KAPITEL 4 Ergebnisse

Die Ergebnisse der Untersuchungen zu den drei Themenkomplexen

- Vorhersage der Schnittstellenregion humaner Signalpeptide
- lokale Sturkturvorhersage von Peptidylprolylbindungen und
- der Entwurf zellprotektiver Peptide

werden im folgenden vorgestellt.

4.a Schnittstellen humaner Signalpeptide

Die genaue Kenntnis der Schnittstellenregion von Signalpeptiden und deren Eigenschaften ist wichtig für das *de novo* Design funktionaler Moleküle. Mit diesem Wissen können neue Schnittstellen oder auch potentielle Inhibitoren für die Signalpeptidase entworfen werden. 12 Aminosäuren umfassende Schnittstellensequenzausschnitte humaner sekretorischer Proteine wurden mit den beschriebenen Verfahren untersucht. Der Vergleich der unterschiedlichen Methoden zeigt, daß Künstliche Neuronale Netze mit adaptiver Kodierung (adaptive coding networks, ACN) wesentliche Vorteile gegenüber den bisher verwendeten KNN haben. Die Analyse des trainierten ACN ermöglicht die Beschreibung einer Konsensussequenz und die detaillierte Beschreibung der Schnittstelleneigenschaften.

Zuerst wurden die Sequenzdaten mit der Informationsanalyse, Häufigkeitsanalyse, Schwerpunktanalyse, dem Bayes-Prediktor und der Hauptkomponentenanalyse untersucht. Mit diesen Methoden ist eine Beurteilung der Daten hinsichtlich ihrer Verwendbarkeit bezüglich einer Klassifikation möglich. Anschließend wurde mittels bioinformatischer Methoden ein Klassifikator entwickelt. Die Möglichkeiten der Interpretation von adaptiv kodierten KNN wurde vorgestellt, um so die Eigenschaften der Schnittstellenregion zu untersuchen.

4.a.i Informationsanalyse

Bei der Informationsanalyse wird ein Ausdruck für die Entropie an den einzelnen Positionen berechnet. Der Wertebereich liegt zwischen Null und Eins. Eins bedeutet, daß die Position vollständig unbestimmt ist und Null, daß sie eindeutig fest steht. Die Information der Schnittstellen-/Nicht-Schnittstellendaten sind in Abbildung 16 als Histogramm über die einzelnen Positionen dargestellt.

Bei den humanen Schnittstellen ist hiernach die erste Position N-terminal zur Schnittstelle (-1) am wichtigsten für die Spaltung der Signalsequenz. Die Information von 0,5 zeigt, daß nur an dieser Position bestimmte Aminosäuren vorkommen. Ähnliches gilt für die Position -3. Die Positionen -7 und -9 scheinen ebenfalls von Bedeutung zu sein. Die Nicht-Schnittstellensequenzen beinhalten keine wesentliche Information. Für alle Positionen der Nicht-Schnittstellensequenzen berechnet sich ein Informationswert um 0,97. Legt man die relativen Häufigkeiten nach McCaldon et al. [11] zugrunde, errechnet sich eine Information von 0,97, was der Verteilung der Aminosäuren in den Nicht-Schnittstellensequenzen entspricht.

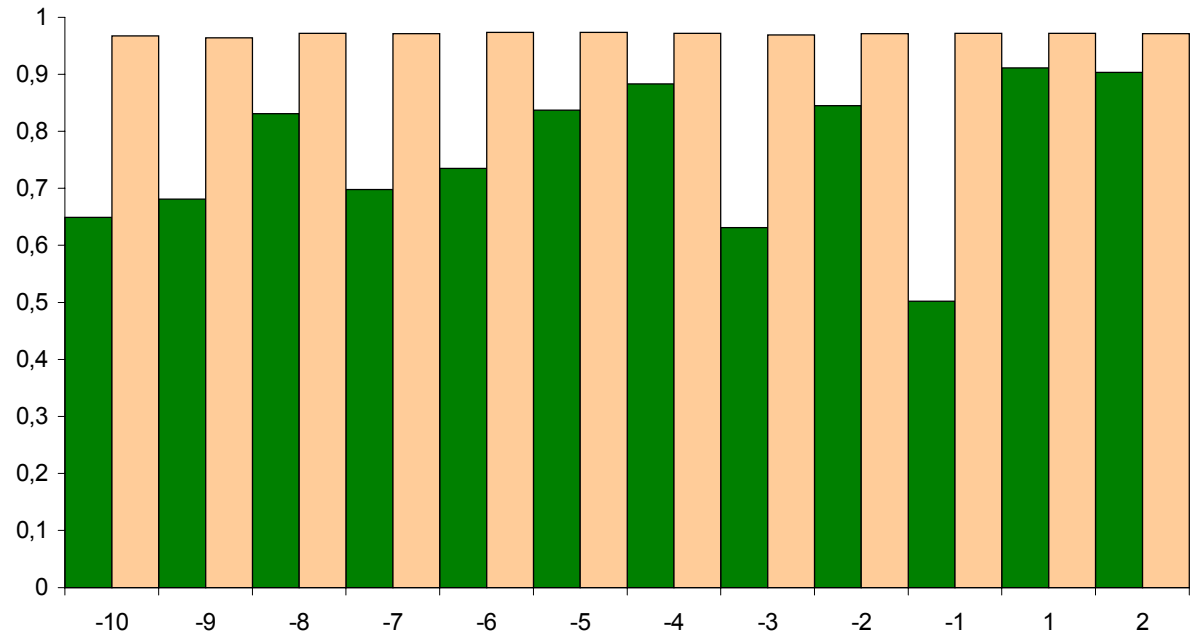

ABBILDUNG 16: Informationsplot der Schnittstellen- (dunkel) und Nicht-Schnittstellensequenzen (hell). Die Schnittstelle befindet sich zwischen Position -1 und 1.

4.a.ii Positionsabhängige Häufigkeitsverteilung von Aminosäuren in Schnittstellenpeptiden

Aus der Informationsanalyse ergibt sich, daß bestimmte Positionen, -1 und -3 einen hohen Informationsgehalt haben. Ohne auf die Nicht-Schnittstellensequenzen einzugehen, werden im folgenden die Verteilungen der Aminosäuren der Schnittstellensequenzen vorgestellt: In Abbildung 17 sind die relativen Häufigkeiten der Aminosäuren an den einzelnen Positionen als Histogramm dargestellt. Die Tabellen 1 und 2 zeigen die relativen Häufigkeiten für die Schnittstellen respektive Nicht-Schnittstellen. An der Position -1 treten nur wenige Aminosäuren auf. In über 70% der Schnittstellensequenzen findet sich entweder Alanin oder Glycin.

An Position -2 wurden weder schwefelhaltige (Methionin und Cystein) noch helixbrechende Aminosäuren (Prolin und Glycin) gefunden.

An der Position -3 sind ebenfalls nur wenig unterschiedliche Aminosäuren vorhanden. Jedoch ist die Variabilität an Aminosäuren etwas höher als an Position -1. So haben knapp 70% der Sequenzen entweder Alanin (28%), Serin (24%) oder Valin (17%) an Position -3.

Die Positionen -6 bis -10 weisen einen hohen Anteil an Leucin auf, welches meistens Bestandteil der hydrophoben Region (H-Region) des Signalpeptids ist.

Somit leisten nach dieser statistischen Untersuchung die Positionen N-terminal der Schnittstelle einen wesentlichen Beitrag für die Erkennung durch die Signalpeptidase.

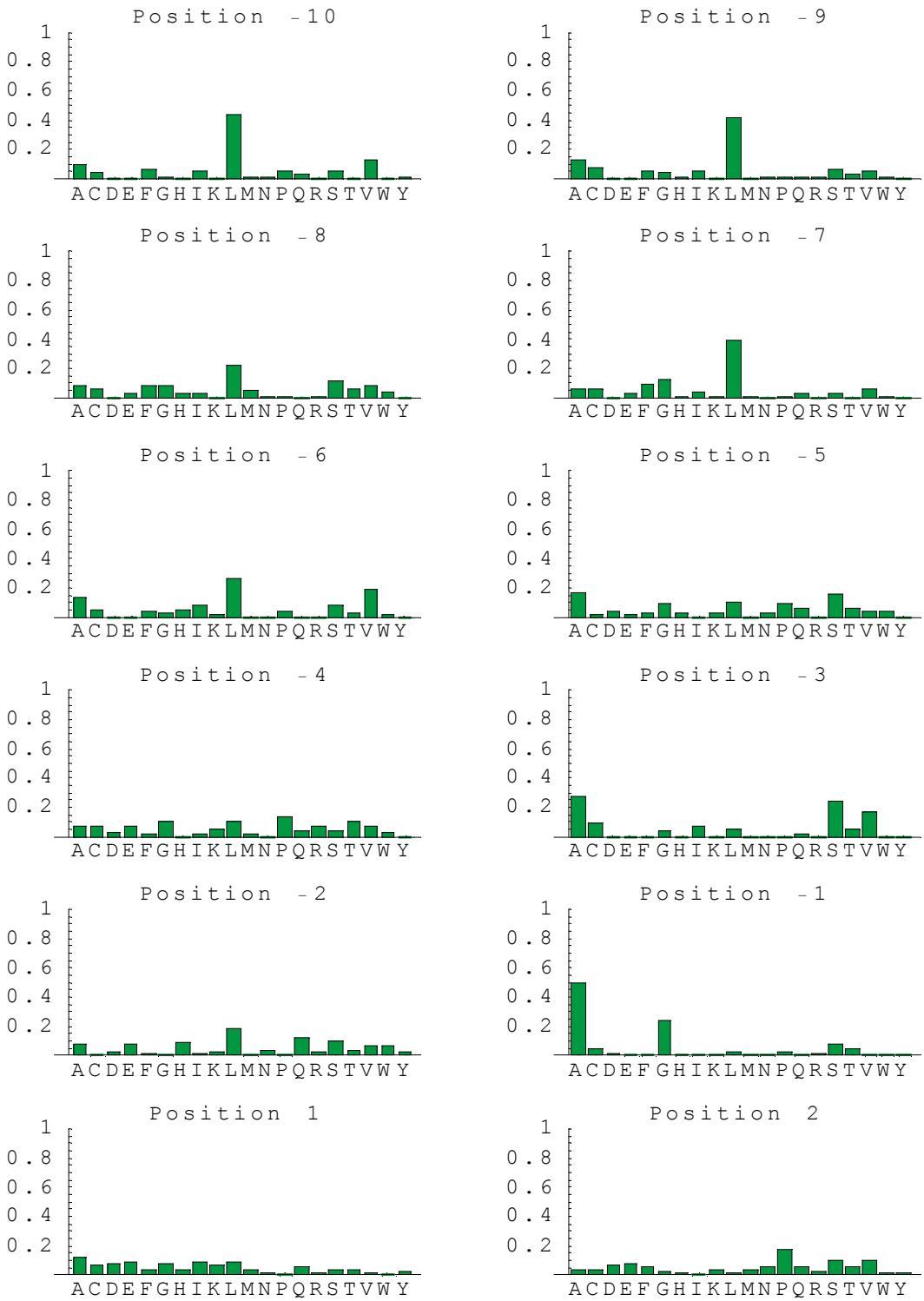

ABBILDUNG 17: Relative Häufigkeiten der Aminosäuren von Schnittstellensequenzen in den angegebenen Sequenzpositionen relativ zur Schnittstelle.

TABELLE 1. Relative Häufigkeit der Aminosäuren an den Positionen relativ zur Schnittstelle für Schnittstellendaten

Pos.	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
-10	0,10	0,04	0	0	0,07	0,01	0	0,05	0	0,43	0,01	0,01	0,05	0,03	0	0,05	0	0,13	0	0,01
-9	0,13	0,08	0	0	0,05	0,04	0,01	0,05	0	0,42	0	0,01	0,01	0,01	0,01	0,07	0,03	0,05	0,01	0
-8	0,08	0,07	0	0,03	0,08	0,08	0,03	0,03	0	0,22	0,05	0,01	0,01	0	0,01	0,12	0,07	0,08	0,04	0
-7	0,07	0,07	0	0,03	0,09	0,13	0,01	0,04	0,01	0,40	0,01	0	0,01	0,02	0	0,03	0	0,07	0,01	0
-6	0,13	0,05	0	0	0,04	0,03	0,05	0,08	0,01	0,26	0	0	0,04	0	0	0,08	0,03	0,18	0,01	0
-5	0,17	0,01	0,04	0,01	0,03	0,09	0,03	0	0,03	0,11	0	0,03	0,09	0,07	0	0,16	0,07	0,04	0,04	0
-4	0,07	0,07	0,03	0,07	0,01	0,11	0	0,01	0,05	0,11	0,01	0	0,13	0,04	0,06	0,04	0,11	0,07	0,03	0
-3	0,28	0,09	0	0	0	0,04	0	0,07	0	0,05	0	0	0	0,01	0	0,24	0,05	0,17	0	0
-2	0,08	0	0,03	0,08	0,01	0	0,09	0,01	0,03	0,18	0	0,04	0	0,12	0,03	0,11	0,04	0,07	0,07	0,03
-1	0,5	0,05	0,01	0	0	0,24	0	0	0	0,03	0	0	0,03	0	0,01	0,08	0,05	0	0	0
1	0,12	0,07	0,08	0,10	0,04	0,08	0,04	0,09	0,07	0,09	0,04	0,01	0	0,05	0,01	0,04	0,04	0,01	0	0,03
2	0,04	0,04	0,07	0,08	0,05	0,03	0,01	0	0,04	0,01	0,04	0,05	0,17	0,05	0,03	0,11	0,05	0,11	0,01	0,01

TABELLE 2. Relative Häufigkeit der Aminosäuren für Nicht-Schnittstellendaten

Pos.	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
-10	0,16	0,06	0,02	0,03	0,04	0,08	0,03	0,04	0,02	0,19	0,01	0,01	0,03	0,03	0,01	0,10	0,05	0,07	0,02	0,01
-9	0,15	0,05	0,03	0,04	0,04	0,08	0,03	0,03	0,02	0,15	0,02	0,01	0,05	0,04	0,02	0,10	0,05	0,08	0,02	0,01
-8	0,15	0,05	0,03	0,04	0,04	0,08	0,03	0,03	0,03	0,13	0,01	0,02	0,06	0,04	0,02	0,09	0,05	0,08	0,02	0,01
-7	0,14	0,05	0,04	0,05	0,03	0,08	0,03	0,03	0,03	0,1	0,02	0,02	0,06	0,05	0,02	0,10	0,05	0,08	0,02	0,01
-6	0,14	0,05	0,04	0,06	0,03	0,08	0,03	0,03	0,04	0,08	0,02	0,02	0,06	0,05	0,03	0,09	0,06	0,06	0,02	0,01
-5	0,12	0,05	0,05	0,07	0,03	0,08	0,03	0,03	0,05	0,07	0,02	0,02	0,07	0,05	0,03	0,08	0,05	0,07	0,02	0,02
-4	0,13	0,05	0,05	0,07	0,03	0,08	0,03	0,04	0,05	0,07	0,02	0,03	0,06	0,05	0,03	0,08	0,04	0,07	0,01	0,03
-3	0,10	0,04	0,06	0,08	0,03	0,08	0,03	0,04	0,05	0,07	0,02	0,03	0,07	0,06	0,04	0,06	0,04	0,06	0,01	0,03
-2	0,10	0,05	0,06	0,08	0,03	0,08	0,03	0,04	0,06	0,06	0,02	0,03	0,08	0,05	0,04	0,06	0,05	0,05	0,01	0,03
-1	0,06	0,05	0,06	0,09	0,03	0,07	0,03	0,04	0,06	0,07	0,02	0,03	0,08	0,05	0,05	0,06	0,05	0,06	0,01	0,03
1	0,05	0,04	0,05	0,09	0,03	0,07	0,03	0,04	0,07	0,06	0,02	0,03	0,09	0,05	0,05	0,06	0,05	0,07	0,01	0,03
2	0,05	0,04	0,05	0,08	0,03	0,08	0,03	0,04	0,08	0,07	0,02	0,03	0,08	0,05	0,05	0,06	0,05	0,07	0,01	0,03

4.a.iii Schwerpunktanalyse

Die Schwerpunktanalyse erlaubt die Klassifikation von Daten aufgrund nicht korrelierter Zusammenhänge, d.h. sie ist ein lineares Trennverfahren. Die Häufigkeitsverteilung der auf die Schwerpunktsgerade projizierten Schnittstellendaten ist in Abbildung 18 wiedergegeben. Tabelle 3 beinhaltet die dazugehörigen Korrelationskoeffizienten (cc-Werte). Die Schwerpunktanalyse als Klassifikationsmethode ergibt, gemittelt über die drei Datensätze der Kreuzvalidierung und einer binären Kodierung, einen Korrelationskoeffizienten von cc = 0,55 für die Testdaten.

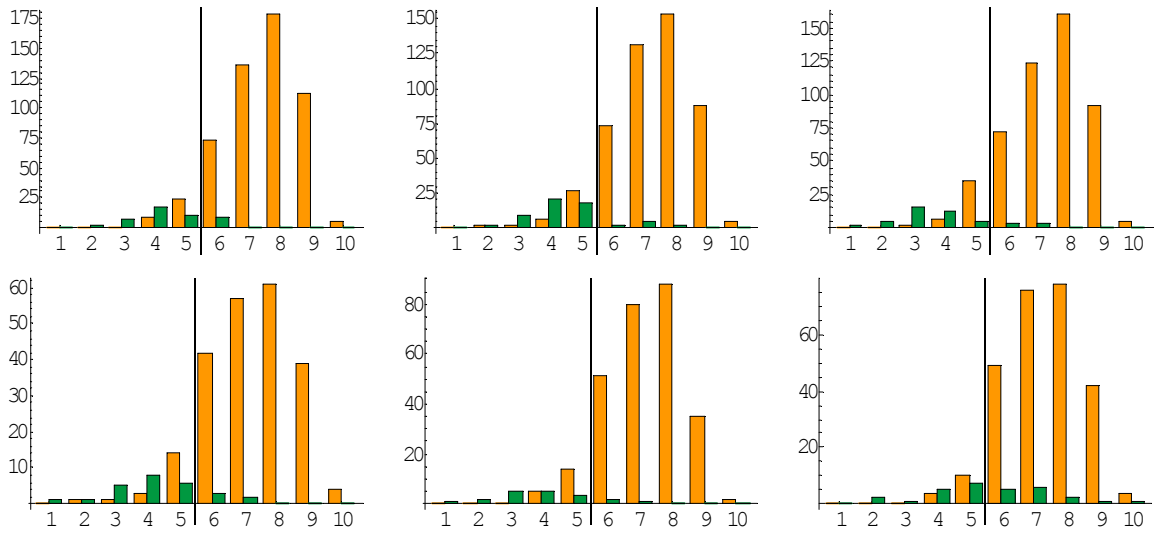

ABBILDUNG 18: Histogramm der Projektionen der Datenpunkte auf die Schwerpunktgerade; dunkelgrau sind die Signalsequenzen, hellgrau die Nicht-Schnittstellensequenzen. In der oberen Zeile sind die drei Trainingssätze in der unteren die entsprechenden Testsätze gezeigt. Die Separationsgerade klassifiziert die Datenpunkte links von ihr als Schnittstellenpunkte, die anderen als Nicht-Schnittstellendatenpunkte.

Somit lassen sich die meisten Schnittstellensequenzen durch ein lineares Trennverfahren von den Nicht-Schnittstellensequenzen unterscheiden. Die Separationsgerade befindet sich in dem 240-dimensionalen Raum, der von den verteilt kodierten Aminosäuren der 12-Aminosäuren langen Sequenzfenster aufgespannt wird. Die Anzahl der zu bestimmenden Parameter ist im Vergleich zur Anzahl der vorhandenen Daten größer als es für eine verläßliche Statistik notwendig ist. Ein kleinerer Datenraum kann durch eine Kodierung mit weniger Dimensionen erreicht werden, jedoch kann dadurch auch keine bessere Klassifikation erreicht werden. Die Aussagen, die in einem kleineren Datenraum erzielt werden, sind statistisch relevanter.

TABELLE 3. Korrelationskoeffizienten (cc) der Schwerpunktanalyse für die drei Datensätze der Kreuzvalidierung und deren Mittelwerte. Es sind die Werte für Training und Test angegeben.

	Teil 1	**Teil 2**	**Teil 3**	**Mittel**
Training	0,61	0,67	0,59	0,62
Test	0,60	0,59	0,45	0,55

4.a.iv Bayes-Prediktor

Eine Auswertung der Daten nach dem Bayes-Prinzip ergab in einem unabhängigen Testdatensatz eine mittlere Vorhersagegenauigkeit von cc = 0,52 (Tabelle 4).

TABELLE 4. Korrelationskoeffizienten (cc) der Bayes Analyse für die drei Datensätze der Kreuzvalidierung (Test 1 bis 3) und deren Mittelwerte. Es sind die Werte für Training und Test angegeben.

	Teil 1	**Teil 2**	**Teil 3**	**Mittel**
Training	0,95	0,94	0,95	0,95
Test	0,39	0,53	0,63	0,52

4.a.v Hauptkomponentenanalyse (PCA)

Die Hauptkomponentenanalyse als Standardverfahren zur Untersuchung von Korrelationen soll Zusammenhänge zwischen einzelnen Aminosäuren an bestimmten Positionen aufzeigen. Abbildung 19 zeigt die Projektion aller Datenpunkte auf die ersten zwei Hauptkomponenten der beiden Klassen (Schnittstellen- und Nicht-Schnittstellendaten).

Der betrachtete Datenraum besteht aus 240 Dimensionen. Für jede der 12 Positionen in der Sequenz wird für alle 20 Aminosäuren jeweils eine Dimension aufgespannt. Die ersten beiden Hauptkomponenten charakterisieren die Ebene mit der maximalen Streuung. Diese Ebene wird in der Regel nicht parallel zu einer ursprünglichen Ebene sein. Die beiden ersten Hauptkomponenten können somit nicht mehr direkt bestimmten Aminosäuren zugeordnet werden; sie repräsentieren vielmehr die Kombination der häufigsten Aminosäuren. Die Streuung der Schnittstellendaten ist wesentlich größer als die der Nicht-Schnittstellen. Ein Vergleich der Eigenwerte (Abbildung 20) mit einer entsprechenden Eigenwertanalyse von Zufallszahlen ergibt, daß die ersten beiden Eigenwerte der Signaldaten um mehr als 10% abweichen, während alle anderen Werte um weniger als 9% abweichen. Wenn, wie in diesem Beispiel, die Ausdehnung der Datenwolken so unterschiedlich ist, heißt das, daß unterschiedliche Aminosäuren oder auch Eigenschaften für die Daten wichtig zur Charakterisierung sind. Die relativ

kleinen Unterschiede zwischen den ersten Eigenwerten der Nicht-Schnittstellendaten läßt auf eine gleichmäßigere Verteilung schließen.

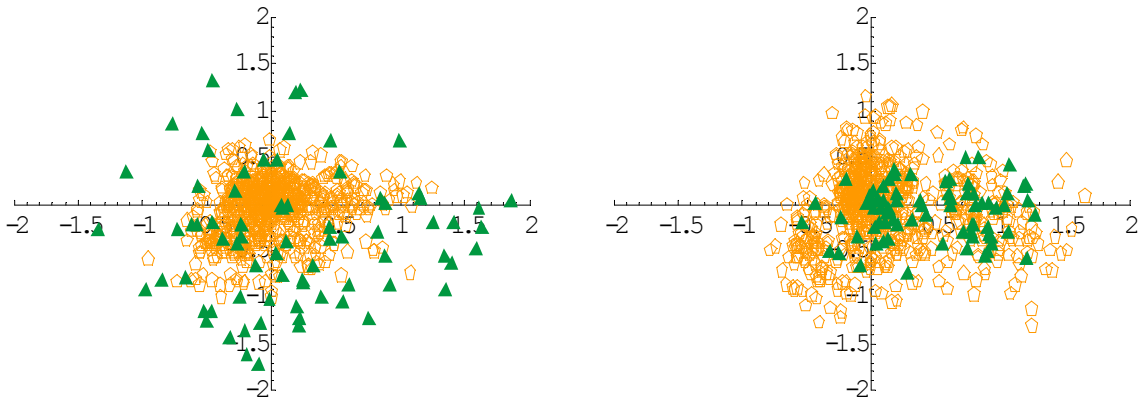

ABBILDUNG 19: Projektion der Datenpunkte (binäre Kodierung) auf die beiden ersten Hauptkomponenten der Signalsequenzen (links) und der Nicht-Signalsequenzen (rechts). Die Signalsequenzen sind durch kleine Kreise dargestellt, die Nicht-Signalsequenzen durch kleine Dreiecke.

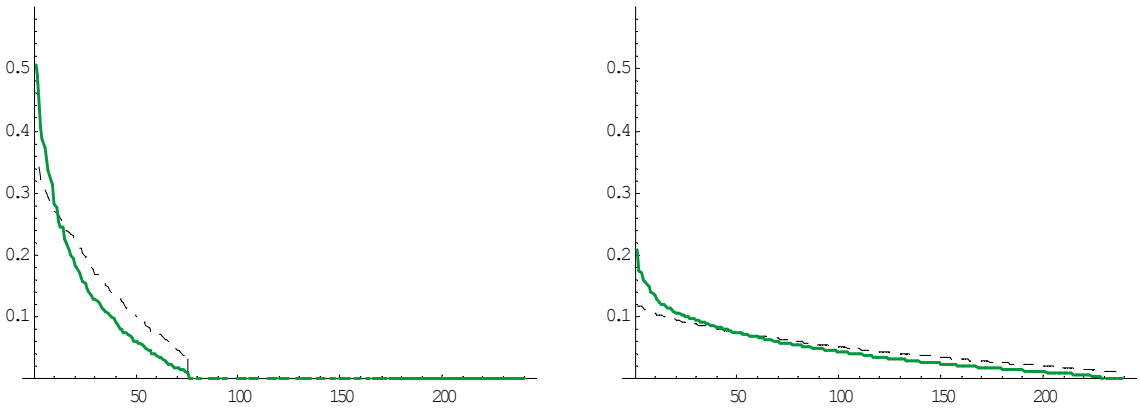

ABBILDUNG 20: Eigenwerte der Schnittstellendaten (links) und Nicht-Schnittstellendaten (rechts) im Vergleich zu jeweils der gleichen Anzahl von Zufalls-Daten (gestrichelte Linien).

Der Betrag der Eigenvektorkomponenten (Abbildung 21) entspricht dem Beitrag der entsprechenden Aminosäure an einer bestimmten Position der Sequenz zu der jeweiligen Eigenschaft (z.B. Schnittstelle). Je größer der Betrag, desto weiter ist ein Datenpunkt vom Koordinatenursprung bei der Projektion auf die Hauptkomponente entfernt, und desto einfacher ist er von z.B. Nicht-Schnittstellendaten zu unterscheiden. Bei den Signalsequenzen sind die Aminosäuren mit dem größten Anteil entlang der ersten Hauptkomponente die hydrophoben Aminosäuren Leucin an Position -9 und Alanin an -3 und -1. Entlang der zweiten Hauptkomponente sind es die ebenfalls hydrophoben Aminosäuren Leucin an -10 und -7 sowie Alanin an -1. Glycin

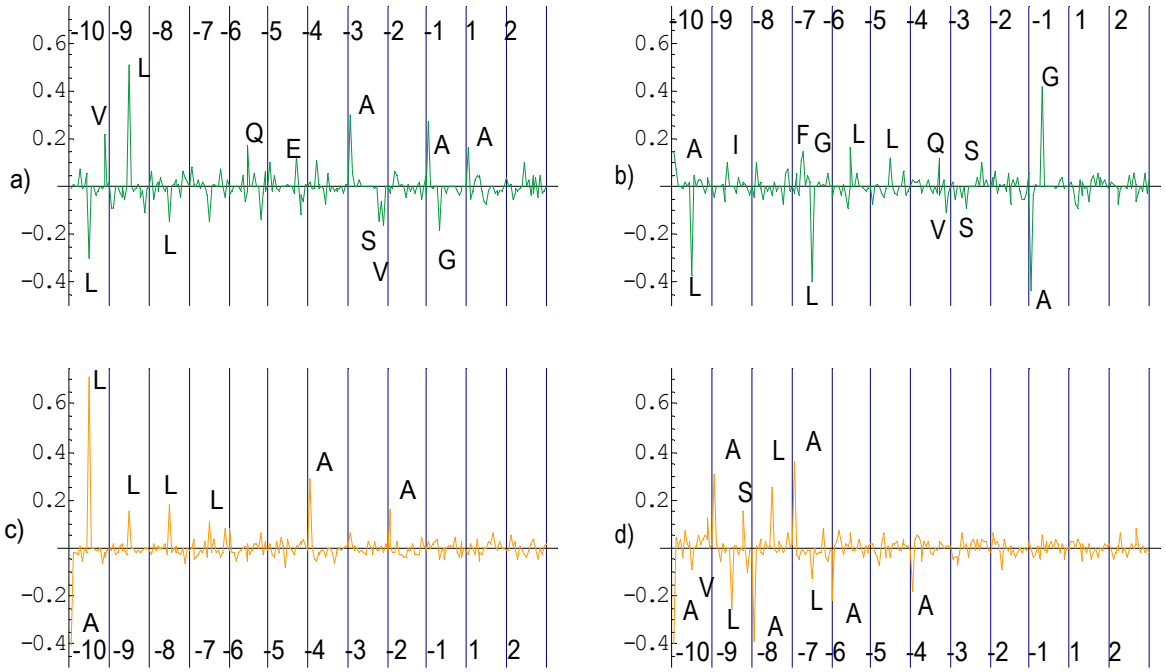

ABBILDUNG 21: Eigenvektoren der ersten (a, c) und zweiten (b, d) Hauptkomponenten der Signalsequenzen (a, b) und der Nicht-Signalsequenzen (c, d) bei einer verteilten Kodierung des 12er-Fensters.

an -1 hat ebenfalls einen großen Anteil, jedoch nicht in Kombination mit Leucin an -10 und -7. Da das Punktprodukt zwischen Datenvektor und Eigenvektor den Koordinatenabschnitt der x-Achse ergibt, wird bei einer Sequenz LXXLXXXXXGXX der Wert niedriger ausfallen. Es können so folgende Sequenzen als typische Schnittstellensequenzen bestimmt werden:

VLTALQEAEA↓AK
LVLLSSKVTG↓LD
AICGLLQSDG↓QV
LGSLIAVILA↓EF

ABBILDUNG 22: Typische Schnittstellensequenzen, die sich aus der Hauptkomponentenanalyse ergeben. "↓" gibt hierbei die Position der Schnittstelle an.

Bei den Nicht-Signalsequenzen fallen vor allem die Aminosäuren Glutamin und Glutaminsäure sowie Alanin, Valin und Leucin auf. Die Werte der Eigenvektorkomponenten sind jedoch nicht so groß wie bei den Signalsequenzen, haben also einen kleineren Beitrag zur Gesamtlage. Aus Abbildung 19 ergibt sich jedoch, daß diese Eigenschaften nicht relevant für die Beschreibung der Signalsequenzen sind, da hier die Ausdehnung entlang der Achsen nur minimal ist. Es können trotzdem noch die Konsensussequenzen, die sich aus den Eigenvektoren ergeben, abgeleitet werden:

LLLLAGAAAQPC
AKVDEKPKIGGG
VALAEDGDCGEQ
ALALAEAEGVGC

ABBILDUNG 23: Typische Nicht-Schnittstellensequenzen, die sich aus der Hauptkomponentenanalyse ergeben. "↓" gibt hierbei die Position der Schnittstelle an.

Diese Sequenzen sollten nicht von der Signalpeptidase gespalten werden.

Eine Trennung der beiden Klassen ist mit dieser Methode jedoch nicht möglich. Die Projektion auf die ersten beiden Hauptkomponenten ermöglicht nur die Visualisierung von paarweisen Korrelationen.

4.a.vi Mahalanobis-Distanzanalyse

Mit der Mahalanobis-Distanzanalyse lassen sich zwei Datenmengen durch ellipsoide Funktionen um die Schwerpunkte der jeweiligen Datenwolken charakterisieren und damit voneinander unterscheiden. Sie ist als eine Erweiterung der Hauptkomponentenanalyse um die Klassifikationsmöglichkeit zu verstehen. Die Daten werden mit den binären Eigenschaften klein und positiv kodiert, um den Datenraum klein zu halten, damit eine Inversion der Korrelationsmatrix möglich ist. Hier lassen sich die Daten gut unterscheiden, die zur Bestimmung des Systems verwendet wurden. Eine Verallgemeinerung ist jedoch nur schlecht möglich (Tabelle 5). Es wurde eine Kreuzvalidierung mit drei Datenteilen durchgeführt und anschließend über die drei Ergebnisse gemittelt.

TABELLE 5. Korrelationskoeffizienten (cc) der Mahalanobis-Distanzanalyse für die drei Datensätze der Kreuzvalidierung und deren Mittelwerte. Es sind die Werte für Training und Test angegeben.

	Teil 1	**Teil 2**	**Teil 3**	**Mittel**
Training	0,81	0,81	0,80	0,81
Test	0,19	0,31	0,24	0,25

4.a.vii Künstliche Neuronale Netze

Bisher konnte keine zufriedenstellende Klassifikation der Schnittstellendaten von den Nicht-Schnittstellendaten erreicht werden. Alle bisher verwendeten Methoden zeigten, daß eine solche Klassifikation nur bis zu einem bestimmten Grad möglich ist. Es wird jetzt eine Methode angewendet werden, die eine geeignete Kodierung der Daten berücksichtigt, um eine für den kleinen Datensatz ausreichende statistische Relevanz zu erlangen.

Für eine weitergehende Analyse zur Merkmalsanalyse der Schnittstellensequenzen wurden Künstliche Neuronale Netze (KNN) verwendet. Dabei wurden sowohl die Lernstrategien (Kapitel 3.g.iv) als auch verschiedene Netztypen angewendet. Es werden vergleichende Untersuchungen zur Lernstrategie, zur Kodierung (KNN vs. ACN) und zur Netzwerkarchitektur durchgeführt. Die Ergebnisse der Netze mit adaptiver Kodierung erlauben auch weitergehende Untersuchungen bezüglich der relevanten Eigenschaften und Positionen in der betrachteten Sequenz. Die Kohonennetze liefern nützliche Resulte, die bei der Beurteilung des Datensatzes von Bedeutung sind. Die gewonnenen Erkenntnisse wurden zur Vorhersage von möglichen Schnittstellensequenzen auf dem World Wide Web (WWW) zur Verfügung gestellt.

4.a.vii.I Lernstrategien

Es wurde die Leistung zweier Lernalgorithmen für Künstliche Neuronale Netze verglichen. Dabei wurden verschiedene binäre Kodierungen verwendet, was auf der einen Seite eine geringe Anzahl von Parametern garantiert, zum anderen den Rechenaufwand gering hält. Die acht physikochemischen Eigenschaften nach Taylor (klein, sehr klein, aliphatisch, aromatisch, hydrophob, negativ, polar und positiv) [118] und alle paarweisen Verknüpfungen werden verwendet.

Eine $(1, 20)$ Evolutionsstrategie und der Backpropagation-Algorithmus wurden über 200 beziehungsweise über 500 Generationen jeweils mit einer Anfangslernrate von $\sigma = 0.5$ trainiert. Die weitaus meisten Netze konvergieren nach dieser Zeit. Nach dem Training eines Netzes wurde das Netz mit dem besten Testergebnis verwendet, um die Korrelationskoeffizienten (cc-Werte) für Training und Test zu berechnen. Es wurde eine dreifache Kreuzvalidierung durchgeführt. Alle Rechnungen werden dreimal wiederholt und über die neun Werte gemittelt. Die gemittelten cc (Test) Werte sind für die verschiedenen Kodierungen in Tabelle 6 und 7 zusammengefaßt.

Die Varianzen (s^2) der Korrelationskoeffizienten liegen zwischen $s^2 = 0{,}003$ und $s^2 = 0{,}026$. Die Trainingswerte sind zum Teil erheblich besser als die Testwerte, die hier aufgeführt sind, jedoch ist das Testergebnis entscheidend für die Bewertung der Generalisierungsfähigkeit.

TABELLE 6. Lernstrategie: (1,20)-Evolutionsstrategie. Korrelationskoeffizienten der Testläufe. s = klein (small); t = sehr klein (tiny); alp = aliphatisch; aro = aromatisch; h = hydrophob; neg = negativ; pol = polar; pos = positiv.

		s	**t**	**alp**	**aro**	**h**	**neg**	**pol**	**pos**
		0,32	0,46	0,43	0,05	0,18	0,04	0,28	0,04
s	0,32	-	0,49	0,40	0,30	0,46	0,45	0,57	0,32
t	0,46		-	0,43	0,30	0,23	0,39	0,10	0,25
alp	0,43			-	0,43	0,37	0,05	0,17	0,23
aro	0,05				-	0,23	0,09	0,31	0,39
h	0,18					-	0,32	0,28	0,33
neg	0,04						-	0,05	0,15
pol	0,28							-	0,40
pos	0,04								-

TABELLE 7. Lernstrategie: Backpropagation-Algorithmus. s = klein (small); t = sehr klein (tiny); alp = aliphatisch; aro = aromatisch; h = hydrophob; neg = negativ; pol = polar; pos = positiv.

		s	**t**	**alp**	**aro**	**h**	**neg**	**pol**	**pos**
		0,24	0,40	0,36	0,10	0,03	0,07	0,20	0,12
s	0,24	-	0,42	0,35	0,21	0,44	0,35	0,45	0,23
t	0,40		-	0,33	0,25	0,14	0,28	0,18	0,10
alp	0,36			-	0,33	0,27	0,05	0,14	0,09
aro	0,10				-	0,10	0,14	0,28	0,22
h	0,03					-	0,21	0,27	0,27
neg	0,07						-	0,09	0,08
pol	0,20							-	0,32
pos	0,12								-

Qualitativ gibt es keine Unterschiede zwischen den Ergebnissen der beiden Lernalgorithmen. So ergaben sich in beiden Fällen mit der Kombination aus „*polar*“ und „*klein“* die besten Testwerte. Auch die Eigenschaften „*sehr klein*“ und „*aliphatisch*“ als Einzelkodierungen wiesen gute Werte auf. Nur wenn die Vorhersagegenauigkeit sehr klein wird, liefert der Backpropagation-Algorithmus bessere Werte. Die Aussage, daß eine korrekte Identifizierung der Testsequenzen nicht möglich ist, wird dadurch jedoch nicht beeinflußt.

Die Evolutionsstrategie liefert etwas bessere Resultate im Vergleich zum Backpropagation-Algorithmus. Für eine vollständige Klassifikation sind jedoch cc-Werte nahe Eins erforderlich.

4.a.vii.II Vergleich von Kodierungsmethoden

Eine Vielzahl von Experimenten wurde durchgeführt, um den Einfluß von unterschiedlichen Kodierungsmethoden auf die Netzgüte zu untersuchen. Die Resultate dieser Experimente sind in Tabelle 8 wiedergegeben. Jedes Experiment besteht aus zwei Schritten: 1. dem Training des Neuronalen Netzes für jeden der drei Kreuzvalidierungs Trainingssätze und 2. der Bestimmung des Korrelationskoeffizienten für die Trainings und Testdaten. Es wurden jeweils 30 Experimente (90 Netzwerke) berechnet.

Die ersten fünf Zeilen zeigen die Ergebnisse, die mit drei unterschiedlichen Kodierungsansätzen und einer Perzeptron Netzarchitektur erzeugt wurden. Die Zeilen sechs bis 15 verwenden den gleichen Kodierungsansatz wie die Zeilen eins bis fünf, jedoch mit einem zwei-lagigen Netz mit drei (Zeilen 5 bis 10) und fünf (Zeilen 11 bis 15) Hiddenschichten. Der Korrelationskoeffizient (cc) und die Standardabweichung, gemittelt über 90 Netze (30 Experimente), sind für Training und Test dargestellt. Desweiteren wird der Rechenaufwand, der sich aus der Anzahl der berechneten Multiplikationen in Tausend ergibt, und die Anzahl der freien Parameter, die optimiert werden mußten, angegeben. Eins, zwei und drei Kodierungsvektoren wurden für die adaptive Kodierung (ACN) verwendet.

TABELLE 8. Ergebnisse der Neuronalen Netze für verschiedene Kodierungsansätze und Netzarchitekturen. Es sind die mittleren Korrelationskoeffizienten (cc) für Training (Train) und Testläufe (Test) sowie deren Standardabweichung (s) angegeben. Der Rechenaufwand wurde als Anzahl der zu berechnenden Multiplikationen in 1000 berechnet. Die Anzahl der freien Parameter gibt die Komplexität des Netzes wieder.

Zeile	Netz-architektur	Kodierung	cc (Train)	s (Train)	cc (Test)	s (Test)	Rechenaufwand [*1000]	Anzahl der freien Parameter
	Perzeptron							
1		verteilt	0,96	0,04	0,52	0,09	77,7	241
2		Vol/Hyd/ Pol	0,78	0,06	0,48	0,08	11,7	37
3		1-Kode	0,78	0,07	0,52	0,08	10,4	33
4		2-Kodes	0,91	0,06	0,55	0,10	20,81	65
5		3-Kodes	0,93	0,06	0,56	0,10	32,1	97
	2-lagig							
6	3-hidden schichten	verteilt	0,91	0,09	0,45	0,10	170,1	728
7		Vol/Hyd/ Pol	0,88	0,08	0,38	0,10	26,4	116
8		1-Kode	0,74	0,11	0,50	0,14	13,9	63
9		2-Kodes	0,88	0,13	0,55	0,15	27,0	119
10		3-Kodes	0,93	0,06	0,54	0,09	37,7	175
11	5-hidden schichten	verteilt	0,92	0,12	0,41	0,11	270,2	1212
12		Vol/Hyd/ Pol	0,90	0,07	0,39	0,11	45,4	192
13		1-Kode	0,77	0,08	0,53	0,11	19,6	91
14		2-Kodes	0,89	0,09	0,54	0,11	37,3	171
15		3-Kodes	0,91	0,08	0,48	0,13	57,1	251

Die besten Vorhersageergebnisse lassen sich mit der Adaptiven Kodierung, Perzeptron Architektur und drei Kodierungsvektoren erzielen (Zeile fünf). Unabhängig von der Netzwerkarchitektur werden mit der ACN Methode bessere Vorhersagen erzielt als mit anderen Kodierungsmethoden. Der Trainingsdatensatz wird am besten mit der verteilten Kodierung wiedergegeben, jedoch sind die Testergebnisse nicht besser als mit der ACN Methode. Hinsichtlich des Rechenaufwandes ist die verteilte Kodierung (Zeilen 1, 6, 11) 8 bis 12 mal aufwendiger als vergleichbare Ansätze mit weniger Parametern (Zeilen 3, 8, 13).

Die Einführung von Hiddenschichten verbessert die Ergebnisse der eindimensionalen Kodierung im Vergleich zur Perzeptron Architektur nicht. Bei der physikochemischen Kodierung verbessern sich die Ergebnisse der Trainingsläufe im Mittel unter Verwendung komplexerer Netzwerke, jedoch verschlechtern sich gleichzeitig die Testergebnisse. (Zeilen 2, 7, 12).

Im allgemeinen erhöhte sich die Zuverläßlichkeit von Vorhersagesystemen mit sinkender Anzahl von freien Parametern. Dies wurde deutlich am kleiner werdenden Unterschied zwischen cc (Train) und cc (Test). Vier Netze besaßen einen kleineren Unterschied als alle anderen Netze; das ist das Netz mit Perzeptron Architektur und physikochemischer Kodierung (Zeile 2) und die drei ACN Architekturen mit einem Kodierungsvektor (Zeilen 3, 8, 13).

Da die bis hierher verwendeten 76 Sequenzdaten nicht ausreichten, um eine zufriedenstellende Ähnlichkeit zwischen cc (Train) und cc (Test) zu erzielen, wurde ein wesentlich größerer Datensatz von Nielsen et al. [80] verwendet. Dieser Datensatz umfaßt 416 Schnittstellensequenzen, die aus der SWISSPROT Datenbank stammen und nicht-homolog sind. Jedoch sind diese Daten nicht explizit auf experimentelle Hinweise untersucht worden, was aus oben genannten Gründen wichtig für eine Aussage ist. Der Zusammenhang zwischen Anzahl der Parameter und Anzahl der Trainingssequenzen wird in Abbildung 24 wiedergegeben. Hier sind die über 9 Berechnungen gemittelten cc-Werte für Training und Test über die Anzahl der Trainingssequenzen für einen und zwei Kodierungsvektoren eines ACN abgebildet. Für den eindimensionalen Fall (ein Kodierungsvektor) mit 33 Gewichten reichen etwa 200 Sequenzen (6/10) als Trainingsdaten, um konsistente Trainings- und Testergebnisse zu produzieren. Mehr freie Parameter (zwei Kodierungen) bedürfen mehr Sequenzbeispiele. Konsistente Ergebnisse werden im Falle der zwei Kodierungvektoren erst mit dem fast vollständigen Datensatz erhalten.

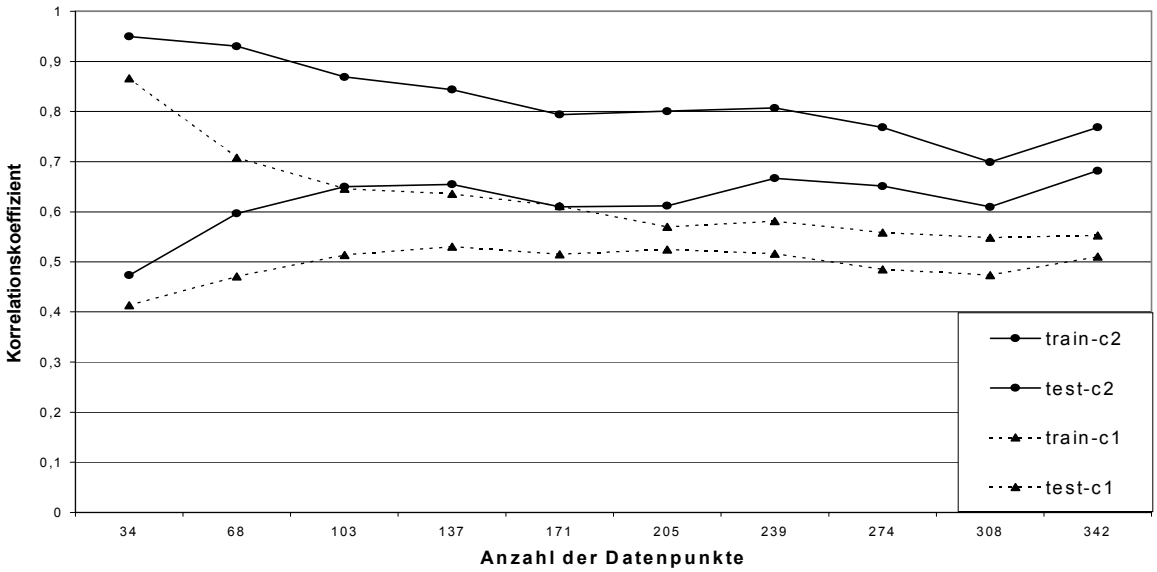

ABBILDUNG 24: Korrelationskoeffizient von Training und Test in Abhängigkeit von der Anzahl der Datenpunkte für ein ACN mit einer und zwei adaptiven Kodierungen. Ein Datensatz von Nielsen et al. [80], bestehend aus 416 Schnittstellensequenzen, wurde verwendet. 4160 Nicht-Schnittstellensequenzen wurden aus den ersten 10 Sliding-window-Positionen der maturen Proteine erzeugt, wie es bereits in Kapitel 2.a für die 76 experimentell verifizierten Schnittstellen gezeigt wurde. Die beiden Datensätze werden in drei Teile aufgeteilt, von denen zwei zum Training verwendet wurden und einer als Testdatensatz dient. Der Trainingsdatensatz wird wiederum in 10 etwa gleich große Teile aufgeteilt und durch sukzessives Hinzufügen werden 10 Datensätze erhalten mit ansteigender Größe. Für jeden dieser Trainingsdatensätze wurden 9 Netze trainiert mit einem und zwei Kodierungsvektoren. Der Testdatensatz blieb immer der gleiche. Es wurde aufgrund des großen Rechenaufwands nur ein Kreuzvalidierungsdurchgang durchgeführt.

Nach von Heijne [34] sind an der Position -1 relativ zur Schnittstelle nur kleine hydrophobe Aminosäuren erlaubt. Deshalb sind die Aminosäuren Asparagin, Glutamin, Phenylalanin, Lysin, Asparaginsäure, Glutaminsäure, Arginin und Tryptophan an dieser Position sehr unwahrscheinlich. In Tabelle 9 sind die Häufigkeiten der Aminosäuren an genau dieser Position aufgelistet. Nimmt man die Häufigkeit der unwahrscheinlichen Aminosäuren an der Position -1 als Qualitätskriterium, so sind 2.6% der experimentell verifizierten Daten fehlerhaft, gegenüber 3.8% im nicht homologen Datensatz von Nielsen et al. und 3.1% in der gesamten SWISSPROT. Asparaginsäure und Tyrosin sind die einzigen Aminosäuren, die in keinem Datensatz an der Position -1 vorkommen.

TABELLE 9. Anzahl der Aminosäuren an Position -1 der Schnittstellensequenzen der verschiedenen Datensätze. red. Homologie bereinigter Datensatz nach [80]. tot.: gesamter Datensatz an Schnittstellensequenzen nach [80]. 76 Sequenzen, die experimentell verifiziert worden sind.

	A	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**	**Σ**
red	186	25	2	2	1	99	0	1	1	5	2	0	15	6	3	49	14	4	1	0	416
tot	326	29	3	2	1	141	0	1	2	5	2	0	15	6	4	54	16	6	1	0	614
76	38	4	1	0	0	18	0	0	0	2	0	0	2	0	1	6	4	0	0	0	76

Eine Hauptkomponentenanalyse der 3*90=270 Kodierungsvektoren der dreidimensionalen adaptiven Kodierung liefert Hinweise auf eine mögliche Konsensussequenz. Die ersten beiden Eigenvektoren der Hauptkomponentenanalyse sind wesentlich größer als der Rest (Abbildung 25a). Wie aus dem Vergleich mit einer einfachen Häufigkeitsanalyse der Schnittstellensequenzen hervorgeht, sind die Unterschiede bei der Analyse der Kodierungsvektoren wesentlich ausgeprägter.

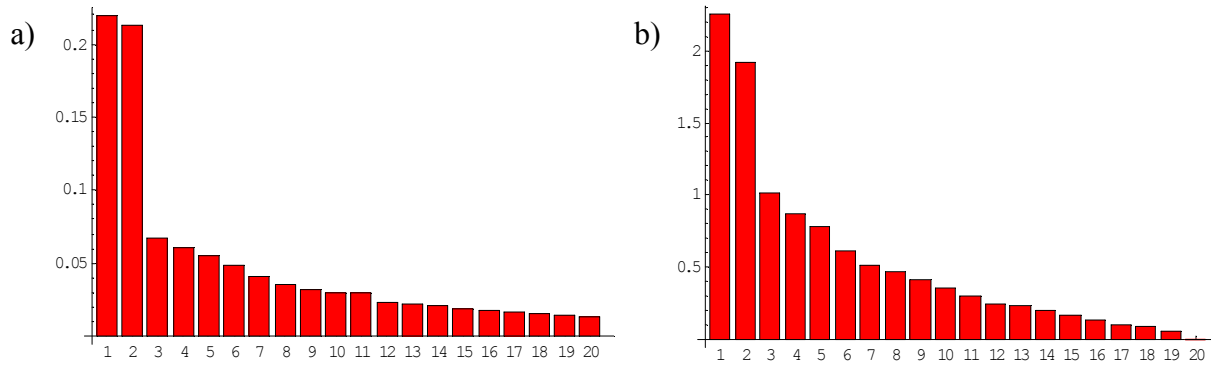

ABBILDUNG 25: Eigenwerte, berechnet aus den 90 Kodierungsvektoren a) und die Eigenwerte der Schnittstellendaten in binärer Kodierung b).

Die Projektion der 20 Aminosäuren auf die entsprechenden Eigenvektoren ist in Abbildung 26 dargestellt. Hier zeigt sich für den Fall der Kodierungsvektoren (a) eine deutlich ausgeprägtere Verteilung der Aminosäuren als für die reine Häufigkeitsuntersuchung (b). Hierbei wird die Hauptkomponentenanalyse für die relativen Häufigkeiten der Aminosäuren an den einzelnen Positionen durchgeführt, wobei nur die Aminosäuren Alanin, Serin und Leucin auffallen.

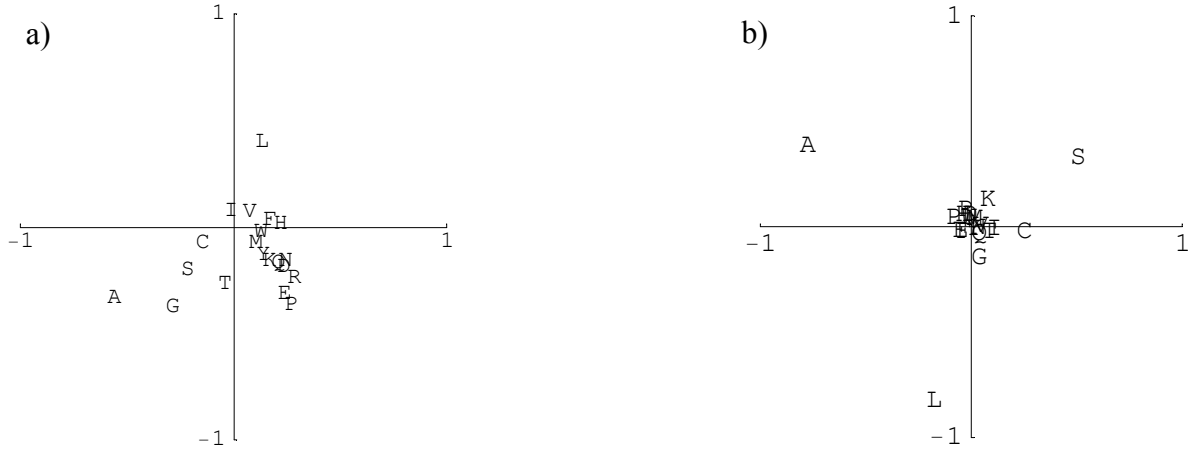

ABBILDUNG 26: Projektion der Aminosäuren auf die ersten beiden Hauptkomponenten der 90 Kodierungsvektoren a) und auf die verteilt kodierten Schnittstellendaten b).

Das Perzeptron mit adaptiver Kodierung, 3 Kodierungsvektoren und der besten Vorhersagegenauigkeit hat die Korrelationswerte cc (Training) = 0,97 und cc (Test) = 0,80.

140 bekannte physikochemische Eigenschaftsvektoren werden mit diesem Basissystem verglichen (vergleiche Kapitel 3.c). Die Eigenschaft mit der besten Übereinstimmung ist „*polar oder Prolin*“ (Tabelle 10). Dieser Eigenschaftsvektor hat die Länge 0,87, das heißt er ist zu 87% im Basissystem der adaptiven Kodierung repräsentiert. Auffällig ist, daß die ersten 10 Eigenschaftsvektoren alle binär kodiert sind.

TABELLE 10. Physikochemische Eigenschaften im Basissystem der adaptiven Kodierung angegeben. In Klammern sind die Verweise auf die Eigenschaften im Anhang (Kapitel 9)

Vektorbetrag	**Physikochemische Eigenschaft**
0.87	Polar oder Prolin (67)
0.86	Sehr klein oder (polar minus aromatisch) oder Prolin (80)
0.84	Sehr klein oder (negativ und hydrophil) oder Threonin oder Prolin (78)
0.84	Geladen oder hydrophil oder Prolin (66)
0.80	Sehr klein oder (polar minus aromatisch) (79)
0.80	(Polar minus aromatisch) oder geladen oder Prolin (68)
0.80	Polar (55)
0.80	((Polar minus aromatisch) minus positiv) oder Prolin (70)
0.80	Sehr klein oder (negativ und hydrophil) oder Threonin (77)
0.80	(Polar minus aromatisch) oder geladen (69)

In Kombination mit anderen physikochemischen Eigenschaften werden jedoch reell kodierte Eigenschaften bevorzugt (Tabelle 11). Hier werden alle dreidimensionalen Basissysteme, die

sich aus Kombination der 140 Eigenschaften ergeben, mit dem Basissystem der Adaptiven Kodierung verglichen. Dabei zeigt sich, daß die Eigenschaftskombination (*Hydrophilizität* (12), *Hydrophobizität* (47) sowie *polar oder Prolin* (67)) die größte Übereinstimmung ergibt. Desweiteren kommen die Eigenschaften *Flexibilität* (5) und *Fraga Recognition Factors* (39) gehäuft vor. Die größten Übereinstimmungen werden erzielt, wenn die Eigenschaften möglichst orthogonal zueinander sind und zudem möglichst gut im Basissystem der Adaptiven Kodierung abgebildet werden. Deshalb tritt hier nur noch die binäre Eigenschaft *polar oder Prolin* (67) auf.

TABELLE 11. Eigenschaftskombinationen mit den besten Übereinstimmungen zur gefundenen Kodierung. In Klammern sind die Verweise auf die Eigenschaften im Anhang (Kapitel 9).

Kodierung 1	**Kodierung 2**	**Kodierung 3**
Hydrophilizität (12)	Hydrophobizität (47)	Polar oder Prolin (67)
Flexibilität (5)	Masse (27)	Fraga Recognition Factors (39)
Flexibilität (5)	Masse (28)	Fraga Recognition Factors (39)
Flexibilität (5)	Umfang(7)	(Klein und hydrophil) oder sehr klein (88)
Flexibilität (5)	Hydrophobizität (50)	Refraktivität (51)
Flexibilität (5)	Volumen (46)	Fraga Recognition Factors (39)
Flexibilität (5)	Versteckte Fläche (3)	Fraga Recognition Factors (39)
Hydrophilizität (12)	Hydrophobizität (47)	((Polar minus aromatisch) minus positiv) oder Prolin (88)
Umfang(7)	pKa des N-Terminus(38)	(Klein und hydrophil) oder sehr klein (88)
Volumen (46)	Polarität (32)	Fraga Recognition Factors (39)
Umfang (7)	Polarität (32)	(Klein und hydrophil) oder sehr klein (88)

Desweiteren ist die Gewichtung der unterschiedlichen 12 Positionen in der Sequenz interessant. Dazu wird wiederum eine Hauptkomponentenanalyse der 90 Gewichtsvektoren des ACN mit Perzeptron Architektur und 3 Kodierungsvektoren durchgeführt. Hierbei ergibt sich nur ein sehr großer Eigenwert (Abbildung 27a), was darauf hindeutet, daß die gelernten Gewichtsvektoren sich nur unwesentlich voneinander unterscheiden. Die Komponenten des Eigenvektors (Abbildung 27b) führen zu der Schlußfolgerung, daß die Positionen -1, -3, -6 und -7 von besonderer Bedeutung sind. Diese Resultate bestätigen die (-1, -3)-Regel von vonHeijne [34] und erweitern sie.

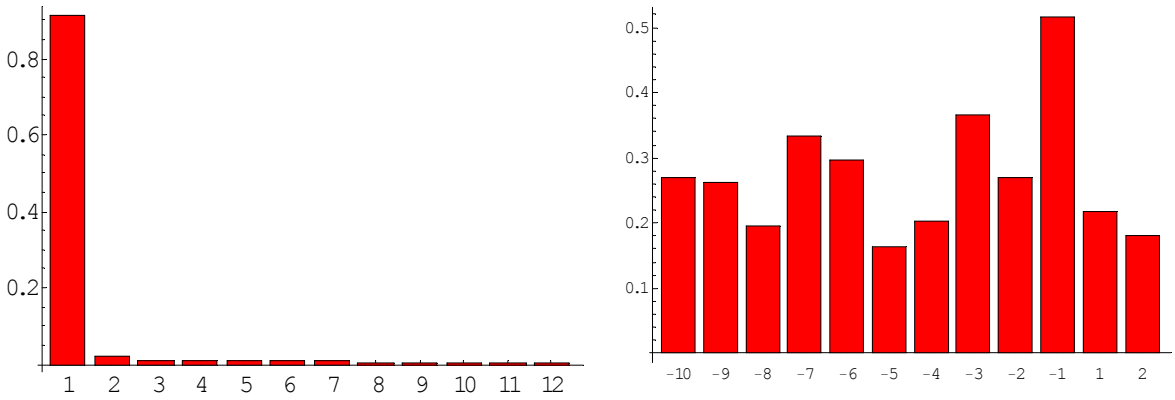

ABBILDUNG 27: Hauptkomponentenanaylse der Gewichtsvektoren: a) Eigenwerte b) der erste Eigenvector.

Im Falle des Perzeptron können die Gewichte des Neuronalen Netzes mit der adaptiv entwikkelten Kodierung multipliziert werden, um einen detaillierten Einblick in Gewichtung der einzelnen Aminosäuren an den einzelnen Positionen zu erhalten. Dies ist graphisch in Abbildung 28 dargestellt und in Tabelle 12 sind die Werte explizit aufgeführt. Die graphische Analyse zeigt, daß die Aminosäuren Alanin, Phenylalanin und Arginin an Position -1 von besonderer Gewichtung sind. Daneben fallen aber auch Leucin an Position -6, Serin an den Positionen -1 und -3, sowie Glycin an -1 und Alanin an Position -3 auf.

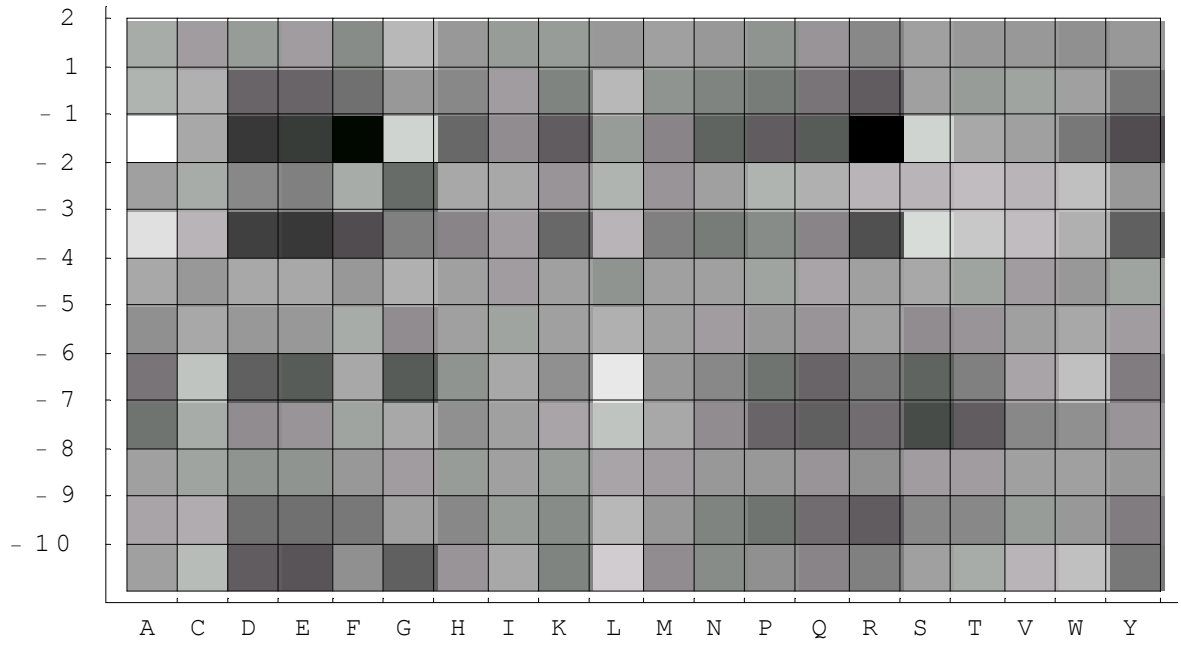

ABBILDUNG 28: Histogramm der Gewichte mit summierten Kodierungen. Die Helligkeit gibt die Tendenz für Schnittstellen an.

Aus der Tabelle 12 können die optimalen Sequenzen für eine Schnittstelle und eine Nicht-Schnittstelle abgeleitet werden. Dazu werden die Aminosäuren ausgewählt, die an der betreffenden Position entweder den minimalen (Schnittstelle) oder maximalen (Nicht-Schnittstelle) Wert besitzen. Die Sequenzen, die sich dabei ergeben sind:
LLLLLLGAWA↓LG (Schnittstellensequenz)
EQRSESLEGF RR (Nicht Schnittstellensequenz).

TABELLE 12. Prozentuale Wahrscheinlichkeit eine Aminosäure an einer Position zu finden, berechnet aus den Gewichts und Kodierungsvektoren.

	-10	**-9**	**-8**	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**1**	**2**
A	6,1	7,9	8,3	3,1	2,6	0,7	8,1	10,0	4,7	10,6	9,2	9,7
C	8,3	8,8	9,4	7,6	10,8	8,0	2,3	7,5	5,7	7,0	8,8	5,6
D	0,6	2,0	0,7	5,1	0,4	3,4	7,2	0,5	2,7	2,3	0,9	5,3
E	0,0	2,3	0,6	5,7	0,0	3,2	7,7	0,0	2,0	2,4	0,8	6,0
F	4,8	3,0	2,8	6,9	8,1	9,3	1,7	1,2	5,6	0,4	2,0	0,7
G	0,9	7,6	5,9	7,4	0,0	0,5	10,1	4,3	0,0	8,8	6,3	12,6
H	5,2	4,8	4,9	5,4	5,9	5,9	4,3	4,6	5,2	4,3	4,7	4,2
I	6,7	6,9	7,2	6,5	8,2	6,8	3,4	6,1	5,4	5,8	6,8	5,1
K	3,8	5,1	4,5	7,1	5,6	6,1	4,6	2,9	3,7	3,9	4,1	5,2
L	9,8	9,9	10,8	9,4	14,5	10,8	0,0	7,5	6,2	6,5	9,6	4,4
M	4,6	6,6	6,1	7,4	6,3	5,9	4,8	4,4	3,6	5,5	5,8	6,5
N	4,3	4,3	4,1	5,1	4,6	5,2	5,0	4,0	4,8	4,1	4,0	4,6
P	4,7	2,4	2,8	2,0	2,5	3,6	5,7	5,0	6,3	3,9	3,3	3,2
Q	3,9	1,8	2,0	1,4	1,0	2,6	6,6	4,6	6,0	3,7	2,6	3,4
R	3,6	0,0	0,0	2,7	3,1	5,8	4,1	1,5	6,4	0,0	0,0	0,0
S	6,1	4,9	5,9	0,0	0,9	0,0	8,1	9,8	6,6	8,7	7,0	6,8
T	7,1	4,8	5,9	1,3	3,8	2,8	5,8	8,8	7,2	7,0	6,6	4,7
V	7,8	6,8	7,6	4,7	7,8	5,9	3,6	8,0	6,6	6,7	7,6	4,8
W	8,7	6,4	7,5	5,5	10,3	8,4	1,5	7,1	7,4	5,0	7,0	2,5
Y	2,9	3,6	2,9	5,7	3,6	5,1	5,5	2,3	3,8	3,2	2,8	4,7

Des weiteren zeigen die Werte der Tabelle, daß die Positionen -8, -5, -4 und 2 den geringsten Einfluß auf die Schnittstelleneigenschaft besitzen.

Die Eigenschaftsvektoren der Tabelle 12 können wieder mit den 140 physikochemischen Eigenschaften nach bekannter Methode verglichen werden. Dabei ergeben sich die Eigenschaften, die in Tabelle 13 aufgelistet sind.

TABELLE 13. Physikochemische Eigenschaften, die am besten mit der gefundenen Kodierung an den jeweiligen Positionen übereinstimmen.

Position	Eigenschaft
-10	Hydrophobizität (16)
-9	Aliphatisch oder (klein und hydrophobisch) (95)
-8	Hydropathizität (10)
-7	Aromatisch oder sehr hydrophobisch (107)
-6	Hydrophobizität (31)
-5	Versteckte Fläche(3)
-4	Sehr klein oder (negativ und hydrophil) oder Threonin oder Prolin (78)
-3	(Klein und hydrophobisch) oder sehr klein (88)
-2	Fraga Recognition Factors (38)
-1	(Klein und hydrophobisch) oder sehr klein (88)
1	Hydrophobizität (47)
2	Sehr klein (56)

Das Netz wurde zur Vorhersage von Schnittstellen in Sequenzen angewendet. Diese Ergebnisse sind in Tabelle 14 wiedergegeben. 78% der Schnittstellen von 19 Proteinen, die eine

TABELLE 14. Anwendung des trainierten KNN auf die 76 experimentell verifizierten Proteine.

	Test	%	Train	%
Identifizierte Schnittstellen	15	78	49	86
Nicht identifizierte Schnittstellen	2	11	6	11
Nicht identifizierte Präpeptide	2	11	2	4
Summe	19	100	57	101

Schnittelle besitzen, konnten korrekt identifiziert werden. In 11% der Sequenzen wurden die Schnittstellen noch nicht einmal als putative Schnittstellen vorhergesagt und in weiteren 11% waren die wirklichen Schnittstellen zwar unter den putativen, jedoch wurde eine falsche ausgewählt.

4.a.viii Kohonen Netze

Kohonen Netze sind selbst organisierende Karten, die unüberwacht lernen. Im Gegensatz zu den bisher behandelten Netzwerken ist hier kein Wissen über die Klassifikation der Sequenzen notwendig zum Lernen. Dieses wird nur nach der Trainingsphase zum Bewerten der Netze

verwendet. Da es nicht sicher ist, wieviele Signalpeptidasen es gibt, soll mit Hilfe der Kohonen Netze eine mögliche Aufteilung der Schnittstellensequenzen in zwei oder mehr Klassen untersucht werden. Da nur sehr wenige Daten zur Verfügung stehen, wurde ein ein-dimensionales Kohonen Netz mit 10 Neuronen verwendet. Hiermit sollte es möglich sein, zwei Gruppen zu unterscheiden.

Ein Histogramm der Anzahl der Sequenzen pro Neuron ist in Abbildung 29 dargestellt. Hier zeigt sich, daß auf das erste und das letzte Neuron die meisten Sequenzen projeziert werden. Insgesamt über 44% der Sequenzen können auf diese beiden Neuronen abgebildet werden. Es ist jedoch keine klare Trennung der Sequenzmuster möglich.

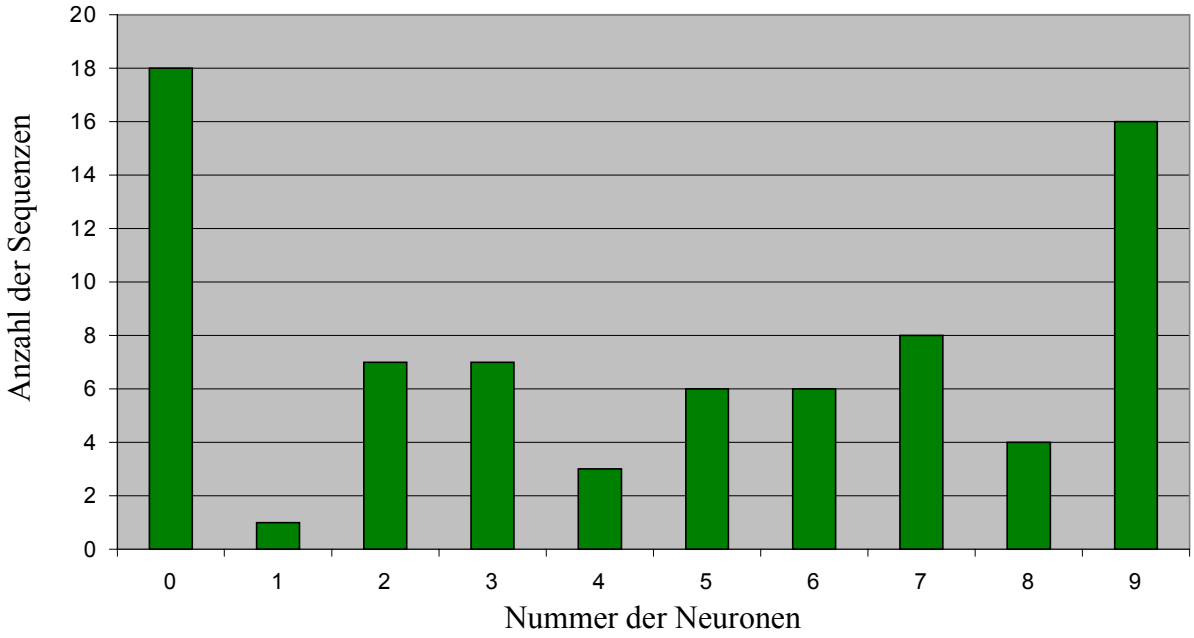

ABBILDUNG 29: Anzahl der Sequenzen an den einzelnen Neuronen des Kohonennetzes der Schnittstellen nach Projektion der Sequenzdaten.

Die Gewichtsvektoren der Neuronen Nr. 0 und Nr. 9 (Abbildung 29) geben Aufschluß über die Sequenzmerkmale, die für eine Unterscheidung der beiden Neuronen verantwortlich sind. Die Gewichte an den Positionen -9, -3, -2 und -1 sind in Abbildung 30 als Histogramm dargestellt, wobei sie entsprechend dem Einbuchstaben-Kode alphabetisch sortiert sind. Hier zeigt sich, daß vor allem die Position -1 entscheidend für die Unterscheidung ist. Hier wird zwischen Alanin und Glycin unterschieden. So werden alle Sequenzen mit einem Glycin im Neuron 0 gefunden. Jedoch ist das Kriterium für Neuron 9 komplexer. So gibt es Sequenzen, die an Position -1 ein Alanin und an Position -9 ein Leucin haben, die nicht in Neuron 9 abgebildet werden.

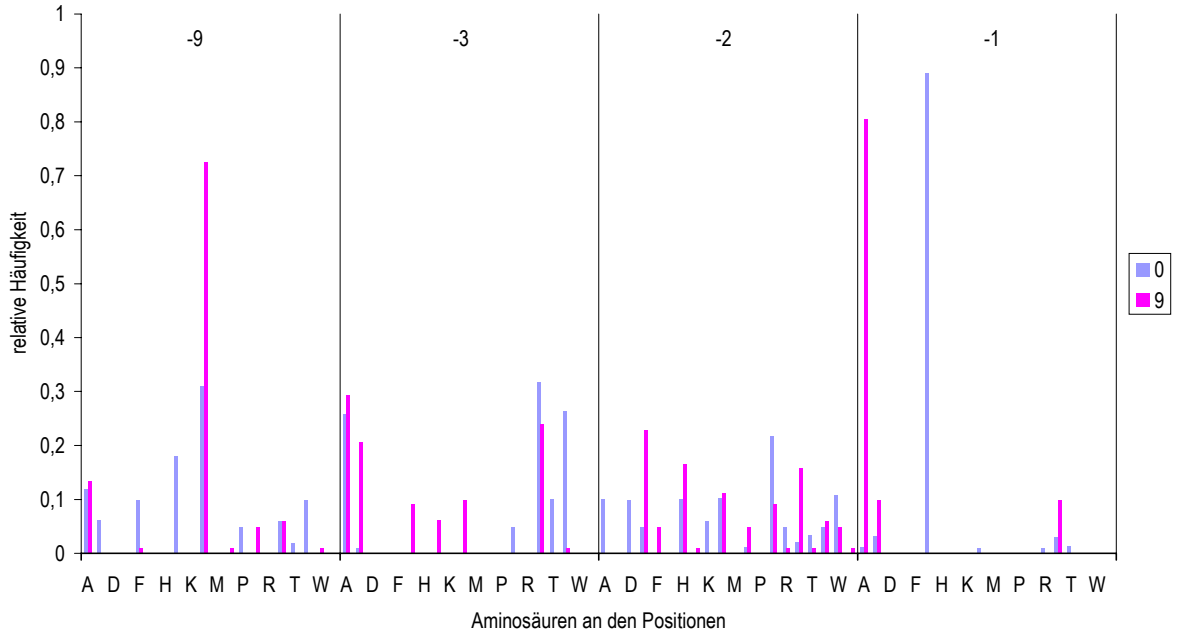

ABBILDUNG 30: Gewichtsvektoren des 0. und 9. Neurons an ausgewählten Positionen (-9, -3, -2 und -1).

Eine Kohonenanalyse der Nicht-Schnittstellensequenzen ergibt das in Abbildung 31 dargestellte Histogramm für die Verteilung der Sequenzen auf die Neuronen. Hier sind die Aminosäuren der Position -10 von besonderer Bedeutung. Die Neuronen 0 und 9 unterscheiden sich durch Leucin (0. Neuron) und Alanin (9. Neuron) an der Position -10. Jedoch werden auf diese Neuronen nur etwa 1/3 der Sequenzen abgebildet.

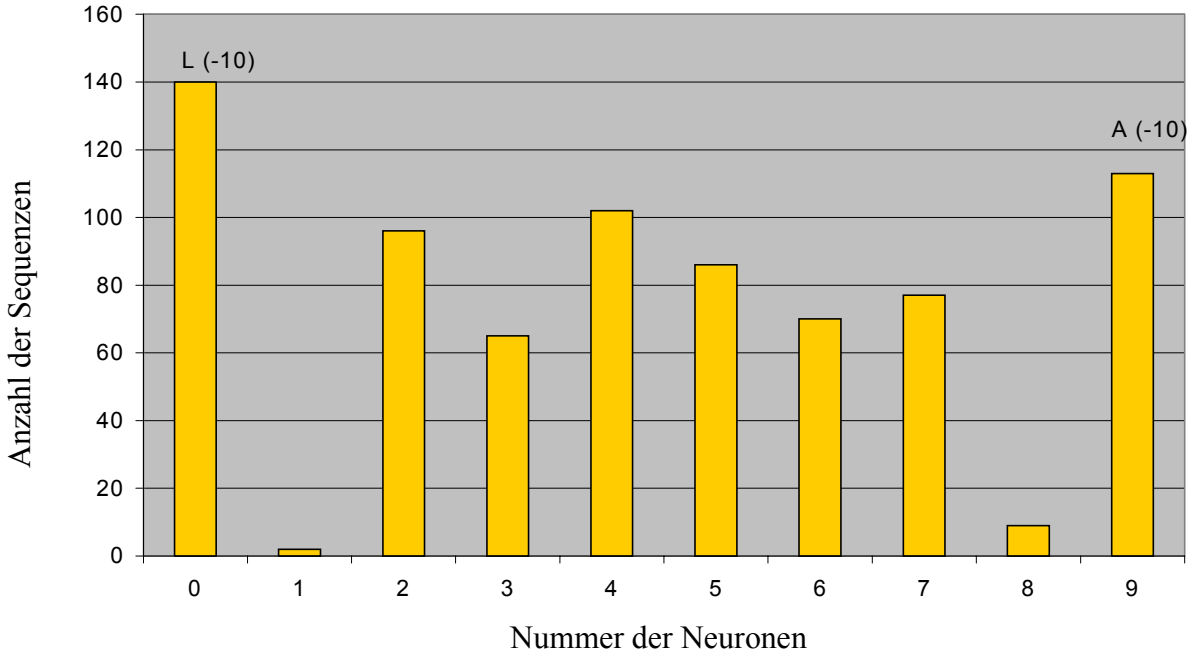

ABBILDUNG 31: Anzahl der Sequenzen an den einzelnen Neuronen des Kohonennetzes der Nicht-Schnittstellen nach Projektion der Sequenzdaten.

Eine Klassifikation mit den Kohonennetzen von Schnittstellen und Nicht-Schnittstellensequenzen ist nicht möglich (Ergebnisse nicht gezeigt).

4.b Vorhersage der *cis*/*trans*- Konformation von Peptidylprolylbindungen

Der hier verwendete SELECT Datensatz [38] wurde mit dem Datensatz der PDB [6] und den nach McCaldon et al. [11] zu erwartenden Häufigkeiten von Aminosäuren in Proteinen verglichen, um einen Eindruck über die Verallgemeinerbarkeit der nachfolgenden Ergebnisse zu vermitteln. Statistische Untersuchungen wurden durchgeführt bezüglich des ω-Winkels von Xaa-Pro Peptidbindungen und bezüglich der relativen Position in den verschiedenen Proteinen. Die lokalen Sequenzbereiche um die Peptidylprolylbindungen wurden mit bioinformatischen Methoden auf eine Klassifizierbarkeit hin untersucht. Die unterschiedlichen Sekundärstrukturumgebungen der Prolylgruppen wurden eingeführt, um eine Klassifikation zu erleichtern. Zum Schluß wurde auch die 3D-Information der benachbarten Umgebung mit in die Untersuchung aufgenommen.

Die in Tabelle 15 dargestellten Häufigkeiten von Aminosäuren zeigen, daß kaum Unterschiede zwischen der Aminosäurezusammensetzung des gesamten PDB-Datensatzes und des SELECT Datensatzes zu finden sind. Die Unterschiede liegen mit maximal 0,3% im Bereich der Fehlertoleranz. Im Vergleich mit der von McCaldon et al. [11] ermittelten Häufigkeit treten die Aminosäuren Glutamin, Arginin, Serin und Leucin in Strukturdatenbanken seltener auf, Asparaginsäure und Glycin treten deutlich häufiger auf.

TABELLE 15. Häufigkeit der 20 Aminosäuren im SELECT- und PDB- Datensatz sowie nach McCaldon et al. (1988).

Aminosäure	Anzahl (SELECT)	rel. Häufigkeit SELECT [%]	rel. Häufigkeit PDB [%]	rel. Häuf. nach McCaldon et al.)[%]	Δ(SELECT - McCaldon et al.[11]) [%]
A	11568	8,3	8,2	8.3	0.0
C	2072	1,5	1,6	1.7	-0.2
D	8320	6,0	5,8	5.3	0.7
E	8471	6,1	5,9	6.2	-0.1
F	5643	4,0	4,0	3.9	0.1
G	11045	7,9	8,1	7.2	0.7
H	3103	2,2	2,3	2.2	0.0
I	7684	5,5	5,4	5.2	0.3
K	8012	5,7	6,1	5.7	0.0
L	11673	8,4	8,4	9.0	-0.6
M	3013	2,2	2,0	2.4	-0.2
N	6624	4,7	4,7	4.4	0.3
P	6707	4,8	4,6	5.1	-0.3
Q	5289	3,8	3,6	4.9	-1.1
R	6575	4,7	4,6	5.7	-1.0
S	8495	6,1	6,4	6.9	-0.8
T	8279	5,9	6,2	5.8	0.1
V	9726	7,0	7,0	6.6	0.4
W	2192	1,6	1,5	1.3	0.3
Y	5243	3,8	3,6	3.2	0.6
Summe	139734				

Die Bestimmung des Winkelbereiches, der eine Peptidbindung als *cis* charakterisiert, wird unterschiedlich durchgeführt. So haben Stewart et al. [117] eine Abweichung des ω-Winkels von 90° toleriert. Jabs et al. haben einen Bereich von +/- 45° verwendet [41]. In der vorliegenden Arbeit wird eine Abweichung von mehr als 20° vom *trans*-Zustand (180°) als *cis*-Bindung bezeichnet (vergleiche hierzu Diskussion). Die Auswirkung der Definition des ω-Winkels auf die Anzahl der Sequenzbeispiele ist in Tabelle 16 zusammengefaßt.

TABELLE 16. Häufigkeiten von Peptidbindungen bei vorgegebenen ω-Winkel und N-ständiger Aminosäure (Spalte AS). Die Daten in "ges." gekennzeichneten Spalten beziehen sich auf die Gesamte PDB, alle anderen auf den SELECT Datensatz. rel. = relative Häufigkeit. Σ = Summe.

AS	0° ±160° (ges.)	rel. (ges.)	0° ±160°	rel.	0°±20° (ges.)	rel. (ges.)	0°±20°	rel.
XA	216	0,03	10	0,02	22	0,01	4	0,01
XC	71	0,01	2	0,00	1	0,00	0	0,00
XD	400	0,05	17	0,03	99	0,02	3	0,01
XE	200	0,02	10	0,02	8	0,00	1	0,00
XF	182	0,02	7	0,01	10	0,00	4	0,00
XG	360	0,04	12	0,02	89	0,02	2	0,01
XH	92	0,01	7	0,01	6	0,00	2	0,01
XI	138	0,02	6	0,01	20	0,01	0	0,00
XK	176	0,02	12	0,02	8	0,00	0	0,00
XL	215	0,03	10	0,02	6	0,00	0	0,00
XM	54	0,01	7	0,01	0	0	0	0,00
XN	248	0,03	11	0,02	24	0,01	4	0,01
XP	4870	0,56	327	0,65	3771	0,88	273	0,89
XQ	159	0,02	5	0,01	2	0,00	0	0,00
XR	148	0,02	6	0,01	3	0,00	0	0,00
XS	427	0,05	19	0,04	120	0,03	6	0,02
XT	274	0,03	17	0,03	30	0,01	4	0,01
XV	276	0,03	10	0,02	2	0,00	0	0,00
XW	84	0,01	2	0,00	2	0,00	0	0,00
XY	169	0,02	7	0,01	44	0,01	4	0,01
Σ	8759	1,03	504	0,96	4267	1,00	307	0,98

Hieraus geht hervor, daß die Peptidylprolylbindung im Vergleich zu den anderen Peptidylpeptidylbindungen am weitaus häufigsten eine *cis*-Konformation aufweist. Dies gilt sowohl für den gesamten PDB- als auch für den SELECT-Datensatz. Der PDB-Datensatz beinhaltet ca. 15 mal mehr *cis*-Xaa-Pro-Bindungungen als der SELECT-Datensatz. Der relative Anteil ist im SELECT-Datensatz jedoch höher. Die relativen Häufigkeiten der anderen Aminosäuren weisen auch bei dem sehr großen Definitionsraum zum Teil noch sehr große Unterschiede zum gesamten Datensatz auf (auch wenn für die Berechnung Prolin nicht berücksichtigt wird). Für ω = ±20° sind 0.03% der Peptidylbindungen in einer *cis*-Konformation (Xaa-non-Pro), 0,22% sind *cis*-Xaa-Pro-Bindungen und 4,6% aller Xaa-Pro-Bindungen liegen in *cis*-Konformation vor. Für ω = ±160° ergeben sich 0,14% für Xaa-(non-Pro), 0,26% für Xaa-Pro und 5,5% der Xaa-Pro-Bindungen sind in *cis*-Konformation.

Die ω-Winkel verteilen sich um die Werte 0° und 180° (Abbildung 32). Die Verteilungen sind an beiden Punkten fast gaussförmig, jeweils mit einer Tendenz zu positiven Winkeln. Dafür finden sich bei negativen Winkelwerten mehr Sequenzen mit einer wesentlich größeren Abweichung.

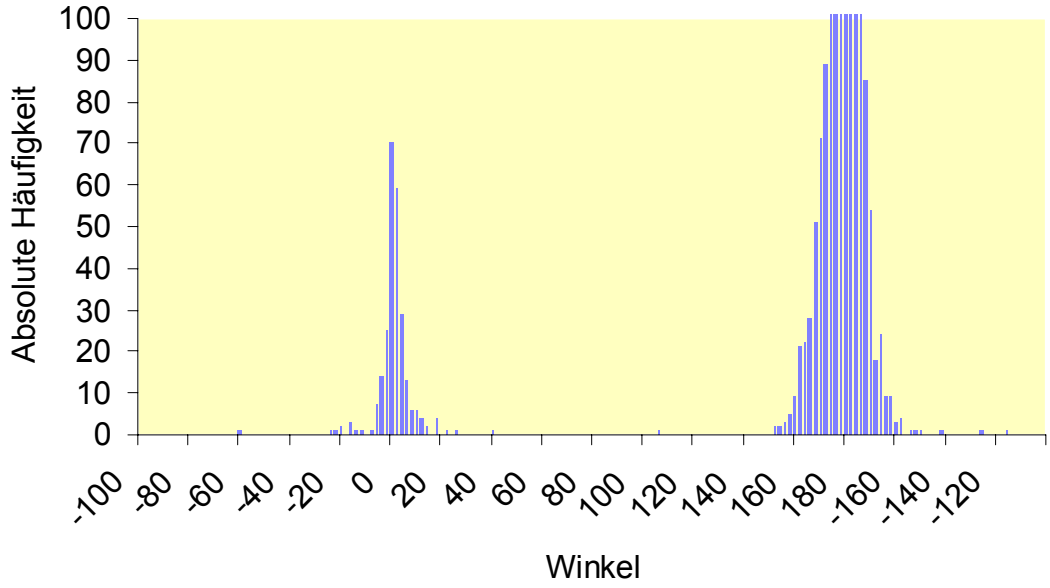

ABBILDUNG 32: Verteilung der ω-Winkel für Xaa-Pro-Bindungen des PDB-SELECT Datensatzes.

Die nicht-homologen Sequenzen des Hobohm Datensatzes [38] beinhalten in einem 4er Fenster (-2, -1, Pro, 1, 2) keine Sequenzen, die sowohl im *cis* als auch im *trans* Datensatz vorkommen. Im gleichen Fenster gibt es nur eine *trans* Sequenz, die zweimal vorkommt. Der Datensatz besteht aus 323 *cis*- und 5682 *trans*-Sequenzen der Länge 21 (±10 Aminosäuren relativ zum zentralen Prolin). Der Einfachheit halber wurden, wenn nicht anders angezeigt, nur die 20 benachbarten Aminosäuren betrachtet, da das zentrale Prolin redundant ist.

4.b.i Informationsanalyse

Der SELECT-Datenstatz als ein nicht-homologer Datensatz weicht in seiner Aminosäurezusammensetzung nicht vom PDB-Datensatz ab, und es gibt nur wenige Abweichungen zur erwarteten Verteilung nach McCaldon et al. [11]. Dieser Datensatz soll die Grundlage zum Studium der Konformationsvorhersage dienen. Die Abweichung von der erwarteten Verteilung der Aminosäuren kann ursächlich in der Beschränkung auf kristallisierbare Proteine lie-

gen. Die Prinzipien, die der Faltung dieser Gruppe von Proteinen unterliegen, werden sich nicht von den anderer Proteine unterscheiden. Daher müssen die Merkmale für die Konformationsvorhersage von Peptidylprolylbindungen auch mit diesem Datensatz beschreibbar sein.

Die Informationsanalyse wurde bereits bei der Analyse der Schnittstellen als ausgezeichnete Methode zur Einschätzung linearer Zusammenhänge angewendet. In Abbildung 33 sind die Werte graphisch aufgetragen. Die Fehlerbalken zeigen, daß sämtliche Besonderheiten im *cis* Datensatz noch innerhalb der Fehlertoleranz liegen. Nach dieser Analyse sind keine nicht korrelierten Eigenarten der beiden Isomere festzustellen.

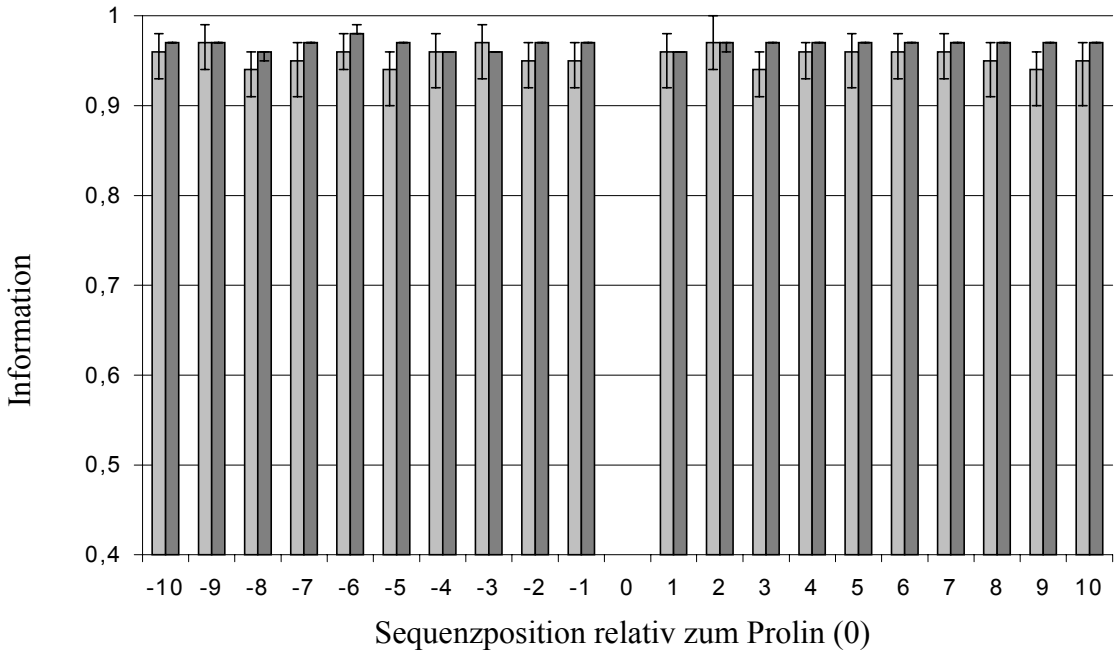

ABBILDUNG 33: Informationsplot der *cis* (hellgrau) und der *trans* (dunkelgrau) Daten. Die Fehlerbalken sind ebenfalls angegeben. Es ist zu sehen, daß es keine wesentlichen Unterschiede zwischen den beiden Isomerensequenzen gibt.

Zur Untersuchung der Unterschiede in der Informationsanalyse müssen die einzelnen relativen Häufigkeiten der Aminosäuren an den jeweiligen Positionen betrachtet werden.

TABELLE 17. Positionen, an denen die Abweichung der relativen Häufigkeit zwischen den *cis*- und *trans*-Sequenzen für die einzelnen Aminosäuren größer als 30% bezogen auf die *cis*-Sequenzen ist. Es sind die Positionen der Reihe nach (1. - 5.) aufgelistet mit den Differenzen (bezogen auf die *cis*-Sequenzen) in Klammern. Negative Prozentangaben bedeuten, daß die *cis*-Sequenzen seltener vorkommen.

Aminosäure	**1.**	**2.**	**3.**	**4.**	**5.**
A	-24 (-37%)				
C	30 (63%)	23 (-55%)	-7 (36%)	7 (32%)	
D	-1 (-68%)				
E	1 (80%)				
F	10 (-48%)	-16 (-46%)	-5 (-33%)		
G	-30 (-114%)	-27 (-91%)	26 (-53%)	22 (-45%)	
H	30 (-101%)	-25 (42%)			
I	-1 (-143%)	1 (-35%)	-26 (-32%)	-10 (-31%)	
K	8 (34%)				
L	9 (-69%)	-12 (-45%)	-2 (-39%)	11 (-36%)	2 (-33%)
M	10 (-207%)	-10 (-166%)	-21 (-150%)	-15 (139%)	-9 (-62%)
N					
P	14 (-31%)				
Q					
R	20 (-73%)	-25 (-38%)			
S					
T	-29 (-72%)				
V	30 (-41%)				
W	22 (-325%)	7 (43%)	14 (-31%)		
Y	21 (-153%)	9 (-46%)			

Abbildungen 34 und 35 zeigen die relativen Häufigkeiten der Aminosäuren Glycin, Valin und Alanin, sowie die Verteilung der Aminosäuren an den Positionen -1, 1 und -30. Der Fehler wurde als Quadratwurzel der Anzahl bestimmt. Aus den Abbildungen geht hervor, daß es zum Teil erhebliche Unterschiede zwischen den beiden Datenmengen gibt. Für eine bessere Auswertung wurden die auffälligsten Daten in Tabelle 17 zusammengefaßt. Hier wurden, unter Berücksichtigung der Fehler, Unterschiede aufgelistet, die mehr als 30% (bezogen auf die *cis*-Sequenzen) betragen. Dabei zeigt sich, daß die betragsmäßig größten Unterschiede nicht wie anzunehmen, in der unmittelbaren Umgebung vom Prolin vorkommen, sondern an weit entfernten Positionen. An Position -30 kommt Glycin wesentlich seltener vor, wenn eine *cis* Peptidylprolylbindung an Position 0 vorliegt, als bei einer entsprechenden *trans*-Bindung. Die meisten Aminosäuren weisen aber an Position -1 einen erheblichen Unterschied auf.

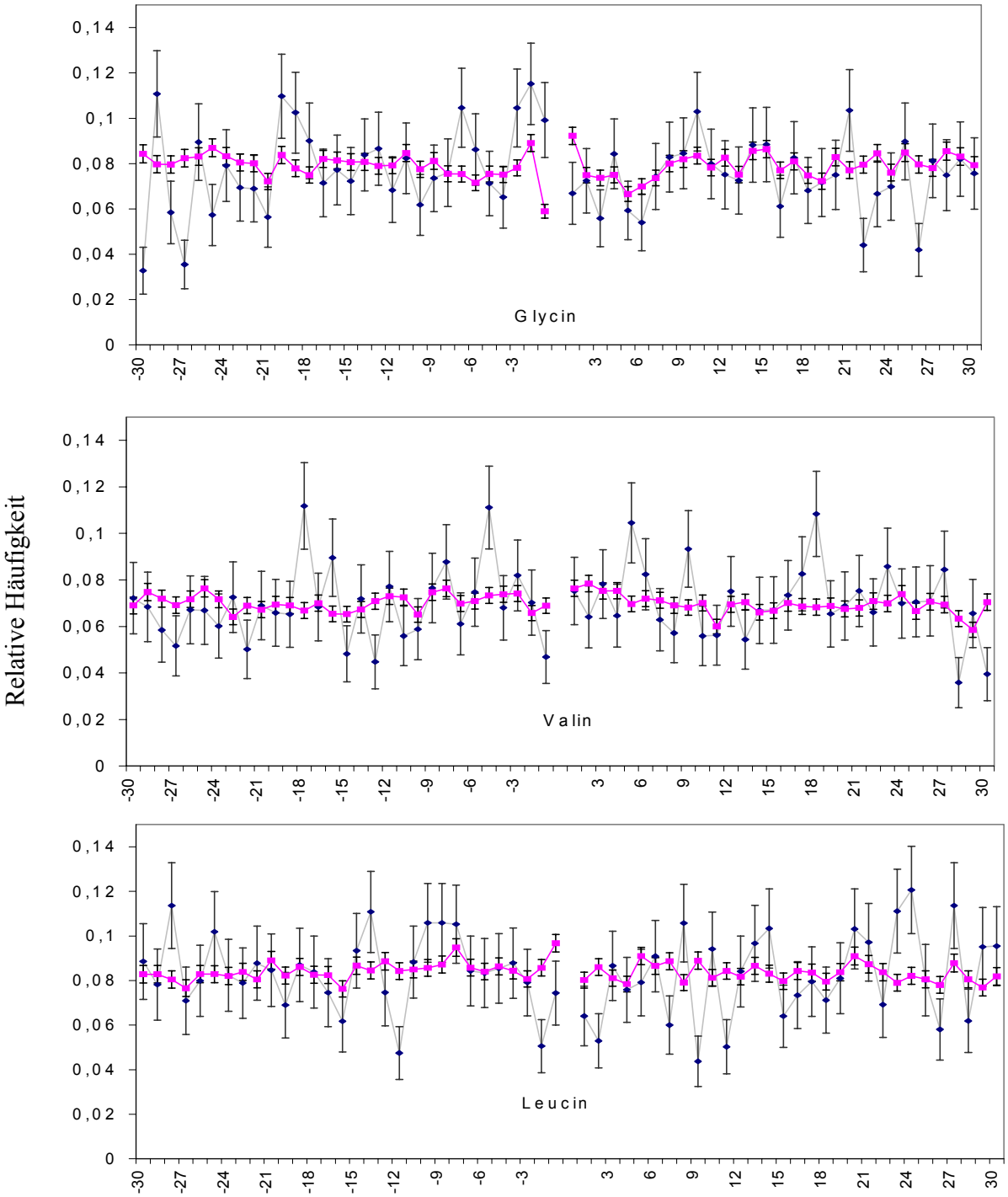

ABBILDUNG 34: Relative Häufigkeit der Aminosäuren Glycin, Valin und Leucin über die Positionen relativ zum zentralen Prolin für die *cis*- (hellgrau) und *trans*- (dunkel) Sequenzen. Die Fehler berechnen sich jeweils aus der Quadratwurzel der Häufigkeit.

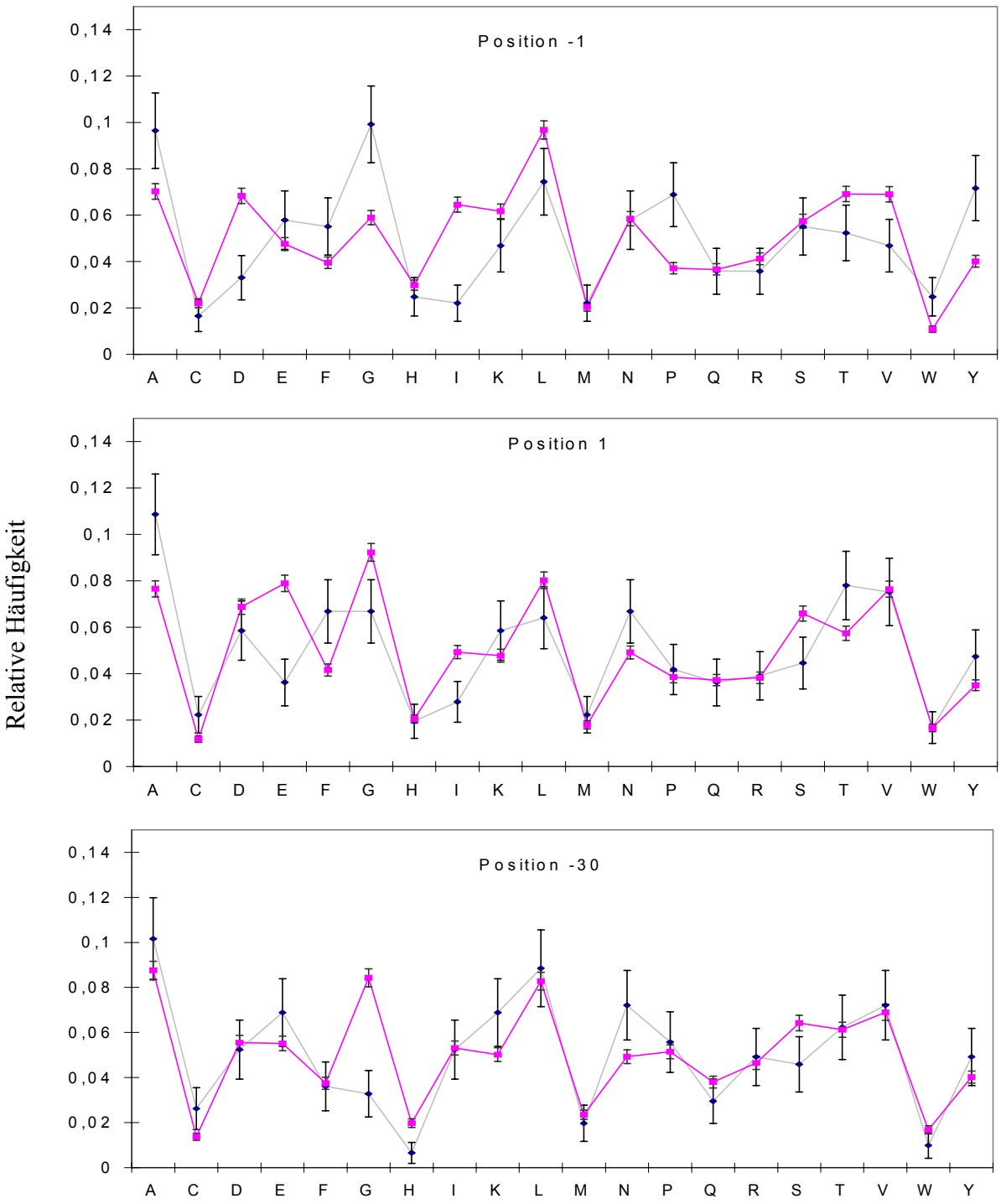

ABBILDUNG 35: Relative Häufigkeit der Aminosäuren an Position -1, 1 und -30 bezogen auf das zentrale Prolin für die *cis*- (hellgrau) und *trans*- (dunkel) Sequenzen. Die Fehler berechnen sich jeweils aus der Quadratwurzel der Häufigkeit.

In Abbildung 36 wird die Position (-1) genauer untersucht. Hier wird gezeigt, daß Prolin, Tyrosin, Phenylalanin und Glycin etwas häufiger in *cis* Sequenzen vorkommen. Asparagin, Isoleucin und Threonin werden hingegen häufiger in *trans* Sequenzen an Position -1 gefunden.

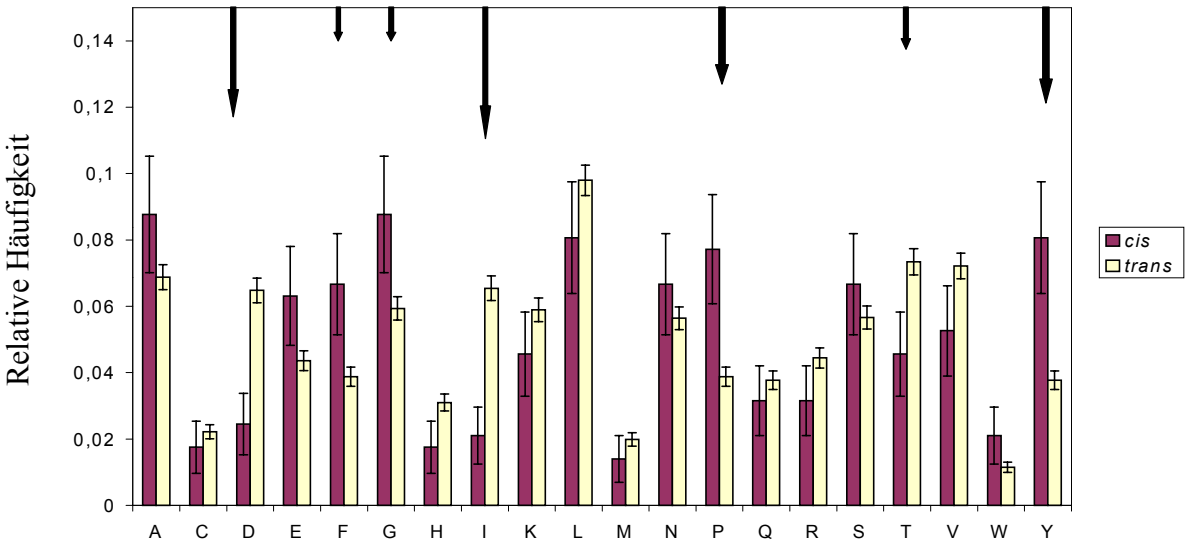

ABBILDUNG 36: Relative Häufigkeit der Aminosäuren an Position -1 relativ zum Prolin für die *cis* (dunkel) und *trans* (hell) Sequenzen. Pfeile zeigen Unterschiede zwischen den beiden Konformeren, deren Länge proportional zum Unterschied ist.

4.b.ii Schwerpunktanalyse

Die Häufigkeitsverteilung der Aminosäuren auf der Schwerpunktgeraden (Abbildung 37) zeigt, daß keine Unterscheidung der Klassen möglich ist. Es wurde über den gesamten Datensatz ein Korrelationskoeffizient von 0,08 errechnet. Dies entspricht den Erwartungen, die aus der Informationsanalyse stammen. Eine leichte Verschiebung der beiden Maxima läßt eine Tendenz erkennen, die jedoch nicht zur Klassifikation ausreicht. Ebenso wie die Häufigkeiten keine klaren Zusammenhänge zwischen der Peptidkonformation und der Sequenz zulassen, ist dies auch nicht mit der Schwerpunktanalyse möglich.

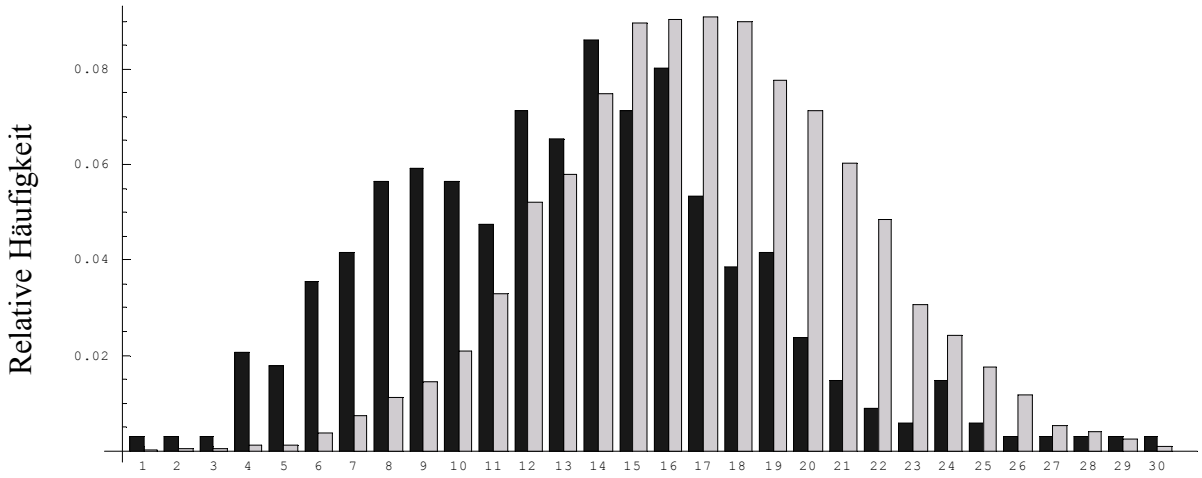

ABBILDUNG 37: Schwerpunktanalyse. Relative Häufigkeit der *cis*-(dunkel) und *trans*-(hell) Daten nach Projektion auf die Schwerpunktgerade.

4.b.iii Hauptkomponentenanalyse

Die Projektion der Datenpunkte auf die ersten beiden Hauptkomponenten (Abbildung 38) zeigt, daß für die meisten Datenpunkte keine Unterscheidung der Konformation möglich ist. Nur etwa 10% der *cis*-Datenpunkte lassen sich von den *trans*-Daten unterscheiden. Diese Daten sind in Tabelle 18 aufgelistet. Sie wurden in drei Klassen aufgeteilt. Die Klassen werden über das Produkt aus einem Eigenvektor und der Sequenz definiert. Alle *cis* Sequenzen, die einen größeren Betrag als die größte *trans* Sequenz aufweisen, wurden in Tabelle 18 aufgenommen. Die Aminosäuren, die mit den größten Eigenvektorelementen (Abbildung 39) übereinstimmen, sind hervorgehoben. Die Sequenzen der Klasse 1 (negative Werte aus dem Produkt mit der 1. Hauptkomponente) werden hauptsächlich aus Immunglobulinen gebildet. Klasse 2 (positive Werte aus dem Produkt mit 2. Hauptkomponente) besteht hauptsächlich aus Hydrolasen. Die Sequenzen der Klasse 3 fallen nicht in eine zusammenhängende Proteinklasse. Das Prolin in Klasse 1 liegt in einer β-VIa1 Struktur vor; alle anderen Klassen besitzen keine strukturellen Gemeinsamkeiten. Die aus den Eigenvektoren abgeleiteten Konsensussequenzen sind in Abbildung 39 wiedergegeben.

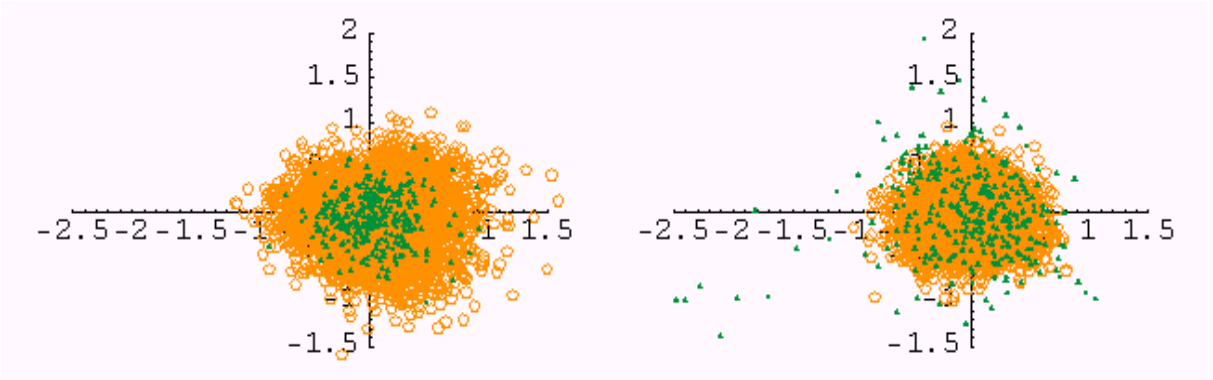

ABBILDUNG 38: Projektion der verteilt kodierten Daten auf die ersten beiden Hauptkomponenten der *trans*-Daten (links, orangene Kreise) und der *cis*-Daten(rechts, grüne Dreiecke)

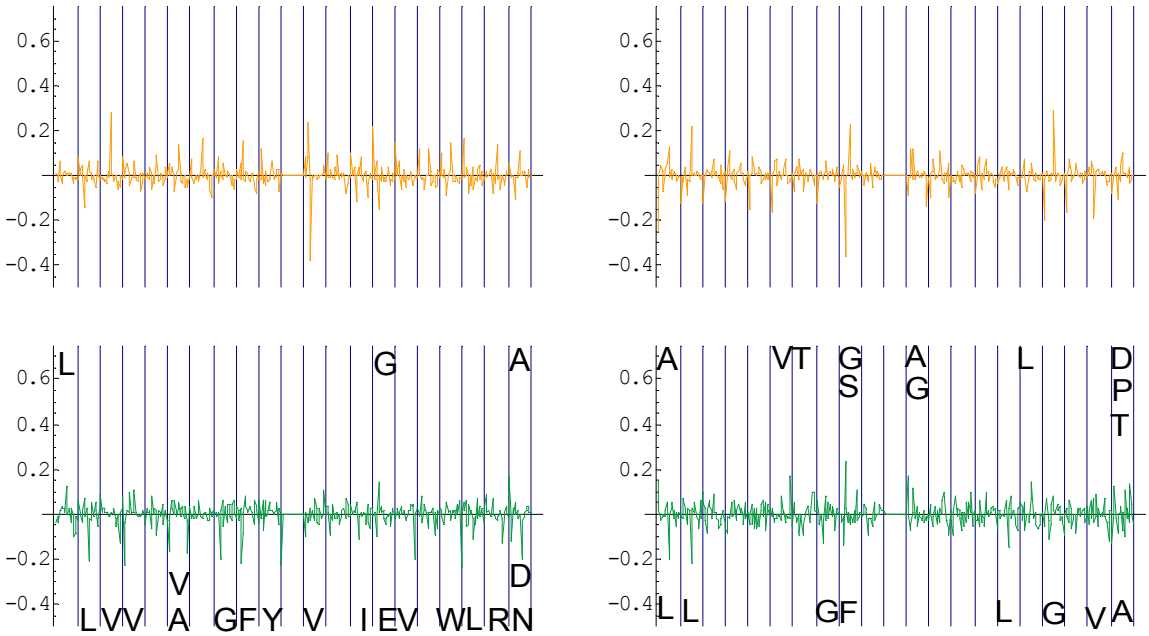

ABBILDUNG 39: Komponenten des 1. und 2. Eigenvektors der *trans*-Daten (oben) und *cis*-Daten (unten). Die Aminosäuren, die zu den größeren Komponenten der *cis*-Daten gehören, wurden an die entsprechenden Positionen geschrieben.

TABELLE 18. *cis*-Sequenzen, die sich nach Projektion auf eine der beiden Hauptkomponenten von allen anderen Sequenzen unterscheiden. Die Sequenzen der Klasse 1 wurden durch Projektion auf die 1. Hauptkomponente der *cis*-Daten erhalten. Die Gruppen 2 und 3 ergeben sich durch Projektion auf die 2. Haupkomponente. Die Sequenzen, PDB-Kodes, der entsprechende ω-Winkel der zentralen Xaa-Pro Einheit, der Wert der Projektion (EV*Sequenz), die Proteinklasse und der CATH [84] -Name (Referenzprotein der Proteinklasse) sind aufgeführt.

Klasse	Sequenz	PDB-Kode	Omega	EV * Sequenz	Proteinklasse	CATH
1	L**LV**CSVS**GFY**PGS**IEV**R**W**F**RN**	1dlhb	-1.441	-2,5	Histokompatibilitäts Protein	1hnf 1
1	S**L**TCLVK**GFY**PSD**I**A**V**E**W**ES**N**	1fc2d	-5.081	-2,4	Immunglobulin	1hnf 1
1	T**L**RCWAL**GFY**PAD**I**TLT**W**QKD	1mhca	0.337	-2,3	MHC Antigen	1hnf 1
1	T**LV**CLAR**GF**FPDHV**E**LS**W**WV**N**	1bec_	-4.176	-2,1	Antigenrezeptor	1hnf 1
1	V**L**ICFIDK**F**TPPVVN**V**T**WLRN**	1dlha	0.287	-2,0	Histokompatibilitäts Protein	1hnf 1
1	SV**V**CFLNN**FY**PKD**I**N**V**K**W**KID	6fabl	-2.906	-1,8	Immunglobulin	1hnf 1
1	V**L**TCA**A**FS**FY**PPELKFRF**LRN**	1frua	0.389	-1,7	Immunglobulin bindendes Protein	1hnf 1
1	A**L**GCLVKDYFPEPVT**V**S**W**NSG	8fabb	-9.437	-1,5	Immunglobulin	1hnf 1
1	QEQHY**A**G**G**NDPANR**E**AT**WL**SG	6taa_	0.060	-1,2	Hydrolase	1nar
	Y**L**SDETQY**F**CPAGL**E**ASQEA**N**	2dln	-2,345	-0,9	Ligase	1iow 2
	L**L**RCKRF**G**RPPTTLAEFSLNQ	1thv	3,848	-0,6	süß schmeckendes Protein (Thaumatin)	1thv
	QPAFS**A**FVTQPADGFTA**W**KYD	1ede	-0,481	-0,5	Dehalogenase	1tat A
	AG**V**FL**A**ANTFPKSR**E**TRAPLV	1yua	-0,333	-0,6	DNA-bindendes Protein	1yua
	EEKRY**A**M**G**DAPDYDRSQ**WL**NE	2gsta	2,249	-0,7	Glutathion Transferase	1aba
2	**A**YITP**V**QI**G**TP**A**QTLN**L**DFD**T**	2er7e	-4.919	1,9	Hydrolase	1fiv A
2	EYAIP**V**SI**G**TP**G**QDFY**L**LFD**T**	1mpp_	-0.094	1,5	Hydrolase	1fiv A
2	**A**FVPF**VT**L**G**DP**G**IEQS**L**KII**D**	1ubsa	0.928	1,4	Lyase Peptid	1pii 2
2	KDVKVLVV**G**NP**A**NTNA**L**IAYK	1bdmb	-3.392	1,3	Oxidoreduktase	1efu A1
	AYPGDI**T**Q**G**SPFDTGI**L**NAL**T**	1crl	8,481	1,1	Hydrolase	1lpp
	ADGGT**V**VIAPP**A**APFRCPPG**P**	1fcd	157,519	1,2	Dehydrogenase	1nhp 1
3	Y**L**RSIKKQLHPSKII**L**ISD**VA**	1vpt_	0.162	-1,3	Methyltransferase	1vpt
	LRLFEYG**GF**PPESNY**L**FLGDY	1fjm	0,332	-1,0	Phosphatase	1fjm A
	T**L**VCLAR**GF**FPDHVE**L**SWW**V**N	1bec	-4,176	-1,4	Antigen Rezeptor	1hnf 1
	LLRCKRF**G**RPPTTLAEFSLNQ	1thv	3,848	-1,1	süß schmeckendes Protein	1thv
	LTPLAYKQ**F**IPNVAEKTLGAS	1vhh	0,025	-1,0	Signal Protein	1vhh

Die Eigenwerte, die aus den Konformationsdaten berechnet wurden, unterscheiden sich nicht von Eigenwerten, die mit einer jeweils gleichen Anzahl von Zufallssequenzen berechnet wurden.

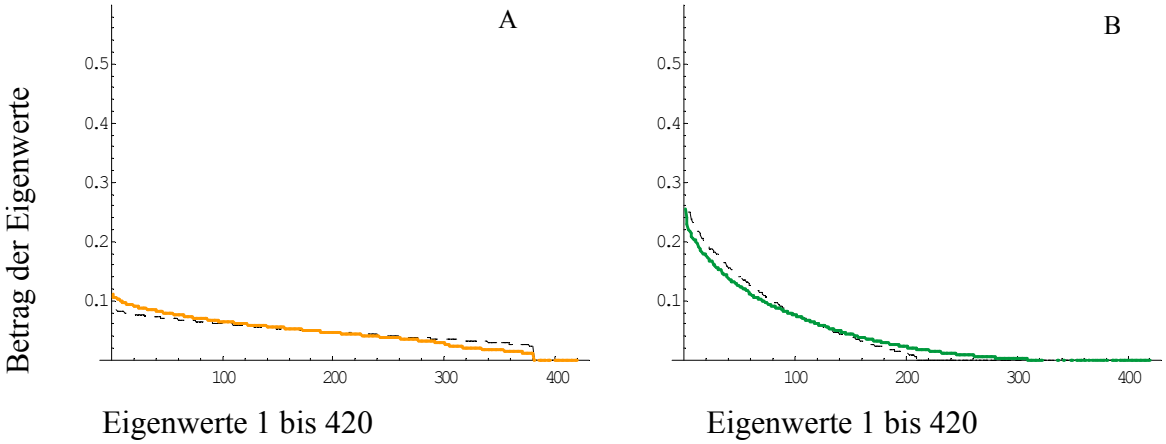

ABBILDUNG 40: Eigenwerte (sortiert nach der Größe) der Hauptkomponentenanalyse der *trans*-Daten (A) und der *cis*-Daten (B) sowie die Eigenwerte für eine entsprechende Anzahl von Zufallssequenzen (gestrichelte Linien). Vergleiche auch Abbildung 20 auf Seite 46.

4.b.iv Mahalanobis-Distanzanalyse

Die Ergebnisse der Hauptkomponentenanalyse lassen sich mit der Mahalanobis-Distanzanalyse auf eine Klassifizierbarkeit hin untersuchen. Dabei ergab sich ein Korrelationskoeffizent von Qcorr = 0,21 für alle Sequenzdaten. Entsprechend der Hauptkomponentenanalyse wurden einige der Daten korrekt klassifiziert. Zwar wurden alle *cis* Daten erkannt, jedoch wurde auch fast die Hälfte der *trans* Sequenzen als *cis* klassifiziert. Aus den Ergebnissen geht hervor, daß keine Trennung mit Hilfe dieser Methode möglich ist.

4.b.v Adaptive Kodierung

Eine Reihe von Neuronalen Netzen mit adaptiver Kodierung wurden trainiert. Es wurde eine Kreuzvalidierung über drei Teile durchgeführt und die Rechnungen jeweils dreimal wiederholt. Für die Perzeptronarchitektur und Netze mit drei, fünf und sieben Hiddenschichten wurde jeweils die Anzahl der Kodierungsvektoren zwischen 1 und sechs variiert. Die Ergebnisse sind in Abbildung 41 graphisch dargestellt. Mit größer werdender Komplexität der Netze werden die Resultate der Trainingsläufe besser. Die Abweichungen hiervon können aufgrund statistischer Schwankungen entstehen. Die Korrelationskoeffizienten für die Testwerte erreichen in keinem Fall einen Wert größer als 0,2. Zwar steigen die Mittelwerte für das Perzeptron und

drei Hidden Neuronen an. Bei fünf Hiddenneuronen und vier Kodierungsvektoren wird jedoch ein Maximum durchlaufen. Alle weiteren Erweiterungen des Parameterraums führen zu einer Verschlechterung der Testergebnisse (Übertrainieren).

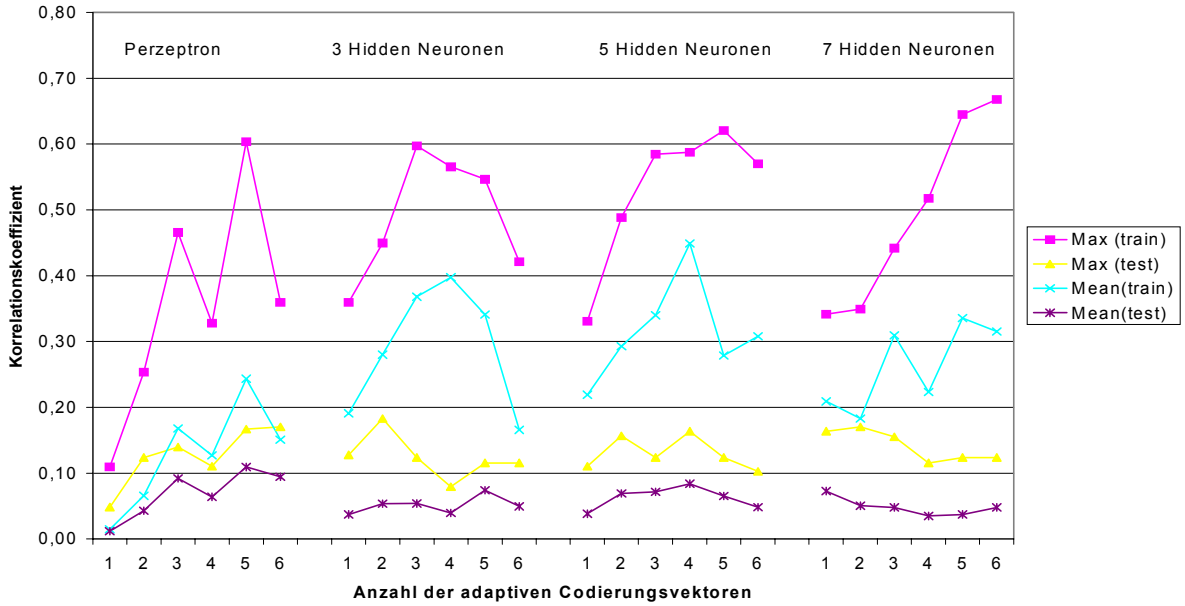

ABBILDUNG 41: Ergebnisse der ACN. Es wurden verschiedene Netzarchitekturen (Perzeptron, drei, fünf und sieben Hiddenschichten) mit ein bis sechs Kodierungsvektoren berechnet. Es sind die über neun Versuche gemittelten sowie Maximalwerte der Korrelationskoeffizienten für Training und Test aufgeführt.

4.b.vi Kohonennetze

Es wurden folgende Untersuchungen mit Kohonennetzen gemacht: Zum einen wurden für verschiedene Kodierungen („*Hydrophobizität & Volumen & Polarität*“, verteilte Kodierung und 8 Eigenschaftsskalen nach Taylor [118]) Netze für den gesamten Datensatz und für die *cis* und *trans* Daten separat berechnet. Außerdem wurden die Sequenzen entsprechend der Sekundärstruktur des zentralen Prolins separat untersucht. Es konnte in keinem Fall eine Trennung durchgeführt werden. Die Ergebnisse sind in Tabelle 19 zusammengefaßt.

TABELLE 19. Ergebnisse der Analyse mit Kohonennetzwerken

Kodierung	**Korrelationskoeefizient**
verteilt	0,06
Hydrophobizität / Volumen / Polarität	-0,02
entsprechend der Sekundärstruktur des Prolins	0,02
8 Eigenschaften nach Taylor [118]	0,03

Die Kohonenexperimente wurden so durchgeführt, daß jede Klasse mit der gleichen Häufigkeit berechnet wurde, was zur Folge hat, daß die *cis*-Daten wesentlich häufiger dem Netz gezeigt wurden als die *trans* Daten. Jedoch konnte so eine Überbewertung der *trans* Daten entsprechend der größeren Anzahl ausgeschlossen werden. Es wurden jeweils 10x10 Netze trainiert.

4.b.vii Klassifikation und Information der Oberflächeneigenschaften

Da die Isomerisierung im allgemeinen ein geschwindigkeitsbestimmender Schritt in der Proteinfaltung ist [25], muß ein lokales Energieminimum erreicht werden, bei dem sich Oberflächenstrukturen ausbilden, bevor eine Isomerisierung stattfindet. Geht man davon aus, daß die meisten lokalen Strukturen des nativen Proteins im lokalen Minimum schon vorhanden sind, kann anhand dieser Sturkturen derjenige Bereich identifiziert werden, der sich auch im lokalen Minimum an der Oberfläche befindet. Nur solche Aminosäuren, die zugänglich sind für Isomerasen, haben auch eine Chance isomerisiert zu werden. Deshalb wurden die Sequenzen identifiziert, die einen gewissen Anteil der Oberfläche des nativen Proteins ausmachen und anschließend die Information dieser Sequenzen bestimmt. Die Informationsplots sind in den folgenden Abbildungen dargestellt:

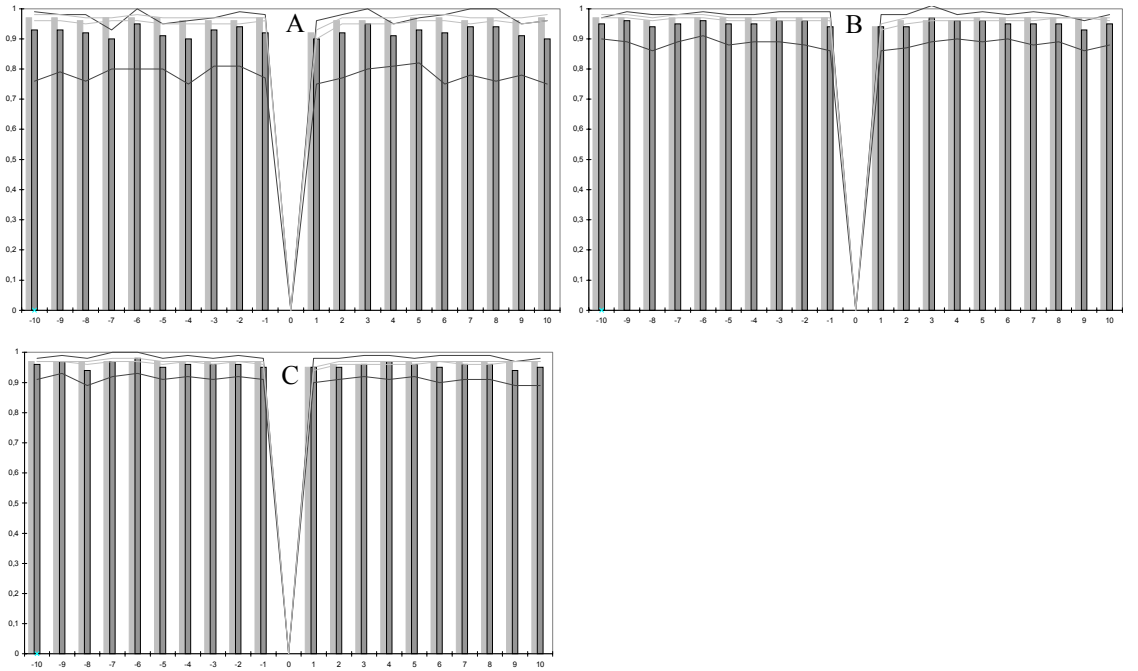

ABBILDUNG 42: Informationsanalyse der Aminosäuren an den Positionen relativ zum Prolin, für Sequenzen bei denen das zentrale Prolin 75Å^2 (A), 40 Å^2 (B) und 10 Å^2 (C) an der Oberfläche des Proteins besitzt (hell *trans*, dunkel *cis*).

Es wurde für 75, 40 und 10 Å^2 die Information der Sequenzen berechnet. Die Unterschiede zwischen den *cis* und *trans* Sequenzen liegen alle im Bereich der Fehlertoleranz. Das heißt, daß auch hier keine Klassifikation möglich ist.

4.b.viii 3D-Umgebung

Nicht nur die Sequenznachbarn können einen bestimmten Einfluß auf die Konformation haben, auch eine Aminosäure deren dreidimensionaler Abstand klein ist, kann einen Einfluß ausüben. Abbildung 43 zeigt die relativen Häufigkeiten der Aminosäuren in der dreidimensionalen Umgebung von 2-3Å um das Prolin für die *cis* und *trans* Sequenzen. Unter Berücksichtigung der Fehler fallen die Aminosäuren Alanin, Methionin und Tryptophan auf, die in *trans* Sequenzen häufiger vorkommen. Hingegen kommt Asparaginsäure etwas häufiger in *cis* Sequenzen vor.

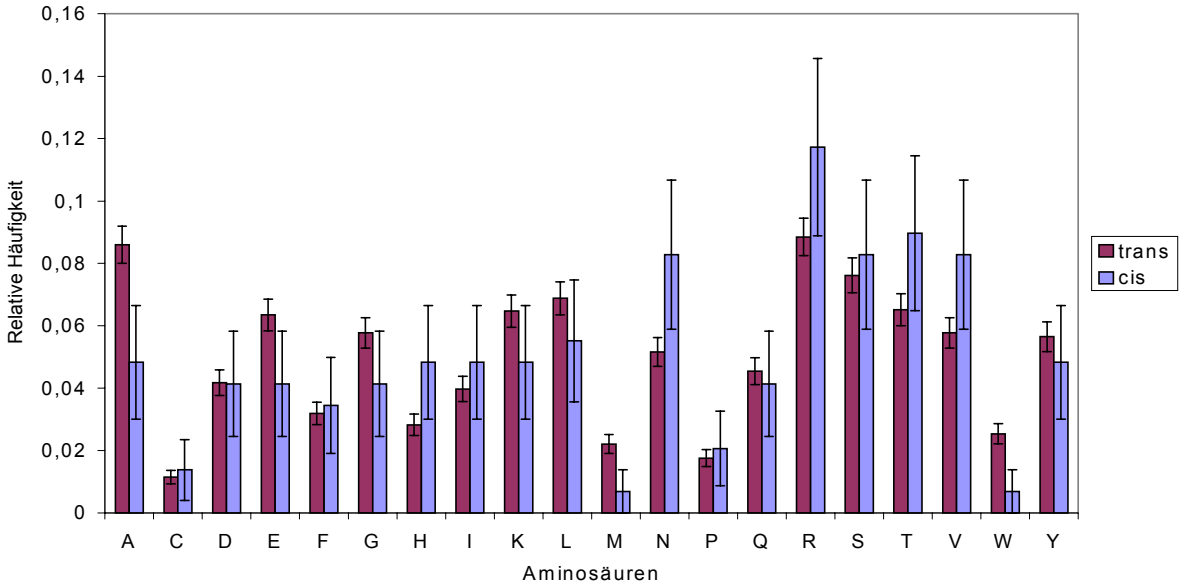

ABBILDUNG 43: Relative Häufigkeit der Aminosäuren im Abstand von 2-3 Å vom Prolin (dunkel *trans*, hell *cis*)

4.b.ix Sekundärstruktur am Prolin

Die Sekundärstruktur der Proteine wurde mit dem Programm DSSP [46] für das Prolin bestimmt. Die relative Häufigkeit der einzelnen Strukturelemente ist in Tabelle 20 aufgeführt. Danach kommen *cis* Sequenzen hauptsächlich in „bends“ (Biegungen, S) und Turn-Strukturen mit einer H-Brücke (T) vor, sowie in random coil Strukturen (_). *Trans* Sequenzen finden sich hauptsächlich in random coil Strukturen (_), wohl aber auch in „bends“ (Biegungen, S) und Turn-Strukturen mit einer H-Brücke (T). Ein nicht zu vernachlässigender Teil kommt in 4-Helixbereichen (α-Helix) vor. Eine Informationsanalyse der Aminosäuren für die Sekundärstrukturelemente T (Abbildung 44), S (Abbildung 45) und _ (Abbildung 46) ist durchgeführt worden.

TABELLE 20. Relative Häufigkeit der Konformere bei vorgegebener Sekundärstruktur (nach DSSP) des Prolins

2° Strukturelement	**2° nach Kabsch et al. [46]**	***cis***	***trans***
erweitertes Faltblat, das sich in einer β-Leiter befindet	E	3,80	0,95
Aminosäure in einer einzelnen β-Brücke	B	0,00	0,00
4-Helix (α-Helix)	H	1,69	14,23
3-Helix (3_{10} Helix)	G	0,84	6,80
5-Helix (π-Helix)	I	0,42	0,03
nicht definiertes 2°-Element (random coil)	–	16,46	43,60
bend	S	45,99	10,63
Turn mit H-Brücke	T	30,80	23,77

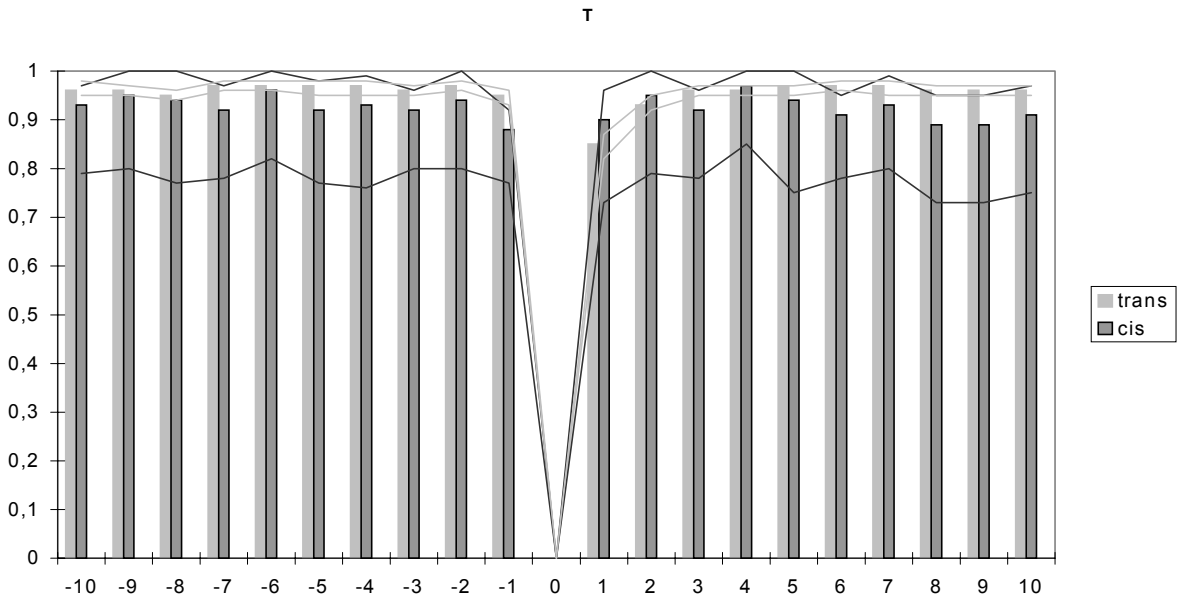

ABBILDUNG 44: Informationsanalyse der Aminosäuren an den Positionen relativ zum Prolin, wenn das 2° Strukturelement nach DSSP des Prolins ein Turn ist (hell *trans*, dunkel *cis*).

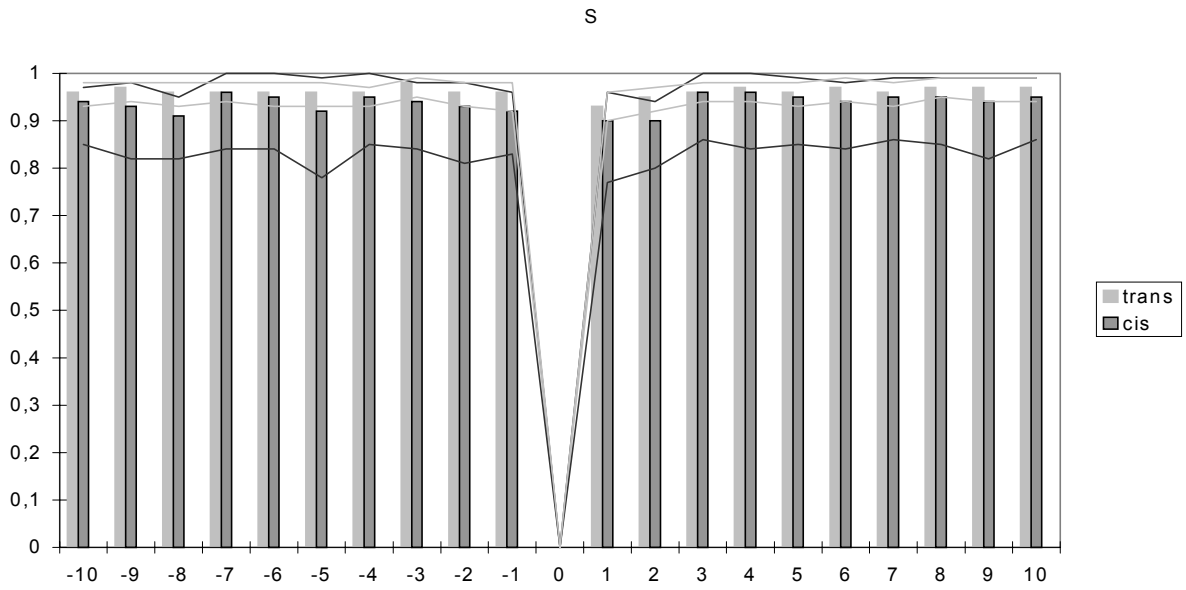

ABBILDUNG 45: Informationsanalyse der Aminosäuren an den Positionen relativ zum Prolin, wenn das 2° Strukturelement nach DSSP des Prolins eine Biegung ist (hell *trans*, dunkel *cis*).

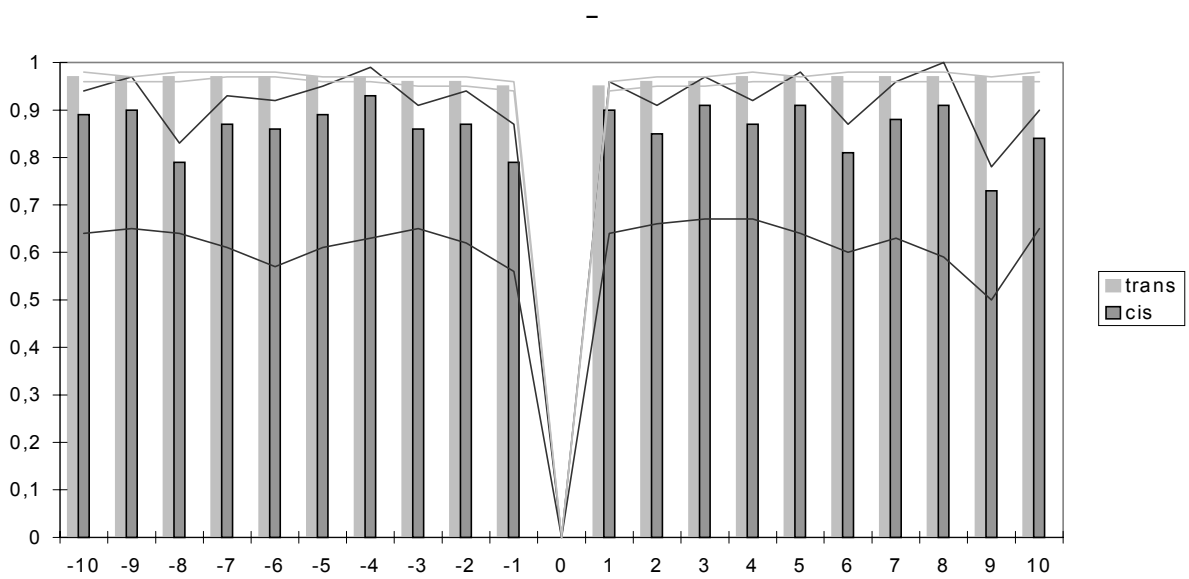

ABBILDUNG 46: Informationsanalyse der Aminosäuren an den Positionen relativ zum Prolin, wenn kein 2° Strukturelement nach DSSP zuzuordnen ist (hell *trans*, dunkel *cis*).

4.c Design von Proteinliganden

Der erste Schritt der Virusinfektion der Wirtszelle durch den Schnupfenvirus HRV 14 ist die Bindung an einen Zelloberflächen-Rezeptor, den ICAM-I Rezeptor. Ziel war es, diese Protein-Protein Wechselwirkung zu inhibieren. Die dreidimensionale Struktur des Virushüllproteins lag mit einer Auflösung von 3Å vor [100]. Für das rationale Design von Proteinliganden, die in der Canyonregion des Rhinovirus HRV 14 binden, wurden zwei Verfahren verwendet. Die Simulierte Molekulare Evolution (SME, [104]) und ein neuer Docking Algorithmus wurden angewendet, um insgesamt 7 Nona-Peptide zu entwickeln, welche anschließend experimentell untersucht wurden.

4.c.i Simulierte Molekulare Evolution

68 Nona-Peptide mit bekannter zellprotektiver Eigenschaft lagen zur Analyse vor (Tabelle 21). Der Grad des Zellschutzes durch ein antivirales Nona-Peptid wurde auf den Wert für die Inhibition der HRV 14-Vermehrung durch einen Inhibitor im Vergleich zur Viruskontrolle in Abwesenheit des Inhibitors bezogen. Dabei geben negative Werte eine Abnahme der Meßsignale im Vergleich zur Viruskontrolle ohne Inhibitor an. Inhibitionswerte ab -30% repräsentieren einen partiellen Zellschutz und werden als signifikant angesehen. Inhibitionswerte zwischen -20% und -29% werden als marginal angesehen. Inhibitionswerte zwischen 0 und -19% werden nicht als Zellschutz gewertet. Positive Werte geben eine Zunahme der Meßsignale im Vergleich zur Viruskontrolle ohne Peptid an.

Die 68 Nona-Peptide wurden in drei etwa gleich große Gruppen für eine Kreuzvalidierung aufgeteilt. Die Daten wurden mit den physikochemischen Eigenschaften Volumen (Zamyatnin [126]), Hydrophobizität (Engelman [20]), Polarität (Jones [45]) und Refraktivität (Jones [45]) kodiert.

TABELLE 21. Nonapeptide, die als Grundlage für die Entwicklung der Neuronalen Netze verwendet wurden. Die zellprotektiven Eigenschaften (Inhibition) sind neben den Sequenzen wiedergegeben. Dabei bedeuten große negative Werte eine hohe Inhibition.

Sequenz	**Inhibition**	Sequenz	**Inhibition**	Sequenz	**Inhibition**	Sequenz	**Inhibition**
GDGVGVGDG	-21%	GPGEGIGEG	-6%	GMGEGIGIG	-24%	GDGDGIGDG	-18%
GDGDGVGDG	13%	GEGGGIGEG	-25%	GYGEGIGIG	-24%	GEGEGIGEG	-18%
GEGEGIGEG	-32%	GEGAGIGEG	-25%	GFGEGIGIG	-24%	GDGIGIGDG	-27%
RGTGTGEGE	-16%	GEGTGIGEG	-25%	GWGEGIGIG	-24%	GEGIGIGEG	-27%
GKEGIGEEG	-15%	GEGSGIGEG	-25%	GEGVGIGEG	-13%	KSTHYESPG	-31%
GEGEGFGFG	-14%	GEGPGIGEG	-25%	GEGIGIGEG	-13%	GSQGPKPWG	-20%
GEGEGEGEG	-8%	GEGEGIGGG	-29%	GEGMGIGEG	-13%	GLKAGVIAV	-51%
KYKHEASAH	-15%	GEGEGIGAG	-29%	GEGYGIGEG	-13%	GILGFFVTL	-9%
GEGEGIGIG	-19%	GEGEGIGTG	-29%	GEGFGIGEG	-13%	GEGEGIGRG	-22%
GIGIGEGEG	-13%	GEGEGIGSG	-29%	GEGWGIGEG	-13%	GEGEGIGHG	-22%
GDGDGVGVG	-18%	GEGEGIGPG	-29%	GEGEGIGVG	-17%	GTGEGIGEG	-6%
GEGIVGIVG	-36%	GKGEGIGEG	-24%	GEGEGIGIG	-17%	GSGEGIGEG	-6%
GIGIGYGEG	-24%	GRGEGIGEG	-38%	GEGEGIGMG	-17%	GAGEGIGEG	-6%
GVGVSYEGE	-21%	GHGEGIGEG	-39%	GEGEGIGYG	-17%	GEGEGIGKG	-22%
GRGRGVGVG	-33%	GEGKGIGEG	-22%	GEGEGIGFG	-17%	GIGEGIGIG	-24%
GRGRGYGYG	-33%	GEGRGIGEG	-22%	GEGEGIGWG	-17%		
GVGEGIGIG	-24%	GEGHGIGEG	-22%	GGGEGIGEG	-6%		

Ein zweilagiges Künstliches Neuronales Netz mit 36 Eingabe- und 10 Zwischen- sowie einer Ausgabeschicht wurde mit einem Lernfaktor von $\sigma = 0{,}2$ und einer (1, 60)-Strategie über 500 Zyklen trainiert. Die Kodierung und Sollausgabewerte wurden entsprechend der im Netz benutzten Sigmoidalfunktion auf einen Wertebereich von -1 bis +1 abgebildet. Der quadratische Fehler wurde minimiert.

Das Netz mit der besten Vorhersagegenauigkeit (beste Testergebnisse) wurde für die Simulierte Molekulare Evolution (SME) verwendet. Eine Ähnlichkeitstabelle, die sich aus den vier physiko-chemischen Eigenschaften berechnete, wurde für die SME verwendet. Die SME wurde mit 20 Nachkommen pro Generation über 200 Generationen durchgeführt. Initialisiert wurde mit der Sequenz GXGXGXGXG (X steht für eine beliebige Aminosäure). Von den circa 100 Sequenzen mit einem optimalen Ausgabewert wurden fünf Sequenzen für die Synthese bestimmt. In Tabelle 22 sind die Sequenzen mit den erzielten zellprotektiven Eigenschaften aufgelistet. Alle neuen Sequenzen führen im virologischen Test zu einer Inhibition, wobei drei der fünf Peptide eine signifikante und zwei eine marginale Zellprotektion bewirkten.

TABELLE 22. 5 Sequenzvorschläge, die mittels der SME erhalten wurden, für die Synthese und zellinhibitorische Wirkung.

Sequenz	Inhibition
GHERGKSHG	-45%
GNGKGKGNG	-39%
PKDRGHIHG	-35%
SRERAHFKE	-22%
GRGRGRGHG	-20%

4.c.ii Peptid-Docking

Der Ansatz dieses Dockingalgorithmus basiert auf der Annahme, daß Protein-Protein-Kontakte in Protein-Komplexen ein typisches Aminosäurebindungsmuster aufweisen. In Zusammenarbeit mit Herrn Dr. Grunert (Institut für Infektionsmedizin, Universitätsklinikum Benjamin Franklin, Freie Universität Berlin) wurden aus der PDB Datenbank verschiedene Datensätze mit Protein-Komplexen erstellt. Durch die Analyse der Häufigkeiten von Aminosäure-Aminosäure-Kontakten für alle 20 Aminosäuren in diesem Datensatz konnten Aminosäurebindungsmuster herausgefiltert werden. Mit Hilfe dieser Aminosäurebindungsmuster wurden 2 Peptide für die tiefste Region im Virus-Canyon konstruiert, synthetisiert und auf ihre zellprotektiven Eigenschaften getestet.

In der Proteindatenbank wurden Homoprotein-Komplexe, Heteroprotein-Komplexe, Enzym-Inhibitor-Komplexe, Antigenbindende Komplexe sowie Virus-Inhibitor Komplexe bestimmt. Eine detaillierte Auflistung der PDB-Bezeichner für die einzelnen Gruppen ist in Kapitel 10 auf Seite 151 aufgeführt. Tabelle 23 führt die Größe der jeweiligen Gruppen auf.

TABELLE 23. Komplex-Klassen und Anzahl der in der PDB gefundenen Proteine

Komplex-Klasse	Anzahl der Komplexe
Homomultimere	849
Heteromultimere	486
Enzym-Inhibitor-Komplexe	151
Antigenbindende-Komplexe	53
Virus-Ligand-Komplexe	2

Die Gruppen der Homo- und Heteromultimere sowie die Enzym-Inhibitor-Komplexe wurden genauer untersucht. Das jeweils kleinste Molekül im Komplex wurde als Ligand bezeichnet. Zu jeder Aminosäure eines Liganden wurden alle räumlich benachbarten Aminosäuren in

einem Abstand von 2 - 3; 2,5 - 3,5; 3 - 4; ... ; 6,5 - 7,5Å bestimmt. Der Abstand zwischen zwei untersuchten Aminosäuren wird über die Atome bestimmt, die am nächsten beieinander liegen. Wasserstoffatome wurden nicht berücksichtigt. Für die 10 Abstandsfenster entstehen so Häufigkeitstabellen von benachbarten Aminosäuren. In Abbildung 47 bis 49 sind die statistischen Auswertungen der Aminosäurewechselwirkungen der verschiedenen Gruppen graphisch dargestellt. Horizontal sind die Aminosäuren des Liganden aufgetragen und die des Rezeptors vertikal. Die Helligkeit gibt die relative Häufigkeit bezogen auf die zu erwartende Häufigkeit für eine Wechselwirkung an.

$$\text{relative Häufigkeit} = \frac{n_{i,j}}{erw_i \cdot erw_j \cdot \sum_{i,j} n_{i,j}} \qquad \text{(GLEICHUNG 44)}$$

Die Werte wurden auf einen Bereich von 0 (schwarz) bis 1 (weiß) abgebildet. Eine Häufigkeit, die einer zufälligen Verteilung entspricht, wird grau dargestellt. Helle Bereiche spiegeln erhöhte Wechselwirkungen wider. Die Linien innerhalb der einzelnen Felder stellen die Verteilung der Aminosäuren auf die 10 Abstände dar. Aufgrund der überlappenden Abstandsbereiche ergeben sich geglättete Kurven. Ein großer Anstieg zwischen zwei Positionen bedeutet, daß der nicht überlappende Bereich die Aminosäuren enthält.

Cystein- Cystein Verbindungen kommen in allen Klassen relativ häufig vor. Im Falle der Homomultimere ist hier das einzige mal in einem Abstand von 2 - 2,5Å die größte Anzahl von Wechselwirkungen zu beobachten. Ansonsten treten die meisten Wechselwirkungen erst in einem Abstand von mehr als 2,5Å auf.

Ebenfalls relativ häufig kommen in dem Abstand 2-3Å Wechselwirkungen zwischen Arginin und Asparagin- und Glutaminsäure vor sowie zwischen Tryptophan und Methionin (Homodimere, Enzym-Inhibitor-Komplexe).

Histidin, Tryptophan und Tyrosin kommen in den Receptorproteinen wesentlich häufiger in der Nähe von Bindungsstellen vor als aufgrund der natürlichen Häufigkeiten zu erwarten ist.

Im Falle der Heteromultimere sind Wechselwirkungen zwischen gleichen Aminosäuren bevorzugt (die Diagonalelemente der Matrizen sind in der Regel hell).

Die Häufigkeitsverteilungen entlang der Abstände zeigen meistens zwei Maxima, bei 3.5-4Å und bei 6.5-7Å.

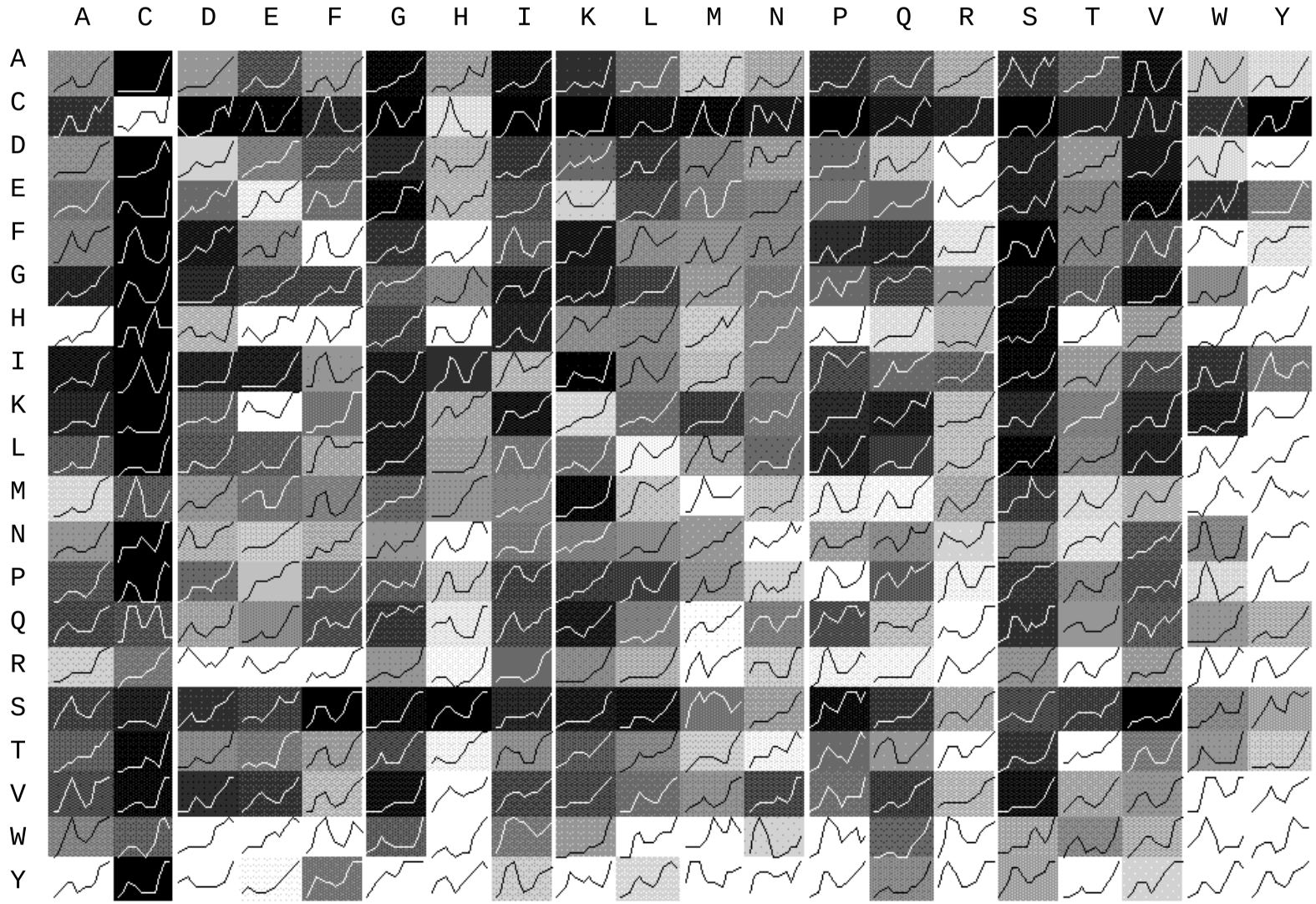

ABBILDUNG 47: Heteromultimer: Aminosäurepartner der Enzym-Inhibitor-Komplexe. Horizontal sind die Aminosäuren des Rezeptors aufgetragen und die des Liganden vertikal. Die Helligkeit gibt die relative Häufigkeit bezogen auf die zu erwartende Häufigkeit für eine Wechselwirkung an. Die Werte wurden auf einen Bereich von 0 (schwarz) bis 1 (weiß) abgebildet. Eine Häufigkeit, die einer zufälligen Verteilung entspricht, wird grau dargestellt. Helle Bereiche spiegeln erhöhte Wechselwirkungen wider. Die Linien innerhalb der einzelnen Felder stellen die Verteilung der Aminosäuren auf die 10 Abstände dar. Aufgrund der überlappenden Abstandsbereiche ergeben sich geglättete Kurven. Ein großer Anstieg zwischen zwei Positionen bedeutet, daß der nicht überlappende Bereich die Aminosäuren enthält.

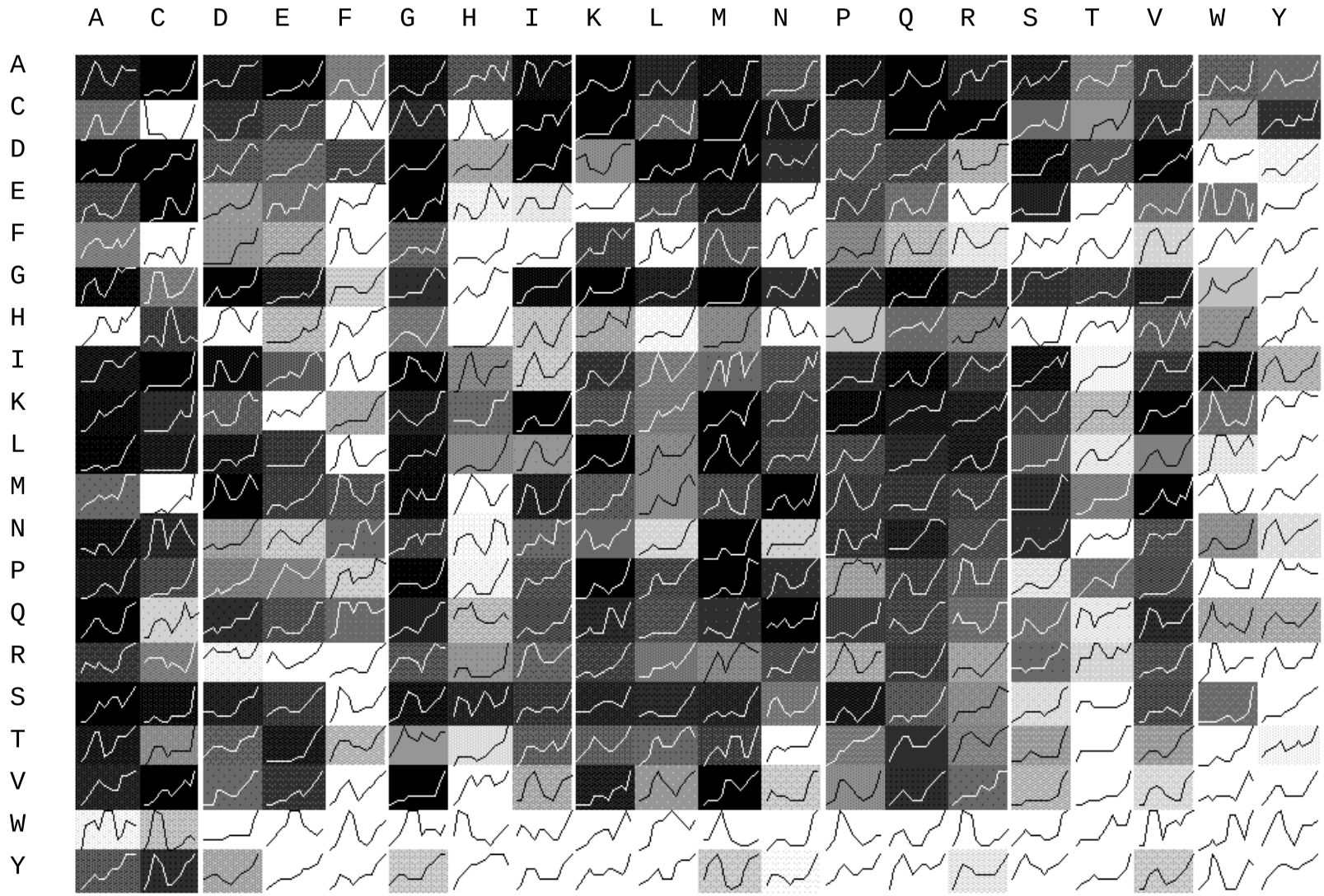

ABBILDUNG 48: Homomultimer: Aminosäurepartner der Enzym-Inhibitor-Komplexe. Horizontal sind die Aminosäuren des Rezeptors aufgetragen und die des Liganden vertikal. Die Helligkeit gibt die relative Häufigkeit bezogen auf die zu erwartende Häufigkeit für eine Wechselwirkung an. Die Werte wurden auf einen Bereich von 0 (schwarz) bis 1 (weiß) abgebildet. Eine Häufigkeit, die einer zufälligen Verteilung entspricht, wird grau dargestellt. Helle Bereiche spiegeln erhöhte Wechselwirkungen wider. Die Linien innerhalb der einzelnen Felder stellen die Verteilung der Aminosäuren auf die 10 Abstände dar. Aufgrund der überlappenden Abstandsbereiche ergeben sich geglättete Kurven. Ein großer Anstieg zwischen zwei Positionen bedeutet, daß der nicht überlappende Bereich die Aminosäuren enthält.

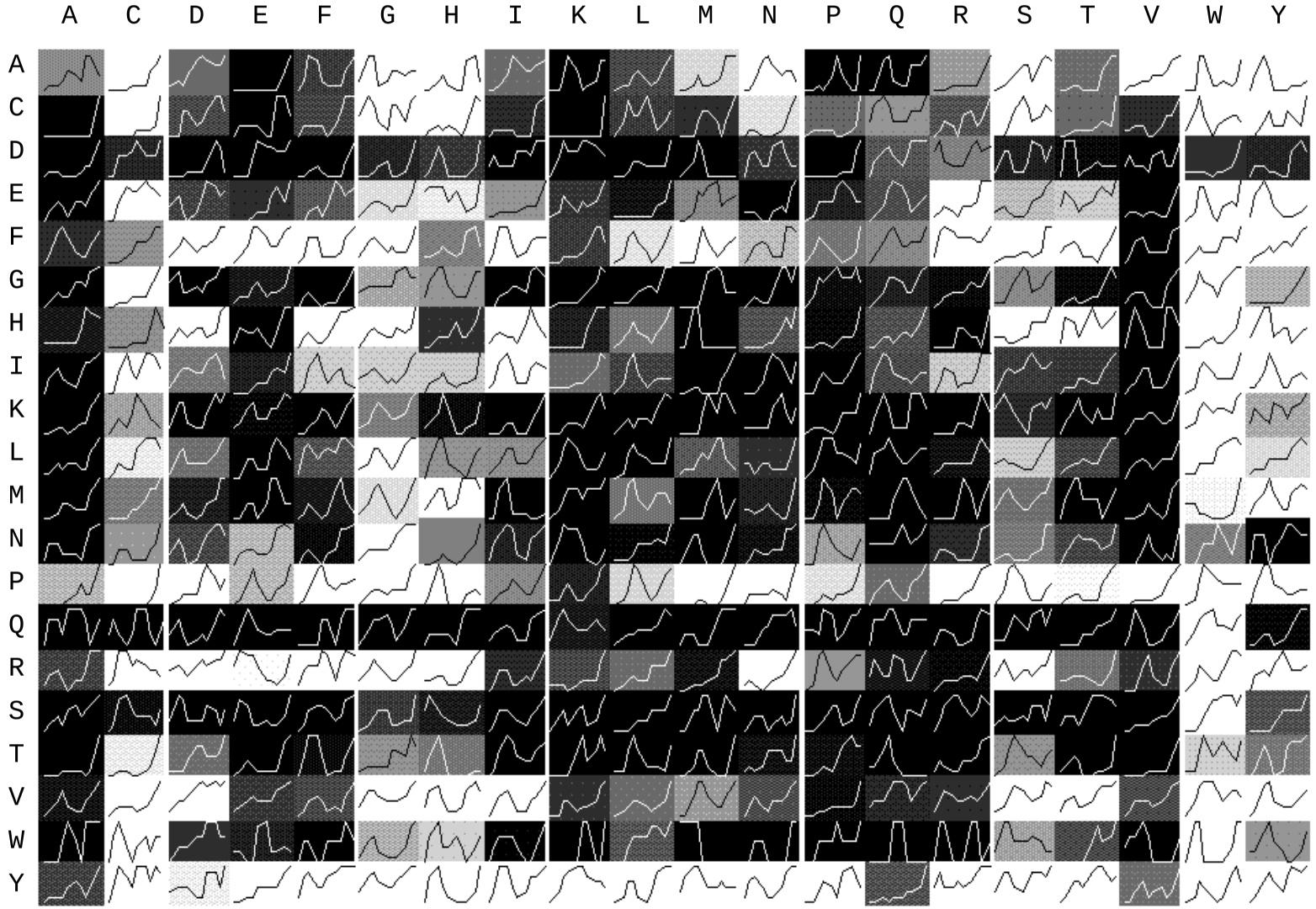

ABBILDUNG 49: Enzym-Inhibitor Komplexe: Aminosäurepartner der Enzym-Inhibitor-Komplexe. Horizontal sind die Aminosäuren des Rezeptors aufgetragen und die des Liganden vertikal. Die Helligkeit gibt die relative Häufigkeit bezogen auf die zu erwartende Häufigkeit für eine Wechselwirkung an. Die Werte wurden auf einen Bereich von 0 (schwarz) bis 1 (weiß) abgebildet. Eine Häufigkeit, die einer zufälligen Verteilung entspricht, wird grau dargestellt. Helle Bereiche spiegeln erhöhte Wechselwirkungen wider. Die Linien innerhalb der einzelnen Felder stellen die Verteilung der Aminosäuren auf die 10 Abstände dar. Aufgrund der überlappenden Abstandsbereiche ergeben sich geglättete Kurven. Ein großer Anstieg zwischen zwei Positionen bedeutet, daß der nicht überlappende Bereich die Aminosäuren enthält.

Für das Design der beiden Nona-Peptide wurden die Enzym-Inhibitor Komplexe weiter ausgewertet. In Tabelle 24 sind für die Aminosäuren des Rezeptors die wahrscheinlichsten Ligandenaminosäuren aufgelistet. Die relativen Wahrscheinlichkeiten sind bezogen auf die nach McCaldon und Argos [11] zu erwartenden Häufigkeiten angegeben (Gleichung 44). Ein Wert größer als Eins bedeutet, daß sich diese Aminosäure als Ligand entsprechend häufiger in der Nähe der Rezeptoraminosäure befindet. Ein Wert kleiner als Eins entspricht einer relativ unwahrscheinlichen Wechselwirkung.

Die beiden zur Synthese vorgeschlagenen Peptide wurden mit Hilfe des Computerprogramms InsightII (BIOSYM) in die Canyonregion des Rhinovirus eingepaßt. Dazu wurden an einer Stelle in der Nähe des Randes der tiefen Canyonregion (Abbildung 50) die Aminosäuren der Oberfläche bestimmt. Da sich an einer Bindungsstelle mehrere Aminosäuren des Rezeptors in näherer Umgebung der Aminosäure des Liganden befinden, wurde die Oberfläche aller Aminosäuren in dieser Region abgeschätzt. Entsprechend der Oberflächenanteile wurde aus Tabelle 24 die wahrscheinlichste Aminosäure bestimmt, indem die Produkte aus relativer Häufigkeit und Oberflächenanteil für die einzelnen Aminosäuren des Liganden berechnet wurden. Diese Aminosäure wurde so gedreht, daß eine möglichst große Wechselwirkung mit dem Rezeptor entsteht. Die nächste Aminosäure wurde nach gleichem Schema an den N-Terminus angehängt. Sollte der Rest der Aminosäure vom Canyon wegragen, so ist ein Serin als hydrophile Aminosäure eingeführt worden. Die resultierende Sequeenz (Nr.1, Tabelle 25) lautet: „GSQGPKPWG“.

TABELLE 24. Wahrscheinlichste Liganden-Aminosäuren bei vorgegebener Rezeptor Aminosäure. Es sind die relativen Häufigkeiten bezogen auf die erwartetete Häufigkeiten der ersten 10 Aminosäuren und die entsprechenden Aminosäuren angegeben. D. h. Methioin kommt als Ligand in der Nähe eines Alanins 6,2 mal häufiger vor als es nach McCaldon et al. [11] zu erwarten ist.

Rezeptor Aminosäure	**1.**		**2.**		**3.**		**4.**		**5.**		**6.**		**7.**		**8.**		**9.**		**10.**	
A	6,2	M	5,2	H	4,6	C	4,3	N	3,8	V	2,6	W	2,4	Y	2,2	A	1,8	S	1,6	F
C	1,8	H	1,6	W	1,1	C	0,7	Y	0,7	Q	0,6	M	0,6	S	0,5	N	0,4	P	0,4	G
D	1,7	Q	1,1	H	1,1	C	1,0	M	0,9	N	0,9	R	0,7	W	0,7	K	0,6	P	0,6	F
E	3,8	C	3,6	M	3,2	R	2,7	H	2,4	W	1,9	Q	1,9	Y	1,4	T	1,3	K	1,3	I
F	8,5	W	3,6	M	3,2	Y	2,4	F	2,3	D	2,2	I	1,6	T	1,6	E	1,5	Q	1,5	R
G	6,0	W	4,2	C	2,4	H	1,7	Q	1,4	M	1,2	Y	1,2	P	0,9	E	0,8	F	0,7	S
H	1,0	W	1,0	D	1,0	F	0,9	I	0,7	C	0,6	T	0,6	Q	0,6	Y	0,4	H	0,4	N
I	3,0	C	2,0	H	1,7	Y	1,7	W	1,7	F	1,7	I	1,5	Q	1,4	K	1,1	R	0,9	M
K	2,6	W	2,0	C	0,9	Y	0,9	H	0,7	P	0,6	E	0,6	N	0,6	M	0,5	Q	0,5	F
L	5,7	W	4,2	M	3,9	C	3,0	H	1,9	I	1,9	F	1,7	Y	1,5	N	1,4	D	1,3	T
M	1,3	H	0,7	W	0,7	Y	0,7	C	0,6	M	0,3	P	0,3	N	0,3	L	0,3	F	0,3	K
N	1,5	P	1,5	C	1,3	H	1,1	E	0,9	W	0,8	M	0,6	T	0,6	F	0,6	N	0,5	D
P	16,	W	5,7	Y	5,6	H	5,6	C	4,4	M	4,0	F	2,7	V	2,7	N	2,3	R	2,1	P
Q	2,4	W	0,9	K	0,8	M	0,7	C	0,6	P	0,6	H	0,5	I	0,4	T	0,4	Q	0,4	N
R	8,4	C	6,4	W	4,9	H	2,9	N	2,5	F	2,2	Y	2,0	D	1,9	E	1,8	P	1,8	M
S	3,6	W	1,3	H	1,1	C	0,8	F	0,8	Y	0,7	K	0,6	M	0,6	Q	0,5	R	0,5	N
T	2,5	C	1,7	H	1,5	W	0,9	P	0,9	D	0,8	M	0,8	F	0,8	Y	0,8	N	0,6	S
V	7,3	W	4,1	M	3,9	C	3,2	H	2,7	I	2,5	Y	2,0	D	1,9	T	1,7	Q	1,4	K
W	1,1	C	0,6	W	0,5	H	0,2	M	0,2	Y	0,1	T	0,1	E	0,1	L	0,1	P	0,1	D
Y	6,9	C	4,2	M	3,5	W	3,5	H	3,2	K	2,2	Y	2,0	F	1,9	P	1,7	E	1,6	N

Das zweite Peptid wurde nach der gleichen Methode wie Peptid 1 konstruiert, nur daß jetzt unwahrscheinliche Aminosäuren an den entsprechenden Positionen aus Tabelle 24 ausgewählt wurden. So ergab sich die Sequenz Nr.2 „KSTHYESPG“.

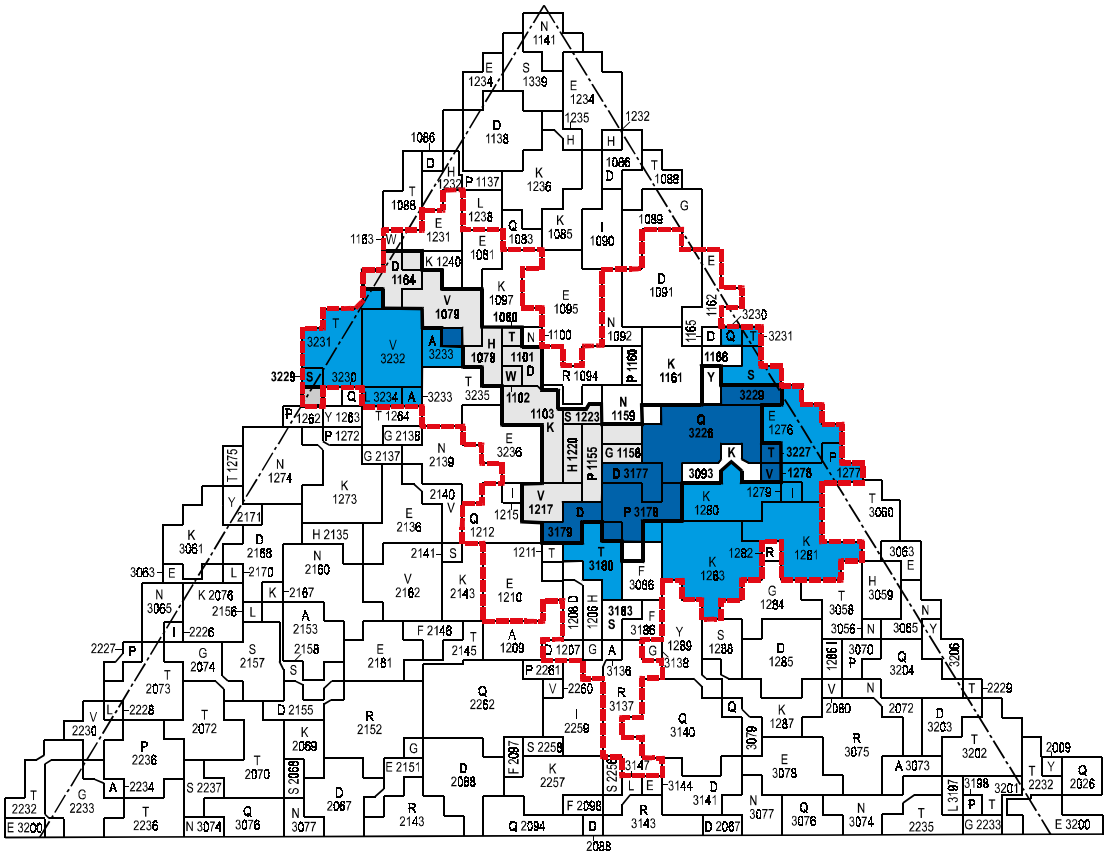

ABBILDUNG 50: Karte eines Teils der Oberfläche des Rhinovirus 14 (frei nach [98]). Der Canyonbereich ist rot umrahmt. Die Aminosäuren werden mit zunehmender Tiefe zum äußeren Grat zunehmend dunkler gezeichnet.

Die experimentell bestimmten inhibitorischen Befunde sind in relativ guter Übereinstimmung mit den vorhergesagten Eigenschaften. Das Peptid mit der gering vorhergesagten Bindungsaffinität besitzt keine inhibitorischen Eigenschaften. Peptid Nr. 1, für das eine zellprotektive Eigenschaft vorausgesagt wurde, besitzt auch eine marginale inhibitorische Wirkung (Tabelle 25).

TABELLE 25. 5 Sequenzvorschläge für die Synthese und zellinhibitorische Wirkung.

Nr.	Sequenz	Inhibition
1	GSQGPKPWG	-25%
2	KSTHYESPG	-16%