Eric Frenzel

Deskriptoren zur Beschreibung eines QSER-Modells

Bachelorarbeit

Hochschule Mittweida

UNIVERSITY OF APPLIED SCIENCES

Mathematik/Naturwissenschaften/Informatik

Mittweida, 2010

Eric Frenzel

Deskriptoren zur Beschreibung eines QSER-Modells

eingereicht als

Bachelorarbeit

an der

Hochschule Mittweida

UNIVERSITY OF APPLIED SCIENCES

Mathematik/Naturwissenschaften/Informatik

Mittweida, 2010

Erstprüfer: Prof. Dr. rer. nat. Dirk Labudde
Zweitprüfer: Dipl. Inf. (FH) Daniel Stockmann

Bibliographische Beschreibung

Frenzel, Eric:
Deskriptoren zur Beschreibung eines QSER-Modells. – 2010. – 109 S. Mittweida, Hochschule Mittweida, Fachbereich Mathematik/Naturwissenschaften/Informatik, Bachelorarbeit, 2010

"Die Wissenschaft kann die letzten Rätsel der Natur nicht lösen. Und das ist so, weil wir letztlich selbst ein Teil des Rätsels sind, das wir zu lösen versuchen."

Max Planck (1858-1947)

Referat

QSAR- und QSPR-Modelle finden bereits seit etwa 40 Jahren in der Arzneimittelherstellung, dem *Drug-Design*, Anwendung. Durch sie lassen sich Verhaltensweisen von chemischen Substanzen vorhersagen, sodass Interaktionen und Reaktionen mit anderen Chemikalien abgeschätzt werden können. Dieser Ansatz lässt sich auch auf Proteine übertragen. Der Faltungsweg eines Proteins hin zu seiner nativen Struktur kann noch nicht genau bestimmt werden. Durch die energetische Beschreibung mit Hilfe von Deskriptoren kann dies allerdings möglich werden. Diese Bachelorarbeit soll sich mit möglichen Deskriptoren befassen, um anschließende mathematische Analysen hin zu einem QSER-Modell (*Quantitative Structure Energy Relationships*) zu vereinfachen. Dabei sollen Zusammenhänge zwischen dem Energieprofil und der Struktur eines globulären Proteins erklärt, erforscht und verstanden werden. Auch Grundgedanken für die weitere Vorgehensweise sollen diskutiert und verglichen werden. Das abschließende Ziel der gesamten Thematik ist die Erklärung struktureller Ausbildungen eines Proteins unter Integration der energetischen Charakterisierungen.

Abstract

Since about 40 years QSAR and QSPR models are used in the medication production – the so-called Drug-Design. Due to the models, it is possible to predict the behaviour patterns of chemical substances, so that interactions and reactions with other chemicals can be estimated. This approach can be used for proteins as well. The folding pathway of a protein towards its native structure cannot yet be defined precisely. This can be realised by means of the energetic specification with the help of descriptors. In this Bachelor Thesis descriptors are analysed which will simplify the subsequent mathematical analysis with regards to the QSER model (*Quantitative Structure Energy Relationships*). In this process the interrelations between the energy profile and the structure of a globular protein will be explained, researched and understood. In addition, fundamental ideas for further procedures shall be discussed and compared. The concluding aim of the whole topic is to explain the structure activities of a protein while integrating energetic characterisations.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Was wir schon alles wissen

Die energetische Betrachtung zur Erforschung und Charakterisierung von Proteinen ist ein sehr junger Ansatz. Die Berechnung der freien Energie aller Aminosäuren ist dabei der Ausgangspunkt aller Ausführungen. Die Bezeichnung *Residuum* wird in der gesamten Arbeit an einer Vielzahl von Stellen benutzt und ist mit dem Ausdruck *Aminosäure* gleichzusetzen. Die ersten Analysen sowie der Algorithmus zur Berechnung der freien Energie aller Residuen in einem globulären Protein wurden von Dr. Frank Dressel entwickelt. Im Zusammenhang mit seiner Dissertationsschrift an der Technischen Universität Dresden formulierte er alle nötigen Vorschriften zur Berechnung der Energie von Proteinen dieser Klasse. Zwei weitere Kommilitonen haben sich ebenfalls dieser Analysestrategie für die Erforschung von Proteinen verschrieben und verfolgen in ihrer Bachelorarbeit andere Ansätze und Ideen. Im Verlauf des letzten halben Jahres sind wir so gemeinschaftlich zu einer Vielzahl von sehr interessanten Erkenntnissen gekommen. Dabei haben sich sowohl die unterschiedliche Herangehensweise an Probleme von jedem Mitarbeitenden positiv auf den Ergebnis-Pool ausgewirkt, als auch differenzierte Interessen in diesem Fachgebiet. Jegliche Analysen und Ergebnisse, welche unter den Gliederungspunkten 1.3 bis 1.6 aufgeführt sind, können in den Belegarbeiten von Florian Heinke sowie Riccardo Brumm und Eric Frenzel nachgelesen werden. Diese Resultate sind für das bessere Verständnis der energetischen Betrachtungen wichtig und werden deshalb in dieser Arbeit aufgeführt.

1.1 Grundlagen über Proteine

Proteine sind Bestandteile aller lebenden Strukturen. Sie nehmen existenzielle Positionen in allen Stoffwechselvorgängen ein und werden nach ihrem Auftreten und ihren Eigenschaften gemäß in zwei Gruppen eingeteilt. Zum einen gibt es die globulären Proteine, welche in Wasser beziehungsweise in einer Umgebung, die den hydrophilen Charakter dieser Proteine unterstützt, löslich sind und zum anderen die Membranproteine. Diese treten in den Bereichen von Biomembranen auf und können an diese an- oder eingelagert

sein. Wenn sie auf der Membran sitzen werden sie als *periphere Membranproteine* bezeichnet, wenn sie die Membran durchziehen als *Transmembranproteine*. Diese Gruppe der Proteine gibt den Zellen Stabilität und sorgt für einen reibungslosen Stoffaustausch mit benachbarten Zellen oder Zellorganellen. Die Globulären Proteine erfüllen einen erheblich größeren Umfang an Funktionen und kommen in einer wesentlich komplexeren Anzahl und Diversität vor. Sie wirken beispielsweise als Enzyme zur Katalyse chemischer Reaktionen, als Antikörper zur Abwehr krankheitsauslösender Erreger oder als molekularer Träger für den Stofftransport. Das wohl bekannteste Beispiel ist dabei das Hämoglobin im menschlichen Blut, welches für die Sauerstoffversorgung verantwortlich ist.

Die Grundeinheiten aller Proteine bilden die 20 L-α-Aminosäuren, welche auch als proteinogene Aminosäuren bezeichnet werden. Diese werden bei der Proteinbiosynthese an den Ribosomen zu der Primärstruktur eines Proteins verknüpft. Außer Prolin besitzen alle proteinogenen Aminosäuren den gleichen grundsätzlichen Aufbau. An das zentrale C_α-Atom sind eine Carboxylgruppe, eine primäre Aminogruppe, ein Wasserstoffatom und die Seitenkette, welche oftmals auch als *Restgruppe* oder *Rest* bezeichnet wird, substituiert. Prolin besitzt statt der primären eine sekundäre Aminogruppe. *Abbildung 1* verdeutlicht schematisch den grundsätzlichen Aufbau.

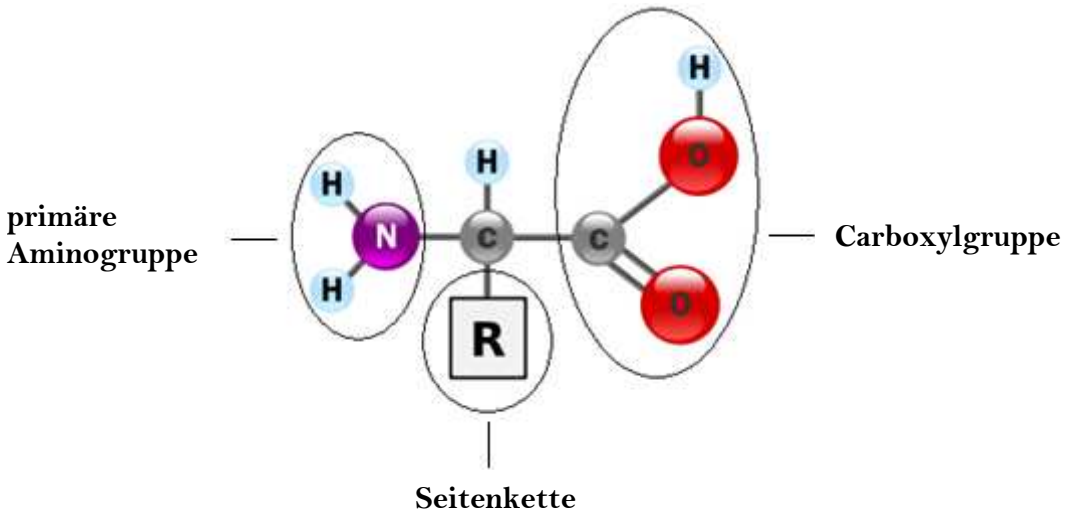

Abbildung 1: *Der allgemeine Aufbau der proteinogenen Aminosäuren aus der Carboxylgruppe, der primären Aminogruppe und der Seitenkette. Einzige Ausnahme bildet Prolin. Dieses besitzt statt einer primären eine sekundäre Aminogruppe. Das zentrale C-Atom wird auch als C_α-Atom bezeichnet.* [1]

Bei der Faltung bilden sich wiederkehrende Strukturen aus, welche man auch als Sekundärstrukturelemente bezeichnet. Dies sind die wohlgeformten Helices und Faltblattstrukturen sowie die Coil-Bereiche. Die Verknüpfung von den Aminosäuren wird über die Peptidbindung erreicht. Dabei wird die Carboxylgruppe einer Aminosäure mit der Aminogruppe einer zweiten Aminosäure, unter Abspaltung eines Wasseratoms, zu einem Dipeptid verknüpft. Die Abfolge dieser Grundeinheiten wird als *backbone* bezeichnet und ist in jedem Protein wiederkehrend und identisch.

Aminosäure 1 + ***Aminosäure 2*** ↔ ***Dipeptid*** + ***Wasser***

Diese Reaktion ist zur besseren Verdeutlichung auf atomarer Ebene in *Abbildung 2* dargestellt. Eine Peptidbindung kann durch verschiedene Möglichkeiten aufgespalten werden. Die Denaturierung durch Wärme ist sicherlich die bekannteste, jedoch gibt es auch hitzestabile Proteine, welche sich erst bei über 100°C in ihre Bausteine auflösen. Desweiteren können auch Peptidasen, welche nach der EC-Systematik der Enzyme vollzählig zu der Gruppe *EC 3.4* gehören, sowie die Veränderung des pH-Wertes oder das Anhaften an eine Oberfläche für die Spaltung verantwortlich sein. [3]

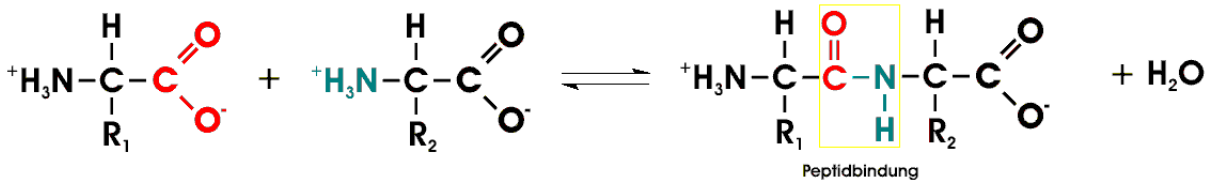

Abbildung 2: *Verknüpfung von 2 Aminosäuren zu einem Dipeptid unter Abspaltung eines Wassermoleküls. Modifiziert nach* [2].

Die drei Sekundärstrukturelemente unterscheiden sich maßgeblich in ihrer Form. Dabei bilden sich auch die Wasserstoffbrückenbindungen differenziert aus, welche im höchsten Maß für die Stabilität des Proteins verantwortlich sind. Dies geschieht immer zwischen den beiden Segmenten „CO“ und „NH“ von zwei Aminosäuren. Bei einer Helix stets entlang des

Stranges und mit in einem Abstand von vier Aminosäuren, demnach die ***i***-te und ***i+4***-te. Diese strukturelle Eigenschaft ist in *Abbildung 3* und *4* abgebildet. Erstere beschreibt es an einer vereinfachten ebenen Darstellung, die nachfolgende in räumlicher Orientierung.

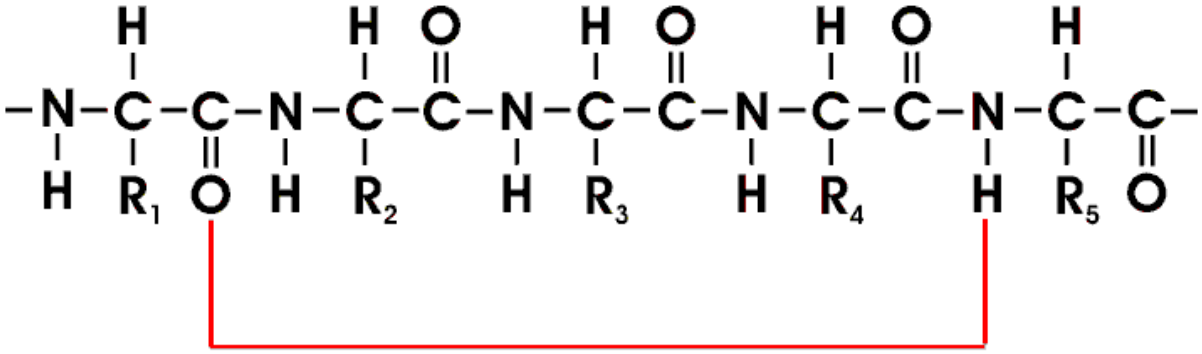

Abbildung 3: *Verdeutlichung einer ebenen Darstellung der Wasserstoffbrückenausbildung in einer Helix.* [4]

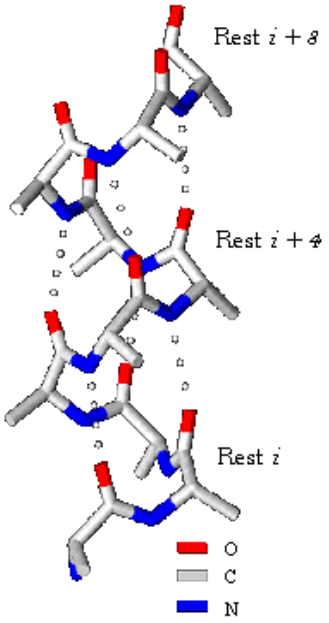

Abbildung 4: *Die räumliche Darstellung der Wasserstoffbrückenausbildung in einer Helix zwischen Stickstoff und Sauerstoff.* [5]

In Faltblattstrukturen bilden sich diese stabilisierenden Bindungen meistens übergreifend zwischen räumlich nebeneinander liegenden Polypeptidketten-Abschnitten in dem Protein aus. Diese bezeichnet man als *strands*. In der Natur gibt es parallele sowie antiparallele Faltblätter. Ihre Orientierung wird bei der Faltung bestimmt. Die Ausbildung der Wasserstoffbrücken ist dabei in *Abbildung 5* dargestellt. In Ausnahmefällen können sie sich auch innerhalb eines *strands* ausbilden.

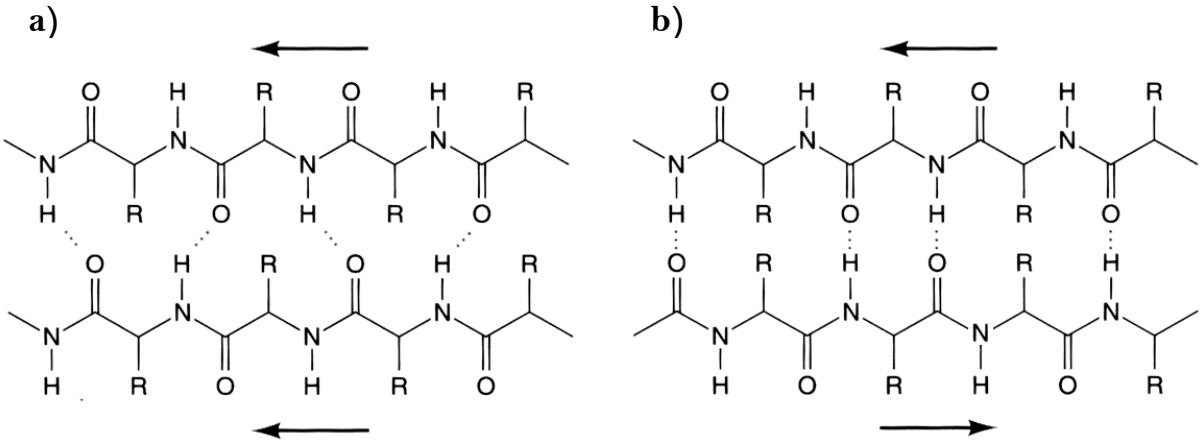

Abbildung 5: *Darstellung der beiden Arten einer Faltblattstruktur.* ***a)*** *Paralleles Faltblatt,* ***b)*** *Antiparalleles Faltblatt.* ***Die*** *einzelnen Stränge werden als* ***strands*** *bezeichnet. Stabilisierende Wasserstoffbrückenbindungen werden in der Regel zwischen räumlich benachbarten Sauerstoff- und Stickstoffatomen ausgebildet.* [6]

Coil-Strukturen verbinden in Form von *Loops* oftmals Helices und Faltblattstrukturen und liegen dabei im Vergleich zu Faltblättern und Helices unstrukturiert vor. Wasserstoffbrückenbindungen bilden sich je nach Länge an unterschiedlichen Positionen aus. Nach der Denaturierung eines Proteins liegen alle Bereiche in dieser Konformation vor und das Protein ist nicht mehr in der Lage seine angestammte Funktion auszuführen, wird dann aber als *Random Coil* bezeichnet. Natürlich gibt es hierbei Ausnahmen, denn besonders robuste Proteine, wie bspw. BMP-2 (*Bone Morphogenetic Protein*), können nahezu nicht dauerhaft denaturiert werden. [7]

1.2 Der Algorithmus zur Bestimmung der freien Energie eines Proteins

Das Programm zum Erstellen eines Energieprofils wurde von Herr Dr. Frank Dressel angefertigt. Bevor es erstmals Anwendung fand, wurde von ihm eine Statistik verfasst, mit welcher sich die Verteilung der Aminosäuren in Proteinen näherungsweise genau beschreiben lässt. Dabei wurde eine große Anzahl an Proteinen untersucht (2500) und in Folge ein Innen-/Außen-Kriterium entwickelt, um festzustellen, ob die jeweils betrachtete Aminosäure an ihrer definierten Position innen oder außen in dem Protein vorliegt. Dafür wurde immer das C_β-Atom der Aminosäure betrachtet, denn dieses ist für die Orientierung der Seitenkette, welche die Ausrichtung im Raum beschreibt, verantwortlich. Dies wurde für jede Aminosäure des Proteindatensatzes durchgeführt, wobei die entstandene Statistik in *Tabelle 1* aufgeführt ist. Die Daten korrelieren mit denen anderer wissenschaftlicher Untersuchungen auf diesem Fachgebiet.

Tabelle 1: *Die von Dr. Frank Dressel aufgestellte Statistik für die Charakterisierung der Lage aller proteinogenen Aminosäuren in Bezug auf ihre Ausrichtung in den Proteinen korrelieren mit anderen wissenschaftlichen Untersuchungen. Kriterium für die Entscheidung war die Lage des C_β-Atoms, da dieses die Orientierung der Aminosäure bzw. der Seitenkette festlegt und somit darüber entscheidet, ob die betrachtete Aminosäure als außen oder innen angesehen wird.*

Aminosäure	innen	außen	Aminosäure	innen	außen
Cys	4582	1016	His	6419	3366
Ile	20370	4141	Gly	16698	14326
Ser	12576	10411	Asp	10001	14327
Gln	7373	7752	Leu	30615	7107
Lys	9285	15193	Arg	11327	10441
Asn	8225	8928	Trp	4001	1193
Pro	9135	9423	Val	23562	6551
Thr	12537	9622	Glu	11165	18091
Phe	13353	2813	Tyr	11228	3529
Ala	22725	11052	Met	7003	1723

Die Vorgehensweise der Charakterisierung gestaltet sich für jede Aminosäure wie nachfolgend verfasst:

1. Man nehme die zu betrachtende Aminosäure *i* und lege eine Kugel mit einem Radius von 10Å herum.
2. Aus allen Aminosäuren in dieser Kugel wird der Schwerpunkt berechnet.
3. Die Aminosäure *i* ist innen, wenn eine der folgenden Bedingung erfüllt ist:

$$|C_\alpha - c| < 5 \vee (C_\alpha - C_\beta)(C_\alpha - c) < 0 \qquad (1)$$

4. Wenn beide obigen Bedingungen nicht erfüllt sind, wird die Aminosäure *i* außen angenommen.

Durch den Term $|C_\alpha - c| < 5$ wird rein der Abstand zu entfernten Aminosäuren betrachtet, wobei sich der Ausdruck $(C_\alpha - C_\beta)(C_\alpha - c) < 0$ auf Abstand sowie Orientierung der Seitenkette bezieht.

Nun kann durch Zuhilfenahme der Boltzmannverteilung die Energiedifferenz dieser beiden Zustände einer Aminosäure ausgedrückt werden:

$$e'_{i0} = -k_B T \ln\left(\frac{n_{\text{innen}}}{n_{\text{außen}}}\right) \qquad (2)$$

e'_{i0}	≙	Energiedifferenz des Innen/Außen-Kriteriums
$-k_B T$	≙	Boltzmannverteilung
n_{innen} / $n_{\text{außen}}$	≙	absolute Anzahlen an spezifisch betrachteter Aminosäure, dass diese innen oder außen vorliegt

Um paarweise Energien zwischen zwei Aminosäuren *i* und *j* zu ermitteln, wurde folgende Formel aufgestellt:

$$e^{*ij} = -k_B T \ln\left(\frac{\mathrm{n_{ij}}}{N_{\text{Kontakt}} * p_i * p_j}\right) \qquad (3)$$

e^{*ij}	≙	Paarenergie zwischen zwei Aminosäuren i und j
n_{ij}	≙	Anzahl an beobachteten Kontakten zwischen Aminosäuren *i* und *j* in dem betrachteten Protein
N_{Kontakt}	≙	Anzahl aller auftretenden Kontakte der Aminosäuren *i* und *j* in allen Proteinen des Datensatzes
p_i / p_j	≙	Wahrscheinlichkeit, dass eine zufällig gewählte Aminosäure gleich der betrachteten Aminosäure *i bzw. j* ist

Die relativen Wahrscheinlichkeiten p_i und p_j der Aminosäuren berechneten sich als Quotient aller gezählten Kontakte der betrachteten Aminosäure durch die komplette Anzahl an Aminosäure-Aminosäure-Kontakten. Die Werte sind in nachstehender *Tabelle 2* vermerkt.

Tabelle 2: *Die Statistik über die relativen Wahrscheinlichkeiten in Bezug auf das Ausbilden von Kontakten zu anderen Residuen wurden für jede Aminosäure als Quotient aus allen ihren gezählten Kontakten zu der Gesamtanzahl an Interaktionen ermittelt.*

Aminosäure	**p_i(in %)**	**Aminosäure**	**p_i(in %)**
Cys	1,8	His	2,4
Ile	5,7	Gly	7
Ser	6,3	Asp	5,8
Gln	4	Leu	8,9
Lys	6,3	Arg	5,1
Asn	4,4	Trp	1,4
Pro	4,5	Val	6,9
Thr	5,5	Glu	6,8
Phe	4,1	Tyr	3,5
Ala	7,5	Met	2

Das Potential der Aminosäure *i* bzw. seiner benachbarten Aminosäure *j* wird zur Berechnung des Paarpotentials zweier Aminosäuren benötigt. Mit nachstehender Formel wird die Energie pro Kontakt von Aminosäure *j* mit Aminosäure *i* berechnet

$$e_{i0} = \left(\frac{1}{\alpha_i}\right) e'_{i0} \tag{4}$$

e_{i0} ≙ Energie pro Kontakt von Aminosäure *j* zu Aminosäure *i*

α_i ≙ Anzahl der räumlich benachbarten Aminosäuren *j* zur Aminosäure *i* in einem maximalen Radius von 8Å

Abschließend kann nun das einfache Paarpotential e_{ij} formuliert werden, welches für jegliche Berechnungen der freien Energie Anwendung findet.

$$e_{ij} = e_{i0} + e_{j0} + e^{*ij} \tag{5}$$

1.3 Energetische Charakterisierung der Aminosäuren

Die erste Analyse der Vergangenheit beschäftigte sich damit, die energetische Diversität der kanonischen Aminosäuren zu ermitteln. Dazu wurde ein Datensatz (*Set*) von 80 globulären Proteinen mit Hilfe der *PDB* (*Protein Data Bank*) erstellt. Die Proteine sollten absolut rein sein. Sie durften also keine Liganden aufweisen, sollten aus einer Polypeptidkette (*chain*) bestehen und mussten bei einer Auflösung (*resolution*) von maximal 2Å (*Angström*) kristallisiert werden. Das *Set* wurde mit voranschreitender Erfahrung und bioinformatischen Möglichkeiten auf 114 Proteine erweitert (≙ kleinem Datensatz) und für einige Analysen stand zusätzlich ein maximierter Datensatz aus 4303 Proteinen zur Verfügung, jedoch sind bei diesem die Ergebnisse auf Grund der Datenmenge nicht ohne Weiteres zu verarbeiten. Darüber hinaus sind zusätzlich Proteine mit einer Auflösung von über 2Å enthalten und auch eine gewisse Anzahl an teilweise ungeeigneten Proteinen ist auf Grund mangelhafter *PDB*-Einträge vorhanden. Begründet durch die große Datenfülle ist

jedoch eine verhältnismäßig korrekte Analyse möglich. Der kleine Datensatz liefert auch aussagekräftige und präzise Erkenntnisse, da globuläre Proteine verschiedener Klassen enthalten sind und auch die Sekundärstrukturelemente in jeweils ausreichender Anzahl vorkommen. Dadurch lassen sich die Ergebnisse auch auf eine größere Datenmenge übertragen und es ist nur mit einer vertretbar geringen Abweichung zu kalkulieren. In dem maximierten *Set* wurde die durchschnittliche freie Energie aller Residuen ermittelt. Diese ist in *Tabelle 3* als *E-Value* bezeichnet. Alle Aminosäuren sind im *3-Letter-Code* dargestellt. Er ist unter Punkt 4.1 für alle proteinogenen Aminosäuren aufgegliedert. Die im mathematischen Mittel stabilste Aminosäure ist gemäß den Energiewerten Cystein, gefolgt von Phenylalanin und Leucin. Die instabilsten sind Lysin, Glutaminsäure sowie Asparaginsäure.

Tabelle 3: *Die energetische Charakterisierung der kanonischen Aminosäuren verdeutlicht die Stabilität. Der E-Value beschreibt die durchschnittliche freie Energie. Cystein, Phenylalanin und Leucin sind die im mathematischen Durchschnitt stabilsten Residuen, Lysin, Glutaminsäure und Asparaginsäure die instabilsten.*

Aminosäure	E-Value	Aminosäure	E-Value
Ala	-14,71	Leu	-25,33
Arg	-5,94	Lys	-0,74
Asn	-4,33	Met	-24,21
Asp	-1,83	Phe	-25,68
Cys	-26,03	Pro	-5,45
Gln	-5,30	Ser	-6,93
Glu	-0,88	Thr	-8,40
Gly	-7,62	Trp	-20,49
His	-12,60	Tyr	-19,64
Ile	-27,33	Val	-22,60

Proteine besitzen in ihrem nativen Zustand zumeist ihre niedrigste freie Energie. Dies dient zur Minimierung ihrer Entropie und hilft zusätzlich eine Abschirmung gegenüber dem Lösungsmittel zu schaffen, sodass sich keine Wechselwirkungen aufbauen. In Ausnahmefällen kann es auch zur Ausbildung der nativen Konformation in einem etwas

höheren Energieniveau kommen, jedoch nur in einem geringen Maße. Dies stellten S. Govindarajan und R. A. Goldstein fest und veröffentlichten ihre Analysen unter dem Titel „On the thermodynamic hypothesis of protein folding, PNAS 95“. [13] Desweiteren gibt es auch spezielle Proteine, welche zwischen mehreren stabilen Tertiärstrukturen wechseln können. Man nennt sie *allosterische Proteine*. Ihre Strukturen sind dabei über Energiebarrieren voneinander abgetrennt. [8]

Anhand aller Energien wurden vier Bereiche, sogenannte Quantile, für eine bessere Klassifizierung und Einschätzung ermittelt. Dazu wurden alle Datenpunkte erfasst und die drei Grenzen jeweils nach einem Viertel der Datenelemente gesetzt. Dadurch ergeben sich folgende Abschnitte:

1. Quantil: $x \geq -3{,}09$
2. Quantil: $-3{,}09 > x \geq -8{,}37$
3. Quantil: $-8{,}37 > x \geq -20{,}27$
4. Quantil: $x < -20{,}27$

Durch diese Einteilung ist es möglich die Stabilität der Aminosäuren zu erläutern. Residuen mit einem Energiewert aus dem 4. Quantil weisen eine hohe Beständigkeit gegen physikochemische Eigenschaften auf, wohingegen jene aus dem 1. Quantil sehr instabil sind und an dieser Stelle die Polypeptidkette eher zu Veränderungen gebracht werden kann. Aus mittleren Energiewerten resultiert ein ambivalentes Verhalten der Aminosäuren.

Um eine Visualisierung der Diversität von Energiewerten in einem Profil zu verdeutlichen, ist in *Abbildung 6* ein Energieprofil dargestellt. Dieses stammt von dem Protein mit der *PDB ID* „1CRN“. Dabei wird ersichtlich, dass es aus allen drei Sekundärstrukturelementen besteht („S“, „H“ und „-“). Die Minima der Energien stellen ausschließlich Aminosäuren aus Faltblättern („S“) und Helices („H“) dar. Dies wurde auch bei der Analyse weiterer Proteine festgestellt. Vermutlich ist dies auf die geordnete Ausbildung der Wasserstoffbrückenbindungen in diesen Strukturen und auf ihre zentrale Lage in den Proteinen zurück zu führen. Helices sind oft von einem „Zick-Zack-Muster“ geprägt. Der gewundene Aufbau der Helix, die resultierende alternierende Anordnung und damit die spezifische Lage der Aminosäuren sind wahrscheinlich dafür verantwortlich. Das

Profil stammt aus dem Programm „EProVis3.1“, welches von meinem Kommilitonen Riccardo Brumm im Zusammenhang mit einer Ausarbeitung in dem Fach „Datenrepräsentation“, entwickelt wurde. Auf die Möglichkeiten dieses nützlichen Tools werde ich in Punkt 2.1 noch einmal ausführlicher zu sprechen kommen.

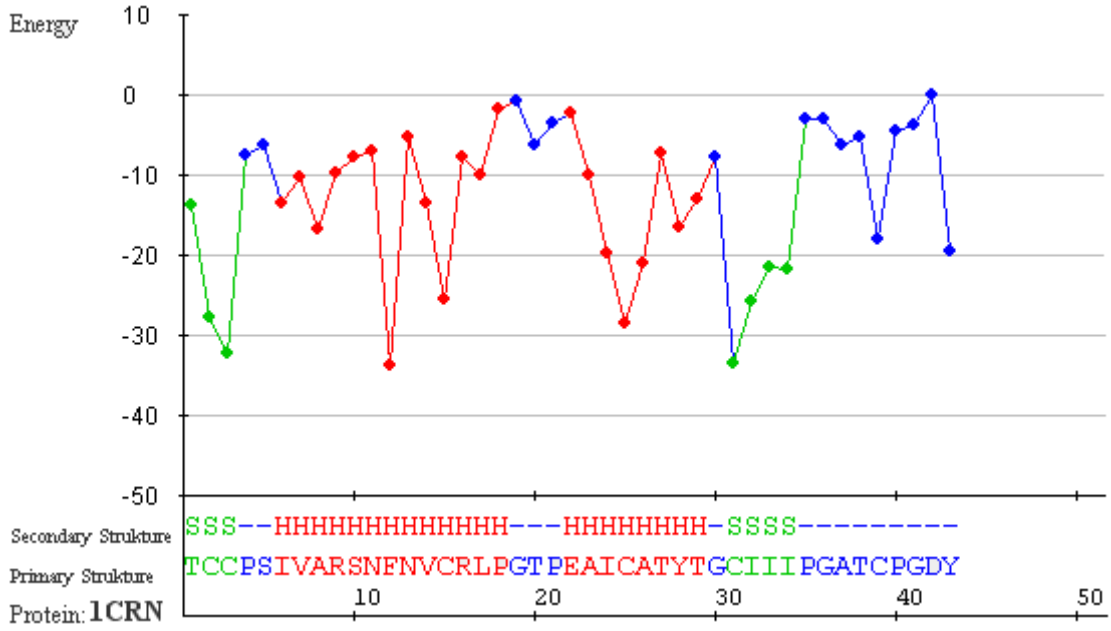

Abbildung 6: *Das Energieprofil eines Crambin-Kristalls mit der PDB ID „1CRN" beinhaltet alle Sekundärstrukturen und die Energieminima, welche die stabilsten Bereiche des Proteins darstellen. Sie sind in den beiden Helices sowie den Faltblättern ausgebildet.*

1.4 Energetische Charakterisierung der Sekundärstrukturelemente

Die *Tabelle 4* verdeutlicht die Verteilung der Datenpunkte des kleinen Datensatzes (19890, 114 Proteine) nach den Sekundärstrukturelementen. Hierbei ist deutlich nachvollziehbar, dass ein Gradient in der Ausbildung der Energien besteht. Coil-Strukturen sind meist im instabilen oder ambivalenten Bereich mit Hang zur Instabilität (1. und 2. Quantil) anzufinden. Faltblätter bilden hingegen meist energetisch stabile Bereich aus. Die

energetische Verteilung in den Helices ist mit der spiralisierten Anordnung der Aminosäuren zu beantworten. Der eine Teil ragt nach außen in das Lösungsmittel hinein, das andere Fragment ist in Richtung des Proteininneren geneigt und besitzt somit eine geringere freie Energie, da weniger Wechselwirkungen auftreten. Auch die bereits erwähnte anzahlmäßige sehr gleichförmige Verteilung der Sekundärstrukturelemente ist erkennbar (Coil = 7868, Sheet = 5074, Helix = 7316).

Tabelle 4: *Nach der Klassifizierung und Zuordnung aller Aminosäuren in ihre zugehörigen Sekundärelementen und die Aufteilung nach den Quantilen lassen sich erste Rückschlüsse auf die energetische Klassifizierung schließen. Die stabilste Sekundärstruktur ist ein Sheet. Im Gegensatz dazu sind Coil-Bereiche sehr oft instabil und treten oft in Wechselwirkung mit dem Lösungsmittel. Helices sind über den gesamten Energiebereich sehr homogen verteilt mit einer Neigung zu einer stabilen Konformation, welche im 3. und 4. Quantil ausgebildet wird.*

	1. Quantil	2. Quantil	3. Quantil	4. Quantil
Coil	2663	2404	1803	996
Sheet	495	798	1398	2017
Helix	1788	1805	1764	1959

1.5 Energetische Korrelation der Torsionswinkel

Der Grundgedanke dieser Analyse war die Annahme, dass durch die Abfolge der Torsionswinkel strukturelle Vorhersagen gemacht werden können. Das Proteinrückgrat, auch *backbone* genannt, wird normalerweise durch die Winkel φ (Phi) und ψ (Psi), gelegentlich zusätzlich durch ω (Omega) beschrieben. Wenn das *backbone* in das Lösungsmittel hinein ragt nimmt die Wechselwirkungsenergie der Residuen zu und sie würden in Richtung das Proteininneren gedrängt werden. Florian Heinke verfasste dahingehend ein Programm und reduzierte wegen dargelegter Tatsache diese Analyse auf die Winkelabfolge zwischen den C_α-Atomen. Dies ist in *Abbildung 7* dargestellt.

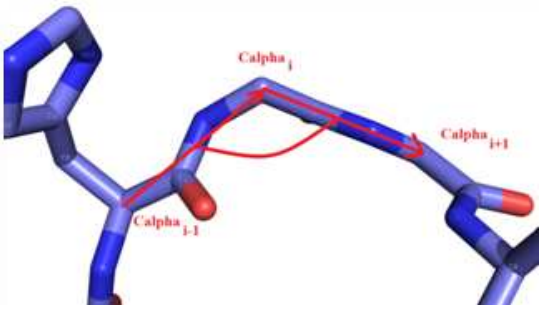

Abbildung 7: *Betrachtung des simplifizierten Torsionswinkels zwischen den C_α-Atomen drei sequenziell aufeinanderfolgender Aminosäuren.* [9]

Es wurde festgestellt, dass für hochenergetische Residuen die Abfolge der Winkel indifferent ist und nicht vorhergesagt werden kann. Für niederenergetische Aminosäuren lässt sich der Verlauf des backbone jedoch relativ zuverlässig beschreiben, denn meist treten Richtungsänderungen von 70-75° auf.

1.6 SASA-Analyse

SASA bedeutet *solvent accessible surface area* und beschreibt die Lösungsmittelzugänglichkeit eines Proteins. Sie wird berechnet, indem eine Kugel mit einem Durchmesser von 1,4Å, welche ein Wassermolekül repräsentiert, über die komplette Van-der-Waals-Oberfläche des Proteins abgerollt wird. Der Hintergrund der Analyse war es, heraus zu finden, wie sich Aminosäuren des gesamten energetischen Spektrums verhalten. Bei der Analyse konnte eine hohe Abhängigkeit zwischen der Energie eines Residuums und seiner Lösungsmittelzugänglichkeit bewiesen werden. Aminosäuren mit einem hohen Energiewert besitzen eine höhere Lösungsmittelzugänglichkeit als niederenergetische. Dies bestätigt die Analysen und Schlussfolgerungen der energetischen Betrachtung der Aminosäuren und der Sekundärstrukturelemente.

2 Was wir schon alles haben – Tools

2.1 eProVis3.1 – Visualisierung von Energieprofilen

Dieser in Java implementierte Viewer ermöglicht es, Energieprofile von Proteinen zu visualisieren. Dabei wird das *PDB*-File benötigt. Dieses muss bei erstmaligem Aufruf von einem Speichermedium abgerufen werden. Über den Button „Save“ kann man dieses Profil dann zu einem späteren Zeitpunkt wieder aufrufen und benötigt das *PDB*-File nicht mehr. Natürlich lassen sich gespeicherte Profile auch wieder löschen. Dies wurde mit dem „Delete“-Button realisiert. Von einem Profil ist sowohl die Sequenz (Primärstruktur), als auch die dazugehörige Sekundärstruktur über die gesamte Struktur ersichtlich. Über „bottom bar“ kann man noch ein weiteres Energieprofil einspeisen, sodass ein Abgleich zwischen den beiden betrachteten Profilen gemacht werden kann. Desweiteren stehen verschiedene Farb-Schemata der Profile zur Verfügung. Eine normale Ansicht ohne Einfärbung, eine „Green-Blue-Coloration“, welche sehr passend bei der Betrachtung von zwei Proteinen ist, um in den beiden übereinander gelegten Profilen die Orientierung zu behalten und abschließend noch eine Farbauswahl, in welcher die verschiedenen Sekundärstrukturbereiche unterschiedlich koloriert sind. Diese Ansicht ist auch in *Abbildung 8* von der C-terminalen Gamma-B Domäne mit der *PDB ID* „1DSL“ dargestellt. Bei der Darstellung der Sekundärstruktur steht „-“ für einen Coil-Bereich, „S“ für ein Faltblatt und „H“ für einen Helix-Abschnitt. Durch die Länge mancher Proteine ist es unabdingbar eine Scrollbar einzufügen, um durch die gesamte Sequenz zu gelangen. Auch wenn zwei Proteine gleichzeitig betrachtet werden ist dies möglich. Um eine druckbare Form eines Profils zu erhalten kann man es als PDF abspeichern und optimal auch auf unterschiedlichen Betriebssystemen verwenden.

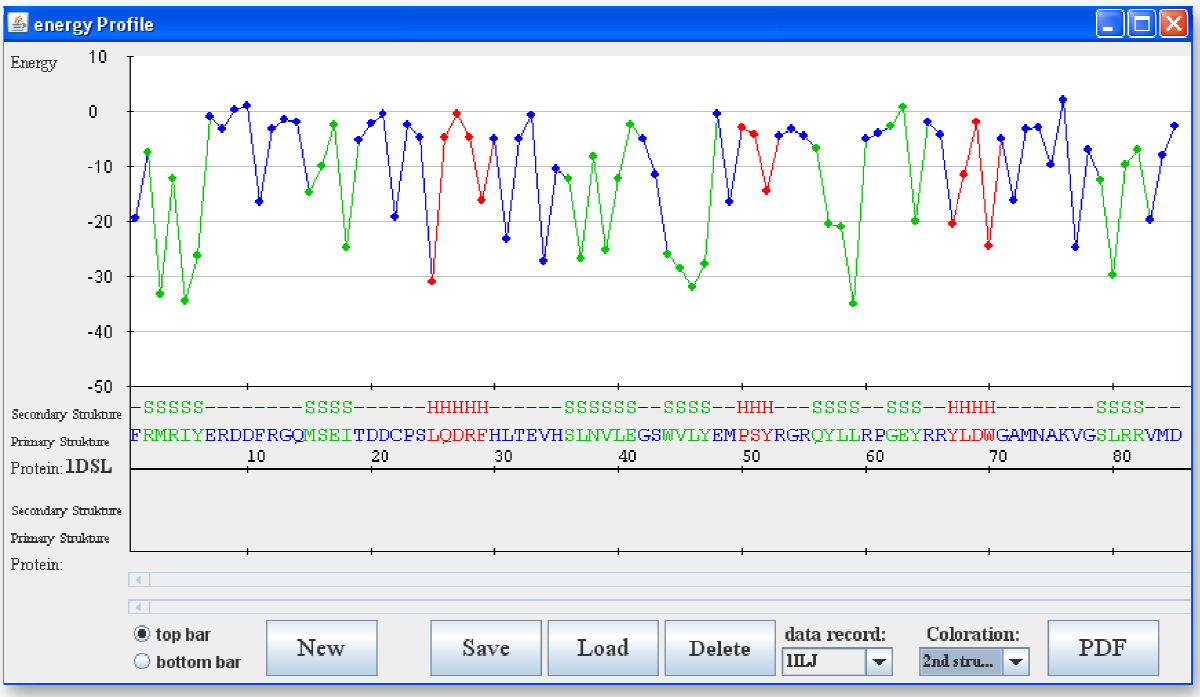

Abbildung 8: *Die Benutzeroberfläche von eProVis3.1 gestaltet sich übersichtlich. Neben dem Energieprofil ist zusätzlich die Primär- sowie die Sekundärstruktur angegeben. Das Profil lässt sich nach verschiedenen Charakteristika, welche auf struktureller sowie sequenzieller Basis gründen, einfärben. Die Ausgabe in ein PDF ist eine hervorragende Variante, um die Profile in ein allgemeingültiges Format zu überführen und damit die Plattformunabhängigkeit zu gewährleisten.*

2.2 Das Superalignment

Jegliche im Internet angebotene Alignment-Tools haben Schwächen, da sie sich nur auf eine spezifische Eigenschaft stützen. Sei es nun die räumliche Struktur (Tertiärstruktur), wie bei dem *DaliLite*-Algorithmus, oder die Betrachtung der Primärstruktur mit Hilfe von *Align* (beide von dem *European Bioinformatics Institute*). Florian Heinke kombinierte drei Charakteristika, welche die Betrachtung von sequenzieller und struktureller Ebene, sowie die berechneten Energiewerte berücksichtigt und schuf somit einen Alignment-Algorithmus mit einer sehr hohen Genauigkeit. Die Ergebnisse sind geprüft besser als jene, die *DaliLite* voraussagt. Der Algorithmus ist auf globuläre Proteine anwendbar.

Der programmspezifische Hintergrund stützt sich auf die dynamische Programmierung unter Zuhilfenahme des Needleman-Wunsch-Algorithmus und einer selbst erstellten Scoring-Funktion, um die Energieprofile werten zu können.

Ein „all-against-all-Durchlauf“ bewies die hohe Sensitivität des Verfahrens. Desweiteren wurde die Homologie für zwei Proteine nachgewiesen, welche nur eine geringe Sequenzidentität aufweisen, unterschiedliche Funktionen ausüben, aber strukturell dennoch fast identisch sind.

Abbildung 9 verdeutlicht den Output des Programms. Dabei sind die beiden Sequenzen der zu vergleichenden Proteine farbig abgebildet, sowie darunter bzw. darüber die Sekundärstrukturelemente (Linie ≙ Coil, Slash ≙ Helix, Balken ≙ Faltblatt). Die schwarze Raute zwischen zwei Aminosäuren wird vermerkt, wenn zwei von drei Charakteristika zutreffen und die rote Raute, wenn alle Betrachtungen repräsentativ übereinstimmen. Der Wert von **SL** repräsentiert einen relativierten Score in Abhängigkeit zur Länge des Alignments (inklusive Gaps).

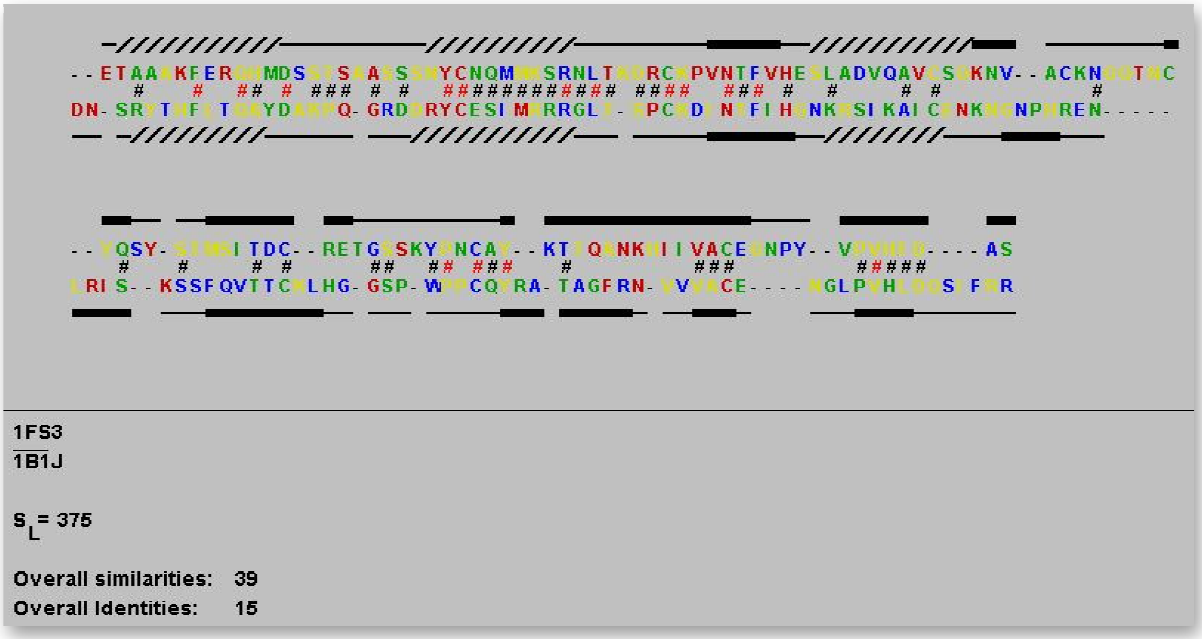

Abbildung 9: *Output des Superalignment-Programms. Dabei werden die zu vergleichenden Sequenzen farbig aufgeführt. Die zugehörigen Sekundärstrukturen sind zusätzlich über bzw. unter der Sequenz visualisiert. Die Übereinstimmungen an sequenziell verglichener Position werden nach Berechnung durch den Algorithmus in Form der farbig markierten Rauten dargestellt.*

2.3 Der eGOR-Algorithmus

Der klassische *Garnier-Osguthorpe-Robson*-Algorithmus (*GOR*-Algorithmus) wurde zur Vorhersage der Sekundärstrukur von Proteinen entwickelt. [10] Es wird dabei mit einem angemessen großen Datensatz eine Statistik der Sekundärstrukturen geschaffen. Diese wird anschließend bei der Vorhersage zur Entscheidung hinzu gezogen. Die Vorhersagegenauigkeit liegt etwa bei 65%.

Der eGOR-Algorithmus dient für die Vorhersage von Energieprofilen. Dahingehend mussten natürlich die Parameter des Inputs angepasst werden. Die für den normalen GOR verwendete Fenstergröße von üblichen 17 Residues wurde auf sieben verkleinert und auch die Wahrscheinlichkeitsmatrizen wurden durch HashMaps abstrahiert.

Das Ergebnis verdeutlichte eindrucksvoll die Korrektheit dieses Ansatzes. In 95% der Fälle wurde ein richtiger Verlauf des Energieprofils vorhergesagt. Somit lässt sich das Tool optimal in weitere Ansätze einbinden und garantiert eine geringe Verarbeitungszeit.

3 Grundlagen für ein QSER-Modell

3.1 Geschichtliches

Die Idee von einem *Quantitative Structure Energy Relationships*-Modell (QSER-Modell) stammt von dem im *Drug-Design* angewendeten QSAR-Modell (*Quantitative Structure Activity Relationships*-Modell). Diese quantitativen Struktur-Wirkungsbeziehungen werden seit etwa 1990 bei der Charakterisierung und der Erklärung von Zusammenhängen zwischen der chemischen Struktur und den biologischen Wirkungen einer chemischen Substanz erfasst. Daraus werden folglich Schlussfolgerungen abgeleitet, um Gesetzmäßigkeiten zu erkennen und aufwendige experimentelle Versuche zu minimieren. Die Grundlagen für die QSER-Analysen stützen sich auf die unterschiedlichen physiko-chemischen Eigenschaften der Aminosäuren und die konträre Verteilung sowie Abfolge der Residuen in den Proteinen und damit zusammenhängende unterschiedliche Funktionsweisen. Die ersten QSAR-Untersuchungen basierten auf der Analyse von Substitutionen verschiedener Liganden an definierte Stoffe, bspw. einem Benzolring, und verarbeiten die Beobachtungen und Messungen der veränderten Eigenschaften nach jedem Substituent. Die Ansätze für diese Modelle schufen 1964 Corwin Hansch und Toshio Fujita mit einem mathematischen Schema für die quantitative Beschreibung der Lipophilie diverser Stoffe. Sie leiteten durch eine große Anzahl an Analysen *Gleichung* (6) (in Ansätzen dargestellt) her.

$$\log\frac{1}{C} = -k_1\,(\log P)^2 + k_2 \log P + k_3\delta + \ldots k \qquad (6)$$

C	≙	molare Konzentration eines Stoffes, welche definierte biologische Reaktion verursacht
$\log P$	≙	Logarithmus des Octanol/Wasser-Verteilungskoeffizienten P

δ	≙	Hammetkonstante (abhängig von Substituent und dessen Stellung an Ausgangssubstanz)
k	≙	Koeffizienten (ermittelt durch Regressionsanalyse)

Der quadratische Term deklariert die quantitative Beschreibung nichtlinearer Lipophilie-Wirkungsbeziehungen. Bei linearer Abhängigkeit entfallen dieser Term sowie alle weiteren, welche nicht relevant für die spezifische Betrachtung sind. In dem gleichen Jahr wurde ein zweites Modell für die Berechnung der Lipophilie verschiedender Substanzen von S. R. Free und J. W. Wilson entwickelt. Der Ansatz war dabei, dass die Ausgangsverbindung einen definierten Beitrag μ zu der biologischen Wirkung liefert und die Substituenten einen additiven Beitrag a_i zu der biologischen Aktivität der Verbindung ausüben. Dies ist in *Formel* (7) verdeutlicht:

$$\log\frac{1}{C} = \sum a_i + \mu \qquad (7)$$

Um die 3D-Struktur eines Moleküls zu erklären bzw. aus einer 3D-Struktur Einflüsse auf berechenbare oder messbare Größen abzuleiten sind Deskriptoren (Zustandsbeschreibungen) für die Beschreibung der 3D-Struktur notwendig. Auch die Vorhersage von spezifischen Bereichen in Proteinen, welche als Motive bezeichnet werden, bzw. die allgemeine energetische Beschreibung der Struktur eines Proteins umfasst die Analyse mit Hilfe von Deskriptoren. Sie werden auf theoretischer Ebene unter Gliederungspunkt 4 dargelegt. [11] [12]

3.2 Multivariate Analysemethoden mit dem Schwerpunkt der Regressionsanalyse

Die Welt der Mathematik hält viele Varianten bereit, um eine große Anzahl von Daten auf einmal zu verarbeiten und Korrelationen zwischen einzelnen Komponenten herzustellen. Um die Vielzahl von Deskriptoren werten zu können und festzustellen, welcher einen höheren sowie niedrigeren Einfluss auf die Ausbildung von Motiven oder die Faltung im dreidimensionalen Raum hat, benötigt man eine multivariate Analysemethode, womit Zusammenhänge ermittelt werden können. Da noch nicht klar ist, welcher Faktor den größten Einfluss ausübt und wie sich nachfolgende qualitativ und quantitativ einordnen, muss das mathematische Gerüst auf diese Analyse abgestimmt werden.

Sicherlich gibt es viele verschiedene Methoden für eine solche praktische Betrachtung, die wohl am besten geeignetste ist jedoch die Regressionsanalyse. Beim Aufstellen von QSAR-Modellen findet sie bereits seit Jahren zumeist Anwendung. [14] Im Nachfolgenden soll sie in Ansätze etwas plausibler erklärt werden. Im Verlauf der Bachelorarbeitszeit ist eine Durchführung solch eines Verfahrens leider nicht realisierbar, da es viel Zeit in Anspruch nimmt und auch eine gewisse Routine und ein ausgeprägtes Verständnis dafür vorhanden sein sollte, um Fehler bestmöglich zu vermeiden.

Die Regressionsanalyse in ihren verschiedenen Ausprägungen ist wahrscheinlich die meist eingesetzte multivariate Analysemethode in Wirtschaft und Wissenschaft. Sie lässt sich beliebig modellieren und eignet sich besonders für „Ursache-Wirkungs-Beziehungen“, welche auch bei der Analyse der Deskriptoren vorliegen.

Dazu ist eine abhängige Variable notwendig, in unserem Fall ein Motiv oder gar die gesamte Tertiärstruktur eines betrachteten Proteins, und mehrere unabhängige Variablen, welche die Deskriptoren darstellen. Diese Zusammenhänge lassen sich formal mit der *Gleichung* (8) ausdrücken.

$$Y = f(X_1, X_2, \dots, X_n) \tag{8}$$

Y	≙	abhängige Variable (bspw. Motiv, Tertiärstruktur)
$X_{1,\dots,n}$	≙	unabhängige Variablen (Deskriptoren)

Für die Lösung und Korrelation mehrerer Einflussgrößen wird im Speziellen die multiple Regressionsanalyse benötigt. Mit einer durchgeführten Regressionsanalyse könnten sich noch mehr Fragen ausbilden, außer jene die erklärt, welcher Deskriptor den höchsten Einfluss hat. Es könnte zusätzlich untersucht werden, wie viele Deskriptoren für die Vorhersage unbedingt nötig sind und welchen eine erhöhte Aufmerksamkeit geschenkt werden sollte. Auch die Abschätzung von Veränderungen in der Struktur kann betrachtet werden. Besonderes Augenmerk ist dabei auf Mutationen zu legen und welche Veränderungen sie in der Ausbildung der Motive oder der räumlichen Struktur hervorrufen. Dies könnte auch für die Modellierung künstlicher Proteine von großem Nutzen sein.

Eine Regressionsanalyse setzt ein metrisches Skalenniveau aller Variablen voraus, d.h. es werden Variablen beschrieben, welche einen diskreten Zahlenwert besitzen. Variablen lassen sich bei Nichterfüllung dieser Bedingung auch durch die *Dummy-Variablen-Technik* in binäre Werte umwandeln und können so auch in eine Regressionsanalyse einbezogen werden, da sie sich wie metrische Variablen verhalten. Die Regressionsanalyse gliedert sich immer in fünf Einzelschritte, welche für eine optimale Lösungsfindung durchlaufen werden müssen. Diese Abfolge ist nach Vorlage des Kapitels der Regressionsanalyse aus dem Buch *Multivariate Analysemethoden – Eine anwendungsorientierte Einführung* [15] in *Abbildung 10* dargestellt.

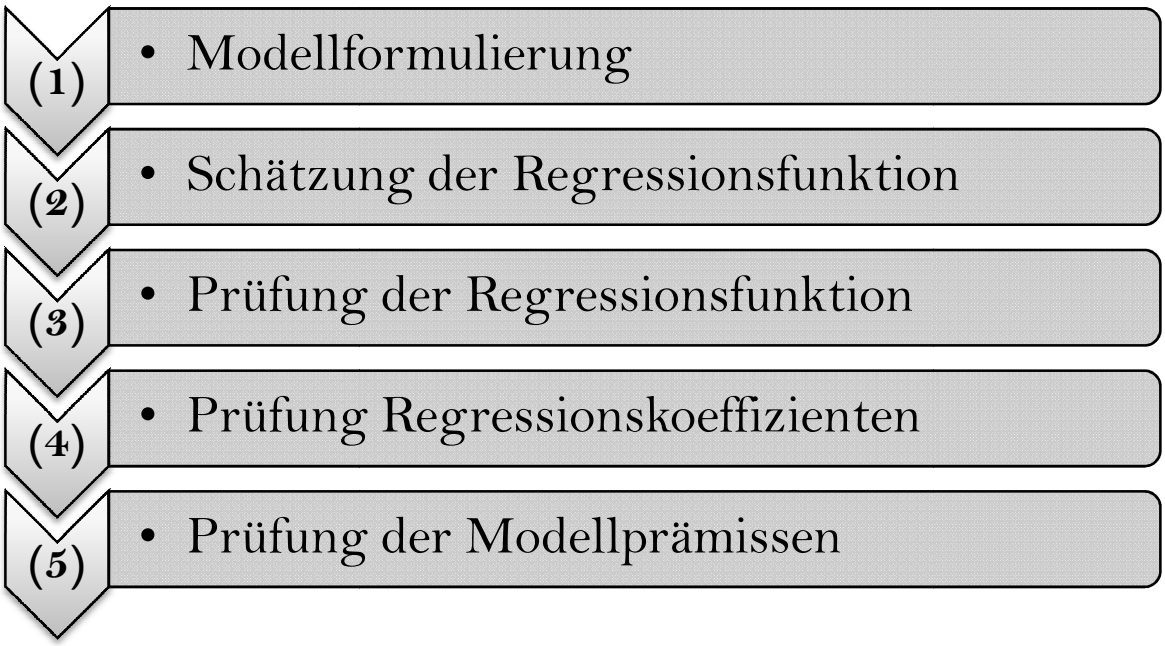

Abbildung 10: *Eine Regressionsanalyse gliedert sich in fünf Bereiche. Die Modellformulierung vermittelt den ersten Schritt. Dabei werden die Deskriptoren festgelegt. Die Schätzung der Regressionsfunktion schließt sich an diesen Schritt an und umfasst die Erstellung von Punktediagrammen, um Abhängigkeiten aufzuzeigen. Anschließend muss die Regressionsfunktion überprüft werden. Die ermittelten Regressionskoeffizienten werden daraufhin auch geprüft. Die Abschließende Untersuchung der Modellprämissen umfasst die Berechnung und Wertung von auftauchenden Störgrößen.*

Die Modellformulierung (1) beinhaltet die Analyse, welche Deskriptoren bzw. Ursache-Wirkungs-Beziehungen einbezogen werden. Um aussagekräftige Ergebnisse in Bezug auf das in der Arbeit formulierte Ziel zu treffen, muss die Auffindung strukturgebender Deskriptoren im Vordergrund stehen. An diese Arbeit anschließende Analysen müssen zeigen, ob die gewählten Deskriptoren sinnvoll sind oder eine bessere Beschreibung mit weniger, aber dafür wesentlicheren Größen exaktere Ergebnisse liefert. Um die Schätzung der Regressionsfunktion zu beginnen, muss ein Deskriptor ausgewählt werden, welcher vermutlich am ehesten einen Zusammenhang vermuten lässt und im besten Fall eine lineare Abhängigkeit verdeutlicht. Dies lässt sich durch ein Punktediagramm evaluieren.

Bei der Schätzung der Regressionsfunktion (2) gilt es heraus zu finden, wie sich unsere abhängige Variable durch die Veränderung der gewählten Deskriptoren verändert. Dazu

werden Punktediagramme aufgestellt und bei allen zwei Koordinatenpunkten ausgewählt, welche am ehesten eine Abhängigkeit aufzeigen. Daraus lässt sich eine Regressionsgerade bestimmen. Diese ist durch das Vorhandensein eines konstanten Gliedes, welches den Schnittpunkt mit der Ordinate markiert, und dem Regressionskoeffizient gekennzeichnet. Dieser beschreibt die Steigung bzw. Neigung der Gerade. Die auftretende Streuung der Datenpunkte um diese muss über die Residualgröße abgedämpft werden. Das Ziel ist es anschließend eine Funktion zu finden, welche Abweichungen möglichst optimal erfasst und verarbeitet. Dies geschieht mit Hilfe der *Methode der kleinsten Quadrate*, welche kurz als *KQ*-Methode bezeichnet wird. Die resultierende Zielfunktion enthält dann, je nach Anzahl an Deskriptoren, Regressionsparameter, welche die Wertung und Einschätzung der Deskriptoren darstellen. Diese müssen über die Lösung eines linearen Gleichungssystems bestimmt werden, was wohlmöglich mit einem großen Rechenaufwand verbunden sein kann. Die Regressionskoeffizienten dürfen nur dann als Wertung angenommen werden, wenn alle in die Rechnung einbezogenen Größen das gleiche Skalenniveau besitzen. Andernfalls muss eine Standardisierung aller Koeffizienten geschaffen werden.

Bei den nachfolgenden Schritten der Prüfung der Regressionsfunktion (3) und der Regressionskoeffizienten (4) wird in die Bereiche der globalen Überprüfung der Regressionsfunktion, wobei die Gleichung als Ganzes kontrolliert wird, und die Prüfung der Regressionsparameter unterschieden. Dabei wird der Sinnhaftigkeit der Deskriptoren besondere Aufmerksamkeit geschenkt und wie gut sie unsere gesuchte Variable beschreiben. Für die beiden Kontrollvorgänge haben sich bestimmte Methoden etabliert. Die globale Prüfung wird besonders mit dem Verfahren der Berechnung des Bestimmtheitsmaßes, der Anwendung der F-Statistik und über den Standardfehler realisiert. Zur Prüfung der Koeffizienten eignen sich der t-Wert und der Beta-Wert.

Bei der Prüfung der Modellprämissen (5) steht die Untersuchung der Störgröße im Vordergrund, welche für die Abweichung in den Daten verantwortlich ist. Sie wird von einer Vielzahl von Komponenten beeinflusst und lässt sich somit nicht trivial ausschließen. Vielmehr ist es sinnvoll, die Störgröße als eine Zufallsvariable zu behandeln und somit die Regressionsanalyse als ein stochastisches Modell anzunehmen. Wenn diese Analysen positiv verlaufen, sollte das aufgestellte Modell letztendlich noch mit reellen Werten abgeglichen werden, um eine einwandfreie Vorhersage zu gewährleisten. [15]

4 Deskriptoren zur Beschreibung des QSER-Modells

Um ein Modell für die Untersuchung von quantitativen Struktur-Energie-Wirkungsbeziehungen in Proteinen aufzustellen, müssen Deskriptoren gefunden werden. Die angeführten Deskriptoren sind vorrangig für globuläre Proteine geeignet. Für Membranproteine müssten noch weitere Beschreibungen ausgewählt werden, bspw. sollten dabei die Wechselwirkungen mit der Biomembran erfasst werden. Mit absoluter Gewissheit kann jedoch nicht gesagt werden, dass diese Deskriptoren korrekt gewählt sind. Dies muss dann die multiple Regressionsanalyse zeigen. Nicht für alle Deskriptoren bietet sich eine energetische Charakterisierung an.

4.1 Chemische Eigenschaften von Aminosäuren

Das chemische Verhalten einer Aminosäure wird maßgeblich durch ihre Seitenkette bestimmt. Dabei weisen sie eine hohe Diversität auf. Die einfachste Unterscheidung und Kategorisierung erreicht man über die vier Eigenschaften:

- Ladung
- Polarität
- Struktur
- Hydrophobizität

Tabelle 5 verdeutlicht die chemischen Eigenschaften der 20 kanonischen Aminosäuren (nach [16], [17], [18], [19]). Die beiden häufig benutzten Alphabete des *3-Letter-* und *1-Letter-Codes* werden in der zweiten und dritten Spalte aufgeführt. Sie sind aus der Bioinformatik nicht mehr wegzudenken und finden in vielen Programmen sowie zur Ausgabe in Textfiles Anwendung und schaffen eine übersichtliche und komprimierte Ansicht.

Tabelle 5: *Die vier markantesten chemischen Eigenschaften der Aminosäuren sind die Ladung, Polarität, das Hydrophobizitätsverhalten sowie die Struktur. Zur besseren Übersichtlichkeit sind die Eigenschaften unterschiedlich farbig hervor gehoben. Neben dem Namen aller Aminosäuren sind zusätzlich noch der zugehörige 1-Letter-Code sowie der 3-Letter-Code angegeben.*

Aminosäure	3-Letter-Code	1-Letter-Code	Ladung	Polarität	Struktur	Hydrophobizität
Alanin	Ala	A	ungeladen	unpolar	aliphatisch	hydrophob
Arginin	Arg	R	geladen	polar	aliphatisch	hydrophil
Asparagin	Asn	N	ungeladen	polar	aliphatisch	hydrophil
Asparaginsäure	Asp	D	geladen	polar	aliphatisch	hydrophil
Cystein	Cys	C	ungeladen	polar	aliphatisch	hydrophob
Glutamin	Gln	Q	ungeladen	polar	aliphatisch	hydrophil
Glutaminsäure	Glu	E	geladen	polar	aliphatisch	hydrophil
Glycin	Gly	G	ungeladen	unpolar	aliphatisch	hydrophob
Histidin	His	H	geladen	polar	aromatisch	hydrophob
Isoleucin	Ile	I	ungeladen	unpolar	aliphatisch	hydrophob
Leucin	Leu	L	ungeladen	unpolar	aliphatisch	hydrophob
Lysin	Lys	K	geladen	polar	aliphatisch	hydrophob
Methionin	Met	M	ungeladen	unpolar	aliphatisch	hydrophob
Phenylalanin	Phe	F	ungeladen	unpolar	Aromatisch	hydrophob
Prolin	Pro	P	ungeladen	unpolar	heterocyclisch	hydrophil
Serin	Ser	S	ungeladen	polar	aliphatisch	hydrophil
Threonin	Thr	T	ungeladen	polar	aliphatisch	hydrophob
Tryptophan	Trp	W	ungeladen	unpolar	aromatisch	hydrophob
Tyrosin	Tyr	Y	ungeladen	polar	aromatisch	hydrophob
Valin	Val	V	ungeladen	unpolar	aliphatisch	hydrophob

Die chemischen Eigenschaften im Zusammenhang mit der Sequenz eines Proteins müssten **theoretisch** zur Beschreibung der Faltung ausreichen, da jegliche Information darin enthalten ist. Dies ist allerdings nach jetzigem wissenschaftlichem Kenntnisstand nicht möglich, da der Zusammenhang der Eigenschaften noch nicht ableitbar und erklärbar ist. Dies wird wohl auch in Zukunft nur schwer möglich werden, weil zu viele Möglichkeiten auftreten können, aus welchen Aminosäuren und in welcher Abfolge sich ein Protein zusammen setzen kann. Aus diesem Grund muss man sich mit Eigenschaften helfen, welche in einem Protein gemessen oder aus Daten berechnet werden können. Dies ist der Grund, warum die angestrebte Analyse mit einer hohen Anzahl an Merkmalsbeschreibungen durchgeführt werden muss.

4.2 3D-Motive und ausgewählte Sequenzmotive

Motive können in Proteinen in verschiedenen Bereichen vorkommen und stellen definierte Sequenzabschnitte dar. Man kann sie deshalb auch als Sequenzmuster bezeichnen. [20] Sie liegen konserviert vor und können entweder strukturell relevant oder funktionell bedeutungsvoll sein. Die 3D-Motive sind Strukturmotive und für die Stabilität der Tertiärstruktur nicht unbedeutend. Die Sequenzmotive sind auch unter der Bezeichnung *PROSITE*-Motive bekannt, da sie auf der *PROSITE*-Datenbank des *Swiss Institute of Bioinformatics* (kurz *SIB*) vermerkt sind. Diese Datenbank verwaltet zusätzlich noch Daten zu Proteinfamilien sowie –domänen. [21] Die Sequenzmotive können in einer definierten Proteinfamilie vorkommen (qualitative Motive), aber auch übergreifend vorhanden sein (quantitative Motive) und werden über einen regulären Ausdruck beschrieben, welcher alle Informationen über das Muster in sich trägt. [22] Ihre Vorhersage gestaltet sich mit der Zuhilfenahme von mehr oder weniger aufwendigen multiplen Alignments.

Ähnlich einem neuronalen Netzwerk haben die Analysen von den Motiven im Zusammenhang mit dieser Bachelorarbeit den Zweck des Lernens. Das Auftreten soll studiert werden, die Lage umliegender Aminosäuren soll geprüft werden, die relativen Häufigkeiten verschiedener Residuen an den definierten Positionen in den Motiven sollen untersucht werden und auch das Ausbilden von „Ersatzresiduen“ an stark konservierten Positionen muss festgestellt werden.

4.3 Sekundärstrukturelemente von Proteinen

Unter Punkt 1.1 wurde bereits der allgemeine Aufbau von Proteinen dargestellt, weshalb an dieser Stelle nicht noch einmal allgemein darauf eingegangen wird. Der Gliederungspunkt 1.3 beschrieb bereits erste Erkenntnisse der energetischen Charakterisierung der Sekundärstrukturen. Die Analysen haben sich in dieser Arbeit noch einmal verfeinert, um eine genaue Einschätzung und Verteilung aller Aminosäuren in den drei Sekundärstrukturelementen ableiten zu können. Dazu wurden die durchschnittlichen Energien aller Residuen in den energetischen Quantilen berechnet und es können somit individuelle Aussagen über die durchschnittliche freie Energie jedes Residuums in den verschiedenen Sekundärstrukturen gemacht werden. Der Hintergrund dieser spezifischen Betrachtung beschränkt sich darauf, dass allgemein die Meinung in der Proteinanalytik besteht, dass besonders die Coil-Strukturen sich in zufälliger Anordnung ausbilden und keinen Anteil an der Reaktivität des Proteins besitzen. Die wesentlichen Informationen und auch die Ausbildung der stabilen Struktur sollen fast ausnahmslos von den Helices und den Faltblattstrukturen ausgehen. Dies ist allerdings nur schwer abzuleiten und einzig das Vorhandensein von Residuen zum Ausbilden des aktiven Zentrums weist derzeit auf ihre existenzielle Wichtigkeit hin. Sekundärstrukturanalysen beschränken sich oftmals nur auf die weitaus bekannteren Helices und Faltblätter.

4.4 Interaktionen der Aminosäuren in globulären Proteinen

Die Van-der-Waals Oberfläche eines Moleküls ist die Fläche über die Van-der-Waals Radien aller im Molekül befindlichen Atome. Die Van-der-Waals Radien kann man sich als Kugeln vorstellen. Die Werte der Radien lassen sich experimentell über Kristallbildungsexperimente oder sehr genaue *ab initio*-Berechnungen ermitteln. [12] Die Van-der-Waals Oberfläche ist ein erstes Maß für die Lösungsmittelzugänglichkeit. *Abbildung 11* verdeutlicht beispielhaft die Ausbildung der Oberfläche und die Vorstellung der Kugeln um jedes Atom. Es handelt es sich um die Aminosäure Serin (R-group: CH_2OH). Das Programm *CMA* (*Contact Map Analysis*) des höchst bekannten und angesehenen *Weizmann Institute of Science* in Rehovot, Israel, kann die molekularen Interaktionen

zwischen Proteinen, einzelnen *chains* oder Ligand-Protein-Komplexen auf Basis der Van-der-Waals Oberfläche jeder Aminosäure berechnen. Somit kann man die Anzahl der Kontakte aller Residuen in einem Protein bestimmen und zusätzlich noch feststellen, welche Aminosäure-Aminosäure-Kontakte bevorzugt vorkommen und welche Kontakte vornehmlich selten auftreten. Zusätzlich wurden noch die Sekundärstrukturelemente verarbeitet und darauf geachtet, ob sich Muster ergeben oder besonders häufige Auftreten feststellbar sind. Da die Analysen mit Hilfe des *CMA*-Programmes sehr rechenintensiv sind wurde ein Programm erstellt, mit welchem in einer virtuellen Kugel mit einem Radius von 8Å die Anzahl an Kontakten aller Aminosäure errechnet wird, in der Hoffnung globalere Aussagen treffen zu können, da diese Methode vereinfacht ist und somit ein größerer Datensatz genutzt werden kann.

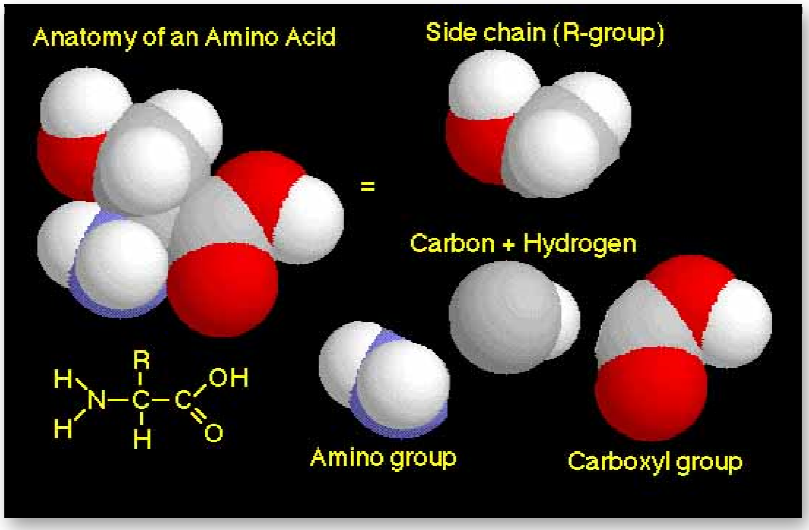

Abbildung 11: *Verdeutlichung der Van-der-Waals Oberfläche einer Aminosäure mit Hilfe eines Kugelmodells. Die Atome der Seitenkette weisen auf Serin hin. [23]*

4.5 Winkelbeziehungen in globulären Proteinen

Das *backbone* eines Proteins kann über die drei Torsionswinkel φ, ψ und ω beschrieben werden, wobei der ω-Winkel meist nicht mit einbezogen wird. Die Abstraktion auf die C_α-

Vektoren und erste Erkenntnisse darüber wurden bereits unter Punkt 1.4 vorgestellt. Die hiesige Analyse beschäftigt sich jedoch mit der Fragestellung, ob bestimmte Winkel für einzelne Aminosäuren in den Sekundärstrukturelementen besonders charakteristisch sind. Dabei werden nur der φ und ψ Winkel betrachtet, also jene, welche immer direkt ein Residuum betreffen. Die positionsspezifischen Energiewerte sollen erfasst werden, sodass Aussagen zu energetischen Ausprägungen getroffen werden können und besonders, um eine eventuell vorhandene Korrelation aller Aminosäuren oder Korrelationen einzelner Residuen festzustellen. Die Diederwinkel werden stets durch vier Atome definiert. Der φ-Winkel beschreibt die Bindung „C-N-C_α-C“ und wird zwischen dem C_α- und dem N-Atom gemessen, wohingegen der ψ-Winkel in der Abfolge der Atome „N-C_α-C-N“ zwischen die Verknüpfung des C_α- und des C-Atoms assoziiert wird. In planarer Abstraktion der verknüpften Aminosäuren, welche stets existent ist, da die Peptidbindung eine planare Disposition aufweist, betragen beide Winkel 180° und eine charakteristische Drehung im Uhrzeigersinn. Dies ist in *Abbildung 12* dargestellt, wobei das zentrale C_α-Atom mit „α-Carbon“ gekennzeichnet ist. Der komplementäre -180°-Winkel kommt durch eine Rotation in die linke Richtung zu Stande. Der φ- sowie ψ-Winkel kann in beiden Konformationen auftreten, jedoch gibt es Ausbildungen, welche aus sterischen Gründen nicht existent sind, da es zu Überlappungen einzelner Atome kommen würde.

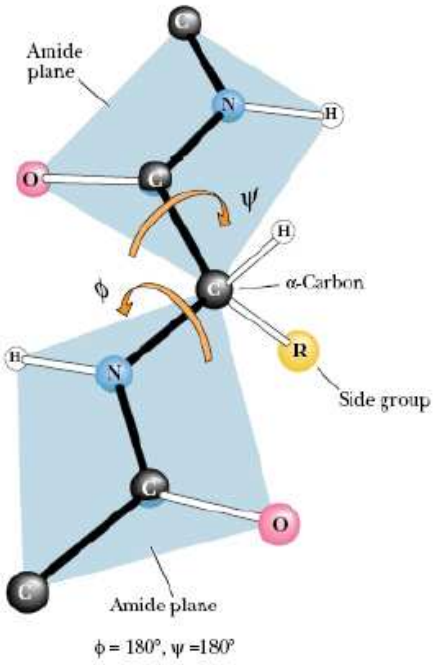

Abbildung 12: *Die Diederwinkel φ und ψ betragen in planarer Konformation der verknüpften Aminosäuren 180° und weisen eine Drehung nach rechts auf. Das zentrale C_α-Atom ist in der Abbildung als „α-Carbon" gekennzeichnet.* [26]

4.6 Disulfidbrücken

Disulfidbrücken werden in Proteinen zwischen räumlich benachbarten Cysteinen ausgebildet. Sie stabilisieren die Tertiärstruktur und sind somit an der korrekten Ausbildung der nativen Struktur und der katalytischen Wirkung eines Proteins beteiligt. Eine Analyse der Lage dieser Bindungen und eine energetische Einordnung drängen sich somit auf und wurden in Bezug auf den kleinen Datensatz durchgeführt. Die Häufigkeiten der Ausbildungen in Bezug auf die Sekundärstrukturelemente sind informativ und auch die energetischen Unterschiede müssen wieder Beachtung finden und sollen Aufschluss über die differenzierte Ausbildung geben.

5 Ergebnisse der Analysen

5.1 Chemische Eigenschaften von Aminosäuren

Für die Analyse der chemischen Eigenschaften wurde der kleine Datensatz mit 114 Proteinen verwendet. Die Präferenzen wurden mit Hilfe des *Chou-Fasman*-Algorithmus bestimmt. Dieser wurde in den 1970er Jahren mit dem eigentlichen Hintergrund der Vorhersage von Sekundärstrukturen in Proteinen entwickelt. Die *Formel* (9), auf welcher die Berechnung aller Präferenzen beruht, ist nachfolgend dargestellt.

$$P_{rs} = \frac{\frac{n_{rs}}{n_s}}{\frac{n_r}{N}} \tag{9}$$

P_{rs}	≙	Präferenz der Aminosäure r in Abhängigkeit des Zustandes s
n_{rs}	≙	Anzahl der Aminosäuren r im Zustand s
n_s	≙	Anzahl aller Aminosäuren im Zustand s
n_r	≙	Anzahl der Aminosäuren r im Datensatz
N	≙	Anzahl aller Aminosäuren im Datensatz

Anhand der Formel wird der Informationsgehalt dieses Algorithmus ersichtlich. Im klassischen Sinne wird für jede Aminosäure in den definierten Sekundärstrukturbereichen die Neigung des Ausbildens dieser Konformation berechnet. Im Zusammenhang mit den betrachteten Eigenschaften der Aminosäuren (Parameter) stellen diese den Zustand s dar und definieren die quantitativen Beschreibungen auf dieser Abstraktionsebene. Für alle 20 kanonischen Aminosäuren r wurden so in den vier Quantilen die Präferenzen für die vier gewählten Eigenschaften der Aminosäuren berechnet. Um eine Vorstellung für das

Auftreten der einzelnen Residuen in den Energiebereichen zu verdeutlichen, sind in *Tabelle 6* alle Werte vermerkt. Der Kopf beschreibt chronologisch die Quantile eins bis vier. Die Anzahl an Ausprägungen reicht von 330 für Tryptophan bis zu 1717 für Alanin. Dafür ist oftmals die unterschiedliche Struktur der Reste verantwortlich und damit die divergente Größe der Aminosäuren. Alanin und Glycin weisen bspw. eine sehr kompakte Seitengruppe auf, weshalb sie auf Grund ihrer Größe an mehr Stellen der Sequenz stehen können, da sie nicht so viel Platz wie ein kompliziert gebautes Residuum wie Histidin einnehmen.

Tabelle 6: *Das Auftreten der Aminosäuren im Datensatz und ihre Verteilung in den vier energetischen Quantilen. Dabei beschreibt 1. Q bis 4. Q die Quantile. Die Gesamtanzahl für jede Aminosäure ist am rechten Rand aufgeführt Die Anzahl an Ausprägungen reicht von 330 für Tryptophan bis 1717 für Alanin.*

Aminosäure	1. Q	2. Q	3. Q	4. Q	
Ala	23	327	1018	349	**1717**
Arg	253	651	203	1	**1108**
Asn	455	452	65	1	**973**
Asp	851	225	6	2	**1084**
Cys	0	3	78	337	**418**
Gln	220	475	102	2	**799**
Glu	1026	139	3	3	**1171**
Gly	397	760	447	12	**1616**
His	5	82	321	18	**426**
Ile	3	12	172	864	**1051**
Leu	7	21	407	1143	**1578**
Lys	942	91	3	3	**1039**
Met	2	8	89	271	**370**
Phe	1	11	147	528	**687**
Pro	393	442	117	0	**952**
Ser	241	692	349	2	**1284**
Thr	122	558	521	5	**1206**
Trp	0	8	148	174	**330**
Tyr	0	29	339	367	**735**
Val	5	21	430	890	**1346**

Da der Algorithmus normiert ist können alle Werte direkt miteinander verglichen werden. Es ist legitim, Rückschlüsse auf das Ausbilden aller Residuen zu ziehen und wie sie sich im Raum der chemischen Eigenschaften verhalten. Im mathematischen Kontext können die Werte dann absolut korrekt beweisen, wie aussagekräftig sie sind und ob durch die Präferenzen eine eindeutige Beschreibung jedes Residuums möglich ist. Die in diesem Abschnitt nachfolgenden Tabellen weisen farbliche Unterscheidungen bei den Residuen auf. Jene mit vergleichbaren Eigenschaften sind dabei in der gleichen Farbe gekennzeichnet.

Ladung

Fünf Aminosäuren weisen eine Ladung in ihrer Seitenkette auf und sind demzufolge gesondert von den verbleibenden zu berechnen. Sie sind in der *Tabelle 7* rot markiert. Dabei wurde bewusst nicht noch einmal zwischen positiver oder negativer Ladung unterschieden, weil allein der Unterschied zu ungeladenen Residuen beschrieben werden sollte.

Tabelle 7: *Präferenzen für die Ladung der Aminosäuren. Fünf Aminosäuren weisen eine Ladung auf. Diese kann positiv bzw. negativ sein. Darauf sollte in der Analyse jedoch nicht geachtet werden. Die Präferenzen weisen bei Vorhandensein der betrachteten Aminosäure Werte zwischen 0,006 und 31,127 auf und sind auf Grund der verwendeten Formel (9) und der daraus resultierenden Residuen-Anzahl nur schlecht vergleichbar.*

Aminosäure	1. Q	2. Q	3. Q	4. Q	1. Q	2. Q	3. Q	4. Q	Aminosäure
Arg	1,476	9,837	6,799	0,665	0,03	0,06	0,735	3,307	**Ile**
Asp	5,075	3,475	0,205	1,359	0,047	0,069	1,158	2,914	**Leu**
Glu	5,664	1,987	0,095	1,887	0,058	0,113	1,08	2,946	**Met**
His	0,076	3,223	28,875	31,127	0,016	0,083	0,961	3,091	**Phe**
Lys	5,861	1,466	0,107	2,127	4,393	2,418	0,552	0	**Pro**
Ala	0,143	0,992	2,663	0,818	1,997	2,807	1,221	0,006	**Ser**
Asn	4,977	2,419	0,3	0,004	1,077	2,41	1,94	0,017	**Thr**
Cys	0	0,037	0,838	3,243	0	0,126	2,014	2,121	**Trp**
Gln	3,93	3,096	0,573	0,01	0	0,206	2,071	2,008	**Tyr**
Gly	2,614	2,449	1,242	0,03	0,04	0,081	1,435	2,66	**Val**

Histidin zeigt aus der Gruppe der geladenen Aminosäuren besonders im dritten und vierten Quantil markante Präferenzen auf. Sie sind auf das sehr hohe Auftreten in diesen Bereichen zurück zu führen. Die verbleibenden Residuen dieser Klassifizierung weisen teilweise annähernd vergleichbare Werte auf, was auf eine näherungsweise homogene Verteilung hinweist. Die extrem hohen Beträge bei Histidin treten auch in nachfolgenden Betrachtungen bei anderen Aminosäuren auf. Dies ist auf die geminderte Anzahl an Residuen in den charakteristischen Gruppen zurück zu führen, denn gemäß dem Algorithmus wird die Präferenz für eine Aminosäure in einer Gruppe mit gleicher Eigenschaft umso größer, je höher die Anzahl n_{rs} ist. Somit steigt auch der Quotient mit n_s.

Polarität

Da polare und unpolare Aminosäuren beinahe in gleicher Verteilung vorliegen ist der Gradient der Präferenzen demzufolge nicht so stark ausgeprägt, wie bei der vorangegangenen Eigenschaft. Zusätzlich muss dabei angemerkt werden, dass die polaren Residuen viel seltener in den energetisch stabilen Quantilen vorkommen als die unpolaren. Die einzigen Ausnahmen dabei liefern, wie in *Tabelle 8* ersichtlich, Cystein und Tyrosin. Im Vergleich dazu fallen die Präferenzen der äquivalenten polaren Aminosäuren sehr gering aus. Die unpolaren Aminosäuren verhalten sich ebenfalls nahezu gleichförmig. Nur Prolin weist im ersten Quantil einen erhöhten und von dem normalen Bereich abweichenden Wert auf. Im vierten Quantil verhalten sie sich bis auf Alanin und Glycin sehr ähnlich. Prolin weist in diesem Energiebereich keine Ausprägungen auf, weshalb der Zahlenwert „0“ beträgt. Dies ist natürlich auch bei allen anderen aufgezeigten Präferenzwerten in diesen Statistiken so, welche mit „0“ bewertet sind.

Tabelle 8: *Präferenzen für die Polarität der Aminosäuren. Auf Grund der ausgeglichenen Aminosäuren-Anzahl sind die ermittelten Parameter sehr gleichmäßig. Sie wurden mit Hilfe der Formel (9) ermittelt.*

Aminosäure	1. Q	2. Q	3. Q	4. Q	1. Q	2. Q	3. Q	4. Q	Aminosäure
Arg	1,104	3,44	1,831	0,024	0,068	0,141	1,094	3,811	Ile
Asp	3,795	1,215	0,055	0,05	0,106	0,164	1,724	4,239	Leu
Glu	4,235	0,695	0,026	0,069	0,129	0,267	1,608	3,395	Met
His	0,057	1,127	7,531	1,134	0,035	0,198	1,431	3,563	Phe
Lys	4,382	0,513	0,029	0,233	9,881	2,003	0,822	0	Pro
Ala	0,321	2,353	3,964	0,956	0,907	3,156	2,717	0,042	Ser
Asn	2,26	2,72	0,029	0,028	0,489	2,709	4,318	0,111	Thr
Cys	0	0,042	1,865	21,641	0	0,3	2,998	2,444	Trp
Gln	1,331	3,481	1,276	0,067	0	0,231	4,61	13,403	Tyr
Gly	5,88	5,81	1,849	0,035	0,089	1,94	2,136	3,065	Val

Struktur

Die Aminosäuren mit einer aromatischen Seitenkette heben sich auf Grund ihrer quantitativen Anzahl natürlich stark in ihren Werten von den aliphatischen ab. Sie kommen im Mittel viel seltener vor. Dies kann man an den Präferenzen in *Tabelle 9* deutlich erkennen. Die Werte bei allen anderen Residuen sind sehr ausgeglichen und weisen erwartungsgemäß nur kleine Unterschiede auf. Prolin stellt mit seinem heterocyclischen Rest und der sekundären Aminogruppe eine individuale Ausnahme dar. Keine andere Aminosäure weist eine solche einzigartige Struktur auf. Die Anzahl an Prolinen in dem Set liegt beinahe am mittleren Durchschnitt aller Residuen. Die große Summe aller Aminosäuren lässt jedoch diese Anzahl verhältnismäßig klein werden, woraus demzufolge auch die kleinen Werte für die Präferenzen resultieren. Umso wichtiger ist Prolin in Bezug auf die Sekundärstrukturausbildung, denn es gilt als sog. Helixbrecher und somit können bei dem Auftreten in Verbindung mit statistischen oder berechneten Werten explizite Aussagen über die Sekundärstruktur an dieser Stelle gemacht werden. [12]

Tabelle 9: *Präferenzen für die Struktur der Aminosäuren. Sie wurden unter Zuhilfenahme der Formel (9) berechnet und sind auf Grund der Charakteristik dieser sehr dispers mit Werten zwischen 0,005 bis 38,909. Prolin nimmt auf Grund seiner einmaligen heterocyclisch en Struktur unter den proteinogenen Aminosäuren bei dieser Klassifizierung eine Sonderstellung ein, weshalb die Werte auch nur sehr gering ausfallen.*

Aminosäure	1. Q	2. Q	3. Q	4. Q	1. Q	2. Q	3. Q	4. Q	Aminosäure
Arg	0,919	2,396	0,909	0,005	0,012	0,047	0,812	4,209	**Ile**
Asp	3,161	0,847	0,028	0,009	0,018	0,054	1,279	3,708	**Leu**
Glu	3,528	0,484	0,013	0,013	0,022	0,088	1,193	3,75	**Met**
His	38,909	29,451	15,694	0,773	4,825	2,45	4,457	14,063	**Phe**
Lys	3,65	0,357	4,497	0,015	0,02	0,022	0,006	0	**Pro**
Ala	0,054	0,777	2,941	1,041	0,756	2,198	1,348	0,008	**Ser**
Asn	1,883	1,895	0,331	0,005	0,407	1,887	2,143	0,021	**Thr**
Cys	0	0,029	0,926	4,128	0	3,709	9,341	9,648	**Trp**
Gln	1,109	2,425	0,633	0,013	0	6,037	9,606	9,137	**Tyr**
Gly	0,989	1,918	1,372	0,038	0,015	0,064	1,585	3,385	**Val**

Hydrophobizität

Auch bei dieser Eigenschaft sind die Unterschiede in den Präferenzen nur gering, was man in *Tabelle 10* sofort erkennen kann. Sogar gruppenübergreifend schwanken die Werte fast ununterbrochen in einem normalen Maße. Einzig Lysin als Vertreter der hydrophoben Aminosäuren weist einen stark erhöhten Wert auf. Ihr Auftreten im ersten Quantil ist jedoch auch überproportional hoch im Vergleich zu den verbleibenden Residuen dieser Gruppe und auch im Vergleich zu ihrer Ausbildung in den anderen drei Quantilen. Nur Alanin, Glutaminsäure und Leucin weisen in jeweils einem Bereich noch einen etwas höheren Wert auf. Im Zusammenhang mit dem Ausbilden nur geringer Unterschiede in den Zahlenwerten könnte die Hydrophobizität wohlmöglich nur einen geringen Einfluss als einer der chemischen Deskriptoren haben. Die mathematische Analyse wird dahingehend jedoch mehr Aussagekraft besitzen als profan formulierte Vermutungen.

Tabelle 10: *Präferenzen für die Hydrophobizität der Aminosäuren. Die homogene Verteilung bewirkt ein sehr enges Spektrum an Werten. Einzig Lysin weist im 1. Quantil einen erhöhten Wert auf. Die Präferenzen wurden unter Zuhilfenahme der Formel (9) berechnet.*

Aminosäure	1. Q	2. Q	3. Q	4. Q	1. Q	2. Q	3. Q	4. Q	Aminosäure
Arg	1,321	3,8	4,313	1,632	0,038	0,118	0,79	3,296	Ile
Asp	4,541	1,342	0,13	3,336	0,059	0,137	1,245	2,904	Leu
Glu	5,068	0,768	0,06	4,632	0,071	0,223	1,161	2,936	Met
His	0,155	1,983	3,638	0,169	0,019	0,165	1,033	3,081	Phe
Lys	11,966	0,902	0,014	0,012	2,388	3,002	2,893	0	Pro
Ala	0,177	1,962	2,862	0,815	1,086	3,485	6,398	2,817	Ser
Asn	2,705	3,004	1,573	1,858	1,335	4,766	2,086	0,017	Thr
Cys	0	0,074	0,901	3,232	0	0,25	2,165	2,114	Trp
Gln	1,593	3,844	3,005	4,526	0	0,406	2,227	2,002	Tyr
Gly	3,242	4,844	1,335	0,03	0,049	0,161	1,542	2,651	Val

5.2 3D-Motive und ausgewählte Sequenzmotive

Zu allen 114 Proteinen des Datensatzes wurden die 3D- sowie spezielle Sequenzmotive analysiert. Diese kann man sich über das Tool *PDBe-Motif* der *PDBe* (*Protein Data Bank Europe*) ermitteln lassen. Der Quellcode dieser Seiten konnte somit ausgelesen werden und die Werte wurden mittels Excel verglichen und aufgezeichnet, um Zusammenhänge, energetische Unterscheidungen und quantitatives Auftreten der Residuen zu erfassen.

Die Fülle an Ergebnissen und statistischen Erhebungen in Form von Tabellen soll nicht der wesentliche Inhalt dieses Arbeitspunktes sein, weshalb sie aufgegliedert im Anhang unter **A1-A3** dieser Arbeit aufgeführt sind.

Die Analysen beschränken sich auf Grund der hohen Anzahl an Sequenzmotiven auf fünf Vertreter dieser Gruppe. Betrachtet wurden jene, welche am häufigsten in den Proteinen vorkommen und auch in verschiedenen Klassen der Proteine auftreten. Die 13 existierenden 3D-Motive wurden komplett analysiert. Da diese teilweise mit einer variierenden Residuen-Anzahl existent sind, wurde in passenden Analysen diese zusätzliche Unterscheidung betrachtet. Ein Motiv muss mindestens aus zwei Aminosäuren bestehen. Dabei reicht die

Anzahl in den analysierten Sequenz- sowie 3D-Motiven bis auf einen Wert von sechs Residuen.

5.2.1 Die 3D-Motive

Das quantitative Auftreten dieser Motive über den gesamten Datensatz ist stark variierend. Die geringste Anzahl an Ausprägungen zeigt der *gammaturn* mit 23 Entsprechungen. Vergleichend dazu stellt die *niche* mit 995 Analogien das höchste Ausmaß dar. Das mittlere Auftreten der 13 Motive im Set liegt bei 242. Diesen Wert erreicht jedoch kein einziges Motiv auch nur näherungsweise. Entweder die Anzahl an Ausprägungen liegt weit unter diesem Wert bis maximal 146 oder entsprechend hoch darüber, beginnend bei 547. *Abbildung 13* verdeutlicht die Verteilung der Motive.

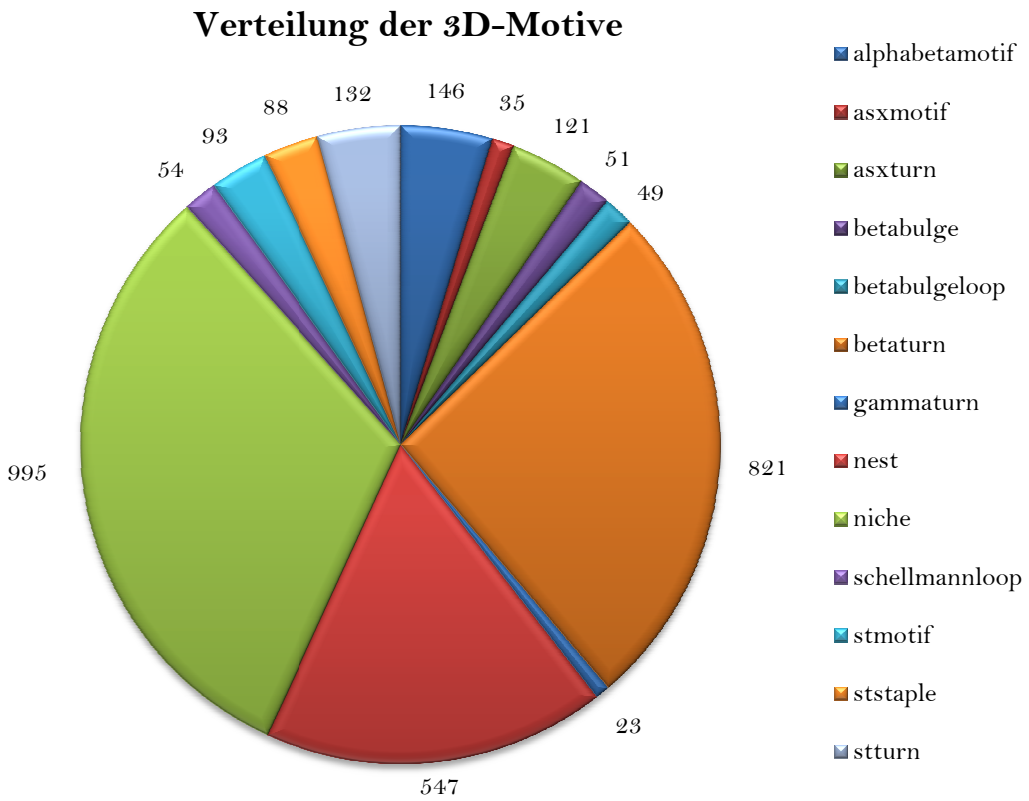

Abbildung 13: *Die Verteilung des Auftretens der 3D-Motive im Set verdeutlicht, dass besonders drei Motive sehr häufig konsistent sind. Es sind die niche, der betaturn und das nest. Sie sind demzufolge auch in vielen verschiedenen Gruppen von Proteinen enthalten.*

Auch die Lage in den verschiedenen Sekundärstrukturbereichen unterscheidet sich graduell. Dabei treten Divergenzen auf, sodass sich einige Motive bevorzugt in einer spezifischen Sekundärstruktur ausbilden. Der *betabulge* tritt sogar nur in Sheets auf. Trotz dieses Ergebnisses soll nicht zwangsläufig behauptet werden, dass er immer dieses Verhalten zeigt, denn auf Grund der im Datensatz befindlichen Anzahl von nur 51 Ausprägungen ist es nicht auszuschließen, dass er bei intensiverer Analyse noch in anderen Sekundärstrukturbereichen vorkommt. In *Tabelle 11* wird die Verteilung vereinfacht wiedergegeben. Dabei werden keine Häufigkeiten dargelegt, sondern das Auftreten durch ein quantitatives und simplifiziertes Bewertungssystem mit den Wertigkeiten von 1-5 dargestellt. Dabei soll sich an folgender Legende orientiert werden:

1 ≙ sehr häufiges Auftreten

2 ≙ häufiges Auftreten

3 ≙ mäßiges Auftreten

4 ≙ seltenes Auftreten

5 ≙ kein Auftreten

Tabelle 11: Die *vereinfachte Bewertung der 3D-Motive für das Vorhandensein in den Sekundärstrukturen wird mit den Wertigkeiten von 1 für sehr häufiges Auftreten bis 5 für kein Auftreten klassifiziert.*

Motiv	Sekundärstruktur		
	Coil	Helix	Sheet
betabulge	5	5	1
stturn	1	3	4
stmotif	4	1	4
betaturn	2	1	4
nest	1	3	4
asxmotif	3	2	4
asxturn	2	2	4
schellmannloop	3	1	4
betabulgeloop	1	4	3
gammaturn	1	4	4
niche	2	2	4
ststaple	4	1	4
alphabetamotif	3	2	4

Die beschriebene Tabelle verdeutlicht, dass sich die Motive bevorzugt in Helices und Coil Strukturen ausbilden, wohingegen Faltblattstrukturen nur selten zur Ausbildung der 3D-Motive beitragen. Dabei muss ergänzend erwähnt werden, dass sich die Residuen in einem betrachteten Motiv auch in verschiedenen, also überlappenden, Sekundärstrukturen ausbilden können. Einzige Ausnahme bildet der schon erwähnte *betabulge*.

Energetische Charakterisierung

Um eine erste Vorstellung für die energetische Vielfalt aufzuzeigen, wird in *Abbildung 14* die durchschnittliche Energie jedes Motivs dargestellt. Dabei wird ersichtlich, dass sich die Motive in einem mittleren Energiebereich von etwa -5 bis -14 bewegen und somit eine recht starke Streuung aufweisen. Den niedrigsten Wert beschreibt das *alphabetamotif* und den höchsten der *asxturn*. Im Anhang unter Punkt **A1** ist zusätzlich eine Tabelle mit den genauen Werten dargestellt.

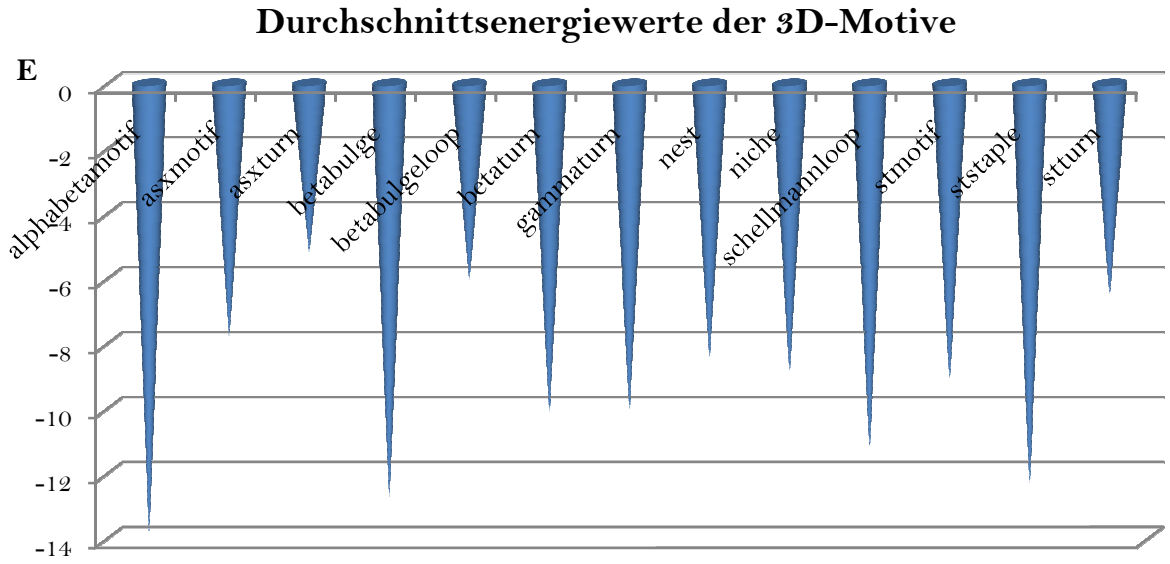

Abbildung 14: *Die durchschnittliche freie Energie aller 3D-Motive vermittelt die Varianz in der Stabilität. Das alphabetamotif ist dabei am stabilsten und der asxturn am instabilsten.*

Für die genaue energetische Charakterisierung wurde jedes Motiv in seinen verschiedenen Ausprägungen beobachtet, um explizite und korrekte Aussagen machen zu können. Wenn alle Motive, auch jene mit einer variierenden Anzahl an Residuen, in einer Statistik erfasst werden würden, sind falsche Aussagen und Schlussfolgerungen sehr wahrscheinlich. Die strukturellen Unterschiede, welche durch zusätzliche Aminosäuren selbstverständlich zu Stande kommen, beeinflussen das energetische Auftreten. Drei Motive können in unterschiedlichen Ausprägungen existent kommen. Dies sind die *niche* mit drei oder vier Aminosäuren sowie der *betabulgeloop* und der *ststaple* in Konformation mit fünf oder sechs Residuen.

Wenn man den mathematischen mittleren Energieverlauf der Residuen in den Motiven betrachtet, stellt man fest, dass man diese in drei differenzierte Gruppen einteilen kann. Bei dieser Charakterisierung kommt es nicht auf die Länge eines Motivs an, sondern rein auf die energetische Abfolge der Residuen. Die meisten Motive folgen der Anordnung, dass die beiden flankierenden Aminosäuren die energetischen Minima darstellen (1). Weiterhin gibt es auch Motive einer Konformation, bei welchen die inneren Residuen energetisch unter den flankierenden liegen oder mindestens ein Teil der sich mittig befindenden Aminosäuren energetisch stabiler ist (2). Die dritte Gruppierung bildet mit der Abfolge der Aminosäuren einen internen energetischen Gradienten aus, wobei jedoch auch eine flankierende Aminosäure das Minimum darstellt (3).

Die Motive in ihrer teilweise variablen Residuen-Anzahl sind nachfolgend aufgezeigt. Die farbig markierten Balken verdeutlichen die durchschnittliche Energie der positionsspezifischen Aminosäuren in den Motiven. Der Name des jeweils energetisch abgebildeten Motivs ist als Diagrammtitel beigefügt. Dabei ist die Anzahl an Residuen in den drei genannten variierenden Motiven in Klammern „()“ angegeben. Die vertikale Achse repräsentiert die Disposition der freien Energie.

Die erste Gruppe (1) umfasst zehn Ausprägungen und ist kumuliert als *Abbildung 15* visualisiert. Die Charakteristik der flankierenden Energieminima ist dabei deutlich ersichtlich. Die Minima lassen sich bei genauer Analyse nicht auf variierende Sekundärstrukturelemente zurück führen. Bei Betrachtung der Tertiärstruktur wird folglich auch ersichtlich, dass an den spezifischen Positionen stets eine geschwungene Form des *backbones* feststellbar ist.

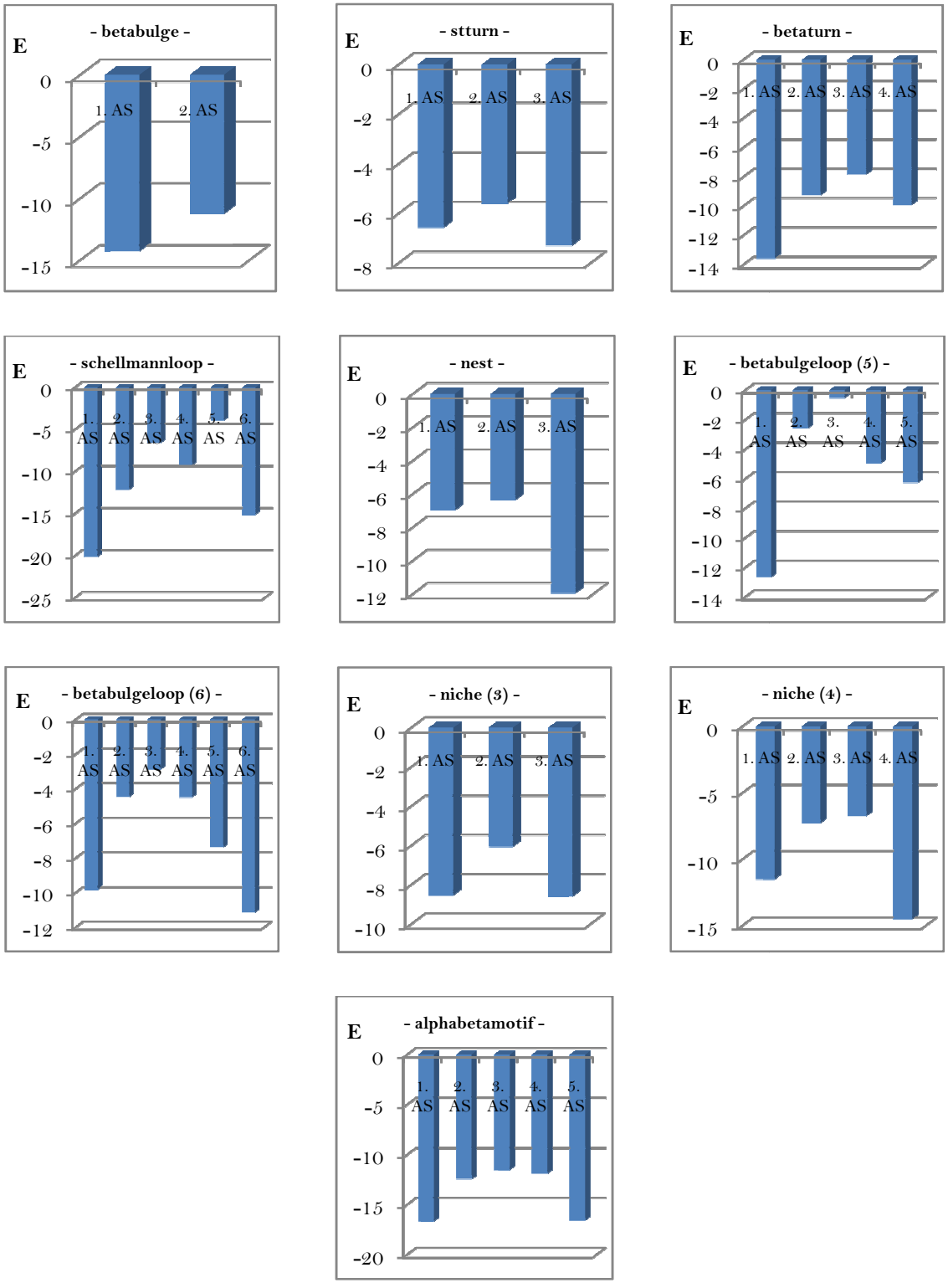

Abbildung 15: *3D-Motive mit flankieren Energieminima bilden die größte Gruppe bei der Charakterisierung durch die energetischen Mittelwerte an jeder Position. Es wurde keine Abhängigkeit zu den spezifischen Sekundärstrukturelementen festgestellt.*

Die Gruppe (2) der Motive mit zentral liegenden Minima ist zusammengefasst in *Abbildung 16* dargestellt. Eine energetische Abhängigkeit dieser Motive ist auch, wie bei den vorhergehenden, nicht zu erkennen. Sie verhalten sich gruppenintern sehr vielseitig und decken ein hohes Energiespektrum ab.

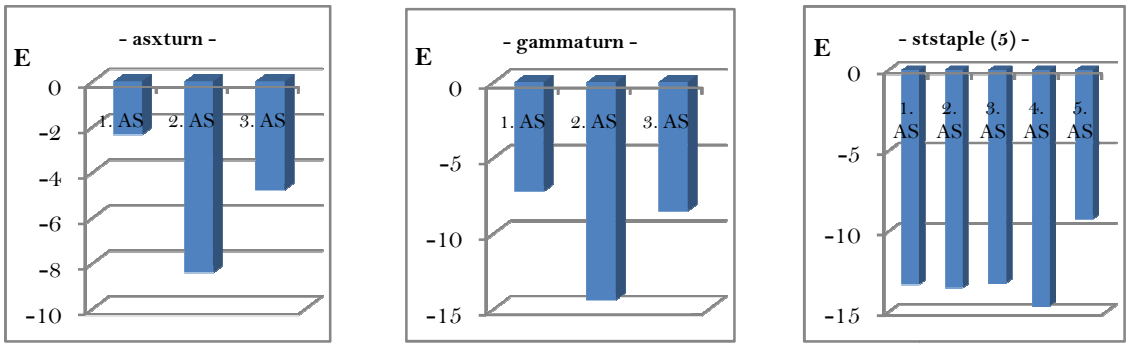

Abbildung 16: *3D-Motive mit internen Energieminima sind verhältnismäßig selten und decken in ihrer energetischen Ausprägung ein hohes Energiespektrum ab.*

Motive der Gruppe (3) und dem identifizierten internen Energiegradienten sind in *Abbildung 17* veranschaulicht. Markant sind der energetische Verlauf und das Minima an erster oder letzter Position in den Motiven. Residuen an dieser Stelle sind außerdem oft inmitten einer Sekundärstruktur liegend und demzufolge nicht an Übergängen aufzufinden.

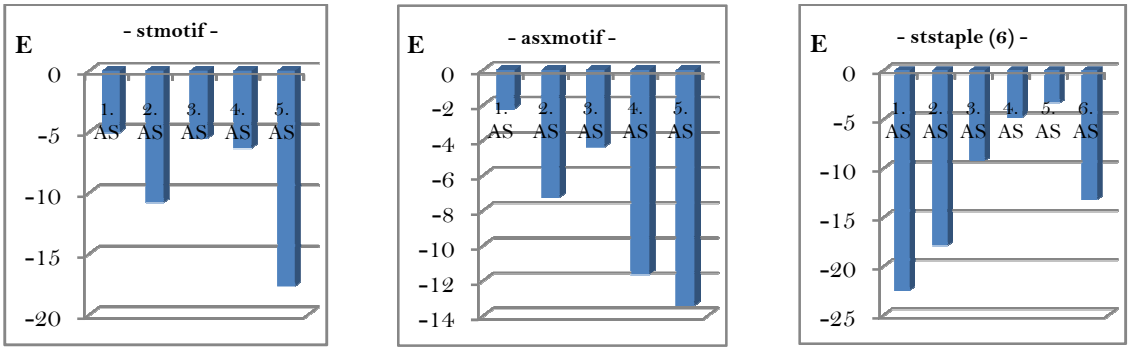

Abbildung 17: *3D-Motive mit einem sequenziellen Energiegradienten treten nur selten auf und sind oft durch das interne Energieminimum an erster oder letzter Position gekennzeichnet.*

Der energetische Verlauf der Motive durch die dargestellten Diagramme ist zweifelsfrei sehr allgemein und zur genauen Beschreibung nicht präzise genug. Eine detaillierte Analyse der energetischen Verteilungen an jeder Position der Motive musste dazu durchgeführt werden. Dafür wurden die energetischen Quantile genutzt und die Anzahlen an Residuen in den Bereichen erfasst. In *Abbildung 18* sollen die Verteilungen einmal beispielhaft an dem *asxturn* verdeutlicht werden. Die Aminosäuren an den Positionen 1 bis 3 sind chronologisch mit a) bis c) bezeichnet. Alle in dem Motiv vorkommenden Residuen wurden energetisch charakterisiert und ihr Wert zugehörig in eins der vier Quantile eingeordnet. Die Anzahlen sind jeweils in den abgebildeten Balken erkenntlich gemacht. Die erste Position in dem Motiv ist von Aminosäuren mit einem hohen Energiewert gekennzeichnet. Der Großteil liegt im 1. und 2. Quantil. Der stabile Energiebereich ist fast gar nicht vertreten. An der zweiten Position verschiebt sich das Verhältnis etwas und es kommt auch eine ausgeglichene Anzahl an Residuen im 3. und 4. Quantil und somit im stabilen Bereich vor. Die letzte Position weist, ähnlich wie Stelle eins, eine verstärkte Anzahl an hochenergetischen Aminosäuren auf. Der energetisch stabilisierende Schwerpunkt muss auf Grund dieser Statistiken vermehrt bei Position zwei liegen.

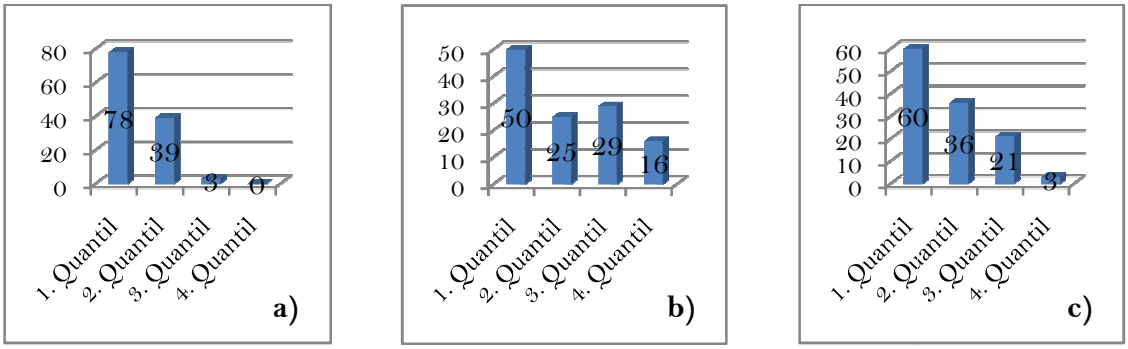

Abbildung 18: *Die energetische Beschreibung des asxturn mit Hilfe der Quantile verdeutlicht eine detaillierte Aufgliederung in den differenzierten Energiebereichen. Teilabbildung a) steht dabei für die erste Aminosäure. Mit einer hohen Auftrittswahrscheinlichkeit weist das an dieser Position vorkommende Residuum einen hohen Energiewert auf und hat somit meist instabilen Charakter. Die energetische Einschätzung der zweiten Aminosäure ist in b) abgebildet. Auch hier weist das Residuum meist einen hohen Energiewert auf, aber auch niederenergetische können vertreten sein. Das letzte Residuum, welches in c) dargestellt ist, stammt auch meist aus dem 1. energetischen Quantil.*

Insgesamt bleibt bei dieser Analysemethode anzumerken, dass sich Unterschiede gut erkennen lassen und somit wahrscheinlich auch eine Charakterisierung eines Motivs an Hand einer Sequenz vereinfacht wird. Die kompletten Werte der Analysen sind im Anhang unter Punkt **A2** in Tabellenform vermerkt. Dafür wurden die relativen Häufigkeiten an jeglichen Positionen bestimmt.

Sequenzielle Charakterisierung

Das Aufkommen spezifischer Residuen an den verschiedenen Positionen in einem Motiv kann Aufschluss über die Lage und die Identifikation geben. Zu allen Motiven wurden demzufolge separate Statistiken erstellt, welche die Häufigkeiten aller Aminosäuren reflektieren. Zur besseren Veranschaulichung kann man sich mit dem Online-Tool „Weblogo“ der *University of California* [24] Sequenzlogos anfertigen, um die Verteilungen der Residuen bildlich zu visualisieren. Im Anhang unter Punkt **A3** werden für alle Motive und ihre differenzierten Ausprägungen die Statistiken in Tabellenform dargestellt. Dabei sind die relativen Häufigkeiten aufgeführt. Sie geben Aufschluss, in welcher Häufigkeit die Aminosäuren an den verschiedenen Positionen der Motive vorkommen.

Wenn man sich über ein 3D-Motiv auf der zugehörigen Seite des *PDBe*-Servers erkundigt, wird man feststellen, dass es zu nahezu jedem Motiv noch mehr Ausprägungen gibt, als die in dieser Arbeit aufgeführten. Sie kommen durch wechselnde Winkel in Folge der Atomstellung zu Stande. Die Unterscheidung und Feststellung, um welche Konformation es sich bei jedem Motiv handelt, ist leider nicht einsehbar und konnte deshalb in die Analysen nicht mit einbezogen werden. Einige Motive weisen feste Residuen an ihrer ***i***-ten Position auf. Dies sind der *asxturn* und das *asxmotif*, bei welchen Asparagin oder Asparaginsäure an dieser Stelle existent sein müssen, sowie der *stturn*, *ststaple* und das *stmotif* mit einem Serin oder Threonin. Demzufolge sind Analysen an diesen Positionen nur von quantitativer Bedeutung. Bei der Betrachtung der weiteren Stellen fällt jedoch besonders eine Aminosäuren auf, welche nur stark vermindert oder gar nicht an bestimmten Positionen in einigen Motiven auftaucht. Dies ist Prolin. Um qualitative Aussagen zu machen, kommen bei dieser Untersuchung fast nur die Motive *betaturn*, *niche* und *nest* in Frage, denn nur diese liegen in einer zahlenmäßig ausreichend großen Anzahl vor, um Fehlbewertungen

bestmöglich auszuschließen. Bei allen ist die Charakteristik des Fehlens von Prolin festzustellen. In dem *betaturn* und der *niche* fehlt es stets an der vierten Position des Motivs, in einem *nest* kommt es sogar an keiner der beiden Stellen des Motivs vor. Dies deutet darauf hin, dass Prolin an diesen Motiv-spezifischen Positionen die strukturelle Stabilität zu sehr limitieren könnte. Auch in dem *asx-turn* kommt es an der dritten Position nicht vor. Markant ist jedoch, dass Prolin an der zweiten Stelle mit Abstand am häufigsten auftritt. Dies unterstreicht die These der Stabilitätsprobleme an den beschriebenen systematischen Positionen.

Auch ein besonders häufiges Auftreten von einem Residuum ist zu beobachten. Es ist die Aminosäure Glycin. In dem *betabulgeloop* weist sie an der dritten Position eine absolute Häufigkeit von 33 Ausprägungen, bei einer Motivanzahl von 49, auf. Damit steht sie bei etwa 67% aller *betabulgeloop*'s an dieser Position. In dem *schellmannloop* kommt Glycin zu über 50% an der fünften Position vor und auch in dem *nest* ist es an nahezu 50% der Ausprägungen an der zweiten Position beteiligt. Man muss das hohe Vorkommen von Glycin jedoch auch in einem gewissen Maße normalisieren, da es, wie in bereits vermerkter *Tabelle 4*, fast am häufigsten im Datensatz vorhanden ist und somit ein Auftreten etwas wahrscheinlicher ist. Jedoch treten die übrigen, sehr häufig vorkommenden Aminosäuren, wie bspw. Alanin und Leucin, nicht in dieser hohen Anzahl auf, weshalb Glycin einen großen Stellenwert bei der Charakterisierung der benannten Motive aufweist.

Abbildung 19 verdeutlicht an dem Beispiel des *alphabetamotif* ein Weblogo. Bei der Betrachtung fällt ad hoc auf, dass es nicht für qualitative Analysen geeignet ist und nur der anschaulichen Visualisierung dient. Nur wenn Residuen konserviert auftreten kann man eindeutige Unterschiede erkennen. Bei dem *alphabetamotif* treten, wie bereits an früherer Stelle der Arbeit erläutert, eine große Anzahl konträrer Aminosäuren in sehr homogener Verteilung auf, weshalb das Logo sehr dicht erscheint und nur die ersten drei bis vier am häufigsten auftretenden Residuen pro Stelle zweifelsfrei identifiziert werden können.

Die Farbgebung steht in der Standardeinstellung auf „Default" und verwendet dabei ein, bei Weblogo, voreingestelltes Schema. Es ist auch möglich die Aminosäuren nach einer spezifischen Eigenschaft farblich zu differenzieren. Dies wurde bei dem abgebildeten Weblogo umgesetzt. Dabei sind alle schwarz eingefärbten Aminosäuren unpolarer Natur und die rot markierten mit polarem Charakter.

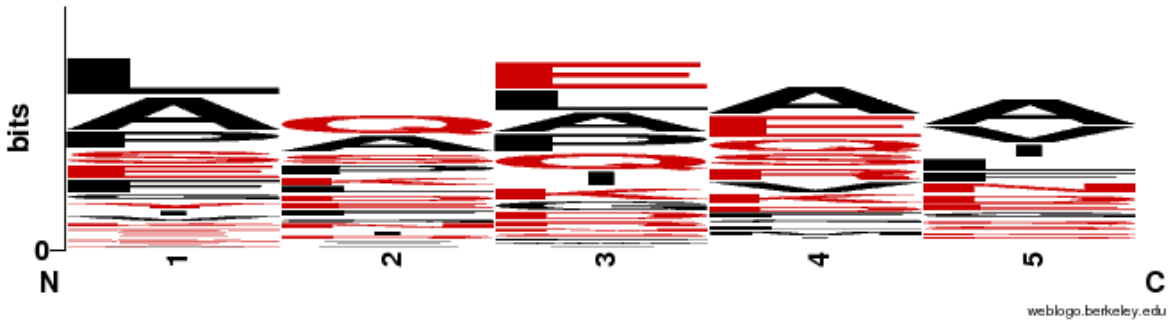

Abbildung 19: *Das Weblogo des alphabetamotif verdeutlicht eine große Ausgeglichenheit in dem Auftreten der Aminosäuren. Keine Aminosäure liegt an einer Position konserviert oder stark bevorzugt vor. Rot markierte Residuen sind polar, schwarz eingefärbte unpolarer Natur. Die Termini sind mit „N" sowie „C" bezeichnet* [24]

5.2.2 Die Sequenzmotive

Die fünf analysierten Sequenzmotive sind in der *PROSITE*-Datenbank unter den Namen

- PS00001
- PS00004
- PS00005
- PS00006
- PS00008

vermerkt. Die Analyse dieser Motive ergibt sich rein aus ihrem verhältnismäßig häufigen Auftreten im Vergleich zu anderen, denn um etwas über die Motive in Erfahrung zu bringen ist natürlich eine ausreichend große Anzahl erforderlich. Die Motive wurden deshalb, genau wie die 3D-Motive, energetisch sowie sequenziell analysiert.

Die energetische Analyse ist stark von der Abfolge der Aminosäuren abhängig. Wenn man die durchschnittliche Energie der Motive betrachtet stellt man fest, dass sie sich oft um das energetische 3. Quantil befindet, was auf eine verhältnismäßig geringe Stabilität hinweist. Wenn man jedoch einmal genauer die sequenzielle Struktur der Motive untersucht, stellt man eine starke Abhängigkeit fest. In allen fünf Motiven kommen an beliebigen Positionen bestimmte Aminosäuren konserviert vor. Diese sind natürlich je nach Funktion des Motivs unterschiedlich. *Tabelle 12* verdeutlicht, in welchen Positionen eine

starke Konservierung auftritt. Die motivischen Positionen sind in Spalte zwei dargestellt und die zugehörigen variierenden Residuen nehmen Spalte drei ein.

Tabelle 12: *Die analysierten Sequenzmotive (PROSITE-Motive) weisen an mindestens einer Position eine sequenzielle Konservierung auf. Die jeweilige Aminosäure ist dabei von der Motiv-spezifischen Funktion abhängig.*

PROSITE-Motiv	**Position**			**konservierte Aminosäuren**		
	a)	**b)**	**c)**	**a)**	**b)**	**c)**
PS00001	1	3		N	S oder T	
PS00004	1	2	4	K oder R	K oder R	S oder T
PS00005	1	3		S oder T	K oder R	
PS00006	1	4		S oder T	D oder E	
PS00008	1			G		

Besonders auffällig ist, dass stets die erste Position konserviert auftritt. Dies hängt allerdings mit der Ausübung ihrer jeweiligen Funktion zusammen. Wenn man sich das globale energetische Verhalten der Motive anschaut, so erkennt man, dass die konservierten Aminosäuren stets einen hohen Energiewert besitzen. Die stabile Struktur der Motive wird dabei vermutlich rein über die variablen Residuen erreicht, denn diese weisen im Mittel Energiewerte im mittleren Bereich des 3. Quantils auf und sind damit schon als verhältnismäßig stabil anzusehen. Die durchschnittliche energetische Verteilung an allen Positionen ist in *Tabelle 13* dargestellt. Die bereits erwähnten konservierten Residuen weisen dabei hohe Energiewerte auf, die verbliebenen verhältnismäßig niedrige, was der großen Diversität an diesen Stellen zuzuschreiben ist.

Tabelle 13: *Die globale energetische Verteilung der PROSITE-Motive repräsentiert meist durchschnittliche Energien im Bereich des 3. Quantils. Konservierte Positionen weisen hohe Energiewerte auf, disperse verhältnismäßig niedrige.*

PROSITE-Motiv	1. AS	2. AS	3. AS	4. AS	5. AS	6. AS	Energetisches Mittel
PS00001	-3,01	-12,38	-7,32	-11,80			**-8,63**
PS00004	-2,49	-1,81	-12,80	-5,48			**-5,65**
PS00005	-6,96	-13,63	-3,42				**-8,00**
PS00006	-6,53	-13,40	-9,48	-0,99			**-7,60**
PS00008	-7,22	-13,40	-10,97	-11,83	-10,89	-14,40	**-11,45**

Die Sequenzmotive wurden in der Vergangenheit schon sehr intensiv erforscht und dies lässt gerade auf sequenzieller Basis nur wenige neue Erkenntnisse zu. Allein die relativen Häufigkeiten des Auftretens aller Aminosäuren können dahingehend charakterisiert werden. Die zugehörigen Tabellen sind im Anhang unter Punkt **A4** dargestellt. Als Deskriptor kann eine solche Auflistung unter Umständen weiterhelfen, weshalb diese Analyse auch durchgeführt wurde. Wenn man die Seite der *PROSITE*-Datenbank besucht kann man sich jedoch sequenzielle Besonderheiten auch sofort zu allen Motiven anzeigen lassen, weswegen an dieser Stelle auf eine genaue Beschreibung verzichtet wurde.

5.3 Sekundärstrukturelemente von Proteinen

Die energetische Analyse aller Residuen in den drei Sekundärstrukturen manifestiert die Tatsache, dass Coil-Bereiche einen wesentlich instabileren Charakter aufweisen. Jedes Residuum hat im Mittel einen höheren Energiewert als in den Helix- und Sheet-Bereichen. *Tabelle 14* verdeutlicht den eindrucksvollen energetischen Gradienten „Coil → Helix → Sheet“ an dem Datensatz über 114 Proteine. Dabei wird die durchschnittliche freie Energie bei den verschiedenen Sekundärstrukturausprägungen in den drei Spalten dargestellt.

Tabelle 14: *Der energetische Gradient für jede Aminosäure beginnt stets mit dem höchsten Wert in dem Coil-Bereich und endet mit dem niedrigsten Betrag in der Sheet-Konformation. Die oftmals stabilisierte Lage der Faltblätter im Inneren eines Proteins stützt diese Charakterisierung.*

Aminosäure	Energie - Coil	Energie - Helix	Energie - Sheet
Ala	-11,40	-14,94	-17,79
Arg	-4,53	-5,78	-7,50
Asn	-3,14	-4,17	-5,67
Asp	-0,85	-1,76	-2,87
Cys	-22,32	-27,56	-28,22
Gln	-3,63	-5,20	-7,07
Glu	0,02	-0,63	-2,02
Gly	-5,06	-7,29	-10,51
His	-10,69	-12,71	-14,41
Ile	-23,16	-27,71	-31,13
Leu	-21,02	-25,66	-29,32
Lys	0,24	-0,85	-1,62
Met	-20,32	-24,91	-27,41
Phe	-21,67	-25,77	-29,61
Pro	-4,04	-4,88	-7,42
Ser	-5,30	-6,44	-9,06
Thr	-6,42	-8,04	-10,75
Trp	-18,18	-21,02	-22,28
Tyr	-16,70	-18,70	-23,53
Val	-19,47	-23,21	-25,13

Um diese Charakteristika in das QSER-Modell einzubringen, wurden die Präferenzen zu allen Aminosäuren in den vier energetischen Quantilen berechnet. Dies wurde mit Hilfe des *Chou-Fasman*-Algorithmus durchgeführt, welcher bereits in *Formel* (9) vorgestellt wurde. An den Präferenzen kann man das Auftreten jeder Aminosäure in allen Quantilen im Verhältnis zur gesamten Anzahl in der jeweiligen Sekundärstruktur erkennen und die Werte somit untereinander vergleichen. Die Werte sind in *Tabelle 15* abgebildet. Dabei repräsentieren die Spalten in chronologischer Abfolge die Quantile 1 bis 4 und sind jeweils in Coil (C), Sheet (E) sowie Helix (H) aufgegliedert. Der höchste Wert jedes Residuums in einem der Sekundärstrukturelemente ist zusätzlich farbig hervorgehoben, um eine bessere Übersicht

zu gewährleisten. Eine Tabelle zur anzahlmäßigen Verteilung der Aminosäuren über den gesamten energetischen Bereich ist im Anhang unter Punkt **A5** dargestellt.

Tabelle 15: *Die Präferenz jedes Residuums im energetischen Bereich in Kombination zu den Sekundärstrukturelementen ist für jede Aminosäure sehr speziell und für eine Charaktersierung essenziell. Die Quantile sind in aufsteigender Reihenfolge von links nach rechts aufgelistet, wobei jeweils noch in die Sekundärstrukturen unterschieden wird. Dabei steht C für Coil, E für Sheet und H für Helix. Die farbig hervorgehobenen Werte markieren die Höchstwerte jedes Chou-Fasman-Parameters in den Sekundärstrukturbereichen.*

	1. Q			2. Q			3. Q			4. Q		
	C	E	H	C	E	H	C	E	H	C	E	H
Ala	0,102	0,062	0,014	0,907	0,649	0,678	2,728	1,624	2,540	0,498	0,937	0,810
Arg	1,037	0,659	0,834	1,809	3,244	2,570	0,419	1,283	0,664	0	0	0,008
Asn	1,704	1,550	1,667	1,258	3,627	2,197	0,161	0,748	0,209	0,015	0	0
Asp	2,538	5,407	3,132	0,454	2,425	0,911	0	0,069	0,031	0,015	0	0,009
Cys	0	0	0	0,039	0	0,032	1,480	0,301	0,327	5,124	2,125	3,411
Gln	1,344	0,853	0,907	1,611	3,174	2,779	0,230	1,254	0,363	0	0	0,019
Glu	2,797	6,605	3,687	0,174	1,721	0,381	0	0,031	0,007	0	0,011	0,012
Gly	0,931	0,742	0,705	1,725	1,464	1,969	0,688	2,126	1,417	0	0,099	0
His	0,073	0	0,026	1,004	0,225	0,718	2,837	3,015	3,176	0,145	0,156	0,189
Ile	0,023	0	0,012	0,117	0	0,034	1,333	0,325	0,609	5,140	2,109	3,144
Leu	0,026	0	0,017	0,123	0	0,023	1,864	0,466	0,909	4,158	2,011	2,880
Lys	2,861	7,381	3,704	0,095	1,291	0,346	0,011	0,018	0,010	0	0	0,026
Met	0,032	0	0,022	0	0	0,022	1,802	0,449	0,860	3,949	2,023	2,920
Phe	0	0	0,015	0,155	0,026	0,015	1,584	0,307	0,885	4,655	2,111	2,909
Pro	1,374	2,002	1,370	1,476	2,484	2,087	0,366	1,241	0,623	0	0	0
Ser	0,873	0,620	0,448	1,728	2,351	2,642	0,763	1,806	0,980	0,013	0	0,009
Thr	0,443	0,113	0,456	1,890	1,598	1,895	1,180	2,385	1,733	0,014	0,021	0,012
Trp	0	0	0	0,162	0	0,086	2,549	1,316	1,606	2,893	1,422	2,209
Tyr	0	0	0	0,286	0,044	0,080	2,638	0,801	2,370	2,432	1,761	1,527
Val	0,025	0	0,021	0,127	0,010	0,064	2,127	0,799	1,205	3,675	1,777	2,571

Bei der Betrachtung der Tabelle und den darin enthaltenen Werten fällt die Verteilung der markierten Werte auf. Die Maxima in den Sekundärstrukturen sind maximal in zwei verschiedenen Quantilen aufzufinden, welche auch nur eine Einheit voneinander entfernt sind. Demzufolge gibt es keine Aminosäure, welche bspw. im 1. und 3. Quantil ihre Höchstwerte aufweist. Nur eine Kombination aus dem 1. und 2., dem 2. und 3. sowie dem 3. und 4. Energiebereich sind vorzufinden. Desweiteren erkennt man an den markierten Größen, dass die Verteilung an Maximalwerten definierten Mustern unterliegt. Die gruppierten Residuen sind nachfolgend aufgegliedert:

1. Asparagin (Asn)
2. Asparaginsäure (Asp), Glutaminsäure (Glu), Lysin (Lys)
3. Arginin (Arg), Glutamin (Gln), Prolin (Pro), Serin (Ser)
4. Glycin (Gly), Threonin (Thr)
5. Alanin (Ala), Histidin (His)
6. Tyrosin (Tyr)
7. Cystein (Cys), Isoleucin (Ile), Leucin (Leu), Methionin (Met), Phenylalanin (Phe), Tryptophan (Trp), Valin (Val)

Wenn man diese Cluster an Aminosäuren vergleicht, stößt man auf ähnliche Eigenschaften der Residuen. Die Verbindung, d.h. welche Eigenschaft es dann in jedem Fall ist, muss die Regressionsanalyse zeigen und lässt sich im Vorherein nur mutmaßen. Teilweise unterscheiden sich die Aminosäuren in ihren Ausprägungen in nur einer Position, weshalb bei diesen eigentlich auch ein kausaler Zusammenhang bestehen muss. Dies legt die Vermutung nahe, dass auch die Kombination aus zwei oder drei Eigenschaften, welche die Energie-Sekundärstruktur-Beziehung beeinflussen, für die ähnlichen Charakteristika verantwortlich ist.

5.4 Interaktionen der Aminosäuren in globulären Proteinen

Für die Analyse wurden zwei Datenreihen aufgestellt. Zum einen die Analysen über den *Weizmann*-Server und das zugehörige Programm *CMA* und andernfalls mittels eines

verfassten Programmes. Dieses berechnet die Anzahl an Wechselwirkungen zwischen allen Aminosäuren in einer Kugel mit dem Radius von 8Å. Die Vergleichbarkeit der Daten ist nur auf Grund ihrer mathematischen Werte möglich, denn für beide Analysen wurden unterschiedliche Datensätze gewählt. Als Ausgangspunkt diente das maximierte Set mit 4303 Proteinen. Wegen der hohen Anzahl an Informationen durch den *CMA*-Server, wurde ein Programm generiert, welches sich zufällig 400 Proteine auswählt, die Informationen dieser Komposition an den Server weiterleitet und dieser anschließend die Berechnungen durchführt. Bei einer größeren Anzahl an Proteinen konnte nicht mehr gewährleistet werden, dass der Computer störungsfrei arbeitet.

Die Ermittlung des relativen Auftretens jeder Aminosäure soll die vergleichbaren Ergebnisse näher veranschaulichen. *Tabelle 16* beschreibt diese Parameter in vergleichender Form von dem *CMA*-Server zu dem verfassten Berechnungsprogramm. Dazu ist für jedes Residuum die Anzahl an Wechselwirkungen mit anderen Aminosäuren angegeben und zusätzlich noch das relative Auftreten im Vergleich zu allen ermittelten Kontakten vermerkt. Die meisten Residuen haben eine geplante und vertretbare Abweichung von bis zu 0,004. Sie ist in der letzten Spalte dargestellt. Dies kann an den beiden verschiedenen Datensätzen liegen, wodurch zweifelsohne Unterschiede entstehen können. Bei einigen treten jedoch auch größere Unterschiede auf, was auf den ersten Blick etwas suspekt wirkt. Wenn die relative Häufigkeit bei dem *CMA*-Experiment geringer ist als bei dem Programm zur Untersuchung der Interaktionen im 8Å Radius, ist die betreffende Aminosäure in grüner Farbe betont, andernfalls ist sie rot gekennzeichnet. Bei Glutamin stimmen beide Werte überein, weshalb es in schwarzer Färbung hervorgehoben ist. Wenn man die Struktur und die atomare Ausprägung der Seitenkette dieser Aminosäuren betrachtet, erkennt man, dass allein durch die Grundidee des Programmes mit dem verarbeiteten 8Å Radius Unterschiede entstehen müssen. Eine vollständig eindeutige Erklärung für alle Aminosäuren ist an Hand der Struktur leider nicht möglich, jedoch lassen sich differenzierte Auffälligkeiten erkennen. Wenn die relative Häufigkeit geringer ist handelt es sich meist um Residuen, welche sehr groß sind und dabei besonders eine verzweigte Anordnung der Atome oder eine Ringstruktur aufweisen. Sie nehmen mehr Platz ein und es können sich in ihrem 8Å Radius nicht so viele andere Aminosäuren aufhalten. Bei der Berechnung über die *Van-der-Waals*-Oberfläche durch *CMA* bilden diese Residuen natürlich eine größere Kontaktfläche aus, da

sie mehr Atome enthalten. An Hand der chemischen Eigenschaften lassen sich keine Aussagen treffen.

Tabelle 16: *Der Vergleich der beiden Programme zur Erfassung von Kontakten zwischen Aminosäuren liefert in den meisten Fällen adäquate Werte. Bei dem Vergleich der relativen Häufigkeiten für das Ausbilden an Kontakten im Vergleich zur Gesamtanzahl aller Residuen weisen die meisten eine normale Abweichung auf Grund zweier unterschiedlicher Datensätze auf. Die Residuen sind je nach Abweichung farblich markiert. Wenn die Berechnung durch das CMA-Weizmann-Programm höher ist als die der Betrachtung durch den 8Å Radius ist die Aminosäure rot markiert, andernfalls grün. Bei Übereinstimmung ist sie schwarz gekennzeichnet. Rot-markierte Residuen weisen wahrscheinlich auf Grund ihrer meist großen und verzweigten Form die erhöhten Häufigkeiten auf, wobei grün hervorgehobene meist kompakt und unverzweigt sind.*

Aminosäure	Auszählung *CMA-Weizmann*		Auszählung im 8Å-Radius		Differenz
	Anzahl	relative Häufigkeit	Anzahl	relative Häufigkeit	
Ala	55772	0,072	755209	0,082	**0,010**
Arg	40332	0,052	440165	0,048	**0,004**
Asn	32840	0,042	396165	0,043	**0,001**
Asp	37911	0,049	466027	0,051	**0,002**
Cys	11937	0,015	179512	0,019	**0,004**
Gln	27657	0,036	328650	0,036	**0,000**
Glu	42489	0,055	523839	0,057	**0,002**
Gly	43688	0,056	662035	0,072	**0,016**
His	17660	0,023	196259	0,021	**0,002**
Ile	54277	0,070	572252	0,062	**0,008**
Leu	85987	0,111	909558	0,099	**0,012**
Lys	42136	0,054	507101	0,055	**0,001**
Met	17528	0,023	206164	0,022	**0,001**
Phe	39514	0,051	391493	0,042	**0,009**
Pro	31488	0,041	392340	0,043	**0,002**
Ser	40135	0,052	550118	0,060	**0,008**
Thr	41648	0,054	519584	0,056	**0,002**
Trp	15366	0,020	143436	0,016	**0,004**
Tyr	34912	0,045	352102	0,038	**0,007**
Val	63673	0,082	727253	0,079	**0,003**

Da die Berechnungen bei der *CMA*-Analyse im biologischen Kontext eine im höchsten Maß fundierte Stellung einnehmen, wurden für genaue Aussagen in Bezug auf die Kontakte aller Aminosäuren in den Sekundärstrukturen die Ergebnisse von diesem Programm genutzt. Das eigen verfasste Programm verarbeitet nicht ausführlich genug die atomaren Kräfte der Residuen, sondern stützt sich rein auf die Größe und die Anzahl an Atomen. *Tabelle* 14 verdeutlichte dieses Problem bereits mit partiell unverhältnismäßig großen Differenzen. Es ist eine Möglichkeit um einen sehr großen Datensatz zu verarbeiten, jedoch muss man mit Abstrichen in der Präzision rechnen.

Bei der vorrangigen Betrachtung der Kontakte aller Aminosäuren in den Sekundärstrukturen wurde der Ansatz gewählt, dass sich nur das jeweils betrachtete Residuum in einer definierten Sekundärstruktur befinden muss. Die Kontakt-Residuen können anschließend in allen Sekundärstrukturbereichen liegen. Es wurde zusätzlich noch zwischen sequenzieller Nähe sowie Ferne unterschieden. Als Grenzwert für sequenzielle Umgebung gilt dabei $i \pm 3$ für eine betrachtete Aminosäure i. [8] Kontakte mit anderen Residuen werden in die Statistik als sequenzielle Ferne aufgenommen. Das *CMA*-Programm macht es möglich, die Interaktionen zwischen Proteinketten in einem Protein zu erkennen. Da die Analysen dieser Bachelorarbeit nur einkettige Proteine umfassen, wurden die internen Wechselwirkungen der *chain* jedes Residuums berechnet. *Abbildung 20* beschreibt an der Endoglucanase *CEL5A* von dem *Bacillus Agaradherans* (PDB ID 1A3H) den Output des Programmes an einem kleinen Ausschnitt, welcher anschließend von einem aufgesetzten Programm registriert und verarbeitet wurde, um die Sekundärstruktur-Analysen zu ermöglichen. Dabei ist an der horizontalen sowie der vertikalen Achse die Sequenz abgebildet, einschließlich der Position jeder Aminosäure und der *chain*, in welcher es enthalten ist. In diesem Fall stets die Kette „A“. Die blau gefärbten Quadrate kennzeichnen auftretende Kontakte zwischen Residuen.

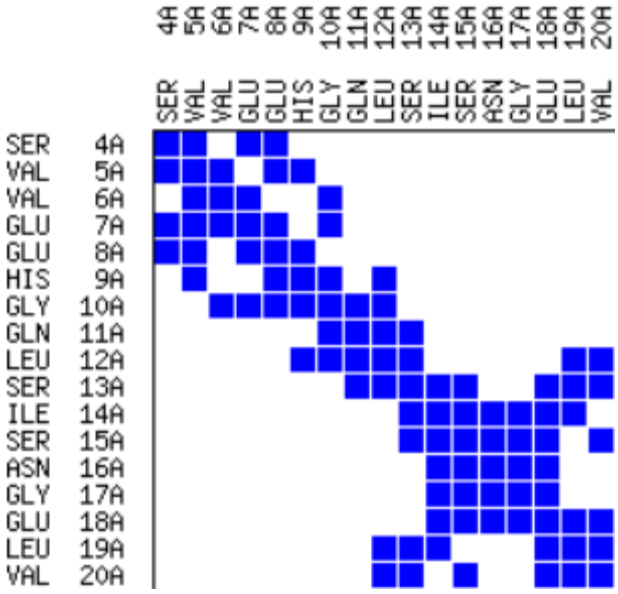

Abbildung 20: *Zur Berechnung von Kontakten nutzt das Programm CMA des Weizmann Institutes die Van-der-Waals Oberfläche jeder Aminosäure. Die Abbildung zeigt einen Ausschnitt der Analyse an der Endoglucanase CEL5A von dem Bacillus Agaradherans (PDB ID 1A3H). Dieses Protein besitzt eine Kette, welche miteinander verglichen werden, sodass Interaktionen in dem Protein auf Basis der Aminosäuren zu detektieren sind. Die Aminosäuren, die zugehörige Position und die chain sind an der horizontalen sowie der vertikalen Achse abgebildet. Blaue Quadrate weisen auf einen Kontakt hin.* [25]

Nach der Zuordnung und Charakterisierung aller Interaktionen lassen sich Aussagen über jede Aminosäure treffen. Der Vergleich zwischen beiden Ausprägungen ist in *Tabelle 17* dargestellt. Alle Aminosäuren, außer Glycin und Asparaginsäure, bilden in sequenzieller Ferne mehr Kontakte aus. Dabei können die Unterschiede nur wenige Hundert Interaktionen umfassen, wie bspw. bei Glutaminsäure, oder aber auch beinahe die doppelte Anzahl aufweisen, wie es bei Leucin der Fall ist. Dieses Residuum bildet zudem mit weitem Abstand die meisten Kontakte aus, gefolgt von Valin und Alanin. Dies liegt vermutlich an der Form der Seitenkette von Leucin. Die Struktur weist eine gegabelte Form mit zwei Methylgruppen auf. Auch Valin besitzt diese Konformation seines Restes, jedoch ist dieser kürzer, besitzt infolgedessen nicht so viele Atome und kann keine so umfangreiche *Van-der-Waals*-Oberfläche ausbilden. Alanin ist verhältnismäßig klein, weshalb sich Residuen stark annähern können und Interaktionen ausbilden können. Cystein geht die wenigsten Beziehungen ein, was wahrscheinlich an seinem verhältnismäßig geringen Aufkommen liegt.

Tabelle 17: *Die meisten Aminosäuren bilden in sequenzieller Ferne mehr Interaktionen mit Residuen aus. Dies geschieht wahrscheinlich wegen der räumlich umfassenden Ausbildung ihrer Seitenketten. Leucin beschreibt insgesamt die meisten Kontakte, Cystein die wenigsten. Dies hängt vermutlich mit der geringen Häufigkeit von Cystein zusammen.*

Aminosäure	sequenzielle Ferne	sequenzielle Nähe	Gesamtanzahl
Ala	30429	25343	**55772**
Arg	23648	16684	**40332**
Asn	17048	15792	**32840**
Asp	18693	19218	**37911**
Cys	7539	4398	**11937**
Gln	14691	12966	**27657**
Glu	21411	21078	**42489**
Gly	20721	22967	**43688**
His	10417	7243	**17660**
Ile	36068	18209	**54277**
Leu	56707	29280	**85987**
Lys	22030	20106	**42136**
Met	11370	6158	**17528**
Phe	27062	12452	**39514**
Pro	16574	14914	**31488**
Ser	20656	19479	**40135**
Thr	23225	18423	**41648**
Trp	10745	4621	**15366**
Tyr	23551	11361	**34912**
Val	40863	22810	**63673**

Sequenzielle Nähe

Die sequenzielle Nachbarschaft wurde, wie bereits erwähnt, von einer betrachteten Aminosäure i mit anderen Residuen bis $i \pm 3$ gewählt. Durch die variierenden Seitenketten entstehen folglich Unterschiede in der Anzahl ihrer Interaktionen. Das unterschiedliche Aufkommen, bezogen auf die Sekundärstrukturbereiche, ist jedoch für jede Aminosäure sehr speziell. Die Divergenzen lassen sich in *Tabelle 18* an Hand der graduellen orangenen Färbungen nachvollziehen. Eine dunkle Kolorierung weist auf ein hohes Vorkommen hin.

Auch die Gesamtanzahlen aller Ausprägungen der Residuen sind in ähnlicher Form gekennzeichnet, nur wurde dabei eine grüne Schattierung eingesetzt. Für eine schnellere und übersichtlichere Bezugnahme zu den Verteilungen ist zusätzlich in der letzten Spalte die relative Häufigkeit zu der Anzahl aller Ausprägungen jeder Aminosäure aufgetragen.

Die meisten Residuen folgen dem Muster, die häufigsten gemessenen Kontakte in den Helix-Bereichen auszubilden. Dies ist auf Grund der spiralisierten Form auch nachvollziehbar, da auf einer kleineren Fläche viel mehr Aminosäuren sind, mit welchen Interaktionen eingegangen werden können. Eine ähnliche Erklärung ist auch für Coil-Bereiche festzuhalten. Sie verbinden sehr oft Helices oder Faltblätter und liegen häufig „scheinbar wahllos“ im äußeren Kompartiment der Proteine. Die Behauptung „scheinbar wahllos“ ist so gewählt, da noch nicht bewiesen werden kann, dass die Ausbildung wirklich zufällig ist. In Bezug auf die Reflexion der Gesamtanzahlen fällt eine relativ homogene Verteilung auf. Keine Aminosäure weist eine absolute Mehrheit in der Anzahl der ausgebildeten Kontakte auf.

Auch bei der Betrachtung der reinen anzahlmäßigen Verteilung der Aminosäuren in den Sekundärstrukturen kann man die Dispersion vermuten. An dem kleinen Datensatz ist diese zwecks der Vollständigkeit und Vergleichbarkeit in *Tabelle 19* abgebildet. Wenn man die graduelle orangene Färbung abgleicht, erkennt man die beinahe identischen Verhaltensweisen.

Um zudem erneut mathematische Werte für eine Analyse bereit zu stellen, wurden die Präferenzen mittels *Formel* (4) berechnet. Sie sind im Anhang unter Punkt **A6** einsehbar. Bei ihrer Betrachtung erkennt man, dass sich nahezu alle gleichmäßig verhalten und nur geringfügige Schwankungen enthalten sind. Die Werte verteilen sich von 0,441 bis 1,882 homogen über den gesamten Bereich und könnten für das große Ziel, der strukturellen Erklärung der Proteine mit Hilfe der energetischen Ansätze entscheidend sein, wenngleich bei dieser Analyse das Energiekriterium nicht mit eingeflossen ist.

Tabelle 18: *Die Gesamtanzahl aller Kontakte jeder Aminosäure bei der Betrachtung in sequenzieller Nähe sind maßgeblich von der Struktur der Seitenkette jedes Residuums abhängig. Die Ausprägungen in den Sekundärstrukturen hängen von ihrer Verteilung in den Proteinen ab. Da die Aminosäuren meist in Helices oder Coil-Bereichen vorliegen verschiebt sich auch ihre Anzahl an Ausprägungen in diese beiden Bereich, wohingegen in Faltblattstrukturen weniger Kontakte zu messen sind.*

Aminosäure	Sekundärstruktur			Gesamtanzahl	relative Häufigkeit
	Coil	Helix	Sheet		
Ala	7337	14258	3748	**25343**	**0,078**
Arg	5799	7693	3192	**16684**	**0,052**
Asn	7744	5817	2231	**15792**	**0,049**
Asp	9192	7622	2404	**19218**	**0,059**
Cys	1574	1663	1161	**4398**	**0,014**
Gln	4432	6467	2067	**12966**	**0,040**
Glu	6775	10778	3525	**21078**	**0,065**
Gly	14479	5735	2753	**22967**	**0,071**
His	2768	2784	1691	**7243**	**0,022**
Ile	4428	7124	6657	**18209**	**0,056**
Leu	7638	14496	7146	**29280**	**0,091**
Lys	7387	9066	3653	**20106**	**0,062**
Met	1704	3235	1219	**6158**	**0,019**
Phe	3716	4753	3983	**12452**	**0,038**
Pro	9884	3624	1406	**14914**	**0,046**
Ser	8365	7723	3391	**19479**	**0,060**
Thr	7454	6179	4790	**18423**	**0,057**
Trp	1521	1870	1230	**4621**	**0,014**
Tyr	3639	3934	3788	**11361**	**0,035**
Val	5739	7885	9186	**22810**	**0,071**
	121575	**132706**	**69221**	***323502***	

Tabelle 19: *Die Sekundärstrukturverteilung der Aminosäuren in dem kleinen Datensatz zeigt, dass die wenigsten Residuen bevorzugt in einem Sheet-Bereich aufzufinden sind. Coil- und Helix-Regionen weisen höhere Aufkommen der Residuen auf.*

Aminosäure	Coil	Helix	Sheet
Ala	523	885	309
Arg	396	481	231
Asn	541	297	135
Asp	533	405	146
Cys	168	127	123
Gln	266	388	145
Glu	338	617	216
Gly	1015	319	282
His	163	158	105
Ile	252	354	445
Leu	454	712	412
Lys	414	433	192
Met	92	188	90
Phe	190	267	230
Pro	643	233	76
Ser	606	402	276
Thr	547	323	336
Trp	101	142	87
Tyr	263	203	269
Val	361	382	603

Sequenzielle Ferne

Die sequenzielle Ferne verschafft dem Verhältnis der Aminosäuren in Bezug auf die Sekundärstrukturen eine ausgewogenere Verteilung. Abgebildet in *Tabelle 20* erkennt man an der graduellen Färbung der Aufkommen in den Sekundärstrukturen, dass nun auch mehr Aminosäuren ihr Maximum an Ausprägungen in dem Faltblatt-Bereich haben. Die Ursache dafür ist sicherlich, dass ein Faltblatt aus mehreren Sheets besteht und meist Bindungen untereinander ausgebildet werden, sodass die Aminosäuren mit mehr Residuen in Kontakt treten können. In Faltblättern steigt die Anzahl der Kontakte aller Residuen kontinuierlich,

wohingegen bei einigen Aminosäuren die Quantität in den beiden anderen Sekundärstrukturbereichen abnimmt. Dies stellt allerdings auch mehr die Ausnahme statt der Regel dar. In diesem Zusammenhang ist Glycin hervorzuheben, denn bei ihr steigt im Sheet-Bereich die Anzahl der Ausprägungen, wohingegen sie in Helices und Coil-Strukturen sinkendes Auftreten beweist. Diese Eigenart lässt sich am plausibelsten mit der Struktur begründen. Glycin ist die kleinste Aminosäure und weist keine übliche Seitenkette auf, sondern besitzt an dieser Position nur ein Wasserstoffatom. Durch die geringe Anzahl an Atomen kann sich keine große *Van-der-Waals*-Oberfläche ausbilden, was zur Folge hat, dass in sequenzieller Ferne bei der Lage in Coil- und Helix-Gebieten die Abstände zu groß sind und sich nicht so viele Kontakte ausbilden können.

Wie bereits bei der ersten vergleichenden Analyse und dahingehend in *Tabelle 15* dargestellt, fallen einige Residuen besonders auf, bei denen die Gesamtanzahlen der Aminosäuren zum einen eine sehr hohe Menge an Kontakten aufweisen und desweiteren auch im Vergleich zu den Ausprägungen in sequenzieller Nähe einen massiven Anstieg zu verzeichnen haben. Es sind Isoleucin, Leucin und Valin. Wenn man nun einmal die einzelnen Sekundärstrukturen betrachtet, ist auch in ihnen stets ein gewichtiger Zuwachs an Interaktionen festzustellen. Bei Isoleucin verschiebt sich zudem die Maximalanzahl an Ausprägungen deutlich von dem Helix- in den Sheet-Bereich. Dies ist mit der bevorzugten Lage dieser Aminosäure in dieser Struktur zu erklären, ersichtlich in *Tabelle 17*, sowie mit derselben Erklärung, wie sie bei Glycin getroffen wurde. Durch die sequenzielle Ferne ist es ihr einfach möglich, näher in Kontakt mit anderen Residuen zu treten, da die dichte Packung und Anordnung eines Sheets dieses sonst verbietet. Die Präferenzen für alle Aminosäuren wurden zusätzlich berechnet und sind im Anhang unter Punkt **A7** ersichtlich.

Tabelle 20: *In sequenzieller Ferne nehmen fast alle Ausprägungen zu, nur in wenigen Ausnahmen ist ein Rückgang zu beobachten. Dabei ist besonders Glycin zu erwähnen, welches durch seine geringe Größe nun weniger Interaktionen in Coil- und Helix-Bereichen zu verzeichnen hat und zudem einen Anstieg in dem Sheet-Kompartiment aufweist. Leucin, Isoleucin und Valin erfahren einen besonders hohen Anstieg. Bei Isoleucin verschiebt sich das Maximum aus dem Helix- in den Sheet-Bereich. Die liegt an der bevorzugten des Isoleucin in dieser Sekundärstruktur.*

Aminosäure	Sekundärstruktur			Gesamtanzahl	relative Häufigkeit
	Coil	Helix	Sheet		
Ala	7867	14736	7826	**30429**	0,067
Arg	7596	9545	6507	**23648**	0,052
Asn	7866	5281	3901	**17048**	0,038
Asp	8648	6183	3862	**18693**	0,041
Cys	2472	2458	2609	**7539**	0,017
Gln	4696	6343	3652	**14691**	0,032
Glu	6286	9438	5687	**21411**	0,047
Gly	11403	4562	4756	**20721**	0,046
His	3551	3509	3357	**10417**	0,023
Ile	7796	11885	16387	**36068**	0,080
Leu	13497	24357	18853	**56707**	0,125
Lys	7346	8372	6312	**22030**	0,049
Met	2884	5483	3003	**11370**	0,025
Phe	7252	9181	10629	**27062**	0,060
Pro	10953	3259	2362	**16574**	0,037
Ser	7958	6696	6002	**20656**	0,046
Thr	8284	6503	8438	**23225**	0,051
Trp	3086	4092	3567	**10745**	0,024
Tyr	6670	7050	9831	**23551**	0,052
Val	8778	11827	20258	**40863**	0,090
	144889	160760	147799	**453448**	

5.5 Winkelbeziehungen in globulären Proteinen

Die Analyse der Torsionswinkel gestaltete sich mit Hilfe eines angefertigten Programmes, welches den φ und ψ Winkel erfasst, dabei nach der Lage in den energetischen Quantilen unterscheidet und die quantitative Anzahl an Ausprägungen in allen Bereichen ermittelt. Als Datensatz fungierte das *Set* aus 114 Proteinen.

Die beiden Diederwinkel kommen sowohl in positiver, als auch in negativer Form vor. Eine sehr bekannte und populäre Variante der Darstellung bei Proteinen ist der *Ramachandran-Plot*, in welchem die Winkel für jedes Residuum vermerkt sind und die Verteilung erkennbar ist. Er ist auch für andere chemische Substanzen anwendbar. *Abbildung 21* zeigt einen vereinfachten Plot der *Endoglucanase CEL5A* von dem *Bacillus Agaradherans* (*PDB ID 1A3H*) und der *Hjm*-DNA-Helicase des Archaebakteriums *Pyrococcus furiosus* (*PDB ID 2ZJ8*), welcher mit einem verfassten Programm erstellt wurde. Es werden Residuen aus unterschiedlichen energetischen Quantilen farblich wie nachfolgend markiert:

1. Quantil: rot
2. Quantil: orange
3. Quantil: grün
4. Quantil: blau

Der Phi-Winkel wird auf der horizontalen, der Psi-Winkel auf der vertikalen Achse aufgetragen. In dem Plot sind betreffend der Sekundärstrukturelemente der Proteine drei Umgebungen charakteristisch. Der Sektor der Faltblattbereiche ist stets im oberen linken Bereich angesiedelt und mit „β-Sheet“ bezeichnet. Die meist vorkommende rechtsgedrehte Helix ist vorwiegend im linken mittleren Abschnitt zu sehen und die selten existierende linksgedrehte Helix im rechten Bereich. Die Gebiete ergeben sich durch die speziellen strukturellen Eigenschaften und damit verbundene sterische Ausprägungen. Man erkennt deutlich, dass Aminosäuren des 3. und 4. Quantils in der linksgedrehten Helix nur sehr selten vorkommen. in den Faltblättern überwiegen jedoch Aminosäuren aus dem 3. und 4. Quantil und sichern die stabile Struktur. Helices weisen sämtliche energetische Ausprägungen auf. Aminosäuren, welche weit außerhalb der Bereiche liegen, beschreiben

oftmals Coil-Bereiche, wobei meist nur einige Residuen zu einer solchen Lage neigen. Es sind meist kleine Aminosäuren wie Glycin, Alanin oder Serin.

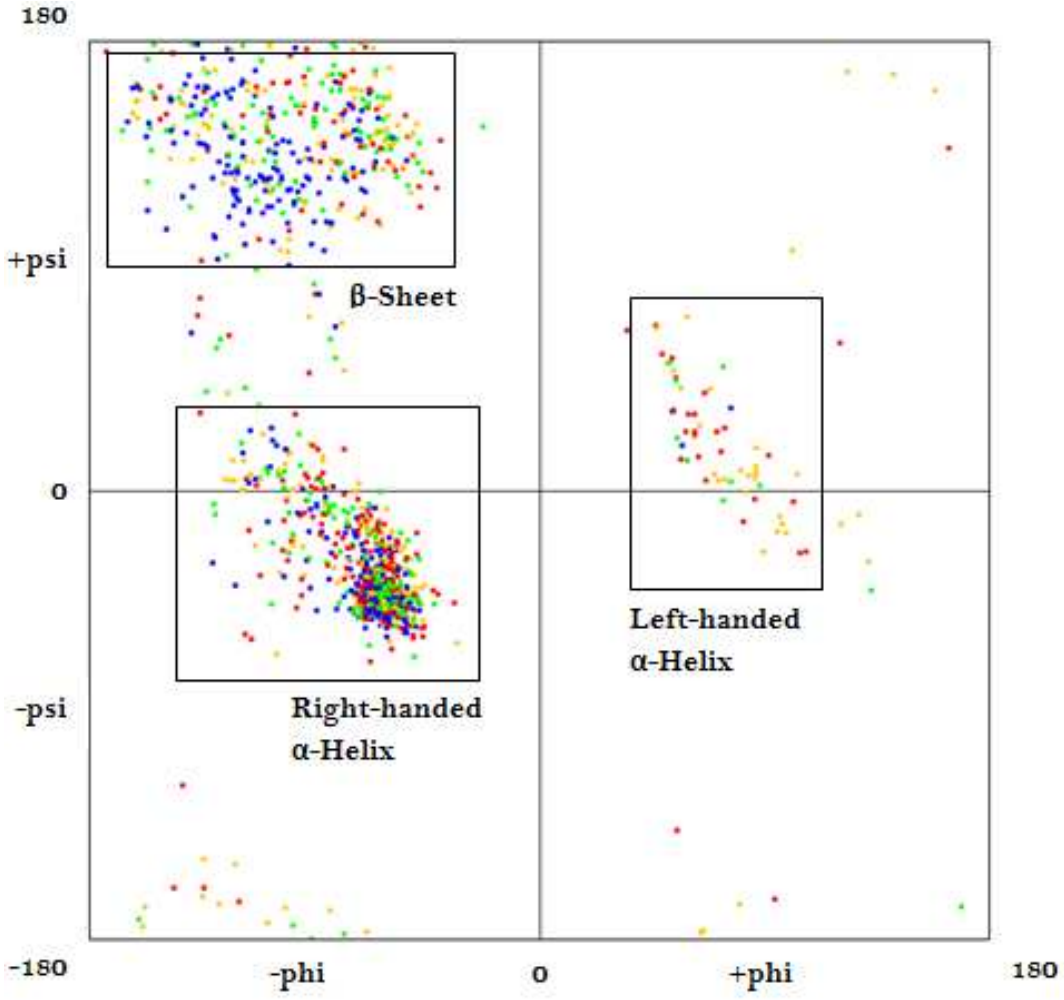

Abbildung 21: *Der Ramachandran-Plot zeigt die Verteilung der Phi- sowie Psi-Winkel der Endoglucanase CEL5A von dem Bacillus Agaradherans (PDB ID 1A3H) und der Hjm-DNA-Helicase des Archaebakteriums Pyrococcus furiosus (PDB ID 2ZJ8) auf. Dabei bilden sich durch die Struktur und die damit verbundenen sterischen Einschränkungen von Helices sowie Faltblattregionen spezifische Ballungsgebiete in dem Plot heraus. Die Residuen wurden zusätzlich nach ihrer Lage in den energetischen Quantilen farblich markiert. Eine rote Färbung steht für das 1. Quantil, orange für das 2., grün für das 3. und blau für den 4. Bereich. Linksgedrehte Helices weisen fast nur Residuen des 1. und 2. Quantils auf, Faltblätter sehr viele aus dem 3. und 4. Quantil und rechtsgängige Helices zeigen eine verhältnismäßig homogene Verteilung aller Energiebereiche. Aminosäuren außerhalb der Bereiche sind oftmals von kleiner Struktur und können durch die damit verbundene geringere strukturelle Einschränkung auch unübliche Konformationen aufweisen.*

Viele Analysen verschiedener Wissenschaftler beschäftigten sich bereits in der Vergangenheit mit den Charakteristika des *backbones* und den Winkelkonformationen, weshalb die Analyse lediglich diese Erkenntnisse vertieft.

Um die Neigungen und Verteilungen der Torsionswinkel im mathematischen Kontext aufzuzeigen wurden Winkel- sowie Sekundärstruktur-spezifische energetische Präferenzen mittels *Formel* (9) berechnet. Die Winkel-spezifischen sind dabei auf die Verteilung der Ausprägungen in den aufgegliederten Winkelbereichen gestützt, d.h. für die positive sowie negative Ausbildung des Winkels wurden für alle Sekundärstrukturen in den Quantilen die Präferenzen ermittelt, sodass ein spaltenweiser Vergleich in gleichen Bereichen möglich ist. Die Sekundärstruktur-spezifischen Präferenzen beziehen sich explizit auf jede Sekundärstruktur, sodass alle Werte für beliebige Winkel und Quantile miteinander verglichen werden können und parallele globale Rückschlüsse möglich sind.

Der Phi-Winkel

Die Winkel-spezifische Charakterisierung wird in *Tabelle 21* dargestellt. Nur im positiven Winkelbereich des 2. Quantils kommt es dazu, dass sich die drei Sekundärstrukturen in ihren Werten ähnlich verhalten, ansonsten ist dies nie der Fall. Am konstantesten reagiert die Helix im negativen Bereich. Die Werte schwanken dabei beständig um 1. Auch im positiven Segment ist nur ein geringer Gradient ersichtlich. Positive Winkel bei Coil-Abschnitten weisen eine ähnliche Charakteristik auf. Negative Daten schwanken stark. Am scheinbar kontroversesten verhält sich der Sheet-Bereich. Dies ist jedoch nicht so, denn durch die strukturelle Invarianz und die somit sterischen Einschränkungen ist nicht in allen Bereichen eine hohe Ausprägung möglich. Die Präferenzen verdeutlichen dies mit einem Gradient von etwa 2 bei den positiven Winkeln.

Tabelle 21: *Die Winkel-spezifischen Präferenzen des Phi-Winkels können Quantil-übergreifend in Bezug auf die positive oder negative Ausprägung hin verglichen werden, denn diese Betrachtung soll die Neigung der Ausprägung im positiven oder negativen Bereich über alle Sekundärstrukturen verdeutlichen. Keine Sekundärstruktur beweist dabei eine durchgängige hohe Neigung, jedoch weist die Helix im negativen Abschnitt einen konstanten Wert um 1 auf. Auch bei der positiven Ausprägung bildet dabei das 3. Quantil die einzige Ausnahme. Der Coil-Bereich verhält sich ähnlich, wenngleich im negativen Bereich ein hoher Gradient besteht. Sheets bilden die Ausnahme, denn dabei treten einige Konformationen sehr häufig auf. Dies ist mit seiner strukturellen Invarianz zu begründen.*

Sekundärstruktur	**Winkelspezifische Präferenzen - φ - Winkel**			
	1. Quantil		**2. Quantil**	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	1,059	1,402	0,994	1,207
Helix	0,868	1,015	0,991	1,019
Sheet	0,494	0,423	1,099	0,684

	3. Quantil		**4. Quantil**	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	0,862	0,949	1,077	0,518
Helix	1,373	0,941	1,011	1,025
Sheet	2,088	1,160	0	1,628

Tabelle 22 präzisiert die Präferenzen auf Sekundärstruktur-spezifischer Ebene. Dabei wird deutlich, dass in jedem Sekundärstrukturbereich ein signifikanter Bereich stark bevorzugt ist. Der Phi-Winkel weist dabei in allen Konstitutionen eine starke Neigung zu einer negativen Ausprägung auf. Die sterische Beschränktheit ist dafür maßgeblich. Den geringsten Gradienten weist der Coil-Bereich auf. Er resultiert aus der strukturellen Varianz dieser Abschnitte. Helix- und Sheet-Bereiche sind deutlich regelmäßiger in ihrem Aufbau und bilden demzufolge auch meist ähnliche Winkel aus.

Tabelle 22: *Die Sekundärstruktur-spezifischen Präferenzen sind in jeder Sekundärstruktur direkt miteinander vergleichbar. Dabei ist eine deutlich höhere Neigung zur Ausbildung von negativen Phi-Winkeln ersichtlich. Die sterische Beschränktheit ist dafür verantwortlich. Sie ist in Coil-Abschnitten am geringsten, weshalb in diesem Bereich zudem der höchste Gradient herrscht, welcher jedoch immer noch beachtlich ist und bis zu 2,5 beträgt. Die Faltblattregionen weisen mit Abstand die höchsten Werte im Bereich der negativen Phi-Winkel auf.*

Sekundärstruktur	**Sekundärstrukturspezifische Präferenzen - φ - Winkel**			
	1. Quantil		**2. Quantil**	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	0,466	2,131	0,539	2,058
Helix	0,082	2,600	0,102	2,580
Sheet	0,146	3,985	0,217	3,914

	3. Quantil		**4. Quantil**	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	0,213	2,383	0,043	2,553
Helix	0,048	2,634	0,003	2,680
Sheet	0,078	4,053	0	4,131

Der Psi-Winkel

Die in *Tabelle 23* abgebildeten Winkel-spezifischen Präferenzen für den Psi-Winkel zeigen abermals eine deutliche divergente Verteilung. In Bezug auf die Sekundärstrukturen lassen sich keine Muster erkennen. Die sterischen Einschränkungen sind auch hierbei für die festen Konformationen verantwortlich.

Tabelle 23: *Die Winkel-spezifischen Präferenzen des Psi-Winkels zeigen, ähnlich wie die des Phi-Winkels, keine auffälligen Übereinstimmungen. Die sterischen Beschränkungen bewirken auch hierbei eine homogene Verteilung der Ausprägungen.*

Sekundärstruktur	**Winkelspezifische Präferenzen - ψ - Winkel**			
	1. Quantil		**2. Quantil**	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	1,421	1,489	1,215	1,294
Helix	1,281	0,843	1,396	0,893
Sheet	0,448	0,558	0,666	1,010

	3. Quantil		**4. Quantil**	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	0,950	0,738	0,520	0,331
Helix	0,821	1,057	0,578	1,258
Sheet	1,096	1,849	1,650	0,654

Die Sekundärstruktur-spezifischen Präferenzen beschreiben deutliche divergente Werte, wobei wiederum die Faltblattstrukturen wegen ihrer sehr konstanten Anordnung und Ausrichtung die höchsten Daten aufweisen. *Tabelle 24* beschreibt die Präferenzen und zeigt eine vergleichbare Charakteristik zu dem Phi-Winkel.

Tabelle 24: *Die Sekundärstruktur-spezifischen Präferenzen für den Psi-Winkel verdeutlichen die strukturelle und sterische Invarianz der Sheet-Konformation mit deutlich höheren Präferenzen und die Coil-Bereiche mit relativ kleinen Quantil-spezifischen Gradienten.*

Sekundärstruktur	Sekundärstrukturspezifische Präferenzen - ψ - Winkel			
	1. Quantil		2. Quantil	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	1,631	0,966	1,727	0,870
Helix	0,391	2,292	0,464	2,218
Sheet	3,748	0,382	3,742	0,389

	3. Quantil		4. Quantil	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	1,997	0,599	2,102	0,495
Helix	0,303	2,380	0,195	2,487
Sheet	3,774	0,356	4,045	0,861

5.7 Disulfidbrücken

Eine Charakterisierung und qualitativ aussagekräftige Beurteilung unter Angabe von mathematischen Werten, welche in das QSER-Modell einfließen können, konnte bei dieser Analyse nicht erbracht werden. Dies hängt zum einen mit der geringen Anzahl an Disulfidbrücken im Datensatz zusammen, denn es sind über alle 114 Proteine nur 103 Disulfidbrücken-Ausbildungen, aber auch zusätzlich mit der Charakteristik der Aminosäure Cystein. Sie weist fast ausnahmslos Energiewerte im sehr stabilen Bereich auf. Nur selten ist der energetische Betrag im 3. Quantil aufzufinden. Der durchschnittliche Wert aller Cysteine im Set beträgt, wie in *Tabelle* 1 bereits aufgezeigt, -26,03. Die Cysteine, welche an den Ausbildungen der stabilisierenden Brücken beteiligt sind, weisen einen mittleren Wert von -25,22 auf und sind somit energetisch nur marginal unter dem Gesamtbetrag. Diese Tatsache macht eine energetische Charakterisierung unsinnig, da sich die Residuen lediglich geringfügig unterscheiden und die minimalen Unterschiede wahrscheinlich keine signifikante Bedeutung haben. Ausschließlich die sequenzielle sowie die räumliche Lage in dem Protein wird Anteil daran haben. Bei der Auszählung und Zuordnung nach den

Sekundärstrukturbereichen manifestierte sich jedoch eine Ausbildung besonders. Dies ist die Bindung zwischen zwei Cysteinen, welche sich in Coil-Strukturen befinden. *Tabelle 25* verdeutlicht das anzahlmäßige Auftreten aller Ausprägungen. Dabei wurde die Lage vom N- zum C-Terminus als Maß des Auszählens betrachtet. Die Sekundärstrukturen sind mit C für Coil, E für Sheet und H für Helix dargestellt.

Tabelle 25: *Die Ausbildung der Disulfidbrücken differenziert stark mit der Lage der Cysteinen in den Sekundärstrukturelementen. Dabei steht C für Coil, E für Sheet und H für Helix. Die häufigste Kombination ist Coil-Coil. Trotz des leicht vermehrten Auftretens von Cysteinen in Coil-Bereichen ist die Anzahl erstaunlich hoch.*

Sekundärstruktur-Komposition	Anzahl
CC	27
CH	15
CE	15
HH	4
HC	12
HE	14
EE	5
EC	5
EH	6

In den Coil-Strukturen treten die Cysteine zwar bevorzugt auf, jedoch nicht mit einer solchen Dominanz, dass es fast das Doppelte an Ausprägungen rechtfertigt. Auch die weiteren Ausprägungen mit einem Coil-Bereich zugehörigen Cystein an N-terminaler Position fallen hoch aus. Nur die Kombinationen Helix-Coil und Helix-Sheet können dazu annähernd aufschließen. Eine Helix-Helix-Kombination kommt im Gegensatz dazu sogar am seltensten vor. Jegliche Konformationen mit einem N-terminalen Cystein in einem Sheet sind auch nur außerordentlich anzutreffen. Eine biologische Begründung für diese Ausbildung ist schwer zu finden. Die Disulfidbrücken bilden sich meist direkt während des Translationsprozesses aus, sodass es sequenziell passieren könnte, dass sich schon zu Beginn mehr Cysteine in Coil-Strukturen befinden und somit bei der aufgeführten Betrachtung vom N- zum C-Terminus ein häufigeres Auftreten zu beobachten ist.

6 Ausblick

Die vorgenommene Aufgabe war auf Grund der Tatsache, verwertbare mathematische Werte für eine Regressionsanalyse zu gewinnen, nicht für jeden Teilbereich trivial. Im Laufe der Zeit und mit voranschreitenden Analysen wurde es immer komplexer und das Ermitteln qualitativ hochwertiger Daten gestaltete sich in der limitierten Dauer als anspruchsvoll. Bei der Disulfidbrücken-Analyse war es auf Grund einer mangelnden Anzahl an Daten nur schwer möglich, aussagekräftige Ergebnisse zu erzielen. Trotz allem sollte die Anzahl an erbrachten Daten und die Resultate jeglicher Berechnungen eine solide Grundlage für weitere Arbeitsschritte darstellen.

Die nun folgende Regressionsanalyse wird die Qualität der Daten erfassen, sodass eine Wichtung und Standardisierung der Deskriptoren festgestellt werden kann. Dabei ist nicht auszuschließen, dass die Ergebnisse unzureichend sind und neue bzw. mehr Deskriptoren gefunden werden müssen. Dazu muss allerdings höchste Akribie in den Details bewiesen werden und besonders die Proteinchemie Betrachtung finden. Dies setzt wahrscheinlich auch spezifische Analysemethoden voraus und evtl. müssen gezielte Experimente durchgeführt werden bzw. die Daten müssen aus fundierten Forschungseinrichtungen eingeholt werden, um die bestmögliche Qualität vorzufinden und diese auch in dem Modell zu gewährleisten.

Die Schaffung eines einheitlichen, nicht redundanten und dabei noch ausreichend großen Datensatzes ist weiterhin ein berechtigtes Ziel für folgende Analysen. Die Einträge in der *PDB* sind sehr oft lückenhaft bzw. fehlerbehaftet, sodass ein allgemeiner Parser sehr oft an die Grenzen stößt und viele Ausnahmen kalkuliert werden müssen. Alle Dateien manuell zu durchsuchen ist jedoch auch keine vorteilhafte Möglichkeit, da die Anzahl an Einträgen viel zu groß ist (Anzahl an *PDB*-Einträgen: 67131, Stand: 16.08.2010). Letztendlich wird es jedoch keine Umsetzung ohne einen korrekt verfassten Parser geben, wobei, gemessen an der Anzahl an heraus gefunden Proteinen, eine stichprobenhafte manuelle Untersuchung gemacht werden sollte. Dies wird mit einem hohen Arbeits- sowie Zeitaufwand verbunden sein, jedoch ist jegliche Grundlage einer perfekten Analyse ein korrekter Datensatz. Dieser muss nicht zwangsläufig einen riesigen Umfang haben, da es, wie bereits an früherer Stelle

erwähnt, teilweise sehr kompliziert wird mit der enormen Anzahl an Aminosäuren und den zugehörigen Messwerten umzugehen und meist zusätzlich verfasste Programme voraus setzt.

Die ausführliche Analyse der Torsionswinkel und die globale Stellung der Atome könnte ein wichtiger Deskriptor sein. Dazu sind bereits eine Vielzahl an Untersuchungen durchgeführt wurden und somit eine breite Masse an Literatur und Messwerten vorhanden. Leider ist diese Thematik enorm tiefgründig und für eine verständliche Vorstellung aller Änderungen müsste ein Programm einbezogen werden bzw. selbst verfasst werden. Dabei sollte anschließend jede Aminosäure explizit untersucht werden und bis ins kleinste Detail analysiert werden. Der globale Zusammenhang und Veränderungen der gesamten Struktur sind jedoch auch von essenzieller Wichtigkeit.

Anlagen

Die nachfolgenden Anlagen umfassen Tabellen, welche im Schriftteil erwähnt wurden, jedoch den laufenden Textfluss behindert hätten und vorrangig der Vollständigkeit dienen. In Tabellen, welche relative Häufigkeiten verdeutlichen, werden Positionen, die einen Wert von 0 haben, als leeres Kästchen gekennzeichnet.

A1 – Anlage zu Gliederungspunkt 5.2.1 – Energetische Charakterisierung

Tabelle 26 beschreibt die energetische Verteilung der 3D-Motive. Dabei wurde auch nach den differenzierten Ausprägungen des *betabulgeloop's*, der *niche* und des *ststaple* unterschieden.

Tabelle 26: *Die 3D-Motive in all ihren genauen energetischen Durchschnittswerten an jeder Position sind dargestellt. E-Value bezeichnet den Energiewert des gesamten Motivs im mathematischen Mittel.*

3D-Motiv	1. AS	2. AS	3. AS	4. AS	5. AS	6. AS	E-Value
betabulge	-14,27	-11,23					**-12,75**
stturn	-6,57	-5,59	-7,30				**-6,49**
stmotif	-5,07	-10,81	-5,55	-6,37	-17,66		**-9,09**
betaturn	-13,55	-9,24	-7,82	-9,90			**-10,13**
nest	-6,98	-6,37	-11,96				**-8,44**
asxmotif	-2,2	-7,25	-4,40	-11,65	-13,43		**-7,79**
asxturn	-2,34	-8,41	-4,79				**-5,18**
schellmannloop	-20,15	-12,16	-6,61	-9,23	-3,92	-15,10	**-11,20**
betabulgeloop (5)	-12,66	-2,58	-0,54	-4,92	-6,27		**-5,39**
betabulgeloop (6)	-9,82	-4,47	-2,87	-4,49	-7,38	-11,14	**-6,70**
gammaturn	-7,22	-14,41	-8,53				**-10,05**
niche (3)	-8,54	-6,06	-8,59				**-7,73**
niche (4)	-11,5	-7,27	-6,73	-14,51			**-10,00**
ststaple (5)	-13,31	-13,51	-13,22	-14,65	-9,24		**-12,79**
ststaple (6)	-22,45	-17,89	-9,24	-4,76	-3,31	-13,20	**-11,81**

alphabetamotif	-16,67	-12,43	-11,52	-11,89	-16,56		**-13,81**

A2 – Anlage zu Gliederungspunkt 5.2.1 – Energetische Charakterisierung

Tabelle 27 verdeutlicht die relativen Häufigkeiten für das Auftreten der Aminosäuren an jeder Position in den 3D-Motiven in Bezug auf die Ausprägung in einem der vier energetischen Quantile. Dabei werden nicht die Residuen als einzelnes betrachtet, sondern nur die jeweilige Position analysiert. Motive mit unterschiedlichen Ausprägungen sind mit der jeweiligen Anzahl an Residuen in Klammern gekennzeichnet. Die Aminosäuren sind von Position 1 bis 6 nummeriert dargestellt.

Tabelle 27: *Die relativen Häufigkeiten für das Auftreten der Aminosäuren an jeder Position in den 3D-Motiven werden gemäß ihrer energetischen Ausprägung in dem 1.-4. Quantil dargestellt. Dabei werden auch Motive mit variierenden Positionen extra betrachtet. Die jeweilige Anzahl an Residuen ist in Klammern dargestellt.*

3D-Motiv	Quantil	AS 1	AS 2	AS 3	AS 4	AS 5	AS 6
betabulge	**1. Q**	0,039	0,078				
	2. Q	0,294	0,392				
	3. Q	0,294	0,431				
	4. Q	0,373	0,098				
stturn	**1. Q**	0,115	0,431	0,466			
	2. Q	0,585	0,323	0,163			
	3. Q	0,290	o,145	0,293			
	4. Q		0,046	0,069			
stmotif	**1. Q**	0,247	0,258	0,473	0,355	0,258	
	2. Q	0,634	0,215	0,269	0,430	0,086	
	3. Q	0,118	0,376	0,215	0,194	0,172	
	4. Q		0,151	0,045	0,022	0,484	
betaturn	**1. Q**	0,239	0,352	0,372	0,282		
	2. Q	0,220	0,259	0,336	0,311		

	3. Q	0,219	0,229	0,179	0,216	
	4. Q	0,322	0,160	0,112	0,191	
nest	**1. Q**	0,386	0,388	0,203		
	2. Q	0,348	0,370	0,256		
	3. Q	0,161	0,185	0,293		
	4. Q	0,104	0,057	0,231		
asxmotif	**1. Q**	0,714	0,457	0,514	0,257	0,229
	2. Q	0,229	0,200	0,314	0,143	0,200
	3. Q	0,057	0,257	0,143	0,343	0,257
	4. Q		0,086	0,029	0,257	0,314
asxturn	**1. Q**	0,650	0,417	0,500		
	2. Q	0,325	0,208	0,300		
	3. Q	0,025	0,242	0,175		
	4. Q		0,133	0,025		
schellmannloop	**1. Q**	0,056	0,296	0,463	0,241	0,444	0,093
	2. Q	0,074	0,148	0,315	0,426	0,444	0,222
	3. Q	0,333	0,296	0,074	0,167	0,111	0,352
	4. Q	0,537	0,259	0,13	0,167		0,315
betabulgeloop (5)	**1. Q**	0,262	0,667	0,905	0,262	0,286	
	2. Q	0,119	0,238	0,095	0,571	0,548	
	3. Q	0,262	0,095		0,119	0,071	
	4. Q	0,357			0,048	0,095	
betabulgeloop (6)	**1. Q**	0,429	0,429	0,571	0,429	0,286	0,286
	2. Q		0,429	0,286	0,286	0,286	0,143
	3. Q	0,286	0,143	0,143	0,286	0,429	0,429
	4. Q	0,286					0,143
gammaturn	**1. Q**	0,273	0,273	0,182		
	2. Q	0,273	0,136	0,409		
	3. Q	0,409	0,136	0,273		
	4. Q	0,045	0,455	0,136		

niche (3)	**1. Q**	0,289	0,440	0,294		
	2. Q	0,344	0,280	0,303		
	3. Q	0,257	0,216	0,289		
	4. Q	0,110	0,064	0,115		
niche (4)	**1. Q**	0,232	0,396	0,391	0,156	
	2. Q	0,320	0,302	0,354	0,178	
	3. Q	0,213	0,212	0,169	0,338	
	4. Q	0,235	0,090	0,086	0,328	
ststaple (5)	**1. Q**	0,286	0,155	0,250	0,143	0,048
	2. Q	0,214	0,238	0,179	0,250	0,393
	3. Q	0,179	0,369	0,298	0,286	0,548
	4. Q	0,321	0,238	0,274	0,321	
ststaple (6)	**1. Q**		0,333	0,333	0,667	0,667	
	2. Q			0,333		0,333	
	3. Q	0,333			0,333		1,000
	4. Q	0,667	0,667	0,333			
alphabetamotif	**1. Q**	0,219	0,226	0,288	0,295	0,144
	2. Q	0,144	0,253	0,281	0,240	0,219
	3. Q	0,205	0,308	0,212	0,151	0,178
	4. Q	0,432	0,212	0,219	0,315	0,452

A3 - Anlage zu Gliederungspunkt 5.2.1 – Sequenzielle Charakterisierung

Tabelle 28 beschreibt die sequenzielle Verteilung der 3D-Motive, d.h. an jeder Position ist die relative Häufigkeit für das Auftreten der jeweils betrachteten Aminosäuren dargestellt. Eine besondere Ausnahme bildet dabei der *ststaple*, welcher in der Analyse mittels *PDBe-Motif* drei Ausprägungen mit einer Aminosäure vor der eigentlich ersten zeigte. Diese wurde als „Aminosäure 0“ aufgenommen und behandelt. Die Aminosäuren sind von Position 1 bis 6 nummeriert dargestellt.

Tabelle 28: *An jeder Position der 3D-Motive ist die relative Häufigkeit für die spezifisch betrachtete Aminosäure aufgeführt. Besonderheit bildet der ststaple, wobei in drei Ausprägungen eine Aminosäure vor der eigentlichen ersten existent ist. Dies wurde bei der Berechnung beachtet und verarbeitet.*

3D-Motiv	Aminosäure	AS 0	AS 1	AS 2	AS 3	AS 4	AS 5	AS 6
alphabetamotif	**A**		0,178	0,130	0,110	0,178	0,178	
	C		0,014			0,007	0,014	
	D		0,014	0,055	0,048	0,048	0,062	
	E		0,075	0,055	0,151	0,137	0,055	
	F		0,075	0,055	0,021	0,041	0,075	
	G		0,041	0,048	0,062	0,034	0,048	
	H		0,014	0,007		0,014	0,014	
	I		0,034	0,041	0,082	0,021	0,089	
	K		0,027	0,068	0,068	0,068	0,027	
	L		0,199	0,062	0,116	0,034	0,089	
	M		0,007	0,021		0,034	0,007	
	N		0,014	0,048	0,027	0,007	0,075	
	P		0,096	0,082	0,103			
	Q		0,014	0,158	0,103	0,089	0,041	
	R		0,014	0,034	0,034	0,082	0,027	
	S		0,082	0,089	0,027	0,089	0,041	
	T		0,014	0,007	0,007	0,014	0,014	
	V		0,034	0,021	0,014	0,075	0,123	
	W		0,014	0,021	0,021	0,014	0,007	
	Y		0,041		0,007	0,014	0,014	
asxmotif	**A**			0,057	0,143	0,171	0,057	
	C			0,029			0,057	
	D		0,743	0,114	0,057	0,114	0,029	
	E			0,086	0,114	0,057	0,029	
	F					0,057	0,029	
	G			0,029	0,057		0,029	
	H					0,029	0,086	
	I			0,086		0,057	0,029	
	K			0,086	0,086		0,057	
	L			0,029		0,086	0,114	
	M							

N	0,257	0,057	0,114	0,057	0,057
P		0,143			0,029
Q			0,057	0,029	0,086
R		0,029	0,086	0,086	0,114
S		0,086	0,114	0,114	
T		0,029	0,057		0,086
V		0,029	0,086	0,029	0,057
W		0,057		0,029	0,029
Y		0,057	0,029	0,086	0,029

asxturn

A		0,066	0,107
C		0,025	
D	0,595	0,008	0,091
E		0,058	0,074
F		0,033	0,041
G		0,099	0,083
H			0,017
I		0,017	
K		0,083	0,066
L		0,050	0,017
M		0,025	
N	0,405	0,025	0,083
P		0,215	
Q			0,025
R		0,025	0,132
S		0,107	0,116
T		0,033	0,050
V		0,050	0,033
W		0,033	0,008
Y		0,050	0,058

betabulge

A	0,039	0,294
C		0,039
D	0,157	
E	0,039	0,059
F	0,020	0,020
G	0,020	0,078

	H	0,020	0,235				
	I	0,137	0,020				
	K	0,020	0,020				
	L	0,137					
	M						
	N	0,020					
	P						
	Q		0,039				
	R	0,039					
	S		0,137				
	T	0,039	0,020				
	V	0,294					
	W	0,020					
	Y		0,039				
betabulgeloop	**A**	0,020	0,143	0,041	0,000	0,041	0,143
	C	0,224			0,020		
	D	0,082	0,082	0,265	0,041	0,020	0,143
	E		0,082	0,082		0,020	
	F					0,020	
	G	0,041	0,020	0,122	0,673	0,082	
	H	0,082				0,020	
	I	0,020			0,041	0,041	0,143
	K	0,102	0,306	0,020	0,102	0,102	
	L	0,102	0,020		0,041	0,041	
	M			0,020			
	N	0,122	0,020	0,286	0,020	0,061	
	P		0,143				
	Q		0,020	0,020	0,020	0,306	0,143
	R	0,020	0,041	0,061		0,163	
	S	0,061	0,082	0,041	0,041	0,041	0,143
	T	0,082		0,020		0,020	
	V	0,020				0,020	0,143
	W	0,020		0,020			
	Y		0,041				0,143
betaturn	**A**	0,096	0,138	0,110	0,094		

C	0,027	0,023	0,005	0,015
D	0,046	0,045	0,089	0,050
E	0,051	0,076	0,072	0,061
F	0,033	0,028	0,018	0,020
G	0,059	0,051	0,156	0,201
H	0,021	0,011	0,015	0,017
I	0,070	0,045	0,029	0,022
K	0,063	0,083	0,054	0,056
L	0,132	0,041	0,067	0,071
M	0,009	0,010	0,009	0,021
N	0,039	0,039	0,080	0,040
P	0,078	0,111	0,023	
Q	0,026	0,076	0,071	0,087
R	0,037	0,068	0,067	0,065
S	0,061	0,078	0,082	0,065
T	0,039	0,017	0,017	0,034
V	0,061	0,037	0,010	0,046
W	0,015	0,011	0,017	0,009
Y	0,039	0,012	0,010	0,028

gammaturn

A	0,261	0,130	0,174
C			0,043
D		0,174	
E	0,043	0,043	0,087
F	0,043		
G	0,217		0,130
H			
I		0,174	
K	0,217	0,043	0,043
L			0,130
M			
N		0,087	
P	0,043	0,043	
Q			0,043
R		0,043	0,261
S			0,043
T			

	V	0,087	0,043	0,043
	W			
	Y	0,087	0,217	

nest	**A**	0,027	0,049	0,062
	C	0,018	0,011	0,011
	D	0,095	0,024	0,044
	E	0,057	0,015	0,029
	F	0,027	0,015	0,055
	G	0,176	0,464	0,101
	H	0,022	0,020	0,016
	I	0,009	0,013	0,071
	K	0,044	0,059	0,068
	L	0,069	0,066	0,112
	M	0,015	0,009	0,009
	N	0,121	0,082	0,029
	P			
	Q	0,033	0,035	0,049
	R	0,106	0,059	0,059
	S	0,060	0,022	0,046
	T	0,059	0,015	0,102
	V	0,007	0,016	0,077
	W	0,027	0,009	0,024
	Y	0,027	0,018	0,037

niche	**A**	0,066	0,089	0,062	0,070
	C	0,021	0,009	0,014	0,045
	D	0,085	0,097	0,095	0,066
	E	0,047	0,064	0,048	0,044
	F	0,029	0,014	0,025	0,048
	G	0,115	0,048	0,158	0,028
	H	0,018	0,021	0,034	0,020
	I	0,054	0,027	0,027	0,070
	K	0,042	0,048	0,061	0,041
	L	0,086	0,064	0,033	0,113
	M	0,015	0,011	0,009	0,044
	N	0,045	0,043	0,103	0,025

	P	0,066	0,121	0,001		
	Q	0,042	0,057	0,032	0,056	
	R	0,055	0,056	0,051	0,037	
	S	0,080	0,096	0,119	0,070	
	T	0,059	0,064	0,055	0,064	
	V	0,033	0,039	0,029	0,088	
	W	0,016	0,010	0,010	0,023	
	Y	0,026	0,021	0,034	0,048	
schellmannloop	**A**	0,185	0,111	0,056	0,037	0,037	0,111
	C	0,037					
	D	0,037	0,019	0,037			0,019
	E		0,056	0,056	0,074		0,019
	F	0,019	0,019	0,019	0,037		0,056
	G	0,019	0,056	0,037	0,019	0,556	0,185
	H	0,019			0,019	0,019	0,019
	I	0,019	0,130	0,037			0,148
	K		0,222	0,130	0,074	0,074	0,056
	L	0,222	0,130	0,019	0,130	0,019	0,148
	M	0,074		0,037	0,019	0,037	0,019
	N	0,111		0,019	0,148	0,074	
	P	0,019					
	Q	0,019	0,056	0,167	0,130	0,056	0,037
	R	0,019	0,074	0,185	0,167	0,111	0,019
	S	0,037		0,074	0,037	0,019	0,019
	T	0,019	0,037	0,019	0,019		
	V	0,037	0,056	0,056	0,000		0,074
	W	0,037			0,037		0,037
	Y	0,074	0,037	0,056	0,056		0,037
stmotif	**A**		0,172	0,312	0,172	0,086
	C		0,022	0,011		0,011
	D		0,022	0,129	0,194	0,011
	E		0,043	0,215	0,151	0,032
	F		0,011			0,022
	G		0,043	0,043	0,032	0,022
	H		0,011	0,022	0,022	

	I		0,022		0,011	0,054
	K		0,011	0,011	0,032	0,215
	L		0,215	0,043	0,011	0,140
	M		0,011		0,011	0,065
	N		0,054	0,054	0,054	
	P		0,161			
	Q		0,043	0,022	0,075	0,022
	R		0,054	0,011	0,054	0,065
	S	0,559	0,032	0,086	0,086	
	T	0,441	0,022	0,011	0,043	
	V		0,043	0,022	0,032	0,204
	W		0,011	0,011		0,043
	Y				0,022	0,011
stturn	**A**		0,045	0,083		
	C		0,030	0,045		
	D		0,098	0,083		
	E		0,038	0,083		
	F		0,008	0,015		
	G		0,220	0,015		
	H		0,023	0,008		
	I		0,008	0,045		
	K		0,045	0,068		
	L		0,053	0,045		
	M		0,008			
	N		0,015	0,114		
	P		0,098			
	Q		0,061	0,038		
	R		0,023	0,053		
	S	0,644	0,121	0,152		
	T	0,356	0,061	0,053		
	V		0,030	0,045		
	W		0,008			
	Y		0,008	0,053		
ststaple	**A**	0,333	0,091	0,170	0,114	0,182	
	C		0,023	0,011		0,023	

D		0,080	0,011	0,080		
E		0,080	0,057	0,125	0,102	
F	0,333	0,023	0,023	0,011	0,011	
G		0,057	0,057	0,057	0,011	
H		0,011	0,045	0,045	0,011	
I		0,080	0,091	0,068	0,091	
K		0,034	0,068	0,023	0,034	
L	0,333	0,091	0,068	0,091	0,148	
M		0,011	0,023		0,011	
N		0,057	0,057	0,011	0,045	
P		0,057	0,011	0,023	0,011	
Q		0,034	0,045	0,068	0,023	
R		0,068	0,045	0,045	0,068	
S		0,023	0,034	0,057	0,057	0,284
T		0,045	0,114	0,057	0,057	0,716
V		0,102	0,011	0,045	0,045	
W		0,023	0,023	0,034	0,034	
Y		0,011	0,034	0,045	0,034	

A4 – Anlage zu Gliederungspunkt 5.2.2

Tabelle 29 verdeutlicht für alle PROSITE-Motive die sequenzielle Verteilung der Aminosäuren unter Angabe der relativen Häufigkeit. Die Aminosäuren sind von Position 1 bis 6 nummeriert dargestellt.

Tabelle 29: *Für jedes PROSITE-Motiv wurden an allen Stellen die relativen Häufigkeiten der Aminosäuren berechnet, welche alle existent heraus gefunden wurden.*

PROSITE-Motiv	Aminosäure	AS 1	AS 2	AS 3	AS 4	AS 5	AS 6
PS00001	**A**		0,092		0,039		
	C		0,026		0,013		
	D		0,079		0,039		
	E		0,053		0,053		
	F		0,026		0,013		
	G		0,092		0,053		

	H		0,013		0,013
	I		0,092		0,079
	K				0,211
	L		0,211		0,105
	M				
	N	1,000	0,013		0,026
	P				
	Q		0,013		
	R		0,026		0,026
	S		0,039	0,487	0,092
	T		0,026	0,513	0,039
	V		0,079		0,118
	W		0,039		0,026
	Y		0,039		0,053
PS00004	**A**			0,056	
	C				
	D			0,111	
	E			0,111	
	F			0,056	
	G			0,056	
	H				
	I			0,056	
	K	0,611	0,611		
	L			0,222	
	M				
	N				
	P			0,056	
	Q				
	R	0,389	0,389	0,056	
	S			0,056	0,556
	T				0,444
	V			0,056	
	W				
	Y			0,111	
PS00005	**A**		0,064		

C		0,064		
D		0,032		
E		0,024		
F		0,036		
G		0,104		
H		0,016		
I		0,068		
K		0,032	0,506	
L		0,096		
M		0,016		
N		0,028		
P		0,036		
Q		0,064		
R		0,044	0,494	
S	0,542	0,112		
T	0,458	0,040		
V		0,076		
W		0,016		
Y		0,036		

PS00006

A		0,036	0,134	
C		0,036	0,004	
D		0,029	0,036	0,543
E		0,033	0,138	0,457
F		0,033	0,022	
G		0,047	0,087	
H		0,011	0,029	
I		0,116	0,033	
K		0,043	0,051	
L		0,214	0,076	
M		0,007	0,007	
N		0,062	0,025	
P		0,058	0,051	
Q		0,011	0,029	
R		0,033	0,098	
S	0,543	0,051	0,022	
T	0,457	0,047	0,069	

	V		0,076	0,043		
	W		0,029	0,014		
	Y		0,029	0,033		
PS00008	**A**		0,155	0,088	0,088	0,307	0,127
	C		0,014	0,021	0,018	0,081	0,042
	D			0,067	0,057		0,042
	E			0,035	0,067		0,039
	F			0,021	0,039		0,032
	G	1,000	0,099	0,088	0,081	0,212	0,134
	H			0,011	0,025		0,011
	I		0,057	0,053	0,028		0,046
	K			0,046	0,032		0,057
	L		0,102	0,053	0,088		0,049
	M		0,021	0,014			0,018
	N		0,095	0,035	0,085	0,102	0,042
	P			0,042	0,046		
	Q		0,078	0,025	0,039		0,028
	R			0,039	0,057		0,039
	S		0,127	0,124	0,067	0,155	0,046
	T		0,117	0,141	0,049	0,141	0,057
	V		0,134	0,042	0,067		0,074
	W			0,018	0,011		0,011
	Y			0,035	0,057		0,106

A5 - Anlage zu Gliederungspunkt 5.3

Tabelle 30 zeigt die absolute Verteilung der Aminosäuren in den drei verschiedenen Sekundärstrukturen. Zusätzlich sind noch die Gesamtanzahl sowie die relative Häufigkeit im Vergleich zum gesamten Datensatz berechnet. Die graduelle Färbung innerhalb jeder Aminosäure verdeutlicht eine schnellere Übersicht, in welcher Sekundärstruktur jedes Residuum am häufigsten festgestellt wurde.

Tabelle 30: *Dargestellt ist die Verteilung jedes Residuums in den spezifischen Sekundärstrukturen mit Hilfe einer graduellen Färbung, um eine schnellere und präzisere Übersicht zu gewährleisten. Auch die Gesamtanzahl jeder Aminosäure sowie deren relative Häufigkeit zur Anzahl aller Residuen ist aufgeführt.*

Aminosäure	Coil	Helix	Sheet	Gesamtanzahl	Relative Häufigkeit
Ala	523	885	309	**1717**	**0,086**
Arg	396	481	231	**1108**	**0,056**
Asn	541	297	135	**973**	**0,049**
Asp	533	405	146	**1084**	**0,054**
Cys	168	127	123	**418**	**0,021**
Gln	266	388	145	**799**	**0,040**
Glu	338	617	216	**1171**	**0,059**
Gly	1015	319	282	**1616**	**0,081**
His	163	158	105	**426**	**0,021**
Ile	252	354	445	**1051**	**0,053**
Leu	454	712	412	**1578**	**0,079**
Lys	414	433	192	**1039**	**0,052**
Met	92	188	90	**370**	**0,019**
Phe	190	267	230	**687**	**0,035**
Pro	643	233	76	**952**	**0,048**
Ser	606	402	276	**1284**	**0,065**
Thr	547	323	336	**1206**	**0,061**
Trp	101	142	87	**330**	**0,017**
Tyr	263	203	269	**735**	**0,037**
Val	361	382	603	**1346**	**0,068**
				19890	

A6 - Anlage zu Gliederungspunkt 5.4 – Sequenzielle Nähe

Tabelle 31 zeigt die Präferenzen für alle Aminosäuren bei der Untersuchung der sequenziellen Nähe. Sie wurden mittels *Formel* (9) ermittelt.

Tabelle 31: *Präferenzen zur Beschreibung der sequenziellen Nähe aller Aminosäuren nach den Sekundärstrukturen.*

Aminosäure	Sekundärstruktur		
	Coil	**Helix**	**Sheet**
Ala	0,770	1,371	0,691
Arg	0,925	1,124	0,894
Asn	1,305	0,898	0,660
Asp	1,273	0,967	0,585
Cys	0,952	0,922	1,234
Gln	0,910	1,216	0,745
Glu	0,855	1,247	0,782
Gly	1,678	0,609	0,560
His	1,017	0,937	1,091
Ile	0,647	0,954	1,709
Leu	0,694	1,207	1,141
Lys	0,978	1,099	0,849
Met	0,736	1,281	0,925
Phe	0,794	0,930	1,495
Pro	1,763	0,592	0,441
Ser	1,143	0,967	0,814
Thr	1,077	0,818	1,215
Trp	0,876	0,986	1,244
Tyr	0,852	0,844	1,558
Val	0,669	0,843	1,882

A7 - Anlage zu Gliederungspunkt 5.4 – Sequenzielle Ferne

Tabelle 32 verdeutlicht die Präferenzen für alle Aminosäuren bei der Untersuchung der sequenziellen Ferne. Sie wurden mittels *Formel* (9) ermittelt.

Tabelle 32: *Präferenzen zur Beschreibung der sequenziellen Ferne aller Aminosäuren nach den Sekundärstrukturen.*

Aminosäure	Sekundärstruktur		
	Coil	Helix	Sheet
Ala	0,809	1,366	0,789
Arg	1,005	1,138	0,844
Asn	1,444	0,874	0,702
Asp	1,448	0,933	0,634
Cys	1,026	0,920	1,062
Gln	1,000	1,218	0,763
Glu	0,919	1,243	0,815
Gly	1,722	0,621	0,704
His	1,067	0,950	0,989
Ile	0,676	0,929	1,394
Leu	0,745	1,212	1,020
Lys	1,044	1,072	0,879
Met	0,794	1,360	0,810
Phe	0,839	0,957	1,205
Pro	2,068	0,555	0,437
Ser	1,206	0,914	0,891
Thr	1,116	0,790	1,115
Trp	0,899	1,074	1,018
Tyr	0,886	0,844	1,281
Val	0,672	0,816	1,521

A8 - Anlage zu Gliederungspunkt 5.5 – Der Phi-Winkel

Tabelle 33 beschreibt die Anzahl an Ausprägungen der Phi-Winkel aller Residuen des kleinen *Sets* bezogen auf Quantil-spezifische Ausprägung, Sekundärstruktur sowie nach der Unterscheidung in einen positiven oder negativen Winkel. Nach diesen Werten wurden alle Präferenzen berechnet.

Tabelle 33: *Anzahl an Ausprägungen der Phi-Winkel aller Residuen des kleinen Sets bezogen auf Quantil-spezifische Ausprägung, Sekundärstruktur sowie nach der Unterscheidung in einen positiven oder negativen Winkel.*

Sekundärstruktur	**Verteilung – φ – Winkel**			
	1. Quantil		**2. Quantil**	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	438	2003	508	1940
Helix	51	1610	72	1819
Sheet	16	437	44	795

	3. Quantil		**4. Quantil**	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	137	1531	15	887
Helix	31	1686	2	1949
Sheet	26	1354	0	2017

A9 - Anlage zu Gliederungspunkt 5.5 – Der Phi-Winkel

Tabelle 34 beschreibt die Anzahl an Ausprägungen der Psi-Winkel aller Residuen des kleinen *Sets* bezogen auf Quantil-spezifische Ausprägung, Sekundärstruktur sowie nach der Unterscheidung in einen positiven oder negativen Winkel. Nach diesen Werten wurden alle Präferenzen berechnet.

Tabelle 34: *Anzahl an Ausprägungen der Psi-Winkel aller Residuen des kleinen Sets bezogen auf Quantil-spezifische Ausprägung, Sekundärstruktur sowie nach der Unterscheidung in einen positiven oder negativen Winkel.*

Sekundärstruktur	**Verteilung – ψ – Winkel**			
	1. Quantil		**2. Quantil**	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	1533	908	1628	820
Helix	242	1420	327	1564
Sheet	412	42	760	79

	3. Quantil		**4. Quantil**	
	positive Winkel	negative Winkel	positive Winkel	negative Winkel
Coil	1283	385	730	172
Helix	194	1523	142	1809
Sheet	1261	119	1975	42

Literaturverzeichnis

[1] URL: http://www.peptide2.com/peptide/Amino_acid_wikipedia_the_free_files/300px-AminoAcidball.png, verfügbar am 16.06.2010

[2] Bauersachs, Guido. <kontakt@guidobauersachs.de>; URL: http://www.guidobauersachs.de/oc/dipep.gif, verfügbar am 16.06.2010

[3] Moss, G.P. <g.p.moss@qmul.ac.uk>; URL: http://www.chem.qmul.ac.uk/iubmb/enzyme/EC3/cont3bb.html, verfügbar am 17.06.2010

[4] Bauersachs, Guido. <kontakt@guidobauersachs.de>; URL: http://www.guidobauersachs.de/oc/hbruecke.gif, verfügbar am 17.06.2010

[5] info@zum.de; URL: http://www.zum.de/Faecher/Materialien/beck/bilder/helix.gif, verfügbar am 17.06.2010

[6] URL: http://www.mcat45.com/images/Beta-Sheets-MCAT.png, verfügbar am 18.06.2010

[7] Duschl, Albert. URL: http://www.uni-salzburg.at/pls/portal/docs/1/550696.PDF, verfügbar am 18.06.2010

[8] Dressel, Frank: *Sequenz, Energie, Struktur - Untersuchungen zur Beziehung zwischen Primär- und Tertiärstruktur in globulären und Membran-Proteinen.* April 2008. Dresden, Technische Universität, Mathematisch-Naturwissenschaftliche Fakultät, Dissertationsschrift, 2008

[9] Heinke, Florian: *Ausarbeitung zur Teilnahme an der Ausschreibung des Carl-Georg-Weitzel Preises 2010.* Juni 2010. Mittweida, Hochschule, Mathematisch-Naturwissenschaftliche-Informatische Fakultät

[10] Holtzhauer, Martin: *Methoden der Proteinanalytik.* – 1. Auflage – Heidelberg: Springer Verlag, 1996

[11] Draber, Wilfried; Fujita, Toshio: *Rational Approaches to Structure, Activity, and Ecotoxycology of Agrochemicals.* – 1. Auflage – Boca Ration, Florida: CRC Press, 1992

[12] Klebe, Gerhard: *Wirkstoffdesign: Entwurf und Wirkung von Arzneistoffen.* – 2. Auflage – Heidelberg: Spektrum Akademischer Verlag, 2009

[13] Govindarajan, Sridhar; Goldstein, Richard A.: *On the thermodynamic hypothesis of protein folding*, *PNAS 95*. May 1998. PNAS Biophysics. Paper

[14] Abraham, Donald J.: *Burger's Medicinal Chemistry and Drug Discovery.* – 6. Auflage – Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA, 2003

[15] Backhaus, Klaus; Erichson, Bernd; Plinke, Wulff; Weiber, Rolf: *Multivariate Analysemethoden – Eine anwendungsorientierte Einführung.* – 11. Auflage – Heidelberg: Springer-Verlag GmbH, 2006

[16] Merkel, Rainer; Waack, Stephan: *Bioinformatik interaktiv.* – 2. Auflage – Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA, 2009

[17] Löffler, Georg: *Basiswissen Biochemie mit Pathobiochemie.* – 6. Auflage – Heidelberg: Springer-Verlag GmbH, 2004

[18] Dose, Klaus: *Biochemie.* – 5. Auflage – Heidelberg: Springer-Verlag GmbH, 1994

[19] Müller-Esterl, Werner: *Biochemie – Eine Einführung für Mediziner und Naturwissenschaftler.* – 1. Auflage – Heidelberg: Spektrum Akademischer Verlag, 2004, korrigierter Nachdruck 2009

[20] Gibas, Cynthia; Jambeck, Per: *Einführung in die praktische Bioinformatik.* – 1. Auflage – Köln: O'Reilly Verlag GmbH & Co. KG, 2002

[21] ExPASy. <prosite@expasy.org>; URL: http://www.expasy.ch/prosite/, verfügbar am 10.07.2010

[22] Hansen, Andrea: *Bioinformatik – Ein Leitfaden für Naturwissenschaftler.* – 2. Auflage – Basel, Schweiz: Birkhäuser Verlag, 2004

[23] URL: http://www.rejuvanex.com/images/rejuv/amino%20acid%20cartoon.jpg, verfügbar am 12.07.2010

[24] Crooks, Gavin E.; Hon, Gary; Chandonia, John-Marc; Brenner, Steven E. <logo@compbio.berkeley.edu>; URL: http://weblogo.berkeley.edu/, verfügbar am 26.07.2010

[25] Potapov, Vladimir. <vladimir.potapov@weizmann.ac.il>; URL: http://ligin.weizmann.ac.il/cma/, verfügbar am 06.08.2010

[26] URL: http://www2.chemie.hu-berlin.de/vbk/lectures/mm2007-2008/v3.pdf, verfügbar am 15.08.2010

Danksagung

An erster Stelle gebührt mein Dank Prof. Dr. Dirk Labudde für seine Grundidee, eine Arbeit zu dieser Thematik zu verfassen. Auch seine Hilfe in schwierigen Momenten habe ich sehr zu schätzen gelernt. Desweiteren möchte ich alle Personen würdigen, welche mich unterstützt haben und bei komplizierten Fragestellungen stets wertvolle Ratschläge geben konnten. Auch bei programmiertechnischen Problemen wurde mir stets gern geholfen. Dabei verdienen insbesondere Florian Heinke und Stefan Schildbach meine höchste Wertschätzung.

Selbstständigkeitserklärung

Ich versichere die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt zu haben. Alle wörtlichen und sinngemäßen Entlehnungen sind unter genauer Angabe der Quelle kenntlich gemacht.

Mittweida, den 24.08.2010

(Eric Frenzel)