Tetrapeptidbasiertes Proteindesign - Ein Lösungsansatz für das inverse Proteinfaltungsproblem

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt der
Naturwissenschaftlichen Fakultät I – Biowissenschaften
der Martin-Luther-Universität Halle-Wittenberg

von
Roman Dallüge
geboren am 30.11.1972 in Eisenach/Thüringen

Gutachter:

Prof. Dr. R. Rudolph
Prof. Dr. F.X. Schmid
Prof. Dr. R. Sterner

Verteidigung der Arbeit am:

14. Februar 2008

urn:nbn:de:gbv:3-000013646
[http://nbn-resolving.de/urn/resolver.pl?urn=nbn%3Ade%3Agbv%3A3-000013646]

Es gibt nichts praktischeres als eine gute Theorie.

Albert Einstein

Inhaltsverzeichnis

Abkürzungen

A	optische Absorption
APS	Ammoniumperoxodisulfat
BLAST	*Basic local alignment search tool*
BLOSUM	*Block substitution matrix*
bp	Basenpaare
CASP	*Critical assessment of protein structure prediction*
CD	Circulardichroismus
COSY	*Correlation Spectroscopy*
DEET	*dead end elimination theorem*
deg	Winkelgrad
DNA	Desoxyribonukleinsäure
E. coli	*Escherichia coli*
EDTA	N,N,N',N'-Ethylendiamintetraacetat
GdmCl	Guanidiniumchlorid
h	Stunde(n)
His-*tag*	Histidin-*tag*
HP-Motiv	Hydrophob-Polares Motiv
IMAC	Immobilisierte Metallchelataffinitätschromatographie
IPTG	β-D-Isopropyl-thio-galactopyranosid
Kan	Kanamycin
LB-Medium	Luria-Bertani-Medium
LMW	*Low molecular weight*
MALDI-TOF	*Matrix-assisted laser desorption ionization time-of-flight*
MWCO	*Molecular weight cut-off*
M_G	Molekulargewicht
Ni-NTA	Nickelnitrilotriacetat
NOESY	*Nuclear-Overhauser-Effect-Spectroscopy*
NMR	*Nuclear Magnetic Resonance*
OD_{600}	Optische Dichte bei 600 nm
p.a.	*pro analysi*
PAGE	Polyacrylamidgelelektrophorese
PDB	*Brookhaven Protein Database*
pI	isoelektrischer Punkt
rpm	*revolutions per minute* (Umdrehungen pro Minute)
rmsd	*root mean square deviation*
S.	Seite
SDS	Natriumdodecylsulfat
TEMED	N,N,N',N'-Tetramethylethylendiamin; 1,2-Bis(dimethylamino)ethan
TOCSY	*Total-Correlation-Spectroscopy*
Tris	Tris-(hydroxymethyl)-aminomethan
U	*unit* (Einheit der Enzymaktivität)
UV	ultraviolett
VIS	*visible*
v/v	Volumen pro Volumen
w/v	Gewicht pro Volumen

Zusammenfassung

Die vorliegende Arbeit beschreibt ein wissensbasiertes System zur Berechnung von alternativen Aminosäuresequenzen für experimentell bestimmte Proteinstrukturen. Die grundlegende Funktionsweise des vorgestellten Algorithmus beruht auf der Assemblierung von Tetrapeptidfragmenten in ihrer bevorzugten Konformation, die im Rahmen einer statistischen Strukturanalyse dieser Fragmente aus experimentell gelösten Proteinstrukturen ermittelt werden konnte. Hierfür wurde ein verbessertes Verfahren zur Bereitstellung nichtredundanter Informationen entwickelt, das in der Lage war, genügend Strukturinformationen bereitzustellen. Die experimentelle Evaluierung des Systems erfolgte am Beispiel der Struktur des Proteins Top7 [Kuhlman *et al.*, 2003].

Die geringe Datenvielfalt innerhalb nichtredundanter Sequenzdatenbanken von Proteinstrukturen und nichtredundanten Strukturdatenbanken ist das fundamentale Hindernis bei der Aufdeckung von Sequenz-Struktur-Korrelationen. Aus diesem Grund konnten bisher nur die $20^3 = 8\,000$ Tripeptide strukturell umfassend charakterisiert werden. Infolge der exponentiellen Zunahme der Anzahl möglicher Sequenzen mit der Länge eines Fragmentes war für die $20^4 = 160\,000$ verschiedenen Tetrapeptide eine statistische Beschreibung ihrer Konformationseigenschaften nicht oder nur sehr eingeschränkt möglich. Die Verwendung nichtredundanter Daten ist bei einer statistischen Analyse eine zwingende Voraussetzung, um verwertbare Ergebnisse zu erzielen. Der Schlüssel zu einem tieferen Verständnis der Konformationseigenschaften von Tetrapeptiden bestand daher in der Entwicklung einer Datenaufbereitung, die ohne Verletzung der Nichtredundanz-Bedingung genügend Strukturinformationen zur Durchführung einer Konformationsbetrachtung zur Verfügung stellen kann.

Die Lösung zu dieser Problemstellung bestand in der Überlegung, dass eine Menge von Objekten nur dann verglichen werden sollte, wenn sie ein bestimmtes Attribut, nämlich das zu untersuchende, gemeinsam haben. Die Beseitigung redundanter Informationen jeweils innerhalb dieser Gruppen führt, im Vergleich zu der Verwendung einer nichtredundanten Datenbasis als Ausgangsdatenbasis, zu einer deutlichen Vergrößerung der Informationsmenge bezüglich der untersuchten Eigenschaft. Dies bedeutet, dass im Rahmen einer statistischen Analyse von Proteineigenschaften (Attribute) nicht mehr nichtredundante Sequenzdatenbanken von Proteinstrukturen (Objekte) als Ausgangsdatensätze verwendet werden, sondern alle verfügbaren Strukturen, die bestimmten Qualitätskriterien genügen. Im Vergleich zu der Verwendung einer nichtredundanten Sequenzdatenbank von Proteinstrukturen als Ausgangsdatensatz, konnte mit Anwendung dieses Verfahrens auf die *Brookhaven Protein Database* (*PDB*) die Anzahl berücksichtigter Strukturen verzehnfacht werden.

Die Konformationsanalyse der Tetrapeptide umfasste die sequenzabhängige Betrachtung des ψ-Winkels der zweiten Aminosäure (ψ_2) und des φ-Winkels der dritten Aminosäure (φ_3). Es wurde eine sehr ausgeprägte strukturelle Präferenz dieser Winkel in einer großen Anzahl an Tetrapeptiden gefunden. Diese bevorzugte Strukturbildung ermöglichte die Entwicklung eines Algorithmus zur Berechnung von alternativen Aminosäuresequenzen zu gegebenen Proteinstrukturen. Das Grundprinzip dieses Verfahrens besteht in der Überlappung von Tetrapeptidfragmenten über drei Aminosäuren hinweg und einer Verlängerung dieser Sequenz bis zur vollständigen Beschreibung der jeweiligen Zielstruktur. Desweiteren konnte gezeigt werden, dass Aminosäuresequenzen, bei denen die einzelnen Tetrapeptide jeweils eine sehr hohe Wahrscheinlichkeit für die wahrscheinlichste Konformation besitzen, die daraus ableitbare wahrscheinlichste

Struktur mit einer größeren Häufigkeit ausbilden, als bei einer unabhängigen Strukturbildung der Einzelfragmente zu erwarten wäre.

Für einen ersten Test des beschriebenen fragmentbasierten Modellierungsverfahrens wurde die Struktur des Proteins Top7 gewählt, das aus zwei βαβ-Motiven mit insgesamt fünf β-Faltblättern und zwei α-Helices besteht. Es wurden acht Aminosäuresequenzen berechnet, welche die Struktur von Top7 kodieren sollten und gleichzeitig Sequenzidentitäten von kleiner als 30 % zu der Originalsequenz von Top7 aufwiesen. Alle acht Proteine konnten rekombinant in *Escherichia coli* exprimiert werden. Bisher wurde die Variante M7 am umfassendsten charakterisiert. Sie wurde als lösliches Konstrukt hergestellt und zeigte eine kooperative Faltung. Die auffälligste Eigenschaft dieses Proteins ist seine äußerst hohe thermodynamische Stabilität, die sich mit einer Freien Entfaltungsenthalpie beim Übergang vom nativen Protein (*n*) zur denaturierten Spezies (*d*) von $\Delta G_{n\to d}^{H_2O} = +69.3$ kJ mit einem Übergangsmittelpunkt von $D_{1/2} = 6.6$ M Guanidiniumchlorid (GdmCl) manifestiert. In gleicher Weise verhält sich dieses Protein bei thermischer Behandlung nahezu indifferent. Eine Entfaltung konnte in 0 M GdmCl bis zu einer Temperatur von T = 383 K (110 °C) nicht beobachtet werden, sondern erst ab einer Konzentration von 5.5 M GdmCl und einer Temperatur von T ≈ 363 K (90 °C). Inzwischen wurde die Struktur von M7 mit Hilfe der NMR-Spektroskopie aufgeklärt. Sie zeigt eine sehr gute Übereinstimmung mit dem Modell.

Die vorliegende Arbeit zeigte experimentell am Beispiel der Struktur des Proteins Top7, dass Aminosäuresequenzen, deren Konformation mit der niedrigsten Freien Energie die Zielstruktur beschreiben soll, sich in sehr einfacher Weise durch ein fragmentbasiertes Design auf Tetrapeptidbasis berechnen lassen. Das Verständnis des Proteinfaltungscodes auf Tetrapeptidebene konnte damit erweitert werden.

I. Einleitung

Computergestützte Methoden finden seit einigen Jahren immer häufiger ihre Anwendung bei der Lösung biologischer Fragestellungen. Ein sehr prominentes Beispiel sind die Erfolge des Humangenomprojektes [Weis, 1990; Deloukas *et al.*, 1998; Venter *et al.*, 2001], die ohne die Entwicklung von geeigneten Algorithmen zur schnellen Assemblierung von DNA-Fragmenten [Ewing *et al.*, 1998; Ewing & Green, 1998] nicht möglich gewesen wären. Neben der Sequenzierung ganzer Genome hat sich die strukturelle Genomik die Strukturaufklärung von Proteinen zum Ziel gesetzt. Bis heute wurden die Strukturen von 36 104 Proteinen bzw. Proteinkomplexen experimentell bestimmt und deren Atomkoordinaten in der *Brookhaven Protein Database* veröffentlicht (*PDB*, Stand 18.07.2006, ftp.rcsb.org) [Berman *et al.*, 2000a]. Trotz der großen Fortschritte in der instrumentellen Strukturbestimmung von Proteinen wächst die Anzahl an bekannten Proteinsequenzen deutlich schneller, als die der gelösten Proteinstrukturen, wie ein Vergleich mit den 3 586 193 Sequenzen der Swissprot-Datenbank zeigt (Version 8.3, 11.07.2006) [Boeckmann *et al.*, 2003]. Neben der experimentellen Strukturaufklärung werden deshalb theoretische Methoden zur Strukturbestimmung von Proteinen erarbeitet. Deren Grundlage wurde bereits 1973 von Anfinsen gelegt, der erkannte, dass in den meisten Fällen ein Protein im nativ gefalteten Zustand eine einzigartige Struktur besitzt, die in seiner Aminosäuresequenz verschlüsselt ist [Anfinsen, 1973; Jaenicke, 1987]. Sternberg und Thornton schlussfolgerten daraus, dass es prinzipiell möglich ist, die dreidimensionale Struktur eines Proteins anhand seiner Aminosäuresequenz vorherzusagen [Sternberg & Thornton, 1978]. Dieser als Proteinfaltungsproblem oder Proteinfaltungscode beschriebene Zusammenhang war der Ausgangspunkt der Entwicklung von Methoden zur rechengestützten Strukturvorhersage (*protein structure prediction*) von Proteinen, basierend auf deren Aminosäuresequenz [Sali *et al.*, 1995; Krieger *et al.*, 2003]. Eine näherungsweise Lösung des Proteinfaltungsproblems wurde jedoch erst durch Chothia und Lesk ermöglicht, die zeigten, dass Proteinstrukturen im Verlauf der Evolution einer wesentlich langsameren Divergenz unterliegen, als die ihr zugrunde liegenden Aminosäuresequenzen [Chothia & Lesk, 1986]. Unterschiedliche Sequenzen können deshalb in die gleiche Struktur (Faltungstopologie) falten. Es wird erwartet, dass zwischen 1000 [Chothia, 1992; Wang, 1998; Leonov *et al.*, 2003] und 2000 [Govindarajan *et al.*, 1999] verschiedene Faltungstopologien während der Evolution entstanden sind. Derzeit sind 1 110 Topologien in der *CATH*-Datenbank (*classification by class, architecture, topology and homology*) gelistet [Pearl *et al.*, 2003] (Version 3.0.0, 04.05.2006). Aufgrund der exponentiellen Zunahme an Proteinstrukturen in der *PDB* konnten Rost und Mitarbeiter die Grenzen der Regel von Chothia und Lesk bestimmen [Rost, 1999]. Die Abbildung I-1 auf Seite 2 zeigt dazu, dass solange der prozentuale Anteil an identischen Aminosäuren in einem paarweisen Alignment von zwei Aminosäuresequenzen in die *safe homology modeling zone* fällt, davon auszugehen ist, dass diese beiden Sequenzen ähnliche Strukturen annehmen werden. Der Prozess der Strukturvorhersage einer Aminosäuresequenz (*query*-Sequenz) mit Hilfe einer Sequenz, deren Struktur bekannt ist, (*template*-Sequenz, *template*-Struktur) wird als Homologiemodellierung (*comparative modeling*) bezeichnet. Die Qualität eines Sequenzalignments ist somit entscheidend für die Güte des Proteinmodelles. Besonders bei Sequenzidentitäten zwischen der *query*-Sequenz und der *template*-Sequenz, die in der *twilight zone* liegen, lässt sich die richtige *template*-Struktur nur schwierig bestimmen und die errechneten Proteinmodelle können in der Folge fehlerbehaftet sein. Eine Zusammenfassung verfügbarer Programme zur Homologiemodellierung von Proteinen findet man z. B. bei Eswar *et al.* [Eswar *et al.*, 2003].

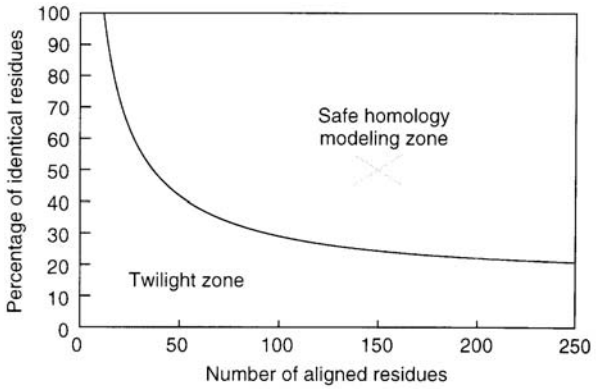


Abbildung I-1 Die beiden Zonen eines Sequenzalignments. Zwei Sequenzen falten mit hoher Wahrscheinlichkeit in die gleiche Struktur, wenn ihre Länge und die paarweise Sequenzidentität in die *safe homology modeling zone* fallen. (Aus *Structural Bioinformatics* [Bourne & Weissig, 2003])

Bowie und Mitarbeiter entwickelten einen alternativen Lösungsansatz zur Proteinstrukturvorhersage, der als das *Inverse Proteinfaltungsproblem* oder als *threading (fold recognition)* bezeichnet wird [Bowie *et al.*, 1991]. Dabei wird versucht die Frage zu beantworten, ob in einer Datenbank Aminosäuresequenzen vorhanden sind, die in eine bekannte Struktur falten können. Der allgemeine Vorgang beim *threading* besteht im „Auffädeln" dieser Aminosäuresequenzen auf eine Struktur und der anschließenden Bewertung dieser Sequenzen mit wissensbasierten Energiefunktionen, den so genannten *database-derived-potentials*. Die Sequenz mit der niedrigsten Energie ist im Ergebnis diejenige, welche mit höchster Wahrscheinlichkeit in die Zielstruktur faltet. Besonders bei Aminosäuresequenzen, die in der *twilight zone* von Abbildung I-1 liegen, erzielt das *threading* sehr gute Erfolge bei der Auswahl geeigneter *template*-Strukturen [Blake & Cohen, 2001, Yona & Levitt, 2002, Pirun *et al.*, 2005]. Sowohl die klassische Homologiemodellierung als auch das *threading* können jedoch nur erfolgreich sein, wenn eine Struktur existiert, die eine Kompatibilität mit der Zielsequenz aufweist, da beide Verfahren nicht in der Lage sind, neue Faltungstopologien vorherzusagen. Die Kenntnis des vollständigen Satzes der natürlichen Faltungstopologien würde die Möglichkeit eröffnen, eine sehr schnelle Strukturabschätzung von natürlichen Proteinsequenzen durchzuführen, wie sie mit experimentellen Methoden nicht möglich ist. Die Vorhersage unbekannter Faltungstopologien (*ab initio folding* oder *de novo folding*) wird derzeit mit sehr gutem Erfolg durch das Programm *Rosetta* ermöglicht [Rohl *et al.*, 2004], wie die Ergebnisse der letzten *CASP*-Wettbewerbe (*critical assessment of protein structure prediction*) gezeigt haben [Bonneau *et al.*, 2001; Chivian *et al.*, 2003; Chivian *et al.*, 2005]. Der Kernalgorithmus von *Rosetta* besteht in der Assemblierung von Tri- und Nonapeptidfragmenten. Eine aus dem Inversen Proteinfaltungsproblem abgeleitete Fragestellung beschreibt das Problem des Designs bzw. der Modellierung von Aminosäuresequenzen zu gegebenen Strukturen [Shakhnovich *et al.*, 1991; Yue & Dill, 1992; Godzik, 1995]. Die Modellierbarkeit (*designability*) einer Struktur wird danach über die Anzahl an Sequenzen definiert, deren Konformation mit der niedrigsten Freien Energie die Zielstruktur ist. Die Plausibilität dieser Definition wird dabei durch die hohe Anzahl an bekannten Aminosäuresequenzen im Vergleich zu den *in natura* gefundenen Faltungstopologien gestützt. Li und Mitarbeiter vermuten die Existenz eines evolutionären *designability principle*, dem eine entscheidende Rolle in der natürlichen Selektion von Proteinsequenzen und Strukturen zugeschrieben wird [Li *et al.*, 1996; Li *et al.*, 1998]. Dieses Prinzip favorisiert Strukturen, die eine relative Stabilität gegenüber Mutationen und eine erhöhte thermodynamische Stabilität gegenüber

anderen möglichen Strukturen besitzen, sowie Sekundärstrukturelemente und Motive [Li *et al.*, 2002]. Es konnte gezeigt werden, dass weniger modellierbare Strukturen (*less designable folds*) häufiger mit krankheitsauslösenden Proteinen assoziiert sind [Wong *et al.*, 2005]. Die Existenz eines *designability principles* würde folglich zumindest die Berechnung von alternativen Proteinsequenzen zu einer im Verlauf der Evolution entstandenen Faltungstopologie erlauben. Da die Funktion eines Proteins über dessen Struktur bestimmt wird, kommt der Entwicklung von Algorithmen zur Berechnung von Aminosäuresequenzen zu vordefinierten Strukturen eine große Bedeutung zu.

I.1. Die Ableitung von wissensbasierten Energiefunktionen aus Datenbanken

Wissensbasierte Energiefunktionen von Proteineigenschaften können zur Bewertung von 3D-Strukturmodellen von Proteinen, beim Prozess des *threadings* oder bei dem Design von Aminosäuresequenzen zu einer gegebenen Struktur verwendet werden. Bisher wurde eine Vielzahl von Proteineigenschaften auf ihre Anwendbarkeit zur Bewertung von nativen Proteinstrukturen hin untersucht. Eine Übersicht dazu findet man u.a. bei [Melo *et al.*, 2002]. Diese Energiefunktionen werden empirisch aus Häufigkeitsverteilungen von Proteineigenschaften ermittelt und können Wechselbeziehungen zwischen diesen Eigenschaften und dem daraus resultierenden Zustand des untersuchten Systems offenbaren [Sippl, 1990; Sippl, 1993; Sippl, 1995; Zhang & Skolnick, 1998; Dehouck *et al.*, 2006]. Die aus diesen Zusammenhängen ableitbaren wissensbasierten Energiefunktionen (*database-derived-potentials*) verfolgen einen deduktiven Ansatz mit der Annahme, dass die einzig zuverlässige Datenquelle bekannte, experimentell bestimmte Proteinstrukturen sind. Die Energiefunktionen werden ausschließlich durch Auswertung dieser Strukturen entwickelt. Nach Sippl ist die grundlegende Annahme hinter dem Konzept der wissensbasierten Energiefunktionen, dass der native Faltungszustand eines Proteins in Lösung im Gleichgewicht dem globalen Minimum der Freien Enthalpie entspricht [Anfinsen, 1973; Jaenicke, 1987] und das die Verteilung von Molekülen auf ihre Mikrozustände durch das Boltzmann-Prinzip beschrieben wird, welches die Energie E dieses Systems mit einer Wahrscheinlichkeitsdichtefunktion p verknüpft [Sippl, 1993]:

I-1
$$p_{ijl} = \frac{1}{Z} e^{-\frac{E_{ijl}}{kT}}$$

k: Boltzmann Konstante

T: Temperatur

Der Parameter Z ist die Zustandssumme über alle n möglichen Zustände des Systems. In den meisten Fällen wird $Z = 1$ definiert [Sippl, 1993]. Die Indizes i, j, l bezeichnen die Variablen des Systems.

I-2
$$Z = \sum_{ijl} e^{-\frac{E_{ijl}}{kT}}$$

Die aus einer Datenbank abgeleiteten relativen Häufigkeiten f_{ijl} einer Eigenschaft lassen sich gemäß der inversen Boltzmann-Verteilung in eine Energie umrechnen, die auch als *potential of mean force* bezeichnet wird [Sippl, 1993; Sippl, 1995]:

I-3 $$E_{ijl} = -kT \ln(f_{ijl}) - kT \ln Z$$

Die relative Häufigkeit f_{ijl} ist dabei in dem Sinne äquivalent zur Wahrscheinlichkeitsdichte p_{ijl}, als dass bei unendlicher Anzahl n von Messungen die relative Häufigkeit f_{ijl} in die Wahrscheinlichkeitsdichte p_{ijl} konvergiert, d.h. es gilt [Sippl, 1993; Sippl, 1995]:

I-4 $$\lim_{n\to\infty} f_{ijl} \equiv p_{ijl}$$

Weiterhin ist zu beachten (I-5), dass die relativen Häufigkeiten und die Wahrscheinlichkeitsdichten auf eins normalisiert sind [Sippl, 1993; Sippl, 1995].

I-5 $$\sum_{ijk} f_{ijk} = \sum_{ijk} p_{ijk} = 1$$

Die Nichtredundanz der verwendeten Daten ist bei einer statistischen Analyse von Proteineigenschaften eine zwingende Voraussetzung um verwertbare Ergebnisse zu erzielen. Der Begriff *Nichtredundanz* lässt sich dabei sowohl auf Sequenzebene, als auch auf Strukturebene definieren. Der Terminus *Sequenzbasierte Nichtredundanz* wird im Allgemeinen bei Gruppen von Aminosäuresequenzen verwendet, die eine maximale gegenseitige Sequenzidentität von 30 % aufweisen [Hobohm *et al.*, 1992]. Die *PDBSELECT*-Datenbank bietet Datensätze von Proteinstrukturen an, die dieser Bedingung genügen. Aktuelle Listen sind im Internet unter http://swift.cmbi.kun.nl/whatif/select/ abrufbar. Demgegenüber fassen nichtredundante Strukturdatenbanken diejenigen Strukturen zusammen, die aus Struktur-Struktur-Alignments hervorgegangen sind. Die *FSSP*-Datenbank (*families of structurally similar proteins*) stellt hierzu Listen von Proteinstrukturen zusammen, die aus dieser Datenselektion hervorgegangen sind [Holm & Sander, 1994; Holm & Sander, 1996; Holm & Sander, 1997]. Gleichzeitig weisen die Proteine in dieser Datenbank eine gegenseitige Sequenzidentität von kleiner als 25 % auf. In Abhängigkeit von der zu untersuchenden Fragestellung können somit die geeigneten Datenbanken gewählt werden.

I.2. Die Konformationseigenschaften von Oligopeptiden

Das Hauptproblem bei der statistischen Analyse der Konformationseigenschaften von Oligopeptiden besteht in der exponentiell zunehmenden Anzahl an Kombinationsmöglichkeiten der 20 proteinogenen Aminosäuren in einem Fragment mit einer bestimmten Länge und der daraus resultierenden dramatisch abnehmenden Häufigkeit, das untersuchte Fragment in einer nichtredundanten Ausgangsdatenbasis zu finden. Für ein Peptid mit einer Länge von *n* Aminosäuren ergeben sich 20^n verschiedene Oligopeptide. Folglich existieren 8 000 verschiedene Tripeptide, 160 000 unterschiedliche Tetrapeptide und 3 200 000 Pentapeptide mit unterschiedlichen Sequenzen. Die exponentielle Zunahme an Kombinationen verhindert bei den zur Zeit zur Verfügung stehenden Strukturdatenbanken von Proteinen eine vertiefende Strukturanalyse von Oligopeptiden ab einer Länge von vier Aminosäuren. Bisher konnten lediglich Tripeptide mit einer ausreichenden statistischen Signifikanz untersucht werden [Anishetty *et al.*, 2002; Betancourt & Skolnick, 2004]. Die Ergebnisse offenbarten das Vorhandensein von Tripeptiden, die strukturelle Präferenzen zeigten, wobei keine allgemeine Korrelation zwischen bevorzugten Strukturen und dem Auftreten der entsprechenden Tripeptide in einem Sekundärstrukturelement beobachtet werden konnte. Die statistische Analyse der Konformationseigenschaften höherer Oligopeptide erfolgte bisher vorwiegend mit einem reduzierten Alphabet der 20 Aminosäuren, in dem diese in Gruppen mit ähnlichen physikalischen Eigenschaften eingeteilt wurden. Mit Anwendung dieses Systems konnte gezeigt werden, dass 73 % der Tetrapeptide eine bevorzugte Konformation aufweisen [Rackovsky, 1995]. Der Autor schloß daher auf die Existenz eines lokalen inversen Proteinfaltungscodes. Weiterführende Arbeiten, die sich mit den Konformationseigenschaften von Tetrapeptiden beschäftigten, erfolgten durch Sudarsanam *et al.* [Sudarsanam & Srinivasan, 1997]. Die Konformationsbetrachtung umfasste dabei die Analyse des ψ-Winkels der zweiten Aminosäure (ψ_2) und des φ-Winkels der dritten Aminosäure (φ_3). Wie die Abbildung I-2 auf Seite 6 zeigt, liegen diese beiden Diederwinkel N- und C-terminal von der mittleren Peptidbindung. Aufgrund der geringen Ausgangsdatenbasis erfolgte eine Klassifizierung der flankierenden ersten und vierten Aminosäure in Gruppen mit ähnlichen Eigenschaften und eine gleichzeitige Erhöhung der gegenseitigen maximalen Sequenzidentität in der nichtredundanten Datenbank auf 90 %. In Einzelfällen konnten statistische Untersuchungen sequenzidentischer Tetrapeptide durchgeführt werden. Die Ergebnisse zeigten, dass eine Konkretisierung der ersten und vierten Aminosäure zu einer restriktiveren Winkelverteilung führt. Diese Fragmente wurden erfolgreich durch die Autoren zur Strukturvorhersage von Polypeptiden verwendet. Bis heute wurden Fragmente mit bis zu neun Aminosäuren Länge untersucht [Kabsch & Sander, 1984; Cohen *et al.*, 1993; Sudarsanam, 1998; Zhou *et al.*, 2000; Kuznetsov & Rackovsky, 2003]. Im Ergebnis dieser Analysen wurde gefunden, dass auch mit zunehmender Anzahl von Aminosäuren in einem Fragment eine strukturelle Ambivalenz identischer Peptide aus verschiedenen Proteinstrukturen zu beobachten war. Im Hinblick auf ein fragmentbasiertes Proteindesign wurde deshalb vorgeschlagen, hierfür höhere Oligopeptide zu verwenden [Sudarsanam, 1998]. Für sequenzidentische Sequenzen, die verschiedene Konformationen ausbilden können, wurde in der Literatur der Begriff *Chamäleonsequenz* geprägt. Die Analyse der Aminosäuren in Chamäleonsequenzen mit einer Länge von fünf bis sieben Aminosäuren offenbarte eine Prävalenz von Alanin, Leucin und Valin in diesen Sequenzen [Mezei, 1998]. Dieses Ergebnis konnte auf Tetra- und Oktapeptide erweitert werden [Zhou *et al.*, 2000]. Zusätzlich wurde dabei das allgemein häufige Auftreten von hydrophoben Aminosäuren mit aliphatischen Seitenketten in diesen Sequenzen verallgemeinert und der große Einfluss des

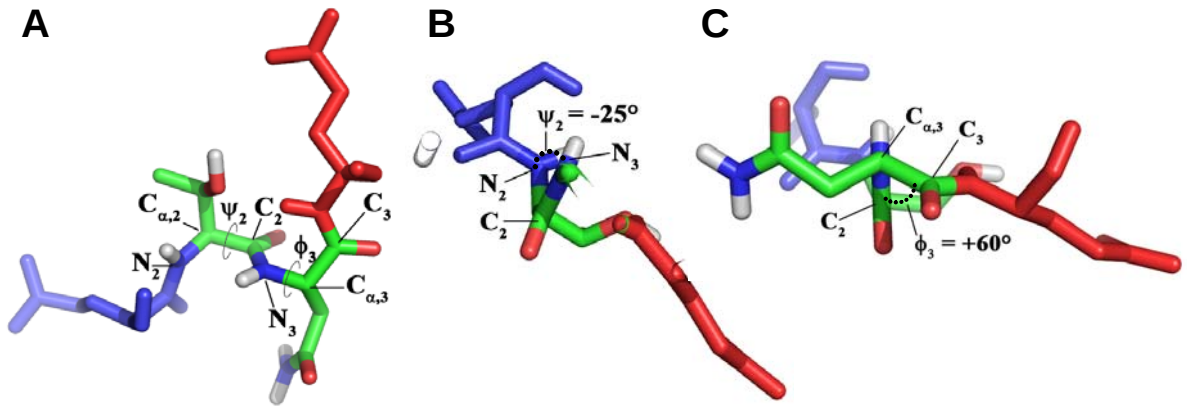


Abbildung I-2 Beschreibung der Konformation eines Tetrapeptides durch Analyse der Diederwinkel ψ_2 und φ_3. Es ist das Tetrapeptid ETNE aus der Kristallstruktur von Agglutinin dargestellt [Transue *et al.*, 1997] (PDB-Code 1JLY, Proteinkette A, Position 238 – 241). Die Wasserstoffatome wurden mit dem *Swiss-PdbViewer 3.7* (*SP5*) angefügt [Guex & Peitsch, 1997]. Die blau bzw. rot dargestellten Aminosäuren kennzeichnen die N- bzw. C-terminale Glutaminsäure. ψ_2 ist durch Rotation um die $C_{\alpha,2}$-C_2-Bindung definiert, φ_3 durch Rotation um die N_3-$C_{\alpha,3}$-Bindung. **A** Gezeigt ist die Peptidbindung zwischen Threonin und Asparagin mit den Atombezeichnungen des Proteinrückgrats. Der Index der Atome bezeichnet die Aminosäurenummer in dem Tetrapeptid. **B** Blick in Richtung der C_2-$C_{\alpha,2}$-Bindung. Der ψ_2-Winkel wird durch die beiden Atome N_3 und N_2 eingeschlossen und errechnet sich zu einem Wert von $\psi_2 = -25°$. **C** Blick in Richtung der $C_{\alpha,3}$-N_3-Bindung. Der φ_3-Winkel wird durch die beiden Atome C_3 und C_2 eingeschlossen. Für φ_3 ergibt sich ein Wert von $\varphi_3 = +60°$. Die Abbildung wurde mit Hilfe des Programms *PyMOL* erstellt [Delano, 2002].

strukturellen Kontextes auf die in einem Protein beobachtete Konformation einer strukturell ambivalenten Sequenz herausgestellt. Der globale und lokale Einfluss auf die Strukturbildung einer Chamäleonsequenz wurde ebenso von anderen Autoren betont [Cohen *et al.*, 1993; Kuznetsov & Rackovsky, 2003]. Eine Klassifizierung beobachteter Strukturen von Sequenzen mit einer Länge von vier bis sieben Aminosäuren erfolgte durch Micheletti und Mitarbeiter [Micheletti *et al.*, 2000]. Die Analyse erfolgte nach binärer Einteilung der einzelnen Aminosäuren hinsichtlich ihrer hydropathischen Eigenschaften (hydrophob/polar), wobei Prolin und Glycin jeweils eine Gruppe darstellten. Diese reduktionistische Betrachtungsweise führte zu Strukturen, so genannten *oligons*, die wiederum durch Verknüpfungen Proteinstrukturen beschreiben können. Dabei zeigte sich, dass für diesen Zweck *oligons* mit einer Länge von fünf und sechs Aminosäuren am besten geeignet sind. *Oligons* mit einer Länge von kleiner als fünf oder größer als sieben stellten sich für eine Beschreibung von Proteinstrukturen als ungeeignet heraus. Fragmentbibliotheken mit einer Länge von vier bis neun Aminosäuren wurden in gleicher Weise zur Strukturmodellierung von Proteinen verwendet [Kolodny *et al.*, 2002; Holmes & Tsai, 2004]. Diese Fragmentbibliotheken beinhalteten ausschließlich Informationen über mögliche sequenzunspezifische Konformationen von Fragmenten mit einer bestimmten Länge. Eine experimentelle Analyse der *in silico* modellierten Proteine erfolgte in keinem der genannten Fälle. Eine beginnende Sequenz-Struktur-Spezifität bei Oligopeptidfragmenten wurde ab einer Länge von 15-20 Aminosäuren beobachtet [Hu *et al.*, 1997]. Dennoch zeigten die Autoren, dass im Allgemeinen 40 % der Gesamtaminosäuresequenz bekannt sein müssen, damit die Sequenz bzw. ein Fragment seine native Struktur erkennt. Melo und Mitarbeiter konnten in Übereinstimmung mit diesen Ergebnissen zeigen, dass statistische Potentiale, die bevorzugte Konformationen von Aminosäuren beschreiben, nur in sehr ungenügende Weise für die Bewertung von 3D-Strukturmodellen von Proteinen verwendet werden können [Melo *et al.*, 2002]. Ebenso wurden die strukturellen Präferenzen bei Dipeptiden oder Tripeptiden hinsichtlich eines bestimmten Sekundärstrukturtyps als nicht signifikant genug bewertet, um sie zur Strukturvorhersage von Proteinen verwenden zu können [Vlasov *et al.*, 2005].

I.3. Das Design von Aminosäuresequenzen

In den letzten Jahren wurden große Fortschritte im Design von Proteinen erzielt. Damit wurde ein Wandel von der reinen Strukturaufklärung hin zu einer kreativen Beschreibung neuartiger Proteine und Proteinstrukturen vollzogen. Die astronomische Anzahl an Kombinationsmöglichkeiten der 20 proteinogenen Aminosäuren, die sich bereits bei einem Protein von 100 Aminosäuren Länge ergibt, schließt jedoch das einfache Ausprobieren von Aminosäuresequenzen mangels Zeit und Ressourcen aus. Einfache Abschätzungen haben gezeigt, dass selbst unter optimalen Bedingungen weniger als 10^{55} unterschiedliche Polypeptidketten auf der Erde entstanden sein können und während der Evolution auf ihre Verwendbarkeit geprüft werden konnten [Wilks *et al.*, 1992]. Im Vergleich zu den theoretisch 10^{130} möglichen Proteindomänen aus 100 Aminosäuren [Zou & Saven, 2000] ist dies nur ein verschwindend geringer Anteil. Das Konzept der inversen Proteinfaltung, also der Suche nach Aminosäuresequenzen, die in eine vorgegebene Tertiärstruktur falten, muss daher einen Lösungsansatz bieten, wie mit endlichen Rechenressourcen eine Abschätzung von möglichen Aminosäuresequenzen für eine Zielstruktur durchgeführt werden kann. Die Identifizierung von Sequenzen, die zu einer niedrigen Energie des gefalteten Proteins führen, stellt dabei das zentrale Problem bei deren Berechnung dar. Zu diesem Zweck wurden von verschiedenen Autoren rechengestützte Methoden entwickelt, die alle zwei Komponenten gemeinsam haben: (i) eine Energiefunktion, welche die Eignung einer bestimmten Sequenz für eine Struktur evaluiert und (ii) einen Algorithmus für die Suche nach Sequenzen, die in die Zielstruktur mit einer niedrigen Energie falten können [Gordon *et al.*, 1999; Voigt *et al.*, 2000; Jaramillo *et al.*, 2001; Mendes *et al.*, 2002; Park *et al.*, 2004; Ventura & Serrano, 2004; Pokala & Handel, 2005]. Im Allgemeinen favorisieren diese Energiefunktionen eine dichte Packung im hydrophoben Kern, die Ausbildung von Wasserstoffbrückenbindungen zwischen den Proteinrückgratatomen verschiedener Peptidbindungen, energiearme Torsionswinkel und die Lokalisation von hydrophoben Aminosäuren im Kern des Proteins bzw. polaren Aminosäuren an dessen lösungsmittelzugänglicher Oberfläche. Die Optimierung der Stabilität eines im Modellierungsprozess befindlichen Proteins erfolgt durch den Vergleich der Energie des Modelles mit einer geeigneten Referenzenergie. Diese Referenzenergien werden empirisch ermittelt oder entsprechen durchschnittlichen Energien von Aminosäuren eines bestimmten Typs aus Modellpeptiden [Wernisch *et al.*, 2000; Kuhlman & Baker, 2000; Liang & Grishin, 2004; Pokala & Handel, 2005]. Eine entscheidende Bedeutung bei der Modellierung von Aminosäuresequenzen zu einer gegebenen Struktur entfällt auf die Vorhersage der Aminosäureseitenkettenkonformationen (Rotamere). Ausgehend von der bekannten Geometrie des Proteinrückgrats sollen unter Verwendung einer Rotamerbibliothek die einzelnen Reste in energetisch günstiger Konformation eingesetzt werden. Dieser äußerst komplexen Problemstellung konnte mit exakten Algorithmen, wie dem *dead-end-elimination-theorem* (*DEET*) [Desmet *et al.*, 1992; Lasters *et al.*, 1997; De *et al.*, 2000; Looger & Hellinga, 2001] oder approximierenden Methoden wie den Monte-Carlo-Algorithmen [Holm & Sander, 1991; Liang & Grishin, 2002] auf wirkungsvolle Weise begegnet werden. Sofern die *DEET*-Algorithmen konvergieren, erreichen sie das globale Minimum der Seitenkettengeometrie. Bei den *DEET*-Verfahren werden iterativ paarweise Kombinationen von solchen Rotameren eliminiert, die nicht zum globalen Minimum führen können. Im Unterschied dazu können Monte-Carlo-Methoden das Auffinden des globalen Minimums nicht garantieren, sie finden jedoch fast immer eine Lösung mit niedriger Energie [Canutescu *et al.*, 2003].

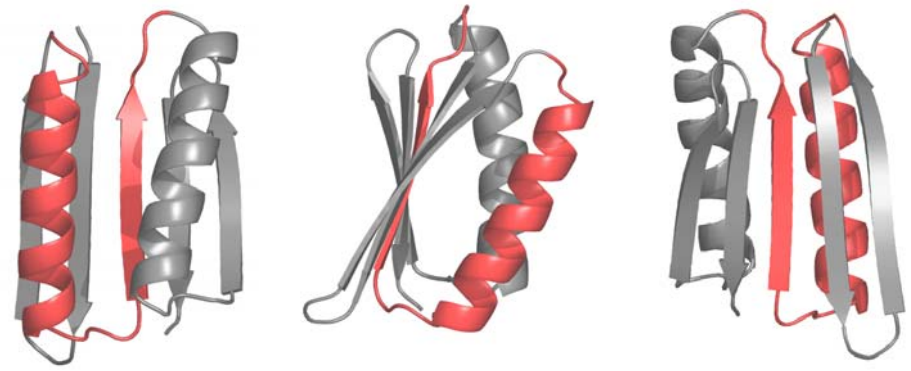


Abbildung I-3 Sekundärstrukturdarstellung des Proteins Top7 (*PDB*-Code 1QYS) [Kuhlman *et al.*, 2003]. Die Abbildung wurde mit Hilfe des Programms *PyMOL* erstellt [Delano, 2002].

Das Design von Aminosäuresequenzen beschränkte sich bisher hauptsächlich auf natürlich vorkommende Proteinstrukturen. Ein erster herausragender Erfolg wurde mit dem computergestützten *Redesign* einer Aminosäuresequenz für die Zinkfingerfaltungstopologie erzielt [Dahiyat & Mayo, 1997; Dahiyat *et al.*, 1997]. In weiterführenden Arbeiten gelang unter Anwendung rechengestützter Methoden die Stabilisierung von Enzymen [Korkegian *et al.*, 2005], die Solubilisierung von Membranproteinen [Slovic *et al.*, 2004], das *Redesign* von Protein-Protein-Interaktionen [Chevalier *et al.*, 2002] und die Generierung neuartiger Enzyme [Dwyer *et al.*, 2004]. Durch ein erneutes Redesign der Sequenz des Zinkfingermotivs konnte die Ausbildung einer dynamischen Struktur erzielt werden, die in Abhängigkeit der Anwesenheit von Zinkionen zwischen zwei Konformationszuständen fluktuieren kann [Ambroggio & Kuhlman, 2006]. Ein Meilenstein im computergestützten Proteindesign wurde mit der Errechnung einer Aminosäuresequenz zu einer artifiziellen Faltungstopologie [Kuhlman *et al.*, 2003] erreicht. Mit dieser Arbeit zeigten die Autoren, dass mit Hilfe der 20 proteinogenen Aminosäuren Sequenzen zu Topologien errechnet werden können, die nicht aus der natürlichen Evolution hervorgegangen sind, eine geeignete Sequenz also jede beliebige modellierbare Struktur kodieren kann. Das von Kuhlman und Mitarbeitern modellierte artifizielle Protein wird als Top7 bezeichnet (*PDB*-Code 1QYS). Wie die Abbildung I-3 zeigt, besteht Top7 aus zwei β/α/β-Motiven mit insgesamt 92 Aminosäuren. Da Top7 eine bis dahin unbekannte Faltungstopologie aufwies, folglich in der *PDB* keine *template*-Struktur vorhanden war, bestand der initiale Schritt vor der Errechnung einer geeigneten Proteinsequenz zu dieser Faltungstopologie in der Berechnung von Strukturmodellen. Dazu wurde das Programm *Rosetta* [Bystroff & Shao, 2002; Bradley *et al.*, 2003; Rohl *et al.*, 2004] verwendet, welches zur *de novo* Strukturvorhersage von Aminosäuresequenzen verwendet wird. Die Erstellung der Strukturmodelle erfolgte dabei durch eine Verknüpfung von Tri- und Nonapeptidfragmenten, die aus Proteinstrukturen der *PDB* entlehnt wurden [Kuhlman *et al.*, 2003]. Die Berechnung von Aminosäuresequenzen zu diesen Modellen wurde unter Verwendung des *RosettaDesign* Monte-Carlo-Suchprotokolles durchgeführt [Kuhlman & Baker, 2000]. Dabei erfolgte eine parallele Optimierung von Aminosäuresequenz und Proteinrückgratgeometrie. Die Annahme der einzelnen Sequenzen erfolgte nach deren Bewertung mit Hilfe des Metropolis-Kriteriums [Metropolis *et al.*, 1953].

Der Monte-Carlo-Algorithmus zählt zu den stochastischen Algorithmen und wird häufig bei Optimierungsproblemen eingesetzt, bei denen das Minimum einer Funktion gesucht wird. Einfache Minimierungsverfahren, bei denen man sich energetisch nur „bergab“ bewegt, versagen häufig bei komplexen Fragestellungen, da sie nicht die Möglichkeit besitzen, lokale Minima einer Funktion zu verlassen. Bei Monte-Carlo-Verfahren verwendet man Zufallszahlen zur

Bearbeitung von Problemstellungen, bei denen eine exakte Berechnung der Lösung schwierig ist. Um mit einem Monte-Carlo-Algorithmus das Minimum einer Funktion mit einer Vielzahl von Variablen zu bestimmen, geht man davon aus, dass man für jede Variable eine Energie berechnen kann und durch Summation dieser Energien die Gesamtenergie $E(X)$ des Systems erhält, wobei X für die Gesamtheit der Variablen steht. Der allgemeine Ablauf eines Monte-Carlo-Algorithmus gestaltet sich dann wie folgt:

i. Erzeuge eine Menge zufälliger Werte für die Einzelvariablen, um eine Ausgangskonfiguration vorzugeben. Berechne die Energie der Konfiguration $E = E(X)$
ii. Störe die Einzelvariablen und erzeuge somit eine Nachbarkonfiguration: $X \rightarrow X'$
iii. Berechne die Energie der neuen Konfiguration $E' = E(X')$
iv. Die Bewertung der neuen Konfiguration erfolgt mit Hilfe des Metropolis-Kriteriums:

   a. Wenn die Energie abgenommen hat, d.h. wenn $E(X) > E(X')$, akzeptiere diesen Schritt und übernimm X' als neue Ausgangskonfiguration
   b. Wenn die Energie zugenommen hat oder gleich geblieben ist, d.h. wenn $E(X) \leq E(X')$, dann berechne $e^{-\frac{\Delta E}{kT}}$ und vergleiche diesen Wert mit einer Zufallszahl λ aus [0…1], hierbei ist k = Boltzmann Konstante und $\Delta E = E(X') - E(X)$. Wenn $\lambda \leq e^{-\frac{\Delta E}{kT}}$, dann wird die neue Konfiguration trotz größerer Energie akzeptiert, sonst wird die Konfiguration X' verworfen und die alte Konfiguration X beibehalten. T beschreibt in diesem Zusammenhang nicht die absolute Temperatur, sondern lediglich einen numerischen Parameter, der die Wahrscheinlichkeit kontrolliert, dass eine neue Konfiguration akzeptiert wird, deren Energie größer ist, als die der aktuellen Konfiguration. Durch Anwendung dieses Prinzips ist es somit möglich, ein lokales Minimum einer Funktion zu verlassen. Beim *simulated annealing*, einer abgewandelten Form des Monte-Carlo-Algorithmus, wird T am Beginn der Rechnung zunächst ein hoher Wert zugewiesen, der dann langsam gesenkt wird. In Analogie zu dem langsamen Abkühlen eines Festkörpers, wird dabei versucht, die niedrigste Energie, d.h. die optimale Lösung des Optimierungsproblems zu erhalten [Kirkpatrick S. *et al.*, 1983].

v. Kehre zu Schritt (ii) zurück. Die Berechnung der Lösung erfolgt somit iterativ. Nach einer zuvor festgelegten Anzahl an Zyklen oder wenn sich nach einer definierten Anzahl von Zyklen die Energie des Systems nicht mehr ändert, wird die Berechnung gestoppt, so dass die aktuelle Konfiguration des Systems als Lösung des Optimierungsproblems erhalten wird.

Es lässt sich beweisen, dass die Folge von Zuständen, die der Metropolisalgorithmus generiert (wenn er lange genug läuft), zum globalen Minimum des Systems konvergiert [Herges T.A., 2003].

Der in *RosettaDesign* implementierte Suchalgorithmus wird von einer Energiefunktion aus 11 Termen kontrolliert, die neben *ab initio* ebenso wissensbasierte Energiefunktionen (*database-derived-potentials*) verwendet [Kuhlman *et al.*, 2003]. Die Anwendung dieses Algorithmus auf eine randomisierte Aminosäuresequenz führte zu verschiedenen Sequenzen, von

denen eine nach rekombinanter Herstellung zu einem Protein führte, das strukturell charakterisiert werden konnte (*PDB*-Code 1QYS). Die Abweichung zwischen Modell und Struktur betrug lediglich 1.17 Å (*rmsd*-Wert Proteinrückgrat). *RosettaDesign* wurde weiterhin bei dem *Redesign* von Proteinen mit bekannter Struktur verwendet [Dantas *et al.*, 2003], der Errechnung von Aminosäuresequenzen, die zwischen zwei Strukturen fluktuieren können [Ambroggio & Kuhlman, 2006] und bei der rechengestützten Identifizierung von Fragmenten aus amyloidogenen Proteinen, die von sich aus amyloidartige Fibrillen bilden können [Thompson *et al.*, 2006].

Die Anwendung heuristischer[1] Algorithmen muss wegen der mit ihnen verbundenen Zufallskomponente nicht immer zu einer geeigneten Lösung oder überhaupt zu einer Lösung führen [Schulze-Kremer, 1995]. Da der Prozess der Proteinfaltung nicht zufallsbehaftet ist [Levinthal, 1968], könnte die Möglichkeit bestehen, deterministische Verfahren zur Berechnung von Aminosäuresequenzen zu einer gegebenen Proteinstruktur zu entwickeln. Im Unterschied zu den nichtdeterministischen Verfahren führen deterministische Prozesse immer zur gleichen Lösung. Das *building block* Faltungsmodell [Lesk & Rose, 1981; Baldwin & Rose, 1999a; Baldwin & Rose, 1999b; Tsai & Nussinov, 2001; Tsai *et al.*, 2002], welches den Vorgang der Proteinfaltung als einen Prozess der kombinatorischen Assemblierung von Protein *building blocks* beschreibt, stellt unter diesem Aspekt ein sehr praktisches Modell dar. Durch eine Assemblierung dieser *building blocks* aus verschiedenen Proteinen konnten neue Proteine modelliert werden, deren Stabilität *in silico* durch Molekulardynamiksimulationen bestätigt werden konnte [Tsai *et al.*, 2004]. Ein Protein *building block* ist nach Tsai über seine Packungsdichte, sein Hydrophobizitätsprofil und über den Grad seiner Unabhängigkeit definiert. Der Begriff *Unabhängigkeit* soll in diesem Zusammenhang die Möglichkeit beschreiben, dass ein *building block* als isolierte strukturelle Einheit existieren kann. Die experimentelle Charakterisierung der Protein *building blocks* erfolgte durch limitierte Proteolyse [Tsai *et al.*, 2002]. Die erhaltenen Fragmente wiesen eine Länge von mindestens 15-20 Aminosäuren auf. Basierend auf diesen Ergebnissen konnte ein Algorithmus entwickelt werden, der Proteinstrukturen in *building blocks* aufspalten kann, die konsistent mit den Fragmenten aus limitierter Proteolyse sind [Tsai *et al.*, 2000]. Es wird vermutet, dass diese Fragmente unabhängig voneinander falten [Haspel *et al.*, 2003].

Die Assemblierung von Fragmenten zu Proteinen wurde ebenso unter Anwendung genetischer Algorithmen durchgeführt [Voigt *et al.*, 2002]. Wie die Monte-Carlo-Algorithmen gehören die genetischen Algorithmen zu den stochastischen Prozessen.

Die einfachste Möglichkeit der Beschreibung eines Proteins und damit ein mögliches Modellierungsschema, besteht in der ausschließlichen Verteilung apolarer Aminosäuren in dessen Kern und der Positionierung polarer Aminosäuren an der lösungsmittelexponierten Oberfläche des Proteins. Dieses Designprinzip nutzt den hydrophoben Effekt der apolaren Reste bei der Strukturbildung eines globulären Proteins, der als eine wichtige Triebkraft bei der Proteinfaltung angesehen wird [Dill, 1990]. Yue und Dill haben jedoch darauf hingewiesen, dass neben einem Minimum an notwendigen hydrophoben Kontakten in einem Protein bei einer Sättigung der Aminosäuresequenz mit apolaren Aminosäuren die Gefahr besteht, dass das Protein alternative Konformationen mit niedriger Freier Enthalpie findet [Yue & Dill, 1992].

[1] „Mit Heuristik bezeichnet man Strategien, die das Finden von Lösungen zu Problemen ermöglichen sollen, zu denen kein mit Sicherheit zum Erfolg führender Algorithmus bekannt ist." (Zitat aus wikipedia.org)

Kamtekar und Mitarbeiter konnten dennoch zeigen, dass die Anwendung eines einfachen binären Musters auf eine Vier-Helix-*bundle* Struktur und deren Expression innerhalb einer Genbibliothek in einigen Fällen zu kompakten α-helikalen Strukturen führt [Kamtekar *et al.*, 1993]. Durch eine binäre Kodierung von Sequenzen innerhalb kombinatorischer Bibliotheken wurden in anderen Arbeiten erfolgreich α-helikale und β-Proteine [Wang & Hecht, 2002; Wei *et al.*, 2003], katalytisch aktive Enzyme [Moffet *et al.*, 2000; Wang & Hecht 2002], amyloidogene Proteine [West *et al.*, 1999] und proteinbasierte Biomaterialien [Brown *et al.*, 2002] hergestellt. Mayo und Marshall konnten zeigen, dass die Kodierung einer Proteinstruktur mit einem hydrophoben Muster unter gleichzeitiger Bewertung der errechneten Aminosäuresequenz mit Hilfe von Kraftfeldern *gezielt* zu Proteinsequenzen führen kann, die in eine stabile Struktur falten [Marshall & Mayo, 2001].

I.4. Die Zielstellung der Arbeit

Das Problem der geringen Datenvielfalt (in nichtredundanten Sequenzdatenbanken von Proteinstrukturen und nichtredundanten Strukturdatenbanken) ist das fundamentale Hindernis bei der Aufdeckung von Sequenz-Struktur-Korrelationen [Solis & Rackovsky, 2002]. Aus diesem Grund sind derzeit keine ausreichenden statistischen Informationen über Konformationen von Oligopeptiden mit einer Mindestlänge von vier Aminosäuren verfügbar. Die Berechnung von alternativen Aminosäuresequenzen zu einer gegebenen Proteinstruktur beschränkt sich daher auf das bloße Verknüpfen von Fragmenten bzw. *oligons* [Micheletti *et al.*, 2000], deren Konformation eine Kompatibilität mit der Zielstruktur zeigt. Die Verwendung von Protein *building blocks* limitiert aufgrund deren Größe die Möglichkeit, Proteinsequenzen vollständig neu zu designen. Eine Anwendung heuristischer Methoden, wie genetische oder Monte-Carlo-Algorithmen, und die mit ihnen verbundene Zufallskomponente müssen nicht immer zu einer geeigneten Lösung oder überhaupt zu einer Lösung führen [Schulze-Kremer, 1995].

Ein deterministisches Verfahren auf Fragmentbasis zur Berechnung von alternativen Aminosäuresequenzen zu einer gegebenen Proteinstruktur muß für jedes Fragment der Sequenz die Wahrscheinlichkeit verschiedener möglicher Konformationszustände bewerten können und daraus folgend in der Lage sein, Sequenzen zu berechnen, die eine maximale Wahrscheinlichkeit für die Zielstruktur besitzen. Die Zielsequenzen dürfen nicht zufällig entstehen, sondern müssen in Abhängigkeit von den gewählten Randbedingungen immer als gleiche Lösung am Ende des Modellierungsprozesses stehen. Bei Tetrapeptiden konnten unter Verwendung eines reduzierten Alphabetes von Aminosäuren bereits ausgeprägte strukturelle Präferenzen nachgewiesen werden [Rackovsky, 1995; Sudarsanam & Srinivasan, 1997]. Eine Strukturspezifität bei Verwendung des vollständigen Satzes der proteinogenen Aminosäuren könnte die Entwicklung eines Modellierungsschemas auf Tetrapeptidfragmentbasis erlauben. Infolge der geringen Größe dieser Fragmente wäre ein vollständiges *Redesign* von Aminosäuresequenzen möglich.

Das erste Ziel der vorliegenden Arbeit war es daher, die Konformationseigenschaften von Tetrapeptiden näher zu charakterisieren. Die Strukturanalyse dieser Fragmente sollte die Betrachtung des ψ-Winkels der zweiten und des φ-Winkels der dritten Aminosäure umfassen (vgl. Abbildung I-2 auf Seite 6) und mögliche strukturelle Präferenzen aufdecken. Vor Beginn dieser Untersuchungen musste eine geeignete Methodik zur Datenaufbereitung von Proteinstrukturen

entwickelt werden, die ohne Verletzung der Nichtredundanz-Bedingung genügend Strukturinformationen zur Durchführung einer statistischen Konformationsanalyse dieser Fragmente bereitstellen konnte.

Im zweiten Abschnitt der Arbeit sollte ein Algorithmus entwickelt werden, der Informationen über mögliche bevorzugte Tetrapeptidkonformationen verwendet und daraus für experimentell bestimmte Proteinstrukturen alternative Aminosäuresequenzen errechnen kann. Die Funktionsfähigkeit des Algorithmus sollte an der Faltungstopologie des Proteins Top7 [Kuhlman *et al.*, 2003] überprüft werden. Dazu sollten Aminosäuresequenzen berechnet werden, die zur Sequenz von Top7 Identitäten von kleiner als 30 % aufweisen sollten.

Im dritten, experimentellen Teil der Arbeit sollten die modellierten Proteine rekombinant in *Escherichia coli* hergestellt werden. Die Charakterisierung der exprimierten Proteine sowie deren Strukturaufklärung sollten abschließend zeigen, inwieweit die errechneten Modelle zu kooperativ faltenden Proteinen führen.

II. Theoretische Methoden

II.1. Programmiersprachen, Entwicklungsumgebung, Daten & Rechner

Alle in der vorliegenden Arbeit untersuchten Proteinstrukturen wurden der *Brookhaven Protein Database* (*PDB*) [Berman *et al.*, 2000b, Berman *et al.*, 2002] am 8.12.2003 entnommen (ftp.rcsb.org).

Die Implementierung der Algorithmen zur Datenaufbereitung bzw. -analyse und zur Bearbeitung aller Problemstellungen, deren Lösung mehrere Tage oder Wochen Rechenzeit benötigten, erfolgte in den Programmiersprachen C/C++. Dies beinhaltete im Besonderen die Algorithmen zur Berechnung der ψ-φ-Diederwinkel des Proteinrückgrats, die Durchführung der *all-against-all* Alignments, die Auswertung der Wahrscheinlichkeitsdichtefunktionen und die durchgeführten Kreuzvalidierungen. Die Umsetzung der Algorithmen zur Berechnung von Aminosäuresequenzen zu einer gegebenen Proteinstruktur erfolgte in der Programmiersprache C#. Die Implementierung der Programme in den Sprachen C/C++ und C# erfolgte in der Entwicklungsumgebung Microsoft Visual Studio .NET Enterprise Edition Version 7.1.3088. Alle Rechungen wurden auf einem Intel Pentium IV HT 3.0 GHz mit 800 MHz FSB und 512 MB PC3200 RAM ausgeführt.

II.2. Durchführung sequenzbasierter *all-against-all* Alignments

Die Durchführung von *all-against-all* Alignments verfolgt das Ziel der Erstellung repräsentativer Datensätze. Dabei werden redundante Informationen entweder auf Sequenzebene (sequenzbasierte Alignments) oder auf Strukturebene (Struktur-Struktur-Alignments) beseitigt. Je nach Fragestellung können so die erforderlichen Datensätze generiert werden.

In der vorliegenden Arbeit wurden sequenzbasierte *all-against-all* Alignments durchgeführt, da die Information benötigt wurde, inwieweit unterschiedliche Sequenzen die gleiche Struktur ausbilden können. Die Erstellung dieser Datensätze erfolgte in Anlehnung an den *select-until-done*-Algorithmus [Hobohm *et al.*, 1992]. Das Prinzip dabei ist, dass aus einer Gruppe von Aminosäuresequenzen eine Sequenz ausgewählt wird (*template*-Sequenz) und alle anderen Sequenzen (*query*-Sequenzen) gegen die *template*-Sequenz alignt werden. Eine zuvor definierte maximale Sequenzidentität zwischen *template*-Sequenz und *query*-Sequenz dient als *cut off*-Wert. Wird dieser Wert überschritten, dann entfernt man die aktuelle *query*-Sequenz aus der Liste mit den *query*-Sequenzen. Dieser Vorgang wird solange wiederholt, bis alle *query*-Sequenzen gegen die *template*-Sequenz alignt wurden. Aus den verbliebenen *query*-Sequenzen wird eine neue *template*-Sequenz gewählt und die verbliebenen *query*-Sequenzen wiederum gegen die neue *template*-Sequenz alignt. Dieser Prozess wird so lange wiederholt, bis keine *query*-Sequenzen mehr vorhanden sind. Alle gewählten *template*-Sequenzen haben eine Sequenzidentität von kleiner oder gleich dem gesetzten *cut off*.

Die Durchführung der Alignments erfolgte gemäß dem Algorithmus von Needleman-Wunsch [Needleman & Wunsch, 1970] unter Verwendung affiner *gap*-Strafen [Gotoh, 1982].

Als *open penalty* wurde ein Wert von -5 und als *extension penalty* ein Wert von -2 definiert. Der *cut off* für die maximal erlaubte gegenseitige Sequenzidentität in einem nichtredundanten Datensatz wurde auf einen Wert von 25 % festgelegt. Es wurden semiglobale Alignments unter Verwendung der BLOSUM62-Substitutionsmatrix [Henikoff & Henikoff, 1992] durchgeführt. Die BLOSUM62-Matrix ist im Anhang VII.9, S. 138, dargestellt.

II.3. Errechnung der Wahrscheinlichkeitsdichtefunktionen

II.3.1. Nichtparametrische Kerndichteschätzung

Bei der Durchführung einer Regression wird versucht, eine funktionelle Beziehung zwischen unterschiedlichen Messgrößen zu finden. In der parametrischen Regression ist diese Beziehung *a priori* festgelegt (z. B. eine Gerade). Bei einer nichtparametrischen Regression hingegen bestimmen die Datenpunkte selbst die funktionelle Form. Es erfolgen keine Annahmen über die Verteilungen von Daten, wodurch eine fehlerhafte Spezifikation des Modelles vermieden wird. Die statistische Analyse der ψ_2-φ_3-Verteilungen in der vorliegenden Arbeit erfolgte mit Hilfe der nichtparametrischen Kerndichteschätzung. Das Ziel dieses Verfahrens ist, mit Hilfe einer Kernfunktion *K* die unbekannte Dichte *f* einer stetigen Zufallsvariablen zu schätzen. Im Ergebnis kann die Struktur der Daten, wie deren Modalität oder Symmetrie, beurteilt werden. Zur Konstruktion des Kerndichteschätzers wird eine Kernfunktion, deren Varianz von einem Parameter *h* kontrolliert wird, über jede Beobachtung gelegt und dann gemittelt. Der Parameter *h* wird als Bandweite bezeichnet. Eine sehr bekannte mathematische Beschreibung eines Kerndichteschätzers ist der *Parzen-Rosenblatt-Kerndichteschätzer* (Gleichung II-1). Wenn *x* den Punkt bezeichnet, an dem die Dichte *P(x)* geschätzt werden soll, dann ist der Kerndichteschätzer definiert als:

II-1 $$\hat{f}_h(x) = \frac{1}{nh}\sum_n K\left(\frac{x - X_i}{h}\right)$$

mit *K* als Kernfunktion. $\{X_1,...,X_n\}$ bezeichnen die beobachteten Daten. Es existiert eine Vielzahl möglicher Kernfunktionen. Diese müssen die Eigenschaft einer Dichtefunktion erfüllen, d.h.:

II-2 $$\int_{-\infty}^{+\infty} K(\zeta)d\zeta = 1 \text{ mit } K(\zeta) \geq 0\,.$$

Kernfunktionen sind in der Regel um Null symmetrisch und unimodal. Beispiele oft verwendeter Kernfunktionen sind die Dichte der Standardnormalverteilung (Gauss-Kern), die Dreiecksdichte, der Rechteck-Kern oder der Epanechnikov-Kern. Der Gauss-Kern ist in der Gleichung II-3 definiert. Er wurde als Kernfunktion in der vorliegenden Arbeit gewählt.

II-3 $$K(\zeta) = \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{\zeta^2}{2\pi}\right)}$$

Die Form von Kerndichteschätzern wird durch die Bandweite beeinflusst. Wenn *h* zu groß gewählt wird, so erfolgt ein „Überglätten“ (*overfitting*) der Funktion. Die besondere Struktur der Daten kann dabei verloren gehen. Zu klein gewählte Bandweiten führen zu einem „Unterglätten“ der Daten und lokale Streuungen bekommen einen zu großen Einfluss auf den Gesamtverlauf der Dichteschätzung. Nach Anwendung von Dichtefunktionen auf eine Problemstellung und anschließender Auswertung der Ergebnisse kann es daher notwendig sein, die Bandweiten anzupassen.

II.3.2. Errechnung von Wahrscheinlichkeitsdichtefunktionen[2]

Die Auswertung der ψ_2-φ_3-Winkelverteilungen der einzelnen Tetrapeptide erfolgte durch nichtparametrische Kerndichteschätzung mit einem Gauss-Kern als Kernfunktion. Dazu wurde das Statistikprogramm „R“ in der Version 1.7.1 verwendet (http://www.r-project.org). Die eigentliche Kerndichteschätzung erfolgte mit der Bibliothek „*sm*“ [Bowman & Azzalini, 1997]. Es wurde eine Bandweite von $h = 15°$ festgelegt. Die errechneten Wahrscheinlichkeitsdichtefunktionen wurden auf 1 normiert und im Textformat gespeichert. Die Auflösung der Dichtefunktionen betrug 2°. Aus der verwendeten Auflösung ergab sich eine Datengröße von 670 KB pro Dichtefunktion. Höhere Auflösungen waren aufgrund des zu großen Speicherplatzbedarfes und zu langer Rechenzeit nicht mehr praktikabel.

II.3.2.1. Das Randproblem

Bei der Berechnung der Wahrscheinlichkeitsdichtefunktionen ist zu beachten, dass die Diederwinkel beim Übergang von Winkeln von +180° nach -180° einer Änderung ihres Vorzeichens unterliegen. Dies bedeutet, dass prinzipiell jeder Winkel ψ_2 und φ_3 Wahrscheinlichkeitsdichte in einen Winkelbereich mit umgekehrtem Vorzeichen streuen kann. Eine Dichtefunktion ist nur dann richtig, wenn sie diesen Umstand beachtet. Die Berechnung der Wahrscheinlichkeitsdichtefunktionen mit der *sm*-Bibliothek für den Bereich von $-180° \le \psi_2,\varphi_3 \le +180°$ führt zu einem Abschneiden von Wahrscheinlichkeitsdichte an den Rändern der Funktionen. Dieser Fehler lässt sich korrigieren, wenn man berücksichtigt, dass man die Diederwinkel ψ und φ einer Aminosäure als Azimut und Höhe auf einer Kugeloberfläche beschreiben kann. Die Darstellung der Streuung eines Punktes auf einer Kugeloberfläche mit den Koordinaten (ψ_2,φ_3) in die Ebene für den Konformationsbereich $-180° \le \psi_2,\varphi_3 \le +180°$ gelingt damit formal durch dessen Erweiterung auf $-360° \le \psi_2,\varphi_3 \le +360°$ und gleichzeitige Transformation jedes Diederwinkelpaares (ψ_2,φ_3) auf die drei Wertepaare:

II-4 $$\left(\psi_2, -360° \cdot \mathrm{sgn}(\phi_3) + \phi_3\right)$$

II-5 $$\left(-360° \cdot \mathrm{sgn}(\psi_2) + \psi_2, \phi_3\right)$$

II-6 $$\left(-360° \cdot \mathrm{sgn}(\psi_2) + \psi_2, -360° \cdot \mathrm{sgn}(\phi_3) + \phi_3\right)$$

[2] Die Termini Wahrscheinlichkeitsdichtefunktion und Dichtefunktion werden in der vorliegenden Arbeit äquivalent verwendet.

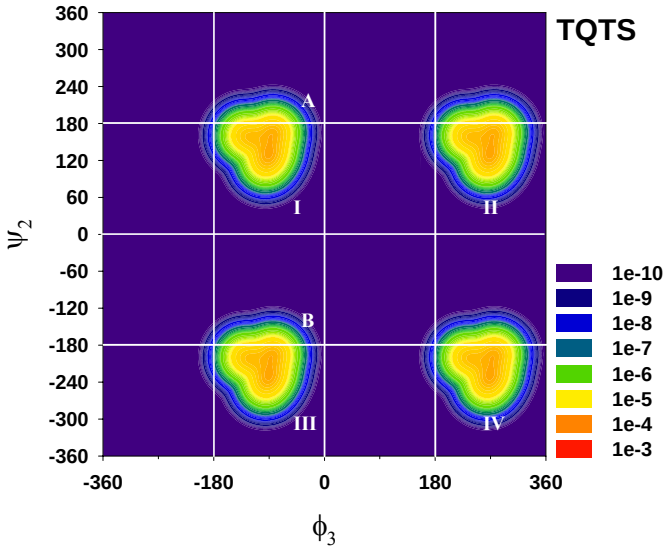


Abbildung II-1 Wahrscheinlichkeitsdichtefunktion der ψ_2-φ_3-Verteilung des Tetrapeptids TQTS für den Konformationsbereich $-360° \leq \psi_2,\phi_3 \leq +360°$. Die Skala gibt die absolute Wahrscheinlichkeit der einzelnen Diederwinkelpaare an. Die vier mittleren weißen Quadrate markieren den normalen Konformationsbereich von $-180° \leq \psi_2,\phi_3 \leq +180°$. Bei **A** streut Wahrscheinlichkeitsdichte aus diesem Bereich heraus. Durch Projektion jedes Diederwinkelpaares (ψ_2,φ_3) aus **I** gemäß (II-4) in den Bereich **II**, gemäß (II-5) in den Bereich **III** und gemäß (II-6) in den Bereich **IV** erfolgt bei **B** eine Streuung in den normalen Konformationsbereich zurück. Es wird nur dieser Konformationsbereich der Dichtefunktion gespeichert.

Die Kerndichteschätzung erfolgt für den Konformationsbereich $-360° \leq \psi_2,\varphi_3 \leq +360°$, die Wahrscheinlichkeitsdichtefunktion wird jedoch nur für den Bereich $-180° \leq \psi_2,\varphi_3 \leq +180°$ gespeichert. Im Ergebnis erfolgt eine Streuung über den „Rand" der „normalen" Dichtefunktion hinaus in den Bereich mit umgekehrtem Drehsinn. Die Abbildung II-1 illustriert diesen Vorgang am Beispiel der Dichtefunktion der ψ_2-φ_3-Verteilung des Tetrapeptids TQTS.

II.3.3. Bestimmung der Wahrscheinlichkeit für einen Konformationszustand

Der Konformationsbereich, den eine Aminosäure i mit ihrem Diederwinkelpaar (ψ_i,φ_i) beschreiben kann, lässt sich nach Tabelle II-1, S. 17, in fünf Bereiche klassifizieren, die als faltblatttypisch (*E*), helixtypisch (*H*), *G-turn* (*G*), *L-turn* (*L*) und *X-turn* (*X*) bezeichnet werden [Bystroff *et al.*, 2000]. Die vorliegende Arbeit verwendet die gleiche Einteilung zur Analyse bevorzugter Tetrapeptidkonformationen. Aus den errechneten Dichtefunktionen *p(t)* der ψ_2-φ_3-Verteilungen lässt sich durch Integration die Wahrscheinlichkeit quantifizieren, mit der ein Tetrapeptid eine Konformation des Typs *E*, *H*, *G*, *L* oder *X* annehmen kann. Die Verteilungsfunktion *F(x)* einer Dichtefunktion gibt die Wahrscheinlichkeit $P(X < x)$ dafür an, dass die Zufallsgröße *X* Werte kleiner als *x* annimmt:

II-7 $$F(x) = \int_{-\infty}^{x} p(t)dt$$

Daraus ableitend errechnet sich die Wahrscheinlichkeit, dass die Zufallsgröße *X* einen Wert in einem Intervall $[x_1,x_2]$ einer Dichtefunktion annimmt, gemäß Gleichung II-8 zu:

II-8 $$P(x_1 \le X < x_2) = P(X < x_2) - P(X < x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} p(t)dt$$

Die errechneten Dichtefunktionen besitzen eine Auflösung von 2°. Somit ergibt sich, unter Berücksichtigung von Tabelle II-1, die Wahrscheinlichkeit für die einzelnen Konformationszustände zu:

II-9 $$P(E) = 4 \cdot \sum_{\psi_2 \ge +40°}^{\psi_2 \le -100°} \sum_{\phi_3 \ge +160°}^{\phi_3 \le +20°} p(\psi_2\varphi_3)\Delta\psi_2\Delta\phi_3 \qquad \textbf{mit} \qquad \Delta\psi_2 = \Delta\phi_3 = 2°$$

II-10 $$P(H) = 4 \cdot \sum_{\psi_2 > -100°}^{\psi_2 \le -10°} \sum_{\phi_3 \ge -180°}^{\phi_3 \le 0°} p(\psi_2\varphi_3)\Delta\psi_2\Delta\phi_3 \qquad \textbf{mit} \qquad \Delta\psi_2 = \Delta\phi_3 = 2°$$

II-11 $$P(G) = 4 \cdot \sum_{\psi_2 > -10°}^{\psi_2 < +40°} \sum_{\phi_3 \ge -180°}^{\phi_3 \le 0°} p(\psi_2\varphi_3)\Delta\psi_2\Delta\phi_3 \qquad \textbf{mit} \qquad \Delta\psi_2 = \Delta\phi_3 = 2°$$

Die Konformationszustände *H* und *G* wurden nicht einzeln betrachtet. Im Sinne einer Zuweisung wurde definiert:

II-12 $$P(H) = P(H) + P(G)$$

II-13 $$P(L) = 4 \cdot \sum_{\psi_2 \ge -90°}^{\psi_2 \le +90°} \sum_{\phi_3 > 0°}^{\phi_3 \le +180°} p(\psi_2\varphi_3)\Delta\psi_2\Delta\phi_3 \qquad \textbf{mit} \qquad \Delta\psi_2 = \Delta\phi_3 = 2°$$

II-14 $$P(X) = 4 \cdot \sum_{\psi_2 > +90°}^{\psi_2 < -90°} \sum_{\phi_3 \ge +20°}^{\phi_3 \le +160°} p(\psi_2\varphi_3)\Delta\psi_2\Delta\phi_3 \qquad \textbf{mit} \qquad \Delta\psi_2 = \Delta\phi_3 = 2°$$

Der Faktor *4* resultiert aus der Verkleinerung des Konformationsbereiches der Dichtefunktionen von $-360° \le \psi_2,\varphi_3 \le +360°$ auf $-180° \le \psi_2,\varphi_3 \le +180°$ beim Speichern der jeweiligen Wahrscheinlichkeitsdichtefunktion.

Tabelle II-1 Konformationsbereiche, die jeweils einen möglichen Strukturtyp einer Aminosäure beschreiben. Nach [Bystroff *et al.*, 2000]. Die Konformationsbereiche *H* und *G* wurden in der vorliegenden Arbeit vereinigt.

Konformationstyp	Konformationsbereich ψ	Konformationsbereich φ
E (faltblatttypisch)	+ 40° < ψ < +260°	-200° < φ < + 20°
H (helixtypisch)	-180° < ψ < 0°	-100° < φ < - 10°
G (*G-turn*)	- 10° < ψ < + 40°	-180° < φ < 0°
L (*L-turn*)	- 90° < ψ < - 90°	0° < φ < +180°
X (*X-turn*)	+ 90° < ψ < - 90°	+ 20° < φ < +160°

II.4. Durchführung einer Kreuzvalidierung

Die Gültigkeit multivariater Modelle lässt sich durch eine Kreuzvalidierung überprüfen (*cross validation* oder auch *boot strapping*). Das prinzipielle Vorgehen bei diesem Verfahren ist, Modelldaten in zwei sich gegenseitig ausschließenden Mengen, die größere Trainingsmenge und eine kleinere Testmenge, aufzuteilen. Die größere Datenmenge wird dazu verwendet, ein Modell aufzustellen, während die kleinere Datenmenge dazu dient, das Modell zu bestätigen oder zu verwerfen, indem man das Modell auf die kleinere Datenmenge anwendet und die Ergebnisse mit den tatsächlichen Werten vergleicht.

Bei der vorliegenden Fragestellung sollte untersucht werden, inwieweit eine Korrelation zwischen dem Auftreten der wahrscheinlichsten Struktur von Fragmenten einer bestimmter Länge und der Wahrscheinlichkeit des wahrscheinlichsten Konformationszustandes der einzelnen Tetrapeptide dieser Fragmente besteht. Zur Vermeidung eines zirkulären Schlusses mussten dazu die Strukturinformationen über das jeweils aktuell betrachtete Protein aus den in der vorliegenden Arbeit errechneten Dichtefunktionen entfernt werden. Die Testmenge würde somit nur eine Proteinstruktur umfassen. Dazu wurde die betrachtete Struktur in ihre *n*-3 Tetrapeptide zerlegt. In den Datensätzen zur Berechnung der Dichtefunktionen der ψ_2-φ_3-Verteilung der einzelnen Tetrapeptide sind die *PDB*-Codes der Proteinstrukturen gespeichert, aus denen jedes Tetrapeptid stammt (vgl. Abbildung IV-2 auf Seite 37). Aus den jeweiligen *PDB*-Codes lässt sich nach Abbildung IV-1 auf Seite 36 die zugehörige Gesamtaminosäuresequenz rekonstruieren. Es wurde jede Gesamtaminosäuresequenz jedes Tetrapeptids aus dem entsprechenden Datensatz zur Berechnung der Dichtefunktionen gegen die Aminosäuresequenz des aktuell betrachteten Proteins alignt. Aus den Datensätzen zur Errechnung der Dichtefunktionen wurden alle Strukturdaten von Tetrapeptiden entfernt, deren Gesamtaminosäuresequenz eine Identität von größer als 25 % gegenüber dem des aktuell betrachteten Proteins zeigte und die gleichzeitig eine Winkelabweichung von $|\Delta\psi_2| < 25°$ und $|\Delta\varphi_3| < 25°$zwischen dem aktuell betrachteten Tetrapeptid aus der Aminosäuresequenz und dem aus dem Datensatz zur Berechnung der Dichtefunktionen aufwiesen. Für jedes untersuchte Protein erfolgte damit eine individuelle Berechnung der Dichtefunktionen der ψ_2-φ_3-Verteilung der einzelnen Tetrapeptide, aus denen die untersuchte Gesamtsequenz besteht.

II.5. Bestimmung des Grades einer Korrelation

Der Grad der Korrelation *g* (*degree of correlation*) zwischen zwei Ereignissen A und B kann als Quotient der bedingten Wahrscheinlichkeit von B, unter der Voraussetzung, dass A eingetreten ist, zu der unbedingten (*a priori*) Wahrscheinlichkeit von B alleine, nach Gleichung II-15 dargestellt werden. Damit lässt sich beschreiben, in welcher Weise A das Ereignis B beeinflusst.

II-15 $$g = \frac{P(B \mid A)}{P(B)} = \frac{P(AB)}{P(A)P(B)}$$

Ist $g > 1$, so sind die Ereignisse A und B positiv korreliert. Ein Wert von $g = 1$ lässt auf die Unabhängigkeit beider Ereignisse schließen. Wenn $g < 1$ ist, so sind die Ereignisse A und B negativ korreliert. Der rechte Teil der Gleichung II-15 folgt aus dem BAYES-Gesetz (Gleichung

II-16), das durch die Berücksichtigung bedingter Wahrscheinlichkeiten eine Verallgemeinerung der Multiplikationsregel für unabhängige Ereignisse darstellt.

II-16 $$P(AB) = P(B \mid A)P(A) = P(A \mid B)P(B)$$

Sind die Ereignisse A und B unabhängig vom Eintreten der Vorbedingung, so gilt $P(B \mid A) = P(B)$ und Gleichung II-16 entspricht der Multiplikationsregel für unabhängige Ereignisse.

Zur Bestimmung der Korrelation zwischen der Wahrscheinlichkeit des wahrscheinlichsten Konformationszustandes der einzelnen Tetrapeptide (*a priori* Wahrscheinlichkeit) in einer Sequenz und der im Ergebnis ausgebildeten wahrscheinlichsten Struktur dieses Fragmentes, werden Peptidfragmente aus einer Datenbank in ihre (*n*-3) Tetrapeptide zerlegt. Die Variable *n* bezeichnet die Anzahl der Aminosäuren in einem Fragment. Die Ereignisse A und B in Gleichung II-15 entsprächen dann der *a priori* Wahrscheinlichkeit zweier überlappender Tetrapeptide für ihren jeweils wahrscheinlichsten Konformationszustand. In diesem Fall würde die Analyse somit ein Pentapeptid umfassen. Bei Berücksichtigung längerer Fragmente lässt sich die Gleichung II-15 zu Gleichung II-17 verallgemeinern. Darin bezeichnen $X_1, X_2 \ldots X_n$ die jeweilige *a priori* Wahrscheinlichkeit für den wahrscheinlichsten Konformationszustand der einzelnen Tetrapeptide.

II-17 $$g = \frac{P(X_1 X_2 X_3 \cdots X_n)}{\prod_{i=1}^{n} P(X_i)}$$

Die bedingte Wahrscheinlichkeit, mit der ein Fragment einer bestimmten Länge (unabhängig von dessen Sequenz) seine wahrscheinlichste Struktur auch tatsächlich ausgebildet hat, ist die beobachtete Häufigkeit in der verwendeten Datenbank. Sie entspricht dem Zähler in Gleichung II-17. Der Nenner ist das Produkt der *a priori* Wahrscheinlichkeiten für den wahrscheinlichsten Konformationszustand aller Tetrapeptide des untersuchten Fragmentes.

II.6. Modellierung der Aminosäureseitenketten

Die Modellierung von Seitenketten an ein Proteinrückgrat erfolgte mit dem Programm *SCWRL 3* [Canutescu *et al.*, 2003]. *SCWRL 3* beruht auf der Verwendung von Rotamerbibliotheken, so dass eine anschließende Energieminimierung notwendig ist, um mögliche Überlappungen von Seitenkettenatomen verschiedener Aminosäuren zu korrigieren.

II.7. Energieminimierung der Proteinmodelle

Ein typisches Kraftfeld beschreibt die Interaktionen auf atomarer Ebene. Die Energieminimierung der Modelle erfolgte mit der Implementierung des GROMOS96 Kraftfeldes unter Verwendung des 43B1-Parameter *set* [van Gunsteren *et al.*, 1996] im *Swiss-PdbViewer 3.7* (*SP5*) [Guex & Peitsch, 1997]. Dabei wurden viermal 20 Zyklen im *steepest descent* Modus durch-

geführt. Die Energieminimierung erfolgte *in vacuo*. Als Ergebnis der Modellierung erhält man die energieminimierte Struktur und die Gesamtenergie jeder Aminosäure mit den sechs aufgeschlüsselten Energietermen. Dies sind die Energie der Bindungslängen, Bindungswinkel, Diederwinkel, der uneigentlichen Diederwinkel (*improper dihedral angles*), der elektrostatischen und der Lennard-Jones-Potentiale. Mögliche energetisch ungünstige Wechselwirkungen können somit auf Aminosäureebene erkannt werden. Aus den Energiebeiträgen der Einzelaminosäuren errechnet sich durch Summation die Gesamtenergie des Modelles.

III. Experimentelle Methoden

III.1. Material

III.1.1. Chemikalien, Enzyme & Kits

Chemikalien	Hersteller
Acrylamid (30%) / 0.8 % N,N'Methylenbisacrylamid	*Carl Roth (Karlsruhe)*
Agarose, für die Mikrobiologie	*Carl Roth (Karlsruhe)*
Ammoniumchlorid	*Sigma-Aldrich (Deisenhofen)*
Ammoniumperoxodisulfat	*Carl Roth (Karlsruhe)*
Ampicillin	*Sigma-Aldrich (Deisenhofen)*
Coomassie Brilliant Blue R250	*Sigma-Aldrich (Deisenhofen)*
EDTA	*ICN (Meckenheim)*
Essigsäure, 96 %, p.a.	*Merck (Darmstadt)*
Glucose, DAB Qualität	*Carl Roth (Karlsruhe)*
Glycerin, wasserfrei	*Merck (Darmstadt)*
GdmCl, *Ultrapure*	*ICN (Meckenheim)*
Imidazol	*Fluka (Bucks, CH)*
IPTG	*Sigma-Aldrich (Deisenhofen)*
Isopropanol	*Carl Roth (Karlsruhe)*
Kaliumchlorid	*Carl Roth (Karlsruhe)*
Kaliumdihydrogenphosphat	*Carl Roth (Karlsruhe)*
Kanamycin	*Carl Roth (Karlsruhe)*
Magnesiumsulfat Heptahydrat	*Carl Roth (Karlsruhe)*
Natriumchlorid	*Carl Roth (Karlsruhe)*
Natriumhydroxid	*Carl Roth (Karlsruhe)*
SDS	*ICN (Meckenheim)*
TEMED	*Carl Roth (Karlsruhe)*
Tris	*Applichem (Darmstadt)*

Alle nicht aufgeführten Chemikalien stammten von den Firmen ICN, Fluka und Sigma und hatten den Reinheitsgrad p.a.. Zur Herstellung von Puffern und Lösungen wurde entionisiertes Wasser aus einer *USF PURELAB Plus* Anlage verwendet.

Enzyme	Hersteller
Nde I	*New England BioLabs (Frankfurt)*
Hind III	*New England BioLabs (Frankfurt)*
Alkalische Phosphatase aus Eismeergarnelen	*Promega (Mannheim)*
T4 DNA Ligase	*New England BioLabs (Frankfurt)*
Thrombin	*Sigma (St.Louis, USA)*

Standards	Hersteller
1 kbp DNA-Längenstandard	*New England BioLabs (Frankfurt)*
LMW-SDS Marker Kit	*Amersham Biosciences (Freiburg)*

Kits	Hersteller
Qiagen Plasmid Mini Kit	*Qiagen (Hilden)*
QIAquick PCR Purification Kit	*Qiagen (Hilden)*
QIAquick MinElute Gel Extraction Kit	*Qiagen (Hilden)*

III.1.2. Bakterienstämme

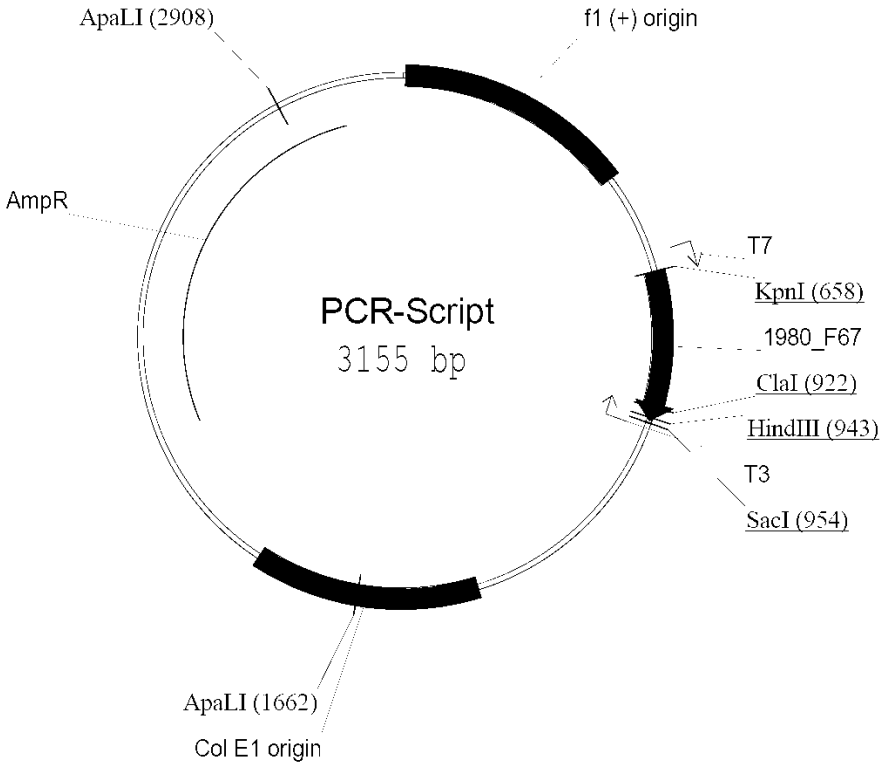
Stämme	Genotyp	Bezugsquelle
E. coli TOP10	F^-, *mcr*A, Δ(*mrr*$^-$ *hsd*RMS$^-$ *mcr*BBC), Φ80*lacZ* ΔM15, Δ*lac*X74 *rec*A1, *deo*R, *ara*D139, Δ(*ara-leu*)7697, *gal*U, *gal*K, *rps*L, (Str^R) *end*A1, *nup*G	*Invitrogen (Carlsbad, USA)*
E. coli BL21 (DE3)	B, F^-, *omp*T, *gal*, [*dcm*], [*lon*], *hsd*SB(rB^-mB^-), galλ(DE3)	*Novagen (Bad Soden)*

III.1.3. Plasmide

Plasmide	Bezugsquelle
pET28a	*Novagen (Bad Soden)*

III.1.4. Gene der Proteine M1-M8

Die codon-optimierte cDNA zu den errechneten Aminosäuresequenzen wurden von der *Geneart AG (Regensburg)* synthetisiert und in einen Klonierungsvektor PCR-Script unter Verwendung der KpnI und SacI Restriktionsschnittstelle kloniert. Die Gene wurden unter Verwendung von NdeI und HindIII aus dem Klonierungsvektor ausgeschnitten und in den Expressionsvektor pET28a umkloniert. Nachfolgend sind die Plasmidkarten des Vektors PCR-Script und des pET28a-Vektors dargestellt. Im Anhang VII.6, S. 129, werden die Gensequenzen der Proteine M1-M8 gezeigt.



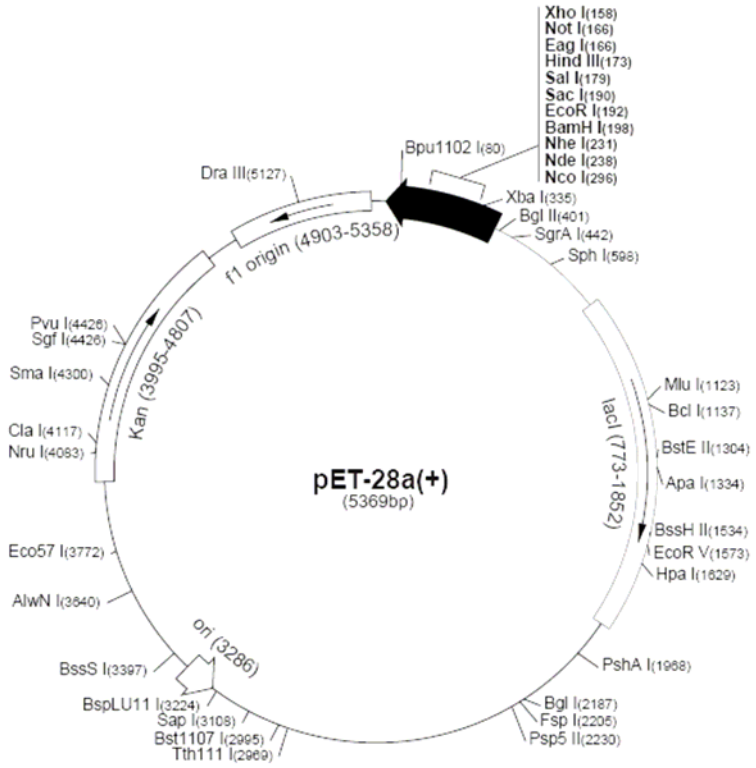


Abbildung III-1 Klonierungsvektor PCR-Script und Expressionsvektor pET-28a(+)

III.1.5. Puffer & Lösungen

Bezeichnung	Zusammensetzung
Laufpuffer Agarose-Gelelektrophorese (1xTAE)	40 mM Tris, 20 mM Essigsäure, 2 mM EDTA, pH 8.0
Laufpuffer Gelfiltration	25 mM Natriumphosphat pH 8.0, 400 mM NaCl
Laufpuffer SDS-Page	50 mM Tris, 380 mM Glycin, 0.1 % (w/v) SDS, pH 8.3 (Essigsäure)
PAGE-Färbelösung	10% (v/v) Essigsäure, 1 g/l Coomassie Brilliant Blue G250
PAGE-Fixierer	10% (v/v) Essigsäure, 25% (v/v) Isopropanol
PAGE-Entfärbelösung	10% (v/v) Essigsäure
PAGE 6 % Sammelgel (für 3 Gele)	1.2 ml Acrylamid (30%), 1.5 ml Sammelgelpuffer (4 x Stammlösung), 3.3 ml Wasser, 6 µl TEMED, 75 µl APS-Lösung (10% w/v)
PAGE 15 % Trenngel (für 3 Gele)	5 ml Acrylamid (30 %), 2.5 ml Trenngelpuffer (4 x Stammlösung), 2.5 ml Wasser, 8 µl TEMED, 150 µl APS-Lösung (10% w/v)
Probenauftragspuffer SDS-PAGE (5x Puffer) (nicht reduzierend)	250 mM Tris, 5 % (w/v) SDS, 50 % Glycerin, 0.005 % Bromphenolblau pH 8.0
Probenpuffer Agarose Gele	10 mM Tris, 1 mM EDTA, 50 % (w/v) Glycerin, 0.05 % (w/v) Bromphenolblau, pH 7.2
Sammelgelpuffer SDS-PAGE (4x Stammlösung)	0.5 M Tris, pH 6.8 (Essigsäure), 0.4 % (w/v) SDS
Trenngelpuffer SDS-PAGE (4x Stammlösung)	1.5 M Tris, pH 8.8 (Essigsäure), 0.4 % (w/v) SDS
Wasch - und Elutionspuffer für Ni-NTA Säule	25 mM Natriumphosphatpuffer pH 7.5, 400 mM NaCl mit jeweils 20 mM, 50 mM (Waschpuffer), 100 mM und 500 mM Imidazol (Elutionspuffer)
Zellaufschlusspuffer	25 mM Natriumphosphatpuffer pH 7.5, 400 mM NaCl,
Puffer zur Abspaltung des His_6-*tags* (Thrombinverdau)	10 mM Tris, pH 7.5 (HCl), 2 mM $MgCl_2$, 150 mM NaCl

III.1.6. Medien & Lösungen für die Kultivierung von *E. coli*

Bezeichnung	Zusammensetzung
Nährmedien	
LB-Medium	10 g/l NaCl, 10 g/l Trypton, 5 g/l Hefeextrakt
LB-Agar	LB-Medium + 15 g/l Agar-Agar
SOC-Medium	20 mM Glucose, 20 g/l Trypton, 5 g/l Hefeextrakt, 0.5 g/l NaCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$
Antibiotika	
Stammlösung Kanamycin: 35 mg/ml in Wasser	Konzentration im Medium: 35 µg/ml
Stammlösung Ampicillin: 100 mg/ml in Wasser	Konzentration im Medium: 100 µg/ml

Fortsetzung Tabelle Medien & Lösungen für die Kultivierung von *E. coli*

Bezeichnung	Zusammensetzung
Medien für Bioreaktorkultivierung (8 l Fermentationsvolumen, 6 l Startvolumen)	
Hefeextrakt-Vollmedium	400 g Hefeextrakt (CMV Hefewerk Hamburg), 4 g NH_4Cl, 1 ml Antischaum in 5 l Wasser lösen
Zusätze für 8 l Volumen (separat autoklaviert)	A) 40g Glucose / 500 ml Wasser B) 88g K_2HPO_4 / 500 ml Wasser C) 5.44 g $MgSO_4 \cdot 7H_2O$ / 500 ml Wasser
Feedinglösung	250 ml/l Glycerin, 300 g/l Hefeextrakt

III.1.7. Geräte & Zubehör

Gerätebezeichnung	Hersteller
Agarosegelelektrophoreseeinheit GNA-100 *submarine unit*	*Amersham Pharmacia Biotech, Freiburg*
Äkta Explorer 100	*Amersham Biosciences (Freiburg)*
Bioreaktor Biostat ED (10 l Arbeitsvolumen) mit digitalem Mess- und Regelsystem C-DCU und Prozessleitsystem MCFSwin	*Sartorius, Göttingen*
CD-Spektrometer Modell J810	*Jasco (Groß-Umstadt)*
Elektroporationsgerät Gene Pulser II	*Biorad (München)*
Elektroporationsküvetten (2 mm)	*Biorad (München)*
Gaulin-Hochdruckhomogenisator Micron Lab 40	*APV, Lübeck*
NMR-DRX 500 Spektrometer	*Bruker (Rheinstetten, Deutschland)*
Quarzglasküvetten Suprasil ®	*Hellma (Müllheim)*
SDS-Gelelektrophoreseeinheit *Mighty small* II SE250/260	*Amersham Pharmacia Biotech, Freiburg*
UV/VIS Spektrometer Ultrospec 4000	*Amersham Pharmacia Biotech, Freiburg*
Zentrifugen Avanti J20, J25 und J30I	*Beckman (München)*
Säulen und Säulenmaterialien	**Hersteller**
Ni-NTA Agarose	*Qiagen (Hilden)*
Superdex S75 Gelfiltrationssäule XK 16/60 *Prepgrade*	*Amersham Biosciences (Freiburg)*
C_{18}-*ZipTips*®	*Millipore (Schwalbach)*
Sonstiges	**Hersteller**
Dialyseschläuche	*Roth (Karlsruhe)*
Zentrifugationskonzentratoren Amicon-Ultra 15 MWCO 5 kDa	*Millipore (Schwalbach)*

III.2. Molekularbiologische Methoden

III.2.1. Isolierung und Aufreinigung von Plasmid-DNA

Die Isolierung und Reinigung von Plasmid-DNA erfolgte mit Hilfe verschiedener Kits der Firma *Qiagen* nach den Anwendungsvorschriften des Herstellers. Zur Gewinnung von Plasmid-DNA aus *E. coli* Übernachtkulturen wurde das *QIAgen Plasmid-Mini Kit* verwendet. Um DNA aus den Ansätzen von Restriktionsspaltungen sowie Dephosphorylierungen zu reinigen, wurde das *QIAquick PCR Purification Kit* angewendet. Die Extraktion von DNA aus Agarosegelstücken wurde mit Hilfe des *QIAquick Minelute Gel Extraction Kits* durchgeführt.

III.2.2. Spaltung von DNA mit Restriktionsendonukleasen

Die Restriktionsendonukleasen NDE I und HIND III wurden gemäß den Angaben des Herstellers unter Verwendung der entsprechend mitgelieferten Puffer eingesetzt. Für präparative Spaltungen wurden in einem Gesamtvolumen von 100 µl pro 1 µg Plasmid-DNA 1 bis 3 U Enzym eingesetzt und der Reaktionsansatz 1-6 h bei der vorgegebenen Temperatur inkubiert. Anschließend erfolgte die Dephosphorylierung der Fragmente, um deren Religation zu minimieren.

III.2.3. Dephosphorylierung von DNA

Für die Dephosphorylierung von DNA-Fragmenten wurde 1 U Alkalische Phosphatase aus Eismeergarnelen und das entsprechende Volumen des mitgelieferten, 10-fach konzentrierten Puffers direkt zum Ansatz pipettiert. Die Mischung wurde für 60 min bei 37 °C inkubiert und die Phosphatase anschließend für 15 min bei 65 °C inaktiviert. Der vollständige Ansatz wurde mit Hilfe einer präparativen Agarosegelelektrophorese aufgereinigt. Es schloss sich die Isolierung der DNA an (vgl. Abschnitt III.2.1, S.26).

III.2.4. Ligation von DNA-Fragmenten

DNA-Fragmente und linearisierte Vektoren, die mit den gleichen Restriktionsendonukleasen behandelt wurden, können über ihre komplementären kohäsiven Enden unter Knüpfung neuer Phosphodiesterbindungen ligiert werden. Diese Reaktion wurde mit Hilfe einer T4-Ligase in einem ATP-haltigen Puffer katalysiert (*10 x buffer for T4 DNA Ligase with 10 mM ATP*; *New England BioLabs (Frankfurt)*). Das jeweilige DNA-Fragment (Gene der Proteine M1…M8) lag im 10-fach molaren Überschuss zum verwendeten, dephosphorylierten Vektor (pET28a-Vektor) im Ansatz vor. Die Inkubation der Ligationsansätze erfolgte für 18 h bei 20 °C.

III.2.5. Agarose-Gelelektrophorese

Die elektrophoretische Trennung von DNA-Fragmenten erfolgte mittels horizontaler Gelelektrophorese bei konstanter Spannung von 8 mV/cm. Die Herstellung der Agarosegele erfolgte mit TAE-Puffer, der gleichzeitig als Laufpuffer verwendet wurde. Es wurden Gele mit 1-1.5 % Agarose verwendet. Um die DNA unter UV-Licht sichtbar zu machen, wurden die Gele 15 min in Ethidiumbromid-Lösung (1 µg/ml in 1x TAE-Puffer) gefärbt und anschließend mit dem GelDoc 2000 System analysiert.

III.2.6. Transformation von *E. coli* mit rekombinanter DNA

Für die Transformation von *E. coli*-Zellen wurde die Methode der Elektroporation angewendet [Dower *et al.*, 1988]. Die Amplifikation von Plasmid-DNA erfolgte in *E. coli*-TOP10. Für die Überexpression rekombinanter Gene wurden *E. coli*-Zellen des Stammes BL21(DE3) eingesetzt. Es wurden 50 µl der elektrokompetenten Zellen auf Eis mit 2 µl eines Ligationsansatzes gemischt und in eine vorgekühlte Elektroporationsküvette mit 0.2 cm Elektrodenabstand pipettiert. Der Stromstoß erfolgte bei einer Spannung von U = 2.5 kV, einem Widerstand von R = 200 Ω und einer eingestellten Kapazität von C = 25 µF. Anschließend wurden die Zellen in 37 °C warmem SOC-Medium resuspendiert und 45 min bei 37 °C geschüttelt. Die Selektion transformierter Zellen erfolgte durch Ausstreichen auf kanamycinhaltige LB-Agarplatten. Die Platten wurden über Nacht bei 37 °C inkubiert. Zur Kultivierung von *E. coli* für die Isolierung von Plasmid-DNA wurde LB-Medium mit einer Einzelkolonie von einer LB-Agarplatte angeimpft und über Nacht bei 37 °C geschüttelt. Das Medium enthielt das zur Selektion von plasmidhaltigen Zellen nötige Antibiotikum (35 µg/ml Kanamycin).

III.2.7. Kultivierung von *E. coli* und Expression von Fremdproteinen

III.2.7.1. Schüttelkultur

Für die Anzucht von *E. coli* im Schüttelkolben wurde zunächst eine Vorkultur hergestellt. Dazu wurden 20 ml LB-Medium mit einer Einzelkolonie beimpft und bei 37 °C und 120 *rpm* über Nacht geschüttelt. Anschließend wurde die Vorkultur im Verhältnis 1:100 in frischem LB-Medium verdünnt und bei 37 °C und 120 *rpm* bis zum Erreichen einer optischen Dichte von OD_{600} = 0.6 - 0.8 geschüttelt. Bei dieser optischen Dichte erfolgte die Expression des Proteins nach Induktion mit 1 mM IPTG über Nacht bei 37 °C unter ständigem Schütteln bei 120 *rpm*.

III.2.7.2. *Fed-Batch*-Fermentationen auf Vollmedium

Die Kultivierung von *E. coli* im Bioreaktor erfolgte im *Fed-Batch*-Verfahren auf Hefeextrakt-Vollmedium. Dazu wurden zunächst fünf Liter Medium für 30 Minuten bei einer Temperatur von 121 °C im Reaktor autoklaviert. Danach erfolgte die Zugabe der separat autoklavierten Zusätze (Glucose, Magnesiumsulfat, Kaliumdihydrogenphosphat). Das Medium wurde mit

500 ml einer Übernachtkultur inokuliert und bei einer Temperatur von 37 °C kultiviert. Der pH-Wert wurde über die Steuerungseinheit durch Zugabe von 25 %-iger Phosphorsäure bzw. 10 %-iger Natronlauge auf 7.0 reguliert. Der Sauerstoffgehalt wurde ebenfalls durch die Steuerungseinheit durch Variation der Rührgeschwindigkeit bzw. durch Zufuhr von Luft bzw. reinem Sauerstoff auf mindestens 20 % Sättigung eingestellt. Zur Unterdrückung der Schaumbildung im Medium wurde gegebenenfalls 50 %-ige (v/v) Polypropylenglycol-Lösung (Antischaum) eingeleitet. Nach Verbrauch der Glucose im Medium (Überprüfung durch Glucosesensorstäbchen) wurde das kontinuierliche *Feeding* durch Zugabe der *Feeding*-Lösung gestartet. Bei einer optischen Dichte von $OD_{600} = 40$ wurde die Wachstumstemperatur auf 30 °C gesenkt und nach 40 Minuten mit 1 mM IPTG induziert. Nach einer Induktionszeit von sechs Stunden wurde die Biomasse bei einer optischen Dichte von $OD_{600} = 50$ durch Zentrifugation geerntet (6 000 *rpm*, 20 min, 4 °C, Rotor JLA 8.1000, Zentrifuge Avanti). Die Biomasse wurde anschließend bei -80 °C bis zu ihrer weiteren Verwendung gelagert.

III.3. Proteinchemische Methoden

III.3.1. Zellernte, -aufschluss und Gewinnung des löslichen Proteinanteils

Die Ernte der Bakterienkulturen erfolgte durch Zentrifugation bei 6 000 *rpm* (20 min, 4 °C, Rotor: JLA 8.100, Zentrifuge Avanti). Das sedimentierte Bakterienpellet wurde in 4-fachem Überschuß (v/v) des Zellaufschlusspuffers resuspendiert und in drei Durchläufen mittels Hochdruckdispersion aufgeschlossen (800-1 000 bar, Micron LAB 40 Gaulin Hochdruck-Homogenisator). Anschließend wurden unlösliche Zellbestandteile durch Zentrifugation bei 20 000 *rpm* (30 min, 4 °C, Zentrifuge Avanti) abgetrennt. Der Überstand (Rohextrakt) mit den enthaltenen löslichen Proteinen wurde 20 Minuten bei 60-80 °C unter Rühren im Wasserbad erhitzt. Die präzipitierten Bestandteile der Lösung wurden 30 Minuten bei 4 °C und 20 000 *rpm* sedimentiert. Die weitere Reinigung der Proteine erfolgte mittels Immobilisierte Metallchelat-Affinitätschromatographie (IMAC).

III.3.2. Proteinreinigung durch Immobilisierte Metallchelat - Affinitätschromatographie

Die Fusion eines Histidin-*tags* an ein Protein ermöglicht dessen Reinigung durch Immobilisierte Metallchelat-Affinitätschromatographie (IMAC) an einer Nickel-NTA Matrix. Diese Methode diente zur weiteren Reinigung des Rohextraktes nach dessen Erhitzung und Zentrifugation. Es wurde ein Hexa-Histidin-*tag* verwendet. Die Ni-NTA Säule wurde mit fünf Säulenvolumina Zellaufschlusspuffer äquilibriert. Anschließend wurde der nach der Hitzepräzipitation erhaltene Rohextrakt auf die Ni-NTA Säule aufgetragen. Die Flußgeschwindigkeit betrug 1 ml/min. Die Säule wurde anschließend mit jeweils fünf Säulenvolumina Ni-NTA-Waschpuffer (25 mM Natriumphosphatpuffer pH 7.5, 400 mM NaCl), die je 20 mM und 50 mM Imidazol enthielten, gewaschen. Die Elution des Proteins erfolgte jeweils durch 2.5 Säulenvolumina Ni-NTA-Elutionspuffer mit je 100 mM und 500 mM Imidazol. Das Eluat wurde vereinigt,

aufkonzentriert und gegen den Puffer für die Abspaltung des His_6-*tags* (Thrombinverdau) dialysiert. Es schloss sich die Abspaltung des His_6-*tags* an.

III.3.3. Aufkonzentrierung von Proteinen und Dialyse

Die Konzentrierung der Proteinlösungen erfolgte durch Ultrafiltration. Hierzu wurden Ultrafiltrationseinheiten der Firma *Millipore* mit einer Molekularausschlußgrenze (MWCO) von 5 000 Da verwendet. Die Zentrifugation erfolgte bei 2 000 *rpm* und 4 °C (Rotor: JLA 20.10, Zentrifuge Avanti).

III.3.4. Spaltung des Hexa-Histidin-*tags*

Die Spaltung des Hexa-Histidin-*tags* (Thrombinverdau) erfolgte mit 30 U Thrombin pro Milligramm Protein über Nacht bei 4 °C in 10 mM Tris pH 7.5 (HCl), 2mM $MgCl_2$ und 150 mM NaCl. Anschließend wurden die Ansätze auf eine Ni-NTA-Säule aufgetragen, um unverdautes von gespaltenem Protein abzutrennen. Es folgte die Endreinigung durch Gelfiltration (Größenausschlusschromatographie).

III.3.5. Proteinreinigung durch Gelfiltration

Die Endreinigung von M7 erfolgte durch Gelfiltration an einer Superdex 75-Säule (V_t = 120 ml) bei einer Flußgeschwindigkeit von 1 ml/min. Der Laufpuffer enthielt 25 mM Natriumphosphat pH 8.0 und 400 mM NaCl. Die Säule wurde mit zwei Säulenvolumina Laufpuffer äquilibriert.

III.3.6. SDS-Polyacrylamid-Gelelektrophorese (PAGE)

Die SDS-PAGE erlaubt die elektrophoretische Auftrennung von Proteinen in denaturierter Form in einem SDS-Polyacrylamidgel anhand des Molekulargewichtes [Laemmli, 1970]. Durch das im Gel und Probenauftragspuffer enthaltene SDS werden die Proteine denaturiert und erhalten eine negative Nettoladung. Man unterscheidet dabei zwischen nicht-reduzierender und reduzierender SDS-PAGE, wobei bei letzterer Methode ein Reduktionsmittel im Probenpuffer vorhanden ist, um Disulfidbrücken vollständig zu reduzieren. In dieser Arbeit wurden Vertikalgele (8 cm x 10 cm x 0,75 mm) eingesetzt, die aus einem 15%-igen Trenngel, überschichtet mit einem 6%-igen Sammelgel, bestanden. Die Prozentangabe bezieht sich dabei auf den Anteil an Polyacrylamid. Als Referenz wurde der LMW-Proteinmolekulargewichtsstandard mitgeführt. Die Elektrophorese erfolgte unter Verwendung des SDS-Laufpuffers bei 35 mA pro Gel für 45 bis 70 min. Anschließend wurden die Gele mit PAGE-Fixierer (15 min), PAGE-Färber (2 h) und PAGE-Entfärber (2 h) behandelt, um die Banden der aufgetrennten Proteine zu visualisieren.

III.4. Biophysikalische Methoden & Strukturaufklärung

III.4.1. UV/VIS-Spektroskopie

Die Konzentrationsbestimmung der gereinigten Proteine in Lösung wurde mittels UV-Absorptionsspektroskopie an einem UV/VIS-Spektrophotometer *Ultrospec 4000* durchgeführt. Die Proteine M1, M2 und M7 enthalten in ihrer Sequenz je eine Tyrosin (vgl. Abbildung IV-18 auf Seite 63). M8 enthält ein Tryptophan. Die Konzentration dieser Proteine lässt sich deshalb mit Hilfe des molaren Absorptionskoeffizienten für Tyrosin bzw. Tryptophan gemäß dem Lambert-Beer'schen Gesetz (III-1) bestimmen.

III-1 $$c = \frac{A}{\varepsilon \cdot d}$$

mit

A	gemessene Absorption
ε	molarer Absorptionskoeffizient ($M^{-1}cm^{-1}$)
c	molare Konzentration der Probe (M)
d	Schichtdicke der Küvette (cm)

III.4.2. CD-Spektroskopie

Die Messung der CD-Spektren erfolgte mit einem Jasco J-810 CD-Spektrometer in Quarzglasküvetten mit einer Schichtdicke von 0.1 cm. Die Proteine waren in 25 mM Natriumphosphatpuffer pH 8.0 und 40 mM NaCl gelöst. Die Spektren wurden mit einer Bandbreite von 0.2 nm und einer Zeitkonstante von 2 s 10-fach akkumuliert. Die gemessene Elliptizität Θ wurde nach Gleichung III-2 in die mittlere residuale Elliptizität $[\Theta]_{MRW}$ umgerechnet. Alle Spektren wurden pufferkorrigiert.

III-2 $$[\Theta]_{MRW} = \frac{100 \cdot M_G \cdot \Theta}{d \cdot c \cdot n_{AS}}$$

mit

$[\Theta]_{MRW}$	Mittlere residuale Elliptizität (deg cm²/dmol)
M_G	Molekulargewicht (Dalton oder g/mol)
Θ	Elliptizität (deg)
n_{AS}	Anzahl Aminosäuren in der Proteinsequenz
d	Küvettenschichtdicke (cm)
c	Proteinkonzentration (mg/ml)

III.4.3. Analyse chemisch induzierter Entfaltungsübergänge

Die Analyse der chemisch induzierten Entfaltungsübergänge erfolgte unter Annahme eines Zweizustandsmodelles, bei dem natives (*n*) und denaturiertes (*d*) Protein im Gleichgewicht vorliegen (keine Intermediate). Die Messung der Elliptizität des Proteins wurde in einem Jasco-810 CD-Spektrometer bei einer Wellenlänge von $\lambda = 220$ nm durchgeführt. Die

numerische Anpassung an die gemessenen Rohdaten erfolgte durch eine nichtlineare Regression der Gleichung III-3 mit Hilfe des Programms SigmaPlot V.8.02. Die Konzentration des Denaturierungsmittels am Übergangsmittelpunkt $D_{1/2}$ wurde mit Gleichung III-4 bestimmt.

III-3
$$Y = \frac{(m_n[D] + b_n) + (m_d[D] + b_d)e^{-\frac{\Delta G_{n\to d}^{H_2O} + m[D]}{RT}}}{1 + e^{-\frac{\Delta G_{n\to d}^{H_2O} + m[D]}{RT}}}$$

III-4
$$D_{1/2} = \frac{\Delta G_{n\to d}^{H_2O}}{m_{n\to d}}$$

mit

Y	angepasster Wert der Messgröße
$[D]$	Denaturierungsmittelkonzentration
m_n	lineare Abhängigkeit der Messgröße des nativen Proteins von der Denaturierungsmittelkonzentration
b_n	spezifischer Wert der Messgröße des nativen Proteins bei 0 M Denaturierungsmittel
m_d	lineare Abhängigkeit der Messgröße des denaturierten Proteins von der Denaturierungsmittelkonzentration
b_d	spezifischer Wert der Messgröße des denaturierten Proteins bei 0 M Denaturierungsmittel
$\Delta G_{n\to d}^{H_2O}$	Freie Entfaltungsenthalpie bei 0 M Denaturierungsmittelkonzentration
R	Gaskonstante (J mol^{-1} K^{-1})
T	Referenztemperatur (T)
$D_{1/2}$	Denaturierungsmittelkonzentration am Übergangsmittelpunkt
$m_{n\to d}$	lineare Abhängigkeit der Freien Entfaltungsenthalpie von der Denaturierungsmittelkonzentration

Für die Bestimmung der Freien Enthalpie der Proteindenaturierung wurden die Proben des jeweiligen Proteins im Verhältnis 1:10 in verschiedenen Konzentrationen von GdmCl verdünnt. Dazu wurden Präzisionsspritzen vom Typ Hamilton 1725 RNR verwendet. Das Gesamtvolumen der einzelnen Proben betrug jeweils 300 µl. Die Messung der Proben erfolgte bei einer Temperatur von T = 293 K (20 °C) nach 18 Stunden Inkubation bei 20 °C. Die Konzentration an Denaturierungsmittel GdmCl wurde refraktometrisch nach Gleichung III-5 bestimmt [Pace, 1986].

III-5
$$[\text{GdmCl}] = 57.147(\Delta N) + 38.68(\Delta N)^2 - 91.6(\Delta N)^3$$

mit

[GdmCl]	Konzentration an GdmCl (mol/l)
ΔN	Differenz der Brechungsindices zwischen Proteinlösung mit und ohne Denaturierungsmittel

III.4.4. Analyse thermisch induzierter Entfaltungsübergänge

Durch Messung thermisch induzierter Entfaltungsübergänge bei verschiedenen GdmCl-Konzentrationen lässt sich die Änderung der Enthalpie $\Delta H_{n\to d}^{H_2O}$, die Änderung der Entropie $\Delta S_{n\to d}^{H_2O}$ und die Änderung der Wärmekapazität $\Delta Cp_{n\to d}^{H_2O}$ durch Denaturierung des nativen Proteins bestimmen. Die Temperaturabhängigkeit der Freien Entfaltungsenthalpie $\Delta G(T)_{n\to d}^{H_2O}$ wird durch die Gibbs-Helmholtz-Gleichung (III-6) beschrieben. Darin bezeichnet *n* das native Protein und *d* die denaturierte Spezies. Die Referenztemperatur wurde auf $T_0 = 293$ K (20 °C) festgelegt.

III-6
$$\Delta G(T)_{n\to d} = \Delta H_{n\to d} - T\Delta S_{n\to d} + \Delta Cp_{n\to d}\left(T - T_0 - T\ln\frac{T}{T_0}\right)$$

mit

T	Messtemperatur
T_0	Referenztemperatur

Die numerische Anpassung an die gemessenen Rohdaten erfolgte nach Gleichung III-7 durch nichtlineare Regression mit einem globalen *fit* (parallele Anpassung aller gemessenen Übergänge). Dazu wurde das Programm SigmaPlot V.8.02 verwendet. In Anlehnung an Myers *et al.* wurde eine lineare Abhängigkeit der Änderung der Enhalpie (III-8), der Änderung der Entropie (III-9) und der Änderung der Wärmekapazität (III-10) von der Denaturierungsmittelkonzentration angenommen [Myers *et al.*, 1995].

III-7
$$Y = \frac{(m_n T + b_n) + (m_d T + b_d)e^{-\frac{\Delta G(T)_{n-d}^{H_2O}}{RT}}}{1 + e^{-\frac{\Delta G(T)_{n-d}^{H_2O}}{RT}}}$$

mit

Y	angepasster Wert der Messgröße
T	Messtemperatur (K)
m_n	lineare Abhängigkeit der Messgröße des nativen Proteins von der Temperatur
b_n	spezifischer Wert der Messgröße des nativen Proteins bei 0 K
m_d	lineare Abhängigkeit der Messgröße des denaturierten Proteins von der Temperatur
b_d	spezifischer Wert der Messgröße des denaturierten Proteins bei 0 K
R	Gaskonstante (J mol^{-1} K^{-1})

III-8
$$\Delta H_{n\to d} = \Delta H_{n\to d}^{H_2O} + m_H[D]$$

III-9
$$\Delta S_{n\to d} = \Delta S_{n\to d}^{H_2O} + m_S[D]$$

III-10
$$\Delta Cp_{n\to d} = \Delta Cp_{n\to d}^{H_2O} + m_C[D]$$

mit

$\Delta H_{n\to d}^{H_2O}$	Änderung der Enthalpie in 0 M Denaturierungsmittel
$\Delta S_{n\to d}^{H_2O}$	Änderung der Entropie in 0 M Denaturierungsmittel
$\Delta Cp_{n\to d}^{H_2O}$	Änderung der Wärmekapazität in 0 M Denaturierungsmittel
$[D]$	Denaturierungsmittelkonzentration
m_H	lineare Abhängigkeit der Änderung der Enthalpie von der Denaturierungsmittelkonzentration
m_S	lineare Abhängigkeit der Änderung der Entropie von der Denaturierungsmittelkonzentration
m_C	lineare Abhängigkeit der Änderung der Wärmekapazität von der Denaturierungsmittelkonzentration

Die Proben des jeweiligen Proteins wurden im Verhältnis 1:10 bei verschiedenen Konzentrationen von GdmCl verdünnt. Dazu wurden Präzisionsspritzen vom Typ Hamilton 1725 RNR verwendet. Das Gesamtvolumen der einzelnen Proben betrug jeweils 300 µl. Nach 18 Stunden Inkubation bei 20 °C wurden die Proben bei einer Heizrate von 1 °C/min in einer 0.1 cm Quarzglasküvette vermessen. Die Konzentration an Denaturierungsmittel GdmCl wurde refraktometrisch nach Gleichung III-5 bestimmt.

III.4.5. Analytische Ultrazentrifugation

Die analytische Ultrazentrifugation wurde verwendet, um den Assoziationsgrad der betreffenden Proteine zu bestimmen. Die Analysen wurden freundlicherweise von PD Dr. Hauke Lilie vom Institut für Biotechnologie der Martin-Luther-Universität Halle-Wittenberg durchgeführt. Die Messungen erfolgten in einer Optima XL-A Zentrifuge mit einem An50Ti-Rotor. Die Proben wurden in Doppelsektorzellen bei einer Detektionswellenlänge von 230 nm vermessen. Die Konzentration des Proteins betrug 180 µg/ml. Die Probenvorbereitung erfolgte durch Dialyse des Proteins über Nacht gegen 20 mM Natriumphosphat pH 8.0 und 400 mM NaCl.

III.4.6. Massenspektrometrie

Die Molekulargewichtsbestimmung mittels Massenspektrometrie wurde freundlicherweise von Frau Dr. A. Schierhorn (Forschungsstelle „Enzymologie der Proteinfaltung“ der Max-Planck-Gesellschaft, Halle/Saale) durchgeführt. Die Proben wurden mit C_{18}-*ZipTips*® nach Angaben des Herstellers aufgearbeitet. Es wurden Spektren von entsalzten Proben durch MALDI-TOF-Massenspektrometrie an einem Esquire-LC-Ionenfallen-Massenspektrometer aufgenommen (Bruker-Franzen Analytik, Bremen).

III.4.7. NMR-spektroskopische Untersuchungen an M7

Die NMR-spektroskopischen Untersuchungen und die Auswertung der Ergebnisse wurden freundlicherweise von Dr. Christian Lücke (Forschungsstelle „Enzymologie der Proteinfaltung“ der Max-Planck-Gesellschaft, Halle/Saale) durchgeführt. Die NMR-Proben enthielten 1.8 mM des Proteins M7 in 15 mM Natriumphosphatpuffer pH 6.5, 250 mM NaCl. Die Messungen erfolgten an einem DRX 500 Spektrometer mit einer ^{1}H-Resonanzfrequenz von

500.13 MHz. Das Spektrometer war mit einem inversen 5 mm Tripleresonanz-Probenkopf versehen, der XYZ Gradientenspulen enthielt. Es wurden Standard 1D- und 2D-NMR Spektren bei 30 °C phasensensitiv mit der TPPI (*time-proportional phase incrementation*) Technik für Quadraturdetektion aufgenommen. Die ^{1}H-chemischen Verschiebungswerte wurden in Bezug auf externes Natrium- 2,2-Dimethyl-2-silapentan-5-sulfonat (DSS) (Cambridge Isotope Laboratories) referenziert. Das Wassersignal wurde durch selektive Sättigung während der Präparationsphase (1.3 s) unterdrückt, wobei die Trägerfrequenz auf die Wasserresonanz plaziert wurde. Die homonuklearen 2D-COSY, 2D-TOCSY (Spinlockzeiten von 6 bzw. 80 ms) und NOESY (Mischzeit von 150 ms) Experimente wurden mit 32 Durchgängen und 6009.6 Hz (12 ppm) spektrale Weite in beiden Zeitdomänen durchgeführt. Dabei wurden 2048 × 512 Datenpunkte (2048 × 1024 beim COSY) gesammelt. Die 2D-Spektren wurden mit 2048 realen Datenpunkten in jeder Dimension prozessiert. Alle Spektren wurden unter Verwendung des XWINNMR 2.6 Softwarepaketes (Bruker) aufgenommen, prozessiert und analysiert.

IV. Ergebnisse

Die vorliegende Arbeit untersucht die Fragestellung, inwieweit sich tetrapeptidbasierte Fragmentbibliotheken zum Proteindesign verwenden lassen. Dafür mussten drei Problemstellungen gelöst werden:

i. Bereitstellung einer geeigneten Datenbasis zur Konformationsanalyse von Tetrapeptiden und Auswertung der statistischen Daten (Abschnitt IV.1, S. 35).
ii. Entwicklung eines Verfahrens zum tetrapeptidbasierten Proteindesign und Berechnung von alternativen Aminosäuresequenzen für die Struktur von Top7 mit einer Sequenzidentität von kleiner als 30 % zu dessen Sequenz (Abschnitt IV.2, S. 53).
iii. Experimentelle Überprüfung der theoretischen Ergebnisse (Abschnitt IV.3, S. 79)

Soweit nicht anderes angegeben, wurden alle nötigen Programme und Algorithmen selbst entwickelt und/oder implementiert.

IV.1. Datenaufbereitung und Datenanalyse

Die Bereitstellung der Datensätze für eine Berechnung der Wahrscheinlichkeitsdichtefunktionen gliedert sich in fünf Punkte, die im Folgenden erläutert werden.

i. Identifizierung der mit Hilfe von Röntgenkristallographie bestimmten Proteinstrukturen der *PDB* (Abschnitt IV.1.1)
ii. Ermittlung von Auflösung und R-Faktor der einzelnen Strukturen (Abschnitt IV.1.1)
iii. Separation der *PDB*-Dateien in einzelne Proteinketten (Abschnitt IV.1.1)
iv. Berechnung des ψ-Winkels der zweiten Aminosäure (ψ_2) und des φ-Winkels der dritten Aminosäure (φ_3) jedes einzelnen Tetrapeptids einer Proteinkette (Abschnitt IV.1.2, S. 36)
v. Beseitigung redundanter Informationen mit Hilfe von *all-against-all* Alignments (Abschnitt IV.1.3, S. 37)
vi. Berechnung der Wahrscheinlichkeitsdichtefunktionen (Abschnitt II.3, S. 14)

IV.1.1. Analyse der Proteinstrukturen der *PDB*

Die zur Konformationsanalyse verwendeten Proteinstrukturen wurden der Proteindatenbank, *PDB* (ftp.rcsb.org), vom 08.12.2003 entnommen [Berman *et al.*, 2000a; Berman *et al.*, 2002]. An die verwendeten Proteinstrukturen wurden definierte Qualitätsanforderungen gestellt. Es fanden ausschließlich durch Röntgenkristallographie aufgeklärte Strukturen Verwendung, die eine Kettenlänge von mindestens 30 Aminosäuren aufwiesen. Kleinere Peptide wurden aus der Konformationsbetrachtung ausgeschlossen. Weiterhin wurden nur diejenigen Strukturen dem Ausgangsdatensatz hinzugefügt, deren Auflösung besser als 3 Å war und deren R-Faktor einen

```
1H7HB        245      2.3      0.15      SKAVIVIPARYGSSRLPGKPLLDIVG…
1HLB_        158      2.5      0.15      XGGTLAIQAQGDLTLAQKKIVRKTWH…
1HVSA         99      2.25     0.15      PQITLWQRPLVTIKIGGQLKEALLDT…
1HVSB         99      2.25     0.15      PQITLWQRPLVTIKIGGQLKEALLDT…
1I4UA        181      1.15     0.15      DKIPDFVVPGKCASVDRNKLWAEQTP…
1I4UB        181      1.15     0.15      DKIPDFVVPGKCASVDRNKLWAEQTP…
1IKGA        349      1.9      0.15      ADLPAPDDTGLQAVLHTALSQGAPGA…
1JCDA         52      1.3      0.15      SSNAKADQASSDAQTANAKADQASND…
```

Abbildung IV-1 Datenformat der Datei zur Rekonstruktion der Aminosäuresequenzen aus den *PDB*-Codes. Die Daten wurden aus *PDB*-Dateien ausgelesen. In der ersten Spalte steht der *PDB*-Code mit der Kettenbezeichnung. War in den *PDB*-Dateien keine Kettenbezeichnung definiert, so wurde die Kette mit „_“ bezeichnet. Die zweite Spalte gibt die Länge der Proteinsequenz gemäß der mit SEQRES in den *PDB*-Dateien gekennzeichneten Reste an. Die dritte Spalte gibt die Auflösung der Kristallstruktur in Ångström an. In der vierten Spalte ist der R-Faktor verzeichnet. War in einer *PDB*-Datei kein R-Faktor verzeichnet so wurde er auf 0.250001 gesetzt. In der letzten Spalte steht die Aminosäuresequenz entprechend der mit SEQRES bezeichneten Reste. Nichtstandardaminosäuren wurden mit einem „*X*“ bezeichnet.

Wert von kleiner oder gleich 0.25 aufwies. War der R-Faktor in den *PDB*-Dateien nicht verzeichnet, so wurde er *per definitionem* auf 0.250001 gesetzt. In vielen Fällen umfasst eine *PDB*-Datei Strukturen von mehreren Proteinketten. Vor Beginn der Konformationsanalyse wurden deshalb die Koordinaten jeder Proteinkette in einer separaten Datei gespeichert, deren Dateiname den vier Zeichen langen *PDB*-Code des Proteins und als fünftes Zeichen die Kettenbezeichnung enthielt, die in der *PDB*-Datei definiert ist. Die verwendete Datenbasis beinhaltete 35 397 Strukturen und diente als Ausgangsdatensatz zur Konformationsbetrachtung der Tetrapeptide.

IV.1.2. Bereitstellung der Ausgangsdatensätze für die Berechnung der Dichtefunktionen

In den heute bekannten Proteinstrukturen finden sich oft längere Bereiche, deren Struktur aus experimentellen Gründen nicht aufgelöst ist (*gaps*). Würden jedoch die Diederwinkel zwischen zwei randständigen Aminosäuren eines *gaps* an den Positionen n und $n + m$ (mit $m > 1$) errechnet werden, dann würde dies offensichtlich zu falschen Ergebnissen führen. Daher war es notwendig, solche *gaps* in Proteinstrukturen sicher zu erkennen. Dazu wurde die maximale Länge einer Peptidbindung mit 1.4 Å als Grenzwert für den Abstand des Carboxy-C-Atoms einer Aminosäure n und des Stickstoffatoms der Aminosäure $n + 1$ festgelegt. Wenn dieser Abstand größer als dieser Grenzwert war, so wurde zwischen diesen Aminosäuren ein *gap* erkannt und demzufolge keine Diederwinkel errechnet. Die Proteinstrukturen wurden nach benachbarten Tetrapeptiden durchsucht, bei deren mittleren Aminosäuren die Proteinrückgratatome N, C_α und C aufgelöst sein mussten. Der Grund hierfür war, dass die Errechnung der Diederwinkel ψ der zweiten Aminosäure (ψ_2) und φ der dritten Aminosäure (φ_3) eine Positionsangabe dieser Atome voraussetzt. Es wurden ebenfalls nur Atompositionen berücksichtigt, deren *occupancy*-Wert ein Betrag von 1.00 zugewiesen wurde. In den Atomdaten nicht verzeichnete Proteinrückgratatome wurden nicht hinzumodelliert, so dass bei Fehlen eines oder mehrerer Atome keine Diederwinkel errechnet und das entsprechende Tetrapeptid somit nicht berücksichtigt wurde. Die Proteinrückgratatome der ersten und der letzten Aminosäure eines Tetrapeptids mussten nicht vollständig aufgelöst sein, da die Diederwinkel zwischen der ersten und der zweiten Aminosäure bzw. zwischen der dritten und der vierten Aminosäure für die spätere Auswertung nicht benötigt

AAAA	36	-43.81	-61.88	2BAA_
AAAA	37	-33.59	-78.57	2BAA_
AAAA	97	-49.9	-55.36	2CCYA
AAAA	221	165.66	-63.52	2POR_
AAAA	120	-3.06	-65.08	1G0UD
AAAA	121	77.6	-81.51	1G0UD
AAAA	270	-53.93	-53.72	1JQLA
AAAA	135	149.23	-125.11	1NEKA
AAAA	147	-44.87	-62.19	1L3LA
AAAA	495	-85.74	-33.81	1QU7A

Abbildung IV-2 Datenformat der Ausgangsdatensätze zur Berechnung der Wahrscheinlichkeitsdichtefunktionen. Gezeigt ist ein Ausschnitt aus der Datei, welche die Strukturinformationen des Tetrapeptids AAAA enthält. Diese Informationen wurden aus *PDB*-Dateien ausgelesen. In der ersten Spalte ist das Tetrapeptid angegeben. Die zweite Spalte zeigt die Position der ersten Aminosäure des Tetrapeptids in der Struktur nach *PDB*-Numerierung. Die dritte und die vierte Spalte zeigen den ψ-Winkel der zweiten Aminosäure (ψ_2) und den φ-Winkel der dritten Aminosäure (φ_3). Die letzte Spalte gibt den *PDB*-Code des Proteins an, in dem das Tetrapeptid gefunden wurde. Das letzte Zeichen im *PDB*-Code bezeichnet die Aminosäurekette aus der *PDB*-Datei. Wenn keine Kettenbezeichnung in einer *PDB*-Datei angegeben war, wurde der Kettenname als „_" definiert**.**

werden. Jeder gefundene Tetrapeptidtyp (z. B. „AAAA") wurde mit dem Index, der seine Position innerhalb der Proteinstruktur bestimmt, den zugehörigen Diederwinkeln ψ_2 und φ_3 und dem *PDB*-Code des Proteins, aus dem dieses Tetrapeptid stammt, in einer separaten Datei gespeichert, die als Ausgangsdatensatz zur weiteren Datenaufbereitung diente. Das Datenformat zeigt die Abbildung IV-2. Bei der Berechnung der Diederwinkel wurden Peptidbindungen mit *cis*-Konformation nicht gesondert behandelt. Ebenfalls wurden Tetrapeptide, die das N-terminale Methionin der Aminosäuresequenz enthalten, bei der Konformationsbetrachtung nicht berücksichtigt.

IV.1.3. Beseitigung redundanter Proteinsequenzen in den Datensätzen

Die Aminosäuresequenzen der Proteinstrukturen in der *PDB* haben teilweise eine hohe gegenseitige Sequenzidentität. Eine Verwendung dieser Daten im Rahmen einer statistischen Analyse würde deshalb zu einem *Bias* auf Strukturen führen, deren Sequenzen häufiger vorkommen, was in der Folge mögliche Sequenz-Struktur-Korrelationen verfälschen würde. Vor der Verwendung dieser Datenbasis mussten daher die redundanten Informationen beseitigt werden. Dies erfolgte durch sequenzbasierte *all-against-all* Alignments.

IV.1.3.1. Durchführung sequenzbasierter *all-against-all* Alignments

Die Durchführung der *all-against-all* Alignments erfolgte in Anlehnung an Abschnitt II.2, S. 13. Bei diesem Verfahren ist jedoch ersichtlich, dass eine große Anzahl an unnötigen Alignments berechnet werden würde, wenn Sequenzen mit einer hohen gegenseitigen Sequenzidentität gegen eine *template*-Sequenz alignt werden, zu der diese eine Sequenzidentität besitzen, die kleiner ist als der zu Beginn definierte *cut off*. Vor der Durchführung eines *all-against-all* Alignments ist es daher günstiger, die Proteinsequenzen in Gruppen zu teilen, deren Sequenzen eine hohe gegen-

seitige Sequenzidentität besitzen und jeweils innerhalb dieser Gruppen[3] ein *all-against-all* Alignment durchzuführen (primäres *all-against-all* Alignment). Anschließend werden die Sequenzen aus den Gruppen in einer Gesamtliste gespeichert, innerhalb derer das finale *all-against-all* Alignment durchgeführt wird. Dieser Vorgang führt zu einer sehr starken Reduktion der Anzahl an vorhandenen Sequenzen in den einzelnen Gruppen und im Ergebnis zu einer erheblichen Zeitersparnis[4] im finalen *all-against-all* Alignment. Eine durchgeführte Analyse der Proteinsequenzen der verwendeten *PDB*-Version zeigte, dass in dieser Datenbasis Sequenzen mit ähnlicher Länge oft auch hohe gegenseitige Sequenzidentitäten aufweisen. Aus diesem Grund wurden die sortierten Sequenzen in Gruppen geteilt, die eine maximale Längendifferenz von 25 Aminosäuren aufwiesen. Innerhalb dieser Gruppen erfolgte das primäre *all-against-all* Alignment. Als *template*-Sequenz wurde immer die Proteinsequenz der qualitativ hochwertigsten Struktur gewählt, die gemäß Auflösung (Å) + R-Faktor [Hobohm & Sander, 1994] bestimmt wurde.

IV.1.3.2. Ergebnisse der Datenaufbereitung

In den Dateien mit den Strukturinformationen der einzelnen Tetrapeptide sind die *PDB*-Codes mit den Kettenbezeichnungen verzeichnet (siehe Abbildung IV-2 auf Seite 37). Mit Hilfe der *PDB*-Codes lässt sich nach Abbildung IV-1 auf Seite 36 die Proteinsequenz rekonstruieren, aus der ein bestimmtes Tetrapeptid stammt. Zur Beseitigung sequenzbasierter redundanter Informationen wurden *all-against-all* Alignments nach dem in IV.1.3.1 beschriebenen Schema innerhalb der Gruppen (Dateien) von Proteinsequenzen durchgeführt, die ein bestimmtes Tetrapeptid enthalten. Ein sequenzbasiertes Alignment berücksichtigt keine strukturellen Eigenschaften von Proteinen. Sequenzidentische Tetrapeptide können jedoch große strukturelle Unterschiede zeigen, auch wenn die Gesamtproteinsequenzen Identitäten von größer als 25 % aufweisen. Es wäre in diesem Fall nicht begründbar, die Strukturinformationen des selektierten Tetrapeptids aus der *template*-Sequenz gegenüber demjenigen aus der *query*-Sequenz zu bevorzugen und die Konformationsdaten des *query*-Tetrapeptids zu löschen. Um die individuellen Konformationen gleicher Tetrapeptide bei höheren Sequenzidentitäten der Gesamtsequenzen zu berücksichtigen, wurde ein *cut off* von |25°| für die erlaubte Winkeldifferenz zwischen dem ψ_2-Winkel bzw. φ_3-Winkel des Tetrapeptids aus der *template*-Struktur und der *query*-Struktur festgelegt. Besitzen also die *template*-Sequenz und die *query*-Sequenz eine gegenseitige Sequenzidentität von größer als 25 % und errechnet sich eine Winkeldifferenz von $|\Delta\psi_2| > 25°$ oder $|\Delta\varphi_3| > 25°$ zwischen dem selektierten Tetrapeptid aus der *template*-Struktur und der *query*-Struktur, so verbleibt die *query*-Sequenz vorläufig in der Liste mit den *query*-Sequenzen und kann am nächsten Zyklus des *all-against-all* Alignments teilnehmen.

In den in der *PDB* aufgeführten Proteinstrukturen, die die unter IV.1.1 genannten Qualitätskriterien erfüllten, wurden von den theoretisch möglichen 160 000 Tetrapeptiden 147 050 Tetrapeptide gefunden. Ein Vergleich mit der Sequenzdatenbank *Swissprot* [Boeckmann *et al.*, 2003] vom Dezember 2003 zeigte, dass alle theoretisch möglichen Tetrapeptide auch als Fragmente in Proteinen existieren. Da nicht alle Aminosäuren mit der gleichen Häufigkeit

[3] vergleichbar mit den *divide-and-conquer* Verfahren
[4] Der Needleman-Wunsch-Algorithmus hat eine (Zeit-) Komplexität von $O(n^2)$.

vorkommen, gilt dies auch für die Häufigkeit einzelner Tetrapeptide. Die statistische Analyse wurde in der vorliegenden Arbeit nur mit Datensätzen durchgeführt, in denen mindestens vier Wertepaare (ψ_2,φ_3) verzeichnet waren. Dies entspricht einer Anzahl von 104 687 Datensätzen von Tetrapeptiden. Zum Zeitpunkt der durchgeführten Analyse waren in der *PDB* Atomkoordinaten aus Strukturen von 35 397 Proteinketten gespeichert, welche die unter IV.1.1 aufgeführten Qualitätskriterien erfüllten. Führt man diese Sequenzen einem wie im Abschnitt IV.1.3.1, S. 37, beschriebenen *all-against-all* Alignment mit einer *open penalty* von -5 und einer *extension penalty* von -2 zu, so erhält man einen nichtredundanten Datensatz von 3 486 Proteinsequenzen mit einer maximalen gegenseitigen Sequenzidentität von 25 % (die *PDBSELECT*-Datenbank wies am 15. Dezember 2003 3 639 Sequenzen mit einer Sequenzidentität von kleiner als 30 % auf, vgl. http://swift.cmbi.kun.nl/whatif/select/). Innerhalb dieses Datensatzes wurden 138 680 verschiedene Tetrapeptide gefunden. Im Rahmen einer statistischen Analyse von Tetrapeptidkonformationen stünden in diesem Fall 78 786 Datensätze mit mindestens vier Wertepaaren

Tabelle IV-1 Übersicht über die Ergebnisse der Datenaufbereitung. ψ_2 bezeichnet den ψ-Winkel der zweiten Aminosäure und φ_3 den φ-Winkel der dritten Aminosäure eines Tetrapeptids. Es wurden nur Sequenzen aus Proteinstrukturen berücksichtigt, die aus mindestens 30 Aminosäuren bestanden, deren Auflösung besser als 3 Å war und deren R-Faktor höchstens 0.25 betrugt. Bei unbekanntem R-Faktor wurde diesem ein Wert von 0.25001 zugewiesen.

[a] Dieser Datensatz wurde aus der Gesamtheit aller berücksichtigten Proteinstrukturen gemäß Abschnitt IV.1.3.1, S. 37, errechnet. Die Begriff *Nichtredundanz* bezieht sich auf zwei Proteinsequenzen mit einer gegenseitigen Sequenzidentität von höchstens 25 % nach Anwendung eines semiglobalen Alignments gemäß dem Algorithmus von Needleman-Wunsch [Needleman & Wunsch, 1970] mit affinen *gap*-Strafen [Gotoh, 1982] unter Verwendung der BLOSUM62-Substitutionsmatrix [Henikoff & Henikoff, 1992]. Als *open penalty* wurde ein Wert von -5 und als *extension penalty* ein Wert von -2 definiert.

[b] Entspricht dem Ausgangsdatensatz zur Konformationsbetrachtung von Tetrapeptiden (alle berücksichtigten Strukturen).

[c] Bei einer Sequenzidentität von größer als 25 % zwischen der *template*-Sequenz und der *query*-Sequenz wurde bei einer Winkeldifferenz von $|\Delta\psi_2| > 25°$ oder $|\Delta\varphi_3| > 25°$ zwischen dem selektierten Tetrapeptid aus der *template*-Struktur und der *query*-Struktur, die *query*-Sequenz aus der Liste mit den *query*-Sequenzen entweder entfernt (vorletzte Spalte) oder nicht entfernt (letzte Spalte).

[d] Der Median bezieht sich auf Datensätze mit einer Mindestanzahl von vier Diederwinkelpaaren (ψ_2,φ_3).

	Nichtredundanter Datensatz[a]	Gesamtheit der berücksichtigten Proteinstrukturen[b]	Nichtredundante Datensätze <*Tetrapeptidtyp*> **ohne** Berücksichtigung von $\|\Delta\psi_2\|$ und $\|\Delta\varphi_3\|$[c]	Nichtredundante Datensätze <*Tetrapeptidtyp*> **mit** Berücksichtigung von $\|\Delta\psi_2\|$ und $\|\Delta\varphi_3\|$[c]
Anzahl berücksichtigter Proteinstrukturen	3 486	35 397	30 803	33 048
Anzahl gefundener Tetrapeptide eines Typs (z. B. „AAAA“)	138 680	147 050	147 050	147 050
Gesamtzahl an Diederwinkelpaaren (ψ_2,φ_3)	861 154	8 617 240	1 280 000	1 486 313
Gesamtzahl an Diederwinkelpaaren (ψ_2,φ_3) in allen Datensätzen mit mindestens vier Diederwinkelpaaren	748 243	8 594 905	1 187 570	1 403 417
Gesamtzahl an Datensätzen mit mindestens vier Diederwinkelpaaren (ψ_2,φ_3)	78 786	138 094	99 901	104 687
Median[d]	7	40	9	10

(ψ_2,φ_3) zur Verfügung. Demgegenüber resultierte die Durchführung von *all-against-all* Alignments zwischen Proteinsequenzen, die ein bestimmtes Tetrapeptid enthalten zu 99 901 Datensätzen, die aus mindestens vier Diederwinkelpaaren (ψ_2,φ_3) bestehen. In diesen Datensätzen sind Strukturinformationen aus 30 803 der 35 397 Proteinstrukturen des Ausgangsdatensatzes enthalten. Die Proteinsequenzen der 30 803 Proteinstrukturen haben zwar untereinander Sequenzidentitäten von größer als 25 %, die Proteinsequenzen in den einzelnen Datensätzen sind jedoch redundanzfrei. An diesem Beispiel ist zu erkennen, dass eine Sortierung von Proteinen bzw. Proteinsequenzen nach der zu untersuchenden Eigenschaft, im vorliegenden Fall *Tetrapeptidtyp*, und die Beseitigung redundanter Informationen jeweils innerhalb dieser Gruppen im Vergleich zu einer nichtredundanten Ausgangsdatenbasis zu einer deutlichen Vergrößerung der Informationsmenge bezüglich der untersuchten Eigenschaft führt. Wie die größere Anzahl an gefundenen Tetrapeptiden nach Anwendung dieses Verfahrens in Tabelle IV-1, S. 39, zeigt, können mit Hilfe dieses Algorithmus Eigenschaften untersucht werden, die bei Verwendung einer nichtredundanten Ausgangsdatenbasis „verborgen" geblieben wären.

Die Berücksichtigung der Strukturunterschiede von Tetrapeptiden ($|\Delta\psi_2| > 25°$ oder $|\Delta\varphi_3| > 25°$) führte zu einer Erhöhung von 99 901 um 4 786 auf 104 687 Datensätze mit mindestens vier Diederwinkelpaaren (ψ_2,φ_3). In diesem Fall wurden 33 048 der 35 397 Strukturen des Ausgangsdatensatzes verwendet, was eine Erhöhung um 2 245 Proteinsequenzen bzw. Proteinstrukturen im Vergleich zu den rein sequenzbasierten *all-against-all* Alignments bedeutet. Es ist erkennbar, dass die Berücksichtigung der Strukturunterschiede von Tetrapeptiden nur zu einer verhältnismäßig geringfügigen Erweiterung der Datenbasis führt, so dass die Bedingung der Nichtredundanz auf Sequenzebene praktisch erhalten bleibt. Die Ergebnisse der Datenaufbereitung sind in Tabelle IV-1 zusammengefasst.

Die statistische Analyse der Konformationseigenschaften von Tetrapeptiden wurde in der vorliegenden Arbeit mit den nichtredundanten Datensätzen durchgeführt, die Diederwinkelabweichungen bei den ψ_2-Winkeln und den φ_3-Winkeln von Tetrapeptiden berücksichtigen.

IV.1.4. Auswertung von Wahrscheinlichkeitsdichtefunktionen

Die berechneten Wahrscheinlichkeitsdichtefunktionen ordnen jedem Diederwinkelpaar (ψ_2,φ_3) eines Tetrapeptids im Konformationsbereich $-180° \leq \psi_2,\varphi_3 \leq +180°$ eine Wahrscheinlichkeit zu. Durch Integration über den Konformationsbereich, der einen bestimmten Konformationstyp (*E*, *H*, *L* oder *X*) beschreibt, lässt sich die Wahrscheinlichkeit quantifizieren, mit der ein Tetrapeptid eine Struktur im selektierten Konformationsbereich einnimmt (vgl. Abschnitt II.3.3, S. 16).

IV.1.4.1. Vergleich der ψ_i-φ_i-Verteilung mit der ψ_i-φ_{i+1}-Verteilung

Der Konformationsbereich, den eine Aminosäure mit ihrem Diederwinkelpaar (ψ_i,φ_i) beschreiben kann, lässt sich in fünf Bereiche einteilen, die mit *E* (faltblatttypisch), *H* (helixtypisch), *G* (*G-turn*), *L* (*L-turn*) und *X* (*X-turn*) beschrieben werden können [Bystroff *et al.*, 2000]. In Tabelle II-1, S. 17, sind die Grenzen der einzelnen Bereiche angegeben. Der Terminus *typisch* beschreibt in diesem Zusammenhang nur die notwendige Bedingung für die Ausprägung einer bestimmten Sekundärstruktur. Im Rahmen der durchgeführten Konformationsanalysen in der vorliegenden Arbeit wurden die Konformationsbereiche *H* und *G* vereinigt. Die Abbildung IV-3A auf Seite 41

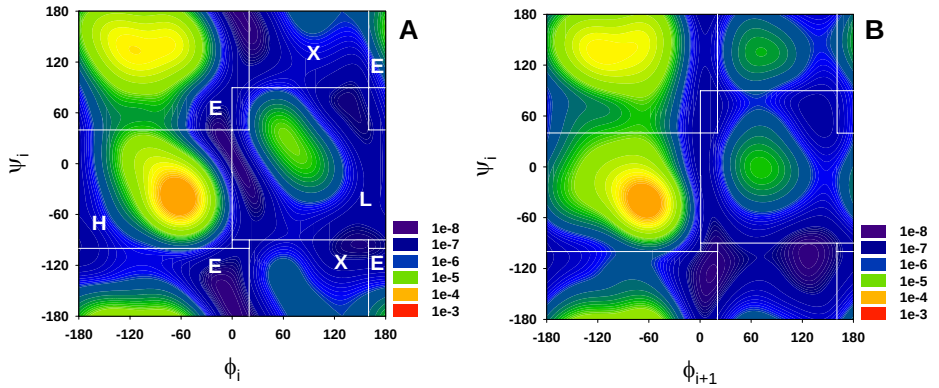


Abbildung IV-3 Vergleich der ψ_i-φ_i-Dichtefunktion mit der ψ_i-φ_{i+1}-Dichtefunktion. In den Abbildungen sind die vier Konformationsbereiche *E* (faltblatttypisch), *H* (helixtypisch), *L* (*L-turn*) und *X* (*X-turn*) eingezeichnet (siehe Tabelle II-1, S. 17). **A** ψ_i-φ_i-Dichtefunktion von 275 638 Aminosäuren aus zufällig gewählten 1 158 Proteinstrukturen der *FSSP*-Datenbank [Holm & Sander, 1997]. Diese Verteilung entspricht dem Ramachandran Diagramm [Ramachandran *et al.*, 1963]. **B** ψ_i-φ_{i+1}-Dichtefunktion von Dipeptiden aus der gleichen Datenbasis mit den Aminosäuren aus A an jeweils erster Position. Es ist eine veränderte Verteilung der Diederwinkel zu sehen, welche die Einteilung in die vier Konformationsbereiche jedoch nur wenig beeinflusst. Man erkennt, dass die Dichtefunktion in jedem der vier Konformationsbereiche zu einem optimalen Diederwinkelpaar (ψ_i ,φ_{i+1}) konvergiert. Dies sind für *E* (+136°,-114°), für *H* (-40°,-64°), für *L* (-2°,+70°) und für *X* (+136°,+72°). Die Abbildung IV-4 zeigt die Strukturen der entsprechenden Dipeptide.

zeigt die Dichtefunktion der ψ_i-φ_i-Verteilung von 275 638 Aminosäuren aus 1 158 zufällig gewählten Strukturen aus der *FSSP*-Datenbank [Holm & Sander, 1997]. Die errechnete Dichtefunktion entspricht dem Ramachandran Diagramm [Ramachandran *et al.*, 1963]. Die eingezeichneten Felder kennzeichnen die vier Konformationsbereiche nach Tabelle II-1, S. 17. Im Vergleich dazu zeigt die Abbildung IV-3B die ψ_i-φ_{i+1}-Verteilung der Dipeptide aus derselben Datenbasis mit den Aminosäuren aus der ψ_i-φ_i-Verteilung an erster Position. Die beiden Verteilungen zeigen bei gleicher Klassifizierung der Konformationsbereiche eine unterschiedliche Struktur. Prinzipiell wäre es bei der ψ_i-φ_{i+1}-Verteilung jedoch nicht erlaubt, die Bezeichnung faltblatttypische Konformation (*E*), helixtypische Konformation (*H*), *L-turn* (*L*) oder *X-turn* (*X*) zu verwenden, da sich diese Klassifizierung auf die Konformation nur einer Aminosäure bezieht. Dessen ungeachtet zeigt die Struktur der ψ_i-φ_{i+1}-Dichtefunktion, dass aufgrund der Peakverteilung diese Einteilung im Wesentlichen ihre Gültigkeit behält, so dass deren Beibehalten sinnvoll erscheint. Die ψ_i-φ_{i+1}-Dichtefunktion beschreibt die erlaubten Konformationen eines Dipeptids. In dieser Funktion ist für jeden Konformationszustand ein Peak, d.h. eine optimale Konformation zu erkennen. Die Abbildung IV-4 zeigt für jedes der daraus ableitbaren vier Dipeptide die Struktur. Diese Dipeptide lassen sich als die vier gemeinsamen Grundbausteine aller Proteinstrukturen auffassen. Die ψ_i-φ_{i+1}-Dichtefunktion lässt die genauen Konformationen der beiden Aminosäuren unbestimmt (φ_i unbekannt, ψ_{i+1} unbekannt). Es sollte daher überprüft werden, welchen Konformationszustand die Aminosäuren eines Dipeptids bei vorgegebenem Konformationszustand des Dipeptids annehmen können.

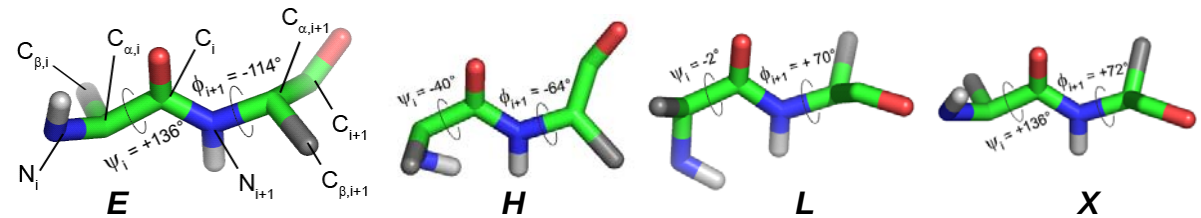


Abbildung IV-4 Dargestellt sind die vier Strukturen eines Dipeptids. Sie entsprechen der jeweils optimalen Konformation in den Konformationsbereichen ***E*** (faltblatttypisch), ***H*** (helixtypisch), ***L*** (*L-turn*) und ***X*** (*X-turn*), die aus den Peaks in Abbildung IV-3B abgeleitet wurde.

Tabelle IV-2 In der Tabelle sind die Häufigkeiten der Konformationszustände (*Zustand*) *E*, *H*, *L* und *X* der zweiten Aminosäure in Abhängigkeit vom Konformationszustand der ersten Aminosäure bei definiertem Konformationszustand des Dipeptids gelistet. Die Konformation des Dipeptids wird mit (ψ_i,φ_{i+1}), die erste Aminosäure wird über das Diederwinkelpaar (ψ_i,φ_i) und die zweite Aminosäure über ihr Diederwinkelpaar (ψ_{i+1},φ_{i+1}) beschrieben. Die Konformationen der ersten und zweiten Aminosäure sind vollständig beschrieben. Für diese Statistik wurden 275 638 Dipeptide aus 1 158 zufällig ausgewählten Strukturen der *FSSP*-Datenbank [Holm & Sander, 1997] verwendet. Die Grenzen der Konformationsbereiche sind in Tabelle II-1, S. 17, gelistet. Hierbei definieren *E* und *H* den faltblatttypischen bzw. den helixtypischen Konformationszustand, *L* und *X* den *L-turn* bzw. den *X-turn*. Exemplarisch entnimmt man der Tabelle, dass bei einem faltblatttypischen Konformationszustand (*E*) eines Dipeptids in 94.9 % der Fälle die erste Aminosäure ebenso in diesem Konformationszustand vorliegt und die zweite Aminosäure in 21.5 % der Fälle einen helixtypischen Konformationszustand ausbildet. Die Kombination aus faltblatttypischem Konformationszustand der ersten Aminosäure und einer Konformation vom Typ *X-turn* bei der zweiten Aminosäure wird nicht beobachtet. Die Ergebnisse in dieser Tabelle lassen den Schluss zu, dass sich die Konformationszustände benachbarter Aminosäuren nicht unabhängig voneinander ausbilden. Die Daten dieser Tabelle entsprechen denjenigen von Tetrapeptiden ohne Berücksichtigung der ersten und vierten Aminosäure.

Zustand Dipeptid	Zustand der ersten Aminosäure	Häufigkeit des Zustandes der ersten Aminosäure (%)	Häufigkeit Zustand der zweiten Aminosäure (%)			
			E	*H*	*L*	*X*
E	*E*	94.9	73.3	21.5	0.1	0
	H	0	0	0	0	0
	L	2.1	1.6	0.5	0.0	0
	X	3.0	1.7	1.3	0.0	0
H	*E*	0	0	0	0	0
	H	94.2	15.2	79.0	0	0
	L	5.8	4.4	1.4	0	0
	X	0.0	0.0	0.0	0	0
L	*E*	4.7	0	0	2.4	2.3
	H	80.7	3.2	0	62.1	15.4
	L	14.6	0.3	0	12.6	1.6
	X	0	0	0	0	0
X	*E*	96.2	0	0	68.9	27.4
	H	0.1	0	0	0.1	0.0
	L	0	0	0	0	0
	X	3.7	0	0	2.0	1.7

Tabelle IV-3 Wahrscheinlichkeiten *P* für die Konformationszustände *E* (faltblatttypisch), *H* (helixtypisch), *L* (*L-turn*) und *X* (*X-turn*) für die ψ_i-φ_i-Dichtefunktion und ψ_i-φ_{i+1}-Dichtefunktion der entsprechenden Verteilungen von 275 638 Aminosäuren bzw. Dipeptiden aus 1 158 zufällig gewählten, analysierbaren Proteinstrukturen der *FSSP*-Datenbank [Holm & Sander, 1997] und Vergleich mit dem Anteil an „echter" Sekundärstruktur nach Auswertung der Proteinstrukturen mit dem Programm *DSSP* [Kabsch & Sander, 1983]. Die Konformationszustände *E*, *H*, *L* und *X* sind in Tabelle II-1, S.17. definiert. Bei der Bestimmung der Sekundärstrukturanteile mit *DSSP* wurde die reduzierte Symbolik verwendet [Jones, 1999] (siehe Text). Bei den Dipeptiden mussten beide Aminosäuren dem Konformationszustand *E* bzw. *H* entsprechen, damit dem entsprechenden Dipeptid der Typ *E* bzw. *H* zugewiesen wurde. Alle Kombinationen verschiedener Symbole entsprechen dem Typ *C*. Die geklammerten Werte in den Spalten *DSSP (E)* und *DSSP (H)* geben den Anteil an Aminosäuren bzw. Dipeptiden an, der bei dem mit *DSSP* bestimmten Sekundärstrukturtyp (*E* oder *H*) ebenso in dem Konformationsbereich *E* oder *H* nach Tabelle II-1 beobachtet wurde.

Verteilung	*P(E)*	*DSSP* (*E*)	*P(H)*	*DSSP* (*H*)	*P(L)*	*P(X)*	*DSSP* (C)
ψ_i-φ_i	0.43	0.22 (96 %)	0.51	0.38 (99 %)	0.046	0.014	0.40
ψ_i-φ_{i+1}	0.43	0.18 (95 %)	0.51	0.34 (99 %)	0.035	0.025	0.48

Diesen Zusammenhang zeigt die Tabelle IV-2, S. 42. Es ist zu erkennen, dass bei einem faltblatttypischen (*E*) oder helixtypischen (*H*) Konformationszustand eines Dipeptids in den meisten Fällen die beiden Aminosäuren ebenso diesen Konformationszustand annehmen, wobei die Häufigkeit für den jeweiligen Konformationszustand bei der ersten Aminosäure deutlich größer ist. Der faltblatttypische Konformationszustand eines Dipeptids kann auch beschrieben werden, wenn keine der beiden Aminosäuren einen faltblatttypischen Konformationszustand annimmt. Besonders auffällig ist dieser Umstand bei den Dipeptiden, die einen Konformationszustand *X-turn* beschreiben. In fast allen Fällen wird dieser Konformationszustand durch eine Kombination von faltblatttypischer Konformation der ersten Aminosäure und einer Konformation vom Typ *L* der zweiten Aminosäure gebildet. Die Einteilung des gesamten Konformationsraumes $-180° \leq \psi_i,\varphi_i \leq +180°$ bei Tetrapeptiden in die Konformationsbereiche *E*, *H*, *L* und *X* hat nur formalen Charakter. Sie ist jedoch für die Selektion geeigneter Tetrapeptide für das Modellierungsschema von Bedeutung (siehe Abschnitt IV.2, S. 53). In der vorliegenden Arbeit wurde die Klassifizierung des Konformationsraumes $-180° \leq \psi_2,\varphi_3 \leq +180°$ in die Konformationszustände *E*, *H*, *L* und *X* gemäß Tabelle II-1, S. 17, beibehalten.

Die ψ- und φ-Winkel von Aminosäuren korrelieren mit bestimmten Sekundärstrukturen und Konformationszuständen. Im Folgenden sollte überprüft werden, ob die Einteilung in diese vier Konformationszustände bei den ψ_i- und den φ_{i+1}-Winkeln von Dipeptiden ebenso Sekundärstrukturelemente beschreiben und damit ihre Gültigkeit behält. Inwieweit die Konformation einer Aminosäure oder eines Dipeptids in einem Protein auch tatsächlich eine Sekundärstruktur beschreibt, lässt sich nur unter Berücksichtigung des lokalen und globalen Kontextes feststellen, in dem sich diese Aminosäure bzw. das Dipeptid befindet. Eine Sekundärstrukturzuordnung einzelner Aminosäuren wird mit dem Programm *DSSP* ermöglicht [Kabsch & Sander, 1983]. Zur Erörterung dieser Fragestellung wurden von allen analysierbaren 275 638 Dipeptiden aus den 1 158 zufällig gewählten Proteinstrukturen der *FSSP*-Datenbank die Sekundärstrukturen der ersten und zweiten Aminosäure bestimmt. Hierzu wurde in Analogie zu Jones die reduzierte *DSSP*-Symbolik verwendet [Jones, 1999]. Dabei werden die acht Strukturtypen (*H, I, G, E, B, S, T,* -) drei Klassen zugeordnet, d.h. die *DSSP*-Symbole *E* (*extended strand*), *B* (*β-bridge*) werden beide dem Symbol *E* zugewiesen, die Symbole *H* (*α-helix*) und *G* (*3-10 helix*) dem Symbol *H* und alle anderen Symbole dem Symbol *C* (*coil*). Da in der vorgestellten Klassifizierung der Dipeptidkonformationen in die Konformationszustände *E*, *H*, *L* und *X* ein bestimmtes Dipeptid immer nur einem Konformationszustand zugeordnet werden kann, wurde bei unterschiedlichen Strukturtypen der beiden Aminosäuren dem jeweiligen Tetrapeptid der Strukturtyp *C* (*coil*) zugewiesen. Die Ergebnisse in Tabelle IV-3, S. 42, zeigen, dass für die Konformationszustände *E* und *H* in der ψ_i-φ_i-Dichtefunktion bzw. in der ψ_i-φ_{i+1}-Dichtefunktion jeweils gleich große Anteile beobachtet werden. Es ergeben sich $P(H) = 0.51$ und $P(E) = 0.43$. Die Analyse der tatsächlichen Sekundärstrukturanteile mit *DSSP* zeigt im Vergleich dazu, dass in der ψ_i-φ_i-Verteilung nur 49 % der Aminosäuren, die nach Tabelle II-1, S. 17, einen faltblatttypischen Konformationszustand annehmen, auch tatsächlich in einer Faltblattstruktur zu finden sind (entspricht dem Anteil von 22 % in Tabelle IV-3, S. 42). Dies sind 96 % der Aminosäuren, die eine gemäß *DSSP* definierte Faltblattkonformation annehmen. Die verbleibenden 4 % werden nicht durch den faltblatttypischen Konformationszustand beschrieben. Im helixtypischen Konformationszustand *H* beschreiben 73 % des Anteils an Aminosäuren in diesem Konformationszustand eine „echte" helikale Struktur nach *DSSP* (38 % in Tabelle IV-3). Die Dimension des helikalen Konformationsbereiches erfasst in diesem Fall 99 % der Aminosäuren, die nach *DSSP* auch eine helikale Konformation annehmen. Für die ψ_i-φ_{i+1}-Verteilung ergeben sich vergleichbare Werte.

Der faltblatttypische Konformationszustand *E* erfasst 95 % der Dipeptide, die nach *DSSP* in einer Faltblattstruktur zu finden sind. Für den helixtypischen Konformationszustand findet man, dass 99 % der Dipeptide, die nach *DSSP* eine helikale Struktur annehmen, durch den helixtypischen Konformationszustand beschrieben werden. Es wurden nur geringe Anteile an Aminosäuren bzw. Dipeptiden gefunden, die in der ψ_i-φ_i-Verteilung und in der ψ_i-φ_{i+1}-Verteilung eine Konformation vom Typ *L* oder *X* aufweisen. Ein Vergleich dieser Werte mit den Werten, die sich nach Analyse der Strukturen mit *DSSP* ergeben, ist nicht möglich, da bei *DSSP* eine Sekundärstruktur nicht nur eine Funktion der entsprechenden Diederwinkel ist, sondern auch durch den globalen Kontext bestimmt wird. Eine Aminosäure bzw. ein Dipeptid kann also beispielsweise eine helikale Konformation ausgebildet haben, aber dennoch als *coil* klassifiziert sein.

Dieser Abschnitt unterstreicht den Befund, dass die einzelnen Konformationszustände nur als notwendige Bedingung für die Ausbildung einer bestimmten Sekundärstruktur zu werten sind. Obwohl die Klassifizierung in die vier Konformationsbereiche ursprünglich die möglichen Sekundärstrukturen *einer* Aminosäure beschreibt, lässt sie sich offenbar ebenso auf die ψ_i-φ_{i+1}-Verteilung anwenden.

IV.1.4.2. Analyse der ψ_2-φ_3-Verteilungen von Tetrapeptiden

Die Analyse der ψ_2-φ_3-Verteilungen von Tetrapeptiden entspricht der Konformationsanalyse von Dipeptiden in Abhängigkeit von den flankierenden Aminosäuren. Bei der ersten und vierten Aminosäure werden jedoch keine strukturellen Betrachtungen durchgeführt, so dass deren Konformation völlig unbestimmt bleibt. Die Abbildung IV-5 auf Seite 45 illustriert am Beispiel von sechs Dichtefunktionen die ψ_2-φ_3-Verteilung verschiedener Tetrapeptide. Das Tetrapeptid AMEY zeigt eine Wahrscheinlichkeit von $P(H) = 1.0$ für eine helixtypische Konformation (*H*). Die entsprechende Dichtefunktion beschreibt eine monomodale Verteilung. Die Dichtefunktion hat ihr Maximum im Diederwinkelpaar (ψ_2,φ_3) mit $\psi_2 = -36°$ und $\varphi_3 = -64°$. Im Gegensatz dazu zeigt das Tetrapeptid AMDY eine hohe Präferenz von $P(E) = 0.78$ für eine faltblatttypische Konformation (*E*), die jedoch eine sehr heterogene Verteilung der einzelnen Diederwinkelpaare (ψ_2,φ_3) aufweist. AVYS ist ein Beispiel für ein Tetrapeptid mit einer bimodalen Verteilung auf die Konformationszustände *H* (helixtypisch) und *E* (faltblatttypisch). Das Diederwinkelpaar mit der höchsten Wahrscheinlichkeit liegt bei $P(E) = P(H) = 0.5$ im helixtypischen Konformationsbereich. Die ψ_2-φ_3-Verteilung des Tetrapeptids ESNE zeigt eine Wahrscheinlichkeit von $P(E) = 0.91$ für eine faltblatttypische Konformation (*E*) mit einer wahrscheinlichsten Konformation (ψ_2,φ_3) von $\psi_2 = +164°$ und $\varphi_3 = -100°$. Daneben findet man eine zweite Konformation mit $\psi_2 = +158°$ und $\varphi_3 = -68°$, die fast gleich wahrscheinlich ist. Interessanterweise beobachtet man bei dem Tetrapeptid ENSE eine sehr starke Änderung des bevorzugten Konformationszustandes hin zu einer helixtypischen Konformation (*H*). Offensichtlich müssen die Dichtefunktionen von ψ_2-φ_3-Verteilungen eines Tetrapeptids hinsichtlich ihres wahrscheinlichsten Konformationszustandes nicht mit denen des inversen Tetrapeptids übereinstimmen. Das Tetrapeptid GGGG weist erwartungsgemäß keine Präferenz für einen bestimmten Konformationszustand auf. In diesem Fall sind alle vier möglichen Konformationszustände fast gleichwahrscheinlich. In Tabelle IV-4, S. 45, sind die Wahrscheinlichkeiten der diskutierten Tetrapeptide für die einzelnen Konformationszustände aufgeführt.

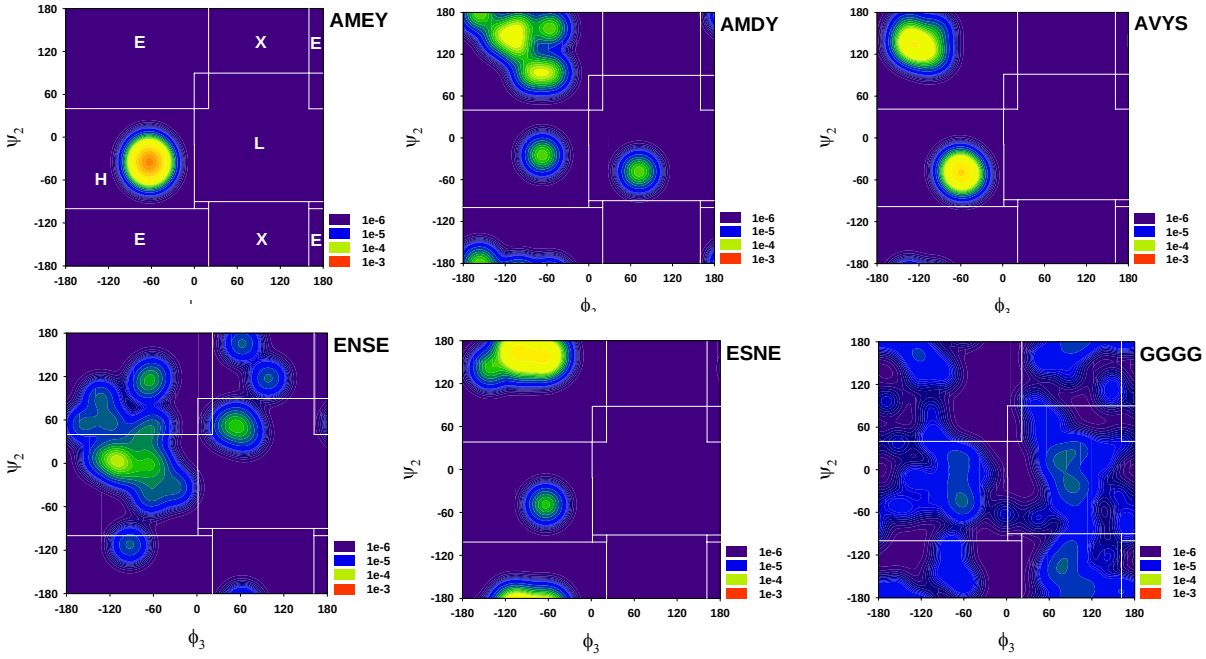


Abbildung IV-5 Darstellung der Dichtefunktionen von ψ_2-φ_3-Verteilungen verschiedener Tetrapeptide. Die weiß umrandeten Bereiche definieren die vier Konformationszustände *E*, *H*, *L* und *X* (siehe Tabelle II-1, S. 17). Die Verteilungen zeigen deutliche Unterschiede in ihrer Modalität und Wahrscheinlichkeit für einen bestimmten Konformationszustand. Die Wahrscheinlichkeiten für die Konformationszustände der einzelnen Tetrapeptide sind in Tabelle IV-4 aufgeführt.

Im Folgenden wurde untersucht, inwieweit eine Präferenz individueller Tetrapeptide für einen Konformationszustand *E*, *H*, *L* oder *X* besteht. Die Kenntnis eines möglichen bevorzugten Konformationszustandes sollte für die Entwicklung eines Systems zum Design von Aminosäuresequenzen zu vorgegebenen Proteinstrukturen verwendet werden. Dazu wurde von allen 104 687 ψ_2-φ_3-Verteilungen der Tetrapeptide die Dichtefunktion errechnet und zu jedem Tetrapeptid der Konformationszustand maximaler Wahrscheinlichkeit bestimmt. Die Abbildung IV-6 auf Seite 46 zeigt die Verteilungsfunktion des Anteils an Tetrapeptiden mit einer bestimmten Wahrscheinlichkeit für den wahrscheinlichsten Konformationszustand und die aufgeschlüsselten Anteile bezüglich der vier Konformationszustände. Danach zeigen 5.6 % der untersuchten Tetrapeptide eine Wahrscheinlichkeit von $P = 1.0$ für einen der Konformationszustände *E*, *H*, *L*, oder *X*, und

Tabelle IV-4 Wahrscheinlichkeiten für die Konformationszustände *E* (faltblatttypisch), *H* (helixtypisch), *L* (*L-turn*) und *X* (*X-turn*) für verschiedene Tetrapeptide mit den wahrscheinlichsten Konformationen für die jeweiligen Konformationszustände. *P(E)*, *P(H)*, *P(L)* und *P(X)* bezeichnen die Wahrscheinlichkeiten für die Konformationszustände nach Tabelle II-1, S. 17. $(\psi_2,\varphi_3)_{MAX,ZUSTAND}$ (Zustand = *E*, *H*, *L* oder *X*) definiert die wahrscheinlichste Konformation für das Tetrapeptid in dem jeweiligen Konformationszustand. Die Spalte *Anzahl* listet die Anzahl der Wertepaare (ψ_2,φ_3), aus denen die jeweilige Dichtefunktion errechnet wurde. Die Wahrscheinlichkeitsdichtefunktionen sind in Abbildung IV-5 dargestellt.

Tetrapeptid	*P(E)*	$(\psi_2,\varphi_3)_{MAX,E}$	*P(H)*	$(\psi_2,\varphi_3)_{MAX,H}$	*P(L)*	$(\psi_2,\varphi_3)_{MAX,L}$	*P(X)*	$(\psi_2,\varphi_3)_{MAX,X}$	Anzahl
AMEY	0.00	-	1.00	(-36°,-64°)	0.00	-	0.00	-	6
AMDY	0.78	(+146°,-110°)	0.11	(-24°,-66°)	0.11	(-48°,+72°)	0.00	-	9
AVYS	0.50	(+134°,-126°)	0.50	(-50°,-60°)	0.00	-	0.00	-	18
ENSE	0.31	(+114°,- 66°)	0.46	(-4°,-110°)	0.14	(+52°,+54°)	0.09	(+166°,+62°)	22
ESNE	0.91	(+164°,-100°)	0.09	(-48°,-64°)	0.00	-	0.00	-	11
GGGG	0.22	(+162°,-124°)	0.23	(-44°,-62°)	0.30	(-22°,+92°)	0.25	(-136°,+78°)	92

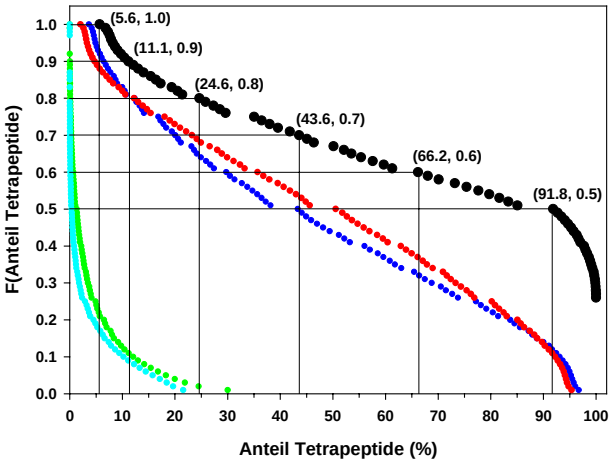


Abbildung IV-6 Verteilungsfunktion des Anteils an Tetrapeptiden mit einer bestimmten Wahrscheinlichkeit für den wahrscheinlichsten Konformationszustand (●). Es ist zu erkennen, dass 43.6 % der 104 687 (=100 %) untersuchten Tetrapeptide eine Mindestwahrscheinlichkeit von $P = 0.7$ für ihren jeweils wahrscheinlichsten Konformationszustand *E* (●), *H* (●), *L* (●) oder *X* (●) haben. Die Präferenz für einen der beiden wahrscheinlichsten Konformationszustände *E* oder *H* ist bei den untersuchten Tetrapeptiden annähernd gleich groß. Für die Zustände *L* und *X* findet man nur geringe Wahrscheinlichkeiten. Die Grenzen der Konformationsbereiche sind in Tabelle II-1, S. 17, definiert.

43.6 % der untersuchten Tetrapeptide haben eine Wahrscheinlichkeit von mindestens $P = 0.70$ für einen dieser Konformationszustände. Diese Abbildung zeigt für die untersuchten Tetrapeptide einen etwa gleich großen Anteil an Tetrapeptiden mit einer ähnlich hohen Wahrscheinlichkeit für einen faltblatttypischen oder helixtypischen wahrscheinlichsten Konformationszustand. Die Konformationszustände *L* und *X* werden nur selten mit einer hohen Wahrscheinlichkeit ausgebildet. Die Einteilung der Tetrapeptide in Gruppen maximaler Wahrscheinlichkeit für die Konformationszustände *E*, *H*, *X* und *L* stellt nur eine Näherung dar. Eine Analyse der Form von Dichtefunktionen von ψ_2-φ_3-Verteilungen zeigte im Folgenden, dass ein Tetrapeptid im Allgemeinen nicht alle Konformationen eines Konformationszustandes mit gleicher Wahrscheinlichkeit annehmen kann. Vielmehr konvergieren die beobachteten Konformationen zu bestimmten wahrscheinlichsten Diederwinkelpaaren. Tetrapeptide, die eine hohe Präferenz für einen bestimmten Konformationszustand besitzen, können trotzdem große Abweichungen in ihrer wahrscheinlichsten Konformation zeigen. Dieses Ergebnis ist in Abbildung IV-7 exemplarisch für die Dichtefunktionen der Tetrapeptide VEYT, VDYT und IDFS gezeigt.

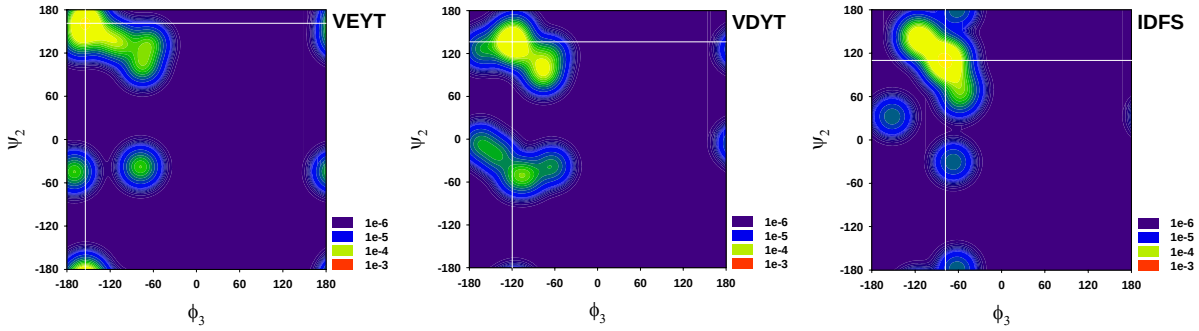


Abbildung IV-7 Es sind die Wahrscheinlichkeitsdichtefunktionen der ψ_2-φ_3-Verteilung von Tetrapeptiden dargestellt, die eine hohe Wahrscheinlichkeit für eine faltblatttypische Konformation (*E*) zeigen. Der entsprechende Konformationsbereich ist in Tabelle II-1, S. 17, definiert. Die Hauptpeaks konvergieren jeweils zu verschiedenen Diederwinkelpaaren (ψ_2,φ_3). **VEYT** Der Hauptpeak konvergiert zu dem Diederwinkelpaar (ψ_2,φ_3) mit $\psi_2 = +162°$ und $\varphi_3 = -154°$. Die Wahrscheinlichkeit für eine faltblatttypische Konformation errechnet sich zu $P(E) = 0.82$. Die Dichtefunktion wurde aus 11 Diederwinkelpaaren (ψ_2,φ_3) berechnet. **VDYT** Der Hauptpeak konvergiert zu dem Diederwinkelpaar (ψ_2,φ_3) mit $\psi_2 = +136°$ und $\varphi_3 = -120°$. Die Wahrscheinlichkeit für einen faltblatttypischen Konformationszustand errechnet sich zu $P(E) = 0.67$. Die Dichtefunktion wurde aus 15 Diederwinkelpaaren (ψ_2,φ_3) berechnet. **IDFS** Der Hauptpeak konvergiert zu dem Diederwinkelpaar (ψ_2,φ_3) mit $\psi_2 = +110°$ und $\varphi_3 = -78°$. Die Wahrscheinlichkeit für den faltblatttypischen Konformationszustand errechnet sich zu $P(E) = 0.89$. Die Dichtefunktion wurde aus 18 Diederwinkelpaaren (ψ_2,φ_3) berechnet.

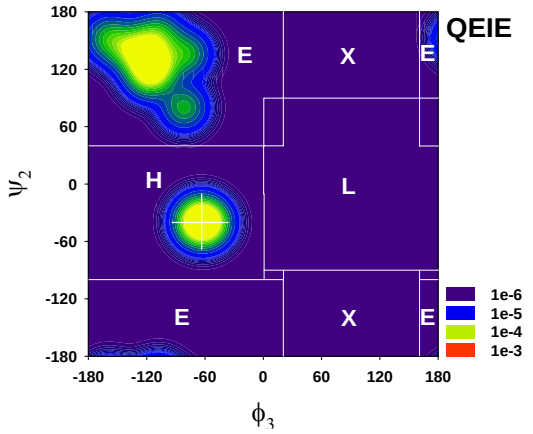


Abbildung IV-8 Wahrscheinlichkeitsdichtefunktion der ψ_2-φ_3-Verteilung des Tetrapeptids QEIE. Das Tetrapeptid besitzt eine Wahrscheinlichkeit von $P(E) = 0.71$ für eine faltblatttypische Konformation und eine Wahrscheinlichkeit von $P(H) = 0.29$ für eine helixtypische Konformation. Das Fadenkreuz markiert die wahrscheinlichste Konformation, die durch $\psi_2 = -40°$ und $\varphi_3 = -62°$ beschrieben wird. Diese Konformation liegt nicht in dem Konformationsbereich maximaler Wahrscheinlichkeit (*E*). Die Dichtefunktion wurde aus 13 Wertepaaren (ψ_2,φ_3) errechnet. Die Grenzen der Konformationsbereiche sind in Tabelle II-1, S. 17, gelistet.

Alle drei Tetrapeptide haben eine hohe strukturelle Präferenz für einen faltblatttypischen Konformationszustand (*E*), ihre wahrscheinlichste Konformation konvergiert jedoch zu unterschiedlichen Diederwinkelpaaren (ψ_2,φ_3). Die wahrscheinlichste Konformation eines Tetrapeptids muss nicht im Konformationsbereich maximaler Wahrscheinlichkeit liegen. Dies ist in Abbildung IV-8 am Beispiel des Tetrapeptids QEIE gezeigt. Obwohl dieses Tetrapeptid mit einer Wahrscheinlichkeit von $P(E) = 0.71$ einen faltblatttypischen Konformationszustand annimmt, ist die wahrscheinlichste Konformation eine helixtypische Konformation (*H*).

IV.1.4.3. Einfluss der Datenaufbereitung auf den wahrscheinlichsten Konformationszustand

Aus den 3 486 nichtredundanten Aminosäuresequenzen von Proteinstrukturen konnten 78 786 Datensätze von Tetrapeptiden mit mindestens vier Wertepaaren (ψ_2,φ_3) bestimmt werden. Diese Datensätze werden im Folgenden als *min_subsets* bezeichnet. Demgegenüber resultierte die Vorsortierung der Aminosäuresequenzen in Gruppen, die jeweils einen bestimmten Tetrapeptidtyp beinhalteten, sowie die Berücksichtigung der strukturellen Divergenz von Tetrapeptiden zu 104 687 Datensätzen mit mindestens vier Wertepaaren (vgl. Tabelle IV-1, S. 39). Diese Datensätze enthalten aufgrund der Datenaufbereitung entweder genau so viele oder mehr Wertepaare (ψ_2,φ_3), als die *min_subsets*. Im Folgenden werden diese Datensätze als *max_subsets* bezeichnet. Es ergibt sich die Frage, inwieweit der größere Strukturinformationsgehalt in den *max_subsets* und damit die Art der Datenaufbereitung einen Einfluss auf den wahrscheinlichsten Konformationszustand der einzelnen Tetrapeptide hat. Zur Erörterung dieser Fragestellung wurde der jeweils wahrscheinlichste Konformationszustand (*E*, *H*, *L* oder *X*) der einzelnen Tetrapeptide aus den *min_subsets* bestimmt und der Quotient aus dieser Wahrscheinlichkeit und der Wahrscheinlichkeit des analogen Tetrapeptids aus dem *max_subset* für den gleichen Konformationszustand (*E*, *H*, *L* oder *X*) ermittelt. Bei gleichem bevorzugten Konformationszustand lässt sich aus dieser Darstellung ableiten, wieviel wahrscheinlicher der wahrscheinlichste Konformationszustand in den *min_subsets* im Vergleich zu denjenigen aus den *max_subsets* ist. In Abbildung IV-9A auf Seite 48 ist jeweils die Anzahl der *min_subsets* mit einem bestimmten relativen Anteil an Wertepaaren (ψ_2,φ_3) bezüglich der Anzahl an Wertepaaren in den *max_subsets* dargestellt. Die Abbildung IV-9B zeigt die daraus abgeleitete Verteilungsfunktion *F*. Es ist zu erkennen, dass 24.5 % der 78 786 *min_subsets* 50 % oder weniger Konformationsdaten (ψ_2,φ_3) von Tetrapeptiden beinhalten, als in den korrespondierenden *max_subsets* zu finden sind.

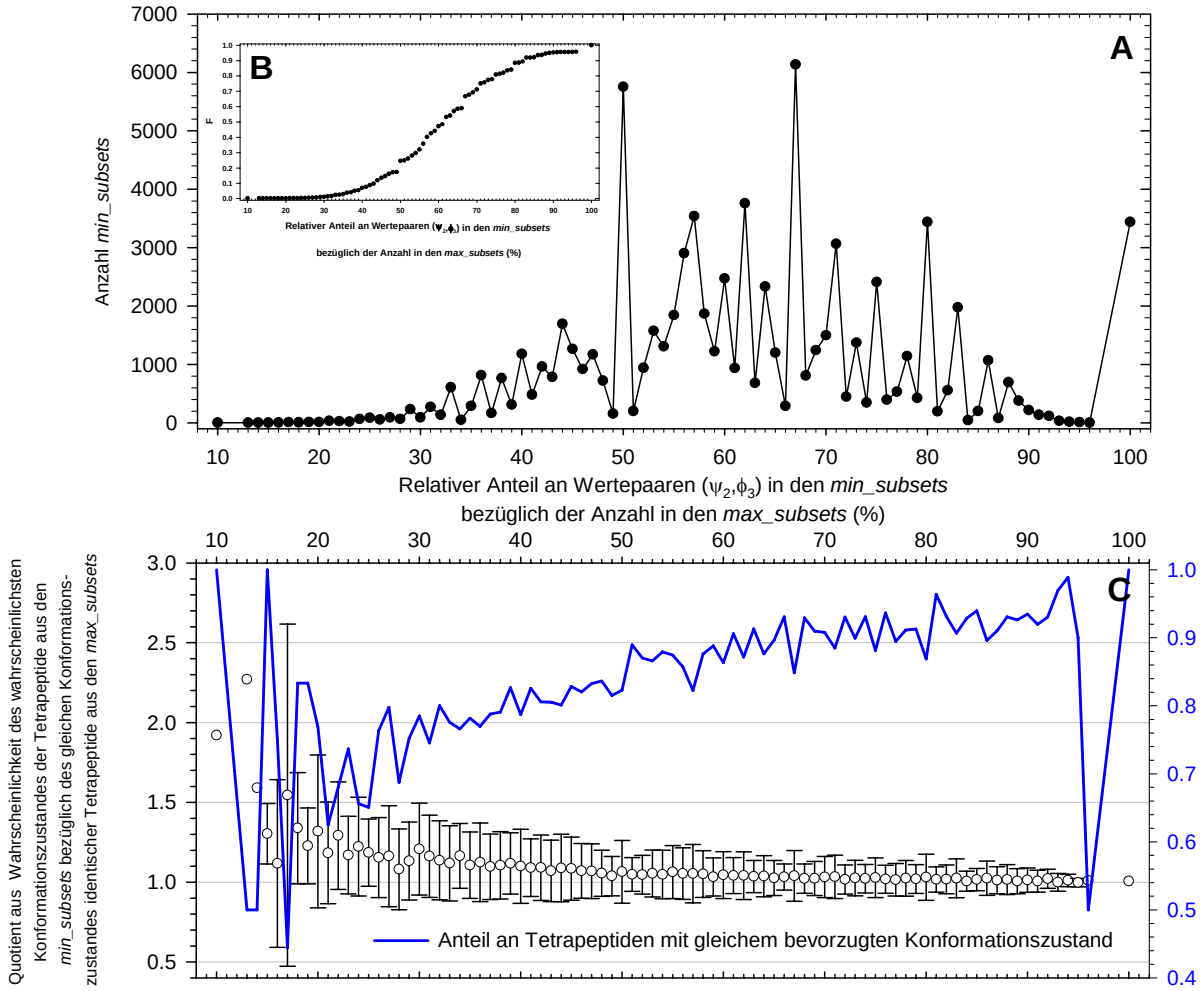


Abbildung IV-9 Einfluss der Datenaufbereitung auf den wahrscheinlichsten Konformationszustand *E*, *H*, *L* oder *X* eines Tetrapeptids. Die vier Konformationszustände sind in Tabelle II-1, S. 17, definiert. Die *min_subsets* entsprechen den 78 786 *subsets* von Tetrapeptidkonformationen mit mindestens vier Wertepaaren (ψ_2,φ_3), die aus der nichtredundanten Sequenzdatenbank von 3 486 Proteinstrukturen abgeleitet wurden. Die *max_subsets* bezeichnen diejenigen *subsets* von Tetrapeptidkonformationen, die nach Sortierung von 35 397 Aminosäuresequenzen von Proteinstrukturen hinsichtlich jeweils eines bestimmten Tetrapeptidtyps erhalten wurden. Bei diesen *subsets* erfolgte die Beseitigung der redundanten Aminosäuresequenzen unter Berücksichtigung der strukturellen Divergenz der Tetrapeptide (vgl. Tabelle IV-1, S. 39). **A** Anzahl an *min_subsets* mit einem bestimmten relativen Anteil an Datensätzen bezüglich der Anzahl an Datensätzen in den *max_subsets*. **B** Verteilungsfunktion von A **C** Quotient aus Wahrscheinlichkeit für den wahrscheinlichsten Konformationszustand für Tetrapeptide aus den *min_subsets* und den Tetrapeptiden aus den *max_subsets* in Abhängigkeit vom relativen Anteil an Datensätzen in den *min_subsets*. Der Typ des wahrscheinlichsten Konformationszustandes (*E*, *H*, *L* oder *X*) der Tetrapeptide ist in beiden *subsets* identisch. Die blaue Linie kennzeichnet den prozentualen Anteil der Tetrapeptide, die bei einem bestimmten relativen Anteil an Wertepaaren (ψ_2,φ_3) in den *min_subsets* den gleichen bevorzugten Konformationszustand besitzen wie die entsprechenden Tetrapeptide aus den *max_subsets*. Exemplarisch ist der Abbildung A zu entnehmen, dass 5 751 *min_subsets* von Tetrapeptidkonformationen gefunden wurden, die im Vergleich zu den korrespondierenden *max_subsets* nur 50 % an Wertepaaren beinhalten. Die Abbildung B zeigt in der Folge, dass 25.5 % der *min_subsets* 50 % oder weniger Wertepaare enthalten, als die korrespondierenden *max_subsets* gleicher Tetrapeptide. In Abbildung C wird ersichtlich, dass bei 82 % der Tetrapeptide in diesen 5 751 *subsets* der wahrscheinlichste Konformationszustand identisch mit demjenigen des analogen Tetrapeptids aus den *max_subsets* ist (blaue Linie). Bei gleichem bevorzugten Konformationszustand hat der Konformationszustand eines Tetrapeptids aus dem *min_subset* im Mittel eine höhere Wahrscheinlichkeit, als der des identischen Tetrapeptids aus dem *max_subset*. An den Datenpunkten ist die zum Mittelwert gehörende Standardabweichung gezeigt.

Die Abbildung IV-9C auf Seite 48 zeigt, dass der geringer werdende Anteil an Datensätzen in den *min_subsets* zu einer erhöhten Wahrscheinlichkeit des bevorzugten Konformationszustandes in diesen Datensätzen führt. In den *max_subsets* sind offensichtlich vermehrt Diederwinkelpaare (ψ_2,φ_3) enthalten, die nicht den wahrscheinlichsten Konformationszustand beschreiben, d.h. die *max_subsets* beschreiben eine größere strukturelle Diversität der einzelnen Tetrapeptide. Die blaue Linie kennzeichnet den prozentualen Anteil der Tetrapeptide, bei denen der wahrscheinlichste Konformationszustand in den *min_subsets* identisch mit demjenigen aus den *max_subsets* ist. Es ist zu erkennen, dass sich mit abnehmendem Anteil der Datensätze in den *min_subsets* der Anteil an Tetrapeptiden mit gleichem bevorzugten Konformationszustand ebenso verringert. Der Median der Anzahl der Wertepaare (ψ_2,φ_3) in den 78 786 *min_subsets* wurde zu einem Wert von 7 bestimmt (vgl. Tabelle IV-1, S. 39). Für die korrespondierenden 78 786 *max_subsets*, welche die Konformationsdaten der gleichen Tetrapeptide beinhalten, errechnete sich der Median zu einem von Wert 13. In beiden Fällen bezieht sich der Median auf eine Anzahl von mindestens vier Wertepaaren in den einzelnen *subsets*. Dies verdeutlicht den großen Informationsgewinn in den *max_subsets*, wie er aufgrund der angewendeten Datenaufbereitung erzielt wurde. Das in Abschnitt IV.1, S. 35, vorgestellte Prinzip der Datenaufbereitung führte zu 104 687 *subsets* von Tetrapeptiden mit mindestens vier Wertepaaren (ψ_2,φ_3). Betrachtet man die Tetrapeptide, die nicht in den *min_subsets* vertreten sind, so errechnet sich der Median für die verbleibenden 25 901 *max_subsets** zu einem Wert von 5. Die Analyse der Verteilungsfunktionen dieser drei *subsets* für die Wahrscheinlichkeit des wahrscheinlichsten Konformationszustandes in Abbildung IV-10 auf Seite 50 zeigt, dass mit steigender Anzahl von berücksichtigten Konformationsdaten aus verschiedenen Proteinen die Wahrscheinlichkeit für den bevorzugten Konformationszustand eines Tetrapeptids abnimmt. Dies beeinflusst im Besonderen diejenigen Tetrapeptide, die eine Wahrscheinlichkeit für einen der Konformationszustände *E*, *H*, *X*, oder *L* von mindestens $P = 0.7$ aufweisen. Die allgemein hohe Präferenz für einen bestimmten Konformationszustand bleibt trotz der großen Unterschiede in der Anzahl der Wertepaare (ψ_2,φ_3), d.h. der Anzahl an berücksichtigten Proteinstrukturen, erhalten. Die Abbildung IV-9C auf Seite 48 zeigt eine Verringerung des Anteils an Tetrapeptiden mit gleichem bevorzugten Konformationszustand mit Abnahme des Anteils an Datensätzen in den *min_subsets*. Berücksichtigt man nun alle Tetrapeptide (unabhängig von der Anzahl an Datensätzen), so findet man in 13.8 % der 78 786 Fälle eine Änderung des bevorzugten Konformationszustandes aufgrund des größeren Strukturdatensatzes. Die Analyse der Konformationseigenschaften dieser 13.8 % der Tetrapeptide in den *min_subsets* und den *max_subsets* zeigte, dass die mittlere Wahrscheinlichkeit für den bevorzugten Konformationszustand in den *min_subsets* von 54 % auf 52 % für die analogen Tetrapeptide aus den *max_subsets* sinkt. Bei 86.2 % der 78 786 Tetrapeptide in den *subsets* führte die größere Anzahl an berücksichtigten Proteinstrukturen zu keiner Änderung des bevorzugten Konformationszustandes. Für die Tetrapeptide in den *max_subsets* wurde hier eine mittlere Wahrscheinlichkeit von 69 % für den bevorzugten Konformationszustand errechnet. Dagegen wurde für die korrespondierenden Tetrapeptide in den *min_subsets* die mittlere Wahrscheinlichkeit zu einem Wert von 72 % bestimmt. Auch hier beobachtet man eine verringerte Wahrscheinlichkeit für den bevorzugten Konformationszustand der Tetrapeptide aus den *max_subsets* im Vergleich zu denjenigen der analogen Tetrapeptide aus den *min_subsets*.

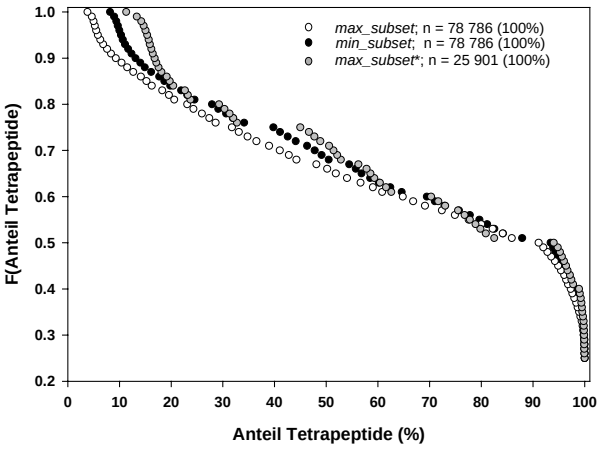


Abbildung IV-10 Vergleich der Verteilungsfunktionen F des Anteils an Tetrapeptiden mit einer bestimmten Wahrscheinlichkeit für den wahrscheinlichsten Konformationszustand aus den *min_subsets*, *max_subsets* und *max_subsets**. Die vier Konformationszustände sind in Tabelle II-1, S. 17, definiert. Exemplarisch entnimmt man dieser Abbildung, dass 8.2 % der Tetrapeptide aus den *min_subsets* eine Wahrscheinlichkeit von $P = 1.0$ für einen der vier Konformationszustände haben. Die *min_subsets* beinhalten die Strukturdaten der 78 786 Tetrapeptide aus der nichtredundanten Sequenzdatenbank von 3 486 Proteinstrukturen. In den *max_subsets* sind die Strukturdaten der äquivalenten Tetrapeptide aus den *min_subsets* enthalten. Die *max_subsets* resultieren aus der in Abschnitt IV.1, S. 35, beschriebenen Datenaufbereitung. Die *max_subsets** beinhalten die Konformationseigenschaften der 25 901 Tetrapeptide, die aufgrund der modifizierten Datenaufbereitung zusätzlich berücksichtigt werden konnten. Der Median der Anzahl von Wertepaaren (ψ_2,φ_3) errechnete sich zu 7 für die *min_subsets*, zu 13 für die *max_subsets* und zu 5 für die *max_subsets**. Der Median bezieht sich auf eine Mindestanzahl von vier Wertepaaren. Die Abbildung zeigt, dass mit steigender Anzahl berücksichtigter Proteinstrukturen in den *subsets* die Wahrscheinlichkeit für den bevorzugten Konformationszustand der Tetrapeptide abnimmt. Dies betrifft im Besonderen die hohen Wahrscheinlichkeiten. Der allgemeine Trend zur Ausbildung eines bevorzugten Konformationszustandes *E*, *H*, *L* oder *X* bleibt erhalten.

Die Ergebnisse zeigen, dass die Erhöhung der Datenvielfalt sowohl bei gleichem als auch bei unterschiedlich bevorzugtem Konformationszustand eines bestimmten Tetrapeptids zu einer tendentiellen Verringerung der Wahrscheinlichkeit des bevorzugten Konformationszustandes in den *max_subsets* im Vergleich zu den bevorzugten Konformationszuständen in den *min_subsets*, führt. Die Änderung des wahrscheinlichsten Konformationszustandes in den *max_subsets* betrifft diejenigen Tetrapeptide, bei denen der bevorzugte Konformationszustand in den *min_subsets* nur schwach ausgeprägt ist.

IV.1.4.4. Die Analyse sequenzähnlicher Tetrapeptide am Beispiel von AMDY

Die Ähnlichkeit von Aminosäuren lässt sich sowohl unter physikalischen als auch unter evolutionären Gesichtspunkten beschreiben. Beide Sichtweisen müssen nicht immer zum gleichen Ergebnis führen. In diesem Sinne führt die evolutionäre Betrachtungsweise der Ähnlichkeit in der BLOSUM62-Matrix [Henikoff & Henikoff, 1992] zum Beispiel dazu, dass Lysin und Glutaminsäure als ähnliche Aminosäuren klassifiziert sind, obwohl sie eine entgegengesetzte Ladung besitzen. Da die während der Datenaufbereitung durchgeführten Alignments unter Verwendung der BLOSUM62-Matrix durchgeführt wurden, bezieht sich der Begriff *Ähnlichkeit* im Folgenden auf Paare von Aminosäuren, die in dieser Substitutionsmatrix einen *score* von größer oder gleich +1 besitzen. Die BLOSUM62-Matrix ist im Angang VII.9 dargestellt. Gemäß diesem Konzept ist zu Alanin die Aminosäure Serin (+1) ähnlich, im Fall von Methionin die Aminosäuren Isoleucin (+1), Leucin (+2) und Valin (+1), zu Asparaginsäure die Aminosäuren Glutaminsäure (+2) und Asparagin (+1) und im Fall von Tyrosin die Aminosäuren

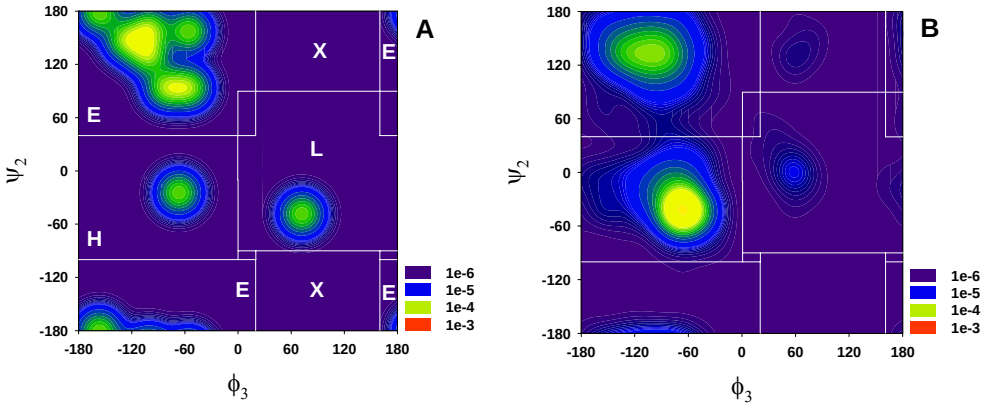


Abbildung IV-11 Vergleich der ψ_2-φ_3-Dichtefunktion der ψ_2-φ_3-Verteilung des Tetrapeptids AMDY (**A**) mit der ψ_2-φ_3-Dichtefunktion der ψ_2-φ_3-Verteilungen von 80 zu AMDY sequenzähnlichen Tetrapeptiden. Der Begriff *Ähnlichkeit* bezieht sich auf Paare von Aminosäuren, die in der BLOSUM62-Matrix [Henikoff & Henikoff, 1992] einen *score* von ≥ +1 besitzen. (**B**). Jeder Datensatz eines berücksichtigten Tetrapeptids beinhaltete mindestens vier Wertepaare (ψ_2,φ_3). In die Dichtefunktionen sind die in Tabelle II-1, S. 17, definierten Grenzen der Konformationsbereiche *E* (faltblatttypisch), *H* (helixtypisch), *L* (*L-turn*) und *X* (*X-turn*) eingezeichnet. Die Wahrscheinlichkeiten für die vier Konformationsbereiche errechnen sich für A bzw. B zu $P(E)_A = 0.78$, $P(H)_A = 0.11$, $P(L)_A = 0.11$, $P(X)_A = 0.00$ und $P(E)_B = 0.45$, $P(H)_B = 0.51$, $P(L)_B = 0.03$, $P(X)_B = 0.01$.

Histidin (+2), Phenylalanin (+3) und Tryptophan (+2). Wie aus Abbildung IV-11 zu entnehmen ist, hat das Tetrapeptid AMDY eine Wahrscheinlichkeit von $P(E) = 0.78$ zur Ausprägung eines faltblatttypischen Konformationszustandes. Die Analyse aller zu diesem Tetrapeptid ähnlichen Tetrapeptide sollte zeigen, ob diese Präferenz auch in der Gruppe der zu AMDY ähnlichen Tetrapeptide vorhanden ist. Zu diesem Zweck wurden alle Diederwinkelpaare (ψ_2,φ_3) aus den Datensätzen, die mindestens vier Wertepaare (ψ_2,φ_3) enthielten, verwendet und daraus die entsprechende Wahrscheinlichkeitsdichtefunktion errechnet. Von den 95 zu AMDY ähnlichen Tetrapeptiden konnten 80 Tetrapeptide mit insgesamt 930 Diederwinkelpaaren (ψ_2,φ_3) berücksichtigt werden. Die Wahrscheinlichkeit für die vier Konformationszustände *E*, *H*, *L* und *X* errechnete sich aus dieser Dichtefunktion zu $P(E) = 0.45$ für einen faltblatttypischen Konformationszustand, $P(H) = 0.51$ für den helixtypischen Konformationszustand, $P(L) = 0.03$ für den *L-turn* und $P(X) = 0.01$ für den *X-turn*. Diese Wahrscheinlichkeitsverteilung ist mit derjenigen der ψ_i-φ_{i+1}-Verteilung von Dipeptiden vergleichbar (vgl. Tabelle IV-3, S. 42). Eine bevorzugte Strukturbildung hinsichtlich eines bestimmten Konformationszustandes ist in der Gruppe der zu AMDY ähnlichen Tetrapeptide nicht zu beobachten. Um die Präferenz der individuellen Tetrapeptide zu beschreiben, vergleicht die Abbildung IV-12 auf Seite 52 die Verteilungsfunktion des Anteils der Tetrapeptide, die zu AMDY sequenzähnlich sind und die eine bestimmte Wahrscheinlichkeit für den wahrscheinlichsten Konformationszustand aufweisen, mit der Verteilungsfunktion aller untersuchten Tetrapeptide, die nicht sequenzähnlich zu AMDY sind. Dies betrifft 104 606 Tetrapeptide. Die beiden Funktionen zeigen einen ähnlichen Verlauf mit etwas geringeren Wahrscheinlichkeiten bei der Verteilungsfunktion der zu AMDY ähnlichen Tetrapeptide. Der Unterschied in den Verteilungsfunktionen der Gesamtheit aller untersuchten Tetrapeptide (104 687 Tetrapeptide) und der um die Anzahl der AMDY ähnlichen Tetrapeptide reduzierten ist vernachlässigbar, so dass ein Vergleich mit den Verteilungsfunktionen für die Konformationszustände *E*, *H*, *L* und *X* aus der Gesamtheit aller untersuchten Tetrapeptide zulässig ist (Abbildung IV-6 auf Seite 46). Dieser Vergleich zeigt, dass die Verteilungsfunktionen für die Konformationszustände *E*, *H*, *L* und *X* einen ähnlichen Verlauf haben, so dass man die Gruppe der zu AMDY ähnlichen Tetrapeptide mit der strukturellen Präferenz ihrer individuellen

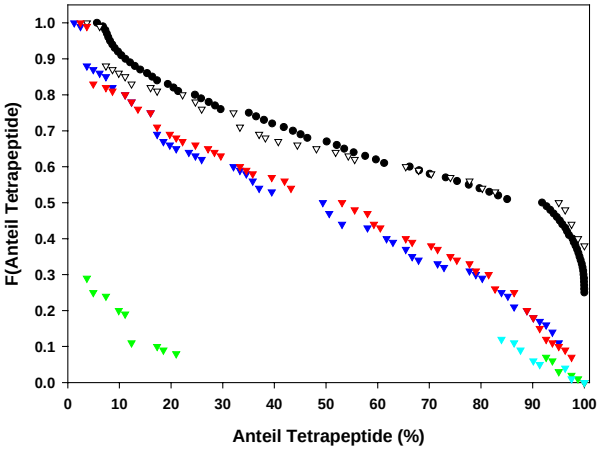


Abbildung IV-12 Vergleich der Verteilungsfunktion F des Anteils an Tetrapeptiden mit einer bestimmten Wahrscheinlichkeit für den wahrscheinlichsten Konformationszustand der zu AMDY sequenzähnlichen Tetrapeptide (▽) mit der Verteilungsfunktion aller Tetrapeptide (●) (ohne Berücksichtigung der AMDY ähnlichen Tetrapeptide und AMDY). Die Ähnlichkeit bezieht sich auf *scores* von ≥+1 zwischen zwei Aminosäuren gemäß der BLOSUM62-Matrix [Henikoff & Henikoff, 1992]. Beide Funktionen haben einen ähnlichen Verlauf. Dies gilt ebenso für die Verteilungsfunktionen der einzelnen wahrscheinlichsten Konformationszustände *E* (▼), *H* (▼), *L*(▼) und *X* (▼) (vgl. Abbildung IV-6 auf Seite 46).

Tetrapeptide als einen repräsentativen Ausschnitt aus der Gesamtmenge der Tetrapeptide darstellen kann.

Das Ergebnis zeigt, dass ausgehend vom wahrscheinlichsten Konformationszustand eines speziellen Tetrapeptids der Schluss auf die strukturellen Präferenzen der Gruppe der zu diesem Tetrapeptid ähnlichen Tetrapeptide offenbar nicht ohne Weiteres möglich ist. Es läßt sich daher schließen, dass die Termini *Sequenzähnlichkeit* und *Strukturähnlichkeit* auf Tetrapeptidebene keinen kausalen Zusammenhang haben müssen.

IV.2. Tetrapeptidbasiertes Proteindesign

Die fragmentbasierte Modellierung einer alternativen Proteinsequenz zu einer experimentell bestimmten Proteinstruktur beruht auf der Vermutung, dass das Überlappen von Peptidfragmenten in ihrer wahrscheinlichsten Konformation zu einer Sequenz führt, deren Konformation mit der niedrigsten Freien Energie die Zielstruktur ist [Holmes & Tsai, 2004]. Die wahrscheinlichste Konformation eines Fragmentes (Tetrapeptids) lässt sich aus den in der vorliegenden Arbeit errechneten Wahrscheinlichkeitsdichtefunktionen der ψ_2-φ_3-Verteilungen ableiten. Bei dem im Folgenden vorgestellten Schema der fragmentbasierten Modellierung mittels Wahrscheinlichkeitsdichtefunktionen wird die Information verwendet, ob ein bestimmtes Fragment mit hoher Wahrscheinlichkeit in der Lage ist, die durch die Zielstruktur vorgegebene Konformation auszubilden und zwar unabhängig davon, ob dieses Fragment die gewünschte Konformation auch schon einmal in dem in der Zielstruktur definierten strukturellen Kontext angenommen hat.

IV.2.1. Die Überlappung von Tetrapeptidfragmenten

Ein wichtiger Aspekt bei der Modellierung eines Proteins mit Hilfe von Peptidfragmenten ist das Ausmaß der Überlappung einzelner Fragmente. Eine Aminosäuresequenz der Länge *n* lässt sich in (*n*-*m*+1) Fragmente der Länge *m* zerlegen. In diesem Fall werden (*m*-1) Aminosäuren eines Fragmentes von dem nachfolgenden Fragment überlappt. Die Verlängerung der modellierten Aminosäuresequenz erfolgt danach stets nur um eine Aminosäure. Das im Folgenden beschriebene fragmentbasierte Modellierungsschema mit Tetrapeptiden verwendet diese maximale Überlappung. Die Abbildung IV-13 auf Seite 54 erläutert die grundlegende Funktionsweise des vorgestellten tetrapeptidbasierten Algorithmus am Beispiel der schematischen Modellierung eines Heptapeptids. Es ist zu erkennen, dass bis auf die ersten und die letzten drei Aminosäuren einer Aminosäuresequenz, jede weitere Aminosäure an vier überlappenden Tetrapeptiden partizipiert. Durch die maximale Überlappung der einzelnen Fragmente lässt sich deshalb die Konformation von sieben aufeinanderfolgenden Aminosäuren kontrollieren. Die Konformation einer Aminosäure ist über ihre beiden Diederwinkel ψ und φ bestimmt. Die errechneten Wahrscheinlichkeitsdichtefunktionen beschreiben jedoch nur den ψ-Winkel der zweiten Aminosäure (ψ_2) und den φ-Winkel der dritten Aminosäure (φ_3), so dass die genaue Konformation der mittleren beiden Aminosäuren in einem Tetrapeptid unbekannt ist. Durch die Überlappung von Tetrapeptiden über drei Aminosäuren hinweg erfolgt bei bekannten ψ_2-und φ_3-Diederwinkeln in jedem Modellierungsschritt eine vollständige Beschreibung der Konformation der drittletzten Aminosäure in der Aminosäuresequenz. Dies folgt daraus, dass die zweite Aminosäure des letzten Tetrapeptids einer Aminosäuresequenz (ψ bekannt, φ unbestimmt) identisch mit der dritten Aminosäure des vorhergehenden Tetrapeptids ist (ψ unbestimmt, φ bekannt) und die Konformation dieser Aminosäure damit vollständig beschrieben wird. Durch die geeignete Wahl von Tetrapeptiden sollte es daher möglich sein, die gewünschte Konformation einer Aminosäure innerhalb einer Sequenz mit hoher Wahrscheinlichkeit zu erzeugen.

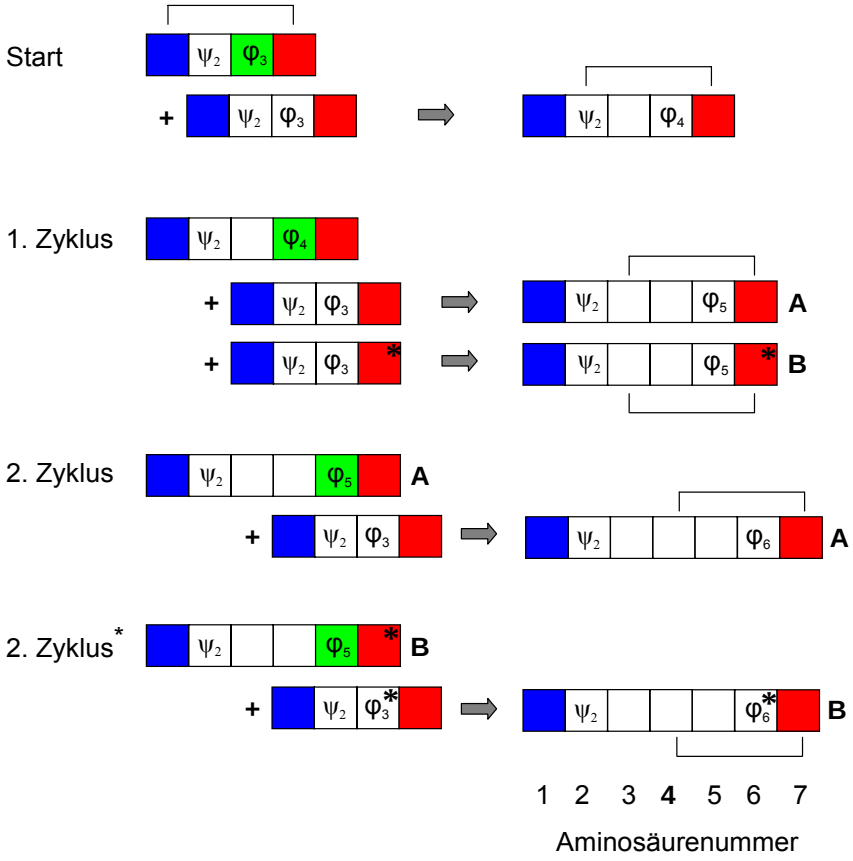


Abbildung IV-13 Modellierung einer Aminosäuresequenz durch rekursives Verknüpfen von Tetrapeptidfragmenten. Das blau markierte Rechteck kennzeichnet den N-Terminus eines Fragmentes, das rot markierte Rechteck den C-Terminus. Bei den verwendeten Tetrapeptidfragmenten ist die Konformation der N-terminalen und C-terminalen Aminosäure undefiniert (ψ unbekannt, φ unbekannt). Die Konformation der zweiten Aminosäure wird nur über ihren ψ-Winkel (ψ_2) beschrieben, φ_2 ist nicht definiert. Die Konformation der dritten Aminosäure ist nur über ihren φ-Winkel (φ_3) determiniert, ψ_3 ist unbestimmt. Das grün dargestellte Rechteck kennzeichnet die Aminosäure einer im Modellierungsprozess befindlichen Aminosäuresequenz, deren Konformation (φ bekannt, ψ unbekannt) im nächsten Schritt durch die Überlappung mit der zweiten Aminosäure eines Tetrapeptidfragmentes (ψ bekannt, φ unbekannt) festgelegt wird. Das Schema zeigt die Modellierung eines Heptapeptids. **Start** Die Modellierung beginnt mit einem Starttetrapeptid. Die Überlappung mit einem zweiten Tetrapeptidfragment erfolgt über drei Aminosäuren hinweg, so dass die modellierte Aminosäuresequenz in jedem Zyklus um eine Aminosäure verlängert wird. Die Tetrapeptidfragmente werden in der Weise gewählt, dass die ψ-Winkel der zweiten Aminosäure und der φ-Winkel der dritten Aminosäure dieses Tetrapeptids bestmöglich die korrespondierenden Diederwinkel aus der Zielstruktur beschreiben (vgl. Abschnitt IV.2.3.3, S. 62). Wurden Aminosäuren in der Zielstruktur als Randbedingungen definiert, so erfolgt die Wahl des Tetrapeptids unter Berücksichtigung dieser Aminosäuren (siehe Abschnitt IV.2.3.2, S. 59). Jeder Zyklus endet mit der vollständigen strukturellen Beschreibung der drittletzten Aminosäure der Sequenz, d.h. beide Diederwinkel ψ und ϕ dieser Aminosäure sind nun bekannt. **1. Zyklus** Bei dem entstandenen Pentapeptid besteht die Möglichkeit, die Sequenz mit zwei verschiedenen Tetrapeptiden zu verlängern, die sich durch die Aminosäure am C-Terminus unterscheiden (dargestellt durch *), was in der Folge zu der Sequenz A und der Sequenz B führt. Im Ergebnis entstehen zwei verschiedene Hexapeptide. Die Sequenz B wird zwischengespeichert. Die weitere Modellierung erfolgt mit Sequenz A. **2. Zyklus** Die Sequenz A wird durch Überlappung mit dem nächsten Tetrapeptid um eine weitere Aminosäure verlängert. Im Ergebnis erhält man ein Heptapeptid. Nachdem die Sequenz von A bestimmt wurde, wird nun die gespeicherte Sequenz B genommen und in einem zweiten Zyklus (markiert durch *Zyklus**) mit einem geeigneten Tetrapeptid um eine weitere Aminosäure verlängert. Der Modellierungsprozess führt in diesem Fall zu zwei Ergebnissequenzen, die sich in der vorletzten Aminosäure unterscheiden. Die Konformation der N-terminalen und C-terminalen Aminosäure einer modellierten Aminosäuresequenz bleibt völlig unbestimmt (ψ unbekannt, φ unbekannt). Bei der zweiten Aminosäure ist φ und bei der vorletzten Aminosäure ψ unbekannt. Es ist ersichtlich, dass die **4.** Aminosäure in einem Heptapeptid an vier überlappenden Tetrapeptiden partizipiert. Bei einer günstigen Wahl der einzelnen Tetrapeptide lässt sich daher die Konformation jeder Aminosäure in einem Heptapeptid direkt kontrollieren.

IV.2.2. Analyse der Strukturbildung von Oligopeptidfragmenten

Das vorgestellte Modellierungsschema verwendet Tetrapeptide in ihrem wahrscheinlichsten Konformationszustand (*E*, *H*, *L* oder *X*, vgl. Abschnitt II.3.3, S. 16) für die Berechnung von Aminosäuresequenzen zu gegebenen Proteinstrukturen (vgl. Abschnitt IV.2.3.3, S. 62). Die wahrscheinlichste Konformation eines Fragmentes der Länge *n* wird dabei durch diejenige Struktur beschrieben, die aus dem wahrscheinlichsten Konformationszustand jedes der *n*-3 überlappenden Tetrapeptide resultiert. Inwieweit ein Zusammenhang besteht (Alternativhypothese) oder nicht besteht (Nullhypothese), nach der eine Sequenz von Tetrapeptiden in ihrem wahrscheinlichsten Konformationszustand die Zielstruktur gegenüber allen anderen möglichen Strukturen mit einer größeren Wahrscheinlichkeit bevorzugt, als statistisch zu erwarten wäre, lässt sich nicht vorhersagen. Ein Vergleich beider Hypothesen gelingt jedoch durch eine statistische Analyse der Konformationen von Fragmenten verschiedener Länge aus experimentell bestimmten Proteinstrukturen (vgl. dazu Abschnitt II.5, S. 18).

Nimmt man an, dass jedes Tetrapeptid einer Aminosäuresequenz seinen wahrscheinlichsten Konformationszustand unabhängig von dem benachbarter Tetrapeptide ausbildet, so entspricht die Wahrscheinlichkeit, das untersuchte Fragment in seiner wahrscheinlichsten Struktur anzutreffen, dem Produkt der Wahrscheinlichkeit des wahrscheinlichsten Konformationszustandes jedes einzelnen Tetrapeptids im Gesamtfragment. Dieses Ergebnis würde implizieren, dass kein Zusammenhang zwischen der Wahrscheinlichkeit der einzelnen Tetrapeptide einer Sequenz für den jeweiligen Zielkonformationszustand und der im Resultat ausgebildeten Struktur des Gesamtfragmentes besteht (Nullhypothese; unabhängige Strukturbildung). Die Alternativhypothese geht davon aus, dass ein solcher Zusammenhang existiert (bedingte bzw. abhängige Strukturbildung). Eine bedingte Strukturbildung würde die Wahrscheinlichkeit drastisch erhöhen, mit dem vorgestellten Modellierungsprinzip Aminosäuresequenzen errechnen zu können, die eine definierte Tertiärstruktur besitzen. Die Überprüfung der Alternativhypothese wird mit Gleichung II-17, S. 19, ermöglicht. Ihre Wahrscheinlichkeit entspricht der beobachteten Häufigkeit, mit der ein Fragment einer bestimmten Länge seine wahrscheinlichste Struktur auch tatsächlich ausgebildet hat. Ist diese beobachtete Häufigkeit größer als bei einer unabhängigen Strukturbildung zu erwarten wäre, dann wird nach Gleichung II-17 $g > 1$ und es liegt eine bedingte (abhängige) Strukturbildung vor (positive Korrelation). Wenn $g = 1$ ist, dann ist die Ausbildung des wahrscheinlichsten Konformationszustandes benachbarter Tetrapeptide unabhängig voneinander. Die wahrscheinlichste Struktur ist in diesem Fall nur mit der statistisch erwarteten Häufigkeit zu beobachten. Eine negative Korrelation würde durch $g < 1$ angezeigt.

Zur Erörterung dieses Problems wurden exemplarisch alle Pentamere, Hexamere, Heptamere, Oktamere und 16-mere aus zufällig gewählten 1 735 Proteinstrukturen der *FSSP*-Datenbank [Holm & Sander, 1997] ermittelt. Die Analyse umfasste 351 087 Pentamere, 329 125 Hexamere, 309 114 Heptamere, 290 209 Oktamere und 177 243 16-mere. Die Konformationsanalyse der einzelnen Fragmente erfolgte nach Kreuzvalidierung, d.h. die verwendeten Dichtefunktionen beinhalteten keine Strukturinformationen über das jeweils aktuell betrachtete Protein (zur Durchführung der Kreuzvalidierung siehe Abschnitt II.4, S. 18). Für jedes untersuchte Protein der *FSSP*-Datenbank wurde ein individueller Satz an Dichtefunktionen errechnet. Um Fragmente mit einer unterschiedlichen Anzahl an Aminosäuren vergleichen zu können, wurde als unabhängige Größe das geometrische Mittel P_{geo} der Wahrscheinlichkeit des wahrscheinlichsten Konformationszustandes der Tetrapeptide eines Fragmentes verwendet. Die Abbildung IV-14A auf Seite 56 zeigt für die untersuchten Fragmente, dass Pentapeptide (zwei überlappende

Tetrapeptide) über den gesamten mittleren Wahrscheinlichkeitsbereich einer Strukturbildung unterliegen, bei der die beiden Tetrapeptide ihren Konformationszustand *unabhängig* voneinander ausbilden (violette Punkte). Exemplarisch ist der Abbildung zu entnehmen, dass bei einer mittleren Wahrscheinlichkeit des wahrscheinlichsten Konformationszustandes von $P_{geo} = 0.7$ der beiden Tetrapeptide für den Zielkonformationszustand der Grad der Korrelation zu $g \approx 1$ bestimmt wurde, was bedeutet, dass die wahrscheinlichste Struktur in dieser Gruppe von Pentapeptiden tatsächlich nur in ca. 49 % der Fälle beobachtet werden konnte. Die Berücksichtigung von Fragmenten mit sechs oder mehr Aminosäuren zeigt ab einer

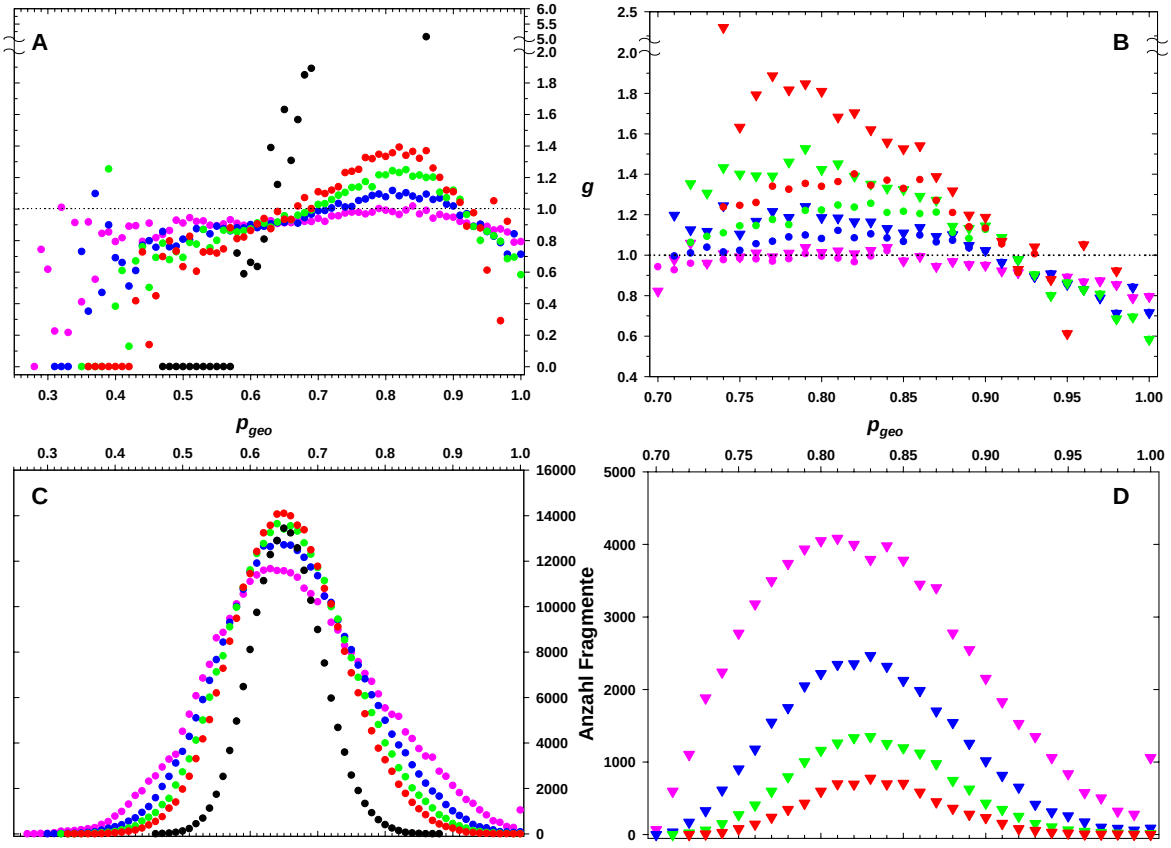


Abbildung IV-14 Bedingte Strukturbildung bei Fragmenten unterschiedlicher Länge. P_{geo} bezeichnet das geometrische Mittel der Wahrscheinlichkeit des wahrscheinlichsten Konformationszustandes aller (n-3) Tetrapeptide eines Fragmentes. g ist das Verhältnis zwischen beobachteter und statistisch erwarteter Häufigkeit der wahrscheinlichsten Struktur eines Fragmentes (Gleichung II-17). Bei $g > 1$ liegt eine bedingte Strukturbildung vor, $g = 1$ indiziert eine unabhängige Strukturbildung, $g < 1$ weist auf eine negative Korrelation hin. Die Analyse umfasste 351 087 Pentamere, 329 125 Hexamere, 309 114 Heptamere, 290 209 Oktamere und 177 243 16-mere und erfolgte nach Kreuzvalidierung. Es wurden nur Fragmente berücksichtigt, die bei einer bestimmten Wahrscheinlichkeit P_{geo} mindestens viermal gefunden wurden. Es wurde eine *a priori* Wahrscheinlichkeit von 0.01 verwendet, wenn die beobachtete Wahrscheinlichkeit einen Wert von kleiner als 0.01 aufwies. Es sind (●) 5-mer, (●) 6-mer, (●) 7-mer, (●) 8-mer, (●)16-mer. Unter der Einschränkung, dass die Wahrscheinlichkeit des wahrscheinlichsten Konformationszustandes aller Tetrapeptide $P \geq 0.7$ ist, definieren (▼) 5-mer, (▼) 6-mer, (▼) 7-mer, (▼) 8-mer. **A** Bei Pentapeptiden ist über den gesamten Wahrscheinlichkeitsbereich nur eine unabhängige Strukturbildung zu beobachten. Längere Fragmente zeigen ab ca. $P_{geo} = 0.7$ eine bedingte Strukturbildung. Aus Gründen der besseren Darstellbarkeit ist der Bereich $2.2 < g < 5$ nicht gezeigt. Dies betrifft die Datenpunkte (P_{geo},g) für die 16-mere: (0.74,2.7); (0.75,2.4); (0.76,2.6); (0.77,2.6); (0.78,2.4); (0.79,3.2); (0.8,2.7); (0.81, 2.1); (0.82,3.9); (0.83,2.4); (0.84,3.1); (0.85,4.3). **B** Berücksichtigt man nur Fragmente bei denen die Bedingung $P \geq 0.7$ erfüllt ist, beobachtet man eine weitere Verschiebung zugunsten einer abhängigen (bedingten) Strukturbildung. Es wurden keine 16-mere gefunden, bei denen alle Tetrapeptide dieser Forderung genügten. **C / D** Anzahl an beobachteten Fragmenten mit einer bestimmten Wahrscheinlichkeit P_{geo} für den wahrscheinlichsten Konformationszustand.

Wahrscheinlichkeit von $P_{geo} \approx 0.7$ eine abhängige Strukturbildung überlappender Tetrapeptide. In diesen Fällen ist eine Tendenz zur Strukturbildung hin zum wahrscheinlichsten Konformationszustand jedes einzelnen Tetrapeptids zu erkennen, die größer ist, als sie bei einer unabhängigen Strukturbildung zu beobachten wäre. Betrachtet man Fragmente, deren einzelne Tetrapeptide ausschließlich eine Wahrscheinlichkeit von $P \geq 0.7$ für den Zielkonformationsbereich aufweisen, so beobachtet man eine weitere Vergrößerung der Tendenz zu einer bedingten Strukturbildung wie aus Abbildung IV-14B auf Seite 56 zu entnehmen ist. „Wahrscheinlichkeits-(dichte)funktionen beschreiben unvollständiges Wissen." [Dill & Bromberg, 2003]. Dies führt dazu, dass die Graphen in Abbildung IV-14A und B bei einer Wahrscheinlichkeit von eins nicht zu einem Wert von eins konvergieren. Infolge der Kreuzvalidierung können seltene Konformationen aus den ψ_2-φ_3-Verteilungen entfernt worden sein, die das jeweilige Zielprotein an einer bestimmten Position ausgebildet hat. Dies kann dazu führen, dass aus Dichtefunktionen, die mehrere Konformationszustände erlauben, Funktionen entstehen, die eine Wahrscheinlichkeit von $P = 1.0$ für einen Konformationsbereich aufweisen. Da diese nun die jeweiligen Zielkonformationszustände nicht mehr richtig beschreiben können und kein anderer Konformationszustand mehr erlaubt ist, nimmt der Graph an dieser Position Werte von kleiner als eins an.

IV.2.3. *Redesign* des Proteins Top7

Für einen ersten Test des beschriebenen tetrapeptidbasierten Modellierungsverfahrens wurde die Kristallstruktur von Top7 verwendet (*PDB*-Code: 1QYS). Top7 besteht aus 92 Aminosäuren. Es wurden acht Aminosäuresequenzen errechnet, die jeweils zu Top7 eine Sequenzidentität von kleiner als 30 % aufwiesen. Die entsprechenden Modelle bzw. Proteine werden mit M1, M2 … M8 bezeichnet. Das Design der Proteinmodelle gliedert sich in vier Abschnitte, die im Folgenden erläutert werden.

i. Definition des hydrophoben Musters der Proteinsequenz (Abschnitt IV.2.3.1, S. 58)
ii. Definition von Aminosäuren auf ausgewählten Positionen innerhalb der Proteinstruktur (Abschnitt IV.2.3.2, S. 59)
iii. Errechnung der Proteinsequenzen (Abschnitt IV.2.3.3, S. 62)
iv. Modellierung der Aminosäureseitenketten und Energieminimierung der Modelle (Abschnitt IV.2.3.5, S. 76)

IV.2.3.1. Definition des hydrophoben Musters der Proteinsequenz

Mit dem hydrophoben Muster (HP-Motiv) einer Proteinsequenz wird definiert, welche Aminosäuren **h**ydrophoben (apolaren) bzw. **p**olaren (hydrophilen) Charakter aufweisen sollen. Eine korrekte Verteilung hydrophober Aminosäuren ist die Voraussetzung für die Bildung des hydrophoben Kerns eines Proteins, dessen Ausbildung während der Proteinfaltung eine entscheidende Bedeutung zufällt. Der Hydrophobe Effekt gilt als eine wesentliche Triebkraft der Proteinfaltung [Dill, 1990]. Die Auswahl einzelner Tetrapeptide muss daher neben dem richtigen Diederwinkel auch das definierte HP-Motiv der Zielstruktur berücksichtigen. Das hydrophobe Muster der modellierten Sequenzen wurde visuell anhand der Struktur von Top7 bestimmt. Als hydrophob wurden die Aminosäuren A, V, L, I, F, W, M und P definiert. Entsprechend wurden C, D, E, G, H, K, R, N, Q, S, T und Y als polare Aminosäuren klassifiziert. Um eine gewisse Variabilität im HP-Motiv zu erlauben wurden in einigen Modellen an den Positionen 35, 41, 66, 74 polare mit hydrophoben Aminosäuren getauscht. Die Änderungen zeigt Tabelle IV-5. Die Sequenzen mit den gekennzeichneten hydrophoben Aminosäuren zeigt die Abbildung IV-18 auf Seite 63.

Tabelle IV-5 Variation des hydrophoben Musters an vier Positionen in den Modellen M1-M8. Die Spalte Position gibt die Nummer der Aminosäure in der Aminosäuresequenz an. Angegeben ist der Typ der hydrophoben Aminosäuren auf den Positionen. Bei den jeweils anderen Modellen befinden sich polare Aminosäuren auf diesen Positionen.

Position	Hydrophobe Aminosäure
35	M4, M5, M6 (ALA)
41	M7 (ALA); M1, M8 (LEU)
66	M1-M7 (ALA)
74	M1, M3 ,M4 ,M5, M6, M8 (MET)

IV.2.3.2. Definition von Aminosäuren als Randbedingungen

Bei der Errechnung der Sequenzen wurden die Aminosäuren Histidin, Prolin und Cystein nicht verwendet, um unerwünschte Nebeneffekte, die mit diesen Aminosäuren verbunden sein könnten, auszuschließen. Dies betrifft die mögliche cis/trans Isomerie von Prolin, die Oxidation von Cystein und die damit in der Folge mögliche Oligomerisierung der Proteine und die pH-abhängige Hydropathie von Histidin. Wie aus Abschnitt IV.2.3, S. 58, zu entnehmen ist, erfolgt die Modellierung der Seitenketten erst nach der Definition der Aminosäuren des Proteinrückgrats, d.h. in diesem Modellierungsstadium sind mögliche Kollisionen zwischen Aminosäureseitenketten nicht zu erkennen. Insbesondere bei großen Aminosäuren, die am hydrophoben Kern beteiligt sein können, kann dies zu einer späteren Überlappung von Seitenketten verschiedener Aminosäuren und in der Folge zu fehlerhaften Modellen führen. Aus diesem Grund wurden die Positionen von Phenylalanin bzw. Tryptophan und die Aminosäuren in deren unmittelbarer räumlicher Umgebung im hydrophoben Kern anhand des Proteinrückgrats von Top7 explizit festgelegt, ansonsten wurden diese Aminosäuren bei der Modellierung ebenso nicht zugelassen. Die Positionen von Alanin im hydrophoben Kern wurden ebenso manuell festgelegt, ansonsten wurde Alanin als Bestandteil des hydrophoben Kerns in gleicher Weise nicht zugelassen. Bei der Modellierung des hydrophoben Kerns von M7 wurden ausschließlich aliphatische hydrophobe Aminosäuren berücksichtigt. Bei den Proteinen M1-M6 und M8 wurden zusätzlich die aromatischen Aminosäuren Phenylalanin und Tryptophan verwendet. Es wurden die Anzahl und die Positionen dieser Aminosäuren in den Modellen variiert. In den Modellen M2, M3 und M4 wurde die Aminosäure Phenylalanin als Teil eines Faltblattes verwendet. Hierbei wurden in M2 zwei Phenylalanine auf den Positionen 48 und 50 definiert. Die Modelle M3 und M4 enthalten je ein Phenylalanin auf der Position 48 bzw. 50. Die Modelle M1, M5, M6 enthalten Phenylalanin als Teil der Helices. Hierbei besetzt Phenylalanin die Position 65 im Modell M1, die Position 38 in Modell M5 und die Positionen 38 und 65 in Modell M6. Das Modell M8 beinhaltet an Position 65 ein Tryptophan als aromatische Aminosäure. Die Größe und die Struktur von Phenylalanin und Tryptophan bedingen in einer engen räumlichen Umgebung nur eine sehr eingeschränkte Drehbarkeit um die C_α-C_β-Bindung. Strukturelle Unterschiede zwischen den experimentell zu bestimmenden Proteinen und den Modellen sollten in diesem Fall nicht zu der Ausbildung eines hydrophoben Kerns führen und im Resultat die Ausbildung kooperativ faltender Proteine verhindern. Die Abbildung IV-15 auf Seite 60 zeigt die Positionen von Phenylalanin und Tryptophan in den Sekundärstrukturdarstellungen der Modelle. In der Abbildung IV-16 auf Seite 60 ist an zwei Beispielen dargestellt, wie anhand der Sekundärstrukturdarstellung von Top7 durch visuelle Analyse Motive definiert wurden, die als Randbedingungen bei der Errechnung der Sequenzen dienten. Es ist zu erkennen, dass durch eine günstige Positionierung von Leucin in den beiden Helices die Ausbildung einer reißverschlussartigen Struktur ermöglicht wird. Die daraus resultierenden hydrophoben Wechselwirkungen könnten zu einer Stabilisierung des Proteins führen. Dieses Motiv wurde in den Modellen M4, M5, M7 und M8 definiert. Die Struktur von Top7 erlaubt die selektive Einführung einer Salzbrücke zwischen den beiden Helices an Position 35 und 71, deren Ausbildung ebenso zu einer Stabilisierung der jeweiligen Proteine führen könnte. Dieses Motiv wurde in den Modellen M1, M3, M8 (LYS35, GLU71) und M2 bzw. M7 (GLU35, LYS71) verwendet.

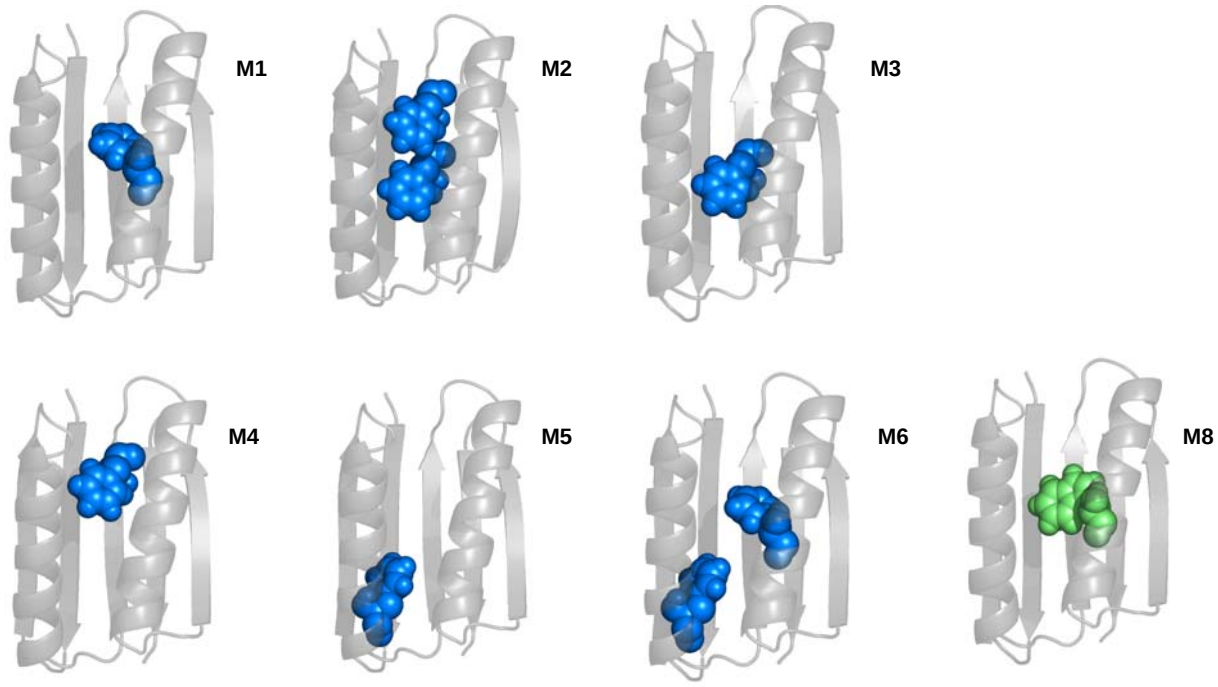


Abbildung IV-15 Sekundärstrukturdarstellungen der Zielstruktur mit den modellierten aromatischen Aminosäuren Phenylalanin und Tryptophan. Diese Aminosäuren wurden in den Modellen M1-M6 und M8 verwendet. Phenylalanin und Tryptophan sind als Kalottenmodelle in blauer bzw. grüner Farbe dargestellt. Die Seitenketten wurden mit dem Programm *SCWRL 3* [Canutescu *et al.*, 2003] modelliert und mit der Implementierung des GROMOS96 Kraftfeldes [van Gunsteren *et al.*, 1996] im *Swiss-PdbViewer3.7* (*SP5*) [Guex & Peitsch, 1997] energieminimiert. Die dargestellten Aminosäuren wurden als Bestandteil des hydrophoben Kerns betrachtet. **M1** Phenylalanin an Position 65, **M2** Phenylalanin an Position 48 und 50, **M3** Phenylalanin an Position 48, **M4** Phenylalanin an Position 50, **M5** Phenylalanin an Position 38, **M6** Phenylalanin an Positionen 38 und 65, **M8** Tryptophan an Position 65. Der hydrophobe Kern von Modell M7 besteht ausschließlich aus hydrophoben aliphatischen Aminosäuren. Die Abbildung wurde mit Hilfe des Programms *PyMOL* erstellt [Delano, 2002].

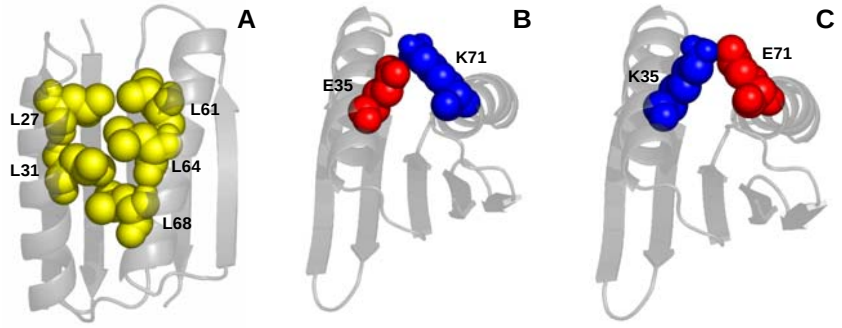


Abbildung IV-16 Sekundärstrukturdarstellungen der Zielstruktur mit definierten Motiven. Die Seitenketten wurden mit dem Programm *SCWRL 3* [Canutescu *et al.*, 2003] modelliert und mit der Implementierung des GROMOS96 Kraftfeldes [van Gunsteren *et al.*, 1996] des *Swiss-PdbViewers* 3.7 (*SP5*) [Guex & Peitsch, 1997] minimiert. Die gelben Kalottenmodelle zeigen die Aminosäure Leucin, rot dargestellt ist Glutaminsäure, blau gezeichnet ist Lysin. **A** Durch eine geeignete Positionierung von Leucin an den beiden Helices lässt sich ein reißverschlussähnliches Motiv erzeugen. Dieses Motiv wurde bei den Modellen M4, M5, M7 und M8 verwendet. **B** In den Modellen M2 und M7 wurde durch die Einführung von Glutaminsäure und Lysin auf den Positionen 35 bzw. 71 eine Salzbrücke definiert. **C** Die Modelle M1, M3 und M8 beinhalten auf der Position 35 und 71 ein Lysin bzw. eine Glutaminsäure. Die Positionierung dieser Aminosäuren führt ebenso zu der Ausbildung einer Salzbrücke. Die Abbildung wurde mit Hilfe des Programms *PyMOL* erstellt [Delano, 2002].

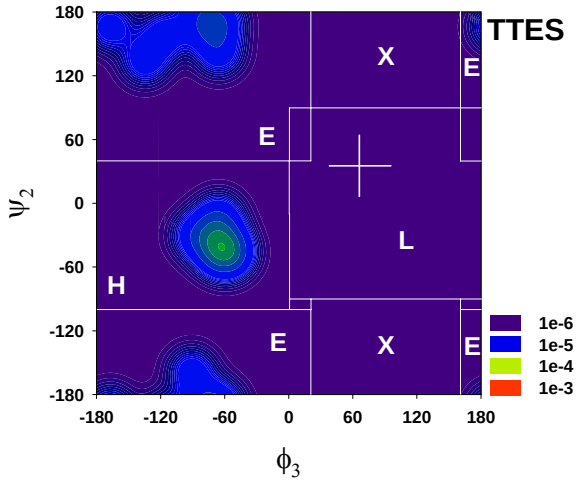


Abbildung IV-17 Dichtefunktion der ψ_2-φ_3-Verteilung des Tetrapeptids TTES. TTES beschreibt an Position 22 in der Top7-Struktur einen *L-turn* mit einer Wahrscheinlichkeit von $P(L) = 10^{-5}$. Das Fadenkreuz markiert die Zielkonformation mit dem Diederwinkelpaar (-36°,+68°). Die Wahrscheinlichkeit zur Ausbildung dieser Konformation ist nur sehr gering. Für die anderen Konformationszustände ergibt sich $P(E) = 0.55$, $P(H) = 0.45$, und $P(X) = 10^{-4}$. Die Dichtefunktion wurde aus 13 Wertepaaren errechnet.

Das Tetrapeptid TTES beschreibt in Top7 an Position 22 einen *L-turn* beim Übergang von einem β-Faltblatt in eine α-Helix. Aus der Struktur wurden für die Diederwinkel ψ_2 und φ_3 Werte von $\psi_2 = -36.93°$ und $\varphi_3 = +68.87°$ bestimmt. Konformationen vom Typ *L-turn* oder *X-turn* werden im Allgemeinen von Tetrapeptiden mit einem Glycin an dritter Position gebildet. In Abbildung IV-17 ist die Wahrscheinlichkeitsdichtefunktion für die Winkel ψ_2 und φ_3 des Tetrapeptids TTES dargestellt. Für die Ausbildung einer Konformation vom Typ *L* und vom Typ *X* ergeben sich formal Wahrscheinlichkeiten von $P(L) = 10^{-5}$ und $P(X) = 10^{-4}$. Im Vergleich dazu errechnen sich die Wahrscheinlichkeiten für eine faltblatttypische (*E*) und eine helixtypische Konformation (*H*) zu $P(E) = 0.55$ und $P(H) = 0.45$. Das vorliegende wissensbasierte System würde deshalb das Tetrapeptid TTES für eine Modellierung dieser Struktur nicht verwenden, da es keine *a priori* Wahrscheinlichkeiten verwendet. Es sollte überprüft werden, inwieweit sich die in der Struktur von Top7 gefundene Konformation des Tetrapeptids TTES reproduzieren lässt. Aus diesem Grund wurde in den Proteinmodellen M2 und M4 dieses Tetrapeptid ebenso

Tabelle IV-6 Übersicht über die verwendeten Tetrapeptide in den Modellen M1-M8 an Position 22. Der Zielkonformationszustand ist *L (L-turn)*. *P(E)*, *P(H)*, *P(L)* und *P(X)* beschreiben jeweils die Wahrscheinlichkeit für die Konformationszustände nach Tabelle II-1, S. 17. Die Sequenz von Top7 besitzt an dieser Position das Tetrapeptid TTES. Die Wahrscheinlichkeit für eine Konformation vom Typ *L* oder Typ *X* errechnete sich für dieses Tetrapeptid aus den korrespondierenden Dichtefunktionen zu 10^{-5} bzw. 10^{-4}. Bei der Modellierung der Proteine M1, M3, M5, M6, M7 und M8 wurden an dieser Position Tetrapeptide verwendet, deren wahrscheinlichster Konformationszustand der Typ *L* ist. Bei den Modellen M2 und M4 wurde das Tetrapeptid TTES für die Modellierung dieser Struktur verwendet. Die Spalte *Anzahl* listet die Anzahl der Wertepaare (ψ_2,φ_3) in den Datensätzen, aus denen die jeweilige Dichtefunktion errechnet wurde.

Modell	Tetrapeptid	*P(E)*	*P(H)*	*P(L)*	*P(X)*	Anzahl
Top7	TTES	0.55	0.45	10^{-5}	10^{-4}	13
M1	RTGE	0.16	0.17	0.50	0.17	18
M2	TTES	0.55	0.45	10^{-5}	10^{-4}	13
M3	TTGE	0.09	0.19	0.54	0.18	30
M4	TTES	0.55	0.45	10^{-5}	10^{-4}	13
M5	ETNE	0.22	0.33	0.45	10^{-5}	9
M6	KTRQ	0.32	0.20	0.40	0.08	10
M7	STGK	0.24	0.13	0.56	0.07	32
M8	ETGE	0.09	0.09	0.75	0.07	34

für die Modellierung dieser lokalen Struktur verwendet. Die Tabelle IV-6, S. 61, zeigt die Zusammenstellung der Tetrapeptide, die bei den Modellen M1-M8 bei der Modellierung dieses *loops* verwendet wurden. Als Ergebnis der Definition von Aminosäuren auf bestimmten Positionen innerhalb einer Proteinstruktur, erhält man Aminosäuresequenzen mit Lücken zwischen den einzelnen Aminosäuren, die im nächsten Schritt mit geeigneten Tetrapeptiden aufgefüllt werden müssen.

IV.2.3.3. Design von Sequenzen für die Top7-Topologie

Das vorgestellte Modellierungsschema mit Hilfe von Dichtefunktionen der ψ_2-φ_3-Verteilungen von Tetrapeptiden versucht, die Zielkonformation $(\psi_2,\varphi_3)_i$ des *i*-ten Tetrapeptids der Zielstruktur bestmöglich nachzubilden. Hierfür wurde die Struktur von Top7 (*PDB-Code* 1QYS) in ihre Tetrapeptide zerlegt und deren Konformationsdaten in einer Datei gespeichert, die als *template* für die Modellierung verwendet wurde. Im Anhang VII.1, S. 110, sind diese Konformationsdaten aufgeführt. Die Selektion der einzelnen Tetrapeptide erfolgte nach den unten genannten Kriterien, bei deren Nichterfüllung die Modellierung abgebrochen oder bei beendeter Modellierung die Aminosäuresequenz verworfen wurde.

i. Die Wahrscheinlichkeit des selektierten Tetrapeptids, den Zielkonformationszustand (*E*, *H*, *L* oder *X*) anzunehmen, muß mindestens 0.7 sein.
ii. Die relative Wahrscheinlichkeit der Zielkonformation (ψ_2,φ_3) bezüglich der wahrscheinlichsten Konformation im Zielkonformationsbereich muß mindestens 0.01 sein.
iii. Das arithmetische Mittel der Wahrscheinlichkeit der einzelnen Tetrapeptide für den Zielkonformationszustand muss für die Gesamtsequenz mindestens 0.7 sein. Es wird diejenige Sequenz als Ergebnis selektiert, deren Wahrscheinlichkeit maximal ist. Es wurde das arithmetische anstatt des geometrischen Mittels gewählt, da bei der Berechnung der Zielsequenzen Tetrapeptide als Randbedingungen verwendet wurden, die nur eine sehr niedrige Wahrscheinlichkeit für den Zielkonformationszustand besitzen. Auf diese Weise konnte auf die Einführung einer *a priori* Wahrscheinlichkeit als Mindestwahrscheinlichkeit für einen Konformationszustand verzichtet werden.
iv. Die Sequenzidentität zur Sequenz von Top7 muss kleiner als 30 % sein.

Der Punkt (i) lässt sich formal als notwendige Bedingung für (ii) interpretieren. Das selektierte Tetrapeptid muss in der Lage sein, die Zielkonformation ausbilden zu können. Dies wird durch die Bedingung (ii) sichergestellt. Erste Modellierungsversuche hatten gezeigt, dass diese Bedingung nicht in jedem Fall erfüllt werden konnte, was in der Folge zu einem Abbruch der Modellierung führte. In diesen Fällen wurde, ausgehend von Bedingung (i), dasjenige Tetrapeptid selektiert, dessen Winkeldifferenz $|\Delta\psi_2|$ und $|\Delta\varphi_3|$ von der Zielkonformation zum nächstliegenden Peak im Zielkonformationsbereich minimal ist. Als Maß für die Winkeldifferenz wurde die Euklidische Distanz von der Zielkonformation zum nächstliegenden Peak verwendet. In der Tabelle IV-9, S. 75, sind für alle Modelle die Positionen angegeben, bei denen die Bedingung (ii) nicht erfüllt werden konnte. Für die einzelnen Tetrapeptide der Sequenz von Top7 errechnete sich

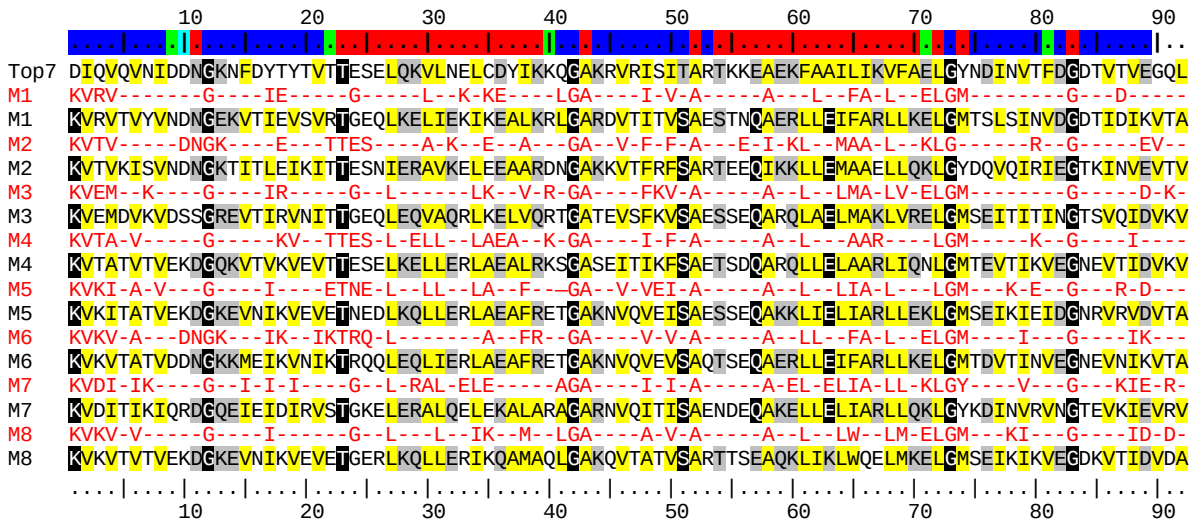


Abbildung IV-18 Errechnete Sequenzen der Modelle M1 bis M8 und Sequenz von Top7. Die rot markierten Aminosäuren wurden als Randbedingungen definiert. Die Lücken wurden während der Modellierung mit geeigneten Tetrapeptiden aufgefüllt. Der farbige Balken am oberen Rand der Abbildung markiert den Zielkonformationszustand ***E*** (faltblatttypisch), ***H*** (helixtypisch), ***L*** (*L-turn*) und ***X*** (*X-turn*) der einzelnen Tetrapeptide. Die Grenzen der Konformationsbereiche sind in Tabelle II-1, S. 17, definiert. Es ist immer nur die erste Aminosäure eines Tetrapeptids farblich kodiert. Dies bedeutet beispielsweise, dass das Tetrapeptid DNGK aus Top7 (Nr. 10) den Zielkonformationszustand *X-turn* ausbildet. Die gelb markierten Aminosäuren kennzeichnen die als hydrophob definierten Aminosäuren (vgl. Abschnitt IV.2.3.1, S. 58).

eine mittlere Wahrscheinlichkeit für den Zielkonformationsbereich von 0.66 ± 0.23 (siehe IV.2.3.4.1). Bei der Modellierung der Alternativsequenzen zu Top7 sollte dieser Wert verbessert werden, weshalb nach Bedingung (iii) eine höhere mittlere Wahrscheinlichkeit für den Zielkonformationsbereich von mindestens 0.7 gefordert wurde. Im Ergebnis sollte diejenige Sequenz selektiert werden, deren mittlere Wahrscheinlichkeit für den Zielkonformationsbereich maximal ist. Das Hauptproblem bei der Modellierung besteht darin, dass aufgrund der sequenziellen Berechnung der Zielsequenzen nicht die Möglichkeit besteht, ungünstige räumliche Verteilungen von Aminosäuren bereits während der Verlängerung der Aminosäuresequenzen zu erkennen. Um den möglichen Sequenzraum einzuschränken, fehlerhafte Modelle infolge Überlappungen von Seitenketten auszuschließen und eine optimale Verteilung von Aminosäuren zu gewährleisten, wurden bei den einzelnen Modellen, anhand der Struktur von Top7, eine unterschiedliche Anzahl an Aminosäuren als Randbedingung definiert. In Abbildung IV-18 sind die errechneten Aminosäuresequenzen und die als Randbedingungen eingefügten Aminosäuren dargestellt. Diese Sequenzen stellen unter diesen Bedingungen die jeweils beste gefundene Lösung dar.

Die Tabelle IV-7, S. 64, zeigt die gegenseitigen Gesamtsequenzidentitäten, die Sequenzidentitäten der hydrophoben Kerne und die Ähnlichkeit der Modelle untereinander und zu Top7. In allen Fällen wurde eine Sequenzidentität von kleiner als 30 % zu Top7 erreicht. Die Sequenzidentitäten der hydropohoben Kerne wurden mit Hilfe der zugänglichen Oberfläche (*ASA, accessible surface area*) errechnet. In Analogie zu Kurochkina & Privalov wurden diejenigen Aminosäuren als begraben definiert, deren zugängliche Oberfläche kleiner als 5 % gegenüber der vollständig zugänglichen Oberfläche dieser Aminosäure war [Kurochkina & Privalov, 1998] und die in Abschnitt IV.2.3.1, S. 58, als hydrophob klassifiziert wurden. Die Abbildung IV-19 auf Seite 64 markiert die nach dieser Definition begrabenen Aminosäuren in den Sequenzen der Modelle M1 bis M8 und von Top7. Die Sequenzidentitäten der einzelnen Modelle zu Top7 bezüglich dieser Aminosäuren sind mehrheitlich kleiner als 50 %, was darauf hindeutet, dass die

Gesamtidentität nicht nur auf eine Variation der lösungsmittelexponierten Aminosäuren zurückzuführen ist. Gleichzeitig wird damit eine Restrukturierung der hydrophoben Kerne angezeigt. Die Modelle weisen bezüglich der BLOSUM62-Matrix Ähnlichkeiten von durchschnittlich 60 % auf.

Tabelle IV-7 Gesamtsequenzidentität, Sequenzidentität der hydrophoben Kerne und Ähnlichkeit der Modelle untereinander und zu Top7. In der oberen Zeile ist jeweils die Gesamtsequenzidentität verzeichnet. Die zweite Zeile beschreibt die Sequenzidentität im hydrophoben Kern. Es wurden alle Aminosäuren als begraben definiert, deren lösungsmittelzugängliche Oberfläche kleiner als 5 % im Vergleich zu der jeweiligen vollständig exponierten Einzelaminosäure ist. Die Sequenzidentitäten zwischen Top7 und den Modellen beziehen sich auf die begrabenen Aminosäuren von Top7. Beim Vergleich der Modelle untereinander wurde immer das Modell mit dem größeren Index mit dem Modell mit dem kleineren Index verglichen. Die lösungsmittelzugängliche Oberfläche (*ASA*) wurde mit dem Programm *CCP4* bestimmt (http://www.ccp4.ac.uk/index.php). Die untere Zeile zeigt die Ähnlichkeit bezüglich der BLOSUM62 Matrix [Henikoff & Henikoff, 1992].

Modell	Top7	M1	M2	M3	M4	M5	M6	M7	M8
Top7	100 100 100	26 45 60	27 30 59	24 50 59	25 45 54	24 50 58	26 40 58	28 55 59	28 75 57
M1	26 45 60	100 100 100	36 42 65	42 55 74	48 63 74	45 53 75	58 72 86	39 58 73	50 75 79
M2	27 30 59	36 42 65	100 100 100	34 47 65	40 39 73	37 44 66	42 52 66	42 47 67	30 37 67
M3	24 50 59	42 55 74	34 47 65	100 100 100	48 53 76	48 47 71	46 68 71	40 47 73	42 60 71
M4	25 45 54	48 63 74	40 39 73	48 53 76	100 100 100	54 53 79	52 59 76	42 58 70	51 60 75
M5	24 50 58	45 53 75	37 44 66	48 47 71	54 53 79	100 100 100	61 77 88	46 68 73	60 55 80
M6	26 40 58	58 72 86	42 52 66	46 63 71	52 59 76	61 77 88	100 100 100	40 53 67	47 65 77
M7	28 55 59	39 58 73	42 47 67	40 47 73	42 58 70	46 68 73	40 53 67	100 100 100	35 47 70
M8	28 45 57	50 75 79	30 37 67	42 60 71	51 60 75	60 55 80	47 65 77	35 47 70	100 100 100

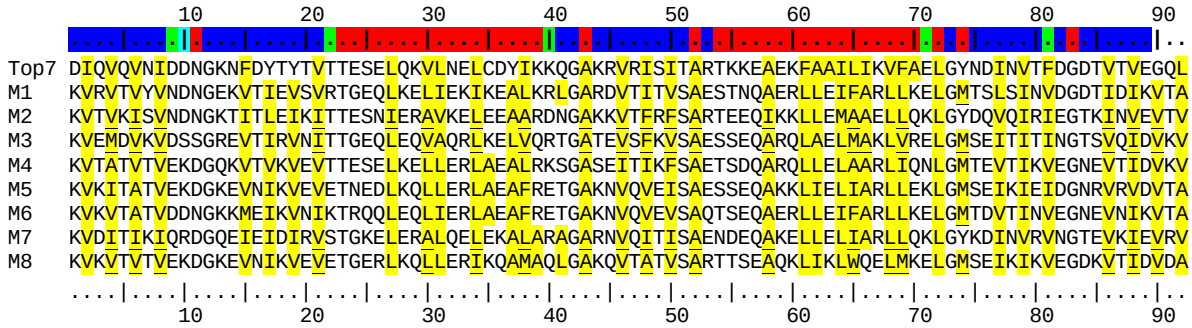


Abbildung IV-19 Begrabene hydrophobe Aminosäuren der Modelle M1 bis M8 und von Top7. Die gelb markierten Aminosäuren wurden in Abschnitt IV.2.3.1, S. 58, als hydrophob klassifiziert. Die unterstrichenen hydrophoben Aminosäuren haben eine zugängliche Oberfläche von kleiner als 5 % gegenüber der vollständig zugänglichen Aminosäure und sind als begraben definiert [Kurochkina & Privalov, 1998]. Die zugänglichen Oberflächen wurden mit dem Programmpaket *CCP4* (http://www.ccp4.ac.uk/index.php) errechnet.

IV.2.3.4. Analyse der Modelle

IV.2.3.4.1. Analyse der Sequenz von Top7

Es sollte untersucht werden, wie gut sich die Struktur von Top7 durch die Dichtefunktionen der ψ_2-φ_3-Verteilungen der entsprechenden Tetrapeptide beschreiben lässt, um mögliche Ergebniskorrelationen zwischen *RosettaDesign* und dem in dieser Arbeit vorgestellten Algorithmus aufzudecken. Die Dichtefunktionen enthielten keine Informationen über die Struktur von Top7. In Abbildung IV-20 sind für jedes der 89 Tetrapeptide der Top7-Sequenz die Wahrscheinlichkeiten dargestellt, eine Konformation im Bereich *E*, *H*, *L* und *X* anzunehmen. Die schwarze Linie kennzeichnet den Konformationszustand, der in der Struktur von Top7 gefunden wurde (Zielkonformationszustand). Zu erkennen ist, dass für eine große Anzahl an Tetrapeptiden der jeweilige Zielkonformationszustand gleichzeitig auch der wahrscheinlichste ist. Bei 18 Tetrapeptiden wird der in der Struktur beobachtete Konformationszustand nicht durch den wahrscheinlichsten Konformationszustand des jeweiligen Tetrapeptids beschrieben. Dies betrifft NIDD (Nr. 7), DDNG (Nr. 9), DNGK (Nr. 10), KNFD (Nr. 13), TTES (Nr. 22), QGAK (Nr. 41), GAKR (Nr. 42), KRVR (Nr. 44), ITAR (Nr. 50), TART (Nr. 51), RTKK (Nr. 53), AILI (Nr. 63), LIKV (Nr. 65), IKVF (Nr. 66), DGDT (Nr. 82), GDTV (Nr. 83), VEGQ (Nr. 88), EGQL (Nr. 89). Besonders in den Übergängen zu den irregulären Strukturbereichen bzw. in den nichtregulären Bereichen kommt es zu einer Verringerung der Wahrscheinlichkeit für den erforderlichen Konformationszustand. Trotzdem kann in einem solchen Fall die jeweilige Zielkonformationen (ψ_2,φ_3) sehr gut beschrieben werden, wie am Beispiel des Tetrapeptids DDNG (Nr. 9) deutlich wird. DDNG nimmt in Top7 den Konformationszustand *L-turn* an, der eine Wahrscheinlichkeit von *P(L)* = 0.16 besitzt. Die in der Struktur beobachtete Konformation wird

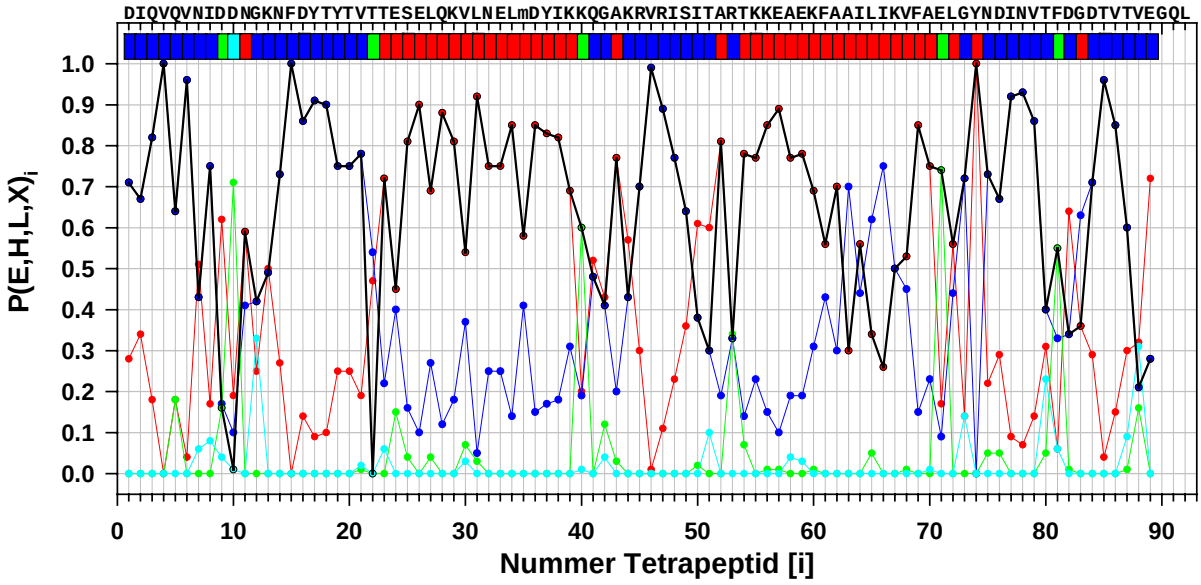


Abbildung IV-20 Wahrscheinlichkeit *P* der Tetrapeptide aus der Sequenz von Top7 für ihren Zielkonformationszustand. Die Grenzen der Konformationsbereiche [■ (faltblatttypisch), ■ (helixtypisch), ■ (*L-turn*) oder ■ (*X-turn*)] sind in Tabelle II-1, S. 17, definiert. Die Farbkodierung der ersten Aminosäure eines Tetrapeptids beschreibt dessen Zielkonformationszustand. Die schwarze Linie kennzeichnet den Zielkonformationszustand aus der Struktur von Top7. Bei 18 Tetrapeptiden ist dieser Konformationszustand nicht der wahrscheinlichste. Die Sequenz von Top7 besitzt an Nummer 35 ein Selenomethionin, dargestellt durch ein „m“. Selenomethionin wurde bei den vier Tetrapeptiden NELm, ELmD, LmDY und mDYI bei der Errechnung der Dichtefunktionen durch Methionin ersetzt. Die angegebenen Wahrscheinlichkeiten beziehen sich daher auf die entsprechenden Tetrapeptide mit einem Methionin.

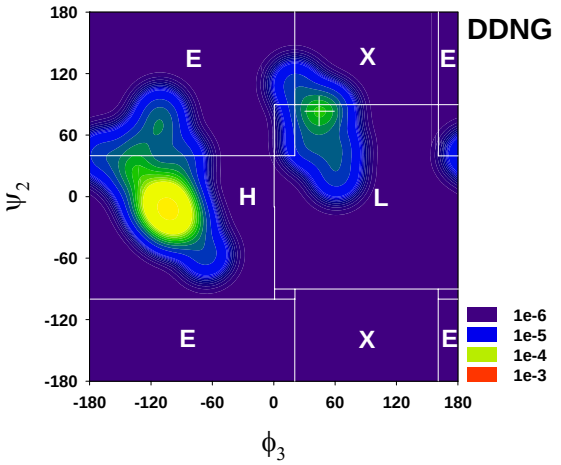


Abbildung IV-21 Dichtefunktion der ψ_2-φ_3-Verteilung des Tetrapeptids DDNG. DDNG beschreibt an Position 9 in der Struktur von Top7 einen *L-turn* mit einer Wahrscheinlichkeit von $P(L) = 0.16$. Das weiße Fadenkreuz markiert das in der Struktur beobachtete Diederwinkelpaar (ψ_2,φ_3) mit $\psi_2 = +80.2°$ und $\varphi_3 = +43.8°$. Die wahrscheinlichste Konformation im Konformationsbereich *L* wird durch das Diederwinkelpaar (ψ_2,φ_3) mit $\psi_2 = +82°$ und $\varphi_3 = +46°$ beschrieben, zu dem sich eine Winkeldifferenz von $|\Delta\psi_2| = 2°$ und $|\Delta\varphi_3| = 2°$ errechnet. Die Konformation von DDNG in Top7 wird sehr gut durch die Dichtefunktion beschrieben.

durch das Diederwinkelpaar (ψ_2,φ_3) mit $\psi_2 = +80.2°$ und $\varphi_3 = +43.8°$ beschrieben. Die Abbildung IV-21 zeigt die Dichtefunktion der ψ_2-φ_3-Verteilung. Die in der Struktur beobachtete Winkelabweichung zur wahrscheinlichsten Konformation im Konformationsbereich *L* zeigt nur sehr geringe Werte von $|\Delta\psi_2| = 2°$ und $|\Delta\varphi_3| = 2°$.

Es sollte an einzelnen Beispielen untersucht werden, wie groß die beobachtete Winkelabweichung von der wahrscheinlichsten Konformation im Zielkonformationsbereich ist. Das Ergebnis ist in Abbildung IV-23 auf Seite 68 dargestellt. Bei den Tetrapeptiden IDDN (Nr. 8), DNGK (Nr. 10) und VTTE (Nr. 21) sind sehr große Abweichungen zu erkennen. In Abbildung IV-22 auf Seite 67 sind die Dichtefunktionen der ψ_2-φ_3-Verteilungen dieser Tetrapeptide dargestellt. Das Tetrapeptid IDDN nimmt in der Struktur von Top7 eine faltblatttypische Konformation mit einer Wahrscheinlichkeit von $P(E) = 0.75$ an. Es wurde ein Diederwinkelpaar von $\psi_2 = +143.85°$ und $\varphi_3 = -168.8$ ausgebildet. Die global wahrscheinlichste Konformation liegt ebenso im Zielkonformationsbereich und wird durch die Diederwinkel $\psi_2 = -176°$ und $\varphi_3 = -60°$ beschrieben. Es ergeben sich große Abweichungen der beobachteten Konformation zu dieser Konformation von $|\Delta\psi_2| = 40°$ und $|\Delta\varphi_3| = 108°$. Diese Abweichungen führen bei diesem Tetrapeptid in der Folge zu einer sehr geringen Wahrscheinlichkeit für diese Konformation (siehe Abbildung IV-24 auf Seite 68). Das Tetrapeptid DNGK beschreibt in der Struktur von Top7 einen *X-turn* mit einem Diederwinkel $\psi_2 = +102.5°$ und $\varphi_3 = +56.5$ (siehe Abbildung IV-22 auf Seite 67). Die Wahrscheinlichkeit für eine Konformation vom Typ *X-turn* errechnet sich zu $P(X) = 0.01$. Diese Wahrscheinlichkeit wird durch das Diederwinkelpaar $\psi_2 = -80°$ und $\varphi_3 = +146$ verursacht, das in dem Konformationsbereich *L* liegt, aber in den *X-turn* Wahrscheinlichkeitsdichte hineinstreut. Dieses Diederwinkelpaar ist in Abbildung IV-22 durch einen weißen Kreis markiert. Der Datensatz zur Berechnung der ψ_2-φ_3-Dichtefunktion beinhaltete keine Diederwinkelpaare, die im *X-turn* Bereich liegen, so dass keine Winkelabweichungen $|\Delta\psi_2|$ und $|\Delta\varphi_3|$ zu der wahrscheinlichsten Konformation in diesem Konformationsbereich angegeben werden können. Aus diesem Grund wurde die Winkeldifferenz zu dem Diederwinkelpaar im Konformationsbereich *L* errechnet, welches Wahrscheinlichkeitsdichte in den Konformationsbereich von *X* hineinstreut. Diese Winkeldifferenz errechnet sich zu $|\Delta\psi_2| = 178°$ und $|\Delta\varphi_3| = 90°$. Die Ursache dieser formal sehr großen Winkelabweichungen liegt in der Einteilung der Konformationsbereiche *X* und *L*. In der Abbildung IV-22 ist zu erkennen, dass das in der Struktur von Top7 gefundene Diederwinkelpaar für DNGK gut durch eine Konformation vom Typ *L-turn* beschrieben werden könnte, deren Wahrscheinlichkeit $P(L) = 0.71$ beträgt.

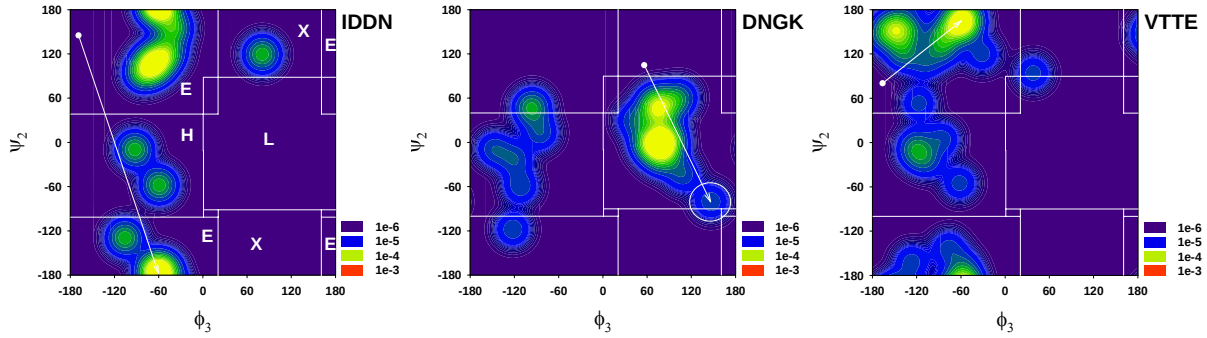


Abbildung IV-22 Dichtefunktionen der ψ_2-φ_3-Verteilungen von Tetrapeptiden aus der Sequenz von Top7. Die Pfeilspitze zeigt auf die wahrscheinlichste Konformation des entsprechenden Tetrapeptids in dem Zielkonformationszustand. Der Startpunkt des Pfeils markiert das in der Struktur gefundene Diederwinkelpaar. Die Tetrapeptide zeigen nur eine geringe Wahrscheinlichkeit für die Ausbildung des in der Struktur beobachteten Diederwinkelpaares (ψ_2,φ_3). **IDDN** (Nr. 8) Die Wahrscheinlichkeit für den faltblatttypischen Konformationszustand ist *P(E)* = 0.75 mit einer wahrscheinlichsten Konformation von $\psi_2 = -176°$ und $\varphi_3 = -60°$. Die Zielkonformation ist $\psi_2 = +143.81°$ und $\varphi_3 = -168.83°$. Die Winkeldifferenz errechnet sich zu $|\Delta\psi_2| = 40°$ und $|\Delta\varphi_3| = 108°$. **DNGK** (Nr. 10) Die Wahrscheinlichkeit für die Konformation Typ *X* beträgt *P(X)* = 0.01 und wird durch die Streuung des Diederwinkelpaares $\psi_2 = -80°$ und $\varphi_3 = +146°$ aus dem Konformationsbereich *L* verursacht. Die Zielkonformation ist $\psi_2 = +102.48°$ und $\varphi_3 = +56.52°$. Die Winkelabweichung zu diesem Diederwinkelpaar beträgt $|\Delta\psi_2| = 178°$ und $|\Delta\varphi_3| = 90°$. **VTTE** (Nr. 21) Die Wahrscheinlichkeit für den faltblatttypischen Konformationszustand ist *P(E)* = 0.78 mit einer wahrscheinlichsten Konformation für diesen Zustand von $\psi_2 = +162°$ und $\varphi_3 = -60°$. Die Zielkonformation ist $\psi_2 = +80.99°$ und $\varphi_3 = -176.77°$. Die Winkeldifferenz errechnet sich zu $|\Delta\psi_2| = 82°$ und $|\Delta\varphi_3| = 116°$.

Die global wahrscheinlichste Konformation liegt mit einem Diederwinkelpaar von $\psi_2 = -2°$ und $\varphi_3 = +78$ ebenso in diesem Konformationsbereich. Das Tetrapeptid VTTE nimmt in der Struktur von Top7 eine faltblatttypische Konformation mit einer Wahrscheinlichkeit von *P(E)* = 0.78 an. Es wurde ein Diederwinkelpaar (ψ_2,φ_3) von $\psi_2 = +80.9°$ und $\varphi_3 = -176.7°$ gebildet. Die global wahrscheinlichste Konformation befindet sich ebenso im Zielkonformationsbereich und beschreibt die Diederwinkel $\psi_2 = +162°$ und $\varphi_3 = -60°$. Auch hier ergibt sich eine große Winkeldifferenz von $|\Delta\psi_2| = 82°$ und $|\Delta\varphi_3| = 116°$. Diese große Abweichung verringert sich auf $|\Delta\psi_2| = 26°$ und $|\Delta\varphi_3| = 58°$, wenn man die Abweichung zum nächstgelegenen Peak berechnet. Für das Tetrapeptid TTES konnte keine Winkelabweichung angegeben werden, da der Konformationszustand *L* unter Berücksichtigung der vorhandenen Daten nicht erlaubt ist (vgl. Abbildung IV-17 auf Seite 61). Die Abbildung IV-23 auf Seite 68 zeigt für alle Tetrapeptide der Top7-Sequenz die Winkelabweichung $|\Delta\psi_2|$ und $|\Delta\varphi_3|$ der beobachteten Konformation zu der wahrscheinlichsten Konformation im Zielkonformationsbereich. Es sollte im Folgenden untersucht werden, inwieweit diese Winkelabweichungen einen Einfluss auf die absolute Wahrscheinlichkeit der beobachteten Konformation der einzelnen Tetrapeptide haben. Die Abbildung IV-24 auf Seite 68 zeigt, dass im Allgemeinen mit zunehmender Differenz $|\Delta\psi_2|$ und $|\Delta\varphi_3|$ die Wahrscheinlichkeit für die beobachtete Konformation abnimmt. Die violett umrandeten schwarzen Punkte markieren die absolute Wahrscheinlichkeit der wahrscheinlichsten Konformation. Die schwarzen Punkte kennzeichnen die absolute Wahrscheinlichkeit der wahrscheinlichsten Konformation der Tetrapeptide im Zielkonformationsbereich. Bei 19 der 89 Tetrapeptide liegt die global wahrscheinlichste Konformation nicht im Zielkonformationsbereich. Dies betrifft, bis auf das Tetrapeptid RTKK (Nr. 53), zusätzlich zu den verbleibenden 17 Tetrapeptiden, bei denen der wahrscheinlichste Konformationszustand nicht der Zielkonformationszustand ist, die beiden Tetrapeptide TFDG (Nr. 80) und TVEG (Nr. 87). Die weißen Punkte markieren die absolute Wahrscheinlichkeit für das beobachtete Diederwinkelpaar (ψ_2,φ_3).

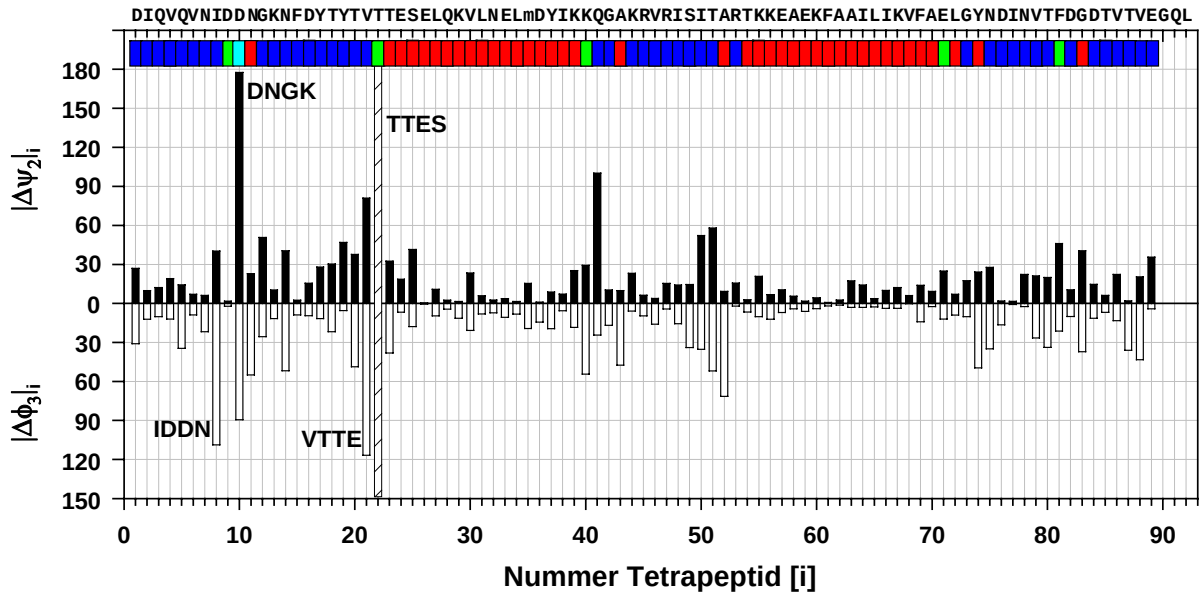


Abbildung IV-23 Winkelabweichungen $|\Delta\psi_2|$ und $|\Delta\varphi_3|$ der Tetrapeptide aus der Sequenz von Top7 von der wahrscheinlichsten Konformation im Zielkonformationsbereich. Die Grenzen der Konformationsbereiche [■ (faltblatttypisch), ■ (helixtypisch), ■ (*L-turn*) oder ■ (*X-turn*)] sind in Tabelle II-1, S. 17, definiert. Die Farbkodierung der ersten Aminosäure eines Tetrapeptids beschreibt dessen Konformationszustand. Für das Tetrapeptid TTES (Nr. 22) kann keine Winkelabweichung angegeben werden, da dieser Konformationszustand nicht erlaubt ist (Abbildung IV-17 auf Seite 61). Die Diederwinkel φ_3 bei dem Tetrapeptid IDDN (Nummer 8), ψ_2 und φ_3 bei DNGK (Nummer 10) und ψ_2 bzw. φ_3 bei VTTE (Nummer 21) zeigen besonders hohe Abweichungen. Die Dichtefunktionen der ψ_2-φ_3-Verteilungen für diese Tetrapeptide sind in Abbildung IV-22 auf Seite 67 dargestellt.

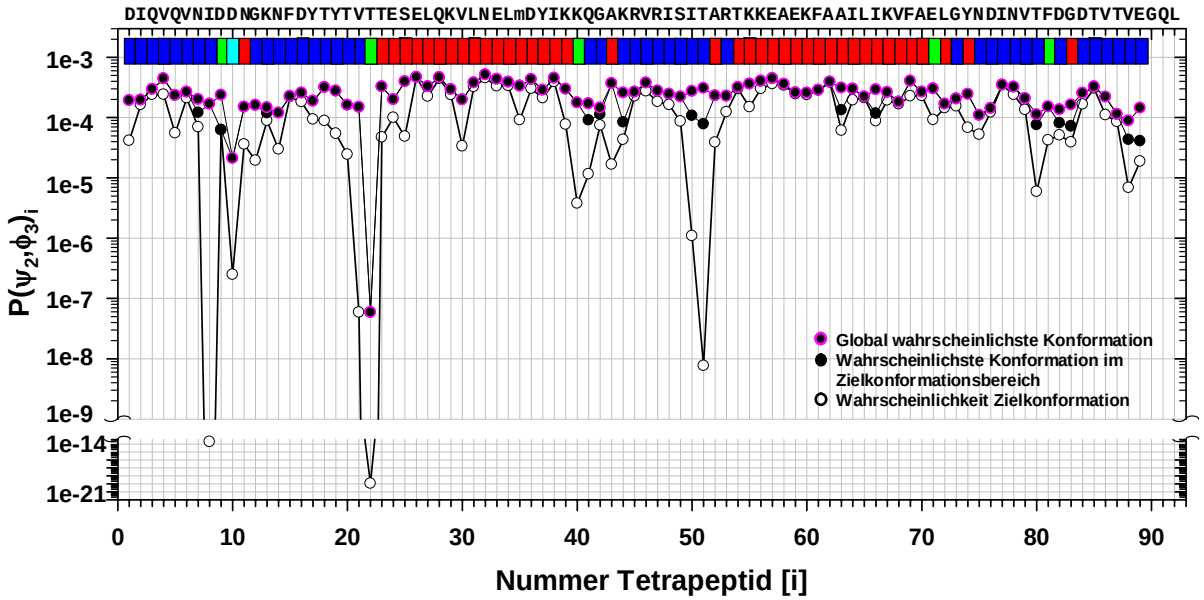


Abbildung IV-24 Absolute Wahrscheinlichkeit *P* für die beobachteten Diederwinkelpaare (ψ_2,φ_3) in der Struktur von Top7. Die Grenzen der Konformationsbereiche[■ (faltblatttypisch), ■ (helixtypisch), ■ (*L-turn*) oder ■ (*X-turn*)] sind in Tabelle II-1, S. 17, definiert. Die Farbkodierung der ersten Aminosäure eines Tetrapeptids beschreibt dessen Konformationszustand. Der Vergleich mit Abbildung IV-23 zeigt, dass die Größe der Winkelabweichungen $|\Delta\psi_2|$ und $|\Delta\varphi_3|$ der beobachteten Konformationen von der wahrscheinlichsten Konformation im Zielkonformationsbereich im Allgemeinen mit der Verringerung der Wahrscheinlichkeit für die in der Struktur beobachteten Konformationen korreliert. Die wahrscheinlichste Konformation im Zielkonformationsbereich bezieht sich bei DNGK (Nr. 10) auf das Diederwinkelpaar (ψ_2,φ_3) mit $\psi_2 = -80°$ und $\varphi_3 = +146°$ im Konformationsbereich *L* (vgl. Abbildung IV-22 auf Seite 67). Das Tetrapeptid TTES besitzt keine wahrscheinlichste Konformation im Zielkonformationsbereich. Aus diesem Grund wurde diese Konformation mit der global wahrscheinlichsten gleichgesetzt.

Dieses Ergebnis unterstreicht den Befund, dass ein Tetrapeptid im Allgemeinen nicht alle erlaubten Konformationen annimmt und die Wahrscheinlichkeitsdichte auf bestimmte Konformationsbereiche beschränkt ist.

Die Analyse der Sequenz von Top7 hat gezeigt, dass die in der vorliegenden Arbeit vorgestellten Wahrscheinlichkeitsdichtefunktionen der ψ_2-φ_3-Verteilungen von Tetrapeptiden die Struktur von Top7 bezüglich des wahrscheinlichsten Konformationszustandes mit einigen Ausnahmen gut beschreiben können.

IV.2.3.4.2. Analyse der Sequenz von Modell M7

Die Abbildung IV-25 zeigt die Wahrscheinlichkeiten der einzelnen Tetrapeptide der Sequenz von Modell M7, den Zielkonformationszustand anzunehmen. Die Sequenz von M7 zeigt hierfür im Vergleich zu der Sequenz von Top7 deutlich höhere Wahrscheinlichkeiten. Das arithmetische Mittel für die Wahrscheinlichkeit der Zielkonformationszustände der einzelnen Tetrapeptide errechnete sich zu 0.75 ± 0.18. Bei sieben Tetrapeptiden entspricht der Zielkonformationszustand nicht dem wahrscheinlichsten Konformationszustand des verwendeten Tetrapeptids. Bei der Sequenz von M7 beobachtet man in Analogie zu der Sequenz von Top7 ein Abfallen der Wahrscheinlichkeiten für den Ziekonformationszustand in den Übergängen zu den irregulären Bereichen bzw. in den irregulären Bereichen, welcher jedoch nicht so stark ausgeprägt ist. Im Besonderen findet man dieses Verhalten bei den Tetrapeptiden RDGQ (Nummer 10, *X-turn*)

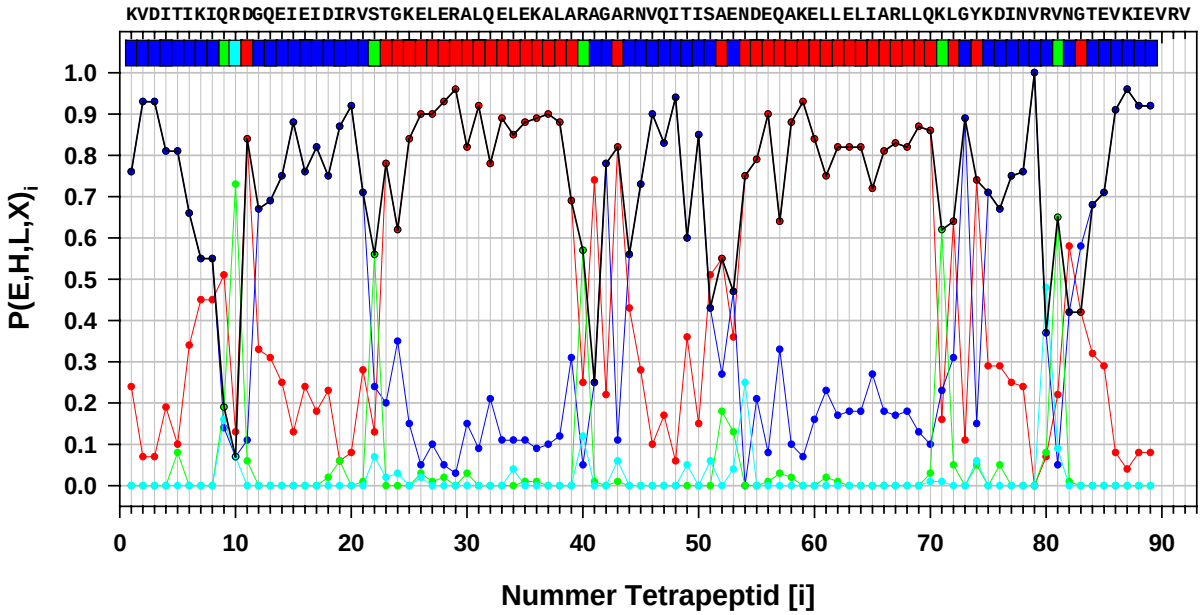


Abbildung IV-25 Wahrscheinlichkeit *P* der Tetrapeptide von Modell M7 für den jeweiligen Zielkonformationszustand. Der Zielkonformationszustand wird von der Struktur von Top7 vorgegeben. Die Grenzen der Konformationsbereiche [■ (faltblatttypisch), ■ (helixtypisch), ■ (*L-turn*) oder ■ (*X-turn*)] sind in Tabelle II-1, S. 17, definiert. Die Farbkodierung der ersten Aminosäure eines Tetrapeptids beschreibt dessen Zielkonformationszustand. Die schwarze Linie kennzeichnet den Zielkonformationszustand. Zu erkennen ist eine sehr gute Übereinstimmung der wahrscheinlichsten Konformationszustände mit den Zielkonformationszuständen. Die Tetrapeptide RDGQ (Nr. 10, *P(X)* = 0.07) und AGAR (Nr. 41, *P(E)* = 0.25) zeigen eine geringe Wahrscheinlichkeit für den Zielkonformationszustand. Die Dichtefunktionen der ψ_2-φ_3-Verteilungen für RDGQ und AGAR sind in Abbildung IV-26 auf Seite 70 dargestellt. Bei den sieben Tetrapeptiden QRDG (Nr. 9, *P(L)* = 0.19), RDGQ (Nr. 10), AGAR (Nr. 41), SAEN (Nr. 51, *P(E)* = 0.43), RVNG (Nr. 80, *P(E)* = 0.37), NGTE (Nr. 82, *P(E)* = 0.42) und GTEV (Nr. 83, *P(H)* = 0.38) ist der wahrscheinlichste Konformationszustand nicht der Zielkonformationszustand.

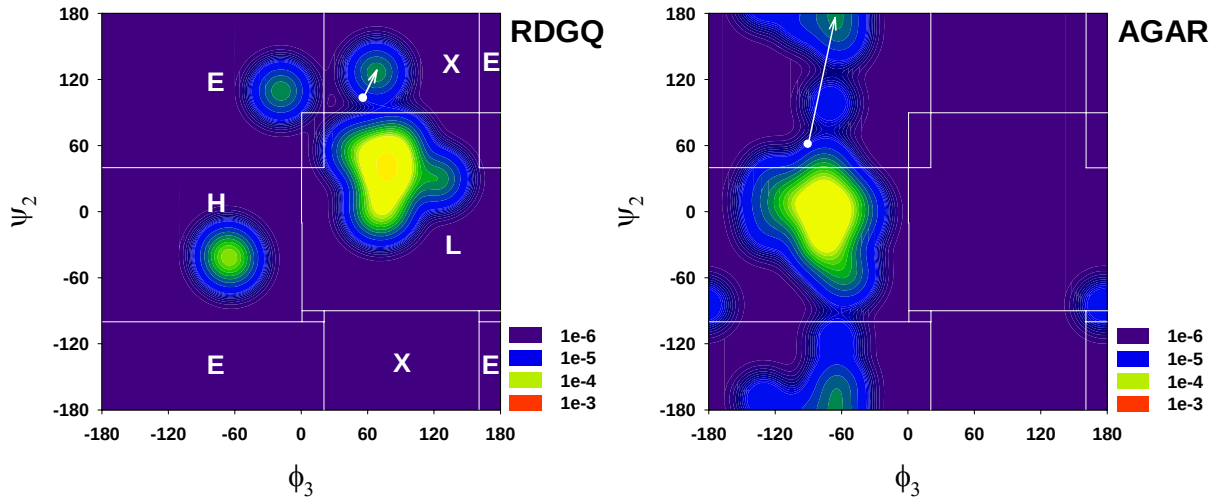


Abbildung IV-26 Dichtefunktionen der ψ_2-φ_3-Verteilungen der Tetrapeptide RDGQ (Nr. 10) und AGAR (Nr. 41) aus der Sequenz von Modell M7. Die Pfeilspitze zeigt auf die wahrscheinlichste Konformation des Zielkonformationszustandes. Der Startpunkt des Pfeils markiert die Zielkonformation. **RDGQ** Die Wahrscheinlichkeit für eine Konformation vom Typ *X-turn* beträgt $P(X) = 0.07$ und für eine Konformation vom Typ *L-turn* $P(L) = 0.73$. **AGAR** Die Wahrscheinlichkeit für eine faltblatttypische Konformation beträgt $P(E) = 0.25$ und für eine helixtypische Konformation $P(H) = 0.74$. Die Wahrscheinlichkeit der Zielkonformation resultiert aus der Streuung von Wahrscheinlichkeitsdichte aus dem helixtypischen Konformationsbereich. Die Dichtefunktionen zeigen, dass in diesen beiden Fällen die formal geringen Wahrscheinlichkeiten für die Zielkonformationszustände aus der Einteilung in die Konformationsbereiche nach Tabelle II-1, S. 17, resultieren.

und AGAR (Nummer 41, faltblatttypisch *E*). RDGQ besitzt eine Wahrscheinlichkeit für den Konformationszustand *X-turn* von $P(X) = 0.07$ und für den *L-turn* von $P(L) = 0.73$. Die Dichtefunktion in Abbildung IV-26 zeigt aber auch in diesem Fall, dass die anscheinend geringe Wahrscheinlichkeit für den Zielkonformationszustand ihre Ursache in der verwendeten Klassifizierung in die vier Konformationsbereiche hat. Das Tetrapeptid AGAR zeigt eine Wahrscheinlichkeit von $P(E) = 0.25$ für den faltblatttypischen Konformationszustand. Die

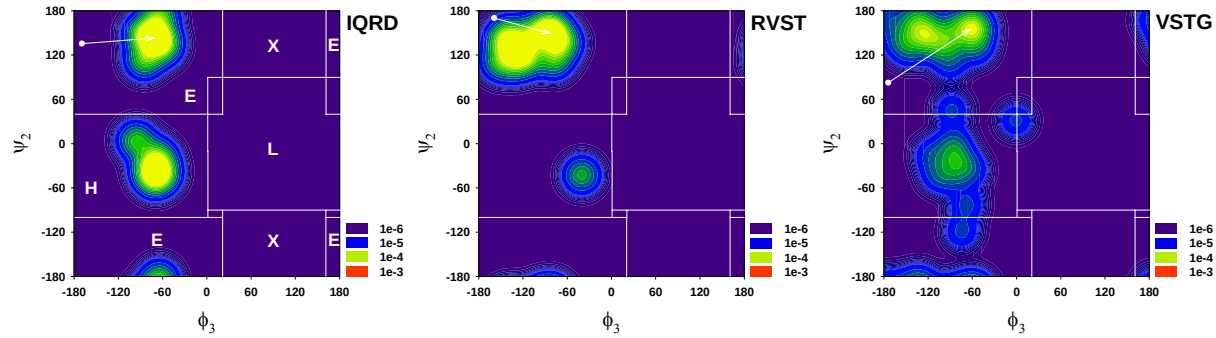


Abbildung IV-27 Dichtefunktionen der ψ_2-φ_3-Verteilungen von Tetrapeptiden aus der Sequenz von M7. Die Pfeilspitze zeigt auf die wahrscheinlichste Konformation des Zielkonformationszustandes. Der Startpunkt des Pfeils markiert die Zielkonformation. Der wahrscheinlichste Konformationszustand ist in allen Fällen faltblatttypisch (*E*). **IQRD** (Nr. 8, $P(E) = 0.55$) Die Zielkonformation ist $\psi_2 = +143.81°$ und $\varphi_3 = -168.83°$. Die wahrscheinlichste Konformation in *E* wird mit $\psi_2 = +144°$ und $\varphi_3 = -70°$ beschrieben. Die Winkeldifferenz zu der wahrscheinlichsten Konformation errechnet sich zu $|\Delta\psi_2| = 0°$ und $|\Delta\varphi_3| = 98°$. Die Dichtefunktion zeigt nur eine geringe Wahrscheinlichkeit für die Ausbildung des beobachteten Diederwinkelpaares. **RVST** (Nr. 20, $P(E) = 0.93$) Die Zielkonformation ist $\psi_2 = +169.58°$ und $\varphi_3 = -158.70°$. Die wahrscheinlichste Konformation in *E* wird mit $\psi_2 = +148°$ und $\varphi_3 = -82°$ beschrieben. Die Winkeldifferenz zu der wahrscheinlichsten Konformation errechnet sich zu $|\Delta\psi_2| = 22°$ und $|\Delta\varphi_3| = 76°$. **VSTG** (Nr. 21, $P(E) = 0.72$) Die Zielkonformation ist $\psi_2 = +80.99°$ und $\varphi_3 = -176.77°$. Die wahrscheinlichste Konformation in *E* wird mit $\psi_2 = +156°$ und $\varphi_3 = -62°$ beschrieben. Die Winkeldifferenz zu der wahrscheinlichsten Konformation errechnet sich zu $|\Delta\psi_2| = 76°$ und $|\Delta\varphi_3| = 114°$.

Winkelabweichung zu der wahrscheinlichsten Konformation in diesem Bereich beträgt $|\Delta\psi_2| = 118°$ und $|\Delta\varphi_3| = 20°$. Trotz dieser großen Winkelabweichung findet man nur eine relativ geringe Verkleinerung der absoluten Wahrscheinlichkeit für die Zielkonformation bezüglich der global wahrscheinlichsten Konformation und der wahrscheinlichsten Konformation im Zielkonformationsbereich *E* (siehe Abbildung IV-29 auf Seite 72). Die Analyse der Dichtefunktion in Abbildung IV-26 auf Seite 70 zeigt, dass die Wahrscheinlichkeit für die Zielkonformation (ψ_2,φ_3) durch die Streuung von Wahrscheinlichkeitsdichte aus dem Bereich der helixtypischen Konformation (*H*) in den Bereich der faltblatttypischen Konformation resultiert. Die Abbildung IV-28 auf Seite 72 zeigt die Winkelabweichungen $|\Delta\psi_2|$ und $|\Delta\varphi_3|$ der von der Zielstruktur vorgegebenen Konformation (ψ_2,φ_3) von der wahrscheinlichsten Konformation des Zielkonformationszustandes. Eine besonders hohe Abweichung von der wahrscheinlichsten Konformation findet man bei den Tetrapeptiden IQRD (Nr. 8, *P(E)* = 0.55), RVST (Nr. 20, *P(E)* = 0.93), VSTG (Nr. 21, *P(E)* = 0.72) und AGAR (Nr. 41, *P(E)* = 0.25), deren Dichtefunktionen ihrer ψ_2-φ_3-Verteilungen in Abbildung IV-27 und Abbildung IV-26 auf Seite 70 gezeigt sind. Bei den 11 Tetrapeptiden **KIQR** (Nr. 7, *E*), **QRDG** (Nr. 9, *L*), RDGQ (Nr. 10, *X*), **QEIE** (Nr. 13, *E*), AGAR (Nr. 41, *E*), RNVQ (Nr. 44, *E*), **TISA** (Nr. 49, *E*), SAEN (Nr. 51, *E*), **ENDE** (Nr. 53, *E*), RVNG (Nr. 80, *E*) und **NGTE** (Nr. 82, *E*), liegt die global wahrscheinlichste Konformation nicht im Zielkonformationsbereich. Zusätzlich liegt bei den hervorgehobenen Tetrapeptiden die global wahrscheinlichste Konformation nicht in dem Konformationsbereich maximaler Wahrscheinlichkeit.

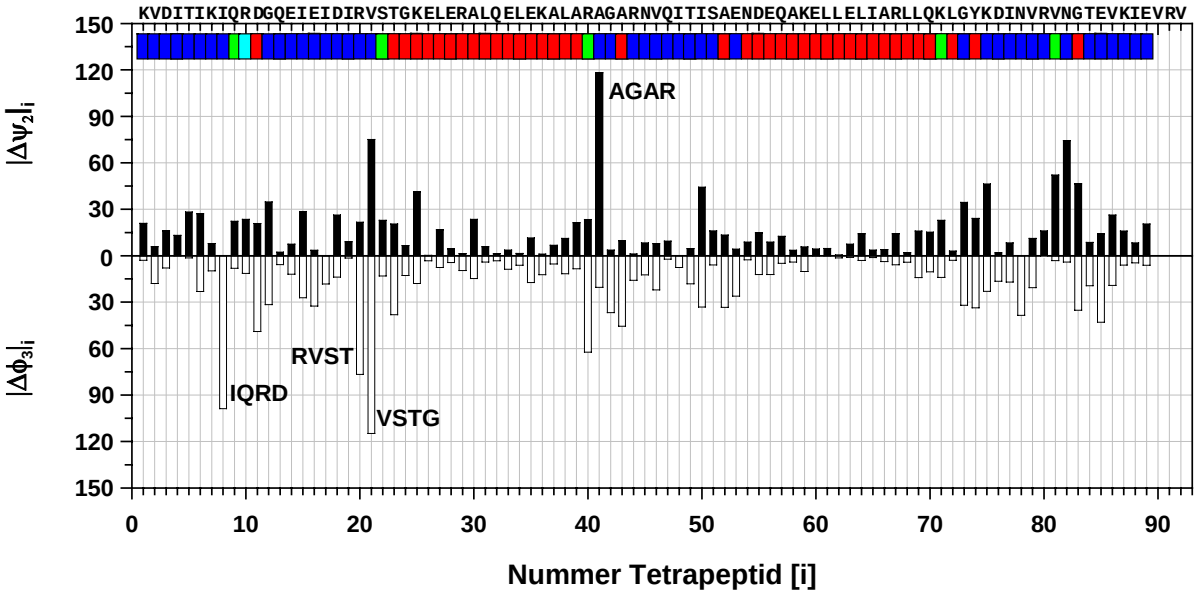


Abbildung IV-28 Winkelabweichungen $|\Delta\psi_2|$ und $|\Delta\varphi_3|$ der Tetrapeptide aus der Sequenz von M7 von der wahrscheinlichsten Konformation im Zielkonformationszustand. Die Grenzen der Konformationsbereiche [(faltblatttypisch), (helixtypisch), (*L-turn*) oder (*X-turn*)] sind in Tabelle II-1, S. 17, definiert. Die Farbkodierung der ersten Aminosäure eines Tetrapeptids beschreibt dessen Zielkonformationszustand. Die Tetrapeptide IQRD (Nummer 8), RVST (Nummer 20), VSTG (Nummer 21) und AGAR (Nummer 41) zeigen große Abweichungen. Die Dichtefunktionen der ψ_2-φ_3-Verteilungen der Tetrapeptide IQRD, RVST, VSTG zeigt die Abbildung IV-27 und für AGAR die Abbildung IV-26 jeweils auf Seite 70.

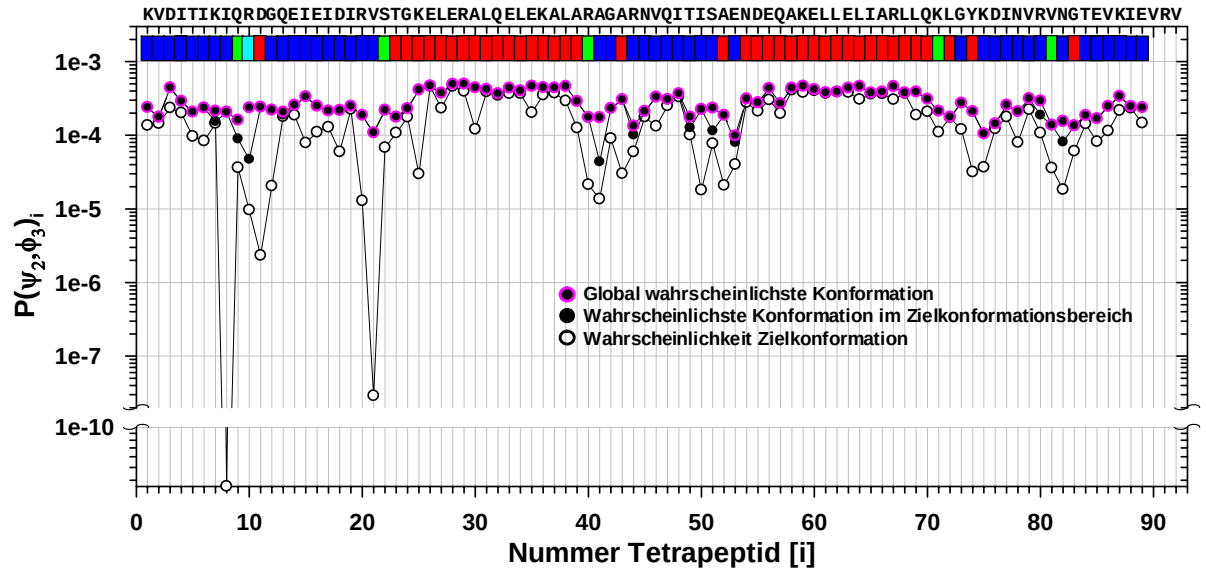


Abbildung IV-29 Absolute Wahrscheinlichkeit *P* der Tetrapeptide aus der Sequenz von Modell M7 für die Zielkonformation (ψ_2,φ_3). Die Grenzen der Konformationsbereiche[(faltblatttypisch), (helixtypisch), (*L-turn*) oder (*X-turn*)] sind in Tabelle II-1, S. 17, definiert. Die Farbkodierung der ersten Aminosäure eines Tetrapeptids beschreibt dessen Zielkonformationszustand. Der Vergleich mit Abbildung IV-28 zeigt, dass die Verringerung der absoluten Wahrscheinlichkeit der Zielkonformation im Allgemeinen mit der Winkelabweichungen $|\Delta\psi_2|$ und $|\Delta\varphi_3|$ zur wahrscheinlichsten Konformation im Zielkonformationsbereich korreliert. Eine starke Verringerung der Wahrscheinlichkeit ist jedoch nicht bei dem Tetrapeptid AGAR (Nr. 41) zu beobachten (vgl. Dichtefunktion in Abbildung IV-26 auf Seite 70).

IV.2.3.4.3. Zusammenfassende Daten für die Sequenzen der Modelle M1-M8

Die Tabelle IV-8 vergleicht das arithmetische und das geometrische Mittel der Wahrscheinlichkeit der einzelnen Tetrapeptide der modellierten Sequenzen für den Zielkonformationszustand mit den entsprechenden Daten der Top7-Sequenz. Das arithmetische Mittel der Wahrscheinlichkeit ist in allen Fällen größer als die geforderte Mindestwahrscheinlichkeit von 0.7. Das geometrische Mittel ist erwartungsgemäß geringer als die arithmetischen Mittel, dennoch konnten bei der Sequenz von M1, M3 und M7 ebenso Werte größer als 0.7 erreicht werden. Die Tabelle IV-8 zeigt weiterhin, dass die mittleren relativen Wahrscheinlichkeiten der Zielkonformationen (ψ_2,φ_3) bezüglich der jeweils wahrscheinlichsten Konformation im Zielkonformationszustand ähnliche, aber dennoch größere Werte zeigen, als bei der Sequenz von Top7. In Tabelle IV-9, S. 75, sind die Tetrapeptide zusammengefasst, bei denen die relative Wahrscheinlichkeit der Zielkonformation bezüglich der wahrscheinlichsten Konformation im Zielkonformationsbereich kleiner ist als die in Abschnitt IV.2.3.3, S. 62, definierte Mindestwahrscheinlichkeit. Man entnimmt der Tabelle, dass bei allen Modellen und der Sequenz von Top7 für das Tetrapeptid Nummer 21 kein alternatives Tetrapeptid verwendet wurde, das die strukturellen Vorgaben hinsichtlich der Zielkonformation erfüllt (vgl. Abbildung IV-22, S. 67, für VTTE bei Top7, Abbildung IV-27, S. 70, für IQRD bei M7). Die Zielkonformation liegt bei dieser Position im faltblatttypischen Konformationsbereich (*E*) und wird mit $\psi_2 = +80°$ und $\varphi_3 = -176°$ beschrieben. Wie die Abbildung IV-30 auf Seite 74 am Beispiel von VNNG zeigt, ist die Nichtexistenz eines geeigneten Tetrapeptids als Ursache für eine ungenügende Beschreibungsmöglichkeit der Zielkonformation auszuschließen. Die global wahrscheinlichste Konformation wird bei diesem Tetrapeptid mit $\psi_2 = +92°$ und $\varphi_3 = -168°$ beschrieben, zu der sich eine Winkeldifferenz zur Zielkonformation von $|\Delta\psi_2| = 12°$ und $|\Delta\varphi_3| = 8°$ errechnet. Die global wahrscheinlichste Konformation liegt ebenso im Zielkonformationsbereich. Die relative Wahrscheinlichkeit der Zielkonformation bezüglich der global wahrscheinlichsten Konformation beträgt 0.76. Trotz dieser sehr guten Daten ist VNNG nicht Bestandteil der Sequenzen von M1-M8. Die Qualität der errechneten Sequenzen hinsichtlich ihrer Wahrscheinlichkeit für den Zielkonformationszustand, als auch der Wahrscheinlichkeit für die Zielkonformation ist direkt von den Aminosäuren

Tabelle IV-8 Mittlere Wahrscheinlichkeiten der modellierten Sequenzen für die Zielstruktur. Das arithmetische Mittel der Wahrscheinlichkeit für den Zielkonformationszustand ist in allen Fällen größer als von Top7. Beim geometrischen Mittel konnten bei M1, M3 und M7 ebenso Werte von größer als die geforderten 0.7 erreicht werden. Die letzte Spalte listet die mittlere relative Wahrscheinlichkeit der Zielkonformation (ψ_2,φ_3), bezogen auf die wahrscheinlichste Konformation des Zielkonformationszustandes. $P(\psi_2,\varphi_3)$ ist der Dichtefunktionswert der Zielkonformation. $P_{MAX}(\psi_2,\varphi_3)_{E,H,L,X}$ beschreibt den Dichtefunktionswert der wahrscheinlichsten Konformation des Zielkonformationszustandes.

Modell	Arithmetisches Mittel Zielkonformationszustand	Geometrisches Mittel Zielkonformationszustand	[a]Mittelwert $P(\psi_2,\varphi_3)/P_{MAX}(\psi_2,\varphi_3)_{E,H,L,X}$
Top7	0.66 ± 0.23	0.61	0.56 ± 0.32
M1	0.74 ± 0.19	0.71	0.64 ± 0.29
M2	0.73 ± 0.21	0.67	0.62 ± 0.30
M3	0.76 ± 0.19	0.74	0.56 ± 0.32
M4	0.74 ± 0.20	0.69	0.61 ± 0.32
M5	0.71 ± 0.18	0.69	0.59 ± 0.30
M6	0.72 ± 0.20	0.67	0.60 ± 0.32
M7	0.75 ± 0.18	0.74	0.59 ± 0.29
M8	0.72 ± 0.20	0.69	0.59 ± 0.32

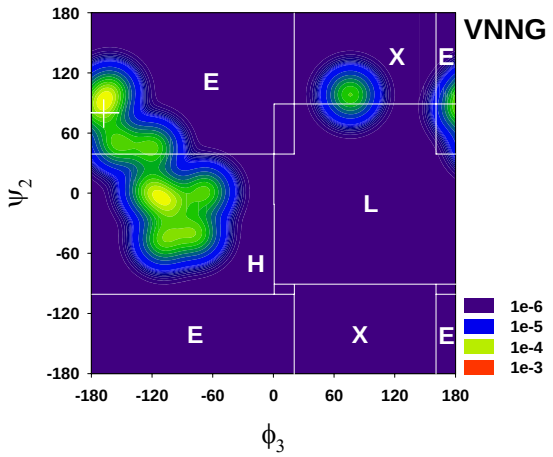


Abbildung IV-30 Dichtefunktion der ψ_2-φ_3-Verteilung des Tetrapeptids VNNG. Das weiße Fadenkreuz markiert die Zielkonformation. VNNG kann an Position 21 zur Modellierung der Zielkonformation mit $\psi_2 = +80°$ und $\varphi_3 = -176°$ verwendet werden. Die relative Wahrscheinlichkeit bezüglich der wahrscheinlichsten Konformation im Zielkonformationszustand errechnet sich zu 0.76. Die Wahrscheinlichkeiten für die Konformationszustände errechnen sich zu *P(E)* = 0.34, *P(H)* = 0.56, *P(L)* = 0.03 und *P(X)* = 0.07. Die Dichtefunktion wurde aus 10 Wertepaaren (ψ_2,φ_3) errechnet.

abhängig, die bei der Errechnung der Sequenzen als Randbedingungen definiert wurden. Dies kann, wie am Beispiel der Position 21 zu erkennen ist, zu der Situation führen, dass Tetrapeptide nicht berücksichtigt werden, obwohl sie die Zielkonformation besser beschreiben, als die tatsächlich verwendeten Tetrapeptide. In gleicher Weise ist der Fall möglich, dass bei der Verwendung eines Tetrapeptids, das die Zielkonformation exakt beschreibt, der Modellierungsprozess trotzdem abgebrochen wird, da keine weiteren Tetrapeptide zur Verlängerung der Proteinsequenz gefunden werden können. Die Abbildung IV-31 zeigt am Beispiel des Tetrapeptids VNDN, dass eine günstige Wahl der Randbedingungen zur Selektion eines Tetrapeptids geführt hat, das die Zielkonformation gut beschreiben kann. VNDN beschreibt an Position 8 im Modell M2 eine faltblatttypische Konformation (*E*) mit der Zielkonformation (ψ_2,φ_3) mit $\psi_2 = +144°$ und $\varphi_3 = -168°$. Die Dichtefunktion der ψ_2-φ_3-Verteilung dieses Tetrapeptids zeigt im Vergleich zu der Dichtefunktion des Tetrapeptids IQRD aus der Sequenz von M7 eine günstige Wahrscheinlichkeitsverteilung (vgl. mit Abbildung IV-27 auf Seite 70). Die relative Wahrscheinlichkeit der Zielkonformation bezüglich der wahrscheinlichsten Konformation im Zielkonformationsbereich beträgt 0.29. Im Anhang VII.3, S. 115, sind für die Aminosäuresequenzen der Modelle M1, M2, M3, M4, M5, M6 und M8 die tetrapeptidbasierten Analysen der Zielsequenzen in Analogie zu den Abschnitten IV.2.3.4.1, S. 65, und IV.2.3.4.2, S. 69, dargestellt.

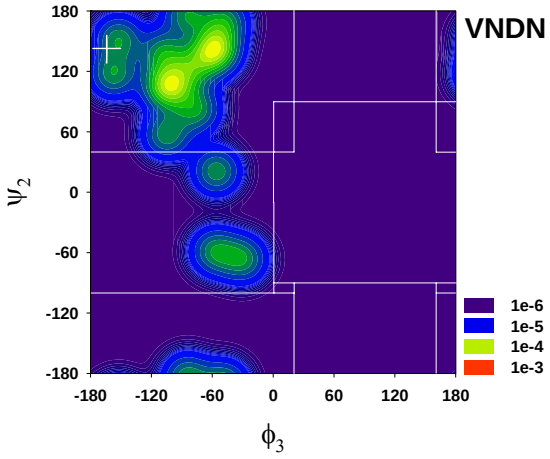


Abbildung IV-31 Dichtefunktion der ψ_2-φ_3-Verteilung des Tetrapeptids VNDN. VNDN wird an Position 8 in der Sequenz von M2 verwendet. Die Zielkonformation ist mit $\psi_2 = +144°$ und $\varphi_3 = -168°$ vorgegeben. Der Vergleich mit dem Tetrapeptid IQRD (Abbildung IV-27 auf Seite 70) aus der Sequenz von M7 zeigt, dass die Zielkonformation deutlich besser beschrieben wird. Die relative Wahrscheinlichkeit bezüglich der wahrscheinlichsten Konformation im Zielkonformationsbereich errechnet sich zu 0.29. Die Wahrscheinlichkeiten für die Konformationszustände errechnen sich zu *P(E)* = 0.8 und *P(H)* = 0.2. Die Dichtefunktion wurde aus 15 Wertepaaren (ψ_2,φ_3) errechnet.

Tabelle IV-9 Übersicht über die Tetrapeptide, bei denen die relative Wahrscheinlichkeit für die Zielkonformation bezüglich der wahrscheinlichsten Konformation im Zielkonformationsbereich kleiner als 0.01 ist. Es wird jeweils auf die Seite verwiesen, auf der die entsprechende Dichtefunktion dargestellt ist. Die Spalte *Position* bezeichnet die Position der ersten Aminosäure des Tetrapeptids in der jeweiligen Sequenz. Die Sequenzen sind in Abbildung IV-18 auf Seite 63 gezeigt. Die Spalte *Ziel* beschreibt den Zielkonformationszustand nach Tabelle II-1, S. 17. Die Spalte *Zielkonformation* listet die Diederwinkelpaare (ψ_2,φ_3) der Zielkonformation. Die Spalten *P(E)*, *P(H)*, *P(L)* und *P(X)* bezeichnen die Wahrscheinlichkeiten der Tetrapeptide für die Konformationszustände *E* (faltblatttypisch), *H* (helixtypisch), *L* (*L-turn*), und *X* (*X-turn*). Die Spalten $|\Delta\psi_2|$ bzw. $|\Delta\varphi_3|$ beschreiben die Winkelabweichungen der Zielkonformation zu dem jeweils nächstliegenden Peak im Zielkonformationsbereich. Die Werte sind hervorgehoben, wenn der nächste Peak der Hauptpeak mit der wahrscheinlichsten Konformation im Zielkonformationsbereich ist. Die Spalte *Anzahl* listet die Anzahl der Diederwinkelpaare (ψ_2,φ_3), aus denen die Dichtefunktionen errechnet wurden.

Modell	Tetrapeptid	S.	Position	Ziel	Zielkonformation	*P(E)*	*P(H)*	*P(L)*	*P(X)*	$\lvert\Delta\psi_2\rvert$	$\lvert\Delta\varphi_3\rvert$	Anzahl
Top7	IDDN	67	8	*E*	(+144°,-168°)	0.75	0.17	0.00	0.08	86°	62°	12
	DNGK	67	10	*X*	(+102°,+ 56°)	0.10	0.19	0.70	0.01	-	-	25
	VTTE	67	21	*E*	(+ 80°,-176°)	0.78	0.19	0.01	0.02	26°	58°	28
	TTES	61	22	*L*	(- 36°,+ 68°)	0.55	0.45	0.00	0.00	-	-	13
	TART	123	51	*E*	(+176°,- 82°)	0.30	0.60	0.00	0.10	34°	56°	10
M1	DNGE	123	10	*X*	(+102°,+ 56°)	0.03	0.25	0.66	0.06	**36°**	**94°**	12
	VRTG	123	21	*E*	(+ 80°,-176°)	0.62	0.38	0.00	0.00	**52°**	**48°**	21
M2	DNGK	67	10	*X*	(+102°,+ 56°)	0.10	0.19	0.70	0.01	-	-	25
	ITTE	123	21	*E*	(+ 80°,-176°)	0.63	0.37	0.00	0.00	64°	42°	21
	TTES	61	22	*L*	(- 36°,+ 68°)	0.55	0.45	0.00	0.00	-	-	13
	DNGA	123	40	*L*	(- 32°,+ 46°)	0.11	0.09	0.74	0.06	**56°**	**30°**	17
M3	SSGR	124	10	*X*	(+102°,+ 56°)	0.09	0.35	0.48	0.08	48°	26°	28
	ITTG	124	21	*E*	(+ 80°,-176°)	0.81	0.19	0.00	0.00	**60°**	**50°**	16
M4	VTTE	67	21	*E*	(+ 80°,-176°)	0.78	0.19	0.01	0.02	26°	58°	28
	TTES	61	22	*L*	(- 36°,+ 68°)	0.55	0.45	0.00	0.00	-	-	13
M5	VETN	124	21	*E*	(+ 80°,-176°)	0.61	0.39	0.00	0.00	50°	14°	10
M6	VDDN	124	8	*E*	(+144°,-168°)	0.67	0.33	0.00	0.00	12°	104°	9
	DNGK	67	10	*X*	(+102°,+ 56°)	0.10	0.19	0.70	0.01	-	-	25
	NIKT	124	20	*E*	(+170°,-158°)	0.53	0.47	0.00	0.00	50°	2°	15
	IKTR	124	21	*E*	(+ 80°,-176°)	0.58	0.42	0.00	0.00	**68°**	**38°**	12
	VSAQ	124	50	*E*	(+110°,- 96°)	0.50	0.50	0.00	0.00	**52°**	**36°**	14
	AQTS	126	52	*H*	(- 36°,-138°)	0.30	0.63	0.07	0.00	**8°**	**76°**	12
M7	IQRD	70	8	*E*	(+144°,-168°)	0.55	0.45	0.00	0.00	**0°**	**98°**	11
	DGQE	126	11	*H*	(+ 16°,-146°)	0.11	0.83	0.06	0.00	**20°**	**48°**	16
	VSTG	70	21	*E*	(+ 80°,-176°)	0.71	0.28	0.01	0.00	68°	52°	31
M8	VETG	126	21	*E*	(+ 80°,-176°)	0.46	0.54	0.00	0.00	**66°**	**40°**	28

IV.2.3.5. Modellierung der Seitenketten und Energieminimierung der Modelle

Nachdem die Aminosäuren des Proteinrückgrats errechnet wurden, erfolgte die Modellierung der Seitenketten mit dem Programm *SCWRL3* [Canutescu *et al.*, 2003]. *SCWRL3* konvergiert, sofern möglich, hin zum globalen Minimum der Aminosäureseitenkettenkonformationen. Anschließend wurden die Seitenketten der Modelle mit der GROMOS96-Implementierung [van Gunsteren *et al.*, 1996] des *Swiss-PdbViewers 3.7* (*SP5*) [Guex & Peitsch, 1997] energieminimiert, um mögliche Überlappungen von Seitenkettenatomen verschiedener Aminosäuren auszuschließen. Die Energieminimierung erfolgte im *steepest descent modus* mit vier mal 20 Zyklen. Die Tabelle IV-10 zeigt die aufgeschlüsselten Gesamtenergieterme für die einzelnen Modelle und für Top7. Um die berechneten Energiewerte von Top7 mit denen der Modelle vergleichen zu können, wurden die Seitenketten von Top7 entfernt, erneut modelliert und anschließend energieminimiert. Zwischen der Kristallstruktur von Top7 und der modellierten Variante ergab sich ein *rmsd*-Wert von 0.10 Å zwischen den C_α-Atomen. Die Energieminimierung führte damit zu praktisch keiner Änderung der Proteinrückgratgeometrie. Unter Berücksichtigung der Aminosäureseitenkettenatome vergrößert sich der *rmsd*-Wert auf 0.16 Å, was zeigt, dass die Seitenketten nahezu in nativer Konformation angefügt wurden. Dabei ist zu beachten, dass dies ebenfalls die lösungsmittelexponierten Seitenketten betrifft. Die Einzelaminosäuren der Modelle M1 bis M8 zeigten bei ihren Energietermen keine Werte, die auf eine Kollision verschiedener Seitenketten haben schließen lassen (Daten nicht gezeigt). Mögliche Überlappungen von Seitenketten wären durch sehr positive Energiewerte des van-der-Waals Energieterms sofort aufgefallen ($E_{nonbonded}$). Die Tabelle zeigt eine große Diversität in den Gesamtenergien. Das Modell M7 ist danach am stabilsten und das Modell M4 am wenigsten stabil. Es ist jedoch zu beachten, dass es sich um *in vacuo* Energien handelt. Die Übertragung dieser Ergebnisse im Sinne einer Stabilitätsreihe für wässrige Bedingungen ist nicht möglich. Die letzte Spalte listet die *rmsd*-Werte der C_α-Atome zwischen den Modellen und der Kristallstruktur von Top7. Es ist auch hier ersichtlich, dass die Energieminimierung nur zu einer geringen Änderung der Proteinrückgratgeometrie geführt hat.

Tabelle IV-10 Energieterme der Modelle M1-M8 und Top7. Die Energieminimierung erfolgte nach der Modellierung der Seitenketten mit der *Implementierung* des GROMOS96 Kraftfeldes [van Gunsteren *et al.*, 1996] des *Swiss-PdbViewers 3.7* (*SP5*) [Guex & Peitsch, 1997] mit vier mal 20 Zyklen im *steepest descent* Modus. Die Energieminimierung wurde angewendet, um Überlappungen von Seitenkettenatomen zwischen verschiedenen Resten infolge einer möglichen fehlerhaften Modellierung auszuschließen. Die Energieminimierung erfolgte *in vacuo*. Es sind die Beiträge der einzelnen Energieterme zur Gesamtenergie E_{Total} angegeben. Es sind E_{Bonds} die Energie der Bindungslängen, E_{Angles} die Energie der Bindungswinkel, $E_{Torsion}$ die Energie der Diederwinkel, $E_{improper}$ die Energie der uneigentlichen Diederwinkel, $E_{Nonbonded}$ die van-der-Waals Energie, $E_{Elektro}$ die Coulombenergie. Die letzte Spalte gibt die *rmsd*-Werte der korrespondierenden C_α-Atome zwischen Modell und Kristallstruktur von Top7 an. Die *rmsd*-Werte wurden mit dem Programm *LSQMAN* bestimmt [Kleywegt, 1999].

Modell	E_{Bonds} (kJ/mol)	E_{Angles} (kJ/mol)	$E_{Torsion}$ (kJ/mol)	$E_{improper}$ (kJ/mol)	$E_{Nonbonded}$ (kJ/mol)	$E_{Elektro}$ (kJ/mol)	E_{Total} (kJ/mol)	rmsd (Å)
Top7	+65.06	+258.99	+303.36	+ 94.72	-2704.89	-3090.53	-5073.28	0.10
M1	+63.01	+278.86	+311.29	+ 96.17	-2343.70	-2909.39	-4503.76	0.12
M2	+67.85	+310.82	+309.52	+116.35	-2316.04	-3171.49	-4683.42	0.13
M3	+63.49	+270.31	+320.91	+ 87.80	-2358.29	-3408.70	-5024.48	0.11
M4	+60.11	+280.99	+310.15	+ 96.69	-2348.70	-2391.65	-3992.42	0.12
M5	+63.17	+273.58	+307.27	+ 95.07	-2421.21	-2807.06	-4489.18	0.11
M6	+63.73	+302.92	+312.56	+100.57	-2411.03	-3600.03	-5231.20	0.11
M7	+67.02	+273.94	+314.42	+ 91.54	-2339.53	-3964.79	-5557.38	0.12
M8	+57.98	+260.99	+296.51	+ 86.54	-2460.60	-2246.38	-4004.96	0.12

IV.2.3.6. Bewertung der Modelle mit *whatcheck*

Das Programmpaket *whatcheck* [Hooft *et al.*, 1996] wird zur Evaluierung von Proteinmodellen verwendet. Eine Analyse der errechneten Modelle M1-M8 sollte zeigen, inwieweit die manuell definierten Randbedingungen und die daraus errechneten Sequenzen zu guten Modellen geführt haben.

Die Tabelle IV-11 zeigt die Zusammenfassung der *whatcheck*-Prüfberichte für die einzelnen Modelle. Als Referenz wurde die Kristallstruktur von Top7 nach Energieminimierung, sowie die Struktur von Top7 nach Entfernen der Seitenketten und erneutem Modellieren selbiger mit anschließender Energieminimierung in der Tabelle aufgeführt. Um einen Vergleich mit natürlichen Proteinen ähnlicher Größe zu erhalten, zeigt die Tabelle die Evaluierung der Kristallstrukturen der N-terminalen Domäne von γ-Kristallin (Aminosäure 1-84, *PDB*-Code 1AMM), Ubiquitin (*PDB*-Code 1UBQ) und Lysozym (*PDB*-Code 153L). Auch in diesen Beispielen erfolgt ein Vergleich des *whatcheck*-Prüfberichts der jeweiligen Kristallstruktur mit dem energieminimierten Protein. Die Daten in der Tabelle entsprechen *Z-scores*, die strukturelle Charakteristika von Proteinen bewerten. Je positiver ein *Z-score* ist, umso günstiger ist diese

Tabelle IV-11 Analyse der Modelle M1 bis M8 mit *whatchek* (Version 20060607-2300) [Hooft *et al.*, 1996] und Vegleich der Parameter mit Proteinstrukturen aus der *PDB*. Die Daten entsprechen einem *Z-score* (0 entspricht dem Mittelwert), bei dem positive Werte eine günstige Konfiguration indizieren und negative *Z-scores* strukturelle Parameter anzeigen, die schlechtere Eigenschaften besitzen, als ein durchschnittliches Protein. Der Zusatz *min* für einige Proteine in der Spalte *Protein* weist auf eine energieminimierte Struktur hin. Die Energieminimierungen erfolgten mit vier mal 20 Zyklen im *deepest descent modus* mit der GROMOS96-Implementierung [van Gunsteren *et al.*, 1996] des *Swiss-PdbViewers* 3.7 (*SP5*) [Guex & Peitsch, 1997]. Der Zusatz *mod* bei dem Protein Top7 in der dritten Zeile weist darauf hin, dass die Seitenketten mit dem Programm *SCWRL3* [Canutescu *et al.*, 2003] neu modelliert und anschließend, wie in Abschnitt IV.2.3.5, S. 76, beschrieben, energieminimiert wurden. In der Originalsequenz von Top7 befindet sich an Position 35 ein Selenomethionin, welches bei der Modellierung der Aminosäureseitenketten durch ein Methionin ersetzt wurde.

Protein	*1st generation packing quality*	*Ramachandran plot appearance*	*chi-1/chi-2 rotamer normality*	*Backbone conformation*
Top7 (1QYS, cryst.)	+0.356	-2.185	-2.581	+0.469
Top7 (min)	+1.766	-1.121	-0.597	+0.520
Top7 (1QYS, mod.)	+1.845	-0.948	+3.562	+0.371
M1	+2.233	-1.044	+2.238	+0.634
M2	+2.062	-1.313	+3.071	+0.307
M3	+2.193	-1.556	+2.160	+0.896
M4	+2.233	-1.505	+2.157	+0.354
M5	+2.320	-1.516	+2.699	+0.784
M6	+1.610	-0.947	+2.808	+0.632
M7	+2.269	-0.930	+2.780	+0.542
M8	+2.062	-1.560	+2.976	+1.169
γ-Kristallin 1-84 (*PDB*-Code 1AMM)	+0.685	-3.329	-0.769	+0.108
γ-Kristallin 1-84 (*min*)	+0.709	-2.000	+0.559	+0.482
Ubiquitin (*PDB*-Code 1UBQ)	+1.191	+0.949	-1.229	+2.812
Ubiquitin (*min*)	+1.104	+1.119	-0.176	+2.756
Lysozym (*PDB*-Code 153L)	-1.239	+0.381	-0.252	-0.981
Lysozym (*min*)	-1.449	+0.536	+0.344	-1.106

Eigenschaft in der jeweiligen Proteinstruktur oder in dem Modell ausgeprägt. Die Tabelle zeigt eine gute Bewertung der Modelle M1 bis M8. Die *1st generation packing quality* [Vriend & Sander, 1993] entspricht den Kontaktpotentialen nach Sippl [Sippl, 1993; Sippl, 1995]. Zu erkennen ist, dass die errechneten Modelle etwas besser bewertet werden als natürliche Proteine oder Top7. Der Ramachandran *Z-score* beschreibt, wie gut die Proteinrückgratkonformation aller Reste durch die erlaubten Regionen im Ramachandran-Diagramm [Ramachandran *et al.*, 1963] beschrieben werden. Auch wenn die *scores* negativ sind, bewertet *whatcheck* die Konformation der einzelnen Reste dennoch als gut. Die Spalte *backbone conformation* zeigt, dass die Gesamtstruktur der Modelle als gut bewertet wird. Mit der *chi-1/chi-2 rotamer normality* bewertet *whatcheck*, inwieweit die Seitenketten der Aminosäuren in den Sekundärstrukturelementen in einer jeweils energetisch günstigen Konformation vorliegen. Der Vergleich der modellierten und energieminimierten Rotamere mit denen, die in der natürlichen Konformation vorliegen zeigt, dass die energieminimierten Strukturen besser bewertet werden, als die Konformation der Seitenketten aus experimentell bestimmten Proteinstrukturen.

IV.3. Experimentelle Charakterisierung der Modelle

IV.3.1. Rekombinante Expression der Proteine M1 bis M8

Die Expressionen der Proteine erfolgten in *E. coli* BL21 (DE3) als Fusionskonstrukt mit einem N-terminalen Hexa-Histidin-*tag* (His_6-*tag*). Die Abbildung IV-32 zeigt den Gesamtzellaufschluß nach Expression der Proteine über sechs Stunden in 20 ml Kulturvolumen. Die Banden entsprechen 10 µl einer Lösung (nach Aufkonzentrierung) mit einer optischen Zelldichte von $OD_{600} = 10$. Die SDS-Gele zeigen eine Überexpression der Proteine, die jedoch unterschiedlich hoch ist. Bei diesem Expressionstest wurden die Varianten M1 und M8 am besten exprimiert.

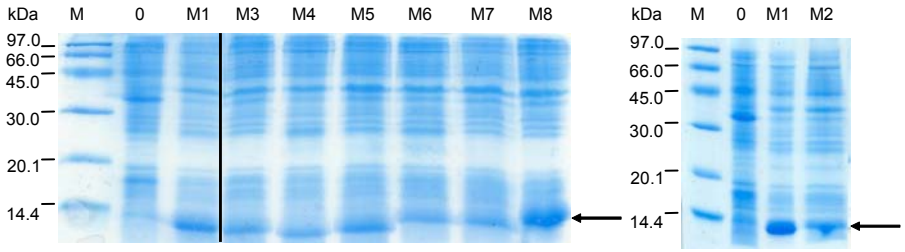


Abbildung IV-32 Überexpression der Proteine M1 bis M8. Die Banden zeigen jeweils den Gesamtzellaufschluß. M-LMW Proteinmarker, 0-Vor Induktion der Proteinexpression. Die Pfeile markieren die Überexpressionsbanden der Zielproteine.

IV.3.2. Rekombinante Herstellung des Proteins M7

Die Expression des Proteins M7 erfolgte in *E. coli* BL21 (DE3) als lösliches Fusionskonstrukt mit einem N-terminalen Hexa-Histidin-*tag* (His_6-*tag*). Eine überwiegende Abtrennung der Wirtsproteine wurde nach Zellaufschluss durch 20 minütiges Erhitzen des Rohextraktes auf 80 °C erzielt. Anschließend erfolgte die Reinigung des Zielproteins durch immobilisierte Metallchelat-Affinitätschromatographie (IMAC) an einer Nickel-NTA Matrix. Nach Elution des Proteins und Entfernen des Imidazols wurde der His_6-*tag* durch einen Thrombinverdau abgespalten. Das Protein wurde anschließend durch eine Gelfiltration bis zur Homogenität gereinigt. Die theoretische Molekularmasse von M7 konnte bis auf eine Abweichnung von 3 Da durch Massenspektrometrie bestätigt werden. Das Massenspektrum ist im Anhang VII.7, S. 137, dargestellt.

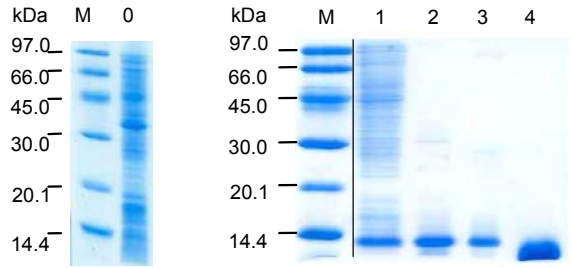


Abbildung IV-33 Dokumentation der Expression und Reinigung des Proteins M7 anhand einer SDS-Gelelektrophorese (15 % Acrylamid, Coomassie-Färbung). **M**-LMW Proteinmarker, **0**-Gesamtzellen vor Induktion der Genexpression, **1**-Gesamtzellen induziert, **2**-Löslicher Anteil des Rohextrakts nach 20 min Erhitzen bei 80 °C. **3**-Eluat nach Immobilisierung an einer Ni-NTA-Matrix. **4**-M7 nach Abspaltung des His_6-*tags* mit Thrombin und anschließender Gelfiltration. Das Protein enthielt nach der Thrombinspaltung zusätzlich zu der modellierten Sequenz (vgl. Abbildung IV-18 auf Seite 63) am N-Terminus die Aminosäuren GSHM.

IV.3.3. Charakterisierung von M7

IV.3.3.1. Spektroskopische Eigenschaften von M7

M7 enthält an Position 74 in seiner Sequenz ein Tyrosin. Das UV-Spektrum zeigte ein Tyrosinspektrum mit einem Maximum bei einer Wellenlänge von $\lambda = 277$ nm (Abbildung IV-34A). Die Konzentrationsbestimmung des Proteins erfolgte über die Absorption bei dieser Wellenlänge. Um Aussagen über die Sekundär- und Tertiärstruktur von M7 treffen zu können, wurde das Protein in seiner nativen und denaturierten Form der Fern-UV-CD-Spektroskopie unterzogen. Ein Vergleich der Spektren zeigte eine Verringerung des Betrages der Amplituden bei der entfalteten Spezies. Die Messung des denaturierten Proteins erfolgte in 7.0 M GdmCl. Das native Protein zeigt das Spektrum eines Proteins mit helikalen und faltblatttypischen Sekundärstrukturelementen. Eine Abschätzung des Sekundärstrukturgehaltes erfolgte mit dem Programm *CDPro* [Sreerama & Woody, 2000]. Damit konnte ein α-helikaler Anteil von 41 %, ein β-Faltblattanteil von 17 % und entsprechend der unstrukturierte Anteil (*coil*) zu 42 % bestimmt werden. Die Tabelle IV-12, S.82, vergleicht die theoretischen Sekundärstrukturanteile mit denjenigen, die mit *CDPro* und mit NMR-Spektroskopie bestimmt wurden (Abschnitt IV.3.3.3, S. 81). Beim β-Faltblattanteil und dem unstrukturierten Anteil zeigen sich dabei erhebliche Unterschiede. Danach ist der unstrukturiere Anteil deutlich zu groß und der β-Faltblattanteil zu gering. Der theoretische Anteil an Sekundärstrukturelementen wurde anhand der Kristallstruktur von Top7 mit dem Programm *DSSP* [Kabsch & Sander, 1983] ermittelt. Dabei wurden die Symbole *E* und *H* den Sekundärstrukturelementen β-Faltblatt bzw. der α-Helix und alle anderen Symbole dem Strukturelement *coil* zugewiesen. Der *DSSP-Output* ist in Anhang VII.2, S. 112, gezeigt.

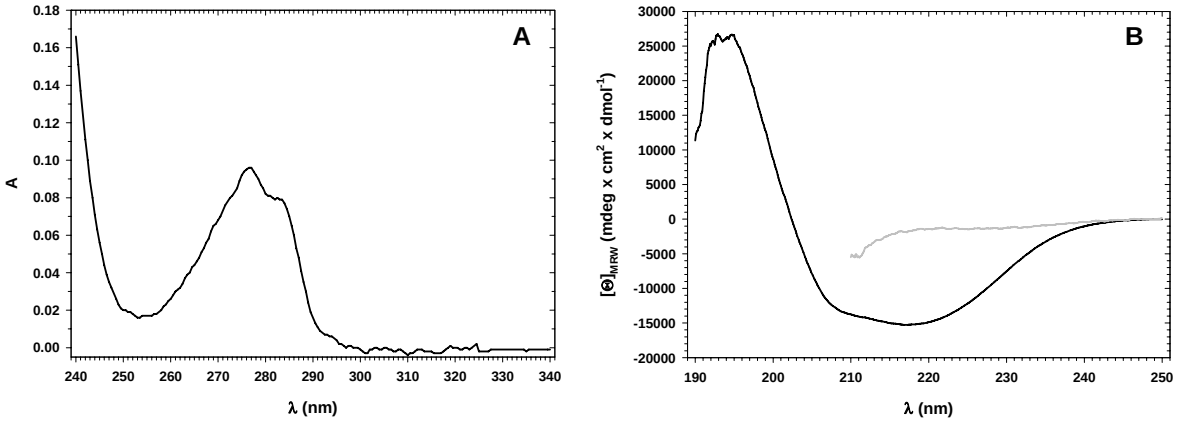


Abbildung IV-34 UV-Spektrum und Fern-UV-CD-Spektrum von M7. Das Protein wurde bei einer Temperatur von 20 °C in 25 mM Natriumphosphatpuffer pH 8.0, 40 mM NaCl vermessen. Das Protein besitzt zusätzlich die vier Aminosäuren GSHM an seinem N-Terminus. **A** UV-Spektrum. Die Form des Spektrums wird durch das Tyrosin an Position 74 der Sequenz von M7 bestimmt. Die Messung erfolgte in einer 1 cm Quarzglasküvette. **B** Circulardichroismus von M7 in nativer (-) und denaturierter (‒) Form. Die Messung erfolgte in einer 0.05 cm Quarzglasküvette bei einer Proteinkonzentration von 0.97 mg/ml. Der denaturierte Zustand wurde durch Inkubation des Proteins über 18 Stunden in 7.0 M GdmCl bei einer Temperatur von 20 °C erzielt.

IV.3.3.2. Analytische Ultrazentrifugation von M7

Zur Bestimmung des Oligomerisierungszustandes von M7 wurde das Protein einer analytischen Ultrazentrifugation unterzogen. Die Abbildung IV-35 zeigt das Ergebnis des Sedimentationsgleichgewichtsexperiments. Das Sedimentationsgleichgewicht in Abbildung A kann durch einen monoexponentiellen *fit* beschrieben werden, was auf eine monomere Spezies hindeutet. Die Abweichungen vom verwendeten *fit* zeigt die Abbildung B. Es wurde ein ungefähres Molekulargewicht von 10.4 ± 0.2 kDa bestimmt, das sehr gut mit der theoretischen Molekularmasse von 10 820.4 Da übereinstimmt.

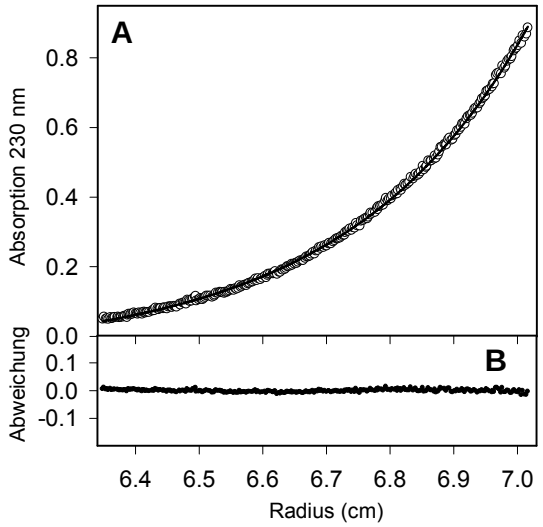


Abbildung IV-35 Analytische Ultrazentrifugation von M7. Die Konzentration des Proteins betrug 180 µg/ml. Die Analyse erfolgte in einem 20 mM Natriumphosphatpuffer pH 8.0, 400 mM NaCl. Das Sedimentationsgleichgewicht des Proteins wurde nach 23 Stunden bei 20 000 *rpm* in einem An50Ti-Rotor erreicht. **A** Optische Absorption als Funktion des Abstandes vom Rotormittelpunkt. **B** Differenzen zwischen angepassten und experimentellen Daten als Funktion des Abstandes vom Rotormittelpunkt

IV.3.3.3. 2D-^{1}H-NMR-Messungen von M7

Die Aufnahme von homonuklearen ^{1}H-NMR-Spektren erlaubte eine Aussage über die Sekundär- und Tertiärstruktur von M7. Die 2D-NMR Daten zeigen, dass M7 im Ganzen ein strukturiertes Protein ist. Dies wird durch die hohe Dispersion der Amidprotonenresonanz des Proteinrückgrats zwischen 9.6-7.1 ppm deutlich, wie im TOCSY-Spektrum zu sehen ist (Abbildung IV-36 auf Seite 82). Da dieses Protein nur zwei aromatische Aminosäuren besitzt (Tyrosin an Position 74 der modellierten Sequenz und Histidin in der Thrombinschnittstelle GSHM am N-Terminus), deren Ringstrom zu Verschiebungseffekten außerhalb der für *random coil* charakteristischen Region (9.0-8.2 ppm) führen kann, deutet die ausgeprägte Verteilung der Amidprotonenresonanzen auf das Vorhandensein eines hohen Anteils an Sekundärstrukturelementen hin. Die NOE-Konnektivitäten im NOESY-Spektrum indizieren das Vorhandensein von α-helikalen und β-faltblatttypischen Sekundärstrukturelementen, wie in Abbildung IV-37 auf Seite 83 zu sehen ist. Ein relativ hoher Anteil an Aminosäuren mit β-Faltblattkonformation lässt sich aus den starken $d_{\alpha N}(i,i+1)$ Konnektivitäten zwischen den H^α und H^N Resonanzen sequentiell benachbarter Reste ableiten. Zusätzlich impliziert die hohe Anzahl an H^N-H^N Konnektivitäten in der Region

zwischen 8.4-7.5 ppm das Vorhandensein von helikalen Strukturelementen. Zur Abschätzung der Sekundärstrukturanteile von M7 wurde eine semi-quantitative Analyse des Sekundärstrukturanteils von M7, basierend auf der H^N-H^α Verteilung in der *fingerprint* Region des 2D-COSY-Spektrums, durchgeführt [Wishart *et al.*, 1991]. Das 2D-COSY-Spektrum und dessen Auswertung ist in Anhang VII.5, S. 128, dargestellt. Die daraus bestimmten Anteile sind, neben den theoretisch zu erwartenden Anteilen und den aus dem CD-Spektrum abgeleiteten, in Tabelle IV-12 aufgeführt. Der Tabelle ist zu entnehmen, dass der Circulardichroismus und die NMR-Spektroskopie in diesem Fall zu unterschiedlichen Vorhersagen in den Anteilen an den verschiedenen Sekundärstrukturelementen geführt haben. Die aus dem COSY-Spektrum abgeleiteten Anteile an β-Faltblatt und *coil* zeigen eine bessere Übereinstimmung zu den erwarteten Werten, als die aus dem CD-Spektrum mit *CDPro* errechneten Anteile.

Tabelle IV-12 Vergleich der mit CD und NMR ermittelten Sekundärstrukturanteile von M7 mit den theoretischen Werten. Die Bestimmung der theoretischen Anteile erfolgte mit dem Programm *DSSP* [Kabsch & Sander, 1983] anhand der Kristallstruktur von Top7. Dabei wurde das Symbol *E* dem Sekundärstrukturelement β-Faltblatt, das Symbol *H* dem Sekundärstrukturelement α-Helix und alle anderen Symbole dem Strukturelement *coil* zugewiesen. Circulardichroismus und NMR zeigen deutliche Unterschiede in den Anteilen an vorhergesagten Sekundärstrukturelementen. Die aus dem 2D-COSY-Spektrum ermittelten Sekundärstrukturanteile zeigen zu den theoretisch erwarteten Anteilen eine bessere Übereinstimmung, als nach Auswertung des CD-Spektrums.

Sekundärstrukturelement	*CD*	*NMR*	*DSSP*
α-Helix	41 %	23 %	36 %
β-Faltblatt	17 %	53 %	40 %
coil	42 %	24 %	24 %

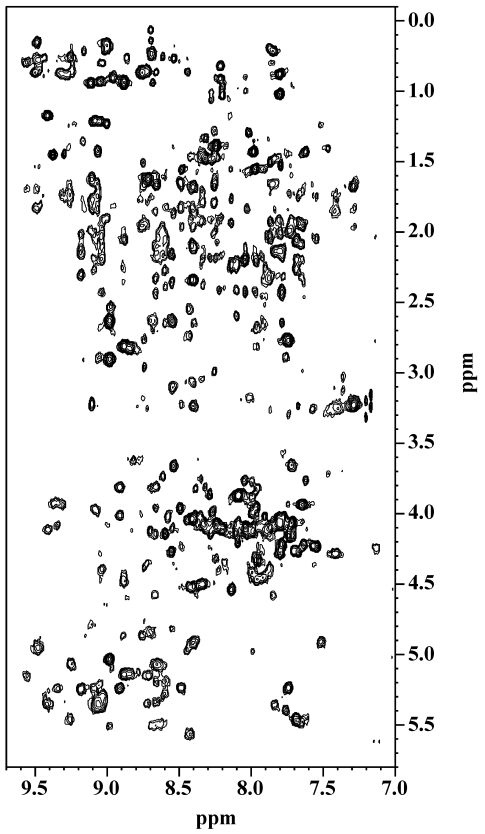


Abbildung IV-36 2D-TOCSY-Spektrum von M7. Es ist die Amidregion des Proteinrückgrats dargestellt. Die Messung erfolgte bei einer Proteinkonzentration von 1.8 mM in 15 mM Natriumphosphatpuffer pH 6.5 und bei einer Temperatur von 30 °C. Die beobachtete Signalverbreiterung (besonders in der für β-Faltblatt charakteristischen Region) könnte ein Indiz für eine Proteindimerisierung aufgrund der Konzentration und/oder des pH-Wertes sein.

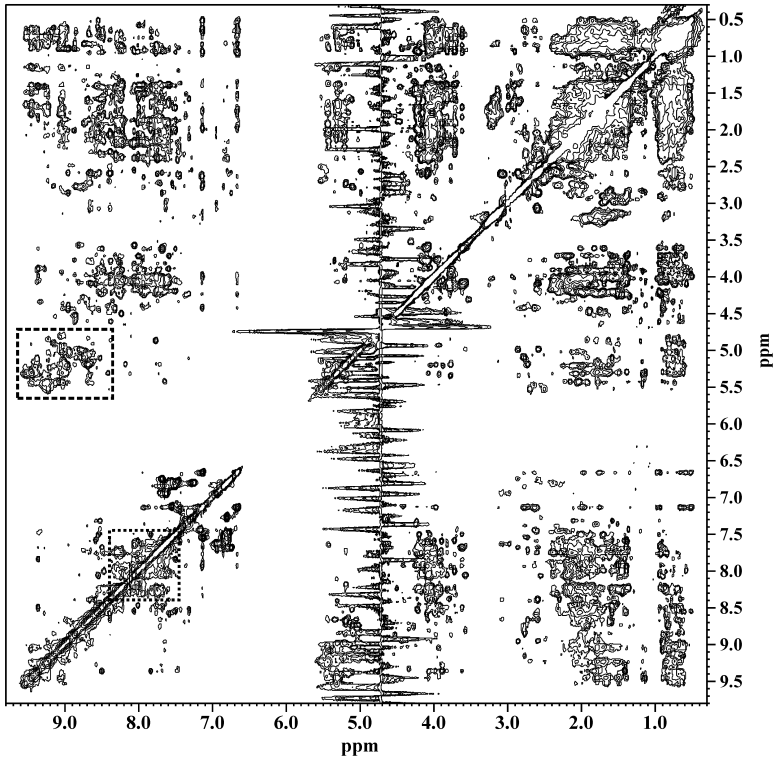


Abbildung IV-37 2D-NOESY-Spektrum von M7. Die Messung erfolgte bei einer Proteinkonzentration von 1.8 mM in 15 mM Natriumphosphatpuffer pH 6.5 und bei einer Temperatur von 30 °C. Die NOE-Konnektivitäten innerhalb der gestrichelten Region weisen auf β-Faltblattstrukturen hin. Die Signale innerhalb des gepunkteten Bereiches weisen auf α-helikale Strukturelemente hin.

IV.3.3.4. Chemische und thermische Stabilität von M7

In Abbildung IV-38A und B auf Seite 84 sind die chemisch bzw. thermisch induzierten Entfaltungsübergänge von M7 in 25 mM Natriumphosphatpuffer pH 8.0 und 40 mM NaCl dargestellt. Das Protein enthielt am N-Terminus zusätzlich zur modellierten Sequenz die vier Aminosäuren GSHM. Als Denaturierungsmittel wurde Guanidiniumchlorid (GdmCl) verwendet. Die Übergänge lassen auf eine außerordentliche Stabilität von M7 schließen. Der Kooperativitätsparameter wurde bei der chemischen Denaturierung zu einem Wert von $m_{n\rightarrow d}$ = -10.5 ± 0.5 kJ/(mol M) bestimmt, woraus sich der Übergangsmittelpunkt zu einem Wert von $D_{1/2}$ = 6.6 M GdmCl ergab. Die Freie Entfaltungsenthalpie $\Delta G_{n\rightarrow d}^{H_2O}$ beim Übergang vom nativen (*n*) Protein zur denaturierten (*d*) Spezies errechnete sich aus der Anpassung der Messpunkte bei der Messtemperatur von T = 293 K (20 °C) zu $\Delta G_{n\rightarrow d}^{H_2O}$ = +69.3 ± 3.0 kJ/mol. Zur Erhebung weiterer thermodynamischer Daten wurde M7 bei unterschiedlichen GdmCl-Konzentrationen thermisch denaturiert. Der Schmelzpunkt des Proteins wurde in denaturierungsmittelfreiem Puffer bis zu einer Temperatur von T = 383 K (110 °C) nicht erreicht. Eine beginnende thermische Entfaltung konnte erst ab einer Konzentration von 5.5 M GdmCl und einer Temperatur von T = 363 K (90 °C) beobachtet werden. Die Abbildung IV-38B zeigt weiterhin, dass ab einer Konzentration von 5.5 M GdmCl das Protein bei niedrigeren Temperaturen einer Kältedenaturierung unterliegt. Bei weiter steigender Konzentration an Denaturierungsmittel

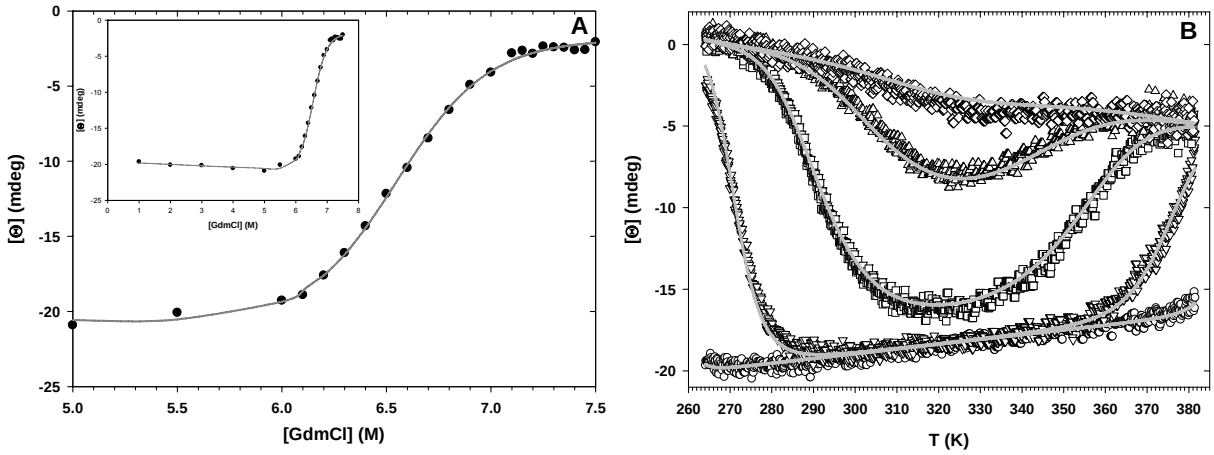
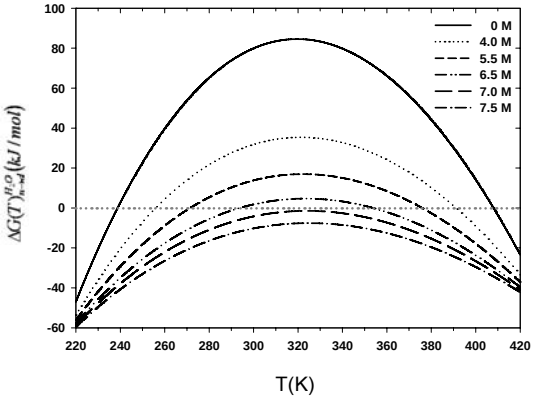


Abbildung IV-38 Chemisch und thermisch induzierte Entfaltung von M7 in 25 mM Natriumphosphatpuffer pH 8.0 mit 40 mM NaCl. Es wurde GdmCl als Denaturierungsmittel verwendet. Die Proteine enthalten zusätzlich zur modellierten Sequenz am N-Terminus die Sequenz GSHM. Die Proteinkonzentration betrug in beiden Messungen 150 µg/ml. Die Proben wurden 18 Stunden bei 20 °C inkubiert und die Messungen bei einer Wellenlänge von 220 nm mit einem Jasco-810 CD-Spektrometer in einer 0.1 cm Quarzglasküvette durchgeführt. Die Anpassung der Daten erfolgte nach einem Zweizustandsmodell (vgl. Abschnitte III.4.3, S. 30, und III.4.4, S. 32). (*n*) bezeichnet das native Protein, (*d*) die denaturierte Spezies. Die *fit*-Parameter aus der Anpassung der Messpunkte sind im Anhang VII.4, S. 127, gelistet. **A** GdmCl induzierte Entfaltung von M7. Die Freie Entfaltungsenthalpie ergibt sich zu einem Wert von $\Delta G_{n\to d}^{H_2O}$ = +69.3 ± 3.0 kJ/mol. Der Kooperativitätsparameter beziffert sich auf $m_{n\to d}$ = -10.5 ± 0.5kJ/(mol M). Daraus bestimmt sich der Übergangsmittelpunkt zu $D_{1/2}$ = 6.6 M. Der Regressionskoeffizient R^2 der angepassten Daten beträgt 0.999 **B** Thermisch induzierte Denaturierung von M7 bei verschiedenen GdmCl-Konzentrationen: (○) 4.0 M, (▽) 5.5 M, (□) 6.5 M, (Δ) 7.0 M, (◇) 7.5 M. Die Heizrate betrug 1 K/min. Der Regressionskoeffizient R^2 der angepassten Daten beträgt 0.996. Die Auswertung der Übergänge ergab eine Freie Entfaltungsenthalpie von $\Delta G(T)_{n\to d}^{H_2O}$ = +76.1 kJ/mol und eine Stabilisierungsenthapie von $\Delta H_{n\to d}^{H_2O}$ = -115.2 ± 2.0 kJ/mol. Die Änderung der Entropie beziffert sich auf einen Wert von $\Delta S_{n\to d}^{H_2O}$ = -0.65 kJ/(mol K). Die Änderung der Wärmekapazität beträgt $\Delta C_{p(n\to d)}^{H_2O}$ = 7.5 kJ/(mol K). Die Werte beziehen sich auf 20 °C.

beobachtet man diesen Effekt ebenso bei höheren Temperaturen. Aus der Anpassung der Messpunkte wurde die Freie Entfaltungsenthalpie zu einem Wert von $\Delta G(T)_{n\to d}^{H_2O}$ = +76.1 kJ/mol bei der Referenztemperatur 20 °C errechnet, was in guter Übereinstimmung mit dem Ergebnis aus der chemischen Denaturierung steht. Die Änderung der Enthalpie $\Delta H_{n\to d}^{H_2O}$ wurde zu einem Wert von -115.2 ± 2.0 kJ/mol und die Änderung der Entropie zu $\Delta S_{n\to d}^{H_2O}$ = -0.65 kJ/(mol K) bestimmt. Die Struktur von M7 wird bei Raumtemperatur netto über entropische Effekte stabilisiert. Die Änderung der Wärmekapazität $\Delta Cp_{n\to d}^{H_2O}$ errechnete sich zu +7.5 kJ/(mol K).

Abbildung IV-39 Temperaturabhängigkeit der Freien Entfaltungsenthalpie von M7. Die Kurven wurden unter Verwendung der aus den thermisch induzierten Entfaltungsübergängen gewonnenen Daten mit Hilfe der Gleichung III-6 errechnet. M7 zeigt seine maximale Stabilität bei einer Temperatur von T = 321 K (48 °C, $\Delta G(T)_{n\to d}^{H_2O}$ = +84.6 kJ/mol). Die Schmelzpunkte in 0 M GdmCl wurden auf eine Temperatur von T = 238 K (-35 °C) und T = 408 K (+135 °C) extrapoliert.



Die Abbildung IV-39 auf Seite 84 zeigt die aus den erhaltenen thermodynamischen Daten der Temperaturübergänge nach Gleichung III-6 berechnete Temperaturabhängigkeit der Freien Entfaltungsenthalpie für verschiedene GdmCl-Konzentrationen und für denaturierungsmittel-freien Puffer. M7 hat danach seine maximale Stabilität bei einer Temperatur von T = 321 K (48 °C, $\Delta G(T)_{n\to d}^{H_2O} = +84.6$ kJ). Durch Extrapolation von ΔG(T) wurde die Temperatur der Kältedenaturierung, T_c, und die Schmelztemperatur, T_m, zu T_c = 238 K (-35 °C) bzw. T_m = 408 K (+135 °C) bestimmt.

V. Diskussion

Im Mittelpunkt der vorliegenden Arbeit stand die Entwicklung eines Modellierungssystems von Aminosäuresequenzen zu gegebenen Proteinstrukturen. Dessen grundlegende Funktionsweise beruht auf der Assemblierung von Tetrapeptidfragmenten in ihrem bevorzugten Konformationszustand, der jeweils durch eine statistische Analyse der Geometrie dieser Fragmente aus experimentell bestimmten Proteinstrukturen ermittelt werden konnte. Hierfür wurde eine verbesserte Methodik zur Datenaufbereitung entwickelt, die genügend Strukturinformationen für eine statistische Analyse dieser Fragmente bereitstellen konnte.

Die Analyse der Tetrapeptidkonformationen erfolgte in Analogie zu Sudarsanam *et al.* durch die Bestimmung des ψ-Winkels der zweiten Aminosäure (ψ_2) und des φ-Winkels der dritten Aminosäure (φ_3) (vgl. Abbildung I-2 auf Seite 6) [Sudarsanam & Srinivasan, 1997]. Diese beiden Diederwinkel eignen sich besonders gut für eine Konformationsanalyse, da die Peptidbindung zwischen diesen beiden Winkeln in den meisten Fällen als strukturell invariant angesehen werden kann. Die Auswertung der Konformationsdaten beschränkt sich daher auf die Berechnung zweidimensionaler Wahrscheinlichkeitsdichtefunktionen. Jede andere Kombination von zwei Diederwinkeln zwischen zwei Aminosäuren führt zu der Einführung von mindestens einer strukturell variablen Position zwischen diesen beiden Torsionswinkeln. Dies müßte bei der Datenauswertung berücksichtigt werden, was in der Folge die Errechnung höherdimensionaler Dichtefunktionen zur Folge hätte. Dafür wäre ein ungleich größerer Rechenaufwand notwendig, der gleichzeitig mit einem Verlust an Anschaulichkeit der erzielten Ergebnisse verbunden wäre.

V.1. Datenaufbereitung und Ergebnisse der Datenanalyse

Statistische Analysen von Proteineigenschaften werden im Allgemeinen mit nichtredundanten Sequenzdatenbanken von Proteinstrukturen durchgeführt [Melo *et al.*, 2002]. Die geringe Datenvielfalt innerhalb dieser Datenbanken ist das fundamentale Hindernis bei der Aufdeckung von Sequenz-Struktur-Korrelationen von Fragmenten [Solis & Rackovsky, 2002]. Besonders die Strukturanalyse von Peptidfragmenten ab einer Länge von vier Aminosäuren ist deshalb nicht oder nur sehr eingeschränkt möglich [Anishetty *et al.*, 2002]. Der Schlüssel zu einem weitergehenden Verständnis der Konformation von Fragmenten mit einer Mindestanzahl von vier Aminosäuren war die Entwicklung einer Methodik zur Datenaufbereitung, die in der Lage war, die vorhandenen Strukturinformationen von Proteinen besser zu nutzen, ohne gleichzeitig die Nichtredundanz-Bedingung zu verletzen. Der Lösungsansatz zu dieser Problemstellung bestand in der Überlegung, dass eine Menge von Objekten nur dann verglichen werden sollte, wenn sie ein bestimmtes Attribut, nämlich das zu untersuchende, gemeinsam haben. Die Nichtredundanz-Bedingung wird durch die Beseitigung redundanter Informationen jeweils innerhalb der Gruppen, die ein bestimmtes Attribut beinhalten, erfüllt. Aminosäuresequenzen von Proteinstrukturen sind danach nur als Unterklassen von Objekten und Tetrapeptide als deren Eigenschaften (Attribute) zu betrachten. Mit der Anwendung dieser Vorschrift konnten 93.3 % der in der *PDB* vorhandenen Strukturen analysiert werden. Im Unterschied dazu, würden bei der Verwendung eines nichtredundanten Datensatzes der *PDB* nur 9.8 % des Gesamtdatensatzes berücksichtigt. Der hohe Informationsgewinn, wie er durch die veränderte Datenaufbereitung erzielt wurde, führte bei 13.8 % der Tetrapeptide zu einer Änderung des bevorzugten Konformationszustandes.

Dieses Ergebnis impliziert, dass die beobachtete strukturelle Präferenz bei Tetrapeptiden nicht ihre Ursache in einer ungenügenden Datenbasis hat, sondern auf physikalische Besonderheiten innerhalb des jeweiligen Tetrapeptids zurückzuführen ist. Diese Vermutung wird durch die Arbeiten anderer Autoren gestützt, die experimentell bereits bei Tripeptiden bevorzugte Strukturen nachweisen konnten [Eker *et al.*, 2002, Motta *et al.*, 2005]. Es ist deshalb nicht davon auszugehen, dass eine weitere Erhöhung der Datenbasis zu einer Nivellierung der statistisch beobachteten bevorzugten Strukturbildung führt. Vielmehr würde die Berücksichtigung weiterer Proteinstrukturen die Datenvielfalt von seltenen Tetrapeptiden vergrößern.

Die Strukturen der errechneten Wahrscheinlichkeitsdichtefunktionen sind das Ergebnis einer Vielzahl von Parametern, die zu Beginn der Datenaufbereitung definiert wurden. Diese Parameter beinhalteten die Auswahl der Proteinstrukturen, den verwendeten Alignmentalgorithmus (Semiglobales Alignment nach Needleman-Wunsch [Needleman & Wunsch, 1970] unter Verwendung affiner *gap*-Strafen [Gotoh, 1982]), die maximale Sequenzidentität (25 %), die Werte der *gap*-Strafen (*open penalty* = -5, *extension penalty* = -2), die verwendete Substitutionsmatrix (BLOSUM62 [Henikoff & Henikoff, 1992]), die Durchführung des primären *all-against-all* Alignments, die Berücksichtigung von Strukturunterschieden zwischen Tetrapeptiden, die Bandweite von $h = 15°$ und die Normierung der Dichtefunktionen. Die optimalen Parameter lassen sich streng logisch nicht ableiten, sondern können nur experimentell verifiziert werden. In Abhängigkeit von den Ergebnissen müssen die Parameter angepasst oder das Modell verworfen werden. Die Analyse des wahrscheinlichsten Konformationszustandes der einzelnen Dichtefunktionen zeigte, dass 94.4 % der untersuchten Tetrapeptide ein strukturell ambivalentes Verhalten bezüglich der Einteilung in die Konformationsbereiche *E*, *H*, *L* und *X* aufweisen. Dieses Resultat steht im Einklang mit den Ergebnissen vieler Autoren, dass sequentiell identische Peptidfragmente unterschiedlicher Länge in verschiedenen Proteinen ungleiche Konformationen annehmen [Kabsch & Sander, 1984; Rackovsky, 1995; Sudarsanam & Srinivasan, 1997; Zhou *et al.*, 2000; Kuznetsov & Rackovsky, 2003]. Es konnte bis heute kein eindeutiges Muster identifiziert werden, auf dass diese Eigenschaft zurückzuführen ist [Mezei, 1998]. Dennoch wurde in der vorliegenden Arbeit eine hohe strukturelle Präferenz bei Tetrapeptiden nachgewiesen, die im Widerspruch zu den Arbeiten von Zhou *et al.* stehen, die bei Tetra- und Pentapeptidfragmenten eine maximale Flexibilität festgestellt haben [Zhou *et al.*, 2000]. Die erzielten Ergebnisse bestätigen vielmehr die Arbeiten von Rackovsky, der mit einem reduzierten Alphabet von Aminosäuren in gleicher Weise eine ausgeprägte strukturelle Präferenz bei Tetrapeptiden nachgewiesen und damit auf die Existenz eines lokalen inversen Faltungscodes bei diesen Fragmenten geschlossen hat [Rackovsky, 1995]. In gleicher Weise konnte die vorliegende Arbeit die Ergebnisse von Sudarsanam und Srinivasan bestätigen und erweitern. Die von diesen Autoren durchgeführten Konformationsbetrachtungen des ψ_2- und des φ_3-Diederwinkels von Tetrapeptiden offenbarten, unter Verwendung einen reduzierten Alphabets von Aminosäuren, strukturelle Präferenzen bei diesen Fragmenten [Sudarsanam & Srinivasan, 1997]. Die nur in Einzelfällen von diesen Autoren durchgeführten Analysen sequenzspezifischer Tetrapeptide zeigten ebenso ausgeprägte strukturelle Präferenzen, wie sie in der vorliegenden Arbeit gefunden wurden.

Die Konformationsanalyse der zweiten und dritten Aminosäure von Tetrapeptiden hatte gezeigt, dass sich deren Konformationen nicht unabhängig voneinander bilden, es demnach nur eine bestimmte Anzahl an verschiedenen lokalen Konformationen geben kann. Die strukturelle

Abhängigkeit benachbarter Aminosäuren wurde ebenso von anderen Autoren beschrieben [Zaman 1999, Pappu 2003, Betancourt 2004] und auf sterische Effekte zurückgeführt.

Ein sehr interessantes Ergebnis ergab sich aus der Analyse der strukturellen Eigenschaften von sequenzähnlichen Tetrapeptiden. Am Beispiel von AMDY wurde gezeigt, dass ähnliche Tetrapeptide keinen gemeinsamen bevorzugten Konformationszustand besitzen müssen. Dies wurde bereits früh erkannt [Sternberg & Islam, 1990] und impliziert, dass jede Vereinfachung einer Aminosäuresequenz durch ein reduziertes Alphabet von Aminosäuren oder ein HP-Motiv immer mit einem Informationsverlust hinsichtlich der Konformationseigenschaften eines spezifischen Fragmentes verbunden ist. Dass es einen Zusammenhang zwischen Sequenzähnlichkeit und Strukturähnlichkeit gibt, bleibt dennoch unbestritten, wie die erfolgreiche Anwendung von Substitutionsmatrizen, wie BLOSUM62 [Henikoff & Henikoff, 1992], in verschiedenen Alignmentverfahren wie z. B. BLAST [Altschul *et al.*, 1990] oder PSIBLAST [Altschul *et al.*, 1997] zeigt. Inwieweit eine Verallgemeinerung von Strukturinformation auf Tetrapeptidsequenzebene im Sinne bevorzugter Strukturen bei diesen Fragmenten möglich ist und wie ausgeprägt diese (bei anderen Tetrapeptiden) ist, könnte nach einer Konformationsanalyse innerhalb aller Gruppen sequenzähnlicher Tetrapeptide formuliert werden. Selbst bei dem Nachweis eines solchen Zusammenhangs ist zu erwarten, dass die Information über die individuellen strukturellen Präferenzen einzelner Tetrapeptide verloren geht. Dies konnte am Beispiel der ähnlichen Tetrapeptide VEYT, VDYT und IDFS gezeigt werden (vgl. Abbildung IV-7 auf Seite 46). Bei allen drei Tetrapeptiden wurde der faltblatttypische Konformationszustand als der wahrscheinlichste identifiziert. Trotz sehr hoher Präferenzen für diesen Konformationszustand zeigten sich große Abweichungen in den wahrscheinlichsten Konformationen zwischen diesen Tetrapeptiden. Der Informationsverlust hinsichtlich spezifischer Konformationen bei Tetrapeptiden nach Verwendung eines reduzierten Alphabets von Aminosäuren wurde ebenso von Sudarsanam und Srinivasan beschrieben [Sudarsanam & Srinivasan, 1997]. Die Autoren beobachteten, dass eine Konkretisierung von Aminosäuren zu einer restriktiveren Verteilung der ψ_2- und φ_3-Winkel bei Tetrapeptiden führte. Besonders im Hinblick auf das in dieser Arbeit vorgestellte Modellierungsschema sind Vereinfachungen jedoch nicht verwendbar, da durch die Zielstruktur genaue Konformationen vorgegeben werden und anschließend versucht wird, diese mit der wahrscheinlichsten Konformation spezifischer Tetrapeptide nachzubilden.

V.2. Modellierungsschema

Für ein Protein mit einer Anzahl von *n* Aminosäuren ergeben sich 20^n verschiedene Aminosäuresequenzen [Zou & Saven, 2000]. Das einfache Ausprobieren von Aminosäuresequenzen zu einer Tertiärstruktur kann deshalb aufgrund der großen Anzahl möglicher Sequenzen nicht durchgeführt werden. Derzeit werden sehr erfolgreich heuristische Verfahren zur Modellierung von Aminosäuresequenzen zu vorgegebenen Proteinstrukturen verwendet [Koehl & Levitt, 1999]. Im Unterschied dazu verwendet das vorgestellte Modellierungsschema einen deterministischen Ansatz, bei dem Tetrapeptide in linearer Abfolge verknüpft werden. Danach soll eine Aminosäuresequenz in die Zielstruktur falten, wenn jedes einzelne Tetrapeptid eine hohe Wahrscheinlichkeit für den Zielkonformationszustand bzw. die Zielkonformation besitzt. Bereits Holmes und Tsai vermuteten, dass das Verknüpfung von Fragmenten in ihrer bevorzugten Konformation zu einer nativen Struktur führen kann [Holmes & Tsai, 2004]. Mit der durchgeführten statistischen Analyse bevorzugter Strukturen von Oligopeptidfragmenten konnte das Verständnis für die Strukturbildung bei Fragmenten vertieft werden. Bei Fragmenten mit einer Länge von 6-16 Aminosäuren und ab einem geometrischen Mittel der Wahrscheinlichkeit des wahrscheinlichsten Konformationszustandes von ca. 70 % der einzelnen Tetrapeptide, wurde die wahrscheinlichste Struktur dieser Fragmente häufiger beobachtet, als statistisch zu erwarten gewesen wäre. In diesen Fällen erfolgte vermehrt eine Strukturbildung hin zum wahrscheinlichsten Konformationszustand jedes einzelnen Tetrapeptids. Es muß an dieser Stelle noch einmal darauf hingewiesen werden, dass die entsprechenden statistischen Analysen Gruppen von Sequenzen umfassten, die nur gleiche mittlere Wahrscheinlichkeiten der wahrscheinlichsten Konformationszustände der einzelnen Tetrapeptide aufwiesen. Diese Gruppen umfassten *unterschiedliche* Aminosäuresequenzen. Inwieweit bei einer *spezifischen* Sequenz eine Tendenz zur einer abhängigen Strukturbildung vorhanden ist, läßt sich nicht vorhersagen, da für Oligopeptide unterschiedlicher Sequenz mit einer Länge von mehr als vier Aminosäuren keine ausreichenden statistischen Informationen hinsichtlich ihrer bevorzugten Struktur vorliegen. Trotz dieser Unsicherheit kann man hier die Schlussfolgerung ziehen, dass ein Verfahren, welches versucht, eine Sequenz mit hohen *a priori* Wahrscheinlichkeiten der einzelnen Tetrapeptide für den jeweiligen Zielkonformationszustand zu erzeugen, bei der Modellierung einer Zielstruktur erfolgreich sein könnte.

Die strukturelle Ambivalenz sequenzidentischer Peptide wird von verschiedenen Autoren als ein Hindernis für die Verwendung von Fragmenten zum Proteindesign diskutiert. Sudarsanam kam nach seiner Beobachtung, dass sequenzidentische Peptide bis zu einer Länge von neun Aminosäuren verschiedene Konformationen annehmen können zu dem Schluss, dass für ein fragmentbasiertes Proteindesign die Verwendung längerer Peptide in Betracht gezogen werden muß [Sudarsanam, 1998]. Hu wies darauf hin, dass eine beginnende Sequenz-Struktur-Spezifität bei Oligopeptidfragmenten erst ab einer Länge von 15-20 Aminosäuren zu beobachten ist, im Allgemeinen jedoch 40 % der Gesamtaminosäuresequenz bekannt sein müssen, damit ein Fragment seine native Konformation erkennt [Hu *et al.*, 1997]. Durch eine Kombination von Tetrapeptiden mit hohen *a priori* Wahrscheinlichkeiten für den jeweiligen Zielkonformationszustand können dennoch, wie bereits dargelegt, Sequenzen erzeugt werden, die eine erhöhte Wahrscheinlichkeit zu einer abhängigen Strukturbildung aufweisen. Ebenso scheint damit eine Verwendung von Protein *building blocks* mit einer Mindestlänge von 15-20 Aminosäuren, wie von Tsai und Mitarbeitern vorgeschlagen, nicht mehr notwendig zu sein [Tsai *et al.*, 2004].

Inwieweit Fragmente mit einer erhöhten Wahrscheinlichkeit zu einer abhängigen Strukturbildung ihre wahrscheinlichste Konformation auch in isolierter Form ausbilden, lässt sich ebenso wie bei isolierten Tetrapeptiden an dieser Stelle nicht vorhersagen. Ho und Dill konnten jedoch mit Hilfe von Molekulardynamiksimulationen zeigen, dass Oktapeptide, die aus Proteinstrukturen entlehnt wurden, ihre native Konformation ebenso in isolierter Form annehmen können [Ho & Dill, 2006]. Dieses Verhalten wurde jedoch nur in 35 % der untersuchten Fälle beobachtet. Die Autoren führen die bevorzugte Strukturbildung in Abwesenheit von Tertiärstrukturwechselwirkungen ausschließlich auf Seitenkettenbehinderungen innerhalb der betrachteten Fragmente zurück. Unter Berücksichtigung der experimentellen Ergebnisse, dass bereits Tripeptide in wässriger Lösung stabile Konformationen annehmen können [Eker *et al.*, 2002; Motta *et al.*, 2005], wäre eine strukturelle Präferenz isolierter Fragmente, bei denen eine erhöhte Wahrscheinlichkeit für eine bedingte Strukturbildung vorliegt, ebenso denkbar. Dies müsste jedoch experimentell überprüft werden. Von verschiedenen Autoren konnte diesbezüglich gezeigt werden, dass kleine Peptide und aus Proteinen isolierte Fragmente in wässriger Lösung bevorzugte Konformationen annehmen können [Baldwin & Rose, 1999a].

Die hohen *a priori* Wahrscheinlichkeiten bei einer Vielzahl von Tetrapeptiden und die gefundene abhängige Strukturbildung bei Fragmenten zeigen, dass ein Gleichgewicht aller erlaubten Strukturen - zumindest im Kontext einer Gesamtaminosäuresequenz - nicht existiert. Das Ungleichgewicht der möglichen Strukturen würde in diesem Fall in direkter Beziehung zu der Überlegung stehen, dass die statistische Suche einer Polypeptidkette durch alle Konformationen nach ihrer nativen Konformation auszuschließen ist [Levinthal, 1968]. Bereits von Zwanzig und Mitarbeitern durchgeführte theoretische Überlegungen zeigten, dass schon ein geringer *Bias* im Faltungspotential ein sehr schnelles Auffinden einer stabilen Struktur zur Folge hat [Zwanzig *et al.*, 1992] und das *Levinthal Paradoxon* [Levinthal, 1968] damit umgangen werden kann [Fetrow *et al.*, 2002]. Die in der vorliegenden Arbeit gefundene bedingte Strukturbildung könnte sich nun als ein Hinweis auf einen solchen *Bias* interpretieren lassen.

Bei Pentapeptiden und längeren Fragmenten mit mittleren Wahrscheinlichkeiten von kleiner als 0.7 für den wahrscheinlichsten Konformationszustand der einzelnen Tetrapeptide, wurde die wahrscheinlichste Konformation der untersuchten Fragmente nur mit einer Häufigkeit - oder seltener - beobachtet, die der statistischen Erwartung entspricht. Dies lässt auf eine strukturelle Unabhängigkeit der einzelnen Tetrapeptide schließen. Bei den erzielten Ergebnissen ist zu beachten, dass die untersuchten Fragmente *per definitionem* ein korrektes hydrophobes Muster aufwiesen. Die binäre Kodierung einer Proteinstruktur mit einem hydrophoben Muster alleine scheint somit nicht ausreichend, um Aminosäuresequenzen zu erzeugen, die nicht nur zufällig in die Zielstruktur falten, was in Konsistenz zu den Arbeiten anderer Autoren steht [Yue & Dill, 1992] und damit eine Anwendung von *high throughput* Methoden, wie kombinatorischer Bibliotheken [Kamtekar *et al.*, 1993; Rojas *et al.*, 1997; Moffet *et al.*, 2000; Wei *et al.*, 2003], sinnvoll erscheinen lässt. An dieser Stelle muss angemerkt werden, dass dem Begriff *unbhängige Strukturbildung* nicht wirklich eine zufällige Strukturbildung der betroffenen Fragmente unterliegt. Vielmehr ist das statistische Modell, welches zur Beschreibung der Strukturbildung in Oligopeptiden in dieser Arbeit aufgestellt wurde, nicht mehr in der Lage, eine bevorzugte Strukturbildung zu erkennen. Dennoch ist davon auszugehen, dass eine solche Präferenz auch weiterhin existiert, sie aber mit komplexeren Analysen untersucht werden muß.

Ein korrekt definiertes hydrophobes Muster ist dennoch wichtig für den Erfolg der Modellierung einer Aminosäuresequenz zu einer (globulären) Proteinstruktur, da der hydrophobe Effekt als eine Haupttriebkraft der Proteinfaltung angesehen wird [Dill, 1990; Dill, 1999]. Der mögliche Sequenzraum, den eine Aminosäuresequenz der Länge *N* mit einem bestimmten hydrophoben Muster beschreiben kann, verringert sich damit bei einer Anzahl *n* an hydrophoben Aminosäuren von 20^N auf $H^n \cdot (20-H)^{(N-n)}$ mögliche Sequenzen. *H* definiert hierbei in der Gruppe der 20 proteinogenen Aminosäuren die Anzahl an hydrophoben Aminosäuren. Damit ergeben sich zum Beispiel für Top7 bei 33 hydrophoben Aminosäuren in der Sequenz eine Anzahl von $8^{33} \cdot 12^{92-33} = 2.9 \cdot 10^{93}$ möglichen Aminosäuresequenzen, die theoretisch in die Zielstruktur falten können. Durch *phage display* oder andere kombinatorische Methoden können nur 10^7-10^8 verschiedene Sequenzen *gescreent* werden [Zou & Saven, 2000], was nur ein geringer Bruchteil der tatsächlich möglichen Varianten darstellt. Ein geeignetes Modellierungsverfahren zur Berechnung von alternativen Aminosäuresequenzen für gegebene Proteinstrukturen muß aus dieser großen Anzahl diejenigen finden, die mit der größten Wahrscheinlichkeit in die Zielstruktur falten. Das zentrale Problem bei dem Design von Aminosäuresequenzen ist die Etablierung von Energiefunktionen, die eine Sequenz oder Sequenzen identifizieren können, deren Konformation mit der niedrigsten Freien Energie die Zielstruktur ist und die eine große Energiedifferenz zu konkurrierenden, falsch gefalteten Strukturen niedriger Energie besitzen [Yue & Dill, 1992; England *et al.*, 2003]. Zu diesem Zweck wurden Energiefunktionen mit einer Vielzahl von Parametern etabliert und optimiert [Kuhlman & Baker, 2000; Liang & Grishin, 2004; Butterfoss & Kuhlman, 2006]. Die Errechnung von alternativen Aminosäuresequenzen zu einer gegebenen Proteinstruktur unter Zuhilfenahme dieser Funktionen durch verschiedene Autoren führte in der Folge zu Sequenzen, die eine Identität von fast 100 % im Bereich der hydrophoben Kerne [Wernisch *et al.*, 2000] oder im allgemeinen hohe Sequenzidentitäten, im Vergleich zu den Wildtypsequenzen aufwiesen [Liang & Grishin, 2004]. Kuhlman und Baker erzielten Sequenzidentitäten von 55 % im hydrophoben Kern der modellierten Proteine im Vergleich zu den Wildtypsequenzen [Kuhlman & Baker, 2000]. Die Autoren schlussfolgerten deshalb, dass native Aminosäuresequenzen ihre Proteinstruktur nahezu optimal kodieren [Kuhlman & Baker, 2000] und die etablierten Energiefunktionen offenbar sehr gut in der Lage sind, native Aminosäuresequenzen zu erkennen. Bei dem Design von Energiefunktionen besteht die Gefahr einer Überanpassung an existierende Proteinstrukturen und deren Sequenzen. In der Folge können die vollautomatischen Methoden daher zu Lösungen konvergieren, die Aminosäuresequenzen beschreiben, welche wildtypähnlich sind, was die Ergebnisse von Wernisch *et al.* bzw. Kuhlman und Baker erklären könnte. Die Gefahr der Berechnung von Sequenzen mit einem *Bias* auf den Wildtyp besteht bei dem in dieser Arbeit vorgestellten System nicht, da die errechneten Wahrscheinlichkeitsdichtefunktionen der ψ_2-φ_3-Verteilungen keine Tertiärstrukturinformationen über die Proteine beinhalten, aus denen die jeweiligen Fragmente stammen und damit unbekannt ist, ob ein gewähltes Fragment seine wahrscheinlichste Konformation auch schon einmal in dem durch die Zielstruktur vorgegebenen lokalen und globalen Kontext angenommen hat.

Die Modellierung der Seitenketten zu den berechneten Sequenzen erfolgte in der vorliegenden Arbeit bei konstanter Proteinrückgratgeometrie. Park und Mitarbeiter haben darauf hingewiesen, dass diese Strategie zu einem *Bias* auf bestimmte Sequenzen führen kann [Park *et al.*, 2004]. Die Berücksichtigung einer Flexibilität im Proteinrückgrat, so die Autoren, würde zu einer größeren Sequenzvariabilität führen, da auf diese Weise sterische Behinderungen ausgeglichen werden

können. Dies ist jedoch mit einem deutlich größeren Rechenaufwand [Park *et al.*, 2004] bei der Modellierung verbunden. Ein solches Verfahren wurde erfolgreich bei dem *Redesign* einer WW-Domäne [Kraemer-Pecore *et al.*, 2003] bzw. bei dem Design der Aminosäuresequenz von Top7 [Kuhlman *et al.*, 2003] angewendet. Das vorgestellte Modellierungssystem besitzt in der Weise einen inhärenten *Bias* auf bestimmte Sequenzen, dass im Idealfall nur Tetrapeptide selektiert werden, die eine sehr hohe Wahrscheinlichkeit für den Zielkonformationszustand besitzen. In gleicher Weise ist zu erwarten, dass der Ausschluss der Aminosäuren Histidin, Cystein und Prolin bzw. Phenylalanin und Tryptophan zu einer erheblichen Einschränkung des theoretischen Sequenzraumes geführt hat.

Die Selektion der einzelnen Modelle erfolgte nach maximaler Wahrscheinlichkeit unter der Bedingung, dass nach optimaler Ausrichtung der Reste keine sterischen Behinderungen zwischen verschiedenen Seitenketten zu beobachten waren. Die errechneten Aminosäuresequenzen sind damit das Ergebnis eines subjektiven Modellierungsprozesses, was gleichzeig bedeutet, dass es *die eine* richtige Lösung für ein Modellierungsproblem nicht geben kann, sondern eine Lösung sich in Abhängigkeit von den gewählten Randbedingungen gestaltet. Dies ist jedoch kein Nachteil, da mit diesem Schema ein echtes rationales Proteindesign ermöglicht wird. Das vorgestellte Modellierungsschema steht in Kontrast zu den vollautomatischen Methoden, die nahezu alle Monte-Carlo-Algorithmen verwenden, deren Energiefunktionen aus einer Vielzahl von Termen bestehen [Koehl & Levitt, 1999; Kuhlman & Baker, 2000; Wernisch *et al.*, 2000; Jaramillo *et al.*, 2001]. Die mit Hilfe von *RosettaDesign* errechnete Sequenz von Top7 ist das Ergebnis der Monte-Carlo-Optimierung einer randomisierten Aminosäuresequenz zu einer gegebenen Proteinstruktur [Kuhlman *et al.*, 2003]. Die Modelle wurden mit einer Energiefunktion aus 11 Termen bewertet. Im Folgenden wird gezeigt, in welcher Weise sich die Randbedingungen des vorgestellten Modellierungsschemas durch die Energieterme von *RosettaDesign* beschreiben ließen. Die Bezeichnung der einzelnen Terme (hervorgehoben) wurde dem *supplement* der Originalpublikation [Kuhlman *et al.*, 2003] entnommen.

Lennard-Jones Potential (E_{atr} and E_{rep})
Bei dem Lennard-Jones Potential erfolgte durch *RosettaDesign* die Optimierung der anziehenden atomaren Kräfte unter gleichzeitiger Kontrolle der abstoßenden Kräfte, die infolge der Überlappung verschiedener Seitenketten auftreten können.

Mit dem vorgestellten Modellierungsschema erfolgt keine automatische Optimierung anziehender atomarer Kräfte. Eine günstige Verteilung von Aminosäuren lässt sich qualitativ durch eine Definition von Randbedingungen festlegen. Die Kontrolle der Seitenkettengeometrie erfolgte in dieser Arbeit durch deren Energieminimierung mit der Implementierung des GROMOS96 Kraftfeldes [van Gunsteren *et al.*, 1996] im *Swiss-PdbViewer 3.7* (*SP5*) [Guex & Peitsch, 1997]. Bei Überlappungen von Seitenkettenatomen wurden die entsprechenden Modelle verworfen. Die erzielte Gesamtenergie stellte insofern kein Selektionskriterium dar, als bei fehlenden Überlappungen von Seitenkettenatomen die errechnete Sequenz eine gültige Lösung des Modellierungsproblems war.

Lazaridis-Karplus solvation model (E_{solv})
RosettaDesign bewertet die Solvatationsenergie des modellierten Proteins nach Lazaridis-Karplus [Lazaridis & Karplus, 1999]. Die Bewertung erfolgt in atomarer Auflösung ausgehend von 24 verschiedenen Atomtypen (z. B. aliphatisches Kohlenstoffatom mit zwei

Wasserstoffatomen). Dieser Term bevorzugt primär apolare Aminosäuren im hydrophoben Kern und polare Aminosäuren auf der lösungsmittelexponierten Oberfläche.

Im vorgestellten Modellierungsschema erfolgte, ausgehend von der Zielstruktur, eine manuelle Definition des hydrophoben Musters. Die Verteilung der polaren und apolaren Aminosäuren erfolgte in subjektiver Weise. Dabei wurden vorwiegend apolare Aminosäuren für die Modellierung des hydrophoben Kerns verwendet und primär polare Aminosäuren auf der lösungsmittelexponierten Oberfläche definiert. Es erfolgte keine Betrachtung der Solvatationsenergie individueller Aminosäuren.

Rotamer Self-energy (E_{rot})
RosettaDesign bewertet explizit die jeweiligen Rotamerenergien der Seitenketten unter Zuhilfenahme der Rotamerbibliothek von Dunbrack und Cohen [Dunbrack, Jr. & Cohen, 1997]. Änderungen von Aminosäuren können zu günstigeren Konformationen von Seitenketten führen, was in der Folge zu einer niedrigeren Gesamtenergie führt.

Die Seitenketten wurden an das Proteinrückgrat der Modelle M1-M8 mit dem Programm *SCWRL3* modelliert [Canutescu *et al.*, 2003], welches ebenso die Rotamerbibliothek von Dunbrack und Cohen verwendet. Es erfolgte jedoch keine Bewertung der erzielten Konformationsenergie. Mögliche abstoßende Wechselwirkungen aufgrund überlappender Reste werden anschließend durch die Energieterme der Energieminimierung erkannt. Sind diese nicht vorhanden, so war die modellierte Aminosäuresequenz eine gültige Lösung des Modellierungsprozesses.

Amino acid preferences for particular regions of phi, psi space ($E_{aa|phi,psi}$)
RosettaDesign bewertet die Wahrscheinlichkeit für jede der 20 proteinogenen Aminosäuren, eine Konformation (φ,ψ) im möglichen Konformationsraum anzunehmen.

Die errechneten Dichtefunktionen der ψ_2-φ_3-Verteilungen erlauben keine Aussagen über bevorzugte Konformationen einer Einzelaminosäure. Die Diederwinkel ψ und φ einer Aminosäure ergeben sich durch Überlappung der einzelnen Tetrapeptide. Inwieweit ein bestimmtes Tetrapeptid in der Lage ist, einen Konformationszustand anzunehmen, der nicht durch die entsprechende Dichtefunktion beschrieben wird, lässt sich mit dem vorgestellten System nicht vorhersagen.

Amino acid dependent torsion potential for phi and psi (E_{rama})
Für jede der 20 Aminosäuren in den drei Sekundärstrukturbereichen Helix, Faltblatt oder *coil* wurde die Häufigkeit von (ψ,φ)-Winkeln für die jeweilige Sekundärstruktur bestimmt. Die Wahrscheinlichkeiten wurden nach Addition von *Pseudocounts* in eine Energie umgerechnet. Die von *RosettaDesign* verwendete Einteilung in die Sekundärstrukturbereiche erfolgte mit dem Programm *DSSP* [Kabsch & Sander, 1983].
Die in der vorliegenden Arbeit verwendeten Dichtefunktionen „kennen“ keine Sekundärstrukturen. Ein im Modellierungsprozess befindliches Tetrapeptid muss nur eine Kompatibilität mit der Zielstruktur aufweisen, um eine valide Lösung des Modellierungsproblems zu sein. Die Wahrscheinlichkeiten der Zielkonformationszustände bzw. Zielkonformationen entsprechen beobachteten Häufigkeiten.

Residue pair potential (E_{pair})
RosettaDesign bewertet die räumliche Umgebung der einzelnen Aminosäuren. Diese Bewertung entspricht den Kontaktpotentialen von Sippl [Sippl, 1993; Sippl, 1995].

Das definierte hydrophobe Muster und die als Randbedingungen definierten Aminosäuren beschreiben qualitativ die Umgebung jeder Aminosäure. Eine quantitative Analyse des *residue pair potential* lässt sich jedoch mit Programmen wie *whatcheck* [Hooft *et al.*, 1996] oder *ANOLEA* [Melo & Feytmans, 1997; Melo *et al.*, 1997; Melo & Feytmans, 1998] durchführen. Die Evaluierung der Modelle mit beiden Programmen zeigte, dass durch die manuelle Definition der Aminosäuren auf den ausgewählten Positionen im Ergebnis eine sehr gute globale Verteilung der Aminosäuren erreicht wurde. (Die Daten für *ANOLEA* wurden nicht gezeigt)

Orientation-dependent hydrogen bonding term (E_{bb_hbond}, E_{sc_hbond}, $E_{bb\_sc\_hbond}$)

Die Energie der Wasserstoffbrücken zwischen den Proteinrückgratatomen (E_{bb_hbond}), zwischen verschiedenen Aminosäureseitenketten und zwischen Proteinrückgratatomen ($E_{bb\_sc\_hbond}$) und Aminosäureseitenkettenatomen (E_{sc_hbond}) wurden explizit durch eine Energiefunktion bewertet [Kortemme *et al.*, 2003].
Das Wasserstoffbrückenbindungsmuster der Sekundärstrukturelemente und die Geometrie der entsprechenden Donor- und Akzeptoratome ist implizit durch das Proteinrückgrat der Zielstruktur festgelegt. Wasserstoffbrückenbindungen zwischen Seitenkettenatomen und anderen Atomen wurden nicht modelliert oder bewertet.

Energy of the unfolded state (E_{ref})

Um die Energie des entfalteten Zustandes zu bestimmen, verwendet *RosettaDesign* für jede Aminosäure eine empirisch ermittelte Referenzenergie. Während der Optimierung der Aminosäuresequenz sollte damit eine möglichst niedrige Energie der Sequenz für die native Struktur erreicht werden.

Das vorgestellte Modellierungsschema verwendete keine energetische Betrachtung des entfalteten Zustandes des Proteins.

Die errechneten Sequenzen der Modelle M1-M8 zeigten in allen Fällen eine Sequenzidentität von kleiner als 30 % gegenüber der Sequenz von Top7. Die Sequenzähnlichkeit bezüglich der BLOSUM62-Matrix [Henikoff & Henikoff, 1992] ergab sich durchschnittlich zu 60 %. Eine Sequenzähnlichkeit impliziert jedoch keine strukturelle Ähnlichkeit, wie am Beispiel des Tetrapeptids AMDY gezeigt wurde. Mit Hilfe der errechneten Dichtefunktionen würden sich Aminosäuresequenzen errechnen lassen, die ebenso sequenzähnlich zu Top7 sind, wie die Modelle M1-M8, die jedoch nur eine geringe Wahrscheinlichkeit für die Zielkonformation aufweisen. Unter Berücksichtigung der Ergebnisse aus der Fragmentanalyse der *FSSP*-Datenbank kann dabei nicht vorhergesagt werden, ob die entsprechenden Sequenzen in die Zielstruktur falten. Die Sequenzidentitäten der hydrophoben Kerne der einzelnen Modelle zu Top7 liegen bei dem definierten *cut off* (< 5 % zugängliche Oberfläche der Aminosäure) mehrheitlich unter 50 %, bei durchschnittlich zehn identischen hydrophoben Aminosäuren. Dies zeigt, dass die geringen Sequenzidentitäten nicht auf eine Variation der lösungsmittelexponierten Aminosäuren zurückzuführen sind, sondern ebenso eine Restrukturierung der hydrophoben Kerne durchgeführt werden konnte.

Die optimale Sequenz zu einer gegebenen Proteinstruktur entspräche nach dem vorgestellten Modellierungsprinzip derjenigen Sequenz, in der jedes Tetrapeptid mit seinem wahrscheinlichsten Konformationszustand die Zielstruktur beschreibt. Weiterhin müsste die

Wahrscheinlichkeit der Zielkonformation $(\psi_2,\varphi_3)_i$ des *i*-ten Tetrapeptids identisch mit der wahrscheinlichsten Konformation sein und es müsste eine optimale Verteilung von Aminosäuren gewährleistet sein. Dieses Ziel wurde bei keinem der acht Modelle mit dem vorgestellten Modellierungsschema vollständig erreicht. Das Entscheidungskriterium zwischen den Konformationszuständen hat sich bei der Selektion geeigneter Tetrapeptide für die Berechnung der Aminosäuresequenzen als nicht optimal erwiesen. Im Ergebnis führt diese Einteilung zu der Situation, dass die Änderung eines Winkels um 2° zu einem anderen Konformationszustand führen kann. An einigen Stellen konnten aus diesem Grund in den modellierten Aminosäuresequenzen formal nur geringe Wahrscheinlichkeiten für den Zielkonformationszustand erreicht werden. Eine Einteilung des Konformationsraumes in einzelne Konformationsbereiche bleibt für eine erfolgreiche Modellierung dennoch notwendig. Die Definition eines Faktors in den Dichtefunktionen im Sinne einer Abklingfunktion, der beispielsweise beschreibt, wie faltblatttypisch eine Konformationszustand noch ist, obwohl er durch den helixtypischen Konformationszustand beschrieben wird, sollte die Errechnung der Aminosäuresequenzen deutlich vereinfachen und zu Sequenzen mit höheren Wahrscheinlichkeiten für die Zielstruktur führen. Bei der Errechnung der Dichtefunktionen wurde eine Bandweite von $h = 15°$ definiert, um eine strukturelle Variabilität (zusätzliche zu der beobachteten) der einzelnen Tetrapeptide zuzulassen. Dennoch konnte an einigen Positionen in den Modellen die exakte Zielkonformation in nur sehr ungenügender Weise beschrieben werden, obwohl das verwendete Tetrapeptid eine sehr hohe Wahrscheinlichkeit für den Zielkonformationszustand aufwies. Inwieweit eine spezifische Konformation in einem Konformationsbereich zugänglich ist, für den ein bestimmtes Tetrapeptid eine hohe Wahrscheinlichkeit besitzt, lässt sich ohne eine detaillierte Analyse nicht vorhersagen. Im Sinne des vorgestellten Modellierungsschemas sind diese Sequenzabschnitte als fehlerhaft zu bewerten, was jedoch nicht auf einen Mangel geeigneter Tetrapeptide zurückzuführen war, sondern auf die definierten Randbedingungen. Die Analyse der Originalsequenz von Top7 hatte ebenso das Vorhandensein von Tetrapeptiden offenbart, die zwar eine hohe Wahrscheinlichkeit für den Zielkonformationszustand aufwiesen, aber gleichzeitig nur eine sehr geringe Wahrscheinlichkeit für die Ausbildung der exakten Zielkonformation zeigten. Besonders im faltblatttypischen Konformationszustand, der in den Modellen M1-M8 und bei Top7 am häufigsten betroffen war, ist es möglich, dass diese Fehler dennoch nicht zu gravierenden Störungen in der Proteinstruktur führen, wenn, wie von Hu & Dill argumentiert, die bevorzugten Konformationseigenschaften von Oligopeptiden primär auf Seitenkettenbehinderungen zurückzuführen sind [Ho & Dill, 2006]. Im Konformationszustand *E* sind die Konformationseinschränkungen der Seitenketten minimal, wie der Vergleich zu den Resten im helikalen Konformationszustand eines Tetrapeptids zeigt (vgl. Abbildung IV-4 auf Seite 41).

An dieser Stelle muß kritisch festgehalten werden, dass das vorgestellte, semiautomatische Modellierungssystem durch die Notwendigkeit einer manuellen Anpassung der Modelle komplizierter ist, als vollautomatische Modellierungssyteme. Besonders die Positionierung geeigneter Aminosäuren im hydrophoben Kern bedarf etwas Übung. Durch eine entsprechende Variation von Aminosäuren auf verschiedenen Positionen in der Struktur lassen sich jedoch sehr schnell die geeigneten Randbedingungen finden, die zu Aminosäuresequenzen führen, die mit einer hohen Wahrscheinlichkeit in die Zielstruktur falten. Dies ist durch die große Anzahl an Tetrapeptiden mit einer hohen Wahrscheinlichkeit für den wahrscheinlichsten Konformationszustand gewährleistet.

Die lineare Berechnung der Aminosäuresequenzen versucht, die Wahrscheinlichkeit für den Zielkonformationszustand und die Zielkonformation bei jedem Schritt zu maximieren. Diese Strategie ist jedoch nicht optimal. Sie kann dazu führen, dass eine in der Modellierung befindliche Sequenz nicht verlängert werden kann, obwohl sie eine bis dahin sehr hohe Wahrscheinlichkeit für die Zielstruktur besitzt. Es ist der Fall denkbar, dass eine etwas geringere Wahrscheinlichkeit eines Tetrapeptids für den Zielkonformationszustand bzw. für die Zielkonformation an einer bestimmten Position in der Sequenz im Ergebnis zu einer deutlich größeren Wahrscheinlichkeit der Gesamtsequenz für die Zielstruktur führt und gleichzeitig ein Abbruch der Modellierung vermieden wird. Dieses Problem lässt sich mit einer sequenziellen Berechnung der Sequenz nicht lösen. Die Anwendung eines Monte-Carlo-Suchprotokolls eröffnet jedoch in diesem Fall die Möglichkeit, die beschriebene Problematik zu umgehen. Die Errechnung der Sequenzen würde in der Weise erfolgen, dass an zufällig ausgewählten Positionen Mutationen eingeführt werden. Da jede Aminosäure (bis auf die ersten und letzten drei) einer Sequenz an vier Tetrapeptiden partizipiert, müsste die Bewertung der Mutation durch eine Analyse des entsprechenden Heptapeptids erfolgen (vgl. Abbildung IV-13 auf Seite 54). In Abhängigkeit vom erzielten Ergebnis würde im Folgenden die Mutation erlaubt oder verworfen werden. Die Anwendung eines heuristischen Verfahrens verändert das grundlegende Ziel des vorgestellten Algorithmus nicht. Die Ergebnissequenz sollte in gleicher Weise eine sehr hohe Wahrscheinlichkeit für die Zielstruktur aufweisen, sie würde aber sowohl lokal zu ausgeglicheneren Werten hinsichtlich der Wahrscheinlichkeit für den Zielkonformationszustand als auch der Wahrscheinlichkeit für die Zielkonformation führen. Der Vorteil einer nichtlinearen Errechnung der Zielsequenz wäre weiterhin, dass gleichzeitig eine Modellierung der Aminosäureseitenketten ermöglicht wird und damit noch während der Optimierung der Sequenz sterische Behinderungen oder ungünstige Verteilungen von Aminosäuren erkannt werden können, was mit dem in dieser Arbeit vorgestellten Schema nicht möglich ist. Die theoretische Evaluierung der aus einem solchen Algorithmus erzielten Ergebnissequenzen könnte sehr schnell zeigen, ob ein solches Verfahren in der Lage ist, qualitativ hochwertige Sequenzen zu erzeugen. In diesem Fall könnte man einen solchen Algorithmus dem in dieser Arbeit beschriebenen Schema vorziehen, da sich, zusätzlich zu den beschriebenen Vorteilen, in gleicher Weise die Möglichkeit eines *rationalen* Proteindesigns implementieren ließe.

V.3. Experimentelle Ergebnisse

Die chemische und thermische Denaturierung von M7 zeigte eine kooperative Entfaltung, was als Hinweis auf definierte Tertiärstrukturen zu interpretieren ist. Eine sehr auffällige Eigenschaft dieses Proteins war seine sehr hohe thermodynamische Stabilität. M7 ist mit einer Freien Entfaltungsenthalpie von $\Delta G_{n\to d}^{H_2O} = +69.3$ kJ (20 °C) im Vergleich zu Top7 ($\Delta G_{n\to d}^{H_2O} = +55.2$ kJ, 20 °C) deutlich stabiler. Offensichtlich hat der Suchalgorithmus von *RosettaDesign* das globale Minimum der Aminosäuresequenz für die Top7-Faltungstopologie nicht gefunden, sondern nur eine Sequenz mit einer niedrigen Energie für die Top7-Struktur. Die Temperatur der Maximalstabilität von M7 wurde zu 48 °C bestimmt. Dies ist ein für Proteine ungewöhnliches Charakteristikum, da gezeigt werden konnte, dass natürlich vorkommende Proteine im Allgemeinen bei einer Temperatur von 20 ± 8 °C ihre Maximalstabilität aufweisen und dies unabhängig von der Schmelztemperatur, ihren strukturellen Eigenschaften oder der Lebenstemperatur der jeweiligen Organismen [Kumar *et al.*, 2002]. Die Änderung der Wärmekapazität pro Aminosäure zeigt für M7 Werte, wie sie für ein strukturiertes Protein dieser Größe zu erwarten sind [Myers *et al.*, 1995].

Ein Grund für die hohe thermodynamische Stabilität von M7 und Top7 lässt sich möglicherweise auf die Struktur selbst zurückführen. Top7 besteht praktisch nur aus Sekundärstrukturelementen, die nur über sehr kurze (minimale) *loops* und *turns* verknüpft sind. Verschiedene Autoren haben darauf hingewiesen, dass eine Verkürzung von *loops* zu einer Stabilisierung von Proteinen führen kann [England *et al.*, 2003]. Weiterhin zeigte die Analyse der Originalsequenz von Top7 in vielen Sequenzabschnitten Tetrapeptide mit sehr hohen Wahrscheinlichkeiten für den Zielkonformationsbereich. Von Dantas und Mitarbeiter wurden mit Hilfe von *RosettaDesign* Proteine *redesignt,* welche im Vergleich zu den Wildtypen in fast allen Fällen eine höhere thermodynamische Stabilität aufwiesen [Dantas *et al.*, 2003]. Die Überprüfung dieser Aminosäuresequenzen und der Wildtypsequenzen mit den in der vorliegenden Arbeit errechneten Dichtefunktionen zeigte, dass die *redesignten* Sequenzen in vielen Teilen aus Tetrapeptiden bestehen, die, ebenso wie Top7, jeweils eine sehr hohe Wahrscheinlichkeit für den wahrscheinlichsten Konformationszustand (den Zielkonformationszustand) besitzen. Die Wildtypsequenzen hingegen zeigten diese Eigenschaft nicht. In gleicher Weise führte die Thermostabilisierung eines Enzyms [Korkegian *et al.*, 2005] zu einer Sequenz, deren optimierte Abschnitte ebenfalls sehr hohe Wahrscheinlichkeiten der einzelnen Tetrapeptide für den Zielkonformationsbereich aufwiesen. Wiederum zeigte die Wildtypsequenz diese Eigenschaft nicht. Diese Ergebnisse könnten ein Indiz dafür sein, dass hohe Wahrscheinlichkeiten für bestimmte Konformationszustände bei Tetrapeptiden eine Folge hoher thermodynamischer Stabilität der betreffenden Tetrapeptide für diese Strukturen sind. Inwieweit einer solchen Korrelation auch ein kausaler Zusammenhang zugrunde liegt läßt sich an dieser Stelle nicht beantworten. Nach Analyse der Proteine aus der *FSSP*-Datenbank wurde gefunden (Daten nicht gezeigt), dass ein Designprinzip, welches die tetrapeptidbasierte Wahrscheinlichkeit einer Sequenz für ihre native Struktur maximiert - obgleich dennoch möglich - *in natura* offenbar nicht favorisiert wird. Keine der untersuchten Sequenzen aus dieser Datenbank hatte für ihre native Struktur eine so hohe Wahrscheinlichkeit, wie die in der vorliegenden Arbeit errechneten Sequenzen für die Top7-Faltungstopologie. Eine Erklärung für diesen Sachverhalt könnte die mit einer erhöhten Wahrscheinlichkeit für eine bedingte Strukturbildung denkbare Zunahme der Proteinstabilität

sein, welche sich negativ auf die Funktion von Proteinen auswirken kann [Ventura & Serrano, 2004].

Die Korrelation der Ergebnissequenzen von *RosettaDesign* und dem in der vorliegenden Arbeit vorgestellten Modellierungsprinzip hinsichtlich ihrer tetrapeptidbasierten Wahrscheinlichkeit für den jeweiligen Zielkonformationszustand könnte sich als Hinweis dafür interpretieren lassen, dass Tetrapeptidkonformationen implizit durch die von *RosettaDesign* verwendete Energiefunktion bewertet werden können. Sofern sich die Korrelation zwischen tetrapeptidbasierter Wahrscheinlichkeit für die Zielkonformationszustände und der im Ergebnis ausgebildeten Struktur in weiteren Experimenten bestätigt, könnte ein zusätzlicher Energieterm, der Tetrapeptidkonformationen explizit evaluiert, möglicherweise das Auffinden des globalen Minimums oder eines lokalen Minimums niedriger freier Energie in stochastischen Verfahren beschleunigen.

Trotz intensiver Versuche war es nicht gelungen, M7 in kristalliner Form zu erhalten. Die nur sehr kurzen *loops* und *turns* in der Struktur von Top7 erlauben nur eine geringe Variabilität in der Orientierung der einzelnen Sekundärstrukturelemente. Man könnte daher erwarten, dass das Modell von M7 seine Struktur mit ähnlich guter Genauigkeit beschreibt, wie das Modell von Top7 seine Struktur. Letztlich können jedoch nur experimentell bestimmte Atomkoordinaten die Qualität des Modells verifizieren. Die Strukturaufklärung von M7 erfolgte nach Beendigung dieser Arbeit durch Dr. Christian Lücke mit Hilfe der NMR-Spektroskopie (Forschungsstelle „Enzymologie der Proteinfaltung“ der Max-Planck-Gesellschaft, Halle/Saale). Es zeigte sich eine sehr gute Übereinstimmung zwischen dem Modell von M7 und der experimentell bestimmten Struktur.

Zusätzlich zu M7 wurde im Rahmen dieser Arbeit die Variante M5 teilweise experimentell charakterisiert. M5 zeigte ebenso eine kooperative Faltung, verhielt sich in 0 M GdmCl bis zu einer Temperatur von $T = 383\,K$ (110 °C) thermisch indifferent und wies mit $\Delta G_{n\rightarrow d}^{H_2O} = +45.2 \pm 4.2\,kJ/mol$ in gleicher Weise eine hohe Stabilität auf. Aufgrund der Aggregationsanfälligkeit dieses Proteins wurden keine NMR-spektroskopischen Untersuchungen zur Abschätzung der Sekundärstrukturanteile durchgeführt, so dass nur formuliert werden konnte, dass M5 ein stabiles, kooperativ faltendes Protein ist. Die Analyse der anderen sechs Varianten muß noch zeigen, inwieweit die verbliebenen Sequenzen zu kooperativ faltenden Proteinen führen. Weiterhin wären Strukturdaten aller modellierten Varianten im Hinblick auf ein *Redesign* funktioneller Proteine äußerst nützlich, da biologische Funktionen sehr empfindlich auf Änderungen der Geometrie des Proteinrückgrats reagieren. Ebenso wirken sich hohe Stabilitäten negativ auf die Funktionalität biologisch aktiver Proteine aus [Ventura & Serrano, 2004]. Da der Grund der hohen Stabilität von M7 derzeit nicht verstanden ist, könnte dies ein erfolgreiches *Redesign* solcher Proteine verhindern. Inwieweit die beobachtete Stabilität von M7 tatsächlich mit der jeweiligen Wahrscheinlichkeit für die Zielkonformationszustände korreliert, ließe sich jedoch durch gezielt eingeführte Mutationen bestimmen, welche die Proteine in der Weise sukzessiv destabilisieren, dass die hohe Wahrscheinlichkeit für den jeweiligen Zielkonformationszustand einzelner Tetrapeptide vermindert wird. Die physikalische Charakterisierung dieser Varianten müßte zeigen, ob und inwieweit eine messbare Verminderung der Stabilität dieser Proteine eingetreteten ist. Das Wissen um einen solchen Zusammenhang könnte für ein *Redesign* oder *de novo* Design funktioneller Proteine verwendet werden.

VI. Literaturverzeichnis

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J.Mol.Biol.* **215**, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.

Ambroggio, X. I. and Kuhlman, B. (2006). Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J.Am.Chem.Soc.* **128**, 1154-1161.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* **181**, 223-230.

Anishetty, S., Pennathur, G., and Anishetty, R. (2002). Tripeptide analysis of protein structures. *BMC.Struct.Biol.* **2**, 9-16.

Baldwin, R. L. and Rose, G. D. (1999a). Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem.Sci.* **24**, 26-33.

Baldwin, R. L. and Rose, G. D. (1999b). Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem.Sci.* **24**, 77-83.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallogr.D.Biol.Crystallogr.* **58**, 899-907.

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., and Westbrook, J. (2000a). The Protein Data Bank and the challenge of structural genomics. *Nat.Struct.Biol.* **7 Suppl**, 957-959.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000b). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242.

Betancourt, M. R. and Skolnick, J. (2004). Local propensities and statistical potentials of backbone dihedral angles in proteins. *J.Mol.Biol.* **342**, 635-649.

Blake, J. D. and Cohen, F. E. (2001). Pairwise sequence alignment below the twilight zone. *J.Mol.Biol.* **307**, 721-735.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365-370.

Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E., and Baker, D. (2001). Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins* **Suppl 5**, 119-126.

Bourne, P. E. and Weissig, H (2003). Structural Bioinformatics. *Wiley Liss*

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

Bowman, A. W. and Azzalini, A. (1997). Applied Smoothing Techniques for Data Analysis. *Oxford Statistical Science Series* **18**,

Bradley, P., Chivian, D., Meiler, J., Misura, K. M., Rohl, C. A., Schief, W. R., Wedemeyer, W. J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C. E., and Baker, D. (2003). Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins* **53 Suppl 6**, 457-468.

Brown, C. L., Aksay, I. A., Saville, D. A., and Hecht, M. H. (2002). Template-directed assembly of a de novo designed protein. *J.Am.Chem.Soc.* **124**, 6846-6848.

Butterfoss, G. L. and Kuhlman, B. (2006). Computer-based design of novel protein structures. *Annu.Rev.Biophys.Biomol.Struct.* **35**, 49-65.

Bystroff, C. and Shao, Y. (2002). Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics.* **18 Suppl 1**, S54-S61.

Bystroff, C., Thorsson, V., and Baker, D. (2000). HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J.Mol.Biol.* **301**, 173-190.

Canutescu, A. A., Shelenkov, A. A., and Dunbrack, R. L., Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**, 2001-2014.

Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D., Monnat, R. J., and Stoddard, B. L. (2002). Design, activity, and structure of a highly specific artificial endonuclease. *Mol.Cell* **10**, 895-905.

Chivian, D., Kim, D. E., Malmstrom, L., Bradley, P., Robertson, T., Murphy, P., Strauss, C. E., Bonneau, R., Rohl, C. A., and Baker, D. (2003). Automated prediction of CASP-5 structures using the Robetta server. *Proteins* **53 Suppl 6**, 524-533.

Chivian, D., Kim, D. E., Malmstrom, L., Schonbrun, J., Rohl, C. A., and Baker, D. (2005). Prediction of CASP6 structures using automated Robetta protocols. *Proteins* **61 Suppl 7**, 157-166.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature* **357**, 543-544.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.

Cohen, B. I., Presnell, S. R., and Cohen, F. E. (1993). Origins of Structural Diversity Within Sequentially Identical Hexapeptides. *Protein Science* **2**, 2134-2145.

Dahiyat, B. I. and Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science* **278**, 82-87.

Dahiyat, B. I., Sarisky, C. A., and Mayo, S. L. (1997). De novo protein design: towards fully automated sequence selection. *J.Mol.Biol.* **273**, 789-796.

Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J.Mol.Biol.* **332**, 449-460.

De, Maeyer M., Desmet, J., and Lasters, I. (2000). The dead-end elimination theorem: mathematical aspects, implementation, optimizations, evaluation, and performance. *Methods Mol.Biol.* **143**, 265-304.

Dehouck, Y., Gilis, D., and Rooman, M. (2006). A new generation of statistical potentials for proteins. *Biophys.J.* **90**, 4010-4017.

Delano, W. L. (2002). The PyMol Molecular Graphics System.

Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T. C., McKusick, K. B., Beckmann, J. S., Bentolila, S., Bihoreau, M., Birren, B. B., Browne, J., Butler, A., Castle, A. B., Chiannilkulchai, N., Clee, C., Day, P. J., Dehejia, A., Dibling, T., Drouot, N., Duprat, S., Fizames, C., Fox, S., Gelling, S., Green, L., Harrison, P., Hocking, R., Holloway, E., Hunt, S., Keil, S., Lijnzaad, P., Louis-Dit-Sully, C., Ma, J., Mendis, A., Miller, J., Morissette, J., Muselet, D., Nusbaum, H. C., Peck, A., Rozen, S., Simon, D., Slonim, D. K., Staples, R., Stein, L. D., Stewart, E. A., Suchard, M. A., Thangarajah, T., Vega-Czarny, N., Webber, C., Wu, X., Hudson, J., Auffray, C., Nomura, N., Sikela, J. M., Polymeropoulos, M. H., James, M. R., Lander, E. S., Hudson, T. J., Myers, R. M., Cox, D. R., Weissenbach, J., Boguski, M. S., and Bentley, D. R. (1998). A physical map of 30,000 human genes. *Science* **282**, 744-746.

Desmet, J., De Maeyer, M., Hazes, B., and Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539-542.

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

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

Dill, K. A. and Bromberg, S. (2003). Molecular driving forces: statistical thermodynamics in chemistry and biology. *Garland Science*

Dower, W. J., Miller, J. F., and Ragsdale, C. W. (1988). High efficiency transformation of E. coli by high voltage electroporation. *Nucleic Acids Res.* **16**, 6127-6145.

Dunbrack, R. L., Jr. and Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661-1681.

Dwyer, M. A., Looger, L. L., and Hellinga, H. W. (2004). Computational design of a biologically active enzyme. *Science* **304**, 1967-1971.

Eker, F., Cao, X., Nafie, L., and Schweitzer-Stenner, R. (2002). Tripeptides adopt stable structures in water. A combined polarized visible Raman, FTIR, and VCD spectroscopy study. *J.Am.Chem.Soc.* **124**, 14330-14341.

England, J. L., Shakhnovich, B. E., and Shakhnovich, E. I. (2003). Natural selection of more designable folds: a mechanism for thermophilic adaptation. *Proc.Natl.Acad.Sci.U.S.A* **100**, 8727-8731.

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V. A., Pieper, U., Stuart, A. C., Marti-Renom, M. A., Madhusudhan, M. S., Yerkovich, B., and Sali, A. (2003). Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* **31**, 3375-3380.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186-194.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175-185.

Fetrow, J. S., Giammona, A., Kolinski, A., and Skolnick, J. (2002). The protein folding problem: a biophysical enigma. *Curr.Pharm.Biotechnol.* **3**, 329-347.

Godzik, A. (1995). In search of the ideal protein sequence. *Protein Eng* **8**, 409-416.

Gordon, D. B., Marshall, S. A., and Mayo, S. L. (1999). Energy functions for protein design. *Curr.Opin.Struct.Biol.* **9**, 509-513.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J.Mol.Biol.* **162**, 705-708.

Govindarajan, S., Recabarren, R., and Goldstein, R. A. (1999). Estimating the total number of protein folds. *Proteins* **35**, 408-414.

Guex, N. and Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714-2723.

Haspel, N., Tsai, C. J., Wolfson, H., and Nussinov, R. (2003). Hierarchical protein folding pathways: a computational study of protein fragments. *Proteins* **51**, 203-215.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc.Natl.Acad.Sci.U.S.A* **89**, 10915-10919.

Herges T.A. (2003). Entwicklung eines Kraftfeldes zur Strukturvorhersage von Helixproteinen. *Dissertation im Fachbereich Physik der Universität Dortmund*

Ho, B. K. and Dill, K. A. (2006). Folding very short peptides using molecular dynamics. *PLoS.Comput.Biol.* **2**, 228-237.

Hobohm, U. and Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522-524.

Hobohm, U., Scharf, M., Schneider, R., and Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409-417.

Holm, L. and Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.* **22**, 3600-3609.

Holm, L. and Sander, C. (1996). The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.* **24**, 206-209.

Holm, L. and Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **25**, 231-234.

Holm, L. and Sander, C. (1991). Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J.Mol.Biol.* **218**, 183-194.

Holmes, J. B. and Tsai, J. (2004). Some fundamental aspects of building protein structures from fragment libraries. *Protein Sci.* **13**, 1636-1650.

Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. (1996). Errors in protein structures. *Nature* **381**, 272-

Hu, W. P., Godzik, A., and Skolnick, J. (1997). Sequence-structure specificity--how does an inverse folding approach work? *Protein Eng* **10**, 317-331.

Jaenicke, R. (1987). Folding and association of proteins. *Prog.Biophys.Mol.Biol.* **49**, 117-237.

Jaramillo, A., Wernisch, L., Hery, S., and Wodak, S. J. (2001). Automatic procedures for protein design. *Comb.Chem.High Throughput.Screen.* **4**, 643-659.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J.Mol.Biol.* **292**, 195-202.

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

Kabsch, W. and Sander, C. (1984). On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc.Natl.Acad.Sci.U.S.A* **81**, 1075-1078.

Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M., and Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262**, 1680-1685.

Kirkpatrick S., Gelatt Jr.C.D., and Vecchi M.P. (1983). Optimization by simulated annealing. *Science* **220**, 671-680.

Kleywegt, G. J. (1999). Experimental assessment of differences between related protein crystal structures. *Acta Crystallogr.D.Biol.Crystallogr.* **55**, 1878-1884.

Koehl, P. and Levitt, M. (1999). De novo protein design. I. In search of stability and specificity. *J.Mol.Biol.* **293**, 1161-1181.

Kolodny, R., Koehl, P., Guibas, L., and Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. *J.Mol.Biol.* **323**, 297-307.

Korkegian, A., Black, M. E., Baker, D., and Stoddard, B. L. (2005). Computational thermostabilization of an enzyme. *Science* **308**, 857-860.

Kortemme, T., Morozov, A. V., and Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes
13. *J.Mol.Biol.* **326**, 1239-1259.

Kraemer-Pecore, C. M., Lecomte, J. T., and Desjarlais, J. R. (2003). A de novo redesign of the WW domain. *Protein Sci.* **12**, 2194-2205.

Krieger, E., Nabuurs, S. B., and Vriend, G. (2003). Homology modeling. *Methods Biochem.Anal.* **44**, 509-523.

Kuhlman, B. and Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc.Natl.Acad.Sci.U.S.A* **97**, 10383-10388.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-1368.

Kumar, S., Tsai, C. J., and Nussinov, R. (2002). Maximal stabilities of reversible two-state proteins. *Biochemistry* **41**, 5359-5374.

Kurochkina, N. and Privalov, G. (1998). Heterogeneity of packing: structural approach. *Protein Sci.* **7**, 897-905.

Kuznetsov, I. B. and Rackovsky, S. (2003). On the properties and sequence context of structurally ambivalent fragments in proteins. *Protein Sci.* **12**, 2420-2433.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680-685.

Lasters, I., Desmet, J., and De, Maeyer M. (1997). Dead-end based modeling tools to explore the sequence space that is compatible with a given scaffold. *J.Protein Chem.* **16**, 449-452.

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

Leonov, H., Mitchell, J. S., and Arkin, I. T. (2003). Monte Carlo estimation of the number of possible protein folds: effects of sampling bias and folds distributions. *Proteins* **51**, 352-359.

Lesk, A. M. and Rose, G. D. (1981). Folding units in globular-proteins. *Proc.Natl.Acad.Sci.U.S.A* **78**, 4304-4308.

Levinthal, C. (1968). Are there pathways for protein folding? *J.Chim.Phys.* **65**, 44-45.

Li, H., Helling, R., Tang, C., and Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666-669.

Li, H., Tang, C., and Wingreen, N. S. (1998). Are protein folds atypical? *Proc.Natl.Acad.Sci.U.S.A* **95**, 4987-4990.

Li, H., Tang, C., and Wingreen, N. S. (2002). Designability of protein structures: a lattice-model study using the Miyazawa-Jernigan matrix. *Proteins* **49**, 403-412.

Liang, S. and Grishin, N. V. (2002). Side-chain modeling with an optimized scoring function. *Protein Sci.* **11**, 322-331.

Liang, S. and Grishin, N. V. (2004). Effective scoring function for protein sequence design. *Proteins* **54**, 271-281.

Looger, L. L. and Hellinga, H. W. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J.Mol.Biol.* **307**, 429-445.

Marshall, S. A. and Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J.Mol.Biol.* **305**, 619-631.

Melo, F., Devos, D., Depiereux, E., and Feytmans, E. (1997). ANOLEA: a www server to assess protein structures. *Proc.Int.Conf.Intell.Syst.Mol.Biol.* **5**, 187-190.

Melo, F. and Feytmans, E. (1998). Assessing protein structures with a non-local atomic interaction energy. *J.Mol.Biol.* **277**, 1141-1152.

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

Melo, F., Sanchez, R., and Sali, A. (2002). Statistical potentials for fold assessment. *Protein Sci.* **11**, 430-448.

Mendes, J., Guerois, R., and Serrano, L. (2002). Energy estimation in protein design. *Curr.Opin.Struct.Biol.* **12**, 441-446.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculation by fast computing machines. *J.Chem.Phys.* **21**, 1087-1092.

Mezei, M. (1998). Chameleon sequences in the PDB. *Protein Eng* **11**, 411-414.

Micheletti, C., Seno, F., and Maritan, A. (2000). Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins* **40**, 662-674.

Moffet, D. A., Certain, L. K., Smith, A. J., Kessel, A. J., Beckwith, K. A., and Hecht, M. H. (2000). Peroxidase Activity in Heme Proteins Derived from a Designed Combinatorial Library. *J.Am.Chem.Soc.* **122**, 7612-7613.

Motta, A., Reches, M., Pappalardo, L., Andreotti, G., and Gazit, E. (2005). The preferred conformation of the tripeptide Ala-Phe-Ala in water is an inverse gamma-turn: implications for protein folding and drug design
9. *Biochemistry* **44**, 14170-14178.

Myers, J. K., Pace, C. N., and Scholtz, J. M. (1995). Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* **4**, 2138-2148.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to search for similarities in the amino acid sequence of two proteins. *J.Mol.Biol.* **48**, 443-453.

Pace, C. N. (1986). Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol.* **131**, 266-280.

Park, S., Yang, X., and Saven, J. G. (2004). Advances in computational protein design. *Curr.Opin.Struct.Biol.* **14**, 487-494.

Pearl, F. M., Bennett, C. F., Bray, J. E., Harrison, A. P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., and Orengo, C. A. (2003). The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.* **31**, 452-455.

Pirun, M., Babnigg, G., and Stevens, F. J. (2005). Template-based recognition of protein fold within the midnight and twilight zones of protein sequence similarity. *J.Mol.Recognit.* **18**, 203-212.

Pokala, N. and Handel, T. M. (2005). Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J.Mol.Biol.* **347**, 203-227.

Rackovsky, S. (1995). On the existence and implications of an inverse folding code in proteins. *Proc.Natl.Acad.Sci.U.S.A* **92**, 6861-6863.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Conformation of polypeptides and proteins. **7**, 95-99.

Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66-93.

Rojas, N. R., Kamtekar, S., Simons, C. T., McLean, J. E., Vogel, K. M., Spiro, T. G., Farid, R. S., and Hecht, M. H. (1997). De novo heme proteins from designed combinatorial libraries. *Protein Sci.* **6**, 2512-2524.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85-94.

Sali, A., Potterton, L., Yuan, F., van, Vlijmen H., and Karplus, M. (1995). Evaluation of comparative protein modeling by MODELLER. *Proteins* **23**, 318-326.

Schulze-Kremer, S. (1995). Molecular bioinformatics: algorithms and applications. *de Gruyter*

Shakhnovich, E., Farztdinov, G., Gutin, A. M., and Karplus, M. (1991). Protein folding bottlenecks: A lattice Monte Carlo simulation. *PHYSICAL.REVIEW LETTERS.* **67**, 1665-1668.

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**, 473-501.

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

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**, 859-883.

Slovic, A. M., Kono, H., Lear, J. D., Saven, J. G., and DeGrado, W. F. (2004). Computational design of water-soluble analogues of the potassium channel KcsA. *Proc.Natl.Acad.Sci.U.S.A* **101**, 1828-1833.

Solis, A. D. and Rackovsky, S. (2002). Optimally informative backbone structural propensities in proteins. *Proteins* **48**, 463-486.

Sreerama, N. and Woody, R. W. (2000). Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal.Biochem.* **287**, 252-260.

Sternberg, M. J. and Islam, S. A. (1990). Local protein sequence similarity does not imply a structural relationship. *Protein Eng* **4**, 125-131.

Sternberg, M. J. and Thornton, J. M. (1978). Prediction of protein structure from amino acid sequence. *Nature* **271**, 15-20.

Sudarsanam, S. (1998). Structural diversity of sequentially identical subsequences of proteins: identical octapeptides can have different conformations. *Proteins* **30**, 228-231.

Sudarsanam, S. and Srinivasan, S. (1997). Sequence-dependent conformational sampling using a database of phi(i)+1 and psi(i) angles for predicting polypeptide backbone conformations. *Protein Eng* **10**, 1155-1162.

Thompson, M. J., Sievers, S. A., Karanicolas, J., Ivanova, M. I., Baker, D., and Eisenberg, D. (2006). The 3D profile method for identifying fibril-forming segments of proteins. *Proc.Natl.Acad.Sci.U.S.A* **103**, 4074-4078.

Transue, T. R., Smith, A. K., Mo, H., Goldstein, I. J., and Saper, M. A. (1997). Structure of benzyl T-antigen disaccharide bound to Amaranthus caudatus agglutinin. *Nat.Struct.Biol.* **4**, 779-783.

Tsai, C. J., Maizel, J. V., Jr., and Nussinov, R. (2000). Anatomy of protein structures: visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc.Natl.Acad.Sci.U.S.A* **97**, 12038-12043.

Tsai, C. J. and Nussinov, R. (2001). Transient, highly populated, building blocks folding model. *Cell Biochem.Biophys.* **34**, 209-235.

Tsai, C. J., Polverino de, Laureto P., Fontana, A., and Nussinov, R. (2002). Comparison of protein fragments identified by limited proteolysis and by computational cutting of proteins. *Protein Sci.* **11**, 1753-1770.

Tsai, H. H., Tsai, C. J., Ma, B., and Nussinov, R. (2004). In silico protein design by combinatorial assembly of protein building blocks. *Protein Sci.* **13**, 2753-2765.

van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hunenberger, P. H., Mark, A. E., and Tironi, I. G. (1996). Biomolecular Simulation: The GROMOS96 Manual and User Guide. *vdf Hochschulverlag AG an der ETH Zurich*

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M.,

Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., bu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di, Francesco, V, Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., and Nodell, M. (2001). The sequence of the human genome. *Science* **291**, 1304-1351.

Ventura, S. and Serrano, L. (2004). Designing proteins from the inside out. *Proteins* **56**, 1-10.

Vlasov, P. K., Vlasova, A. V., Tumanyan, V. G., and Esipova, N. G. (2005). A tetrapeptide-based method for polyproline II-type secondary structure prediction. *Proteins* **61**, 763-768.

Voigt, C. A., Gordon, D. B., and Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J.Mol.Biol.* **299**, 789-803.

Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L., and Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nat.Struct.Biol.* **9**, 553-558.

Vriend, G. and Sander, C. (1993). Quality control of protein models: directional atomic contact analysis. *J.Appl.Cryst.* **26**, 47-60.

Wang, W. and Hecht, M. H. (2002). Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins. *Proc.Natl.Acad.Sci.U.S.A* **99**, 2760-2765.

Wang, Z. X. (1998). A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng* **11**, 621-626.

Wei, Y., Kim, S., Fela, D., Baum, J., and Hecht, M. H. (2003). Solution structure of a de novo protein from a designed combinatorial library. *Proc.Natl.Acad.Sci.U.S.A* **100**, 13270-13273.

Weis, J. H. (1990). Usefulness of the Human Genome Project. *Science* **248**, 1595-

Wernisch, L., Hery, S., and Wodak, S. J. (2000). Automatic protein design with all atom force-fields by exact and heuristic optimization. *J.Mol.Biol.* **301**, 713-736.

West, M. W., Wang, W., Patterson, J., Mancias, J. D., Beasley, J. R., and Hecht, M. H. (1999). De novo amyloid proteins from designed combinatorial libraries. *Proc.Natl.Acad.Sci.U.S.A* **96**, 11211-11216.

Wilks, H. M., Cortes, A., Emery, D. C., Halsall, D. J., Clarke, A. R., and Holbrook, J. J. (1992). Opportunities and limits in creating new enzymes. Experiences with the NAD-dependent lactate dehydrogenase frameworks of humans and bacteria. *Ann.N.Y.Acad.Sci.* **672**, 80-93.

Wishart, D. S., Sykes, B. D., and Richards, F. M. (1991). Simple techniques for the quantification of protein secondary structure by 1H NMR spectroscopy. *FEBS Lett.* **293**, 72-80.

Wong, P., Fritz, A., and Frishman, D. (2005). Designability, aggregation propensity and duplication of disease-associated proteins. *Protein Eng Des Sel* **18**, 503-508.

Yona, G. and Levitt, M. (2002). Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J.Mol.Biol.* **315**, 1257-1275.

Yue, K. and Dill, K. A. (1992). Inverse protein folding problem: designing polymer sequences. *Proc.Natl.Acad.Sci.U.S.A* **89**, 4163-4167.

Zhang, L. and Skolnick, J. (1998). How do potentials derived from structural databases relate to "true" potentials? *Protein Sci.* **7**, 112-122.

Zhou, X., Alber, F., Folkers, G., Gonnet, G. H., and Chelvanayagam, G. (2000). An analysis of the helix-to-strand transition between peptides with identical sequence. *Proteins* **41**, 248-256.

Zou, J. and Saven, J. G. (2000). Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. *J.Mol.Biol.* **296**, 281-294.

Zwanzig, R., Szabo, A., and Bagchi, B. (1992). Levinthal's paradox. *Proc.Natl.Acad.Sci.U.S.A* **89**, 20-22.

VII. Anhang

VII.1. Tetrapeptidbasierte Strukturanalyse des Proteins Top7

Tabelle VII-1 Es sind die einzelnen Tetrapeptide der Sequenz des Proteins Top7 (*PDB*-Code 1QYS) [Kuhlman *et al.*, 2003] aufgeführt. Die erste und zweite Spalte geben die Position und die Sequenz des Tetrapeptids an. In der dritten und vierten Spalte sind die Werte des ψ-Winkels der zweiten Aminosäure (ψ_2) bzw. des φ-Winkels der dritten Aminosäure (φ_3) jedes Tetrapeptids aufgeführt, wie sie aus der Struktur von Top7 bestimmt wurden. Die letzte Spalte beschreibt den Konformationszustand des jeweiligen Tetrapeptids, der sich aus dem ψ_2-Winkel und dem φ_3-Winkel ergibt. Die Grenzen der Konformationsbereiche sind in Tabelle II-1, S. 17, definiert. Da die errechneten Dichtefunktionen der ψ_2-φ_3-Verteilungen nur eine Auflösung von 2° besitzen, erfolgte die Zuordnung der Konformationszustände nach Runden der Winkel auf 2° Genauigkeit. Die Diederwinkel und die korrespondierenden Konformationszustände in dieser Tabelle dienten als Vorlage zur Modellierung der alternativen Aminosäuresequenzen zu der Struktur von Top7.

Nummer	Tetrapeptid	ψ_2	φ_3	Konformationszustand
1	**D**IQV	+111.06°	-101.00°	*E*
2	**I**QVQ	+116.17°	-105.91°	*E*
3	**Q**VQV	+115.83°	-112.13°	*E*
4	**V**QVN	+114.87°	-111.95°	*E*
5	**Q**VNI	+117.76°	-122.64°	*E*
6	**V**NID	+108.89°	-106.83°	*E*
7	**N**IDD	+117.89°	-109.79°	*E*
8	**I**DDN	+143.81°	-168.83°	*E*
9	**D**DNG	+ 80.22°	+ 43.86°	*L*
10	**D**NGK	+102.49°	+ 56.52°	*X*
11	**N**GKN	+ 16.87°	-146.94°	*H*
12	**G**KNF	+107.24°	-115.63°	*E*
13	**K**NFD	+125.66°	-112.31°	*E*
14	**N**FDY	+132.56°	-125.89°	*E*
15	**F**DYT	+103.47°	-100.82°	*E*
16	**D**YTY	+126.41°	-124.47°	*E*
17	**Y**TYT	+123.97°	-128.20°	*E*
18	**T**YTV	+162.32°	-139.81°	*E*
19	**Y**TVT	+113.06°	-118.53°	*E*
20	**T**VTT	+169.59°	-158.71°	*E*
21	**V**TTE	+ 81.00°	-176.78°	*E*
22	**T**TES	- 36.94°	+ 68.87°	*L*
23	**T**ESE	- 7.45°	-100.10°	*H*
24	**E**SEL	- 50.62°	- 76.82°	*H*
25	**S**ELQ	+ 1.46°	- 83.94°	*H*
26	**E**LQK	- 40.25°	- 62.67°	*H*
27	**L**QKV	- 25.19°	- 73.61°	*H*
28	**Q**KVL	- 44.51°	- 68.42°	*H*
29	**K**VLN	- 45.40°	- 54.49°	*H*
30	**V**LNE	- 65.42°	- 47.25°	*H*
31	**L**NEL	- 45.86°	- 59.91°	*H*
32	**N**ELM	- 41.32°	- 69.30°	*H*
33	**E**LMD	- 36.35°	- 53.26°	*H*
34	**L**MDY	- 40.58°	- 72.20°	*H*
35	**M**DYI	- 51.47°	- 48.67°	*H*
36	**D**YIK	- 43.07°	- 76.30°	*H*
37	**Y**IKK	- 35.24°	- 56.58°	*H*
38	**I**KKQ	- 53.19°	- 54.40°	*H*

Tabelle VII-1 (Fortsetzung)

Nummer	Tetrapeptid	ψ_2	φ_3	Konformationszustand
39	**K**KQG	- 57.24°	- 65.59°	*H*
40	**K**QGA	- 31.30°	+ 45.66°	*L*
41	**Q**GAK	+ 59.79°	- 86.41°	*E*
42	**G**AKR	+152.40°	- 65.22°	*E*
43	**A**KRV	- 49.76°	-115.58°	*H*
44	**K**RVR	+132.82°	-135.86°	*E*
45	**R**VRI	+129.65°	-121.61°	*E*
46	**V**RIS	+135.82°	-138.03°	*E*
47	**R**ISI	+124.59°	-126.27°	*E*
48	**I**SIT	+136.14°	-125.61°	*E*
49	**S**ITA	+121.43°	- 90.11°	*E*
50	**I**TAR	+109.80°	- 95.20°	*E*
51	**T**ART	+176.01°	- 82.04°	*E*
52	**A**RTK	- 35.34°	-137.38°	*H*
53	**R**TKK	+150.27°	- 55.95°	*E*
54	**T**KKE	- 35.13°	- 66.63°	*H*
55	**K**KEA	- 23.19°	- 78.17°	*H*
56	**K**EAE	- 46.85°	- 51.87°	*H*
57	**E**AEK	- 31.47°	- 70.98°	*H*
58	**A**EKF	- 47.52°	- 57.90°	*H*
59	**E**KFA	- 36.26°	- 72.14°	*H*
60	**K**FAA	- 37.62°	- 64.06°	*H*
61	**F**AAI	- 37.24°	- 64.07°	*H*
62	**A**AIL	- 39.41°	- 65.32°	*H*
63	**A**ILI	- 51.40°	- 62.82°	*H*
64	**I**LIK	- 27.84°	- 67.05°	*H*
65	**L**IKV	- 45.69°	- 64.76°	*H*
66	**I**KVF	- 36.00°	- 60.27°	*H*
67	**K**VFA	- 56.16°	- 60.21°	*H*
68	**V**FAE	- 39.99°	- 63.80°	*H*
69	**F**AEL	- 55.93°	- 51.88°	*H*
70	**A**ELG	- 49.23°	- 71.59°	*H*
71	**E**LGY	+ 16.90°	+ 80.04°	*L*
72	**L**GYN	- 2.95°	- 78.97°	*H*
73	**G**YND	+121.57°	-120.10°	*E*
74	**Y**NDI	- 3.84°	-117.66°	*H*
75	**N**DIN	+ 91.68°	-105.00°	*E*
76	**D**INV	+138.11°	-108.44°	*E*
77	**I**NVT	+127.63°	-116.93°	*E*
78	**N**VTF	+121.82°	-128.48°	*E*
79	**V**TFD	+128.84°	-105.37°	*E*
80	**T**FDG	+136.00°	-133.87°	*E*
81	**F**DGD	+ 90.05°	+ 70.78°	*L*
82	**D**GDT	-113.59°	-106.05°	*E*
83	**G**DTV	+ 6.46°	-101.26°	*H*
84	**D**TVT	+123.35°	-106.54°	*E*
85	**T**VTV	+134.21°	-122.96°	*E*
86	**V**TVE	+105.72°	-100.75°	*E*
87	**T**VEG	+116.03°	-117.98°	*E*
88	**V**EGQ	+127.64°	-123.39°	*E*
89	**EGQL**	+158.51°	-124.10°	*E*

VII.2. DSSP-Output der Kristallstruktur von Top7

```
==== Secondary Structure Definition by the program DSSP, updated CMBI version by ElmK / April 1,2000 ==== DATE=21-OCT-2006     .
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637                                                            .
HEADER    DE NOVO PROTEIN                         11-SEP-03   1QYS                                                        .
COMPND   2 MOLECULE: TOP7;                                                                                                   .
SOURCE   2 ORGANISM_SCIENTIFIC: COMPUTATIONALLY DESIGNED SEQUENCE;                                                           .
AUTHOR    B.KUHLMAN,G.DANTAS,G.C.IRETON,G.VARANI,B.L.STODDARD,D.BAKER                                                        .
   92  1  0  0  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)              .
  5181.0   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)                                                                     .
   83 90.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES                          .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES                          .
   33 35.9   TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES                          .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PER 100 RESIDUES                         .
    2  2.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES                         .
   15 16.3   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES                         .
    5  5.4   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES                         .
   29 31.5   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES                         .
    1  1.1   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES                         .
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30     *** HISTOGRAMS OF ***       .
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0    RESIDUES PER ALPHA HELIX     .
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    PARALLEL BRIDGES PER LADDER  .
 0  0  0  0  0  0  3  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    ANTIPARALLEL BRIDGES PER LADDER  .
 0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    LADDERS PER SHEET            .
 #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
   1    3 A D              0   0  146      0, 0.0     2,-0.7     0, 0.0    21,-0.1   0.000 360.0 360.0 360.0-165.5   -3.1   18.2   17.1
   2    4 A I  E     -A   21   0A  19     19,-1.3    19,-1.9    48,-0.1     2,-0.6  -0.762 360.0-165.7 -91.4 111.1   -0.8   15.6   15.6
   3    5 A Q  E     -AB  20  51A  61     48,-1.4    48,-2.4    -2,-0.7     2,-0.6  -0.860   0.7-165.3-101.0 116.2   -1.6   12.1   16.8
   4    6 A V  E     -AB  19  50A   5     15,-2.1    15,-2.1    -2,-0.6     2,-0.6  -0.892   4.5-174.0-105.9 115.8    0.9    9.3   16.1
   5    7 A Q  E     -AB  18  49A  53     44,-3.1    44,-2.5    -2,-0.6     2,-0.5  -0.937   1.2-175.3-112.1 114.9   -0.2    5.7   16.6
   6    8 A V  E     -AB  17  48A   4     11,-3.1    11,-3.0    -2,-0.6     2,-0.5  -0.948   1.2-176.8-112.0 117.8    2.5    3.1   16.2
   7    9 A N  E     -AB  16  47A  21     40,-3.2    40,-2.2    -2,-0.5     2,-0.6  -0.915   1.2-177.5-122.6 108.9    1.4   -0.6   16.4
   8   10 A I  E     -AB  15  46A   4      7,-2.3     7,-2.9    -2,-0.5     2,-0.4  -0.906   4.4-171.5-106.8 117.9    4.0   -3.3   16.2
   9   11 A D  E     +AB  14  45A  83     36,-1.5    36,-0.8    -2,-0.6    35,-0.3  -0.864  15.7 156.4-109.8 143.8    2.8   -7.0   16.2
  10   12 A D        -     0   0   41      3,-1.4    -2,-0.0    -2,-0.4     4,-0.0  -0.395  63.6 -57.3-168.8  80.2    5.1   -9.9   16.5
  11   13 A N  S    S-     0   0  146      1,-0.2     3,-0.0    -2,-0.0    -2,-0.0   0.901 120.3 -12.1  43.9 102.5    3.9  -13.3   17.8
  12   14 A G  S    S+     0   0   95      1,-0.1     2,-0.3     0, 0.0    -1,-0.2   0.651 122.2  90.8  56.5  16.9    2.4  -13.0   21.3
  13   15 A K        +     0   0   51      2,-0.0    -3,-1.4     0, 0.0     2,-0.5  -0.862  48.9 175.6-146.9 107.2    3.8   -9.5   21.6
  14   16 A N  E     -A    9   0A  97     -2,-0.3     2,-0.5    -5,-0.2    -5,-0.2  -0.955   9.5-165.4-115.6 125.7    1.9   -6.4   20.6
  15   17 A F  E     -A    8   0A  31     -7,-2.9    -7,-2.3    -2,-0.5     2,-0.5  -0.934   2.2-170.0-112.3 132.6    3.3   -2.9   21.2
  16   18 A D  E     -A    7   0A  81     -2,-0.5     2,-0.5    -9,-0.2    -9,-0.2  -0.918   4.3-173.2-125.9 103.5    1.1    0.2   20.9
```

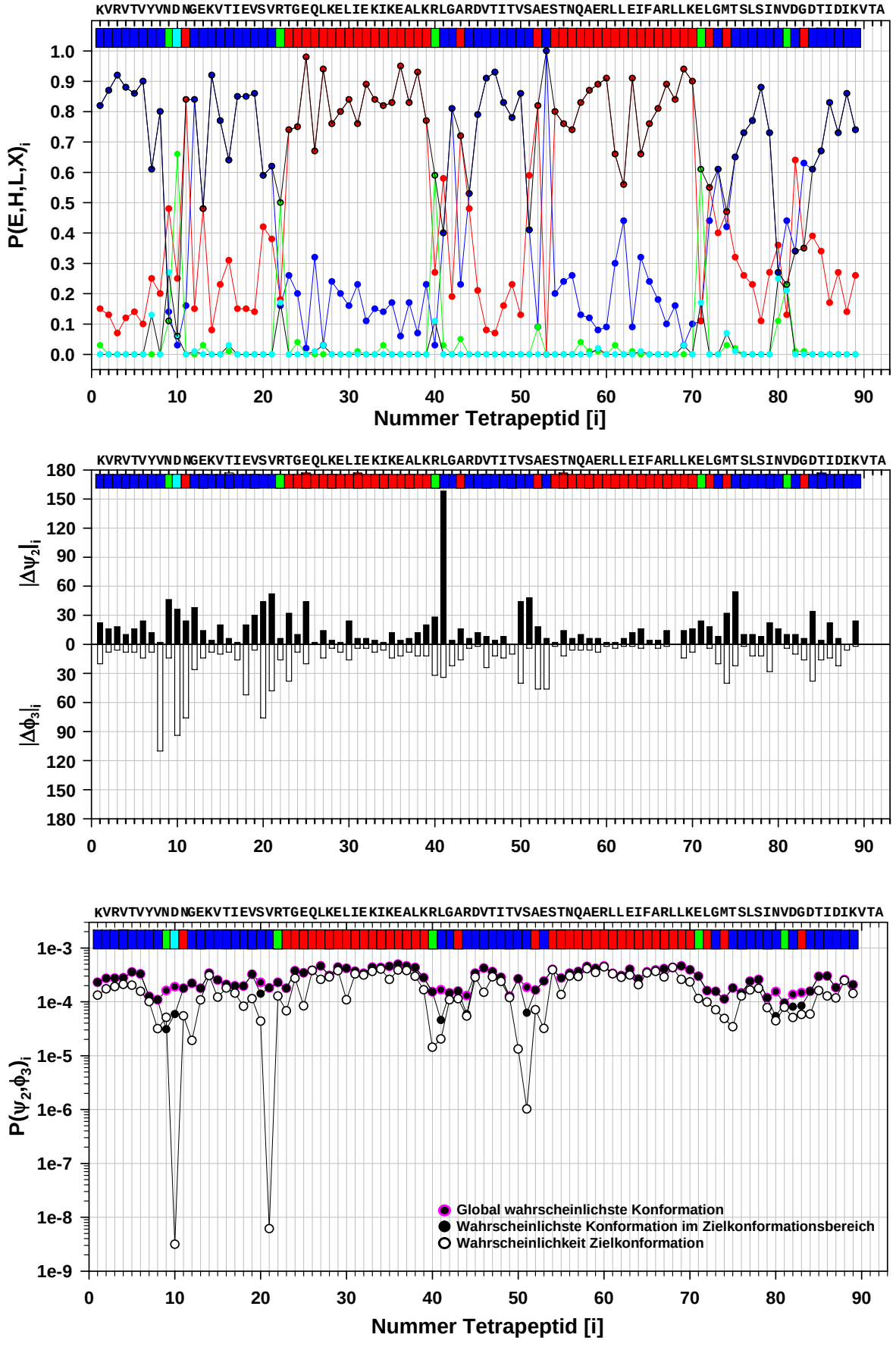
```
17   19 A Y  E     -A    6   0A  73    -11,-3.0   -11,-3.1    -2,-0.5     2,-0.5  -0.850   1.2-174.8-100.8 126.4    2.9    3.6   21.0
18   20 A T  E     +A    5   0A  74     -2,-0.5     2,-0.3   -13,-0.2   -13,-0.2  -0.986   8.6 169.6-124.5 124.0    0.9    6.8   21.1
19   21 A Y  E     -A    4   0A  38    -15,-2.1   -15,-2.1    -2,-0.5     2,-0.4  -0.879  21.8-154.2-128.2 162.3    2.5   10.2   20.9
20   22 A T  E     +A    3   0A  93     -2,-0.3     2,-0.2   -17,-0.2   -17,-0.2  -0.957  24.1 176.0-139.8 113.1    1.1   13.7   20.5
21   23 A V  E     -A    2   0A   4    -19,-1.9   -19,-1.3    -2,-0.4     2,-0.1  -0.736  24.7-172.9-118.5 169.6    3.5   16.2   18.9
22   24 A T  S    S+     0   0   79     -2,-0.2    -2,-0.0   -21,-0.1     0, 0.0  -0.566  72.2  55.7-158.7  81.0    3.4   19.8   17.8
23   25 A T  S    S-     0   0   78     -2,-0.1    -2,-0.0     2,-0.1     0, 0.0   0.212  72.6-138.6-176.8 -36.9    6.6   20.8   16.0
24   26 A E  S    S+     0   0  121      1,-0.1     4,-0.2   -22,-0.1    -3,-0.0   0.381 107.4  70.4  68.9  -7.5    7.3   18.5   13.0
25   27 A S  S >> S+     0   0   62      2,-0.2     4,-0.7     3,-0.1     3,-0.5   0.786  97.4  40.3-100.1 -50.6   10.6   19.1   14.6
26   28 A E  H 3> S+     0   0   83      1,-0.2     4,-1.1     2,-0.2     5,-0.0   0.449 109.3  67.6 -76.8   1.5    9.9   17.1   17.8
27   29 A L  H 3> S+     0   0    3      2,-0.2     4,-3.8     1,-0.2    -1,-0.2   0.825  93.1  54.0 -83.9 -40.2    8.3   14.8   15.1
28   30 A Q  H <> S+     0   0   56     -3,-0.5     4,-2.4    -4,-0.2     5,-0.2   0.788 107.5  54.3 -62.7 -25.2   11.7   14.1   13.7
29   31 A K  H  X S+     0   0   56     -4,-0.7     4,-2.0     2,-0.2    -1,-0.2   0.905 111.7  40.8 -73.6 -44.5   12.5   13.1   17.3
30   32 A V  H  X S+     0   0   12     -4,-1.1     4,-2.8     2,-0.2    -2,-0.2   0.918 116.7  51.6 -68.4 -45.4    9.6   10.7   17.5
31   33 A L  H >X S+     0   0   32     -4,-3.8     4,-2.3     2,-0.2     3,-0.6   0.981 112.2  42.9 -54.5 -65.4   10.3    9.5   14.0
32   34 A N  H 3X S+     0   0   50     -4,-2.4     4,-1.4     1,-0.3    -1,-0.2   0.899 115.0  53.7 -47.3 -45.9   14.0    8.8   14.6
33   35 A E  H 3X S+     0   0   71     -4,-2.0     4,-1.8     1,-0.2    -1,-0.3   0.867 108.1  47.3 -59.9 -41.3   13.0    7.2   17.9
34   36 A L  H <X S+     0   0   10     -4,-2.8     4,-2.4    -3,-0.6     5,-0.3   0.877 103.0  64.1 -69.3 -36.3   10.5    4.8   16.3
35   37 A X  H  X S+     0   0   47     -4,-2.3     4,-1.9    -5,-0.2    -1,-0.2   0.883 106.4  44.6 -53.3 -40.6   13.0    3.8   13.7
36   38 A D  H  X S+     0   0   96     -4,-1.4     4,-2.6     2,-0.2     5,-0.3   0.964 108.2  53.8 -72.2 -51.5   15.2    2.4   16.5
37   39 A Y  H  X S+     0   0   96     -4,-1.8     4,-2.4     1,-0.2    -2,-0.2   0.900 115.8  42.5 -48.7 -43.1   12.4    0.6   18.4
38   40 A I  H  X S+     0   0   11     -4,-2.4     4,-2.8     2,-0.2     5,-0.4   0.848 108.5  55.7 -76.3 -35.2   11.5   -1.2   15.2
39   41 A K  H  < S+     0   0  129     -4,-1.9     -2,-0.2    -5,-0.3    -1,-0.2   0.984 113.8  44.9 -56.6 -53.2   15.0   -1.9   14.0
40   42 A K  H  < S+     0   0  188     -4,-2.6     -2,-0.2     1,-0.2    -1,-0.2   0.937 117.2  41.0 -54.4 -57.2   15.5   -3.7   17.3
41   43 A Q  H  < S-     0   0   90     -4,-2.4     -1,-0.2    -5,-0.3    -2,-0.2   0.834  84.8-162.3 -65.6 -31.3   12.2   -5.6   17.4
42   44 A G     <  -     0   0   37     -4,-2.8      2,-0.2    -5,-0.2    -3,-0.1   0.934  19.5-174.1  45.7  59.8   12.4   -6.5   13.7
43   45 A A        -     0   0    0     -5,-0.4     49,-0.2     1,-0.1   -33,-0.1  -0.518  32.8-131.2 -86.4 152.4    8.7   -7.3   13.4
44   46 A K  S    S+     0   0   61    -35,-0.3     48,-2.3     1,-0.2     2,-0.4   0.955  94.3  35.1 -65.2 -49.8    7.0   -8.8   10.4
45   47 A R  E     +BC   9  91A  34    -36,-0.8   -36,-1.5    46,-0.2     2,-0.4  -0.915  65.0 175.4-115.6 132.8    4.2   -6.3   10.5
46   48 A V  E     -BC   8  90A   4     44,-2.5     44,-2.9     -2,-0.4     2,-0.4  -0.996   7.2-168.3-135.9 129.6    4.3   -2.6   11.5
47   49 A R  E     +BC   7  89A 106    -40,-2.2   -40,-3.2     -2,-0.4     2,-0.4  -0.948   6.0 178.5-121.6 135.8    1.5   -0.0   11.4
48   50 A I  E     -BC   6  88A  10     40,-2.5     40,-2.9     -2,-0.4     2,-0.4  -0.986   2.1-179.1-138.0 124.6    1.8    3.7   11.8
49   51 A S  E     -BC   5  87A  18    -44,-2.5   -44,-3.1     -2,-0.4     2,-0.5  -0.985   6.2-166.2-126.3 136.1   -1.1    6.1   11.5
50   52 A I  E     -BC   4  86A   2     36,-2.3     36,-2.5     -2,-0.4     2,-0.9  -0.984  13.1-147.1-125.6 121.4   -0.9    9.9   11.9
51   53 A T  E     -BC   3  85A  50    -48,-2.4   -48,-1.4     -2,-0.5    34,-0.3  -0.809  27.6-173.2 -90.1 109.8   -3.9   12.1   12.4
52   54 A A        -     0   0    5     32,-3.1     3,-0.1     -2,-0.9    32,-0.1  -0.383  34.6-116.0 -95.2 176.0   -3.1   15.3   10.6
53   55 A R  S    S+     0   0   74      1,-0.2     2,-0.3     -2,-0.1    -1,-0.1   0.828  97.9  18.1 -82.0 -35.3   -5.0   18.6   10.5
54   56 A T  S >> S-     0   0   79     30,-0.1     4,-2.2      1,-0.1     3,-1.5  -0.979  76.0-118.9-137.4 150.3   -5.6   18.5    6.8
55   57 A K  H 3> S+     0   0  109     -2,-0.3     4,-1.4      1,-0.3     5,-0.1   0.845 115.7  58.1 -55.9 -35.1   -5.5   15.8    4.1
56   58 A K  H 3> S+     0   0  155      1,-0.2     4,-1.2      2,-0.2    -1,-0.3   0.772 107.9  48.7 -66.6 -23.2   -2.8   17.6    2.3
57   59 A E  H <> S+     0   0   56     -3,-1.5     4,-2.2      2,-0.2    -2,-0.2   0.916 104.5  56.9 -78.2 -46.9   -0.8   17.4    5.5
58   60 A A  H  X S+     0   0    0     -4,-2.2     4,-1.9      1,-0.2    -2,-0.2   0.795 105.7  54.6 -51.9 -31.5   -1.5   13.7    5.9
59   61 A E  H  X S+     0   0   76     -4,-1.4     4,-2.4      2,-0.2    -1,-0.2   0.928 103.9  50.8 -71.0 -47.5    0.1   13.3    2.4
60   62 A K  H  X S+     0   0   51     -4,-1.2     4,-1.5      1,-0.2    -2,-0.2   0.883 113.9  47.9 -57.9 -36.3    3.3   15.0    3.4
61   63 A F  H  X S+     0   0    2     -4,-2.2     4,-3.1      2,-0.2    -1,-0.2   0.858 106.9  54.3 -72.1 -37.6    3.4   12.6    6.4
62   64 A A  H  X S+     0   0    7     -4,-1.9     4,-1.8      1,-0.2    -2,-0.2   0.885 107.0  52.4 -64.1 -37.2    2.7    9.6    4.2
63   65 A A  H  X S+     0   0   46     -4,-2.4     4,-1.5      2,-0.2    -1,-0.2   0.904 112.0  46.9 -64.1 -39.4    5.6   10.6    2.1
64   66 A I  H  X S+     0   0   53     -4,-1.5     4,-1.9      1,-0.2    -2,-0.2   0.952 112.3  47.9 -65.3 -51.4    7.7   10.7    5.3
65   67 A L  H  X S+     0   0    0     -4,-3.1     4,-2.5      1,-0.2     5,-0.2   0.788 106.0  59.9 -62.8 -27.8    6.5    7.4    6.6
66   68 A I  H  X S+     0   0   70     -4,-1.8     4,-2.5     -5,-0.2    -1,-0.2   0.952 106.4  45.0 -67.1 -45.7    7.1    5.8    3.3
```

```
67   69 A K  H  X S+     0   0  155     -4,-1.5     4,-2.8     2,-0.2     5,-0.3   0.865 112.0  54.8 -64.8 -36.0   10.8    6.6    3.3
68   70 A V  H  X S+     0   0   20     -4,-1.9     4,-2.4     1,-0.2    -2,-0.2   0.970 112.6  40.4 -60.3 -56.2   11.0    5.4    6.9
69   71 A F  H  <>S+     0   0    5     -4,-2.5     5,-2.6     2,-0.2    -2,-0.2   0.852 116.6  51.3 -60.2 -40.0    9.5    2.0    6.2
70   72 A A  H ><5S+     0   0   44     -4,-2.5     3,-2.3    -5,-0.2    -2,-0.2   0.977 109.8  47.3 -63.8 -55.9   11.5    1.7    3.0
71   73 A E  H 3<5S+     0   0   78     -4,-2.8    -2,-0.2     1,-0.3    -1,-0.2   0.942 109.7  55.1 -51.9 -49.2   14.8    2.5    4.6
72   74 A L  T 3<5S-     0   0   27     -4,-2.4    -1,-0.3    -5,-0.3    -2,-0.2   0.263 123.9-104.1 -71.6  16.9   14.0    0.0    7.4
73   75 A G  T < 5S+     0   0   27     -3,-2.3     2,-0.6     1,-0.3    -3,-0.2   0.462  77.1 136.3  80.0  -3.0   13.5   -2.7    4.9
74   76 A Y      < +     0   0    2     -5,-2.6    -1,-0.3    -6,-0.2    18,-0.2  -0.735  23.1 167.3 -79.0 121.6    9.7   -2.9    4.9
75   77 A N        +     0   0  123     16,-1.4     2,-0.6    -2,-0.6    -1,-0.1   0.491  51.7  66.8-120.1  -3.8    8.9   -3.1    1.2
76   78 A D  E     S-D   91   0A 105     15,-1.1    15,-0.6     2,-0.0     2,-0.4  -0.833  72.7-180.0-117.7  91.7    5.2   -4.0    0.7
77   79 A I  E     -D    90   0A  46     -2,-0.6     2,-0.5    13,-0.2    13,-0.2  -0.768  24.1-157.7-105.0 138.1    3.4   -0.9    2.0
78   80 A N  E     -D    89   0A  83     11,-2.9    11,-2.6    -2,-0.4     2,-0.5  -0.927   9.0-167.7-108.4 127.6   -0.3   -0.0    2.3
79   81 A V  E     +D    88   0A  54     -2,-0.5     2,-0.4     9,-0.2     9,-0.2  -0.985  10.7 178.0-116.9 121.8   -1.1    3.7    2.5
80   82 A T  E     -D    87   0A  76      7,-1.8     7,-3.2    -2,-0.5     2,-0.5  -0.990  14.9-157.2-128.5 128.8   -4.7    4.4    3.5
81   83 A F  E     +D    86   0A  72     -2,-0.4     2,-0.4     5,-0.2     5,-0.2  -0.891  10.5 179.3-105.4 136.0   -6.2    7.8    4.0
82   84 A D  E >   -D    85   0A 121      3,-2.9     3,-2.1    -2,-0.5     2,-0.7  -0.826  66.8 -62.2-133.9  90.1   -9.3    8.1    6.2
83   85 A G  T 3  S-     0   0   55     -2,-0.4   -25,-0.1     1,-0.3   -29,-0.0  -0.564 120.6 -15.9  70.8-113.6  -10.2   11.7    6.4
84   86 A D  T 3  S+     0   0   63     -2,-0.7   -32,-3.1    -3,-0.1     2,-0.5   0.307 120.3  95.3-106.1   6.5   -7.3   13.4    8.0
85   87 A T  E <   -CD   51  82A  37     -3,-2.1    -3,-2.9   -34,-0.3     2,-0.5  -0.867  57.2-162.9-101.3 123.4   -5.8   10.2    9.3
86   88 A V  E     -CD   50  81A   0    -36,-2.5   -36,-2.3    -2,-0.5     2,-0.6  -0.902   1.3-164.3-106.5 134.2   -3.1    8.5    7.2
87   89 A T  E     -CD   49  80A  34     -7,-3.2    -7,-1.8    -2,-0.5     2,-0.6  -0.912   6.1-176.6-123.0 105.7   -2.3    4.8    7.9
88   90 A V  E     -CD   48  79A   0    -40,-2.9   -40,-2.5    -2,-0.6     2,-0.5  -0.893   7.6-178.9-100.7 116.0    0.9    3.4    6.5
89   91 A E  E     -CD   47  78A  63    -11,-2.6   -11,-2.9    -2,-0.6     2,-0.3  -0.975   4.6-174.8-118.0 127.6    1.3   -0.3    7.2
90   92 A G  E     -CD   46  77A   0    -44,-2.9   -44,-2.5    -2,-0.5     2,-0.4  -0.858  16.1-149.9-123.4 158.5    4.3   -2.2    6.1
91   93 A Q  E      CD   45  76A  92    -15,-0.6   -16,-1.4    -2,-0.3   -15,-1.1  -0.967 360.0 360.0-124.1 140.8    5.6   -5.8    6.1
92   94 A L              0   0  104    -48,-2.3   -17,-0.2    -2,-0.4    -1,-0.1   0.909 360.0 360.0 -64.6 360.0    9.3   -6.8    6.3
```

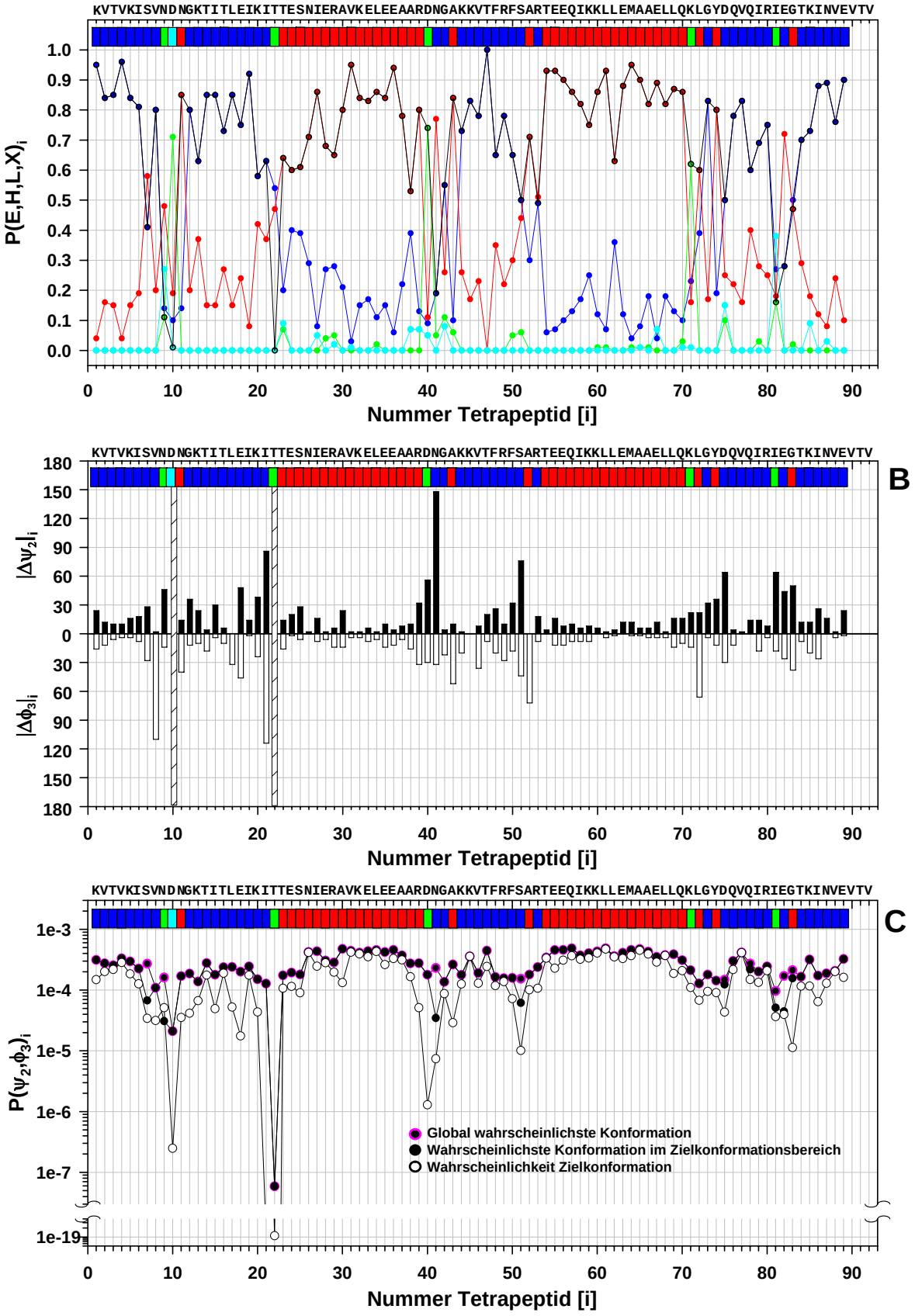
VII.3. Daten der Modelle M1 bis M6 und M8

Die folgenden Diagramme bewerten die einzelnen Tetrapeptide der Aminosäuresequenzen der Modelle M1, M2, M3, M4, M5, M6 und M8 hinsichtlich ihrer Eigenschaft, die Zielstruktur zur beschreiben. Das Diagramm **A** zeigt jeweils für jedes Tetrapeptid *i* der betrachteten Sequenz die Wahrscheinlichkeit $P(E,H,L,X)_i$ zur Ausbildung der Konformationszustände *E* (●), *H* (●), *L* (●) und *X* (●). Die Grenzen dieser Konformationsbereiche sind in Abschnitt II.3.3, S. 16, definiert. Die Wahrscheinlichkeit errechnet sich durch Integration der Dichtefunktionen der ψ_2-φ_3-Verteilungen der entsprechenden Tetrapeptide innerhalb Grenzen dieser Konformationsbereiche. Es wurden diejenigen Punkte schwarz umrandet, die dem Zielkonformationszustand entsprechen. In Diagramm **B** sind jeweils die Winkelabweichungen $|\Delta\psi_2|_i$ oder $|\Delta\varphi_3|_i$ des *i*-ten Tetrapeptids zur wahrscheinlichsten Konformation im Zielkonformationsbereich dargestellt. Das Diagramm **C** zeigt die absoluten Wahrscheinlichkeiten $P(\psi_2,\varphi_3)_i$ der Zielkonformation $(\psi_2,\varphi_3)_i$ des *i*-ten Tetrapeptids im Kontext zur global wahrscheinlichsten Konformation und der wahrscheinlichsten Konformation im Zielkonformationsbereich.

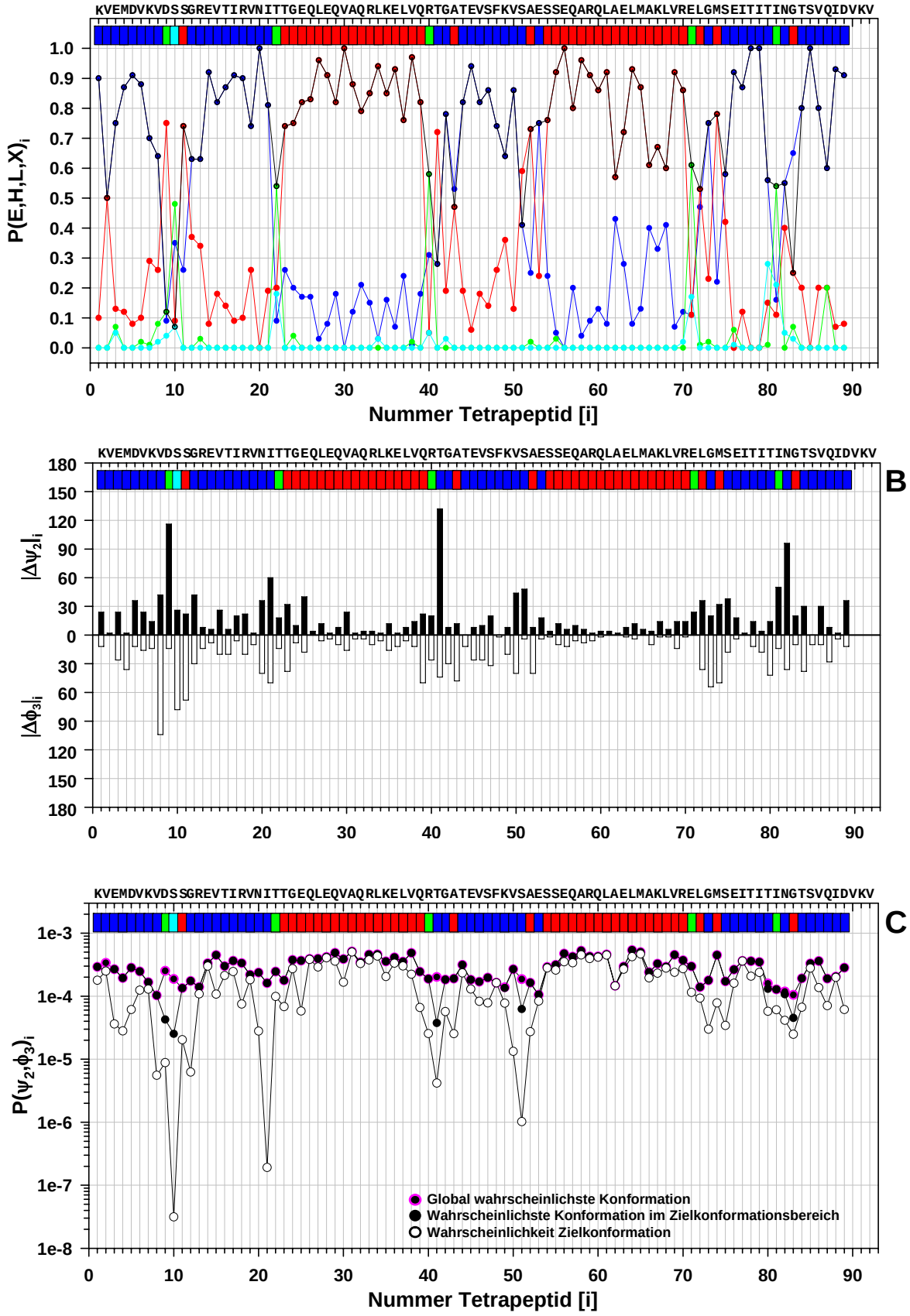
VII.3.1. M1


KVRVTVYVNDNGEKVTIEVSVRTGEQLKELIEKIKEALKRLGARDVTITVSAESTNQAERLLEIFARLLKELGMTSLSINVDGDTIDIKVTA
P(E,H,L,X)i
Nummer Tetrapeptid [i]
B
|Δψ2|i
|Δϕ3|i
C
P(ψ2,ϕ3)i
Global wahrscheinlichste Konformation
Wahrscheinlichste Konformation im Zielkonformationsbereich
Wahrscheinlichkeit Zielkonformation

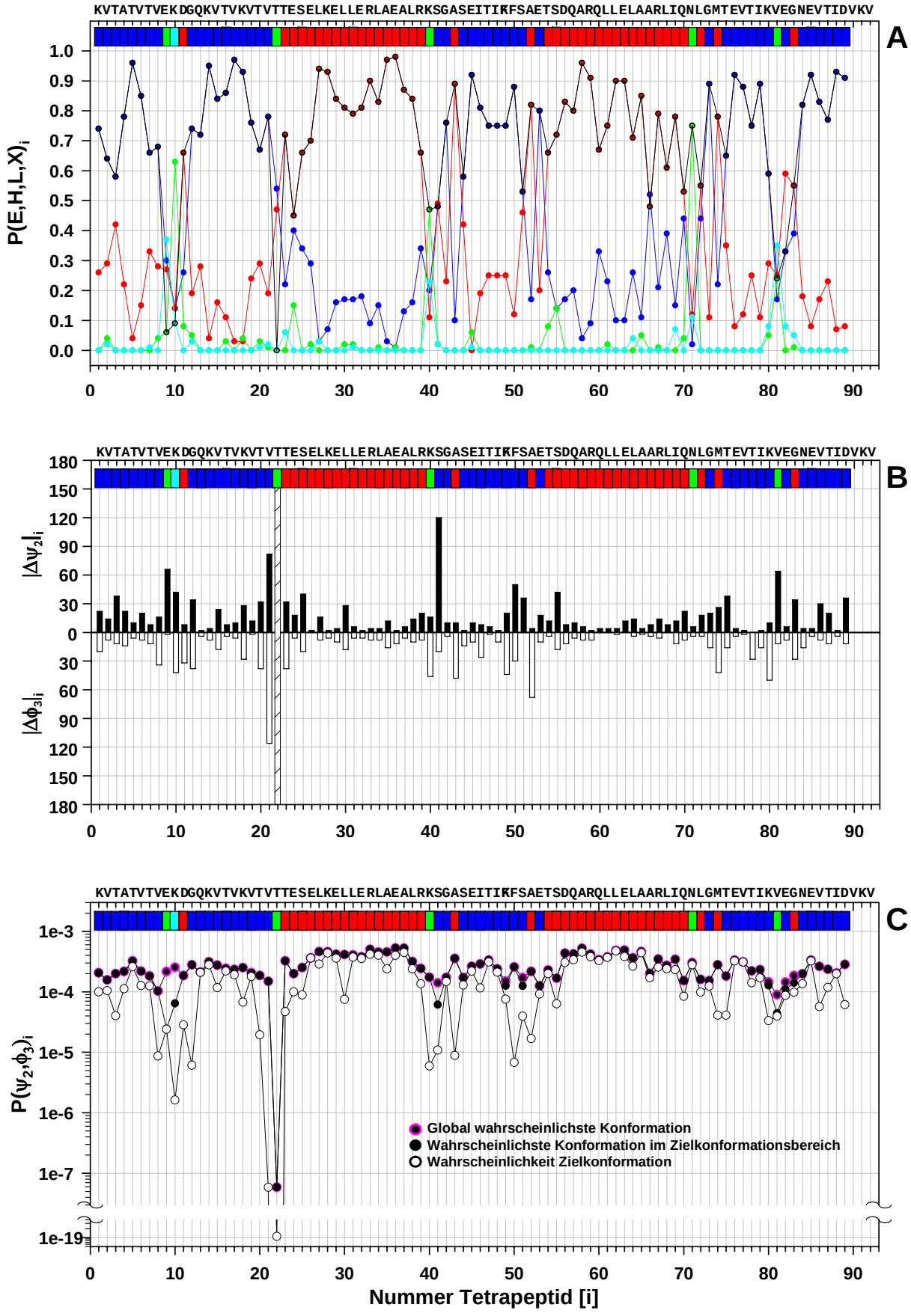
VII.3.2. M2


KVTVKISVNDNGKTITLEIKITTESNIERAVKELEEAARDNGAKKVTFRFSARTEEQIKKLLEMAAELLQKLGYDQVQIRIEGTKINVEVTV
P(E,H,L,X)i
Nummer Tetrapeptid [i]
B
|Δψ2|i
|Δϕ3|i
C
P(ψ2,ϕ3)i
Global wahrscheinlichste Konformation
Wahrscheinlichste Konformation im Zielkonformationsbereich
Wahrscheinlichkeit Zielkonformation

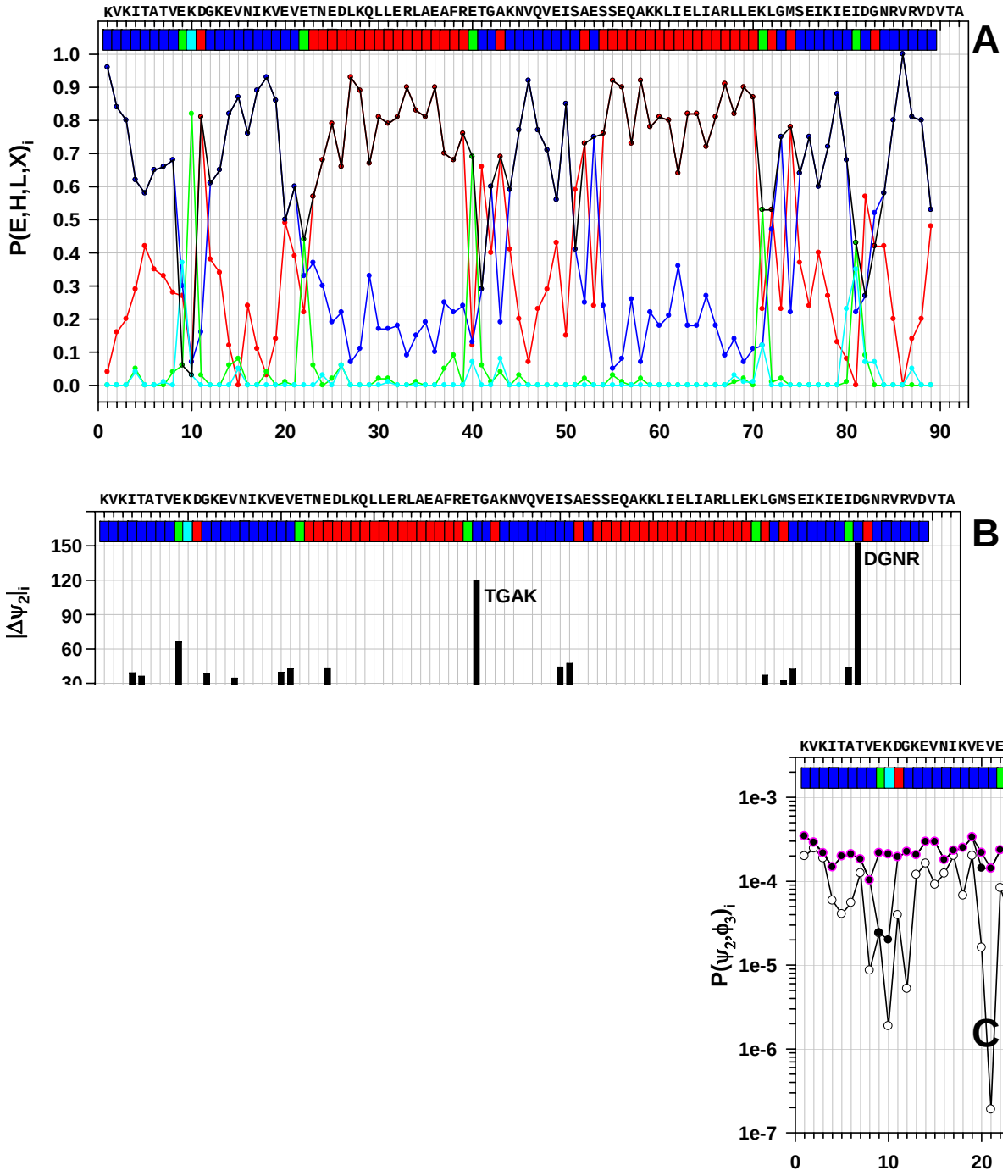
VII.3.3. M3


KVEMDVKVDSSGREVTIRVNITTGEQLEQVAQRLKELVQRTGATEVSFKVSAESSEQARQLAELMAKLVRELGMSEITITINGTSVQIDVKV
P(E,H,L,X)i
Nummer Tetrapeptid [i]
B
|Δψ2|i
|Δϕ3|i
C
P(ψ2,ϕ3)i
Global wahrscheinlichste Konformation
Wahrscheinlichste Konformation im Zielkonformationsbereich
Wahrscheinlichkeit Zielkonformation

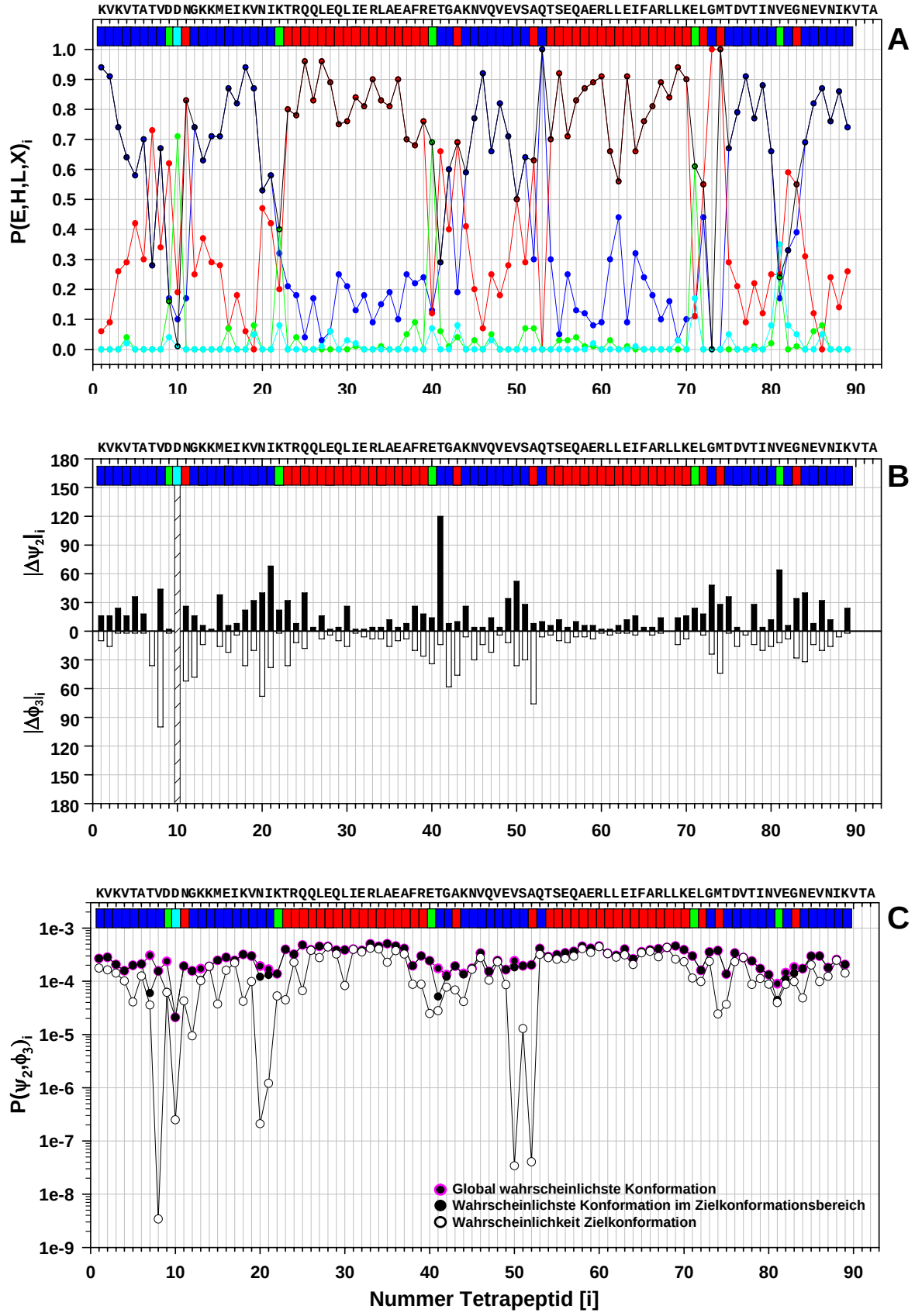
VII.3.4. M4


KVTATVTVEKDGQKVTVKVTVTTESELKELLERLAEALRKSGASEITIKFSAETSDQARQLLELAARLIQNLGMTEVTIKVEGNEVTIDVKV
A
P(E,H,L,X)i
B
|Δψ2|i
|Δϕ3|i
C
P(ψ2,ϕ3)i
Global wahrscheinlichste Konformation
Wahrscheinlichste Konformation im Zielkonformationsbereich
Wahrscheinlichkeit Zielkonformation
Nummer Tetrapeptid [i]

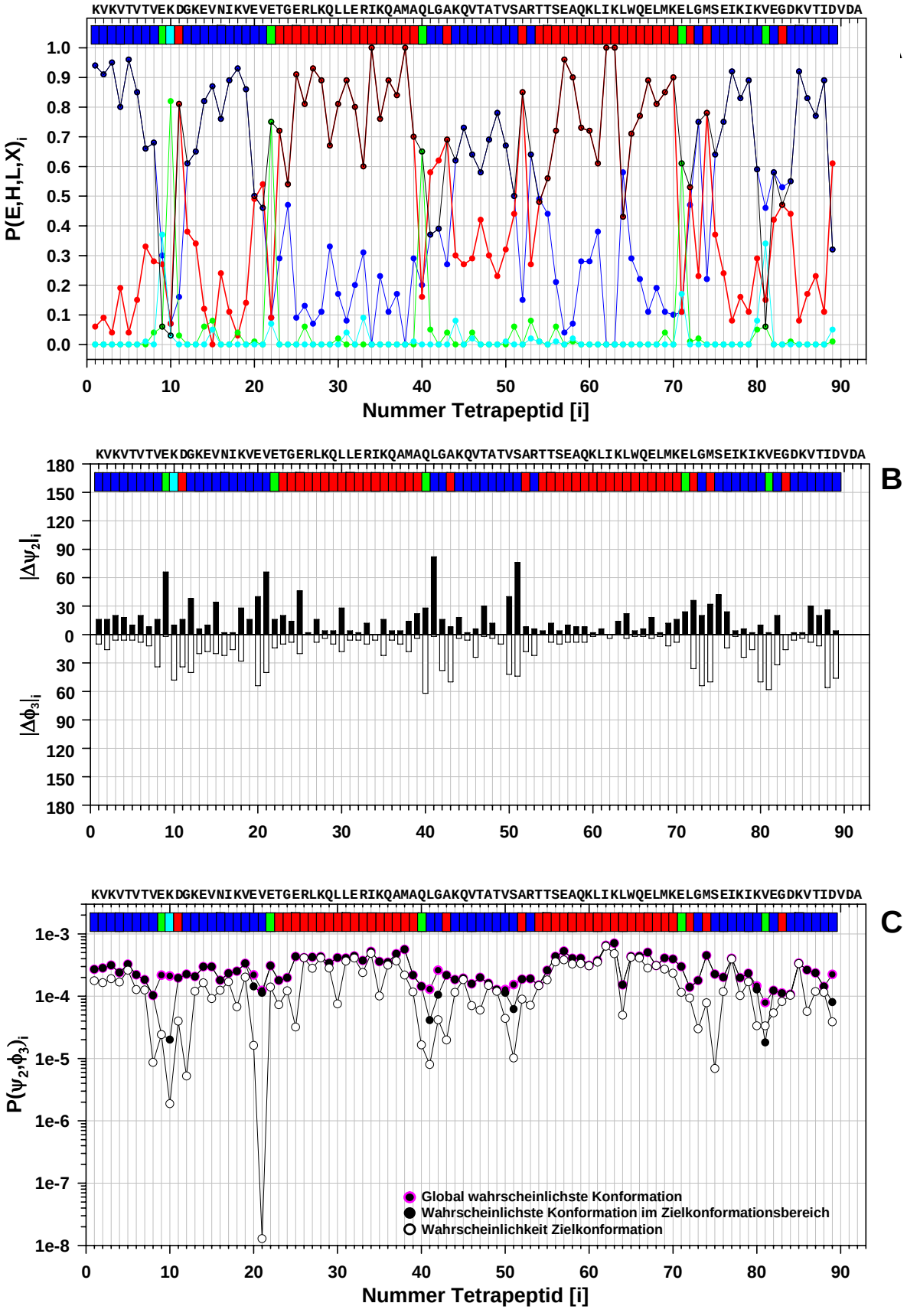
VII.3.5. M5


KVKITATVEKDGKEVNIKVEVETNEDLKQLLERLAEAFRETGAKNVQVEISAESSEQAKKLIELIARLLEKLGMSEIKIEIDGNRVRVDVTA
A
P(E,H,L,X)i
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
KVKITATVEKDGKEVNIKVEVETNEDLKQLLERLAEAFRETGAKNVQVEISAESSEQAKKLIELIARLLEKLGMSEIKIEIDGNRVRVDVTA
B
|Δψ2|i
150
120
90
60
30
TGAK
DGNR
KVKITATVEKDGKEVNIKVEVE
C
P(ψ2,ϕ3)i
1e-3
1e-4
1e-5
1e-6
1e-7
0
10
20
Global wahrscheinlichste Konformation
Wahrscheinlichste Konformation im Zielkonformationsbereich
Wahrscheinlichkeit Zielkonformation

VII.3.6. M6


KVKVTATVDDNGKKMEIKVNIKTRQQLEQLIERLAEAFRETGAKNVQVEVSAQTSEQAERLLEIFARLLKELGMTDVTINVEGNEVNIKVTA
A
P(E,H,L,X)i
B
|Δψ2|i
|Δφ3|i
C
P(ψ2,φ3)i
Global wahrscheinlichste Konformation
Wahrscheinlichste Konformation im Zielkonformationsbereich
Wahrscheinlichkeit Zielkonformation
Nummer Tetrapeptid [i]

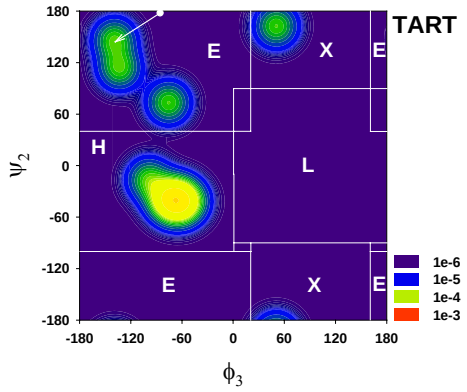
VII.3.7. M8


KVKVTVTVEKDGKEVNIKVEVETGERLKQLLERIKQAMAQLGAKQVTATVSARTTSEAQKLIKLWQELMKELGMSEIKIKVEGDKVTIDVDA
P(E,H,L,X)i
Nummer Tetrapeptid [i]
B
|Δψ2|i
|Δϕ3|i
C
P(ψ2,ϕ3)i
Global wahrscheinlichste Konformation
Wahrscheinlichste Konformation im Zielkonformationsbereich
Wahrscheinlichkeit Zielkonformation

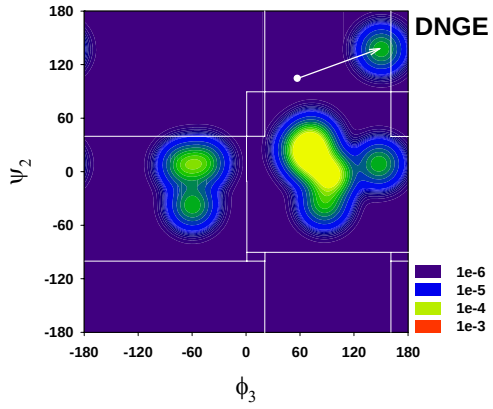
VII.3.8. Abweichungen von den Zielkonformationen

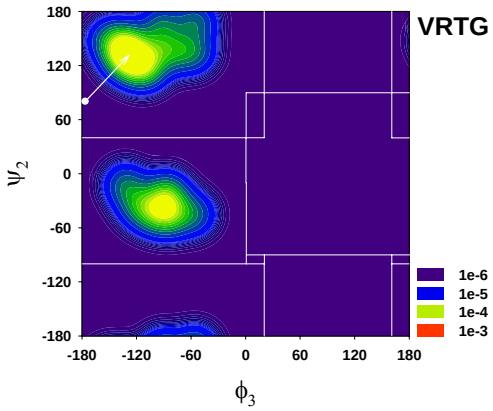
Die folgenden Abbildungen zeigen die Dichtefunktionen der ψ_2-φ_3-Verteilungen von Tetrapeptiden der modellierten Sequenzen, bei denen die relative Wahrscheinlichkeit für die Zielkonformation bezüglich der wahrscheinlichsten Konformation im Zielkonformationsbereich kleiner als 0.01 ist. Die Darstellungen beziehen sich auf die Tabelle IV-9, S. 75. Die Pfeilspitze zeigt jeweils, ausgehend von der Zielkonformation, auf den nächstliegenden Peak im Zielkonformationsbereich.

VII.3.8.1. Top7

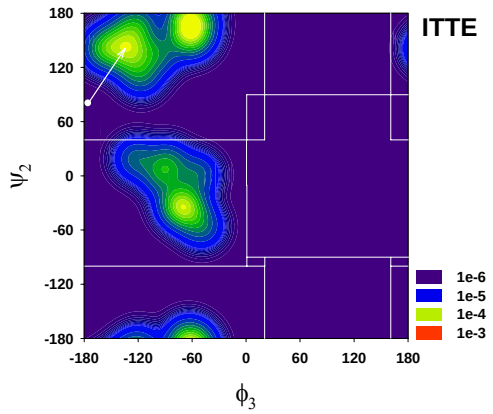


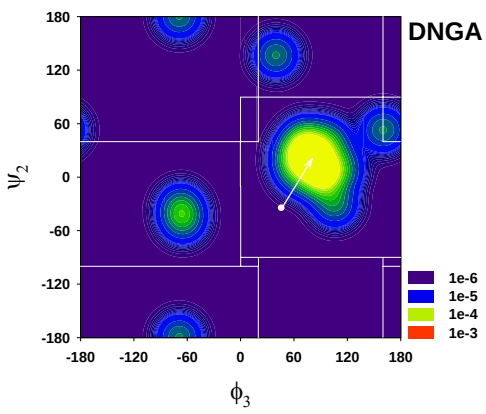
VII.3.8.2. M1



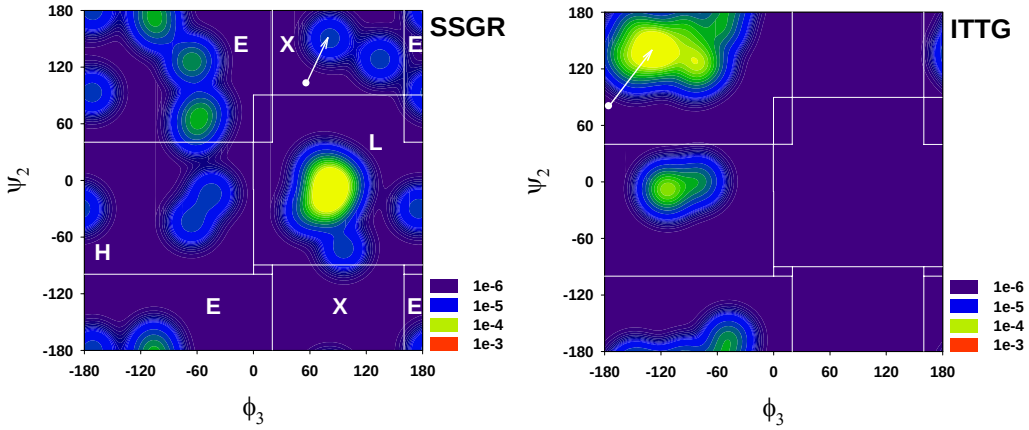


VII.3.8.3. M2

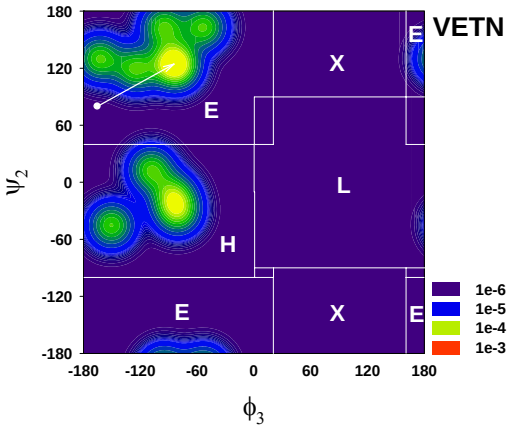




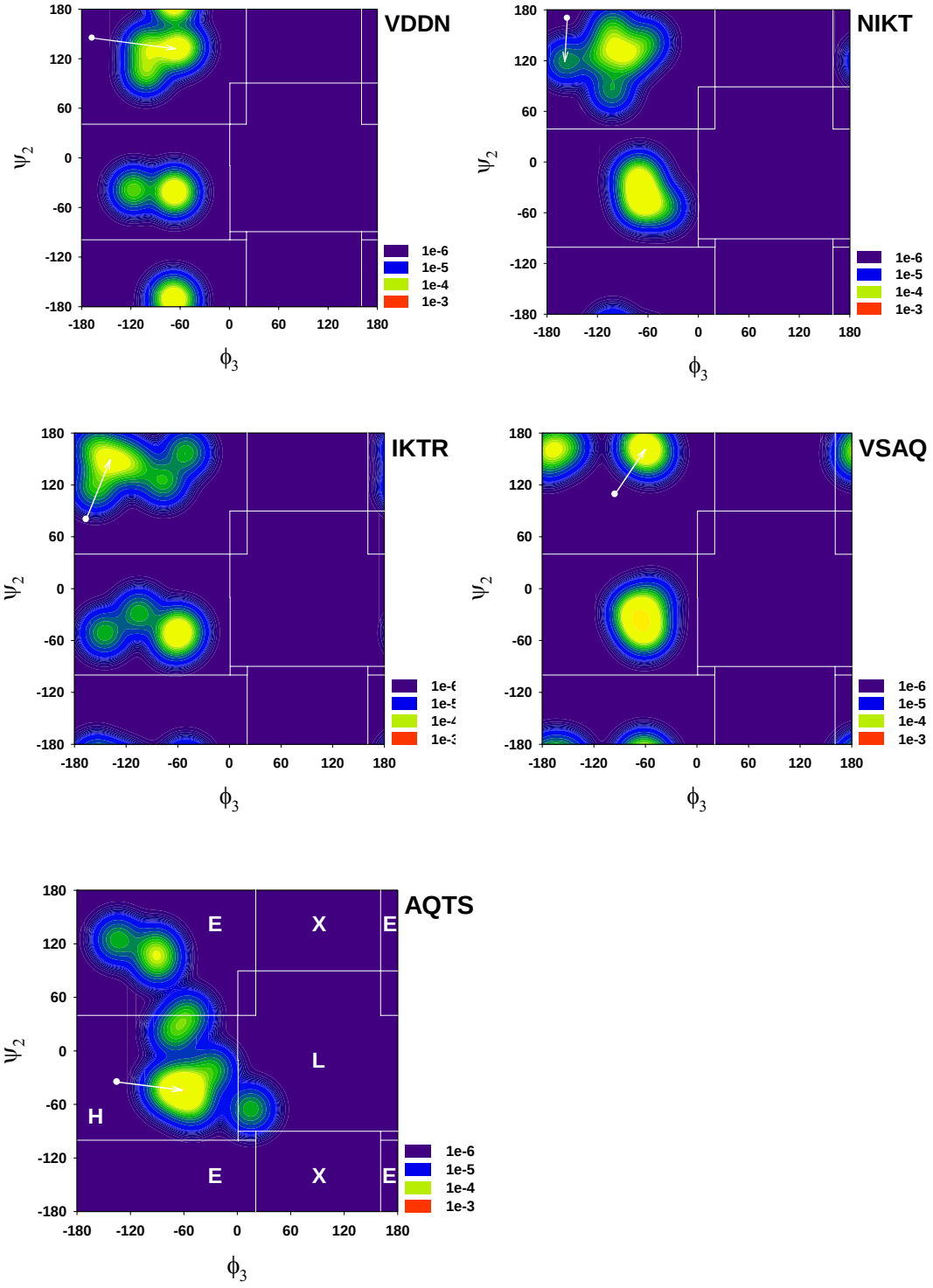
VII.3.8.4. M3



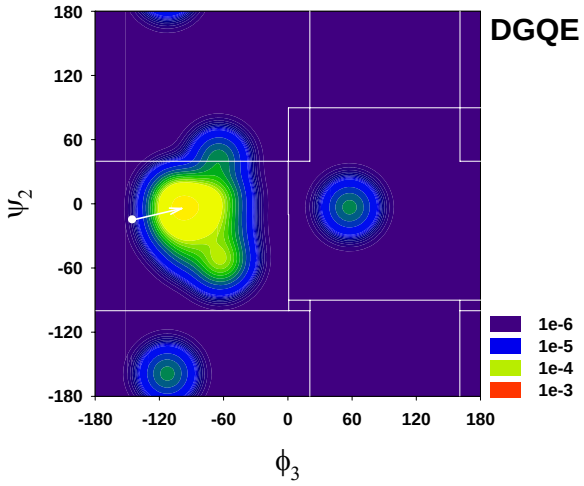
VII.3.8.5. M5



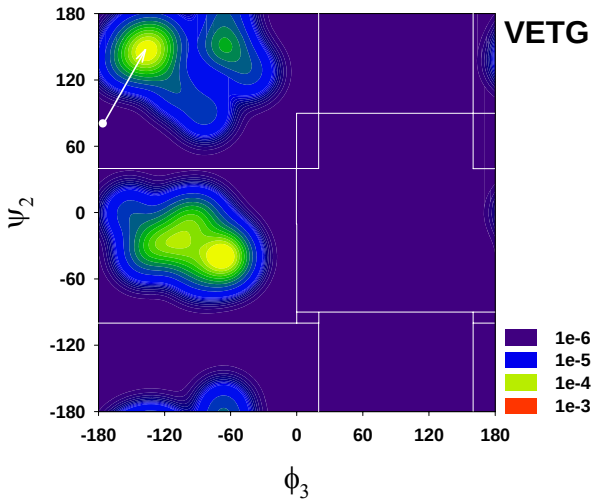
VII.3.8.6. M6


VDDN
NIKT
IKTR
VSAQ
AQTS
E
X
L
H
1e-6
1e-5
1e-4
1e-3
ψ2
ϕ3
180
120
60
0
-60
-120
-180

VII.3.8.7. M7



VII.3.8.8. M8



VII.4. fit-Parameter aus der Anpassung der Datenpunkte bei chemischer und thermischer Denaturierung

Tabelle VII-2 *fit-Parameter* aus der Anpassung Datenpunkte bei der thermischen Entfaltung des Proteins M7. Die Parameter sind in Abschnitt III.4.4, S. 32, definiert. Die Referenztemperatur beträgt $T_0 = 293$ K (20 °C)

Parameter		M7
$\Delta H_{n\to d}^{H_2O}(J\cdot mol^{-1})$	-	115 167.9 ± 2 042.9
$m_H(J\cdot mol^{-1}\cdot M^{-1})$	+	2 860.6 ± 363.1
$\Delta S_{n\to d}^{H_2O}(J\cdot mol^{-1}\cdot K^{-1})$	-	652.4 ± 6.8
$m_S(J\cdot mol^{-1}\cdot K^{-1}\cdot M^{-1})$	+	49.8 ± 1.2
$\Delta Cp_{n\to d}^{H_2O}(J\cdot mol^{-1}\cdot K^{-1})$	+	7 539.0 ± 73.3
$m_C(J\cdot mol^{-1}\cdot K^{-1}\cdot M^{-1})$	-	642.8 ± 12.5
$b_n(\mathrm{deg})$	-	27.7 ± 0.1
$m_n(\mathrm{deg}\cdot K^{-1})$	+	0.03
$b_d(\mathrm{deg})$	+	11.7
$m_d(\mathrm{deg}\cdot K^{-1})$	-	0.04
Regressionskoeffizient R^2		0.996

Tabelle VII-3 *fit-Parameter* aus der Anpassung der Datenpunkte bei der chemischen Enfaltung des Proteins M7. Die Parameter sind in Abschnitt III.4.3, S. 30, definiert. Die Messungen erfolgten bei einer Temperatur von $T = 293$ K (20 °C)

Parameter		M7
$\Delta G_{n\to d}^{H_2O}(J\cdot mol^{-1})$	+	69 266.1 ± 2 974.8
$m_{n\to d}(J\cdot mol^{-1}\cdot M^{-1})$	-	10 496.0 ± 479.8
$b_n(\mathrm{deg})$	-	19.5 ± 0.2
$m_n(\mathrm{deg}\cdot M^{-1})$	-	0.24 ± 0.06
$b_d(\mathrm{deg})$	+	17.0 ± 9.1
$m_d(\mathrm{deg}\cdot M^{-1})$	-	2.5 ± 1.2
Regressionskoeffizient R^2		0.999

VII.5. Fingerprintbereich des 2D-COSY Spektrums von M7

Die folgende Grafik zeigt den Fingerprint Bereich des 2D-COSY Spektrums von M7. Zu sehen sind die H^N-H^α Konnektivitäten aller Aminosäurereste. Die Signale unterhalb der roten Linie (> 4.85 ppm) werden β-Faltblattsträngen zugeordnet. Signale, die zwischen der blauen und der violetten Linie liegen (4.1-3.4 ppm) werden helikalen Strukturen angerechnet. Oberhalb der violetten Linie (< 3.4 ppm) sind Signale von H^ε Amidresonanzen der Argininseitenketten zu erkennen. Der Bereich zwischen den beiden grünen Linien (9.0-8.2 ppm) wird den ungeordneten Strukturelementen (*coil*) zugeschrieben. Hieraus ergeben sich, durch Vergleich des Fingerprint Bereichs in den COSY und TOCSY Spektren, die folgenden Zahlen:

Signale in den α-Helix Sektoren: $A = 30$
Signale in den β-Strand Sektoren: $B = 41$
Signale in den Coil Sektoren: $C = 42$

Man berechnet nun, ausgehend von der Signalverteilung in den verschiedenen Sektoren des Fingerprint Bereichs, nach Wishart *et al.* die Anteile der Sekundärstrukturelemente wie folgt [Wishart *et al.*,1991]:

α-Helix Anteil = $2 \times (A - 2 \times \langle Gly \rangle) = 36$ [bei 6 Glycinresten in M7]
β-Faltblatt Anteil = $2 \times B = 82$
coil Anteil = $0.9 \times C = 38$
Gesamtsumme = 156

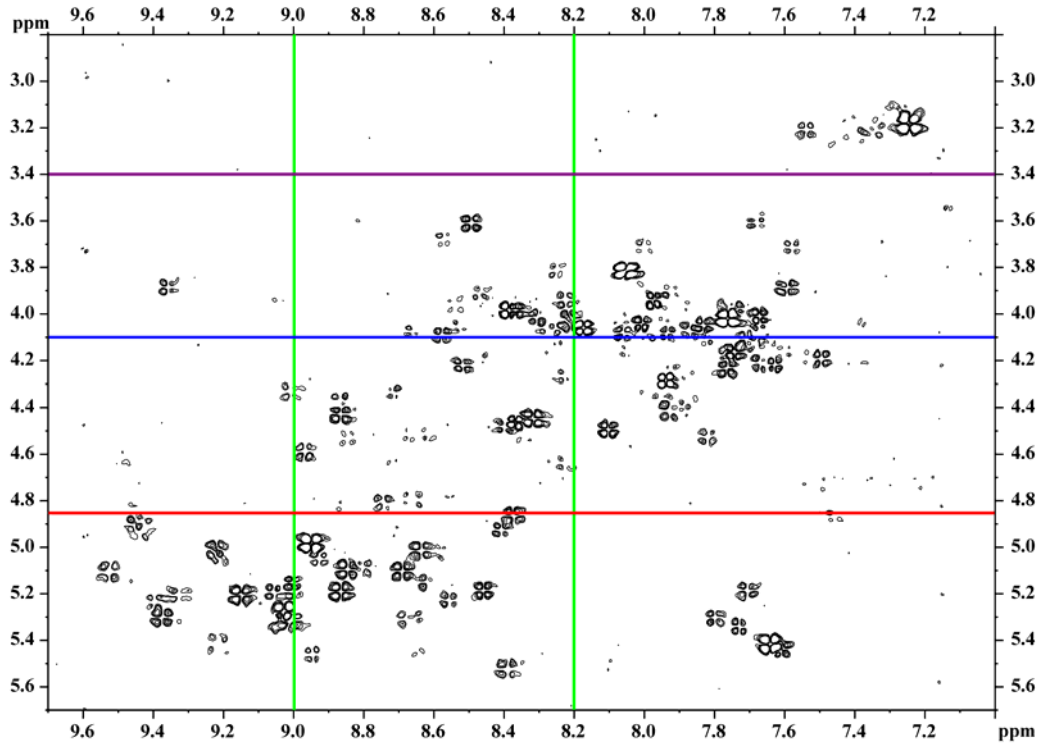


Abbildung VII-1 Fingerprintbereich des 2D-COSY Spektrums von M7

VII.6. Gensequenzen der Proteine M1 bis M8

VII.6.1. Gensequenz M1

```
            KpnI  NdeI                          HincII
      GGGCGAATTGGGTACCCATATGAAAGTGCGCGTGACCGTTTACGTTAACGATAACGGCGA
1     ---------+---------+---------+---------+---------+---------+
      CCCGCTTAACCCATGGGTATACTTTCACGCGCACTGGCAAATGCAATTGCTATTGCCGCT
                        M__K__V__R__V__T__V__Y__V__N__D__N__G__E_

                                              PvuII
      AAAAGTGACCATCGAAGTTAGCGTTCGTACCGGCGAACAGCTGAAAGAACTGATCGAAAA
61    ---------+---------+---------+---------+---------+---------+
      TTTTCACTGGTAGCTTCAATCGCAAGCATGGCCGCTTGTCGACTTTCTTGACTAGCTTTT
      _K__V__T__I__E__V__S__V__R__T__G__E__Q__L__K__E__L__I__E__K_


      AATCAAAGAAGCGCTGAAACGTCTGGGTGCGCGTGATGTTACCATTACCGTGAGCGCGGA
121   ---------+---------+---------+---------+---------+---------+
      TTAGTTTCTTCGCGACTTTGCAGACCCACGCGCACTACAATGGTAATGGCACTCGCGCCT
      _I__K__E__A__L__K__R__L__G__A__R__D__V__T__I__T__V__S__A__E_

                                            BssHII
      AAGCACCAATCAGGCGGAACGTCTGCTGGAAATTTTTGCGCGCCTGCTGAAAGAACTGGG
181   ---------+---------+---------+---------+---------+---------+
      TTCGTGGTTAGTCCGCCTTGCAGACGACCTTTAAAAACGCGCGGACGACTTTCTTGACCC
      _S__T__N__Q__A__E__R__L__L__E__I__F__A__R__L__L__K__E__L__G_

                                              ClaI                HindIII
      TATGACCAGCCTGAGCATTAACGTGGATGGCGATACCATCGATATTAAAGTGACCGCGTA
241   ---------+---------+---------+---------+---------+---------+
      ATACTGGTCGGACTCGTAATTGCACCTACCGCTATGGTAGCTATAATTTCACTGGCGCAT
      _M__T__S__L__S__I__N__V__D__G__D__T__I__D__I__K__V__T__A__*_

            SacI
      AGCTTGGAGCTCCAGCTTTTGTTCCC
301   ---------+---------+------
      TCGAACCTCGAGGTCGAAAACAAGGG
```

VII.6.2. Gensequenz M2

```
          KpnI  NdeI                          HincII
      GGCGAATTGGGTACCCATATGAAAGTGACCGTGAAAATCAGCGTTAACGATAACGGCAAA
1     ---------+---------+---------+---------+---------+---------+
      CCGCTTAACCCATGGGTATACTTTCACTGGCACTTTTAGTCGCAATTGCTATTGCCGTTT
                       M__K__V__T__V__K__I__S__V__N__D__N__G__K__


      ACCATCACCCTGGAAATCAAAATCACCACCGAAAGCAACATTGAACGCGCGGTTAAAGAA
61    ---------+---------+---------+---------+---------+---------+
      TGGTAGTGGGACCTTTAGTTTTAGTGGTGGCTTTCGTTGTAACTTGCGCGCCAATTTCTT
      T__I__T__L__E__I__K__I__T__T__E__S__N__I__E__R__A__V__K__E__

               BssHII                                        BssHII
      CTGGAAGAAGCGGCGCGCGATAACGGCGCGAAAAAAGTGACCTTTCGTTTTAGCGCGCGT
121   ---------+---------+---------+---------+---------+---------+
      GACCTTCTTCGCCGCGCGCTATTGCCGCGCTTTTTTCACTGGAAAGCAAAATCGCGCGCA
      L__E__E__A__A__R__D__N__G__A__K__K__V__T__F__R__F__S__A__R__

                                                      PstI
      ACCGAAGAACAAATCAAAAAACTGCTGGAAATGGCGGCGGAACTGCTGCAGAAACTGGGC
181   ---------+---------+---------+---------+---------+---------+
      TGGCTTCTTGTTTAGTTTTTTGACGACCTTTACCGCCGCCTTGACGACGTCTTTGACCCG
      T__E__E__Q__I__K__K__L__L__E__M__A__A__E__L__L__Q__K__L__G__

        BclI                                                  HindIII
      TATGATCAGGTGCAGATTCGTATCGAAGGCACCAAAATCAACGTTGAAGTGACCGTGTAA
241   ---------+---------+---------+---------+---------+---------+
      ATACTAGTCCACGTCTAAGCATAGCTTCCGTGGTTTTAGTTGCAACTTCACTGGCACATT
      Y__D__Q__V__Q__I__R__I__E__G__T__K__I__N__V__E__V__T__V__*__

           SacI
      GCTTGGAGCTCCAGCTTTTGTTCCC
301   ---------+---------+-----
      CGAACCTCGAGGTCGAAAACAAGGG
```

VII.6.3. Gensequenz M3

```
           KpnI  NdeI
      GGGCGAATTGGGTACCCATATGAAAGTGGAAATGGATGTTAAAGTTGATAGCAGCGGTCG
1     ---------+---------+---------+---------+---------+---------+
      CCCGCTTAACCCATGGGTATACTTTCACCTTTACCTACAATTTCAACTATCGTCGCCAGC
                         M__K__V__E__M__D__V__K__V__D__S__S__G__R_

                    HincII               PvuII
      CGAAGTTACCATTCGCGTTAACATTACCACCGGCGAACAGCTGGAACAGGTTGCGCAGCG
61    ---------+---------+---------+---------+---------+---------+
      GCTTCAATGGTAAGCGCAATTGTAATGGTGGCCGCTTGTCGACCTTGTCCAACGCGTCGC
      _E__V__T__I__R__V__N__I__T__T__G__E__Q__L__E__Q__V__A__Q__R_


      TCTGAAAGAACTGGTTCAGCGTACCGGCGCGACCGAAGTGAGCTTTAAAGTTAGCGCGGA
121   ---------+---------+---------+---------+---------+---------+
      AGACTTTCTTGACCAAGTCGCATGGCCGCGCTGGCTTCACTCGAAATTTCAATCGCGCCT
      _L__K__E__L__V__Q__R__T__G__A__T__E__V__S__F__K__V__S__A__E_

                     PvuII
      AAGCAGCGAACAGGCGCGTCAGCTGGCGGAACTGATGGCGAAACTGGTTCGTGAACTGGG
181   ---------+---------+---------+---------+---------+---------+
      TTCGTCGCTTGTCCGCGCAGTCGACCGCCTTGACTACCGCTTTGACCAAGCACTTGACCC
      _S__S__E__Q__A__R__Q__L__A__E__L__M__A__K__L__V__R__E__L__G_

                                                              HindIII
      CATGAGCGAAATTACCATCACCATTAACGGCACCAGCGTGCAGATTGATGTGAAAGTGTA
241   ---------+---------+---------+---------+---------+---------+
      GTACTCGCTTTAATGGTAGTGGTAATTGCCGTGGTCGCACGTCTAACTACACTTTCACAT
      _M__S__E__I__T__I__T__I__N__G__T__S__V__Q__I__D__V__K__V__*_

           SacI
      AGCTTGGAGCTCCAGCTTTTGTTCCC
301   ---------+---------+------
      TCGAACCTCGAGGTCGAAAACAAGGG
```

VII.6.4. Gensequenz M4

```
          KpnI  NdeI
      GGGCGAATTGGGTACCCATATGAAAGTGACCGCGACCGTTACCGTTGAAAAAGATGGCCA
1     ---------+---------+---------+---------+---------+---------+
      CCCGCTTAACCCATGGGTATACTTTCACTGGCGCTGGCAATGGCAACTTTTTCTACCGGT
                         M__K__V__T__A__T__V__T__V__E__K__D__G__Q_


      GAAAGTTACCGTTAAAGTTGAAGTTACCACCGAAAGCGAACTGAAAGAACTGCTGGAACG
61    ---------+---------+---------+---------+---------+---------+
      CTTTCAATGGCAATTTCAACTTCAATGGTGGCTTTCGCTTGACTTTCTTGACGACCTTGC
      _K__V__T__V__K__V__E__V__T__T__E__S__E__L__K__E__L__L__E__R_


      TCTGGCGGAAGCGCTGCGTAAAAGCGGCGCGAGCGAAATCACCATTAAATTCAGCGCGGA
121   ---------+---------+---------+---------+---------+---------+
      AGACCGCCTTCGCGACGCATTTTCGCCGCGCTCGCTTTAGTGGTAATTTAAGTCGCGCCT
      _L__A__E__A__L__R__K__S__G__A__S__E__I__T__I__K__F__S__A__E_


                       PvuII
      AACCAGCGATCAGGCGCGTCAGCTGCTGGAACTGGCGGCGCGTCTGATTCAGAATCTGGG
181   ---------+---------+---------+---------+---------+---------+
      TTGGTCGCTAGTCCGCGCAGTCGACGACCTTGACCGCCGCGCAGACTAAGTCTTAGACCC
      _T__S__D__Q__A__R__Q__L__L__E__L__A__A__R__L__I__Q__N__L__G_


                                                 ClaI            HindIII
      CATGACCGAAGTGACCATCAAAGTGGAAGGCAACGAAGTTACCATCGATGTTAAAGTTTA
241   ---------+---------+---------+---------+---------+---------+
      GTACTGGCTTCACTGGTAGTTTCACCTTCCGTTGCTTCAATGGTAGCTACAATTTCAAAT
      _M__T__E__V__T__I__K__V__E__G__N__E__V__T__I__D__V__K__V__*_


            SacI
      AGCTTGGAGCTCCAGCTTTTGTTCCC
301   ---------+---------+------
      TCGAACCTCGAGGTCGAAAACAAGGG
```

VII.6.5. Gensequenz M5

```
          KpnI  NdeI
     GGGCGAATTGGGTACCCATATGAAAGTGAAAATCACCGCGACCGTTGAAAAAGATGGCAA
1    ---------+---------+---------+---------+---------+---------+
     CCCGCTTAACCCATGGGTATACTTTCACTTTTAGTGGCGCTGGCAACTTTTTCTACCGTT
                       M__K__V__K__I__T__A__T__V__E__K__D__G__K_

    HincII                              BglII     PvuII
     AGAAGTTAACATCAAAGTTGAAGTTGAAACCAACGAAGATCTGAAACAGCTGCTGGAACG
61   ---------+---------+---------+---------+---------+---------+
     TCTTCAATTGTAGTTTCAACTTCAACTTTGGTTGCTTCTAGACTTTGTCGACGACCTTGC
     _E__V__N__I__K__V__E__V__E__T__N__E__D__L__K__Q__L__L__E__R_

                                            BspMI
     TCTGGCGGAAGCGTTTCGCGAAACCGGCGCGAAAAATGTGCAGGTGGAAATTAGCGCGGA
121  ---------+---------+---------+---------+---------+---------+
     AGACCGCCTTCGCAAAGCGCTTTGGCCGCGCTTTTTACACGTCCACCTTTAATCGCGCCT
     _L__A__E__A__F__R__E__T__G__A__K__N__V__Q__V__E__I__S__A__E_


     AAGCAGCGAACAGGCGAAAAAACTGATTGAACTGATCGCGCGTCTGCTGGAAAAACTGGG
181  ---------+---------+---------+---------+---------+---------+
     TTCGTCGCTTGTCCGCTTTTTTGACTAACTTGACTAGCGCGCAGACGACCTTTTTGACCC
     _S__S__E__Q__A__K__K__L__I__E__L__I__A__R__L__L__E__K__L__G_

                          ClaI                                  HindIII
     CATGAGCGAAATCAAAATCGAAATCGATGGCAACCGTGTGCGTGTTGATGTTACCGCGTA
241  ---------+---------+---------+---------+---------+---------+
     GTACTCGCTTTAGTTTTAGCTTTAGCTACCGTTGGCACACGCACAACTACAATGGCGCAT
     _M__S__E__I__K__I__E__I__D__G__N__R__V__R__V__D__V__T__A__*_

          SacI
     AGCTTGGAGCTCCAGCTTTTGTTCCC
301  ---------+---------+------
     TCGAACCTCGAGGTCGAAAACAAGGG
```

VII.6.6. Gensequenz M6

```
          KpnI  NdeI
      GGGCGAATTGGGTACCCATATGAAAGTGAAAGTGACCGCGACCGTTGATGATAACGGCAA
1     ---------+---------+---------+---------+---------+---------+
      CCCGCTTAACCCATGGGTATACTTTCACTTTCACTGGCGCTGGCAACTACTATTGCCGTT
                         M__K__V__K__V__T__A__T__V__D__D__N__G__K_

                                                PvuII    PvuII
      AAAAATGGAAATCAAAGTGAACATCAAAACCCGTCAGCAGCTGGAACAGCTGATTGAACG
61    ---------+---------+---------+---------+---------+---------+
      TTTTTACCTTTAGTTTCACTTGTAGTTTTGGGCAGTCGTCGACCTTGTCGACTAACTTGC
      _K__M__E__I__K__V__N__I__K__T__R__Q__Q__L__E__Q__L__I__E__R_

                                                  BspMI         BssHII
      TCTGGCGGAAGCGTTTCGCGAAACCGGCGCGAAAAATGTGCAGGTTGAAGTTAGCGCGCA
121   ---------+---------+---------+---------+---------+---------+
      AGACCGCCTTCGCAAAGCGCTTTGGCCGCGCTTTTTACACGTCCAACTTCAATCGCGCGT
      _L__A__E__A__F__R__E__T__G__A__K__N__V__Q__V__E__V__S__A__Q_

                                                BssHII
      GACCAGCGAACAGGCGGAACGTCTGCTGGAAATTTTTGCGCGCCTGCTGAAAGAACTGGG
181   ---------+---------+---------+---------+---------+---------+
      CTGGTCGCTTGTCCGCCTTGCAGACGACCTTTAAAAACGCGCGGACGACTTTCTTGACCC
      _T__S__E__Q__A__E__R__L__L__E__I__F__A__R__L__L__K__E__L__G_

                                                HincII                HindIII
      CATGACCGATGTTACCATTAACGTGGAAGGCAACGAAGTTAACATCAAAGTTACCGCGTA
241   ---------+---------+---------+---------+---------+---------+
      GTACTGGCTACAATGGTAATTGCACCTTCCGTTGCTTCAATTGTAGTTTCAATGGCGCAT
      _M__T__D__V__T__I__N__V__E__G__N__E__V__N__I__K__V__T__A__*_

            SacI
      AGCTTGGAGCTCCAGCTTTTGTTCCC
301   ---------+---------+------
      TCGAACCTCGAGGTCGAAAACAAGGG
```

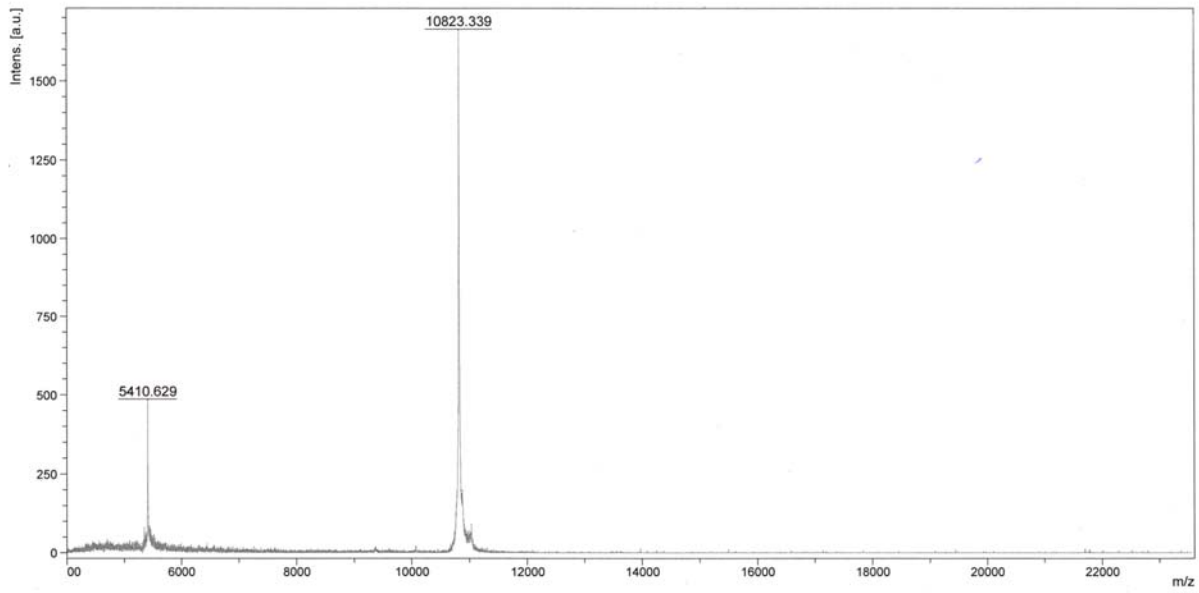
VII.6.7. Gensequenz M7

```
           KpnI   NdeI       EcoRV
      GGGCGAATTGGGTACCCATATGAAAGTGGATATCACCATCAAAATTCAGCGTGATGGCCA
1     ---------+---------+---------+---------+---------+---------+
      CCCGCTTAACCCATGGGTATACTTTCACCTATAGTGGTAGTTTTAAGTCGCACTACCGGT
                        M__K__V__D__I__T__I__K__I__Q__R__D__G__Q_

                  EcoRV          AgeI                       PstI
      GGAAATCGAAATTGATATCCGCGTGAGCACCGGTAAAGAACTGGAACGTGCGCTGCAGGA
61    ---------+---------+---------+---------+---------+---------+
      CCTTTAGCTTTAACTATAGGCGCACTCGTGGCCATTTCTTGACCTTGCACGCGACGTCCT
      _E__I__E__I__D__I__R__V__S__T__G__K__E__L__E__R__A__L__Q__E_

                    BssHII    BssHII
      ACTGGAAAAAGCGCTGGCGCGCGCGGGCGCGCGTAATGTTCAGATTACCATCAGCGCGGA
121   ---------+---------+---------+---------+---------+---------+
      TGACCTTTTTCGCGACCGCGCGCGCCCGCGCGCATTACAAGTCTAATGGTAGTCGCGCCT
      _L__E__K__A__L__A__R__A__G__A__R__N__V__Q__I__T__I__S__A__E_

                                                        PstI
      AAATGATGAACAGGCGAAAGAACTGCTGGAACTGATTGCGCGTCTGCTGCAGAAACTGGG
181   ---------+---------+---------+---------+---------+---------+
      TTTACTACTTGTCCGCTTTCTTGACGACCTTGACTAACGCGCAGACGACGTCTTTGACCC
      _N__D__E__Q__A__K__E__L__L__E__L__I__A__R__L__L__Q__K__L__G_

       EcoRV                                                    HindIII
      CTATAAAGATATCAACGTGCGCGTTAATGGCACCGAAGTGAAAATCGAAGTTCGCGTTTA
241   ---------+---------+---------+---------+---------+---------+
      GATATTTCTATAGTTGCACGCGCAATTACCGTGGCTTCACTTTTAGCTTCAAGCGCAAAT
      _Y__K__D__I__N__V__R__V__N__G__T__E__V__K__I__E__V__R__V__*_

      SacI
      AGCTTGGAGCTCCAGCTTTTGTTCCC
301   ---------+---------+------
      TCGAACCTCGAGGTCGAAAACAAGGG
```

VII.6.8. Gensequenz M8

```
          KpnI  NdeI
      GGGCGAATTGGGTACCCATATGAAAGTGAAAGTGACCGTTACCGTTGAAAAAGATGGCAA
1     ---------+---------+---------+---------+---------+---------+
      CCCGCTTAACCCATGGGTATACTTTCACTTTCACTGGCAATGGCAACTTTTTCTACCGTT
                         M__K__V__K__V__T__V__T__V__E__K__D__G__K_

    HincII                                              PvuII
      AGAAGTTAACATCAAAGTTGAAGTTGAAACCGGCGAACGTCTGAAACAGCTGCTGGAACG
61    ---------+---------+---------+---------+---------+---------+
      TCTTCAATTGTAGTTTCAACTTCAACTTTGGCCGCTTGCAGACTTTGTCGACGACCTTGC
      _E__V__N__I__K__V__E__V__E__T__G__E__R__L__K__Q__L__L__E__R_

                        PvuII           BstEII           BssHII
      TATTAAACAGGCGATGGCGCAGCTGGGTGCGAAACAGGTTACCGCGACCGTTAGCGCGCG
121   ---------+---------+---------+---------+---------+---------+
      ATAATTTGTCCGCTACCGCGTCGACCCACGCTTTGTCCAATGGCGCTGGCAATCGCGCGC
      _I__K__Q__A__M__A__Q__L__G__A__K__Q__V__T__A__T__V__S__A__R_


      TACCACGAGCGAAGCGCAGAAACTGATTAAACTGTGGCAGGAACTGATGAAAGAACTGGG
181   ---------+---------+---------+---------+---------+---------+
      ATGGTGCTCGCTTCGCGTCTTTGACTAATTTGACACCGTCCTTGACTACTTTCTTGACCC
      _T__T__S__E__A__Q__K__L__I__K__L__W__Q__E__L__M__K__E__L__G_

                                                    ClaI           HindIII
      CATGAGCGAAATCAAAATCAAAGTGGAAGGCGATAAAGTTACCATCGATGTTGATGCGTA
241   ---------+---------+---------+---------+---------+---------+
      GTACTCGCTTTAGTTTTAGTTTCACCTTCCGCTATTTCAATGGTAGCTACAACTACGCAT
      _M__S__E__I__K__I__K__V__E__G__D__K__V__T__I__D__V__D__A__*_

        SacI
      AGCTTGGAGCTCCAGCTTTTGTTCCC
301   ---------+---------+------
      TCGAACCTCGAGGTCGAAAACAAGGG
```

VII.7. Massenspektrum des Proteins M7



VII.8. Biophysikalische Eigenschaften der Proteine

Tabelle VII-4 Berechnete physikochemische Eigenschaften der Proteine M1-M8. Die Daten wurde mit dem Programm *ProtParam* (http://www.expasy.org/tools/protparam.html) bestimmt. *Vor Thrombinspaltung* Die Daten beziehen sich auf die errechneten Sequenzen mit dem N-terminalen Hexa-Histidin-*tag* (MGSS**HHHHHH**SSGLVPRGSHM). *Nach Thrombinspaltung* Die Angaben beziehen sich auf die Aminosäure-sequenzen nach Spaltung des Hexa-Histidin-*tags*. Nach der Thrombinspaltung befinden sich am N-Terminus zusätzlich zu der modellierten Sequenz die Aminosäuren GSHM.

Protein	Vor Thrombinspaltung			Nach Thrombinspaltung		
	MG (Da)	Theoretischer pI	Extinktionskoeffizient ($M^{-1}cm^{-1}$) 280 nm	*MG* (Da)	Theoretischer pI	Extinktionskoeffizient ($M^{-1}cm^{-1}$) 280 nm
M1	12480.2	7.07	1490	10598.2	5.86	1490
M2	12669.4	8.96	1490	10787.4	8.18	1490
M3	12427.1	6.65	-	10545.0	5.41	-
M4	12365.0	6.37	-	10483.0	5.21	-
M5	12585.3	6.38	-	10703.3	5.28	-
M6	12606.3	7.07	-	10724.3	5.89	-
M7	12702.5	7.08	1490	10820.4	5.90	1490
M8	12545.5	9.22	5500	10663.4	8.90	5500

VII.9. Die BLOSUM62-Matrix

Die BLOSUM62-Matrix [Henikoff & Henikoff, 1992] wurde dem Programm *PyMOL* [Delano, 2002] entnommen und diente als Substitutionsmatrix für die in dieser Arbeit durchgeführten Alignments.

	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**	**B**	**Z**	**X**	*****
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*****	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit ausschließlich unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe. Die aus anderen Werken wörtlich oder inhaltlich übernommenen Stellen sind als solche gekennzeichnet. Ich habe diese Arbeit bisher weder an der Martin-Luther-Universität Halle-Wittenberg, noch an einer anderen Bildungseinrichtung zur Erlangung eines akademischen Grades eingereicht.

Halle (Saale), den 1. September 2007

Roman Dallüge

Danksagung

Ich möchte mich ganz herzlich bei Prof. Rainer Rudolph, Dr. Christian Lange, PD Dr. Hauke Lilie, Dr. Christian Lücke, Jan Oschmann, Olaf Birkenmeier, Sabine Bergelt, Hagen Hofmann, Franzi Leich, Christoph Parthier und Björn Schott für ihre Unterstützung bei der Anfertigung dieser Arbeit, die sehr fruchtbaren Diskussionen, vielen Zigaretten und für die unvergessliche Zeit am Institut für Biotechnologie der Martin-Luther-Universität Halle-Wittenberg bedanken.

Lebenslauf

PERSÖNLICHE DATEN

Name:	Dallüge
Vorname:	Roman
Anschrift:	An der Magistrale 37 06124 Halle (Saale)
Geburtsdatum:	30.11.1972
Geburtsort:	Eisenach/Thüringen
Familienstand:	ledig

AUSBILDUNG

Seit 09/2007	*Datamanager Clinical Trial* bei MDS Pharma Services Central Lab GmbH (Hamburg)
03/2007-08/2007	Weiterbildung zum *Clinical Datamanager* beim mibeg-Institut-Medizin in Köln
01/2005-02/2007	Doktorand am Institut für Biotechnologie der Martin-Luther-Universität Halle-Wittenberg
01/2001-12/2004	Wissenschaftlicher Mitarbeiter und Doktorand in der Firma ACGT Progenomics AG in Halle (Saale)
07/2000-12/2000	Wissenschaftlicher Mitarbeiter am Max-Delbrück-Zentrum für Molekulare Medizin in Berlin-Buch
10/1998-07/2000	Fortführung und Beendigung des Chemiestudiums an der Humboldt-Universität zu Berlin Diplomarbeit: „Untersuchungen zum targetierten nicht-viralen *in vitro* Gentransfer“
10/1995-07/1998	Fortführung des Chemiestudiums an der Universität Potsdam
10/1993-07/1995	Beginn eines Chemiestudiums an der Technischen Universität Chemnitz-Zwickau
09/1992-07/1993	Projektleiter AG-Chemie an der EOS „Ernst Abbe“
09/1990-07/1992	EOS „Ernst Abbe“, Eisenach/Thüringen
09/1979-07/1990	5. POS „Liselotte Hermann“, Eisenach/Thüringen

PUBLIKATIONEN

Haberland, A., **Dallüge, R**., Zaitsev, S., Stahn, R., Böttger, M.
Ligand-histone H1 conjugates: Increased solubility of DNA complexes, but no enhanced transfection activity
Somat. Cell. Mol. Genet. 25 (1999), 237-245

Haberland, A., **Dallüge, R**., Erdmann, B., Zaitsev, S., Cartier, R., Schäfer-Korting, M., Böttger, M.
Polycation-mediated transfection: How to overcome undesirable side effects of sticky complexes
Somat. Cell Mol. Genetics 25 (1999), 327-332

Haberland, A., Zaitsev, S., **Dallüge, R**., Schäfer-Korting, M., Böttger, M.
New peptides for efficient gene transfer
New Drugs 4 (2001), 44-47

Lucius, H., Haberland, A., **Dallüge, R**., Zaitsev, S., Schneider, M., Böttger, M.
Structure of transfection-active histone H1/DNA complexes
Molec. Biol. Rep. 28 (2001), 157-165

Dallüge, R., Haberland, A., Zaitsev, S., Schneider, M., Zastrow, H., Sukhorukov, G., Böttger, M.
Characterization of structure and mechanism of Transfection-active peptide-DNA complexes
Biochim. Biophys. Acta 1576 (2002), 45-52

Haberland, A., **Dallüge, R**., Zaitsev, S., Schneider, M., Zastrow, H., Sukhorukov, G., Böttger, M.
Peptide-mediated gene transfer: Effect of complex size on transfection efficiency and targeting behaviour
Biol. Membrany 20 (2003), 278-287

Dallüge, R., Oschmann, J., Birkenmeier, O., Lücke, C., Lilie, H., Rudolph, R. Lange, C.
"A Tetrapeptide Fragment-based Design Method Results in Highly Stable Artificial Proteins"
Proteins (2007), 68(4): 839-849

POSTER

Dallüge, R., Haberland, A., Zaitsev, S., Schneider, M., Sukkhorukov, G., Akari, S., Möhwald, H., Böttger, M.
Characterization of structure and mechanism of transfection-active peptide-DNA complexes
SXM4, Intenat. Conference on the Development and Application of Scanning Probe Methods
28.-29. 9. 2000, Münster, Deutschland

Böttger, M., Haberland, A., **Dallüge, R**., Zaitsev, S., Sukorukov, G., Schneider, M., Zastrow, H., Möhwald, H.
Complex aggregation interferes with receptor-specific uptake at peptide-mediated gene transfer
8. Meeting of European Society of Gene Therapy
7. – 10. 10. 2000, Stockholm, Schweden

Böttger, M., Haberland, A., Zaitsev, S., **Dallüge, R**.
Nuclear protein-mediated gene transfer
43. Symposium of the Society for Histochemistry
26. – 29. 9. 2001, Wien, Österreich

Böttger, M., Haberland, A., Zaitsev, S., **Dallüge, R**., Sukhorukov, G.
From polycationic transfection systems to aggregation-free polyelectrolyte nanoparticles
9. Meeting of European Society of Gene Therapy
2. – 4. 11. 2001, Antalya, Türkei

PATENTE

„Peptide for gene transfer, useful e.g. in gene therapy, comprises polylysine segment with C-terminal extension, provides high transfection efficiency without receptor targeting“ — DE 100 27 414 A1
Böttger, M., Haberland, A., **Dallüge, R**., Zaitsev, S.
(2000)

„Gene transfer to animal cells, useful in vitro or for gene therapy, using vector DNA treated sequentially in water with oppositely charged polyelectrolytes“ — DE 101 57 799 A1
Böttger, M., Zaitsev, S., Haberland, A., **Dallüge, R**., Sukhorukov, G., Schneider, M., Zastrow, H., Möhwald, H.
(2000)

„Methods for establishing and analyzing the Conformation of amino acid sequences"
Dalluege Roman, Boehm Gerald
(2003)

WO05027009

Halle (Saale), den 1. September 2007

Roman Dallüge