendung des BPMPD-Programmes.
3. Globale Geometrieoptimierung der in Punkt 2 eingesetzten nativen Sequenzen unter Verwendung eines genetischen Algorithmus, in welchem der Potentialansatz, der die optimierten Koeffizienten enthält, implementiert wurde. Hierdurch optimierte nicht-native Strukturen, die eine niedrigere Energie als der native Zustand besitzen, werden der Menge der falschen Strukturen hinzugefügt.
4. Beende Programm, wenn keine energetisch niedrigeren Strukturen mittels des genetischen Algorithmus gefunden wurden. Ansonsten gehe zu Punkt 5.
5. Sortierung und Auswahl aller falschen Strukturen anhand ihrer Energie für eine erneute Parameteroptimierung, in der die aus dem genetischen Algorithmus erhaltenen Strukturen mitberücksichtigt werden. Die energetisch niedrigsten Strukturen sind in der Parameteroptimierung wichtiger als diejenigen, deren Energie sehr viel größer als die des nativen Zustandes ist, da diese leicht als nicht-native Strukturen erkannt werden können (im Sinne einer hohen potentiellen Energie).
6. Gehe zu Punkt 2.

Zusammengefasst besteht der Iterationsprozess darin, die aus der globalen Optimierung erhaltenen falschen Strukturen für eine neue Parameteroptimierung zu verwenden, solange bis eine Konvergenz erreicht ist, die dadurch definiert ist, dass im globalen Optimierungsprozess keine oder sehr wenige Strukturen erzeugt werden, deren Energie niedriger als die des nativen Zustandes ist, wobei dies davon abhängig sein kann, ab welcher geometrischen Distanz eine Struktur als nativ angesehen wird. Daher können trotzdem noch Proteine erzeugt werden, deren Energie niedriger ist als die der Datenbankstruktur, sofern sie dieser ähnlich genug sind. Im wesentlichen soll so vermieden werden, dass die Energiehyperfläche ein tiefliegendes energetisches Minimum fernab von der nativen Struktur besitzt.

Während dieser Iterationen wurden alle Parameter des Kraftfeldes konstant gehalten und nur die Auswahl an falschen Strukturen angepasst. Da wie oben erwähnt, das Programm auf einen Hauptspeicher von 20 GB beschränkt war und insgesamt 591 Koeffizienten verwendet wurden, konnten nach Gl. 4.78 mehr als 700000 falsche Strukturen während eines Parameteroptimierungsschrittes gleichzeitig behandelt werden, was im Vergleich zu vielen anderen publizierten Kraftfeldern, die ebenfalls gegen falsche Strukturen optimiert wurden, eine große Anzahl darstellt und häufig über deren Größe hinausgeht. Tatsächlich wurden exakt 705739 falsche Strukturen behandelt, was eine kleinere Anzahl darstellt, als die nach Gl. 4.78 möglich wäre. Dies ist darin begründet, dass die Grenze von 20 GB für die Programmausführung um 5 % reduziert wurde, um Ungenauigkeiten bei der empirischen Bestimmung von Gl. 4.78 auszugleichen, da im Falle eines zu groß gewählten Ungleichungssystems der Lösungsalgorithmus abgebrochen wurde. Da im gesamten Prozess jeweils die Vorbereitung der Daten in Form der Erstellung der MPS-Datei der zeitaufwändigste Schritt war, hätte bei einem Programmabbruch aufgrund eines zu großen vorgegebenen Problems die vollständige Input-Datei neu geschrieben werden müssen, da das MPS-Format das (Un-)Gleichungssystem spaltenweise enthält. Während die Lösung des linearen Problems auf einem Großrechner der Universität Kiel stattfand, wurde die notwendige Eingabedatei auf einem lokalen Arbeitsplatzrechner erstellt, dessen Hauptarbeitsspeicher 8 GB betrug. Da dieser zu klein war, um das gesamte Gleichungssystem dort zu speichern, musste es auf die Festplatte in Teilen ausgelagert werden und in einem abschließenden Schritt zu einer Gesamtdatei zusammengefügt werden. Aufgrund dieser Tatsache, war dieser Prozess zeitlich langwierig. Die Größe der Eingabedatei im MPS-Format betrug 14 GB.

4.6 Ergebnisse der Parameteroptimierung

4.6.1 Koeffizienten und Energien

Wie im vorherigen Abschnitt beschrieben, wurde die Parameteroptimierung mit insgesamt 48 nativen Proteinen (siehe Tab. 4.5) und zugehörigen 705739 falschen Strukturen mit dem Programm BPMPD durchgeführt. Da aufgrund der selbst durchgeführten Produktion der falschen Strukturen von diesen eine größere Anzahl zur Verfügung stand als aufgrund der Hauptarbeitsspeicherbeschränkung des Programmes verwendet werden konnte, wurden für die erste Parameteroptimierung die falschen Strukturen vollkommen zufällig aus der Gesamtmenge aller falschen Strukturen ausgewählt. Dies waren die verzerrten nativen Strukturen, die Strukturen aus der Sequenzübertragung und die falschen Strukturen des LKF2-Satzes (siehe Abschnitt 4.3.2). Die Konvergenzkriterien des BPMPD-Programmes wurden unverändert verwendet, so dass eine Konvergenz bzw. ein Optimalpunkt bei Erfüllung der Bedingungen Gl. 4.66 gegeben war.
Das Optimierungsprogramm zur Lösung des linearen Problems wurde auf dem NEC-SX8-Großrechner der Universität Kiel ausgeführt. Für die Optimierung der Parameter wurden in jeder Iteration ca. 20 bis 30 Innere-Punkte-Schritte benötigt, bis ein optimaler Punkt gefunden wurde. Alle im iterativen Prozess erstellten lineare Probleme (siehe Seite 136) erwiesen sich als lösbar, woraus zunächst gefolgert werden konnte, dass die angesetzten Potentialfunktionen ausreichend waren, die gewählten falschen Strukturen von den nativen zu unterscheiden. Insgesamt wurden vier Iterationen durchgeführt, wobei eine Iteration aus einer Parameteroptimierung, der globalen Geometrieoptimierung und dem Hinzufügen der neuen falschen Strukturen aus der Geometrieoptimierung zur Gesamtmenge aller falscher Strukturen bestand.
Die effektive Anzahl der neuen falschen Strukturen aus den GA-Optimierungsläufen ließ sich generell vor Beendigung der Programmausführung nicht vorhersagen, da die Optimierung mittels genetischen Algorithmen ein stochastischer Prozess ist. Ebenfalls konnte die reale Laufzeit des Programms nicht vorhergesagt werden, da diese vom detaillierten Verlauf des Algorithmus abhing. In dem in dieser Arbeit verwendeten Programm ist beispielsweise die lokale Kraftfeldoptimierung der zeitaufwändigste Schritt, so dass die Gesamtlaufzeit stark von der Anzahl an auszuführenden lokalen Optimierung mittels des Kraftfeldes und damit auch von der benötigten Anzahl an Schritten innerhalb der lokalen Optimierung bis zum Erreichen des (numerischen) lokalen Minimums. Dies hängt davon ab, wie dicht die Anfangsstruktur an einem lokalen Minimum liegt. Aus diesen Gründen und aufgrund der Tatsache, dass nicht alle nativen Proteine aufgrund der begrenzten computertechnischen Ressourcen gleichzeitig bearbeitet werden konnten, wurde der Gesamtsatz an nativen Proteinen in kleinere Unter-

gruppen unterteilt, die jeweils Proteine mit unterschiedlichen Größen enthielten, wobei aber die Größenverteilung der Proteine nach Möglichkeit so gewählt wurde, dass alle Gruppen eine gleichmäßige Verteilung von kleinen zu großen Proteinen hin beinhalteten, um zu vermeiden, dass beispielsweise eine Gruppe lediglich kleine Proteine enthielt, während eine andere rein aus großen Proteinen bestand.
Der genetische Algorithmus wurde mit diesen Proteingruppen so durchlaufen, dass jeweils eine Gruppen auf einem Prozessor seriell bearbeitet wurde. Wenn eine Gruppe beendet war, indem alle der Gruppe zugehörigen Proteine den GA durchlaufen hatten, wurde die Gruppe neu gestartet. Es wurde zur Generierung der neuen falschen Strukturen somit keine bestimmte Anzahl an Läufen pro Gruppe festgelegt, sondern es wurde eine Gesamtzeit angesetzt, in welcher die Gruppen gestartet wurden. Die Häufigkeit, mit der ein Protein den GA durchlief, hing somit von der Gesamtrechenzeit der Gruppe ab, der es zugehörte. Hiermit sollte erreicht werden, dass die größeren Proteine, deren globale Optimierung bei gleichen voreingestellten Parametern für den genetischen Algorithmus, wie beispielsweise Anzahl an Generationen und Individuen, wesentlich mehr Realzeit zur Berechnung benötigt, eine mit den kleineren Proteinen vergleichbare Anzahl an GA-Läufen zugeordnet bekommen, um eine Unausgewogenheit des resultierenden Strukturdatensatzes zu vermeiden. Dennoch konnte die Unausgewogenheit nicht vermieden werden, da unter der Voraussetzung, dass für alle Proteinen die gleichen Parametern des GA gelten, bei langen Sequenzen (ca. 200 oder mehr Aminosäuren) sehr wenige bis gar keine neuen falschen Strukturen gefunden wurden, was vermutlich darauf zurückzuführen ist, dass bei diesen Proteingrößen der GA mit den voreingestellten Parametern den Konformationsraum der Proteine nur unzureichend abgesucht hat. Ein Indiz hierfür war, dass die finalen niedrigsten Energien in den entsprechenden Läufen häufig wesentlich größer als die der nativen Strukturen waren.

Die Anzahl an hinzukommenden falschen Strukturen aus der globalen Optimierung war in jeder Parameterbestimmungsiteration und für jedes Protein aus den oben genannten Gründen unterschiedlich. So kamen in den drei Iterationen 1389, 10891 und schließlich 2729 neue Strukturen hinzu. Die größere Anzahl von 10891 im Vergleich zu den anderen beiden Iterationen ist darauf zurückzuführen, dass bei dieser Iteration dem GA mehr Gesamtrechenzeit zugeordnet wurde.

Im folgenden werden die Ergebnisse der Parameteroptimierung beschrieben, indem auf die resultierenden Parameter und deren Einfluss auf die Gesamtenergie der nativen Proteine eingegangen wird. Hierzu werden exemplarisch vier Proteine dargestellt, die aufgrund der unterschiedlichen Zusammensetzung der Sekundärstruktur ausgewählt wurden und repräsentativ für die Gesamtheit stehen. Hierbei handelt es sich um die Proteine 1a92(A), das größtenteils eine α-helicale Struktur besitzt, 1ifc mit einem großen Anteil an β-Faltblatt-Strukturen

ohne Helices, 1dhn, das annähernd gleiche Anteile an Helices und Faltblättern enthält und schließlich 1chd, welches ebenfalls ein gemischtes α/β-Protein ist, das aber zusätzlich einen erhöhten Anteil an Windungen enthält (siehe Tab. 4.5).

Die nach der Parameteroptimierung erhaltenen Energien E^* für die nativen Proteine, die sich nach

$$E^* = \sum_{i=1}^{n_c} c_i \cdot \phi_i(\mathbf{X}^*) \tag{4.79}$$

berechnen, sind für die vier oben erwähnten Proteine in Abb. 4.31 dargestellt. In dieser sind die Gesamtenergien der unterschiedlichen Funktionsklassen enthalten. Die Energien sind zum direkten Vergleich der Beiträge relativ zur Energie des Nahwechselwirkungsterms dargestellt, welcher für die Darstellung auf den Wert -1 normiert wurde. Die Energien werden in Abhängigkeit von der Iteration I_n der Parameteroptimierung dargestellt, wobei für die Berechnung der Energien zu I_0 die optimierten Parameter verwendet wurden, die auf Grundlage der drei anfänglich verwendeten Sätze falscher Strukturen bestimmt wurden, ohne Ergänzung durch falsche Strukturen aus der globalen Geometrieoptimierung. Die weiteren Iteration $I_{>0}$ verwenden Parameter, die jeweils nach Hinzufügung der GA-optimierten Strukturen bestimmt wurden.
Zusätzlich ist in Abb. 4.31 der Mittelwert $W^{(k)}$ der Energie für jede Funktionsklasse k gezeigt, der über alle nativen Proteine gemittelt wurde

$$W^{(k)} = \frac{1}{n_{\text{nat}}} \sum_{p=1}^{n_{\text{nat}}} E_p^{*(k)} \tag{4.80}$$

wobei n_{nat} die Anzahl an nativen Proteinen ($n_{\text{nat}} = 48$), $E_p^{*(k)}$ die Energie der Wechselwirkungsklasse k mit $k \in \{1, \ldots, 5\}$ für das p-te native Protein ist. k ist hierbei identisch mit den Klassen "Nahwechselwirkung", "Seitenkettenpotential" usw.

Es zeigte sich, dass alle Energiebeiträge mit Ausnahme des Oberflächenpotentials stets einen stabilisierenden (negativen) Beitrag zur Gesamtenergie lieferten, während das Oberflächenpotential destabilisierend (positiv) eingeht. Dieses Verhalten beruht beim Oberflächenpotential auf dessen mathematischer Definition, welche nur positive Funktionswerte zulässt (siehe Abschnitt 4.4.5), welches mit dem Ziel so implementiert wurde, den hydrophoben Effekt zu simulieren und die Faltung zu kompakten Strukturen zu treiben, indem diese durch einen niedrigeren Funktionswert gegenüber den weniger kompakten Geometrien bevorzugt werden. Der nach jeder Parameteriteration erhaltene destabilisierende Anteil dieser Funktion wird aber stets durch die anderen Energiebeiträge überkompensiert, so dass die Gesamtenergie der nativen Proteine negativ ist. Dies ist aber keine Voraussetzung für ein erfolgreiches Potential zur Strukturvorhersage, da der Nullpunkt der Energieskala durch Hinzufügung einer Konstanten beliebig gewählt werden kann.

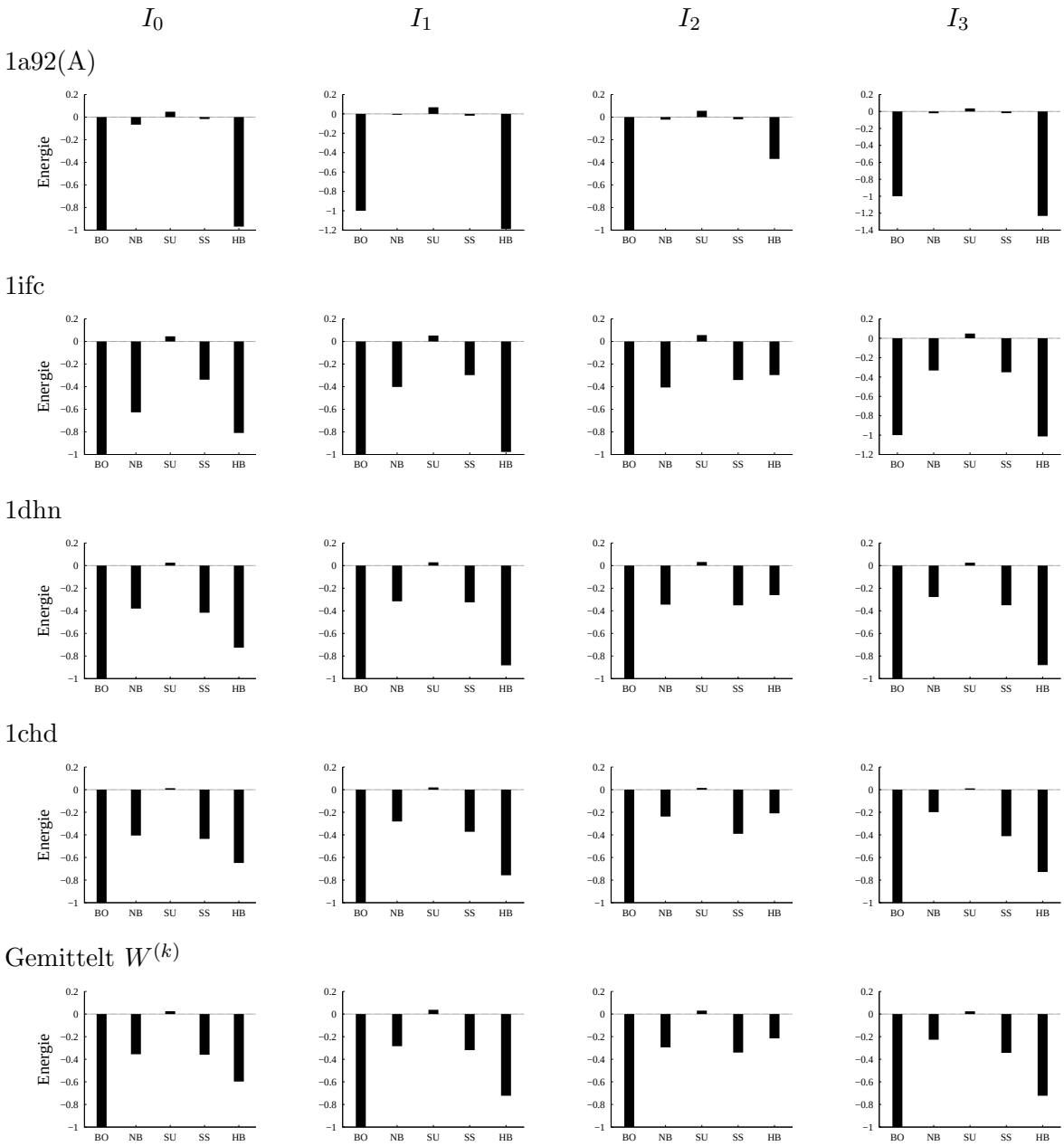


Abbildung 4.31: Gesamtenergiebeiträge der definierten Wechselwirkungsklassen relativ bezogen auf die auf -1 normierte Nahwechselwirkungsenergie. Über den Grafiken ist die jeweilige Parameteroptimierungsiteration I_n angegeben. Auf der Abszisse sind die Funktionstypen mit folgenden Bedeutungen aufgetragen: BO = bindend bzw. Nahwechselwirkung, NB = nicht-bindend, SU = Oberfläche, SS = Seitenketten und HB = Wasserstoffbrücken. Die unterste Grafik-Zeile enthält die über alle Proteine gemittelten Energiewerte (siehe Gl. 4.80).

Bei den Funktionen, die zum Wasserstoffbrückenpotential beitragen, ist die Situation ähnlich dem Oberflächenpotential, da diese per Definition nur rein negative Funktionswerte liefern können (siehe Abschnitt 4.4.6).
Da in den nativen Strukturen die Abstände der C^{α}-Atome und die inneren Produkte der Bindungsvektoren, welche als Variablen in die Nahwechselwirkungsterme eingehen, im Bereich ihrer "natürlichen" Werte sind, die über ihre statistische Verteilung zur Definition der Minima der Potentiale verwendet wurden (siehe Abschnitt 4.4.2), ist die Summe der Nahwechselwirkungsterme ebenfalls negativ. Teilt man die Nahwechselwirkungsfunktionen in zwei Gruppen auf, wobei die eine Gruppe nur Funktionen enthält, die vom Abstand zwischen den C^{α}-Atomen abhängen (siehe Gl. 4.27 bis 4.30), während die andere Gruppe nur die Funktionen enthält, die vom inneren Produkt der Bindungsvektoren abhängen (siehe Gl. 4.31 bis 4.34), und anschließend die (absoluten) Energiebeiträge vergleicht, so zeigt sich dass in allen Iterationen der Beitrag der Abstandsfunktionen zur Gesamtenergie dieser Klasse einen Anteil von ca. 70 % beträgt. Während der einzelnen Iterationen nahm dieser Anteil weiter zu, so dass dieser von anfänglichen 69 % in I_0 zu 72 % in I_3 anstieg. Diese Werte sind auf die gemittelten Energien bezogen, die nach Gl. 4.80 erhalten wurden. Betrachtet man die über alle Proteine gemittelten einzelnen Energiebeiträge der Funktionen $\phi_1^{(B)}$ bis $\phi_8^{(B)}$ separat, welche analog zu Gl. 4.80 hier für einzelne Klassen k bestimmt wurden, so zeigt sich, dass sich $\phi_2^{(B)}$ über die Iterationen hinweg zur dominierenden Funktion entwickelt, während die Funktionen $\phi_1^{(B)}$, $\phi_4^{(B)}$, $\phi_5^{(B)}$ und $\phi_8^{(B)}$ eher untergeordnete Rollen einnehmen, da deren Beiträge zwei Größenordnungen kleiner sind als der Beitrag durch $\phi_2^{(B)}$ (siehe Tab. 4.19). Die dominierende Funktion $\phi_2^{(B)}$ beschreibt den Abstand zweier direkt benachbarter C^{α}-Atome und enthält ein Minimum an der Position einer *trans*-C_i^{α}-C_{i+1}^{α}-Bindung. Weiterhin ist auffällig, dass die Energiewerte der Funktion $\phi_4^{(B)}$ als einzige einen positiven Mittelwert aufweisen. Die Einzelenergien, die zum Mittelwert von $\phi_4^{(B)}$ in Tab. 4.19 führen, zeigen ein uneinheitliches Bild von positiven und negativen Funktionswerten. Die positiven Werte beruhen auf der schon in Abschnitt 4.4.2 angesprochenen Schwierigkeit, die Häufigkeitsverteilung der $(i, i+2)$-Abstandsverteilung (siehe Abb. 4.19) für Abstände > 6 Å entsprechend zu modellieren, da die Häufigkeitsverteilung hier breiter und flacher verläuft. Die positiven Funktionswerte entstehen vornehmlich durch C_i^{α}-C_{i+2}^{α}-Abstände, die in gestreckten Konformationen auftreten, im Bereich größer als 7 Å, in welchem die Potentialfunktion repulsiv wird (vergleiche Abb. 4.22 b). Dieses könnte dadurch korrigiert werden, dass entweder direkt eine entsprechende andere Potentialfunktion gewählt wird, dessen Minimum ein entsprechend breiteres Einzugsgebiet besitzt, oder dass ähnlich zu den nicht-bindenden Funktionen eine Linearkombination verschiedener Basisfunktionen mit dem Ziel angesetzt wird, durch die Parameteroptimierung eine Linearkombination zu erhalten, die die notwendige funktionale Form besitzt. Dieser zweite Ansatz wurde im Laufe der Kraftfeldentwicklung getestet, indem lediglich diese Funktion als Linearkombination verschiedener anderer Basisfunktionen angesetzt wurde, während alle anderen Funktionen

I	$\phi_1^{(B)}$	$\phi_2^{(B)}$	$\phi_3^{(B)}$	$\phi_4^{(B)}$	$\phi_5^{(B)}$	$\phi_6^{(B)}$	$\phi_7^{(B)}$	$\phi_8^{(B)}$
0	$-1.9\cdot10^{-5}$	$-7.6\cdot10^{-4}$	$-1.0\cdot10^{-4}$	$3.5\cdot10^{-5}$	$-5.5\cdot10^{-5}$	$-1.3\cdot10^{-4}$	$-1.7\cdot10^{-4}$	$-3.7\cdot10^{-5}$
1	$-3.0\cdot10^{-5}$	$-8.1\cdot10^{-4}$	$-1.5\cdot10^{-4}$	$3.8\cdot10^{-5}$	$-3.5\cdot10^{-5}$	$-1.6\cdot10^{-4}$	$-2.1\cdot10^{-4}$	$-4.8\cdot10^{-5}$
2	$-3.4\cdot10^{-5}$	$-1.7\cdot10^{-3}$	$-2.4\cdot10^{-4}$	$6.1\cdot10^{-5}$	$-5.7\cdot10^{-5}$	$-2.3\cdot10^{-4}$	$-4.0\cdot10^{-4}$	$-8.7\cdot10^{-5}$
3	$-2.7\cdot10^{-5}$	$-1.5\cdot10^{-3}$	$-2.1\cdot10^{-4}$	$6.3\cdot10^{-5}$	$-4.6\cdot10^{-5}$	$-2.1\cdot10^{-4}$	$-3.6\cdot10^{-4}$	$-6.9\cdot10^{-5}$

Tabelle 4.19: Energiebeiträge der Nahwechselwirkungsfunktionen $\phi_j^{(B)}$ gemittelt über alle Proteine für die jeweilige Iterationen I. $\phi_1^{(B)}$ bis $\phi_4^{(B)}$ sind die abstandsabhängigen Funktionen, während $\phi_5^{(B)}$ bis $\phi_8^{(B)}$ von den inneren Produkten der Bindungsvektoren abhängen.

und Konstanten unverändert blieben. Hier wurden Funktionen verwendet, die eine ähnliche Form wie die nicht-bindenden Basisfunktionen besaßen (siehe Gl. 4.36), mit dem Unterschied jedoch, dass diese für große Werte für r_{ij} repulsiv werden und für den Trennpunkt (siehe Abb. 4.22) zwischen den zusammengehörenden Nahwechselwirkungstermen gegen Null streben. Die mit diesen Funktionen resultierenden linearen Probleme waren stets lösbar. Die Linearkombinationen zeigten jedoch keine Minimumstruktur, sondern zumeist nur ein Verhalten ähnlich einer nach unten geöffneten Halbparabel, so dass für die $r_{ij} \to 0$ die Funktionswerte gegen $-\infty$ strebten, was prinzipiell zur reinen Erkennung von nativen Strukturen in einem größeren Satz falscher Strukturen weniger problematisch ist, während es dagegen aber in einer lokalen Optimierung mit Gradienteninformationen, wie sie im genetischen Algorithmus angewendet wurde, zur Folge hat, dass die C^α-Atome dissoziieren. Da dieses Problem auch nicht mit einer Beschränkung bestimmter Funktionskoeffizienten während der linearen Optimierung auf positive Werte behoben werden konnte, wurde die ursprüngliche Funktion verwendet.

Die nicht-bindenden Wechselwirkungen zentriert auf den C^α-Atomen zusammen mit den nicht-bindenden Wechselwirkungen der Seitenketten tragen zu der gemittelten Energie (Abb. 4.31) in allen Iterationen ungefähr den gleichen Anteil an stabilisierender Energie bei, der in Summe in der Größenordnung von ca. 2/3 der Energie der Nahwechselwirkungsterme liegt. Diese Wechselwirkungen wurden in der Definition des Kraftfeldes als Linearkombination mehrerer Basisfunktionen angesetzt, um die Form des effektiven nicht-bindenden Potentials ohne Voraussetzungen optimieren zu können (siehe Abschnitt 4.4.3). Hierzu wurden, um bestimmte Randbedingungen zu erfüllen, modifizierte Funktionen des Grundtyps $1/r_{ij}^k$ angesetzt und die Gewichtungskoeffizienten dieser Funktionen für unterschiedliche Werte des Exponenten k bestimmt. Wie unter anderem aus Tab. 4.16 ersichtlich ist, führt dies, die nicht-bindenden Potentiale der C^α-Atome und die der Seitenketten zusammengenommen, zu insgesamt $2\cdot36 = 72$ Linearkombinationen, die jeweils für ein Aminosäurepaar spezifisch sind[2]. Jede Linearkom-

[2] Aus y Aminosäureklassen lassen sich $\frac{y^2+y}{2}$ von der Reihenfolge in der Sequenz unabhängige Aminosäure-Paare bilden. Für $y = 8$ ergeben sich somit 36 Paare.

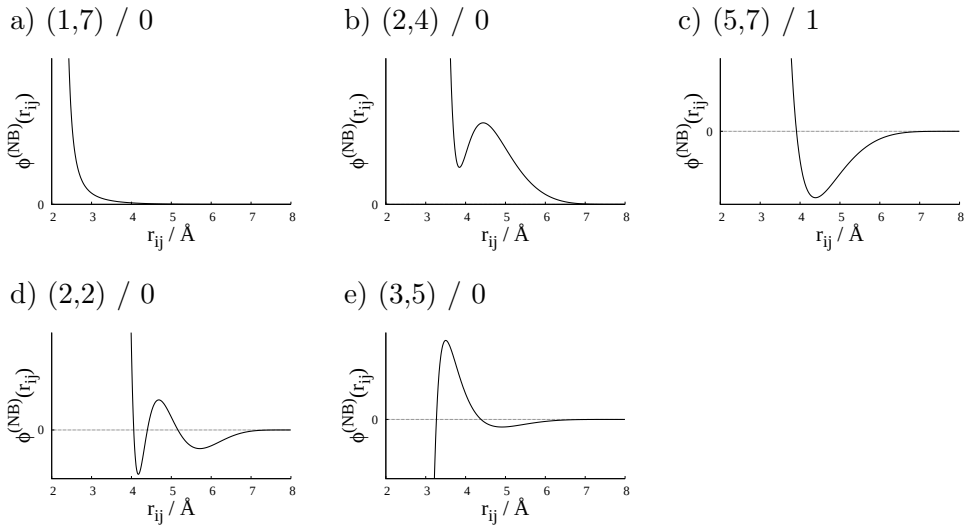


Abbildung 4.32: Beispielpotentiale der nicht-bindenden Wechselwirkungen zur Darstellungen der unterschiedlichen resultierenden funktionellen Typen nach der Linearkombination der Basisfunktionen. Die Titelzahlen "(a, b) / c" der Grafiken geben das zugehörige (codierte) Aminosäurepaar a, b und die jeweilige Iteration c an.

bination setzt sich für die Wechselwirkungen zwischen den C^{α}-Atomen aus 6 bzw. zwischen den Seitenketten aus 4 kombinierten Basisfunktionen zusammen. Die resultierenden funktionalen Formen dieser Linearkombinationen variieren von Iteration zu Iteration abhängig vom betrachteten Aminosäurepaar unterschiedlich stark, wobei beispielsweise die Positionen der Extrempunkte oder die Krümmung der Funktion angepasst wird.
Im folgenden werden hierzu einige ausgewählte Linearkombination repräsentativ dargestellt. Viele nicht-bindende Potentiale weisen sehr ähnliche funktionale Formen auf, so dass an dieser Stelle auf die am häufigsten auftretenden Formen eingegangen wird. Die hierfür verwendeten Wechselwirkungen sind in Abb. 4.32 dargestellt. Die Ordinaten sind in dieser Abbildung nicht skaliert, da diese Grafiken beispielhaft für verschiedene Potentiale mit unterschiedlichen Gewichtungskoeffizienten stehen. Die in den Grafiken verwendeten codierten Aminosäureklassen, die jeweils mehreren Aminosäuren entsprechen können, sind in Tab. 4.8 aufgeschlüsselt.

Wie aus der Abb. 4.32 ersichtlich ist, enthalten die optimierten Linearkombinationen somit Potentiale, die folgende Formen besitzen:
• Rein repulsiv (Abb. 4.32a). Hier durch das Aminosäurepaar (1,7) gegeben, das einer Wechselwirkung zwischen polaren und unpolaren Aminosäuren entspricht.
• Repulsive Potentiale, die Minima bzw. metastabile Zustände enthalten (in Abb. 4.32b). Das Paar (2,4) entspricht z. B. einer Cystein-Histidin-Interaktion.
• Bindende Potentiale, die ein Minimum enthalten und bei kleinen Abständen repulsiv werden, ähnlich zu einem klassischen Lennard-Jones-Potential (Abb. 4.32c). Das Paar (5,7) gehört zu einem Kontakt zwischen zwei polaren Aminosäuren.

• Bindende Potentiale, die zwei Minima enthalten (Abb. 4.32d). Zwei ist die höchste Anzahl an Minima, die in den optimierten Funktionen beobachtet wurde. Typischerweise enthalten diese Funktionen ein tiefliegendes Minimum bei kleineren Abständen und ein etwas flacheres Minimum bei größeren Abständen. Für einige Aminosäureklassen liegen beide Minima so dicht beieinander, dass das Potential in die Form Abb. 4.32c übergeht. Zu der hier betrachteten Aminosäureklasse 2 gehören die Aminosäuren Cystein und Tyrosin.
• Vereinzelte Linearkombinationen führten zu der Funktionsform wie sie in Abb. 4.32e gezeigt ist. Hier strebt der Funktionswert für $r_{ij} \to 0$ gegen $-\infty$. Der Verlauf dieser Potentiale zwischen den Randwerten kann stark unterschiedlich sein. Ein Teil dieser Potentiale läuft von $-\infty$ bei kleinem r_{ij} zum Funktionswert Null, ohne Extrempunkte im Zwischenbereich aufzuweisen (quasi als Funktionstyp $-1/r$), während andere Formen mehrere Extremwerte enthalten können, wie in Abb. 4.32e dargestellt. In der Parameteroptimierung wurde der Wertebereich für die Koeffizienten der Linearkombinationen so festgelegt, dass die Basisfunktion, die den größten Exponenten für r_{ij}^k besitzt, auf positive Werte beschränkt wurde, um ein repulsives Verhalten für $r_{ij} \to 0$ zu erzeugen. Wenn aber durch die Parameteroptimierung der zu dieser Funktion zugehörige Koeffizient einen sehr kleinen Wert zugeordnet bekam, konnten Potentiale mit dieser physikalisch falschen Form resultieren. Hierbei lässt sich für diese Funktion vermuten, dass ein Minimum bei den entsprechenden kleinen Abständen (z. B. bei ca. 3 Å in Abb. 4.32e) die richtigere Form für dieses Potential wäre. Es existiert aber auch die Möglichkeit, dass die Koeffizienten so optimiert wurden, um andere Zwänge auszugleichen bzw. zu erfüllen. Beispielsweise könnte es in der Parameteroptimierung notwendig sein, diese Funktion so zu wählen, um durch eine zusätzliche stark negative Energie die Optimierungsnebenbedingungen $E_{d,p} - E_p^* > 0$ zu erfüllen, während die Summe der anderen Funktionen dazu nicht ausreichend war. Ebenso könnte ein Mangel an Informationen über diesen Abstandsbereich dazu führen, dass das Potential eine falsche Form erhält. Solche problematischen Funktionsverläufe wurden auch von anderen Autoren publiziert, die ebenfalls Parameter mittels eines linearen Programms gegen falsche Strukturen optimiert haben (siehe z. B. [116]). Diese Art der Potentiale bereitet besonders auch in der globalen Optimierung mittels des genetischen Algorithmus Probleme. Da die Gesamtenergie eines Proteins als Bewertungsfunktion verwendet wurde, führte die Optimierung mit dem Kraftfeld, das diese Art von Funktionen beinhaltete, zu Strukturen, in denen die nicht-bindenden Wechselwirkungen aufgrund der stark negativen Energie bei kleinen Abständen dominierten, welche um Größenordnungen größer waren als die Summe der anderen Funktionen. Dies führt zu verwickelten Strukturen mit sehr eng benachbarten C^α-Atomen, die unrealistisch kurze nicht-bindende Abstände enthielten. Hierdurch blieben die Optimierungen in physikalisch falschen Minima weit ab von der richtigen nativen Struktur hängen. Dieses Problem wurde im genetischen Algorithmus umgegangen, indem zusätzliche repulsive Potentiale eingeführt wurden. Die Details hierzu sind in Abschnitt 4.7.3 beschrieben.

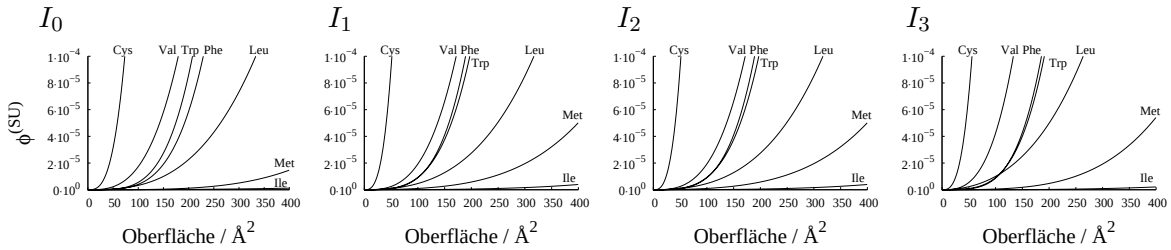


Abbildung 4.33: Oberflächenfunktionen der hydrophoben Aminosäuren in der jeweiligen Parameter-Iteration I_n.

Wie in Abschnitt 4.4.5 dargelegt wurde, wurden für das Potential insgesamt sieben Aminosäuren bestimmt, die als hydrophobe Aminosäuren angesetzt wurden. Für diese Aminosäuren wurde eine "Straffunktion" definiert, die zu einer Minimierung der Oberfläche dieser Aminosäuren führen sollte, wodurch sie im inneren des Proteins eingelagert werden sollten. Hierzu wurde ein Potenzreihenansatz der Form

$$\phi^{(SU)}(A_i) = \sum_{k=1}^{4} c_{k,s_i}^{(SU)} A_i^k \tag{4.81}$$

gemacht, wobei A die lösungsmittelzugängliche Oberfläche für die i-te Aminosäure und $c_{k,s_i}^{(SU)}$ die Koeffizienten sind. Der Wertebereich für die Koeffizienten $c_{k,s_i}^{(SU)}$ wurde zunächst nicht beschränkt (also $c_{k,s_i}^{(SU)} \in \mathbb{R}$), um die Bildung von Extrempunkten durch eine entsprechende Linearkombinationen der Potenzfunktionen zu ermöglichen. Hierdurch wurden zunächst wie beabsichtigt einigen Koeffizienten in der Parameterbestimmung negative Werte zugeordnet wurden, was jedoch für einige Aminosäuren dazu führte, dass die Linearkombinationen Gl. 4.81 wie auch schon bei den nicht-bindenden Potentialen in einer Funktion resultierten, deren Funktionswerte für große A gegen $-\infty$ strebte, was für hydrophobe Reste eine falsche Beschreibung bedeutet, da dies einer Bevorzugung einer maximal großen zugänglichen Oberfläche bzw. maximalen Exposition zur Umgebung entspricht. In der globalen Optimierung mittels des genetischen Algorithmus führte dies zu langen entfalteten Ketten. Aus diesem Grund wurden die Koeffizienten $c_{k,s_i}^{(SU)}$ in den folgenden Parameterbestimmungen auf positive Werte beschränkt ($c_{k,s_i}^{(SU)} \in \mathbb{R}_{\geq 0}$). Die Linearkombinationen mit diesen bestimmten Koeffizienten erzeugten nach oben geöffnete parabelförmige Funktionen unterschiedlicher Steigung für die einzelnen Aminosäuren, die schließlich zu einer Tendenz für alle Aminosäuren führte, die Oberfläche zu minimieren, da die Minima aller Funktionen bei $A = 0$ lagen. Die Steigung bzw. Krümmung dieser Funktionen bestimmt wie "tolerant" eine Aminosäure gegenüber einer exponierten Oberfläche ist.
Die resultierenden Linearkombination für die sieben hydrophoben Aminosäuren Cystein, Isoleucin, Leucin, Phenylalanin, Tryptophan und Valin mit den Parametern aller Iterationen sind in Abb. 4.33 dargestellt.

Die Funktionen in dieser Abbildung sind zum besseren Vergleich der Funktionsverläufe untereinander bis zu einer Oberfläche von 400 $Å^2$ gezeigt. Es sei an dieser Stelle darauf hingewiesen, dass selbstverständlich die Aminosäuren eine begrenzte maximale Oberfläche besitzen, die im Fall einer isolierten Aminosäure erreicht werden kann (zusätzlich abhängig von der internen Geometrie der Aminosäure), die für die verwendeten Aminosäuren kleiner als die hier angegebenen 400 $Å^2$ sind. Eine Übersicht, welche realen Oberflächen in nativen Proteinen zu erwarten sind, gibt Abb. 4.24.
Wie aus Abbildung 4.33 ersichtlich ist, unterscheiden sich die Potentialfunktionen der Oberflächenfunktionen von der ersten bis zur letzten Iterationen nur in Details, indem sich die Steigungen der Funktionen ändern. Besonders für Methionin von I_0 zu I_1 ebenso wie für Valin und Leucin von I_2 zu I_3 nimmt die Steigung im Vergleich zu den anderen Oberflächenfunktionen zu. Dies kann dahin gedeutet werden, dass für diese Aminosäuren solche falschen neuen Strukturen aus der globalen Optimierung hinzugekommen sind, für die eine höhere Gewichtung dieser Funktion die Unterscheidung zwischen den falschen und der nativen Struktur verbessert, indem beispielsweise diese Aminosäure eine im nativen Protein und in den falschen Strukturen eine ähnlich dichte Umgebung besitzen, wodurch die kleineren Unterschiede einer stärkeren Betonung bedürfen, was eine Skalierung der Funktionen notwendig macht.

Die letzte optimierte Funktionsklasse diente der Beschreibung der Wasserstoffbrücken zwischen den Rückgrateinheiten der Aminosäuren. Hierfür wurden drei Funktionen $\phi^{(HB)}$ angesetzt, die den Abstand der virtuellen Zentren Z, die Projektion des Wasserstoffbrückenvektors auf die entsprechenden Bindungsvektoren und die Antiparallelität zwischen zwei Wasserstoffbrücken modellieren sollten (siehe Abschnitt 4.4.6).
Die zugehörigen Koeffizienten wurden unabhängig von den an einer Wasserstoffbrücke beteiligten Aminosäuren gewählt, da zum einen davon ausgegangen wurde, dass alle Aminosäure ein ähnliche Tendenz besitzen, Wasserstoffbrücken auszubilden und zum anderen deswegen, weil die Präferenzen, welche Aminosäuren zur Zusammenlagerung neigen und somit die Möglichkeit zur Ausbildung einer Wasserstoffbrücke bieten, über die nicht-bindenden und die Seitenketten-Wechselwirkungen erfasst wurden. Aufgrund der speziellen Formulierungen der Basisfunktionen zu diesen Wasserstoffbrücken-Wechselwirkungen wurden die Koeffizienten für die Parameteroptimierung auf negative Werte beschränkt.
Die nach Gl. 4.80 durchschnittliche Energien für ein Protein, welche diese drei Funktionen einzeln beitragen, ist für jede Parameteroptimierungsiteration in Tab. 4.20 dargestellt. Diese zeigt, dass die Wasserstoffbrückenenergien in unterschiedlichen Größenordnungen zur Gesamtenergie beitragen. Den dominantesten Beitrag liefert die abstandsabhängige Funktion $\phi_1^{(HB)}$. Die Funktion $\phi_2^{(HB)}$ trägt Energien bei, die bereits, abhängig von der betrachteten Iteration um zwei bis vier Größenordnungen kleiner sind als die Energien der Funktion $\phi_1^{(HB)}$. Der Beitrag der Funktion $\phi_3^{(HB)}$ ist ebenfalls noch einmal um eine bis vier Größenordnungen klei-

I	$\phi_1^{(HB)}$	$\phi_2^{(HB)}$	$\phi_3^{(HB)}$
0	-7.482·10^{-4}	-5.524·10^{-8}	-8.382·10^{-10}
1	-1.014·10^{-3}	-1.878·10^{-7}	-6.318·10^{-11}
2	-5.676·10^{-4}	-1.437·10^{-6}	-1.746·10^{-8}
3	-1.694·10^{-3}	-1.645·10^{-7}	-3.929·10^{-8}

Tabelle 4.20: Gesamtenergien der einzelnen Wasserstoffbrücken-Basisfunktionen gemittelt über alle verwendeten nativen Proteine in Abhängigkeit von der Parameter-Iteration I.

ner. Somit besitzen $\phi_2^{(HB)}$ und $\phi_3^{(HB)}$ einen vernachlässigbaren Einfluss auf die Gesamtenergie eines Proteins.

Betrachtet man nun die relativen Beiträge der eben beschriebenen Wechselwirkungsklassen zur Gesamtenergie in Abhängigkeit von den durchlaufenen Iterationen I (siehe Abb. 4.31), so ändern sich über die Iterationen die relativen Beiträge der einzelnen Terme bis auf eine Ausnahme nur in kleinem Maßstab. Grundsätzlich trägt auf der Seite der negativen Energien der Nahwechselwirkungsterm am meisten Energie bei, gefolgt von den Wasserstoffbrückenbindungen, während die nicht-bindenden Wechselwirkungen zentriert auf den C^{α}-Atomen und die nicht-bindenden Wechselwirkungen zentriert auf den Seitenketten zusammengenommen in etwa einen ähnlich großen Beitrag liefern wie die Wasserstoffbrückenbindungen.
Der Term der Wasserstoffbrückenbindungen zeigt einen Sprung zur Iteration I_2, in welcher deren relativer Beitrag stark reduziert wird, so dass diese einen kleineren Wert als die nicht-bindenden Wechselwirkungsterme beitragen. Zu I_3 hingegen nimmt der relative Anteil dieses Potentials wieder zu und erreicht das Niveau vor I_2. Diese Sprung entsteht in den Parametern durch einen weniger negativen Koeffizienten der Funktion ϕ_1^{HB} bei gleichzeitig stärker negativen Koeffizienten beispielsweise der Nahwechselwirkungsterme wie ϕ_3^{BO} (vergleiche Tab. 4.19 und 4.20). Ein möglicher Grund hierfür könnte darin liegen, dass im Sortierungsprozess der falschen Strukturen für die Parameteroptimierung der folgenden Iteration ein oder mehrere Proteine hinzukommen, die eine im Vergleich zu den nativen Proteinen größere Anzahl an Wasserstoffbrückenbindungen aufweisen, wodurch deren Beitrag and der Gesamtenergie durch eine entsprechende Parameteranpassung reduziert werden musste, da diese falschen Strukturen ansonsten energetisch günstiger wären, wodurch die Optimierungsnebenbedinungen verletzt wären. Diese Strukturen kommen hinzu da vor der Parameteroptimierung alle falschen Strukturen, mit Ausnahme der aus dem GA stammenden Proteine, anhand ihrer Energie sortiert werden und nur die energetisch niedrigsten verwendet werden. Zu der Folgeiteration I_3 könnten diese Strukturen im Sortierungsprozess wieder entfernt worden sein.
Des weiteren zeigt sich im Verlauf der Iterationen, dass der Anteil der nicht-bindenden Wechselwirkungen zwischen den C^{α}-Atomen relativ zu den Nahwechselwirkungstermen abnimmt.

So liegt deren Anteil in I_0 bei ca. 40 % der Nahwechselwirkungsenergie und sinkt zu I_3 auf rund 20 %. Der Anteil der nicht-bindenden Energie zwischen den Seitenketten bleibt dagegen bei ca. 40 % relativ konstant über alle Iterationen. Ebenso bleibt auch der relative Anteil der Oberflächenenergie annähernd konstant.
Betrachtet man zusätzlich noch die relativen Energien bezogen auf die Sekundärstrukturmerkmale anhand der für Abb. 4.31 verwendeten Proteinen, so zeigt sich hier ebenfalls ein Zusammenhang, der besonders für das Protein 1a92(A) deutlich zu erkennen ist, bei welchem der Anteil der nicht-bindenden Energie sowohl zwischen den C^{α}-Atomen als auch zwischen den Seitenketten an der Gesamtenergie im Vergleich zu den anderen drei Termen vernachlässigbar ist. Dies ist darin begründet, dass dieses Protein zwei Untereinheiten enthält, die jeweils in der α-helicalen Konformation vorliegen, die aber nur eine kleine gemeinsame Kontaktfläche besitzen, wodurch die Seitenketten gegenüber der Umgebung exponiert sind. Zusätzlich sind durch die ausgedehnte helicale Konformation die Seitenketten der in der Sequenz benachbarten Aminosäuren räumlich weit voneinander entfernt, wodurch insgesamt die nicht-bindenden Wechselwirkungen zwischen den Seitenketten und zwischen den C^{α}-Atomen sehr gering sind. Da in diesem Protein fast alle Aminosäuren an einer α-Helix beteiligt sind, ist der Beitrag durch die Wasserstoffbrückenbindungen so groß, dass diese den sonst dominierenden Nahwechselwirkungsterm übersteigt (z. B. in I_1 und I_3). Die anderen drei Proteine 1ifc, 1dhn und 1chd weisen dagegen nicht diese großen Unterschiede zwischen den Potentialfunktionen auf, sondern besitzen dagegen die oben beschriebene Verteilung. Diese unterscheidet sich auch zwischen diesen drei Proteinen nur in Details, so dass beispielsweise 1dhn eine annähernd ausgeglichene Verteilung zwischen den nicht-bindenden Funktionen zwischen den C^{α}-Atomen und zwischen den Seitenketten besitzt, während in 1chd die Seitenkettenwechselwirkung größer ist. Dies lässt sich auch wieder mit der Struktur dieser Proteine erklären, da 1dhn eine große β-Faltblattstruktur besitzt, deren Seitenketten direkt an die Umgebung angrenzen, wodurch diese wenige Interaktionspartner besitzen. 1chd besitzt eine hiermit vergleichbare ausgedehnte Faltblattstruktur, die aber im Unterschied zu 1dhn im inneren des Proteins eingelagert und von α-Helices umgeben ist, wodurch die Struktur insgesamt kompakter ist.

Weiterhin lassen sich die Energien zwischen den nativen und den falschen Strukturen vergleichen. Hierzu wurden die Energien aller falschen Proteine aus allen Strukturs̈atzen d berechnet und die Differenzen der Energien $\Delta E_{p,d}$ zu dem zugehörigen nativen Protein p bestimmt. Für diese Differenzen wurde der Mittelwert $\mu(\Delta E_{p,d})$ berechnet, indem über alle falschen Proteine eines Satzes nach folgender Gleichung gemittelt wurde:

$$\mu(\Delta E_{p,d}) = \frac{1}{\nu_{p,d}} \sum_{i=1}^{\nu_{p,d}} E_{p,d,i} - E_p^* \tag{4.82}$$

wobei p die Nummer des nativen Proteins mit $p \in \{1, \ldots, 48\}$ und d die Klasse bzw. Satz der falschen Strukturen ist. Es ist $d \in \{1, 2, 3, 4\}$, was identisch mit den Sätzen "Verzerrte

Strukturen", LKF2, GA-Strukturen und Proteine aus der Sequenzübertragung ist. Weiterhin ist $\nu_{p,k}$ die Gesamtanzahl an falschen Strukturen der Klasse k, $E_{p,d,i}$ die Energie der i-ten falschen Struktur und E_p^* die Energie des nativen Proteins. Gl. 4.82 gibt den gemittelten Gesamtenergieunterschied zwischen den Strukturen wieder. Dies lässt sich weiterhin in die einzelnen Funktionsklassen $k \in \{1, \ldots, 5\}$ des Potentials aufschlüsseln, wobei Gl. 4.82 dann übergeht in

$$\mu\left(\Delta E_{p,d}^{(k)}\right) = \frac{1}{\nu_{p,d}} \sum_{i=1}^{\nu_{p,d}} E_{p,d,i}^{(k)} - E_p^{*(k)} \tag{4.83}$$

Die Potentialklassen k entsprechen hierbei Nahwechselwirkungen (BO), nicht-bindende Wechselswirkungen (NB), Oberflächenpotential (SU), Seitenkettenwechselwirkungen (SS) und Wasserstoffbrückenbindungen (HB). Die Mittelwerte, die sich aus Gl. 4.83 ergeben, sind in Abb. 4.35 für die bereits für Abb. 4.31 verwendeten Proteine dargestellt. Positive Energien bedeuten in dieser Darstellung, dass die falsche Struktur eine höhere (schlechtere) Energie besitzt als die native. Wie aus dieser Abbildung ersichtlich ist, sind die durchschnittlichen Beiträge, die die Strukturen unterscheiden, abhängig sowohl vom betrachteten Protein als auch von der Klasse der falschen Strukturen, so dass es keinen generell dominanten Term gibt, der alle Strukturen gleichermaßen unterscheidet.

Betrachtet man zunächst die Klasse der verzerrten Strukturen, so ist für diese der Term der Wasserstoffbrückenbindungen entscheident. Dies lässt sich dadurch erklären, dass die falschen Strukturen so erzeugt wurden, dass zufällig gewählte C^α-Atome um eine Achse rotiert wurden, die durch das in der Sequenz vorangehende und folgende C^α-Atom definiert war. Hierdurch änderte sich ebenfalls dich Position der Zentren Z zwischen den C^α-Atomen, die zur Ausbildung von Wasserstoffbrückenbindungen dienen. Durch eine Änderung der Position der Zentren Z ändern sich entsprechend die Abstände zu den anderne Zentren sowie die Orientierungen zueinander. Da Wasserstoffbrücken im gewählten Modell nur berechnet werden, wenn die Zentren zueinander bestimmte Abstands- und Winkelkriterien erfüllen, kann die Verzerrung leicht zu einem Bruch einer Wasserbrücke führen, woraus sich die größeren Energieunterschiede ergeben. Gefolgt wird dieser Energieunterschied der Wasserstoffbrückenbindungen von den Seitenketten-Wechselwirkungen. Zunächst ist hier bemerkenswert, dass der Unterschied in den Seitenketten-Wechselwirkungen wesentlich größer als der Unterschied in nicht-bindenden C^α-Wechselwirkungen ($\phi^{(NB)}$) ist. Dies ergibt sich zum einen dadurch, dass die Seitenkette des rotierten C^α-Atoms am weitesten vom Rotationszentrum entfernt, wodurch die Positionsänderung der Seitenketten am größten und zum anderen dadurch, dass bei Änderung der Position des i-ten C^α-Atoms auch die Seitenketten der Aminosäuren $i-1$ und $i+1$ verändert werden, wodurch insgesamt drei Seitenketten neu gesetzt werden. Als drittgrößter Term dient bei der Unterscheidung die Nahwechselwirkungen, wobei hier ebenfalls zu beachten ist, dass die Rotation eines C^α-Atoms zu einer Änderung von bis zu sechs Nahwechselwirkungstermen führen kann ($\phi_3^{(B)}$ bis $\phi_8^{(B)}$).

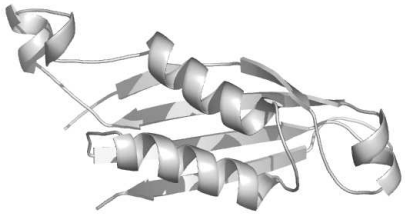


Abbildung 4.34: Native Struktur des Proteins 1dhn in der *Cartoon*-Darstellung.

Betrachtet man als nächstes den LKF2-Datensatz, so ist für diesen das Wasserstoffbrückenpotential in der Utnerscheidung der Strukturen ebenfalls sehr wichtig, was wieder auf eine unterschiedliche Gesamtanzahl an Wasserstoffbrücken und/oder auf unterschiedlich gut orientierte Zentrenverbindungsvektoren zurückzuführen ist. Bei den anderen Termen zeigt sich kein einheitliches Bild. So ist für 1a92(A) die Nahwechselwirkungen für die Unterscheidung noch wichtig, weil die native Struktur von 1a92(A) fast ausschließlich in der helicalen Konformation vorliegt, welche als Minimum für die unterschiedlichen Potentialfunktionen definiert wurde, wodurch andere Konformationen, die beispielsweise Zufallsgeometrien (*random coil*) oder Windungen enthalten, die nicht im Bereich des Minimums liegen, zu schlechteren Nahwechselwirkungsenergien führen. Bei dem Protein 1dhn sind zudem hier die Unterschiede in den nicht-bindenden Wechselwirkungen zwischen den C^{α}-Atomen markant. Zum Verständnis muss die Struktur des Proteins näher betrachtet werden. Die native Struktur besteht aus einer Mischung von α- und β-Konformationen, wobei der Zentralteil der Struktur daraus besteht, dass zwei α-Helices mit vier β-Faltblättern eine kompakte Gesamteinheit bilden (siehe Abb. 4.34). Der große energetische Unterschied ist auf einige bestimmte Wechselwirkungen zwischen C^{α}-Atomen zurückzuführen, die ausschließlich in der ausgedehnten β-Struktur des Proteins liegen, wobei es sich um räumlich direkt-benachbarte C^{α}-Atome handelt, die aber in der Sequenz weit voneinander entfernt sind, so zum Beispiel die Paarungen der Reste (16, 70), (31, 98) und (35, 94). Diese Wechselwirkungen fehlen ganz oder teilweise in den LKF2-Strukturen, indem diese Paarungen dort entweder sehr weit voneinander entfernt sind oder nicht den "optimalen" Abstand zwischen zwei Strängen mit β-Konformation besitzen. Hieraus kann gefolgert werden, dass für dieses Protein, ähnlich deutlich wie bei 1a92(A), das eine reine α-helicale Konformation besitzt, die Strukturunterscheidung, neben der abweichenden Anzahl an Wasserstoffbrücken, auf die richtige Zuordnung der entsprechenden Reste zu einer β-Konformation zurückzuführen ist.

Vergleicht man weiterhin die Energiebeiträge zur Unterscheidung für die falschen Strukturen, die aus dem genetischen Algorithmus stammen, so zeigt sich bei diesen, dass, mit Ausnahme der helicalen Struktur des Proteins 1a92(A), die anderen Proteine anhand der schlechteren Nahwechselwirkungen unterschieden werden. Die schlechtere Nahwechselwirkugnsnergie in diesen Strukturen resultiert aus Geometrien, in denen der C^{α}_{i-1}-C^{α}_{i}-C^{α}_{i+1}-Winkel sehr groß wird, wodurch der C^{α}_{i-1}-C^{α}_{i+1}-Abstand sehr groß wird (> 7 Å) und dadurch in den repulsiven Teil des entsprechenden Nahwechselterms liegt. Dies ist ein Effekt, der daraus resultiert, dass

in den späteren Generationen des genetischen Algorithmus die Geometrien der Proteine kompakter werden, worduch die nicht-bindenden Wechselwirkungen wichtiger werden, was dazu führen, dass eine Geometrie mit lokal schlechteren Nahwchselwirkungstermen durch günstigere nicht-bindende Wechselwirkungen kompensiert wird. Hierbei müssen alle Terme wie auch die nicht-bindenden Potentiale zwischen den Seitenketten und das Oberflächenpotential mitberücksichtigt werden (Siehe hierzu auch Abschnitt 4.8.2). Weiterhin ist an dieser Stelle zu beachten, dass die gezeigten Grafiken auf Parametern basieren, die nach der Generierung der verwendeten GA-Strukturen erhalten, so dass die gezeigten Energien darauf optimiert wurden, sich von den nativen zu unterscheiden, während dagegen dieselben Strukturen in den GA-Läufen insgesamt energetisch niedriger (besser) als die nativen Strukturen waren. Wären sie nicht energetisch niederiger gewesen, wären sie nicht zur Parameteroptimierung herangezogen worden.
Die letzte zu betrachtende Strukturklasse enthält die aus der Sequenzübertragung erhaltenen falschen Geometrien. Hier ist zunächst auffällig, dass sich das Muster der Energiedifferenzen beim Protein 1a92(A) von der anderen drei betrachteten Proteinen unterscheidet, die sehr ähnliche Differezen aufweisen. Bemerkenswert ist hierbei, dass zu erwarten gewesen wäre, dass eine Unterscheidung auch anhand der nicht-bindenden Wechselwirkung sowohl der C^{α}-Atome als auch der Seitenketten stattfinden müsste, da diese in 1a92(A) aufgrund der fast durchgängigen α-helicalen Struktur nur eine untergeordente Rolle spielen, da es nur eine geringe Berührungsfläche zwischen unterschiedlichen Proteinsegmenten gibt. Überträgt man die Sequenz auf andersartig aber kompakt gefaltete native Strukturen, so sind für diese Terme größere Wechselwirkungsenergien zu erwarten. Die Grafik 4.35 zeigt jedoch nur eine im Vergleich zu den Nahwechselwirkungen und den Wasserstoffbrückenbindungen geringe Differenz der nicht-bindenden Wechselwirkungen. Durch eine Analyse der Energiebeiträge der durch Sequenzübertragung erhaltenen Strukturen wurde dies auf eine doppelte Kompensation an Energien zurückgeführt. Die Sequenz von 1a92(A) ist reich an hydrophilen Aminosäuren. Mehr als 2/3 aller Aminosäuren sind dieser Klasse zuzuordnen. Nicht-bindende Wechselwirkungen zwischen diesen Aminosäuren sind häufig mit einem repulsiven Potential verbunden. Durch die kompakte Faltung der verwendeten Datenbank-Proteine führt dies, bei bestimmten Strukturen, zu einer großen repulsiven (positiven) nicht-bindenden Energie, sowohl zwischen den Seitenketten als auch zwischen den C^{α}-Atomen. Dies kann dazu führen, dass die gesamte nicht-bindende Energie sich zu Null mittelt, was dem ersten hier erwähnten Kompensationseffekt entspricht, oder dass diese Energie sogar positiv ist. Viele der Strukturen aus der Sequenzübertragung besitzen aber tatsächlich eine sehr große stabilisierende (negative) nicht-bindende Energie. Neben diesen finden sich weniger häufig Proteine, die eine sehr stark destabilisierende (positive) nicht-bindende Energie besitzen, die aus den erwähnten ungünsten kurzen Abständen der hydrophilen Aminosäuren resultiert. Diese kann so groß sein, dass sie die negativen Energien für die Mittelwertbildung (siehe Gl. 4.83) zu Grafik 4.35 zu einem

sehr kleinen Gesamtwert kompensieren (zweiter oben erwähnter Effekt). Neben diesen beiden Kompensationseffekten ist für das Protein 1a92(A) weiterhin zu beachten, dass nahezu die maximal mögliche Anzahl an Wasserstoffbrücken ausgebildet ist, wodurch dieser Term bei der Unterscheidung sehr dominant ist. Die Energie des Wasserstoffbrückenpotentials ist in der nativen Struktur vergleichen mit den falschen Strukturen um bis zu zwei Größenordnungen energetisch günstiger als die nicht-bindenden Wechselwirkungen. Die anderen drei aufgeführten Proteine zeigen im Vergleich mit 1a92(A) ein anderes Muster bei der Unterscheidung. Zunächst ist für diese Proteine die Nahwechselwirkungen weniger entscheident, da native Strukturen verwendet werden, die nah an den Idealwerten für die bekannten Sekundärstrukturen liegen. Besonders wichtig sind bei der Unterscheidung die Seitenketten-Potentiale. Dies beruht auf der Methode der Erzeugung dieser Strukturen. Da die Sequenz der originalen Datenbank-Proteine mit den Sequenzen des gewählten TOP500H-Datensatzes überschrieben wurden, konnten die ursprünglichen Datenbank-Seitenketten-Positionen nicht mitverwendet werden. Daher war es nötig, diese neu zu positionieren (siehe Abschnitt 4.3.2). Da hierfür aber mit Ausnahme der erlaubten Minimalabstände zwischen den Seitenketten und zu den C^{α}-Atomen keine externen Informationen wie beispielsweise eine anderes Kraftfeld verwendet wurde, wurden die Strukturen nicht relaxiert, so dass die Seitenketten nicht die energetisch günstigste Position einnehmen konnten. Hierdurch sind destabilisierende Wechselwirkungen möglich. Ebenso sind bei dieser Strukturklasse, abgesehen von Wasserstoffbrückenbindungen, auch die nicht-bindenden Potentiale zwischen den C^{α}-Atomen relevant, die wie schon bei 1a92(A) beschrieben aus ungünstigen engen Packungen zwischen C^{α}-Atomen mit repulsiven Potentialen oder aus auch aus dem Fehlen einer entsprechenden Wechselwirkungsumgebung durch Auslassen der Sequenz (siehe Abschnitt 4.3.2) für ein C^{α}-Atom resultieren können.

Insgesamt zeigt sich, dass zur Unterscheidung der falschen von den nativen Strukturen weder lediglich ein Potential-Term noch ein wiederkehrendes Muster von mehreren Termen generell auftritt, sondern dass die Unterscheidungsterme stark von der Struktur des nativen Proteins und von der Methode zur Erzeugung der falschen Strukturen abhängt, wobei sich innerhalb einer Klasse von falschen Strukturen ähnliche Energiemuster ergeben.

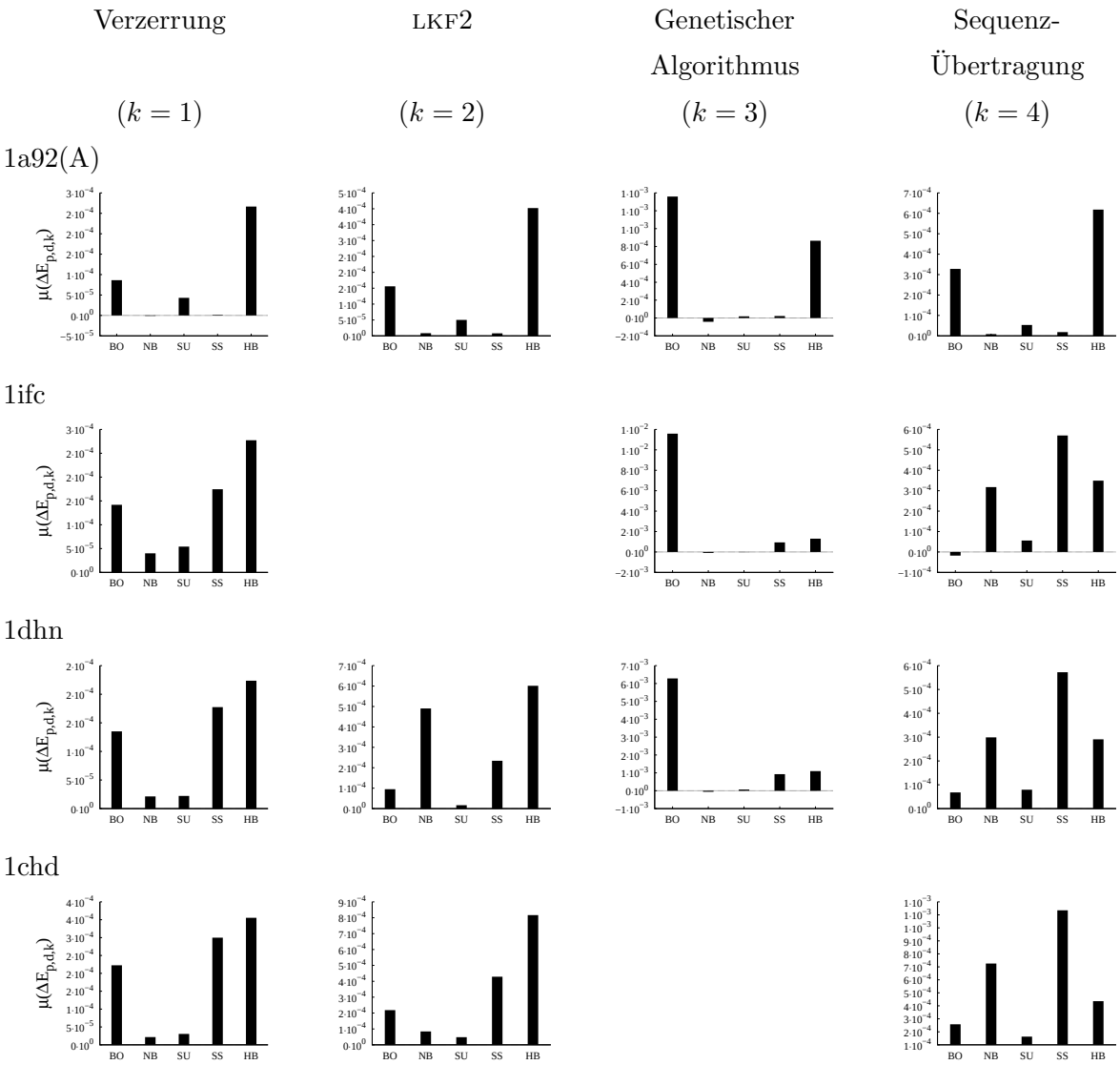


Abbildung 4.35: Vergleich der Energien zwischen falschen und nativen Strukturen. $\mu(\Delta E_{p,l,k})$ ist die über alle falschen Strukturen eines Datensatzes k gemittelte Energiedifferenz zur nativen Struktur p basierend auf den Parametern der Iteration I_3 (siehe Gl. 4.83). Auf der Ordinate sind die Funktionsklassen aufgetragen, mit BO = bindend bzw. Nahwechselwirkung, NB = nicht-bindend, SU = Oberflächenpotential, SS = Seitenkettenpotential, HB = Wasserstoffbrückenbindungen.

4.6.2 Erkennungstest

Potentialfunktionen für die *ab-initio*-Proteinfaltung werden mit dem Ziel erstellt, durch die in der Funktion und dem Modell enthaltenen Informationen Vorschläge für die native Struktur einer unbekannten Sequenz zu erhalten bzw. um die Faltung bekannter Sequenzen besser verstehen zu können. Dazu wird die gewählte Potentialfunktion bzw. die in ihr enthaltenen Parameter anhand eines bestimmten ausgewählten Proteinstruktursatzes optimiert, was auch als "Training" bezeichnet wird. Bevor eine Potentialfunktion zur Strukturvorhersage verwendet werden kann, muss sie an bekannten Proteinen getestet werden. Das Ziel einer Potentialkonstruktion zur Faltungsvorhersage ist bekanntlich, dem nativen Zustand einen ausgezeichneten Funktionswert des Potentials zuzuordnen, was in der Regel einem tiefen und im Idealfall dem globalen Minimum der Funktion entspricht, während alle anderen Strukturen, die eine gewisse geometrische Distanz zum nativen Zustand besitzen, einen höheren Funktionswert einnehmen. Eine hierzu häufig angewendete Prozedur ist zu prüfen, inwieweit das gegebene Potential in der Lage ist, in einem großen Satz an möglichen Alternativstrukturen das native Protein zu identifizieren, wobei dieser Testsatz generell keine Proteine enthält, die bereits im Optimierungssatz ("Trainingssatz") enthalten waren. Dazu werden alle gegebenen Strukturen zu einer Testsequenz anhand ihrer Funktionswerte des Potentials sortiert und der Rang (Platz) der nativen Struktur in der Gesamtreihenfolge bestimmt. Dies wird in der Literatur als *ranking* bezeichnet. Hierdurch erhält man Informationen darüber, ob das Potential im durch die alternativen Strukturen abgedeckten Konformationsraum die richtigen Vorhersagen liefert.
Im Idealfall ordnet somit ein Potential allen nativen Strukturen den ersten bzw. letzten Rang zu, in Abhängigkeit von der Definition des Sortierungsschemas, wobei in der Literatur praktisch ausschließlich die energieärmste Struktur den ersten Rang erhält. Neben dieser einfachen Aufstellung der Reihenfolge von Strukturen, wird in der Regel auch der Z-Wert (siehe Gl. 4.48) dieser Reihenfolge angegeben, der ein statistisches Merkmal einer Verteilung ist, und Informationen darüber enthält, ob die Identifizierung der nativen Struktur statistisch signifikant ist oder ob eine zufällige Zuordnung zu einem ähnlichen Ergebnis führen würde. Dieser Wert enthält zusätzlich Informationen darüber, inwieweit die native Struktur und die Alternativstrukturen energetisch voneinander getrennt sind, da es für eine Strukturvorhersage wichtig ist, ob der native Zustand energetisch so weit von den anderen Geometrien getrennt ist, dass eine Identifizierung möglich ist, oder ob sehr viele energetisch ähnliche Strukturen existieren, die im ungünstigsten Fall auch im wesentlichen voneinander abweichende Geometrien besitzen. Für diese Untersuchung der Qualität der Potentialfunktionen werden in der Regel bereits bekannte publizierte Sätze falscher Strukturen verwendet (siehe zur Übersicht Tab. 4.6), wobei viele Autoren meist nur eine kleine Auswahl aller Sätze verwenden und aus

bestimmten Gründen teilweise auch einzelne Proteine aus den verwendeten Sätzen entfernen, weil diese zum Beispiel entweder für das gewählte Proteinmodell problematisch wären oder weil diese ungewöhnliche oder fehlerhafte native Strukturen besitzen.

Zum Testen des in dieser Arbeit entwickelten Potentials wurden vier verschiedene Struktursätze verwendet. Es handelt sich hierbei um die Datensätze Rosetta, Lmds, Lattice_ssfit und 4state_reduced. Der Rosetta-Satz wurde ausgewählt, da er mit 56 Proteinen und ca. 1000 falschen Strukturen je Protein-Sequenz eine sehr große Anzahl an Vergleichsgeometrien bereitstellt, die durch Neuanordnung von bekannten Strukturelementen aus Datenbank-Proteinen generiert und im Anschluss durch ein *Simulated-annealing*-Verfahren relaxiert wurden, wodurch die Strukturen kompakte und nativ-ähnliche Geometrien besitzen. Dieser Satz wird hier dazu verwendet, die Rangbestimmung und die erhaltenen Z-Werte in Abhängigkeit von der Parameteroptimierungsiteration zu betrachten, und zusätzlich zu zeigen, wodurch bestimmte native Proteine richtig erkannt wurden und wodurch andere native Proteine falsch bestimmt wurden. Neben diesem Satz wurden auch die Datensätze Lmds, Lattice_ssfit und 4state_reduced verwendet. Diese sind sowohl von der Anzahl an Proteinsequenzen wie auch von der Gesamtanzahl an falschen Strukturen kleiner als der Rosetta-Satz (vergleiche Tab. 4.6). Diese Struktursätze wurden verwendet, um die Ergebnisse direkt mit anderen publizierten Literatur-Ergebnissen zu vergleichen. Im folgenden werden zunächst die Ergebnisse des Rosetta-Satzes diskutiert und im Anschluss daran der Vergleich mit den anderen Datensätzen.

Zur Berechnung der Reihenfolge (*ranking*) der Strukturen und der Z-Werte war es nötig, den Funktionswert des Potentials zu jeder gegebenen Struktur zu berechnen.
Bevor auf die Ergebnisse eingegangen wird, müssen an dieser Stelle zunächst einige Hinweise darauf gegeben werden, wie die Test-Datensätze verwendet wurden, da einige der enthaltenen Proteine für das in dieser Arbeit gewählte Modell und Potential ungeeignet waren. Diese, im Anschluss aufgeführten Punkte, gelten für alle im folgenden besprochenen Struktur-Datensätze:
• Es wurden aus den Datensätzen Rosetta, Lmds, Lattice_ssfit und 4state_reduced keine Proteine verwendet, die im Parameteroptimierungssatz enthalten waren.
• Einige Proteine der Test-Datensätze enthielten Disulfidbrücken zwischen zwei Cystein-Seitenketten, was im hier gewählten Proteinmodell nicht miterfasst wurde, da keine entsprechenden Funktionen implementiert waren. Bei diesen Proteinen wurde so verfahren, dass die nicht-bindenden Wechselwirkungen zwischen den Disulfidbrückenpartnern, sowohl zwischen den C^{α}-Atomen als auch zwischen den Seitenketten, aus der Potentialberechnung herausgenommen wurden. Alle anderen Wechselwirkungen wurden regulär berechnet. Hierbei wurde die Annahme getroffen, dass diese stabilisierende (bindende) Wechselwirkung für alle Proteine zu einer Sequenz ungefähr gleich groß ist bzw. dass die Anzahl an Disulfidbrücken in den

falschen und in den nativen Strukturen gleich ist.

• Die aus den TOP500H-Datenbank-Strukturen ermittelten kürzesten Abstände zwischen den Wechselwirkungszentren zur Definition der repulsiven Potentiale für die Faltung mittels des genetischen Algorithmus wurden bei der Energieberechnung nicht verwendet, um Situationen zu vermeiden, in welcher native Proteine der Testsätze Abstände enthielten, die kürzer als die Minimalabstände des TOP500H-Satzes sind. Dies hätte zur Folge, dass diese nativen Strukturen eine zusätzliche destabilisierende Wechselwirkung enthalten würden, die in der realen Energiefläche des Proteins nicht enthalten ist und lediglich auf den unvollständigen und begrenzten statistischen Informationen des TOP500H-Satzes beruht.

• Eine große Anzahl an nativen Proteinstrukturen wurde mittels NMR-Spektroskopie bestimmt, so dass für diese gewöhnlich mehrere alternative Modelle existieren, deren Geometrien sich nicht unerheblich voneinander unterscheiden. Da alle diese Modelle zu den experimentellen Daten passen, ist es unklar, welches Modell die reale Struktur am besten nähert und als Vergleich mit den falschen Strukturen verwendet werden sollte. In diesem Fall wurde jeweils die erste Struktur einer PDB-Datenbank-Datei verwendet.

• Einige Testsatz-Strukturen enthielten Fremdatome oder Fremdmoleküle. Beispielsweise waren einige Proteinstrukturen während der Bindung des Proteins an einen DNA-Strang bestimmt worden oder sie enthielten Metallionen wie Zink oder Eisen. Da das gewählte Modell und Potential für solche Systeme nicht ausgelegt war, wurden die Wechselwirkungen mit den Fremdatomen und -molekülen ignoriert und nur die proteininterne Energie bestimmt.

• Einige falsche Strukturen enthielten Unterbrechungen im Rückgrat. Diese wurden über einen C_i^α-C_{i+1}^α-Abstand > 4.5 Å definiert. Eine Kettenunterbrechung hat zur Folge, dass die Nahwechselwirkungsterme, die von den C^α-Abständen abhängen, aufgrund des zu langen Abstands eine sehr große destabilisierende Energie beitragen. Daher wurde in einem solchen Fall nur der Teil des Proteins bis zur Unterbrechung zur Potentialauswertung verwendet. Andere Proteine enthielten von vornherein eine von der nativen Struktur abweichende Anzahl an Aminosäuren. Hierbei wurde bei der Berechnung des Potentials nur der zwischen dem nativen Protein und den falschen Strukturen gemeinsame Teil der Sequenz berücksichtigt. Beide Situationen können in den schlechtesten Fällen nur als Schätzung für die echte Energie gelten, da durch das Auslassen von Aminosäuren ein großer Einfluss auf die proteininterne Stabilisierung ausgeübt werden kann, indem beispielsweise eine neuen Grenzfläche zur Umgebung von sonst eingelagerten hydrophoben Seitenketten erzeugt wird oder indem Partner von Wasserstoffbrückenbindungen entfernt werden. Beide Arten der Sequenzverkürzung traten bei den falschen Strukturen am Ende bzw. am Anfang der Sequenz auf, so dass beispielsweise lediglich die letzten beiden Aminosäure ausgelassen (siehe z. B. Protein 4rxn weiter unten). Es trat kein Fall auf, in welchem wesentliche Teile des Protein wegfielen.

Rangfolge im Rosetta-Datensatz

Die Berechnung der Reihenfolge und die Z-Werte erfolgte in Abhängigkeit von der Parameteriteration n (mit $n \in \{0, 1, 2, 3\}$). Die Ergebnisse hierfür für den Rosetta-Struktursatz sind in Tab. 4.22 angegeben, wobei R_n der Rang nativen Struktur unter allen falschen Strukturen und der Z-Wert Z_n angegeben sind. Zusätzlich sind Informationen über die Sekundärstruktur-Zusammensetzung des nativen Proteins, die experimentelle Methode zur Strukturaufklärung und das Verhältnis $\bar{A}_{p,d}$ der gesamten lösungsmittelzugänglichen Oberfläche von falscher zu nativer Struktur enthalten, die sich aus

$$\bar{A}_{p,d} = \frac{\mu\big(A_{p,d}\big)}{A_p^*} \tag{4.84}$$

ergibt, wobei $\mu\big(A_{p,d}\big)$ der Mittelwert der lösungsmittelzugänglichen Oberflächen $A_{p,d}$ für alle falschen Strukturen des Datensatzes d (hier identisch mit dem Rosetta-Satz) zu einer Sequenz bzw. zu einem nativen Protein p ist, welches eine Oberfläche von A_p^* besitzt. Der Oberflächenmittelwert für die falschen Strukturen berechnet sich nach:

$$\mu\big(A_{p,d}\big) = \frac{1}{\nu_{p,d}} \sum_{i=1}^{\nu_{p,d}} A_{p,d,i} \tag{4.85}$$

Hier ist $\nu_{p,d}$ die Gesamtanzahl an falschen Strukturen zum Protein p im Rosetta-Struktursatz und $A_{p,d,i}$ die absolute Oberfläche (in $\mathrm{\AA}^2$) der i-ten falschen Struktur. Das Verhältnis $\bar{A}_{p,d}$ dient hier als Maß für die relative Kompaktheit der Strukturen, wobei für $\bar{A}_{p,d} > 1$ im Mittel die falschen Strukturen weniger kompakt als die native Struktur sind bzw. deren mittlere Oberfläche um den Faktor $\bar{A}_{p,d}$ größer als die der nativen Struktur sind. Dieser Unterschied kann als Maß für die Kompaktheit verwendet werden, da die Sequenzen der falschen und der nativen Strukturen identisch sind, wodurch die unterschiedlichen Oberflächen nur durch verschiedene Geometrien, aber nicht beispielsweise durch eine andere Zusammensetzung der Seitenketten entstehen kann. Die zur Berechnung verwendeten Oberflächen wurden mittels der statistischen Näherung bestimmt, die in Abschnitt 4.4.5 dargestellt wurde.

Das in der Tab. 4.22 aufgeführte Ergebnis zeigt, dass für ca. 33 der 53 Proteine (ca. 57 %) die native Struktur richtig erkannt wurde, indem dieser die niedrigste Energie zugeordnet wurde. Für weitere drei Proteine (insgesamt 68 %) ist der native Zustand unter den zehn niedrigsten Strukturen. Daneben zeigen sich besonders für sechs Proteine große Probleme, den nativen Zustand richtig zuzuordnen, so dass bei diesen das native Protein auf Platz 600 bis 1000 eingeordnet wird (siehe z. B. 1aa3, 1bor, 1cc5, 1fbr, 1fwp, 1pft). Wie sich über die verschiedenen Iterationen zu sehen ist, zeigen insgesamt die Reihenfolgen der unterschiedlichen Proteine keinen eindeutigen Trend, so dass sich für einige Proteine der Rang der nativen Struktur zwar ändert, der allgemeine Trend innerhalb einer Reihenfolge aber erhalten bleibt.

Für einige Proteine wird die Zuordnung zu $n = 3$ besser (siehe z. B. 1ark, 1fbr, 1roo), während sie aber auch gleichzeitig für einige Proteine schlechter wird (siehe z. B. 1dec, 1gpt, 1leb, 1nxb, 1ptq). Die Proteine, deren native Strukturen in der Iteration Null bereits auf Platz 1 gesetzt waren, werden zumeist auch in den anderen Iterationen so bewertet, was darauf hindeutet, dass diese Proteine starke stabilisierende Wechselwirkungen enthalten, die bei der Parameteroptimierung eine entsprechende Gewichtung erhalten haben. Hierzu wurde untersucht, welche Potentiale maßgeblich zu der Stabilisierung beitragen und so die native Struktur von den falschen Strukturen unterscheidet. Ohne auf die Gesamtheit einzugehen sind als repräsentative Vertreter für die richtig erkannten Sequenzen die beiden Proteine 1aj3 und 1orc aufgeführt. Das erste Protein besteht aus drei zusammengelagerten α-Helices, während das zweite aus einer Kombination von α-Helices und β-Faltblättern aufgebaut ist. Als Vergleich zu diesen wurden weiterhin die sehr schlecht erkannten Proteine 1cc5 und 1fwp ausgewählt, die eine ähnliche Sekundärstruktur-Zusammensetzung wie 1aj3 und 1orc besitzen. Die Geometrien dieser Proteine zusammen mit beispielhaften falschen Strukturen aus dem Rosetta-Datensatz sind in Abb. 4.36 dargestellt.
Zu diesen vier Proteinen p wurden die Mittelwerte der Energiedifferenzen $\mu\left(\Delta E_{p,d}^{(k)}\right)$ der einzelnen Potentialklassen k berechnet über

$$\mu\left(\Delta E_{p,d}^{(k)}\right) = \frac{1}{\nu_{p,d}} \sum_{i=1}^{\nu_{p,d}} E_{p,d,i}^{(k)} - E_p^{*(k)} \tag{4.86}$$

wobei $E_{p,d,i}^{(k)}$ die Energie der i-ten falschen Struktur, entnommen aus dem Datensatz d (hier Rosetta), für den k-ten Potentialterm, $E_p^{*(k)}$ die Energie der nativen Struktur und $\nu_{p,d}$ die Anzahl an verwendeten falschen Strukturen ist. Die Mittelwertbildung erfolgte für die Proteine 1aj3 und 1orc über alle falschen Strukturen im Rosetta-Datensatz ($\nu_{\text{1aj3,Rosetta}} = \nu_{\text{1orc,Rosetta}} = 999$). Die Mittelwerte für 1cc5 und 1fwp wurden dagegen nur über die falschen Proteine berechnet, die energetisch niedriger als die native Struktur waren, um Mittelwerte darüber zu erhalten, wodurch die falschen Strukturen gegenüber der nativen Struktur stabilisiert wird, und um die Kompensation der Potentialterme mit energetisch höheren Strukturen zu verringern ($\nu_{\text{1cc5,Rosetta}} = 912$ und $\nu_{\text{1fwp,Rosetta}} = 964$). Die Ergebnisse sind in Tab. 4.23 dargestellt. Eine positive Energie bedeutet hierbei, dass die native Struktur gegenüber den falschen im Mittel um den entsprechenden Betrag energetisch günstiger ist.
Wie aus dieser Tabelle zu entnehmen ist, wird das Protein 1aj3 gegenüber den falschen Strukturen vornehmlich, wie auch oben für 1a92(A) beschrieben, durch die große Anzahl an Wasserstoffbrückenbindungen stabilisiert, während die anderen Mittelwerte um bis zu zwei Größenordnungen kleiner sind. Die Begründung hierfür ist entsprechend zu 1a92(A) durch die Struktur des Proteins 1aj3 gegeben, welches fast ausschließlich aus α-Helices aufgebaut ist, wodurch sehr viele Wasserstoffbrücken ausgebildet werden.

P	M	R_0	R_1	R_2	R_3	Z_0	Z_1	Z_2	Z_3	$\bar{A}_p$	w_α	w_β	w_t	F
1aa2	X	1	1	1	1	-3.0	-2.6	-1.7	-2.8	1.22	0.55	0.00	0.11	
1aa3	N	832	894	769	941	0.3	1.1	0.3	1.4	1.07	0.33	0.06	0.14	
1acf	X	1	1	1	1	-1.2	-1.8	-1.2	-1.8	1.41	0.32	0.33	0.14	
1ag2	N	1	1	1	1	-0.4	-0.5	-0.3	-0.3	1.20	0.52	0.04	0.12	
1aho	X	1	1	1	1	-5.7	-5.9	-6.2	-7.9	0.62	0.16	0.33	0.13	
1ail	X	1	1	1	1	-3.7	-5.1	-3.7	-6.4	1.07	0.84	0.00	0.04	
1aj3	N	1	1	1	1	-1.9	-2.5	-1.2	-2.7	1.04	0.89	0.00	0.05	
1ajj	X	403	188	316	268	-0.1	-0.5	-0.1	-0.2	0.59	0.00	0.11	0.27	C
1apf	N	1	1	1	3	-0.6	-0.9	-0.3	-0.3	1.17	0.00	0.12	0.20	
1ark	N	692	9	4	25	0.0	-1.1	-0.7	-0.8	1.12	0.00	0.30	0.13	
1ayj	N	1	1	1	1	-0.9	-1.6	-0.6	-0.8	1.18	0.22	0.27	0.00	
1bdo	X	1	1	1	1	-2.3	-3.1	-2.1	-3.4	1.26	0.00	0.45	0.15	BT
1bgk	N	1	1	1	1	-0.8	-1.2	-1.2	-1.8	1.06	0.49	0.00	0.22	
1bor	N	928	599	969	979	0.6	0.1	1.7	2.3	1.22	0.00	0.00	0.09	Z
1btb	N	3	124	15	19	-1.7	-0.9	-1.0	-1.4	1.39	0.42	0.18	0.20	
1c5a	N	7	4	10	28	-1.0	-1.8	-1.1	-1.3	1.06	0.71	0.00	0.06	
1cc5	X	936	503	835	917	0.4	-0.1	0.1	0.5	1.24	0.47	0.00	0.19	H
1cmr	N	1	1	1	1	-0.8	-1.6	-1.1	-1.2	0.99	0.00	0.13	0.16	
1csp	X	1	1	1	1	-2.0	-2.5	-1.5	-2.6	1.20	0.00	0.54	0.15	
1ctf	X	1	1	1	1	-4.7	-7.7	-5.0	-6.1	1.16	0.51	0.26	0.03	S
1ddf	N	1	13	1	4	-1.2	-1.4	-1.8	-2.0	0.69	0.49	0.00	0.16	
1dec	N	5	5	45	165	-1.4	-2.0	-1.0	-0.8	0.97	0.00	0.21	0.05	
1eca	X	1	1	1	1	-1.4	-1.5	-1.1	-1.8	1.08	0.71	0.00	0.13	H
1erd	N	1	1	1	1	-1.2	-2.0	-1.6	-1.9	0.86	0.45	0.00	0.15	
1fbr	N	859	786	791	261	0.3	0.2	0.0	-0.3	1.29	0.00	0.28	0.06	
1fwp	N	950	900	948	965	0.9	0.7	1.0	1.3	1.29	0.30	0.25	0.06	
1gpt	N	5	31	32	32	-0.6	-1.1	-0.9	-1.0	1.18	0.23	0.36	0.15	
1hev	N	1	1	1	1	-1.8	-2.9	-3.2	-2.7	0.84	0.09	0.09	0.07	
1hlb	X	1	1	1	1	-3.4	-3.1	-2.6	-3.3	0.93	0.65	0.00	0.13	H
1hsn	N	1	1	1	1	-2.6	-2.8	-2.3	-2.8	0.83	0.57	0.00	0.19	BM
1jvr	N	1	144	1	1	-1.8	-0.5	-1.3	-1.8	0.43	0.28	0.00	0.16	
1kte	X	1	1	1	1	-1.9	-4.4	-3.4	-4.6	1.15	0.48	0.17	0.10	
1leb	N	5	7	144	68	-0.8	-1.0	-0.4	-0.8	0.92	0.53	0.08	0.04	
1lfb	X	1	1	1	1	-1.7	-2.4	-2.1	-2.6	0.98	0.52	0.00	0.18	
1lis	X	1	1	1	1	-2.6	-2.3	-1.2	-2.5	0.91	0.67	0.00	0.03	
1mbd	X	1	1	1	1	-0.9	-1.3	-0.5	-0.8	1.15	0.74	0.00	0.08	H
1mzm	X	1	1	1	1	-1.8	-2.9	-2.6	-3.0	0.84	0.55	0.00	0.10	P
1nkl	N	1	1	1	1	-4.6	-4.6	-4.3	-5.0	1.07	0.73	0.00	0.06	
1nre	N	1	1	1	1	-3.9	-4.0	-5.0	-4.1	0.95	0.68	0.00	0.06	
1nxb	X	71	261	545	609	-0.5	-0.6	0.0	0.0	1.07	0.00	0.42	0.16	S
1orc	X	1	1	1	1	-2.7	-4.0	-2.7	-3.7	0.94	0.39	0.30	0.11	
1pal	X	1	1	1	1	-1.3	-2.3	-1.9	-2.3	1.15	0.50	0.04	0.11	C
1pce	N	1	1	1	1	-2.5	-2.8	-3.4	-3.9	0.57	0.18	0.15	0.15	
1pdo	X	1	1	1	1	-2.5	-3.1	-2.0	-3.1	1.15	0.47	0.15	0.12	

1pft	N	983	857	984	783	1.8	0.9	3.5	0.6	0.62	0.00	0.16	0.14	Z
1pgx	X	1	1	1	1	-1.5	-1.7	-0.9	-1.8	0.81	0.20	0.40	0.09	
1pou	N	5	4	3	3	-0.9	-1.6	-1.1	-1.6	1.15	0.68	0.00	0.17	
1ptq	X	2	5	1	29	-0.5	-0.8	-0.4	-0.4	1.00	0.08	0.20	0.16	Z
1qyp	N	130	608	116	146	-0.4	0.0	-0.3	-0.4	0.77	0.00	0.33	0.05	Z
1r69	X	1	1	1	1	-3.8	-3.4	-2.4	-3.4	1.12	0.63	0.00	0.14	
1roo	N	997	1	1	1	5.3	-2.5	-1.9	-2.2	0.82	0.31	0.00	0.23	
1uxd	N	1	1	1	1	-3.5	-3.2	-3.4	-3.2	0.71	0.44	0.00	0.19	

Tabelle 4.22: Übersicht über die Reihenfolge (*ranking*) der nativen und falschen Strukturen bezogen auf die Energie. Gegeben sind die Proteine P mit ihrer PDB-Kennung, M die Methode der Strukturbestimmung (X = Röntgenkristallographie und N = NMR-Spektroskopie), R_n der Rang der nativen Struktur in der Parameter-Iteration n, der Z-Wert Z_n der Verteilung (siehe Gl. 4.48), die relative Oberfläche $\bar{A}_p$ (siehe Gl. 4.84) und die Anteile der einzelnen Sekundärstrukturen mit w_α Anteil der α-Helix, w_β Anteil der β-Konformation und w_t Anteil der Schlaufen und Windungen (jeweils bestimmt mit dem DSSP-Programm). F gibt Fremdatome oder Moleküle in den jeweiligen nativen Strukturen mit Ausnahme von Wasser an. Es bedeutet: BM β-Mercaptoethanol, BT Biotin, C Calcium (Ca^{2+}), H Häm, P Palmitinsäure, S Sulfat-Anion (SO_4^{2-}) und Z Zink (Zn^{2+})

Für das Protein 1orc sind aufgrund der gemischten Sekundärstruktur drei Terme in Mittel in der gleichen Größenordnung, die zur Unterscheidung beitragen. Dies sind die nicht-bindenden Wechselwirkungen zwischen den C^α-Atomen und den Seitenketten sowie die Wasserstoffbrückenbindungen. Die Beiträge zu den nicht-bindenden Wechselwirkungen, sowohl zwischen den C^α-Atomen als auch zwischen den Seitenketten, werden nicht durch einzelne Paarwechselwirkungen bestimmt, sondern ergeben sich aus einer Summe vieler ähnlich großer Beiträge. Da die falschen Strukturen und die native Struktur zum einen gemessen an der relativen Oberfläche von $\bar{A}_{\mathrm{1orc,Rosetta}} = 0.94$ durchschnittlich eine vergleichbare Kompaktheit besitzen und zum anderen auch ähnliche Sekundär- und Tertiärstrukturen einnehmen (siehe Abb. 4.36), lässt sich die richtige Einordnung der nativen Struktur somit durch die Erkennung von wichtigen nativen Kontakten über die Potentialfunktion erklären.

Neben diesen richtig identifizierten Proteinen finden sich auch viele kaum erkannte native Strukturen im Rosetta-Satz, so dass für diese Proteine sehr viele falsche Strukturen als energetisch ärmer berechnet wurden. Als Beispiele hierfür wurden die Proteine 1cc5 und 1fwp ausgewählt und die Mittelwerte der Energiedifferenzen zu Strukturen berechnet, die energetisch günstiger als die native Struktur sind (siehe Tab. 4.23). Die Energieterme, die im wesentlichen dazu beitragen, dass die Rosetta-Strukturen für das Protein 1cc5 energetisch günstiger sind, sind die Nachwechselwirkungen, die nicht-bindenden Paarpotentiale zwischen den C^α-Atomen und die Wasserstoffbrückenbindungen. Für das Protein 1fwp stabilisieren

Protein p	$\mu(\Delta E_{p,d}^{(BO)})$	$\mu(\Delta E_{p,d}^{(NB)})$	$\mu(\Delta E_{p,d}^{(SU)})$	$\mu(\Delta E_{p,d}^{(SS)})$	$\mu(\Delta E_{p,d}^{(HB)})$
1aj3	$9.386{\cdot}10^{-5}$	$2.215{\cdot}10^{-6}$	$2.944{\cdot}10^{-5}$	$7.161{\cdot}10^{-5}$	$5.956{\cdot}10^{-4}$
1orc	$5.251{\cdot}10^{-5}$	$1.500{\cdot}10^{-4}$	$3.499{\cdot}10^{-5}$	$3.255{\cdot}10^{-4}$	$3.340{\cdot}10^{-4}$
1cc5	$-2.199{\cdot}10^{-4}$	$-4.083{\cdot}10^{-4}$	$2.837{\cdot}10^{-4}$	$2.129{\cdot}10^{-5}$	$-1.024{\cdot}10^{-4}$
1fwp	$-2.306{\cdot}10^{-5}$	$-5.577{\cdot}10^{-4}$	$4.077{\cdot}10^{-5}$	$1.630{\cdot}10^{-4}$	$-2.585{\cdot}10^{-5}$

Tabelle 4.23: Vergleich der über alle falschen Strukturen des Rosetta-Datensatzes (d) gemittelten Energiedifferenzen $\mu(\Delta E_{p,d}^{(k)})$ der Funktionsklasse k für das Protein zur nativen Struktur p (siehe Gl. 4.86). Die Potentiale k sind: BO Nahwechselwirkung, NB nicht-bindend, SU Oberfläche, SS Seitenketten und HB Wasserstoffbrücken.

die gleichen Terme die falschen Strukturen gegenüber der nativen, wobei für dieses Protein den nicht-bindenden Wechselwirkungen mehr Bedeutung zukommt. Betrachtet man für beide Proteine 1cc5 und 1fwp die Gesamtenergiebeträge der Nahwechselwirkungen und der Wasserstoffbrücken, so sind die Gesamtenergien jeweils eine Größenordnung größer als die zugehörigen Differenzbeträge aus Tab. 4.23, während der Differenzbetrag der nicht-bindenden Wechselwirkungen bezogen auf die beigetragene Gesamtenergie wesentlich größet ist. Dies zeigt, dass der Unterschied in den Termen Nahwechselwirkung und Wasserstoffbrücken im Vergleich weniger bedeutend ist. Dies ist darauf zurückzuführen, dass das native Protein und die falschen Strukturen ähnliche Geometrien besitzen, was auch aus Abb. 4.36 ersichtlich ist. Aufgrund der Gewichtung der Basisfunktionen resultiert der energetische Unterschied in diesen beiden Potentialtermen durch eine unterschiedliche Anzahl an Aminosäuren in der α-Helix-Konformation bzw. in der Anzahl an ausgebildeten Wasserstoffbrücken. Der größte Unterschied liegt demnach in den nicht-bindenden Wechselwirkungen. Vergleicht man hier die gesamte Energie des nativen Zustandes mit der gemittelten Gesamtenergie der falschen Strukturen, so sind diese Werte für das Protein 1cc5 $+3.260{\cdot}10^{-4}$ zu $-8.235{\cdot}10^{-5}$ und für 1fwp $+4.724{\cdot}10^{-4}$ zu $-8.539{\cdot}10^{-5}$. Die nativen Energien sind also stets positiv, während die Energien der falschen Strukturen negativ sind. Dieser Unterschied ist nicht darauf zurückzuführen, dass die Anordnung der C^{α}-Atome in den falschen Strukturen wesentlich günstiger wäre, sondern darauf, dass die nativen Strukturen einige wenige sehr ungünstige Wechselwirkungen enthalten, die die nicht-bindende Energie bestimmen. Für 1cc5 finden sich beispielsweise lediglich drei nicht-bindende C^{α}-Paare, deren Wechselwirkungsenergie zusammen $+4.413{\cdot}10^{-4}$ beiträgt. Beteiligt sind hierbei die Aminosäurepaare mit den Sequenzindices[3] (55,61), (7,76) und (19,28) in 1cc5. Diese drei Paarungen tragen 78 % der gesamten auftretenden positiven (repulsiven) Energie der nicht-bindenden Wechselwirkungen für 1cc5 bei. Für 1fwp sind es nur die zwei Aminosäurepaare (2,67) und (4,65), die zusammen 77 % der gesamten positiven

[3]Entsprechend der Nummerierung der Sequenz in der zugehörigen DSSP-Datei, die stets bei eins beginnt. Die Nummerierung in der PDB-Datei kann hiervon abweichen.

Protein	Paar	AS	m	r_0 / Å	±0.2Å	±0.4Å	±0.6Å	±0.8Å	±1.0Å
1cc5	(7,76)	Gly,Ala	287	3.51	0	2	3	8	15
1cc5	(55,61)	Gly,Pro	131	3.52	2	3	4	6	13
1cc5	(19,28)	His,Lys	31	4.39	3	5	6	6	6
1fwp	(2,67)	Arg,Thr	86	3.47	0	0	0	0	1
1fwp	(4,65)	Ile,Phe	120	3.88	0	0	2	4	7

Tabelle 4.24: Auftreten der Paarungen, die zur Destabilisierung der nativen Zustände der Proteine 1cc5 und 1fwp maßgeblich beitragen. "Paar" gibt die Indizes der Aminosäuren im Protein an, "AS" die Namen der Aminosäuren im 3-Buchstaben-Code, r_0 den Abstand des Paares im Protein und m die Gesamtanzahl des Auftretens des Paares im TOP500H-Satz (innerhalb von maximal 8 Å Distanz). Die Spalten $\pm x$Å geben die Anzahl des Auftreten des Paares im Abstandsintervall $r_0 \pm x$Å an.

nicht-bindenden Energie beitragen. Zu diesen fünf Paarungen gehören fünf unterschiedliche Potentialfunktionen, so dass dies nicht das Ergebnis einer ungünstig optimierten Linearkombination ist. Betrachtet man die funktionalen Verläufe, so entsprechen die Potentiale der Paarungen (2,67), (7,76) und (55,61) der Funktionsklasse wie in Abb. 4.32e dargestellt, wobei die Abstände, die diese Paarungen in den Proteinen besitzen jeweils sehr nahe dem dargestellten funktionalen Maximum sind, welches in Abb. 4.32e beispielhaft bei ca. 3.5 Å liegt. Das Potential zum Paar (4,65) entspricht dem rein repulsiven Potential Abb. 4.32a, und das Paar (19,28) entspricht Abb. 4.32c, wobei bei diesem C^{α}-Paar der Abstand so kurz ist, dass es im repulsiven Bereich der Potentialkurve liegt. Bezogen auf Abb. 4.32c entspricht dies einem Abstand kleiner als 4 Å. Der Grund für die ungünstige Form dieser Potentiale, die einen wichtigen Einfluss auf die Energie der nativen Rosetta-Strukturen nehmen, liegt vermutlich in dem seltenen Auftreten der zugehörigen Aminosäurepaare in dem zur Parametrisierung genutzten TOP500H-Datensatz, der auf die ausgewählten 48 Proteine reduziert wurde. Nimmt man als Ausgangspunkt die Abstände r_0 der oben beschriebenen fünf Aminosäurepaare in den beiden Rosetta-Proteinen 1cc5 und 1fwp, die die ungünstigen Wechselwirkungen erzeugen und analysiert davon ausgehend, wie häufig diese Paarungen in einem Abstandsintervall bis zu einem Abstand von $r_0 \pm 1.0$ Å im TOP500H-Datensatz auftreten, so zeigt sich, dass die ungünstigen Paarungen sehr selten sind (siehe Tab. 4.24).

Für die Parametrisierung bedeutet dies, dass im entsprechenden Intervall wenige Informationen über diese Aminosäurepaare vorliegen. Dies kann entweder darauf zurückgeführt, dass diese Paarungen bei den entsprechenden Abständen tatsächlich eher ungünstig sind oder darauf, dass der Satz an ausgewählten Proteinen und die damit verbundene Informationsbasis zu klein ist und diese Paarungen und deren bevorzugte Abstände nicht repräsentativ wiedergibt.

Somit ist es wahrscheinlich, dass im Datensatz, der zur Parametrisierung verwendet wurde, diese Paarungen in den gegebenen Abständen in größerer Anzahl in den falschen Strukturen enthalten sind, während sie in den nativen Proteinen kaum vertreten sind, wodurch diese als "ungünstig" optimiert werden. Verallgemeinert man dies weiter, könnte ein einfacher Grund auch sein, dass diese Paarungen insgesamt zu selten auftreten, wodurch deren Beiträge zur Gesamtenergie und zur Unterscheidung der Strukturen für den Optimierungsstrukturdatensatz weniger ins Gewicht fallen als andere Wechselwirkungen, was dazu führen kann, dass die gewählte funktionale Form während des Optimierungsprozesses weniger relevant ist.

Entfernt man die fünf Aminosäurepaarungen aus Tab. 4.24 aus der Potentialberechnung, entweder für alle Proteine oder lediglich für die native Struktur, was keinen Einfluss auf das Ergebnis ausübt, so verbessern sich die nativen Zustände auf Rang 165 für 1fwp und Rang 425 für 1cc5 (unter Verwendung der Parameter der dritten Iteration). Für beide Proteine sind danach die resultierenden Energien des nativen Proteins und der falschen Strukturen sehr ähnlich. Die falschen Strukturen sind meist aufgrund einer höheren Anzahl an Wasserstoffbrückenbindungen und günstiger Nahwechselwirkungen durch eine größere Anzahl an Aminosäuren in einer α-Helix weiterhin energetisch niedriger.

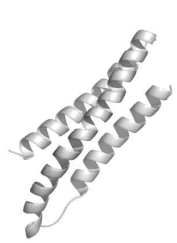
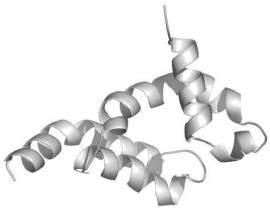
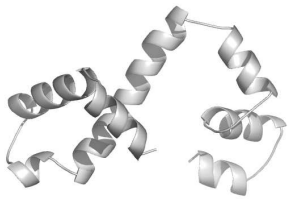
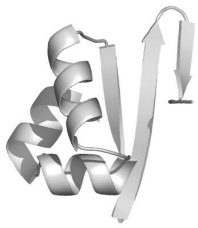
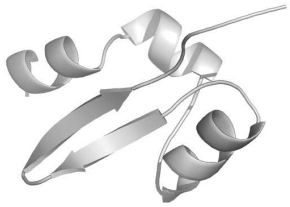
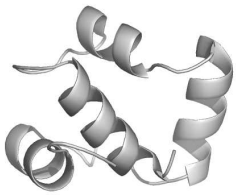
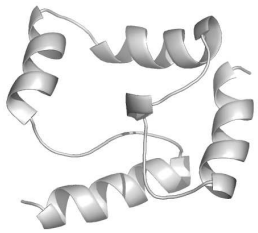
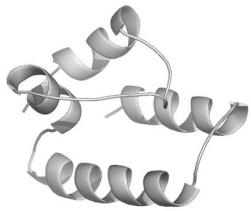
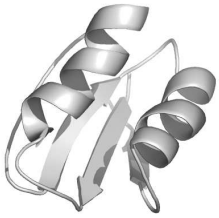
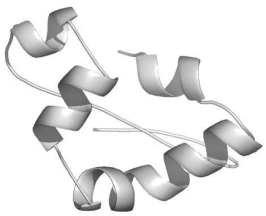
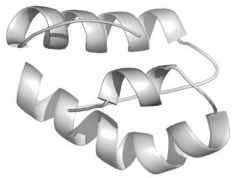
Protein	Native Struktur	Rosetta-Struktur 1	Rosetta-Struktur 2
1aj3		qa1aj3140-low	qa1aj3999-low
			
1orc		qa1orc165-low	qa1orc119-low
			
1cc5		qa1cc5030-low	qa1cc5326-low
			
1fwp		qa1fwp897-low	qa1fwp897-low
			

Abbildung 4.36: Vergleich der nativen Strukturen der Proteine 1aj3, 1orc, 1cc5 und 1fwp mit jeweils zwei falschen Strukturen des Rosetta-Datensatzes. Die Geometrien sind in der *Cartoon*-Form dargestellt. Die Zeichenkette über den Rosetta-Strukturen (z. B. qa1cc5326-low) gibt den Namen der verwendeten Datei bzw. Struktur an.

Rangfolge in anderen Datensätzen

Der Rosetta-Datensatz wurde zunächst zur Erstellung der Rangfolgeliste herangezogen, da dieser Struktursatz eine große Anzahl an unterschiedlichen Sequenzen und zusätzlich zu jeder Sequenz eine große Anzahl an falschen Strukturen bereitstellt. Anhand dieses Datensatzes wurde an Beispielen gezeigt, welche Wechselwirkungen für eine erfolgreiche bzw. nicht erfolgreiche Identifizierung des nativen Zustandes wichtig sind. Für einen direkten Vergleich mit anderen publizierten Potentialen wurden die Datensätze Lmds, Lattice_ssfit und 4state_reduced herangezogen (siehe Tab. 4.6), zu denen verschiedene Vergleichsergebnisse vorhanden waren.
Die Bewertung dieser Struktur-Datensätze erfolgte ebenfalls auf Basis der Energiereihenfolge, wobei zum Vergleich wieder zusätzlich zugehörigen die Z-Werte berechnet wurden. Der Vergleich der Ergebnisse, die in der Tab. 4.25 zusammengefasst sind, erfolgt mit drei in 2007 publizierten Bewertungsfunktionen. Diese basieren alle auf stark vergröberten Proteinmodellen, für die zum Teil nur bis zu nur einem Wechselwirkungszentrum pro Aminosäure angesetzt wurde. Bei den ersten beiden Potentialen (siehe [120, 127] für weitere Details) handelt es sich um statistische Potentiale, die für die Parameter-Bestimmung die quasi-chemische Näherung verwenden (siehe Abschnitt 4.5.1) und zu Standard-Kraftfeldern ähnliche Terme wie beispielsweise nicht-bindende Paarwechselwirkungen oder Winkel- und Torsionswinkelterme beinhalten. Das dritte Vergleichspotential [242] enthält lediglich eine stark vereinfachte Näherung der Solvatationsenergie bzw. der hydrophoben Wechselwirkung als alleinige Funktion zur Unterscheidung der Strukturen. Insgesamt sind diese drei Potentiale in ihrer Komplexität der Energiefunktionen und des Proteinmodells vergleichbar mit dem in dieser Arbeit verwendeten Ansatz.

Die Ergebnisse (siehe Tab. 4.25) zeigen, dass die drei Literaturpotentiale sehr effektiv in der Identifizierung des nativen Zustandes der drei Struktursätze sind. Das in dieser Arbeit erstellte Potential ordnet für 9 von insgesamt 24 Proteinen (bzw. 38 %) den nativen Zustand richtig zu und wertet fünf weitere Proteine (somit insgesamt 58 %) unter den zehn besten Strukturen. Die höchste Erkennungsrate wird hierbei für den Lattice_ssfit-Satz erreicht. Insgesamt ist die Verteilung der Ränge für das in dieser Arbeit erstellte Potential gemischter als im Rosetta-Satz, in welchem die Proteine hauptsächlich entweder besonders gut oder besonders schlecht erkannt wurden. Bei einigen Proteinen, die in diesen drei Sätzen nicht richtig erkannt wurden, zeigt sich, dass auch die anderen Potentiale teilweise Probleme in der richtigen Identifizierung besitzen (siehe z. B. 1bba, 1fc2, 2ovo oder 4rxn), wobei das in dieser Arbeit verwendete Potential einige dieser Proteine besser erkennt als einige Vergleichspotentiale (1fc2, 4rxn). Die Z-Werte der Literaturpotentiale [120] und [242] liegen gemittelt bei -3.7 bzw. -3.1, wobei für [120] das Protein 1bba, dessen Verteilung einen Z-Wert von 21.4

besitzt, in der Berechnung nicht mitberücksichtigt wurde. (Die Z-Werte für [127] sind in Tab. 4.25 nicht aufgeführt, da der Literaturquelle keine eindeutigen nach Proteinen aufgeschlüsselten Werte entnommen werden konnte.) Die Z-Werte der Verteilungen zu dem in dieser Arbeit verwendeten Potential liegen im Bereich von -0.5 bis -2.2 mit einem Mittelwert von -1.0. Vergleicht man die Identifizierung der nativen Zustände zusammen mit den erhaltenen Z-Werten zu den anderen Potentialen, so zeigt sich, dass das Potential in der Lage ist, für mehr als die Hälfte der Proteine eine gute Voraussage für den nativen Zustand zu treffen, dass aber die anderen Potentiale eine höhere Erfolgsquote erbringen. Die erhaltenen Z-Werte zeigen negative Werte, die indizieren, dass die Zuordnung des nativen Zustandes durch Informationen in der Potentialfunktion erreicht wurde und dass eine zufällige Zuordnung mit einer geringen Wahrscheinlichkeit zu dem gleichen Ergebnis gekommen wäre. Allerdings sind die negativen Z-Werte noch nicht so eindeutig wie in den anderen Potentialen, die teilweise sehr gute Ergebnisse mit Z-Werten bis zu -6 oder -7 erreichen.

Betrachtet man die Energiebeiträge des Potentials der schlecht zugeordneten Proteine, so zeigt sich, dass es, anders als oben für 1cc5 und 1fwp beschrieben, keinen dominanten Term gibt, der den nativen Zustand gegenüber den falschen Strukturen destabilisiert. Die niedrigere Energie der falschen Strukturen ergibt sich zumeist aus einer Summe mehrerer Terme, die jeder für sich genommen lediglich geringfügig energetisch günstiger für die falsche als für die native Struktur sind.
Das Protein 1bba (aus Lmds), das mit dem hier vorgestellten Potential sowie von anderen Literaturpotentialen sehr schlecht erkannt wurde, besitzt einige ungewöhnliche verzerrte C^{α}-Bindungswinkel, die dazu führen, dass es zwei sehr kurze C^{α}_{i}-C^{α}_{i+2}-Abstände enthält, in denen diese C^{α}-Atome lediglich ca. 4.1 Å voneinander entfernt sind. Die mittels des TOP500H-Struktursatzes bestimmte Untergrenze (Minimalabstand) für eine $(i, i+2)$-Nahwechselwirkung betrug dagegen ca. 5.0 Å (vergleiche hierzu auch Abb. 4.19). Dementsprechend erhielt das Nahwechselpotential für solch kurze Abstände einen stark repulsiven Anteil, was in diesem Fall zu einer Destabilisierung des nativen Zustandes führte. Analog zu dem vorher beschriebenen Vorgehen wurden die entsprechenden Nahwechselterme zu den sehr kurzen C^{α}-Abständen von der Energieberechnung ausgeschlossen, was in diesem Fall dazu führte, dass die Energiedifferenz zwischen dem nativen Protein und den falschen Strukturen sehr viel kleiner wurde. Die Gesamtenergie für die Nahwechselwirkungsterme änderte sich von einem stark destabilisierenden Werte von $+5.905 \cdot 10^{-4}$ auf $-3.365 \cdot 10^{-4}$, wodurch sich die Gesamtenergie des nativen Proteins ebenfalls von $+2.758 \cdot 10^{-4}$ auf $-6.513 \cdot 10^{-4}$ verringerte. Die Gesamtenergie der falschen Proteine betrug aber im Mittel ca. $-1.065 \cdot 10^{-3}$, wodurch sich trotz der Verbesserung der Rang des nativen Proteins nicht änderte.
Die native Struktur des ebenfalls schlecht erkannten Proteins 4rxn (aus 4state_reduced) enthielt eine C^{α}-C^{α}-Bindungslücke zwischen den Resten 52 und 53, da deren Abstand in der

Datenbank-Datei ca. 5.3 Å beträgt. Daher wurden bei der Energieberechnung für alle zu dieser Sequenz gehörenden Proteine die letzten beiden Reste (Glu_{53} und Glu_{54}) nicht mitberücksichtigt. Die PDB-Datei zu 4rxn zeigt auch für die Seitenketten dieser beiden Aminosäuren ungewöhnliche Geometrien, was eventuell auf Probleme in der Strukturaufklärung des Proteins bei diesen Aminosäuren zurückzuführen ist.

Fast man die Ergebnisse aller Rangfolgebestimmungen zusammen, so zeigt sich, dass die Erfolgsquote für eine Identifizierung des nativen Zustandes für das in dieser Arbeit verwendete Potential bei ca. 60 % liegt, was für die erste Generation eines völlig neu entwickelten Potentials ein sehr guter Startwert ist. Ebenso weisen die Z-Werte in die Richtung einer eindeutigen Zuordnung des nativen Proteins über die Erkennung von wichtigen nativen Kontakten aufgrund der funktionalen Form des Potentials und der optimierten Parameter. Dennoch ist das Identifizierungsgesamtergebnis (noch) nicht auf der Höhe der Literaturpotentiale, wobei aber klare Hinweise dahingehend bestehen, wie das Potential zu verbessern ist, um die Erfolgsquote weiter zu steigern. Dies betrifft zum einen die Nahwechselwirkungen, die auf $(i, i+2)$-Abständen beruhen, weil diese häufig die native Struktur destabilisieren, wenn das Protein eine größere Anzahl an Aminosäuren in der β-Faltblatt-Konformation enthält. Falsche Strukturen sind häufig dadurch energetisch günstiger, dass sie einen größeren Anteil an α-Helices enthalten. Dies ist ein Problem der funktionalen Form der zugehörigen Basisfunktion, die für den entsprechenden Bereich der β-Konformation angepasst werden muss. Des weiteren sind einige nicht-bindende Wechselwirkungen für die nativen Proteine ungünstig, was, wie oben gezeigt wurde, auf eine zu kleine Informationsbasis für die entsprechenden Aminosäuren bei den ungünstigen Abständen zurückzuführen ist. Dies kann dadurch verbessert, indem der Satz an nativen Proteinen verändert bzw. vergrößert wird, wobei diesen kritischen Aminosäurepaaren mehr Rechnung getragen wird. Da vor der Optimierung praktisch keine Annahmen über die funktionale Form der nicht-bindenden Wechselwirkungen gemacht werden, ist es unwahrscheinlich, dass bei diesen eine andere mathematische Form des Potentials eine Verbesserung erbringt.

Satz	Protein	D	R_3	R[127]	R[120]	R[242]	Z_3	Z[120]	Z[242]
Lmds									
	1bba	(500)	501	-	501	-	30.8	21.4	-
	1ctf	(497)	1	1	1	1	-1.8	-3.4	-3.4
	1dtk	(215)	210	6	2	-	3.3	-2.5	-
	1fc2	(500)	251	359	53	409	-0.1	-1.3	0.9
	1igd	(500)	3	1	1	1	-0.7	-4.0	-2.8
	1shf	(437)	11	1	1	1	-0.6	-5.3	-2.9
	2cro	(500)	5	2	1	1	-2.2	-7.7	-3.4
	2ovo	(347)	123	61	1	16	-0.2	-3.2	-1.7
	4pti	(343)	80	-	1	6	-0.6	-3.5	-2.2
Lattice_ssfit									
	1beo	(1997)	1	-	1	-	-1.1	-5.6	-
	1ctf	(1998)	3	1	1	1	-1.1	-6.0	-5.0
	1dkt	(1994)	1	1	1	8	-0.4	-3.1	-2.7
	1fca	(2000)	72	1	1	1	-0.7	-4.7	-7.4
	1nkl	(1995)	4	1	1	1	-1.1	-4.1	-4.5
	1pgb	(1998)	1	1	1	1	-1.1	-4.7	-4.0
	1trl	(1998)	1	1	1	101	-0.8	-3.6	-1.6
	4icb	(1997)	1	-	1	1	-2.1	-4.4	-3.5
4state_reduced									
	1ctf	(631)	1	1	1	1	-1.0	-3.4	-3.5
	1r69	(676)	1	1	1	1	-1.4	-4.0	-3.7
	2sn3	(660)	6	1	1	1	-1.2	-3.6	-2.5
	2cro	(674)	1	3	1	2	-1.2	-3.2	-3.0
	3icb	(654)	117	-	1	1	-0.9	-2.9	-2.2
	4pti	(687)	59	1	1	5	-0.8	-3.1	-2.5
	4rxn	(677)	247	1	667	4	55.3	2.5	-2.7

Tabelle 4.25: Vergleich des Potentials in der Parameter-Iteration $n = 3$ mit Literaturergebnissen. Gegeben sind die Anzahl an verwendeten falschen Strukturen D, der Rang der nativen Struktur R_n basierend auf dieser Arbeit, die Ränge R[X] entnommen der Publikation X, der Z-Wert Z_n basierend auf dieser Arbeit und die Z-Werte der entsprechenden Publikationen.

4.7 Globale Geometrieoptimierung

4.7.1 Einleitung

In vielen wissenschaftlichen oder ökonomisch orientierten Bereichen treten Probleme bzw. Fragestellungen auf, deren möglicher Lösungsraum aufgrund einer großen Anzahl an Variablen oder Parametern sehr groß ist, wodurch die im Regelfall interessierenden charakteristischen Punkte einer mathematischen Funktion, die das gegebene Problem beschreibt, nur mit aufwendigeren Methoden zu bestimmen sind.
In der Chemie, sowohl in der Grundlagenforschung als auch an der Schnittstelle zu Anwendung, beziehen sich solche Optimierungsprobleme häufig auf die Vorhersage oder Verbesserung bzw. problemspezifische Anpassung von Molekülstrukturen oder Molekülkomplexen, was sowohl zum Vergleich mit dem Experiment wie auch zur Analyse experimentell schwer zugänglicher Systeme genutzt werden kann. Hierzu wären zum Beispiel die Bestimmung von Clusterstrukturen [243, 244], die Kristallstrukturvorhersage [245], das *Docking*, bei welchem die ideale Geometrie eines Substrates oder einer Bindungstasche gesucht wird, und eben die auch in dieser Arbeit behandelte Proteinfaltung zu nennen. Abhängig von der gewählten Fragestellung oder von den modellierten Versuchsbedingungen ist hierbei häufig das Auffinden der Struktur mit der minimalen Energie das Ziel der Analyse, welche dem globalen Minimum der Energiehyperfläche entspricht, die entweder z. B. durch die rein internen Wechselwirkungen oder durch die freie Energie des Gesamtsystems mit Umgebung gegeben sein kann. Da aber reale Experimente nicht am absoluten Nullpunkt der Temperaturskala durchgeführt werden können, sind im Regelfall auch die oberhalb des globalen Minimums liegenden energetisch höheren Niveaus statistisch besetzt bzw. populiert, wodurch auch diese lokalen Minima der Energiehyperfläche für das Ergebnis relevant sein können [246, 247]. Ebenso können aber auch in einer Simulation der Dynamik von chemischen Reaktionen die lokalen Maxima eine wesentliche Rolle spielen, da diese beim Übergang von einem Minimum in ein anderes den Übergangszuständen zwischen diesen entsprechen, deren energetische Niveaus direkten Einfluss auf den Verlauf und die Geschwindigkeit der Reaktion nehmen.
Diese Bedeutung der unterschiedlichen charakteristischen Punkte auf der Energiehyperfläche lässt sich auch direkt auf die Proteinforschung übertragen. Wie in Abschnitt 3.2 beschrieben wurde, falten Proteine in ihrer natürlichen Umgebung spontan oder mit Unterstützung von Hilfsmolekülen in ihren nativen Zustand, der in der Regel die biologisch aktive Form ist. Dieser Zustand entspricht dem globalen Minimum der freien Energie des Proteins und seiner Umgebung, wodurch dieser Punkt der freien Energiefläche eine besondere Bedeutung zukommt, da die entsprechende Geometrie des Proteins zur Bestimmung und Analyse der Funktion und der Reaktion viele Hinweise geben kann und für eine weitergehende Simula-

tion ein vernünftiger Startpunkt darstellt. Andererseits verändern bestimmte Proteine bei Ausführung der Funktion ihre dreidimensionale Gestalt, beispielsweise beim Vorgang des Membrantransportes anderer Moleküle. Diese Konformationsänderungen entsprechen einem Pfad auf der Energiefläche des Proteins entlang einer Reihe lokaler Minima und Sattelpunkte bis zur Vollendung der Reaktion, beispielsweise durch Verschieben von Domänen. Nach Beendigung des Prozesses kehrt das Protein in seinen Ausgangszustand zurück, welcher wiederum im Normalfall dem globalen Energieminimum entspricht. Eine ähnliche Situation liegt bei einer Enzym-Reaktion vor, bei der ein Substrat an das aktive Zentrum eines Proteins bindet (*docking*). Hier finden ebenfalls Geometrieänderungen statt, die aber wesentlich kleiner ausfallen als vergleichsweise beim Transmembrantransport, da sich hierbei im wesentlichen die Strukturänderungen auf die Seitenkettenorientierungen nahe an bzw. in der Bindungsstelle beschränken, welche anteilig nur einen kleinen Teil des gesamten Proteins ausmachen, während der Rest des Proteins nahezu unverändert bleibt.
Diese Beispiele zeigen, dass die Suche nach dem globalen Minimum der freien Energie, welche sich, wie in Abschnitt 4.4 dargelegt wurde, auf eine Suche nach dem globalen Minimum der internen Wechselwirkungen beschränken lässt, für eine Simulation der Proteinreaktion ein geeigneter Startpunkt ist, da dies sehr häufig die biologisch aktive Form ist bzw. die Reaktion bei diesem startet.

Die Schwierigkeit der Vorhersage des globalen Minimums in der Proteinfaltung besteht in der hochdimensionalen Komplexität des Problems. Die Energiefläche eines Proteins enthält in der Regel viele ausgedehnte zugängliche Bereiche, besonders in den entfalteten Geometrien, in denen die Aminosäuren relativ weit voneinander entfernt sind, während in den hochgeordneten kompakten Zuständen große Bereiche nicht mehr zugänglich sind, und die erlaubten Geometrien auf schmale Bereiche begrenzt werden, da es selbst bei einer kleiner Geometrieverzerrung schnell zu einer Atomkollision mit sehr hoher potentieller Energie kommen kann. Andererseits behalten aber auch beispielsweise einige Seitenketten eine große Rotationsfreiheit im gefalteten Zustand. Diese Eigenschaften der Protein-Energiefläche erschweren die Suche nach den charakteristischen Punkten. Hier kommt weiterhin hinzu, dass die Fläche eine sehr große Anzahl an lokalen Minima enthält. In der Literatur wird deren Anzahl, die von der Länge Sequenz des Proteins abhängt, mit verschiedenen Größenordnungen abgeschätzt. Der größte Wert wird hier mit 10^N lokalen Minima für N Aminosäuren angesetzt [248], was zeigt, dass selbst wenn man diesen Wert als eine obere Grenze hierfür annimmt und kleinere Werte als zehn verwendet, dass schon bei einer durchschnittlichen Größe des Proteins von 100 bis 200 Aminosäuren die Anzahl der lokalen Minima sehr groß ist. Diese Anzahl an Punkten, unter denen das globale Minimum zu finden ist, ist zu groß, um sie mit den heutigen verfügbaren Computer-Kapazitäten in vernünftiger Zeit berechnen zu können. Dies wird in der Literatur auch als "kombinatorische Explosion" bezeichnet [184]. Aufgrund dieser Komplexi-

tät und dem damit verbundenen Zeitproblem ist es nicht möglich, den Konformationsraum eines mittelgroßen Proteins bzw. die zugehörige Energiehyperfläche mit deterministischen Verfahren durch eine inkrementelle systematische Variation aller Freiheitsgrade zu berechnen (siehe hierzu auch das Beispiel unten).
Aus diesem Grund werden in der Praxis für derartige Probleme andere Methoden zur Durchsuchung des Variablenraumes verwendet, die eher statistischen Ansätzen entsprechen. Für die Proteinfaltung haben hier die Moleküldynamik (MD), *simulated annealing* und Monte-Carlo-Methoden (MC) besondere Bedeutung und Verbreitung, auf welche aber im Rahmen dieser Arbeit nicht näher eingegangen wird. Angewendet wurden statt diesen sogenannte genetische Algorithmen, welche sich in ihrer Methodik wie auch in ihrer Sprache an der Biologie orientieren, indem sie Prinzipien der Evolution in systematischer Form in einen Optimierungsalgorithmus umsetzen. Der Ansatz dieser Methode ist sehr allgemein und kann für viele verschiedene Fragestellungen angewendet werden, so dass dieser nicht auf Probleme der theoretischen Chemie beschränkt ist. In der Praxis werden genetische Algorithmen beispielsweise zur Konstruktion und dem Design von Maschinen verwendet oder bei der Optimierung von Transportwegen, Prozessführungen und Produktionsabläufen zur Kostenreduktion.
Im Folgenden werden die allgemeinen Grundlagen und Eigenschaften eines genetischen Algorithmus näher erläutert, wobei auf eine ausfürhliche Diskussion der Detailfragen wie beispielsweise zum Hintergrund und zur Implementation an dieser Stelle verzichtet wird. Hier wird auf die entsprechende Literatur verwiesen [249–255]

Genetische Algorithmen basieren auf einer zufallsgesteuerten bzw. statistischen Variablenvariation, was zur Folge hat, dass die Ergebnisse und deren Qualität nicht vorhersagbar sind, sowie dass die Ergebnisse selbst bei exakt gleichen Anfangsbedingungen nicht zwangsweise reproduzierbar sind. Durch das Wahrscheinlichkeitsmoment führen die gleichen Rechnungen im allgemeinen zu unterschiedlichen Ergebnissen, außer wenn ein gut konvergierendes System vorliegt oder die Simulationsdauer sehr lang wird. Schließlich kann ebenso keine Garantie dafür gegeben werden, dass das erhaltene Ergebnis die beste oder überhaupt eine gute Lösung des Problems darstellt. Eine Einstufung oder Beurteilung des Ergebnisses ist nur dann möglich, wenn das beste Ergebnis bereits vorher bekannt wäre, was beispielsweise in Testfällen durch einen Vergleich mit deterministischen Methoden und durch die allgemeine Kenntnis über Eigenschaften des Problems erreicht wird. In dem in der Praxis aber in der Regel eintretenden Fall ist dieses nicht bekannt, so dass die Bestimmung der optimalen Lösung oder einer, der dieser nahe kommt, nur durch den Vergleich mehrerer unabhängiger Rechnungen möglich ist.

Beispiel kombinatorische Explosion

Um die Unmöglichkeit der Berechnung aller lokalen Minima zu zeigen, sei hier eine Beispielrechnung für die Parallelschaltung von m Prozessoren gegeben, die jeweils 10^k lokale Minima pro Sekunde berechnen können. Für ein Protein mit $N = 100$ Aminosäuren und 10^N lokalen Minima würde die Zeit t zur Berechnung aller Minima

$$t = \frac{10^N}{m\,10^k s^{-1}} = \frac{10^{N-k}}{m} s = \frac{10^{100-k}}{m} s \tag{4.87}$$

benötigt. Als Abschätzung mit $k = 9$ für einen 1 GHz-getakteten Prozessor und $m = 1000$ parallelen Prozessoren ergibt sich so

$$t = \frac{10^{100-9}}{10^3} s = 10^{88} s \approx 10^{80}\,\text{Jahre} \tag{4.88}$$

Ein genetischer Algorithmus arbeitet mit einem Satz aus möglichen Lösungsansätzen für ein Problem. Für die Geometrieoptimierung von Molekülen sind diese Lösungsansätze unterschiedliche geometrische Anordnungen. Der gesamte Satz an Lösungsansätzen wird als Population und die einzelnen Teillösungen als Individuen bezeichnet.
Im Optimierungsprozess werden die Individuen oft codiert dargestellt, wobei die codierte Form als Genotyp bezeichnet wird. Dieses Verfahren wird eingesetzt, da es die Anwendung der genetischen Operatoren vereinfachen kann, welche den Genotyp und somit die Individuen verändern, um so den Variablenraum abzusuchen. Die Codierung kann und sollte so gewählt werden, dass sie dem Problem angepasst ist und einen eindeutigen Zusammenhang mit der vollständigen Darstellung des Individuums liefert. Die nicht-codierte Darstellung wird als Phänotyp bezeichnet und kann für Moleküle beispielsweise die Darstellung in kartesischen Koordinaten sein. Für Proteine könnte eine Codierung zum Beispiel auf einer Diskretisierung des Konformationsraumes basieren, bei welcher jedem diskreten Zustand eine (ganze) Zahl zugeordnet wird.
Die Individuen werden iterativ verändert, wobei eine Iteration als Generation bezeichnet wird. Die erste Generation bzw. Generation Null werden in der Regel vollkommen zufällig generiert. Auf die Generation Null und auf die folgenden Generationen wirken dann die genetischen Operatoren, die die Individuen in bestimmter Weise verändern und ihre Qualität bestimmen. Die Individuen am Anfang einer Generation werden als Eltern, die Produkte, die aus deren Veränderungen hervorgehen, als Kinder bezeichnet.

Im Allgemeinen werden bei der Implementation eines GA mehrere Standard-Operatoren verwendet. Dies sind Operatoren zur Variation der Individuen und zur Bewertung und Selektion dieser für einen Folge-Zyklus. Welche Operatoren dies sind und wie diese im Programm umgesetzt wurden wird in den nächsten Abschnitten beschrieben (siehe 4.7.2).

Diese Operatoren sollen eine Evolution der Lösungen simulieren, so daß diese sich weiterentwickeln können und nur die besten Individuen überleben, welche im Rahmen der Population die besten Lösungen zu einem Problem liefern.
Eine wichtige Eigenschaft eines genetischen Algorithmus ist, dass die Operatoren so konzipiert sind, dass die Optimierung auch größere Barrieren zwischen unterschiedlichen Bereichen überwinden kann, um so nach Möglichkeit ein Festsitzen in einem lokalen Minimum zu vermeiden. Dies ist bzw. kann mit großen Sprüngen im Variablenraum verbunden sein. Trotz dieser Eigenschaft ist es nicht garantiert, dass die so erhaltenen Lösungen ausgezeichneten Punkten auf der Fläche entsprechen, da es in einem großen Konfigurationsraum sehr unwahrscheinlich ist, dass direkt eine Minimumsposition getroffen wird. Aus diesem Grund wird ein genetischer Algorithmus häufig mit einer lokalen Gradientenoptimierung verbunden (Hybridmethode), die direkt aus der Form der Fläche ihre Informationen für die Schritte der Optimierung gewinnt und so ein Individuum zum nächsten Minimum optimieren kann. Durch diese Kombination kann erreicht werden, dass zwar große Sprünge im Variablenraum vollzogen werden, die erreichten Strukturen aber stets Minima der Fläche entsprechen.
Das Ziel hierbei ist es nun, durch die genetischen Operatoren, die die in den Genotypen enthaltenen Informationen über die lokalen Minima auszutauschen und zu kombinieren, so dass im Idealfall die weiteren erzeugten Individuen energetisch tiefer liegenden Minima entsprechen, um sich so über eine Serie von Sprüngen entlang lokaler Minima zum globalen Minimum zu bewegen.

4.7.2 Der genetische Algorithmus

Der für diese Arbeit implementierte genetische Algorithmus basiert in seiner prinzipiellen Form auf dem bereits vorher publizierten PROFET-Algorithmus (*Protein Folding with Evolutionary Techniques*) [160]. Dieser ist dahingehend konzipiert, ein Protein in einer vollständigen atomaren Darstellung zu behandeln und dieses auf einer durch etablierte Potentiale, wie beispielsweise das ECEPP2/3- [221–224] oder das FLEX-Kraftfeld [256, 257], gegebenen Energiefläche global zu optimieren, wobei die internen Koordinaten Bindungslängen, -winkel und -torsionswinkel der Rückgratatome und die Torsionswinkel der Seitenketten als die suchraumaufspannenden Variablen verwendet werden. Das Protein wurde in diesem Programm ohne eine explizite Umgebung im Vakuum bei Null Kelvin angenommen, wobei allerdings die Möglichkeit besteht, Lösungsmitteleffekte mit einer abstandsabhängigen Dielektrizitätskonstanten zu approximieren.
Da der Programmquellcode des PROFET-Algorithmus auf die oben beschriebene Proteindarstellung optimiert war, wurde der genetische Algorithmus größtenteils vollständig neu programmiert und nur einige bestimmte Unterroutinen des originären Programms wieder verwendet, um diesen dem in dieser Arbeit gewählten Proteinmodell anzupassen. Hierzu wurde, wie in folgenden Abschnitten ausgeführt wird, der generelle Programmablauf beibehalten, dieser aber um einige zusätzliche Aspekte erweitert.
Die Basis des genetischen Algorithmus bildete das zur Kraftfeld-Parametrisierung angesetzte Modell, welches auf einer C^{α}-Rückgrat-Darstellung beruhte, die durch die Zentren zur Simulation der Wasserstoffbrücken erweitert wurden sowie die eingeführten Seitenkettenschwerpunkte, die durch den Cluster-Ansatz beschrieben werden. Die Variablen für den Konformationsraum eines Proteins sind die Längen der (Pseudo-) Bindungen zwischen den C^{α}-Atomen, daneben die durch drei sequentielle C^{α}-Atome gegebenen Winkel und die durch vier sequentielle C^{α}-Atome gegebenen Torsionswinkel sowie die Positionen der Seitenketten-Schwerpunkte.
In den folgenden Abschnitten wird zunächst der grundsätzliche Ablauf des für diese Arbeit programmierten genetischen Algorithmus ausgeführt, worauf im Anschluss die wichtigsten Programmabschnitte im Detail erläutert werden.

Programmablauf

Das Programm lässt sich grundsätzlich in zwei hierarchische Ebenen einteilen. Die zu Grunde liegende Ebene wird durch den aus dem PROFET-Algorithmus übernommenen Teil des Programmes gebildet, der unter Verwendung evolutionärer Methoden den Konformationsraumes einer Population an Proteinen nach Minima der vorgegebenen (parametrisierten) Bewertungsfunktion absucht. Dieser Ebene übergeordnet steht der Repliken-Algorithmus, der

verschiedene dieser Populationen parallel behandeln kann und wahlweise einen Informationsaustausch zwischen den einzelnen Repliken ermöglicht. Ebenso können unterschiedliche Proteine bearbeitet werden. In der jetzigen Form des Programmes müssen jedoch alle Programmabschnitte sequentiell abgearbeitet bearbeitet werden, da noch keine Parallelisierung des Quellcodes erfolgt ist. Der generelle Verlauf des Programmes für ein Protein ist in Abb. 4.37 dargestellt und wird detailliert im folgenden zunächst für nur eine Replik beschriebenen. Das Programm wurde auf Basis des FORTRAN90/95-Standards programmiert. Die zur Ausführung benötigten Variablen und Felder werden im Programm dynamisch erstellt, so dass die maximale Problemgröße von den zur Verfügung stehenden Computerressourcen beschränkt wird. Das Programm enthält keine expliziten Parallelisierungen.

Die Basis des genetischen Algorithmus ist die iterative Veränderung einer Population bzw. einer Menge an Proteinstrukturen, welche im übergeordneten Rahmen hier auch als Replik bezeichnet wird, wenn davon mehrere behandelt werden. Zu Beginn steht die Generierung der Anfangsstrukturen der Population mit der Anzahl *MaxPop*. Diese werden beim N-Terminus beginnend aminosäureweise aufgebaut, wobei zuerst des nächste C^{α}-Atom mit einem zufällig gewählten Bindungswinkel und zufällig gewähltem Torsionswinkel an die Kette angehängt wird. Der Abstand zweier sequentieller C^{α}-Atome wurde zunächst mit 3.8 Å angesetzt, da dies der dominierende Abstand in bekannten experimentellen Strukturen ist (siehe hierzu auch Abschnitt 4.4.1). Konnte dieses C^{α}-Atom kollisionsfrei platziert werden, wurde die Seitenkette des vorherigen C^{α}-Atoms gesetzt und auf Kollisionen getestet. Im Falle von Kollisionen wurden die Positionen des zuletzt angehängten C^{α}-Atoms bzw. der zuletzt gesetzten Seitenkette variiert (siehe für Details weiter unten im Abschnitt Kollisionstests). Bei nicht behebbaren Fehlern wurde die Struktur verworfen. Nach der Vollendung der Generierung einer Zufallsanfangsstruktur wird diese lokal optimiert (siehe Abschnitt Evolutionsoperatoren: Optimierung). Alle erfolgreichen Proteine werden am Ende der Routine gesammelt und nach ihrer potentiellen Energie geordnet, bis deren Anzahl *MaxPop* betrug und somit die Population vollständig war. Dieser Satz an Proteinen wird dann an das Hauptprogramm übergeben, in dem sich die Entwicklung mittels evolutionären Methoden anschließt. Diese läuft über insgesamt *MaxGen* Generationen. Bei dem Erreichen dieser Zahl an Iterationen, wird das Programm beendet.
Für die Evolution der Strukturen werden aus den *MaxPop*-Proteinen alle möglichen Paarungen gebildet, wobei jede Paarung nur einmal vorkommt. Die Paarbildung war unabhängig von der Reihenfolge der Proteine, so dass das Proteinpaar mit den Ordnungsnummern k, j gleich dem Paar j, k ist, wobei $k, j \in \{1, 2, ..., MaxPop\}$. Dies führt zu insgesamt $0.5 MaxPop\,(MaxPop - 1)$ Paarungen. Zunächst wird anhand der voreingestellten Wahrscheinlichkeit *XProb* mit $0 \leq XProb \leq 1$ entschieden, ob eine Kreuzung zwischen den gepaarten Proteinen stattfindet. Für den Austausch an Informationen über eine Kreuzung wurden im Pro-

gramm zwei unterschiedliche Methoden realisiert (siehe hierzu Evolutionsoperatoren: Kreuzung). Unabhängig von der Entscheidung über eine Kreuzung, wird im folgenden mit einer Wahrscheinlichkeit von *MutProb* mit $0 \leq MutProb \leq 1$ für jedes Protein einzeln eine Mutation durchgeführt. Hierfür stehen wiederum zwei unterschiedliche Verfahren zur Verfügung (siehe hierzu Evolutionsoperatoren: Mutation). Wenn entweder durch eine Kreuzung oder durch eine Mutation die Proteinstruktur verändert worden ist, wird diese nach Abschluss aller Veränderungen unter Verwendung des Kraftfeldes lokal optimiert (siehe hierzu Evolutionsoperatoren: Optimierung).
Die optimierten Strukturen werden, ebenso wie die Paare oder Einzelproteine, auf welche weder der Kreuzungs- noch der Mutationsoperator angewendet wurde, zu einer Gesamtmenge zusammengefasst, bis dieser Prozess für alle Paarungen beendet ist. Hierauf folgend werden aus der Gesamtmenge die Proteine für die Folgegeneration selektiert, wobei sie dann auch als Eltern bezeichnet werden. Für die Selektion können verschiedene Kriterien angewendet werden (siehe hierzu Evolutionsoperatoren: Selektion).
Dieser Ablauf von Veränderungen der Proteinstruktur durch Kreuzung und Mutation mit anschließender lokaler Optimierung und Selektion wird solange wiederholt, bis die voreingestellte Anzahl an Generationen durchlaufen wurde. Nach deren Erreichen wird das Programm beendet.

Dieser Einzelevolution einer Proteinmenge übergeordnet ist es möglich, mehrere Populationen parallel zu behandeln, wobei diese dann auch als Repliken bezeichnet werden. Die Parallelität ist hierbei nicht auf die programmtechnische Realisation bezogen, sondern darauf, dass die Populationen unabhängig voneinander den Evolutionszyklus des genetischen Algorithmus durchlaufen. Ohne einen Informationsaustausch zwischen den Repliken entspricht das Programm in dieser Form zunächst nichts anderem als unabhängigen parallelen Einzelläufen des Programmes mit jeweils nur einer Replik bzw. einer Population. Das Ziel bzw. der Vorteil des Repliken-Algorithmus besteht darin, die Informationen, die jede Replik über die zu Grunde liegende Energiefläche mit sich trägt, zu einem oder mehreren Zeitpunkten auszutauschen, um durch die Kombination der Daten eine effizientere Suche des Konformationsraumes zu ermöglichen, indem Repliken Zugang zu strukturellen Informationen bekommen, die in der Entwicklung der Einzelreplik bisher nicht enthalten waren. Diese auch als Repliken-Austausch bekannte Methode wird in dieser oder anderen problemorientierten Realisierung auf unterschiedliche Fragestellungen angewendet, wie unter anderem auch in der Proteinfaltung (siehe hierzu bspw. [258, 259]).
Wie oben bereits erwähnt ist die Implementation des Repliken-Austausches seriell, so dass alle Repliken generationenweise nacheinander bearbeitet werden (siehe auch Abb. 4.37).
Jede Replik startet mit unterschiedlichen Initialisierungsdaten für den Zufallszahlen-Generator, wodurch sie verschiedene Bereiche des Konformationsraumes erfassen, da sich dadurch im

Regelfall sowohl die Startstrukturen wie auch deren weitere Evolution auf der gegebenen Energiefläche unterscheiden. Sofern der Repliken-Algorithmus nicht bereits zu einem lokalen oder globalen Minimum konvergiert ist, enthalten die Repliken somit unterschiedliche Informationen. Der Zeitpunkt bzw. die Generation, zu dem der Austausch zwischen den Repliken stattfindet (*RepExGen*), wird zu Anfang Programmes festgelegt. Die Möglichkeiten erstrecken sich hier von maximal einem Austausch nach jeder beendeten Generation, über eine Austausch zu einigen bestimmten Generationen bis hin zu keinem Austausch. Erreicht das Programm im Verlauf eine dieser Generationen, so werden, die Proteine zwischen den einzelnen Repliken ausgetauscht. Diese geschieht dergestalt, dass zunächst alle Proteine aus allen Repliken zu einer neuen (ungeordneten) Gesamtmenge zusammengefasst werden, aus welcher dann mittels Zufallsprinzip die neuen Repliken zusammengestellt werden. Die Zuordnung der Proteine zu den Repliken erfolgt also ohne weitere Kriterien wie bspw. energetische oder geometrische Ähnlichkeiten. Nach Beendigung der Neuverteilung, durchlaufen die einzelnen Repliken wieder unabhängig voneinander den Evolutionsalgorithmus, bis eine weitere Austausch-Generation oder das Programmende erreicht wird.
Im ungünstigsten Fall kann es dazu kommen, dass eine energetisch tief liegende nicht-native Struktur nach Vermischung der Repliken die unterschiedlichen Population dominiert, wodurch der Algorithmus strukturell und energetisch schnell konvergiert und einem lokalen Minimum gefangen bleibt.

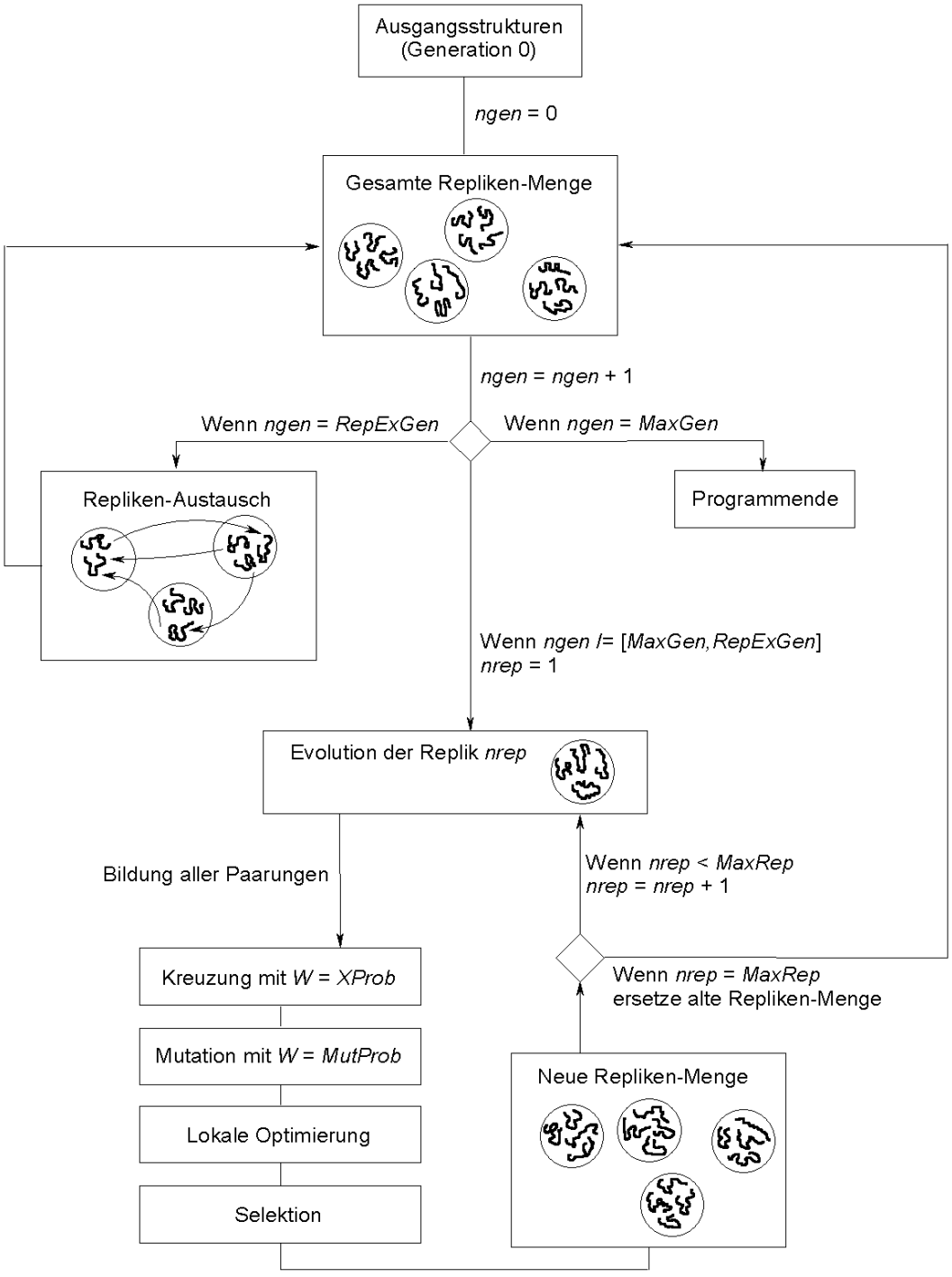


Abbildung 4.37: Verlauf des genetischen Algorithmus mit Repliken-Austausch. Es bedeuten $ngen$ die Nummer der Generation, $MaxGen$ die maximale Anzahl an Generationen, die durchlaufen wird, $MaxRep$ die Zahl an verwendeten Repliken, $nrep$ die Nummer der Replik, $RepExGen$ die Generation ("Zeitpunkt") des Austausches zwischen den Repliken, W eine Wahrscheinlichkeit, $XProb$ die Wahrscheinlichkeit für eine Kreuzung und $MutProb$ die Wahrscheinlichkeit für eine Mutation.

4.7.3 Programmteile

In den folgenden Abschnitten werden einige wichtige Programmteile eingehender dargestellt. Diese sind:

- Kollisionstests
- Kreuzungsoperator
- Mutationsoperator
- lokale Optimierung
- Selektion

Kollisionstests

Für der Veränderung einer Proteinstruktur bietet die Verwendung von internen Koordinaten einige Vorteile. Aus diesem Grund wird diese Darstellung der Proteingeometrie in verschiedenen Programmteilen angewendet. Nachteilig hierbei ist allerdings, dass durch eine Änderung der internen Koordinaten die resultierende Proteingeometrie Kollisionen enthalten kann, die erst nach der Umwandlung in kartesische Koordinaten festgestellt wird. Eine Kollision, welche einen unnatürlich kurzen Abstand beschreibt, der in nativen Strukturen nicht auftritt, ist aus unterschiedlichen Gründen problematisch. Aus einer exakten, fundamentalen Sichtweise heraus, ist eine Kollision zu vermeiden, da sie unphysikalisch und nur ein Effekt des genäherten Modells bzw. der unzulänglichen Methode ist. Dagegen kann allerdings aus rein praktischer Sicht heraus bzw. im Licht des alleinigen Ziels, den RMSD-Wert zur nativen Struktur zu minimieren, eine Kollision weniger ins Gewicht fallen, wenn ansonsten die Zielgeometrie gut reproduziert wird. (Dies kann beispielsweise Proteingeometrien vorliegen, die auf NMR-Datenbasis bestimmt wurden.) Neben diesen Aspekten kann sich eine Kollision in einem unerwarteten Sinne negativ auf Bewegungen auf einer Potentialenergiefläche während der Suche nach charakteristischen Punkten auswirken, wie sie hier für die globale Optimierung notwendig ist. Dies ist ein Effekt der Methode der Parameteroptimierung und des funktionalen Ansatzes des Potentials: Die verwendeten falschen Strukturen wurden so konstruiert, dass sie keine Kollisionen enthielten, um das Informationsgewicht auf die anderen Wechselwirkungen zu legen. Damit fehlen aber in der Parameteroptimierung die Informationen über diesen Bereich der Energiefläche. Dies kann, wie im Abschnitt der Ergebnisse der Parameteroptimierung gezeigt, dazu führen, dass beispielsweise nicht-bindende Potentiale für zu kurze Abstände stark attraktiv wirken, was konträr zur physikalischen Realität steht, in der diese Potentiale stark repulsiv wirken. Werden somit bei sehr kurzen Abstände dort attraktive Funktionen nicht korrigiert, verfälscht dies die Gesamtenergie, indem eine unrealistische Stabilisierungsenergie

hinzugefügt wird. Dies kann in Abhängigkeit des funktionalen Verlaufes der entsprechenden nicht-bindenden Funktion soweit gehen, dass die Gesamtenergie durch einen einzigen zu kurzen Abstand dominiert wird, wodurch die globale Optimierung in diesem lokalen Minimum gefangen bleibt, welches ein reines Artefakt der Methode ist.
Aus diesem Grund wird im Programmablauf an bestimmten Stellen auf Kollisionen getestet, um realistische Proteinstrukturen zu erhalten. Als Minimalabstände zwischen den Seitenketten, den Seitenketten und dem Rückgrat und zwischen den Rückgratatomen wurden die auf Basis des TOP500H-Proteindatensatzes ausgewerteten kürzesten Abstände verwendet, welche im Anhang aufgeführt sind (siehe Tab. 6.1 bis 6.3). Diese Tests werden bspw. nach Anwendung eines Mutations- oder Kreuzungsoperators durchgeführt.
Für den gegebenen Fall einer Kollision wurden zunächst Routinen geschrieben, die das Proteinrückgrat und die Seitenkettenpositionen verändern, um nach Möglichkeit eine kollisionsfreie Anordnung zu finden. Hierzu wurde zunächst der Teil der Sequenz des Proteins beginnend am N-Terminus vollständig aufgebaut bzw. unverändert verwendet, der keine Kollision enthielt. Das restliche Protein wurde anschließend aminosäureweise angehängt, und bei jedem neuen Rest auf eine Kollision getestet. Lag eine solche vor, wurde das letzte angehängte C^{α}-Atom auf einer Kugeloberfläche rotiert, deren Zentrum das vorhergehende C^{α}-Atom war, was einer Änderung des Bindungswinkels und des Torsionswinkels entsprach, wobei aber der Bindungsabstand bzw. der Kugelradius konstant gehalten wurde. Gleichzeitig mit der Veränderung des letzten C^{α}-Atoms wurde auch die Position der Seitenkette der vorletzten Aminosäure jeweils neu berechnet, da diese nicht unabhängig von der Position des zuletzt angehängten C^{α}-Atoms ist. Die abgesuchten Positionen auf der Kugeloberfläche entsprachen inkrementellen Änderungen des zugehörigen Torsionswinkels um jeweils fünf Grad und des zugehörigen Bindungswinkels um zwei Grad. Sofern mehrere erlaubte Positionen erhalten wurden, wurde die neue Position des letzten C^{α}-Atoms nach drei einstellbaren Kriterien gewählt: a) Dichteste Position zu Ursprünglichen, b) kompakteste Gesamtgeometrie oder c) zufällige Auswahl einer Position aus allen erlaubten. Wenn es von vornherein nur zu einer Kollision verursacht durch die Seitenkette der Aminosäuren kam, dann wurden zunächst auch nur die anderen Seitenkettenpositionen in Form der Clusterzentren getestet. Konnte hierbei keine erlaubte Seitenketten-Anordnung gefunden werden, wurde anschließend dann auch die Position des folgenden C^{α}-Atoms verändert.
Lagen viele Kollisionen im Protein vor, so war diese Art der Problembehebung relativ zeitaufwendig im Vergleich zu anderen Programmteilen. Daher wurde der Schritt der Kollisionsbehebung in einer späteren Programmversion in die lokale Optimierung miteinbezogen und das separate Testen von erlaubten Positionen weggelassen. Statt dessen wurden zu dem ursprünglichen Kraftfeld Kollisionspotentiale $V^{(k)}$ hinzugefügt, die den minimalen Abstand

zwischen zwei Interaktionspunkten gewährleisten sollten. Diese hatten die generelle Form

$$V^{(k)}(r_{i,j}) = \begin{cases} c^{(k)}_{s_i,s_j}(r_{i,j} - \delta_{s_i,s_j})^2 & \text{für } r_{i,j} < \delta_{s_i,s_j} \\ 0 \end{cases} \tag{4.89}$$

Der Abstand $r_{i,j}$ kann hierbei die Abstände Seitenkette-zu-Seitenkette, Seitenkette-zu-C^α-Atom oder C^α-zu-C^α-Atom repräsentieren. Die Konstante δ_{s_i,s_j} ist der Minimalabstand der zwei Interaktionszentren, wie er aus den TOP500H-Proteindaten erhalten wurde und $c^{(k)}_{s_i,s_j}$ ist der Gewichtungskoeffizient dieser Funktion im Gesamtpotential, welcher manuell so eingestellt wurde, dass diese Funktion im Falle eines zu kurzen Abstands die entsprechende nicht-bindende Funktion überwog. Zusätzlich wurden für die nicht-bindenden Funktionen eine Dämpfungsfunktionen eingeführt, um unrealistische, bei kleinen Abständen stark attraktive Potentiale besser handhabbar zu machen. Die Dämpfung, welche Funktionswerte zwischen Null und eins annehmen kann, wurde innerhalb eines bestimmten Intervalls mit den Funktionswerten der nicht-bindenden Funktion multipliziert. Wie in Abschnitt 4.4.3 erläutert wird, wurde für die nicht-bindenden Potentialfunktionen eine untere Grenze (unteres Limit) l_1 definiert, unterhalb derer sie nicht mehr definiert sind. Die aus den Proteindaten gewonnen Minimalabstände lagen stets oberhalb dieser Funktionsgrenze, so dass stets $\delta_{s_i,s_j} > l_1$ gilt. Für den Bereich zwischen diesen beiden Abständen wurde die Dämpfungsfunktion η definiert zu:

$$\eta(r_{i,j}) = \begin{cases} 1 & \text{für } r_{i,j} \geq \delta_{s_i,s_j} \\ \frac{1}{2}\sin\left(\frac{\pi(r_{i,j}-\delta_{s_i,s_j})}{l_1-\delta_{s_i,s_j}} - \frac{\pi}{2}\right) + \frac{1}{2} & \text{für } l_1 < r_{i,j} < \delta_{s_i,s_j} \\ 0 & \text{für } r_{i,j} \leq l_1 \end{cases} \tag{4.90}$$

Diese Funktion besteht im Intervall $l_1 < r_{i,j} < \delta_{s_i,s_j}$ aus einer Halbwelle der Sinus-Funktion, die für $r_{i,j}$ gegen l_1 gegen Null geht, wodurch das nicht-bindende Potential verschwindet, während sie für $r_{i,j}$ gegen δ_{s_i,s_j} gegen eins geht, so dass sie für $r_{i,j} \geq \delta_{s_i,s_j}$ die nicht-bindenden Potentiale nicht verändert.

Evolutionsoperatoren: Kreuzung

Der erste Schritt bei der Kombination zweier Eltern bestand in der Verwendung eines Kreuzungsoperators (*crossover*) zwischen beiden Eltern. Für diesen Operator wurden zwei unterschiedliche Varianten implementiert. Die erste Version einer Kreuzung bestand aus einem direkten Austausch interner Koordinaten zwischen zwei Eltern. Diese internen Koordinaten umfassten die C^α-Abstände, -Pseudo-Bindungswinkel und -Pseudo-Torsionswinkel. Nicht enthalten waren die Seitenkettenkoordinaten (wie für diese verfahren wurde siehe Evolutionsoperatoren: Optimierung). Bei Ausführung dieses Kreuzungsoperators bestand entweder die

Wahl, die internen Koordinaten vollständig ab einem bestimmten Punkt entlang der Sequenz auszutauschen (Einzelpunkt-Kreuzung) oder die internen Koordinaten zwischen zwei Punkten in der Sequenz auszutauschen (Mehrpunkt-Kreuzung), wobei die erste Möglichkeit als Spezialfall der zweiten aufgefasst werden kann, indem einer der beiden Punkte entweder mit der ersten oder der letzten Aminosäure zusammenfällt. Der Vorgang der Mehrpunkt-Kreuzung wird in folgender Gleichung 4.92 veranschaulicht, wobei "×" in dieser für den Kreuzungsoperator steht: Gegeben sei ein Satz an Proteinen P_k mit der Ordnungsnummer k für das Protein und deren Darstellung in einer Matrix der internen Koordinaten I, die Spaltenweise die C^α-Abstände $d_{k,i}$ der Atome i und $i+1$ enthält, sowie die C^α-Winkel $\kappa_{k,i}$ und C^α-Torsionswinkel $\tau_{k,i}$. Die Darstellung eines Proteins P_k in internen Koordinaten $I(P_k)$ ist somit:

$$I(P_k) = \begin{pmatrix} r_{k,1} & \kappa_{k,1} & \tau_{k,1} \\ r_{k,2} & \kappa_{k,2} & \tau_{k,2} \\ & \dots & \\ r_{k,N-1} & \kappa_{k,N-1} & \tau_{k,N-1} \end{pmatrix} \tag{4.91}$$

Der Kreuzungsoperator tauscht die internen Koordinaten zwischen zwei solchen Matrizen zeilenweise, beispielsweise zwischen den Positionen j_1 und j_2 in der Sequenz. Die Anwendung des Kreuzungsoperators führt dann zu:

$$\begin{aligned} I(P_1) \times I(P_2) &= \begin{pmatrix} r_{1,1} & \kappa_{1,1} & \tau_{1,1} \\ r_{1,2} & \kappa_{1,2} & \tau_{1,2} \\ & \dots & \\ r_{1,j_1} & \kappa_{1,j_1} & \tau_{1,j_1} \\ & \dots & \\ r_{1,j_2} & \kappa_{1,j_2} & \tau_{1,j_2} \\ & \dots & \\ r_{1,N-1} & \kappa_{1,N-1} & \tau_{1,N-1} \end{pmatrix} \times \begin{pmatrix} r_{2,1} & \kappa_{2,1} & \tau_{2,1} \\ r_{2,2} & \kappa_{2,2} & \tau_{2,2} \\ & \dots & \\ r_{2,j_1} & \kappa_{2,j_1} & \tau_{2,j_1} \\ & \dots & \\ r_{2,j_2} & \kappa_{2,j_2} & \tau_{2,j_2} \\ & \dots & \\ r_{2,N-1} & \kappa_{2,N-1} & \tau_{2,N-1} \end{pmatrix} = \\ &= \begin{pmatrix} r_{1,1} & \kappa_{1,1} & \tau_{1,1} \\ r_{1,2} & \kappa_{1,2} & \tau_{1,2} \\ & \dots & \\ r_{2,j_1} & \kappa_{2,j_1} & \tau_{2,j_1} \\ & \dots & \\ r_{2,j_2} & \kappa_{2,j_2} & \tau_{2,j_2} \\ & \dots & \\ r_{1,N-1} & \kappa_{1,N-1} & \tau_{1,N-1} \end{pmatrix} + \begin{pmatrix} r_{2,1} & \kappa_{2,1} & \tau_{2,1} \\ r_{2,2} & \kappa_{2,2} & \tau_{2,2} \\ & \dots & \\ r_{1,j_1} & \kappa_{1,j_1} & \tau_{1,j_1} \\ & \dots & \\ r_{1,j_2} & \kappa_{1,2,j_2} & \tau_{1,j_2} \\ & \dots & \\ r_{2,N-1} & \kappa_{2,N-1} & \tau_{2,N-1} \end{pmatrix} = I(P_1') + I(P_2') \end{aligned} \tag{4.92}$$

Im Falle des Einzelpunkt-Operators wäre $j_1 = 1$ oder $j_2 = N$. Durch diese Kreuzung werde komplette Abschnitte unter vollständiger Erhaltung der Struktur des getauschten Abschnittes übertragen.

Neben dem Operator zum Austausch der internen Koordinaten wurden zusätzlich noch der kartesische Kreuzungsoperator implementiert [260]. Dieser Operator wirkt direkt auf die kartesischen Koordinaten zweier Proteine, so dass hier eine Umwandlung in interne Koordinaten entfällt. Zunächst werden die zwei Proteine optimal durch eine Minimierung des cRMSD-Wertes überlagert. Dann werden die kartesischen Koordinaten der beiden Ausgangsproteine $\mathbf{X_1}$ und $\mathbf{X_2}$ linear kombiniert, so dass ein neues Protein $\mathbf{X_3}$ entsteht:

$$\mathbf{X_3} = u \cdot (\mathbf{X_2} - \mathbf{X_1}) + \mathbf{X_1} \tag{4.93}$$

Hierbei ist u ein Gewichtungsfaktor mit $0 \leq u \leq 1$. Für $u = 0$ wird $\mathbf{X_1}$ erhalten, für $u = 1$ wird $\mathbf{X_2}$ erhalten. Die Werte zwischen Null und eins entsprechen Mischgeometrien beider Proteine, wobei die anteilige Information der ursprünglichen Proteine vom Wert von u abhängt. Dieser Operator lässt sich so auffassen, dass ein Ausgangsprotein unter Verwendung eines anderen Proteins in Richtung der Koordinaten des anderen Proteins verzerrt wird. Für die Wahl für u werden in [260] zwei verschiedene Mengen an Werten angegeben bzw. vorgeschlagen: Zum einen eine Menge, bei der das resultierende Protein nah an der Ursprungsgeometrie eines der Ausgangsproteine bleibt. Bei dieser ist $u \in \{0.05, 0.1, 0.15, 0.85, 0.90, 0.95\}$. Damit wird der Konformationsraum dicht bei $\mathbf{X_1}$ bzw. $\mathbf{X_2}$ erreicht. In der zweiten Menge werden mehr Strukturinformationen übertragen, wobei eher der Konformationsraum zwischen beiden Ausgangsproteinen erreicht wird. Hierbei ist $u \in \{0.1, 0.2, 0.35, 0.65, 0.8, 0.9\}$. Durch den Benutzer kann festgelegt werden, welche der beiden Mengen im Programm verwendet werden soll. Bei dem Aufruf des kartesischen Kreuzungsoperators werden stets alle Proteine zu den gewählten u-Werten gebildet, so dass im Unterschied zum Kreuzungsoperator in internen Koordinaten, bei dem aus zwei Ausgangsproteinen zwei neue Proteine entstehen, hier sechs neue Proteine gebildet werden. Die weiteren Unterschiede zwischen beiden Operatoren bestehen darin, dass nach einem Tausch der internen Koordinaten im Regelfall komplett neue Proteine entstehen, da sich durch die Änderung der internen Koordinaten, die Sekundärstruktur und damit auch größtenteils die gesamte Tertiärstruktur durch eine Neuarrangierung der Proteinsegmente verändern kann. Dies entspricht großen Sprüngen im Konformationsraum des Proteins. Der kartesische Operator verändert in dieser Hinsicht die Gesamtgeometrie weit weniger, denn es werden alle Punkte lediglich in Richtung der anderen Koordinaten verschoben, wodurch keine vollständige Neufaltung entsteht. Der große Nachteil bei dieser Methode ist, dass meist stark verzerrte Strukturen mit unnatürlichen Bindungslängen oder nicht-bindenden Abständen entstehen, die korrigiert werden müssen.
Dieser Operator eignet sich dafür, ihn in im fortgeschrittenen Stadium einer globalen Opti-

mierung einzusetzen, wenn die Strukturen bereits nahe eines Minimums sind, da der Kreuzungsoperator wie oben beschrieben die Strukturen meist zu stark verändert, wodurch die resultierende Struktur wieder weit entfernt vom Minimum ist. Der kartesische Operator ist dagegen besser geeignet, die lokale Umgebung des Proteins zu erreichen.

Evolutionsoperatoren: Mutation

Neben den Kreuzungsoperatoren wurden auch zwei verschiedene Mutationsoperatoren implementiert, die sich ähnlich zu den beiden Kreuzungsoperatoren unterschiedlich stark auf die Veränderung der Proteinstruktur auswirken.
Der erste Mutationsoperator $\hat{\mathbf{M}}_1$ wirkt auf die internen Koordinaten des Proteins und verändert diese innerhalb eines bestimmten Segmentes. Innerhalb dieses werden die Bindungslängen, -winkel und -torsionswinkel mit neuen Werten überschrieben. Schematisch lässt sich dieser Vorgang wie folgt darstellen, wobei die Mutation auf die Positionen zwischen den Resten j_1 und j_2 wirkt:

$$\hat{\mathbf{M}}_1 I(P_k) = \hat{\mathbf{M}}_1 \begin{pmatrix} r_{k,1} & \kappa_{k,1} & \tau_{k,1} \\ r_{k,2} & \kappa_{k,2} & \tau_{k,2} \\ & \dots & \\ r_{k,j_1} & \kappa_{k,j_1} & \tau_{k,j_1} \\ & \dots & \\ r_{k,j_2} & \kappa_{k,j_2} & \tau_{k,j_2} \\ & \dots & \\ r_{k,N-1} & \kappa_{k,N-1} & \tau_{k,N-1} \end{pmatrix} = \begin{pmatrix} r_{2,1} & \kappa_{2,1} & \tau_{k,1} \\ r_{2,2} & \kappa_{2,2} & \tau_{k,2} \\ & \dots & \\ r'_{k,j_1} & \kappa'_{k,j_1} & \tau'_{k,j_1} \\ & \dots & \\ r'_{k,j_2} & \kappa'_{k,j_2} & \tau'_{k,j_2} \\ & \dots & \\ r_{k,N-1} & \kappa_{k,N-1} & \tau_{k,N-1} \end{pmatrix} \tag{4.94}$$

Die neuen Werte $r'_{k,i}$, $\kappa'_{k,i}$ und $\tau'_{k,i}$ werden dabei entweder zufällig generiert oder aus einer Liste vorgegebener Werte ausgewählt. Dies wurde im Programm so realisiert, dass zunächst die Art der Mutation ausgewählt wurde, wobei zwischen drei Typen gewählt werden konnte. Diese waren die α-Helix, das β-Faltblatt oder eine Zufallsgeometrie (*random coil*). Für die ersten beiden Typen wurden für den Bindungswinkel κ' und den Torsionswinkel τ' feste idealisierte Literatur-Mittelwerte $\bar{\kappa}$ und $\bar{\tau}$ verwendet, zu denen zur Variation eine kleine Abweichung $\Delta\bar{\kappa}$ und $\Delta\bar{\tau}$ hinzuaddiert wurde, wobei $0 \leq \Delta\bar{\kappa} \leq \Delta\bar{\kappa}_{max}$ und $0 \leq \Delta\bar{\tau} \leq \Delta\bar{\tau}_{max}$ ist. Die hierfür verwendeten Werte ebenso wie die minimale und maximale Länge L_{min} und L_{max} eines mutierten Segmentes sind in Tab. 4.26 aufgelistet. Für eine Mutation mit Zufallsgeometrie wurden die Winkel und die Torsionswinkel für das ausgesuchte Segment aus den in der Tabelle gegebenen Intervallen zufällig gewählt. Nach der Veränderung der Matrix der internen Koordinaten wurde diese in die kartesischen Koordinaten umgerechnet und auf Kollisionen getestet sowie die Seitenketten gesetzt.

Struktur	$\bar{\kappa}/^\circ$	$\Delta\bar{\kappa}_{max}/^\circ$	$\bar{\tau}/^\circ$	$\Delta\bar{\tau}_{max}/^\circ$	L_{min}	L_{max}
α-Helix	88	10	130	15	1	10
β-Faltblatt	130	10	178	15	1	4
Zufallsgeometrie	[60,170]	-	[-178,178]	-	1	5

Tabelle 4.26: Vorgegebene Werte für Mutationen der internen Koordinaten. τ ist der Torsionswinkel für vier sequentielle C^α-Atome und κ der Bindungswinkel zwischen drei sequentiellen C^α-Atomen.

Der oben beschriebene Mutationsoperator in internen Koordinaten hat ähnlich wie der Kreuzungsoperator in internen Koordinaten den Effekt, dass kleinste Änderungen der internen Koordinaten große Auswirkungen auf die dreidimensionale Struktur des Proteins haben können. Dies eignet sich gerade zu Beginn einer Optimierung, um größere Bereiche des Konformationsraumes zu erfassen. Am Ende einer Optimierung nahe eines Minimums zeigt sich jedoch zumeist, dass durch diese Art der Mutation die Veränderungen zu groß sind und sich die neue Struktur weitab vom erreichten Minimum in Bereichen mit hoher potentieller Energie befindet. Dies führt dazu, dass viele Strukturen wieder verworfen werden müssen, weil nur Strukturen mit niedrigerer Gesamtenergie weiterverwendet werden (in einer Monte-Carlo-Simulation beispielsweise hätten diese Strukturen dagegen eine Überlebenswahrscheinlichkeit proportional zum Boltzmannfaktor). Dieser Mutationsoperator verliert dadurch in den späten Stadien der Optimierung an Effektivität. Daher wurde ein zweiter Mutationsoperator implementiert, der weniger drastische Veränderungen an der Gesamtstruktur eines Proteins vornimmt. Dieser führt nur lokal zu einer Modifikation des Rückgrates ohne die relative Anordnung des restlichen, nicht mutierten Proteins zu verändern, weshalb er auch als lokaler Twist-Operator (LT-Operator) bezeichnet wird [141, 261].
Die Mutation mittels des LT-Operators wird in verschiedenen Schritten durchgeführt, welche sowohl interne wie kartesische Koordinaten benötigen, wobei schließlich aber die finale Mutation des Proteins in kartesischen Koordinaten stattfindet. Zunächst wird ein Segment mit einer maximalen Länge von zehn Aminosäuren zur Mutation ausgewählt. Dieses Segment wird analog zum vorher beschriebenen Mutationsoperator neu gefaltet, wobei die drei oben beschriebenen Mutationstypen in internen Koordinaten angewendet werden. Anschließend werden die kartesischen Koordinaten des Segmentes berechnet. Die neuen Koordinaten des veränderten Segments werden nun mit seinen Koordinaten vor der Mutation unter Minimierung des cRMSD-Wertes übereinandergelegt, so dass das Segment eine zur ursprünglichen möglichst ähnliche Orientierung erlangt, um Kollisionen zu vermeiden. Nach der Überlagerung wird das gesamte Segment so verschoben, dass die Position des ersten C^α-Atoms mit seiner Position vor der Mutation zusammenfällt. Da nur in Ausnahmefällen die überlagerten Segmente eine gleiche oder sehr ähnliche Länge besitzen, tritt im Regelfall eine Lücke

am Ende des Segmentes zum benachbarten C^{α}-Atom des unveränderten Proteinrestes auf. Diese wird unter Verwendung des FCCD-Algorithmus geschlossen [161, 162], welcher nochmal die Bindungs- und Torsionswinkel des mutierten Segmentes verändert. Wenn es hierbei nach dem Schließen der Lücke und dem Setzen der Seitenketten zu Kollisionen kam, wurde das Segment neu gefaltet und wieder eingesetzt. Dieser Zyklus wurde bis maximal 50 mal wiederholt. Wurde innerhalb dieser Anzahl keine erlaubte Anordnung gefunden, wurde die Mutation verworfen.
Dieser Operator zeigt eine große Erfolgsquote, was die Erzeugung von kollisionsfreien Geometrien betrifft, so dass auch in späten Stadien der Optimierung bei Verwendung dieses Operators nur wenige Proteine verworfen werden müssen. Mit Hilfe des LT-Operators können Proteingeometrien abschnittsweise neu gefaltet werden, ohne die Gesamtfaltung während der Mutation zu verändern. In der anschließenden lokalen Optimierung konnten selbstverständlich Anpassungen der Proteingeometrie an die neue lokale Struktur stattfinden, welche aber verhältnismäßig gering ausfielen.
Da der LT-Operator im Gegensatz zum Mutationsoperator in internen Koordinaten ungeeignet ist, große Konformationsbereiche abzusuchen, wurde er so implementiert, dass er in einer späteren Generation des Programmes als zweiter alternativer Mutationsoperator dazu genommen wurde. Der LT-Operator wurde aktiviert, wenn $ngen \geq 0.75\ MaxGen$ war.

Evolutionsoperatoren: Optimierung

Nach einer erfolgreichen Kreuzung oder Mutation zweier Eltern erfolgte die lokale Optimierung der resultierenden Proteine. Dieser Schritt setzt sich aus zwei Phasen zusammen. Im ersten Schritt werden die Seitenkettenpositionen mittels eines weiteren genetischen Algorithmus verbessert, während im zweiten Schritt die gesamte Struktur auf Basis des entwickelten Kraftfeldes zu einem lokalen Minimum der potentieller Energie optimiert wird.

Beim Eintritt in die Optimierungsroutine besitzen die Aminosäuren der Proteine bereits Seitenkettenpositionen, wie sie in den vorher durchlaufenen und oben beschriebenen Routinen wie Mutation oder Kreuzung gesetzt wurden. Das Ziel in diesen Programmabschnitten war darauf ausgerichtet gewesen, eine Seitenkettenposition zu finden, die nicht im Konflikt mit anderen Seitenketten oder C^{α}-Atomen stand, um dem Hauptprogramm kollisionsfreie Strukturen zurückzugeben. Die Positionierung der Seitenkette beruhte somit auf rein geometrischen Erwägungen. Energetische Aspekte wurden dabei nicht berücksichtigt. Diesem wird zu Beginn der Optimierungsroutine Rechnung getragen. Durch Testrechnungen zeigte sich im Vorwege, dass sich bei einer lokalen Geometrieoptimierung mit Gradienteninformation die Seitenketten-Positionen nur sehr geringfügig änderten, so dass sie nicht in der Lage waren,

sich vom einem Clusterzentrum zu einem anderen zu bewegen. Gleichzeitig hängt aber die Gesamtenergie eines Proteins nicht unbedeutend von der Seitenketten-Konfigurationen ab, da z. B. zwei gleichartig geladene Seitenketten ein stark repulsives Paarpotential besitzen. Daher wurde vor dem Schritt der Gradientenoptimierung des gesamten Proteins ein kurzer genetischer Algorithmus implementiert, der eine optimale bzw. verbesserte Anordnung der Seitenketten sucht. Bei diesem Algorithmus wurden die Koordinaten der C^{α}-Atome als fixiert angenommen. Die Seitenketten-Konfigurationen wurden über einen Vektor codiert, der die Ordnungsnummer der Clusterpositionen zu jeder Seitenketten enthielt (Genotyp). Der zugehörige Phenotyp entspricht den kartesischen Koordinaten der Seitenketten im Rahmen des festen Rückgrates. Die Anfangskonfigurationen wurden zufällig generiert, und durch die Konfiguration der Seitenketten des Proteins bei Eintritt in diese Routine ergänzt. Wie für den übergeordneten GA bereits beschrieben, durchlaufen die Seitenketten-Konfigurationen eine Anzahl an Generationen, wobei in jeder Generation eine Kreuzung oder Mutation der Konfigurationen eintreten kann. Die Wahrscheinlichkeiten hierfür waren identisch zu denen des Hauptprogramms. Nach den Mutationen wurden die Konfigurationen in kartesische Koordinaten umgewandelt und das Wechselwirkungspotential zwischen den Seitenketten, zwischen den Seitenketten und dem Rückgrat sowie das Oberflächenpotential berechnet. Da das Rückgrat hierbei fixiert war, wurden die bindenden und nicht-bindenden Wechselwirkungen der C^{α}-Atome untereinander nicht mit berücksichtigt. Zusätzlich zu diesen Energien wurden noch die repulsiven Potentiale hinzugenommen, die den Minimalabstand der Seitenketten zu anderen Punkten beinhalten. Diese Energieberechnung geschah ohne Verwendung von Gradienteninformationen als Einzelenergiewert-Rechnung (*single-point energy*). Um diesen Teil des Programmes kurz zu halten, wurden die Anzahl an Individuen (Seitenketten-Konfigurationen) und Generationen separat in einigen Testläufen bestimmt, wobei ein Kompromiss zwischen Zeitaufwand und Energieminimierung getroffen wurde. Es wurden insgesamt fünf Individuen und fünf Generationen für den Seitenketten-GA verwendet.

Nach Beendigung der Optimierung der Seitenkettenpositionen, wurde das gesamte Protein lokal mittels des parametrisierten Kraftfeldes optimiert. Für die Gradientenberechnung wurden die analytischen Ableitungen der Potentialfunktionen verwendet. Die Richtungen der einzelnen Optimierungsschritte auf der Energiefläche wurden mittels eines Quasi-Newton-Algorithmus bestimmt, wobei die Hesse-Matrix über die BFGS-Methode aktualisiert wurde [262].

Evolutionsoperatoren: Selektion

Nach der lokalen Optimierung aller veränderten Proteine in einer Generation folgte der letzte Schritt, welcher zur Vorbereitung der nächsten Generation diente. In diesem wurde aus der vollständigen Menge aller in einer Generation vorhandenen Proteine ein Satz bestimmter Strukturen ausgewählt, welche an die nächste Generation weitergegeben wurde (Eltern). Die nicht ausgewählten Proteine wurden dagegen verworfen und nicht weiter verwendet.
Für die Selektion der Proteine, die die Eltern der nächsten Generation bilden, wurden insgesamt drei unterschiedliche Methoden bzw. Kriterien implementiert, die vom Benutzer eingestellt werden können. Zusätzlich zu diesen drei Verfahren, die im folgenden beschrieben werden, kann eingestellt werden, ob im Programm eine Elite-Strategie verfolgt werden soll oder nicht. Dies bedeutet, dass, im Falle einer gewünschten Elite-Strategie, unabhängig davon, welche Proteine durch die Selektionskriterien ausgewählt wurden, das Protein mit der niedrigsten potentiellen Energie garantiert an die nächste Generation weitergegeben wird. Ein Nachteil der Elite-Strategie ist, dass dadurch eine globale Optimierung schnell in einem lokalen Minimum gefangen bleiben kann. Andererseits aber verliert man Informationen über energetisch günstige Anordnungen, wenn die Elite-Strategie nicht verfolgt wird, und dadurch das am besten bewertete Protein aus der Menge entfernt wird.

Neben der Festlegung der Elite-Strategie, welche lediglich ein Protein für die Folgegeneration auswählt, muss auch das Kriterium zur Bestimmung der restlichen Proteine einer Population, die an die nächste Generation weitergegeben wird, bestimmt werden. Hierzu kann im Programm weiterhin entschieden werden, ob die unveränderten Eltern, die am Anfang einer Generation standen, mit in den Auswahlprozess aufgenommen werden sollen, oder ob nur aus den neuen, mutierten Proteinen gewählt werden soll. Dies wurde stets so eingestellt, dass die Eltern mit verwendet wurden, um zu verhindern, dass deren Informationen verloren gehen würden, wenn alle mutierten Proteine schlechtere Kandidaten darstellen.
Die Selektion der Proteine für die nächste Generation konnte schließlich mit Hilfe dreier verschiedener Kriterien durchgeführt werden:
a) Auswahl der besten Strukturen. Hierbei werden sämtliche Proteine nach ihrer potentiellen Energie geordnet und diejenigen mit der niedrigsten Energie übernommen. Hierbei muss sich die Energie der Proteine um jeweils mindestens einen Wert von $1 \cdot 10^{-8}$ Energieeinheiten unterscheiden. Ansonsten wird angenommen, dass die Proteine zu ähnlich sind, so dass nur eines verwendet wird.
b) Zufällige Auswahl. Hier werden die Proteine völlig zufällig aus der gesamten Menge ausgewählt. Diese Methode gewährleistet nicht, dass die besten oder guten Strukturen weitergegeben werden.
c) Nischen-Kriterium. Dies ist das Kriterium der Wahl, welches durchweg in den Rechnun-

gen benutzt wurde. Die Proteine werden zunächst wie unter a) nach ihrer Energie geordnet. Danach werden alle Proteine untereinander auf ihre geometrische Ähnlichkeit hin untersucht, welche durch den cRMSD-Wert nach einer idealen Überlagerung beider Strukturen gegeben ist. Zwei Strukturen werden hierbei als ähnlich eingestuft und der gleichen Kategorie zugeordnet, wenn der cRMSD < 3 Å ist. Ist ein Protein keinem anderen Protein einer Kategorie ähnlich wird für dieses Protein eine neue Kategorie erstellt. Wenn alle Proteine Kategorien zugeordnet wurden, werden beginnend bei der ersten Kategorie und in der Reihenfolge, in der die anderen Kategorien erstellt wurden, jeweils die ersten Proteine dieser Kategorien ausgewählt, bis genügend Proteine für die nächste Generation selektiert wurden. Dieses Vorgehen führt dazu, dass jeweils die energetisch niedrigsten Strukturen weiterverwendet werden, die zudem unterschiedliche strukturelle Informationen beinhalten, wodurch eine Diversität der Population gewährleistet wird. Sollten nicht genügend Kategorien zur Verfügung stehen, um die nächste Population zu füllen, so werden die fehlenden Proteine unter Verwendung der Strukturerzeugungsroutinen der Anfangsgeneration neu erstellt.

4.7.4 Zusammenhang zwischen der globalen Geometrieoptimierung und der Parameteroptimierung

Der genetische Algorithmus wurde im Rahmen dieser Arbeit dazu verwendet, auf der Fläche des nach der Optimierung der Parameter vorhandenen Kraftfeldes die verwendeten Sequenzen global zu optimieren. Hiermit sollte zum einen die Energiefläche dahingehend geprüft werden, ob der native Zustand das globale Minimum der Potentialfunktion ist, während zum anderen gleichzeitig diejenigen Proteinstrukturen, die in diesem Prozess erzeugt wurden und eine niedrigere Energie als das native Protein besaßen, für die nächste Iteration der Parameter-Optimierung verwendet wurden.
Die diesem Vorgehen zugrunde liegende Idee ist es, in einem iterativen Prozess Informationen über die Umgebung des globalen Minimums zu erhalten, wobei davon ausgegangen wird, dass die geometrisch entfernten Strukturen eine höhere Energie besitzen, als die Strukturen, die nahe dem globalen Minimum liegen. Verschiedene Untersuchungen zeigen, dass diese Annahme für Proteine im Allgemeinen gerechtfertigt ist, da die freie Energiefläche, die sowohl die Enthalpie wie auch die Entropie berücksichtigt, für natürliche Proteine eine Trichterform besitzt, wobei der native Zustand das globale Minimum bzw. die Spitze des Trichters bildet [263–265]. Diese Trichterform beschreibt in einer vergröberten Art die Form der Energiefläche als ganzes. Betrachtet man die Energiefläche detailliert, so ist diese dennoch von vielen lokalen Minima und Maxima gekennzeichnet.
Aufgrund der computertechnischen Beschränkungen können zur Parameteroptimierung nur eine begrenzte Anzahl an Informationen über die gesamte Energiefläche der Proteine verwen-

det werden. Diese Informationen werden durch die dem Optimierungsprozess zur Verfügung gestellten Proteingeometrien erhalten. Hierbei entspricht eine Proteingeometrie lediglich einem isolierten Punkt auf der zu der gegebenen Proteinsequenz gehörenden hochdimensionalen Energiefläche. Daher ist insgesamt die Informationsbasis verglichen mit der Gesamtgröße der Fläche sehr klein. Verwendet man das grundlegende experimentell bekannte Wissen über die nativen Strukturen für die Proteinstrukturvorhersage, so lässt sich der benötigte Bereich der Energiefläche reduzieren. So sind die entfalteten Strukturen oder Strukturen mit sehr unregelmäßiger Sekundär- und/oder unregelmäßiger Tertiärstruktur im Allgemeinen nicht relevant. Auch die Geometrien mit hoher interner Energie sind weniger aussichtsreiche Kandidaten für den nativen Zustand. Aber selbst durch eine solche Reduktion bleibt der mögliche Konformationsraum sehr groß. Es sei an dieser Stelle noch einmal darauf hingewiesen, dass die Anzahl der lokalen Minima mit bis zu 10^N für N Aminosäuren skaliert, wobei diese Anzahl für eine vollständige atomare Darstellung des Proteins gilt. Für eine vergröberte Darstellung ist die Skalierung kleiner.

Aus diesem Grund müssen die der Parameter-Optimierung zur Verfügung gestellten Geometrien auf den wichtigen Bereich der Energiefläche beschränkt werden. Dieser umfasst diejenigen Proteins, die aufgrund ihrer Struktur als Möglichkeit für den nativen Zustand in Betracht kommen. Dabei handelt es sich im Normalfall um Strukturen, die kompakt gefaltet sind und die ein Mindestmaß an wohl definierter Sekundär- und Tertiärstruktur besitzen. Als Quelle für diese Proteine dienten die festen Sätze der falschen Strukturen und die mittels des genetischen Algorithmus erzeugten Geometrien. Der Unterschied zwischen den falschen Strukturen und der mittels GA-produzierten Geometrien besteht darin, dass die falschen Strukturen zu einem beliebigen Energiewert auf der Fläche gehören, während die GA-Strukturen (numerische) Minima der Fläche repräsentieren. Das Ziel des iterativen Parameter-Optimierungsprozesses ist es, mit Hilfe dieser Proteine die wichtigen Bereiche der Energiefläche abzusuchen und Minima zu finden, die energetisch niedriger als die native Struktur sind, was unter Voraussetzung der Gültigkeit der thermodynamischen Hypothese einem Fehler in der Energiefläche entspricht. Im Idealfall könnte somit der Bereich um den nativen Zustand erkundet und alle gefundenen Minima durch die Parameter-Optimierung energetisch gegenüber dem nativen Zustand destabilisiert werden. Diesem Prozess sind aber in der praktischen Durchführung bestimmte Schranken gesetzt. Zum einen ist die Anzahl an Strukturen, die zur Parameter-Optimierung verwendet werden können, durch die Programm- und/oder durch die Rechner-Architektur nach oben beschränkt. Zum anderen ist der genetische Algorithmus ein stochastischer Prozess, so dass nicht garantiert werden kann, dass bestimmte Minima überhaupt gefunden werden. Beide Faktoren beschränken die Größe der verfügbaren Informationsbasis. Zusammen mit der gewählten Methodik zur Parameter-Optimierung kommt hierzu eine weitere Schwierigkeit hinzu, welche die Form der Energiefläche an sich betrifft. Als Folge des iterativen Vorgehens ändert sich die Form der Potentialfläche von einer Iteration zur nächsten, da die

Parameter an die ausgewählten Proteine bzw. an die mittels GA gefundenen Minimumstrukturen, neu angepasst werden mussten. Aufgrund der oben erwähnten begrenzten Anzahl an verwendbaren Proteinen sind nur die in einer Iteration verwendeten Punkte (Proteinstrukturen) der Fläche bekannt, jedoch nicht die Form der Fläche zwischen diesen. Durch diese Tatsache ist das iterative Vorgehen komplizierter als eine reine Suche und Korrektur zu niedriger nicht-nativer Minima auf der Potentialfläche, da sich von einer Generation zur nächsten die energetische Reihenfolge der Proteine zusammen mit dem Aussehen der Fläche ändern kann. Dies führt zu einer Art von direktem Wechselspiel zwischen der Parameteroptimierung und der globalen Geometrieoptimierung, indem die Parameteroptimierung die Energiefläche bereitstellt, auf der in der globalen Geometrieoptimierung neue Punkte erzeugt werden. Die Form der Fläche und somit die Parameteroptimierung hat aber einen sehr wichtigen Einfluss darauf, welche Proteine im genetischen Algorithmus erzeugt werden. Besonders wichtig sind hierbei die Minima der Fläche, da der genetische Algorithmus in der lokalen Optimierung unter Verwendung des Gradienten des Potentials bzw. der Krümmung der Fläche zum nächsten (numerischen) Minimum läuft und im globalen Schritt zwischen verschiedenen Minima springt. Diese Minima aber beeinflussen wiederum in der anschließenden Iteration die Parameter-Optimierung, wodurch sich die Folgefläche von der vorhergehende Energiefläche unterscheiden wird. Diese Wechselspiel ist in Abb. 4.38 symbolisch für ein Protein entlang einer eindimensionalen Geometriekoordinate $\mathbf{X}$ dargestellt. Vor der Parameter-Optimierung ist, was in Abb. 4.38a gezeigt ist, keine Energiefläche gegeben, bzw. nur eine Art "Grundfläche", die durch die Basisfunktionen ohne optimierte Parameter gegeben ist. Hierdurch ist die Energie der Proteine nicht bekannt bzw. nicht angepasst. Es besteht lediglich die Forderung, dass die falschen Proteine $\mathbf{X}_i$ energetisch höher als die native Struktur $\mathbf{X}^*$ sein sollen. Nach der Parameter-Optimierung kann den Geometrien $\mathbf{X}$ eine Energie zugeordnet werden, was in Abb. 4.38b dargestellt ist. Da aber nur für die bekannten Geometrien $\mathbf{X}_i$ Informationen vorlagen, hängt der Bereich der Fläche zwischen diesen Punkten von der Optimierung und den resultierenden Potentialfunktionen ab und ist im Grunde nicht auf einen bestimmten Wertebereich beschränkt, so dass zwischen den Punkten $\mathbf{X}_i$, die per Konstruktion eine höhere Energie als die native Struktur $\mathbf{X}^*$ besitzen, viele energetisch niedrigere Minima existieren können, was in der Grafik 4.38b beispielsweise durch die Minima zwischen $\mathbf{X}_1$ und $\mathbf{X}_2$ sowie zwischen $\mathbf{X}^*$ und $\mathbf{X}_3$ gegeben ist. Im folgenden Schritt wird diese Fläche mittels des genetischen Algorithmus abgesucht, um nach Möglichkeit tiefliegende Minima wie die eben erwähnten zu entdecken. Die mit dieser Methode gefundenen Minima, die eine niedrigere Energie als die native Struktur besitzen, werden für die nächste Generation ausgewählt. Dies ist in Abb. 4.38c verdeutlicht, wobei die leeren Quadrate die mittels genetischem Algorithmus gefunden und ausgewählten energiearmen Strukturen darstellen. Das Minimum zwischen $\mathbf{X}_1$ und $\mathbf{X}_2$ wurde in diesem Beispiel aufgrund des statistischen Charakters des globalen Optimierungsalgorithmus nicht erreicht, so dass es nicht verwendet kann. Aufgrund der für dieses

Beispiel hypothetischen computertechnischen Beschränkung können nur vier falsche Strukturen verwendet werden, wodurch die Punkte $\mathbf{X}_1$ und $\mathbf{X}_4$, die energetisch am weitesten vom nativen Zustand entfernt sind, aussortiert werden. Somit werden für die nächste Parameter-Optimierung lediglich die Punkte $\mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_5$ und $\mathbf{X}_6$ verwendet. Mit der erneuten Anpassung der Parameter, was in Abb. 4.38d gezeigt ist, bei der das Ziel ist, die Basisfunktionen so zu parametrisieren, dass die Strukturen $\mathbf{X}_5$ und $\mathbf{X}_6$ energetisch angehoben bzw. $\mathbf{X}^*$ energetisch relativ zu diesen abgesenkt wird, verändert sich die Form der Fläche, da über die Form der Fläche prinzipiell keine weiteren Annahmen gemacht werden, insbesondere auch, was die relativen Energien der falschen Strukturen untereinander betrifft. Dadurch können beispielsweise neue Minima entstehen oder vorher existierende Minima verschwinden. Diese Änderung geschieht zum einen durch die neu hinzugekommenen Informationen durch die Geometrien der Punkte $\mathbf{X}_5$ und $\mathbf{X}_6$, aber ebenso auch durch eine Abnahme an Informationen über die Fläche, indem $\mathbf{X}_1$ und $\mathbf{X}_4$ nicht mehr mitberücksichtigt wurden. Abb. 4.38d zeigt auch, dass eine erneute Geometrie-Optimierung auf dieser Fläche andere Minimumstrukturen erzeugen wird, als mit Fläche aus Abb. 4.38b erhalten wurden.
Hier ist die Frage zu stellen, ob diese Prozedur konvergiert, so dass ein Satz an Proteinstrukturen erhalten werden kann, auf dessen Basis eine Potentialfunktion parametrisiert werden kann, die tatsächlich den nativen Zustand als globales Minimum liefert oder ob aufgrund der begrenzten Anzahl an Punkten auf der Fläche die Fluktuationen von einer Iteration zur nächsten erhaltenen bleiben, so dass stets andere Minima erzeugt werden, die eine nativartige, aber falsche Geometrie besitzen, die energetisch niedriger als der native Zustand sind.

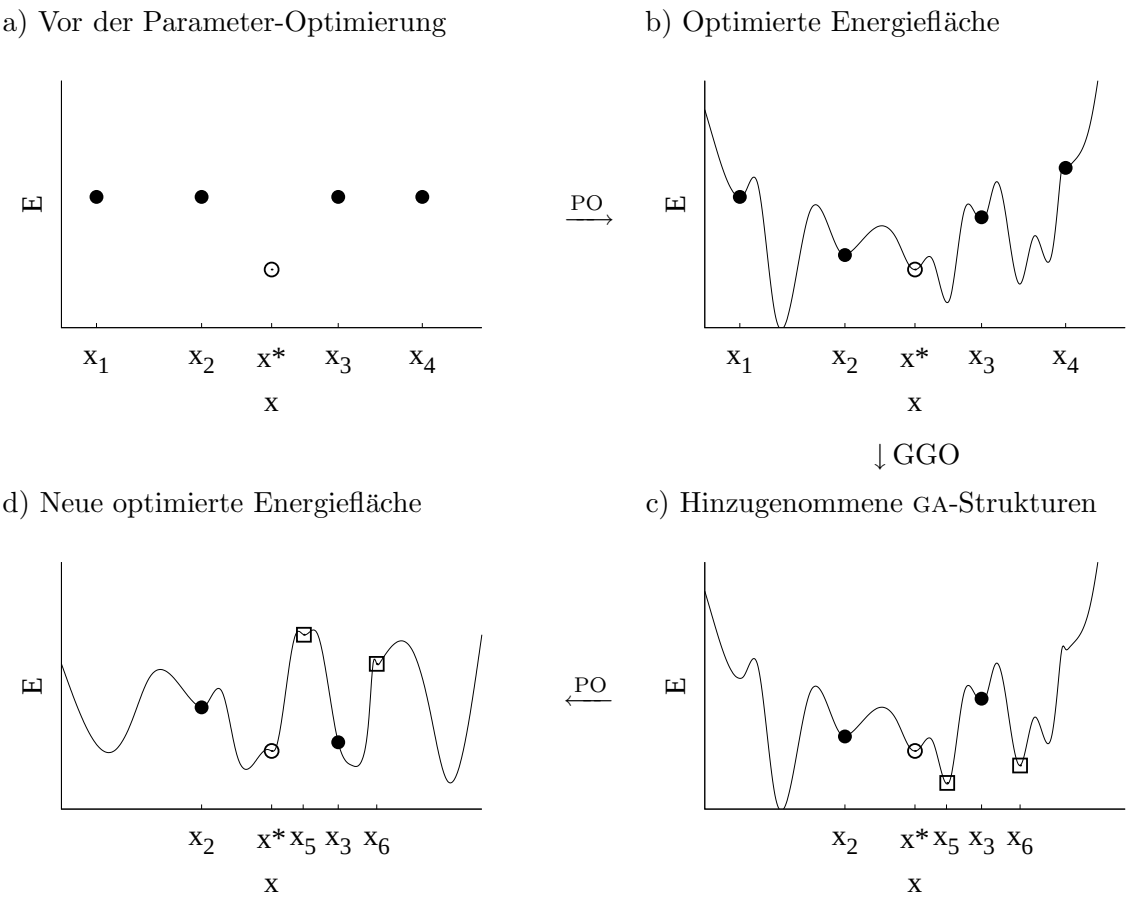


Abbildung 4.38: Schematische Darstellung der Abhängigkeit zwischen der Parameter-Optimierung (PO) und der globalen Geometrie-Optimierung (GGO). In den Grafiken ist die Energie E des Proteins in Abhängigkeit einer Geometriekoordinate $\mathbf{X}$ aufgetragen. $\mathbf{X}^*$ steht für die native Geometrie, die $\mathbf{X}_i$ für die falschen, nicht-nativen Strukturen. In a) sind durch die Punkte die ausgewählten Geometrien vor Parameter-Optimierung dargestellt, wobei der hohle Kreis für die native Geometrie steht. Durch die unterschiedliche Höhe der Punkte ist die Nebenbedingung verdeutlicht, dass alle falschen Strukturen eine höhere Energie besitzen sollen. b) zeigt die erhaltene Energiefläche nach der Parameter-Optimierung. c) enthält die ausgewählten Proteine nach der globalen Geometrieoptimierung für die nächste Parameter-Optimierung. Die Quadrate stehen für Geometrien, die aus dem genetischen Algorithmus erhalten wurden. d) zeigt die Energiefläche nach der erneuten Parameter-Optimierung. Die verwendeten falschen Strukturen besitzen eine höhere Energie, da aber keine Informationen zwischen den Punkten vorliegen, können dort energetisch niedrigere Bereiche entstehen.

4.8 Ergebnisse der globalen Geometrieoptimierung

4.8.1 Vortest des Programms

Ziel dieser Arbeit war es, unter Verwendung des parametrisierten Kraftfeldes Proteinsequenzen global zu optimieren, um so im Idealfall die Struktur des nativen Zustandes bestimmen zu können. Bevor dieses Verfahren auf unbekannte Sequenzen angewendet werden kann, musste getestet werden, inwieweit bekannte Strukturen reproduziert werden.
Bevor hierzu der genetische Algorithmus für die globale Optimierung der Proteinsequenzen verwendet wurde, wurde er vorher darauf getestet, ob er generell unabhängig von einem bestimmten Kraftfeld in der Lage sei, reale Proteinstrukturen zu erzeugen und sich dem nativen Zustand zu nähern. Hierzu wurden die 15 kürzesten Proteine des reduzierten TOP500H-Datenbanksatzes ausgewählt und testweise durch den genetischen Algorithmus optimiert. Da hierfür kein Kraftfeld verwendet wurde, wurden die C^{α}-Bindungslängen auf 3.8 Å fixiert und nur die Bindungswinkel und -torsionswinkel variiert. Die Energie- bzw. Bewertungsfunktion in diesen Testläufen war die direkte cRMSD-Abweichung von der nativen Struktur nach einer optimalen Überlagerung. Es fand keine lokale Optimierung unter Verwendung von Gradienteninformationen statt. Dies bedeutet, dass es leicht zu Kollisionen bzw. zu geometrischen Anordnungen kommen konnte, die in realen Proteinen nicht möglich oder sehr ungünstig sind, da keinerlei repulsive (oder attraktive) Wechselwirkungen zwischen den Interaktionszentren vorhanden waren. Diese Art der globalen Optimierung ist eine anspruchsvolle Aufgabe für den genetischen Algorithmus, da keine Informationen über bevorzugte Geometrien oder Wechselwirkungen innerhalb des Proteins verwendet wurden, wodurch der zugängliche Konformationsraum des Proteins wesentlich größer ist als bei einer physikalisch richtigen Beschreibung. Die einzige Triebkraft war nur durch die Minimierung der strukturellen Differenz zum nativen Zustand gegeben.
Die 15 für diesen Vortest ausgewählten Proteine besaßen eine Länge zwischen 50 und 99 Aminosäuren und deckten die wichtigsten Sekundär- und Tertiärstrukturen ab. Es waren sowohl reine α-Helix- oder β-Faltblatt-Strukturen, aber auch Proteine mit einer gemischten Struktur enthalten. Jedes Protein wurde in insgesamt zwanzig Läufen global optimiert, wobei für jedes Protein eine Population von zwanzig Individuen und eine Laufzeit von fünfzig Generationen angesetzt wurde. Bestimmt wurde zu allen Proteinen der niedrigste erhaltene und der über alle Läufe gemittelte cRMSD-Wert. Die Ergebnisse sind in Tab. 4.27 und die minimalen cRMSD-Werte zusätzlich grafisch in Abb. 4.39 gezeigt.

Die Ergebnisse der minimalen cRMSD-Abweichungen zeigen, dass der Algorithmus in der Lage ist, trotz Mangels an strukturellen Informationen Geometrien zu erzeugen, die dem nativen

Zustand nahe sind. Die erhaltenen Werte decken einen Bereich von 1.58 bis 5.99 Å ab. Wie aus Abb. 4.39 ersichtlich ist, korrelieren diese Werte schwach mit der Länge N der Proteine. Diese Korrelation ergibt sich daraus, dass die Anzahl an möglichen Konformationen eines Proteins, wie in Abschnitt 3 beschrieben, mit der Anzahl an Aminosäuren N skaliert, wobei die Obergrenze dieser Skalierung mit N^{10} abgeschätzt wird. Da die Anzahl an Individuen und Generation und somit die mögliche Gesamtanzahl an berechenbaren Punkten auf der Energiefläche für alle Proteine konstant war, ist die Abdeckung des Konformationsraumes für die größeren Proteine schlechter, wodurch die minimalen und mittleren cRMSD-Werte schlechter sind. Zudem ist die Korrelation weniger deutlich, da es sich um einen statistischen Prozess handelt.

Betrachtet man zusätzlich noch die Anteile der Sekundärstrukturen der nativen Strukturen, so zeigt sich dass deren Zusammensetzung ebenfalls Auswirkungen auf die Vorhersage haben können. Dies ist beispielsweise bei den Proteinpaaren 1nkd und 1tud sowie 1lmb(4) und 1fna deutlich. Die Geometrien von 1nkd und 1lmb(4) enthalten lediglich α-Helices, aber keine β-Faltblattstrukturen, während die Proteine 1tud und 1fna diesbezüglich gemischte Geometrien enthalten. Die Paarweisen cRMSD-Werte dieser Proteine zeigen jeweils ein Verhältnis von 2.2 zu 4.0 Å für 1nkd und 1tud sowie 4.0 zu 6.0 Å für 1lmb(4) und 1fna, wobei jeweils das reine α-Helix-Protein zuerstgenannt ist. Der Unterschied in den cRMSD-Werten beträgt somit für beide Paarungen ca. 2 Å, obwohl sich in jedem Paar die Länge der Proteine jeweils nur um eine Aminosäure unterscheidet. Dies bedeutet, dass bei gleicher Länge die Qualität der Vorhersage ebenso auch mit der Komplexität der dreidimensionalen Struktur zusammenhängt, was darauf zurückzuführen, dass β-Faltblattstrukturen (oder andere wohl definierte Supersekundär- oder Tertiärstrukturen) algorithmisch wesentlich schwieriger zu erzeugen sind als α-Helices, welche durch eine ausgeprägte Nahordnung stabilisiert werden. Sofern es sich bei β-Faltblattstrukturen nicht um sog. β-*hairpins* handelt, sind die Segmente, die sich zu einem Faltblatt zusammenlagern in der Sequenz meist weit voneinander entfernt, wodurch eine Erzeugung dieser Strukturen ohne beispielsweise Potentialinformationen zu verwenden, die eine Triebkraft in Richtung dieser Strukturen durch eine Wasserstoffbrückenformation erzeugen, oder ohne eine direkte Generierung der Strukturen durch ein Programm, sehr unwahrscheinlich ist, da die Anzahl an kombinatorischen Möglichkeiten in der Abfolge der Bindungs- und der Torsionswinkeln zwischen den C^{α}-Atomen, die keine β-Strukturen erzeugen, sehr viel größer ist als diejenige, die diese Geometrien enthält.

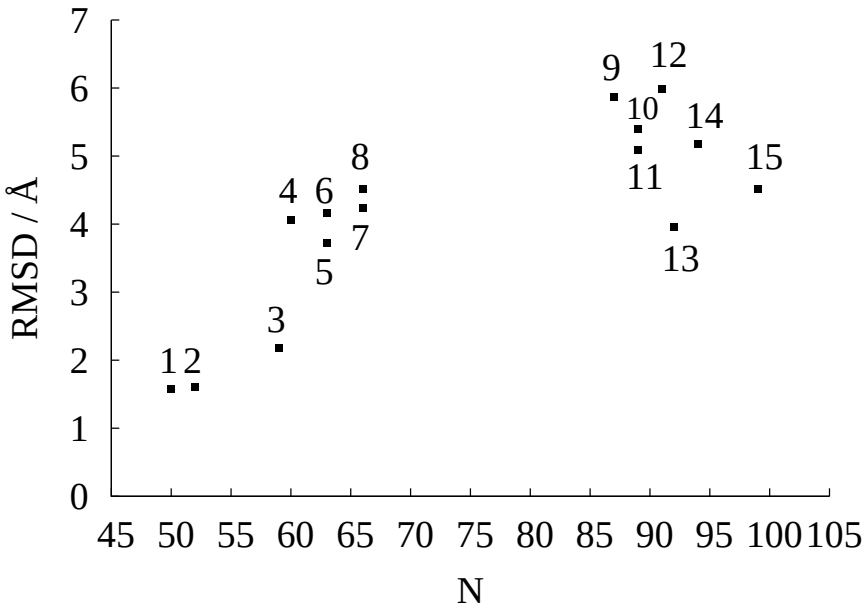


Abbildung 4.39: Darstellung der minimalen cRMSD aus Tab. 4.27 mit Anzahl der Aminosäuren N. Die angegebenen Nummern entsprechen den Proteinen der Tab. 4.27

Nr.	PDB-Code	N	RMSD_{min}/Å	$\text{RMSD}_{\text{mittel}}$/Å
1	1a92(A)	50	1.58	1.98
2	1hcr(A)	52	1.60	4.45
3	1nkd	59	2.18	3.93
4	1tud	60	4.06	4.83
5	1a1y(I)	63	3.73	4.67
6	1bf4(A)	63	4.16	6.02
7	1msi	66	4.24	4.91
8	1pcf(A)	66	4.51	5.09
9	1gvp	87	5.87	7.09
10	1ten	89	5.39	7.22
11	1ay7(B)	89	5.09	5.98
12	1fna	91	5.99	6.90
13	1lmb(4)	92	3.96	5.51
14	1mol(A)	94	5.17	6.54
15	1bm8	99	4.51	6.17

Tabelle 4.27: Ergebnisse der Testläufe ohne Kraftfeld auf Basis der Minimierung des cRMSD-Wertes. N ist die Anzahl an Aminosäuren, RMSD_{min} der erreichte minimale cRMSD-Wert und $\text{RMSD}_{\text{mittel}}$ der gemittelte cRMSD-Wert aus allen Läufen.

4.8.2 Ergebnisse mit der Potentialfunktion

RMSD-Abweichungen und -Energiekorrelation

In diesem Abschnitt werden im folgenden die Strukturen der Proteine beschrieben, die mittels des genetischen Algorithmus' erhalten wurden. Aufgrund der Vielzahl an erzeugten Geometrien wird aber nicht auf alle erhaltenen Proteine eingegangen, sondern im Wesentlichen die Gruppeneigenschaften der zur einer Sequenz gehörenden Proteinmenge gezeigt oder einzelne ausgewählte Proteine als Beispiele aufgeführt.

Für die Optimierungsläufe wurden 27 der insgesamt 48 Proteinsequenzen des verwendeten TOP500H-Datensatzes ausgewählt, wobei das Auswahlkriterium die Sequenzlänge war. Es wurden die kleinsten Proteine ausgewähl, da der einem Protein zugängliche Konformationsraum mit der Anzahl an Aminosäuren skaliert, so dass die globale Suche in diesem für größere Proteine generell längere Simulationszeiten erfordert, welchem im genetischen Algorithmus durch eine größere Anzahl an Individuen und/oder durch eine größere Anzahl an Generationen Rechnung getragen werden muss. Zusätzlich steigert eine größere Zahl an Aminosäuren die benötigte Dauer für eine lokale Optimierung, da mehr Wechselwirkungen aufgrund der höheren Anzahl an Interaktionszentren zu berechnen sind. Diese Gründen führen dazu, dass die benötigte reale Rechenzeit stark mit der Anzahl der Aminosäuren anwächst. Um in möglichst kurzer Zeit dennoch viele Informationen über die Energieflächen aus der globalen Optimierung zu erhalten, wurden die kürzesten Sequenzen verwendet. Dies hat den Vorteil, dass so bei vorgegebener Gesamtrechenzeit mehrere unterschiedliche Sequenzen optimiert werden können, während im Fall der Wahl von großen Proteinen sehr viel weniger Optimierungen durchgeführt werden könnten. Die verwendeten Proteine sind in der ersten Spalte der Tab. 4.28 aufgelistet.

In der globalen Geometrieoptimierung sind im Hinblick auf das Ziel der Strukturvorhersage des nativen Zustandes zwei Merkmale der erhaltenen Geometrien wichtig: Zum einen ist es zunächst notwendig zu bestimmen, ob die native Struktur reproduziert werden kann bzw. wie ähnlich die erzeugten Geometrien der Zielstruktur sind. Dieser Unterschied wird hier mit dem RMSD-Wert der C^{α}-Atome quantifiziert. Eine solche Einschätzung wurde bereits im Vortest für den genetischen Algorithmus ohne das optimierte Potential durchgeführt (siehe Abschnitt 4.8.1). Weiterhin ist die Kenntnis wichtig, ob die Strukturen mit den kleinsten RMSD-Werten zu den die energetisch niedrigsten Individuen gehören bzw. welcher funktionaler Zusammenhang im Allgemeinen zwischen den RMSD-Werten und der Energie der Proteine besteht.
Zur Beantwortung dieser Fragestellungen werden zunächst für die jeweilige Parameter-Iteration die minimalen und gemittelten RMSD-Werte betrachtet, die die Strukturen aus der

globalen Optimierung zur nativen Geometrie besitzen. Diese sind Tab. 4.28 aufgelistet. Die gemittelten RMSD-Werte ($\text{RMSD}_{\text{mittel,n}}$) für die n-te Parameter-Iteration geben den durchschnittlichen RMSD-Wert aller energieärmsten Individuen aus allen Optimierungsläufen bei Beendigung des Algorithmus wieder. Die minimalen RMSD-Werte ($\text{RMSD}_{\text{min,n}}$) sind die kleinsten Abweichungen zur nativen Struktur, die im Laufe aller Simulationen gefunden wurden. Sie sind absolut und nicht gemittelt. Die minimalen RMSD-Werte sind zusätzlich in Abb. 4.42 für alle Iterationen grafisch aufgetragen. Diese Werte geben die Information darüber, ob der genetische Algorithmus in Verbindung mit dem optimierten Kraftfeld in der Lage ist, nahnative Strukturen zu erzeugen. Die Abbildung 4.42 zeigt zunächst, dass auch die RMSD-Werte für die Simulationen mit dem optimierten Potential eine lineare Abhängigkeit von der Größe der Proteine besitzen. Die Korrelationskoeffizienten für die (nicht abgebildeten) Regressionsgeraden für die drei Graphen betragen 0.851, 0.911 und 0.911. Diese Linearität kann darauf hindeuten, dass für die Proteine die Suche im Konformationsraum noch nicht ausreichend war. Hier wäre als Verbesserung vorzunehmen, den genetischen Algorithmus mit mehr Individuen und/oder Generationen laufen zu lassen, bzw. auf eine Parallel-Version umzustellen, in der zusätzlich mehrere Populationen gleichzeitig bearbeitet werden, so dass zwischen diesen zusätzlich Informationen ausgetauscht werden können. Dieses Vorgehen hat selbstverständlich zur Folge, dass dadurch die realen Rechenzeiten weiter zunehmen.

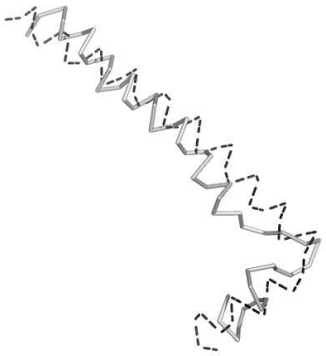
Die Werte in Tab. 4.28 zeigen, dass keine eindeutige Tendenz zu erkennen ist, ob die Zuordnung des nativen Zustandes im Verlauf der Parameter-Iterationen besser oder schlechter wird. Im Durchschnitt verbleiben die RMSD-Werte auf einem ähnlichen Niveau. So können die Schwankungen in den RMSD-Werten, die von einer Parameter-Iteration zur nächsten im Bereich von ein bis zwei Å liegen, auf den stochastischen Charakter des genetischen Algorithmus zurückgeführt werden, wodurch nicht erkenntlich ist, ob eine Entwicklung über die Parameter-Iterationen zu verzeichnen ist, wenn diese kleiner als die statistischen Schwankungen des genetischen Algorithmus sind.
Insgesamt sind die erreichten RMSD-Abweichungen noch nicht zufriedenstellend. Selbst für die kürzesten Proteine 1a92(A), 1hcr(A) und 1nkd sind die RMSD-Werte mit 4.4, 5.6 und 5.4 Å für eine erfolgreiche Vorhersage zu groß. Die zu diesen RMSD-Werten gehörenden Strukturen sind in ihrer optimalen Überlagerung mit der nativen Struktur in Abb. 4.40 dargestellt. Die Proteine 1a92(A) und 1nkd sind in ihrer Sekundär- und Tertiärstruktur relativ gut reproduziert, wobei diese zum größten Teil lediglich aus angeordneten α-Helices bestehen. Das Protein 1hcr ist generell schwieriger zu reproduzieren, da der α-Helix-Anteil mit ca. 48 % geringer ist und der Rest dieses Proteins keine regelmäßige Sekundärstruktur enthält, wodurch es im Hinblick beispielsweise auf die Nahwechselwirkungsterme, die darauf ausgelegt sind, sich wiederholende Sekundärstrukturen zu erzeugen, unwahrscheinlicher ist, die entsprechende native Struktur mit einer geringen Energie zu erzeugen.

Protein	N	$\text{RMSD}_{\text{min},0}$	$\text{RMSD}_{\text{min},1}$	$\text{RMSD}_{\text{min},2}$	$\text{RMSD}_{\text{mittel},0}$	$\text{RMSD}_{\text{mittel},1}$	$\text{RMSD}_{\text{mittel},2}$
1a1y(I)	63	7.897	8.987	8.188	11.282	12.445	11.956
1a3a(D)	144	14.034	15.317	13.698	17.982	20.840	18.872
1a92(A)	50	4.376	6.433	7.142	12.507	12.734	12.878
1ay7(B)	89	8.175	10.534	10.695	13.540	14.109	13.755
1bf4(A)	63	9.500	10.152	8.182	12.479	12.612	11.066
1bfg	126	12.703	13.872	12.425	18.319	17.240	17.106
1bgf	124	13.353	12.921	11.197	19.396	18.410	19.915
1bkr	108	11.245	11.421	10.692	15.821	16.034	15.427
1bm8	99	10.419	11.062	11.636	13.112	15.346	14.624
1dhn	121	14.880	14.285	14.540	21.046	18.735	18.138
1dpt(A)	117	13.395	13.739	13.557	16.713	16.483	16.001
1erv	105	11.272	10.693	11.806	14.247	14.157	14.608
1fna	91	12.885	12.587	12.253	17.525	17.107	16.675
1gvp	87	11.833	11.736	12.258	15.947	14.743	15.370
1hcr(A)	52	5.556	6.298	7.572	9.132	10.052	10.151
1ifc	131	14.354	14.139	13.755	17.383	18.251	18.713
1iib(A)	103	11.523	12.203	11.764	13.705	14.805	14.126
1lmb(4)	92	9.898	10.425	9.888	14.593	14.329	14.001
1mol(A)	94	13.797	13.319	13.111	16.390	19.401	16.009
1msi	66	9.441	9.082	8.873	12.271	12.641	12.208
1nkd	59	5.423	7.547	7.556	11.632	11.189	11.840
1pcf(A)	66	9.426	9.695	9.093	12.245	13.244	13.421
1stn(H)	136	14.579	13.760	13.323	19.198	17.386	17.048
1ten	89	13.492	12.504	12.609	17.556	16.202	15.901
1tfe	142	12.926	13.631	13.854	19.014	18.177	17.521
1ttb(A)	127	14.905	15.529	15.032	19.947	19.020	18.830
1tud	60	8.923	8.454	8.977	12.651	12.113	11.234

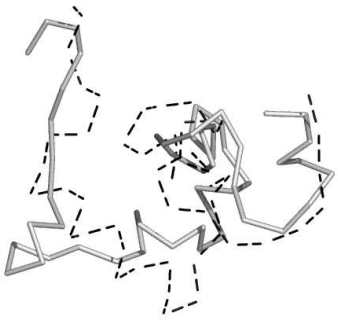
Tabelle 4.28: Übersicht über die erreichten cRMSD-Werte des genetischen Algorithmus für jeweils das beste (energetisch niedrigste) Individuum in Abhängigkeit von der Parameter-Iteration n mit $n = 0, 1, 2$. Angegeben sind die Anzahl an Aminosäuren N, $\text{RMSD}_{\text{min,n}}$ der minimale cRMSD-Wert und $\text{RMSD}_{\text{mittel,n}}$ der gemittelte cRMSD-Wert aus den Menge der besten Individuen.

Die anderen größeren Proteine erreichen lediglich noch höhere RMSD-Werte. Betrachtet man die zugehörigen Strukturen, so zeigt sich, dass diese zwar kompakt gefaltet sind, deren Faltungsschema aber weit vom nativen Zustand entfernt ist. Eine Möglichkeit dies zu begründen könnte in der ungenügend langen Laufzeit bzw. an einer zu kleinen Anzahl an durchlaufenen Iterationen und/oder zu wenigen Individuen im genetischen Algorithmus bestehen, was einer unzureichenden Exploration der Energiefläche entspricht. Diesem sprechen aber einige Argumente entgegen. Zum einen sind diese optimierten Strukturen, die weit vom nativen Zu-

a) 1a92(A)



b) 1hcr(A)



c) 1nkd

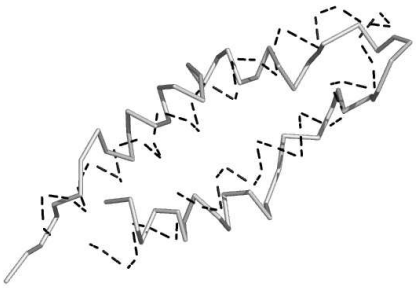


Abbildung 4.40: Optimale Überlagerung der mit dem genetischen Algorithmus erhaltenen Strukturen mit der kleinsten RMSD-Abweichung zur nativen Struktur für die Proteine 1a92(A), 1hcr(A) und 1nkd. Dargestellt sind jeweils die C^{α}-Atome des Proteinrückgrats. Das gestrichelt dargestellte Protein ist die native Struktur, das andere das optimierte Protein. Für die Überlagerung wurden nur die C^{α}-Atome verwendet.

stand entfernt sind, bereits energetisch günstiger als dieser. Daher wäre es zu erwarten, dass eine Veränderung der Parameter des genetischen Algorithmus wenige Vorteile bringen würde, sich dem nativen Zustand weiter zu nähern, da nur noch energetisch günstigere Minima erkundet werden würden, wobei die Wahrscheinlichkeit dafür, dass sich ein unentdecktes Minimum in der unmittelbaren Nähe zum nativen Zustand befindet, eher gering ist. Die geringe Wahrscheinlichkeit hierfür lässt sich wiederum folgendermaßen begründen: Zur Parameter-Optimierung wurden die auf den nativen Strukturen beruhenden verzerrten Proteine verwendet, die die unmittelbare geometrische Nachbarschaft des nativen Proteins erfassen. Diese Punkte der Energiefläche sind per Definition nach der Parameter-Optimierung energetisch höher als das native Proteine. Die Anzahl dieser falschen Strukturen ist in der Parameter-Optimierung zwar begrenzt und verteilt sich über einen hochdimensionalen Raum, anderer-

seits ist jedoch der native Zustand in der Regel sehr kompakt gefaltet, so dass viele Bereiche im direkt angrenzenden Konformationsraum durch Atomüberlappungen verboten sind, wodurch diese in der Parameter-Optimierung vernachlässigt werden können, da diese Kollisionen mit entsprechenden Wechselwirkungen im Potential erfasst werden können. Zusätzlich sind die Basisfunktionen zur Berechnung der Energie wiederum relativ einfache Funktionen, die auf der Ångström-Skala keine starken Oszillationen besitzen. Durch diese Faktoren ist es unwahrscheinlich, dass sich geometrisch betrachtet zwischen den verzerrten Strukturen und dem nativen Protein weitere tiefliegende Minima befinden, die in der Simulation nicht entdeckt wurden. Die Folgerung hieraus ist, das die nativen Zustände in der optimierten Form des Potentials lediglich ein lokales Minimum einnehmen, während die Energieflächen viele weitere energetisch günstige Minima enthalten, wodurch die Otimierung zu diesen läuft und nicht in der Nähe des nativen Zustandes endet.
Das andere oben bereits angeführte und nun weiter ausgeführte Argument für die schlechten RMSD-Werte könnte in einem schlecht parametrisierten genetischen Algorithmus liegen, wodurch die Energiefläche unzureichend erforscht würde. Da der genetische Algorithmus ein stochastischer Prozess ist, ließe sich dies lediglich mittels einer großen Anzahl an Testläufen ermitteln, bei welchen die Abdeckung des Problemraumes untersucht wird. Dies ist ein sehr aufwendiger Prozess, der hier nicht durchgeführt wurde. Ein Hinweis jedoch, ob der Explorationsvorgang zu kurz angesetzt wurde, liefert die Entwicklung der Proteinenergien in Abhängigkeit von der Generation. Dies wird in Abb. 4.41 für vier exemplarisch ausgewählte Proteine 1a92(A), 1bgf, 1nkd und 1ten dargestellt. In dieser Abbildung ist erkenntlich, dass die Energien zu Beginn der Simulation stark abnehmen und dann im Bereich der Generationen 10 bis 20 in ein Plateau übergehen, was einem typischen Verlauf eines genetischen Algorithmus entspricht. Die Erfahrung zeigt, dass, wenn das Plateau der Bewertungsfunktion erreicht ist, nur noch wenige bis gar keine Verbesserungen des Funktionswertes zu erwarten sind. Dies liegt typischerweise daran, dass meist relativ zu Beginn einer Simulation eine Geometrie erzeugt wird, die einige günstige Wechselwirkungen enthält, die die Gesamtenergie im Protein dominieren. Aufgrund der Komplexität des Konformationsraumes ist es anschließend unwahrscheinlich, dass in den folgenden Generationen diese Wechselwirkungen aufgelöst und durch andere mindestens genauso günstige Wechselwirkungen ersetzt werden, da dieser Vorgang meist mit größeren geometrischen Veränderungen verbunden wäre (Dies ist eine Folge des Selektionsoperators des genetischen Algorithmus). Dies führt dazu, dass diese dominierenden Wechselwirkungen an die Folgegenerationen vererbt werden und lediglich die anderen Wechselwirkungen bzw. die restliche Geometrie weiter optimiert wird, bis auch hier keine Verbesserungen mehr erzielt werden, wodurch die energetisch Konvergenz erreicht wird. Dieser Prozess der dominanten Wechselwirkungen, die die Optimierungsrichtung bestimmen, wird in den Simulationsläufen zwar häufig beobachtet, ist aber kein hervorstechendes Merkmal des implementierten genetischen Algorithmus, so dass ebenso auch viele Läufe beendet werden, in

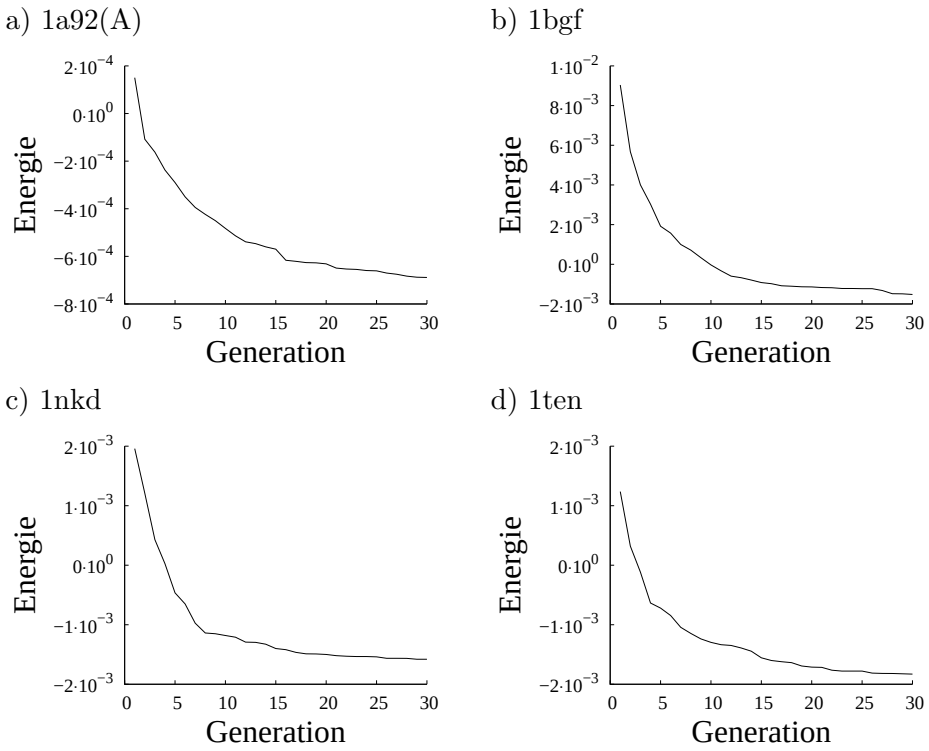


Abbildung 4.41: Mittelwert der Energien der besten Individuen in Abhängigkeit von der Generation des genetischen Algorithmus. Die dargestellten Energien wurden über alle Optimierungsläufe gemittelt.

denen bis zum Schluss merkliche geometrische Veränderungen stattfinden. Aber trotz dieser Veränderungen, fallen die energetischen Differenzen zum Ende einer Simulation gering aus. Abb. 4.41 zeigt nun, dass für die vier ausgewählten Proteine die energetische Konvergenz erreicht wurde, so dass es nicht zu erwarten ist, dass längere Simulationsläufe im Wesentlichen bessere Ergebnisse erzeugt hätten. Hieraus ließe sich zumindest schließen, dass die Parameter des genetischen Algorithmus ausreichend sind, die erreichten lokalen Minima zu erkunden.

Diese Argumente im Kombination mit den Ergebnissen des Vortests führen zu der Schlussfolgerung, dass nicht der genetische Algorithmus an sich ungenügend implementiert ist, um nah-native Strukturen zu erhalten, sondern, dass die Parametrisierung des Potentials verbessert werden sollte, um die falschen Minima der Energiefläche zu minimieren, die sich im Bereich möglicher Kandidaten für einen nativen Zustand befinden.

Neben der Erzeugbarkeit der Geometrie des nativen Proteins für die Strukturvorhersage ist ein weiterer wichtiger Punkt wie bereits oben angesprochen, der Zusammenhang zwischen der Energie und der geometrischen Distanz zum nativen Zustand. Für eine wie in dieser Arbeit angelegte Vorhersagemethode ist es wichtig, dass ein zur nativen Struktur niedriger RMSD-Wert mit einer sehr niedrigen Energie korrespondiert. Dies ist wichtig, da ohne weitere

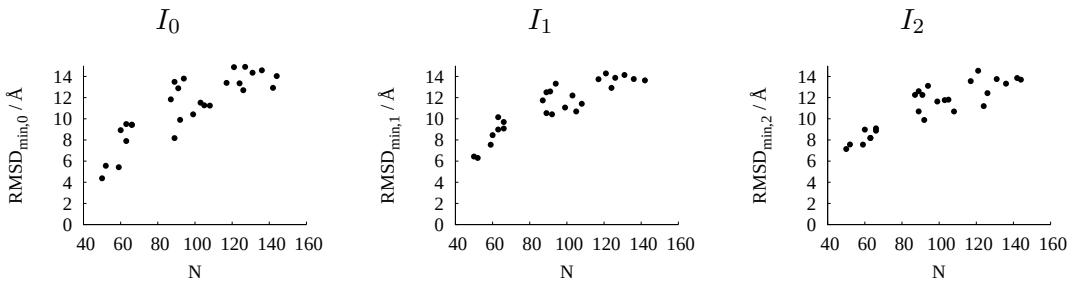


Abbildung 4.42: Zusammenhang zwischen dem erreichten minimalen RMSD$_{min,n}$-Werten in Abhängigkeit von der Anzahl der Aminosäuren N zu jeder Parameter-Iteration I_n.

Kenntnisse über die Struktur, was das Ziel einer *ab-initio*-Strukturvorhersage ist, die Energie das einzige Bewertungskriterium einer Struktur ist.

Wie oben bereits gezeigt, wurden mit dem genetischen Algorithmus Geometrien erzeugt, die energetisch niedriger als der native Zustand waren. Hierzu stellt sich die Frage, ob diese Geometrien in ihrer Anzahl in der Gesamtheit aller Geometrien eine untergeordnete oder eine wichtige Rolle einnehmen. Dies geht mit der Fragestellung, ob die Energiefläche viele Minima besitzt, die energetisch günstiger als das native Protein sind. Sofern diese Proteine selten auftreten ließen sich diese für die praktische Anwendung der Strukturvorhersage in einem zweiten eventuell anhand anderer Kriterien eliminieren. Für die Untersuchung des Zusammenhangs zwischen der RMSD-Abweichung und der Energie werden die entsprechenden Verteilungen im folgenden exemplarisch für einige Proteine betrachtet. Zur Darstellung wurden alle im genetischen Algorithmus erzeugten Strukturen verwendet, da sie in ihrer Gesamtheit Informationen über die gleiche Energiefläche liefern.

Für eine für die Proteinstrukturvorhersage ideal generierte Energiefläche wäre als Grundvoraussetzung zu erwarten, dass mit einem größeren RMSD-Wert auch die Energie größer als der native Zustand ist, wobei für diesen Zusammenhang keine bestimmte funktionale Form verbunden ist, da in der Parameter-Optimierung und bei der Auswahl der Basisfunktionen keinerlei Annahmen über die Form der Fläche oberhalb der nativen Energie gemacht wurden. Wie eine solche Verteilung der RMSD-Werte in Abhängigkeit von der Energie für ein idealisierte Fläche im Prinzip aussehen kann bzw. sollte, zeigt Abb. 4.43. Zu erwarten wäre, dass sich die erhaltenen Datenpunkte im Gebiet G befinden, dass jeweils nach unten durch eine Schranke g_1 und nach links durch eine Schranke g_2 begrenzt ist. Die explizite Verteilung der Punkte im Gebiet G und die Form der Schranken hängen stark von der Form der Fläche ab. Abb. 4.43 soll hierzu verdeutlichen, dass für eine erfolgreiche Vorhersage keine Strukturen mit einem niedrigen RMSD-Wert existieren sollten, die eine hohe Energie besitzen, was einer Überquerung der Schranke g_2 entspräche und dass keine Strukturen mit sehr niedriger Energie, aber hohem RMSD-Wert existieren sollten, was einer Überquerung der Schranke g_1 entspräche. Generell sollte ab einem gewissen geometrischen Abstand zum nativen Zustand

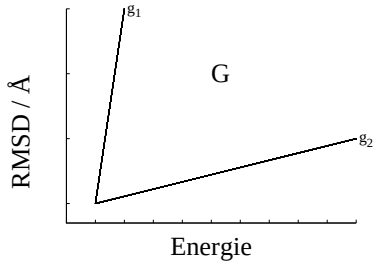


Abbildung 4.43: Idealisierte zu erwartende Verteilung der RMSD-Werte in Abhängigkeit von der Energie im Gebiet G, das durch g_1 und g_2 beschränkt wird.

gelten, dass je kleiner der RMSD-Wert ist, desto kleiner ist auch die Energie. Die gekennzeichneten Schranken g_1 und g_2 enden in der Abbildung nicht bei einem RMSD-Wert von Null, um zu verdeutlichen, dass eine Abhängigkeit vom Modell und vom Potential existiert, inwieweit man sich der nativen Geometrie nähern kann. Im einem vergröberten Modell können beispielsweise bestimmte Wechselwirkungen oder Struktureigenschaften verloren gehen, die eine falsche von der nativen Struktur gerade noch unterscheiden, wodurch die falsche Struktur eine zur nativen sehr ähnliche Energie besitzt, sich geometrisch aber noch unterscheidet.

Im Folgenden werden hierzu für einige ausgewählte Proteine diese Korrelationsgrafiken der RMSD-Werte dargestellt. Hierzu wurden die Proteine 1a92(A), 1pcf(A), 1ten und 1tfe ausgewählt. Die RMSD-Energie-Zusammenhänge für diese Proteine sind in Abb. 4.44 dargestellt. Die zusätzliche vertikale Linie in den Abbildungen gibt die native Energie an. Aus dieser Grafik ist grundsätzlich zu entnehmen, dass zu niedrigeren Energien hin die Dichte der Punkte zunimmt, während bei höheren Energien nur sehr wenige Punkte zu finden sind. Dies ist eine Folge des exponentiellen Verlaufs der Proteinenergien in der globalen Optimierung wie sie in Abb. 4.41 dargestellt wurde. Durch die schnelle Abnahme der Energien zu Beginn der Simulation werden nur wenige Minima in diesem Bereich der Energiefläche erreicht, während zum Ende der Simulation kaum noch energetische und strukturelle Verbesserungen erzielt werden, wodurch im erreichten Energie- und RMSD-Bereich viele dicht beieinanderliegende Punkte erzeugt werden. Da Abb. 4.44 eine Kombination mehrerer unabhängiger Simulationen mit unterschiedlichen Startbedingungen ist, können die Häufungspunkte mit hoher Punktdichte Informationen über die Minimastruktur der Energiefläche liefern. So entsprechen Gebiete mit der gleichen Energie und der gleichen RMSD-Abweichung im Regelfall auch einer ähnlichen Geometrie. Über die Ausdehnung der Punktwolke um die Minima können Informationen über die Einzugsgebiete der Minima gewonnen werden und damit über die Form den Energiefläche. An dieser Stelle sei aber wieder angemerkt, dass diese Darstellung der Energiefläche nicht vollständig ist, da bestimmte Minima aufgrund der Methode der globalen Optimierung nicht erreicht worden sein können. Betrachtet man im einzelnen die erhaltenen Punkteverteilungen aus Abb. 4.44, so ist die Verteilung für 1a92(A) am diffusten. Sie zeigt keine deutliche Ten-

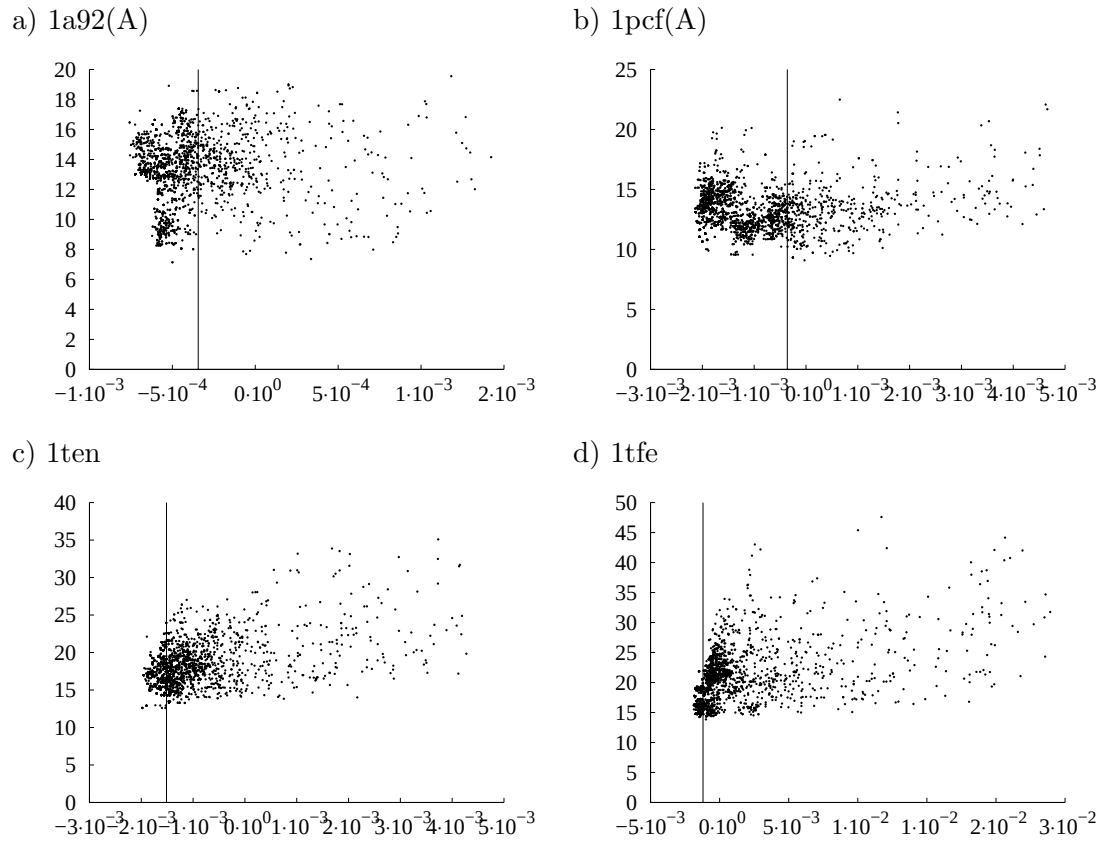


Abbildung 4.44: Streuung der RMSD-Werte in Abhängigkeit von der Energie des Proteins für vier ausgewählte Proteine des genetischen Algorithmus. Die vertikale Linie markiert die Energie des nativen Proteins.

denz zur Bevorzugung bestimmter Bereiche, was auf eine schlechte statistische Abdeckung oder auch auf eine relativ flache Energiefläche schließen lässt. Die anderen drei Proteine zeigen größere Tendenzen, in bestimmten Bereiche Häufungen zu bilden, wobei dieses für 1tfe am deutlichsten ist. Ebenso zeigt die Punkteverteilung zu diesem Protein die größte Ähnlichkeit zur idealen Verteilung, wie sie in Abb. 4.43 dargestellt ist. Über die eingezeichnete vertikale Linie, die die Energie des nativen Proteins angibt, lässt sich schließen, dass die Energieflächen durch viele Minima charakterisiert sind, die eine niedrigere Energie als das native Protein besitzen. Zur Darstellung, wie viele Punkte sich oberhalb der nativen Energie befinden, die für die Strukturvorhersage als richtig eingestuft werden können, im Vergleich zu der Anzahl an Punkten, die eine niedrigere Energie als das native Protein besitzen, wurde das Verhältnis n_l/n_h für jedes Protein des Optimierungssatzes berechnet. Hierbei sind n_l die Anzahl an Geometrien, die eine niedrigere als die native Energie besitzen und n_h die Anzahl an Proteinen, die eine höhere Energie besitzen. Dies wurde gegen die Anzahl an Aminosäuren N der Proteine aufgetragen und ist in Abb. 4.45 dargestellt. Diese Grafik zeigt, dass ein gewisser Zusammenhang zwischen der Anzahl an Aminosäuren und dem Verhältnis n_l/n_h besteht, wobei für kleinere Proteine mehr energetisch niedrigere Minima gefunden werden

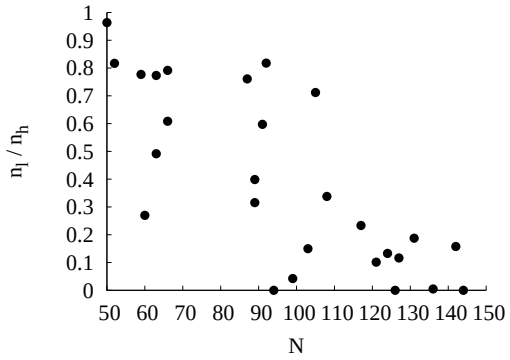


Abbildung 4.45: Verhältnis der Anzahl an energetisch niedrigeren n_l zu energetisch höheren n_h Proteinen relativ zur nativen Energie in Abhängigkeit von der Anzahl an Aminosäuren.

als für größere. Dies kann zwei Ursachen haben: Zum einen kann für kleinere Proteine die Erkundung der Energiefläche vollständiger sein als für die größeren, wodurch die Simulation für die kurzen Proteine in den energetisch niedrigen Minima endet, wobei die entsprechenden Minima für die längeren Proteine aufgrund der gleichen Simulationsdauer noch nicht erreicht wurden, da der Konformationsraum komplexer ist. Eine alternative oder zusätzliche Erklärung hierzu wäre, dass das Potential für größere Proteine besser funktioniert, da bei den sehr kleinen Proteinen die Lösungsmitteleffekte, welche im Potential nicht explizit erfasst sind, eine wichtige Rolle spielen, während bei den größeren Proteinen die Gesamtenergie mehr durch die proteininternen Wechselwirkungen bestimmt werden.

Des weiteren kann Abb. 4.44 auf ein weiteres Problem des Potentials hindeuten. Wie ersichtlich ist, fallen viele Minimumstrukturen auch in den Bereich der nativen Energien, besitzen aber einen deutlich zu großen RMSD-Wert um als nativ-ähnlich qualifiziert zu werden. Es wurden im Bereich der nativen Strukturen keine Minima entdeckt, die einen niedrigen RMSD-Wert besitzen. Dies könnte zum einen daran liegen, dass der Einzugsbereich des nativen Minimums sehr schmal ist, wodurch die Wahrscheinlichkeit, durch eine Mutationen mit anschließender lokaler Optimierung in diesen Bereich zu gelangen, sehr klein sein kann. Zum anderen könnte das native Minimum entweder lediglich nur durch sehr kleine Übergangsenergien von den nächsten Minima getrennt sein oder es könnte kein echtes Minimum darstellen, in welchen der (numerische) Gradient gegen Null strebt. In der Parameter-Optimierung wurden keine Bedingungen dafür aufgestellt, dass die Summe der ersten Ableitungen der Potentialbasisfunktionen am Punkt der nativen Gesamtgeometrie $\mathbf{X} = \mathbf{X}^*$ eine Nullstelle aufweisen muss. Als Bedingung wurde lediglich festgelegt, dass der Funktionswert niedriger als alle anderen Strukturen zu sein hat. Sofern im hochdimensionalen Raum genügend nah-benachbarte Strukturen zur Parameter-Optimierung verwendet werden und das Potential nicht stark oszilliert, ist eine solche reine Funktionswert-Bedingung näherungsweise aber auch eine Forderung da-

nach, dass die native Struktur nahe einem Minimum sein muss. Jedoch wäre selbst durch eine Einbeziehung der Forderung nach einem verschwindenden Gradienten für die nativen Strukturen nicht gewährleistet, dass die zugehörigen Minima gut von anderen Minima getrennt sind. Zur Untersuchung dieser Frage wurden die nativen Strukturen, wie sie in den PDB-Dateien enthalten sind, direkt einer lokalen Kraftfeldoptimierung unterzogen. Hierbei wurde festgestellt, dass die lokal optimierten Strukturen RMSD-Abweichungen von kleiner 0.01 Å zu den ursprünglichen Koordinaten nach einer optimalen Überlagerung besitzen, woraus geschlossen werden kann, dass diese Strukturen zumindest sehr nah einem numerischen Minimum sind. Fasst man dies kurz zusammen, so können sowohl niedrige Übergangsenergien wie auch das mögliche Nichtvorhandensein eines Minimums bei $\mathbf{X}^*$ dafür sorgen, dass die lokale Optimierung nicht bei $\mathbf{X}^*$ konvergiert, sondern zu einem anderen Minimum weiterläuft, wodurch die native Struktur nicht erreicht wird.

Insgesamt kann hieraus gefolgert werden, dass die aus der globalen Optimierung resultierenden Energien (noch) nicht als Maß für die Güte einer Struktur verwendet werden können, da es keine deutliche Struktur-Energie-Korrelation im unteren Bereich der RMSD-Werte gibt. Da zusätzlich die RMSD-Werte an sich für eine Strukturvorhersage zu groß sind, lässt sich daraus schließen, dass die Parametrisierung des Potentials und/oder dessen funktionale Form unzureichend ist, was dazu führt, dass das globale Minimum nicht dem nativen Zustand entspricht, sondern dass andere Bereiche der Energiefläche energetisch günstiger als dieser sind, was als Folge des limitierten Datensatzes zur Parameter-Optimierung aufgefasst werden kann, indem diese Bereiche der Energiefläche nicht repräsentiert werden. Hierdurch kann das erstellte Programm noch nicht ohne zusätzliches Wissen über die native Struktur zur Vorhersage verwendet werden.

Struktur- und Energiebeiträge

Wie im Text oben gezeigt wurde, ist die Vorhersage des nativen Zustandes mittels der Potentialfunktion bei den ausgewählten Proteinen noch nicht gelungen. Hierbei ist es nützlich zu betrachten, welche Geometrien die optimierten Proteine besitzen und welche Energiebeiträge maßgeblich zu der Stabilisierung beitragen. Auf diese beiden Punkte wird im folgenden näher eingegangen.

Wie in der Einleitung zu dieser Arbeit beschrieben wurde (siehe Abschnitt 3.1), setzt sich die Sekundärstruktur natürlich vorkommender Proteine in ihrem nativen Zustand mit bis zu 90 % aus den regelmäßigen sich wiederholenden Einheiten der α-Helix und des β-Faltblattes sowie aus den Schlaufen und Windungen zusammen. Die tatsächliche Zusammensetzung und die

relativen Anteile dieser Strukturen können sich stark von Protein zu Protein unterscheiden (siehe hierzu beispielsweise Tab. 4.5 für die Zusammensetzung der Proteine des TOP500H-Datensatzes). Für eine erfolgreiche Proteinstrukturvorhersage ist es somit notwendig, dass im Laufe des genetischen Algorithmus Proteine mit einem hohen Anteil an Sekundärstruktur erzeugt werden. Dies sollte sich mit der Ausbildung einer kompakten Tertiärstruktur fortsetzen, die ebenfalls native Motive enthalten sollte. Da die Tertiärstruktur ein Arrangement bestehend aus den oben erwähnten regelmäßigen Sekundärstrukturen ist, ist die Bildung der Sekundärstruktur eine notwendige Voraussetzung zur Erzeugung einer nah-nativen Tertiärstruktur, weshalb an dieser Stelle diese geometrische Eigenschaft der Proteine näher betrachtet wird. Verkompliziert wird diese Sicht allerdings dadurch, dass einige Sekundärstrukturen, wie beispielsweise die β-Konformation oder die Collagen-Helix, in Isolation nicht stabil sind und nur durch Bildung einer Supersekundär- oder Tertiärstruktur entstehen können.

Betrachtet werden hier wie bereits zuvor die Proteine 1a92(A), 1pcf(A), 1ten und 1tfe. Zu diesen Proteinen wurde der Anteil θ der Aminosäuren an der Gesamtlänge berechnet, die zu einer der beiden Sekundärstrukturen α-Helix oder β-Faltblatt gehören. Zur Definition dieser Sekundärstrukturen über die C^{α}-Atome wurden die Mittelwerte aus Tab. 4.26 verwendet, wobei jeweils vier sequentielle C^{α}-Atome als strukturdefinierende Einheit angesetzt wurden. Ein C^{α}-Atom i wurde einer bestimmten Sekundärstruktur k zugeordnet, wenn der Torsionswinkel τ der Atome $i-1, i, i+1, i+2$ in das Intervall $\bar{\tau}(k) \pm \Delta\bar{\tau}_{max}(k)$ und wenn gleichzeitig die beiden C^{α}-Bindungswinkel κ der Atome $i-1, i, i+1$ sowie $i, i+1, i+2$ in das Intervall $\bar{\kappa}(k) \pm \Delta\bar{\kappa}_{max}(k)$ fielen (siehe Tab. 4.26 für die Definition dieser Werte). Insgesamt wurden drei verschiedene Anteile θ dieser so zugeordneten Aminosäuren bestimmt. Zum einen wurde über alle Individuen gemittelt, die zu einer Population einer Generation nach dem Sortierungs- und Auswahlprozess gehörten, um so einen Gesamtdurchschnittswert für den Sekundärstrukturanteil zu erhalten. Des weiteren wurde jeweils das Individuum mit dem größten sowie mit dem kleinsten Anteil an Sekundärstrukturelementen bestimmt, um eine obere und untere Schranke für den Sekundärstrukturanteil zu erhalten. Die Ergebnisse hierzu sind in Abhängigkeit von der Generation des genetischen Algorithmus in Abb. 4.46 dargestellt.

In dieser Grafik ist ersichtlich, dass, beginnend bei der ersten Generation, der Anteil an Sekundärstruktur im Verlaufe der Generationen zunimmt. Dies geschieht durch die Einführung der Sekundärstrukturen über den Mutationsoperator. Nach der anfänglichen Zunahme erreicht der Anteil ab ca. Generation 15 bis 20 ein Plateau, ab welchem nur noch kleine Änderungen im Anteil auftritt. Der Minimalanteil an Sekundärstruktur beträgt für die dargestellten Proteine zwischen 5 und 15 %. Der gemittelte Wert erreicht das Plateau bei ca. 20 % und der maximale Sekundärstrukturanteil liegt zwischen 30 und 40 %. Vergleicht man diese Grafiken mit der

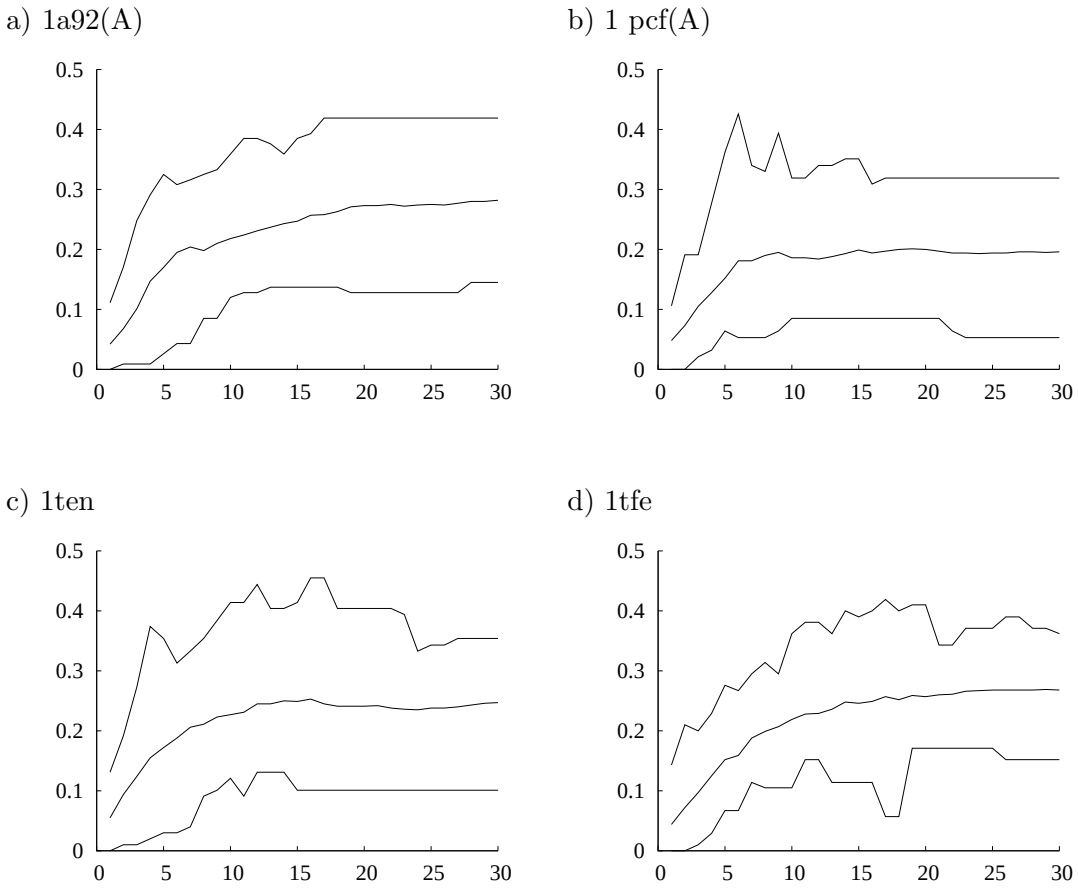


Abbildung 4.46: Anteil θ der Aminosäuren in einer α- oder β-Konformation in Abhängigkeit von der Anzahl an GA-Iterationen. Die unterste Linie stellt den Minimalanteil, die mittlere Linie den durchschnittlichen Anteil und die obere Linie den größten erhaltenen Anteil dar.

Abb. 4.41, welche den abnehmenden exponentiellen Verlauf der Gesamtenergie darstellt, so zeigt sich, dass die Zunahme an Sekundärstrukturelementen zu dieser umgekehrt proportional ist. Die Konstanz in der Zusammensetzung der Sekundärstruktur korrespondiert demnach aus dem erreichen eines lokalen Minimums, welches der genetische Algorithmus nicht mehr verlässt.
Zusätzlich zu diesen drei Anteilen wurde noch der Sekundärstrukturanteil desjenigen Individuums betrachtet, das in jeder Generation die niedrigste Energie besitzt. Dessen Anteil verhält sich sehr ähnlich dem gemittelten Anteil, wodurch es nicht explizit in der Abb. 4.46 aufgeführt wurde. Vergleicht man hierzu wieder Tab. 4.5, so sind die erreichten 20 % Sekundärstrukturanteil des energetisch günstigsten Individuums für die große Mehrzahl an den Proteinen unzureichend, einen nativen Zustand darzustellen.

Da der Anteil der Sekundärstrukturen einen zur Energie umgekehrt proportionalen Verlauf besitzt, wurden zusätzlich die Energieanteile der einzelnen Basisfunktionsklassen analysiert, die zur Gesamtenergie beitragen. Die Aufteilung der zur Gesamtenergie beitragenden Potenti-

alklassen Nahwechselwirkung, nicht-bindende Wechselwirkung zwischen den C^{α}-Atomen und zwischen den Seitenketten, Oberflächenterm und die Wasserstoffbrücken ist in Abb. 4.47 in Abhängigkeit von der Anzahl an Generationen dargestellt. Da die Energie ab ca. Generation 15 bis 20 nahezu konstant wird, wurden die letzten zehn Generationen nicht mit abgebildet. Die dargestellten Energien sind Mittelwerte, die über alle Optimierungsläufe für das jeweils energetisch niedrigste Individuum berechnet wurden.
Anhand der Abbildung 4.47 ist zu erkennen, dass für alle Proteine eine ähnliche Entwicklung der Energieanteile gegeben ist. Wie auch schon in Abschnitt 4.6 gezeigt wurde, zeigt lediglich der Oberflächenterm eine positive Energie, während die anderen Terme stabilisierende negative Energien beitragen. Am Verlauf der Energien in Abhängigkeit von der Anzahl an Generationen zeigt sich, dass die Abnahme der Gesamtenergien, wie in Abb. 4.41 dargestellt, zum überwiegenden Anteil aus der Abnahme der Oberflächenwechselwirkung resultiert. Dieser Term ist, besonders zu Anfang einer Simulation, relativ zu den anderen Termen meist sehr groß, was auf die geringe Kompaktheit und die damit verbundene Exposition der hydrophoben Aminosäuren der zufällig generierten Anfangsgeometrien zurückzuführen ist. Über ca. die ersten zehn Generationen hinweg erfolgt dann eine starke Reduktion des Oberflächenterms, so dass dieser schließlich vom Betrag her vergleichbar oder kleiner als die anderen wichtigen Energieterme wie die Nahwechselwirkungen oder die nicht-bindenden Potentiale zwischen den C^{α}-Atomen ist. Damit erfüllt dieser Potentialterm die intendierte Aufgabe, eine kompakte Faltung zu erzeugen und Proteine mit großer hydrophober Oberfläche zu destabilisieren.
Vergleicht man hierzu die anderen Energiebeiträge, so ist zunächst ersichtlich, dass die Wasserstoffbrücken und die nicht-bindenden Wechselwirkungen zwischen den Seitenketten nur einen geringen Anteil an der Gesamtenergie beitragen. Der Term der Wasserstoffbrückenbindungen trägt einen Anteil bei, der vom Betrag her eine Größenordnung kleiner als beispielsweise die nicht-bindenden Wechselwirkungen zwischen den C^{α}-Atomen ist. Das Seitenkettenpotential ist lediglich für das Protein 1a92(A) vom Betrag her mit den anderen Funktionen vergleichbar, während es in den drei anderen Proteinen von untergeordneter Bedeutung ist. Die wesentlichen negativen Energiebeiträge stammen somit aus den Nahwechseltermen und den nicht-bindenden Funktionen zwischen den C^{α}-Atomen. Betrachtet man die Änderung dieser Terme in Abhängigkeit von der Anzahl der Generationen, so zeigen die nicht-bindenden Terme zumeist die größte Änderung (siehe 1pcf(A) und 1ten und weniger deutlich für 1tfe in Abb. 4.47), während die anderen Potentiale auf der gegebenen Skala relativ konstant bleiben. Fasst man diese Ergebnisse zusammen, so kann für die globale Optimierung gefolgert werden, dass die Energiereduktion hauptsächlich durch die Minimierung des Oberflächenterms getrieben wird. Dessen Funktionen sind auf den C^{α}-Atomen zentriert, wodurch dies mit einer dichteren Packung der C^{α}-Atome einhergeht, was zur Folge hat, dass die entsprechenden nichtbindenden Wechselwirkungen vom Betrag her leicht zunehmen. Die vergleichsweise geringere Abnahme der nicht-bindenden Energie basiert darauf, dass durch die dichtere Packung, ge-

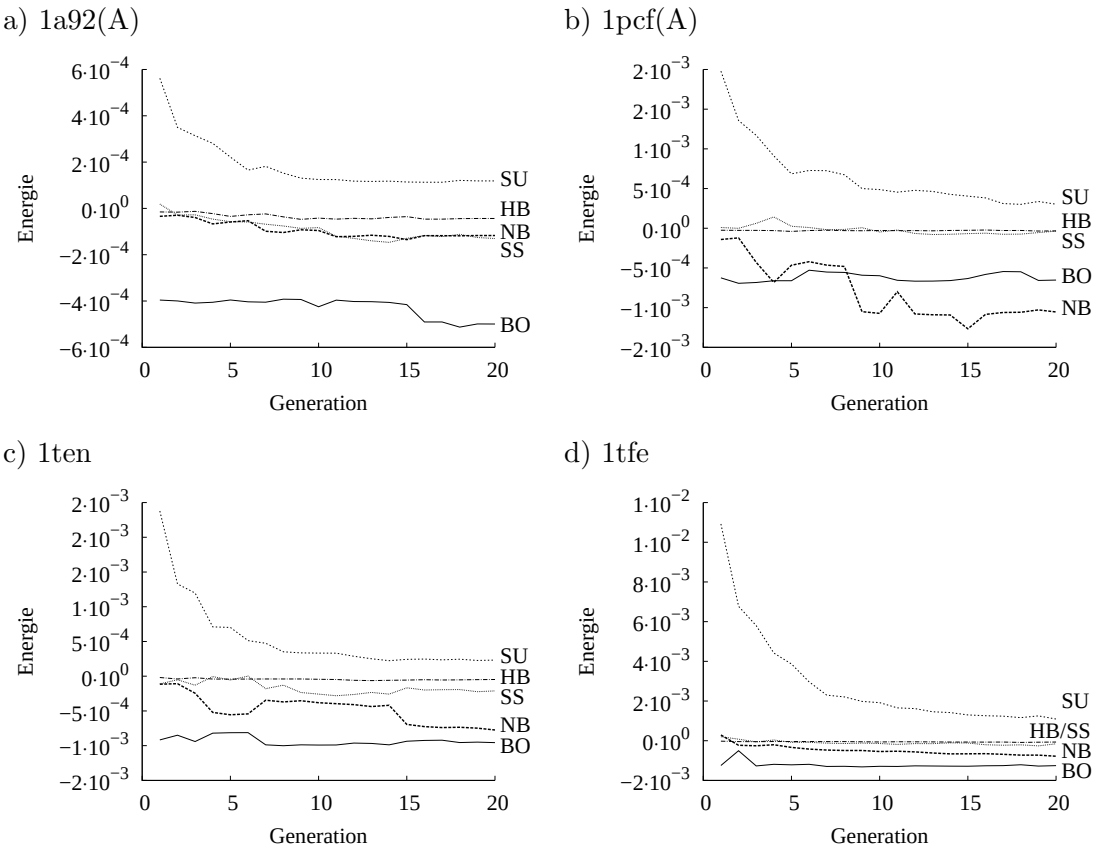


Abbildung 4.47: Beiträge der einzelnen Potentialklassen zur Gesamtenergie in Abhängigkeit von der Anzahl an Generationen im genetischen Algorithmus. Die Potentialklassen sind BO = Nahwechselwirkung (durchgezogene Linie), NB = nicht-bindend zwischen C^{α}-Atomen (breitere gestrichelte Linie), SS = nicht-bindend zwischen Seitenketten (gepunktete Linie), SU = Oberfläche (gestrichelte Linie) und HB = Wasserstoffbrücken (gestrichelte und gepunktete Linie).

trieben durch die Oberflächenreduktion, auch Aminosäuren bzw. C^{α}-Atome nah zueinander angeordnet werden, die eine ungünstige bzw. repulsive Wechselwirkung miteinander besitzen, was eine günstigere nicht-bindende Energie zwischen anderen Partnern kompensieren kann. Da das Oberflächenpotential die Koordinaten bzw. Abstände der C^{α}-Atome verwendet, hat dies nur einen indirekten Effekt auf die Seitenkettenpositionen, indem diese der Position der C^{α}-Atome "folgen" müssen, aber nicht zwangsweise eine günstige und/oder dichte Packung einnehmen, obwohl dieses während des Schrittes der lokalen Optimierung mittels eines untergeordneten genetischen Algorithmus angestrebt wird. Hierdurch und durch die optimierten Koeffizienten trägt die Seitenkettenwechselwirkung kaum zur Gesamtenergie bei. Ebenso bleiben die die Nahwechselwirkungen konstant, was damit einhergeht, dass nur wenige Aminosäuren eine Konformation einnehmen, die den regelmäßigen Sekundärstrukturen enstspricht. Dies scheint ebenfalls ein Effekt des Oberflächenpotentials zu sein, da diese Sekundärstrukturen zwar im Laufe des genetischen Algorithmus erzeugt werden, aber energetisch ungünstiger

sind und dadurch meist verworfen werden. Das Problem hierbei könnte sein, dass es keinen Prozess im Programm gibt, der eine Sekundärstruktur erzeugt und diese gleichzeitig im Proteininneren einlagert, um die Oberfläche zu reduzieren. Diese wäre besonders für ein β-Faltblatt ein wichtiger Vorgang, da dieses in einer sehr gestreckten Konformation vorliegt, wodurch auch die Abschirmung der untereinander direkt benachbarten Aminosäuren sehr gering ist. Dies kann den niedrigen Gesamtanteil an Aminosäuren in einer regelmäßigen Sekundärstruktur erklären, was im Zusammenhang mit Abb. 4.46 besprochen wurde. Ebenso resultiert hieraus der geringe Beitrag der Wasserstoffbrückenenergie, da sich ohne wiederholende Sekundärstrukturen nur wenige Wasserstoffbrücken ausbilden. Dies begründet auch den energetischen Unterschied, der sich beispielsweise deutlich zu Abb. 4.31 zeigt, in welcher die relativen Energiebeiträge der nativen Proteine dargestellt wurden, wobei für diese die Wasserstoffbrückenenergie einen wesentlichen Beitrag lieferte.

5 Zusammenfassung und Ausblick

Obwohl seit einigen Jahrzehnten bereits in den verschiedensten Arbeitsbereichen intensive Forschung am Proteinfaltungsproblem von unterschiedlichst orientierten Wissenschaftlern mit verschiedenen Intentionen und Zielsetzungen betrieben wird, entzieht sich dieses Problem immer noch einer umfassenden Lösung. Und obwohl in bestimmten spezielleren Fragestellungen, wie beispielsweise der Aufklärung unbekannter Proteinstrukturen oder die Erklärung spezifischer Reaktionen, an denen Proteine beteiligt sind, große Fortschritte erzielt wurden, die sowohl für das theoretische Verständnis wie auch für praktische Anwendungen wichtige Erfolge darstellten, fehlt eine umfassende Erklärung für die sehr wichtige Grundlagenfrage über den Zusammenhang zwischen der dreidimensionalen Struktur eines Proteins und seiner Aminosäuresequenz, welche in der DNA codiert gespeichert vorliegt und die Funktion und die Eigenschaften eines Proteins bestimmt. Mathematisch lässt sich dieses Problem verallgemeinert über das Fehlen einer Projektionsvorschrift vom eindimensionalen Sequenzvektorraum, dessen entscheidende Information in der Reihenfolge der Vektorelemente liegt, auf den dreidimensionalen Geometrieraum beschreiben. Prinzipiell existiert diese Projektionsvorschrift bzw. ein entsprechendes Konversionsverfahren, da über das Wissen der Quantenmechanik die gesamten Wechselwirkungen innerhalb des Proteins und mit seiner Umgebung bekannt sind. Ebenso sind prinzipiell hinreichende Verfahren bekannt, über die Quantenmechanik den nativen Zustand des Proteins bestimmen zu können. Die Probleme, die einer Lösung hier aber im Weg stehen, sind hauptsächlich technischer Natur, die in der Architektur heutiger Computersysteme begründet liegen. Für eine umfassende Berechnung des Proteins und Analyse der freien Energiefläche unter Einbeziehung der Umgebung, welche für eine realistische Beschreibung notwendige Voraussetzung wäre, da die Thermodynamik der Faltung auf der Minimierung der freien Energie des Gesamtsystems beruht, stehen zwar effiziente mathematische Algorithmen zur Verfügung, aber aufgrund der sehr hohen Komplexität, die durch die Anzahl an Freiheitsgraden der Moleküle gegeben ist, ist es mit heutiger Computertechnik aussichtslos, hochgenaue quantenmechanische Rechnungen an derart großen Systemen in vernünftiger Zeit durchführen zu können. Aufgrund dieser technischen Schranken ist man in der Frage der Proteinfaltung und der Identifikation des nativen Zustandes wie auch in vielen anderen Bereichen auf Vereinfachungen des Problems angewiesen, wodurch die Dimensionalität des zu explorie-

renden Variablenraumes erniedrigt wird. Die Umsetzung dieser Reduktion der Freiheitsgrade wird mit unterschiedlichen Methoden erreicht und hängt in der Regel in direkter Form mit der gewählten Energie- oder Bewertungsfunktion zur Beschreibung der Wechselwirkungen zusammen. Dieses Vorgehen führt zumeist dahin, die Umgebungseffekte zu ignorieren oder sie in einer impliziten Form zu erfassen und die Anzahl an Interaktionspunkten im Protein zu verkleinern, indem Atome ausgelassen oder mit anderen zusammengefasst werden. Dies entspricht einer Glättung der der Faltung eines Proteins zugrundeliegenden Energiefläche, wodurch die Anzahl an möglichen charakteristischen Punkte auf der Fläche kleiner wird. Die für die vereinfachten Proteinmodelle benutzten Energiefunktionen sind in ihrer mathematischen Formulierung sehr unterschiedlich. Diese reichen von elaborierten Kraftfeldern, über diskretisierte Abstandspotentiale bis hin zu einfachen Kontaktpotentialen. Diese Kombination von vereinfachten Modellen und Energiefunktionen ermöglicht es, auch für größere Proteine die bekannten mathematischen Verfahren zur Analyse und Vorhersage der Proteinfaltung anzuwenden. Die Leistung dieser Verfahren wird heutzutage in regelmäßig stattfindenden Wettbewerben verglichen (siehe z. B. CASP [266]). Besonders erfolgreich zeigten sich hierbei immer wieder statistische Verfahren, die auf der massiven Auswertung von heutzutage leicht zugänglichen Proteinstrukturdaten beruhen. Ein Aspekt des Erfolges dieser Verfahren resultiert aus der Tatsache, dass, obwohl kontinuierlich viele Proteinstrukturen experimentell aufgeklärt werden, nach einer gegenwärtigen Einschätzung auf Basis der zur Zeit zur Verfügung stehenden Daten nur eine begrenzte Anzahl an Faltungsklassen existieren. Hierdurch ist es oft möglich, aus den Informationen der Geometrie bekannter Sequenzen auf die Geometrie ähnlicher Sequenzen zu schließen. Diese Methoden, so erfolgreich sie für die große Anzahl an Anwendungen sind, haben zwei markante Nachteile. Zunächst funktionieren sie wie bereits erwähnt nur für Strukturen, die bereits bekannt sind. Neue Faltungsklassen lassen sich somit nur unwahrscheinlich vorhersagen. Zweitens lösen sich nicht das Grundproblem zur Erklärung des kausalen Zusammenhangs zwischen der Sequenz und der dreidimensionalen Proteinstruktur, sondern reduzieren es auf eine phänomenologische Beschreibung.

Im Hinblick auf diese beschriebenen möglichen unterschiedlichen Extrema zur Erfassung des Proteinfaltungsproblems, die zwischen einer hochgenauen quantenmechanischen Beschreibung und den sehr stark vergröberten Modellen liegen, wurde in dieser Arbeit eine Vorgehensweise gewählt, die ebenfalls eine dem Problem angemessene notwendige Modellvereinfachungen beinhaltet, um längere Sequenzen technisch zugänglich zu manchen. Für die Energiefunktion wurde ein neuartiger Kombinationsansatz bestehend aus der Verwendung statistischer Datenbank-Informationen und einer Potentialentwicklung über eine Linearkombination von Basisfunktionen realisiert. Im folgenden werden hierzu die wichtigsten Punkte dieser Arbeit zusammengefasst.

Das generelle Ziel war es, einen Algorithmus zu entwickeln, der für eine vorgegebene Aminosäuresequenz im Idealfall eine Vorhersage für den nativen Zustand erzeugt bzw. für diesen mögliche Kandidaten anbietet, was mit der Entwicklung einer algorithmischen Methode einherging, die native Geometrie in einer möglichst großen Menge falscher oder alternativer Proteinstrukturen identifizieren zu können. Als Basis für jegliche verwendete Information musste zunächst ein Proteindatensatz mit verlässlichen Strukturen ausgewählt werden. Hierfür wurde der TOP500H-Proteindatensatz benutzt, welcher sehr viele hochaufgelöste Kristallstrukturen enthält, die anhand vieler Gütekriterien ausgewählt wurden. Aus diesem Satz wurden schließlich 48 Proteine verwendet, die die unterschiedlichen Zusammensetzungen an Sekundär- und Tertiärstruktur repräsentierten sowie kompatibel mit dem gewählten Proteinmodell waren. Für dieses Modell wurde aufgrund der angestrebten Größe der Zielproteine, welche einhundert oder mehr Aminosäuren enthalten sollten, damit viele biologisch relevante Sequenzen behandelt werden können, eine vereinfachte bzw. vergröberte Darstellung verwendet. Das Modell wurde in Anlehnung an andere in der Literatur publizierter Proteinmodelle konstruiert und enthielt schließlich drei Wechselwirkungszentren für eine Aminosäure, welche das C^{α}-Atom, den Schwerpunkt der Seitenkette und das Zentrum der Verbindungslinie zwischen zwei sequentiellen C^{α}-Atomen beinhalten. Um hierbei die Anzahl an möglichen Freiheitsgraden klein zu halten, wurde zunächst untersucht, ob die Position des Seitenkettenschwerpunktes direkt über die Koordinaten der C^{α}-Atome definierbar sei. Da aber, wie sich zeigte, die Informationen der Rückgratkoordinaten nicht ausreichend für eine Positionierung des Seitenkettenschwerpunktes waren, wurde das Modell in zwei Schritten auf eine flexible Seitenkettenpositionierung über ein proteininternes Koordinatensystem erweitert, das auf einer Clusteranalyse experimentell bekannter Strukturen basierte. Dieses Seitenkettenmodell kam ohne aufwendigere Berechnungen von Bindungswinkeln oder Torsionswinkeln aus, wodurch die numerische Behandlung schneller ist. Zusätzlich zu diesem Clustermodell wurden aus dem TOP500H-Proteindatensatz weitere statistische Informationen gewonnen, die beispielsweise bevorzugte oder verbotene Abstände in den verwendeten Proteinen angaben, die in den späteren Programmteilen Verwendung fanden.

Da das Ziel der Arbeit eine Unterscheidung zwischen nativen und nicht-nativen Strukturen zum Ziel hatte, wurde die Literatur nach erhältlichen Strukturdatensätzen von falschen Geometrien hin überprüft. Hier zeigte sich, dass, obwohl einige Datensätze existieren, eine nicht geringe Anzahl der darin enthaltenen Proteine nicht verwendet werde konnten. Die auftretenden Probleme waren dabei, dass die Anzahl an falschen Strukturen zu gering war, dass Proteine häufig nicht mit dem angesetzten Proteinmodell kompatibel waren, indem beispielsweise Disulfidbrücken und/oder Fremdmoleküle enthalten waren, und dass die zur Verfügung gestellten Strukturen häufig Fehler wie Kettenunterbrechungen, abweichende Anzahl an Aminosäuren im Vergleich zur nativen Struktur oder Atomkollisionen enthielten. Aus diesen Gründen wurden in dieser Arbeit Programme zur Erzeugung eigener falscher Strukturen erstellt.

Hierbei wurden zwei verschiedene Sätze an falschen Geometrien mit zwei unterschiedlichen Methoden generiert. Diese wurden zusätzlich durch einen Satz an publizierten Literaturstrukturen ergänzt, um eine möglichst breite Informationsbasis bereitstellen zu können. Das Ziel zur Verwendung dieser drei Sätze, die im späteren Verlauf um einen vierten Satz erweitert wurden, war eine möglichst umfassende Abdeckung der unterschiedlichen Konformationsräume der Energiefläche der Proteine, die vom nah-nativen Bereich bis hin zu den schlecht gefalteten und wenig kompakten Strukturen reichen.
Zur Unterscheidung von falschen und nativen Strukturen ist eine Bewertungsfunktion erforderlich, die in der Lage ist, in mathematischer Form die nativen Strukturelemente zu identifizieren. Hierzu ist es notwendig, dass die Bewertungsfunktion einer bestimmten Aminosäuresequenz die richtige Sekundärstruktur zuordnet, welche wie in der Literatur gezeigt wurde, hauptsächlich über Wechselwirkungen von in der Sequenz benachbarten Aminosäuren bestimmt wird und somit einer lokalen Codierung entspricht. Zusätzlich muss die Anordnung der Sekundärstrukturen zur Supersekundär- und Tertiärstruktur erkannt werden, wobei in der Sequenz weit entfernte Aminosäuren zusammengelagert werden. Für die Gestaltung einer solchen Bewertungsfunktion bzw. eines solchen Potentials wurde ein gemischter Ansatz aus statistischer Datenanalyse und Entwicklung von Potentialfunktionen mittels Linearkombination bestimmter Basisfunktionen gemacht. In der Sequenz nah benachbarte Aminosäuren zeigten bestimmte bevorzugte geometrische Anordnungen, die statistisch signifikant waren, so dass diese zur Definition von Potentialfunktionen genutzt wurden, welche als Nahwechselwirkungen im Kraftfeld enthalten waren. Je größer der Abstand zwischen den Aminosäuren in der Sequenz war, um so unkorrelierter waren deren geometrische Anordnungen, so dass für in der Sequenz entfernte Aminosäuren keine signifikanten statistischen Muster erhalten wurden. Da zusätzlich aufgrund der Vergröberung des Modells und somit der Mittelung über mehrere unterschiedliche atomare Wechselwirkungen die effektive Form der Potentiale nicht bekannt war, wurden die Potentiale zwischen den entfernten Aminosäuren als zu bestimmende Linearkombinationen von Basisfunktionen angesetzt, um die optimale funktionale Form dieser Wechselwirkungen bestimmen zu können. Dieser Ansatz wurde für die Wechselwirkungen zwischen den C^{α}-Atomen und zwischen den Seitenketten gemacht. Neben diesen Funktionen wurden weiterhin noch ein Oberflächenterm, um die kompakte Faltung und die Einlagerung der hydrophoben Aminosäuren im Proteininneren zu erreichen, sowie ein Potential zur Beschreibung der Wasserstoffbrücken zwischen Rückgratamideinheiten eingeführt, das eine besondere Relevanz zur Bildung der Sekundärstrukturen besitzt.
Die für dieses Potential notwendigen Parameter in den Basisfunktionen, wie beispielsweise die Gleichgewichtsabstände, wurden direkt an die vorhandenen statistischen Daten angepasst, während die Gewichtungskoeffizienten der Energieterme bzw. Funktionsklassen untereinander über die Formulierung und Lösung eines linearen Optimierungsproblems bestimmt wurden. Hierzu wurde ein Ungleichungssystem aufgestellt, in welchem die Elemente der Koeffizien-

tenmatrix die Energien der einzelnen Potentialklassen waren, welche durch die Auswertung der Basisfunktionen berechnet wurden. Der Lösungsvektor des Ungleichungssystems enthielt die zu bestimmenden Gewichtungskoeffizienten der Basisfunktionsklassen. Die zum linearen Optimierungsproblem gehörende zu minimierende Zielfunktion war durch die Summe der Gewichtungskoeffizientenbeträge gegeben. Die Differenz der Energien der verwendeten falschen Proteine in Vergleich zu den nativen Strukturen wurden als Nebenbedingungen des linearen Problems eingebunden. Diese Nebenbedingungen forderten, dass die Energie der nativen Struktur stets niedriger als alle Energien der falschen Strukturen zu sein hatte, was im Sinne der thermodynamischen Hypothese ist, nach welcher der native Zustand das globale Minimum der freien Energie des Proteinsystems ist, so dass alle andersartig gefalteten Geometrien eine höhere Energie besitzen sollten. Insgesamt konnten für die Nebenbedingungen über 700000 falsche Proteinstrukturen verwendet werden, was im Vergleich mit der Literatur eine wesentlich größere Anzahl ist als durchschnittlich für eine solche Aufgabe verwendet wird bzw. verwendet werden kann. Für die Parameter-Optimierung wurde das Programm BPMPD verwendet, das über einen Innere-Punkte-Algorithmus die Lösung des linearen Problems bestimmt. Die bearbeiteten Ungleichungssysteme waren dabei stets lösbar, so dass hieraus gefolgert werden konnte, das im Rahmen der für das Problem gegebenen Strukturen, das Proteinmodell in Verbindung mit der Potentialfunktion genau genug war, die falschen von den nativen Geometrien zu unterscheiden.

Über die Optimierung der Gewichtungskoeffizienten und über die Definition der Basisfunktionen wurden für die Proteine Energieflächen erzeugt, auf welchen im Anschluss die zugehörigen Sequenzen global optimiert wurden. Dies diente der Analyse, ob die generierten Energieflächen tatsächlich den nativen Zustand als globales Minimum enthielten. Denn trotz der großen Anzahl an verwendeten falschen Geometrien, konnte aufgrund der Komplexität des möglichen Konformationsraumes einer Proteinsequenz nicht garantiert werden, dass dieser hinreichend abgedeckt wurde, wodurch in den nicht erfassten Bereichen die Potentialfunktion energetisch günstigere Minima enthalten kann. Hierfür wurde ein genetischer Algorithmus zusammen mit den notwendigen Berechnungsroutinen für das Potential und den Gradienten sowie der lokalen Optimierung vollständig neu implementiert. Aufgrund rechenzeitlicher Erwägungen wurden von den 48 zur Parameter-Optimierung genutzten Sequenzen nur die 27 kürzesten Proteine für die globale Geometrieoptimierung verwendet. Die im genetischen Algorithmus explorierten Minima, die eine niedrigere Energie als die zugehörige native Struktur besaßen, wurden zu einer weiteren Gruppe falscher Strukturen zusammengefasst, womit zusammen mit den oben angegebenen drei Strukturdatensätzen dann insgesamt vier Datensätze mit falschen Geometrien für die Parameter-Optimierung zur Verfügung standen.

Dieses Vorgehen wurde zu einem iterativen Prozess erweitert, wobei ein Zyklus aus den Schritten Parameter-Optimierung auf Basis der ausgesuchten falschen Strukturen, der globalen Geometrieoptimierung auf den generierten Flächen und Zusammenstellung der neuen Menge

der falschen Strukturen für die folgende wiederholte Parameter-Optimierung bestand. Durch diesen Prozess wurden die im genetischen Algorithmus erhaltenen neuen niedrigen Minima über das Niveau des nativen Zustandes gehoben. Dieser Zyklus wurde mit dem Ziel implementiert, zu einer Konvergenz zu gelangen, bei welcher im Schritt der globalen Geometrieoptimierung keine relevanten Minima mehr gefunden werden, die eine niedrigere Energie als der native Zustand besitzen, wodurch eine Unterscheidung bzw. Identifikation dieses Zustandes von anderen möglich wäre. Insgesamt wurden drei Iterationen durchgeführt, wobei die Ergebnisse keine eindeutige Tendenz zeigten, ob sich die Vorhersageergebnisse verbesserten oder verschlechterten, da die Fluktuationen in den Ergebnissen ebenso auch auf den statistischen Charakter des genetischen Algorithmus zurückgeführt werden konnten. Betrachtet man die strukturellen Differenzen der resultierenden Geometrien so sind diese für die längeren Proteinsequenzen noch zu weit vom nativen Zustand entfernt, als dass eine Vorhersage als erfolgreich hätte qualifiziert werden können. Für die kurzen Sequenzen konnten aber Geometrien erzeugt werden, die Übereinstimmungen mit der nativen Geometrien zeigten, so dass für diese RMSD-Abweichungen im Bereich von 4 bis 5 Å erreicht wurden. Hier bestand allerdings das Problem, dass es keine eindeutige Energie-Struktur-Korrelation gab, wodurch nicht wie in der Zielstellung formuliert anhand der Potentialfunktion auf die native Geometrie geschlossen werden konnte. Hier konnte gezeigt werden, dass mehrere Probleme dieses Ergebnis bedingten. So enthielten die im iterativen Prozess erzeugten Energieflächen sehr viele energetisch günstigere Minima, wodurch die Geometrieoptimierung nicht beim nativen Zustand endete. Dies schien ein Problem der hochdimensionalen Energiefläche zu sein, so dass aufgrund der computertechnischen Beschränkung nicht genügend falsche Strukturen verwendet werden konnten, um die vorhandenen energetisch günstigeren Minima abzudecken. In direktem Zusammenhang hiermit wurde ein Sekundärstrukturerzeugungsproblem erkannt, wodurch die optimierten Strukturen einen zu großen Anteil an nicht-regelmäßigen Sekundärstrukturen enthielten, was eine Folge des Solvatations- bzw. Oberflächenpotentials der hydrophoben Aminosäuren zu sein schien, indem dieses die Energieminimierung im Verlauf des genetischen Algorithmus dominierte und so eine Sekundärstrukturerzeugung durch die hierfür angesetzten anderen Terme wie beispielsweise die Nahwechselwirkungen verhinderte. Neben der globalen Geometrieoptimierung wurde zur Bestimmung der Unterscheidungsleistung des entwickelten Proteinpotentials noch ein Erkennungstest (*ranking*) an vier publizierten Strukturdatensätzen durchgeführt, die nicht für die Parameter-Optimierung verwendet wurden, und dies mit Literaturergebnissen verglichen. Hier zeigte sich insgesamt, dass das Potential in der Lage war, eine große Anzahl an nativen Proteinen in den Mengen der falschen Strukturen zu identifizieren. Markant an diesen Ergebnissen war ein fehlendes qualitatives Mittelfeld in der Bestimmung des Ranges der nativen Struktur, da diese entweder sehr gut oder nur sehr schlecht erkannt wurde. Eine Analyse der den nativen Zustand von den anderen unterscheidenden Energiebeiträgen zeigte, dass vor allem die Nahwechselterme sowie

das Wasserstoffbrückenpotential für ein Nicht-erkennen verantwortlich waren. Dies beruhte hauptsächlich auf einer unterschiedlichen Zusammensetzung der Sekundärstrukturen im nativen Protein und in den falschen Strukturen. So enthielten die falschen Strukturen häufig ideale geometrische Anordnungen oder eine größere Anzahl an Wasserstoffbrückenbindungen, wodurch diese insgesamt energetisch günstiger waren. Der Vergleich der Erkennungstestergebnisse mit anderen publizierten Potentialen zeigte, dass das hier entwickelte Kraftfeld ein Identifikationsniveau im oberen Mittelfeld besitzt. Die Schwächen des Kraftfeldes bei der vergleichenden Identifikation des nativen Zustandes beruhten aber nicht auf größeren prinzipiellen Unzulänglichkeiten, sondern eher auf Detailfragen der Formulierung und Parametrisierung der Basisfunktionen, so dass hier durch Änderungen noch bessere Erfolge erzielt werden könnten.

Für die Weiterentwicklung des in dieser Arbeit grundlegend implementierten iterativen Parameter-Optimerungsprozesses und globalen Geometrieoptimierung wurden einige Punkte identifiziert, die in einer zukünftigen Arbeit angegangen werden könnten, wofür im folgenden einige Anstöße gegeben werden.
Die Grundlage der Parameter-Optimierung bildeten die eigengenerierten, die aus externen Quellen hinzugenommenen und die aus der Geometrie-Optimierung stammenden Struktur-Datensätze. Für deren Verwendung könnte eine weitere Analyse von Vorteil sein, indem beispielsweise die eigengenerierten Strukturen einer Kräfterelaxation entweder mit dem vorhandenen oder einem anderen Kraftfeld unterzogen werden, um direkt Geometrien zu erhalten, die lokalen Minima auf der Energiefläche entsprechen. Gleichzeitig könnte die geometrische Distanz der Strukturen untersucht werden, um so beispielsweise redundante Strukturen zu eliminieren, die die gleichen oder sehr ähnliche Bereiche auf der Energiefläche abdecken, da die bisherige Analyse nur die relativen Energien als vergleichendes Kriterium verwendete. Hier könnte beispielsweise auch eine Cluster-Analyse der Proteine hilfreich sein, indem Strukturen mit einem ähnlichen oder redundanten Informationsgehalt entfernt werden, wodurch andere Proteingeometrien zur Parameter-Optimierung hinzugenommen werden können, die bisher aufgrund der beschränkten Anzahl an Nebenbedingungen nicht einbezogen werden konnten. Dies könnte auch zu einer Analyse erweitert werden, die nicht nur rein strukturelle Differenzen berücksichtigt, sondern ebenso überprüft, welche Strukturen zu einem Minimum bzw. zu dessem Einzugsgebiet auf der Energiefläche gehören.
Die Zielfunktion für die Parameter-Optimierung beinhaltete die Minimierung der Gewichtungskoeffizienten der Basisfunktionen unter Berücksichtigung der Nebenbedingungen, die falschen Strukturen gegenüber den nativen Proteinen energetisch zu destabilisieren. Hier könnten zusätzlich Gradienten-Informationen hinzugenommen werden, um zu gewährleisten, dass die Datenbank-Strukturen der nativen Proteine Minima auf der Energiefläche entsprechen. Ein Ansatz hierzu wäre die Forderung, dass die Summe der Matrixelemente des Gra-

dienten verschwindet bzw. dass diese Summe numerisch kleiner als ein vorgegebener Wert zu sein hat.

Bei der Analyse der Energieanteile der Funktionsklassen zu den Erkennungstests zeigten sich für die Definition der Potentialbasisfunktionen ebenfalls mögliche Verbesserungsansätze. Die Nahwechselwirkungsterme destabilisierten einige natürlich vorkommende β-Faltblattstrukturen, wenn diese in einer sehr gestreckten Konformation vorlagen bzw. wenn der Bindungswinkel dreier sequentieller C^{α}-Atome sehr groß war. Hierdurch verschlechterte sich die Energie des nativen Proteins gegenüber den Alternativgeometrien. Dies war ein Problem der Doppelminimumdefinition der entsprechenden Nahwechselwirkungsbasisfunktion, da sich die gestreckte Konformation aufgrund der optimierten funktionalen Form im repulsiven Bereich dieses Potentials befand. Hier sollte die funktionale Form angepasst werden, indem beispielsweise das Minimum für die β-Konformation verbreitert oder durch eine andere funktionale Form ersetzt wird.

Für das Wasserstoffbrücken-Potential wurden insgesamt drei Basisfunktionen angesetzt, wobei sich zeigte, dass lediglich die abstandsabhängige Funktion nach der Parameter-Optimierung für die Gesamtenergien eine Rolle spielte. Hier könnten die zugehörigen Basisfunktionen eventuell verbessert werden, indem die nach der Parameter-Optimierung als weniger relevant gewichteten Winkelfunktionen direkt in die Abstandsfunktion miteinbezogen, durch andere Funktionen ersetzt oder ausgelassen werden. Ebenso ist es möglich die abstands- oder winkelabhängigen Wasserstoffbrücken-Funktionen analog zu den nicht-bindenden Wechselwirkungen über einen Linearkombinationsansatz zu formulieren, dessen effektive funktionale Form sich erst über die Parameter-Optimierung ergibt. Bisher wurden die Wasserstoffbrückenfunktionen direkt an die statistischen Daten angepasst.

Das Oberflächen- oder Solvatationspotential erfüllte die ihm zugedachte Aufgabe, nichtkompakte Strukturen zu destabilisieren und hydrophobe Aminosäuren im Inneren des Proteins abzuschirmen. Hier stellte sich das Problem, dass dieser Energieterm im Vergleich zu den anderen Energiebeiträgen sehr dominant war. Verbesserungen könnten hier erreicht werden, indem beispielsweise eine genauere Methode zur Berechnung der lösungsmittelzugänglichen Oberfläche angewendet wird, von welcher der Funktions- bzw. Energiewert abhängt. Der Vorteil des in dieser Arbeit verwendeten probabilistischen Ansatzes liegt in dessen einfach zu implementierender Methodik und vor allem in dem benötigten geringen rechentechnischen Aufwand. Der Nachteil dieser Methode besteht darin, dass, während aus Fehlerkompensationsgründen die Gesamtoberfläche eines Proteins meist sehr gut im Bereich von ein bis zwei Prozent Ungenauigkeit reproduziert wird, die Oberflächenwerte für einzelne Aminosäuren stärker fehlerbehaftet sein können. Da aber die Potentialberechnung auf den zugänglichen Oberflächen einzelner Aminosäuren beruht, könnten die durch die Methode erhaltenen Fehler Auswirkungen auf die Parametrisierung des Potentials haben. Zur Berechnung der Oberfläche existieren in der Literatur viele Methoden, die exaktere Ergebnisse liefern, wobei aber

der Gewinn an Genauigkeit gewöhnlich durch einen wesentlich größeren rechentechnischen und damit zeitlich größeren Aufwand erkauft wird. Hierzu müsste untersucht werden, ob eine quantitativ korrektere Oberflächenmethode sich auf die Parameter-Optimierung und die Leistung des Potentials auswirkt.
Neben diesen Proteinmodell- und Energiefunktionsverbesserungsmöglichkeiten besitzt auch der genetische Algorithmus Elemente, die genauer analysiert werden könnten, um die Leistung des Programmes zu steigern. Generell können die vor Beginn einer Rechnung festgelegten Programmparameter wie Individuen- und Generationenanzahl oder Mutationswahrscheinlichkeit selbst einer Voroptimierung unterzogen werden, um die Programmparameter zu bestimmen, die in Verbindung mit dem Kraftfeld die besten Vorhersageergebnisse erzielen. Auch die Operatoren, die für die Kreuzung oder Mutation von Individuen zuständig sind, könnten verbessert werden. Wie im vorangegangen Kapitel dargestellt wurde, bestand hier beispielsweise das Problem, dass die aus diesen Operatoren resultierenden Geometrien zu unkompakt waren, wodurch sie aus energetischen Gründen verworfen wurden, was in der Konvergenzphase des genetischen Algorithmus zu Strukturen mit einem niedrigen Sekundärstrukturgehalt führte. Hier könnten Operatoren implementiert werden, die direkt auch auf die Tertiärstruktur wirken und kompakte Geometrien produzieren.

Neben diese Änderungen könnten weitere umfangreichere Modifikationen des Programmes vorgenommen werden. Beispielsweise ließe sich das Proteinmodell bzw. -potential und der genetische Algorithmus auf die Behandlung von Disulfidbrücken erweitern, wie ebenso auf Fremdatome und -moleküle oder Proteinkomplexe.
Der genetische Algorithmus wie auch Teile der Potential- und Gradientenberechnung lassen sich parallelisieren, um deren Berechnung auf mehrere Prozessoren verteilen zu können. Hierzu ließen sich dann beispielsweise auch für ein Protein mehrere Populationen gleichzeitig behandeln, wobei zwischen diesen dann ein Informationsaustausch stattfinden könnte.
Ein weiteres Ziel sollte in zukünftigen Arbeiten die verbesserte Erzeugung von Sekundär- und Tertiärstrukturen im globalen Optimierungsprozess sein. Ein Möglichkeit hierzu wäre beispielsweise ein Potential einzuführen, dass auf mehrere Aminosäuren gleichzeitig wirkt und die Ausbildung von größeren Bereichen mit regelmäßiger Struktur bevorzugt.

Generell sind die Möglichkeiten, das vorhandene Potential und die Optimierungsmethoden zu verändern, vielfältig, so dass hier nur Einzelbeispiele gegeben werden konnten. Welche Änderungen sinnvoll und erfolgreich sind, muss durch die praktische Ausführung der Methoden getestet und mit der Literatur abgeglichen werden. Die Erfahrungen dieser Arbeit zur Definition des Potentials zeigten, dass vermeintlich logische Ansätze häufig aufgrund der komplexen Zusammenhänge zwischen der Wahl der Basisfunktionen, der Parameter-Optimierung und der globalen Geometrieoptimierung zu unbrauchbaren Resultaten führten, welche, um

den Rahmen der präsentierten Arbeit nicht zu überspannen, an den entsprechenden Stellen nicht mitaufgeführt wurden.
Insgesamt konnte in dieser Arbeit somit ein von Grund auf bisher noch nicht verwendetes Potential gestaltet werden, das einen neuartigen kombinierten Ansatz aus statistischen Informationen und linearkombinierten Basisfunktionen verwendete, welche eine glatte Energiefläche erzeugten, die an allen Punkten differenzierbar ist. Hinzu kam eine umfassende Analyse experimentell bekannter Datenbankstrukturen, die Generierung neuer falscher Proteinstrukturen und mehrerer neuer Ansätze zur Beschreibung der Seitenkettenposition. Diese Arbeiten und Entwicklungen wurden genutzt, um hierauf ein globales Geometrieoptimerungsprogramm zu implementieren, welches basierend auf der Sequenz eines Proteins Vorschläge für den nativen Zustand erzeugt. Zusätzlich wurde die Leistung des Potentials ausführlich gegen publizierte Literaturpotentiale getestet, die auf vergleichbaren Proteinmodellen beruhten. Zusammengefasst zeigten die Resultate bereits teilweise eine gute Vergleichbarkeit mit Literaturergebnissen, sind aber in den aufgezeigten problematischen Bereichen entwicklungsfähig und verbesserbar. Für ein von Null auf vollständig neu entwickeltes Programm sind die erhaltenen Ergebnisse aber sehr ermutigend, so dass zukünftige Entwicklungen dieses Programm auf ein vollständig vergleichbares Niveau heben könnten.

6 Anhang

Dieser Anhang enthält zum einen die ausgewerteten Minimalabstände zwischen den Wechselwirkungszentren, welche im genetischen Algorithmus für die Definition der repulsiven Potentiale verwendet wurden (siehe Abschnitt 4.7.3) sowie zum anderen die Daten der für die Seitenketten-Positionierung bestimmten Clusterzentren (siehe Abschnitt 4.2.3).

6.1 Minimalabstände

Dieser Abschnitt führt die aus dem TOP500H-Proteindatensatz ermittelten Mindestabstände zwischen zwei Seitenketten (Tab. 6.1), zwischen einer Seitenkette und einem C^{α}-Atom (Tab. 6.2) und zwischen zwei C^{α}-Atomen (Tab. 6.3) auf. Diese werden für jedes der 20·20 Aminosäurepaare im folgenden in tabellarischer Matrixform angegeben, wobei jeweils nur das untere Dreieck und die Diagonale ausgeführt ist. Das obere Dreieck ergibt sich jeweils aus der Symmetrie der Matrix.

Ala	3.42																			
Cys	3.52	3.58																		
Asp	3.17	3.75	3.06																	
Glu	3.36	3.16	3.61	3.46																
Phe	3.44	3.50	3.72	3.60	3.64															
Gly	-	-	-	-	-	-														
His	3.39	3.33	3.19	3.70	3.59	-	3.75													
Ile	3.62	3.86	3.85	3.76	3.70	-	3.72	4.17												
Lys	3.29	3.87	3.26	3.44	3.62	-	3.76	4.01	3.64											
Leu	3.75	3.85	3.81	4.01	3.69	-	3.95	4.02	4.03	4.08										
Met	3.53	3.55	3.73	3.91	3.99	-	3.88	4.01	3.87	3.98	4.32									
Asn	3.35	3.46	3.36	3.52	3.63	-	3.47	3.73	3.67	3.83	3.62	3.62								
Pro	3.60	3.78	3.49	3.78	3.84	-	3.73	4.15	3.92	4.23	4.05	3.29	4.00							
Gln	3.37	3.85	3.38	3.59	3.70	-	3.80	4.01	3.60	3.79	3.88	3.48	3.89	4.20						
Arg	3.30	3.59	3.33	3.48	3.60	-	3.58	3.93	3.38	3.73	3.86	3.46	3.91	3.88	3.70					
Ser	3.17	3.34	3.19	3.24	3.52	-	3.43	3.71	3.42	3.74	3.63	3.37	3.58	3.01	3.27	3.06				
Thr	3.45	3.55	3.41	3.52	3.36	-	3.55	3.73	3.13	3.71	3.78	3.45	3.81	3.41	3.34	3.13	3.26			
Val	3.53	3.84	3.77	3.60	3.80	-	3.71	4.08	3.99	4.04	3.96	3.65	3.93	3.90	3.45	3.35	3.72	3.98		
Trp	3.37	3.80	3.49	4.11	3.80	-	3.45	3.84	3.51	3.82	3.91	3.44	3.65	3.53	3.54	3.53	3.58	3.71	4.58	
Tyr	3.29	3.86	3.71	3.79	3.49	-	3.51	3.77	3.69	3.96	3.73	3.70	3.67	3.74	3.52	3.31	3.82	3.51	3.79	3.69
	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr

Tabelle 6.1: Minimaler Abstand in Å zweier Seitenketten im TOP500H-Proteindatensatz.

$C^{\alpha} \rightarrow$																				
Ala	3.50																			
Cys	3.62	3.68																		
Asp	3.39	3.77	3.25																	
Glu	3.44	3.70	3.71	3.73																
Phe	3.56	3.61	3.28	3.77	3.88															
Gly	0.00	0.00	0.00	0.00	0.00	0.00														
His	3.61	3.85	3.34	3.57	3.77	0.00	3.82													
Ile	3.55	3.88	3.77	3.75	3.72	0.00	4.03	3.90												
Lys	3.53	3.96	3.67	3.51	3.46	0.00	3.58	3.91	3.66											
Leu	3.40	3.81	3.72	3.83	3.49	0.00	3.78	3.89	3.92	4.14										
Met	3.56	3.68	3.59	3.75	3.53	0.00	3.84	4.30	4.09	4.23	4.02									
Asn	3.65	3.86	3.64	3.65	3.62	0.00	3.84	4.05	3.89	4.02	3.74	3.71								
Pro	3.22	3.71	3.52	3.54	3.34	0.00	3.62	3.94	3.49	4.12	3.98	3.43	3.84							
Gln	3.61	3.46	3.51	3.73	3.56	0.00	4.09	4.23	4.09	4.04	4.09	3.60	3.82	3.68						
Arg	3.56	3.64	3.72	3.79	3.82	0.00	3.82	4.03	3.63	3.72	3.82	3.65	3.82	3.76	3.54					
Ser	3.57	3.56	3.54	3.70	3.37	0.00	3.57	3.85	3.83	3.88	3.82	3.70	3.81	3.59	3.59	3.35				
Thr	3.55	3.69	3.44	3.58	3.65	0.00	3.97	3.83	3.89	4.10	3.74	3.77	3.82	3.72	3.72	3.44	3.78			
Val	3.62	3.72	3.71	3.69	3.74	0.00	3.84	4.28	4.06	4.11	4.04	3.76	3.84	3.61	3.62	2.64	3.70	3.81		
Trp	3.76	3.99	3.93	3.86	4.03	0.00	3.42	4.19	4.21	4.24	3.73	3.99	3.88	4.02	3.74	3.54	3.76	3.83	3.97	
Tyr	3.61	3.73	3.67	3.92	3.72	0.00	3.85	4.03	3.41	4.20	3.93	3.72	3.76	3.85	3.70	3.51	3.74	3.97	3.71	3.63
SC↓	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr

Tabelle 6.2: Minimaler Abstand in Å zwischen einem C^{α}-Atom (Zeilen) und einer Seitenkette (Spalten) im TOP500H-Proteindatensatz.

Ala	3.78																			
Cys	4.10	3.97																		
Asp	3.98	4.43	4.28																	
Glu	3.95	4.24	4.10	3.90																
Phe	3.86	4.06	4.48	4.18	4.20															
Gly	3.50	3.67	3.62	3.85	3.78	3.51														
His	3.96	4.13	4.23	4.09	3.98	3.64	4.10													
Ile	3.94	4.19	3.85	3.91	4.05	3.71	4.07	4.22												
Lys	3.94	4.64	3.97	4.11	4.11	3.78	4.19	4.06	4.00											
Leu	3.75	4.13	4.16	4.04	4.13	3.74	4.36	4.32	4.06	4.29										
Met	4.21	4.42	4.23	4.33	4.18	3.66	4.25	4.18	4.25	4.28	4.35									
Asn	4.00	4.33	4.23	3.94	4.11	3.58	3.96	4.12	4.07	3.63	4.11	4.36								
Pro	3.77	4.11	4.06	4.06	3.99	3.59	3.97	3.81	3.79	4.01	4.08	4.07	3.97							
Gln	4.20	4.62	4.16	4.15	4.28	3.56	3.88	4.13	3.96	4.38	4.31	3.74	4.16	4.19						
Arg	4.01	3.97	3.98	4.03	4.18	3.61	3.91	3.99	4.25	3.98	4.02	4.09	3.95	4.13	4.14					
Ser	3.55	4.03	3.83	4.15	3.97	3.76	4.03	3.98	3.94	4.02	3.81	3.87	4.01	4.04	3.79	3.75				
Thr	3.82	4.22	3.99	3.43	4.00	3.57	4.15	4.21	3.88	4.13	4.22	3.86	3.89	4.18	3.95	3.97	3.77			
Val	3.96	4.15	4.20	3.92	3.97	3.76	4.02	4.11	4.05	4.16	4.16	4.15	4.01	4.23	3.96	3.53	3.90	4.25		
Trp	4.10	4.95	4.32	4.28	4.24	3.75	4.37	3.86	4.16	3.84	4.04	4.30	3.95	4.12	4.18	4.17	4.23	4.29	4.31	
Tyr	3.93	3.99	3.94	4.11	4.16	3.62	4.05	4.20	3.98	4.03	3.80	4.08	3.86	4.27	4.27	4.08	4.13	4.14	3.93	3.93
	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr

Tabelle 6.3: Minimaler Abstand in Å zweier C^{α}-Atome i und j mit $|i-j| \geq 5$ im TOP500H-Proteindatensatz.

6.2 Seitenketten-Clusterdaten

Dieser Abschnitt zeigt die Daten der Clusterzentren zur Approximation der Position des Seitenkettenschwerpunktes (siehe Abschnitt 4.2.3). Die folgende Übersicht beinhaltet für jede Aminosäure *AS* (mit Ausnahme von Glycin) die Ordnungsnummer *Nr.* des Clusterzentrums, die Koeffizienten der Seitenketten-Positionierungsbasisvektoren λ_1, λ_2 und λ_3 (siehe Gl. 4.13 und 4.14), die Population *Pop.* bzw. die Anzahl an Punkten, die zu einem Clusterzentrum gehören in Prozent, die Varianz *Var* der Distanzen der zu einem Clusterzentrum gehörenden Punkte, sowie der kürzeste $d_{min}^{(c)}$, der mittlere $d_{av}^{(c)}$ und der größte Abstand eines Punktes zu seinem Clusterzentrum $d_{max}^{(c)}$.

AS	Nr.	λ_1/Å	λ_2/Å	λ_3/Å	Pop./%	Var/Å	$d_{min}^{(c)}$/Å	$d_{av}^{(c)}$/Å	$d_{max}^{(c)}$/Å
Ala	1	-0.9709	0.2739	1.1348	0.6772	0.0104	0.0023	0.1140	0.8330
	2	-1.3308	0.0897	0.7065	0.3228	0.0176	0.0110	0.1902	1.9769
Cys	1	-1.7772	-0.7110	1.3672	0.3324	0.0450	0.0486	0.3289	1.0808
	2	-2.0094	1.1042	0.5165	0.2874	0.0373	0.0533	0.3618	1.0574
	3	-2.1966	-0.7813	0.3087	0.1959	0.0592	0.0561	0.3463	2.0826
	4	-0.9251	0.3496	2.1396	0.1843	0.0444	0.0262	0.4264	0.9262
Asp	1	-2.1957	1.1235	0.4665	0.3139	0.0648	0.0455	0.4494	1.4760
	2	-1.9842	-0.7658	1.3825	0.3022	0.0348	0.0214	0.2773	1.1833
	3	-0.8614	0.3323	2.3604	0.1697	0.0571	0.0266	0.4085	1.4708
	4	-1.9872	-1.2775	0.8836	0.1100	0.0247	0.0260	0.3379	1.1228
	5	-2.4259	-0.6379	0.2130	0.1041	0.0392	0.0394	0.3783	1.2180
Glu	1	-2.5118	1.4864	1.8374	0.1409	0.0379	0.0345	0.3588	1.1545
	2	-2.6666	-0.6480	2.0818	0.1162	0.0225	0.0233	0.3187	0.7402
	3	-2.3640	-0.2949	2.5119	0.1088	0.0276	0.0346	0.2955	0.9328
	4	-3.2198	-0.4656	0.7557	0.0959	0.0363	0.0813	0.4659	1.1046
	5	-2.5055	-1.0482	0.2685	0.0959	0.0455	0.0879	0.4209	1.1874
	6	-3.1165	1.0858	0.8993	0.0783	0.0351	0.0386	0.4169	0.9438
	7	-2.5641	0.9304	-0.2509	0.0733	0.0643	0.0452	0.5556	1.3871
	8	-2.7529	0.1879	2.0409	0.0576	0.0318	0.0474	0.4109	0.8257
	9	-1.3620	-1.5947	2.0297	0.0557	0.0643	0.0588	0.4079	1.1839
	10	-2.6080	-0.4997	-0.7839	0.0507	0.0917	0.1098	0.5421	2.6211
	11	-2.6392	-1.4341	1.3385	0.0436	0.0333	0.0809	0.5008	1.0459
	12	-1.4655	0.7419	2.9459	0.0365	0.0917	0.1029	0.5737	1.8569
	13	-0.3874	-0.3396	2.9147	0.0281	0.0855	0.0652	0.4910	1.1792
	14	-1.6690	2.1732	1.0623	0.0184	0.0882	0.1361	0.5626	1.3665

Fortsetzung

Fortsetzung Tabelle 6.4

AS	Nr.	λ_1/Å	λ_2/Å	λ_3/Å	Pop.	Var/Å	$d_{min}^{(c)}$/Å	$d_{av}^{(c)}$/Å	$d_{max}^{(c)}$/Å
Phe	1	-2.4469	2.2451	0.6754	0.2495	0.0351	0.0325	0.3495	1.2582
	2	-2.9013	-1.7332	-0.0230	0.1620	0.0267	0.0626	0.3711	0.8530
	3	-2.9136	1.6907	0.0325	0.1183	0.0564	0.0583	0.5042	1.5083
	4	-3.1790	-1.0714	-0.3615	0.1066	0.0761	0.0671	0.4440	2.1995
	5	-2.8885	-1.3051	1.1881	0.1033	0.0419	0.0551	0.4255	1.2460
	6	-2.3716	-2.2855	0.6483	0.0825	0.0369	0.0567	0.4041	1.3277
	7	-2.2977	-1.8420	1.6369	0.0808	0.0368	0.0305	0.4712	1.2400
	8	-1.1806	0.0510	3.1810	0.0729	0.0799	0.0412	0.5235	1.7202
	9	-0.3956	1.0606	3.1744	0.0242	0.0825	0.1007	0.5815	1.4983
His	1	-2.3262	2.0304	0.7072	0.1842	0.0419	0.0274	0.3414	1.4208
	2	-2.8849	-1.2692	-0.1091	0.1661	0.0556	0.0607	0.4466	1.5931
	3	-2.7152	1.5935	0.1684	0.1584	0.0627	0.0851	0.4432	1.4691
	4	-2.5944	-1.2375	1.3430	0.1514	0.0437	0.0332	0.3710	1.2934
	5	-2.4016	-1.9786	0.4507	0.1361	0.0236	0.0664	0.4047	0.9685
	6	-2.0845	-1.8973	1.3637	0.0935	0.0389	0.0570	0.4229	1.1535
	7	-0.3682	0.7136	3.0546	0.0565	0.0757	0.0547	0.5179	1.4598
	8	-1.1292	0.0202	2.9455	0.0537	0.0732	0.1467	0.5300	1.4440
Ile	1	-1.7705	0.0641	1.5543	0.4331	0.0365	0.0104	0.2503	1.3737
	2	-2.2462	-0.2480	0.3722	0.4331	0.0299	0.0731	0.3049	3.2441
	3	-1.9130	0.4220	0.9339	0.1339	0.0212	0.1196	0.4494	1.2677
Lys	1	-2.8272	1.9092	1.5533	0.1187	0.0333	0.0186	0.4161	1.4972
	2	-2.8688	-1.0221	2.1954	0.1176	0.0256	0.0754	0.4030	1.0775
	3	-2.7328	-0.3038	2.4874	0.0978	0.0320	0.0997	0.4612	0.9418
	4	-3.4496	-0.9510	0.3292	0.0872	0.0258	0.1153	0.5047	0.9620
	5	-3.2409	1.4973	0.7641	0.0754	0.0251	0.1400	0.5032	1.0206
	6	-3.3202	-0.6255	1.3335	0.0748	0.0283	0.0771	0.4855	0.8983
	7	-2.7870	-1.7515	1.2739	0.0677	0.0483	0.1622	0.5597	1.7596
	8	-2.8490	0.8911	2.0052	0.0562	0.0349	0.0557	0.5462	1.0170
	9	-2.6152	-1.7863	0.1727	0.0559	0.0600	0.0744	0.5719	1.5466
	10	-2.9198	-0.9357	-0.8405	0.0488	0.1296	0.1411	0.6418	3.0246
	11	-3.3843	0.3857	0.2880	0.0488	0.0381	0.1328	0.6204	1.2483
	12	-2.7611	1.5509	-0.3035	0.0476	0.0864	0.0945	0.6168	1.6616
	13	-1.4573	0.7806	3.2686	0.0459	0.0993	0.0755	0.6596	2.9061
	14	-1.8873	2.0385	2.1150	0.0358	0.0882	0.0996	0.6685	1.6876
	15	-1.5125	-0.9388	2.8083	0.0218	0.1367	0.0722	0.8493	2.1335
Leu	1	-2.1475	-0.4355	1.3934	0.3146	0.0566	0.0139	0.2661	2.3108
	2	-2.5222	-0.5463	0.2225	0.2252	0.0474	0.0276	0.3404	2.3245
	3	-2.0969	1.1417	1.0362	0.1823	0.0461	0.0091	0.2790	1.9593
	4	-2.4099	0.9501	0.3249	0.1404	0.0305	0.0314	0.3004	1.5099
	5	-2.1974	-0.9388	0.9790	0.1374	0.0237	0.0144	0.3113	0.8290

Fortsetzung

Fortsetzung Tabelle 6.4

AS	Nr.	λ_1/Å	λ_2/Å	λ_3/Å	Pop.	Var/Å	$d_{min}^{(c)}$/Å	$d_{av}^{(c)}$/Å	$d_{max}^{(c)}$/Å
Met	1	-2.5177	-0.1444	2.3237	0.1780	0.0390	0.0323	0.3433	1.5192
	2	-2.7169	-0.5740	0.1816	0.1117	0.0272	0.0403	0.3882	0.9641
	3	-3.2387	-0.2821	0.7152	0.0926	0.0254	0.0777	0.4101	0.7673
	4	-2.6728	-0.7129	1.9825	0.0854	0.0753	0.0733	0.4480	1.6017
	5	-2.3365	-1.3326	0.3175	0.0814	0.0543	0.0501	0.3729	1.4716
	6	-3.1331	0.9901	0.8587	0.0782	0.0355	0.0959	0.4196	1.0881
	7	-2.9925	-1.1345	1.0022	0.0758	0.0373	0.1025	0.4407	1.1323
	8	-2.8120	0.7053	0.0931	0.0686	0.0394	0.0607	0.3851	1.0139
	9	-2.5789	-0.7428	-0.8096	0.0583	0.1093	0.0816	0.4887	2.2113
	10	-2.6448	1.2910	1.7285	0.0519	0.0297	0.0600	0.3887	0.8957
	11	-1.8133	1.9635	1.7835	0.0479	0.0570	0.2441	0.5812	1.1839
	12	-2.4570	1.0294	-0.7497	0.0351	0.0955	0.1152	0.5799	1.5606
	13	-1.6405	0.7038	2.8632	0.0351	0.1032	0.1607	0.6128	1.6370
Asn	1	-1.9969	-0.8088	1.2943	0.2492	0.0284	0.0146	0.2867	1.0163
	2	-2.1808	1.1485	0.3620	0.2369	0.0493	0.0419	0.4056	1.6571
	3	-1.9619	-1.3935	0.6845	0.1741	0.0219	0.0442	0.3306	0.8524
	4	-0.9456	0.3009	2.3033	0.1432	0.0635	0.0448	0.4114	1.6402
	5	-2.3532	-0.7839	0.1624	0.1392	0.0872	0.0368	0.3875	2.5463
	6	-2.2315	0.4730	1.0292	0.0574	0.0424	0.0821	0.3921	1.1830
Pro	1	-0.3723	-0.7706	1.6458	0.4952	0.0158	0.0132	0.2363	0.7080
	2	-1.0298	-1.0478	1.1358	0.4583	0.0223	0.0228	0.2603	1.8527
	3	-1.4210	0.1715	1.1603	0.0464	0.0368	0.0601	0.4089	0.7741
Gln	1	-2.4426	-0.3092	2.3750	0.2066	0.0320	0.0395	0.3337	1.0090
	2	-3.1549	-0.5344	0.6712	0.1063	0.0336	0.1435	0.4895	1.0667
	3	-2.4075	-1.1198	0.2795	0.1063	0.0452	0.0587	0.4599	1.1343
	4	-2.5792	1.3776	1.7639	0.0985	0.0369	0.0362	0.3725	0.8838
	5	-2.5091	0.9327	-0.2073	0.0958	0.0680	0.0231	0.5026	1.2674
	6	-2.9003	-0.3469	1.7146	0.0731	0.0292	0.0608	0.4187	0.9023
	7	-2.4723	-0.6421	-0.8032	0.0699	0.0947	0.0361	0.4877	2.4304
	8	-3.0991	1.0208	0.7947	0.0686	0.0362	0.0711	0.4447	1.0243
	9	-2.5964	-1.3155	1.6538	0.0581	0.0380	0.1204	0.4926	1.2627
	10	-1.6638	0.6054	2.8393	0.0368	0.0674	0.1016	0.5344	1.4124
	11	-1.3436	-1.6490	1.8526	0.0345	0.0956	0.0866	0.5638	1.4279
	12	-1.5934	2.0645	1.4830	0.0268	0.0816	0.1052	0.7392	1.5854
	13	-0.5323	-0.4524	2.8583	0.0186	0.1109	0.0970	0.4932	1.7709

Fortsetzung

Fortsetzung Tabelle 6.4

AS	Nr.	λ_1/Å	λ_2/Å	λ_3/Å	Pop.	Var/Å	$d_{min}^{(c)}$/Å	$d_{av}^{(c)}$/Å	$d_{max}^{(c)}$/Å
Arg	1	-3.9241	0.2960	1.4125	0.0734	0.0474	0.1082	0.5892	1.2581
	2	-3.8233	-1.2012	1.9059	0.0730	0.0371	0.0817	0.5233	0.9839
	3	-3.3175	-1.2516	2.8784	0.0675	0.0285	0.0698	0.4732	0.8630
	4	-4.2130	-1.1142	0.2622	0.0638	0.0509	0.1382	0.6415	1.4046
	5	-3.4265	2.4337	1.5727	0.0631	0.0674	0.0676	0.6489	1.8622
	6	-2.2901	-0.8728	3.4397	0.0624	0.1286	0.1175	0.5561	2.1396
	7	-2.4154	0.1032	3.4380	0.0594	0.0484	0.1041	0.5301	1.1446
	8	-3.1672	0.9168	2.5399	0.0594	0.0407	0.1220	0.5706	1.0684
	9	-3.5041	-0.1884	2.4604	0.0550	0.0265	0.1039	0.5277	0.8887
	10	-2.1001	1.7132	3.2006	0.0525	0.0521	0.1035	0.6490	1.2297
	11	-3.7885	-2.0135	0.9038	0.0484	0.0453	0.0894	0.5786	1.1579
	12	-3.9284	0.2287	-0.1522	0.0466	0.0493	0.1040	0.6255	1.1904
	13	-4.0073	1.6425	0.5572	0.0448	0.0415	0.1110	0.6451	1.1530
	14	-3.0253	-1.9580	-1.1237	0.0418	0.1103	0.1323	0.6947	2.0040
	15	-1.6841	2.7120	2.4727	0.0400	0.0847	0.1027	0.6861	1.3835
	16	-2.7262	-2.3803	2.1700	0.0367	0.0917	0.1889	0.7086	2.1254
	17	-2.9922	1.9904	-0.7646	0.0363	0.1144	0.0809	0.7901	1.7556
	18	-0.9053	0.6850	4.1763	0.0271	0.2644	0.0836	0.8486	2.9858
	19	-3.2220	-0.2279	-1.2781	0.0257	0.1069	0.2719	0.8465	2.2680
	20	-2.7295	-2.5099	0.2812	0.0231	0.0597	0.1868	0.7297	1.4399
Ser	1	-1.6107	0.0405	0.8441	0.5345	0.0422	0.1533	0.7209	2.5459
	2	-0.9006	0.3899	1.6629	0.4655	0.0403	0.0235	0.3377	1.2088
Thr	1	-1.3797	0.0613	1.3287	0.7677	0.0284	0.0139	0.3750	1.0926
	2	-1.8833	0.0575	0.4201	0.2323	0.0346	0.0103	0.2595	1.3474
Val	1	-1.5137	0.2163	1.1706	0.5628	0.0356	0.0129	0.3198	1.1766
	2	-1.9105	0.0962	0.3546	0.4372	0.0194	0.0044	0.1770	1.8145

Fortsetzung

Fortsetzung Tabelle 6.4

AS	Nr.	λ_1/Å	λ_2/Å	λ_3/Å	Pop.	Var/Å	$d_{min}^{(c)}$/Å	$d_{av}^{(c)}$/Å	$d_{max}^{(c)}$/Å
Trp	1	-2.3530	2.9335	0.7178	0.0862	0.0298	0.0386	0.3557	0.9847
	2	-3.3762	-1.0434	-1.2215	0.0862	0.1303	0.0484	0.4805	2.4827
	3	-2.9765	2.1766	-0.3067	0.0748	0.0224	0.0887	0.4793	0.8653
	4	-1.9112	3.1036	1.4363	0.0697	0.0264	0.0704	0.4074	0.9095
	5	-3.4293	-1.5880	0.0229	0.0684	0.0245	0.1028	0.4285	0.7975
	6	-3.8086	-0.7297	-0.1265	0.0672	0.0213	0.1747	0.4655	0.8795
	7	-3.0226	-2.1286	0.6727	0.0570	0.0254	0.1031	0.4266	0.7695
	8	-2.9797	-1.3541	2.5278	0.0558	0.0507	0.1409	0.5944	1.0425
	9	-3.6914	-1.0206	1.1392	0.0545	0.0527	0.1918	0.5628	1.2977
	10	-2.7856	2.5524	1.4709	0.0520	0.0434	0.0486	0.3814	1.0461
	11	-3.5815	1.3116	-0.0015	0.0520	0.0380	0.0970	0.4644	1.0353
	12	-0.1457	2.0760	3.2108	0.0418	0.0666	0.1548	0.6240	1.2405
	13	-3.4914	1.7962	0.8325	0.0406	0.0423	0.2166	0.5200	1.1667
	14	-2.3016	-2.5748	1.5202	0.0342	0.0899	0.3409	0.6905	1.3031
	15	-2.7547	-2.2624	-0.6701	0.0304	0.0437	0.2073	0.5573	0.9759
	16	-1.0441	1.0072	3.6087	0.0304	0.1201	0.0860	0.6392	1.3690
	17	-3.5310	0.7299	-1.0677	0.0279	0.0436	0.2482	0.5283	0.9128
	18	-0.5537	-0.7596	3.7276	0.0266	0.0422	0.1549	0.4380	1.0030
	19	-1.6268	-0.5557	3.6302	0.0228	0.0822	0.0796	0.5134	1.0922
	20	-2.5624	1.9670	-1.5923	0.0215	0.1689	0.2555	0.7611	1.7366
Tyr	1	-2.4567	2.8755	0.7743	0.1207	0.0513	0.0225	0.3588	1.9591
	2	-3.0458	2.3406	0.5605	0.1091	0.0333	0.0293	0.3287	1.2703
	3	-2.9294	-2.4671	-0.1005	0.0823	0.0246	0.0736	0.3820	1.2044
	4	-3.4285	-1.8056	0.0403	0.0823	0.0166	0.0702	0.3322	0.7199
	5	-2.8160	2.5871	-0.0421	0.0712	0.0287	0.0798	0.3516	1.1167
	6	-2.4152	-2.8414	0.8337	0.0662	0.0390	0.0443	0.4182	1.0064
	7	-3.5987	-1.2594	-0.5998	0.0626	0.0390	0.0574	0.3909	1.1683
	8	-3.1860	-1.9749	-0.7577	0.0591	0.1366	0.0815	0.4030	2.7460
	9	-3.0686	-2.1403	0.9469	0.0566	0.0159	0.0585	0.3572	0.6231
	10	-2.9781	-1.7558	1.7975	0.0525	0.0247	0.0928	0.3651	0.9552
	11	-1.4123	0.3358	3.6121	0.0500	0.0607	0.0747	0.5058	1.1844
	12	-3.3830	1.7960	-0.2893	0.0495	0.0886	0.1221	0.5389	1.8123
	13	-2.3697	-2.3462	1.8903	0.0460	0.0791	0.0470	0.4623	1.8295
	14	-3.4689	-1.3512	1.1463	0.0333	0.0407	0.1491	0.4081	1.0537
	15	-0.5713	-0.2210	3.8164	0.0318	0.0909	0.0535	0.5655	1.7420
	16	-0.2397	1.3200	3.6192	0.0268	0.0794	0.1015	0.6109	1.3705

Tabelle 6.4: Tabelle der optimierten Cluster, mit Aminosäure AS, der Clusternummer $Nr.$, den Vektorfunktionskoeffizienten λ_k, der anteiligen Population $Pop.$ des Clusters an allen Punkten, der Varianz Var des Clusters und die Abstände der Punkte, die zu jedem Cluster gehören: kleinster Abstand $d_{min}^{(c)}$, mittlerer Abstand aller Punkte, die zu einem Cluster gehören $d_{av}^{(c)}$ und größter Abstand $d_{max}^{(c)}$.

Literaturverzeichnis

[1] A. L. Lehninger, *Biochemie*, (Verlag Chemie, Weinheim, 1977).

[2] D. Voges, P. Zwickel, and W. Baumeister, Annu. Rev. Biochem. **68**, 1015 (1999).

[3] A. L. Goldberg, Nature **426**, 895 (2003).

[4] M. H. Glickman and A. Ciechanover, Physiol. Rev. **82**, 373 (2002).

[5] R. Gronostajski, A. B. Pardee, and A. L. Goldberg, J. Biol. Chem. **260**, 3344 (1985).

[6] R. Schlögl, Angewandte Chemie **115**, 2050 (2003).

[7] K. U. Linderström-Lang and J. A. Schellmann, *Protein Structure and Enzyme Activity. The Enzymes, Vol. 1, 2nd ed.*, (Academic Press, New York, 1960).

[8] C. B. Anfinsen, Science **181**, 223 (1973).

[9] P. Lengyel and D. Söll, Bacteriological Rev. **33**, 264 (1969).

[10] N. Nameki and. M. Yoneyama and. S. Koshiba, N. Tochio, M. Inoue, E. Seki, T. Matsuda, Y. Tomo, T. Harada, K. Saito, N. Kobayashi, T. Yabuki, M. Aoki, E. Nunokawa, N. Matsuda, N. Sakagami, T. Terada, M. Shirouzu, M. Yoshida, H. Hirota, T. Osanai, A. Tanaka, T. Arakawa, P. Carninci, J. Kawai, Y. Hayashizaki, K. Kinoshita, P. Guntert, T. Kigawa, and S. Yokoyama, Prot. Sci. **13**, 2089 (2004).

[11] L. Pauling, R. B. Corey, and H. R. Branson, Proc. Natl. Acad. Sci. USA **37**, 205 (1951).

[12] A. Wada, Adv. Biophys. Mol. Biol. **9**, 1 (1976).

[13] W. G. H. Hol, Prog. Biophys. Mol. Biol. **45**, 149 (1985).

[14] L. Pauling, *The Nature of the chemical bond*, (Cornell University Press, Ithaca, New York, 1960).

[15] I. K. McDonald and J. M. Thornton, J. Mol. Biol. **238**, 777 (1994).

[16] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson, Proteins **40**, 389 (2000).

[17] P. A. Karplus, Prot. Sci. **5**, 1406 (1996).

[18] IUPAC-IUB Commission in Biochemical Nomenclature, Biochemistry **9**, 3471 (1969).

[19] G. E . Schulz und R. H. Schirmer, *Principles of protein structure*, (Springer-Verlag, Heidelberg, 1979).

[20] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton, Proteins **12**, 345 (1992).

[21] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, J. Mol. Biol. **7**, 95 (1963).

[22] G. N. Ramachandran and V. Sasisekharan, Adv. Prot. Chem. **23**, 283 (1968).

[23] N. Mandel, G. Mandel, B. L. Trus, J. Rosenberg, G. Carlson, and R. E. Dickerson, J. Biol. Chem. **252**, 4619 (1977).

[24] D. A. Brant, W. G. Miller, and P. J. Flory, J. Mol. Biol. **23**, 47 (1967).

[25] G. Nemethy and H. A. Scheraga, Quart. Rev. Biophys. **10**, 239 (1977).

[26] G. N. Ramachandran, C. M. Venkatachalam, and S. Krimm, Biophys. J. **6**, 849 (1966).

[27] O. Herzberg and J. Moult, Proteins **11**, 223 (1991).

[28] N. J. Darby and T. E. Creighton, *Protein Structure*, (Oxford University Press, New York, 1993).

[29] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).

[30] http://swift.cmbi.ru.nl/gv/dssp/.

[31] L. Jin, S. L. Briggs, S. Chandrasekhar, N. Y. Chirgadze, D. K. Clawson, R. W. Schevitz, D. L. Smiley, A. H. Tashjian, and F. Zhang, J. Biol. Chem. **275**, 27238 (2000).

[32] W. G. H. Hol, P. T. van Duijnen, and H. J. Berendsen, Nature **273**, 443 (1978).

[33] D. E. Engel and W. F. DeGrado, J. Mol. Biol. **337**, 1195 (2004).

[34] C. M. Anderson, R. C. McDonald, and T. A. Steitz, J. Mol. Biol. **123**, 1 (1978).

[35] S. Miller, Protein Eng. **3**, 77 (1989).

[36] M. Levitt and C. Chothia, Nature **261**, 552 (1976).

[37] G. Nemethy, D. C. Philips, S. J. Leach, and H. A. Scheraga, Nature **214**, 363 (1967).

[38] R. Srinivasan, R. Balasubramanian, and S. S. Rajan, Science **194**, 720 (1976).

[39] P. Bornstein and H. Sage, Annu. Rev. Biochem. **49**, 957 (1980).

[40] D. R. Eyre, M. A. Paz, and P. M. Gallop, Annu. Rev. Biochem. **53**, 717 (1984).

[41] K. Nagano, J. Mol. Biol. **109**, 235 (1977).

[42] J. S. Richardson, Nature **268**, 495 (1977).

[43] S. Lifson and C. Sander, J. Mol. Biol. **139**, 627 (1980).

[44] A. G. Street and S. L. Mayo, Proc. Natl. Acad. Sci. USA **96**, 9074 (1999).

[45] D. L. Minor and P. S. Kim, Nature **371**, 264 (1994).

[46] M. Parisien and F. Major, Proteins **68**, 824 (2007).

[47] C. M. Venkatachalam, Biopolymers **6**, 1425 (1968).

[48] P. Y. Chou and G. D. Fasman, J. Mol. Biol. **115**, 135 (1977).

[49] M. Levitt and J. Greer, J. Mol. Biol. **114**, 181 (1977).

[50] I. D. Kuntz, J. Am. Chem. Soc. **94**, 4009 (1972).

[51] G. Nemethy and M. P. Printz, Macromolecules **5**, 755 (1972).

[52] B. W. Matthews, Macromolecules **5**, 818 (1972).

[53] P. N. Lewis, F. A. Momany, and H. A. Scheraga, Biochimica et Biophysica Acta **303**, 211 (1973).

[54] G. D. Rose, Nature **272**, 586 (1978).

[55] S. S. Zimmermann and H. A. Scheraga, Proc. Natl. Acad. Sci. USA **74**, 4126 (1977).

[56] C. Chothia and A. V. Finkelstein, Annu. Rev. Biochem. **59**, 1007 (1990).

[57] S. B. Ozkan, G. A. Wu, J. D. Chodera, and Ken A. Dill, Proc. Natl. Acad. Sci. USA **104**, 11987 (2007).

[58] C. Brandon and J. Tooze, *Introduction to Protein Structure*, (Garland Publishing, New York, 1999).

[59] J. M. Thornton, D. T. Jones, M. W. MacArthur, C. M. Orengo, and M. B. Swindels, Phil. Trans. R. Soc. Lond. B **348**, 71 (1995).

[60] R. F. Doolittle, Annu. Rev. Biochem. **64**, 287 (1995).

[61] M. Gangloff, A. Murali, J. Xiong, C. J. Arnot, A. N. Weber, A. M. Sandercock, C. V. Robinson, R. Sarisky, A. Holzenburg, C. Kao, and N. J. Gay, J. Biol. Chem. **283**, 14629 (2008).

[62] B. Meusser, C. Hirsch, E. Jarosch, and T. Sommer, Nat. Cell Biol. **7**, 766 (2005).

[63] J. Payandeh, C. Li, M. Ramjeesingh, E. Poduch, C. E. Bear, and E. F. Pai, J. Biol. Chem. **283**, 11721 (2008).

[64] C. M. Dobson, Nature **426**, 884 (2003).

[65] B. Gutte and R. B. Merrifield, J. Am. Chem. Soc. **91**, 501 (1969).

[66] R. Hirschmann, R. F. Nutt, D. F. Veber, R. A. Vitali, S. L. Varga, T. A. Jacob, F. W. Holly, and R. G. Denkewalter, J. Am. Chem. Soc. **91**, 507 (1969).

[67] K. A. Dill, Biochemistry **29**, 7133 (1990).

[68] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. USA **89**, 8721 (1992).

[69] M.-H. Hao and H. A. Scheraga, J. Phys. Chem. **98**, 9882 (1994).

[70] J. N. Onuchic, P. G. Wolynes, and Z. Luthey-Schulten, Proc. Natl. Acad. Sci. USA **92**, 3626 (1995).

[71] A. Sali, E. Shakhnovich, and M. Karplus, Nature **369**, 248 (1994).

[72] A. Sali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).

[73] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).

[74] R. Elber and M. Karplus, Science **235**, 318 (1987).

[75] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).

[76] A. Neumaier, SIAM Rev. **39**, 407 (1997).

[77] C. Levinthal, J. Chim. Phys. **65**, 44 (1968).

[78] U. H. E. Hansmann, J. Chem. Phys. **120**, 417 (2004).

[79] D. Baker and D. A. Agard, Biochemistry **33**, 7505 (1994).

[80] A. E. Franke, D. E. Danley, F. S. Kaczmarek, S. J. Hawrylik, R. D. Gerard, S. E. Lee, and K. F. Geoghegan, Biochim. Biophys. Acta **1037**, 16 (1990).

[81] S. S. Jaswell, J. L. Sohl, J. H. Davis, and D. A. Agard, Nature **415**, 343 (2002).

[82] D. J. Wales, *Energy Landscapes. With Applications to Clusters, Biomolecules and Glasses*, (Cambridge University Press, Cambridge, 2003).

[83] F. M. Richards, J. Mol. Biol. **82**, 1 (1974).

[84] W. Kauzmann, Adv. Prot. Chem. **14**, 1 (1959).

[85] H. Li, C. Tang, and N. S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).

[86] G. J. Lesser and G. D. Rose, Proteins **8**, 6 (1990).

[87] D. Xu, S. L. Lin, and R. Nussinov, J. Mol. Biol. **265**, 68 (1997).

[88] C. N. Pace, Biochemistry **40**, 310 (2001).

[89] Y. Harano and M. Kinoshita, Chem. Phys. Lett. **399**, 342 (2004).

[90] T. Imai, Y. Harano, M. Kinoshita, A. Kovalenko, and F. Hirata, J. Chem. Phys. **126**, 225102 (2007).

[91] C. M. Dobson, A. Sali, and M. Karplus, Angew. Chem. Int. Ed. **37**, 868 (1998).

[92] D. A. Snyder, Y. Chen, N. G. Denissova, T. Acton, J. M. Aramini, M. Ciano, R. Karlin, J. F. Liu, P. Manor, P. A. Rajan P. Rossi G. V. T. Swapna, R. Xiao B. Rost, J. Hunt, and G. T. Montelione, J. Am. Chem. Soc. **127**, 16505 (2005).

[93] A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo, Proc. Natl. Acad. Sci. USA **104**, 9615 (2007).

[94] G. M. Clore and A. M. Gronenborn, Proc. Natl. Acad. Sci. USA **95**, 5891 (1998).

[95] L. Whitmore and B. A. Wallace, Biopolymers **89**, 392 (2007).

[96] C. Fenselau, Annu. Rev. Biophys. **20**, 205 (1991).

[97] A. Kolinski and J. Skolnick, Polymer **45**, 511 (2004).

[98] A. Neumaier, S. Dallwig, W. Huyer, and H. Schichl, New Techniques for the Construction of Residue Potentials for Protein Folding, In Computational Molecular Dynamics: Challenges, Methods, Ideas, Springer, 1999.

[99] A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, J. Comput. Chem. **19**, 849 (1997).

[100] A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, S. Oldziej, and H. A. Scheraga, J. Comput. Chem. **18**, 874 (1997).

[101] A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, and H. A. Scheraga, J. Comput. Chem. **19**, 259 (1998).

[102] B. Park and M. Levitt, J. Mol. Biol. **258**, 367 (1996).

[103] http://www.gnuplot.info.

[104] W. H. Press, B. P. Flannery, S. A. Teulsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing (Fortran Edition)*, (Cambridge University Press, Cambridge, second edition, 1992).

[105] J. Hartigan and M. Wong, Applied Statistics **28**, 100 (1979).

[106] W. Martinez and A. Martinez, *Computational Statistics Handbook with MATLAB*, (Chapman and Hall / CRC, 2002).

[107] D. Sparks, Applied Statistics **22**, 126 (1973).

[108] https://people.scs.fsu.edu/∼burkardt/f_src/f_src.html.

[109] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, Nucl. Acid. Res. **28**, 235 (2000).

[110] http://www.rcsb.org/pdb.

[111] G. Wang and R. L. Dunbrack, Jr, Bioinformatics **19**, 1589 (2003).

[112] http://dunbrack.fccc.edu/pisces.

[113] S. C. Lovell, I. W. Davis, W. B. Arendall III, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, Proteins **50**, 437 (2003).

[114] U. Hobohm and C. Sander, Prot. Sci. **3**, 522 (1994).

[115] http://kinemage.biochem.duke.edu/databases/top500.php.

[116] J. Meller, M. Wagner, and R. Elber, J. Comput. Chem. **23**, 111 (2002).

[117] J. Qui and R. Elber, Proteins **61**, 44 (2005).

[118] G. M. Crippen, J. Mol. Graph. Mod. **19**, 87 (2001).

[119] J. Zhang, R. Chen, and J. Liang, Conf. Proc. IEEE Eng. Med. Biol. Soc. **4**, 2976 (2004).

[120] F. Fogolari, L. Pieri, A. Dovier, L. Bortolussi, G. Giugliarelli, A. Corazza, G. Esposito, and P. Viglino, BMC Struc. Biol. **7**, 15 (2007).

[121] L. Wroblewski and J. Skolnick, J. Comput. Chem. **28**, 2059 (2007).

[122] M. Chhajer and G. M. Crippen, BMC Struc. Biol. **2**, 4 (2002).

[123] B. A. Reva, A. V. Finkelstein, M. F. Sanner, and A. J. Olson, Pacific Symposium on Biocomputing, 373 (1997).

[124] J. Meller and R. Elber, Proteins **45**, 241 (2001).

[125] B. H. Park, E. S. Huang, and M. Levitt, J. Mol. Biol. **266**, 831 (1997).

[126] L. Zhang and J. Skolnick, Prot. Sci. **7**, 112 (1998).

[127] Y. Wu, M. Lu, M. Chen, J. Li, and J. Ma, Prot. Sci. **16**, 1449 (2007).

[128] Y. Dehouck, D. Gilis, and M. Rooman, Biophys. J. **90**, 4010 (2006).

[129] L. A. Mirny and E. I. Shakhnovich, J. Mol. Biol. **264**, 1164 (1996).

[130] C. Loose, J. L. Klepeis, and C. A. Floudas, Proteins **54**, 303 (2004).

[131] R. Rajgaria, S. R. McAllister, and C. A. Floudas, Proteins **65**, 726 (2006).

[132] E.-A. D. Amir, N. Kalisman, and C. Keasar, Proteins **72**, 62 (2008).

[133] P. Ren and J. W. Ponder, J. Phys. Chem. **107**, 5933 (2003).

[134] P. Ren and J. W. Ponder, J. Comput. Chem. **23**, 1497 (2002).

[135] R. K. Hart R. V. Pappu and J. W. Ponder, J. Phys. Chem. **102**, 9725 (1998).

[136] J. W. Ponder M. E. Hodsdon and D. P. Cistola, J. Mol. Biol. **264**, 585 (1996).

[137] J. W. Ponder C. E. Kundrot and F. M. Richards, J. Comput. Chem. **12**, 402 (1991).

[138] J. W. Ponder and F. M. Richards, J. Comput. Chem. **8**, 1016 (1987).

[139] http://dasher.wustl.edu/tinker.

[140] D. Gilis, J. Biomol. Struct. Dyn. **21**, 725 (2004).

[141] Edited by D. M. Webster, *Methods in Molecular Biology, vol. 143: Protein Structure Prediction: Methods and Protocols*, (Humana Press, Totowa, 2000).

[142] http://dd.compbio.washington.edu.

[143] R. Samudrala and J. Moult, J. Mol. Biol. **279**, 287 (1998).

[144] http://dd.stanford.edu.

[145] K. T. Simons, C. Kooperberg, E. S. Huang, and D. Baker, J. Mol. Biol. **268**, 209 (1997).

[146] R. Samudrala, E. S. Huang, and M. Levitt, unpublished results.

[147] Y. Xia, E. S. Huang, M. Levitt, and R. Samudrala, J. Mol. Biol. **300**, 171 (2000).

[148] M. R. Lee, J. Tsai, D. Baker, and P. A. Kollmann, J. Mol. Biol. **313**, 417 (2001).

[149] http://depts.washington.edu/bakerpg/decoys.

[150] http://titan.princeton.edu/Decoys.

[151] http://titan.princeton.edu/HRDecoys.

[152] C. Keasar and M. Levitt, J. Mol. Biol. **329**, 159 (2003).

[153] L. Holm and C. Sander, J. Mol. Biol. **225**, 93 (1992).

[154] C. I. Brandon and T. A. Jones, Nature **343**, 687 (1990).

[155] Die ursprüngliche Seite http://prostar.carb.nist.gov ist nicht mehr verfügbar. Eine Kopie der Daten ist erhältlich unter: http://www.mat.univie.ac.at/∼neum/protein.html.

[156] R. Samudrala and M. Levitt, BMC Struc. Biol. **2**, 3 (2002).

[157] J. Tsai, R. Bonneau, V. Morozov, B. Kuhlmann, C. A. Rohl, and D. Baker, Proteins **53**, 76 (2003).

[158] F. Fogolari and S. C. E. Tosatto, Prot. Sci. **14**, 889 (2005).

[159] Y. Wang, H. Zhang, and R. A. Scott, Proc. Natl. Acad. Sci. USA **92**, 709 (1995).

[160] F. Koskowski and B. Hartke, J. Comput. Chem. **26**, 1169 (2005).

[161] W. Boomsma and T. Hamelryk, BMC Bioinf. **6:159**, (2005).

[162] A. Canutescu and R. Dunbrack, Prot. Sci. **12**, 963 (2003).

[163] D. S. Riddle, J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova amd Q. Yi, and Bake D, Nat. Struct. Biol. **4**, 805 (1997).

[164] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).

[165] M. Cieplak, N. S. Holter, A. Maritan, and J. R. Banavar, J. Chem. Phys. **114**, 1420 (2001).

[166] P. G. Wolynes, Nat. Struct. Biol. **4**, 871 (1997).

[167] J. Wang and W. Wang, Phys. Rev. E **65**, 41911 (2002).

[168] M. Cheon and I. Chang, J. Korean Phys. Soc. **44**, 1577 (2004).

[169] B. Robson and E. Suzuki, J. Mol. Biol. **107**, 327 (1976).

[170] B. C. Orcutt, R. M. Schwartz, and M. O. Dayhoff, Atlas Protein Sequence Struct. **5**, 345 (1978).

[171] R. W. Taylor, J. Theor. Biol. **119**, 205 (1986).

[172] J. Wang and W. Wang, Nat. Struc. Biol. **6**, 1033 (1999).

[173] R. L. Murphy, A. Wallqvist, and M. R. Levy, Protein Eng. **13**, 149 (2000).

[174] N. Cannata, S. Toppo, C. Romualdi, and G. Valle, Bioinformatics **18**, 1102 (2002).

[175] T. Li, K. Fan, J. Wang, and W. Wang, Protein Eng. **16**, 323 (2003).

[176] H. N. Buttimore, N. Goldman, and C. Koisol, J. Theor. Biol. **228**, 97 (2004).

[177] M. Cheon, M. Heo, and I. Chang, J. Korean Phys. Soc. **47**, 901 (2005).

[178] A. Luthra, A. N. Jha, G. K. Ananthasuresh, and S. Vishveswara, J. Biosci. **32**, 883 (2007).

[179] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).

[180] L. J. Smith, C. Redfield, R. A. G. Smith, C. M. Dobson, G. M. Clore, A. M Gronenborn, M. R. Walter, T. L. Naganbushan, and A. Wlodawer, Nat. Struc. Biol. **1**, 301 (1994).

[181] W. Braun, M. Vasak, A. H. Robbins, C. D. Stout, G. Wagner, J. H. R. Kägi, and K. Wüthrich, Proc. Natl. Acad. Sci. USA **89**, 10124 (1992).

[182] M. Billeter, J. Vendrell, G. Wider, F. X. Aviles, M. Coll, A. Guasch, R. Huber, and K. Wüthrich, J. Biomol. NMR **2**, 1 (1992).

[183] H. B. Cole, S. W. Sparks, and D. A. Torchia, Proc. Natl. Acad. Sci. USA **85**, 6362 (1988).

[184] F. Jensen, *Introduction to Computational Chemistry*, (WILEY-VCH-Verlag GmbH, Winheim, 1999).

[185] J. W. Ponder and D. A. Case, Adv. Prot. Chem. **66**, 27 (2003).

[186] M. Berrera, H. Molinari, and F. Fogolari, BMC Bioinf. **4:8**, (2003).

[187] X. Li and J. Liang, Proteins **60**, 46 (2005).

[188] M. R. Ejtehadi, S. P. Avall, and S. S. Plotkin, Proc. Natl. Acad. Sci. USA **101**, 15088 (2004).

[189] Y. Feng, A. Kloczkowski, and R. L. Jernigan, Proteins **68**, 57 (2007).

[190] H. H. Gan, A. Tropsha, and T. Schlick, Proteins **43**, 161 (2001).

[191] S. Mayewski, Proteins **59**, 152 (2005).

[192] B. Rost and C. Sander, Proc. Natl. Acad. Sci. USA **90**, 7558 (1993).

[193] B. Rost, J. Struct. Biol. **134**, 204 (2001).

[194] D. Shortle, Prot. Sci. **11**, 18 (2002).

[195] D. Kihara, Prot. Sci. **14**, 1955 (2005).

[196] Z.-H. Wang and H. C. Lee, Phys. Rev. Lett. **84**, 574 (2000).

[197] D. Shortle, W. E. Stites, and A. K. Meeker, Biochemistry **29**, 8033 (1990).

[198] T. Lazaridis and M. Karplus, Curr. Opin. Struct. Biol. **10**, 139 (2000).

[199] T. Lazaridis and M. Karplus, J. Mol. Biol. **288**, 477 (1999).

[200] E. Huang, S. Subbiah, and M. Levitt, J. Mol. Biol. **252**, 709 (1996).

[201] G. Cassari and M. Sippl, J. Mol. Biol. **224**, 725 (1992).

[202] R. Roux and T. Simonson, Biophys. Chem. **78**, 1 (1999).

[203] D. Chandler, Nature **437**, 640 (2005).

[204] M. S. Lin, N. L. Fawzi, and T. Head-Gordon, Structure **15**, 727 (2007).

[205] A. G. Street and S. L. Mayo, Fold. Des. **3**, 253 (1998).

[206] D. Eisenberg and A. D. McLachlan, Nature **319**, 199 (1986).

[207] C. Chothia, Nature **254**, 304 (1975).

[208] V. Viswanadhan, Int. J. Biol. Macromol. **9**, 39 (1987).

[209] B. Fain, Y. Xia, and M. Levitt, IBM J. Res. & Dev. **45**, 525 (2001).

[210] J. Vondrasek, L. Bendova, V. Klusak, and P. Hobza, J. Am. Chem. Soc. **127**, 2615 (2005).

[211] J. L. Cornette, K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky, and C. DeLisi, J. Mol. Biol. **195**, 659 (1987).

[212] http://www.mat.univie.ac.at/∼neum/software/protein/aminoacids.html, Unpublizierte Ergebnisse.

[213] B. Lee and F. M. Richards, J. Mol. Biol. **55**, 379 (1971).

[214] T. J. Richmond, J. Mol. Biol. **178**, 63 (1984).

[215] S. J. Wodak and J. Janin, Proc. Natl. Acad. Sci. USA **77**, 1736 (1980).

[216] L. Cavallo, J. Kleinjung, and F. Fraternali, Nucl. Acids Res. **31**, 3364 (2003).

[217] W. Hasel, T. F. Hendrickson, and W. C. Still, Tetrahedron Comput. Methodol **1**, 103 (1988).

[218] M. Chinchio, C. Czaplewski, S. Oldziej, and H. A. Scheraga, Multiscale Model. Simul. **5**, 1175 (2006).

[219] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollmann, J. Am. Chem. Soc. **117**, 5179 (1995).

[220] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).

[221] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, J. Phys. Chem. **79**, 2361 (1975).

[222] G. Nemethy, M. S. Pottle, and H. A. Scheraga, J. Phys. Chem. **87**, 1883 (1983).

[223] M. J. Sipple, G. Nemethy, and H. A. Scheraga, J. Phys. Chem. **88**, 6231 (1984).

[224] G. Nemethy et al., J. Phys. Chem. **96**, 6472 (1992).

[225] N. L. Allinger, J. Am. Chem. Soc. **99**, 8127 (1977).

[226] E. Krieger, T. Darden, S. B. Nabuurs, A. Finkelstein, and G. Vriend, Proteins **57**, 678 (2004).

[227] A. V. Finkelstein, A. Y. Badretdinov, and A. M. Gutin, Proteins **23**, 142 (1995).

[228] S. H. Bryant and C. E. Lawrence, Proteins **9**, 108 (1991).

[229] J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik, Prot. Sci. **6**, 676 (1997).

[230] H. Zhou and Y. Zhou, Prot. Sci. **11**, 2714 (2002).

[231] Y. Xia and M. Levit, J. Phys. Chem. **113**, 9318 (2000).

[232] M. Vendruscolo, L. A. Mirny, E. I. Shakhnovich, and E. Domany, Proteins **41**, 192 (2000).

[233] H. A. Eiselt and C.-L. Sandblom, *Linear Programming and its Applications*, (Springer Verlag, New York Berlin Heidelberg, 2007).

[234] G. B. Dantzig and M. N. Thapa, *Linear Programming 1: Introduction*, (Springer Verlag, New York Berlin Heidelberg, 1997).

[235] G. B. Dantzig and M. N. Thapa, *Linear Programming 2: Theory and Extensions*, (Springer Verlag, New York Berlin Heidelberg, 2003).

[236] C. Mészáros, *The efficient implementation of interior point methods for linear programming and their applications*, Ph.D. thesis, Eötvös Loránd, University of Sciences, 1996.

[237] C. Mészáros, Opt. Meth. Soft. **11**, 431 (1999).

[238] BPMPD Manual, http://www.doc.ic.ac.uk/research/technicalreports/1997/DTR97-8.pdf.

[239] http://www.sztaki.hu/ meszaros/bpmpd/.

[240] http://lpsolve.sourceforge.net/5.5/mps format.htm.

[241] http://miplib.zib.de/miplib3/mps_format.txt.

[242] A. Bhattacharyay, A. Trovato, and F. Seno, Proteins **67**, 285 (2007).

[243] B. Hartke, J. Comput. Chem. **20**, 1752 (1999).

[244] F. Schulz and B. Hartke, Comput. Phys. Commun. **3**, 98 (2002).

[245] G. Trimarchi and A. Zunger, Phys. Rev. B **75**, 104113 (2007).

[246] A. E. Howard and P. A. Kollmann, J. Med. Chem. **31**, 1669 (1988).

[247] H. A. Scheraga, Chem. Rev. **71**, 195 (1970).

[248] T. E. Creighton, *Proteins. Structure and Molecular Principles*, (Freeman, New York, 1984).

[249] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, (Addison-Wesley, Reading, Massachusetts, 1989).

[250] L. Davis (Ed.), *Genetic Algorithms and Simulated Annealing*, (Pitman, London, 1987).

[251] L. David (Ed.), *Handbook of Genetic Algorithms*, (Van Nostrand Reinhold, New York, 1991).

[252] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, (Addison-Wesley, Reading, Massachusetts, 1989).

[253] R. S. Judson, Rev. Comput. Chem. **10**, 1 (1997).

[254] M. A. Duncan, Annu. Rev. Phys. Chem. **48**, 69 (1997).

[255] D. M. Deaven and K. M. Ho, Phys. Rev. Lett. **75**, 288 (1995).

[256] V. B. Zhurkin, V. I. Poltiev, and V. L. Florent'ev, Mol. Biol. (Moscow) **14**, 1116 (1980).

[257] J. Rammstein and R. Lavery, Proc. Natl. Acad. Sci **85**, 7231 (1988).

[258] M. Nanias, M. Chinchio, S. Oldziej ans C. Czaplewski, and H. A. Scheraga, J. Comput. Chem. **26**, 1472 (2005).

[259] C. Thachuk, A. Shmygelska, and H. H. Hoos, BMC Bioinf. **8:342**, (2007).

[260] A. A. Rabow and H. A. Scheraga, Prot. Sci. **5**, 1800 (1996).

[261] N. Go and H. A. Scheraga, Macromolecules **3**, 178 (1970).

[262] J. Stoer, *Numerische Mathematik 1*, (Springer-Verlag, Heidelberg, 1999).

[263] S. S. Plotkin and J. N. Onuchic, Quart. Rev. Biophys. **35**, 111 (2002).

[264] S. S. Plotkin and J. N. Onuchic, Quart. Rev. Biophys. **35**, 205 (2002).

[265] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[266] http://www.predictioncenter.org.

Danksagung

Hiermit danke ich meinem Betreuer und Doktorvater Prof. Dr. B. Hartke, der mir die Möglichkeit eröffnet hat, auf diesem Thema arbeiten zu können und stets eine offene Tür hatte. Ebenso danke ich Prof. Dr. A. Neumaier von der Universität Wien für viele Gespräche und ausführliche E-Mails zu aufgekommenen Problemen.

Ich danke meinen Eltern und meinem Bruder für die große Unterstützung vor und während des Studiums, die vieles erleichtert hat.

Ich danke dem Arbeitskreis für eine schöne Arbeitsatmosphäre und vielen lustigen Sitzungen.

Schließlich danke ich noch Verena und Steffi für ihre Hilfe und die schöne Zeit.

Vita

Florian Koskowski

Geburtsdatum:	13.10.1978
Geburtsort:	Kiel
Staatsangehörigkeit:	Deutsch
Anschrift:	Gravelottestr. 5
	24116 Kiel

1986 - 1989	Grundschule Kronsburg in Kiel
1989 - 1998	Käthe-Kollwitz-Gymnasium Kiel. Abschluss: Abitur
1998 - 1999	Grundwehrdienst
1999 - 2000	Studium der Materialwissenschaft an der Christian-Albrechts-Universität zu Kiel
2000 - 2004	Studium der Chemie an der Christian-Albrechts-Universität zu Kiel
2004 - 2005	Diplomarbeit im Fachbereich Chemie an der Christian-Albrechts-Universität zu Kiel Titel: ”Proteinfaltung mit evolutionären Algorithmen” Betreuer: Prof. Dr. B. Hartke
2005 - 2009	Promotion im Fachbereich Chemie an der Christian-Albrechts-Universität zu Kiel Titel: ”Entwicklung und Optimierung einer neuartigen Potentialfunktion mit Anwendung in der globalen Geometrie-Optimierung zur Vorhersage der Proteinfaltung” Betreuer: Prof. Dr. B. Hartke

Eidesstattliche Erklärung

Hiermit versichere ich, diese Dissertationsschrift unter Anleitung meines Betreuers Prof. Dr. B. Hartke selbständig verfasst und keine weiteren Hilfsmittel als die angegebenen benutzt zu haben. Die verwendeten Quellen sind deutlich gekennzeichnet.
Die Arbeit ist unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden.
Diese Arbeit hat weder in gleicher noch in ähnlicher Form im Rahmen irgendeines Prüfungsverfahrens vorgelegen. Frühere Promotionsversuche sind von mir nicht unternommen worden.

Kiel, den .....