Entwicklung einer richtungs– und abstandsabhängigen wissensbasierten Bewertungsfunktion für die Vorhersage der Thermostabilität von Proteinen

INAUGURAL – DISSERTATION

zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Christian Hoppe
aus Hannover

Köln 2002

1. Referent:	Prof. Dr. D. Schomburg
2. Referent:	Prof. Dr. S. Waffenschmidt
Eingereicht:	4.12.2002
Disputation:	

The distribution of distances for any given pair was determined by evolution, or by God, not by Boltzmann.

A. Ben-Naim, 1997

Danksagung

Diese Arbeit wurde bei Prof. Schomburg am Institut für Biochemie der Universität Köln von Mai 2000 bis November 2002 durchgeführt.
Mein besonderer Dank gilt Prof. Schomburg für die interessante Themenstellung und ständige Diskussionsbereitschaft.

Vera Grimm, Oliver Hofmann, Guido Hansen, Martin Weinand, Marcel Kevic und Daniel Wetzler für die nette Arbeitsatmosphäre und fachlichen Diskussionen, der Blast Gruppe, allen Kollegen und Kolleginnen des Arbeitskreises von Prof. Schomburg für die anregenden Unterhaltungen.
Nicht zu vergessen Sebastian, Christian, Guido, Heinrich, Jan und Oliver die Spieler von Midgard.

Herrn Prof. Dr. Gregor Tyrchan und Frau Uta Tyrchan für ihre Mühen.

Renan für dies und das. Meinen Eltern und Tina insbesondere.

Inhaltsverzeichnis

Abkürzungsverzeichnis

Aminosäuren

Alanin	Ala	A
Arginin	Arg	R
Asparagin	Asn	N
Aspartat	Asp	D
Cystein	Cys	C
Glutamat	Glu	E
Glutamin	Gln	Q
Glycin	Gly	G
Histidin	His	H
Isoleucin	Ile	I
Leucin	Leu	L
Lysin	Lys	K
Methionin	Met	M
Phenylalanin	Phe	F
Prolin	Pro	P
Serin	Ser	S
Threonin	Thr	T
Trypthophan	Trp	W
Tyrosin	Tyr	Y
Valin	Val	V

Sonstige Abkürzungen

σ	Standardabweichung
ρ	Wahrscheinlichkeitsdichtefunktion
φ	Phi
ψ	Psi
ω	Relative Orientierung in Grad
$\Delta\Delta G$	Stabilisierungsenergie
ΔE	Potential
a	Gewichtungsfaktor
Å	Angstroem (10^{-10} Meter)
aa	Die zwanzig Aminosäuren
aap	Aminosäure-Atom-Potential
all	Durchschnittsbildung über alle zwanzig Aminosäuren
AS	Aminosäure
at	Die gewählten Atomtypen
b	Gewichtungsfaktor
C	Kohlenstoff
C	Carbonylkohlenstoff des Aminosäure Rückrats
CA	C_α-Atom der Aminosäure
C_{ali}	Aliphatischer Kohlenstoff
C_{arom}	Aromatischer Kohlenstoff
CB	C_β-Atom der Aminosäure
CD	Circulardichroismus
C_p	Wärmekapazität
CZ	Geometrisches Zentrum der Seitenkette
d	Abstand (in Å)
DSC	*Differential Scanning Microcalorimetry*
E	Energie
EEEF	Empirisch effektive Energiefunktion

eq	Equilibrium
E_{ww}	Wechselwirkungsenergie
F	Faktor
G	Freie Enthalpie
g	Verteilungsfunktion
H	Enthalpie
HKA	Hauptkomponentenanalyse
k	Boltzmann-Konstante
K	Gleichgewichtskonstante
m	Wichtungsparameter
N	Stickstoff
N	Stickstoff des Aminosäure Backbones
n	Anzahl
NC	Entspricht V_{NC}
N_{ij}	Anzahl der Atompaare
O	Sauerstoff
O_{as}	Carbonylsauerstoff des Aminosäure Backbones
O_K	Kristallwasser
O_m	Modellierter Sauerstoff und Kristallwasser
O_{mo}	Modellierter Sauerstoff
P	Punkt
PDB	*Protein data bank*
PEEF	Physikalisch effektive Energiefunktion
r	Radius
R	Universelle Gaskonstante
r_{cor}	Korrelationskoeffizient
r_{hd}	Korrelationskoeffizient
rv	Richtige Vorhersage
S	Entropie
s.	Siehe
SAS	Lösungsmittelzugänglichkeit
SEEF	Statistisch effektive Energiefunktion

Sens	Sensibilität
Spez	Spezifität
T	thermodynamische Temperatur
tp	Torsionswinkelpotential
V	Kugelvolumen
V_{NC}	Vektor
W	*Potential of mean force*
Z	Zustandssumme

1. Einleitung

1.1. Einführung in die Problemstellung

Proteine sind Heteropolymere, deren Monomereinheit als Aminosäure bezeichnet wird. Die Faltungs- und Funktionsvielfalt der Proteine beruht auf der Kombination von zwanzig verschiedenen Aminosäuren. Diese Aminosäuren sind über Peptidbindungen miteinander zu einer Polypeptidkette verknüpft. Der Faltungsvorgang einer eindimensionalen Aminosäuresequenz zu der dreidimensionalen funktionellen Struktur des Proteins ist bis heute ein ungelöstes Problem der Proteinbiochemie [Sippl, 1999].

Im Gegensatz zu den meisten synthetischen Polymeren, die unterschiedliche Konformationen annehmen können, liegt ein Protein in nur einem definierten nativen Zustand vor [Robertson und Murphy, 1997]. Dieser Umstand und die Beobachtung, dass die Faltung eines Proteins nur von seiner Sequenz abhängig ist, wird auch als *thermodynamische Hypothese* (s. Kapitel 1.2.6) der Proteinbiochemie bezeichnet [Anfinsen, 1972]. Die biologische Funktion hängt in großem Maße von der Konformation des gefalteten Proteins ab, weniger von seiner Sequenz [Frauenfelder et al., 1988]. Unter physiologischen Bedingungen liegt ein Protein in seinem nativen Zustand vor. Werden diese Bedingungen geändert, kommt es zu einer Entfaltung, die als *Denaturierung* (s. Kapitel 1.2.5) bezeichnet wird.

Über den Faltungsvorgang und seine Zwischenzustände wird in der Literatur kontrovers diskutiert, da eine Beobachtung mit dem heutigen Stand der Technik nicht möglich ist [Klimov und Thirumalai, 2001; Nakamura, 2001]. Es wird auch vom *protein folding problem* gesprochen [Banavar und Maritan, 2001].

Die Kräfte, die bei der Faltung eines Proteins sowie bei seiner Stabilisierung eine Rolle spielen, sind Gegenstand zahlreicher wissenschaftlicher Untersuchungen. Eine Vielzahl von Experimenten wurde in den letzten zwei Dekaden durchgeführt, um die Faktoren, welche eine Erhöhung der Thermostabilität in thermophilen Proteinen (z.B. den Thermozymen) bewirken, zu identifizieren [Kannan und Vishveshwara, 2000].

Quantenmechanische Simulationen und *ab initio* Berechnungen sind für Makromoleküle noch zu aufwendig [Leach, 2001; Dinner et al., 2000]. Versuche, den Faltungsvorgang und die damit eng verknüpfte Frage nach der Proteinstabilität und Thermostabilität mit grundlegenden Kräften der Physik zu beschreiben und zu beantworten, scheitern an der Komplexität des Problems und der Unkenntnis über den Faltungsablauf.

Gute Ergebnisse werden mit weniger generellen theoretischen Methoden erzielt, dem *homology modelling* im Fall der Faltungsvorhersage [Novotny et al., 1984] oder den wissensbasierten Potentialen. Es wurde gezeigt, dass die Korrektheit einer Faltung [Maiorov und Crippen, 1992; Hendlich et al., 1990], die Faltung selbst [Lazaridis und Karplus, 2000; Moult, 1999; Moult et al., 1999; Jones und Thornton, 1996] und die Auswirkung einer Veränderung der Aminosäuresequenz eines Proteins (Mutation) auf seine Stabilität [Kocher et al., 1994; Topham et al., 1997; Miyazawa und Jernigan, 1994] mit wissensbasierten Potentialen richtig erkannt und vorhergesagt werden kann.

In dieser Arbeit wird ein wissensbasiertes Potential entwickelt, das eine quantitative Beschreibung der Wechselwirkung einer Aminosäure mit ihrer Umgebung ermöglicht. Die Implementierung der Ableitungen, Berechnungen und Bewertungen dieser wissensbasierten Bewertungsfunktionen für die Vorhersage der Thermostabilität von Proteinen erfolgte mit den Programmiersprachen C++ und Python (s. Kapitel 2.2.8).

1.2. Proteinfaltung

Globuläre Proteine nehmen in Lösung spontan eine wohl definierte dreidimensionale Konformation ein, den nativen Zustand.

Nach seiner Synthese an den Ribosomen liegt das Protein in einem ungefalteten Zustand vor. Dieser umgebungsabhängige Zustand zeigt keine definierte Struktur, eher flexible Formen, die schnell ineinander übergehen und verschiedene lokale Eigenschaften besitzen [Dinner et al., 2000]. Während des Faltungsprozesses gewinnt die Polypeptidkette an Kompaktheit, und die verschiedenen polaren und dissoziierbaren Seitenketten der Aminosäuren richten sich

entsprechend ihrer Eigenschaften aus. Das Protein durchläuft verschiedene Konformationen mit zunehmender Stabilität. Nicht kovalente Bindungen innerhalb des Proteins oder mit dem umgebenden Lösungsmittel werden gebildet und wieder gebrochen. Insbesondere Wasserstoffbrückenbindungen zwischen dem gelösten Protein und dem Solvens Wasser sowie im Lösungsmittel selbst spielen eine wichtige Rolle [Brady und Sharp, 1997; Eisenberg und McLachlan, 1986]. Bei der Zusammenlagerung von hydrophoben Aminosäureseitenketten zu einem Proteininneren werden Lösungsmittelmoleküle freigesetzt [Dill, 1990; Kauzmann, 1993]. Das gefaltete Protein besitzt einen dichtest gepackten rigiden und hydrophoben Kern mit einem Solvens exponierten flexiblen und variablen Mantel an geladenen Aminosäuren [Lee und Richards, 1971; Richards, 1974].

Die Faltung erfolgt in wenigen Millisekunden, manchmal in noch kürzerer Zeit. Trotz der in einer Zelle vorkommenden Faltungshelfermoleküle [Jaenicke und Seckler, 1997; Ellis und Hartl, 1999] bilden viele Proteine auch ohne diese Helfer *in vitro* korrekt ihre nativen Strukturen.

Die für die Proteinfaltung und Proteinstabilität grundlegenden Begriffe und Modelle werden in den folgenden Kapiteln vorgestellt und erläutert.

1.2.1. Modelle der Proteinfaltung

Die Aufklärung des Faltungsmechanismus, durch den ein Protein seinen nativen Zustand adaptiert, und die Frage nach dem Beitrag der Sequenz an der Strukturbildung und nach der Art dieses Beitrages ist Gegenstand zahlreicher Untersuchungen [Itzhaki *et al.*, 1995; Alm und Baker, 1999; Nölting und Andert, 2000; Klimov und Thirumalai, 2001]. Mit theoretischen und experimentellen Methoden können geschwindigkeitsbestimmende Schritte identifiziert und die Struktur von Faltungsintermediaten aufgeklärt werden. Experimentell lässt sich der Faltungsvorgang mit Hilfe der Fluoreszenz-, NMR-, Circulardichroismus-Spektroskopie und mit Aktivitätstests verfolgen [Galzitskaya *et al.*, 2000; Chakraborty *et al.*, 2001].

Es wurden verschiedene Theorien entwickelt, um die Proteinfaltung und die Geschwindigkeit ihrer Faltung erklären zu können. Einige der bekanntesten sind das *framework* Modell [Udgaonkar und Baldwin, 1988; Kim und Baldwin, 1990], der Kern-Wachstums-Mechanismus [Wetlaufer, 1973], der Diffusions-Kollisions-Mechanismus [Karplus und Weaver, 1994], das Modell des hydrophoben Kollapses [Dill, 1985; Dill, 1990] und das *funnel* Modell [Leopold *et al.*, 1992; Dill, 1999; Dinner *et al.*, 2000]. Das Kern-Kondensations-Modell [Fersht, 1995; Itzhaki *et al.*, 1995], eine Kombination des Modells des hydrophoben Kollapses mit dem Kern-Wachstumsmodell, führt zu einer besseren Beschreibung des Faltungsvorganges [Nölting und Andert, 2000].

1.2.2. Proteinstruktur

Im Zusammenhang mit Proteinkonformationen werden vier Strukturebenen definiert. Die Primärstruktur bezeichnet die eindimensionale Aminosäuresequenz. Die Sekundärstruktur beschreibt die räumliche Anordnung von eng in der linearen Sequenz beieinander liegenden Aminosäureresten. Einige dieser Anordnungen führen aufgrund von sterischen Beziehungen zu sich wiederholenden Strukturmustern, z. B. zum Element der α-Helix und des β-Faltblattes. Die dritte Strukturebene ist die der Tertiärstruktur, welche die räumliche Beziehung von auf der linearen Sequenz weit entfernt liegenden Aminosäureresten sowie das Muster der Disulfidbrücken beschreibt.
Die letzte Strukturebene, die Quartärstruktur, bezieht sich auf die räumliche Anordnung von Polypeptidketten, sogenannte Untereinheiten, innerhalb eines Proteins.

1.2.3. Molten Globule

Viele Proteine liegen unter physiologischen oder leicht denaturierenden Bedingungen nicht in ihrer nativen Form vor, sondern in einer Teilfaltung, die als *molten globule* bezeichnet wird [Gursky und Atkinson, 1996; Zhang und Matthews, 1998]. Proteine im *molten globule* Zustand weisen einen hohen Grad an Se-

kundärstruktur und vorgebildeter Tertiärstruktur auf, zeigen aber keine optimale und rigide Seitenkettenpackung. Der Zustand des *molten globule* stellt bei vielen Proteinen einen gut beobachteten Übergangszustand dar. Studien über den *molten globule* Zustand eines Proteins können, bei Nichtbeachtung von Seitenketteninteraktionen, Faktoren für die Proteinfaltung und die Proteinstabilität ermitteln. In der Literatur werden außerdem noch andere biologische Funktionen diskutiert: So dienen nicht vollständig gefaltete Proteine als konformationelle Schalter oder als Ziel für die Genregulation. Mutationen, welche die native Form eines Proteins in eine *molten globule* ähnliche Form umstrukturieren, scheinen mit der genetischen Prädisposition von Krankheiten zu korrelieren [Chakraborty *et al.*, 2001].

1.2.4. Proteinstabilität

Ein ungefaltetes Protein interagiert in zahlreichen Wechselwirkungen mit dem Lösungsmittel Wasser. Während der Faltung werden die nichtkovalenten Wechselwirkungen zu dem Lösungsmittel abgebrochen und neue nichtkovalente Wechselwirkungen der Proteinelemente untereinander eingegangen. Die hydrophoben Seitenketten tendieren dazu, sich vom polaren und protischen Solvens abzuwenden, um miteinander in Wechselwirkung zu treten. Zahlreiche Wasserstoffbrücken, Donatoren und Akzeptoren bauen hoch vernetzte Wasserstoffbrückenbindungssysteme auf, welche in Bildung von Sekundärstrukturelementen der Helix oder des Faltblattes münden können. Jede dieser dabei auftretenden Interaktionsenergien ist vom Betrag klein, doch aufgrund ihrer großen Anzahl ist die gesamte Interaktionsenergie in der nativen Faltung und dem denaturierten, dem ungefalteten Zustand riesig, d.h. jeweils mit einigen tausend Kilojoule pro Mol.

Ob ein Protein seine native Struktur annimmt, hängt von der Differenz der Gesamtinteraktionsenergien des gefalteten und des ungefalteten Zustandes ab. Es handelt sich um jeweils große und ähnliche, eng beieinander liegende Größen, die schwierig zu berechnen sind.

Eine Erhöhung der Proteinstabilität kann durch eine Stabilisierung des gefalteten Proteins, durch eine Destabilisierung des entfalteten Zustands, oder aber durch Kombination von beiden erzielt werden.
Die Stärke einer gegebenen Wechselwirkung, einer Stabilisierung oder einer Destabilisierung in einem gefalteten Protein, ist von seiner Umgebung abhängig. Es werden zahlreiche Experimente, z. B. Mutationsstudien, durchgeführt, um die Faktoren, welche einen stabilitätsverändernden Einfluss haben, zu identifizieren. Diese Ergebnisse fließen in die theoretischen Betrachtungen ein. Die experimentell gewonnenen Daten dienen als *Benchmark* für die Entwicklung, Verifizierung und Optimierung theoretischer Funktionen.
Die nichtkovalenten Interaktionen basieren auf vier grundlegenden Kräften, den hydrophoben-, den elektrostatischen-, den van der Waals Kräften und den Wasserstoffbrückenbindungen.

1.2.4.1. Hydrophober Effekt

Nicht polare Moleküle weisen eine vorteilhafte, negative freie Transferenthalpie von Wasser zu organischen Lösungsmittelsystemen auf. Bei Raumtemperatur haben nicht polare, aliphatische Ketten eine positive Transferenthalpie und Transferentropie und eine negative Transferwärmekapazität [Ratnaparkhi und Varadarajan, 2000]. Der absolute Wert dieser vier thermodynamischen Größen steigt mit der Vergrößerung der unpolaren Fläche an [Dill, 1990]. Bei der Faltung eines Proteins werden die anfangs Lösungsmittel exponierten unpolaren Seitenketten so bewegt, dass sie möglichst Lösungsmittel unzugänglich liegen. Diese hydrophoben Gruppen lagern sich spontan zusammen, es kommt zu einem hydrophoben Kollaps. Die negative Entropie bei der Hydration von unpolaren Gruppen im ungefalteten Protein gilt als eine der Haupttriebfedern für die Proteinfaltung und als einer der Faktoren für die Stabilisierung des *molten globule* [Dill, 1990].

1.2.4.2. Elektrostatische Kräfte

Elektrostatische Kräfte scheinen bei der Proteinfaltung eine untergeordnete Rolle zu spielen. Sie können in einem Protein zwischen Ladungen oder Polaritäten

auftreten oder auch zwischen Solvens und Protein. In speziellen Fällen wie bei nicht kompensierten Ladungen innerhalb des Proteins hat die Stärke der elektrostatischen Kräfte einen kritischen Einfluss auf die Faltung. Polare Atome unpolarer Aminosäuren tragen mit ihren Wechselwirkungen ebenfalls zur Interaktionsenergie bei [Nakamura, 1996].
Kompensierte Ladungen, sogenannte Salzbrücken, zeigten in einer Reihe von Mutationsstudien keinen oder nur sehr schwachen Einfluss auf die Proteinstabilität. Allerdings ist ihre Bedeutung für die Thermostabilität umstritten [Erwin *et al.*, 1990; Dao-pin *et al.*, 1991; Matthews, 1993; Kumar *et al.*, 2001].

1.2.4.3. Van der Waals Kräfte

Unter den nach van der Waals benannten zwischenmolekularen Kräften werden die Wechselwirkungen auf atomarer (Dispersionskräfte oder London-Kräfte) und molekularer Ebene, mittels induzierter oder permanenter Dipolmomente, verstanden. Ihr Anteil an der Proteinstabilisierung wächst mit der Größe des Proteinkerns [Lins und Brasseur, 1995].

1.2.4.4. Wasserstoffbrückenbindungen

Wie kovalente Bindungen besitzt die Wasserstoffbrückenbindung eine Vorzugsrichtung. Sie spielt eine zentrale Rolle bei der Proteinfaltung und der Stabilisierung von spezifischen Strukturelementen bei Proteinen [Chothia und Finkelstein, 1990; Richardson *et al.*, 1992; Sharp und Englander, 1994].

1.2.5. Denaturierung von Proteinen

Eine Denaturierung von Proteinen kann durch Änderung verschiedenster Umgebungsvariablen hervorgerufen werden. Die üblichen Methoden zur Denaturierung sind das Erhitzen, die Zugabe chemischer Denaturanten wie z. B. Harnstoff oder Guanidin Chlorid, eine Änderung des pH-Wertes oder die Anwendung von hohem Druck. Bei vielen, vor allem kleinen, Proteinen ist die Denaturierung ein reversibler Prozess, der von einer spontanen Rückfaltung begleitet werden kann, wenn die Umgebungskonditionen dies zulassen. Grosse Proteine hinge-

gen denaturieren oft irreversibel, da sie im ungefalteten Zustand aggregieren und aus der Lösung ausfallen. Das Problem der Aggregation kann auch bei kleinen, denaturierten Proteinen auftauchen bei entsprechender Erhöhung der Proteinkonzentration.

1.2.6. Thermodynamische Beschreibung der Proteinfaltung und Proteinentfaltung

Der native Zustand eines Proteins ist nur bedingt stabil, die freie Faltungsenthalpie $\Delta G_{Faltung}$ liegt im Bereich von 20-60 kJ/mol [Dill, 1990; Jaenicke und Seckler, 1997]. Dieser geringe Betrag ergibt sich aus der Differenz der freien Bildungsenthalpien $G_{Bildung}$ des ungefalteten Zustandes und des nativen, gefalteten Zustandes mit Gleichung 1.1:

$$\Delta G_{Faltung} = G_{Bildung}^{gefaltet} - G_{Bildung}^{entfaltet} \qquad (1.1)$$

Die freie Faltungsenthalpie dient als Maß für die Proteinstabilität. Nach der *thermodynamischen Hypothese* entspricht die niedrigste $\Delta G_{Faltung}$, bezogen auf das gesamte System, der nativen Form eines Proteins unter normalen physiologischen Bedingungen [Anfinsen, 1972].

Im Falle einer thermischen Denaturierung lässt sich der Prozess der Proteinfaltung oder der Proteinentfaltung mit der klassischen Thermodynamik beschreiben. Diese Prozesse müssen unter gleichen Bedingungen miteinander übereinstimmen [Galzitskaya *et al.*, 2000].

Der denaturierte oder ungefaltete Zustand, unter dem im Folgenden genau der Zustand gemeint ist, in dem sich ein Protein nach reversibler thermischer Zersetzung befindet, zeichnet sich durch einen hohen Grad an konformationeller Freiheit aus, die einer hohen konformationellen Entropie S entspricht (Gleichung 1.2). Das Protein liegt nicht als rigide, im Inneren eng gepackte Struktureinheit vor, sondern ganze Bereiche der Polypeptidkette können sich frei zueinander bewegen und die Seitenketten sind frei um die Einzelbindungen rotierbar.

$$S = k \cdot \ln W \quad (1.2)$$

W entspricht der Anzahl möglicher Zustände des Systems, k ist die Boltzmannkonstante

Im Gegensatz dazu besitzt die native Form eines Proteins nur eine geringe Entropie. Während der spontanen Faltung des Poteins muss der Verlust an konformationeller Entropie durch einen Zugewinn der Beträge an Enthalpie H kompensiert werden (s. Gleichung 1.3). Es wird davon ausgegangen, dass der Faltungsvorgang große negative Werte der Enthalpie sowie der Entropie aufweist. Um die thermodyamischen Größen der Faltung zu bestimmen, muss neben den Protein spezifischen enthalpischen und entropischen Größen die Thermodynamik des Solvens Wasser betrachtet werden. Entropie und Enthalpie des physiologischen Lösungsmittels addieren sich zu der Entropie und der Enthalpie des Proteins, um damit die thermodynamischen Eigenschaften des nativen und des denaturierten Zustandes adäquat zu beschreiben.
Die freie Faltungsenthalpie ergibt sich aus der klassischen Thermodynamik zu:

$$\Delta G_{\text{Faltung}} = \Delta H_{\text{Faltung}} - T\Delta S_{\text{Faltung}} \quad (1.3)$$

$$\Delta G_{\text{Faltung}(T)} = [\Delta H_{(T_2)} - \Delta C_p (T_2 - T_1)] - T_1 \cdot [\frac{\Delta H_{(T_2)}}{T_2} - \Delta C_p \cdot \ln(\frac{T_2}{T_1})] \quad (1.4)$$

T ist die absolute Temperatur

Die Wärmekapazität ΔC_p ist wie folgt definiert:

$$\Delta C_p = \frac{d(\Delta H)}{dT} = T \cdot \frac{d(\Delta S)}{dT} \quad (1.5)$$

Das Gleichgewicht zwischen denaturiertem (D) und nativem (N) Zustand des Proteins ist wie folgt definiert:

$$K = \frac{[N]}{[D]} \quad (1.6)$$

K ist die Gleichgewichtkonstante

[X] bedeutet: Die Konzentration von X

Daraus folgt nach der klassischen Thermodynamik :

$$\Delta G_{\text{Faltung}} = - RT \cdot \ln K \qquad (1.7)$$

R entspricht der universellen Gaskonstanten

1.2.6.1. Experimentelle Bestimmung thermodynamischer Größen

Die zur Stabilitätsbestimmung eines Proteins benötigten Daten (Schmelztemperaturen, thermodynamische Größen) lassen sich mit unterschiedlichen experimentellen Verfahren bestimmen. Die meisten dieser Messungen werden mit optischen oder kalimetrischen Methoden durchgeführt, einige wenige mit direkten oder indirekten Aktivitätstests.

Die wichtigste experimentelle Methode zur Bestimmung thermodynamischer Daten ist die *Differential Scanning Microcalometry* (DSC). Dabei wird die benötigte Wärmemenge zur Entfaltung eines Proteins relativ zu einer Pufferlösung als Funktion der Temperatur gemessen. Im Prinzip lassen sich aus einem DSC-Experiment alle thermodynamischen Daten für die Faltung oder die Entfaltung eines globulären Proteins ermitteln und berechnen (s. Gleichung 1.4).

Die optische Methode des Circulardichroismus (CD) nimmt das CD-Signal einer Proteinlösung bei einer Wellenlänge während des Schmelzvorganges auf. Dabei wird der Anteil von unterschiedlichen Sekundärstrukturelementen eines Proteins ermittelt. Die Gleichgewichtskonstante *K* kann über die Bestimmung der Konzentrationen von nativem und denaturiertem Protein mit der Gleichung 1.6 errechnet werden. Die freie Faltungsenthalpie ergibt sich bei gegebener Temperatur aus Gleichung 1.6 und Gleichung 1.7.

Die erhaltenen thermodynamischen Größen können nur bedingt miteinander verglichen werden und sind mit wechselnden Fehlern behaftet, deren Größe nur geschätzt werden kann. Dies gilt sowohl für den Vergleich der mit unterschiedlichen Methoden als auch der innerhalb einer Methode gewonnenen Größen [Robertson und Murphy, 1997].

1.3. Thermostabilität von Proteinen

Das Problem der Proteinfaltung (s. Kapitel 1.2) ist eng mit der Frage nach den ursächlichen Faktoren für die Proteinstabilität verbunden (s. Kapitel 1.1). Bei zahlreichen Vergleichsstudien hinsichtlich Sequenz und Strukturen homologer Proteine aus mesophilen (optimale Wachstumstemperatur: 15-45°C), thermophilen (optimale Wachstumstemperatur: 50-80°C) oder hyperthermophilen (optimale Wachstumstemperatur mehr als 80°C) Organismen konnten keine systematischen Unterschiede in Sequenz(-alignment), Kristallstruktur, Aminosäurezusammensetzung, Funktion und chemischen Eigenschaften gefunden werden. Für die Anpassung von Enzymen an hohe Temperaturen, sogenannte thermostabile Enzyme, werden kleinere Wechselwirkungen zwischen Strukturelementen und zahlreiche untergeordnete Sequenzunterschiede verantwortlich gemacht, die Protein oder Organismen spezifisch sein können [Adams und Kelly, 1998; Xiao und Honig, 1999; Kannan und Vishveshwara, 2000; Kumar *et al.*, 2000; Szilagyi und Zavodszky, 2000; Kumar *et al.*, 2001]. Sie resultieren in einer gesteigerten konformationellen Stabilität, die durch Erhöhung der Packungsdichte, eine optimierte Aminosäuresequenz in Abfolge polar/unpolar, durch Erhöhung der Anzahl von Ionenpaaren an der Oberfläche, durch eine Minimierung der Solvens zugänglichen hydrophoben Oberfläche, eine Helix-Stabilisierung oder eine optimierte Anordnung von Untereinheiten erzielt werden kann [Kumar *et al.*, 2000].

Die Rigidität von thermostabilen Enzymen und damit ihre Stabilität gegenüber denaturierenden Bedingungen steht im Gegensatz zu der für die Proteinfunktion notwendigen Flexibilität [Tuena de Gomez-Puyou und Gomez-Puyou, 1998; Bruins *et al.*, 2001]. Die Adaption von Proteinen an extreme Temperaturen scheint das Ergebnis eines Kompromisses zwischen diesen beiden Faktoren Rigidität und Flexibilität zu sein.

1.3.1. Anwendung von thermostabilen Enzymen in der Praxis

Thermostabile Enzyme zeigen häufig zusätzlich eine erhöhte Stabilität, bei konservierter Funktion, gegenüber pH-Werten oder extremen chemischen Bedingungen. Dies macht sie attraktiv für den Einsatz in einer Vielzahl von industriellen und biotechnologischen Prozessen (s. Tabelle 1.1). Dort können sie chemische Katalysatoren oder mesophile Enzyme in schon etablierten Verfahren ersetzen oder neue effiziente Möglichkeiten für bisher unzugängliche Hochtemperaturbereiche erschließen. Durch die Verwendung von thermostabilen Enzymen und folglich deren Einsatzmöglichkeit bei erhöhten Temperaturen kann die Ausbeute und Wirtschaftlichkeit von chemischen Prozessen positiv beeinflusst werden.

Es ergibt sich eine Reihe von Vorteilen:

- Bei der Durchführung von biotechnologischen Prozessen bei erhöhter Temperatur verringert sich das Risiko von Kontaminationen. Temperaturen über 70° C töten zuverlässig die meisten pathogenen Bakterien und verhindern die Verunreinigung mit Keimen bei Verfahren der Nahrungsmittelherstellung
- Erhöhte Temperaturen bergen für großtechnische Prozesse eine Reihe von verfahrenstechnischen Vorteilen: Erniedrigung der Viskosität von Reaktionslösungen, Erhöhung von Diffusionsraten, erhöhte Löslichkeit von Substraten, welche das chemische Gleichgewicht der Reaktion hin zu den Produkten verschieben kann, und die thermische Kontrolle der Reaktion
- Thermostabile Enzyme sind über einen größeren Temperaturbereich hinweg aktiv und bei konstanter Temperatur länger aktiv als ihre mesophilen Verwandten. Dadurch lässt sich der Enzymverbrauch verringern
- Bei endothermen Reaktionen kann durch Erhöhung der Reaktionstemperatur eine chemische Gleichgewichtsverschiebung zugunsten der Produkte erwirkt werden

Eine Erhöhung der katalytischen Geschwindigkeit wird allerdings nicht beobachtet [Bruins *et al.*, 2001].

Thermostabile, Polymer verdauende Enzyme spielen schon jetzt eine wichtige Rolle in Prozessen der Herstellung von Nahrungsmitteln, Textilien, Papier und Pharmazeutika (s. Tabelle 1.1).

Tabelle 1.1: Einige wichtige thermostabile Enzyme und ihre industrielle Anwendung

Enzym	Anwendung
Amylasen, Pullulanasen, Glukoamylasen	Auflösen von Stärke, Zuckerherstellung
Cellulase	Papierherstellung
DNA Polymerase	DNA Amplifikation und Sequenzierung
Lipasen	Waschmittelindustrie
Xylanasen, Laccasen	Bleichen von Textilien
Proteasen	Vielfältiger Einsatz in der Nahrungsmittel- ,Textil- und Waschmittelindustrie
Glukose/Xylose-Isomerasen	Zuckerherstellung

1.3.2. Methoden zur Erhöhung der Thermostabilität

Es existieren zahlreiche Möglichkeiten zur Erhöhung von Proteinstabilitäten, die experimentell untersucht und angewendet werden. Dazu gehören chemische Modifikationen durch kovalente Bindungen zwischen Proteinmolekülen, dem *crosslinking* [White und Olsen, 1987], an einzelnen Aminosäuren [Minotani *et al.*, 1979] oder an Aminosäuregruppen [Munch und Tritsch, 1990], durch Fixierung an einer inaktiven Matrix zur Immobilisierung des Enzyms [Fukui *et al.*, 1975; Toraya *et al.*, 1975; Saleemuddin, 1999]; außerdem der teilweise oder vollständige Einsatz organischer Lösungsmittel und stabilisierender Reagentien [Garza-Ramos *et al.*, 1992; Jiang und Dalton, 1994; Tuena de Gomez-Puyou und Gomez-Puyou, 1998].

Diese Methoden sind jedoch nicht *in vivo* oder unter ähnlichen Bedingungen anwendbar. Eine solche intrinsische Stabilitätserhöhung kann nur durch genetische Veränderung des Zielproteins erzielt werden. Für die Mutation eines Pro-

teins (Mutagenese) stehen die Methoden der direkten Evolution [Giver *et al.*, 1998; Song und Rhee, 2000; Lehmann und Wyss, 2001] oder des gezielten Aminosäureaustausches durch das *Protein Engineering* [Bruins *et al.*, 2001] zur Verfügung. Jedoch wird die gezielte Erhöhung der Thermostabilität durch den Austausch einer oder mehrerer Aminosäuren durch die hohe Anzahl der Austauschmöglichkeiten (Sequenzlänge *n* mit 19 multipliziert) und durch die Unkenntnis über den Beitrag der auszutauschenden Aminosäuren an der Proteinstabilität erschwert. Die dadurch notwendigen Mutationsscreens sind nur mit großem Aufwand durchführbar.

1.3.3. Untersuchung von Proteinstabilitäten durch gezielte Mutagenese

Die gezielte Mutagenese, das *Protein Engineering,* ermöglicht es, Faktoren für die Stabilität von Proteinen zu bestimmen und zu identifizieren. Der Einfluss einzelner Aminosäuren auf die Stabilität eines Proteins und ihr Anteil daran wird untersucht und die verantwortlichen Kräfte beschrieben [Matthews, 1993].

Die stetig anwachsenden Zahl von Mutationsexperimenten findet größtenteils Eingang in die über das Internet (http://www.rtc.riken.go.jp/jouhou/Protherm/protherm.html) veröffentlichte Datenbank ProTherm [Gromiha *et al.*, 1999; Gromiha *et al.*, 2000]. Sie bietet Übersicht über mehr als 11.000 Einträgen von Einzelmutationsstudien und Mehrfachmutationen von über 500 Proteinen unter verschiedensten Bedingungen. Dabei sind systematische Studien von größerem Interesse, da Untersuchungsbedingungen und Fragestellung meist gleichbleibend sind.

Durch die Einführung einer Mutation in ein Protein ändern sich die Eigenschaften der Mutante im Vergleich zu denen des Wildtyps. Die Signifikanz dieser Änderung auf struktureller und funktioneller Ebene gilt es vorauszusagen. Ein erfolgreiches Mutagenese Experiment, welches zur Verbesserung von thermischen Eigenschaften des Enzyms führen soll, hat immer den Erhalt von Struktur und Funktion zum Ziel.

Das Hinzufügen einer oder mehrerer Aminosäuren in den Proteinkern oder ihr Entfernen daraus bringt oft wesentliche Änderungen der Struktur, unter dem Verlust von Funktion, mit sich [Sondek und Shortle, 1992; Shortle und Sondek, 1995].

Austauschmutationen hingegen beeinflussen oder treffen nur selten Schlüsselstellen, können aber entscheidende Stabilitätsänderungen herbeiführen. Der Austausch einer Aminosäure, sprich das Entfernen und Zufügen von Atomen, bewirkt eine lokale strukturelle und elektronische Umgebungsänderung [Sondek und Shortle, 1992; Shortle und Sondek, 1995]. Daraus resultiert eine Änderung der freien Bildungsenthalpien des gefalteten und ungefalteten Zustandes. Die Mutation kann sich stabilisierend oder destabilisierend auf die beiden Zustandsformen auswirken. Eine einzelne Quantifizierung des Effektes an der Mutationsstelle ist nicht möglich, da sich nur $\Delta G_{Faltung}$ experimentell bestimmen lässt.

Aus der Differenz der freien Faltungsenthalpien von Wildtyp und mutiertem Protein berechnet sich die (de-)stabilisierende Wirkung der Mutation nach Gleichung 1.8:

$$\Delta\Delta G_{(De\text{-})stabilisierung} = \Delta G^{Faltung}_{Wildtyp} - \Delta G^{Faltung}_{Mutante} \qquad (1.8)$$

Bei einem mehrfachen Aminosäureaustausch kann der Beitrag der einzelnen Mutationen additiv zur Stabilität gewertet werden, auch wenn dies nicht streng quantitativ gilt [Serrano *et al.*, 1993; Zhang *et al.*, 1995].

1.3.3.1. Design von thermostabilen Proteinen

In einer Reihe von Experimenten wurden der Mechanismus der Thermostabilisierung und der Zusammenhang von Struktur und Stabilität untersucht. Sie erlauben Rückschlüsse auf mögliche Stabilisierungsstrategien für Enzyme. Im Folgenden sollen verschiedene Ansätze vorgestellt werden.

Die Änderung oder der Ersatz von flexiblen Bereichen, den *Loops*, kann zu einer Stabilisierung führen [Kumar *et al.*, 2000; Li *et al.*, 2002]. Die Einführung

oder Stabilisierung von Oberflächenladungen durch Salzbrücken [Erwin *et al.*, 1990; Matthews, 1993] oder Wasserstoffbrückenbindungen [Alber *et al.*, 1987] führen zu konformationeller Stabilität.

Die Beeinflussung der Packungsdichte des Proteinkerns durch gezielten oder systematischen Austausch von Aminosäureresten zur Eliminierung oder Vergrößerung von Hohlräumen [Kellis *et al.*, 1988; Matsumura *et al.*, 1988; Sandberg und Terwilliger, 1989; Shortle *et al.*, 1990] und die Stabilisierung von α-Helices durch Ladungsausgleich an ihren polaren Enden [Serrano und Fersht, 1989; Matthews, 1993; Marshall *et al.*, 2002] zeigen unterschiedlich erfolgreiche Ergebnisse.

Die Einführung zusätzlicher Disulfidbrücken ist eine der ältesten Techniken zur Beeinflussung der Thermostabilität. Dabei kann eine Aminosäure an passender Stelle gegen ein Cystein ausgetauscht oder dort ein Cysteinpaar eingebaut werden [Matsumura *et al.*, 1989]. An vielen Stellen kann das Cysteinpaar zu ungünstigen sterischen Wechselwirkungen führen. Darüber hinaus zerfällt Cystein oxidativ bei Temperaturen größer als 100° C [Bruins *et al.*, 2001].

Eine Reduzierung von konformationellen Freiheitsgraden der Polypeptidkette durch eine Einschränkung ihrer Flexibilität bewirkt eine Destabilisierung des entfalteten Zustandes und eine Stabilisierung der nativen Form [Imanaka *et al.*, 1986; Matthews *et al.*, 1987].

1.4. Computer gestützte Vorhersagen für das *Protein Engineering*

Rechnergestützte Modelle für die Berechnung von Struktur–Stabilitäts-Beziehungen umfassen Simulationsexperimente, bei denen atomare Strukturmodelle mit semi-empirischen Kraftfeldern verknüpft werden [Karplus *et al.*, 1991; Prevost *et al.*, 1991]. Diese Modelle erreichen gute Korrelationen der experimentellen mit den theoretisch berechneten Faltungsenthalpien. Sie beziehen sich aber oft auf ausgesuchte einzelne Proteine oder sogar nur auf einzelne Aminosäuren. Detailliertere Modelle führen zu abstrakten Ansätzen [Lee, 1994; Sippl, 1995].

Für die Berechnung von Faltungsenthalpien bieten sich Thermodynamische Störungsrechnungen oder die Thermodynamische Integration an.

Diese spezifischen Modelle sind schon für einzelne Mutationen oder ein einzelnes Protein langwierig und kompliziert [Leach, 2001]. Sie erfordern ähnlich den *per Hand* Vorhersagen viel (Rechen-)Zeit und Expertenwissen. Die im vorhergehenden Kapitel 1.3.3.1 vorgestellten Regeln und Methoden sind nicht ausreichend genau, da die Auswirkungen einer Mutation auf ihre Umgebung nur unzureichend berücksichtigt werden.

Es besteht der Bedarf an einem generelleren Ansatz, einer automatischen Methode, die, ausgehend von der dreidimensionalen Struktur eines Proteins, ein Mutationsprofil erstellen kann. Bei einer hohen Sensitivität der Erkennung von (de-)stabilisierenden Mutationen sollte eine Reduzierung möglicher Mutageneseziele von $n \cdot 19$ (n=Sequenzlänge) auf unter hundert möglich sein.

Die Vorhersagemethode soll an einer möglichst großen Anzahl von Proteinen bzw. Mutationen getestet werden. Auswertung und Bewertung der Vorhersagegüte von Stabilitätswerten erfolgen über statistische Methoden. Neben der Klassifizierung der vorhergesagten Mutation in richtig oder falsch (de-)stabilisierend kann auch bei gegebenen linearen Zusammenhang zwischen vorhergesagten und experimentellen Werten der Korrelationskoeffizient r (s. Kapitel 2.3.8) bestimmt werden.

Die erste automatische, Computer gestützte Methode, die zu einer Vorhersage erhöhter Thermostabilität führen sollte, war das Programm MODIP von Sowdhamini *et al*. [Sowdhamini *et al*., 1989]. Sie versuchten die Platzierung von möglichen Disulfidbrücken in einem Protein unter dem Erhalt der Ausgangsstruktur vorherzusagen. Dieser Ansatz zur Thermostabilitätserhöhung ist ein indirekter, denn es wird nicht eine mögliche Stabilitätsänderung bewertet.

Die erste direkte automatische Vorhersage der Thermostabilität wurde 1995 von Ota *et al*. [Ota *et al*., 1995] publiziert. Sie verwendeten eine aus einem 3D-Profil nach Bowie *et al.* [Bowie *et al.*, 1991; Lüthy *et al.*, 1992] empirisch abgeleitete Energiepotentialfunktion, ein wissensbasiertes Potential (s. Kapitel 1.4.1), um damit die Wechselwirkungsenergie zwischen einer Umgebung und jeder der an dieser Stelle potentiell einsetzbaren zwanzig Aminosäuren zu be-

schreiben. Ihr Potential setzt sich additiv aus den vier Termen: Seitenkettenpackung, Lösungsmittelzugänglichkeit, Wasserstoffbrückenbindungen und einer Beschreibung der lokalen Konformation zusammen.

Wang *et al.* [Wang *et al.*, 1996] erstellten ein Mutationsprofil aus einem wissensbasierten Potential, welches die Hauptkettenpolarität beurteilt.

Der Ansatz von Gilis und Rooman [Gilis und Rooman, 1996; Gilis und Rooman, 1997] beruht ebenfalls auf einem wissensbasierten Potential. Aus einem Strukturdatensatz wurden Aminosäure-Aminosäure-Potentiale und Torsionswinkelpotentiale bestimmt. Unterschiedliche Kombinationen der beiden Potentiale für verschiedene Lösungsmittelzugänglichkeiten wurden berechnet und überprüft.

Topham *et al.* [Topham *et al.*, 1997] wählten für ihre Vorhersage einen anderen Ansatz. Sie definierten 216 Aminosäureumgebungen, in die Hauptkettenkonformationen, Solvenszugänglichkeiten, Wasserstoffbrückenbindungen von Seitenketten und Wasserstoffbrückenbindungen von Hauptkettengruppen einflossen. Aminosäureaustauschwahrscheinlichkeiten wurden aus einem Strukturdatensatz homologer Proteine einer Familie bestimmt und einzeln oder kombiniert mit der Umgebungsbeschreibung für die Stabilitätsvorhersage eingesetzt.

1.4.1. Wissensbasierte Potentialfunktionen

Wissensbasierte Systeme beruhen auf der Annahme, dass auf der Grundlage einer repräsentativen Sammlung von Wissen eine Ableitung darin inhärent enthaltender Regeln und Gesetzmäßigkeiten möglich ist [Tanaka und Scheraga, 1976; Sippl, 1990]. Durch Anwendung statistischer Mechanik auf die auf molekularer Ebene beobachteten Häufigkeitsfunktionen von charakteristischen Wechselwirkungen in experimentell bestimmten Systemen lassen sich wissensbasierte Potentialfunktionen ableiten. Sie bewerten ein beobachtetes System von Wechselwirkungen dann als günstig, wenn dieses in der Nähe von Häufigkeitsmaxima einer repräsentativen Wissensbasis liegt. Mit der *inversen Boltzmann-Gleichung* (s. Kapitel 2.1.1) [Sippl, 1995] lassen sich Häufigkeitsverteilungen für interatomare Wechselwirkungen aus Proteinstrukturen in Freie

Energie Beiträge, *potentials of mean force*, bzw. wissensbasierte Potentiale umsetzen. Die Thermodynamische Grundlage sowie ihre Terminologie sind in der Literatur umstritten [Finkelstein und Gutin, 1995; Thomas und Dill, 1996; Ben-Naim, 1997; Koppensteiner und Sippl, 1998]. Dennoch erzielen wissensbasierte Potentiale bessere Ergebnisse als das entsprechende semi-empirische Kraftfeld oder *ab initio* Rechnungen [Jones und Thornton, 1996; Finkelstein, 1997; Moult, 1997]. Die Anwendbarkeit wissensbasierter Potentiale auf Probleme der Proteinstabilität und Thermostabilität wurde in Publikationen von Sippl [Sippl, 1990] und Gilis und Rooman [Gilis und Rooman, 1996] oder Wang *et al*. [Wang *et al.*, 1996] hinreichend gezeigt.

1.5. Aufgabenstellung

Für die Vorhersage der Thermostabilität von Proteinen ist die Erstellung eines Mutationsprofils der entscheidende Schritt während des *Protein Engineering*. Eine Vorhersage per Hand ist zeitaufwendig und erfordert ein großes Expertenwissen. Das Ziel dieser Arbeit ist es, einen Ansatz zur automatischen Erstellung eines Mutationsprofils für ein zu thermostabilisierendes Enzym zu entwickeln. Die einer Stabilisierung zugrunde liegenden physikalischen Kräfte sind noch nicht vollständig bekannt. Es soll eine Energiefunktion entwickelt werden, welche die Umgebung einer Aminosäure quantifizieren und die Wirkung eines Aminosäureaustausches bewerten kann. Dazu wird eine richtungs- und abstandsabhängige wissensbasierte Aminosäure-Atom-Potentialfunktion beschrieben, die mit einem wissensbasierten Torsionswinkelpotential kombiniert werden soll [Leven, 1999]. Die Differenz aus dem Wechselwirkungsbeitrag für das native, unveränderte Protein, den Wildtyp und das mutierte Protein, die Mutante, stellt ein Maß für die Stabilitätsänderung dar.

Die Vorhersagefunktion wird an einem durch Literaturrecherche möglichst großen Satz an experimentellen Daten entwickelt und ihre statistische Auswertung durchgeführt.

2. Theorie und Methoden

Die vollständige Beschreibung der physikalischen Kräfte innerhalb eines Proteins, ihres ineinander Wirkens während der Proteinfaltung sowie ihres Einflusses auf die Proteinstabilität steht noch aus. Dies erschwert die Simulierung und quantitative Erfassung dieser Kräfte durch *ab initio* Rechnungen, physikalisch effektive Energiefunktionen (PEEF), semi-empirische Kraftfelder, molekulare Mechanik und Methoden der molekularen Dynamik.

PEEF-Methoden sind äußerst zeitaufwendig und erfordern einen massiven Rechenaufwand. Für die molekulare Faltungssimulation von 1 µs des Villin-Kopfstückes, mit seinen 36 Aminosäuren und ungefähr 3000 modellierten Wassermolekülen, zu einem stabilen Faltungsintermediat wurden vier Monate Rechenzeit und ein Parallelrechner mit 256 Prozessoren benötigt [Duan und Kollman, 1998; Duan *et al.*, 1998; Leach, 2001]. Die Rechenzeit kann durch eine implizite Beschreibung von Solvatationsenergien und der Seitenkettenentropie verringert werden. Trotzdem können nur freie Faltungsenthalpien von einzelnen und wenigen Mutationen berechnet werden [Kollman *et al.*, 2000].

Dem stehen die statistischen Energiefunktionen (SEEF) gegenüber. Sie extrahieren aus einem repräsentativen Satz von Proteinstrukturen effektive, wissensbasierte Potentiale [Tanaka und Scheraga, 1976; Sippl, 1990; Lazaridis und Karplus, 2000]. Mit ihnen kann die Wechselwirkung zwischen verschiedenen Strukturelementen eines Proteins beschrieben werden. Sie beinhalten jede Krafteinwirkung, die für eine Zustandsabweichung von einer zufälligen, statistischen Verteilung der Strukturmotive verantwortlich ist, ohne diese Wechselwirkungsbeiträge zu spezifizieren.

Für die Stabilitätsvorhersage in der vorliegenden Arbeit werden die lokalen und die nicht lokalen Wechselwirkungen einer jeden Aminosäure in einem Protein mit ihrer Umgebung erfasst und in einem wissensbasierten Potential beschrieben. Ein solches Potential dient als Bewertungsfunktion für die Stabilität von möglichen Mutationen. Ein Mutationsprofil ergibt sich aus der Anwendung der extrahierten Potentialfunktion auf die Bestimmung der Wechselwirkung einer jeden Aminosäureumgebung einer Struktur mit allen dort einsetzbaren, definier-

ten Aminosäuren und enthält die vorhergesagten energetischen Auswirkungen dieser möglichen Mutagenese Experimente.
Die Ableitung und die Bewertung des wissensbasierten Potentials erfolgt an einem repräsentativen Satz von Proteinstrukturen. Sie sollten die durch das Potential abzuleitenden Kräfte möglichst genau enthalten und entsprechend der Problemstellung gewählt werden [Furuichi und Koehl, 1998].
Die im Folgenden vorgestellten Methoden zur Validierung und Optimierung der wissensbasierten Bewertungsfunktion zur Thermostabilitätsvorhersage benötigen einen weiteren Datensatz an experimentellen Mutationsdaten (s. Kapitel 2.2.6).

2.1. Wissensbasierte Potentiale

Die expliziten Beschreibungen enthalpischer intermolekularer Wechselwirkungsbeiträge und der entropischen Effekte (der konformationelle Freiheitsgrad der Polypeptidkette, (De-)Solvatationen der Polypeptidkette bzw. Reorganisation des Lösungsmittels um die Polypeptidkette) sind für die Bewertung des Stabilitätseinflusses einer Mutation nicht möglich [Leach, 2001].
Unter der Annahme, dass sich experimentell bestimmte native Proteinstrukturen in einem globalen energetischen Minimum befinden [Anfinsen, 1972], unterscheiden sich die Verteilungen von Strukturelementen eines Proteins in diesem Zustand von einer zufälligen, statistischen Verteilung.
Die Ableitung und die Beschreibung der implizit in den Strukturen enthaltenden Effekte erfolgen durch die wissensbasierten Potentiale. Die physiko-chemische Grundlage des in diesem Zusammenhang verwendeten Begriffes eines *potential of mean force* geht auf die statistisch-mechanische Theorie von Flüssigkeiten zurück [Hill, 1986; Ben-Naim, 1987]. Die mikroskopische Struktur eines molekularen Systems wird durch eine n-Teilchen Korrelationsfunktion $g^{(n)}(r_1,..,r_i,..,r_n)$, mit r_i als Ortskoordinaten des *i*-ten Teilchens, beschrieben und kann in ein *potential of mean force W* übergeführt werden, es gilt:

$$W^{(n)}(r_1,..,r_i,..,r_n) = -RT \ln g^{(n)}(r_1,..,r_i,..,r_n) \quad (2.1)$$

Die Paarkorrelationsfunktion für zwei wechselwirkende Teilchen $g^{(2)}(r_1, r_2)$ geht unter der Annahme sphärischer Symmetrie in die radiale Verteilungsfunktion $g^{(2)}(r_{12})$, wobei $r_{12} \equiv |r_1 - r_2|$, über [Gohlke *et al.*, 2000]. Durch Auszählen der Kontakthäufigkeiten von Paaren wechselwirkender Atome im Distanzintervall $[r_{12}, r_{12} + dr]$ kann die radiale Paarverteilungsfunktionen und das *potential of mean force* aus Proteinkristallstrukturen abgeleitet werden, s. Gleichung 2.2 und Gleichung 2.3. Dabei entspricht $n(r_{12})$ der beobachteten Besetzung und $n_{id}(r_{12})$ der statistisch erwarteten Besetzung von Paaren bei einem Abstand r_{12} eines Schalenelements dr.
Damit ergibt sich für die Paarkorrelation:

$$g^{(2)}(r_{12}) = \frac{n(r_{12})}{n_{id}(r_{12})} \quad (2.2)$$

und das *potential of mean force*:

$$W^{(2)}(r_{12}) = -RT \ln g^{(2)}(r_{12}) \quad (2.3)$$

Dies gilt unter folgenden Annahmen [Thomas und Dill, 1996; Gohlke *et al.*, 2000]:

- Die Verteilungen von Paarhäufigkeiten sind voneinander unabhängig.
- Die Verteilungen von interatomaren Distanzen eines Atompaares sind ähnlich für verschiedene Umgebungen.
- Die Verteilungen der Distanzen für Paare sind hinreichend scharf und voneinander getrennt.
- Die Wechselwirkung zwischen zwei Paaren ist symmetrisch.

2.1.1. Beschreibung der inversen Boltzmanngleichung

Befindet sich ein physikalisches System im thermodynamischen Gleichgewicht, dann kann die Verteilung von Molekülen über die Mikrozustände mit dem Gesetz von Boltzmann beschrieben werden. Die Wahrscheinlichkeitsdichtefunktion ρ eines Systems wird mit seiner Energie E in Bezug gesetzt. Die folgende Herleitung der *inversen Boltzmanngleichung* folgt der Darstellung von Sippl [Sippl, 1990; Sippl, 1993]. Für ein betrachtetes Paar (i,j) im Abstand r_{ij} gilt:

$$\rho_{ij}(r_{ij}) = Z_{ij}^{-1} \cdot \exp(-E_{ij}(r_{ij}) / kT) \tag{2.4}$$

Die Zustandssumme Z_{ij} ist definiert als:

$$Z_{ij} = \sum_{ij} \exp\left(-E_{ij}(r_{ij}) / kT\right) \tag{2.5}$$

Wenn die Energien $E_{ij}(r_{ij})$ aller Zustände des Systems bekannt sind, können die entsprechenden Verteilungen von i und j bestimmt werden. Umgekehrt kann aus der Wahrscheinlichkeitsdichtefunktion $\rho_{ij}(r_{ij})$ eines Systems die Energie berechnet werden.

$$E_{ij}(r_{ij}) = -kT \ln(\rho_{ij}(r_{ij})) - kT \ln Z_{ij} \tag{2.6}$$

Gleichung 2.6 wird als die inverse Boltzmanngleichung bezeichnet.
Der Vergleich von Energien der Subsysteme mit einem Referenzsystem E(r), resultiert in einem Netto-Potential $\Delta E_{ij}(r_{ij})$ [Sippl, 1990].

$$\Delta E_{ij}(r_{ij}) = E_{ij}(r_{ij}) - E(r) = -kT \ln\left[\frac{\rho_{ij}(r_{ij})}{\rho(r)}\right] - kT \ln\left[\frac{Z_{ij}}{Z}\right] \tag{2.7}$$

Die Zustandssummen sind nur schwer zugänglich und können nicht aus den Wahrscheinlichkeitsdichtefunktionen bestimmt werden. Sind Z und Z_{ij} konstant, gilt in Näherung $Z \approx Z_{ij}$ und damit $-kT\ln\left[\frac{Z_{ij}}{Z}\right] \approx 0$.

Die radiale Verteilungsfunktion eines Paares (i,j) im Abstand r_{ij} $g_{ij}(r_{ij})$ konvergiert nach dem Gesetz der großen Zahlen mit der Wahrscheinlichkeitsdichte:

$$\rho_{ij}(r_{ij}) : \lim_{n\to\infty} g_{ij}^{(2)}(r_{ij}) \equiv \rho_{ij}(r_{rj}) \qquad (2.8)$$

Daraus folgt:

$$\Delta E_{ij}(r_{ij}) = E_{ij}(r_{ij}) - E(r) = -kT\ln\left[\frac{g_{ij}^{(2)}(r_{ij})}{g^{(2)}(r)}\right] \qquad (2.9)$$

2.1.1.1. Berücksichtigung der Direktionalität in den Paarpotentialen

Die Existenz sowie die Stärke spezifischer Wechselwirkungen zwischen polaren funktionellen Gruppen, aromatischen Systemen oder Wasserstoffbrückenbindungen hängen von deren Entfernung und gegenseitigen relativen Orientierung ab [Böhm *et al.*, 1996; Cole *et al.*, 1998; Grzybowski *et al.*, 2000; Kannan und Vishveshwara, 2000]. Eine distanzabhängige Paarverteilungsfunktion weist einen kugelsymmetrischen Potentialverlauf auf, in dem die Wechselwirkungsstärke unabhängig vom Winkel beschrieben wird. Eine winkel- und abstandsabhängige Paarverteilungsfunktion $g^{(2)}(r_{12}\omega_{12})$ umfasst zusätzlich die relative Orientierung ω_{12} des einen Strukturelementes zum anderen.

$$\Delta E_{ij}(r_{ij}\omega_{ij}) = E_{ij}(r_{ij}\omega_{ij}) - E(r\omega) = -kT\ln\left[\frac{g_{ij}^{(2)}(r_{ij}\omega_{ij})}{g^{(2)}(r\omega)}\right] \qquad (2.10)$$

Die räumliche Umgebungscharakterisierung erfordert die Definition eines für alle Aminosäuren (s. Kapitel 2.2) anwendbaren Bezugssystems. In einem drei-

dimensionalen Raum ergibt sich ein solches Bezugssystem als Ebene E mit den drei Punktrepräsentanten $P_0(r_0)$, $P_1(r_1)$, $P_2(r_2) \in \mathrm{E}$, wobei r_i die Ortskoordinaten des *i*-ten Punktes darstellen. Diese Repräsentanten (s. Tabelle 2.1) finden sich in der Literatur und den dort verwendeten wissensbasierten Potentialfunktionen (CA,CB [Bryant und Lawrence, 1993; Huang *et al.*, 1995];CZ [Kocher *et al.*, 1994; Ota *et al.*, 1995]). Der Normalenvektor n der Ebene berechnet sich aus dem Vektorprodukt (s. Kapitel 2.3.3) von $\mathbf{P_0P_1}$x$\mathbf{P_0P_2}$, wobei $\mathbf{P_0P_1}$, $\mathbf{P_0P_2}$, $\mathbf{P_0P_1}$x$\mathbf{P_0P_2}$ ein Rechtssystem bilden. P_0 ist genau der Punkt, zu dem der diskrete Abstand der Strukturelemente im Intervall $[r_0,r_0+dr)$ bestimmt wird. Der Vektor $\mathbf{P_0P_1}$ entspricht dem *Richtungsvektor*, der die Bezugsgröße für die relative Orientierung ω zweier betrachteter Strukturelemente darstellt (s. Abbildung 2.1). Die räumliche Anordnung der Strukturelemente zueinander wird durch Polarkoordinaten repräsentiert, d.h. alle Strukturelemente, die auf einem Richtungsvektor mit $|\mathbf{P_0P_1}| \to \infty$ liegen, entsprechen einer Orientierung von $\omega_{ij}=(\varphi=0°; \psi=0°)$ (s. Kapitel 2.3.5) zu dem zentralen Strukturelement.

Tabelle 2.1: Repräsentanten der Punkte P_0,P_1 und P_2 (+: entsprechenden Atomtyp verwendet)

Atomtyp [a)]	P_0 und P_1	P_2
CA	+	+
CB	+	-
CZ [b)]	+	-
O	-	+
N	+	+
C	-	+
V_{NC} [c)]	-	+

[a)] Die Bezeichnung der Atomtypen folgt, wenn nicht anders angegeben, der PDB-Notation (PDB). [b)] Entspricht dem geometrischen Mittelpunkt einer Aminosäure. Für die Aminosäure Glycin entspricht der geometrische Mittelpunkt dem modellierten CB. [c)] Diese Bezeichnung entspricht dem Vektor $\mathbf{V_{NC}}= (\mathbf{N}+\mathbf{C})/2$

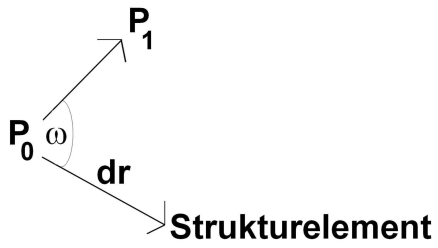


Abbildung 2.1: Darstellung der relativen Orientierung und des Abstandes eines Strukturelementes zu dem Richtungsvektor P_0P_1

2.1.2. Referenzzustand

Bei Benutzung eines Referenzzustandes können die Verteilungen der Strukturelementpaare in Bezug gesetzt und verglichen werden. Betrachtete Strukturelementpaare können Aminosäure-Aminosäure-Kontakthäufigkeiten, Torsionswinkel-, Sekundärstruktur-Verteilungen, Atom-Atom-Kontakthäufigkeiten oder Solvatisierungsgrade darstellen. Die Wahl eines geeigneten Referenzzustandes ist schwierig. Im Falle der Stabilitätsbewertung von Mutationen entspricht der Referenzzustand dem nicht nativen, also dem entfalteten oder denaturierten Zustand. Es gibt kein schlüssiges Konzept für diesen Proteinzustand, aber nicht alle möglichen Proteinkonformationen werden *a priori* zugänglich sein (s. Kapitel 1.2). Der Referenzzustand muss physikalisch sinnvoll allen möglichen, zufälligen Verteilungen der betrachteten Proteinstrukturelemente entsprechen. Dieser Zustand wird über ein Ensemble von Elementen über viele, verschiedene Proteinstrukturen angenähert [Finkelstein und Gutin, 1995]. Diese statistische, als zufällig betrachtete Verteilung sollte keine Präferenz für ein gewähltes Strukturelement aufweisen.

Durch Auszählen der Häufigkeit des Auftretens der Strukturelemente i und j, summiert über ein Protein *P* für alle *k* gewählten Proteine *GP* im Intervall $[r_d, r_d+dr)$ mit der relativen Orientierung ω_d zueinander, ergibt sich die Anzahl der Atompaare $N_{ij}(r_d\omega_d)$ wie folgt:

$$N_{ij}(r_d\omega_d) = \sum_{k\in GP}\sum_{p\in P}\delta(d_{ij}, r_d, \omega_d) \qquad (2.11)$$

wobei $\delta(d_{ij},r_d,\omega_d)$ =,1 wenn $d_{ij} \in [r_d,r_d+dr)$, sonst ist $\delta(d_{ij},r_d,\omega_d)=0$.

Der Referenzzustand $g(r\omega)$ (s. Gleichung 2.10) entspricht:

$$g(r\omega) = \frac{\sum_{k\in K}\sum_{p\in P} g_{ij}^{(2)}(r_d\omega_d)}{N_{ij}(r_d\omega_d)} \qquad (2.12)$$

2.1.3. Problem geringer Datenmengen

Die experimentell zugängliche radiale Verteilungsfunktion $g_j(r_d)$ (s. Gleichung 2.2) kann nur bei einer entsprechend hoher Datenmenge an die Wahrscheinlichkeitsdichtefunktion $\rho_{ij}(r_d)$ (s. Gleichung 2.4) angenähert werden. Die Kontakthäufigkeit für die Paare ist aber oft zu gering, um statistisch relevante Aussagen treffen zu können. Sippl [Sippl, 1990] gibt einen Grenzwert von 50 Beobachtungen an, darunter trifft die Annahme der Konvergenz (s. Gleichung 2.8) nicht mehr zu. Allerdings stellen auch einzelne Beobachtungen eines Strukturpaares in Abhängigkeit von Abstand und Orientierung eine wichtige Information da, die nicht verworfen werden darf. Die Potentialkurve soll eine Wiedergabe günstiger Strukturpaaranordnungen, mit vielen Beobachtungen und ungünstigen Strukturpaaranordnungen, mit keiner oder nur geringer Kontakthäufigkeit sein. Um die Information *kein Kontakt* oder *geringer Kontakt* in die Potentialfunktion ΔE_{ij} einfließen lassen zu können, ergibt sich nach Sippl [Sippl, 1990]:

$$\Delta E_{ij}(r_{ij}) = kT\ln[1+m\sigma] - kT\ln\left[1+m\sigma\frac{g_{ij}^{(2)}(r_{ij})}{g(r)}\right] \qquad (2.13)$$

m und σ stellen experimentelle Wichtungsparameter dar, nach Sippl [Sippl, 1990] entspricht σ = 0,002 und *m* der Anzahl von Beobachtungen eines Paares

Die Beobachtung eines Kontaktes hat zwei Auswirkungen. Zum einen werden alle Energien $\Delta E_{ij}(r_{ij})$ zu höheren Energien um den Beitrag $kT\ln[1+m\sigma]$ ver-

schoben, während die Energie eines Zustandes um $-kT\ln\left[1+m\sigma\frac{1}{g(r)}\right]$ verringert wird.

In dieser Arbeit werden mit Gleichung 2.14 die Kontakthäufigkeiten in Abhängigkeit von Abstand und Richtung beobachtet:

$$\Delta E_{ij}(r_{ij}\omega_{ij}) = kT\ln[1+m\sigma] - kT\ln\left[1+m\sigma\frac{g_{ij}^{(2)}(r_{ij}\omega_{ij})}{g(r\omega)}\right] \qquad (2.14)$$

Dieses Potential nach Sippl [Sippl, 1990] ist eine Beschreibung der Präferenz des Auffindens eines Strukturelementes zu allen anderen betrachteten Strukturelementen. Das Sippl-Potential wird auch als statistische Netto-Präferenz bezeichnet [Gohlke *et al.*, 2000]. Mit diesen Potentialfunktionen (s. Gleichung 2.13 und Gleichung 2.14) lassen sich die Fälle geringer Datenmengen lösen.

2.2. Die wissensbasierte Bewertungsfunktion zur Vorhersage der Thermostabilität von Proteinen

2.2.1. Beschreibung der Aminosäureumgebung und der in ihr wirkenden Kräfte

Die Struktur eines Proteins wird durch die Wechselwirkungen der Aminosäuren mit ihrer proteinogenen Umgebung und dem Solvens Wasser bestimmt. Den Einfluss einer Aminosäure auf jede andere eines Proteins sowie auf das Lösungsmittel gleichzeitig zu berücksichtigen, entspricht einem komplexen und aufwändig zu behandelnden Mehrkörperproblem. Die Wechselwirkungsbeschreibung vereinfacht sich, wenn das Protein in Umgebungen eingeteilt wird. Es wird zuerst nur noch der Einfluss einer definierten direkten Umgebung auf die Aminosäure betrachtet, und die Summe dieser Umgebungen soll dem Gesamtwechselwirkungsbeitrag entsprechen. In einer weiteren Näherung wird der Einfluss von nicht direkter Umgebung, also der restlichen Struktur, als konstant betrachtet. Die Umgebung einer Aminosäure kann mit Sequenzinformation und

mit Strukturinformation beschrieben werden. Auf die Aminosäure wirken die zwei prinzipiellen Kräfte nicht lokaler und lokaler Natur, die sich in der räumlichen dreidimensionalen Anordnung der Polypeptidkette widerspiegeln. Die Umgebung einer Aminosäure kann als räumliche Größe oder entlang der Aminosäuresequenz genauso individuell und flexibel definiert werden wie die zu betrachtenden Strukturelemente. Aus den Strukturelementen setzt sich die Umgebungsbeschreibung einer Aminosäure zusammen. Eine Charakterisierung der Umgebung geschieht demnach durch die richtungs- und entfernungsabhängige Bestimmung der Anzahl in Beziehung zu setzender Strukturelemente.

Das wissensbasierte Potential zur Beschreibung der Stabilitätsänderung durch eine Mutation sollte nicht nur die zwei prinzipiellen, in der Literatur diskutierten Kräfte mit stabilisierender Wirkung auf die native Struktur eines Proteins erfassen, sondern auch die reduzierte Umgebung möglichst detailliert beschreiben. Dazu werden in dieser Arbeit zwei Potentiale, ein Aminosäure-Atom-Potential und ein Torsionswinkelpotential [Leven, 1999]. Das Aminosäure-Atom-Potential soll durch die richtungs- und abstandsabhängige Beschreibung der dreidimensionalen Häufigkeitsverteilung von Atomen um eine Aminosäure nicht lokale Kräfte erfassen. Über die Torsionswinkel φ und ψ der betrachteten Aminosäure werden die direkten, lokalen Wechselwirkungen mit ihren Nachbarn und die entstehenden Restriktionen des Konformationsraumes durch kovalente Bindungen innerhalb der Polypeptidkette berücksichtigt.

2.2.2. Das richtungs- und abstandsabhängige Aminosäure-Atom-Potential

Als klassisches wissensbasiertes Potential zur Faltungserkennung und Stabilitätsvorhersage dient ein Aminosäure-Aminosäure-Potential. Mit der Beschreibung der Umgebung durch Atome statt Aminosäuren soll eine bessere Auflösung der Umgebungswechselwirkungen erreicht werden. Ein rein abstandsabhängiges Aminosäure-Atom-Potential wurde von Dengler und Leven entwickelt [Dengler, 1998; Leven, 1999]. In dieser Arbeit hingegen wird die Anzahl der in Beziehung zu setzenden Aminosäure-Atom-Strukturelementpaare entfernungs-

und richtungsabhängig um gewählte Repräsentanten der Aminosäure bestimmt (s. Kapitel 2.1.1.1). Bei einer distanzabhängigen Betrachtung mit diskreten Radien ergeben sich Schalen um die Aminosäure, deren Strukturelemente erfasst und anteilig gezählt werden. Die wechselwirkende Aminosäure wird als Kugel approximiert. Für jedes Radienintervall berechnet sich ein individueller Wechselwirkungsbeitrag zwischen Aminosäure und Atom [Leven, 1999]. Bei einer richtungsabhängigen Betrachtung werden bei definiertem Öffnungswinkel des Richtungsvektors Kegel erhalten, die mit einer festgelegten Schrittweite eine Kugel durchfahren. Die abstands- und richtungsabhängige Betrachtung resultiert in Kugelschalen, die von Kegeln mit definiertem Öffnungswinkel und definierter Schrittweite geschnitten werden. Im Fall eines Öffnungswinkels von 180° wird eine nur abstandsabhängige Betrachtung erreicht, ein Öffnungswinkel von 90° ergibt zwei aus Schalen aufgebaute Halbkugeln (s. Abbildung 3.1). Die Atome werden in diskreten Abständen und Winkelintervallen erfasst und ein spezifischer abstands- und richtungsabhängiger Wechselwirkungsbeitrag zwischen Aminosäure und Atom berechnet (s. Gleichung 2.14).

2.2.3. Anwendung der Aminosäure-Atom-Potentialfunktion

Das abstands- und richtungsabhängige Aminosäure-Atom-Potential bestimmt den energetischen Wechselwirkungsbeitrag eines Atoms zu einer Aminosäure im Abstand r und der relativen Orientierung ω. Die Erstellung des Mutationsprofils erfolgt durch eine Strukturanalyse des zu stabilisierenden Proteins. Dabei wird die Umgebung jeder Aminosäure den gewählten Kriterien entsprechend nach Atomtyp, Radien- und Winkelintervall ausgewertet und festgehalten. Es ergibt sich eine Charakterisierung der Umgebung, aus der sich die Präferenz, einen Atomtyp in bestimmtem Abstand und bestimmter Richtung zu einer Aminosäure zu finden, ableiten lässt. Die Atomtypen werden anteilig gezählt und ihre Häufigkeit mit den entsprechenden Potentialwerten für das Aminosäure-Atom-Paar multipliziert. Daraus ergeben sich die Energiebeiträge der Atomtypen in Abhängigkeit von Abstand und Richtung zu den Repräsentanten der A-

minosäure. Diese werden über beliebig gewählte Abstands- und Richtungsintervalle zum Wechselwirkungsbeitrag eines Atomtyps aufsummiert.

$$\Delta E_{ij}(r_{ij}\omega_{ij}) = kT\ln[1+m\sigma] - kT\ln\left[1+m\sigma\frac{g_{ij}^{(2)}(r_{ij}\omega_{ij})}{g(r\omega)}\right] \quad (2.14)$$

$$\Delta E_{at,r,\omega}^{aa} = kT\ln[1+m\sigma] - kT\ln\left[1+m\sigma\frac{n_{at,r,\omega}^{aa}}{n_{at,r,\omega}^{all}}\right] \quad (2.15)$$

$$E_{at}^{aa} = \sum\left(n_{at,r,\omega}^{aa}\cdot\Delta E_{at,r,\omega}^{aa}\right) \quad (2.16)$$

Variablen werden in Tabelle 2.2 erläutert

Die Gesamtwechselwirkungsenergie zwischen der Umgebung und einer Aminosäure berechnet sich aus der Kombination der Atomtyp abhängigen Wechselwirkungsbeiträge:

$$E^{aa} = \sum E_{at}^{aa} \quad (2.17)$$

Wird dies an jeder Stelle der Proteinsequenz mit allen zwanzig Aminosäuren durchgeführt, ergibt sich das Mutationsprofil eines Proteins. Es stellt einen energetischen Abdruck im möglichen Zustandsraum da:

$$\Delta G_{Faltung} = E_{WW} = \sum\left(n_{at,r,w}^{aa}\cdot\Delta E_{at,r,w}^{aa}\right) \quad (2.18)$$

Berechnung der Wechselwirkungsenergie E_{ww}, die der freien Faltungsenthalpie entspricht, durch Multiplikation der Häufigkeit n der Umgebungselemente mit der statistischen Netto-Präferenz $\Delta E_{at,r,\omega}^{aa}$ über alle Aminosäuren und Atomtypen, in Abhängigkeit von Abstand und Richtung.

Tabelle 2.2: Erläuterung der Parameter in den Formeln 2.15 – 2.18

Variable	Beschreibung
aa	Die zwanzig Aminosäuren
all	Durchschnittsbildung über alle Aminosäuren
at	Die gewählten Atomtypen
n	Atomhäufigkeiten
r	Abstand
ω	Richtung

Die Anwendung des Aminosäure-Atom-Potentials für die Vorhersage der Thermostabilität eines Proteins oder zur Überprüfung eines experimentellen Datensatzes ist in Abbildung 2.2 dargestellt.

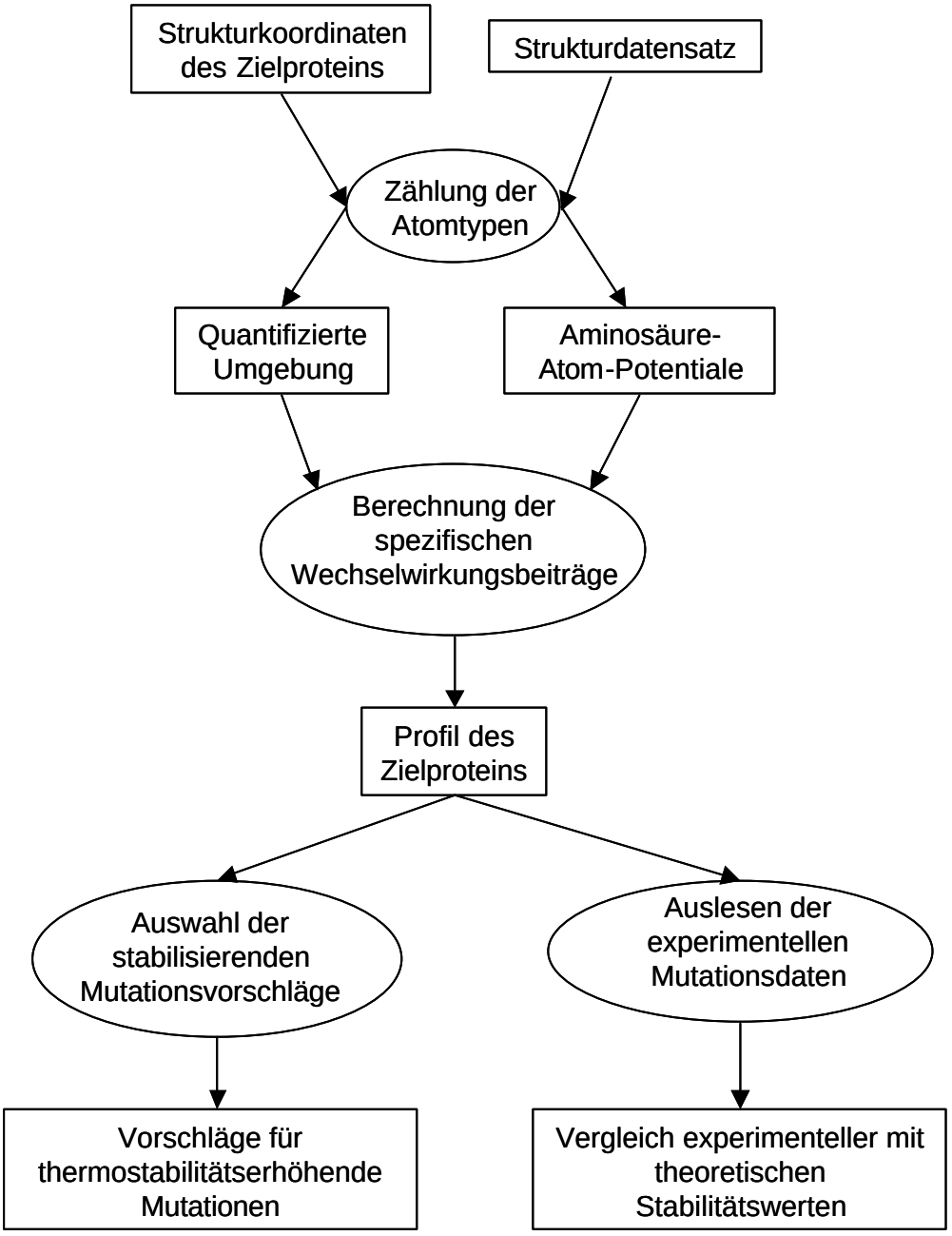


Abbildung 2.2: Schematische Darstellung der Thermostabilitätsvorhersage mit dem Aminosäure-Atom-Potential. Das zu stabilisierende Protein wird als Zielprotein bezeichnet, nach Leven [Leven, 1999]

2.2.3.1. Variation über Richtungs- und Winkelintervalle

Die Aminosäure-Atom-Wechselwirkungsenergien werden über Richtungs- und Winkelintervalle zu einem Gesamtwechselwirkungsbeitrag aufsummiert (s. Gleichung 2.16-2.18). Die Summation kann für unterschiedliche Atomtypen über kleine Abstände, große Abstände oder große Intervalle erfolgen. Günstigste Abstands- oder Winkelbereiche sowie Intervalllängen sind *a priori* nicht bekannt. Durch eine systematische Suche, mit Überprüfung möglichst vieler Kombinationen von Abstands- oder Winkelbereichen und deren Auswirkungen auf den Entwicklungsdatensatz, soll eine Bestimmung der optimalen Parameter durchgeführt werden.

Der initiale Parametersatz wurde im Hinblick auf die vorhandene Literatur und rechenzeitlimitierende Faktoren ausgewählt. So wurde der maximale Abstand zum gewählten Repräsentanten der Aminosäure auf 20 Å festgelegt. In der Literatur finden sich für nicht lokale wissensbasierte Potentiale optimale Abstände von 8 Å [Gilis und Rooman, 1997], 10 Å [Huang *et al.*, 1995; Ota *et al.*, 1995; Meeker *et al.*, 1996; Melo und Feytmans, 1997], 14 Å [Sippl *et al.*, 1996] und 15 Å [Leven, 1999]. Die Schrittlänge der Intervalle wurde auf 0,5 Å oder 1 Å gesetzt [Leven, 1999]. Öffnungswinkel sowie Schrittweite des Kegels wurden frei gewählt (s. Tabelle 2.3).

Tabelle 2.3 : Auflistung der zu optimierenden Faktoren und der gewählten Werte für die Aminosäure-Atom-Potentialfunktion

Variablen	Gewählte Werte
Schalen	3-20 Å; in Schritten von 0,5 Å bzw. 1 Å
Öffnungswinkel	30°, 45°, 60°, 90° und 180°
Schrittweite (φ,ψ)	0°, 30°, 45°, 60°, 90° und 180°

2.2.3.2. Verwendete Atomtypen

Die Umgebungsbeschreibung einer Aminosäure erfolgt auf atomarer Ebene. Um in dieser Arbeit einen Kompromiss zwischen Redundanz von Informationen,

chemischer Komplexität und statistisch relevanter Vorkommenshäufigkeit im Protein zu finden, werden, dem Aminosäure-Atom-Potential von Dengler und Leven entsprechend, fünf verschiedene Atomtypen gewählt (s. Tabelle 2.4) [Dengler, 1998; Leven, 1999]. Der ersten Klassifizierung in unpolare (Kohlenstoff) und polare Elemente (Sauerstoff und Stickstoff) folgt eine weitere Unterteilung in sp^2 und sp^3 hybridisierten Kohlenstoff. Sauerstoff wird nach Herkunft aus Protein oder Lösungsmittel getrennt behandelt. Kristallwasser wird als Lösungsmittel gewertet und wurde, anders als bei Leven [Leven, 1999], berücksichtigt.

Tabelle 2.4: Darstellung der verschiedenen Atomtypen und der verwendeten Abkürzungen

Atomtyp	Abkürzung
sp^2-hybridisierter (aromatischer) Kohlenstoff	C_{arom}
sp^3-hybridisierter (aliphatischer) Kohlenstoff	C_{ali}
Stickstoff	N
Sauerstoff aus Aminosäuren	O_{as}
Sauerstoff aus Kristallwasser	O_k
Sauerstoff aus modellierten Wasser	O_{mo}
Sauerstoff aus modellierten Wasser und Kristallwasser	O_m

Die verschiedenen Atomtypen können bei der Summierung der Aminosäure-Atom-Beiträge zu der Gesamtwechselwirkungsenergie unterschiedlich zueinander gewichtet werden (s. Gleichung 2.17). Allerdings sind die gewählten Atomtypen nicht unabhängig, sondern über implizite physikalische und chemische Abhängigkeiten miteinander verflochten. Über eine Faktorenanalyse (s. Kapitel 2.3.9) wird die Abhängigkeit der Atomtypen aufgelöst und in ein unabhängiges System überführt. Die unterschiedlichen atomaren Energiebeiträge (E_a, E_b, E_c, ..., E_i) der abhängigen Variablen (s. Gleichung 2.19) werden in voneinander unabhängige Faktoren F_i überführt, die Linearkombinationen der Ausgangsvariablen darstellen (s. Gleichung 2.20).

$$E_{ww} = E_a + E_b + E_c + ... + E_i \quad (2.19)$$

$$F_i = f_i^a E_a + f_i^b E_b + f_i^c E_c . + ... + f_i^i E_i \quad (2.20)$$

f_i entsprechen den Linearitäts-Faktoren

Die Berechnung des Wechselwirkungsbeitrages erfolgt statt über die Summierung der Einzelenergien der Atomtypen über diese Faktoren:

$$E_{WW} = \sum_i E_i \Rightarrow E_{WW}^{unab.} = \sum_i F_i \quad (2.21)$$

Für eine vorzunehmende Optimierung können diese Faktoren zueinander gewichtet werden (s. Kapitel 2.3.9).

2.2.3.3. Modellierung des Lösungsmittels

Die Wechselwirkung von Aminosäuren eines Proteins mit dem Lösungsmittel ist notwendig für die Proteinfaltung und die Stabilität der Proteinstruktur (s. Kapitel 1.2)[Kang *et al.*, 1987; Petukhov *et al.*, 1999]. Um eine Stabilitätsvorhersage treffen zu können, muss eine möglichst genaue Beschreibung des Protein-Wasser-Systems mit seinen Interaktionen erfolgen. Mit der strukturaufklärenden Methode der Röntgenkristallographie lassen sich bei gegebener atomarer Auflösung nur besonders spezielle Wassermoleküle (Fixierung durch z.B. polare Gruppen) finden. Um eine potentielle Beschreibung der Protein-Wasser-Wechselwirkung zu erreichen, wird eine theoretisch berechnete Wasserhülle um und in das Protein modelliert. Dies geschieht mit der Methode des *kontinuierlichen Lösungsmittels* [Dengler, 1998]. Es wird postuliert, dass Wassermoleküle nicht fixiert, sondern frei beweglich sind und jeder entsprechend große und freie Raum von Wassermolekülen besetzt ist [Colonna-Cesari und Sander, 1990; Gerstein und Levitt, 1998; Lazaridis und Karplus, 1999]. Als Volumen für ein Wassermolekül wurden annähernd 30 $Å^3$ gewählt [Lee und Richards, 1971; Richards, 1974; Dengler, 1998]. Das Protein wird in ein kubisches, nicht opti-

miertes Gitter mit einer Gitterkonstanten von 0,6 Å gelegt. Mit Protein oder Kristallwasser besetzte Zellen werden markiert, während die freien Zellen bei ausreichendem Platz mit Wasser aufgefüllt werden (s. Abbildung 2.3).

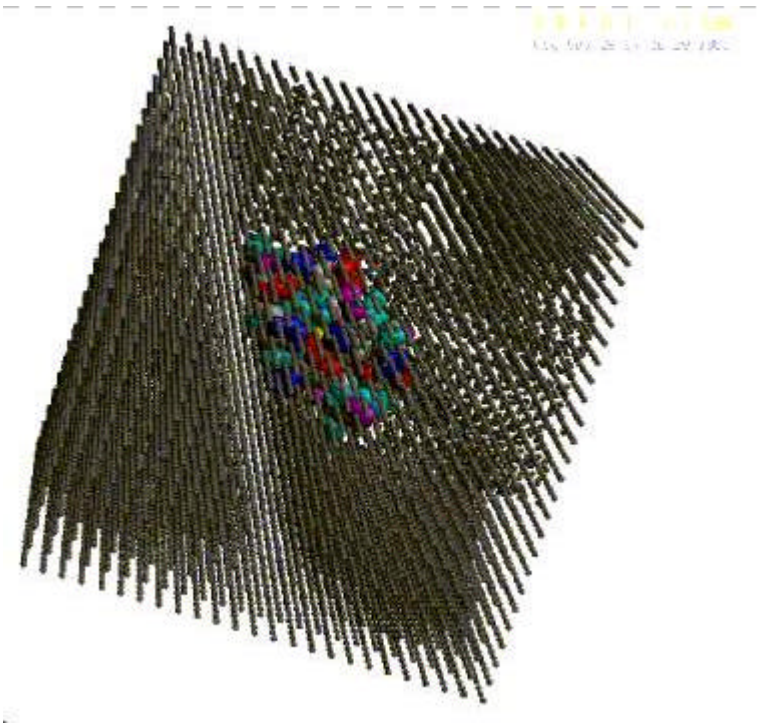


Abbildung 2.3: Das Protein 2Cl2 mit 65 Aminosäuren und 16810 Wassermolekülen (graue Kugeln) bei einer Gitterkonstanten von 0,6 Å und 2.348 * 10^6 Gitterzellen (erstellt mit BRAGI)

2.2.3.4. Probleme bei der Verwendung der Aminosäure-Repräsentanten CB und CZ

Die Wechselwirkungskräfte, die auf eine Aminosäure in einem Protein wirken, lassen sich in Kräfte lokaler und nicht lokaler Natur einteilen (s. Kapitel 1.2.4 und Kapitel 2.2.1). Nicht lokale Wechselwirkungen sind von Aminosäureseitenketten dominiert [Dill, 1990; Dill, 1999]. Um ihre Wichtigkeit für die Stabilitätsbewertung darzustellen, werden entsprechende Strukturelemente in der wissensbasierten Potentialfunktion abgeleitet. Für eine Vergleichbarkeit der Paarverteilungsfunktionen müssen allerdings die betrachteten Paare von Strukturelementen identisch sein (s. Kapitel 2.1). Bei der richtungs- und abstandsab-

hängigen Beschreibung der Umgebung werden für die Aminosäure drei Repräsentanten festgelegt (s. Kapitel 2.1.1.1). Allerdings besitzt Glycin kein CB, während bei Alanin die Position von CB mit CZ, dem geometrischen Mittelpunkt der Seitenkette, zusammenfällt. Für Glycin wurde eine virtuelle Position für ein CB-Atom bestimmt. Bei einer Festlegung von CB oder CZ als P_0, oder P_1 gilt für beide Aminosäuren: P_0 = CB und P_1=CA.

2.2.3.5. Modellierung eines virtuellen CB-Atoms für Glycin

Um Glycin-Mutationen entsprechend bewerten zu können, wird eine virtuelle Atomposition für das CB-Atom theoretisch berechnet. Seine Position wird durch zwei Winkel (θ (**N,CB**) und θ (**C**,**CB**)) und den Abstand CA-Atom zu CB-Atom bestimmt. Diese Werte werden aus den gemittelten CB-Positionen der anderen Aminosäuren bestimmt. Als Vorlage wurde das Protein mit der PDB - Bezeichnung 1IGN nach Schneckner gewählt (s. Tabelle 2.5) [Schneckner, 1998].
Der Vektor **c** zwischen CA-Atom und CB-Atom ergibt sich aus den Vektoren **a,** zwischen CA-Atom und dem N-Atom der Aminogruppe, und Vektor **b,** zwischen CA-Atom und dem Kohlenstoff der Carboxylgruppe C (s. Tabelle 2.1).

Tabelle 2.5: Verwendete Winkel und Vektorlängen zur Berechnung der Glycin CB- Koordinaten

θ (**N,CB**)	110,98 °
θ (**C**,**CB)**	110,63 °
Länge CAC	1,53 Å

Die Berechnung des Vektors **c** liefert zwei Lösungen. Die gesuchte Lösung entspricht einem Rechtssystem der Vektoren **a**, **b**, **c** (s. Kapitel 2.3.4), da sich in biologischen Systemen ausschließlich L-Aminosäuren finden lassen. Der erhaltene Vektor **c** wird auf die Länge von 1,53 Å normalisiert.

2.2.4. Das Torsionswinkelpotential

Die Hauptkette eines Peptids ist auf beiden Seiten der starren Peptidbindung drehbar [Edison, 2001]. Der Rotationswinkel um die Bindung zwischen N und CA wird als phi (φ) bezeichnet, derjenige um CA und das Carbonylkohlenstoffatom C als psi (ψ). Sind alle Werte von φ und ψ aus einer Struktur bekannt, ist die Konformation der Hauptkette vollständig definiert und die Proteinfaltung beschrieben. Ein Aminosäurerest in der Polypeptidkette kann nicht jedes beliebige φψ-Wertepaar annehmen. Der konformationelle Freiheitsgrad des Peptidrückrades ist aufgrund von sterischen Hinderungen eingeschränkt.

Die Torsionswinkel φ und ψ resultieren aus den lokalen Wechselwirkungen einer Aminosäure mit ihrem direkten sequentiellen Nachbarn und erlauben so Rückschlüsse auf diese Wechselwirkungen. Jede Aminosäure hat bevorzugte Torsionswinkelpaare und weist ein spezifisches Verteilungsmuster des φψ-Wertepaares auf. Daraus lässt sich ein wissensbasiertes Potential ableiten [Niefind und Schomburg, 1991; Dengler, 1998; Leven, 1999].

Die Darstellung der φψ-Wertepaarverteilung erfolgt in einem Stereokonturdiagramm, dem sogenannten *Ramachandran* Diagramm. In einem solchen Diagramm werden die φψ-Winkelpaare in 1° Schritte aufgeteilt und es werden 360^2 = 129600 Felder erhalten. Um eine stetige Verteilung aus den diskreten φψ-Werten zu erhalten, werden die φψ-Wertepaare durch Normalverteilungen ersetzt.

Die einem repräsentativen Strukturdatensatz entnommenen und normierten Verteilungshäufigkeiten $n_{(\varphi,\psi)}$ werden in die inverse Boltzmanngleichung eingesetzt. Daraus ergibt sich der Energiewert $\Delta E^{aa}_{(\varphi,\psi)}$ für die betrachtete Aminosäure mit gegebenem Winkelpaar (φ,ψ).:

$$\Delta E^{aa}_{(\varphi,\psi)} = -kT \ln\left(\frac{n^{aa}_{(\varphi,\psi)}}{n^{all}_{(\varphi,\psi)}}\right) \qquad (2.22)$$

Die Energiewerte für alle φψ-Winkelpaarkombinationen sind vorberechnet, von Leven [Leven, 1999] übernommen und werden bei der Vorhersage einer Tabelle entnommen.

2.2.5. Kombination der beiden wissensbasierten Potentialfunktionen

Das Aminosäure-Atom-Potential und das Torsionswinkelpotential beschreiben unterschiedliche Wechselwirkungsbeiträge einer Umgebung auf die betrachtete Aminosäure. Es wird untersucht, ob eine Kombination der beiden Potentiale zu einer besseren Gesamtbeschreibung der Umgebungswechselwirkung führen kann. Wenn beide Potentiale unabhängig voneinander sind, können ihre Energiewerte mit unterschiedlicher Gewichtung bewertet werden. Unter dieser Annahme gilt:

$$E_{WW}^{komb} = a \cdot E_{ww}^{aap} + b \cdot E_{WW}^{tp} \qquad (2.23)$$

Kombination des richtungs- und abstandsabhängigen Aminosäure-Atom-Potentials (aap) mit dem Torsionswinkelpotential (tp) zu einem Gesamtpotential E_{WW}^{komb}. Die Gewichtung erfolgt über die Variablen *a* und *b*, wobei *a*+*b*=1; *a*=0.1,0.2,0.3, ...,0.9,1

2.2.6. Verwendete experimentelle Datensätze zur Ableitung, Entwicklung und Prüfung der wissensbasierten Bewertungsfunktion

Wissensbasierte Potentiale und insbesondere Vorhersagemethoden sind auf experimentelle Daten angewiesen. Die Ableitung der wissensbasierten Potentialfunktionen erfolgt über einen der Problemstellung angepassten Strukturdatensatz, der im Idealfall alle zu beschreibenden Kräfte implizit enthält.
Eine Qualitätsüberprüfung sowie die Entwicklung einer Vorhersagemethode wird an experimentellen Daten durchgeführt. Dies geschieht mit Hilfe eines Entwicklungsdatensatzes und eines abschließenden Testdatensatzes. Alle diese Datensätze unterliegen zu definierende Kriterien, die bei der Auswertung der

Vorhersageergebnisse und bei der Vorhersagegüte berücksichtigt werden müssen. Diese Kriterien werden bei der Auswahl des Strukturdatensatzes weitgehend Ausschlusskriterien sein, während die Entwicklungs- und Testdatensätze nur im Hinblick auf ihre Qualität charakterisiert werden können [Leven, 1999].

2.2.6.1. Strukturdatensatz

Nach der thermodynamischen Hypothese befinden sich native Proteine im globalen energetischen Minimum. Die Kräfte, die diesen Zustand stabilisieren, können extrahiert und mit wissensbasierten Potentialen beschrieben werden. In dieser vorliegenden Arbeit zur Thermostabilitätsvorhersage werden ausschließlich Aminosäure-Mutationen betrachtet. Mit den gewählten wissensbasierten Potentialfunktionen soll der Einfluss einer Aminosäure auf ihre definierte Umgebung bewertet und so genau wie möglich beschrieben werden. Die Ableitung der Kräfte soll nur aus Proteinstrukturen erfolgen, an denen Aminosäure-Aminosäure-Interaktionen studiert werden können. Schwermetallionen sowie prosthetische Gruppen mit ausgedehnten aromatischen Systemen oder Eisen-Schwefel-Cluster üben eine nur schwer zu beschreibende Wirkung auf die Struktur eines Proteins aus. Die Beschreibung und richtige Erfassung ihrer Kräfte stellen für wissensbasierte Potentiale ein besonderes Problem dar [Lazaridis und Karplus, 2000]. Somit werden Strukturen mit solchen Gruppen hier ausgeschlossen. Der Strukturdatensatz soll möglichst groß bei geringer struktureller Redundanz sein. Die verwendeten Proteinstrukturen sollen durch Röntgenstrukturanalyse aufgeklärt sein. Diese bietet eine bessere Strukturauflösung als NMR und liefert eine genauere Abbildung der Struktur.
Die Kriterien für den Strukturdatensatz können wie folgt zusammengefasst werden [Leven, 1999]:

- Verwendung von ausschließlich mit Röntgenstrukturspektroskopie aufgelösten Strukturen, die eine Auflösung von $\leq$ 2,5 Å aufweisen

- Die Proteinstrukturen sollten vollständig aufgeklärt sein

- Der Strukturdatensatz sollte möglichst umfangreich sein und eine geringe strukturelle Redundanz aufweisen, dadurch wird eine Verbesserung des Signal-Rausch-Verhältnisses erreicht

- Die Strukturen dürfen keine Kofaktoren, prosthetische Gruppen oder Schwermetallionen enthalten. Die von diesen ausgehenden Effekte führen bei der Ableitung mit wissensbasierten Potentialen zu falschen Beschreibungen und zu einer Verschlechterung des Signal-Rausch-Verhältnisses

- Membran-, Virus-Coating- oder DNA- bindende Proteine sollten ebenfalls ausgeschlossen werden. Ihre spezifischen Charakteristika würden bei einer Anwendung auf globuläre Proteine zu falschen Potentialbeschreibungen führen

Ein den oben aufgeführten Kriterien entsprechender Datensatz ist im Anhang wiedergegeben (s. Anhang 6.1).

2.2.6.2. Entwicklungs- und Testdatensatz

Wenn die verwendeten Datensätze zur Entwicklung und Überprüfung der wissensbasierten Bewertungsfunktion möglichst groß, umfangreich und fehlerfrei sind, lassen sich daraus zuverlässige Aussagen über die Qualität der Vorhersage machen.
Die erfassten experimentellen Daten für eine Thermostabilitätsmessung sind thermodynamische Daten wie Wärmekapazität, Enthalpie, freie Faltungsenthalpie oder Schmelztemperaturen. Diese werden jeweils vom Wildtyp und der Mutante gemessen und miteinander in Bezug gesetzt (s. Kapitel 1.2.6). Im Idealfall sind die experimentell festgestellten Daten homogen, d.h. mit bekannten und abschätzbaren Fehlern behaftet, vergleichbar und reproduzierbar. Unter diesen Bedingungen sollte ein richtiger Vorhersagealgorithmus die Vorgabe des

Überprüfungsdatensatz präzise erreichen. Bei schlechter Datenqualität kann selbst eine richtige Methode keine guten Ergebnisse liefern.

In nur wenigen experimentellen Arbeiten werden Fehler und Vertrauensgrenzen angegeben, bei vielseitigen Möglichkeiten von Fehlerquellen. Die durchgeführten Experimente unterscheiden sich in Zielsetzung und Bedingungen, wie z.B. pH-Wert, Salzkonzentration oder Proteinkonzentration. Alle diese Faktoren haben Auswirkungen auf die Stabilität und sind nur in wenigen Fällen in den Arbeiten angegeben. Darüber hinaus sind die verschiedenen Messmethoden, optische oder kalorimetische Verfahren (s. Kapitel 1.2.6.1), mit unterschiedlichen Fehlern behaftet, die zusätzlich von der Arbeitsgruppe oder sogar einzelnen Personen, welche die Messungen vorgenommen haben, abhängen. Manche experimentelle Daten werden mit geschätzten, fehlerbehafteten theoretischen Daten errechnet.

In der Realität ist die Anzahl an Thermostabilitätsdaten für Einzelmutationen bei thermodynamischer Reversibilität jedoch gering, so dass die Qualitätsanforderungen diesem Umstand Rechnung tragen müssen. Für experimentelle Thermostabilitätsdaten gibt es, anders als bei aufgeklärten Proteinstrukturen mit den Angaben der Auflösung, des R-Faktors oder von Ramachandran Diagrammen, keine Qualitätsmerkmale. Dadurch existiert kein aussagekräftiges Maß für die Güte der Daten. Diese sind sehr inhomogen und mit unterschiedlichen und schwer zu erfassenden Fehlern belegt. Der Einschätzung von Qualität und Fehlern kommt demnach eine besondere Bedeutung zu.

In dieser Arbeit wurde der Gesamtdatensatz für die experimentellen Mutationsdaten zu Teilen einer Internetdatenbank, der ProTherm-Datenbank (s. Kapitel 1.3.3), sowie einigen schon in der Literatur angegebenen Datensätzen entnommen. Damit ist ein Vergleich mit den in der Literatur durchgeführten Stabilitätsvorhersagen möglich. Neben Zahlenwerten, die aus systematischen Austauschen stammen, fließen zahlreiche Einzelexperimente mit weniger generellen Ansätzen, diese Arbeiten behandeln meistens Modellproteine zur Faltungsuntersuchung, in den Gesamtdatensatz ein. Dieser wurde dann nach Redundanz und Plausibilitätstests in zwei Datensätze aufgeteilt [Leven, 1999]: Den Entwicklungsdatensatz, an dem die wissensbasierte Bewertungsfunktion trai-

niert und optimiert werden soll, und einen Testdatensatz für die abschließende Bewertung.

2.2.7. Kriterien für die Bewertung der Stabilitätsvorhersage durch das wissensbasierte Potential

Die Entwicklung und Optimierung der wissensbasierten Bewertungsfunktion für die Thermostabilitätsvorhersage durch Einzelmutationen erfolgt an dem Entwicklungsdatensatz. Die Überprüfung der Methode geschieht mit dem Testdatensatz. Beide Datensätze enthalten experimentelle Daten für die Thermostabilität einer Mutante im Vergleich zu der des Wildtyps. Theoretisch berechnete Stabilitätswerte werden mit den experimentellen Werten verglichen. Der durch die wissensbasierte Potentialfunktion errechnete Wert soll der freien Faltungsenthalpie entsprechen (s. Gleichung 2.18). Die Differenz der theoretisch bestimmten freien Faltungsenthalpien $\Delta G_{Faltung}$ für den Wildtyp und die Mutante entspricht den experimentell bestimmbaren freien Faltungsenthalpien (s. Kapitel 1.2.6). Diese Energiewerte können miteinander verglichen werden. Bei exakter Vorhersage sollte ein linearer Zusammenhang zwischen den theoretischen und experimentellen Faltungsenthalpien bestehen, der durch die Bestimmung des Korrelationskoeffizienten r (s. Kapitel 2.3.8) überprüft werden kann. Die Auswertung der Vorhersage kann auch quantitativ erfolgen, indem die errechneten und die experimentellen Energiewerte einzeln miteinander verglichen und in vier Vorhersageklassen eingeteilt werden. Die Klasse *richtig positive*, eine im Experiment stabilisierende Mutation wird durch die Bewertungsfunktion als stabilisierend beurteilt, die *falsch positive*, eine im Experiment destabilisierende Mutation wird durch die Bewertungsfunktion als stabilisierend beurteilt, die *richtig negative*, eine im Experiment destabilisierende Mutation wird durch die Bewertungsfunktion als destabilisierend beurteilt und die *falsch negative*, eine im Experiment stabilisierende Mutation wird durch die Bewertungsfunktion als destabilisierend beurteilt. Eine *richtige Vorhersage* rv umfasst die richtig positiven und die richtig negativen Klassifizierungen.

Die wissensbasierte Bewertungsfunktion stellt einen Vorfilter dar, mit der mögliche Ziele in einem *Protein Engineering* Experiment für eine thermostabilisierende Mutation identifiziert und quantifiziert werden sollen. Damit kommt der Klassifizierung in eine *falsch negative* eine schwerwiegendere Bedeutung zu als eine *falsch positive* Vorhersage. Während die *falsch positive* Vorhersage einen (vermeidbaren) Mehraufwand bedeutet, hat eine *falsch negative* Vorhersage zur Konsequenz, dass eine eigentlich stabilisierende Mutation nicht erkannt wird und das Experiment überflüssig macht. Daher soll eine Bewertungsfunktion, die als Vorfilter dient, im Idealfall keine *falsch negativen* Vorhersagen treffen.
Als Maß für Qualität und Güte der Bewertungsfunktion dienen die zwei Faktoren Sensibilität und Spezifität [Sachs, 1997].

2.2.7.1. Sensibilität

Die Sensibilität Sens stellt ein Maß für das Verhältnis *richtig positiv* zu *falsch negativ* dar und ist das entscheidende Kriterium für eine Funktion, die als Vorfilter dienen soll. Im Idealfall, wenn keine *falsch negative* Bewertung vorgenommen worden ist, liegt der Wert für die Sensibilität bei 1. Es gilt:

$$\mathrm{Sens} = \frac{\mathrm{n}_{richtig\,positiv}}{(\mathrm{n}_{richtig\,positiv} + \mathrm{n}_{falsch\,negativ})} \qquad (2.24)$$

n ist die Anzahl der entsprechend getroffenen Vorhersagen

2.2.7.2. Spezifität

Die Spezifität Spez stellt ein Maß für das Verhältnis *richtig positiv* zu *falsch positiv* dar. Im Idealfall, wenn keine *falsch positive* Bewertung vorgenommen worden ist, liegt der Wert für die Spezifität bei 1. Es gilt:

$$\mathrm{Spez} = \frac{\mathrm{n}_{richtig\,positiv}}{(\mathrm{n}_{richtig\,positiv} + \mathrm{n}_{falsch\,positiv})} \qquad (2.25)$$

2.2.8. Verwendete Programme und programmiertechnische Hilfsmittel

Die Berechnung der Lage von Wassermolekülen um ein und in einem Protein sowie das Zählen von Aminosäure-Atom-Kontakten wurden in der Programmiersprache C++ implementiert. Die Ableitung, Berechnung und die Bewertung der wissensbasierten Potentialfunktion wurden in der Programmiersprache Python (Version 2.1.1) (http://www.python.org/) implementiert.
Die Entwicklung erfolgte unter dem Betriebssystem Linux (Mandrake 8.1) auf einem Pentium-III-PC (600 MHz) sowie unter dem Betriebsystem Solaris 8 auf einer Sunfire 880v (Sechs UltraSparc III-Prozessoren, 750 MHz).
Folgende Programme wurden verwendet:

- Die Sekundärstrukturberechnung erfolgte mit dem Programm DSSP (*Dictionary of Secondary Structures in* Proteins) CMBI Version 1.4.2000 (http://www.cmbi.kun.nl/gv/dssp/) [Kabsch und Sander, 1983]

- Die Lösungsmittelzugänglichkeiten wurden mit dem Programm *psa* v.1.1c (*protein surface accessibility*) aus dem Programmpaket JOY berechnet (http://www-cryst.bioc.cam.ac.uk/~joy/)

- Die Speicherung und Auswertung der Vorhersageergebnisse erfolgte in der Datenbank MySQL (http://www.mysql.com/)

- GNU C/C++-Compiler Version 2.96

- Swiss-Pdbviewer Version 3.7 (http://www.expasy.org/spdbv)

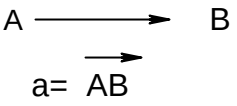
- BRAGI (http://www.uni-koeln.de/math-nat-fak/biochemie/ds/dsbrag_e.htm)

2.3. Statistische und Mathematische Methoden

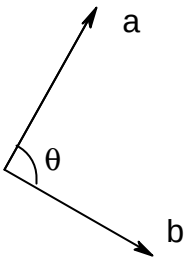
Im Folgenden sollen die benötigten mathematischen und statistischen Grundlagen bereitgestellt werden.

2.3.1. Vektorrechnung

Ein Vektor in Raum oder Ebene wird durch Richtung und Länge charakterisiert. Die Strecke A nach B entspricht dem Vektor **a**, mit der Länge $|\mathbf{a}|$.



Der eingeschlossene Winkel zwischen zwei Vektoren **a** und **b** wird mit θ bezeichnet.



2.3.2. Skalarprodukt

Das Skalarprodukt zweier Vektoren **a** und **b** ist definiert als [Rade und Westergren, 1997]:

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| \cdot |\mathbf{b}| \cdot \cos \Theta = a_x b_x + a_y b_y + a_z b_z \qquad (2.26)$$

Das Ergebnis ist ein Skalar.

2.3.3. Vektorprodukt

Das Vektorprodukt **c** = **a** x **b** der zwei Vektoren **a** und **b** ist definiert durch:

$$\mathbf{c} = |\mathbf{a} \times \mathbf{b}| = |\mathbf{a}| \cdot |\mathbf{b}| \cdot \sin \Theta \qquad (2.27)$$

Das Ergebnis ist ein Vektor **c**, der orthogonal zu **a** und **b** steht und mit diesen ein Rechtssystem bildet.

2.3.4. Spatprodukt

Für das Spatprodukt [**a**, **b**, **c**] der Vektoren **a**, **b** und **c** gilt:

$$[\mathbf{a},\mathbf{b},\mathbf{c}] = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) \qquad (2.28)$$

Die Vektoren bilden ein Rechtssystem, wenn [**a**, **b**, **c**] >0 ist.

2.3.5. Umwandlung von kartesischen Koordinaten in Polarkoordinaten

Die Lage eine Punktes P im Raum kann durch die kartesischen Koordinaten x, y und z oder durch den Abstand r vom Ursprung, dem Winkel ψ (psi) zwischen Ortsvektor und z-Achse und dem Winkel φ (phi) beschrieben werden (s. Abbildung 2.4).

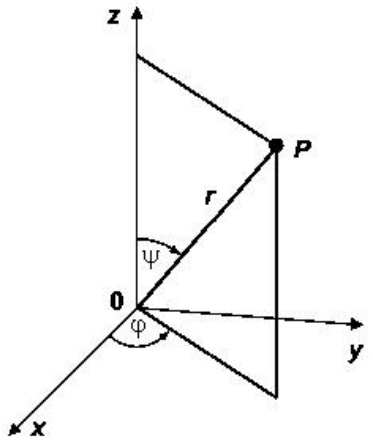


Abbildung 2.4: Darstellung der räumlichen Polarkoordinaten

Für die Beziehungen dieser Parameter untereinander ergibt sich:

$$\begin{aligned} x &= r \sin y \cos j \\ y &= r \sin y \sin j \\ z &= r \cos y \end{aligned} \qquad (2.29)$$

bzw.

$$\begin{aligned} r &= \sqrt{x^2 + y^2 + z^2} \\ j &= \arctan \frac{y}{x} \\ y &= \arctan \sqrt{\frac{x^2 + y^2}{z}} \end{aligned} \qquad (2.30)$$

r, ψ und φ werden als die Kugelkoordinaten oder Polarkoordinaten des Punktes P bezeichnet. Für den Definitionsbereich gilt:

$r \geq 0$,

$0 \leq \psi \leq \pi$

und $0 \leq \varphi \leq 2\pi$.

2.3.6. Kugelvolumen

Das Kugelvolumen berechnet sich nach:

$$V = \frac{4}{3} \cdot \pi \cdot r^3 \qquad (2.31)$$

2.3.7. Z-Transformation

Variabeln, die sich um mehrere Größenordungen oder in Einheiten unterscheiden, können mit der Z-Transformation in Beziehung gesetzt werden [auf der Heyde, 1990]. Nach Gleichung 2.32 ist der sogenannte *Z-Score* z dimensionslos, während die Merkmalsverteilung d_i dimensionsbehaftet ist. Es

wird eine standardisierte Normalverteilung mit dem Mittelwert null und der Standardabweichung eins erhalten.

$$z = \frac{d_i - \overline{d}}{\sigma} \tag{2.32}$$

$$s^2 = \frac{1}{m-1} \cdot \sum_{i=1}^{m} (d_i - \overline{d})^2 \tag{2.33}$$

$$\overline{d} = \frac{1}{m} \cdot \sum_{i}^{m} d_i \tag{2.34}$$

$$\sigma = \sqrt{s^2} \tag{2.35}$$

d_i ist der i'te Wert von d, $\overline{d}$ entspricht dem arithmetischen Mittelwert von d_i, σ ist die Standardabweichung, s^2 die Varianz und m die Anzahl von d

2.3.8. Korrelationsanalyse

Für zwei Paarverteilungen X,Y kann der lineare Zusammenhang durch den Korrelationskoeffizienten r beschrieben werden. Die Abstände zwischen den Beobachtungen zweier Merkmale und deren arithmetischem Mittel werden zueinander in Beziehung gesetzt [Sachs, 1997]:

$$r_{cor} = \frac{\sum \left(x_i - \overline{x}\right) \cdot \left(y_i - \overline{y}\right)}{\sqrt{\sum \left(x_i - \overline{x}\right)^2 \cdot \sum \left(y_i - \overline{y}\right)^2}} \tag{2.36}$$

Der Korrelationskoeffizient ist ein dimensionsloses Maß und behandelt beide Merkmale symmetrisch, es gilt:

$$r_{cor}(X,Y) = r_{cor}(Y,X) \tag{2.37}$$

Bei mehr als zwei Merkmalen errechnet sich die Korrelationsmatrix aus der Kovarianzmatrix der Z-Scores (s. Kapitel 2.3.7) [auf der Heyde, 1990].
Ein exakt linearer Zusammenhang liegt vor, wenn der Korrelationskoeffizient die Werte +1 oder –1 annimmt, bei $r_{cor}=0$ sind die Merkmale linear unabhängig voneinander. Liegt r_{cor} im Wertebereich von –1 und 0 sowie 0 und +1, wird ein stochastischer Zusammenhang angenommen, der durch Überprüfung der Signifikanz bestätigt werden muss.
Ein exakt linearer Zusammenhang ergibt sich auch für zwei Verteilungen, deren Werte sich um einen konstanten positiven Faktor unterscheiden. Der Korrelationskoeffizient beschreibt nur die Ähnlichkeit der Verteilung zweier Merkmale, nicht die Ähnlichkeit der Größenordnung ihrer einzelnen Werte. Zur Beurteilung der Ähnlichkeit von Form und Größenordnung zweier Verteilungen schlagen Hodgkin und Richards [Hodgkin und Richards, 1987] für die Berechnung des Korrelationskoeffizienten Gleichung 2.38 vor:

$$r_{hr} = \frac{2\sum \left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sum \left(x_i - \bar{x}\right)^2 + \sum \left(y_i - \bar{y}\right)^2} \qquad (2.38)$$

Die Überprüfung der Signifikanz des Korrelationskoeffizienten im Wertebereich von $-1 < r < 1$, wobei $r \neq 0$, kann durch verschiedene mathematische Methoden oder auf Standardwerten basierend erfolgen. Die prozentuale Irrtumswahrscheinlichkeit, die überschritten werden muss, damit die Korrelation als signifikant angesehen werden kann, hängt von der Datenmenge ab. Eine Übersicht der Standardwerte bei gegebenen Datenpunkten für die 0,1 %, 1 % und 5 % Irrtumswahrscheinlichkeit der Mindestkorrelation wird in Tabelle 2.6 wiedergegeben [Sachs, 1997; Leven, 1999].

Tabelle 2.6: Prüfung des Korrelationskoeffizienten $|r|$ auf Signifikanz

Für jeden Freiheitsgrad (FG) wird die zu erreichende Mindestkorrelation für die jeweilige Irrtumswahrscheinlichkeit angegeben. Der Freiheitsgrad entspricht der Anzahl erfasster Datenpunkte n minus zwei: FG=n-2 [Sachs, 1997].

	Irrtumswahrscheinlichkeit		
FG	**0,1 [%]**	**1 [%]**	**5 [%]**
1	0.9999	0.9998	0,9969
2	0,9990	0,9900	0,9500
4	0,974	0,917	0,811
8	0,872	0,765	0,632
10	0,823	0,708	0,576
15	0,725	0,606	0,482
20	0,652	0,537	0,423
25	0,597	0,487	0,381
30	0,554	0,449	0,349
40	0,490	0,393	0,304
50	0,443	0,354	0,273
80	0,357	0,283	0,217
100	0,321	0,254	0,195
500	0,146	0,115	0,0875
1000	0,104	0,0813	0,0619
2000	0,0734	0,0575	0,0438

2.3.9. Faktorenanalyse

Viele physikalisch-chemische Größen sind voneinander abhängig. Diese Abhängigkeit kann sich unter anderem durch das Wechselwirken von Atomen über lokale oder nicht lokale Kräfte über diskrete Schalen und Kegel hinweg ergeben (s. Kapitel 2.2.2). Auch die Tatsache, dass ein definierter freier Raum nur durch ein Atom belegt werden kann, trägt zur Abhängigkeit der betrachteten Variablen *Atomtypen* bei. Die Koordinatenachsen des von ihnen aufgespannten Merk-

malsraums sind nicht orthogonal zueinander. Bei Verschiebung einer Variablen entlang einer Achse, eines Merkmals werden auch die Koordinaten auf den anderen Achsen verändert.

Um die Abhängigkeit von Variablen in einem hochdimensionalen Raum interpretieren und verstehen zu können, erfolgt eine Analyse der Datenmuster durch eine lineare Transformation der entsprechenden Variablen in ein unabhängiges System. Die Hauptkomponentenanalyse (HKA) [auf der Heyde, 1990] überführt also durch eine Koordinatentransformation Variablen von einem abhängigen in einen unabhängigen Merkmalsraum. Die Achsen in dem neuen Parameterraum werden in Richtung größtmöglicher Varianz gelegt. Im *n*-dimensionalen Fall werden *n* orthogonale Faktoren erhalten, die Linearkombinationen der *n* Ausgangsvariablen darstellen. Eine Verkleinerung der Dimensionalität kann erreicht werden, wenn der Anteil der Eigenwerte der betrachteten Variablen an der Gesamtvarianz bei weniger als *n*-Achsen für ausreichend befunden wird.

Das Verfahren beruht auf dem Eigenwertproblem, das für die Kovarianzmatrix C, aus den zu untersuchenden Variablen bestehend, gelöst wird. Aus der Matrix C errechnen sich die Eigenwerte λ und die Eigenvektoren e, es gilt [auf der Heyde, 1990]:

$$(\mathrm{C} - \lambda \mathrm{I}) \cdot e = 0 \tag{2.39}$$

Wird die Matrix der Eigenvektoren E mit der Wurzel ihrer Eigenwerte $\Lambda^{1/2}$ multipliziert, ergeben sich die sogenannten Faktorladungen oder Hauptkomponenten.

$$\mathrm{F} = \mathrm{E} \cdot \Lambda^{1/2} \tag{2.40}$$

$$\mathrm{C} = \mathrm{F} \cdot \mathrm{F}^{\mathrm{T}} \tag{2.41}$$

Diese entsprechen den Multiplikatoren in Gleichung 2.20, mit denen die Energiewerte in Faktoren umgerechnet werden.

3. Ergebnisse

3.1. Experimentelle Datensätze

Die Ableitung der wissensbasierten Bewertungsfunktion zur Vorhersage der Thermostabilität von Proteinen geschieht über einen Strukturdatensatz. Die Validierung und Optimierung der Bewertungsfunktion erfolgt über einen Entwicklungs- und einen Testdatensatz. Diese drei experimentellen Datensätze werden im Folgenden näher betrachtet und erläutert.

3.1.1. Strukturdatensatz

Der Strukturdatensatz bildet die Grundlage für die wissensbasierte Bewertungsfunktion. Aus gewählten repräsentativen Proteinstrukturen werden die zu beschreibenden Kräfte abgeleitet und mit einem wissensbasierten Potential beschrieben (s. Kapitel 2.1). Bei der Selektion von Strukturen aus der Gesamtmenge von zur Verfügung stehenden Proteinstrukturen wurden die Kriterien aus Kapitel 2.2.6.1 angewandt. Als Grundlage diente die Proteindatenbank [Bernstein *et al.*, 1978], die zum Zeitpunkt der Auswahl des Strukturdatensatzes durch Leven [Leven, 1999] im Januar 1998 ca. 8500 Struktureinträge aufwies. Aus dieser Gesamtmenge wurden 284 Strukturen extrahiert, die im Anhang 6.1 aufgeführt werden.

3.1.2. Entwicklungsdatensatz

Die Daten für den Entwicklungsdatensatz wurden der Literatur entnommen. Da die Suche nach einzelnen Mutationsdaten sehr aufwendig ist, wurde auf die bereits von Gilis, Topham und Leven [Gilis und Rooman, 1996; Gilis und Rooman, 1997; Topham *et al.*, 1997; Leven, 1999] zusammengestellten experimentellen Datensätze zurückgegriffen. Zudem ermöglicht dies einen Vergleich zwischen den in der Literatur von Gilis, Topham und Leven beschriebenen Methoden zur Stabilitätsvorhersage und der in dieser Arbeit vorgestellten Methode. Nach dem Entfernen von redundanten Daten oder Mehrfachmutationen umfasste der Datensatz elf Proteine mit 646 Angaben zur Änderung der freien Faltungsenthal-

pie. Bis auf einige wenige Ausnahmen finden sich weder bei Gilis, Topham und Leven noch in den von ihnen angegebenen Literaturstellen die genauen experimentellen Bedingungen. Beschreibungen der Datensätze finden sich in Tabelle 3.1 und Tabelle 3.2.

Die ausgewählten Proteine liegen in der Proteinstrukturdatenbank teilweise in mehreren aufgeklärten Strukturen vor. Diese Strukturen unterscheiden sich in den experimentellen Bedingungen, unter denen sie gelöst wurden. So wurden teilweise Mutationen eingefügt (1l63 unterscheidet sich durch die zwei Mutationen C54T und C97A von dem Wildtyp 3lzm), die Strukturen verfeinert (2lzm und 3lzm) oder mit Inhibitoren, Antikörpern etc. kokristallisiert (1bgs ist ein Barstar Komplex, 1rnb eine Nuclease mit Inhibitor).

Tabelle 3.1: Beschreibung des Entwicklungsdatensatzes, mit Anzahl der jeweiligen Mutationsdaten der freien Faltungsenthalpie

Enzym	PDB	E.C.	n_{DDG}	n_{AS}
Barnase	1bgs	3.1.27.-	60	597
t4 Lysozyme	1l63	3.2.1.17	94	162
t4 Lysozyme	1lyd	3.-	78	164
t4 Lysozyme	1lz1	3.2.1.17	5	130
Barnase	1rnb	3.1.27.-	118	111
Staphylococcal nuclease	1stn	3.1.31.1	83	136
Chymotrypsin Inhibitor	2ci2		79	65
t4 Lysozyme	2lzm	3.2.1.17	16	164
Tryptophan Synthase	2wsy	4.2.1.20	18	612
t4 Lysozyme	3lzm	3.2.1.17	59	164
Hen egg white lysozyme	4lyz	3.2.1.17	36	129
		Summe:	646	

$n_{\Delta\Delta G}$ entspricht der Anzahl der experimentellen Stabilitätsdaten, n_{AS} entspricht der Anzahl der Aminosäuren im Protein, E.C. bezieht sich auf die NC-IUBMB *Enzym Nomenclature*, PDB bezeichnet den PDB-Code, die Enzymbezeichnungen wurden der Originalliteratur entnommen.

Tabelle 3.2: Aufschlüsselung des Entwicklungsdatensatzes in die Anzahl stabilisierender und destabilisierender Mutationsdaten

Protein	n_{DDG}	$n_{destabilisierend}$	$n_{stabilisierend}$
1bgs	60	56	4
1l63	94	66	28
1lyd	78	37	41
1lz1	5	5	0
1rnb	118	112	6
1stn	83	81	2
2ci2	79	72	7
2lzm	16	16	0
2wsy	18	0	18
3lzm	59	35	24
4lyz	36	19	17
Summe:	646	499	147

3.1.3. Testdatensatz

Der Testdatensatz wurde im Februar 2002 aus der ProTherm-Datenbank [Gromiha *et al.*, 2000] extrahiert. Ein Teil dieser Daten war redundant zum Entwicklungsdatensatz und wurde entfernt. Es blieben 918 Angaben aus 27 verschiedenen Proteinen zur Änderung der freien Faltungsenthalpie übrig. Der Testdatensatz ist in Tabelle 3.3 und Tabelle 3.4 dargestellt.

Tabelle 3.3: Darstellung des Testdatensatzes mit der jeweiligen Anzahl von Mutationsdaten zur thermischen Stabilitätsänderung

Enzym	PDB	E.C.	n_{DDG}	n_{AS}
Manganese superoxide dismutase	1abm	1.15.1.1	1	198
Adenylate kinase (adk)	1ank	2.7.4.3	4	214
Barnase	1bni	3.1.27.-	21	108
Bovine pancreatic trypsin inhibitor	1bpi		51	58
Major cold shock protein	1csp		4	67
Cytochrome b5	1cyo		3	88
Catabolite gene activator protein	3gap		2	401
Human lysozyme	1lhm	3.2.1.17	14	130
Human lysozyme	1lz1	3.2.1.17	33	130
Myoglobin	1mbn		38	153
Arc repressor	1myl		3	45
Ribonuclease t1	1rn1	3.1.27.3	16	103
Cole1 repressor of primer	1rop		21	56
Ribonuclease a	1rtb	3.1.27.3	11	124
Ribonuclease sa	1sar	3.1.4.8	3	96
Staphylococcal nuclease	1stn	3.1.31.1	12	136
Subtilisin	1sup	3.4.21.62	6	275
Tailspike-protein	1tyu		7	543
Cytochrome c	1ycc		13	103
Chymotrypsin inhibitor 2	2ci2		12	65
t4 Lysozyme	2lzm	3.2.1.17	405	164
Ribonuclease h	2rn2	3.1.26.4	122	155
Thioredoxin	2trx		2	108
Tryptophan Synthase	2wsy	4.2.1.20	16	612
Serine protease inhibitor	3ssi		50	108
Dihydrofolate reductase	1dyj	1.5.1.3	10	159
Hen egg white lysozyme	4lyz	3.2.1.17	38	129
	Summe:		918	

Tabelle 3.4: Aufschlüsselung des Testdatensatzes in die Anzahl stabilisierender und destabilisierender Mutationsdaten

PDB	n_{DDG}	$n_{destabilisierend}$	$n_{stabilisierend}$
1abm	1	1	0
1ank	4	3	1
1bni	21	20	1
1bpi	51	49	2
1csp	4	3	1
1cyo	3	2	1
3gap	2	1	1
1lhm	14	13	1
1lz1	33	29	4
1mbn	38	31	7
1myl	3	0	3
1rn1	16	13	3
1rop	21	9	12
1rtb	11	11	0
1sar	3	3	0
1stn	12	6	6
1sup	6	0	6
1tyu	7	7	0
1ycc	13	9	4
2ci2	12	11	1
2lzm	405	284	121
2rn2	122	37	85
2trx	2	2	0
2wsy	16	14	2
3ssi	50	36	14
1dyj	10	10	0
4lyz	38	21	17
Summe:	918	625	293

3.2. Die wissensbasierte Bewertungsfunktion für die Vorhersage der Thermostabilität

Es soll eine wissensbasierte Bewertungsfunktion für die Vorhersage der Thermostabilität gefunden werden. Für das Aminosäure-Atom-Potential können, neben einer unterschiedlichen Beschreibung der Umgebung (z.B. einer atomaren Betrachtung) und dem methodischen Aufbau des Potentials, mehrere Parameter gewählt und optimiert werden. Danach soll das Aminosäure-Atom-Potential mit einem von Leven [Leven, 1999] entwickelten Torsionswinkelpotential kombiniert werden. Die Entwicklung der Potentiale erfolgt durch eine Überprüfung der geleisteten Vorhersage an dem Entwicklungsdatensatz (s. Kapitel 3.1).

3.2.1. Das wissensbasierte Aminosäure-Atom-Potential

3.2.1.1. Anwendung des Aminosäure-Atom-Potentials

Die Erstellung einer Stabilitätsvorhersage erfolgt durch die im Methodenteil skizzierte (s. Kapitel 2.2.3 Abbildung 2.1) Anwendung des wissensbasierten Aminosäure-Atom-Potentials. Für das zu stabilisierende Protein werden die Umgebungen jeder Aminosäure durch Auszählen der unterschiedlichen Atomtypen nach definierten Parametern bestimmt (s. Kapitel 2.1). Die erhaltenen Häufigkeitswerte werden mit den korrespondierenden Energiewerten aus den abgeleiteten Potentialkurven multipliziert. Für das Protein wird so ein Mutationsprofil für alle Aminosäuren erhalten.

Zur Überprüfung der Vorhersagequalität können die entsprechenden theoretischen Energiewerte aus dem Mutationsprofil ausgelesen und mit experimentellen Daten verglichen werden. Für ein *Protein Engineering* Experiment kann eine sortierte Ausgabe des gesamten erstellten Mutationsprofils oder nur die der stabilisierenden Mutationen erfolgen.

Die Ergebnisse für die Entwicklung und Optimierung des Aminosäure-Atom-Potentials werden im Vergleich mit einem Satz von experimentellen Mutationsdaten, dem Entwicklungsdatensatz, erhalten. Aus dem theoretischen Mutati-

onsprofil für jedes dieser Proteine werden die experimentell charakterisierten Mutationen ausgelesen und den experimentellen Stabilitätswerten gegenübergestellt. Die Auswertung erfolgt entsprechend den im Methodenteil beschriebenen Kriterien (s. Kapitel 2.2.7).

3.2.1.2. Suche nach den optimalen Parametern für die wissensbasierte Aminosäure-Atom-Potentialfunktion

Die Charakterisierung einer Aminosäure-Umgebung kann unter verschiedenen Gesichtspunkten erfolgen. Neben der Wahl der Atomtypen kann eine rein abstandsabhängige oder eine richtungs- und abstandsabhängige Betrachtung gewählt werden. Um die optimalen Parameter für das Aminosäure-Atom-Potential zur Vorhersage der Thermostabilität zu finden, werden die im Folgenden beschriebenen Untersuchungen durchgeführt. Die Vorhersagen erfolgen mit einem initialen Parametersatz:

- Es wurden die in Kapitel 2.2.3.2 aufgeführten Atomtypen verwendet
- Die atomaren Energiebeiträge wurden gleichgewichtet

3.2.1.3. Untersuchung der Abstandsabhängigkeit des Aminosäure-Atom-Potentials

Die Umgebung einer Aminosäure wird bei der abstandsabhängigen und richtungsunabhängigen Beschreibung in Kugelschalen aufgeteilt und die dort gefundenen Atome entsprechend ihrem Typus zusammengefasst. Die Lage des Kugelmittelpunktes ist zu Beginn nicht festgelegt. Für eine Beschreibung der Aminosäure-Atom-Wechselwirkungen sollte die Kugel jedoch möglichst genau die Seitenkette der zu betrachtenden Aminosäure erfassen: Es wurden, wie in Kapitel 2.1.1.1 beschrieben, drei Punkte (P_0, P_1, P_2) festgelegt, welche die Lage der zur Kugel approximierten Aminosäure definieren. In Tabelle 3.6 sind die Ergebnisse für die unterschiedlichen abstandsabhängigen Aminosäure-Atom-Potentiale aufgeführt. Die besten Ergebnisse für die Vorhersage der Thermo-

stabilität durch die jeweiligen entsprechenden Potentiale sind grau unterlegt. Die Ergebnisse für die besten Vorhersagen der angewendeten Kriterien Korrelation r_{cor}, richtige Vorhersage rv, Spezifität Spez und Sensibilität Sens im Vergleich mit allen untersuchten Potentialen sind grau unterlegt und **fett** wiedergegeben (s. Tabelle 3.6).

Die Ergebnisse dort zeigen eine deutliche Abstandsabhängigkeit (in den Spalten dargestellt), aber nur geringe Unterschiede zwischen den verschiedenen gewählten Potentialen (in den Zeilen dargestellt). Beim Vergleich der gewählten Potentiale variiert die Korrelation zwischen 0,655 und 0,672, die richtig vorhergesagten Mutationen variieren zwischen 66,41 % und 67,49 % und die Sensibilität variiert zwischen 0,64 und 0,76. Die Kugeln mit dem geometrischen Mittelpunkt CZ der Seitenkette als Zentrum schneiden am besten ab. Werden die Abstände unabhängig vom Proteinrückrat (nur Berücksichtigung der Aminosäure-Repräsentanten CB und CZ) bestimmt, verbessern sich die Ergebnisse entsprechend. Die nicht lokalen Interaktionen der Aminosäure mit ihrer Umgebung und deren Beschreibung mit dem Aminosäure-Atom-Potential geschieht über die Seitenkette.

Es zeigt sich, dass es nicht *ein bestes* Potential gibt: Die Qualität des Aminosäure-Atom-Potentials hängt von der Bewertung der angeführten Kriterien ab. Das beste Potential für eine richtige Vorhersage ist (CZ,CA,O) mit dem Intervall [3,11] oder dem Intervall [4,11], die beste Korrelation liefert (CZ,CB,O) mit dem Intervall [3,13] oder dem Intervall [4,13] und für die Sensibilität fällt die Wahl auf (CZ,CB,O) mit dem Intervall [4,12]. Diese unterschiedlichen Beurteilungen des Potentials geben die prinzipielle Leistungsfähigkeit des verwendeten Algorithmus für den Entwicklungsdatensatz wieder. Bei detaillierterer Untersuchung zeigt sich, dass diese Leistungsfähigkeit für die verschiedenen Proteine und Proteinstrukturen schwankt. In Tabelle 3.5 sind die Werte für die Korrelation und die richtig vorhergesagten Mutationen für die einzelnen Proteine des Entwicklungsdatensatzes wiedergegeben. Ein statistisch signifikanter linearer Zusammenhang zwischen theoretischen und experimentellen Daten hängt bei gegebenen Datenpunkten von der Höhe des Korrelationskoeffizienten ab (s. Kapitel 2.3.8). Bei einer geringen Anzahl von Datenpunkten muss der Korrelations-

wert entsprechend höher sein. Während die Gesamtkorrelation für die Betrachtung aller Mutationen als statistisch signifikant angenommen werden kann, gilt dies nicht mehr für kleine Mengen von Datenpunkten.

Die Anzahl der experimentellen Mutationswerte für die humanen Lysozyme 1lz1, 2lzm und 3lzm sowie der Tryptophan Synthase sind für eine Beurteilung eines linearen Zusammenhanges zwischen den Stabilitätswerten nicht ausreichend. Die verwendete experimentelle Datenmenge zur Überprüfung eines linearen Zusammenhanges mit den Werten des hier eingesetzten Algorithmus ist nicht für alle Proteine genügend groß.

Tabelle 3.5: Bestimmung der richtigen Vorhersage (rv) und des Korrelationskoeffizienten r_{cor} mit verschiedenen Potentialen für die einzelnen Proteine des Entwicklungsdatensatzes

		(CZ,CB,O) [3,11]		(CZ,CB,O) [3,13]		(CZ,CB,O) [4,12]	
Protein	**n_{DDG}**	**rv [%]**	**r_{cor}**	**rv [%]**	**r_{cor}**	**rv [%]**	**r_{cor}**
1bgs	60	63,33	**0,45**	66,67	**0,46**	66,67	**0,47**
1l63	94	65,96	**0,75**	58,51	**0,78**	56,38	**0,77**
1lyd	78	67,95	**0,45**	56,41	**0,46**	66,67	**0,46**
1lz1	5	100	0,83	100	0,82	100	0,83
1rnb	118	58,47	**0,34**	57,63	**0,37**	57,63	**0,36**
1stn	83	98,8	**0,65**	95,18	**0,64**	97,59	**0,64**
2ci2	79	62,03	**0,54**	69,62	**0,52**	64,56	**0,54**
2lzm	16	50,00	-0,35	18,75	-0,29	37,5	-0,33
2wsy	18	88,89	0,51	88,89	0,5	88,89	0,51
3lzm	59	50,85	0,27	57,63	0,31	54,24	0,28
4lyz	36	63,89	**0,54**	58,33	**0,51**	66,67	**0,53**
Summe:	**646**						

Die mit einer Irrtumswahrscheinlichkeit von einem Prozent signifikant linear abhängigen Korrelationen sind **fett** wiedergegeben.

Tabelle 3.6: Vorhersage-Ergebnisse aus dem Entwicklungsdatensatz für verschiedene Kombinationen der Punkte P_0, P_1 und P_2 für gewählte Abstandsintervalle (Orginalübernommen)

P_0,P_1,P_2	Radius-intervall [Å]:	[3,6]	[3,8]	[3,9]	[3,10]	[3,11]	[3,12]	[3,13]	[3,14]	[3,15]	[3,20]	[4,6]	[4,8]	[4,9]	[4,10]	[4,11]	[4,12]	[4,13]	[4,14]	[4,15]	[4,20]
CA,CZ,O	Spezifität	0.26	0.36	0.36	0.35	0.33	0.32	0.33	0.31	0.28	0.21	0.26	0.36	0.36	0.35	0.33	0.32	0.33	0.31	0.28	0.21
	Sensibilität	0.63	0.63	0.64	0.63	0.58	0.59	0.63	0.6	0.52	0.51	0.62	0.63	0.64	0.63	0.59	0.59	0.63	0.6	0.52	0.51
	rv [%]	50.93	65.79	66.41	65.48	63.31	62.69	62.54	60.84	58.51	44.89	50.62	65.79	66.25	65.48	63.47	62.69	62.38	60.84	58.51	44.89
	r_{cor}	0.326	0.577	0.603	0.606	0.623	0.646	0.655	0.648	0.618	0.017	0.325	0.576	0.603	0.606	0.623	0.646	0.655	0.648	0.618	0.017
CB,CA,O	Spezifität	0.31	0.36	0.34	0.35	0.35	0.34	0.33	0.31	0.3	0.22	0.31	0.36	0.35	0.35	0.35	0.34	0.33	0.31	0.31	0.22
	Sensibilität	0.61	0.66	0.63	0.65	0.65	0.67	0.65	0.58	0.56	0.52	0.63	0.66	0.63	0.65	0.66	0.67	0.65	0.58	0.56	0.52
	rv [%]	59.6	65.94	64.24	64.4	64.4	63.16	62.23	61.76	60.68	47.06	60.22	66.1	64.4	64.55	64.55	63.16	62.23	61.76	60.84	47.06
	r_{cor}	0.486	0.621	0.617	0.631	0.649	0.661	0.666	0.657	0.622	0.033	0.49	0.622	0.618	0.631	0.65	0.662	0.666	0.657	0.622	0.033
CA,CB,O	Spezifität	0.26	0.36	0.36	0.35	0.33	0.32	0.33	0.31	0.28	0.21	0.25	0.36	0.36	0.35	0.33	0.32	0.33	0.31	0.28	0.21
	Sensibilität	0.63	0.63	0.64	0.63	0.58	0.59	0.63	0.61	0.52	0.52	0.62	0.63	0.64	0.63	0.59	0.59	0.63	0.61	0.52	0.52
	rv [%]	50.15	65.79	66.41	65.48	63.31	62.69	62.54	60.99	58.51	45.36	50.0	65.79	66.25	65.48	63.47	62.69	62.38	60.99	58.51	45.36
	r_{cor}	0.326	0.577	0.603	0.606	0.623	0.646	0.655	0.648	0.618	0.016	0.325	0.577	0.603	0.606	0.623	0.646	0.655	0.648	0.618	0.016
CZ,CB,O	Spezifität	0.33	0.35	0.36	0.38	**0.39**	0.37	0.36	0.36	0.33	0.24	0.33	0.35	0.36	0.38	**0.39**	0.38	0.36	0.36	0.33	0.24
	Sensibilität	0.63	0.64	0.65	0.69	0.74	0.74	0.66	0.68	0.62	0.57	0.64	0.64	0.65	0.69	0.74	**0.76**	0.67	0.68	0.62	0.57
	rv [%]	62.54	65.02	65.33	66.72	67.34	65.94	65.02	65.79	62.85	50.15	62.54	65.02	65.33	66.72	67.34	66.25	65.17	65.79	62.85	50.15
	r_{cor}	0.49	0.615	0.648	0.654	0.663	0.67	**0.672**	0.658	0.622	0.077	0.493	0.615	0.648	0.654	0.663	0.67	**0.672**	0.658	0.622	0.077
CB,CZ,O	Spezifität	0.3	0.37	0.35	0.35	0.35	0.34	0.32	0.31	0.29	0.22	0.3	0.37	0.35	0.35	0.34	0.34	0.32	0.31	0.29	0.22
	Sensibilität	0.61	0.66	0.63	0.65	0.63	0.67	0.64	0.56	0.55	0.51	0.62	0.66	0.63	0.65	0.63	0.67	0.63	0.56	0.56	0.51
	rv [%]	58.2	66.56	64.71	64.24	64.4	63.47	61.46	62.07	59.13	46.75	58.82	66.41	64.71	64.09	64.09	63.31	61.61	61.76	59.6	46.75
	r_{cor}	0.411	0.627	0.625	0.63	0.65	0.666	**0.672**	0.665	0.63	0.043	0.445	0.631	0.627	0.632	0.651	0.667	0.671	0.665	0.63	0.043
CZ,CA,O	Spezifität	0.32	0.35	0.36	0.38	**0.39**	0.37	0.36	0.36	0.34	0.25	0.33	0.35	0.36	0.38	**0.39**	0.37	0.36	0.36	0.34	0.25
	Sensibilität	0.63	0.64	0.65	0.69	0.73	0.71	0.66	0.67	0.63	0.58	0.64	0.64	0.65	0.69	0.73	0.73	0.67	0.67	0.63	0.58
	rv [%]	61.61	65.17	65.33	67.18	**67.49**	65.33	65.17	64.86	63.93	50.62	61.61	65.33	65.33	67.18	**67.49**	65.79	65.17	64.86	63.93	50.62
	r_{cor}	0.476	0.608	0.643	0.65	0.661	0.667	0.67	0.657	0.621	0.079	0.479	0.609	0.643	0.651	0.661	0.667	0.67	0.657	0.621	0.079

rv entspricht einer richtigen Vorhersage in Prozent, bei einer Gesamtmenge von 646 Mutationen

3.2.1.4. Suche nach der optimalen Lage der von P_0, P_1 und P_2 definierten Ebene im abstandsabhängigen Aminosäure-Atom-Potential

Die betrachtete Aminosäure wird durch die drei Punkte P_0, P_1 und P_2 repräsentiert. P_0 und P_1 liegen in der definierten Ebene und legen den Richtungsvektor fest. Damit bleibt P_2 frei wählbar. Das einzige Kriterium, dem P_2 genügen muss, ist seine Existenz in allen zu betrachtenden Aminosäuren. Aus diesem Grund wurden die in Kapitel 2.1.1.1 angegebenen P_2-Repräsentanten verwendet. Die Ergebnisse der Vorhersagen aus dem Entwicklungsdatensatz sind in Tabelle 3.7 wiedergegeben. Ausgehend von dem Richtungsvektor **CZCB**, mit CZ als gewähltem Mittelpunkt für das abstandsabhängige Aminosäure-Atom-Potential, weisen die Vorhersagen nur geringe Unterschiede auf. Die deutlichste Verbesserung wird bei der Sensibilität von 0,76 auf 0,82 unter Verwendung von NC (s. Tabelle 2.1 entspricht V_{NC}) als P_2 erzielt. Der höchste Korrelationskoeffizient von 0,673 wird mit C als P_2, die beste richtige Vorhersage von 67,94 % mit NC als P_2 erhalten. Die anfangs für das abstandsabhängige wissensbasierte Potential eingesetzte Kombination von (CZ,CB,O) liegt im Mittel in ihren Ergebnissen am günstigsten (s. Tabelle 3.6 und Tabelle 3.5).

3.2.1.5. Untersuchung der Richtungsabhängigkeit des abstandsabhängigen Aminosäure-Atom-Potentials

Im Gegensatz zu einer nur distanzabhängigen Beschreibung bietet eine richtungs- und abstandsabhängige Betrachtung einen weiteren Informationsgewinn. Die Strukturelemente werden in ihrer relativen Orientierung zu einem der Aminosäure zugeordneten Richtungsvektor wiedergegeben. Die Einteilung der Umgebung erfolgt nicht nur in Abstandsintervallen, sondern zusätzlich in Winkelintervallen. Neben dem Abstand r wird die Lage der Atomtypen durch die Winkel φ und ψ beschrieben (s. Polarkoordinaten Kapitel 2.3.5).

Die Erfassung der Atome erfolgt in von P_0 ausgehenden Kegeln, welche die Winkelintervalle definieren. Die möglichen umgebungsbeschreibenden Parameter entsprechen der Schrittweite der Abstandsintervalle, der Schrittweite der

Winkelintervalle sowie dem Öffnungswinkel der Kegel. Die Suche nach den besten Winkelparametern wurde weitgehend an dem (CZ,CB,O) Aminosäure-Atom-Potential durchgeführt, da dieses die besten Ergebnisse in der abstandsabhängigen Untersuchung lieferte. Es wurden verschiedene Öffnungswinkel und Winkelintervallschrittweiten definiert und miteinander kombiniert (s. Tabelle 3.8). Eine Winkelschrittweite von 45° und ein Öffnungswinkel von 30° bedeuten, dass jeweils die Atome alle 45° ± 30° vom Richtungsvektor in den drei Raumrichtungen entfernt zusammengefasst werden. Eine Winkelschrittweite von 180° und ein Öffnungswinkel von 90° soll zwei Halbkugeln beschreiben, die oberhalb und unterhalb der eingangs definierten Ebene liegen (s. Abbildung 3.1).

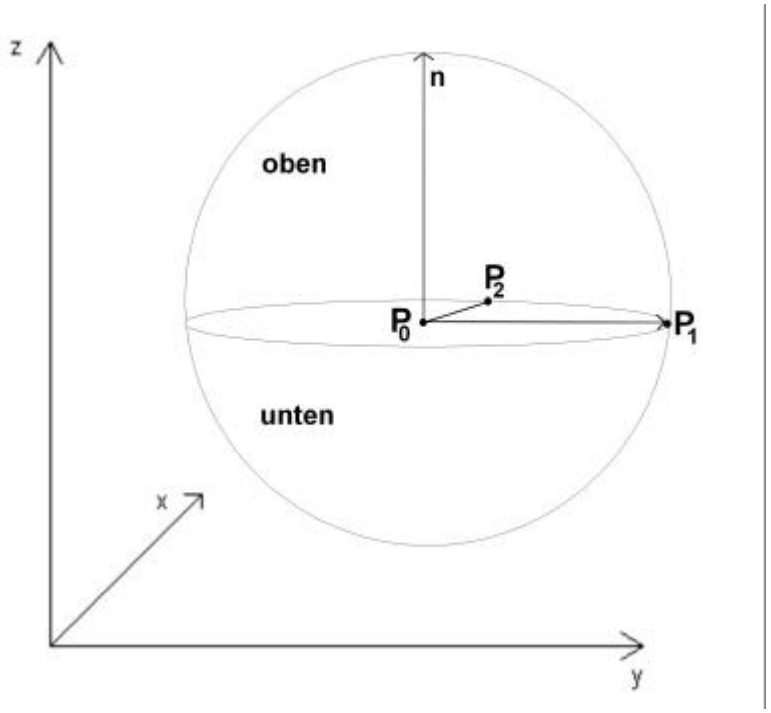


Abbildung 3.1: Räumliche Darstellung der durch P_0, P_1 und P_2 gebildeten Ebene und des Richtungsvektors P_0P_1, n entspricht dem Normalenvektor

Die Ergebnisse in Tabelle 3.8 zeigen nur kleine Unterschiede zwischen den unterschiedlich gewählten Bedingungen für ein betrachtetes Abstandsintervall. Die Korrelation und die richtige Vorhersage sind nur gering gegenüber dem abstandsabhängigen Aminosäure-Atom-Potential verbessert (vgl. Tabelle 3.6). Der Korrelationskoeffizient steigt für (CZ,CB,O) [3,13] von 0,672 auf 0,676, die

richtige Vorhersage für (CZ,CB,O) mit dem Intervall [3,11] und dem Intervall [4,11] von 67,34 % auf 68 %. Die Sensibilität bleibt unverändert bei 0,76. Die leicht verbesserten Ergebnisse werden alle für (CZ,CB,O) und den Winkelintervallschrittweite von 180° sowie 90° und einem Öffnungswinkel von 90° sowie 45° erhalten.

Tabelle 3.7: Vorhersage-Ergebnisse für die Variation des Ebenenrepräsentanten P_2 für die Abstandsintervallschrittweite von 1 Å. (Orginalübernommen)

(CZ,CB,P_2)	Abstandsintervalle [Å]:	[3,8]	[3,9]	[3,10]	[3,11]	[3,12]	[3,13]	[3,14]	[3,15]	[3,20]	[4,8]	[4,9]	[4,10]	[4,11]	[4,12]	[4,14]	[4,14]	[4,15]
C	Spezifität	0,35	0,36	0,38	0,38	0,37	0,35	0,36	0,33	0,24	0,35	0,36	0,38	0,38	0,38	0,36	0,36	0,33
	Sensibilität	0,64	0,65	0,69	0,73	0,74	0,66	0,68	0,63	0,57	0,64	0,65	0,69	0,73	0,76	0,67	0,68	0,63
	rv [%]	65,02	65,63	66,72	66,87	65,63	64,71	65,63	63,16	49,69	65,02	65,63	66,72	66,87	65,94	64,86	65,63	63,16
	r_{cor}	0,61	0,646	0,653	0,663	0,671	**0,673**	0,66	0,624	0,078	0,611	0,646	0,653	0,664	0,671	**0,673**	0,66	0,624
N	Spezifität	0,35	0,36	0,38	0,39	0,37	0,36	0,36	0,34	0,25	0,35	0,36	0,38	0,39	0,38	0,36	0,36	0,34
	Sensibilität	0,64	0,65	0,71	0,73	0,73	0,68	0,68	0,63	0,58	0,64	0,65	0,71	0,73	0,75	0,69	0,68	0,63
	rv [%]	65,17	65,63	67,18	67,49	65,63	65,02	65,02	63,62	50,46	65,33	65,63	67,18	67,49	65,94	65,17	65,02	63,62
	r_{cor}	0,603	0,638	0,646	0,657	0,664	0,667	0,654	0,619	0,079	0,603	0,638	0,646	0,657	0,664	0,667	0,654	0,619
NC	Spezifität	0,4	0,4	0,39	0,39	0,38	0,36	0,34	0,31	0,29	0,4	0,4	0,39	0,39	0,38	0,36	0,34	0,31
	Sensibilität	0,74	0,79	0,81	**0,82**	0,8	0,75	0,7	0,6	0,57	0,74	0,79	0,81	**0,82**	0,8	0,75	0,7	0,6
	rv [%]	**67,97**	67,45	66,67	66,41	65,36	63,02	61,46	59,64	58,07	**67,97**	67,45	66,67	66,41	65,36	63,02	61,46	59,64
	r_{cor}	0,6	0,624	0,63	0,638	0,639	0,631	0,608	0,575	0,253	0,599	0,624	0,63	0,638	0,639	0,631	0,608	0,575
O	Spezifität	0,35	0,36	0,38	0,39	0,37	0,36	0,36	0,33	0,24	0,35	0,36	0,38	0,39	0,38	0,36	0,36	0,33
	Sensibilität	0,64	0,65	0,69	0,74	0,74	0,66	0,68	0,62	0,57	0,64	0,65	0,69	0,74	0,76	0,67	0,68	0,62
	rv [%]	65,02	65,33	66,72	67,34	65,94	65,02	65,79	62,85	50,15	65,02	65,33	66,72	67,34	66,25	65,17	65,79	62,85
	r_{cor}	0,615	0,648	0,654	0,663	0,67	0,672	0,658	0,622	0,077	0,615	0,648	0,654	0,663	0,67	0,672	0,658	0,622

Tabelle 3.8: Vorhersage-Ergebnisse aus dem Entwicklungsdatensatz für verschiedene Kombinationen gewählter Abstands- und Richtungsintervalle für die Abstandsintervallschrittweite von 1 Å (Orginalübernommen)

P_0,P_1,P_2	Öffnungswinkel [°]	Schrittweite [°]	Radiusintervall [Å]:	[3,8]	[3,9]	[3,10]	[3,11]	[3,12]	[3,13]	[3,14]	[3,15]	[4,8]	[4,9]	[4,10]	[4,11]	[4,12]	[4,13]	[4,14]	[4,15]
CZ,CB,O	90	180	Spezifität	0,35	0,36	0,36	0,38	0,38	0,34	0,36	0,3	0,35	0,36	0,36	0,38	0,38	0,35	0,36	0,3
			Sensibilität	0,63	0,65	0,67	0,72	0,73	0,65	0,66	0,6	0,63	0,65	0,67	0,72	**0,76**	0,67	0,66	0,6
			rv [%]	64,7	65,5	65,6	66,4	66,1	63,8	65	62	64,9	65,5	65,6	66,4	66,7	63,9	65	62
			r_{cor}	0,615	0,648	0,654	0,664	**0,672**	**0,676**	0,661	0,622	0,616	0,649	0,654	0,665	**0,672**	**0,673**	0,662	0,623
CZ,CB,O	90	90	Spezifität	0,35	0,36	0,38	0,38	0,36	0,35	0,36	0,3	0,35	0,36	0,38	0,38	0,37	0,36	0,36	0,3
			Sensibilität	0,64	0,65	0,7	0,73	0,71	0,66	0,67	0,6	0,64	0,65	0,7	0,73	0,73	0,68	0,67	0,6
			rv [%]	64,9	65,6	67,2	67	64,7	64,4	65	64	65	65,6	67,2	67	65,2	64,9	65	64
			r_{cor}	0,602	0,638	0,645	0,656	0,662	0,666	0,653	0,618	0,602	0,638	0,645	0,656	0,662	0,666	0,653	0,618
CZ,CB,O	45	90	Spezifität	0,37	0,37	**0,39**	**0,39**	**0,39**	0,36	0,34	0,3	0,37	0,37	**0,39**	**0,39**	**0,39**	0,36	0,34	0,3
			Sensibilität	0,67	0,66	0,71	**0,76**	0,73	0,69	0,65	0,6	0,67	0,66	0,71	**0,76**	0,73	0,7	0,65	0,6
			rv [%]	66,6	67	**68,4**	68	68	64,7	63,6	64	66,6	67	**68,4**	68	67,8	64,9	63,6	64
			r_{cor}	0,525	0,567	0,573	0,588	0,595	0,6	0,581	0,541	0,525	0,567	0,573	0,588	0,595	0,6	0,581	0,541
CZ,CB,O	60	60	Spezifität	0,34	0,36	**0,39**	0,37	0,36	0,35	0,33	0,3	0,34	0,36	**0,39**	0,37	0,32	0,35	0,33	0,3
			Sensibilität	0,62	0,63	0,73	0,71	0,71	0,68	0,65	0,6	0,63	0,64	0,73	0,72	0,72	0,68	0,65	0,6
			rv [%]	63,9	65,6	67,5	65,9	64,6	63,6	62,5	63	64,2	65,6	67,5	66,3	63	63,6	62,5	63
			r_{cor}	0,586	0,624	0,629	0,638	0,64	0,635	0,607	0,548	0,587	0,625	0,629	0,638	0,64	0,635	0,607	0,548
CZ,CB,O	30	30	Spezifität	0,34	0,35	0,36	0,34	0,34	0,32	0,31	0,3	0,34	0,35	0,36	0,34	0,34	0,32	0,31	0,3
			Sensibilität	0,63	0,67	0,72	0,73	0,73	0,68	0,66	0,7	0,63	0,67	0,72	0,73	0,73	0,68	0,66	0,7
			rv [%]	63,2	64,6	65	62,2	61,9	60,2	59,3	58	63,2	64,6	65	62,2	61,9	60,2	59,3	58
			r_{cor}	0,46	0,491	0,486	0,486	0,469	0,432	0,36	0,276	0,462	0,492	0,486	0,487	0,47	0,433	0,361	0,277
CZ,CA,O	90	180	Spezifität	0,36	0,36	0,38	0,38	0,37	0,35	0,35	0,3	0,36	0,36	0,38	0,38	0,37	0,35	0,35	0,3
			Sensibilität	0,64	0,64	0,69	0,73	0,73	0,67	0,66	0,6	0,64	0,64	0,69	0,73	0,74	0,67	0,66	0,6
			rv [%]	65,5	65,9	67,2	67	65,6	64,7	64,1	63	65,5	66,1	67,2	67,2	65,9	64,7	64,1	63
			r_{cor}	0,607	0,641	0,649	0,66	0,668	0,67	0,657	0,623	0,608	0,641	0,649	0,66	0,668	0,67	0,657	0,623

3.2.1.6. Untersuchung verschiedener Schrittlängen für das Abstandsintervall einer richtungs- und abstandsabhängigen Aminosäure-Atom-Potentialfunktion

Für ein festgelegtes richtungs- und abstandsabhängiges Aminosäure-Atom-Potential mit den drei Aminosäurerepräsentanten (CZ,CB,O) wurde die Schrittweite der Abstandsintervalle einmal mit großer Schrittlänge (einer Schale entsprechend, 10 Å, s. Tabelle 3.9) und einmal mit kleiner Schrittlänge (0,5 Å, s. Tabelle 3.10) gewählt.

Tabelle 3.9: Darstellung des richtungs- und abstandsabhängigen Aminosäure-Atom-Potentials für (CZ,CB,O) für verschiedene Öffnungswinkel und Winkelschrittweiten

(P_0, P_1, P_2)	Öffnungswinkel [°]	Winkelschrittweite [°]	Radienintervall [Å]:	[3,13]
(CZ,CB,O)	45	45	Spezifität	0,37
			Sensibilität	**0,74**
			rv [%]	65,17
			r_{cor}	0,647
(CZ,CB,O)	30	60	Spezifität	0,35
			Sensibilität	0,72
			rv [%]	63,31
			r_{cor}	**0,664**
(CZ,CB,O)	45	90	Spezifität	0,35
			Sensibilität	0,71
			rv [%]	63,78
			r_{cor}	0,626
(CZ,CB,O)	60	60	Spezifität	0,35
			Sensibilität	0,73
			rv [%]	63,47
			r_{cor}	0,656
(CZ,CB,O)	90	90	Spezifität	0,35
			Sensibilität	0,72
			rv [%]	63,16
			r_{cor}	0,635
(CZ,CB,O)	90	180	Spezifität	0,36
			Sensibilität	**0,74**
			rv [%]	63,62
			r_{cor}	0,618

Die Vorhersageergebnisse für die große Schrittweite der Abstandsintervalle von 10 Å sind durchgehend schlechter, verglichen mit der Ausgangsradienschrittweite von 1 Å. Die Vorhersagergebnisse für eine gewählte Schrittweite der Ra-

dienintervalle von 0,5 Å zu 1 Å sind geringfügig besser (s. Tabelle 3.8 und Tabelle 3.10).

3.2.1.7. Untersuchung der gewählten richtungs- und abstandsabhängigen wissensbasierten Aminosäure-Atom-Potentialfunktion

Die richtungsabhängige Beschreibung der Umgebung führt für das verwendete abstandsabhängige Aminosäure-Atom-Potential zu geringfügigen Verbesserungen in den Werten für den Korrelationskoeffizienten und bezüglich der richtigen Vorhersage. Diese Ergebnisse werden für die folgenden Parameter erhalten:

- Schrittweite des Abstandsintervalls 1 Å oder 0,5 Å
- Öffnungswinkel 90°
- Schrittweite 90° oder 180°
- Die Repräsentanten (CZ,CB,O)

Im Folgenden wird im Hinblick auf Kapitel 3.2.3 die richtungs- und abstandsabhängige Potentialfunktion mit (CZ,CB,O), einer Schrittweite des Abstandsintervalls von 0,5 Å, dem Öffnungswinkel 90° und der Schrittweite des Winkelintervalls von 180° näher betrachtet.

Tabelle 3.10: Vorhersage-Ergebnisse aus dem Entwicklungsdatensatz für verschiedene Kombinationen gewählter Abstands- und Richtungsintervalle für die Abstandsintervallschrittweite von 0,5 Å (Orginalübernommen)

P_0,P_1,P_2	Öffnungswinkel [°]	Schrittweite [°]	Radiusintervall [Å]:	[3,8]	[3,9]	[3,10]	[3,11]	[3,12]	[3,13]	[3,14]	[3,15]	[4,8]	[4,9]	[4,10]	[4,11]	[4,12]	[4,13]	[4,14]	[4,15]
CZ,CB,O	90	180	Spezifität	0,35	0,36	0,37	0,38	0,38	0,35	0,36	0,32	0,35	0,37	0,37	0,38	0,38	0,36	0,36	0,32
			Sensibilität	0,62	0,65	0,67	0,72	0,73	0,67	0,67	0,61	0,63	0,65	0,67	0,72	0,76	0,68	0,67	0,61
			rv [%]	65,33	66,25	66,4	66,9	66,3	64,7	65	62,1	65,3	66,3	66,1	66,6	66,6	65,2	65	62,07
			r_{cor}	0,613	0,652	0,655	0,664	0,671	0,673	0,66	0,619	0,617	0,65	0,655	0,67	0,671	0,672	0,66	0,62
CZ,CB,O	90	90	Spezifität	0,35	0,37	0,37	0,38	0,37	0,35	0,36	0,33	0,35	0,37	0,36	0,38	0,38	0,36	0,36	0,33
			Sensibilität	0,62	0,65	0,67	0,71	0,73	0,66	0,67	0,63	0,63	0,65	0,67	0,71	0,76	0,69	0,68	0,63
			rv [%]	65,33	66,41	66,4	**67,5**	65,5	64,9	65,3	63,2	65,3	66,4	65,8	67,2	66,3	65,6	65,2	63,16
			r_{cor}	0,611	0,647	0,653	0,665	0,672	0,673	0,662	0,622	0,615	0,649	0,654	0,665	0,672	**0,674**	0,662	0,622
CZ,CB,O	45	90	Spezifität	0,36	0,37	0,38	**0,39**	0,38	0,37	0,35	0,35	0,36	0,37	**0,39**	0,38	**0,39**	0,37	0,35	0,34
			Sensibilität	0,65	0,66	0,68	0,73	0,74	0,69	0,65	0,64	0,65	0,66	0,69	0,73	0,75	0,69	0,66	0,64
			rv [%]	66,1	66,25	68	67,5	67	66,3	63,9	64,2	65,6	66,3	68,3	67,2	67,3	66,1	63,9	63,93
			r_{cor}	0,589	0,627	0,631	0,64	0,649	0,648	0,628	0,574	0,593	0,629	0,632	0,641	0,649	0,649	0,629	0,574
CZ,CB,O	180	0	Spezifität	0,35	0,36	0,37	0,38	0,37	0,35	0,35	0,33	0,35	0,36	0,37	0,38	0,38	0,35	0,35	0,33
			Sensibilität	0,62	0,65	0,68	0,73	0,74	0,65	0,65	0,62	0,63	0,65	0,68	0,73	**0,78**	0,66	0,66	0,62
			rv [%]	65,33	66,25	66,1	67,2	65,6	65	64,7	62,7	65,3	66,1	65,9	67	66,4	64,9	64,9	62,69
			r_{cor}	0,613	0,648	0,655	0,666	0,673	**0,674**	0,663	0,624	0,617	0,65	0,655	0,666	0,673	**0,674**	0,663	0,625

3.2.1.8. Füllung der Umgebungsschalen durch die erfassten Atomtypen und das modellierte Wasser

Bei einer abstandsabhängigen Betrachtung der Umgebung einer als Kugel approximierten Aminosäure ergeben sich aus den verwendeten Radiusintervallen Kugelschalen. Die unterschiedlichen Atomtypen werden in den Umgebungsschalen erfasst und gezählt (s. Kapitel 2.2.3.2). Im Gegensatz zu dem im Kristallstrukturdatensatz definierten Protein muss das Lösungsmittel erst noch modelliert werden. Der nicht von Protein besetzte Raum soll vollständig mit Wasser gefüllt sein, ohne das Gesamtvolumen zu überschreiten. Zur Untersuchung des Füllungsgrades der Umgebungsschalen wurde die Anzahl der Atomtypen in gegebenen Radiusintervallen bestimmt und ihr Volumen berechnet. Das Kugelvolumen wurde nach Kapitel 2.3.6 bestimmt und auf 100% gesetzt. Der Volumen der Atomtypen wurde mit dem Kugelvolumen in Bezug gesetzt. Die Radien der Atomtypen wurden entsprechend den vom Programm *psa* (s. Kapitel 2.2.8) und von Richards und Lee [Lee und Richards, 1971; Richards, 1974] angegebenen und benutzen mittleren Atomradien gewählt (r_{cali}=1,9 Å, r_{carom}=1,9 Å, r_{n}=1,7 Å, r_{oas}=1,4 Å, r_{om}=1,86 Å). Die Anzahl der in den Radiusintervallen gefundenen Atomtypen wurde über alle Aminosäuren gemittelt. Atome, die zu einer betrachteten Aminosäure gehören, wurden nicht erfasst. Packungseffekte und sich dadurch ändernde Atomradien fließen nicht in die Berechnung mit ein. Die Abbildung 3.2 zeigt den prozentualen Anteil eines Atomtyps an der theoretisch maximal möglichen Schalenfüllung sowie die relative Gesamtfüllung der Umgebungsschalen. Mit steigendem Kugelschalenradius nimmt die Menge an vorhergesagtem Lösungsmittel zu. Eine maximale Füllung der Umgebungsschalen wird im Mittel zu 99,3 ± 7,3 % erreicht. Bis 4 Å ist die Füllung der Schalen unvollständig. Während zwischen 4-6 Å und ab 16 Å die Schalen überfüllt sind, lässt sich eine Unterbesetzung der Umgebungsschalen zwischen 6-14 Å beobachten (s. Kapitel 4.1.1.1).

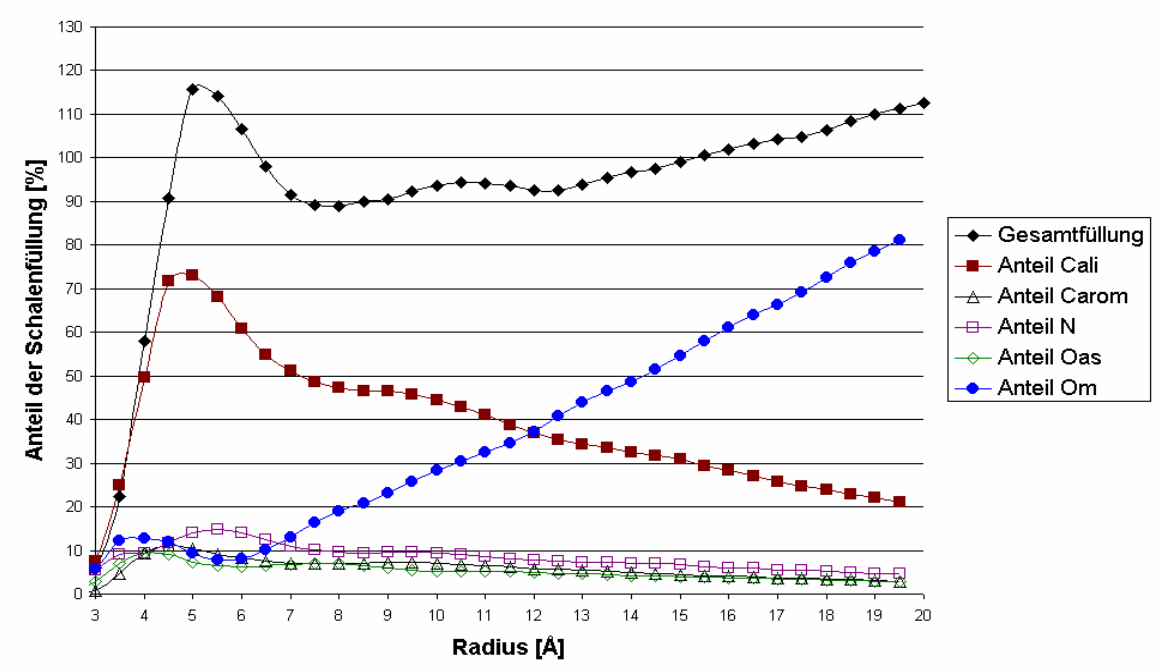


Abbildung 3.2: Abschätzung des Füllungsgrades der Umgebungsschalen um die mittlere Aminosäure für (CZ,CB,O)

3.2.1.9. Aminosäureabhängige Verteilungen der erfassten Atomtypen im Strukturdatensatz

Das verwendete Aminosäure-Atom-Potential beruht auf der grundlegenden Annahme, dass die Umgebung jeder Aminosäure spezifisch für diese Aminosäure ist. Die Verteilung der Atomtypen ist in den erfassten Umgebungsschalen abhängig von der betrachteten Aminosäure. Bei der Ableitung des Potentials wird die unterschiedliche Zusammensetzung der Umgebung aus Atomtypen in einen Aminosäure-Atomtyp-Wechselwirkungsbeitrag umgesetzt. Die abstands- und richtungsabhängigen Verteilungen der Atomtypen für (CZ,CB,O) in zwei Halbkugeln (s. Abbildung 3.1) sind jeweils für eine kleine, eine hydrophobe, eine positiv und eine negativ geladene sowie eine aromatische Aminosäure in Abbildung 3.3 dargestellt. Die Berechnung der Verteilungen erfolgt wie in Kapitel 2.2 angegeben.

Die Umgebungsschalen sind für alle dargestellten Aminosäuren bei 5 Å und ab 15 Å stärker gefüllt, teilweise sogar überfüllt, während sie im Zwischenbereich eher unvollständig gefüllt sind. Den größten Besetzungsanteil an allen Kugelschalen für alle Aminosäuren stellen die aliphatischen Kohlenstoffatome C_{ali} und der Sauerstoff aus dem Lösungsmittel O_m dar. Während der Anteil für aliphatische Kohlenstoffe mit dem Radius abnimmt, füllt der Sauerstoff aus dem Lösungsmittel die Schalenelemente in immer stärker werdendem Maße aus. Die Anteile für aromatischen Kohlenstoff C_{arom}, Stickstoff N und Sauerstoff aus den Aminosäuren O_{as} weisen ab 9 Å keine starken Schwankungen und Unterschiede in ihren Verteilungen mehr auf.

Die Verteilungskurven der Atomtypen zeigen für die einzelnen dargestellten Aminosäuren einen unterschiedlichen Verlauf. Der Anteil der aliphatischen Kohlenstoffe hat für alle Aminosäuren bis auf Glycin (dort 5 Å) das Maximum bei 6 Å und fällt im weiteren Verlauf für die Aminosäuren Leucin und Glycin steil ab, zeigt dann aber für Leucin ein Plateau bis 11 Å, um schließlich auch langsam abzuflachen. Der Anteil des Lösungsmittelsauerstoffs an der Schalenfüllung zeigt für die hydrophoben und die aromatischen Aminosäuren einen zunächst flachen Verlauf, um später überproportional zuzunehmen (hängender Kurvenverlauf). Im Fall der polaren Aminosäuren steigt der Lösungsmittelsauerstoffanteil zunächst ab 7 Å steil an, um später abzuflachen. Die Füllung der letzten Schale mit dem Lösungsmittelsauerstoff liegt bei allen Aminosäuren um 80 %.

Die geladenen Aminosäuren weisen einen deutlich höheren Anteil von Stickstoff und strukturellem Sauerstoff O_{as}, neben Lösungsmittelsauerstoff im Bereich um 6 Å, auf als hydrophobe oder aromatische Aminosäuren. Die Umgebung polarer Aminosäuren ist ebenfalls polar. Der Atomtyp des aromatischen Kohlenstoffes erreicht bei den Aminosäuren Tryptophan und Leucin einen Anteil von bis zu 16 %, während dieser bei den polaren Aminosäuren bei ungefähr 5 % liegt. Insgesamt zeigen die Kurven ein hohes Maß an Spezifität für jede Aminosäure und nur eine geringe zwischen den Halbkugeln.

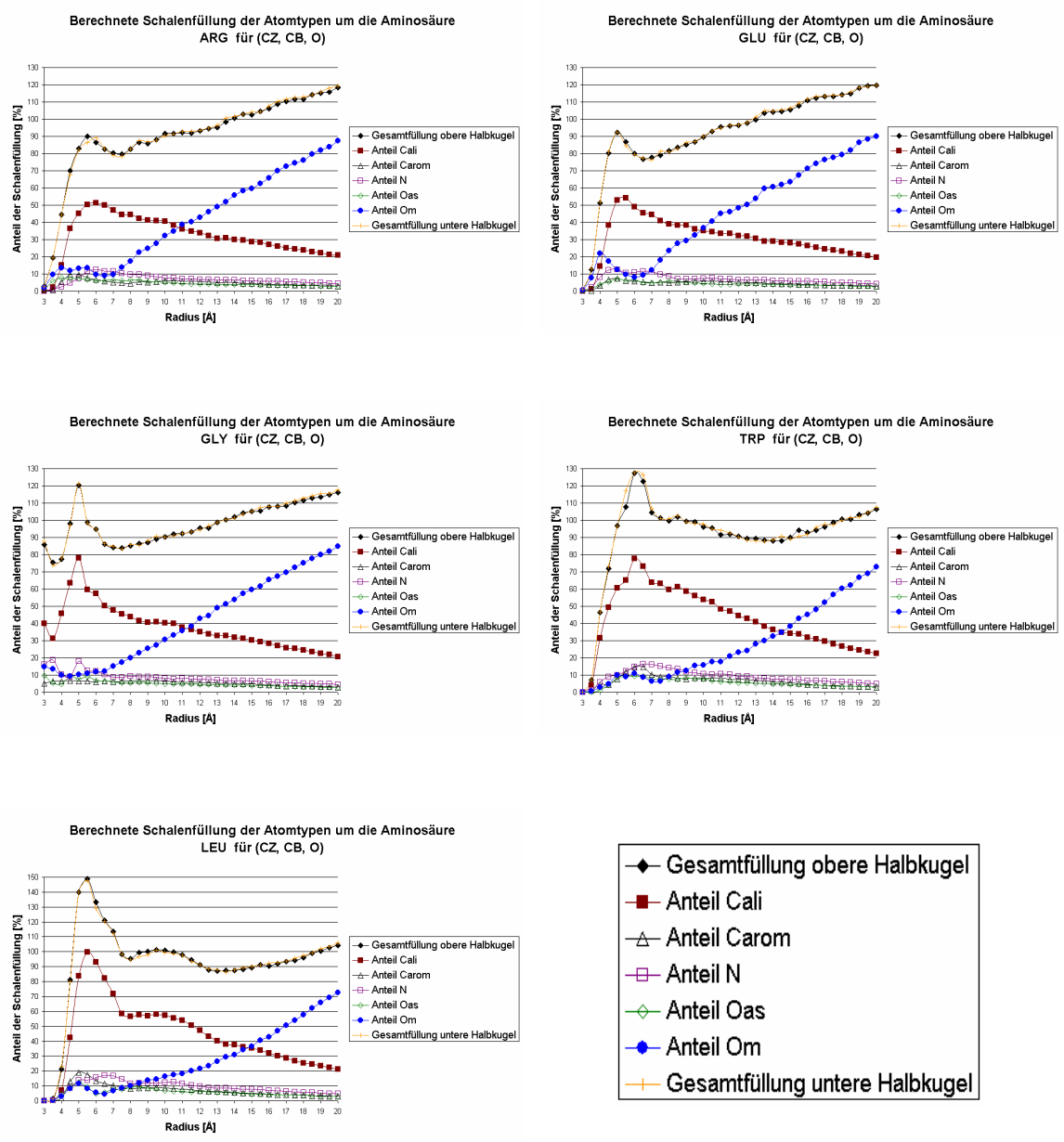


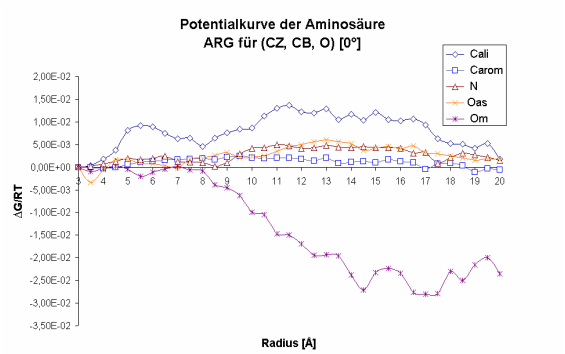
Abbildung 3.3: Darstellung der prozentualen Besetzung der Umgebungsschalen durch die fünf Atomtypen (x-Achse: Radius [Å], y- Achse: Anteil der Schalenfüllung [%])

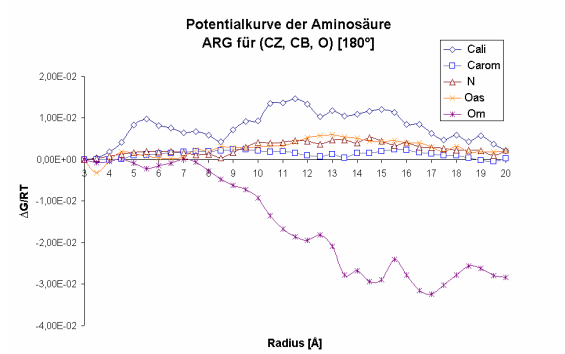
3.2.1.10. Aminosäure-Atom-Potentiale

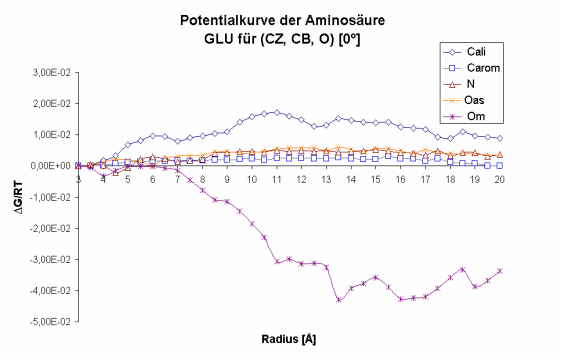
Die Häufigkeitsverteilungen der Atomtypen um eine betrachtete Aminosäure werden durch Anwendung der inversen Boltzmanngleichung (s. Kapitel 2.1.1) in Energiewerte umgerechnet. Die graphische Darstellung der abstands- und richtungsabhängigen Potentialkurven ist für einige ausgesuchte Aminosäuren in Abbildung 3.4 wiedergegeben.

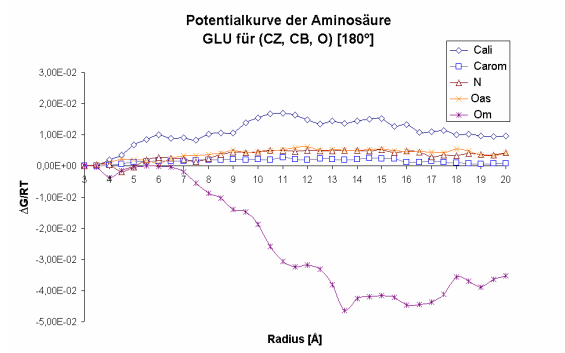
Die Potentialkurven beschreiben den abstands- und richtungsabhängigen energetischen Wechselwirkungsbeitrag zwischen einer Aminosäure und den Umgebungselementen. Die Umgebungselemente entsprechen den in Kapitel 2.2.3.2 definierten Atomtypen. Die Kurvenverläufe schwanken für kleine Abstände stark um das Nullniveau, während sie für große Abstände gegen Null konvergieren oder um einen konstanten Wert pendeln.

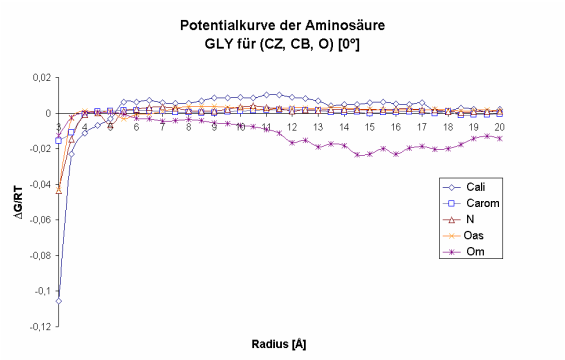
Die Kurvenscharen (s. Abbildung 3.4) der unterschiedlichen Aminosäuren sowie das Verhalten der Potentiale in den Halbkugeln zeigen einen charakteristischen Verlauf. Außer nach diesen Aminosäure-spezifischen Einzelheiten lassen sich die Wechselwirkungen nach Aminosäureeigenschaften (z.B. Hydrophobizität oder Größe) gruppieren. So zeigen hydrophobe und aromatische Aminosäuren durchgehend einen Kurvenverlauf unterhalb der Nulllinie für die Wechselwirkung mit aliphatischen Kohlenstoffen, dagegen polare oder geladene Aminosäuren oberhalb der Nulllinie, entsprechend einer stabilisierenden oder destabilisierenden Wechselwirkung. Die Kurvenverläufe für Sauerstoff aus dem Lösungsmittel sind denen für aliphatischen Kohlenstoff entgegengesetzt.

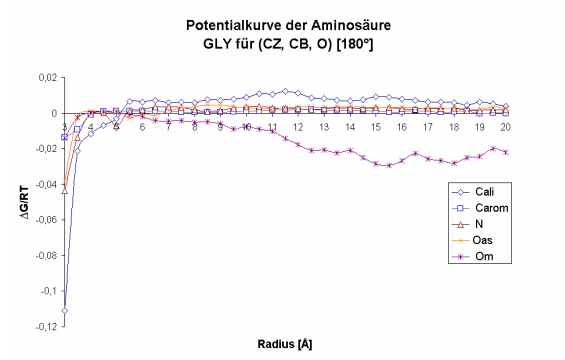

Potentialkurve der Aminosäure
ARG für (CZ, CB, O) [0°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]

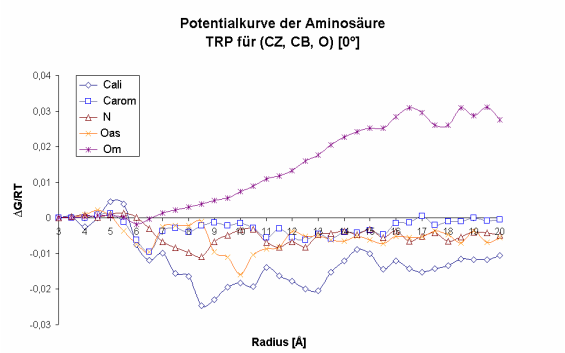

Potentialkurve der Aminosäure
ARG für (CZ, CB, O) [180°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]

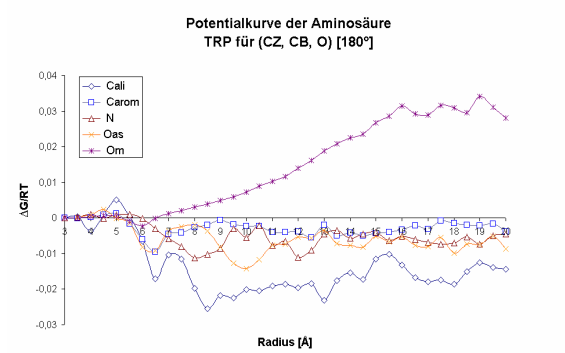

Potentialkurve der Aminosäure
GLU für (CZ, CB, O) [0°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]


Potentialkurve der Aminosäure
GLU für (CZ, CB, O) [180°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]


Potentialkurve der Aminosäure
GLY für (CZ, CB, O) [0°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]


Potentialkurve der Aminosäure
GLY für (CZ, CB, O) [180°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]


Potentialkurve der Aminosäure
TRP für (CZ, CB, O) [0°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]


Potentialkurve der Aminosäure
TRP für (CZ, CB, O) [180°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]

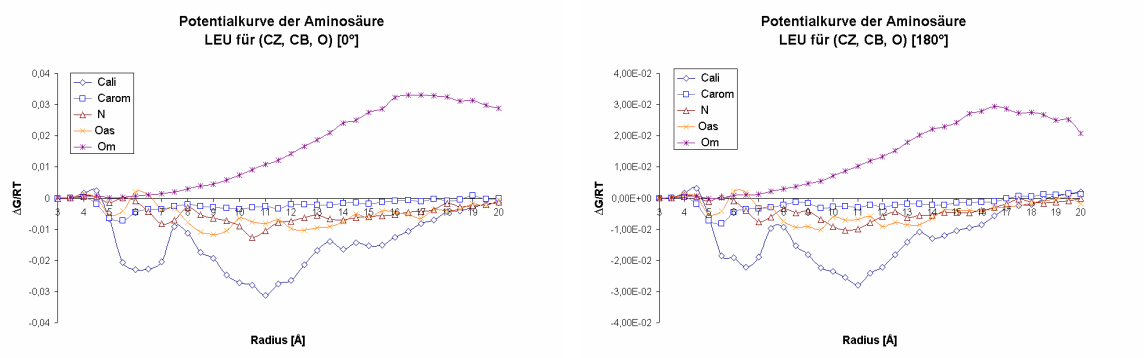


Abbildung 3.4: Darstellung des abstands- und richtungsabhängigen Aminosäure-Atom-Potentials (CZ,CB,O) für die Aminosäuren ARG,GLU,GLY, TRP und LYS (auf dieser und der vorhergehenden Seite, x-Achse: Radius[Å], y-Achse: DG/RT)

3.2.1.11. Qualitative Analyse der Kurvenverläufe für das Aminosäure-Atom-Potential

Die Potentialkurvenverläufe für die Atomtypen (s. Abbildung 3.4) zeigen Charakteristika für einzelne Aminosäuren und bezüglich der Eigenschaften von Aminosäuregruppen, wie Polarität, Hydrophobizität oder der Größe der Aminosäuren. Im Folgenden sollen diese Potentialkurvenverläufe für die jeweiligen Atomtypen genauer beschrieben werden.

Aliphatische Kohlenstoffatome

In ihrer Wechselwirkung mit hydrophoben Aminosäuren zeigen aliphatische Kohlenstoffatome eine stark stabilisierende Wirkung, wobei der Bereich der maximalen Stabilisierung mit zunehmender Größe der Aminosäure wächst. Die Kurven zeigen zwei Minima für 5 Å und um 11 Å. Die Aminosäuren Methionin und Cystein zeigen einen durchgehend konstanten stabilisierenden Verlauf in ihrem Wechselwirkungsbeitrag mit dem Atomtyp des aliphatischen Kohlenstoffes. Für Serin und Threonin zeigt der anfängliche Potentialkurvenverlauf Ähnlichkeit mit dem von kleinen hydrophoben Aminosäuren, mit dem Minimum bei 5 Å. Bei zunehmendem Radius wird diese Wechselwirkung destabilisierend. Aromatische Aminosäuren zeigen eine leichte Destabilisierung bei kleinen Ra-

dien, die sich bei wachsendem Radius in eine Stabilisierung mit einem Minimum bei 8-9 Å wandelt.
Für polare Aminosäuren ist die Interaktion mit aliphatischen Kohlenstoffatomen destabilisierend. Dieser Effekt kann für Radien kleiner als 5 Å stark (Lysin oder Arginin) oder schwach (Asparagin oder Asparaginsäure) ausgeprägt sein.

Aromatische Kohlenstoffatome

In ihrer Wechselwirkung ähneln aromatische Kohlenstoffe aliphatischen Kohlenstoffen und dem Stickstoff. Es zeigt sich eine stabilisierende Interaktion mit hydrophoben Aminosäuren sowie eine destabilisierende für polare Aminosäuren. Der maximale stabilisierende und destabilisierende Wechselwirkungsbeitrag liegt um 6 Å (s. z.B. LEU oder TRP in Abbildung 3.4).

Stickstoffatome

Die Potentialkurven zeigen für hydrophobe Aminosäuren einen stabilisierenden Beitrag, mit einem Minimum bei 3-6 Å. Der Wechselwirkungsverlauf für die Aminosäuren Serin und Threonin erinnert an den der aliphatischen Kohlenstoffatome. Für kleine Radien findet sich eine Stabilisierung, die sich mit zunehmendem Radius in eine Destabilisierung wandelt. Die stabilisierende Wechselwirkung mit aromatischen Aminosäuren zeigt ihr Minimum bei 8 Å.
Eine destabilisierende Interaktion mit Stickstoffatomen kann für polare Aminosäuren beobachtet werden. Die stabilisierende Wechselwirkung kann bei Radien kleiner als 5 Å unterschiedlich stark sein.

Sauerstoffatome

Grundsätzlich kann die Wechselwirkung mit Sauerstoffatomen für hydrophobe Aminosäuren als destabilisierend, für polare Aminosäuren als stabilisierend beschrieben werden.

3.2.1.12. Quantitative Analyse der Kurvenverläufe für das Aminosäure-Atom-Potential

Für die quantitative Untersuchung der Zusammenhänge zwischen den Potentialkurven der Atomtypen wurden die jeweiligen Korrelationsmatrizen für die zwanzig Aminosäuren berechnet (s. Tabelle 3.11, Anhang 6.4.3 und Kapitel 2.3.8).

Tabelle 3.11: Die Korrelationskoeffizienten r_{hr} zwischen den gemittelten 20 Aminosäure-Atom-Potentialen. Für die 36 Datenpunkte wird ein signifikant linearer Zusammenhang für eine Irrtumswahrscheinlichkeit von 0,1 % bei r > 0,519 festgestellt.

	Obere Halbkugel					Untere Halbkugel				
	C_{ali}	C_{arom}	N	O_{as}	O_m	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,00					1,00				
C_{arom}	0,28	1,00				0,21	1,00			
N	0,62	0,31	1,00			0,60	0,26	1,00		
O_{as}	0,64	0,11	0,45	1,00		0,63	0,03	0,44	1,00	
O_m	-0,25	-0,14	-0,12	-0,09	1,00	-0,26	-0,16	-0,11	-0,09	1,00

Die entsprechenden Korrelationskoeffizienten der Potentialkurven für die Halbkugeln unterscheiden sich nur geringfügig. Die größten Unterschiede zwischen den Korrelationskoeffizienten weisen Methionin, Cystein und Histin auf.

Die Potentiale für aliphatischen Kohlenstoff korrelieren bei fast allen Aminosäuren signifikant mit einer Irrtumswahrscheinlichkeit von 0,1 % mit denen von Stickstoff und strukturellem Sauerstoff, unabhängig von dem gewählten Richtungselement. Die Ausnahmen sind Alanin, Prolin und Valin.

Polare Aminosäuren zeigen höhere Korrelationskoeffizienten als hydrophobe oder aromatische Aminosäuren. Serin und Threonin verhalten sich wie kleine hydrophobe Aminosäuren.

Die Korrelationskoeffizienten zwischen Lösungsmittelsauerstoff und den meisten Aminosäuren sowie Atomtypen weisen negative Werte auf.

3.2.1.13. Versuche zur Optimierung des gewählten Aminosäure-Atom-Potentials

Es werden Möglichkeiten zur weiteren Verbesserung des Parametersatzes des Aminosäure-Atom-Potentials gesucht. In den vorhergehenden Kapiteln war die Bewertungsfunktion im Hinblick auf die Korrelation zwischen experimentellen Stabilitätsdaten und theoretisch berechneten noch nicht optimiert. Die Gesamtwechselwirkungsenergie der Aminosäure mit ihrer Umgebung wird durch Summation der Wechselwirkungsbeiträge der Atomtypen über die Abstandsintervalle, die Umgebungsschalen, erhalten. Es werden die fünf definierten Atomtypen verwendet und ein gemeinsamer Gesamtabschnitt betrachtet. Im folgenden werden optimale Radienintervalle für die einzelnen Atomtypen gesucht, die zu einem Gesamtpotential kombiniert werden. Um die Energiebeiträge der Atomtypen bei ihrer Kombination gewichten zu können, müssen diese voneinander abhängigen Variablen in unabhängige Faktoren überführt werden. Dazu wurde eine Hauptkomponentenanalyse (HKA) durchgeführt (s. Kapitel 2.3.9).

3.2.1.14. Suche nach optimalen Radienintervallen

Die Wechselwirkungsenergie der Umgebung wurde durch Summation der abstands- und richtungsabhängigen Intervalle über die fünf Atomtypen ermittelt. Im Folgenden werden für den Entwicklungsdatensatz die Gesamtkorrelationen für einen Atomtyp für alle möglichen Intervalle bestimmt. Die Intervallgrenzen liegen zwischen 3 Å und 20 Å, daraus ergeben sich 171 Kombinationen von 3-3.5, 3-4, 3-5, ..., bis 19.5-20 für die einzelnen Atomtypen. In Tabelle 3.12 sind die vom Betrag her höchsten Korrelationen der einzelnen Atomtypen und die jeweiligen Intervalle dargestellt.

Die Korrelationen der einzelnen Atomtypen sind nicht besser als die bislang erreichte maximale Korrelation (r_{cor}=0,67). Der hohe Korrelationskoeffizient von r_{cor}=0,68 für den Atomtyp der aromatischen Kohlenstoffatome wurde für 266 Datenpunkte bestimmt. Nicht für jede Mutation lassen sich in den gewählten Radienintervallen aromatische Kohlenstoffatome finden.

Kleine Intervalllängen führen für die einzeln betrachteten Atomtypen zu besseren Ergebnissen. Die optimalen Intervallgrenzen für alle Atomtypen liegen zwischen 3 Å und 13 Å. Auffallend sind die vom Betrag her niedrigen oder sogar negativen Korrelationswerte für den Atomtyp O_m. Dies ist im Einklang mit der Literatur [Leven, 1999] und führt bei der Kombination der Atomtypen zu einer Verbesserung der entsprechenden Korrelation (s. Tabelle 3.14).

Tabelle 3.12: Die vier besten Intervalllängen für die Bestimmung der Gesamtkorrelation des Entwicklungsdatensatzes für die einzelnen Atomtypen

Atomtyp	Radienintervall [Å]	Datenpunkte	rv [%]	r_{cor}
C_{ali}	[8,11]	646	65	0,653
	[8,12]	646	65	0,653
	[7,12]	646	65	0,651
	[8,13]	646	65,5	0,651
C_{arom}	[5,6]	261	79,7	0,69
	[3,6]	266	79	0,68
	[4,6]	266	79	0,68
	[3,7]	368	72,8	0,64
N	[6,11]	646	67	0,646
	[7,11]	646	66,4	0,646
	[4,11]	646	66,7	0,643
	[5,11]	646	66,3	0,64
O_{as}	[7,10]	646	67	0,642
	[8,10]	646	66,7	0,642
	[7,11]	646	65,5	0,64
	[7,12]	646	65,3	0,64
O_m	[5,5.5]	165	67,9	0,3
	[3,5]	381	55,6	0,01
	[4,5]	337	56,7	0,04
	[6,6.5]	131	37,41	-0,42

3.2.1.15. Kombination von optimalen Radienintervallen für die Atomtypen

Im Kapitel 3.2.1.14 wurden die besten Wechselwirkungsbereiche für die Atomtypen mit den Aminosäuren gesucht. Um die Umgebung möglichst exakt zu beschreiben, wurden die Wechselwirkungspotentiale der Atomtypen zu einem Potential zusammengefasst. Der sich ergebende Kombinationsraum ist für eine systematische Suche zu groß (171 mögliche Intervalle für fünf Atomtypen ergeben 171^5 Kombinationen). Für die Zusammenfassung der Wechselwirkungsbeiträge wurden die Intervalle der fünf Atomtypen mit den vier vom Betrag her größten Korrelationskoeffizienten gewählt.

Tabelle 3.13: Die besten Ergebnisse für die Intervallkombination

Intervalle						
C_{ali}	C_{arom}	N	O_{as}	O_m	rv [%]	r_{cor}
8-12	3-6	6-11	8-10	3-5	65,79	0,655
8-12	3-6	6-11	8-10	4-5	65,79	0,655
8-12	3-7	6-11	8-10	3-5	65,79	0,655
8-12	3-7	6-11	8-10	4-5	65,79	0,655
8-12	4-6	6-11	8-10	3-5	65,79	0,655
8-12	4-6	6-11	8-10	4-5	65,79	0,655
8-12	3-6	7-11	8-10	3-5	65,48	0,655
8-12	3-7	7-11	8-10	3-5	65,48	0,655
8-12	4-6	7-11	8-10	3-5	65,48	0,655
8-12	5-6	7-11	8-10	3-5	65,48	0,655
8-12	3-6	7-11	7-10	3-5	65,33	0,655

Die beste erreichte Korrelation von r_{cor}=0,655 für die Kombination der idealen Intervalle liegt unter der bisher erreichten von r_{cor}=0,672 (s. Tabelle 3.14). Die richtige Vorhersage rv kann von 64,71 auf 65,79 gesteigert werden (s. Tabelle 3.13). Der Atomtyp O_m führt, trotz schlechter Vorhersageergebnisse und vom Betrag her deutlich niedrigeren Korrelationskoeffizienten von r_{cor}= -0,069 im Vergleich mit den anderen Atomtypen, für das Aminosäure-Atom-Potential zu einer Verbesserung der Korrelation zwischen berechneten und experimentellen Werten von r_{cor}=0,645 zu r_{cor}=0,672 (s. Tabelle 3.14).

Tabelle 3.14: Die Gesamtkorrelation des Entwicklungsdatensatzes für die einzelnen Atomtypen, (AT_{Gesamt}) und für alle Atomtypen ohne O_m mit dem Radienintervall [3,13]

Atomtyp	Datenpunkte	rv [%]	r_{cor}
O_m	529	39,89	-0,069
O_{as}	646	65,17	0,632
N	646	65,94	0,630
C_{arom}	624	66,67	0,548
C_{ali}	646	65,33	0,647
AT_{Gesamt} ohne O_m	646	65,33	0,645
AT_{Gesamt}	646	64,71	0,672

3.2.1.16. Hauptkomponentenanalyse des Gesamtwechselwirkungspotentials

Die Einzelwechselwirkungsenergien der Atomtypen sollen bei ihrer Kombination zum Gesamtwechselwirkungspotential gewichtet werden. Dadurch könnte eine Verbesserung der Vorhersage und der Korrelation erreicht werden. Die ist allerdings nur zulässig, wenn die zu gewichtenden Variabeln linear unabhängig voneinander sind. Die Hauptkomponentenanalyse HKA überführt voneinander abhängige Variabeln in unabhängige Faktoren. Diese Faktoren können dann zueinander gewichtet werden (s. Kapitel 2.3.9).

Für die HKA wurden die zwanzig Potentialkurven der Aminosäuren zu einer mittleren Kurve zusammengefasst. Für jeden Atomtyp wurde der Durchschnitt über die abstandsabhängigen Werte von 3 Å bis 13 Å gebildet.

Die Faktorladungen n_i ergeben nach Multiplikation mit den entsprechenden Potentialwerten die jeweiligen Faktoren. Diese werden als Linearkombination der Potentialkurven für die Stabilitätsvorhersage eingesetzt und für eine Optimumssuche gegeneinander gewichtet [Leven, 1999]. Die Korrelation zwischen den berechneten Stabilitätswerten und den experimentellen Werten des Entwicklungsdatensatzes wurde für jede mögliche Kombination der Gewichte –1, -0.5, 0, 0.5, 1 bestimmt (s. Tabelle 3.15).

Tabelle 3.15: Faktorladungen n_i

Faktor	n_{ali}	n_{arom}	n_n	n_{oas}	n_{om}	V [%]
a	0,00090035	-0,00018736	0,00019864	-0,000015771	0,000032878	8,97
b	0,00028726	0,00016753	-0,00010719	-0,00056318	-0,00010928	89,84
c	-0,00051192	0,00016423	0,00020780	-0,000015771	-0,00002931	1,12
d	0,00035148	0,00001.1063	-0,000084636	0,000027105	-0,00017684	0,03
e	0,00018245	0,0018898	-,0001926602	-0,0000082216	0,00010297	0,42

V ist die erklärte Varianz der Eigenwerte für die jeweiligen Eigenvektoren in Prozent

Bei der Verwendung aller Faktoren und einer in Tabelle 3.16 angezeigten Gewichtung konnte im Vergleich mit der Ausgangskonfiguration oder mit der Gleichgewichtung aller Faktoren eine leichte Verbesserung des Korrelationskoeffizienten erreicht werden, nämlich von r_{cor}=0,672 auf r_{cor}=0,677.

Tabelle 3.16: Tabellarische Übersicht über die Ergebnisse der Hauptkomponentenanalyse

	Faktoren Gewichtung					rv [%]	r_{cor}
	a	b	c	d	e		
Ohne Faktoren						64,71	0,672
mit allen Faktoren	1	1	1	1	1	65,2	0,632
mit allen Faktoren	0,5	-1	-1	-1	-1	64,4	0,677
Faktoren mit den beiden höchsten erklärten Varianzen	0,5	-1				65,8	0,664
Faktor mit der höchsten erklärten Varianz						64.86	0.322

3.2.2. Das wissensbasierte Torsionswinkelpotential

Die Häufigkeitsverteilungen des Torsionswinkelpaares wurden, wie in Kapitel 2.2.4 erläutert, aus dem Strukturdatensatz extrahiert. Die Berechnung der inversen Boltzmannenergien erfolgte Gleichung 2.22 entsprechend und analog zu Leven [Leven, 1999].

Die Torsionswinkelpotentiale wurden für die 20 Aminosäuren bestimmt und tabellarisch aufgeführt. Die Darstellung der Energieverteilungen für die 20 Aminosäuren und ihre Diskussion finden sich bei Leven [Leven, 1999].

Für eine Vorhersage der Thermostabilität kann der Energiewert des Torsionswinkelpaares der zu mutierenden Aminosäure aus dieser Potentialtabelle ausgelesen werden. Durch das Beibehalten der Proteinstruktur und der bekannten (φ,ψ)- Winkelkombination kann der Energiewert für die einzutauschende Aminosäure ermittelt werden. Die Differenz der Torsionswinkelenergien für den Wildtyp und die Mutante entspricht einer Änderung der freien Faltungsenthalpie. Wurde diese Vorhersagemethode auf den Entwicklungsdatensatz und seine experimentellen Werte zur Änderung der freien Faltungsenthalpie für entsprechende Mutationen angewendet, konnte eine Gesamtkorrelation von r_{cor}= 0,264 (646 Datenpunkte), eine richtige Vorhersage von 67,81 % und eine Sensibilität von 0,75 erzielt werden. Für die Korrelation kann ein signifikanter linearer Zusammenhang bei einer Irrtumswahrscheinlichkeit von 0,1 % festgestellt werden. Bei detaillierter Betrachtung der Ergebnisse für den Entwicklungsdatensatz auf Struktur- bzw. Proteinebene zeigt sich, dass die Korrelationen im Vergleich mit dem Aminosäure-Atom-Potential zwischen berechneten und experimentellen Werten mit einer Signifikanz des linearen Zusammenhangs bei einer Irrtumswahrscheinlichkeit von 0,1 % deutlich abnehmen. Nur die Korrelation für die Barnase 1rnb erreicht ähnlich gute Werte wie bei der Verwendung des Aminosäure-Atom-Potentials. Besser sieht es bei den richtigen Vorhersagen für stabilisierende und destabilisierende Mutationen aus. Bei Verwendung des Torsionswinkelpotentials kann die richtige Vorhersage für fünf von elf Proteinen des Entwicklungsdatensatzes, die T4-Lysozyme (1lyd, 2lzm und 3lzm) und die Barnase (1rnb) im Vergleich mit dem Aminosäure-Atom-Potential deutlich verbessert werden (s. Tabelle 3.17 und Tabelle 3.5).

Tabelle 3.17: Vorhersage-Ergebnisse mit dem Torsionswinkelpotential für die Proteine des Entwicklungsdatensatzes

Protein	n_{DDG}	rv [%]	r_{cor}
1bgs	60	58,33	-0,045
1l63	94	62,77	0,295
1lyd	78	76,62	0,338
1lz1	5	80,00	-0,425
1rnb	118	72,03	**0,348**
1stn	83	75,9	**0,360**
2ci2	79	67,09	0,079
2lzm	16	93,75	0,150
2wsy	18	77,78	0,184
3lzm	59	59,65	0,265
4lyz	36	41,67	-0,142
Summe:	646	67,81	0,264

Die mit einer Irrtumswahrscheinlichkeit von einem Prozent signifikant linear abhängigen Korrelationen sind **fett** wiedergegeben.

3.2.3. Die kombinierte wissensbasierte Aminosäure-Atom- und Torsionswinkel-Potentialfunktion

Die durch Kombination der Aminosäure-Atom-Potentialfunktion (CZ,CB,O; Intervalllänge 0,5 Å; Intervall [3,13]; Schrittweite 180°; Öffnungswinkel 90°) mit der Torsionswinkelpotentialfunktion erhaltene Gesamtbewertungsfunktion sollte zu einer besseren Gesamtbeschreibung der Umgebungswechselwirkung führen. Gilis und Rooman [Gilis und Rooman, 1997] konnten zeigen, dass die Vorhersageleistung eines Aminosäure-Aminosäure-Potentials durch die Kombination mit dem Torsionswinkelpotential verbessert wurde. Die Vorhersagequalität der Gesamtbewertungsfunktion wurde an Hand des Entwicklungsdatensatzes getestet.

Die Korrelation zwischen berechneten und experimentellen Werten und die Vorhersagequalität verbessern sich im Vergleich von Aminosäure-Atom-Potential und Torsionswinkelpotential mit der kombinierten Bewertungsfunktion deutlich. (s. Tabelle 3.18, Tabelle 3.17 und Tabelle 3.5).

Tabelle 3.18: Vorhersage-Ergebnisse für den Entwicklungsdatensatz mit der kombinierten Aminosäure-Atom- und Torsionswinkelpotentialfunktion

Protein	n_{DDG}	rv [%]	r_{cor}
1bgs	60	76,67	**0,478**
1l63	94	69,15	**0,805**
1lyd	78	67,95	**0,545**
1lz1	5	100	0,843
1rnb	118	74,58	**0,542**
1stn	83	96,39	**0,708**
2ci2	79	79,75	**0,568**
2lzm	16	50	-0,090
2wsy	18	88,89	0,521
3lzm	59	59,32	0,397
4lyz	36	61,11	0,415
Summe:	646	74,46	0,72

Unter der Vorraussetzung, dass beide Potentiale unabhängig voneinander sind, konnten ihre Energiewerte mit unterschiedlicher Gewichtung bewertet werden (s. Kapitel 2.2.5). Dabei wurde der Anteil des Aminosäure-Atom-Potentials in 5 % Schritten von 0 auf 100 % erhöht, während der Anteil des Torsionswinkelpotentials entsprechend verringert wurde (s. Tabelle 3.19).
Die beste Korrelation für die kombinierte Bewertungsfunktion ergibt sich aus einer Gleichgewichtung der beiden Potentiale. Die beste richtige Vorhersage wird mit einem Anteil des Aminosäure-Atom-Potentials von 25 % an der Gesamtbewertungsfunktion erhalten. Die Sensibilität des Aminosäure-Atom-Potentials konnte von 0,67 auf 0,75 für die Gesamtbewertungsfunktion gesteigert werden.

Eine weitere Steigerung der Sensibilität von 0,75 auf 0,82 erfolgt bei einem 15%- oder 35%-Anteil der Aminosäure-Atom-Potentialfunktion am Gesamtpotential.

Tabelle 3.19: Vorhersage-Ergebnisse für die anteilig gewichteten Potentialfunktionen

Anteil Aminosäure-Atom-Potential an der Gesamtbewertungsfunktion [%]	Spezifität	Sensibilität	rv [%]	r_{cor}
0	0,39	0,75	67,81	0,264
5	0,42	0,78	70,28	0,349
10	0,45	0,81	72,91	0,434
15	0,48	**0,82**	75,54	0,513
20	0,49	0,81	76,32	0,58
25	**0,50**	0,81	**76,93**	0,632
30	0,49	0,81	76,32	0,67
35	0,48	**0,82**	76,01	0,695
40	0,47	0,8	74,77	0,71
45	0,47	0,77	74,92	0,717
50	0,46	0,75	74,46	**0,72**
55	0,45	0,74	73,53	0,719
60	0,44	0,74	72,76	0,716
65	0,43	0,74	72,14	0,711
70	0,43	0,74	71,36	0,706
75	0,42	0,73	70,9	0,7
80	0,41	0,73	69,5	0,694
85	0,39	0,71	68,27	0,688
90	0,37	0,66	67,03	0,683
95	0,36	0,65	65,48	0,677
100	0,35	0,67	64,71	0,672

Die Standardabweichung $\sigma_{Entwicklungsdatensatz}$ der Differenz aus berechneten und experimentellen Werten des Entwicklungsdatensatzes für die gleichgewichteten Potentialfunktionen entspricht $\sigma_{Entwicklungsdatensatz}$ = 1,44 kcal/mol. Werden alle Mutationen ausgeschlossen, deren Differenz zwischen berechneten und experimentellen Werten größer als die dreifache Standardabweichung ist, ergibt sich eine Korrelation von r_{cor}= 0,76, eine richtige Vorhersage rv = 75,4 % und eine Standardabweichung von 1,00 kcal/mol für 612 Datenpunkte (dies entspricht

einem Datenausschluss von 5,2 %) (s. Abbildung 3.5 und Abbildung 3.6). Noch besser wird die Vorhersage, wenn Mutationen ausgeschlossen werden, deren Differenz zwischen berechneten und experimentellen Werten größer als die zweifache Standardabweichung ist. Dann ergibt sich eine Korrelation von r_{cor}= 0,81, eine richtige Vorhersage rv = 76 % und eine Standardabweichung von 0,77 kcal/mol für 561 Datenpunkte (dies entspricht einem Datenausschluss von 13,2 %) (s. Tabelle 3.21 und Tabelle 3.22).

Um für eine richtig stabilisierende Vorhersage die Sensibilität zu steigern, kann die Standardabweichung als einseitiges Kriterium dienen. Von allen als destabilisierend bewerteten Vorhersagewerten kleiner als die Standardabweichung wird $\sigma_{\text{Entwicklungsdatensatz}}$ = 1,44 kcal/mol subtrahiert. Es werden bei 646 Datenpunkten von 147 möglichen stabilisierenden Mutationen 141 als richtig erkannt und nur noch sechs als falsch destabilisierend (falsch negativ) bewertet (s. Tabelle 3.20). Die Sensibilität steigt auf 0,96, die Korrelation zwischen modifizierten berechneten Werten und experimentellen Werten beträgt r_{cor}=0,72 und für die richtige Vorhersage wird rv = 69,1 % erreicht (s. Kapitel 4.3).

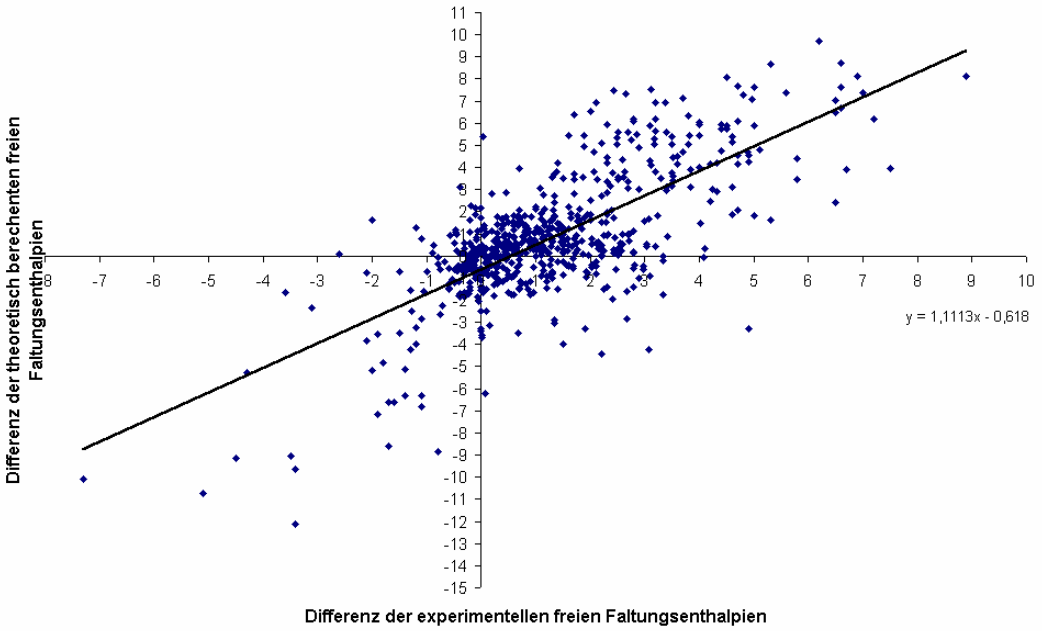


Abbildung 3.5: Auftragung der vorhergesagten Stabilitätsänderungen für das Gesamtpotential gegen die experimentellen Stabilitätsänderungen für den Entwicklungsdatensatz

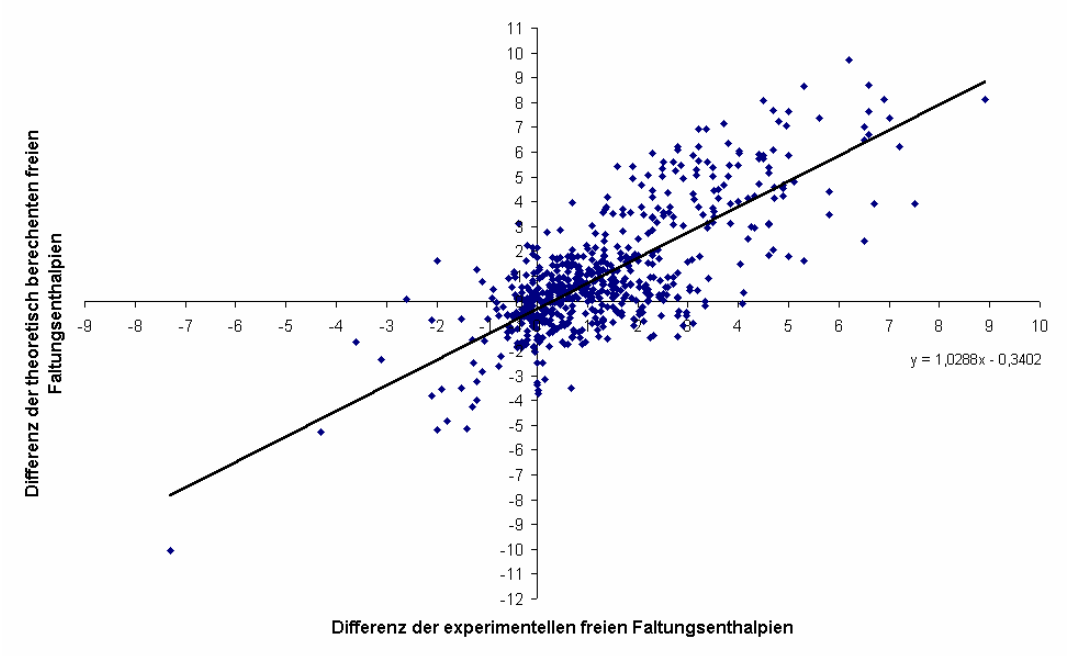


Abbildung 3.6: Auftragung der vorhergesagten Stabilitätsänderungen für das Gesamtpotential gegen die experimentellen Stabilitätsänderungen für den Entwicklungsdatensatz. Mutationen mit einem berechneten Wert > 3*s (s=1,44 kcal/mol) werden ausgeschlossen

Tabelle 3.20: Sechs mit der Gesamtbewertungsfunktion als falsch negativ bewertete Aminosäuremutationen

Protein	Wildtyp	AS-Nr.[a)]	Mutante	$\Delta\Delta G_{exp}$	$\Delta\Delta G_{theo}$	Solvenszugänglichkeit[b)]	Sekundär-Struktur[c)]
1lyd	GLY	30	ALA	-0,1	0,43	20 %<X<40 %	Turn
2ci2	LYS	37	ALA	-0,21	0,2	>40 %	Helix
2wsy	GLU	49	PRO	-2	0,19	<20 %	β-Faltblatt
3lzm	LYS	60	PRO	-0,1	0,71	>40 %	Helix
4lyz	PHE	34	TYR	-0,19	0,82	20 %<X<40 %	Helix
4lyz	GLY	102	ARG	-0,38	1,67	>40 %	Random

[a)] AS-Nr. entspricht der Aminosäurenummer aus der PDB-Datei

[b)] die Solvenszugänglichkeit wurde mit dem Programm *psa* bestimmt (s. Kapitel 2.2.8)

[c)] die Sekundärstruktur wurde mit dem Programm DSSP ermittelt (s. Kapitel 2.2.8)

3.2.4. Anwendung der optimierten wissensbasierten Bewertungsfunktion auf den Testdatensatz

Die in Kapitel 3.2.3 vorgestellte Bewertungsfunktion wurde in dieser Arbeit weiter untersucht. Dieses Potential sowie ihre Parameter wurden mit Hilfe des Entwicklungsdatensatzes ausgewählt, überprüft und optimiert. Um eine Überanpassung der Methoden oder ein Zufallsergebnis ausschließen zu können, wurde die entwickelte Bewertungsfunktion an einem Testdatensatz abschließend überprüft. Der Testdatensatz beinhaltet, wie in Kapitel 3.1.3 und Kapitel 2.2.6.2 vorgestellt, insgesamt 918 Stabilitätsdaten aus 27 Proteinen.

Die Vorhersage wurde für das Aminosäure-Atom-Potential, das Torsionswinkelpotential und die kombinierte Bewertungsfunktion wiederholt (s. Tabelle 3.21). Die Korrelation zwischen den vorhergesagten Stabilitätswerten und den experimentellen Werten zur Änderung der freien Faltungsenthalpie weist für alle Methoden einen auf dem 0,1 %-Irrtumsniveau signifikanten linearen Zusammenhang auf. Die auf diese Weise in dieser Arbeit bestimmten Korrelationen für die verschiedenen Methoden sind durchgehend für den Testdatensatz schlechter als für den Entwicklungsdatensatz. Die Kombination des Aminosäure-Atom-Potentials mit dem Torsionswinkelpotential führt auch für den Testdatensatz zu einer verbesserten Korrelation und Vorhersage. Die beste richtige Vorhersage rv in Bezug auf den Entwicklungsdatensatz findet sich für das Torsionswinkelpotential. Dieses Ergebnis findet sich im Testdatensatz nicht wieder (s. Tabelle 3.21).

Werden alle Mutationen für den Testdatensatz verworfen, deren Differenz zwischen vorhergesagten und experimentellen Werten eine aus dem Entwicklungsdatensatz errechnete dreifache Standardabweichung übertreffen ($\sigma_{Entwicklungsdatensatz}$ =1,44 kcal/mol) (s. Kapitel 3.2.3), steigt die Korrelation auf r_{cor}=0,64, die richtige Vorhersage steigt auf rv=70 % und die Standardabweichung entspricht σ= 1,06 kcal/mol. Es wurden 67 Datenpunkte ausgeschlossen, d.h. 7,2 % einer Gesamtmenge von 918 Datenpunkten (s. Abbildung 3.7 und Tabelle 3.22). Dieses Kriterium des Datenausschlusses bewährt sich also auch beim Testdatensatz.

Tabelle 3.21: Übersicht über die Vorhersage-Ergebnisse der unterschiedlichen Potentiale für Entwicklungs- und Testdatensatz. Angegeben sind der Korrelationskoeffizient r_{cor}, die richtig vorhergesagten rv Mutationen sowie die Anzahl aller betrachteten Mutationen (in Klammern)

Methode	Testdatensatz r_{cor}	Testdatensatz rv [%]	Entwicklungs-datensatz r_{cor}	Entwicklungs-datensatz rv [%]
Aminosäure-Atom-Potential	0,415 (918)	62,85	0,670 (646)	64,71
Torsionswinkelpotential	0,161 (918)	60,9	0,264 (646)	67,81
Kombiniertes Aminosäure-Atom- und Torsions-winkelpotential	0,462 (918)	69,39	0,720 (646)	74,46

Tabelle 3.22: Übersicht der Ergebnisse für die Gesamtbewertungsfunktion Angegeben sind der Korrelationskoeffizient r_{cor}, die richtig vorhergesagten rv Mutationen sowie die Anzahl aller betrachteten Mutationen (in Klammern)

	Testdatensatz		Entwicklungsdatensatz	
Methode	r_{cor}	rv [%] (s [kcal/mol])	r_{cor}	rv [%] (s [kcal/mol])
Gesamtbewertungsfunktion[a)]	0,46 (918)	69,4 (1,67)	0,72 (646)	74,5 (1,44)
Gesamtbewertungsfunktion mit Datenausschluss > 3*σ[b)]	0,64 (851)	70 (1,06)	0,76 (612)	75 (1,00)
Gesamtbewertungsfunktion mit Datenausschluss > 2*σ[b)]	0,74 (747)	71 (0,75)	0,81 (561)	76 (0,77)

[a)] Gesamtbewertungsfunktion entspricht dem gewählten kombinierten Aminosäure-Atom- und Torsionswinkelpotential

[b)] σ = 1,44 kcal/mol

Werden nun Mutationen ausgeschlossen, deren Differenz zwischen berechneten und experimentellen Werten im Betrag größer als die zweifache Standardabweichung ist, wird eine Korrelation von r_{cor}=0,74, eine richtige Vorhersage von rv=71,4 % und eine Standardabweichung von σ = 0,75 kcal/mol erreicht. Es wurden 171 Datenpunkte ausgeschlossen, d.h. 19 % der Gesamtmenge von 918 Stabilitätsdaten.

Wird von jedem berechneten Stabilitätswert, entsprechend $\sigma_{Entwicklungsdatensatz}$ =1,44 kcal/mol abgezogen, erhöht sich die Sensibilität für die 918 Stabilitätswerte des Testdatensatzes von 0,67 auf 0,93 (22 von 312 stabilisierenden Mutationen werden als destabilisierend vorhergesagt) (s. Kapitel 4.3).

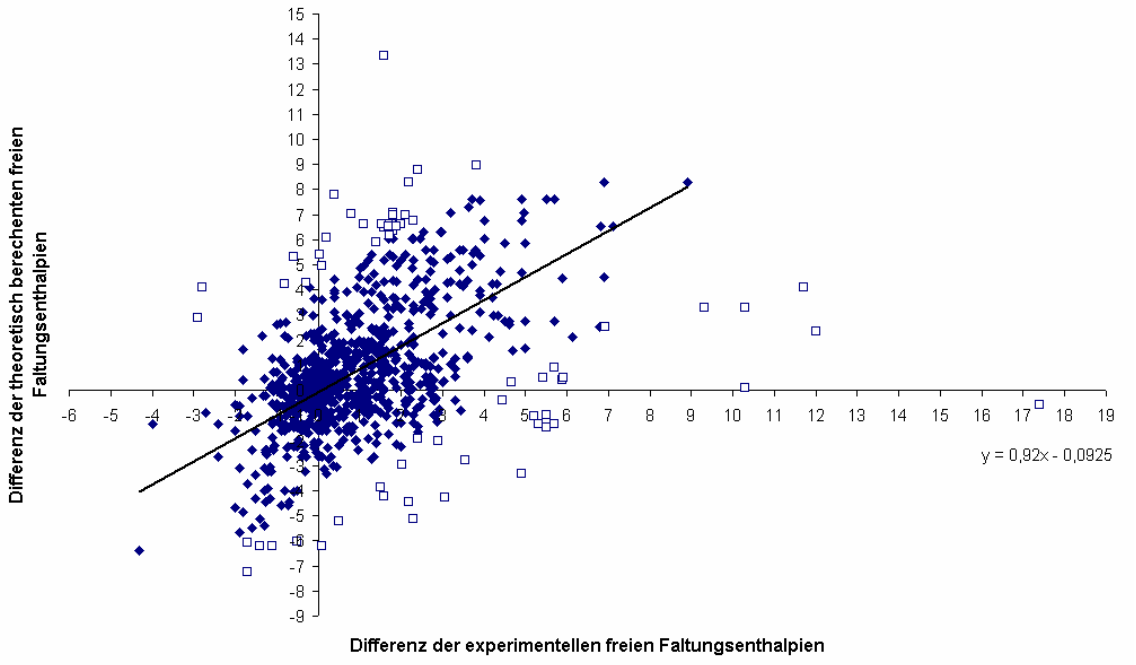


Abbildung 3.7: Auftragung der vorhergesagten Stabilitätsänderungen für das Gesamtpotential gegen die experimentellen Stabilitätsänderungen für den Testdatensatz. Mutationen mit einem berechneten Wert > 3*s (s=1,44 kcal/mol) werden ausgeschlossen und sind als leere Quadrate dargestellt

3.2.5. Charakterisierung der Ergebnisse in Abhängigkeit von Strukturcharakteristika und von der Art der Mutation

Die Wirkung eines Aminosäureaustausches auf die Stabilität eines Proteins hängt von der Art der auszutauschenden und/oder ausgetauschten Aminosäure sowie ihrer Umgebung ab. Der Einfluss der Umgebung auf eine Mutation im hydrophoben Kern ist ein anderer als an der polaren Oberfläche eines Proteins. Deshalb können daraus für einen gleichgearteten Austausch andere Auswir-

kungen für die Proteinstabilität resultieren. Die Proteine des Entwicklungsdatensatzes und des Testdatensatzes liegen in einer gelösten Kristallstruktur vor. Mit diesen Strukturen lassen sich die Orte der experimentell durchgeführten Mutationen beschreiben. Für eine solche Beschreibung bieten sich die Lösungsmittelzugänglichkeiten (SAS) an. Hohe Lösungsmittelzugänglichkeiten charakterisieren die Lage einer Aminosäure an der Oberfläche eines Proteins, während niedrige Lösungsmittelzugänglichkeiten ein Indiz sind für die Lokalisierung der Aminosäure im Proteininneren. Weitere wichtige, die Umgebung beschreibende Elemente sind die Sekundärstrukturanordnungen um die Mutationsstelle. Der Austausch einer Aminosäure in einer Helix oder einem Faltblatt kann zu einer Zerstörung dieser Struktur und damit zu einer Destabilisierung führen. Eng mit der Beschreibung des Umgebungseinflusses ist die Betrachtung der Art der ausgetauschten und/oder auszutauschenden Aminosäure verknüpft.

Die von den Potentialen getroffenen Vorhersageergebnisse werden im Folgenden weiter in Bezug auf die Strukturcharakteristika und die Art der Mutation analysiert.

3.2.5.1. Abhängigkeit der Vorhersage von der Lösungsmittelzugänglichkeit

Die Lösungsmittelzugänglichkeit ist ein Maß für die relative räumliche Lage einer Aminosäure in einem Protein. Die Oberfläche eines Proteins wird mit einer Lösungsmittelkugel abgetastet und der Anteil der Oberfläche einer Aminosäure in Kontakt mit der Lösungsmittelkugel kann in $Å^2$ bestimmt werden. Dieser Wert wird mit der maximal zugänglichen Oberfläche der betrachteten Aminosäure AS in einem Tripeptid Gly-AS-Gly in Beziehung gesetzt. Daraus ergibt sich eine relative Lösungsmittelzugänglichkeit in % [Chothia, 1984]. Die Bestimmung der SAS erfolgte mit dem Programm *psa* (s. Kapitel 2.2.8).

Für die Beschreibung der Lage der Mutationen im Entwicklungs- und Testdatensatz wurde die Lösungsmittelzugänglichkeit in drei Intervalle (<20 %, 20 %< AS <40 % und >40%) geteilt [Gilis und Rooman, 1997]. Für jedes Intervall wur-

den die Mutationen bestimmt und die Korrelation zwischen vorhergesagten und experimentellen Stabilitätswerten für das Aminosäure-Atom-Potential, das Torsionswinkelpotential und das kombinierte Aminosäure-Atom- und Torsionswinkelpotential (Gesamtbewertungsfunktion) bestimmt (s. Tabelle 3.23).

Das Aminosäure-Atom-Potential zeigt für den Entwicklungs- und Testdatensatz die höchste Korrelation r_{cor}= 0,73 für eine Lösungsmittelzugänglichkeit von < 20 %, die mit Zunahme der Lösungsmittelzugänglichkeit abnimmt. Das Torsionswinkelpotential weist die schlechteste Korrelation r_{cor}=0,11 für den Entwicklungsdatensatz bei einer Lösungsmittelzugänglichkeit von 20 %<AS<40 % auf.

Tabelle 3.23: Korrelationskoeffizienten für alle Potentialfunktionen in Abhängigkeit von der Lösungsmittelzugänglichkeit

	Entwicklungsdatensatz			**Testdatensatz**		
Methode	**<20 %**	**20 %<AS<40 %**	**> 40%**	**<20 %**	**20 %<AS<40 %**	**> 40%**
	r_{cor}	**r_{cor}**	**r_{cor}**	**r_{cor}**	**r_{cor}**	**r_{cor}**
Aminosäure-Atom-Potential	**0,73**	**0,54**	**0,22**	**0,52**	0,18	0,1
Torsionswinkel-potential	**0,34**	0,11	**0,52**	0,15	0,23	**0,18**
Gesamtbewertungs-funktion	**0,76**	**0,56**	**0,54**	**0,58**	**0,29**	**0,19**
Datenpunkte	239	128	279	401	168	349

Die mit einer Irrtumswahrscheinlichkeit von 0,1 % signifikant linear abhängigen Korrelationen sind **fett** wiedergegeben.

Dieses Ergebnis kehrt sich im Testdatensatz um, wo dieses Zugänglichkeitsintervall die beste Korrelation von r_{cor}=0,23 aufweist. Die Vorhersagequalität wird bei der Kombination beider Potentiale für beide Datensätze durchgehend besser, der beste Korrelationskoeffizient von r_{cor}=0,58 wird für eine Lösungsmittelzugänglichkeit von <20 % erreicht und sinkt, analog zu dem Aminosäure-Atom-Potential, entsprechend mit einer Zunahme der Solvenszugänglichkeit. Für die Gesamtbewertungsfunktion lässt sich ein signifikanter linearer Zusammenhang mit einer Irrtumswahrscheinlichkeit von 0,1 % zwischen theoretischen und ex-

perimentellen Daten für alle Zugänglichkeitsintervalle und Datensätze feststellen.

3.2.5.2. Abhängigkeit der Vorhersage von der Lage der Mutation in Sekundärstrukturelementen

Für die vorliegendenden Mutationsdaten wurde untersucht, ob sie Teil eines bestimmten Strukturelementes sind. Für die Sekundärstrukturvorhersage wurde das Programm DSSP verwendet (s. Kapitel 2.2.8). Die Mutationen wurden in vier Strukturklassen unterteilt und der jeweilige Korrelationskoeffizient bestimmt. Die Strukturklassen entsprechen einer Zugehörigkeit zu Helices (α-Helices oder 3-10 Helices), β-Faltblättern, Turns (Krümmung > 70°, wasserstoffverbrückte Turns und isolierte β-Brücken) und zu Random (entspricht einer Nichtzuweisung durch DSSP) (s. Tabelle 3.24 und Tabelle 3.8).

Tabelle 3.24: Vorhersage-Ergebnisse der Potentiale für den Entwicklungsdatensatz in Abhängigkeit von der Sekundärstrukturzugehörigkeit der Mutationen

Methode	Helices	b-Faltblätter	Turns	Random
	r_{cor}	r_{cor}	r_{cor}	r_{cor}
Aminosäure-Atom-Potential	0,69	0,79	0,37	0,11
Torsionswinkelpotential	0,26	0,45	0,30	0,39
Gesamtbewertungsfunktion	0,72	0,83	0,51	0,27
Datenpunkte	296	119	135	96

Die betrachteten Potentiale weisen Präferenzen für unterschiedliche Sekundärstrukturelemente auf. Das Aminosäure-Atom-Potential liefert gute Ergebnisse für Helices und β-Faltblätter. Die Vorhersagegüte bricht aber bei Turns und *Random* Residuen, bei denen keine Zuordnung zu einem Sekundärstrukturelement vorliegt, ein.

Die Korrelation zwischen theoretischen und experimentellen Werten für die Sekundärstrukturzuordnung der β-Faltblätter sinkt bei Verwendung des Testdaten-

satzes anstelle des Entwicklungsdatensatzes zum Teil deutlich (s. Tabelle 3.25). Das Torsionswinkelpotential liefert in dieser Untersuchung generell schlechtere Korrelationswerte als das Aminosäure-Atom-Potential, eine Ausnahme stellen die Vorhersagen für Mutationen in nicht zugeordneten Sekundärstrukturelementen dar. Hier kann im Entwicklungsdatensatz eine Verbesserung der Vorhersage erzielt werden. Die Kombination des Aminosäure-Atom-Potentials mit dem Torsionswinkelpotential zeigt eine höhere Korrelation zwischen theoretischen und experimentellen Stabilitätsdaten als bei den Einzelpotentialen. Für Mutationen, die als *Random* klassifiziert wurden, nähert sich die Vorhersage für die Gesamtbewertungsfunktion (r_{cor}=0,27) der des Torsionswinkelpotentials (r_{cor}=0,39) an (s. Tabelle 3.24).

Tabelle 3.25: Vorhersage-Ergebnisse der Potentiale für den Testdatensatz in Abhängigkeit von der Sekundärstrukturzugehörigkeit der Mutationen

Methode	Helices	b-Faltblätter	Turns	Random
	r_{cor}	r_{cor}	r_{cor}	r_{cor}
Aminosäure-Atom-Potential	0,63	0,35	0,13	0,01
Torsionswinkelpotential	0,24	0,13	0,26	0,12
Gesamtbewertungsfunktion	0,67	0,38	0,31	0,07
Datenpunkte	485	110	189	134

Im Testdatensatz zeigen das Aminosäure-Atom-Potential und das Torsionswinkelpotential eine deutlich niedrigere Vorhersagegüte für Mutationen in β-Faltblättern als im Entwicklungsdatensatz (s. Tabelle 3.24 und Tabelle 3.25). Für Residuen, die keinem Strukturelement zugeordnet wurden, konnten im Testdatensatz mit keinem Potential ein signifikanter linearer Zusammenhang zwischen errechneten Werten und experimentellen Stabilitätsdaten gefunden werden.

3.2.5.3. Abhängigkeit der Vorhersage von der mutierten Aminosäure

Es soll der Einfluss einer Aminosäure auf die Vorhersage vor ihrem Austausch untersucht werden. Die Potentialfunktionen werden hinsichtlich ihrer Fähigkeit,

die Umgebung einer Aminosäure vor der Mutation zu beschreiben, charakterisiert. Der Einfluss einer Aminosäure auf ihre Umgebung ist nicht klar zu definieren, er ist von genereller Natur, so sind hydrophobe Aminosäuren oft von anderen hydrophoben Aminosäuren umgeben, weniger häufig von polaren oder geladenen Aminosäuren.

Die Datensätze wurden in 20 Segmente, den 20 Aminosäuren entsprechend, zerlegt und die jeweiligen Korrelationen zwischen den von den drei Potentialen berechneten Werten und experimentellen Stabilitätsdaten bestimmt (s. Tabelle 3.25)[Leven, 1999].

Das Aminosäure-Atom-Potential zeigt im Entwicklungsdatensatz eine Präferenz für hydrophobe oder aromatische Aminosäuren (Tyrosin, Valin, Phenylalanin, Methionin, Leucin). Diese zeigen eine hohe Korrelation ($r_{cor} \geq 0{,}75$ bei mehr als 15 Datenpunkten, Methionin mit 14 Datenpunkten) zwischen theoretischen und experimentellen Werten. Die Korrelationswerte für polare oder geladene Aminosäuren sind niedriger, teilweise kann kein linearer Zusammenhang nachgewiesen werden. Eine Ausnahme stellt die Glutaminsäure mit einem Korrelationskoeffizienten von $r_{cor}= 0{,}72$ (36 Datenpunkte) dar. Diese Präferenz für hydrophobe Aminosäuren schwächt sich für den Testdatensatz ab.

Für das Torsionswinkelpotential sind die Korrelationen durchgehend geringer, teilweise kann kein signifikant linearer Zusammenhang zwischen den betrachteten Werten festgestellt werden. Im Entwicklungsdatensatz stellt Glycin eine Ausnahme mit einem Korrelationskoeffizienten von $r_{cor}=0{,}84$ dar. Diese Sonderstellung geht im Testdatensatz verloren.

Durch die Kombination des Aminosäure-Atom-Potentials mit dem Torsionswinkelpotential zu der Gesamtbewertungsfunktion verbessern sich die Korrelationskoeffizienten durchgehend. Für den Entwicklungsdatensatz auf teilweise hohem Niveau, für den Testdatensatz auf deutlich niedrigerem Niveau, z.B. ergeben sich für Valin die Werte r_{cor}(Entwicklungsdatensatz)=0,79, r_{cor}(Testdatensatz)=0,39.

Tabelle 3.26: Vorhersagegenauigkeit der Potentiale auf den Entwicklungs- und Testdatensatz in Abhängigkeit von der auszutauschenden Aminosäure

	Entwicklungsdatensatz				Testdatensatz			
Aminosäure	Potential A	Potential B	Potential C		Potential A	Potential B	Potential C	
	r_{cor}	r_{cor}	r_{cor}	n	r_{cor}	r_{cor}	r_{cor}	n
Alanin	0,07	0,26	0,10	38	0,57	0,2	0,59	71
Arginin	-0,23	0,34	0,04	14	-0,02	-0,14	-0,08	31
Asparagin	0,15	0,58	0,55	31	0,23	0,67	0,62	45
Aspartat	0,04	0,43	0,28	30	0,31	0,23	0,34	92
Cystein					0,03	0,22	0,07	16
Glutamat	0,72	0,48	0,76	36	0,28	-0,06	0,28	25
Glutamin	0,15	0,51	0,40	19	0,42	0,48	0,54	21
Glycin	0,00	0,84	0,73	20	-0,03	0,23	0,22	52
Histidin	0,76	0,33	0,77	10	0,38	0,19	0,40	14
Isoleucin	0,69	0,14	0,69	88	0,59	-0,08	0,58	103
Leucin	0,75	0,3	0,82	52	0,77	-0,23	0,79	67
Lysin	-0,21	0,32	0,04	26	-0,13	0,49	0,23	38
Methionin	0,90	0,12	0,89	14	0,75	0,26	0,79	56
Phenylalanin	0,78	0,46	0,78	22	0,62	0,13	0,67	30
Prolin	-0,77	0,23	-0,8	4	0,02	0,19	0,14	26
Serin	0,37	0,32	0,43	58	0,36	0,36	0,45	57
Threonin	-0,31	0,45	0,03	90	-0,37	0,55	0,06	73
Tryptophan					0,00	0,00	0,00	2
Tyrosin	0,75	0,31	0,74	26	0,63	0,66	0,67	20
Valin	0,79	0,41	0,79	68	0,26	0,52	0,39	79

n ist die Anzahl der betrachteten Datenpunkte

Potential A: Aminosäure-Atom-Potential

Potential B: Torsionswinkelpotential

Potential C: Gesamtbewertungsfunktion

3.2.5.4. Abhängigkeit der Vorhersageleistung von der eingefügten Aminosäure

Durch den Einbau einer Aminosäure in eine Proteinstruktur kann sich diese in einer Weise ändern, die mit der in dieser Arbeit vorgestellten Potentialen nicht behandelt werden kann (Prinzip der unveränderten Proteinstruktur).

Die Datensätze wurden wie in Kapitel 3.2.5.3 in 20 Segmente, entsprechend den 20 Aminosäuren, zerlegt und die jeweiligen Korrelationen zwischen den von den drei Potentialen berechneten Werten und den experimentellen Stabilitätsdaten bestimmt (s. Tabelle 3.27).

Die Vorhersagen der Potentiale für den Entwicklungsdatensatz in Abhängigkeit von der einzubauenden Aminosäure zeigen insgesamt gute Ergebnisse (bis auf Prolin, Tryptophan und Tyrosin) und größtenteils einen signifikant linearen Zusammenhang zwischen berechneten und experimentellen Stabilitätswerten. Die Kombination des Aminosäure-Atom-Potentials mit dem Torsionswinkelpotential führt zu einer verbesserten Vorhersage für den Entwicklungsdatensatz.

Die durchgehend guten Vorhersagen brechen bis auf Aspartat, Glutamat und Threonin im Testdatensatz ein. Gute Korrelationswerte, die im Entwicklungsdatensatz erzielt wurden, sind im Testdatensatz niedriger oder nicht mehr signifikant. Die Gesamtbewertungsfunktion weist für die Mehrzahl der Aminosäuren eine bessere Vorhersage auf als die einzelnen Potentiale.

Tabelle 3.27: Vorhersagegenauigkeit der Potentiale für den Entwicklungs- und Testdatensatz in Abhängigkeit von der eingesetzten Aminosäure.

	Entwicklungsdatensatz				Testdatensatz			
Aminosäure	Potential A	Potential B	Potential C		Potential A	Potential B	Potential C	
	r_{cor}	r_{cor}	r_{cor}	n	r_{cor}	r_{cor}	r_{cor}	n
Alanin	0,64	0,16	0,69	185	0,51	0,14	0,55	260
Arginin	0,25	0,69	0,70	13	0,53	0,55	0,81	20
Asparagin	0,14	0,48	0,50	24	0,09	0,33	0,25	32
Aspartat	0,43	0,52	0,65	36	0,79	0,16	0,77	44
Cystein	0,71	0,66	0,74	8	0,12	-0,26	0,06	9
Glutamat	0,49	0,43	0,61	30	0,62	0,17	0,68	52
Glutamin	0,38	0,66	0,74	14	0,25	-0,17	0,17	10
Glycin	0,78	0,23	0,80	85	0,29	0,48	0,51	47
Histidin	0,65	0,55	0,67	10	0,68	0,1	0,14	31
Isoleucin	0,64	0,57	0,72	22	0,44	0,41	0,51	40
Leucin	0,78	0,13	0,82	21	0,08	0,48	0,21	56
Lysin	0,86	0,21	0,93	12	0,65	-0,09	0,66	27
Methionin	0,87	0,25	0,89	13	0,21	-0,04	0,19	36
Phenylalanin	0,67	0,54	0,77	23	0,39	-0,16	0,36	38
Prolin	0,29	-0,34	0,16	14	0,52	0,16	0,48	42
Serin	0,34	0,45	0,54	34	0,05	-0,13	0,01	43
Threonin	0,74	0,29	0,82	24	0,68	0,29	0,69	31
Tryptophan	0,69	0,19	0,71	5	-0,11	-0,15	-0,13	6
Tyrosin	0,43	0,35	0,43	9	-0,42	-0,25	-0,47	13
Valin	0,25	0,16	0,31	64	0,24	0,11	0,31	81

n ist die Anzahl der betrachteten Datenpunkte

Potential A: Aminosäure-Atom-Potential

Potential B: Torsionswinkelpotential

Potential C: Gesamtbewertungsfunktion

3.2.5.5. Abhängigkeit der Vorhersageleistung von dem betrachteten Protein

Die Vorhersageleistung soll in Abhängigkeit von dem betrachteten Protein untersucht werden. Die Ergebnisse sind in Tabelle 3.28 und Tabelle 3.29 dargestellt. Im Entwicklungsdatensatz zeigen die Korrelationskoeffizienten für die T4-

Lysozyme 2lzm, 3lzm und 4lyz auf einem 0,1 % Irrtumsniveau keinen linearen Zusammenhang zwischen theoretischen und experimentellen Stabilitätswerten. Werden diese drei Proteine aus der Betrachtung ausgeschlossen, ergibt sich für die Gesamtbewertungsfunktion ein Korrelationskoeffizient von r_{cor}=0,75 (für 535 Datenpunkte) und für die richtige Vorhersage rv= 77,76 %.

Tabelle 3.28: Vorhersage-Ergebnisse für die Gesamtbewertungsfunktion, nach Proteinen des Entwicklungsdatensatzes differenziert

Protein	n_{DDG}	rv [%]	r_{cor}
1bgs	60	76,67	0,48
1l63	94	69,15	0,80
1lyd	78	67,95	0,55
1lz1	5	100	0,84
1rnb	118	74,58	0,54
1stn	83	96,39	0,71
2ci2	79	79,75	0,57
2lzm	16	50,00	-0,09
2wsy	18	88,89	0,52
3lzm	59	59,32	0,39
4lyz	36	61,11	0,42

Während die richtigen Vorhersagen für den Testdatensatz, bis auf die Proteine 3gap und 1dyj, ein gleichmäßiges Bild zeigen, schwanken die Korrelationskoeffizienten zwischen den einzelnen Proteinen stark. Größtenteils weisen die Proteine mit einem Korrelationskoeffizienten kleiner als 0,3 nur geringe Datenmengen n mit n<25 auf.

Tabelle 3.29: Vorhersage-Ergebnisse für die Gesamtbewertungsfunktion, nach Proteinen des Testdatensatzes differenziert

Protein	n_{DDG}	rv [%]	r_{cor}
1abm	1	100,00	0,00
1ank	4	75,00	0,63
1bni	21	80,95	0,53
1bpi	51	80,39	0,19
1csp	4	75,00	0,88
1cyo	3	66,67	0,15
3gap	2	0,00	-1,00
1lhm	14	92,86	0,41
1lz1	33	78,79	0,71
1mbn	38	73,68	0,54
1myl	3	33,33	1,00
1rn1	16	62,50	0,36
1rop	21	57,14	0,25
1rtb	11	81,82	0,69
1sar	3	100,00	0,26
1stn	12	100,00	0,89
1sup	6	50,00	-0,12
1tyu	7	57,14	-0,12
1ycc	13	61,54	0,11
2ci2	12	91,67	0,81
2lzm	405	64,44	0,64
2rn2	122	77,87	0,55
2trx	2	100,00	-1,00
2wsy	16	75,00	0,27
3ssi	50	74,00	0,64
1dyj	10	0,00	0,37
4lyz	38	60,53	0,41

4. Diskussion

Ziel dieser Arbeit war es, einen Ansatz zur automatischen Erstellung eines Mutationsprofils für ein zu thermostabilisierendes Enzym zu erstellen. Es sollte ein wissensbasiertes Potential entwickelt werden, welches die Umgebung einer Aminosäure quantifizieren und die Wirkung eines Aminosäureaustausches bewerten kann. Dazu wurde eine richtungs- und abstandsabhängige wissensbasierte Aminosäure-Atom-Potentialfunktion (s. Kapitel 4.1.1) beschrieben, die mit einem wissensbasierten Torsionswinkelpotential (s. Kapitel 4.1.1.4) zu einer Bewertungsfunktion kombiniert wurde.
Diese Vorhersagefunktionen wurden an einem durch Literaturrecherche möglichst großen Satz an experimentellen Daten validiert und optimiert und einer statistische Auswertung unterworfen (s. Kapitel 4.2).
Zu diskutieren bleiben die Methodenentwicklung und die Ergebnisse zur Vorhersage der Thermostabilität sowie deren Analyse.

4.1. Methodenentwicklung

4.1.1. Das Aminosäure-Atom-Potential

4.1.1.1. Füllungsgrad der erfassten Schalenelemente

Die wissensbasierte Bewertungsfunktion wird aus der Häufigkeitsverteilung der ausgewählten Atomtypen in der in Kugelschalen differenzierten Aminosäure-Umgebung abgeleitet. Diese Umgebung enthält implizit alle für eine Faltungsstabilisierung verantwortlichen Kräfte. Da die Wechselwirkungen von Aminosäuren eines Proteins mit dem Lösungsmittel notwendig für die Proteinfaltung sowie für die Stabilität der Proteinstruktur sind (s. Kapitel 1.2) [Kang *et al.*, 1987; Petukhov *et al.*, 1999], wurde eine theoretisch berechnete Wasserhülle um und in das Protein modelliert. Die Qualität der Raumfüllung durch Wasser soll anhand des Füllungsgrades abgeschätzt werden.

Ein Problem bei der Verwendung kubischer Modelle für die Raumfüllung mit Wasser entsteht durch die weiche und kurvige Proteinoberfläche, die mit diskreten Kuben angenähert wird [Kang *et al.*, 1987]. Je feiner das Gitter gewählt wird, desto genauer kann die wirkliche Oberfläche angenähert werden. In dieser Arbeit wurde eine Gitterkonstante von 0,6 Å gewählt, die einen Kompromiss zwischen der Genauigkeit der Oberflächen-Beschreibung und der Rechenzeit darstellen soll.

Ein weiteres Problem bei der Berechnung des Füllungsgrades stellen die gewählten Werte der van der Waals-Radien für die Atomtypen dar. Dabei fällt in eine beliebige Atomtypklasse eine Vielzahl von unterschiedlichen Atomen oder Atomgruppen, z.B. umfasst die Klasse der aliphatischen Kohlenstoffatome gleichermaßen Methyl-, Methylen- und Methingruppen. Die experimentell gemessenen van der Waals-Radien für die Klasse der aliphatischen Kohlenstoffatome liegen zwischen 1,61 Å und 2,00 Å und schwanken für die Gruppen in dieser Klasse in einer vergleichbaren Größenordnung [Tsai *et al.*, 1999; Rossmann und Arnold, 2001]. Für die Berechnung des Füllungsgrades, bzw. dem benötigten Volumen der gefundenen Atome in der entsprechenden Schale, wurde als van der Waals-Radius der aliphatischen Kohlenstoffatome 1,9 Å festgelegt. Insgesamt wurden für die Atomtypen mittlere Radien entsprechend den vom Programm *psa* (s. Kapitel 2.2.8) und von Richards und Lee [Lee und Richards, 1971; Richards, 1974; Rossmann und Arnold, 2001] angegebenen und benutzten mittleren Atomradien gewählt (r_{cali}=1,9 Å, r_{carom}=1,9 Å, r_n=1,7 Å, r_{oas}=1,4 Å, r_{om}=1,86 Å). Diese Atomradien sind gemittelte Radien über unterschiedliche Atomgruppen sowie deren Ladungs- oder Protonierungszustände.

Weitere Ungenauigkeiten in der Berechnung des Füllungsgrades können in der Strukturbestimmung von Proteinen (z.B. durch Kryobedingungen) oder in Unsicherheiten bei der Strukturlösung liegen.

Die Füllung der Schalen durch das modellierte Wasser für die mittlere Aminosäure (s. Kapitel 3.2.1.8 und Abbildung 3.2) ist zufriedenstellend. Bei einem Radius von 5 Å findet sich die zweite Koordinationssphäre eines betrachteten Zentralatoms. Die daraus resultierende dichteste Packung von Atomen wird durch die gewählten mittleren Radien, vor allem von aliphatischen Kohlenstoff-

atomen (mit einem Anteil von 70 % an der Gesamtfüllung) und bei Stickstoff (Anteil von 15 % an der Gesamtfüllung), falsch beschrieben (Gesamtfüllung ca. 120 %, d.h. das berechnete Schalenvolumen ist kleiner als das berechnete Volumen für die entsprechend gezählten Atomtypen). Der Anteil von Wasser an der Gesamtfüllung in diesem Strukturbereich ist kleiner als 10 %.

Mit steigendem Radius nimmt der Anteil des modelliertem Wassers zu. Zwischen 6,5 Å und 14 Å lässt sich eine Unterbesetzung der Umgebungsschalen beobachten, die auf die ungenaue Beschreibung der glatten und weichen Proteinoberfläche durch Kuben zurückzuführen ist. Die Überbesetzung der Schalen ab 18 Å kann durch die grobe Näherung des Wassermolekülvolumens von 30 $Å^3$ und durch die mittleren Atomradien für strukturellen Sauerstoff und Stickstoff erklärt werden. Das Wasser wird lose Wasserstoffbrückennetzwerke innerhalb der Wassermoleküle aufbauen und nicht eng und dicht gepackt vorliegen. Die Carbonylgruppen werden zum Teil geladen und die Aminogruppen zum Teil protoniert vorliegen. Die sich dadurch ergebenden Unterschiede in den Atomradien (geladene Gruppen zu ungeladenen Gruppen, protonierte Gruppen zu nicht protonierten Gruppen) werden in der Berechnung nicht berücksichtigt.

Für die Bestimmung der Umgebungswechselwirkungen werden nur Atome gezählt, die nicht zu der betrachteten Aminosäure gehören. Daraus und aus der Festlegung der van der Waals-Radien der Aminosäuren lassen sich die schlechten Ausfüllungen der Umgebungsschalen für Radien kleiner als 4,5 Å erklären. Für Glycin ist dieser Bereich mit ungefähr 80 % besser gefüllt als für Tryptophan mit 40 % (s. Abbildung 3.3).

Insgesamt liegt die mittlere Füllung für die mittlere Aminosäure bei 99,3 ± 7,3 %. Das in dieser Arbeit verwendete Wassermodell kann somit für die betrachteten Radienintervalle von 3 Å bis 13 Å als hinreichend gut bezeichnet werden.

4.1.1.2. Unterschiede in den Schalenbesetzungen zwischen Aminosäure-Atomtyp-Kombinationen und zwischen den jeweiligen Aminosäuren

Unterschiede in den Umgebungsschalenbesetzungen zwischen betrachteter Aminosäure und entsprechendem Atomtyp sowie zwischen den Aminosäuren selbst bilden die Grundlage des Aminosäure-Atom-Potentials. Dieses Potential soll die Umgebung einer spezifischen Aminosäure beschreiben und im Hinblick auf eine einzufügende Aminosäure bewerten. Dies wird hinfällig, wenn sich keine Unterschiede in den Schalenbesetzungen der Aminosäuren finden lassen.
Die Atomtypen der aliphatischen Kohlenstoffatome und der Lösungsmittelatome dominieren in der Schalenfüllung (s. Abbildung 3.3). Entscheidend sind aber nicht die generellen Verläufe der Anteilskurven oder der Anteil an der Gesamtfüllung selbst, sondern die Unterschiede in der Besetzung der einzelnen Schalen mit Atomtypen. Dort lassen sich individuelle Schwankungen in den Aminosäure-Atomtyp-Häufigkeiten sowohl bei schwach wie stärker vertretenen Aminosäure-Atomtyp-Kombinationen auffinden.
Die Beobachtung von Atomtyp-Häufigkeiten bei kleinen Radien ist abhängig von der Größe der Aminosäure. Eine kleine Aminosäure wie Glycin kann schon bei kleinen Abstandsradien zu CZ vollkommen von anderen Aminosäuren umgeben sein, während dies für große Aminosäuren wie Tryptophan oder Asparagin erst bei größeren Abständen zu CZ der Fall sein wird.

4.1.1.3. Diskussion der Aminosäure-Atom-Potentialkurven

Die Aminosäure-Atom-Potentialkurven werden mit der Boltzmanngleichung berechnet (s. Kapitel 2.1 und Kapitel 3.2.1.10). Die Häufigkeitsverteilungen der Atomtypen in einer definierten Umgebung werden mit einem Referenzsystem in Beziehung gesetzt (s. Kapitel 2.1.2). Für den Referenzzustand von Verteilungskurven gibt es kein schlüssiges Konzept. In dieser Arbeit wurde der Referenzzustand über ein Ensemble von Elementen über viele, verschiedene Proteinstrukturen angenähert [Finkelstein und Gutin, 1995; Leven, 1999]. Er entspricht der Repräsentation einer Aminosäure allgemeinen Typs und enthält hauptsäch-

lich unspezifische Informationen, bedingt durch generelle Packungseffekte [Sippl, 1990; Sippl, 1993; Gohlke *et al.*, 2000].

Bei der Interpretation einzelner Merkmale in den Potentialkurven ist zu berücksichtigen, dass die Paarverteilungsfunktionen einer Mittelung über den gesamten ausgewählten Strukturdatensatz entsprechen (s. Kapitel 2.2.6 und Kapitel 3.1). Die implizit darin enthaltenen Wechselwirkungsinformationen können von unterschiedlichem physikalischen Ursprung sein (z.B. elektrostatische oder sterische Wechselwirkungen) oder aus in kondensierten Systemen auftretenden Vielkörperwechselwirkungen stammen. Daher ist die Beschreibung einzelner Merkmale für die Betrachtung von Wechselwirkungspaaren und somit eine Separierung in ein Zweikörperproblem nur bedingt möglich [Moult, 1997; Gohlke *et al.*, 2000].

Die abstandsabhängige und richtungsunabhängige Betrachtung für das Aminosäure-Atom-Potential zeigt ein hohes Maß an Individualität für die Aminosäure-Atom-Kombinationen und zwischen den Aminosäuren. Die Abbildung 4.1 zeigt beispielhaft die Potentialkurven für den Atomtyp C_{ali} in Abhängigkeit von der Aminosäure. Die Aminosäuren wurden gemäß ihrer Eigenschaften jeweils Gruppen zugeordnet (aromatische, hydrophobe, geladene oder polare Aminosäuren sowie Hydroxyalkyl-Aminosäuren und Aminosäuren, die sich nicht einer dieser Gruppen zuordnen lassen). Die teilweise erhöhten Potentiale um den Wert von 4 Å herum können vermutlich auf aromatische Wechselwirkungen, die π-π-Wechselwirkungen aromatischer Systeme [Hunter *et al.*, 1991] (s. Abbildung 4.2), sowie auf Wechselwirkungen mit sekundärstrukturartigen Faltblattanordnungen zurückgeführt werden. Die Strukturierung bei Abständen um 5 Å und 11 Å lässt sich mit dem Auftreten von Wechselwirkungsmustern in den Koordinationssphären kondensierter Systeme um ein betrachtetes Zentralatom erklären. Die Potentialkurven für die Aminosäure-C_{ali}-Wechselwirkungen, die in Abbildung 4.1 dargestellt sind, verlaufen für die Gruppe aromatischer und hydrophober Aminosäuren größtenteils unter dem Nullniveau, was einer stabilisierenden Wechselwirkung entspricht, während die Kurven für polare oder geladenen Aminosäuren hauptsächlich über dem Nullniveau verlaufen, entsprechend einer destabilisierenden Wechselwirkung.

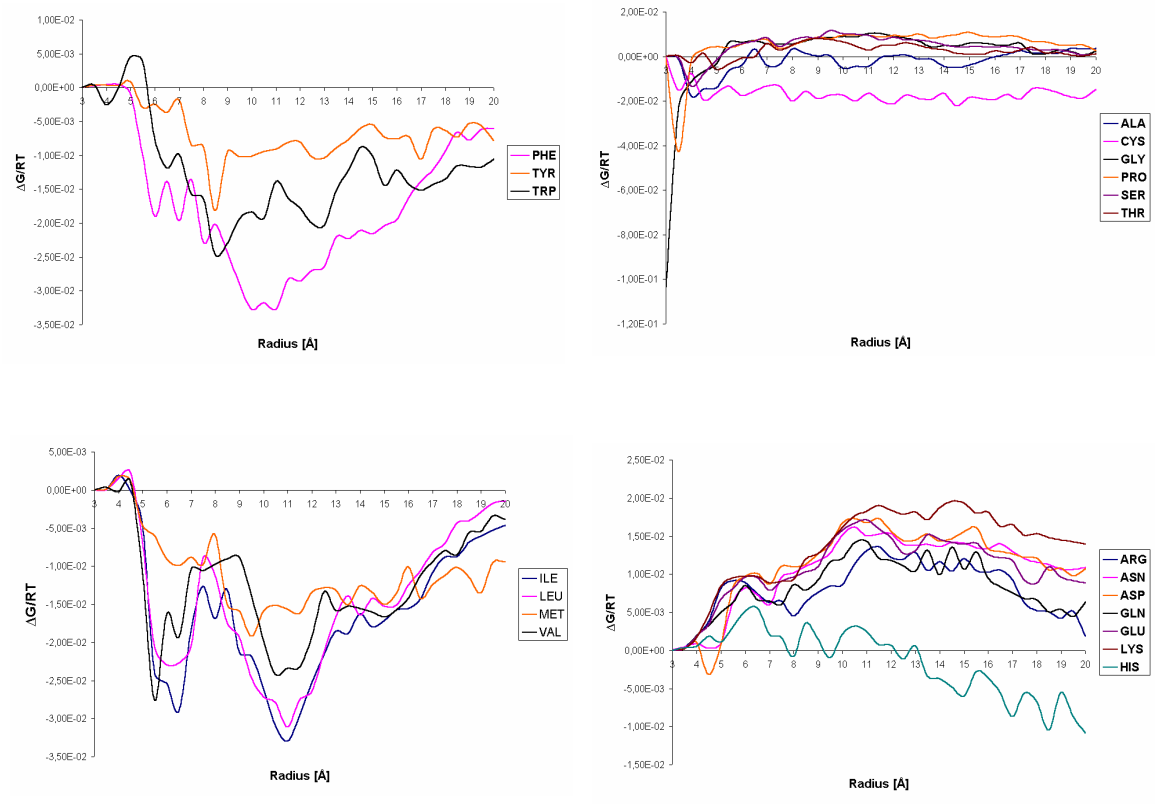


Abbildung 4.1: Darstellung der Potentialkurven für das Aminosäure-Atom-Potential in der oberen Halbkugel von zwanzig Aminosäuren und dem Atomtyp aliphatischer Kohlenstoff (x-Achse: Radius [Å], y-Achse: DG/RT)

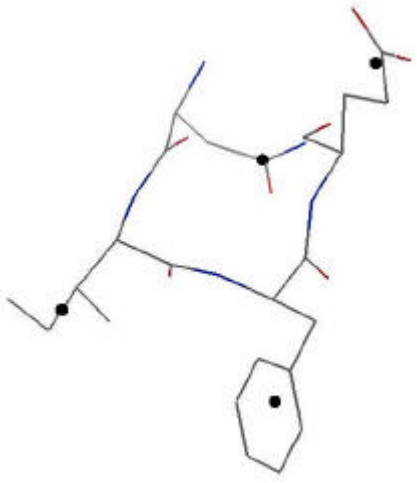


Abbildung 4.2: Darstellung von CZ (schwarze Punkte) für die Aminosäuren des Tetrapeptids ASN,ILE,PHE und GLU (erstellt mit dem Swiss-Pdbviewer, rot entspricht Sauerstoff, blau Stickstoff und grau Kohlenstoff)

Auf dieser Grundlage können auch die Unterschiede in den Kurvenverläufen zweier Aminosäuren diskutiert werden. Im Falle der aromatischen Aminosäuren sind die Potentialkurven sehr ähnlich. Sie unterscheiden sich nur in der Stärke ihrer Minima und Maxima, deren Lage leicht verschoben ist. Ihre Potentialkurven scheinen nach der Polarität der Aminosäuren sortiert zu sein, Tyrosin zeigt die geringste, Phenylalanin die stärkste Interaktion mit aliphatischen Kohlenstoffatomen.

Auch die Kurvenverläufe für die hydrophoben Aminosäuren ähneln sich. Für sie gilt wie schon bei den aromatischen Aminosäuren, dass sie sich nur in der Ausprägung ihrer Minima und Maxima unterscheiden, deren Lage auf der Ordinate verschoben ist. Für Valin, als kleinste hydrophobe Aminosäure, sind diese in Richtung kleinerer Abstände vorgezogen. Andere Aminosäuren können sich vom Abstand her früher um Valin gruppieren. Isoleucin mit seinem verzweigten CB-Atom kann wesentlich dichter gepackt werden als Leucin und zeigt dementsprechend eine stärkere stabilisierende Wechselwirkung mit den aliphatischen Kohlenstoffatomen. Geladene Aminosäuren zeigen, bis auf Histidin mit seiner Ringstruktur und π-Elektronen, durchgehend eine destabilisierende Wechselwirkung mit aliphatischen Kohlenstoffatomen.

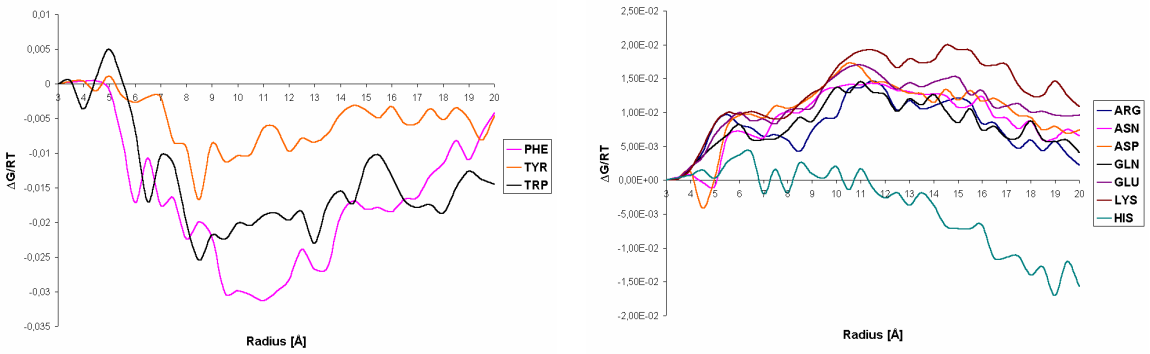


Abbildung 4.3: Darstellung der Potentialkurven für das Aminosäure-Atom-Potential in der unteren Halbkugel von aromatischen und geladenen oder polaren Aminosäuren und dem Atomtyp aliphatischer Kohlenstoff (x-Achse: Radius [Å], y-Achse: DG/RT)

In der richtungsabhängigen Betrachtung, dem Vergleich von Aminosäure-Atom-Potentialkurven zwischen den Halbkugeln, zeigen sich nur geringe Unterschiede, die meistens auf die Stärke der jeweiligen Wechselwirkung zurückgehen. Diesen Eindruck geben die in Abbildung 4.3 dargestellten Potentialkurven für die Wechselwirkungen zwischen aromatischen und polaren oder geladenen Aminosäuren mit aliphatischen Kohlenstoffatomen wieder. Eine Vorzugsrichtung kann unter den gewählten Bedingungen (s. Kapitel 2.2) nicht gefunden werden. Die Verwendung von zwei Halbkugeln und den in dieser Arbeit ausgewählten Atomtypen lässt keine genaue Beschreibung der Richtung von wechselwirkenden Strukturelementen zu. Eine genauere Winkeleinteilung der Umgebung führte durchgehend zu schlechteren Ergebnissen (s. Kapitel 3.2). Somit kann ein Einfluss der in Kapitel 2.1.1.1 dargestellten spezifischen Wechselwirkungen auf die Potentialkurven nicht gefunden werden.

4.1.1.4. Das Torsionswinkelpotential

Das in dieser Arbeit verwendete Torsionswinkelpotential entspricht dem entwickelten und diskutierten Torsionswinkelpotential von Leven [Leven, 1999]. Die (φ,ψ)-Häufigkeitsverteilungen sind abhängig von der Art der Aminosäure. Diese Abhängigkeiten spiegeln sich in den abgeleiteten Potentialen wider. Die besondere Rolle von Glycin und Prolin sowie Ähnlichkeiten aufgrund von Eigenschaften (z.B. hydrophobe oder polare Aminosäuren) werden erkannt und in der Energielandschaft abgebildet. Beobachtet werden Ähnlichkeitsmuster auf struktureller Basis (a.a.O), wie sie zwischen Isoleucin und Valin auftreten (Aminosäuren mit einem tertiären CB-Atom) oder die durch Seitenkettenvolumen verursachten Einschränkungen des Konformationsraumes.
In dieser Arbeit wurden die Torsionswinkel bei dem Austausch einer Aminosäure konstant gehalten. Die Stabilisierungsenergie entspricht der Differenz der Energiewerte der strukturellen Torsionswinkel (φ,ψ) für den Wildtyp und die Mutante. Eine mögliche Winkeloptimierung durch Energieminimierungen oder empirische Kraftfeldmethoden wurden in dieser Arbeit nicht durchgeführt.

4.2. Validierung und Optimierung der wissensbasierten Bewertungsfunktion

In dieser Arbeit sollte eine Methode zur automatischen Thermostabilitätsvorhersage entwickelt werden. Ein Aminosäure-Atom-Potential kombiniert mit einem Torsionswinkelpotential soll als Bewertungsfunktion dienen (s. Kapitel 3.2.3). Die notwendige Validierung und Optimierung dieser wissensbasierten Bewertungsfunktion muss an einem Satz experimenteller Daten durchgeführt werden (s. Kapitel 2.2.6). Diesem Datensatz mit experimentellen Stabilitätsdaten kommt eine besondere Rolle zu. Er soll mögliche Kriterien für eine Verbesserung der Bewertungsfunktion liefern, eine kritische Bewertung der Vorhersagequalität ermöglichen und darüber entscheiden, ob der entwickelte Vorhersagealgorithmus in der Realität anwendbar ist (s. Kapitel 2.2.7). Alle verwendeten experimentellen Datensätze sollen möglichst groß, umfangreich und fehlerfrei sein. Nur so lassen sich zuverlässige Aussagen über die Qualität der Vorhersage machen. Die Datensätze sollen die möglichen Eigenschaften eines zu betrachtenden Systems repräsentieren. Ihre Auswahl sollte objektiv diesen Kriterien gehorchen und nicht durch statistische, willkürliche oder chemisch-physikalisch begründbare Kriterien soweit eingeschränkt werden, dass der Vorhersagealgorithmus sein Ziel erreichen kann. Auf der Basis einer solchen Optimierung kann keine Entscheidung darüber getroffen werden, ob der entwickelte Ansatz generell einsetzbar ist oder den geforderten Ansprüchen genügt.

In dieser Arbeit wurde die Validierung und Optimierung der wissensbasierten Bewertungsfunktionen zur Vorhersage der Thermostabilität an einem aus der Literatur gewählten Datensatz durchgeführt, der größtenteils den von Gilis [Gilis und Rooman, 1996; Gilis und Rooman, 1997], Topham [Topham *et al.*, 1997] und Leven [Gerk *et al.*, 2000] verwendeten experimentellen Stabilitätsdaten entspricht. Der Entwicklungsdatensatz enthält 646 Thermostabilitätsdaten aus elf Proteinen (s. Kapitel 2.2.6 und Kapitel 3.1). Die optimierten Bewertungsfunktionen wurden anschließend auf einen Testdatensatz angewandt, dessen Stabilitätsdaten aus einer wesentlich größeren Datensammlung stammen (s. Kapitel 1.3.3 und Kapitel 2.2.6.2) und nur durch die Verfügbarkeit der Daten und zum

Teil durch die in Kapitel 2.2.6.1 angegebenen Selektionskriterien für die Entwicklung und Ableitung der Potentiale eingeschränkt ist. Der Testdatensatz enthält mehr als 900 Stabilitätsdaten aus 29 unterschiedlichen Proteinen (s. Kapitel 3.1).

4.2.1. Experimentelle Thermostabilitätsdaten

Die der Literatur entnommenen experimentellen Daten für die thermische Denaturierung sind aus mehreren Gründen kritisch zu betrachten. Sie besitzen eine hohe Ungenauigkeit und sind mit schwer zu schätzenden Fehlern behaftet [Robertson und Murphy, 1997]. In den experimentellen Beschreibungen fehlen häufig Angaben über die Versuchsbedingungen, wie z.B. pH-Wert, Protein- oder Salzkonzentration, die einen wichtigen Einfluss auf die Thermostabilität eines Proteins ausüben können (s. Kapitel 1.3.2). Bei den Experimenten zur thermischen Stabilität ist oft nicht klar, ob die thermodynamische Reversibilität gegeben und damit eine Berechnung der thermodynamischen Größen zulässig ist (s. Kapitel 1.2.6). Für die Ermittlung der freien Enthalpie (s. Kapitel 1.2.6) wird der Wert der Wärmekapazität häufig nicht aus dem Experiment gewonnen, sondern theoretischen Berechnungen entnommen. Die theoretischen Werte sind nicht proteinspezifisch und mit schwer zu fassenden Fehlern behaftet [Shoichet *et al.*, 1995; Robertson und Murphy, 1997; Xu *et al.*, 1998; Rees und Robertson, 2001]. Die experimentellen Werte der Wärmekapazität, der Enthalpie und der Temperatur wiederum hängen danach stark von den Versuchsbedingungen, der Art der Messung (DSC oder CD-Spektroskopie) und der Qualität der Messung ab.

4.3. Ergebnisse dieser Arbeit

Die Leistung der in dieser Arbeit entwickelten Bewertungsfunktionen für die Vorhersage der Thermostabilität von Proteinen wurde an experimentellen Literaturdaten überprüft. Die Ergebnisse für den Vorhersagealgorithmus zeigen deutliche Unterschiede zwischen Entwicklungsdatensatz und Testdatensatz (s. Tabelle 4.1 und Tabelle 4.2). Während die Vorhersagekriterien für den Entwicklungsdatensatz die besten Werte r_{cor}=0,81, rv=76 %, Sens=0,96 und Spez = 0,48 erreichen, betragen sie für den Testdatensatz r_{cor} =0,74, rv =71,4 %, Sens=0,93 und Spez=0,57. Die Verbesserungen, die für die Vorhersagefunktionen auf der Basis des Entwicklungsdatensatzes getroffen wurden, finden sich im Testdatensatz wieder. Dies gilt auch für das Verhältnis zwischen Aminosäure-Atom-Potential und Torsionswinkelpotential sowie für die Kombination der beiden zu einer Gesamtbewertungsfunktion. Dies kann als Indiz dafür gewertet werden, dass die modellierten physikalisch-chemischen Abhängigkeiten vom Modell gut beschrieben und wiedergegeben werden.

Tabelle 4.1: Zusammenfassung der Vorhersage-Ergebnisse der Potentialfunktionen für den Entwicklungsdatensatz

	Entwicklungsdatensatz			
Methode	**r_{cor}**	**rv [%] (s [kcal/mol])**	**Sens**	**Spez**
Aminosäure-Atom-Potential	0,67 (646)	64,7 (1,44)	0,67	0,35
Torsionswinkelpotential	0,26 (646)	67,8 (1,48)	0,75	0,39
Gesamtbewertungsfunktion[a)]	0,72 (646)	74,5 (1,44)	0,75	0,48
Gesamtbewertungsfunktion mit Datenausschluss > 3*σ[b)]	0,76 (612)	75 (1,00)	0,73	0,46
Gesamtbewertungsfunktion mit Datenausschluss > 2*σ[b)]	0,81 (561)	76 (0,77)	0,72	0,47
Gesamtbewertungsfunktion mit n-σ[b)]	0,72 (646)	52 (1,37)	0,96	0,32

Tabelle 4.2: Zusammenfassung der Vorhersage-Ergebnisse der Potentialfunktionen für den Testdatensatz

	Testdatensatz			
Methode	**r_{cor}**	**rv [%] (s [kcal/mol])**	**Sens**	**Spez**
Aminosäure-Atom-Potential	0,42 (918)	62,9 (1,69)	0,64	0,46
Torsionswinkelpotential	0,16 (918)	60,9 (1,63)	0,63	0,44
Gesamtbewertungsfunktion[a)]	0,46 (918)	69,4 (1,67)	0,67	0,53
Gesamtbewertungsfunktion mit Datenausschluss > 3*σ[b)]	0,64 (851)	70,0 (1,06)	0,68	0,56
Gesamtbewertungsfunktion mit Datenausschluss > 2*σ[b)]	0,74 (747)	71,4 (0,75)	0,67	0,57
Gesamtbewertungsfunktion mit n-σ[b)]	0,46 (918)	59,0 (1,68)	0,93	0,45

[a)] Gesamtbewertungsfunktion entspricht dem gewählten kombinierten Aminosäure-Atom- und Torsionswinkelpotential

[b)] σ = 1,44 kcal/mol

n theoretisch berechnete Werte

Die Bewertung der zwischen Entwicklungs- und Testdatensatz gemittelten Standardabweichung (σ_{Mittel} = 1,55 kcal/mol) im Hinblick auf die möglichen experimentellen Fehler für die Gesamtbewertungsfunktion fällt schwer. Der errechnete Fehler ist deutlich höher als ein geschätzter mittlerer Fehler für die experimentelle freie Enthalpie von 0,2-0,5 kcal/mol. Mögliche Gründe hierfür können der gewählte Referenzzustand sowie die Nicht-Berücksichtigung des Faktors *R*T in der Boltzmanngleichung sein. Andere Gründe können darin liegen, dass die Potentiale aus Proteinen abgeleitet wurden, die keinerlei Kofaktoren, prosthetische Gruppen oder Schwermetallionen enthalten durften (s. Kapitel 2.2.6.1), die Überprüfung der wissensbasierten Bewertungsfunktion im Fall des Testdatensatzes wurde aber teilweise an Enzymen mit solchen Gruppen durchgeführt. Werden Mutationen ausgeschlossen, deren Differenz zwischen experimentellen und theoretischen Werten sich um mehr als die dreifache Standardabweichung unterscheidet, wird eine mittlere Standardabweichung (Entwicklungsdatensatz zu Testdatensatz) von σ = 1,03 kcal/mol erhalten. Dieser Wert

liegt einem geschätzten experimentellen Fehler deutlich näher. In den Datensätzen sind Stabilitätswerte aus verschiedenen Studien mit unterschiedlichen Zielsetzungen zusammengefasst. Die Vergleichbarkeit dieser Werte kann nur geschätzt werden, so dass die Vorhersagegenauigkeiten für den Entwicklungsdatensatz und den Testdatensatz wohl den Fehlerrahmen der Literaturdaten wiederspiegeln.

Werden die falsch negativ vorhergesagten Mutationen für den Entwicklungsdatensatz (s. Tabelle 3.19) betrachtet, kann festgestellt werden, dass bis auf zwei Mutationen alle experimentellen Stabilitätswerte im geschätzten Fehlerbereich von 0,2-0,3 kcal/mol liegen und damit eine Aussage über die Stabilitätsänderungen durch die Mutationen zweifelhaft ist. Die Ausnahmen betreffen eine Mutation von Glycin im Protein 4lyz und einen Austausch von Lysin durch Prolin im Protein. In diesen beiden Fällen scheinen die sterischen Ansprüche falsch modelliert worden zu sein, denn Glycin ist die Aminosäure mit dem größten, Prolin die Aminosäure mit dem kleinsten erlaubten Bereich von Torsionswinkelkombinationen. Eine Mutation von Glycin oder ein Einsatz von Prolin bewirken daher meist eine lokale Änderung im Verlauf des Peptidrückrates, doch diese wird durch das in dieser Arbeit angenommene Prinzip der unveränderten Proteinstruktur nicht berücksichtigt.

4.3.1. Abhängigkeit der Vorhersagegenauigkeit von den betrachteten Proteinen

Die wissensbasierte Bewertungsfunktion zeigte Unterschiede in der Vorhersagegenauigkeit bei verschiedenen Proteinen sowie bei verschiedenen Strukturen eines Proteins (s. Tabelle 3.5, Tabelle 3.28 und Tabelle 3.29). Die Unterschiede in der Qualität der Vorhersage für unterschiedliche Proteine sind erklärbar, nicht jedoch die Unterschiede in den Ergebnissen für verschiedene Strukturen des gleichen Proteins (s. Tabelle 3.6 und Tabelle 3.28). Wurden die Strukturen der T4-Lysozyme während der Vorhersage vertauscht, hatte dies keinen Einfluss auf die Vorhersagequalität, die Unterschiede müssen in den Mutationsdaten liegen [Leven, 1999]. Die Stabilitätsdaten für das Protein 2lzm stammen größ-

tenteils von Alber *et al*. [Alber *et al.*, 1987]. In ihrer Arbeit wurde der Einfluss von Thr 157, einer Aminosäure in einer Oberflächen-Loop-Region, auf die Proteinstabilität untersucht. Dazu wurden verschiedene Mutanten hergestellt und ihre Kristallstruktur bestimmt. Dabei zeigte sich, dass alle Mutationen strukturelle Änderungen, durch Neubildung oder Abbruch von Wasserstoffbrückenbindungen, im Verlauf der Protein-Hauptkette verursachten. In der Arbeit von Alber *et al*. wurden keine Angaben zu den experimentellen Bedingungen gemacht, nur die Verwendung von CD-Spektroskopie bei den thermodynamischen Studien wurde erwähnt.

Die Vorhersageleistung hängt also im großen Maße nicht nur von der Proteinstruktur, sondern auch von den Mutationsdaten ab. Diese Abhängigkeit geht so weit, dass es für eine *de novo* Vorhersage nicht möglich sein wird, zwischen diesen Einflüssen zu differenzieren und die Ergebnisse zu interpretieren.

Die für den Testdatensatz mit der Gesamtbewertungsfunktion getroffenen Vorhersage-Ergebnisse für die einzelnen Proteine zeigen für die richtige Vorhersage rv durchgehend gute Ergebnisse, während die Korrelationskoeffizienten r_{cor} stark schwanken (s. Tabelle 3.29). Die Gründe für dieses Verhalten können im Hinblick auf die große Datenmenge nicht nachvollzogen werden. An dieser Stelle soll auf die Diskussion der Ergebnisse des Entwicklungsdatensatzes verwiesen und noch einmal die Ungenauigkeiten bei experimentellen Messungen erwähnt werden.

4.3.2. Abhängigkeit der Vorhersageleistung von der Mutation

Die Vorhersageergebnisse können im Hinblick auf die betrachtete Aminosäure näher untersucht werden. Dazu wurden die Datensätze nach Kriterien der Lösungsmittelzugänglichkeit, der Sekundärstrukturzugehörigkeit, der ausgetauschten Aminosäure und der eingesetzten Aminosäure segmentiert.

4.3.2.1. Lösungsmittelzugänglichkeit

Die Vorhersage-Ergebnisse für das Aminosäure-Atom-Potential zeigen mit abnehmender Lösungsmittelzugänglichkeit eine Steigerung der Vorhersageleis-

tung (s. Kapitel 3.2.51). In der Literatur werden Aminosäure-Aminosäure-Potentiale auch als *hydrophobe Potentiale* bezeichnet [Casari und Sippl, 1992; Bryant und Lawrence, 1993; Thomas und Dill, 1996]. Dies scheint auch für die in dieser Arbeit verwendeten Aminosäure-Atom-Potentiale eine zutreffende Beschreibung zu sein (s. Kapitel 3.2.5.1). Die Torsionswinkelpotentiale zeigen bei durchgehend geringeren Werten der Korrelationskoeffizienten r_{cor} im Vergleich zum Aminosäure-Atom-Potential eine umgekehrte Präferenz im Hinblick auf die Lösungsmittelzugänglichkeit. Die Vorhersagequalität steigt bei den Torsionswinkelpotentialen mit zunehmender Lösungsmittelzugänglichkeit. Durch die geringeren Packungseffekte und die zunehmende Flexibilität mit zunehmender Lösungsmittelzugänglichkeit nimmt der lokale Einfluss der direkten Nachbarn in der Sequenz auf die betrachtete Aminosäure zu, während im Kern Packungseffekte dominieren [Gilis und Rooman, 1996]. Die Gesamtbewertungsfunktion gleicht in ihrem Verhalten dem Aminosäure-Atom-Potential bei aber durchgehend besseren Korrelationskoeffizienten und richtigen Vorhersagen.

4.3.2.2. Sekundärstrukturelemente

Das Aminosäure-Atom-Potential und die Gesamtbewertungsfunktion weisen im Gegensatz zu dem Torsionswinkelpotential eine deutliche Beeinflussung der Vorhersageleistung durch die Lage der Mutation in Sekundärstrukturelementen auf (s. Kapitel 3.2.5.2). Die Korrelationskoeffizienten für die Sekundärstrukturelemente Helix und Faltblatt sind deutlich gegenüber einer Nichtzuordnung der Mutation zu einem Sekundärstrukturelement erhöht. Sekundärstrukturelemente erstrecken sich durch den Proteinkern, sind dichter gepackt und durchschnittlich stärker von Strukturelementen bedeckt. Demnach ist die Umgebung von Sekundärstrukturelementen besser definiert und eine Umgebungsbeschreibung kann mehr mögliche Informationen enthalten. Gegen eine mögliche implizite Abhängigkeit von der Lösungsmittelzugänglichkeit spricht, dass sich die einem Sekundärstrukturelement nicht zugeordneten Mutationen durch das gesamte Protein ziehen und trotz hoher Anzahl einen nicht signifikanten linearen Zusammenhang zwischen experimentellen und theoretischen Werten aufweisen.

Das Torsionswinkelpotential weist keine so klare Abhängigkeit von Sekundärstrukturelementzugehörigkeiten auf. Die für den Entwicklungsdatensatz angeführten beiden höchsten Korrelationskoeffizienten entsprechen jenen, die für Faltblätter und eine Nicht-Zuordnung erreicht werden. Die Erhöhung der Vorhersageleistung für die keinem Sekundärstrukturelement zugeordneten Mutationen lässt sich auf einen größeren Einfluss der lokalen Sequenznachbarn zurückführen. Dieses Ergebnis findet sich für den Testdatensatz nicht wieder. Die Korrelationskoeffizienten liegen alle auf einem vergleichbar niedrigen Niveau.
Der erhöhte Korrelationswert für Faltblätter, aber nicht für Helices, könnte auf eine höhere Flexibilität der Faltblätter gegenüber den Helices hinweisen. Dieser Effekt wurde schon bei Leven diskutiert [Leven, 1999]. Dort wird auf die makroskopisch beobachtbare höhere Flexibilität von z.B. Seidenfibrinogen hingewiesen.

4.3.2.3. Betrachtung der Vorhersagegenauigkeit in Abhängigkeit von der ausgetauschten und der eingefügten Aminosäure

Die in Kapitel 3.2.5.3 vorgestellten Ergebnisse zeigen für das Aminosäure-Atom-Potential und die Gesamtbewertungsfunktion eine im Trend deutlich bessere Vorhersageleistung für hydrophobe und aromatische Aminosäuren. Diese sind häufig im Inneren und damit in den hydrophoben Bereichen eines Proteins lokalisiert. Die Präferenz des Aminosäure-Atom-Potentials und der Gesamtbewertungsfunktion für den Proteinkern bestätigt sich in dieser Untersuchung. Dies entspricht den Beobachtungen in der Literatur [Gilis und Rooman, 1996; Gilis und Rooman, 1997; Leven, 1999].
Das Torsionswinkelpotential weist in der Vorhersage der Thermostabilität für die Aminosäuren Glycin, Glutamin und Asparagin erhöhte Korrelationskoeffizienten auf. Glutamin und Asparagin liegen in einem Protein häufig Lösungsmittelexponiert vor. Glycin kommt aufgrund seiner fehlenden Seitenkette eine Sonderrolle zu. Im Vergleich der Ergebnisse zwischen Entwicklungsdatensatz und Testdatensatz können die oben vorgestellten Abhängigkeiten wiedergefunden werden.
Die Abhängigkeiten der Vorhersageleistungen von der einzufügenden Aminosäure sind nicht so deutlich. Alle Aminosäure zeigen bis auf Prolin, Tryptophan

und Tyrosin durchgehend unter Berücksichtigung aller angewendeten Potentiale gute Korrelationskoeffizienten für den Entwicklungsdatensatz. (s. Kapitel 3.2.5.4). Die Ergebnisse brechen für den Testdatensatz ein und sind für fast alle Aminosäuren durchgehend schlechter. Die Abhängigkeiten der Korrelationen scheinen unabhängig von der Art oder den Eigenschaften der Aminosäuren zu sein. Während der Einfluss der Wildtyp-Aminosäure durch Struktur und Umgebungsbestimmung definiert ist, können diese Informationen nicht in die Betrachtung der einzubauenden Aminosäure eingehen. Eine Mutation kann die Struktur eines Proteins, durch z.B. sterische Effekte oder Bildung sowie Abbruch von Wasserstoffbrückenbindungen, verändern. Mögliche Strukturänderungen fließen in den entwickelten Vorhersagealgorithmus nicht mit ein.

4.3.2.4. Benötigte Rechenzeit für die Erstellung eines Mutationsprofils

Für die Erstellung eines Mutationsprofils mit vollständiger Beschreibung der Umgebung mit den fünf Atomtypen wird auf einem Pentium-III-PC (600 MHz, 512 RAM Arbeitsspeicher) eine mittlere Laufzeit von 0,28 s /Aminosäure benötigt. Die Bestimmung der benötigten Rechenzeit wurde mit dem Unix-Befehl *time* und exemplarisch für die Enzyme 1lyd (164 Aminosäuren) und 2wsy (612 Aminosäuren) dargestellt.

```
Read in THERM_Protein ... 1lyd
Erstelle Gitter mit einer Gitterkonstante von 0.6 A
Apexangle: 90° Maxdst: 20
Es wurden 23632 Wasser und 164 AS eingelesen
Berechnung der Stabilisierungsenergien für [3,13]
Program READY and FINISH ...
47.01user 0.52system 0:47.74elapsed 99%CPU

Read in THERM_Protein ... 2wsy
Erstelle Gitter mit einer Gitterkonstante von 0.6 A
Apexangle: 90° Maxdst: 20
4632 Atome aus 4648 pdb-Zeilen
Es wurden 51174 Wasser und 612 AS eingelsen
Berechnung der Stabilisierungsenergien für [3,13]
Program READY and FINISH ...
171.85user 1.31system 2:55.55elapsed 98%CPU
```

4.4. Diskussion der Ergebnisse mit bekannten Verfahren zur Vorhersage der Thermostabilität

Die Arbeiten von Topham [Topham *et al.*, 1997], Gilis und Rooman [Gilis und Rooman, 1996; Gilis und Rooman, 1997], Leven [Leven, 1999; Gerk *et al.*, 2000] und Guerois [Guerois *et al.*, 2002] stellen die in der Literatur veröffentlichten Methoden zur Vorhersage der Thermostabilität mit einer systematischen Überprüfung der Vorhersage-Ergebnisse dar. Während Topham zur Vorhersage Aminosäure-Austausch-Tabellen verwendete, entwickelten Gilis und Rooman, Leven und Guerois wissensbasierte Energiefunktionen.

4.4.1. Aminosäure-Austausch- und Eigenschafts-Tabellen zur Vorhersage der Thermostabilität von Proteinen

Die in der Arbeit von Topham et al [Topham *et al.*, 1997] angewendete Methode zur Thermostabilitätsvorhersage leitet eine Aminosäureaustauschwahrscheinlichkeit her, die auf Strukturalignments von ca. 500 Proteindomänen aus 113 Proteinfamilien beruht. Die Austauschwahrscheinlichkeiten können mit umgebungsbeschreibenden Parametern (z.B. Lösungsmittelzugänglichkeit) kombiniert werden (s. Kapitel 1.4), um eine Stabilitätsbewertung durch die Mutation einer Aminosäure zu treffen.

Die berechneten Tabellen wurden an der Stabilitätsvorhersage von 217 Datenpunkten des Lysozyms aus *T4 Phage* getestet und die aus der Lysozym-Auswertung gewonnene Regressionsgeradengleichung auf 68 Mutationen der Barnase sowie auf 83 Mutationen der *Staphylococcal nuclease* angewendet. Die erreichte Korrelation für die 217 Stabilitätsdaten des Lysozymes der optimierten Methode betrug r=0,77 und es wurden 73,3 % aller Mutationen richtig stabilisierend sowie destabilisierend vorhergesagt. Für 151 Daten aus Barnase und *Staphylococcal nuclease* wurde eine Korrelation von r=0,79 mit den experimentellen Werten erhalten und 86 % aller stabilisierenden und destabilisierenden Mutationen richtig vorhergesagt.

Diese Vorhersagemethode hat seine Leistungsfähigkeit an einem kleinen ausgesuchten Satz von Stabilitätsdaten bewiesen. Allerdings ist der vorgestellte Ansatz nur bedingt oder gar nicht für eine generelle Anwendung geeignet, da jede zu bewertende Mutation als entsprechende Kristallstruktur vorliegen muss.
Die in dieser Arbeit entwickelte Bewertungsfunktion erreicht für 143 Stabilitätswerte von Barnase und *Staphylococcal nuclease* eine Korrelation von r_{cor}=0,72 und eine richtige Vorhersage von rv= 88,2 %. Für 198 Stabilitätswerte von T4-Lysozyme erreicht die Gesamtbewertungsfunktion einen Korrelationskoeffizienten zwischen berechneten und experimentellen Werten von r_{cor}=0,74 und eine richtige Vorhersage von rv=68 %. Die Vorhersagequalität kann als gleichwertig angesehen werden, wobei der technische Aufwand der hier vorgestellten Bewertungsfunktion deutlich geringer ist, da nur die Struktur des Wildtypproteins benötigt wird.

4.4.2. Wissensbasierte Energiefunktionen zur Vorhersage der Thermostabilität von Proteinen

Die wissensbasierten Energiefunktionen umfassen zwei Klassen von Potentialen, die statistisch effektiven Energiefunktionen (SEEF) und die empirisch effektiven Energiefunktionen (EEEF) (s. Kapitel 2). Die SEEF werden aus Häufigkeitsverteilungen von Strukturelementen abgeleitet, während die EEEF auf aus Proteinversuchen gewonnene experimentelle Daten zurückgreifen.
Diese vorliegende Arbeit sowie die Arbeiten von Gilis und Rooman [Gilis und Rooman, 1996; Gilis und Rooman, 1997] und Leven [Leven, 1999] beschreiben eine spezielle Methode zur Vorhersage der Thermostabilität mit SEEF, während Guerois *et al*. [Guerois *et al.*, 2002] die EEEF verwenden.

4.4.2.1. Statistisch effektive Energiefunktionen zur Vorhersage der Thermostabilität von Proteinen

Gilis und Rooman [Gilis und Rooman, 1996; Gilis und Rooman, 1997] setzten erstmals wissensbasierte Potentiale für die Stabilitätsvorhersage ein. Sie entwi-

ckelten ein Aminosäure-Aminosäure-Potential und zwei Torsionswinkelpotentiale. Das Paarpotential wurde über die Abstände von 3-8 Å der Aminosäureseitenkettenschwerpunkte abgeleitet. Für die Vorhersage wurde ein Datensatz mit 238 experimentellen Stabilitätswerten aus sieben Proteinen verwendet. Die Daten wurden in drei Lösungsmittel-zugängliche Bereiche (0-20 %, 20-40 % und 40-50 %) aufgeteilt. Für die 121 Mutationen, die im kleinsten Lösungsmittel-zugänglichen Bereich lokalisiert waren, erzielte das Aminosäure-Aminosäure-Potential eine Korrelation von r=0,78, durch Addition eines Torsionswinkelpotentials konnte dieser Wert auf r=0,8 gesteigert werden. Für 69 Mutationen mit einer Lösungsmittelzugänglichkeit zwischen 20-40 % wurden für ein Torsionswinkelpotential und das Aminosäure-Aminosäure-Potential die Korrelationswerte r= 0,57 und r=0,58 erreicht. Durch Kombination und geeignete Wichtung eines Torsionswinkelpotentials mit dem Aminosäure-Aminosäure-Potential konnte die Korrelation auf r=0,7 gesteigert werden. Für die 48 Mutationen zwischen 40-50 % konnte keines der eingesetzten Potentiale oder eine geeignete Kombination aus zweien (nach Handselektion der Mutationsdaten) mehr als die maximale Korrelation von r=0,55 erreichen.

Leven [Leven, 1999] entwickelte ein abstandsabhängiges Aminosäure-Atom-Potential und ein Torsionswinkelpotential, die er anschließend miteinander kombinierte. Das Torsionswinkelpotential entspricht nicht dem von Gilis und Rooman verwendeten und ist in Kapitel 2.2.4 vorgestellt. Das Aminosäure-Atom-Potential ermöglicht eine genauere Umgebungsbeschreibung. Der in der Arbeit von Leven betrachtete Radius, mit dem die Umgebung erfasst wird, ist mit 4-15 Å deutlich größer als jener, der in der Methode von Gilis und Rooman verwendet worden ist. Die Methode von Leven erreichte mit einem kombinierten Potential für den Datensatz von Gilis und Rooman eine Gesamtkorrelation für alle Zugänglichkeitsbereiche von r=0,67, mit den entsprechenden Werten für die Lösungsmittelzugänglichkeitsbereiche die Korrelationswerte von $r_{0\text{-}20\%}$=0,72, $r_{20\text{-}40\%}$=0,71 und $r_{40\text{-}50\%}$=0,62. Für den auch in dieser Arbeit verwendeten Entwicklungsdatensatz erreicht die Methode von Leven mit einem daran optimierten abstandsabhängigen Aminosäure-Atom-Potential einen Korrelationskoeffizienten von r_{cor}= 0,69 (rv= 65,2 %) und für das Torsionswinkelpotential einen

Wert für die Korrelation von r_{cor}=0,32 (rv=62,7 %). Ein optimiertes Aminosäure-Atom-Potential wurde auf einen Testdatensatz angewendet und erzielte einen Korrelationskoeffizienten von r_{cor}=0,51 für 831 Mutationsdaten (aus 876 Stabilitätsdaten von 29 unterschiedlichen Proteinen). Die beste Gewichtung der Potentiale in ihrer Kombination erreichte einen Korrelationskoeffizienten von r=0,74 (645 Datenpunkte) für den verwendeten Entwicklungsdatensatz und einen Wert von r=0,57 (831 Datenpunkte) für den verwendeten Testdatensatz. Wurden alle experimentellen und theoretischen Werte ausgeschlossen, deren Betrag kleiner als 0,5 war, was einem Datenausschluss von 74 % aller Daten entsprach, konnten 198 Mutationen aus 224 Mutationen richtig vorhergesagt werden (rv=88 %).

Nicht alle Werte können mit der in dieser Arbeit entwickelten Potentialfunktion erreicht werden, allerdings können auch für wesentliche größere Datensätze leicht Ergebnisse gezeigt werden, welche die Methode von Leven oder Gilis und Rooman übertreffen können. Der in dieser Arbeit erreichte Gesamtkorrelationskoeffizient für den Datensatz für Gilis und Rooman beträgt r_{cor}=0,7 (237 Datenpunkte) mit rv=75,1 %. Dieser Wert liegt über dem von Leven erreichten. Die Einzelkorrelationen sind teilweise besser oder auch schlechter und erzielen die Korrelationswerte $r_{0\text{-}20\%}$=0,72, $r_{20\text{-}40\%}$=0,52 und $r_{40\text{-}50\%}$=0,6. Werden nur die Mutationen betrachtet, die in der Sekundärstruktur des β-Faltblatts liegen, wird ein Korrelationskoeffizient von r_{cor}=0,83 (52 Datenpunkte) und rv=83 % erreicht. Dieser Datensatz von 52 Stabilitätsdaten liegt in der Größenordnung der von Gilis und Rooman verwendeten 69 Mutationen zwischen 20-40% Lösungsmittelzugänglichkeit und der 48 Mutationen mit einer Lösungsmittelzugänglichkeit zwischen 40-50 %. Noch besser schneidet die in dieser Arbeit entwickelte Bewertungsfunktion für den Entwicklungsdatensatz ab, wenn Mutationen, deren berechnete Stabilitätswerte vom Betrag her um ein Vielfaches größer sind als eine Standardabweichung, ausgeschlossen werden. Die Standardabweichung wurde für die Gesamtdatenmenge des Entwicklungsdatensatzes zu σ = 1,44 kcal/mol bestimmt (s. Tabelle 4.3 und Kapitel 3.2.3 und Kapitel 3.2.4). Die erreichten Korrelationen liegen weit über denen von Leven oder Gilis und Rooman.

Für den Testdatensatz wird mit der hier entwickelten Gesamtpotentialfunktion ein Korrelationskoeffizient von r_{cor}=0,46 und ein rv=69 % erreicht. Werden alle Mutationen ausgeschlossen, deren Differenz zwischen berechneten und experimentellen Werten größer als die dreifache Standardabweichung von 1,44 kcal/mol ist, steigt die Korrelation auf r_{cor}=0,64 für immer noch 851 Stabilitätsdaten (entsprechend 93 % der Gesamtdatenmenge) (s. Tabelle 4.4).

Tabelle 4.3: Vorhersage-Ergebnisse der Gesamtbewertungsfunktion für den Entwicklungsdatensatz in Abhängigkeit von der Lösungsmittelzugänglichkeit

Methode	0-50 % r_{cor}	$rv_{0-50\%}$ (s [kcal/mol])	0-20% r_{cor}	20-40% r_{cor}	40-50% r_{cor}
Gesamtbewertungsfunktion mit Datenausschluss > 3*σ[a)]	0,76 (614)	75 (1,00)	0,78 (295)	0,56 (122)	0,54 (275)
Gesamtbewertungsfunktion mit Datenausschluss > 2*σ[a)]	0,81 (561)	76 (0,77)	0,86 (181)	0,68 (111)	0,58 (269)

[a)] σ = 1,44 kcal/mol

n entspricht der betrachteten Datenmenge (in Klammern)

Dieser Korrelationswert ist deutlich besser als der mit der Methode von Leven erreichte. Diese Werte können noch weiter für einen Datensatz verbessert werden, bei dem alle Mutationen ausgeschlossen werden, deren Differenz zwischen berechneten und experimentellen Wert größer als die zweifache Standardabweichung von 1,44 kcal/mol ist (s. Tabelle 4.4).

Die Bezeichnung des entwickelten Aminosäure-Atom-Potentials und der von diesem dominierten Gesamtbewertungsfunktion als *hydrophobes Potential* findet eine weitere Bestätigung in den in Tabelle 4.3 und Tabelle 4.4 aufgeführten Ergebnissen (s. Kapitel 4.3.2.1).

Tabelle 4.4: Vorhersage-Ergebnisse der Gesamtbewertungsfunktion für den Testdatensatz in Abhängigkeit von der Lösungsmittelzugänglichkeit

Methode	0-50 % r_{cor}	$rv_{0-50\%}$ (s [kcal/mol])	0-20% r_{cor}	20-40% r_{cor}	40-50% r_{cor}
Gesamtbewertungsfunktion mit Datenausschluss > 3*σ[a)]	0,64 (851)	70 (1,06)	0,72 (361)	0,51 (161)	0,29 (329)
Gesamtbewertungsfunktion mit Datenausschluss > 2*σ[a)]	0,74 (747)	71 (0,75)	0,80 (295)	0,55 (148)	0,43 (304)

[a)] σ = 1,44 kcal/mol

4.4.2.2. Empirisch effektive Energiefunktionen zur Vorhersage der Thermostabilität von Proteinen

Die von Guerois et al [Guerois *et al.*, 2002] entwickelte Methode beruht auf der Verwendung eines EEEF-Potentials. Die von ihnen so benannte FOLD-X-Energiefunktion besteht aus acht energetischen Termen, mit denen die freie Enthalpie berechnet werden soll. Für die Optimierung ihrer Energiefunktion und die Wichtung der einzelnen Parameter zueinander verwendeten sie einen Entwicklungsdatensatz mit 339 Mutationen. Anschließend wurde die entwickelte Methode an einem Testdatensatz aus 625 Mutationen aus 27 verschiedenen Proteinen getestet. Dabei erreichte die auf den Entwicklungsdatensatz optimierte Methode von Guerois für den Entwicklungsdatensatz eine Korrelation von r=0,7 (339 Datenpunkte), mit einer Standardabweichung von σ=0,97 kcal/mol. Nach dem Ausschluss von 5 % aller Daten, die eine Differenz zwischen den berechneten und experimentellen Stabilisierungswerten größer als die zweifache Standardabweichung aufwiesen, stieg die Korrelation auf r=0,8 (323 Datenpunkte), mit einer Standardabweichung von σ=0,75 kcal/mol. Entsprechend erzielte ihre Methode für den Testdatensatz eine Korrelation von r=0,73 (625 Datenpunkte), mit einer Standardabweichung von σ=1,02 kcal/mol. Werden auch hier alle Mutationswerte ausgeschlossen, deren Differenz zwischen den be-

rechneten und experimentellen Daten größer als die zweifache Standardabweichung ist, wird eine Korrelation von r=0,8 (591 Datenpunkte) mit einer Standardabweichung von σ=0,84 kcal/mol erreicht. Die in den Tabellen (s. Tabelle 4.3 und Tabelle 4.4) aufgeführten Ergebnisse für die Gesamtbewertungsfunktion dieser Arbeit sind für eine deutlich größere Datenmenge der Methode von Guerois mindestens gleichwertig, wenn nicht überlegen. Eine weitergehende Diskussion fällt schwer, da die Vorhersage-Ergebnisse in der Arbeit von Guerois nicht weiter nach Kriterien wie z.B. der Lösungsmittelzugänglichkeit untersucht oder die Datensätze näher erläutert werden.

4.5. Schlussfolgerung

Ziel dieser Arbeit war es, einen Ansatz zur automatischen Erstellung eines Mutationsprofils für ein zu thermostabilisierendes Enzym zu entwickeln. Eine solche Vorhersage stellt einen Sortierungsprozess da, der als Ausgangspunkt für ein *Protein Engineering* Experiment dienen kann. Die Methode soll die Auswahl möglicher Versuche einschränken, im besten Fall stellt das angewendete Verfahren das einzige Kriterium für die Durchführung eines thermostabilisierenden Experimentes da. Die Grundlage für eine solche Anwendung wird in dieser Arbeit aus den erreichten Vorhersage-Ergebnissen für experimentelle Stabilitätsdaten, die aus einem Entwicklungsdatensatz und einem Testdatensatz stammen, abgeleitet. Die angewandten Kriterien zur Bewertung der Qualität der vorgestellten Bewertungsfunktion waren die richtige Vorhersage, der Korrelationskoeffizient, die Spezifität und die Sensibilität. Ein zusätzliches Kriterium für eine Anwendung wäre die benötigten Zeitdauer oder die entsprechend gebrauchte Rechenzeit für die Erstellung eines Mutationsprofils.

Die Sensibilität stellt das entscheidende Kriterium für die Verwendung der Methode als Vorfilter dar. Sie gibt das Verhältnis der als richtig erkannten stabilisierenden Mutationen zu den nicht erkannten stabilisierenden Mutationen (den falsch negativen) wieder. Je näher ihr Wert bei eins liegt, desto besser ist die Güte der Methode. Die in dieser Arbeit erreichten Sensibilitäten von 0,96 (646 Daten des Entwicklungsdatensatzes) und 0,93 (918 Daten des Testdatensat-

zes) kann als sehr gut klassifiziert werden. Für eine Vorhersagefunktion, die als Vorfilter dienen und den Arbeitsaufwand so gering wie möglich halten soll, muss die niedrige Spezifität mit 0,32 (646 Daten des Entwicklungsdatensatzes) und 0,45 (918 Daten des Testdatensatzes) noch verbessert werden.

In dieser Arbeit wurde erstmalig ein richtungs- und abstandsabhängiges Potential für die Vorhersage der Thermostabilität entwickelt und beschrieben. Die Optimierung und die Validierung sowie die systematische Analyse der Vorhersage-Ergebnisse der Bewertungsfunktion wurde an der bisher größten dafür verwendeten Stabilitätsdatensammlung durchgeführt. Die erreichten Korrelationskoeffizienten und richtigen Vorhersagen sind den bisher in Literatur beschriebenen Methoden gleichwertig oder überlegen (s. Kapitel 4.4). Einige Mutationen zeigen deutliche Abweichungen von der Standardabweichung (s. Kapitel 3.2.3 und Kapitel 3.2.4). Wurden diese Mutationen ausgeschlossen, verbesserten sich die Ergebnisse drastisch. Unter anderem aufgrund der Vielzahl dieser Mutationen (mehr als 200) war es in dieser Arbeit nicht möglich, eindeutige Gründe für dieses Verhalten zu finden. Die Unterschiede werden aber vermutlich auf strukturelle Veränderungen oder kooperative Wechselwirkungen zurückgehen. Diese Wechselwirkungen beruhen auf Vielkörperinteraktionen, die in dem angewandten Modell nicht berücksichtigt werden. Eine Strukturänderung wird mit dem hier vorgestellten Modell nicht vorhergesagt, und eine tatsächliche Strukturänderung kann den experimentellen Daten nicht entnommen werden.

4.5.1. Ausblick

Das richtungsabhängige Aminosäure-Atom-Potential hat noch nicht zur erhofften Beschreibung der physikalischen Wechselwirkungen geführt, die im Kapitel 2.1.1.1 dargestellt sind, wie zwischen polaren funktionellen Gruppen, aromatischen Systemen oder Wasserstoffbrückenbindungen, deren Existenz und Stärke von ihrer Entfernung und ihrer gegenseitigen relativen Orientierung abhängen.

Mögliche Verbesserungen wären:

- Die Einführung neuer Atomtypen, die mögliche oder tatsächliche umgebungsbeschreibende Parameter enthalten (z. B. Sekundärstrukturelementzugehörigkeit oder Wasserstoffbrückenbindungs-Akzeptor oder -Donator)

- Die Beschreibung seltener Atomtypen und Ableitung entsprechender Potentiale, z.B. für S, Br oder Cl

Die Sensiblität erreichte sehr gute Werte für die Vorhersage der Thermostabilität hinsichtlich thermostabilisierender Mutationen. Die in dieser Arbeit entwickelte Potentialfunktion stellt damit einen zufriedenstellenden Vorfilter für *Protein Engineering* Versuche dar. Eine noch nötige Verbesserung der Spezifität könnte durch eine Energieminimierung der eingesetzten Aminosäure im Hinblick auf die Torsionswinkelpotentiale oder durch einen Kraftfeld-Ansatz erfolgen.

5. Zusammenfassung

In der vorliegenden Arbeit wurde eine wissensbasierte Bewertungsfunktion für die Vorhersage der Thermostabilität entwickelt und getestet. Diese Methode stellt einen Ansatz zur automatischen Erstellung eines Mutationsprofils für ein zu thermostabilisierendes Enzym dar.

Die verwendete Bewertungsfunktion quantifiziert die Umgebung einer Aminosäure und bewertet die Wirkung eines Aminosäureaustausches. Diese Bewertungsfunktion setzt sich aus zwei wissensbasierten Potentialfunktionen zusammen, die eine Beschreibung von nicht-lokalen Wechselwirkungen und eine sequentiell lokale Wechselwirkungsbeschreibung, ermöglichen. Die nicht-lokalen Interaktionen werden mit einem Aminosäure-Atom-Paarpotential erfasst, das eine richtungs- und abstandsabhängige atomare Beschreibung einer Aminosäureumgebung, basierend auf fünf definierten Atomtypen, liefert. Eine sequentiell lokale Wechselwirkungsbeschreibung wird durch ein Torsionswinkelpotential erreicht. Aus der Kombination dieser beiden Potentiale wird die wissensbasierte Bewertungsfunktion erhalten.

Die Vorhersagefunktion wurde an einem durch Literaturrecherche möglichst großen Satz an experimentellen Daten (646 Stabilitätsdaten von elf Proteinen) validiert und optimiert sowie ihre statistische Auswertung durchgeführt. Die Vorhersageleistung der optimierten Bewertungspotentialfunktion wurde an dem größten bisher verwendeten Testdatensatz mit 918 Mutationen aus 27 Proteinen überprüft.

Die Vorhersagequalität hing in großem Maße von der Qualität der experimentellen Daten ab, deren systematische oder unsystematische Fehler selten angegeben und schwer zu schätzen waren. Die im Datenmaterial vorliegenden Schwankungen wurden bei der detaillierten Analyse der Vorhersage-Ergebnisse deutlich.

Die in dieser Arbeit entwickelte Methode zur Erstellung von Mutationsprofilen für mögliche Kandidaten eines Experiments mit dem Ziel der Thermostabilitätserhöhung von Proteinen erzielt hinsichtlich ihres Einsatzes als möglicher Vorfilter oder als alleiniges Kriterium für das *Protein Engineering* gute Ergebnisse.

6. Anhang

6.1. Strukturdatensatz

Angabe der 282 Strukturen, aus denen das Aminosäure-Atom-Potential und das Torsionswinkelpotential angeleitet wurden. Die Kürzel entsprechen den Identifizierungscodes der Proteindatenbank (PDB) [Bernstein *et al.*, 1978].

1aap	1cfy	1gow	1lbe	1szt	2rhe	1dlc	1le4	1wba
1aar	1cgj	1gpr	1lcl	1tbg	2spc	1ede	1lis	1wbc
1ab0	1chk	1gta	1leh	1tdt	2trh	1edg	1lit	1wer
1abq	1cka	1gto	1leo	1tfe	2vab	1enh	1lki	1xnc
1acf	1cns	1hhl	1lgy	1tfg	3app	1enj	1mil	1xnd
1ade	1col	1hia	1lmk	1tmc	3blm	1er8	1mla	1ypb
1ag8	1cpj	1hij	1ltb	1udi	3rp2	1eur	1mpb	1zon
1anu	1ctn	1hin	1mif	1vad	3ssi	1fas	1mzl	1zym
1apa	1cyw	1hjr	1mml	1vhi	3wrp	1ffd	1nar	2cpl
1avp	1dco	1hrt	1msp	1vie	3xin	1fnf	1noa	2era
1avy	1dei	1htm	1mss	1vpn	4ake	1gbs	1ntn	2exo
1bar	1dfn	1htr	1myk	1vqd	4gpd	1ghr	1odd	2hvm
1bec	1dol	1hul	1nap	1vsc	4mon	1gln	1orc	2liv
1bed	1dsu	1hyl	1ncf	1whi	4pgm	1gnd	1plr	2ovo
1beo	1dup	1icw	1oib	1who	6rlx	1gob	1psn	2pcy
1bet	1dyn	1idm	1onr	1xkj	7ahl	1gsd	1ptd	2pec
1bhs	1ecp	1ift	1opa	1xyz	1aaj	1gzi	1ptf	2plc
1bkl	1edt	1igp	1opg	1ycq	1agi	1hcv	1ptx	2sfa
1ble	1egp	1iib	1plg	1ydv	1arb	1hey	1pzc	2sil
1bms	1elp	1ilk	1pmk	1ypi	1arl	1hoe	1r69	3pte
1bnf	1epa	1ilr	1pp2	1ypr	1ayd	1hva	1rbx	4rnt
1boy	1erw	1jkw	1qil	1ypt	1bas	1hyp	1reg	1dpd
1bpl	1esc	1jud	1rfb	1zik	1bmg	1iad	1rkm	1stm
1brb	1f36	1kba	1ris	2chs	1bre	1idk	1rpl	2cas
1bro	1f3g	1klo	1rsu	2cro	1cem	1ifc	1sfe	2rmu
1bth	1fle	1kob	1rve	2fb4	1chd	1igd	1thv	4rhv
1bv1	1for	1kpc	1sei	2gmf	1clp	1iib	1thx	
1cbg	1fps	1kpt	1sip	2hnt	1cms	1jpo	1tmy	
1cdc	1ftp	1kra	1sjs	2nck	1cnv	1kaa	1u9a	
1cdh	1fvj	1kve	1smn	2not	1csk	1kid	1vcc	
1cei	1gfl	1kxi	1sph	2nul	1csp	1kte	1vin	
1cfr	1ghs	1l01	1sry	2pol	1dkz	1kxf	1vmo	

6.2. Entwicklungsdatensatz

Der Entwicklungsdatensatz liegt als Textdatei auf CDROM im Institut für Biochemie vor.

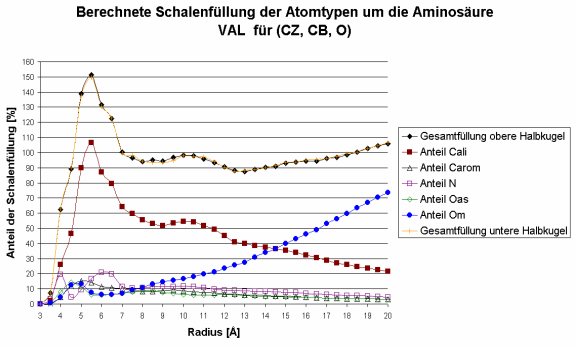
6.3. Testdatensatz

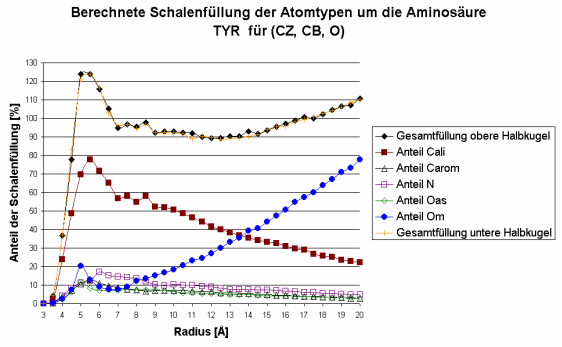
Der Testdatensatz liegt als Textdatei auf CDROM im Institut für Biochemie vor.

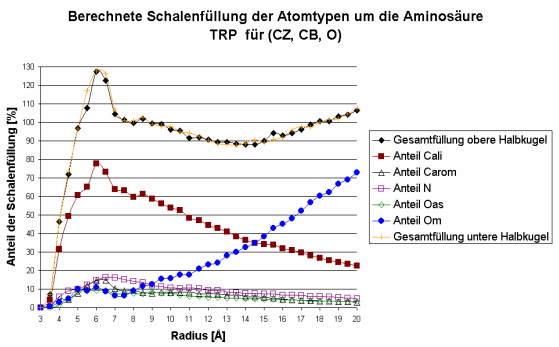
6.4. Das Aminosäure-Atom-Potential

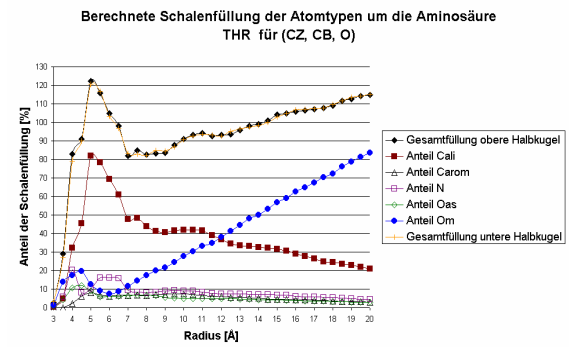
6.4.1. Füllungsgrad der Umgebungsschalen

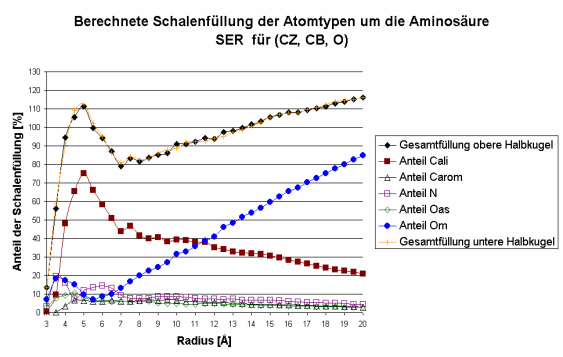
Die Abbildungen zeigen den Füllungsgrad der Umgebungsschalen für die fünf Atomtypen in Abhängigkeit von den zwanzig Aminosäuren (s. Kapitel 2.2 und Kapitel 3.2.1.9).

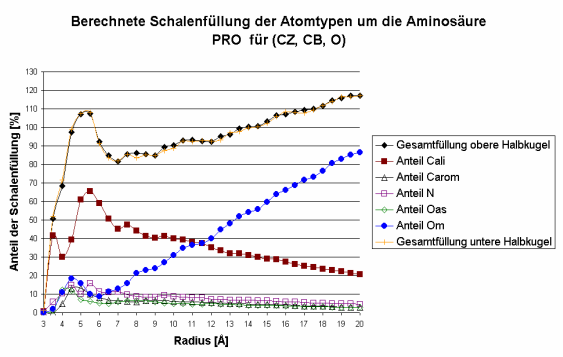


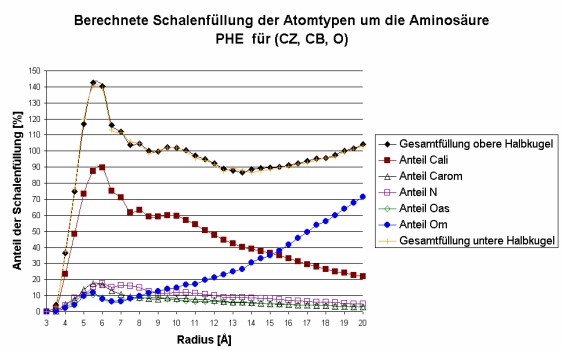


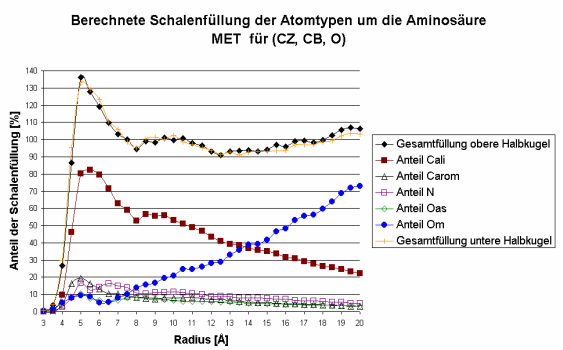


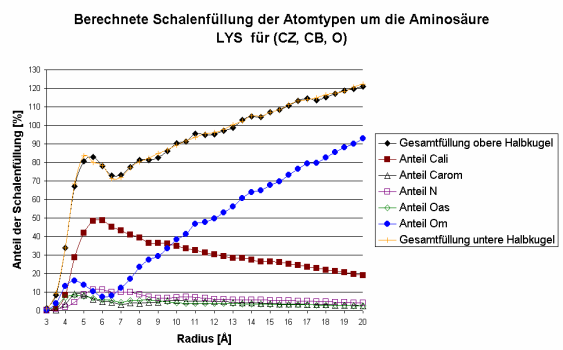


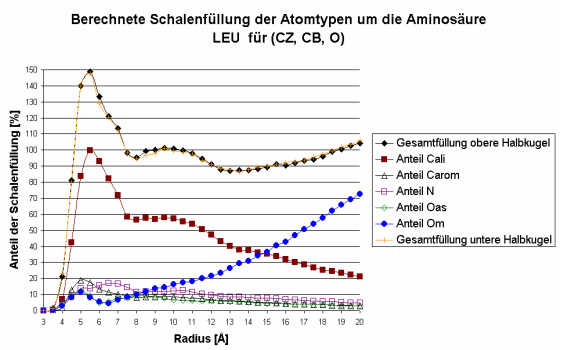

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
SER für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

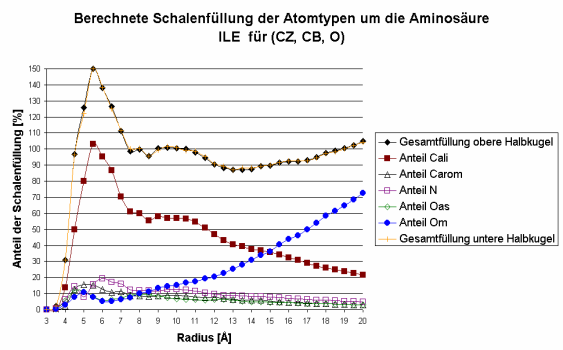

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
PRO für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

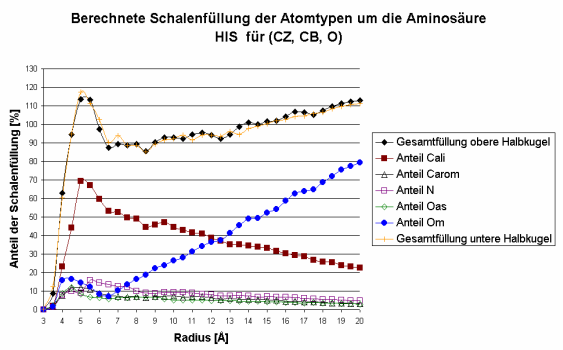

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
PHE für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

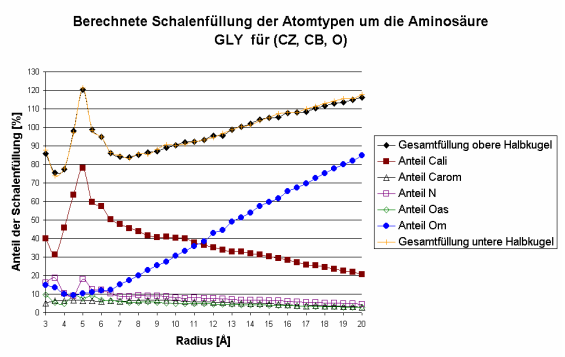

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
MET für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

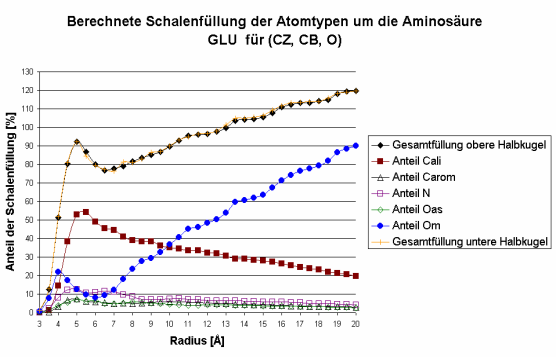

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
LYS für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

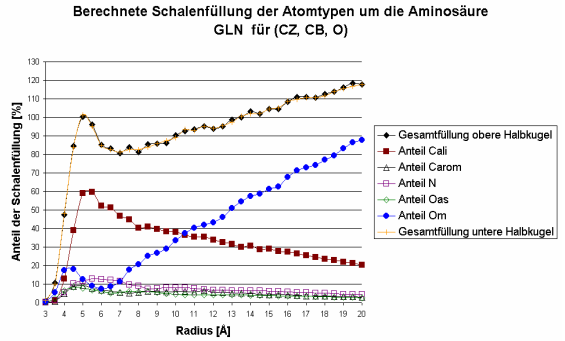

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
LEU für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

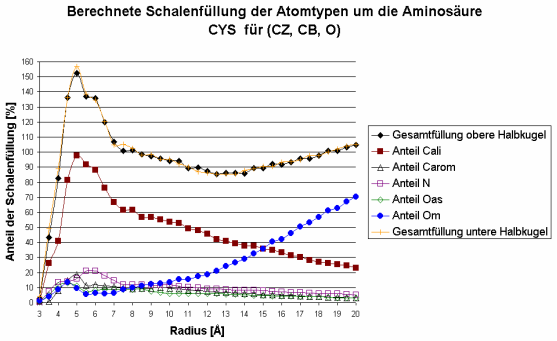

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
ILE für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

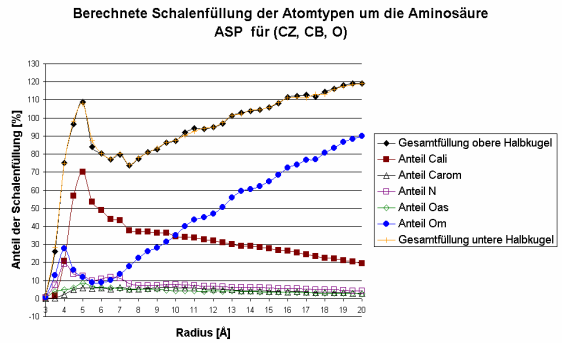

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
HIS für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

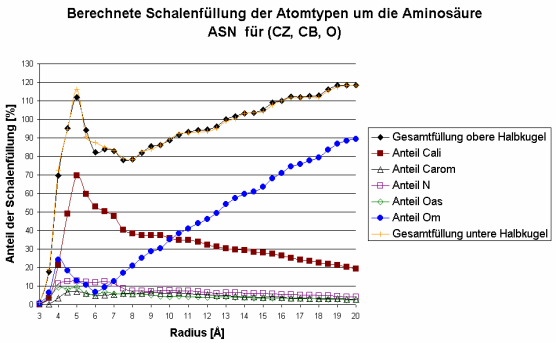

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
GLY für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

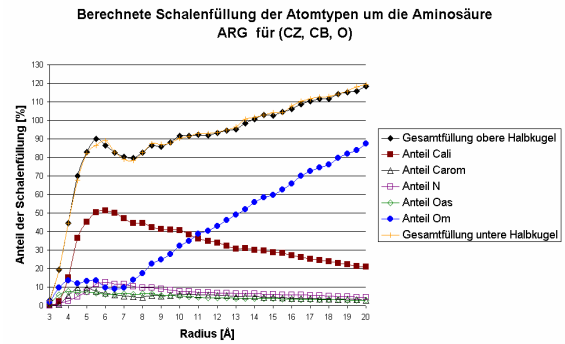

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
GLU für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

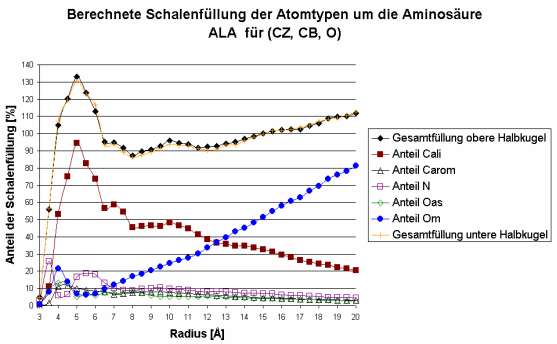

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
GLN für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel


Berechnete Schalenfüllung der Atomtypen um die Aminosäure
CYS für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel


Berechnete Schalenfüllung der Atomtypen um die Aminosäure
ASP für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

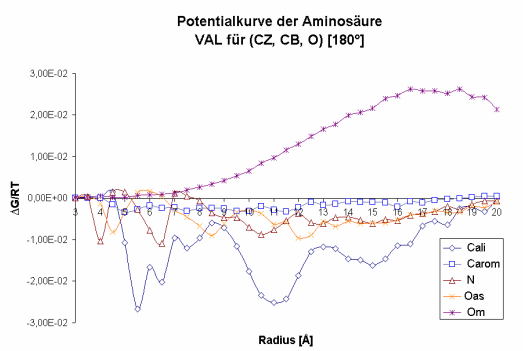

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
ASN für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

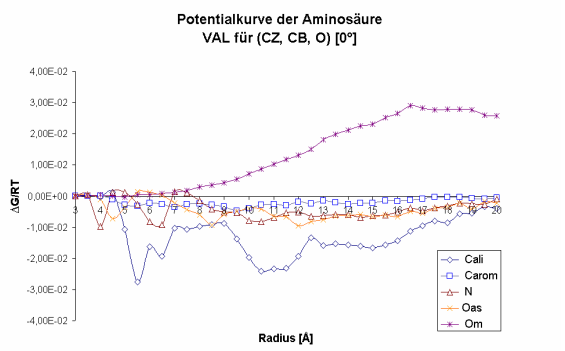

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
ARG für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

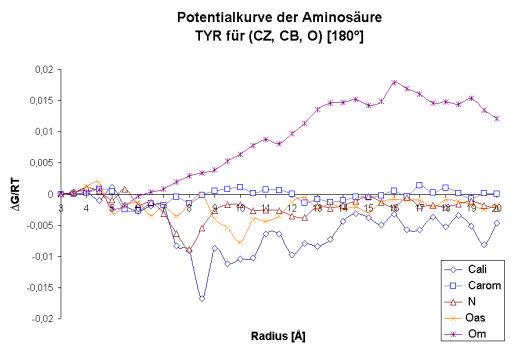

Berechnete Schalenfüllung der Atomtypen um die Aminosäure
ALA für (CZ, CB, O)
Anteil der Schalenfüllung [%]
Radius [Å]
Gesamtfüllung obere Halbkugel
Anteil Cali
Anteil Carom
Anteil N
Anteil Oas
Anteil Om
Gesamtfüllung untere Halbkugel

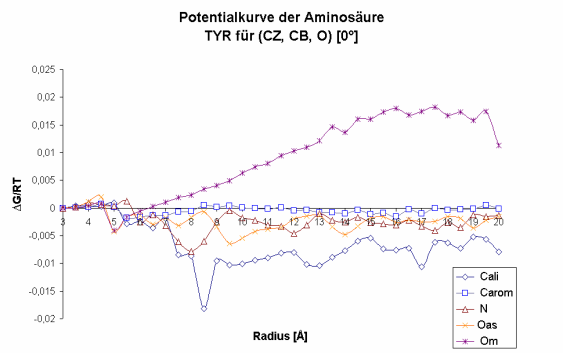
6.4.2. Potentialkurven für das Aminosäure-Atom-Paarpotential

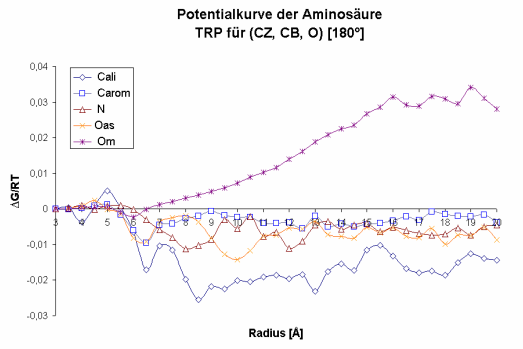
Die Abbildungen zeigen die Aminosäure-Atom-Potentialkurven für die gewählten Bedingungen (CZ,CB,O). Die Angabe [O°] bezeichnet die obere Halbkugel und die Angabe [180°] die untere Halbkugel (s. Abbildung 3.1 und Kapitel 3.2.1.10).

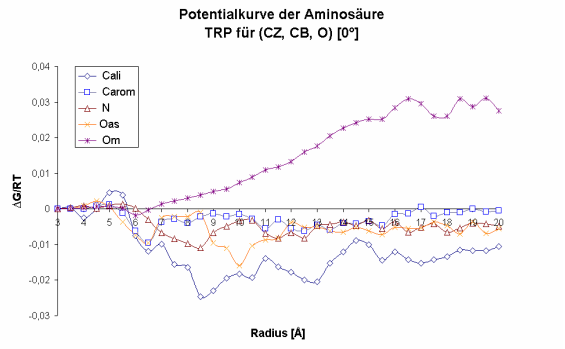


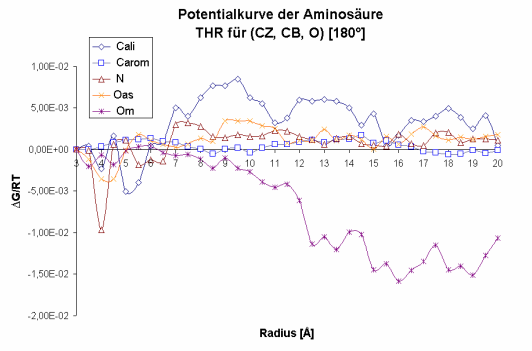


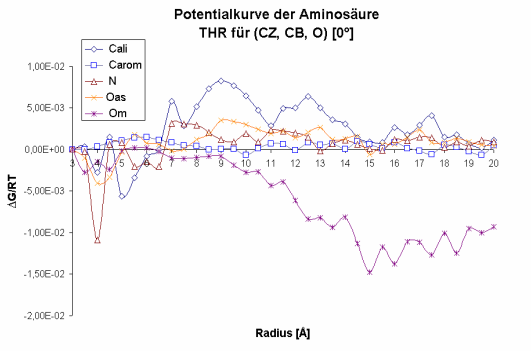


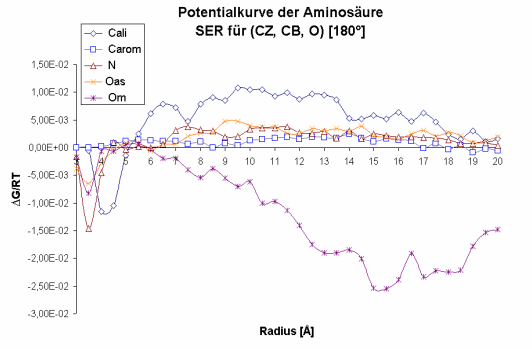


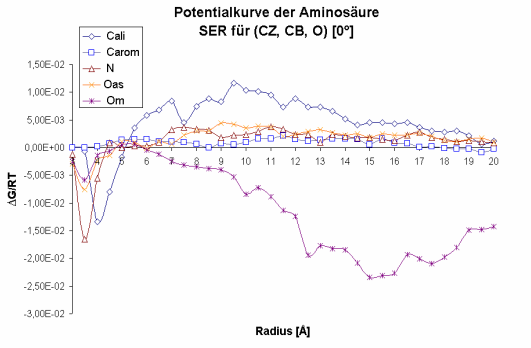


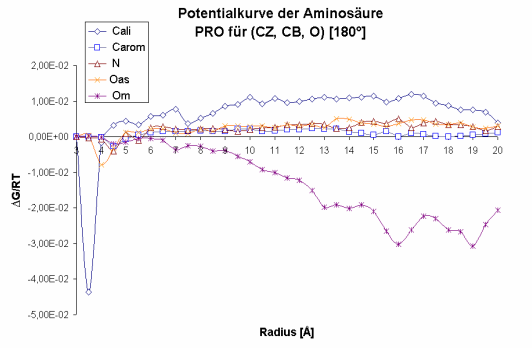


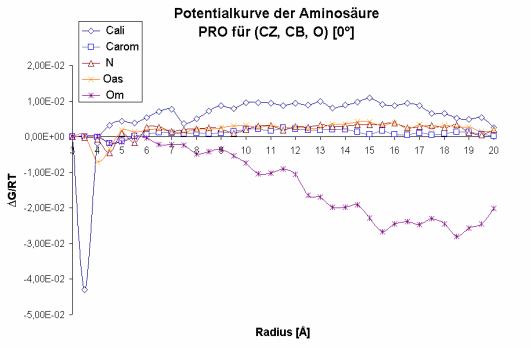

Potentialkurve der Aminosäure
THR für (CZ, CB, O) [180°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]

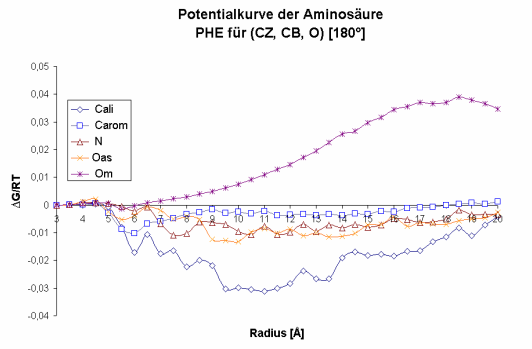

Potentialkurve der Aminosäure
THR für (CZ, CB, O) [0°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]

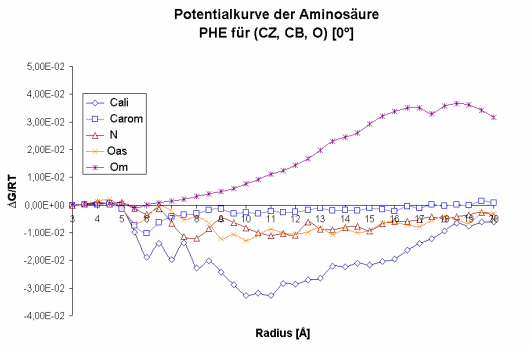

Potentialkurve der Aminosäure
SER für (CZ, CB, O) [180°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]

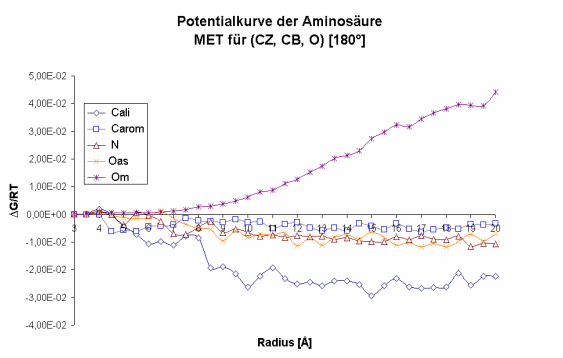

Potentialkurve der Aminosäure
SER für (CZ, CB, O) [0°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]

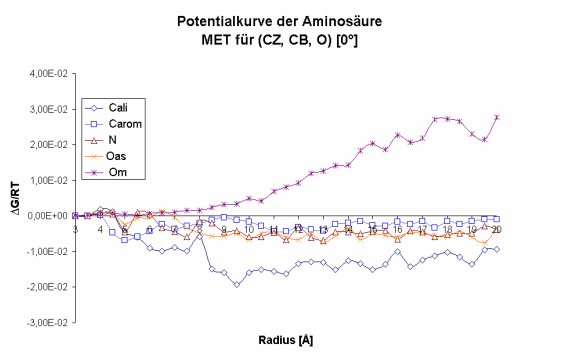

Potentialkurve der Aminosäure
PRO für (CZ, CB, O) [180°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]

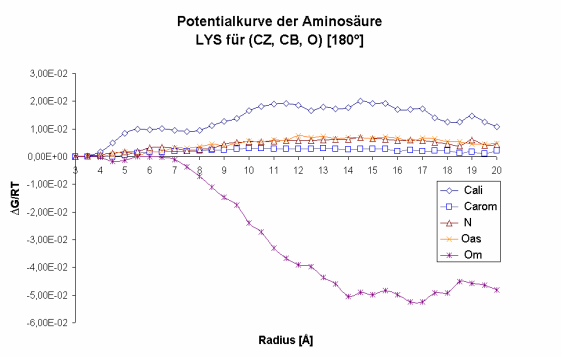

Potentialkurve der Aminosäure
PRO für (CZ, CB, O) [0°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]

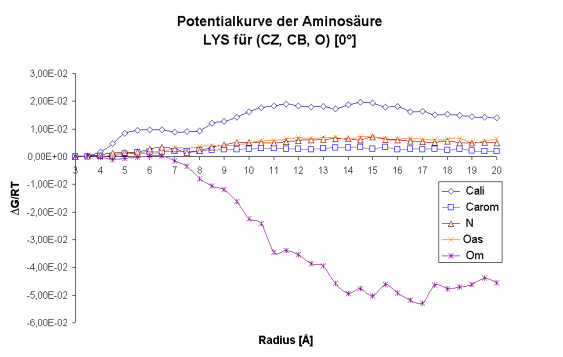

Potentialkurve der Aminosäure
PHE für (CZ, CB, O) [180°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]

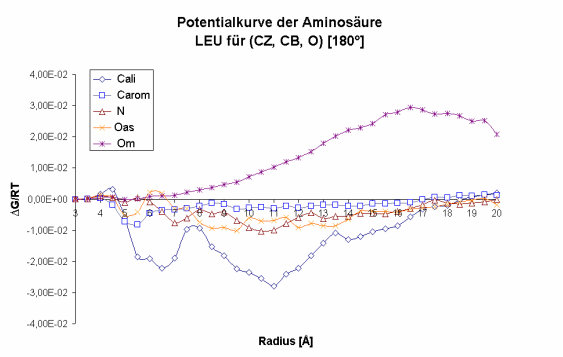

Potentialkurve der Aminosäure
PHE für (CZ, CB, O) [0°]
Cali
Carom
N
Oas
Om
ΔG/RT
Radius [Å]

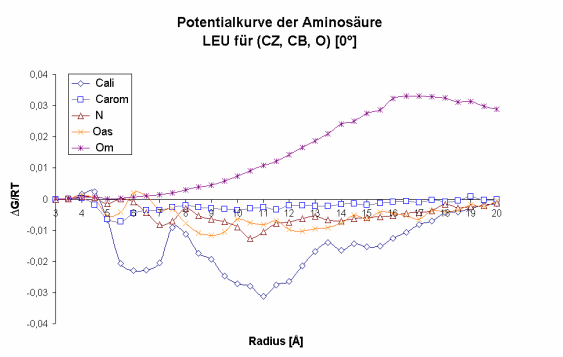

Potentialkurve der Aminosäure
MET für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

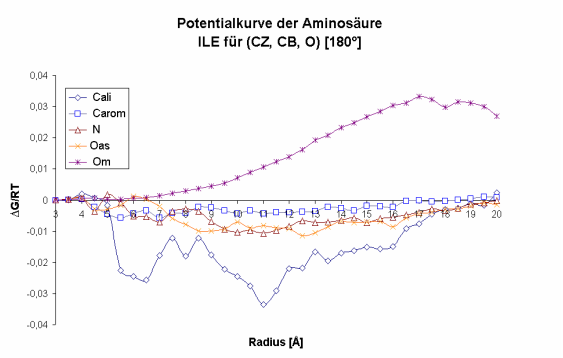

Potentialkurve der Aminosäure
MET für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

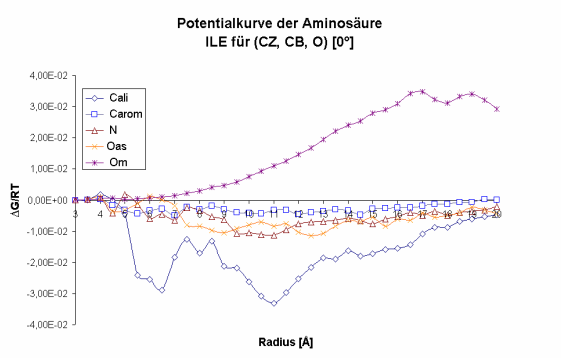

Potentialkurve der Aminosäure
LYS für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

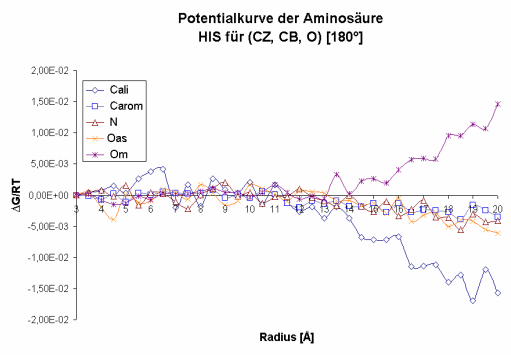

Potentialkurve der Aminosäure
LYS für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

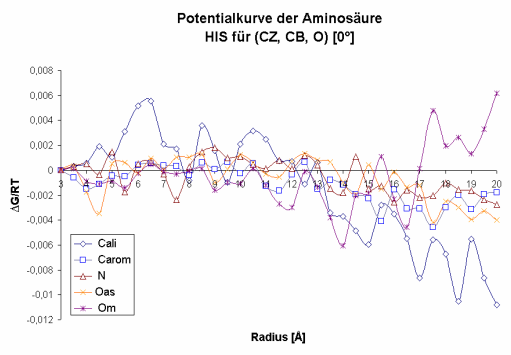

Potentialkurve der Aminosäure
LEU für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

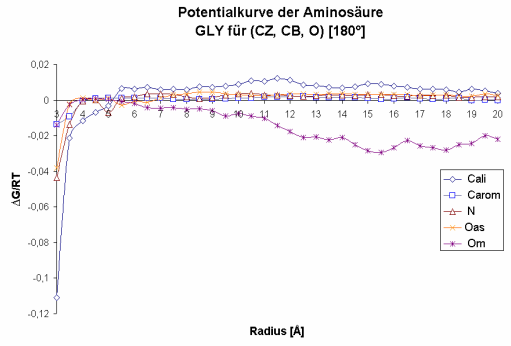

Potentialkurve der Aminosäure
LEU für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

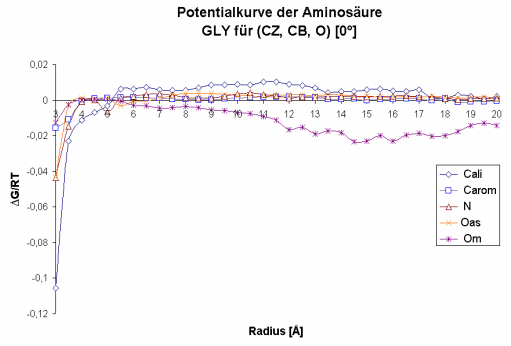

Potentialkurve der Aminosäure
ILE für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

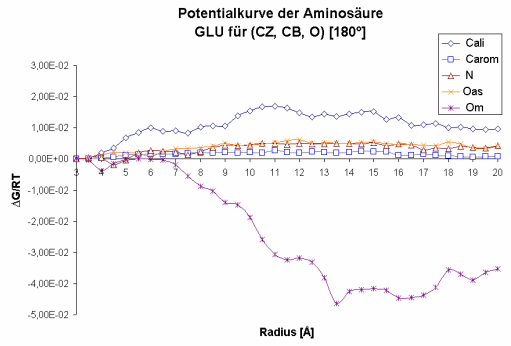

Potentialkurve der Aminosäure
ILE für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

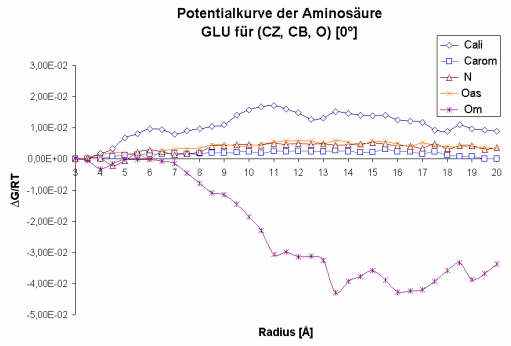

Potentialkurve der Aminosäure
HIS für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

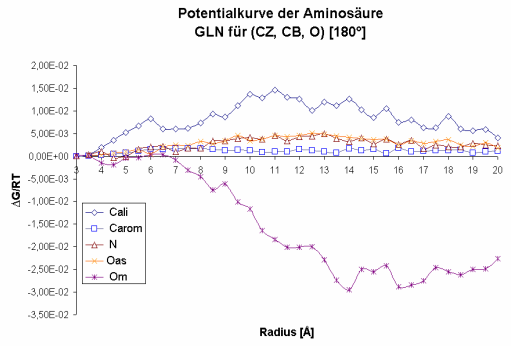

Potentialkurve der Aminosäure
HIS für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

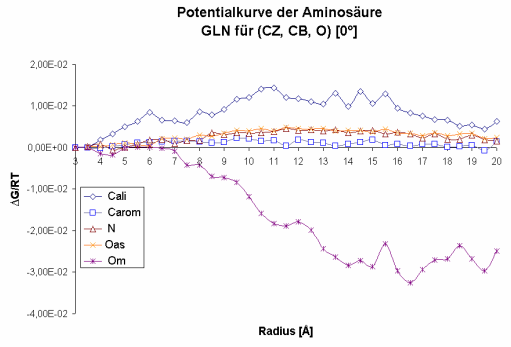

Potentialkurve der Aminosäure
GLY für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

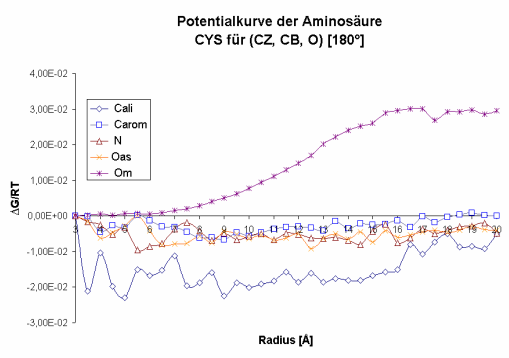

Potentialkurve der Aminosäure
GLY für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

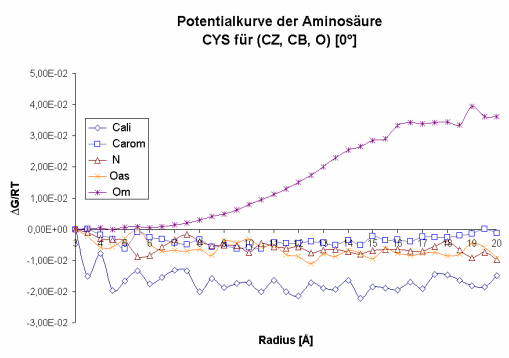

Potentialkurve der Aminosäure
GLU für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

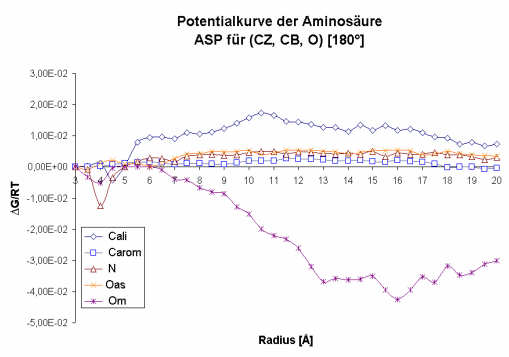

Potentialkurve der Aminosäure
GLU für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

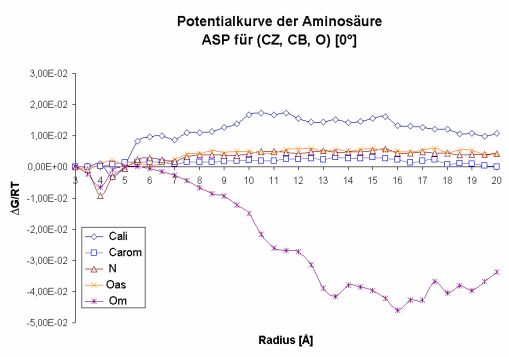

Potentialkurve der Aminosäure
GLN für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

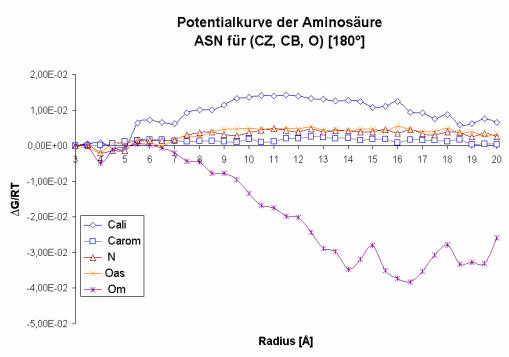

Potentialkurve der Aminosäure
GLN für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

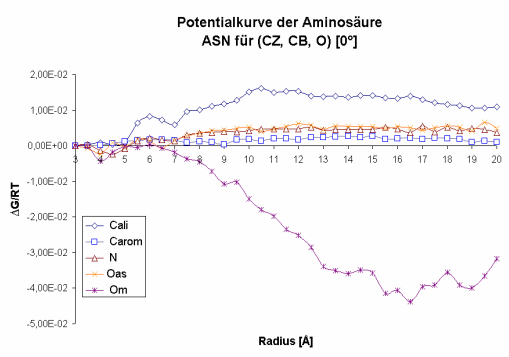

Potentialkurve der Aminosäure
CYS für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

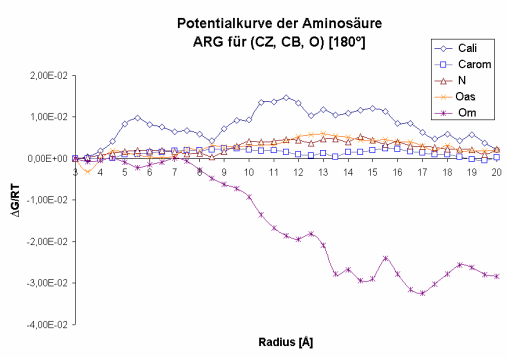

Potentialkurve der Aminosäure
CYS für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

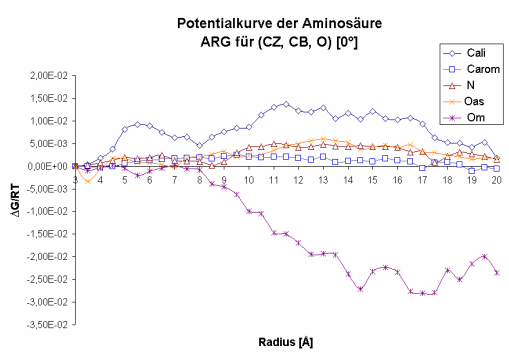

Potentialkurve der Aminosäure
ASP für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

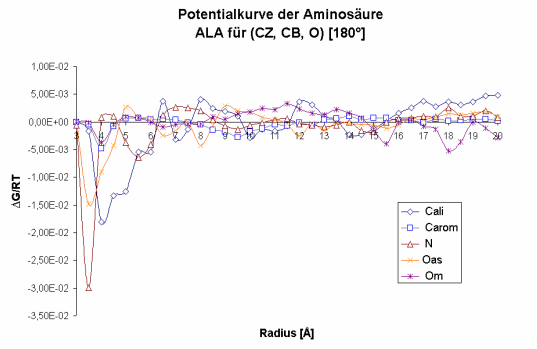

Potentialkurve der Aminosäure
ASP für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om

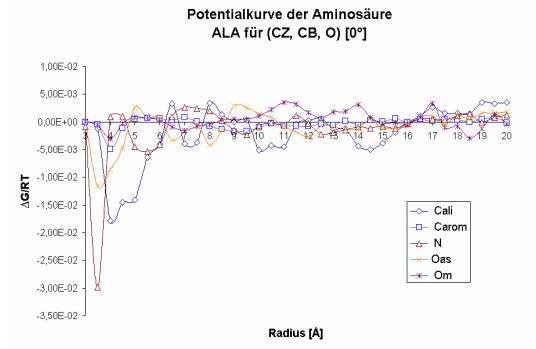

Potentialkurve der Aminosäure
ASN für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om


Potentialkurve der Aminosäure
ASN für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om


Potentialkurve der Aminosäure
ARG für (CZ, CB, O) [180°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om


Potentialkurve der Aminosäure
ARG für (CZ, CB, O) [0°]
ΔG/RT
Radius [Å]
Cali
Carom
N
Oas
Om





6.4.3. Die quantitative Analyse der Potentialkurven

Die Korrelationsmatrizen der Potentialkurven für die fünf Atomtypen werden in den folgenden Tabellen aufgeführt.

GLU	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,840	0,884	0,874	-0,651
C_{arom}	0,840	1,000	0,690	0,739	-0,388
N	0,884	0,690	1,000	0,836	-0,767
O_{as}	0,874	0,739	0,836	1,000	-0,780
O_m	-0,651	-0,388	-0,767	-0,780	1,000

TYR	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	-0,076	0,655	0,437	-0,357
C_{arom}	-0,076	1,000	0,140	-0,055	0,081
N	0,655	0,140	1,000	0,201	-0,166
O_{as}	0,437	-0,055	0,201	1,000	-0,140
O_m	-0,357	0,081	-0,166	-0,140	1,000

ASN	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,724	0,921	0,884	-0,618
C_{arom}	0,724	1,000	0,646	0,637	-0,539
N	0,921	0,646	1,000	0,901	-0,728
O_{as}	0,884	0,637	0,901	1,000	-0,763
O_m	-0,618	-0,539	-0,728	-0,763	1,000

HIS	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,748	0,753	0,738	-0,739
C_{arom}	0,748	1,000	0,578	0,685	-0,456
N	0,753	0,578	1,000	0,607	-0,649
O_{as}	0,738	0,685	0,607	1,000	-0,701
O_m	-0,739	-0,456	-0,649	-0,701	1,000

TRP	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,424	0,791	0,653	-0,358
C_{arom}	0,424	1,000	0,330	0,446	-0,036
N	0,791	0,330	1,000	0,315	-0,362
O_{as}	0,653	0,446	0,315	1,000	-0,334
O_m	-0,358	-0,036	-0,362	-0,334	1,000

VAL	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,676	0,557	0,211	0,004
C_{arom}	0,676	1,000	0,226	0,286	0,382
N	0,557	0,226	1,000	0,134	-0,141
O_{as}	0,211	0,286	0,134	1,000	-0,293
O_m	0,004	0,382	-0,141	-0,293	1,000

ARG	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,609	0,860	0,767	-0,372
C_{arom}	0,609	1,000	0,397	0,420	0,045
N	0,860	0,397	1,000	0,782	-0,619
O_{as}	0,767	0,420	0,782	1,000	-0,635
O_m	-0,372	0,045	-0,619	-0,635	1,000

LEU	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,623	0,798	0,603	0,196
C_{arom}	0,623	1,000	0,271	0,301	0,487
N	0,798	0,271	1,000	0,684	-0,085
O_{as}	0,603	0,301	0,684	1,000	-0,111
O_m	0,196	0,487	-0,085	-0,111	1,000

MET	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,435	0,862	0,874	-0,640
C_{arom}	0,435	1,000	0,395	0,340	-0,267
N	0,862	0,395	1,000	0,813	-0,712
O_{as}	0,874	0,340	0,813	1,000	-0,701
O_m	-0,640	-0,267	-0,712	-0,701	1,000

ALA	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,401	0,115	0,304	0,020
C_{arom}	0,401	1,000	-0,010	0,352	0,041
N	0,115	-0,010	1,000	0,570	-0,071
O_{as}	0,304	0,352	0,570	1,000	0,183
O_m	0,020	0,041	-0,071	0,183	1,000

PHE	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,382	0,857	0,846	-0,069
C_{arom}	0,382	1,000	0,203	0,187	0,423
N	0,857	0,203	1,000	0,742	-0,204
O_{as}	0,846	0,187	0,742	1,000	-0,303
O_m	-0,069	0,423	-0,204	-0,303	1,000

THR	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	-0,282	0,549	0,557	-0,147
C_{arom}	-0,282	1,000	-0,142	-0,168	0,149
N	0,549	-0,142	1,000	0,521	-0,183
O_{as}	0,557	-0,168	0,521	1,000	-0,178
O_m	-0,147	0,149	-0,183	-0,178	1,000

CYS	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,029	-0,145	-0,165	0,198
C_{arom}	0,029	1,000	0,216	0,473	0,246
N	-0,145	0,216	1,000	0,419	-0,361
O_{as}	-0,165	0,473	0,419	1,000	-0,372
O_m	0,198	0,246	-0,361	-0,372	1,000

GLY	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,879	0,975	0,550	-0,112
C_{arom}	0,879	1,000	0,910	0,644	-0,101
N	0,975	0,910	1,000	0,594	-0,127
O_{as}	0,550	0,644	0,594	1,000	-0,067
O_m	-0,112	-0,101	-0,127	-0,067	1,000

PRO	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,361	0,464	0,415	-0,387
C_{arom}	0,361	1,000	0,608	0,547	-0,197
N	0,464	0,608	1,000	0,739	-0,607
O_{as}	0,415	0,547	0,739	1,000	-0,565
O_m	-0,387	-0,197	-0,607	-0,565	1,000

LYS	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,920	0,966	0,931	-0,797
C_{arom}	0,920	1,000	0,883	0,871	-0,650
N	0,966	0,883	1,000	0,944	-0,865
O_{as}	0,931	0,871	0,944	1,000	-0,896
O_m	-0,797	-0,650	-0,865	-0,896	1,000

GLN	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,576	0,867	0,867	-0,471
C_{arom}	0,576	1,000	0,433	0,467	-0,058
N	0,867	0,433	1,000	0,896	-0,630
O_{as}	0,867	0,467	0,896	1,000	-0,676
O_m	-0,471	-0,058	-0,630	-0,676	1,000

SER	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,434	0,535	0,677	-0,216
C_{arom}	0,434	1,000	0,339	0,358	-0,041
N	0,535	0,339	1,000	0,847	-0,212
O_{as}	0,677	0,358	0,847	1,000	-0,355
O_m	-0,216	-0,041	-0,212	-0,355	1,000

ASP	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,724	0,790	0,803	-0,578
C_{arom}	0,724	1,000	0,517	0,591	-0,387
N	0,790	0,517	1,000	0,671	-0,559
O_{as}	0,803	0,591	0,671	1,000	-0,750
O_m	-0,578	-0,387	-0,559	-0,750	1,000

ILE	C_{ali}	C_{arom}	N	O_{as}	O_m
C_{ali}	1,000	0,794	0,821	0,532	0,132
C_{arom}	0,794	1,000	0,588	0,446	0,337
N	0,821	0,588	1,000	0,677	-0,122
O_{as}	0,532	0,446	0,677	1,000	-0,219
O_m	0,132	0,337	-0,122	-0,219	1,000

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Adams, M. W. und R. M. Kelly (1998). Finding and using hyperthermophilic enzymes. *Trends Biotechnol.*, **16**(8), 329-332.

Alber, T., D. P. Sun, K. Wilson, J. A. Wozniak, S. P. Cook und B. W. Matthews (1987). Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme. *Nature*, **330**(6143), 41-46.

Alm, E. und D. Baker (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. U S A*, **96**(20), 11305-11310.

Anfinsen, C. B. (1972). Studies on the principles that govern the folding of protein chains. *Nobel Lecture*.

auf der Heyde, T. P. E. (1990). Analysing Chemical Data in More Than Two Dimensions. *J. chem. Education*, **67**(6), 461-469.

Ben-Naim, A. (1987). Solvation Thermodynamics. New York, Plenum Press.

Ben-Naim, A. (1997). Statistical potentials extracted from protein structures: Are these meaningful potentials ? *J. Chem. Phys.*, **107**(9), 3698-3706.

Bernstein, F. C., T. F. Koetzle, G. J. Williams, E. F. J. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi und M. Tasumi (1978). The protein data bank: a computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.*, **185**(2), 584-591.

Böhm, H.-J., G. Klebe und H. Kubinyi (1996). Wirkstoffdesign. Heidelberg-Berlin-Oxford, Spektrum

Bowie, J. U., R. Luthy und D. Eisenberg (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**(5016), 164-170.

Bruins, M. E., A. E. Janssen und R. M. Boom (2001). Thermozymes and their applications: a review of recent literature and patents. *Appl. Biochem. Biotechnol.*, **90**(2), 155-186.

Bryant, S. H. und C. E. Lawrence (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins*, **16**(1), 92-112.

Casari, G. und M. J. Sippl (1992). Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.*, **224**(3), 725-732.

Chakraborty, S., V. Ittah, P. Bai, L. Luo, E. Haas und Z. Peng (2001). Structure and dynamics of the alpha-lactalbumin molten globule: fluorescence studies using proteins containing a single tryptophan residue. *Biochemistry*, **40**(24), 7228-7238.

Chothia, C. (1984). Principles that determine the structure of proteins. *Annu. Rev. Biochem.*, **53**, 537-572.

Chothia, C. und A. V. Finkelstein (1990). The classification and origins of protein folding patterns. *Annu. Rev. Biochem.*, **59**, 1007-1039.

Cole, J. C., R. Taylor und M. L. Verdonk (1998). Directional preferences of intermolecular contacts to hydrophobic groups. *Acta Crystallogr. D. Biol. Crystallogr.*, **54**(1 (Pt 6)), 1183-1193.

Colonna-Cesari, F. und C. Sander (1990). Excluded volume approximation to protein-solvent interaction. The solvent contact model. *Biophys. J.*, **57**(5), 1103-1107.

Dao-pin, S., D. E. Anderson, W. A. Baase, F. W. Dahlquist und B. W. Matthews (1991). Structural and thermodynamic consequences of burying a charged residue within the hydrophobic core of T4 lysozyme. *Biochemistry*, **30**(49), 11521-11529.

Dengler, U. (1998). Kristallstruktur der D-2-Hydroxyisocaproat-Dehydrogenase aus *Lactobacillus casei*: Verfeinerung, Interpretation und Anwendung in einem Verfahren zur Erkennung der Proteinfaltung. Dissertation. Technische Universität Carolo-Wilhelmina zu Braunschweig.

Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry*, **24**(6), 1501-1509.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**(31), 7133-7155.

Dill, K. A. (1999). Polymer principles and protein folding. *Protein Sci.*, **8**(6), 1166-1180.

Dinner, A. R., A. Sali, L. J. Smith, C. M. Dobson und M. Karplus (2000). Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.*, **25**(7), 331-339.

Duan, Y. und P. A. Kollman (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, **282**(5389), 740-744.

Duan, Y., L. Wang und P. A. Kollman (1998). The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. *Proc. Natl. Acad. Sci. U S A*, **95**(17), 9897-9902.

Edison, A. S. (2001). Linus Pauling and the planar peptide bond. *Nat. Struct. Biol.*, **8**(3), 201-202.

Ellis, R. J. und F. U. Hartl (1999). Principles of protein folding in the cellular environment. *Curr. Opin. Struct. Biol.*, **9**(1), 102-110.

Erwin, C. R., B. L. Barnett, J. D. Oliver und J. F. Sullivan (1990). Effects of engineered salt bridges on the stability of subtilisin BPN'. *Protein Eng.*, **4**(1), 87-97.

Fersht, A. R. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl. Acad. Sci. U S A*, **92**(24), 10869-10873.

Finkelstein, A. V. (1997). Protein structure: what is it possible to predict now? *Curr. Opin. Struct. Biol.*, **7**(1), 60-71.

Finkelstein, A. V. und A. M. Gutin (1995). Why do protein architectures have Boltzmann-like statistics? *Proteins*, **23**(2), 142-150.

Fukui, S., S. Ikeda, M. Fujimura, H. Yamada und H. Kumagai (1975). Comparative studies on the properties of tryptophanase and tyrosine phenol-lyase immobilized directly on Sepharose or by use of Sepharose-bound pyridoxal 5'-phosphate. *Eur. J. Biochem.*, **51**(1), 155-164.

Furuichi, E. und P. Koehl (1998). Influence of protein structure databases on the predictive power of statistical pair potentials. *Proteins*, **31**(2), 139-149.

Galzitskaya, O. V., A. V. Skoogarev, D. N. Ivankov und A. V. Finkelstein (2000). Folding nuclei in 3D protein structures. *Pac. Symp. Biocomput.*, 131-142.

Garza-Ramos, G., D. A. Fernandez-Velasco, L. Ramirez, L. Shoshani, A. Darszon, M. Tuena de Gomez-Puyou und A. Gomez-Puyou (1992). Enzyme activation by denaturants in organic solvent systems with a low water content. *Eur. J. Biochem.*, **205**(2), 509-517.

Gerk, L. P., O. Leven und B. Muller-Hill (2000). Strengthening the dimerisation interface of Lac repressor increases its thermostability by 40 deg. C. *J. Mol. Biol.*, **299**(3), 805-812.

Gerstein, M. und M. Levitt (1998). Simulating water and the molecules of life. *Sci. Am.*, **279**(5), 100-105.

Gilis, D. und M. Rooman (1996). Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials. *J. Mol. Biol.*, **257**(5), 1112-1126.

Gilis, D. und M. Rooman (1997). Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.*, **272**(2), 276-290.

Giver, L., A. Gershenson, P. O. Freskgard und F. H. Arnold (1998). Directed evolution of a thermostable esterase. *Proc. Natl. Acad. Sci. U S A*, **95**(22), 12809-12813.

Gohlke, H., M. Hendlich und G. Klebe (2000). Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, **295**(2), 337-356.

Gromiha, M. M., J. An, H. Kono, M. Oobatake, H. Uedaira, P. Prabakaran und A. Sarai (2000). ProTherm, version 2.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **28**(1), 283-285.

Gromiha, M. M., M. Oobatake, H. Kono, H. Uedaira und A. Sarai (1999). Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. *Protein Eng.*, **12**(7), 549-555.

Grzybowski, B. A., A. V. Ishchenko, R. S. DeWitte, G. M. Whitesides und E. I. Shakhnovich (2000). Development of a Knowledge-Based Potential for Crystals of Small Organic Molecules: Calculation of Energy Surfaces for C=0...H-N Hydrogen Bonds. *J. Phys. Chem. B*, **104**, 7293-7298.

Guerois, R., J. E. Nielsen und L. Serrano (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**(2), 369-387.

Gursky, O. und D. Atkinson (1996). Thermal unfolding of human high-density apolipoprotein A-1: implications for a lipid-free molten globular state. *Proc. Natl. Acad. Sci. U S A*, **93**(7), 2991-2995.

Hill, T. L. (1986). An Introduction to Statistical Thermodynamics. New York, Dover Publications, Inc.

Hodgkin, E. E. und W. G. Richards (1987). Molecular Similarity Based on Electrostatic Potentials an Electric Field. *Int. J. Quant. Chem: Quant. Biol. Symp.*, (14), 105-110.

Huang, E. S., S. Subbiah und M. Levitt (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.*, **252**(5), 709-720.

Hunter, C. A., J. Singh und J. M. Thornton (1991). Pi-pi interactions: the geometry and energetics of phenylalanine-phenylalanine interactions in proteins. *J. Mol. Biol.*, **218**(4), 837-846.

Imanaka, T., M. Shibazaki und M. Takagi (1986). A new way of enhancing the thermostability of proteases. *Nature*, **324**(6098), 695-697.

Itzhaki, L. S., D. E. Otzen und A. R. Fersht (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.*, **254**(2), 260-288.

Jaenicke, R. und R. Seckler (1997). Protein misassembly in vitro. *Adv. Protein Chem.*, **50**, 1-59.

Jiang, Y. und H. Dalton (1994). Chemical modification of the hydroxylase of soluble methane monooxygenase gives one form of the protein with significantly increased thermostability and another that functions well in organic solvents. *Biochim. Biophys. Acta*, **1201**(1), 76-84.

Jones, D. T. und J. M. Thornton (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.*, **6**(2), 210-216.

Kabsch, W. und C. Sander (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**(12), 2577-2637.

Kang, Y. K., G. Nemethy und H. A. Scheraga (1987). Free Energies of Hydration of Solute Molecules. 1. Improvement of the Hydration Shelöl Model by Exact Computations of Overlapping Volumes. *J. Phys. Chem*, **91**, 4105-4109.

Kannan, N. und S. Vishveshwara (2000). Aromatic clusters: a determinant of thermal stability of thermophilic proteins. *Protein Eng.*, **13**(11), 753-761.

Karplus, M., M. Prevost, B. Tidor und S. Wodak (1991). Simulation analysis of the stability mutants R96H of bacteriophage T4 lysozyme and I96A of barnase. *Ciba Found Symp*, **161**, 63-74.

Karplus, M. und D. L. Weaver (1994). Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci.*, **3**(4), 650-668.

Kauzmann, W. (1993). Reminiscences from a life in protein physical chemistry. *Protein Sci.*, **2**(4), 671-691.

Kellis, J. T., Jr., K. Nyberg, D. Sali und A. R. Fersht (1988). Contribution of hydrophobic interactions to protein stability. *Nature*, **333**(6175), 784-786.

Kim, P. S. und R. L. Baldwin (1990). Intermediates in the folding reactions of small proteins. *Annu Rev. Biochem.*, **59**, 631-660.

Klimov, D. K. und D. Thirumalai (2001). Multiple protein folding nuclei and the transition state ensemble in two-state proteins. *Proteins*, **43**(4), 465-475.

Kocher, J. P., M. J. Rooman und S. J. Wodak (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.*, **235**(5), 1598-1613.

Kollman, P. A., I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case und T. E. Cheatham, 3rd (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, **33**(12), 889-897.

Koppensteiner, W. A. und M. J. Sippl (1998). Knowledge-based potentials--back to the roots. *Biochemistry (Mosc)*, **63**(3), 247-252.

Kumar, S., C. J. Tsai und R. Nussinov (2000). Factors enhancing protein thermostability. *Protein Eng.*, **13**(3), 179-191.

Kumar, S., H. J. Wolfson und R. Nussinov (2001). Protein flexibility and electrostatic interactions. *IBM J. Res. & Dev.*, **45**(3/4), 499-512.

Lazaridis, T. und M. Karplus (1999). Effective energy function for proteins in solution. *Proteins*, **35**(2), 133-152.

Lazaridis, T. und M. Karplus (2000). Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.*, **10**(2), 139-145.

Leach, A. R. (2001). Molecular Modelling - Principles and Applications, Prentice Hall.

Lee, B. und F. M. Richards (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**(3), 379-400.

Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.*, **236**(3), 918-939.

Lehmann, M. und M. Wyss (2001). Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Curr. Opin. Biotechnol.*, **12**(4), 371-375.

Leopold, P. E., M. Montal und J. N. Onuchic (1992). Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. U S A*, **89**(18), 8721-8725.

Leven, O. (1999). Vorhersage der Thermostabilität von Proteinen. Dissertation. Universität zu Köln.

Li, B., D. O. Alonso und V. Daggett (2002). Stabilization of Globular Proteins via Introduction of Temperature-Activated Elastin-Based Switches. *Structure (Camb)*, **10**(7), 989-998.

Lins, L. und R. Brasseur (1995). The hydrophobic effect in protein folding. *Faseb J.*, **9**(7), 535-540.

Lüthy, R., J. U. Bowie und D. Eisenberg (1992). Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83-85.

Marshall, S. A., C. S. Morgan und S. L. Mayo (2002). Electrostatics significantly affect the stability of designed homeodomain variants. *J. Mol. Biol.*, **316**(1), 189-199.

Matsumura, M., W. J. Becktel und B. W. Matthews (1988). Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile 3. *Nature*, **334**(6181), 406-410.

Matsumura, M., G. Signor und B. W. Matthews (1989). Substantial increase of protein stability by multiple disulphide bonds. *Nature*, **342**(6247), 291-293.

Matthews, B. W. (1993). Structural and genetic analysis of protein stability. *Annu. Rev. Biochem.*, **62**, 139-160.

Matthews, B. W., H. Nicholson und W. J. Becktel (1987). Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. U S A*, **84**(19), 6663-6667.

Meeker, A. K., B. Garcia-Moreno und D. Shortle (1996). Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. *Biochemistry*, **35**(20), 6443-6449.

Melo, F. und E. Feytmans (1997). Novel knowledge-based mean force potential at atomic level. *J. Mol. Biol.*, **267**(1), 207-222.

Minotani, N., T. Sekiguchi, J. G. Bautista und Y. Nosoh (1979). Basis of thermostability in pig heart lactate dehydrogenase treated with O-methylisourea. *Biochim. Biophys. Acta*, **581**(2), 334-341.

Moult, J. (1997). Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.*, **7**(2), 194-199.

Munch, O. und D. Tritsch (1990). Irreversible thermoinactivation of glucoamylase from Aspergillus niger and thermostabilization by chemical modification of carboxyl groups. *Biochim Biophys. Acta*, **1041**(2), 111-116.

Nakamura, H. (1996). Roles of electrostatic interaction in proteins. *Q. Rev. Biophys.*, **29**(1), 1-90.

Niefind, K. und D. Schomburg (1991). Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles. *J. Mol. Biol.*, **219**(3), 481-497.

Nölting, B. und K. Andert (2000). Mechanism of Protein Folding. *Proteins*, **41**, 288-298.

Ota, M., S. Kanaya und K. Nishikawa (1995). Desk-top analysis of the structural stability of various point mutations introduced into ribonuclease H. *J. Mol. Biol.*, **248**(4), 733-738.

Petukhov, M., D. Cregut, C. M. Soares und L. Serrano (1999). Local water bridges and protein conformational stability. *Protein Sci*, **8**(10), 1982-1989.

Prevost, M., S. J. Wodak, B. Tidor und M. Karplus (1991). Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96-Ala mutation in barnase. *Proc. Natl. Acad. Sci. U S A*, **88**(23), 10880-10884.

Rade, L. und B. Westergren (1997). Springers Mathematische Formeln, Springer.

Ratnaparkhi, G. S. und R. Varadarajan (2000). Thermodynamic and structural studies of cavity formation in proteins suggest that loss of packing interactions rather than the hydrophobic effect dominates the observed energetics. *Biochemistry*, **39**(40), 12365-12374.

Rees, D. C. und A. D. Robertson (2001). Some thermodynamic implications for the thermostability of proteins. *Protein Sci.*, **10**(6), 1187-1194.

Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.*, **82**(1), 1-14.

Richardson, J. S., D. C. Richardson, N. B. Tweedy, K. M. Gernert, T. P. Quinn, M. H. Hecht, B. W. Erickson, Y. Yan, R. D. McClain, M. E. Donlan und et al. (1992). Looking at proteins: representations, folding, packing, and design. Biophysical Society National Lecture, 1992. *Biophys. J.*, **63**(5), 1185-1209.

Robertson, A. D. und K. P. Murphy (1997). Protein Structure and the Energetics of Protein Stability. *Chem. Rev.*, **97**, 1251-1267.

Rossmann, M. G. und E. Arnold, Eds. (2001). International Tables for Crystallography. Volume F: Crystallography of biological macromolecules, Kluwer Academic Publishers.

Sachs, L. (1997). Angewandte Statistik, Springer Verlag.

Saleemuddin, M. (1999). Bioaffinity based immobilization of enzymes. *Adv Biochem Eng Biotechnol*, **64**, 203-226.

Sandberg, W. S. und T. C. Terwilliger (1989). Influence of interior packing and hydrophobicity on the stability of a protein. *Science*, **245**(4913), 54-57.

Schneckner, S. (1998). Positionsgenaues Alignment von Proteinsequenzen. Dissertation. Universität zu Köln.

Serrano, L., A. G. Day und A. R. Fersht (1993). Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J. Mol. Biol.*, **233**(2), 305-312.

Serrano, L. und A. R. Fersht (1989). Capping and alpha-helix stability. *Nature*, **342**(6247), 296-299.

Sharp, K. A. und S. W. Englander (1994). How much is a stabilizing bond worth? *Trends Biochem. Sci.*, **19**(12), 526-529.

Shoichet, B. K., W. A. Baase, R. Kuroki und B. W. Matthews (1995). A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U S A*, **92**(2), 452-456.

Shortle, D. und J. Sondek (1995). The emerging role of insertions and deletions in protein engineering. *Curr. Opin. Biotechnol.*, **6**(4), 387-393.

Shortle, D., W. E. Stites und A. K. Meeker (1990). Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry*, **29**(35), 8033-8041.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**(4), 859-883.

Sippl, M. J. (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided Mol. Des.*, **7**(4), 473-501.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, **5**(2), 229-235.

Sippl, M. J. (1999). Who solved the protein folding problem? *Structure Fold. Des.*, **7**(4), R81-83.

Sippl, M. J., M. Ortner, M. Jaritz, P. Lackner und H. Flockner (1996). Helmholtz free energies of atom pair interactions in proteins. *Fold Des.*, **1**(4), 289-298.

Sondek, J. und D. Shortle (1992). Structural and energetic differences between insertions and substitutions in staphylococcal nuclease. *Proteins*, **13**(2), 132-140.

Song, J. K. und J. S. Rhee (2000). Simultaneous enhancement of thermostability and catalytic activity of phospholipase A(1) by evolutionary molecular engineering. *Appl. Environ. Microbiol.*, **66**(3), 890-894.

Sowdhamini, R., N. Srinivasan, B. Shoichet, D. V. Santi, C. Ramakrishnan und P. Balaram (1989). Stereochemical modeling of disulfide bridges. Criteria for introduction into proteins by site-directed mutagenesis. *Protein Eng.*, **3**(2), 95-103.

Szilagyi, A. und P. Zavodszky (2000). Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure Fold. Des.*, **8**(5), 493-504.

Tanaka, S. und H. A. Scheraga (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, **9**(6), 945-950.

Thomas, P. D. und K. A. Dill (1996). Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.*, **257**(2), 457-469.

Topham, C. M., N. Srinivasan und T. L. Blundell (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.*, **10**(1), 7-21.

Toraya, T., K. Oashi und S. Fukui (1975). Immobilized diol dehydrase and its use in studies of cobalamin binding and subunit interaction. *Biochemistry*, **14**(19), 4255-4260.

Tsai, J., R. Taylor, C. Chothia und M. Gerstein (1999). The packing density in proteins: standard radii and volumes. *J. Mol. Biol.*, **290**(1), 253-266.

Tuena de Gomez-Puyou, M. und A. Gomez-Puyou (1998). Enzymes in low water systems. *Crit. Rev. Biochem. Mol. Biol.*, **33**(1), 53-89.

Udgaonkar, J. B. und R. L. Baldwin (1988). NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A. *Nature*, **335**(6192), 694-699.

Wang, Y., L. Lal, S. Li, Y. Han und Y. Tang (1996). Position-dependent protein mutant profile based on mean force field calculation. *Protein Eng.*, **9**(6), 479-484.

Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U S A*, **70**(3), 697-701.

White, F. L. und K. W. Olsen (1987). Effects of crosslinking on the thermal stability of hemoglobin. I. The use of bis(3,5-dibromosalicyl) fumarate. *Arch. Biochem. Biophys.*, **258**(1), 51-57.

Xiao, L. und B. Honig (1999). Electrostatic contributions to the stability of hyperthermophilic proteins. *J. Mol. Biol.*, **289**(5), 1435-1444.

Xu, J., W. A. Baase, E. Baldwin und B. W. Matthews (1998). The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Sci.*, **7**(1), 158-177.

Zhang, J. und C. R. Matthews (1998). Ligand binding is the principal determinant of stability for the p21(H)-ras protein. *Biochemistry*, **37**(42), 14881-14890.

Zhang, X. J., W. A. Baase, B. K. Shoichet, K. P. Wilson und B. W. Matthews (1995). Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive. *Protein Eng.*, **8**(10), 1017-1022.

Kurzzusammenfassung

In der vorliegenden Arbeit wurde eine wissensbasierte Bewertungsfunktion für die Vorhersage der Thermostabilität von Proteinen an einem Entwicklungsdatensatz, mit 646 Stabilitätsdaten aus elf Proteinen, entwickelt und abschließend an 918 Stabilitätsdaten aus 27 Proteinen getestet. Die zu entwickelnde Bewertungsfunktion soll die Umgebung einer Aminosäure quantifizieren und die Wirkung eines Aminosäureaustausches bewerten. Dazu wurden zwei wissensbasierte Potentialfunktionen, ein richtungs- und abstandsabhängiges Aminosäure-Atom-Potential sowie ein Torsionswinkelpotential, verwendet und miteinander kombiniert.
Diese Methode stellt einen Ansatz zur automatischen Erstellung eines Mutationsprofils dar und erzielt hinsichtlich ihres Einsatzes als möglicher Vorfilter oder als alleiniges Kriterium für das *Protein Engineering* gute Ergebnisse.

Abstract

A knowledge-based discrimination function was developed and evaluated for the prediction of protein thermostability. This function should be able to quantifiy the effect of every possible mutation on the stability of a protein and consists of two parts, the knowledge-based aminoacid-atom potential and the knowledge-based torsion angle energy function. The pairwise aminoacid-atom energy function is based on a direction- and distance-dependant atomic description of the aminoacid environment. A training set of eleven proteins with 646 single mutations was used to find and optimize the best set of parameters. The resulting potential function was then tested using a blind test mutant database consisting of 27 proteins with 918 single mutations.
The present potential function creates possible candidates for point mutations and act as a fast reliable guide in *protein engineering* and enzyme optimisation.

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. D. Schomburg betreut worden.

Christian Hoppe

Teile dieser Arbeit wurden bereits veröffentlicht:

Hoppe C., Schomburg D, Prediction of Protein Thermostability, Poster präsentiert bei der *ECCB* (European Conference of Computational Bioinformatics) 2002, 6-9. Oktober 2002.

Lebenslauf

7. Mai 1974	geboren in Hannover Staatsangehörigkeit deutsch
1980-1984	Grundschule Köln-Lövenich
1984-1993	Georg Büchner Gymnasium, Weiden
Juni 1993	Abitur
1993	Beginn des Chemiestudiums an der Universität zu Köln
1996	Vordiplomprüfung im Fachbereich Chemie
1998	Diplomprüfung im Fachbereich Chemie
September 1999	Abgabe der Diplomarbeit bei Herrn Prof. Dr. U. Fuhr, Institut für Pharmakologie, Universität zu Köln „Bedeutung humaner fremdstoffmetabolisierender Enzyme beim Metabolismus des Zytostatikums Procarbazin“
10/1999 – 2/2000	Forschungsaufenthalt bei Herrn Prof. Dr. Leumann im Institut für Biochemie und organischer Chemie an der Universität Bern
seit Mai 2000	Doktorarbeit bei Herrn Prof. Dr. D. Schomburg, Institut für Biochemie, Universität zu Köln „Entwicklung einer richtungs – und abstandsabhängigen wissensbasierten Bewertungsfunktion für die Vorhersage der Thermostabilität von Proteinen“