HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

# BACHELORARBEIT

Herr
**Christoph Leberecht**

## Konzeption und Implementierung einer Softwareanwendung zur Visualisierung evolutionär konservierter Kontaktinformationen, alpha-helikaler Membranproteine in einer 2,5d Darstellung

2013

**Fakultät Mathematik / Naturwissenschaften / Informatik**

# BACHELORARBEIT

# Konzeption und Implementierung einer Softwareanwendung zur Visualisierung evolutionär konservierter Kontaktinformationen, alpha-helikaler Membranproteine in einer 2,5d Darstellung

Autor:
**Christoph Leberecht**

Studiengang:
Biotechnologie und Bioinformatik

Seminargruppe:
BI10-w2b

Erstprüfer:
M. Sc. Steffen Grunert

Zweitprüfer:
Prof. Dr. rer. nat. Dirk Labudde

Mittweida, 20.08.2013

**Bibliografische Angaben**

Leberecht, Christoph: Konzeption und Implementierung einer Softwareanwendung zur Visualisierung evolutionär konservierter Kontaktinformationen, alpha-helikaler Membranproteine in einer 2,5d Darstellung, 47 Seiten, 16 Abbildungen, DVD, Hochschule Mittweida, University of Applied Sciences, Fakultät Mathematik / Naturwissenschaften / Informatik

Bachelorarbeit, 2013



Satz: LaTeX

**Referat**

Membranproteine sind essentiell für viele lebenswichtige Prozesse in allen Organismen. Es ist hilfreich ihre Struktur aufzuklären, um ihre Funktionen und die diesen zu Grunde liegenden Mechanismen zu verstehen. Untersuchungen haben gezeigt, dass es kurze Motive gibt, die in $\alpha$-helikalen transmembranen Proteinen allgegenwärtig sind. Es wird vermutet, dass diese Sequenzabschnitte elementare Eigenschaften besitzen, die die Stabilität und eventuell auch die Funktionen dieser Proteine ausmachen.

In dieser Arbeit wurden die Interaktionen und die Beschaffenheit dieser Motive untersucht. Auf der Grundlage verschiedener Datenbanken konnte ein Informationsalmanach erstellt werden, der relevante Informationen zu jedem Motiv enthält. Mit einer Applikation können diese Daten visualisiert werden. Die so generierte Darstellung ist eine Möglichkeit, Proteine auf Grundlage von Motiven genauer zu analysieren. Eine Untersuchung von Motiven aus verschiedenen Proteinfamilien erbrachte weitere Einblicke. Es hat sich gezeigt, dass es Motive gibt, die vorrangig für die Aufrechterhaltung der Struktur verantwortlich sind und solche, die für eine Funktion wichtig erscheinen. Es existieren Motive, die in allen Proteinfamilien vorkommen, jedoch in jeder eine veränderte Zusammensetzung zeigen. Trotzdem scheinen sie in allen Familien die gleichen Auswirkungen zu haben. Andere Motive sind nur in bestimmten Proteinfamilien hoch konserviert und wahrscheinlich für eine Funktion spezifisch. Auf dieser Grundlage wurde eine Methode entwickelt, die einen Vergleich zwischen den Familien ermöglicht.

# I. Inhaltsverzeichnis

**Anhang**

# II. Abbildungsverzeichnis

# III. Tabellenverzeichnis

# IV. Abkürzungsverzeichnis

| | |
|---|---|
| 1BRR | Bacteriorhodopsin |
| CAMs | Cell adhesion molecule |
| CSU | Contacts of Structural Units |
| DALI | Distance Alignment Matrix Method |
| GO | Gene Ontology |
| MSA | Multi Sequenz Alignments |
| PDBTM | Protein Data Bank of Transmembrane Proteins |
| Pfam | Protein Families Database |
| TMPad | Transmembrane Protein Helix-Packing Database |
| UPGMA | Unweighted Pair Group Method with Arithmetic Mean Algorithmus |

# V. Danksagung

Zunächst möchte ich mich an dieser Stelle bei all denjenigen bedanken, die mich während der Anfertigung dieser Bachelorarbeit unterstützt und motiviert haben.

Ganz besonders gilt mein Dank Herrn Steffen Grunert, der meine Arbeit und somit auch mich betreut hat. Ohne unseren regelmäßigen Austausch und seine moralische Unterstützung wäre mir die Anfertigung dieser Arbeit sicher nicht gelungen. Er bewegte mich dazu, vieles aus einem anderen Blickwinkel zu sehen und auch für seine unzähligen Anregungen bin ich ihm sehr dankbar.

Bei Professor Dirk Labudde möchte ich mich außerdem vielmals bedanken. Mein gesamtes Studium hindurch inspirierte er mich zu Höchstleistungen. Nicht nur seine thematische Kompetenz sondern auch Erfahrungen auf den verschiedensten Gebieten der Naturwissenschaften erwiesen sich als äußerst hilfreich.

Ein großes Dankeschön gilt meiner Freundin, Marlene Gruttke-Rölke, die mich während der Arbeit und besonders während des Endspurtes seelisch unterstützt hat. Ebenso möchte ich meinen Eltern und Großeltern herzlich für ihren Beistand und ihr Engagement danken.

# 1 Grundlagen

## 1.1 Transmembrane Proteine

Proteine werden auf Grundlage ihrer typischen Tertiärstruktur in drei große Gruppen eingeteilt. Neben Membranproteinen gibt es globuläre Proteine und fasrige Proteine. Nahezu alle globulären Proteine sind in Wasser löslich und fungieren als Enzyme zur Katalyse von Reaktionen. Dazu zählen z.B. Amylasen die Polysaccharide abbauen und Peptidasen die Peptide spalten können. Fasrige Proteine wie Collagen und Elastin dienen zur Aufrechterhaltung von Geweben. Membranproteine sind in eine Lipid-Doppelschicht eingebettet. Sie können sich dauerhaft innerhalb der Membran befinden oder nur temporär an der Membran angelagert sein. Die Membran muss dabei nicht die Zellmembran sein, es kann sich auch um die Membran anderer Zellorganellen handeln.

Membranproteine erfüllen eine große Bandbreite an Funktionen, die für das Leben aller Organismen unerlässlich sind. Eine Studie zeigte, dass Membranproteine der Wirkort von etwa 60 % der untersuchten Medikamente sind [2]. Membranrezeptoren regeln die Interaktion von Zellen mit ihrer Umwelt. Sie kommunizieren mit anderen Molekülen und übermitteln Reize ins Innere der Zelle. Transporter sind involviert in die Weiterleitung von Ionen, kleinen Molekülen und sogar Makromolekülen. Dabei lässt sich aktiver Transport, bei dem Energie verbraucht wird, und passiver Transport unterscheiden. Weiterhin gibt es viele membrangekoppelte Enzyme. Beispielhaft stehen dafür ATPasen wie die $Na^+/K^+$-ATPase, die mitochondrialen Systeme für den Elektronentransfer und die Reaktionszentren bei der Photosynthese. Zelladhäsionsmoleküle (auch CAMs, englisch für cell adhesion molecule) sind integrale Membranproteine, die zum Zusammenhalt von Geweben und zur Kommunikation zwischen Zellen beitragen.

Aufgrund ihrer 3d-Struktur und Lage in der Membran lassen sich Membranproteine wiederum in verschiedene Gruppen einteilen (Abb. 1.1). Dabei werden integrale und membranständige Membranproteine unterschieden. Außerdem kann nach Anzahl der transmembranen Segmente sowie deren Orientierung innerhalb der Membran klassifiziert werden. Membranständige oder periphere Membranproteine sind nur temporär an eine Membran gebunden. Sie können mit Hilfe von nicht-kovalenten Bindungen wie hydrophoben oder elektrostatischen Wechselwirkungen in der Membran gehalten werden, oder kovalent an ein Lipid in der Membran gekoppelt sein. Die Koppelung an die Membran kann unter anderem dazu dienen, eine Konformationsänderung hervorzurufen und somit die Funktion des Proteins zu aktivieren [3]. Integrale Membranproteine oder auch transmembrane Proteine sind permanent in der Membran gebunden. Von transmembranen Proteinen lassen sich zwei Untergruppen bilden, zum einen die $\beta$-Barrel und zum anderen die $\alpha$-helikalen transmembranen Proteine. Die Art der $\beta$-Barrel Proteine bildet Poren innerhalb der Membran. Prinzipiell durchspannt eine große zylinderför-

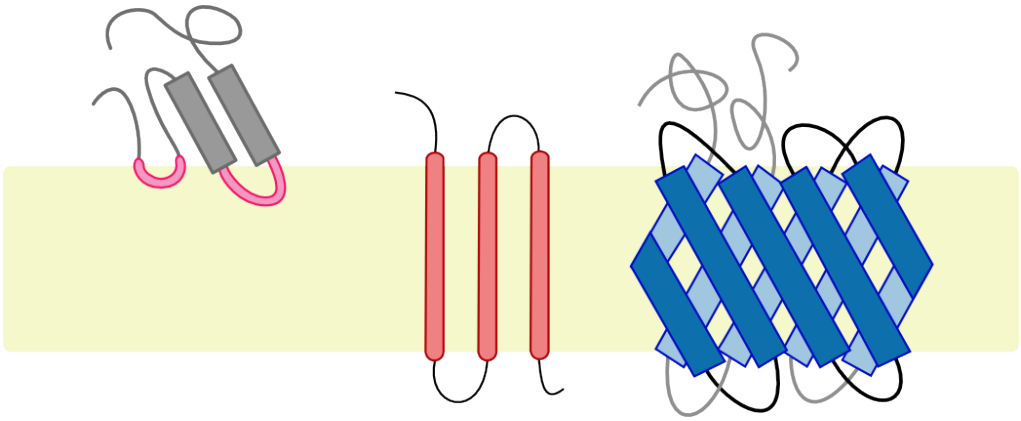

Abbildung 1.1: Schematische Abbildung der Einteilung von Membranproteinen nach deren Struktur: Membranständige Proteine (links), $\alpha$-helikale transmembrane Proteine (mittig) und $\beta$-Barrel Proteine (rechts).

mige Faltblattstruktur die gesamte Lipid-Doppelschicht. Die hydrophilen Aminosäure-Seitenketten sind meistens auf der Innenseite des Zylinders und bilden den wasserdurchlässigen Kanal. Hydrophobe Seitenketten befinden sich auf der Zylinderaußenseite und haben Kontakt zu hydrophoben Bereichen der Membran. Die einzelnen Stränge des Faltblattes sind durch von Wasserstoffbrücken miteinander verbunden [4].

Helikale Sekundärstrukturelemente machen den größten Teil der membrandurchspannenden Bereiche von Membranproteinen [5] aus. Proteine mit einem transmembranen $\alpha$-helikalen Bereich werden zur Gruppe der $\alpha$-helikalen transmembranen Proteine zusammengefasst. Die helikalen Bereiche bestehen im Durchschnitt aus 26 Aminosäuren mit hydrophobem und unpolarem Charakter. Besonders häufig treten dabei Alanin, Leucin und Glycin auf [6]. Durch Interaktionen zwischen den einzelnen Aminosäuren und der lipophilen Umgebung neigen die Helices dazu, eng gepackt senkrecht nebeneinander in der Lipid-Doppelschicht zu liegen. In wasserlöslichen Proteinen wird eine parallele Anordnung der Helices bevorzugt, wohingegen in transmembranen Helixkomplexen eine anti-parallele Anordnung vorgezogen wird (Abb. 1.2). Der Winkel in dem benachbarte Helices zueinander liegen, wird Kreuzungswinkel genannt. Dieser liegt in transmembranen Bereichen im Durchschnitt zwischen $0^\circ$ und $+30^\circ$. In transmembranen Proteinen liegen nur etwa 10 % der Winkel außerhalb dieses Bereiches. Dies beruht auf der geringeren Anzahl an Möglichkeiten zur Anordnung innerhalb der Membran. Bei globulären Proteinen hingegen liegt der Kreuzungswinkel im Schnitt bei $-35^\circ$ und ist gleichmäßiger über das gesamte Spektrum verteilt [6]. Weiterhin sind die Abstände zwischen den Helices transmembraner Proteine geringer. Der durchschnittliche Abstand zweier C$\alpha$-Atome liegt für transmembrane Helices bei 5,5 $\pm$ 1,1 Å. Bei nicht transmembranen Helix-Paaren sind es 6,0 $\pm$ 1,1 Å. [7]

Trotz ihrer großen Vielfalt sind nur etwas über 2 % der strukturell aufgeklärten Proteine auf der Proteindatenbank (PDB) Membranproteine [8]. Die experimentelle Aufklärung von transmembranen Proteinen wird erschwert durch die vielen hydrophoben Amino-

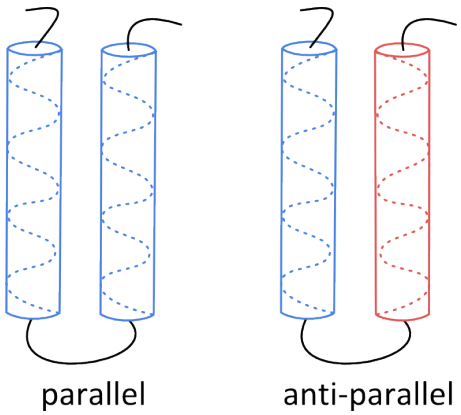


Abbildung 1.2: Schematische Darstellung der parallelen (links) und anti-parallelen Anordnung (rechts) von Helices.

säuren in den transmembranen Bereichen. Bei dem vorwiegend verwendeten Verfahren zur Aufklärung von Proteinstrukturen, der Kristallstrukturanalyse, muss das Protein in einer wässrigen Lösung auskristallisieren. Die hydrophoben Aminosäuren, die normalerweise mit der Lipid-Doppelschicht interagieren, werden durch diesen Schritt ins Zentrum des Proteins gezwungen. Dadurch faltet sich das Membranprotein oft zu einem unauflösbaren Aggregat [9].

Es wurde versucht, die geringe Anzahl an experimentellen Daten mit einer Fülle an theoretisch ermittelten Daten zu ergänzen. Vom theoretischen Standpunkt aus sind transmembrane Proteine leichter zu untersuchen. Sie besitzen strukturell gesehen eine geringere Vielfalt als wasserlösliche Proteine, was statistische Analysen vereinfacht. Tatsächlich gibt es viele genaue Algorithmen zur Vorhersage von helikalen Bereichen und helikalen Interaktionen. Der SOSUI Algorithmus beispielsweise beruht auf der Hydrophobizität von Aminosäuren. Er wurde an der Universität Tokyo entwickelt und konnte in einem Evaluierungsdatensatz etwa 97 % der transmembranen helikalen Bereiche sicher vorhersagen [10]. Ein Hydrophobizitätsdiagramm des membranständigen Breast Cancer Resistance Proteins mit vorhergesagter 2d-Struktur ist in der Abbildung 1.3 zu sehen. HMMTOP ist ein auf einem Hidden-Markov-Modell beruhender Algorithmus, der in 89 % der betrachteten Proteine alle Helices richtig vorhersagt [11]. Zur Prognose von Interaktionen zwischen Helices wurde 2009 ein Algorithmus entwickelt. Mit Hilfe einer Support Vector Machine können basierend auf Kontakten von Aminosäuren Helix Interaktionen vorhergesagt werden [12]. Auch viele Datenbanken beschäftigen sich mit den bereits vorhandenen Daten zu transmembranen Proteinen. Auf der Protein Data Bank of Transmembrane Proteins (PDBTM) befinden sich experimentell ermittelte 3d-Strukturen von transmembranen Proteinen [13]. Die TMPAD beinhaltet unter anderem Informationen über Helix-Helix Interaktionen, Winkel und Abstände zwischen Helices [14].

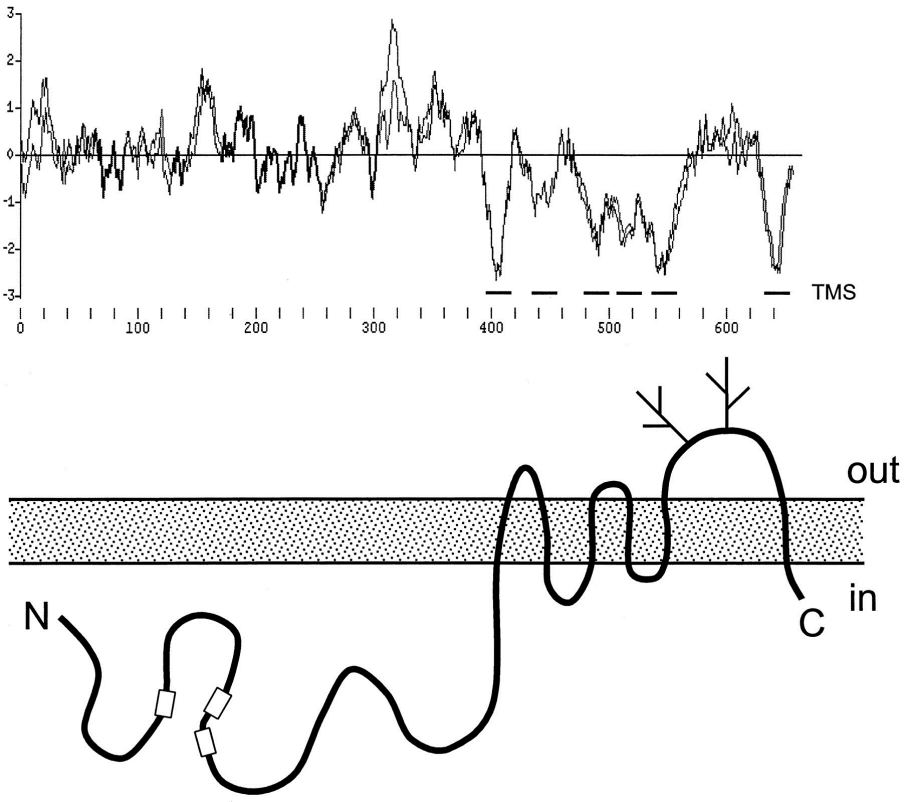


Abbildung 1.3: Der obere Bereich zeigt das Hydrophobizitäts-Diagramm von Breast Cancer Resistance Protein (BCRP/ABCG2). Die x-Achse zeigt die Aminosäureposition und die y-Achse die Hydrophobizität des Bereiches. Unterstrichen sind die etwaigen Bereiche der transmembranen Helices. Im unteren Bereich ist eine hypotetische 2d-Struktur des Proteins zu sehen. Modifiziert nach [1].

## 1.2 Interaktionen und Faltung von transmembranen Proteinen

Zwei Aminosäuren stehen in Kontakt zu einander, wenn der kürzeste Abstand zwischen ihren Atomen einen festgelegten Grenzwert unterschreitet [15]. Einige andere Definitionen beinhalten, dass die Summe der Van-der-Waals Radien zweier Atome als Abstandskriterium gilt [16]. Um allerdings zu bestimmen, welcher Wechselwirkungs- oder Kontakttyp vorliegt, müssen weitere Informationen einbezogen werden. Eine Vorhersage der Kontakttypen wurde im Jahr 1999 von Sobolev et al. [17] umgesetzt. Als Kontakttypen wurden dabei Wasserstoffbrücken, Stapelwechselwirkung (pi stacking interaction), hydrophobe Interaktion sowie destabilisierende hydrophobe-hydrophile Wechselwirkung berücksichtigt.

Wasserstoffbrücken liegen vor, wenn ein Protonendonor und ein Protonenakzeptor miteinander in Kontakt stehen. Der Donor ist dabei ein elektronegatives Atom, meistens Sauerstoff, an dem ein Wasserstoffatom kovalent gebunden ist. Ein anderes Atom mit freien Elektronenpaaren ist der Akzeptor. Es bildet sich eine elektrostatische Dipol-Dipol Wechselwirkung aus [18]. Wasserstoffbrückenbindungen sind eine Form der Nebenvalenzbindungen, die erheblich schwächer sind als kovalente Bindungen - in Proteinen überschreiten sie 17 $\frac{kJ}{mol}$ nur selten [19].

Stapelwechselwirkungen oder $\pi$-$\pi$-Wechselwirkungen bezeichnen nicht kovalente Interaktionen zwischen aromatischen Ringen. Durch delokalisierte Elektronen entsteht eine Polarisation des aromatischen Ringes. Innerhalb des Systems ist er negativ, außerhalb positiv polarisiert. Analysen der aromatischen Aminosäuren Phenylalanin, Tyrosin, Histidin und Tryptophan zeigten, dass Dimere derer Seitenketten eine stabilisierende Wirkung auf die Struktur von Proteinen haben. Diese Wechselwirkung wirkt über eine größere Distanz als Van-der-Waals Kräfte [20]. In Benzolringen wurden Bindungen mit Stärken zwischen 8 und 12 $\frac{kJ}{mol}$ gemessen, die über einen Abstand von 4,96 Å wirken . [21]

Als hydrophobe Interaktion wird ein Entropie getriebener Prozess bezeichnet, der die Bildung von Wasserstoffbrücken zu unpolaren Molekülen verhindert. Durch diese Unterbrechung wird die Bewegung der Moleküle der hydrophilen und der hydophoben Phase eingeschränkt. Aufgrund der Einschränkung lagern sich die unpolaren Moleküle durch hydrophobe Assoziation näher zusammen und die Gesamtoberfläche verringert sich. Die hydrophobe Wechselwirkung funktioniert indirekt dadurch, dass auf Grund der verringerten Oberfläche, zusammen mit einer Verringerung der möglichen Anordnungen der Moleküle, günstigere Bindungen eingegangen werden [18].

Die treibende Kraft bei der Faltung von nicht-transmembranen Proteinen ist das Zusammenlagern der hydrophoben Seitenketten, die einer wässrigen Lösung ausgesetzt sind. Dieser Prozess wird als hydrophober Kollaps bezeichnet [22]. Bei transmembra-

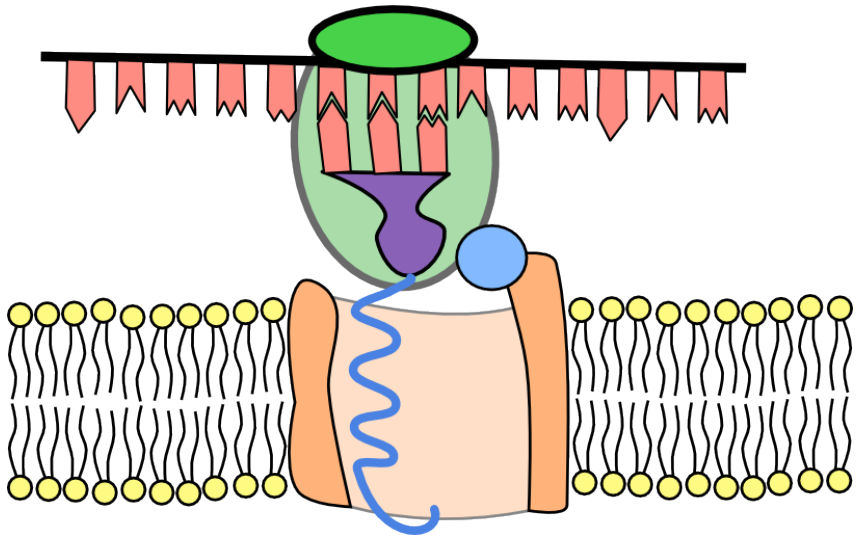

Abbildung 1.4: Bei der Faltung an Translokationsporen lagert sich das Ribosom (grün) mit der zu übersetzenden RNA (rot) an eine Translokationspore (orange) an. Die Translation wird durchlaufen und die Helices falten sich in der Pore. Nachdem sich die Helices ausgebildet haben, löst sich die Pore aus der Membran. Nun lagern sich die Helices zu einem Komplex zusammen.

nen Proteinen ergeben sich Probleme bei der Faltung in wässrigem Milieu. Die Ausbildung eines hydrophoben Inneren kann bei der Faltung von Membranproteinen keine treibende Kraft darstellen, da ein großer Teil der ”Hülle” des Proteins in einer Lipid-Doppelschicht liegt. Dieses Problem wird durch sogenannte Translokationsporen gelöst (Abb. 1.4). Das Ribosom dockt bei der Translation an diese Poren, die einen Tunnel in der Membran aufspannen, an und das Protein faltet sich bereits im Inneren der Membran. Nach der Translation löst sich das Ribosom und die Translokationspore verlässt die Membran [23]. Da sich nun kein hydrophober Kern bilden kann, spielt sich ein anderes Prinzip der Proteinfaltung ab. Es wird vermutet, dass sich zuerst die transmembranen Helices ausbilden, die sich anschließend durch die bereits beschriebenen Wechselwirkungen zwischen Aminosäuren zusammen lagern. Dieser Prozess wird als Zwei-Stufen Modell (Two stage model) bezeichnet [24]. Das Modell wurde 2003 erweitert zu einem Drei-Stufen-Prozess, in dem auf die ersten beiden Schritte die Einlagerung von Liganden folgt [25]. Es lässt sich also zeigen, dass die Interaktion zwischen Aminosäuren, in Bezug auf die Faltung und Stabilisierung transmembraner Proteine, eine große Rolle spielt.

## 1.3 Motive in transmembranen Helices

Membranproteindomänen wurden auf Grundlage ihrer Sequenz- und Topologieähnlichkeiten von der Pfamdatenbank klassifiziert [26]. Durch Studien dieser Proteinfamilien wurden interessante Kompositionen von Aminosäuren innerhalb der transmembranen Helices entdeckt. Es zeigte sich eine Tendenz dafür, dass Cystein, Threonin und Tryp-

tophan häufig in unmittelbarer Nähe voneinander liegen [27]. Besonderes Augenmerk wurde auf das Vorkommen von Glycinen gelegt, die im Abstand von drei Aminosäuren voneinander vorkommen. Dieses GxxxG Motiv wurde vor allem mit der Dimerisation von Glycophorin in Verbindung gebracht [28]. Die Studie wurde im Jahr 2000 fortgeführt und mit Hilfe der SwissProt Datenbank untersucht. Es zeigte sich, dass das GxxxG Motiv im untersuchten Datensatz 32% häufiger vorkam als in einer zufälligen Verteilung angenommen. Zusätzlich wurde ein starker Bezug dieser Motive zu in benachbarten Helices vorkommenden Isoleucin und Valin Resten beobachtet. Weiterhin waren alle Kombinationen von Paaren kleiner Aminosäuren (Glycin, Alanin, Leucin) überrepräsentiert [29]. Eine besonders hohe Konserviertheit dieser Motive wurde in Proteinfamilien gezeigt, die als Transporter, Symporter und Kanäle klassifiziert sind [26]. Es wurde weiterhin beobachtet, dass große Aminosäuren wie Phenylalanin, Tryptophan und Histidin eine höhere Wahrscheinlichkeit haben in transmembranen Taschen aufzutauchen als kleinere Aminosäuren [30].

# 2 Zielstellung

Das primäre Ziel dieser Arbeit soll sein, eine Visualisierung der Interaktionen zwischen Helices in transmembranen Proteinen zu ermöglichen. Dies erfolgt mit Hilfe von Motiven mit definierten Start- und Endaminosäuren. Auf Grundlage der vorhandenen Interaktionsdaten von der ”TransMembrane Protein Helix-Packing Database” (TMPad) [14] sowie Informationen, die mithilfe der Software ”Contacts of Structeral Units” (CSU) [17] berechnet wurden, sollen diese Motive genauer definiert werden. Zusätzliche Informationen wird der Bereich zwischen den Start- und Endaminosäuren liefern, der in bisherigen Veröffentlichungen als eher unbeachtet blieb. Der Bereich der zu betrachtenden Motive soll allerdings eingeschränkt werden. Es sollen nur Motive berücksichtigt werden, die sich in der Nähe eines Kontaktes zweier Helices befinden.

Aus diesen Daten soll zuerst ein Informationsalmanach erstellt werden, der die folgenden Werte der Kontaktmotive enthält. Es sollen die Konserviertheit und die Kontaktwahrscheinlichkeit der Aminosäuren sowie deren Partnermotive festgehalten werden. Außerdem soll der Bindungstyp, durch den diese Kontakte aufrechterhalten werden, in den Daten erhalten bleiben. Weiterhin sollen diese Informationen nicht nur für die Gesamtheit der transmembranen Proteine berechnet werden, sondern auch für ausgewählte Proteinfamilien. Mithilfe dieser Daten soll ein Schema erstellt werden, das die Interaktion von ausgewählten Motiven darstellen kann. Die relevanten Informationen zu den dargestellten Motiven sollen dabei übersichtlich angeordnet werden.

In einem weiteren Teil der Arbeit soll versucht werden, mit den generierten Motivdaten die Ähnlichkeit zweier Proteinfamilien zu bestimmen. Als Grundlage dafür soll die Klassifizierung der Proteinfamilien von der Pfam Datenbank [31] dienen. Diese Ähnlichkeitsbeziehung beruht dabei nicht auf der Ähnlichkeit von Sequenzen wie bei Multi Sequenz Alignments (MSA), sondern auf der Ähnlichkeit vom Kontaktmotiven, die in den jeweiligen Familien für die Interaktion zwischen Helices verantwortlich sind. Dies könnte neue Einblicke gewähren in die strukturelle Konserviertheit und Evolution von Proteinen.

# 3 Methoden

## 3.1 Datenaggregation

Die Analysen dieser Arbeit beruhen auf Daten, die von verschiedenen Datenbanken zusammengeführt wurden. Allgemeine Sequenzdaten, wie die Sequenz und die Annotation von Sekundärstrukturelementen wurden von der TMPad Datenbank bezogen. Des Weiteren wurden die Interaktionsdaten dieser Datenbank genutzt, um Aminosäuren zu kennzeichnen, die Interaktionen zu anderen Aminosäuren in anderen Helices eingehen. Diese Kontakte wurden mit Informationen über die Distanz der Atome und den Bindungstyp aus der CSU Software ergänzt. Von der Pfam Datenbank wurde erfasst, welcher Bereich der Sequenz für welche Proteinfamilie charakteristisch ist.

Der grundlegende Datensatz umfasst 1.107 transmembrane $\alpha$-helikale Proteine, mit insgesamt 4.061 Ketten und 17.413 Helix-Helix Interaktionen. Dieser wurde in zwei Clustern redundanter und nicht-redundanter Proteine eingeteilt. Im ersten redundanten Cluster befinden sich alle 1.107 Proteine. Der zweite Cluster mit nicht redundanten Daten beinhaltet ausschließlich Proteine, deren Sequenzen mit Hilfe des CD-HIT Algorithmus [32] ausgewählt wurden. Dabei wurden die Wortlänge auf 2 und die prozentuale Identität auf 40% festgelegt. Sequenzen mit weniger als 40 Aminosäuren wurden verworfen. Dieselben Parameter fanden auch beim Erstellen nicht redundanter Cluster auf der PDBTM Verwendung [33].

Die resultierenden Proteindaten wurden in einem XML-Format abgelegt. In diesem Format sind alle Eigenschaften der Proteine gespeichert. Die später daraus berechneten Werte zu den Motiven wurden ebenfalls im XML-Format festgehalten.

## 3.2 Analyse der Motive

Zuerst wurden Wörter bzw. Teilsequenzen um die Positionen der Aminosäuresequenz extrahiert, an der ein Kontakt zwischen zwei Helices vorhergesagt wurde. Es fanden alle möglichen Subwörter der Längen drei bis zehn Berücksichtigung. Weiterhin wurde bei jedem Subwort die Position der interagierenden Aminosäure vermerkt.

Im Anschluss erfolgte die Einteilung der Wörter in verschiedene Gruppen. Diese Gruppen wurden auf Grundlage der Länge der Sequenz sowie der Start- und Endaminosäure erstellt. Eine Aminosäure $A_S$ wird dabei von $n$ Aminosäuren von der Aminosäure $A_E$ getrennt. Aufgrund der zuvor genannten Bedingung gilt $2 \geq n \leq 10$ für alle Sequenzen die analysiert wurden. Die Bezeichnung einer Gruppe bzw. eines Motivs ergibt sich aus $A_S A_E(n+1)$. Die Gesamtlänge $l$ eines Motivs, d.h. die Anzahl der Elemente aus

denen das Wort besteht, ist demzufolge $n+2$. Beispielsweise wurden alle Sequenzen der Form LxxxG der Gruppe LG4 zugeteilt, hierbei stehen die ”x” für die variablen Aminosäuren. Die Gesamtheit der ”Platzhalter”-Aminosäuren ist der *variable Bereich* eines Motivs. Diese Gruppen werden als Mengen von Strings behandelt und mit $M$ bezeichnet. Bei den einzelnen Sequenzen innerhalb dieser Gruppen handelt es sich um *Ausprägungen* $m \in M$ des jeweiligen Motivs. Jede dieser Gruppen wurde anschließend einem Analyseverfahren zu Bestimmung von Konserviertheit und Kontaktwahrscheinlichkeit unterzogen.

### 3.2.1 Analyse der Konserviertheit

Die Konserviertheit $Kons(a,i,M)$ gibt an, wie hoch die Wahrscheinlichkeit ist, eine Aminosäure $a$ an einer bestimmten Position $i \in [1,n]$ im variablen Bereich eines Motivs $M$ anzutreffen. Die Konserviertheit einer Position ist immer mit dem natürliche Vorkommen einer Aminosäure im transmembranen Bereich gewichtet. Zur Berechnung der Konserviertheit dient ein Quotenverhältnis (auch Odds Ratio) bei dem zwei Odds miteinander verglichen werden. Beim Quotenverhältnis handelt es sich um eine statistische Maßzahl, die über den Zusammenhang zweier Merkmale Auskunft gibt. Ein Quotenverhältnis von 1 bedeutet, dass es keinen Unterschied zwischen den Odds gibt. Bei einem Quotenverhältnis, das größer als 1 ist, sind die Odds der ersten Gruppe größer, bei einem Ergebnis unter 1 sind die Odds der ersten Gruppe kleiner. Dieses Verfahren wurde gewählt, um die natürlichen Schwankungen der Vorkommen der Aminosäuren auszugleichen. Die Konserviertheit der Start- und Endaminosäuren ist aufgrund der Klassifikation der Motive $A_S = A_E = \infty$. Die Aminosäuren $a$ sind Elemente der Menge $A$, die alle 20 kanonischen Aminosäuren enthält. Die Wahrscheinlichkeit $P(a,i,M)$:

$$P(a,i,M) = \frac{\sum_{m \in M} f(a,i,m)}{|M|} \tag{3.1}$$

mit

$$f(a,i,m) = \begin{cases} 1 & \text{wenn } i \text{ in } m \text{ ist } a \\ 0 & \text{sonst,} \end{cases} \tag{3.2}$$

gibt an, wie hoch die Wahrscheinlichkeit ist, eine Aminosäure $a$ an der Position $i$ im Motiv $M$ vorzufinden. Die bedingten Wahrscheinlichkeiten für das natürliche Vorkommen $P(a|Natur)$ ist in der Tabelle A.1 einzusehen. Diese Werte wurden von der statistischen Übersicht der TMPad [14] übernommen. Aus den Wahrscheinlichkeiten ergeben sich die Odds

$$R(a,i,M) = \frac{P(a,i,M)}{1 - P(a,i,M)} \tag{3.3}$$

und

$$R(a|Natur) = \frac{P(a|Natur)}{1 - P(a|Natur)}. \tag{3.4}$$

Das Quotenverhältnis $R(a,i,M:a|Natur)$ ergibt sich aus

$$R(a,i,M:a|Natur) = \frac{R(a,i,M)}{R(a|Natur)}. \tag{3.5}$$

Die Konserviertheit einer Aminosäure $a$, an der Position $i$ im Motiv $m$, wird demnach mit Hilfe von

$$Kons(a,i,M) = R(a,i,M:a|Natur) = \frac{\frac{P(a,i,M)}{1-P(a,i,M)}}{\frac{P(a|Natur)}{1-P(a|Natur)}} = \frac{P(a,i,M)\cdot(1-P(a|Natur))}{P(a|Natur)\cdot(1-P(a,i,M))} \tag{3.6}$$

berechnet.

Die Gesamtkonserviertheit $Kons(i,M)$, einer Position $i$, ergibt sich aus der Summe der Konserviertheiten an dieser Position

$$Kons(i,M) = \sum_{a\in A} Kons(a,i,M). \tag{3.7}$$

### 3.2.2 Analyse der Kontaktpositionen

Bei der Kontaktwahrscheinlichkeit $Kont(i,M)$ handelt es sich um die Wahrscheinlichkeit, dass eine bestimmte Position $i$ im Motiv $M$ einen Kontakt zu einem anderen Motiv aufbaut. Die Menge $M$ enthält alle Ausprägungen eines Motivs. Die Position $i \in [0,n+1]$ geht eine Bindung zu einem anderen Motiv ein, wenn die Kontaktposition $pos_K = i$. Die Wahrscheinlichkeit $P(pos_K,i,M)$:

$$Kont(i,M) = P(pos_K,i,M) = \frac{\sum_{m\in M} g(pos_K,i,m)}{|M|} \tag{3.8}$$

mit

$$g(pos_K,i,m) = \begin{cases} 1 & \text{wenn } i = pos_K \text{ in } m \\ 0 & \text{sonst,} \end{cases} \tag{3.9}$$

gibt dies an.

Bei der aminosäurenspezifischen Kontaktwahrscheinlichkeit wird die Wahrscheinlichkeit nicht nur für eine Position, sondern für jede Aminosäure an dieser Position berechnet. Die aminosäurenspezifische Kontaktwahrscheinlichkeit wird demzufolge äquivalent mit

$$Kont(a,i,M) = P(a,pos_K,i,M) = \frac{\sum_{m\in M} h(a,pos_K,i,m)}{|M|} \tag{3.10}$$

mit

$$h(a,pos_K,i,m) = \begin{cases} 1 & \text{wenn } i = pos_K \text{ in } m, \text{ und} \\ & \text{wenn } i \text{ in } m \text{ ist } a \\ 0 & \text{sonst,} \end{cases} \tag{3.11}$$

berechnet. Die Aminosäuren $a$ sind wiederum Elemente der Menge $A$, die alle 20 kanonischen Aminosäuren enthält. Da jede aminosäurenspezifische Kontaktwahrscheinlichkeit an den gesamten Kontakten gemessen wird, gilt $\sum_{a\in A}\sum_{i=0}^{n+1} P(a,pos_K,i,M) = 1$.

### 3.2.3 Analyse der Bindungsarten

Das Bindungsprofil enthält Daten über die Art der Bindung, die eine bestimmte Position des Motivs eingeht. Die Bindungsarten sind beschränkt auf hydrophob-hobdrophob, hydrophob-hydrophil, hydrophil-hydrophil (mögliche Wasserstoffbrücken) sowie aromatische Wechselwirkungen. Unbekannte Bindungsarten werden ebenfalls als solche gekennzeichnet.

Die Menge $B$ enthält 5 Elemente $B = \{HH, Hh, hh, Ar, Uk\}$, die jeweils für eine Bindungsart stehen. ”HH” steht für hydrophob-hydrophoben Zusammenhalt, ”Hh” für hydrophob-hydrophile Abstoßung, ”hh” für potenzielle Wasserstoffbrücken bzw. hydrophil-hydrophile Interaktion, ”Ar” für aromatische Wechselwirkungen und ”Uk” für unbekannte Bindungen. Jeder Position $i \in I = [0, n+1]$ eines Motivs $m$ wird eine Bindungsart $b \in B$ zugewiesen. Das entspricht der Funktion $\varphi$, die mit

$$\varphi : I \to B, i \to b \tag{3.12}$$

definiert ist. Die Wahrscheinlichkeit $P(b,i,M)$

$$P(b,i,M) = \frac{\sum_{m\in M} k(b,i,m)}{|M|} \tag{3.13}$$

mit

$$k(b,i,m) = \begin{cases} 1 & \text{wenn } \varphi(i) = b \text{ in } m \\ 0 & \text{sonst,} \end{cases} \tag{3.14}$$

gibt an, wie wahrscheinlich es ist, dass eine Position eine bestimmte Bindung aufbaut. Das Bindungsprofil eines Motivs $M$ beinhaltet die Werte zu jeder Position $i$ und jeder Bindungsart $b$.

$$Bnd_M = \begin{pmatrix} P(HH|1|M) & P(Hh|1|M) & \cdots & P(Uk|1|M) \\ P(HH|2|M) & P(Hh|2|M) & \cdots & P(Uk|2|M) \\ \vdots & \vdots & \ddots & \vdots \\ P(HH|n|M) & P(Hh|n|M) & \cdots & P(Uk|n|M) \end{pmatrix} \tag{3.15}$$

## 3.3 Analyse der Interaktionen zwischen Motiven

Zwei Motive interagieren miteinander, wenn sich innerhalb eines Motivs eine Aminosäure befindet, die mit einer Aminosäure in einem anderen Motiv in Kontakt steht. Dieses Verhältnis zweier Motive wird weiterhin als Co-Occurrence bezeichnet. Das gemeinsame Auftreten von Motiven entspricht einem ungerichteten gewichteten Graphen ohne mehrfach Kanten (siehe Abb. 3.1). Die Motive bilden die Menge der Knoten $V$. Sollten zwei Motive durch einen Kontakt verbunden sein, wird eine Kante zwischen den beiden Knoten eingefügt. Die Menge der Kanten entspricht der Menge $E$. Jeder der Kanten $e \in E$ kann mit Hilfe von $f(e) = g$ ein Gewicht zugewiesen werden. Für diesen Graphen $G = (V, E, g)$ existiert eine Adjazenzmatrix $A = [a_{ij}]$ wobei $i, j \in V$, die über ihre Einträge definiert ist.

$$a_{ij} = \begin{cases} g_{ij} & \text{falls die Kante } (i,j) \in E \\ 0 & \text{sonst,} \end{cases} \tag{3.16}$$

Bei der ersten Co-Occurrence zweier Motive $m_1$ und $m_2$ wird die Kante $e = M_1, M_2$ hinzugefügt und deren Gewichtung auf 1 gesetzt $f(e) = g = 1$. Sollten noch mehr Co-Occurrences zwischen den selben Motiven auftreten, würde die Gewichtung jeweils um 1 erhöht $f(e) = f(e) + 1$.

Bei der Extraktion der Co-Occurrences wird in einem Datensatz jedes Protein untersucht. Für jede Interaktion zwischen zwei Helices werden alle möglichen Motive auf beiden Seiten der Interaktion erzeugt und als Knoten hinzugefügt sowie mit einer Kante im Graphen verbunden.

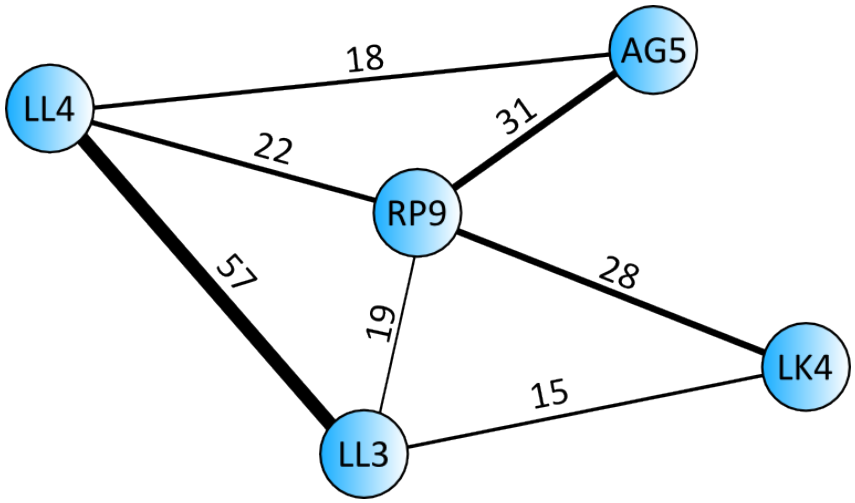


Abbildung 3.1: Beispiel eines Graphen zur Analyse der Co-Occurrences. Die Motive sind in blauen Kreisen abgebildet, die die Knoten des Graphen darstellen. Die Linien zwischen den Knoten symbolisieren die Kanten mit den zugehörigen Gewichtungen. Die Dicke der Kanten korrespondiert mit den Gewichtungen.

## 3.4 Vergleich von Motiven aus verschiedenen Proteinfamilien

Aus dem grundlegenden Datensatz wurden Proteine ausgewählt, die einer bestimmten Proteinfamilie zugeordnet werden konnten. Dabei wurden die folgenden 13 Familien ausgesucht: PF00001, PF00520, PF00528, PF00654, PF00664, PF00689, PF00860, PF01036, PF01127, PF02932, PF03595, PF07885, PF10320. Diese Proteinfamilien repräsentieren verschiedene Protein-Funktionen und es sind ausreichend viele Daten zur Analyse vorhanden. Damit ein Protein aus diesen Familien ausgewählt und zum Datensatz hinzugefügt wird, muss es mindestens einen Bereich besitzen, der für diese Proteinfamilie typisch ist. Für jede dieser Familien wurde eine Berechnung der Konserviertheit, Kontaktwahrscheinlichkeit, Bindungsprofile und Co-Occurrences durchgeführt.

Um die Ähnlichkeit zweier Familien $F_1$ und $F_2$ zu bestimmten, wurden zuerst die Motive ausgewählt, die in beiden Familien vorkamen und zur Menge der Motive $M$ hinzugefügt. Anschließend sollen alle Motive nacheinander verglichen und jedem Motiv ein Ähnlichkeitsmaß zugewiesen werden. Es wurden die zuvor berechneten aminosäure- und positionsspezifischen Konserviertheitswerte der Motive ausgewählt. Um zwei Motive miteinander zu vergleichen, wurden deren variable Bereiche gegenübergestellt.

Es sollen für jede Position $i$ die fünf Aminosäuren mit den höchsten Konserviertheits-Werten in einer Menge $B$ zusammengefasst werden.

Es sei $K = K_0$ die Menge aller $Kons(a,i,M) \forall a \in A$ und

$$B_1 := \{b \in K_0 : b \geq a \forall\, a \in K_0\}. \tag{3.17}$$

Außerdem gilt, $K_1 := K_0 \setminus B_1$. Für alle weiteren $K_j \forall\, j \in \mathbb{N}$ sei

$$B_{j+1} := \{b \in K_j : b \geq a \forall\, a \in K_j\}, \tag{3.18}$$

sowie $K_{j+1} := K_j \setminus B_{j+1}$.

Die Menge $B_5$ wäre die Menge mit den fünf am höchsten konservierten Aminosäuren. Für ein bestimmtes Motiv $M$, das in den Ausprägungen der Familien $M_{F1}$ und $M_{F2}$ vorkommt und jede Position $i$ im variablen Bereich $i \in [1,n]$, wurden diese fünf Aminosäuren so in Mengen überführt. Wenn beispielsweise an der ersten Position des Motivs LL4 aus der Pfamfamilie PF00780 die Aminosäuren A, G, S, V und W am höchsten konserviert waren, wurden diese in der Menge $B = \{A,G,S,V,W\}$ zusammengefasst. Dasselbe geschieht mit der ersten Position des Motivs der zu vergleichenden Proteinfamilie. Die beiden so erhaltenen Mengen werden nun mit Hilfe einer Kosinus-Ähnlichkeitsfunktion verglichen [34]. Die Kosinus-Ähnlichkeit wurde gewählt, da sie ein Ergebnis liefert, das bei den vorhandenen Werten eleganter Weise zwischen 1 und 0 liegt und somit direkt

in ein prozentuales Ähnlichkeitsmaß umgerechnet werden kann. Die Mengen werden dabei zuerst in zwei Vektoren $\mathbf{r}$ und $\mathbf{s}$ überführt. Dafür wird die Vereinigungsmenge $C$ beider Wörter ermittelt:

$$C = B_{5_{F1_i}} \cup B_{5_{F2_i}}. \tag{3.19}$$

Die Gesamtanzahl der sich unterscheidenden Buchstaben gibt die Anzahl der Dimensionen der Vektoren an $u = |C|$. Demzufolge sind $\mathbf{r}$ und $\mathbf{s}$ Elemente des $\mathbb{N}^u$. Anschließend wird die Menge $C$ in ein Tupel überführt,

$$C = \{c_1, c_2, \ldots, c_u\} := T = (t_1, t_2, \ldots, t_u) \tag{3.20}$$

um einen definierten Zugriff auf die Elemente zu ermöglichen.

Wenn ein Element $t_j$ wobei $j \in \mathbb{N}$ in $B_{5_{F1_i}}$, bzw. $B_{5_{F2_i}}$ vorhanden ist, wird im korrespondierenden Vektor $\mathbf{r}$, respektive $\mathbf{s}$, eine 1 eingetragen.

$$\mathbf{r} = (r_1, r_2, \ldots, r_u) \tag{3.21}$$

$$r_j = \begin{cases} 1 & \text{wenn } t_j \in B_{5_{F1_i}} \\ 0 & \text{sonst,} \end{cases} \tag{3.22}$$

Dasselbe geschieht mit dem Vektor $\mathbf{s}$ und Familie 2.

Mithilfe des Skalarproduktes und der Norm wird der Kosinus des von den Vektoren eingeschlossenen Winkels berechnet:

$$Sim(\mathbf{r}, \mathbf{s}) = \cos(\Theta) = \frac{\mathbf{r} \cdot \mathbf{s}}{\| \mathbf{r} \| \cdot \| \mathbf{s} \|} = \frac{\sum_{i=1}^{n} \mathbf{r}_i \times \mathbf{s}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{r}_i)^2} \times \sqrt{\sum_{i=1}^{n} (\mathbf{s}_i)^2}} \tag{3.23}$$

Das Ergebnis liegt immer zwischen 0 und 1, da die Werte im Vektor nicht negativ werden können. Eine 1 gibt die Identität der Vektoren an. Bei 0 zeigen sie keine Übereinstimmungen. Dieses Verfahren wird auf jede der variablen Positionen angewendet, um anschließend einen Mittelwert aus allen Werten zu bilden.

$$Sim(M_{F1}, M_{F2}) = \frac{1}{n} \sum_{i=0}^{n} Sim(\mathbf{r}, \mathbf{s}) \tag{3.24}$$

Dieser Mittelwert ist ein Maß für die Ähnlichkeit der Motive. Nachdem diese Ähnlichkeit für alle Motive der Familien berechnet wurde, kann durch die erneute Bildung des Mittelwertes all dieser Werte die Ähnlichkeit zweier Familien abgeschätzt werden.

$$Sim(F_i, F_j) = \frac{1}{n} \sum_{m \in M} Sim(m_{F1}, m_{F2}) \tag{3.25}$$

Dieser Wert kann Auskunft darüber geben, wie ähnlich sich zwei Proteinfamilien sind. Um das Ähnlichkeitsmaß zu einem Distanzmaß umzuwandeln, wird der Wert lediglich von 1 subtrahiert. Wenn mehr als zwei Familien verglichen werden sollen, kann mit den

Werten der Ähnlichkeiten eine euklidische Distanzmatrix gefüllt werden.

$$D_{n,n} = \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n,1} & d_{n,2} & \cdots & d_{n,n} \end{pmatrix} \tag{3.26}$$

mit

$$d_{i,j} = 1 - Sim(F_i, F_j)\ i, j, n \in \mathbb{N} \tag{3.27}$$

## 3.5 Vorhersage von helikalen Bereichen auf Grundlage der Sequenz

Es soll neben einer Darstellung von Proteinen, deren helikalen Bereiche bereits bekannt sind, auch eine Darstellung von Proteinen ermöglicht werden, bei denen dies nicht der Fall ist. Auf Grundlage der Sequenz kann eine Vorhersage der helikalen Bereiche getroffen werden. Die hier beschriebene Methode basiert auf der Konserviertheit von Motiven [35].

Als Grundlage dient ein Datensatz, in dem nicht nur die Motive um die Kontaktpositionen betrachtet wurden. Es sind alle Motive, die in helikalen Bereichen vorkommen, einbezogen worden. Für die Beschränkung der Länge der variablen Bereiche wurde $2 \geq n \leq 10$ gewählt. Danach wurden sogenannte LogOdd-Profile generiert, in denen die Konserviertheit der variablen Bereiche beschrieben wird [36]. Mithilfe dieser Profile kann eine eingegebene Aminosäuresequenz auf helikale Bereiche untersucht werden. Die Sequenz wird mit den im Profil enthaltenen Motiven abgeglichen. Wenn ein Motiv an einer Stelle in der Sequenz ”passt”, werden die darunterliegenden Aminosäuren mit den LogOdd-Werten gewichtet. Sollten mehrere Motive an einer Position übereinstimmen, wird der Durchschnitt der beiden Gewichtungen gewählt. Nachdem der Algorithmus die Sequenz durchlaufen hat, sollten sich die helikalen Bereiche durch eine höhere Wertung herauskristallisieren. Zum Abschluss werden noch Verfeinerungen durchgeführt. Es werden kurze Lücken zwischen Helices geschlossen und Helices mit zu wenigen Aminosäuren werden verworfen.

## 3.6 Darstellung der Motive in einem 2,5d Schema

Als Vorlage für die von einem Algorithmus zu erstellenden Schemas diente die von Hand erstellte Darstellung 3.2. Primär soll dargestellt werden, welche Motive miteinander interagieren können. Dabei sind die Motive natürlich von höherer Bedeutung, die eine hohe Anzahl an Co-Occurrences aufweisen. Die Auswahl der möglichen Nachbarmotive zu einem Hub-Motiv sollte demnach eingeschränkt werden. Zu jedem Motiv sollen weiterhin Informationen zur Verfügung stehen, um mögliche Funktionen des Motivs an dieser Stelle besser zu verstehen.

Algorithmisch wurde die wie folgt beschriebene Herangehensweise gewählt, die auch in einem Programmablaufplan im Anhang unter B.1, B.2 und B.3 festgehalten ist. Als Grundlage für die Darstellung eines Interaktionsschemas dient ein beliebiges transmembranes $\alpha$-helikales Protein. Weist das Protein charakteristische Bereiche auf, die sich einer Proteinfamilie zuweisen lassen, so wird der jeweils passende Motivdatensatz geladen. Wenn das Protein beispielsweise einen Bereich besitzt, der zur Proteinfamilie PF01036 gehört, werden auch deren Konserviertheits-, Kontaktwahrscheinlichkeits- und Co-Occureence-Werte verwendet. Sollte der Datensatz nicht vorhanden sein, oder

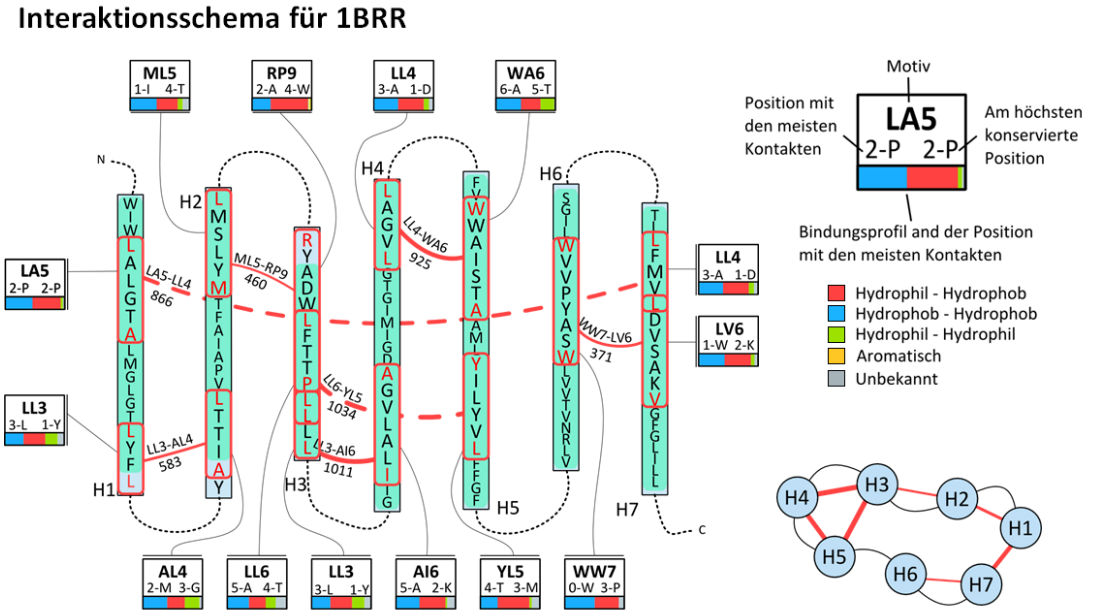


Abbildung 3.2: Das Schema zeigt die helikalen Bereiche des Proteins 1BRR. Ausgewählte Motiven wurden hervorgehoben und mit hoch gewichteten Co-Occurrence Partnern verbunden. Die Zahl unter der Verbindungslinie gibt an, wie oft diese Verbindung im Datensatz zur Proteinfamilie PF01036 vorkam. In den Kästen am Rande der Struktur sind die Informationen zu den Motiven enthalten. Unter diesen Informationen befinden sich die Position und die Aminosäure mit den meisten Kontakten und die am höchsten konservierte Position und Aminosäure. Darunter ist das Bindungsprofil in einem Balkendiagramm dargestellt. In der rechten unteren Ecke ist das Protein von Oben betrachtet innerhalb der Membran abgebildet.

das Protein keine solchen charakteristischen Bereiche haben, wird der Datensatz gewählt, dem die nicht redundanten Proteine als Berechnungsgrundlage dienten. Sollte das gewählte Protein eines der 1.107 Proteine sein, die im Ausgangsdatensatz enthalten sind, so wird der helikale Bereich dargestellt, der im Datensatz bereits vorliegt. Wenn dies nicht der Fall ist, wird auf Grundlage der Primärstruktur eine Vorhersage der helikalen Bereiche durchgeführt. In der Darstellung werden die Helices in Boxen nebeneinander eingezeichnet. Eine gebogene Linie zwischen den Boxen symbolisiert das Rückgrat des Proteins. Innerhalb der Boxen stehen die Aminosäuresequenzen der Helices in der jeweiligen Ausrichtung zur Membran. Nun erfolgt die Auswahl eines ”Hub-Motivs”. Das Hub-Motiv stellt das Zentrum der Interaktion dar, alle weiteren hinzugefügten Motive werden im Bezug auf dieses Motiv betrachtet. Das Hub-Motiv wird immer in Schwarz dargestellt. Nach der Auswahl des Hub-Motivs werden 50 weitere Motive vorgeschlagen, die absteigend nach Häufigkeit der Co-Occurrence zu diesem Motiv sortiert sind. Nach der Auswahl der Co-Occurrence Motive werden zuerst die Ausprägungen der Hubmotive in den dargestellten Helices gesucht und markiert. Anschließend werden auch alle Ausprägungen der gewählten Co-Occurrence Motive markiert und in die Abbildung eingetragen. Die Farben für die Darstellung der Ausprägung der Motive

sind zufällig gewählt, jedoch einheitlich für jedes Motiv. Anschließend werden die Hub-Motive mit den Co-Occurrence Motiven verbunden. Es ist möglich, die Daten die zu den Motiven berechnet wurden, durch einen Klick aufzurufen und in einer überblicksartigen Form anzusehen.

# 4 Ergebnisse

## 4.1 Gesammelte Daten

Die bisher erfassten Daten der verschiedenen Datenbanken wurden zu zwei Teilen einer Datenbank zusammengefügt. Der erste Teil besteht aus allgemeinen Proteindaten wie der Sequenz, den helikalen Bereichen und den Interaktionsdaten. Der zweite Teil beinhaltet die Motive. In diesem Abschnitt sind zu jedem Motiv die Konserviertheit, die Kontaktwahrscheinlichkeit, Bindungsprofile und die Co-Occurrences gespeichert. Diese Motiv-Daten wurden sowohl aus der Gesamtheit der transmembranen Proteine als auch aus den Proteinfamilien berechnet.

## 4.2 Struktur von Motiven

Es existieren Motive, die häufig und in allen Proteinfamilien vorkommen und solche, die hoch konserviert und nur in bestimmten Proteinfamilien auftauchen. Allgegenwärtig sind kurze Motive mit hydrophoben Start und Endaminosäuren wie LL4, GG4, LL5 und AL4. Das bestätigt die bisherigen Ergebnisse von Senes [29] und Liu [26]. Interessanterweise sind die variablen Bereiche innerhalb der Motive zwar hoch konserviert, jedoch unterscheiden sie sich zwischen den Proteinfamilien stark, wie in Abbildung 4.1 auszugsweise zu erkennen. Es befinden sich überdurchschnittlich oft die großen aromatischen Aminosäuren Tryptophan und Tyrosin zwischen den hydrophoben Start- und Endaminosäuren. Dies deckt sich mit den Entdeckungen, die Adamian und Kollegen machten [30]. Mehrere kleine Aminosäuren (Alanin, Glycin, Leucin), die sehr häufig in transmembranen Bereichen auftauchen, scheinen eine Tasche für wenige große Aminosäuren zu bilden. In spannungsgesteuerten Ionenkanälen (z.B. PF00654) könnten diese großen Aminosäuren die Pore versperren bis das Membranpotenzial eine bestimmte Schwelle erreicht. Danach könnte es zu einer Konformationsänderung kommen, durch die diese Pore geöffnet bzw. bei einem Abfall des Potentials wieder geschlossen wird. In der untersuchten Proteinfamilie PF00654, zu der spannungsgesteuerte Chloridkanäle zählen, sind im LL4 Motiv an zwei der drei variabeln Positionen Tyrosine hoch konserviert (Abb. 4.1).

Besonders Augenmerk wurde außerdem auf Motive gelegt, in denen bestimmte Aminosäuren außergewöhnlich hoch konserviert waren. Das Motiv RP9 kommt besonders hoch konserviert in der Familie PF01036 vor, welche bacteriorhodopsin-ähnliche Proteine beinhaltet. Das Motiv ist Teil des Bacteriorhodopsin Signaturmotivs der PROSITE [37] und sticht auch im HMM-Logo [38] der Proteinfamilie heraus. Im Datensatz zur PF01036 ist Tyrosin immer an Position 1 und Tryptophan immer an Position 4 des Motivs anzutreffen (Abb. 4.2). Die restlichen Positionen erweisen sich ebenfalls als hoch

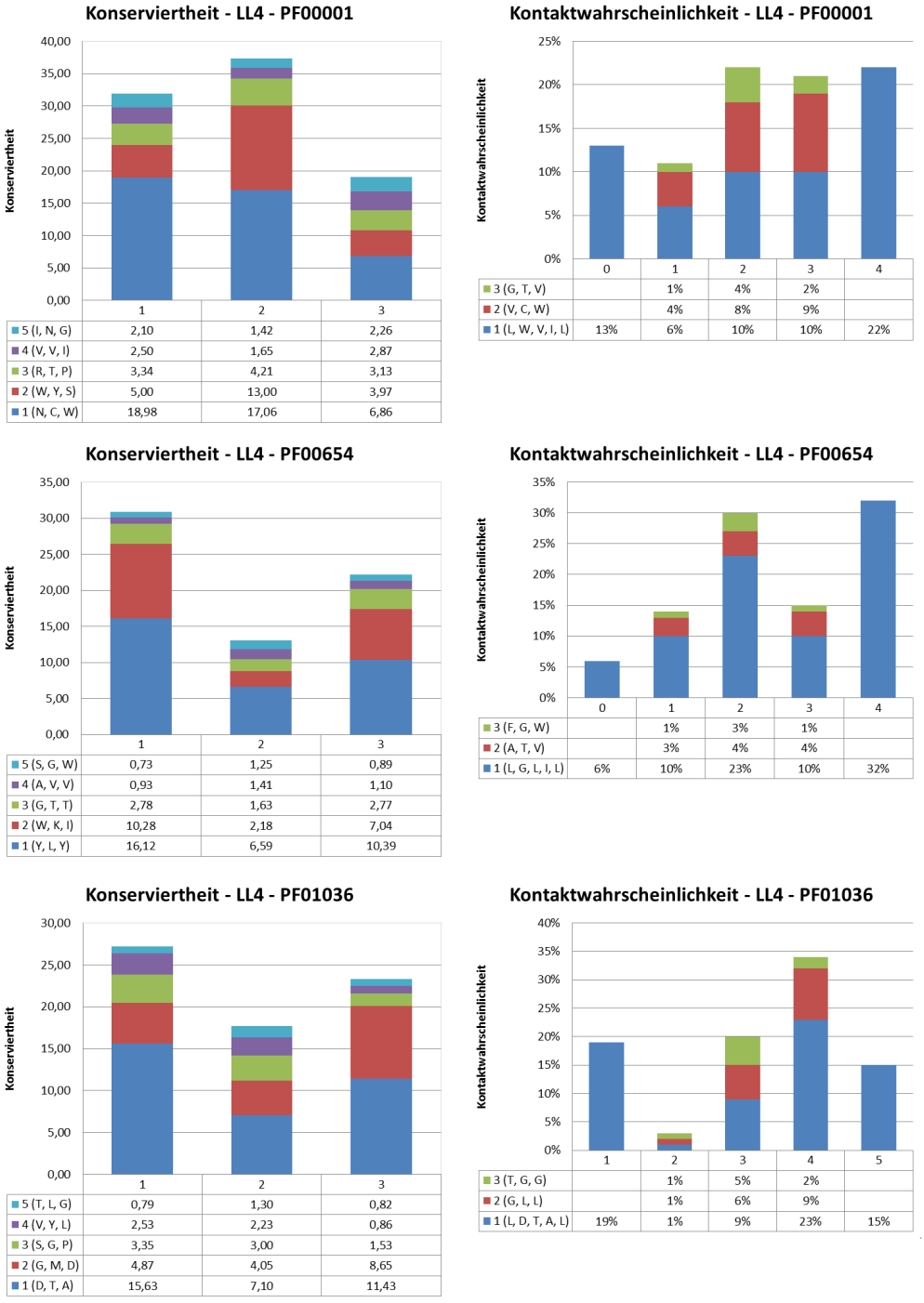


Abbildung 4.1: Diagramme und Werte des Motivs LL4 aus den Proteinfamilien PF00001, PF01036 und PF00654. In der linken Hälfte sind die Konserviertheiten der variablen Bereiche grafisch dargestellt. Rechts sind die Kontaktwahrscheinlichkeiten aller Positionen abgebildet. Auf der x-Achse sind die Positionen $i$ im Motiv aufgetragen. In der Tabelle unter den Diagrammen sind die am höchsten konservierten Aminosäuren bzw. die Aminosäuren mit den höchsten Kontaktwahrscheinlichkeiten aufsteigend eingetragen. In der ersten Spalte sind in Klammern die Aminosäuren aufgezeigt, welche diesen Wert an der jeweiligen Position ausmachen.

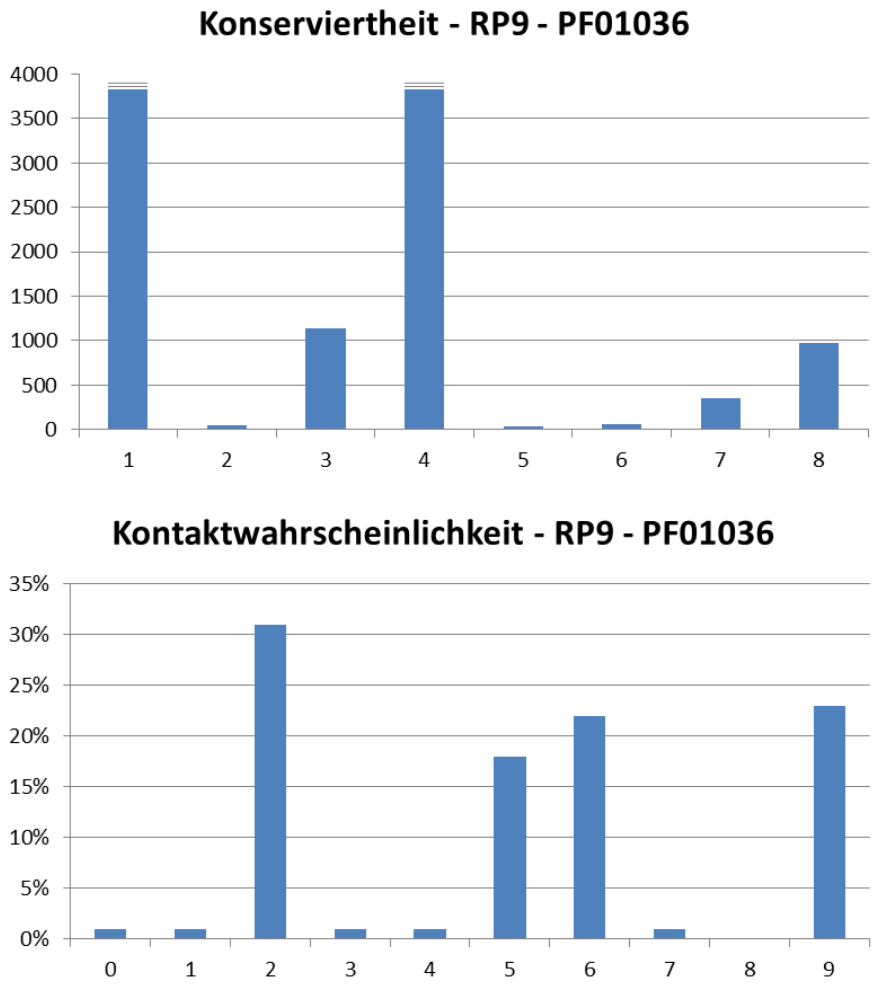


Abbildung 4.2: Diagramme und Werte des Motivs RP9 aus PF01036. Das obere Diagramm stellt die Konserviertheit der variablen Bereiche grafisch dar. Im unteren sind die Kontaktwahrscheinlichkeiten aller Positionen abgebildet. Auf der x-Achse sind die Positionen $i$ im Motiv aufgeführt. Die Konserviertheit der Werte 1 und 4 ist theoretisch $\infty$ und wurde durch einen abgeschnittenen Balken dargestellt.

konserviert. Interessanterweise sind die Kontaktwahrscheinlichkeiten beider Positionen 1 und 4 sehr gering. Wenn die Aminosäuren nicht zur Aufrechterhaltung von Interaktionen zwischen Helices dienen, und dennoch so hoch konserviert sind, müssen sie zu einem anderen Zweck evolutionär begünstigt worden sein. Diese beiden Aminosäuren könnten demzufolge für die Funktion des Proteins konserviert sein und nicht für die Aufrechterhaltung der Struktur.

## 4.3 Analyse der Ähnlichkeiten zwischen Proteinfamilien

Die 13 genannten Proteinfamilien wurden miteinander mit dem vorgestellten Algorithmus verglichen und die Ergebnisse in eine Distanzmatrix eingetragen. Mithilfe einer Implementierung des Unweighted Pair Group Method with Arithmetic Mean Algorithmus (UPGMA) [39], die vom Institut Pasteur Biology IT Center [40] zur Verfügung gestellt wird, konnte ein Ähnlichkeitsbaum erstellt werden (Abb. 4.3). Die zugehörige Distanzmatrix ist im Anhang A.1 zu sehen.

Eine sehr hohe Ähnlichkeit und demzufolge einen kleine Distanzwert haben die Proteinfamilien PF00520 und PF07885. Bei PF00520 handelt es sich um Proteine mit sechs transmembranen Helices, die zum Transport von Anionen wie Kalium, Natrium und Calcium dienen. PF07885 ist eine Proteinfamilie, die lediglich zwei Helices besitzt, aber ebenfalls für den Transport von Anionen zuständig ist. Mit einem Abstand von nur 0.49 heben sich diese beiden Proteinfamilien vom Rest der Vergleiche ab, die im Schnitt um die 0.83 liegen. Beim Betrachten einiger Vertreter dieser Gruppen fällt auf, dass die beiden Helices, die in PF07885 den transmembranen Bereich formen, so ähnlich aussehen, wie die zwei Helices in PF00520 (siehe Abb. 4.4).

Eine paarweise Analyse zweier Proteine aus diesen Familien, mittels distance alignment matrix method (DALI) [41], ergab ein gutes strukturelles Alignment beider Strukturen mit

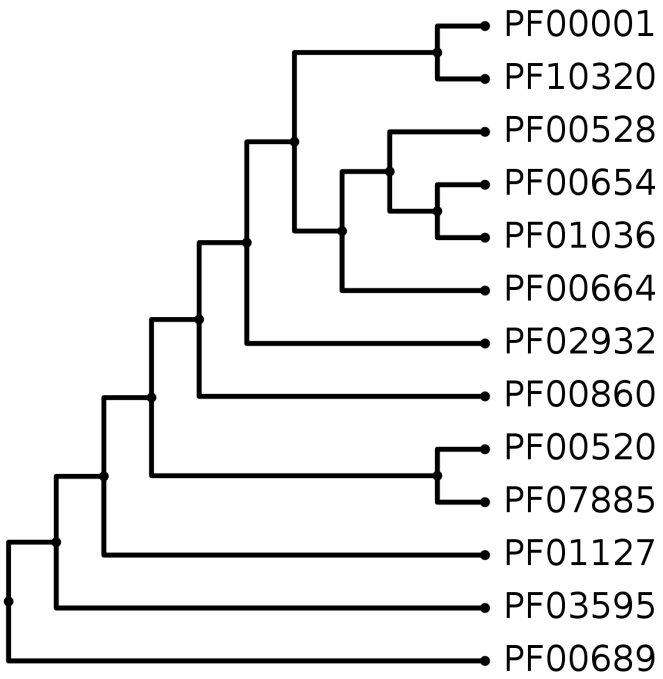


Abbildung 4.3: Der abgebildete Ähnlichkeitsbaum wurde mit UPGMA berechnet. Abgebildet sind die 13 Pfamfamilien und ihre Beziehungen zueinander.

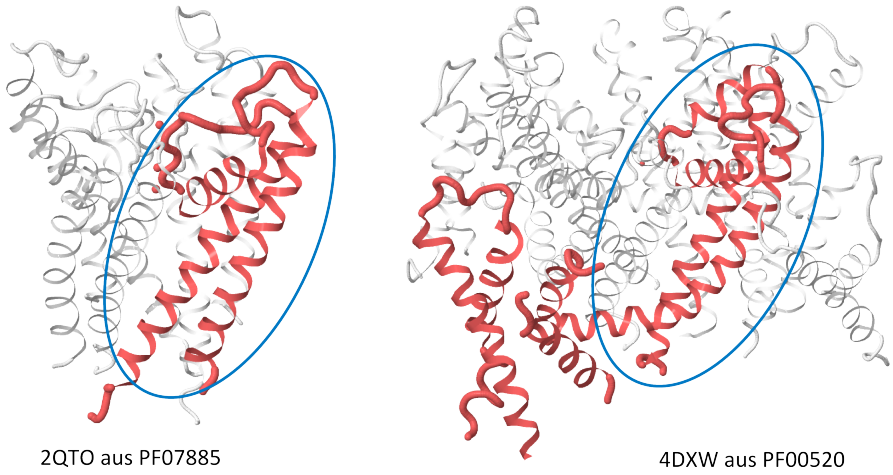


Abbildung 4.4: Abgebildet sind die beiden Proteine 2QTO aus PF07885 und 4DWX aus PF00520. Rot hervorgehoben ist die Kette A beider Proteine. Eingekreist sind die sich ähnelnden Abschnitte.

einem Z-Wert von 7,8. Eine Überlagerung der beiden Helices ist in Abbildung 4.5 zu sehen. Die Ähnlichkeiten der Strukturen könnte darauf schließen lassen, dass sich beide Proteinfamilien durch die Evolution aus einem Vorgänger entwickelt haben. Dafür spricht außerdem die Tatsache, dass die Proteinfamilie PF00520 vor allem in Eukaryoten vorkommt [42], wohingegen bei PF07885 der Anteil an Bakterien überwiegt [43]. PF07885 könnte als einfachere Variante von PF00520 betrachtet werden.

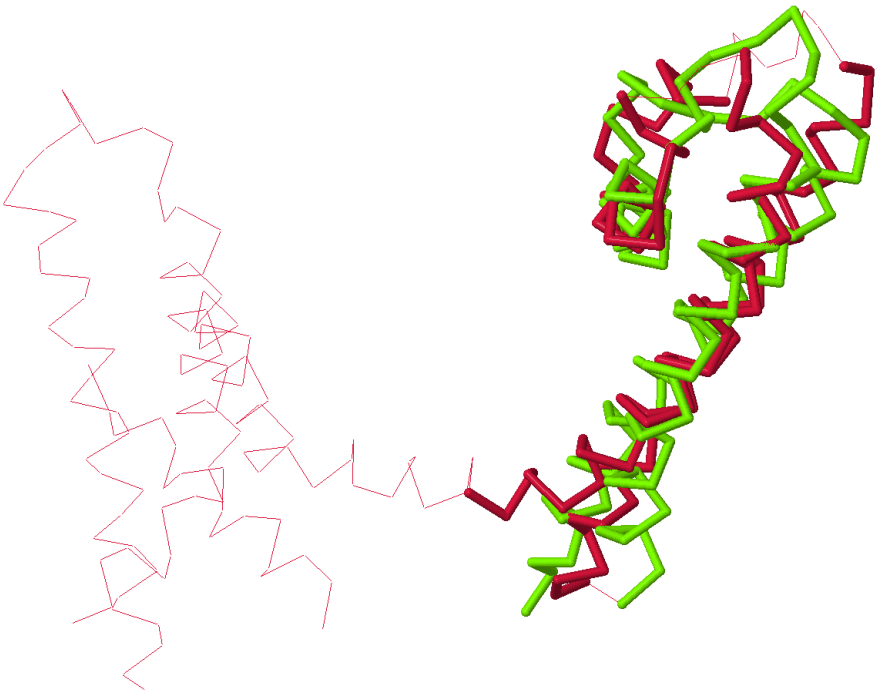

Abbildung 4.5: Abgebildet ist eine überlagerte Darstellung der Proteine 2QTO aus PF07885 und 4DWX aus PF00520. Rot ist das Rückgrat von 4DWX dargestellt, grün das von 2QTO.

Die beiden Proteinfamilien PF00001 und PF10320 zeigen eine erstaunliche Ähnlichkeit. Der Abstand der beiden Proteinfamilien beträgt lediglich 0.12 und ist somit der geringste aller verglichenen Proteinfamilien. Beide Proteine gehören zu den Rezeptorproteinen. PF00001 ist eine Gruppe G-Protein-gekoppelter Rezeptoren, die ein Signal ins Innere der Zelle übermitteln können. In PF10320 sind Proteine die hauptsächlich in Nematoden vorkommen und als chemosensorische Rezeptoren dienen [44]. Beide Proteinfamilien erfüllen die Funktion der Weiterleitung von Signalen mit Hilfe von sogenannten Guaninnucleotid-bindenden Proteinen (kurz G-Proteine). Bereits die Funktion lässt darauf schließen, dass diese Proteine sehr nah verwandt sein sollten, wenn sie dieselben Liganden besitzen. Leider sind bisher keine Strukturen der Familie PF10320 aufgeklärt, was einen direkten Vergleich verhindert.

Einen eher durchschnittlichen Abstand mit der Distanz von 0,81 haben die Familien PF00654 und PF01036. PF00654 ist wie bereits erwähnt, eine Proteinfamilie in der verschiedene spannungsgesteuerte Chloridkanäle zu finden sind. PF01036 klassifiziert Bacteriorhodopsine. Um den Verwandtheitsgrad der beiden Proteinfamilien näher zu bestimmen, wurden deren Funktion bzw. Aktivität mithilfe der Gene Ontology (GO) [47] untersucht (Abb. 4.6). Beide Familien sind Ionenkanäle. Bei PF00654 ist dies noch et-

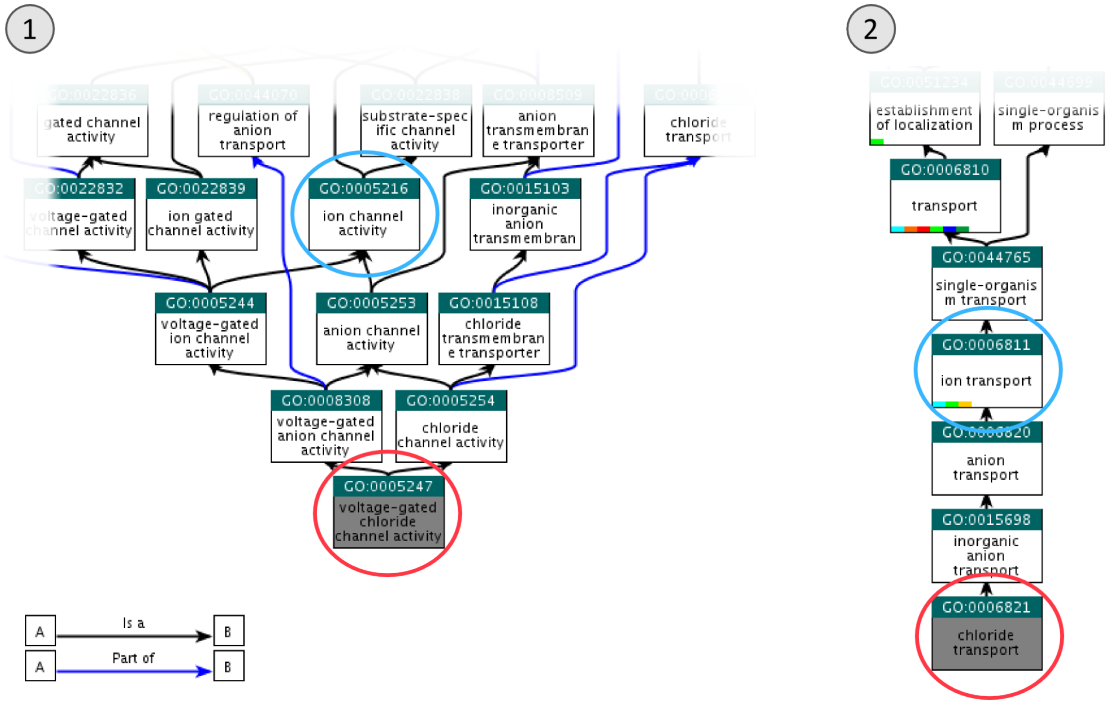


Abbildung 4.6: Die abgebildeten Kästen zeigen jeweils einen GO Term mit dem zugehörigen Identifier. Es sind Baumstrukturen zu sehen, die die Beziehungen der GO Terms darstellen. Links sind die molekularen Funktionen abgebildet, rechts ist der biologische Prozess dargestellt, in dem beide Proteinfamilien zu finden sind. In den roten Kreisen sind die Terme dargestellt, die der Familie PF00654 zugeordnet sind. In den blauen Kreisen wurden die Terme von PF01036 hervorgehoben. Abbildung modifiziert nach [45] und [46].

was genauer spezifiziert. Es handelt sich hierbei um einen spannungsgesteuerten Chloridkanal. Ebenso sind beide im Ionentransport involviert. Wiederum ist bei PF00654 genauer angegeben, dass es sich um den Transport von Chloriden handelt. Ein struktureller Vergleich mittels DALI der Vertreter 1BRR aus PF01036 und 1OTS aus PF0654 ergab ein Alignment mit einem Z-Wert von 2,7. Es sind gewisse Ähnlichkeiten zu erkennen. Sie sind jedoch nicht so stark ausgeprägt wie in den andern betrachteten Paarungen.

Die ausgewählten Beispiele zeigen, dass der Algorithmus bereits funktionstüchtig ist. Es is eine Abstufung zwischen größerer und geringerer Ähnlichkeit der Proteinfamilien erkennbar, die mit den berechneten Werten korreliert.

# 5 Zusammenfassung und Ausblick

## 5.1 Erreichte Ergebnisse

Es konnte gezeigt werden, dass die bereits beschriebenen Motive noch weitaus mehr Informationen beinhalten, als auf den ersten Blick erkennbar sind. Der variable Bereich zwischen den beiden konservierten Aminosäuren gibt weiteren Aufschluss über die Interaktion von Helices in transmembranen Bereichen. So können durch einen Vergleich der Konserviertheit und der Kontaktwahrscheinlichkeit eines Motivs, Aussagen über dessen ”Natur” gemacht werden. Einige Motive dienen vorrangig zur Aufrechterhaltung der Struktur, andere sind durch ihre Funktion konserviert. Der erstellte ”Interaktionsalmanach” bildet eine Grundlage für weiterführende Forschungen.

Eine Applikation mit graphischer Benutzeroberfläche ermöglicht eine vereinfachte Handhabbarkeit der gesammelten Daten. Es besteht die Möglichkeit, helikale Bereiche aus dem bestehenden Datensatz zu laden. Außerdem kann er auf Grundlage von Aminosäuresequenzen vorhergesagt werden. Wenn der helikale Bereich eines Proteins vorhanden ist, können Motive in diesem Protein untersucht werden. Durch die Möglichkeit der Vorhersage können auch neu entdeckte Proteine mit dem Programm ausgewertet werden. Es ist möglich, die Co-Occurence von Motiven auf den jeweiligen Proteinen

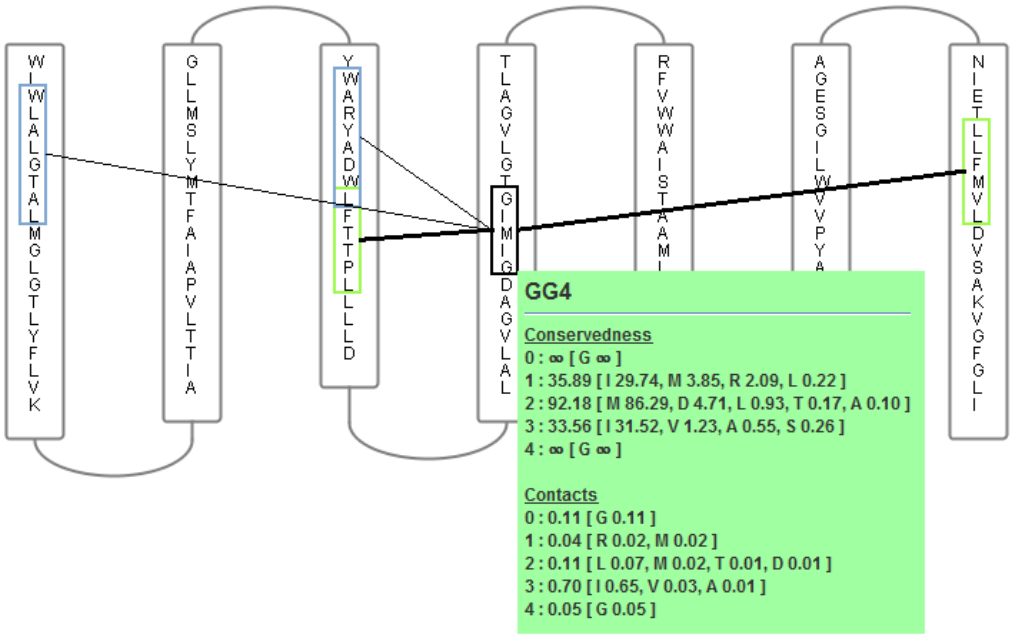


Abbildung 5.1: Interaktionschema der Kette A des Proteins 1BRR. Als Hub-Motiv wurde GG4 ausgesucht und als Co-Occurence Motive wurden WL7 und LL5 gewählt. Die Eigenschaften des ausgewählten Motivs sind in einem grünen Kasten dargestellt.

zu beobachten. Dabei werden Ausprägungen der Motive in den dargestellten Proteinen hervorgehoben und Verbindungen zu ausgewählten Nachbarmotiven hergestellt. Eine Beispielausgabe ist in Abbildung 5.1 zu sehen.

Ein Algorithmus zum Vergleich von Proteinfamilien auf Grundlage von Motiven wurde implementiert. Die Methode kann genutzt werden, Proteine mit ähnlichen Funktionen oder Strukturen zu finden sowie Distanzmatrizen für Ähnlichkeitsbäume zu erstellen. Erste Ergebnisse sind bereits erzielt worden und versprechen weitere mögliche Vorhersagen zur Ähnlichkeit von Proteinen.

Um eine Fortführung und Aktualisierung der Forschung zu gewährleisten, wurden graphische Benutzeroberflächen geschaffen, die das Erfassen neuer Proteindaten, das Generieren von Motiven aus Proteinen und das Erstellen von Distanzmatrizen vereinfachen. Die ausgegebenen Daten können im XML-Format gespeichert werden und sind so zur Weiterverarbeitung bereit.

## 5.2 Kritische Wertung

Es könnte eine Optimierung der Laufzeit und des Speicherbedarfs bei der Berechnung der Datensätze durchgeführt werden. Derzeit werden alle gefundenen Motive für eine weitere Analyse zwischengespeichert, was eine große Arbeitsspeicherbelastung darstellt. Eine temporäre Speicherung aller Motive und das jeweilige Laden der benötigten Motive könnten eine Entlastung bringen. Es würde jedoch zu einer längeren Laufzeit führen. Durch Parallelisierung könnten mehrere Motive gleichzeitig analysiert werden, was eine bessere Auslastung der Rechenkapazität von Mehrkernprozessoren zur Folge hätte.

Die vorhergesagten Ähnlichkeiten bei den Proteinfamilien lassen darauf schließen, dass der Ansatz des Algorithmus funktioniert. Um eine richtige Funktionsweise zu bestätigen, müssten noch weitere Evaluierungen durchgeführt werden. Da viele der berechneten Werte nahe beieinander liegen, sollte die Sensitivität des Algorithmus verfeinert werden. Beispielsweise könnte die Anzahl der verglichenen Aminosäuren variiert werden. Derzeit werden immer fünf Aminosäuren verglichen. Außerdem sollte ein Grenzwert gewählt werden, der Motive ausschließt, die nur selten in den Proteinfamilien gefunden wurden. Dadurch würde die Anzahl nicht-repräsentativer Motive verringert.

## 5.3 Ausblick

Bei der Extraktion von Motiven um eine Kontaktposition werden längere Motive zwangsläufig öfter vorkommen als kurze. Beispielsweise können nur drei Motive vom Typ XY2 generiert werden, die jeweils eine sich unterscheidende Kontaktposition beinhalten, je-

doch sieben vom Typ XY6. Daher gibt es auch häufig lange Motive, die in der Co-Occurrence hoch repräsentiert sind. Da jedoch kurze Motive auch in längeren Motiven beinhaltet seien können, ist es nicht ratsam diese Häufigkeiten durch eine Gewichtung zu vermeiden. Denn möglicherweise sind diese Repräsentationen der Motive nur in dieser Konstellation von kleineren Motiven wirksam. Diese verschachtelten- oder ”Matrjoschka”-Motive könnten Gegenstand weiterer Untersuchungen sein.

Es existieren mit Sicherheit weitere Motive, die für die Aufrechterhaltung der Struktur oder der Funktion unerlässlich sind. Eine weitere Erfassung von besonders hoch konservierten Motiven oder solchen Motiven, die auf wenige Positionen konzentriert, hohe Kontaktwahrscheinlichkeiten besitzen, ist von besonderem Interesse. Eventuell kann für jede Proteinfamilie eine kleine Anzahl von Motiven gefunden werden, die die Funktion bzw. die Struktur dieser Familie definieren. Anhand dieser Singnaturmotive könnten auch neue Proteine diesen Proteinfamilien zugeordnet werden.

# Anhang A: Tabellen

| Aminosäure | Absolute Häufigkeit | Relative Häufigkeit (%) |
|---|---|---|
| Ala (A) | 27.154 | 11,87 |
| Arg (R) | 2.945 | 1,29 |
| Asn (N) | 3.414 | 1,49 |
| Asp (D) | 1.961 | 0,86 |
| Cys (C) | 2.794 | 1,22 |
| Glu (E) | 2.523 | 1,10 |
| Gln (Q) | 2.383 | 1,04 |
| Gly (G) | 20.809 | 9,10 |
| His (H) | 3.251 | 1,42 |
| Ile (I) | 24.489 | 10,71 |
| Leu (L) | 38.166 | 16,69 |
| Lys (K) | 2.361 | 1,03 |
| Met (M) | 9.072 | 3,97 |
| Phe (F) | 20.079 | 8,78 |
| Pro (P) | 5.522 | 2,41 |
| Ser (S) | 11.030 | 4,82 |
| Thr (T) | 12.912 | 5,65 |
| Trp (W) | 6.811 | 2,98 |
| Tyr (Y) | 6.737 | 2,95 |
| Val (V) | 24.280 | 10,62 |

Tabelle A.1: Natürliches Vorkommen von Aminosäuren in transmembranen Bereichen von Proteinen [14].

| Familie | PF00001 | PF00520 | PF00528 | PF00654 | PF00664 | PF00689 | PF00860 | PF01036 | PF01127 | PF02932 | PF03595 | PF07885 | PF10320 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PF00001 | 0.0000 | 0.8583 | 0.8395 | 0.8361 | 0.8410 | 0.8652 | 0.8618 | 0.8186 | 0.8640 | 0.8395 | 0.8584 | 0.8557 | 0.1204 |
| PF00520 | 0.8583 | 0.0000 | 0.8612 | 0.8499 | 0.8566 | 0.8815 | 0.8693 | 0.8504 | 0.8625 | 0.8753 | 0.8663 | 0.4913 | 0.8525 |
| PF00528 | 0.8395 | 0.8612 | 0.0000 | 0.8195 | 0.8438 | 0.8653 | 0.8410 | 0.8223 | 0.8585 | 0.8515 | 0.8589 | 0.8403 | 0.8394 |
| PF00654 | 0.8361 | 0.8499 | 0.8195 | 0.0000 | 0.8230 | 0.8675 | 0.8215 | 0.8176 | 0.8419 | 0.8298 | 0.8487 | 0.8295 | 0.8360 |
| PF00664 | 0.8410 | 0.8566 | 0.8438 | 0.8230 | 0.0000 | 0.8700 | 0.8449 | 0.8212 | 0.8442 | 0.8457 | 0.8649 | 0.8406 | 0.8369 |
| PF00689 | 0.8652 | 0.8815 | 0.8653 | 0.8675 | 0.8700 | 0.0000 | 0.8850 | 0.8595 | 0.8935 | 0.8813 | 0.8908 | 0.8797 | 0.8691 |
| PF00860 | 0.8618 | 0.8693 | 0.8410 | 0.8215 | 0.8449 | 0.8850 | 0.0000 | 0.8362 | 0.8545 | 0.8485 | 0.8670 | 0.8490 | 0.8589 |
| PF01036 | 0.8186 | 0.8504 | 0.8223 | 0.8176 | 0.8212 | 0.8595 | 0.8362 | 0.0000 | 0.8333 | 0.8259 | 0.8290 | 0.8184 | 0.8234 |
| PF01127 | 0.8640 | 0.8625 | 0.8585 | 0.8419 | 0.8442 | 0.8935 | 0.8545 | 0.8333 | 0.0000 | 0.8624 | 0.8695 | 0.8453 | 0.8627 |
| PF02932 | 0.8395 | 0.8753 | 0.8515 | 0.8298 | 0.8457 | 0.8813 | 0.8485 | 0.8259 | 0.8624 | 0.0000 | 0.8722 | 0.8612 | 0.8429 |
| PF03595 | 0.8584 | 0.8663 | 0.8589 | 0.8487 | 0.8649 | 0.8908 | 0.8670 | 0.8290 | 0.8695 | 0.8722 | 0.0000 | 0.8595 | 0.8631 |
| PF07885 | 0.8557 | 0.4913 | 0.8403 | 0.8295 | 0.8406 | 0.8797 | 0.8490 | 0.8184 | 0.8453 | 0.8612 | 0.8595 | 0.0000 | 0.8557 |
| PF10320 | 0.1204 | 0.8525 | 0.8394 | 0.8360 | 0.8369 | 0.8691 | 0.8589 | 0.8234 | 0.8627 | 0.8429 | 0.8631 | 0.8557 | 0.0000 |

Tabelle A.2: Distanzmatrix der 13 Proteinfamilien, die mit dem vorgestellten Ähnlichkeitsalgorithmus verglichen wurden.

# Anhang B: Diagramme

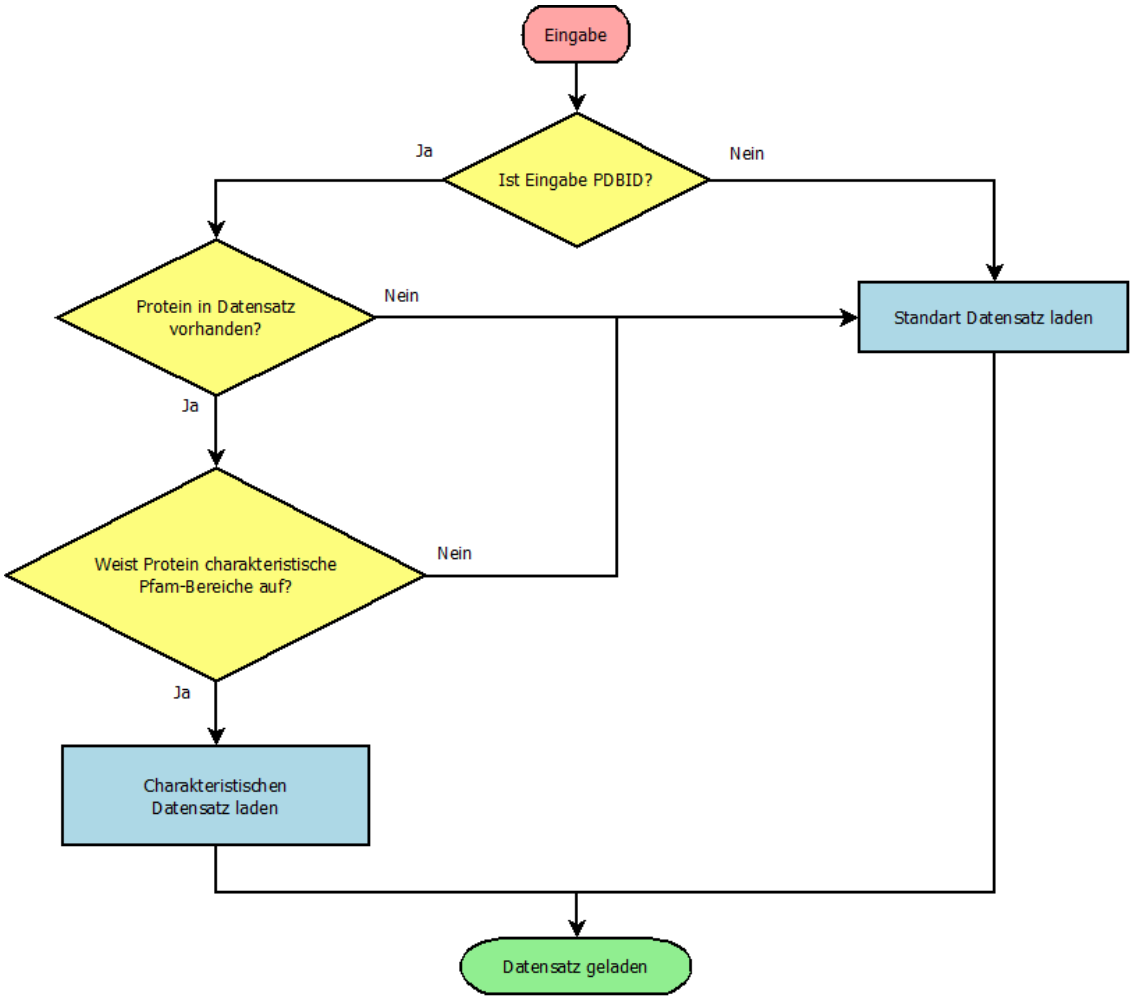


Abbildung B.1: Ablaufplan zur Auswahl eines Datensatzes. Der Start des Ablaufdiagrammes ist rot dargestellt. Anweisungen sind gelb und Befehle blau hervorgehoben. Das Ende des Ablaufplanes wurde grün betont.

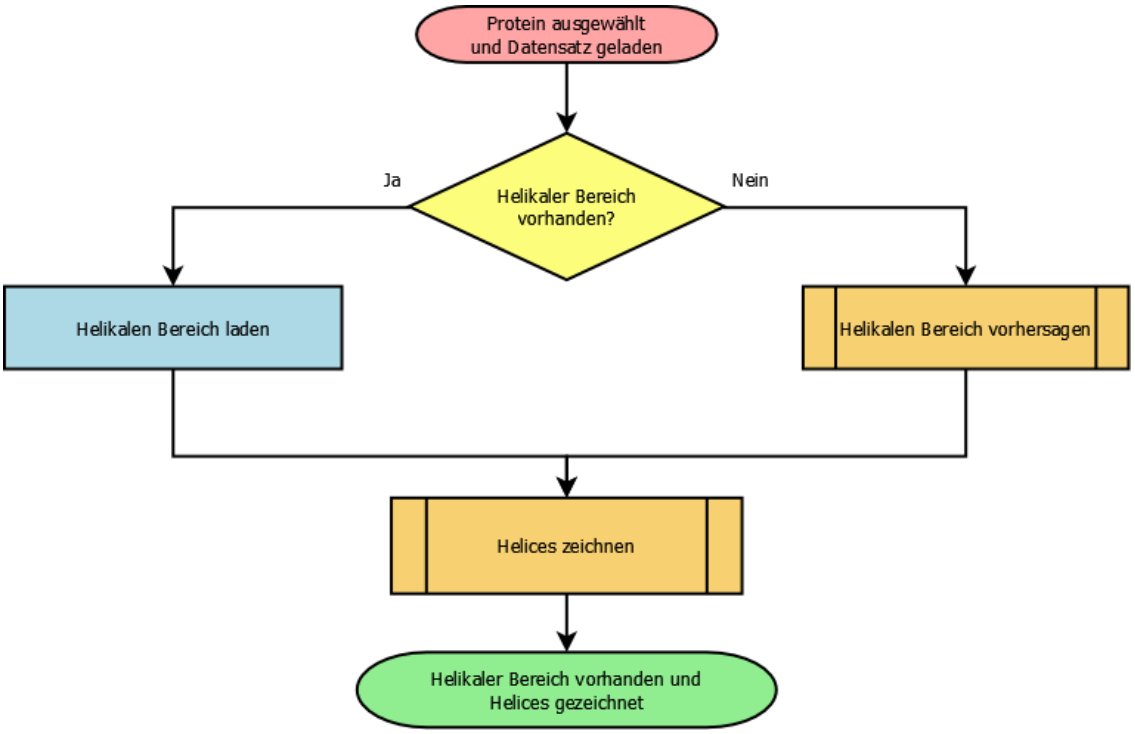


Abbildung B.2: Ablauf zum Laden oder Vorhersagen eines helikalen Bereiches. Der Start des Ablaufdiagrammes ist rot dargestellt. Anweisungen sind gelb, Befehle blau und Unterprogrammen orange hervorgehoben. Das Ende des Ablaufplanes wurde grün betont.

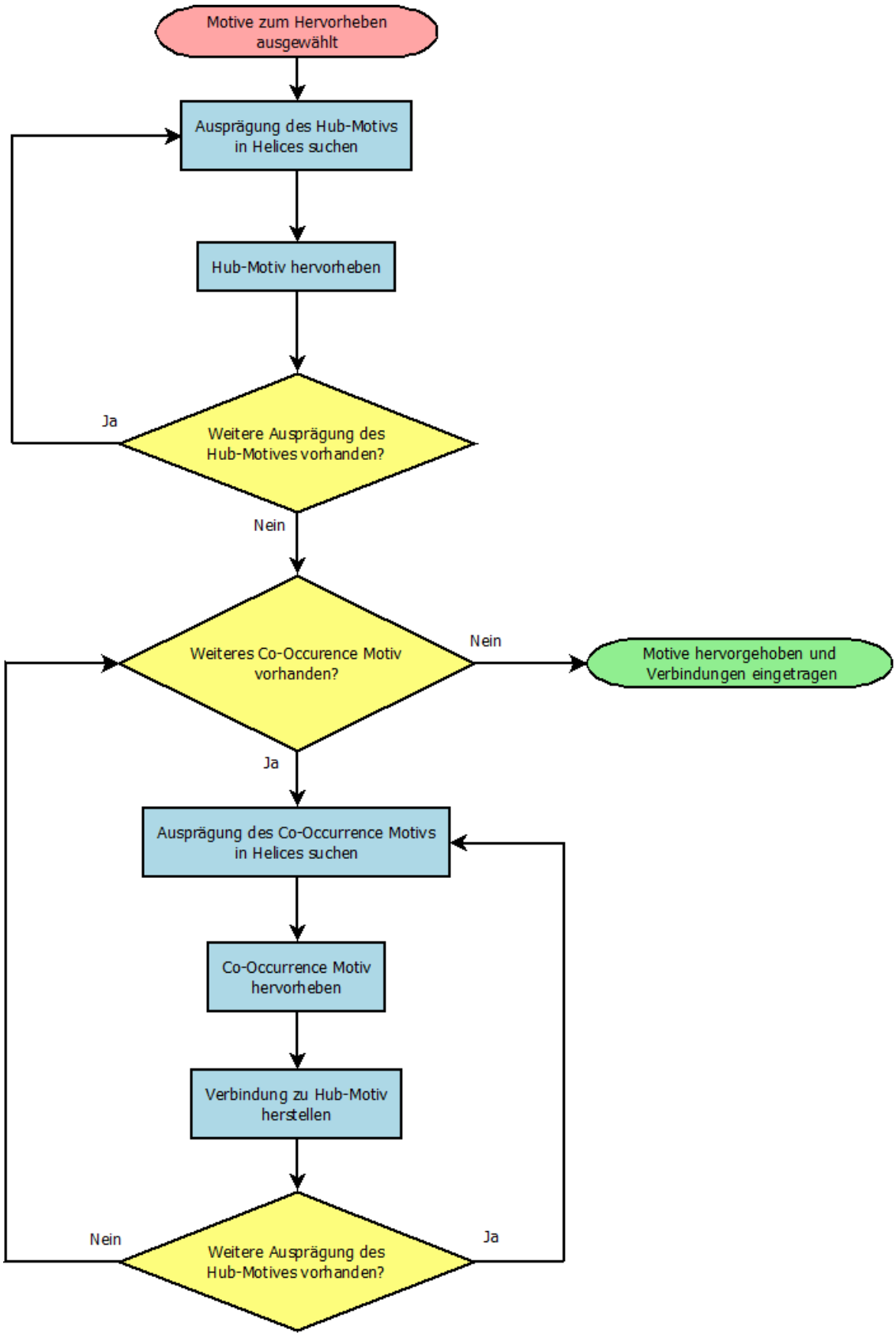


Abbildung B.3: Ablaufplan zur Hervorhebung von ausgewählten Motiven. Der Start des Ablaufdiagrammes ist rot dargestellt. Anweisungen sind gelb und Befehle blau hervorgehoben. Das Ende des Ablaufplanes wurde grün betont.

# Anhang C: Software

Eine DVD mit der Software und den Daten liegt bei. Auf ihr enthalten sind:

- Eine digitale Version der Bachelorarbeit
- Die Software
  - Spacial Motif Interaction 2.5d (`smi.jar`)
  - Distance Matrix Creator (`matrix.jar`)
  - Motif Collection Creator (`motifs.jar`)
  - Protein Collection Creator (`proteins.jar`)
- Generierte Daten
  - Proteindaten
  - Motivdaten
  - Co-Occurrence-Daten
- Ausgangsdaten
  - TMPad Daten (`tmpad.zip`)
  - Weizmann CSU Daten (`csu.zip`)
  - Pfam Mapping (`pfam_mapping.txt`)
  - Liste mit nicht redundanten Proteinen (`list_nr.txt`)
  - Liste mit allen Proteinen der PDBTM (`list_r.txt`)

Um die Software auszuführen, wird folgendes benötigt:

- Java Runtime Environment (JRE) 7

Um die Applikation "`smi.jar`" zu starten, muss sich der Ordner `data` (zu finden unter `/Software/data`) im selben Verzeichnis befinden wie die jar-Datei.

# Literaturverzeichnis

[1] John D. Allen and Alfred H. Schinkel. Multidrug resistance and pharmacological protection mediated by the breast cancer resistance protein (bcrp/abcg2). *Mol Cancer Ther*, 1(6):427–434, Apr 2002.

[2] John P. Overington, Bissan Al-Lazikani, and Andrew L. Hopkins. How many drug targets are there? *Nat Rev Drug Discov*, 5(12):993–996, Dec 2006.

[3] Anthony Watts. *Protein-lipid interactions*. Access Online via Elsevier, 1993.

[4] Carl Branden, John Tooze, et al. *Introduction to protein structure*, volume 2. Garland New York, 1991. Access Online via Google Books.

[5] Ursula Lehnert, Yu Xia, Thomas E. Royce, Chem-Sing Goh, Yang Liu, Alessandro Senes, Haiyuan Yu, Zhao Lei Zhang, Donald M. Engelman, and Mark Gerstein. Computational analysis of membrane proteins: genomic occurrence, structure prediction and helix interactions. *Q Rev Biophys*, 37(2):121–146, May 2004.

[6] J. U. Bowie. Helix packing in membrane proteins. *J Mol Biol*, 272(5):780–789, Oct 1997.

[7] Marina Gimpelev, Lucy R. Forrest, Diana Murray, and Barry Honig. Helical packing patterns in membrane and soluble proteins. *Biophys J*, 87(6):4075–4086, Dec 2004.

[8] Gábor E. Tusnády, Zsuzsanna Dosztányi, and István Simon. Pdbtm statistics - growth. via pdbtm.enzim.hu, 06 2013.

[9] Martin Caffrey. Membrane protein crystallization. *J Struct Biol*, 142(1):108–132, Apr 2003.

[10] T. Hirokawa, S. Boon-Chieng, and S. Mitaku. Sosui: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14(4):378–379, 1998.

[11] G. E. Tusnády and I. Simon. The hmmtop transmembrane topology prediction server. *Bioinformatics*, 17(9):849–850, Sep 2001.

[12] Allan Lo, Yi-Yuan Chiu, Einar Andreas Rødland, Ping-Chiang Lyu, Ting-Yi Sung, and Wen-Lian Hsu. Predicting helix-helix interactions from residue contacts in membrane proteins. *Bioinformatics*, 25(8):996–1003, Apr 2009.

[13] Gábor E. Tusnády, Zsuzsanna Dosztányi, and István Simon. Pdbtm: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*, 33(Database issue):D275–D278, Jan 2005.

[14] Allan Lo, Cheng-Wei Cheng, Yi-Yuan Chiu, Ting-Yi Sung, and Wen-Lian Hsu. Tmpad: an integrated structural database for helix-packing folds in transmembrane proteins. *Nucleic Acids Res*, 39(Database issue):D347–D355, Jan 2011.

[15] S. Miyazawa and R. L. Jernigan. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. *Protein Eng*, 6(3):267–278, Apr 1993.

[16] M. B. Swindells. A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci*, 4(1):93–102, Jan 1995.

[17] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, and M. Edelman. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–332, Apr 1999.

[18] Paula Yurkanis Bruice. *Organische Chemie: Studieren kompakt*. Pearson Deutschland GmbH, 2011.

[19] Lin Jiang and Luhua Lai. Ch...o hydrogen bonds at protein-protein interfaces. *J Biol Chem*, 277(40):37732–37740, Oct 2002.

[20] G. B. McGaughey, M. Gagné, and A. K. Rappé. pi-stacking interactions. alive and well in proteins. *J Biol Chem*, 273(25):15458–15463, Jun 1998.

[21] Mutasem Omar Sinnokrot, Edward F. Valeev, and C David Sherrill. Estimates of the ab initio limit for pi-pi interactions: the benzene dimer. *J Am Chem Soc*, 124(36):10887–10893, Sep 2002.

[22] C. N. Pace, B. A. Shirley, M. McNutt, and K. Gajiwala. Forces contributing to the conformational stability of proteins. *FASEB J*, 10(1):75–83, Jan 1996.

[23] E. Bibi. The role of the ribosome-translocon complex in translation and assembly of polytopic membrane proteins. *Trends Biochem Sci*, 23(2):51–55, Feb 1998.

[24] J. L. Popot and D. M. Engelman. Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, 29(17):4031–4037, May 1990.

[25] Donald M. Engelman, Yang Chen, Chen-Ni Chin, A Rachael Curran, Ann M. Dixon, Allison D. Dupuy, Albert S. Lee, Ursula Lehnert, Erin E. Matthews, Yana K. Reshet-

nyak, Alessandro Senes, and Jean-Luc Popot. Membrane protein folding: beyond the two stage model. *FEBS Lett*, 555(1):122–125, Nov 2003.

[26] Yang Liu, Donald M. Engelman, and Mark Gerstein. Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol*, 3(10):research0054, Sep 2002.

[27] I. T. Arkin and A. T. Brunger. Statistical analysis of predicted transmembrane alpha-helices. *Biochim Biophys Acta*, 1429(1):113–128, Dec 1998.

[28] K. R. MacKenzie, J. H. Prestegard, and D. M. Engelman. A transmembrane helix dimer: structure and implications. *Science*, 276(5309):131–133, Apr 1997.

[29] A. Senes, M. Gerstein, and D. M. Engelman. Statistical analysis of amino acid patterns in transmembrane helices: the gxxxg motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol*, 296(3):921–936, Feb 2000.

[30] Larisa Adamian, Ronald Jackups, Jr, T Andrew Binkowski, and Jie Liang. Higher-order interhelical spatial interactions in membrane proteins. *J Mol Biol*, 327(1):251–272, Mar 2003.

[31] Marco Punta, Penny C. Coggill, Ruth Y. Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L L. Sonnhammer, Sean R. Eddy, Alex Bateman, and Robert D. Finn. The pfam protein families database. *Nucleic Acids Res*, 40(Database issue):D290–D301, Jan 2012.

[32] W. Li, L. Jaroszewski, and A. Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, Mar 2001.

[33] Gábor E. Tusnády, Zsuzsanna Dosztányi, and István Simon. Pdbtm statistics - manual. via pdbtm.enzim.hu, 06 2013.

[34] Ning Liu, Benyu Zhang, Jun Yan, Qiang Yang, Shuicheng Yan, Zheng Chen, Fengshan Bai, and Wei-Ying Ma. Learning similarity measures in non-orthogonal space. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 334–341. ACM, 2004.

[35] Steffen Grunert, Christoph Leberecht, and Labudde Dirk. Evolution conserved spatial residue interactions of alpha-helical membrane protein structures to 2.5d chart information presentation. *International Conference on Applied Informatics for Health and Life Sciences*, 09 2013.

[36] Steffen Grunert, Florian Heinke, and Dirk Labudde. Structure topology prediction of discriminative sequence motifs in membrane proteins with domains of unknown functions. *Structural Biology*, 2013:10, 2013.

[37] Christian J A. Sigrist, Edouard de Castro, Lorenzo Cerutti, Béatrice A. Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. New and continuing developments at prosite. *Nucleic Acids Res*, 41(Database issue):D344–D347, Jan 2013.

[38] Benjamin Schuster-Böckler, Jörg Schultz, and Sven Rahmann. Hmm logos for visualization of protein families. *BMC Bioinformatics*, 5:7, Jan 2004.

[39] Joe Felsenstein. Phylip (phylogeny inference package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, 1993.

[40] Bertrand Néron, Hervé Ménager, Corinne Maufrais, Nicolas Joly, Julien Maupetit, Sébastien Letort, Sébastien Carrere, Pierre Tuffery, and Catherine Letondal. Mobyle: a new full web bioinformatics framework. *Bioinformatics*, 25(22):3005–3011, Nov 2009.

[41] Hitomi Hasegawa and Liisa Holm. Advances and pitfalls of protein structural alignment. *Curr Opin Struct Biol*, 19(3):341–348, Jun 2009.

[42] PfamDatabase. Pf00520, 08 2013. http://pfam.sanger.ac.uk/family/PF00520.

[43] PfamDatabase. Pf07885, 08 2013. http://pfam.sanger.ac.uk/family/PF07885.

[44] E. R. Troemel, J. H. Chou, N. D. Dwyer, H. A. Colbert, and C. I. Bargmann. Divergent seven transmembrane receptors are candidate chemosensory receptors in c. elegans. *Cell*, 83(2):207–218, Oct 1995.

[45] EMBL-EBI. Go:0006821 chloride transport anchestors, 08 2013.

[46] EMBL-EBI. Go:0005247 voltage-gated chloride channel activity, 08 2013.

[47] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

# Erklärung

Hiermit erkläre ich, dass ich meine Arbeit selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die Arbeit noch nicht anderweitig für Prüfungszwecke vorgelegt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Mittweida, 20.08.2013