KAPITEL 5 Diskussion

In der vorliegenden Arbeit werden Verfahren beschrieben, mit denen es möglich ist, aus vorhandenen Sequenz- und Strukturdaten von Proteinen den Zusammenhang zwischen diesen Daten und einer Funktion zu beschreiben. Dabei wird ein neues Verfahren, die adaptiv kodierten KNN [43], zur Vorhersage von Schnittstellen humaner Signalpeptide vorgestellt. Es wird gezeigt, daß sich funktionale Zusammenhänge von Proteinen aus der Primärstruktur ableiten lassen.

Die Untersuchung der vorliegenden Peptidylprolylsequenzen ergab keine Anhaltspunkte für einen Zusammenhang zwischen lokaler Sequenzinformation und Konformation der Peptidbindung.

Für die Entwicklung neuer Peptide in eine gegebene dreidimensionale Umgebung wird ein weiterer Algorithmus erfolgreich vorgestellt. Ausgehend von der 3-D-Struktur des Rhinovirus wurden Peptide entwickelt, die inhibitorische Effekte besitzen.

5.a Schnittstellen humaner Signalpeptide

Für die Bestimmung der Sequenz-Aktivitäts-Relation (SAR) bedarf es Verfahren, die möglichst wenig freie Parameter aufweisen. Ziel ist es, komplexe Zusammenhänge anhand weniger Beispiele darzustellen.

Die vorliegende Arbeit stellt ein KNN mit wenigen freien Parametern für das Beispiel der Schnittstellensequenzen humaner sekretorischer Proteine vor. Für die Merkmalsextraktion wurden ausschließlich experimentell verifizierte Daten verwenden.

Für die Untersuchung der vorliegenden Schnittstellendaten wurden verschiedene bekannte sowie die hier neu vorgestellte Methode angewendet. Es kann nun ein Vergleich der Ergebnisse und eine Bewertung der Methoden erfolgen.

Die **experimentell verifizierten**, **nicht-homologen** Sequenzdaten aus der SWISSPROT wurden mit den nachstehenden Verfahren untersucht:

1) Informationsanalyse,

2) Lineare Trennverfahren (Schwerpunktsanalyse, Bayes-Analyse),

3) Hauptkomponentenanalyse/Mahalanobisanalyse

4) Künstliche neuronale Netze (KNN) und adaptiv kodierte neuronale Netze.

zu 1) Die Ergebnisse der Informationsanalyse der Schnittstellensequenzen zeigen, daß die Positionen -1, -3 und -6, -7 die wesentliche Information enthalten. Damit wurde die bisherige (-1, -3)-Regel [35] um die Positionen -6 und -7 erweitert. Diese Information konnte anschließend mit komplexeren Methoden zur Klassifikation verwendet werden. Für die Peptidylprolylsequenzen ergab sich keine positionsspezifische Information. Hier konnten auch die anderen bisherigen Methoden keine zufriedenstellende Klassifikation erreichen.

Die Informationsanalyse erweist sich somit als eine zuverlässige Methode, um in einem gegebenen Datensatz Positionen von Interesse aufzuzeigen. Insbesondere kann relativ einfach ermittelt werden, ob sich ein Datensatz für eine Klassifikation eignet. Die Klassifikation selbst kann jedoch damit ebensowenig wie der Zusammenhang zwischen den Sequenzpositionen erkannt werden. Für die Informationsanalyse werden pro Sequenzposition nur 20 Parameter bestimmt (die relativen Häufigkeiten der Aminosäuren), was als günstig anzusehen ist.

2) Lineare Trennverfahren (Schwerpunktanalyse und Bayesanalyse) ermöglichten die partielle Trennung der Sequenzen. Die Ergebnisse zeigen, daß eine gute Klassifikation schon mit diesen Methoden möglich ist. Nachteilig bei diesen beiden Methoden ist, daß immer noch 240 Parameter bestimmt werden müssen. Es gibt eine Faustregel, nach der für eine zuverlässige Statistik mindestens dreimal so viele Beispieldaten wie freie Parameter zur Verfügung stehen müssen. Die Ergebnisse der ACN zeigen sogar, daß erst die fünffache Menge an Daten ausreicht. Demzufolge müßten bei 240 freien Parametern 1200 Daten zur Verfügung stehen. Die 76 Beispiele sind nach diesen Kriterien jedoch nur für 15 bis 25 Parameter ausreichend.

3) Die Hauptkomponentenanalyse lieferte insgesamt acht Sequenzen mit vorhergesagter Schnittstelleneigenschaft. Auch wenn mit dieser Methode keine Trennung der Daten möglich ist, so konnten doch Sequenzen mit bestimmten Eigenschaften abgeleitet werden.

Die Entwicklung eines Vorhersagesystems für unbekannte Sequenzen mittels der Mahalanobis-Analyse erfordert bei einer verteilten Kodierung und dem 12er Fenster die Anpassung von 240 * 240 = 57600 Parametern. Dieser große Unterschied zwischen der Anzahl der freien Parameter und vorhandenen Daten kann die niedrige Vorhersagequalität erklären. Diese Methode eignet sich unter den gegebenen Vorraussetzungen nur schlecht als Klassifikationsmethode.

4) Künstliche Neuronale Netze (KNN) sind schon erfolgreich unter anderem zur Vorhersage von mitochondrialen Schnittstellen [105, 107] und von transmembranen Regionen [74] angewendet worden. Mit KNN ist es möglich, die Anzahl der freien Parameter durch eine entsprechend angepaßte Architektur des Netzes zu reduzieren. Hier sind zum einen die Arbeiten von Lohman et al. [69] zu erwähnen, die die komplette Netzwerkarchitektur in einem Evolutionsalgorithmus optimieren. Ähnliche Ansätze verfolgen Pruningalgorithmen, die die Anzahl der Verbindungen in einem Neuronalen Netz sukzessive reduzieren [48]. Der Eingaberaum läßt sich durch die Verwendung physikochemischer Eigenschaften verkleinern. Da die optimalen Eigenschaften in der Regel nicht bekannt sind, müssen alle Kombinationen ausprobiert werden, um die optimalen Eigenschaften zu finden. Damit wird das Problem der großen Datenräume in ein kombinatorisches verlagert. Eine adaptiven Kodierung wurde bereits von Riis et al. zur Sekundärstrukturvorhersage verwendet [61, 95]. Diese Ansätze sind aber nicht weiterverfolgt worden.

Der in dieser Arbeit vorgestellte *weight-sharing* Ansatz [33, 43] zur adaptiven Bestimmung der Kodierung reduziert die Anzahl der freien Parameter, ähnlich wie die physikochemischen Eigenschaften, jedoch braucht kein Vorwissen über die Eigenschaften einzufließen. Die Methode der ACN lieferte die besten Resultate für Training (cc = 0.93) und Test (cc = 0.56, Tabelle 8). Im Vergleich zu anderen *weight-sharing* Methoden [102] besitzt das hier neu vorgestellte ACN keinen Bias in der Kodierungsschicht. Dadurch ist eine einfache Interpretation in Hinblick auf physikochemische Eigenschaften erleichtert. Bisher wurden *weight-sharing* Ansätze auch nicht zur Analyse von Eigenschaften oder Positionen verwendet. Wie in dieser Arbeit gezeigt werden konnte, liefern solche Analysen aber wichtige Hinweise auf die biologische Funktion.

Um die Zuverlässigkeit der adaptiv kodierten KNN (ACN) zu testen, ist eine statistische Analyse durchgeführt worden. Hierzu wurden die Ergebnisse des letzten Trainingsschrittes des konvergierten Netzes verwendet.

Es ist allgemein üblich, das Netz, das die meisten Sequenzen des Testdatensatzes richtig zuordnet, als Grundlage für die Klassifikation unbekannter Sequenzen zu verwenden. Das bedeutet, daß auch Netze verwendet werden können, bei denen rein zufällig die Testdaten gut klassifiziert werden.

In der vorliegenden Arbeit wurde deshalb der Lernprozeß nicht vorzeitig abgebrochen, um diese Fehlerquelle auszuschließen. Das Training wurde solange durchgeführt, bis sich die Gewichte nicht mehr wesentlich veränderten (Konvergenzkriterium: $\sigma < 0.001$).

Drei Indikatoren zur Bestimmung der Leistung des ACN sind in Tabelle 8 zusammengefaßt: Die durchschnittliche Vorhersagegenauigkeit wird über die gemittelten cc(test)-Werte bestimmt. Die Standardabweichung gibt die Zuverlässigkeit der mittleren Vorhersagegenauigkeit an. In allen praktischen Anwendungen wird der Rechenaufwand ein limitierender Faktor sein. Er ist deshalb in die Tabelle mitaufgenommen worden. Zu beachten ist, daß die Standardabweichung die erwartete Fluktuation der individuellen cc-Werte ist und nicht die Fluktuation der Mittelwerte. Die Fluktuation der Mittelwerte wird als $\text{std.dev.}/\sqrt{\text{No. of runs}}$ bestimmt. Für Tabelle 8 ergibt sich eine Variation von ungefähr 0,01.

Die beste Vorhersagegenauigkeit (cc-test) wurde mit dem adaptiven Algorithmus erzielt. Der Unterschied zur verteilten Kodierung ist nicht sehr groß, aber unter Berücksichtigung des Rechenaufwandes und der Anzahl der zu bestimmenden Parameter, wird der Vorteil der adaptiven Kodierung deutlich. Es konnte der Rechenaufwand auf weniger als die Hälfte reduziert werden (32.1 : 77.7).

Um zuverlässige Netze zu bekommen, die sich für eine Vorhersage von unbekannten Sequenzen eignen, muß ein Auswendiglernen ausgeschlossen werden. Auswendiglernen kann an der Abnahme der Testergebnisse gemessen werden. Es tritt immer dann auf, wenn zu viele Parameter eingeführt werden (wie Hidden Units) und das Netz anfängt, die Trainingsdaten exakt zu reproduzieren, wobei die Fähigkeit verlorengeht zu verallgemeinern.

Im Falle der Vorhersage von Schnittstellensequenzen humaner sekretorischer Proteine konnte gezeigt werden, daß für eine verteilte Kodierung mit der einfachsten Netzarchitektur keine zuverlässigen Netze erzeugt werden konnten, was an dem Unterschied der entsprechenden cc-Werte für Training und Test zu erkennen ist.

Mit den physikochemischen Eigenschaften können nur dann gute Ergebnisse erzielt werden, wenn Hidden Schichten eingeführt werden. Da vergleichbare Ergebnisse mit einer einfacheren Netzarchitektur unter Verwendung der ACN erzielt werden konnten, sind die physikochemischen Eigenschaften nicht optimal, um in dem hier behandelten Sinne gute Ergebnisse zu erzielen. Die Adaptive Kodierung dagegen ermöglicht die Entwicklung von effizienten Kodierungen, die einfachere Netzwerkarchitekturen erlauben und das bei einer hohen Recheneffizienz.

Das beste gefundene Kodierungsschema ist in guter Übereinstimmung mit der binären Eigenschaft „*polar oder Prolin*“. Sie steht auch nicht in Widerspruch mit der (-1, -3)-Regel von von Heijne, bei der sich kleine hydrophile Aminosäuren an Position -1 und -3 relativ zur Schnittstelle befinden sollten.

Im Gegensatz zu den vordefinierten physikochemischen Eigenschaften werden die Eigenschaften, die sich für das ACN ergeben, direkt aus den Daten bestimmt. Dadurch wird eine problemspezifische Kodierung entwickelt. Die Anzahl der Kodierungsvektoren läßt sich einfach über eine Hauptkomponentenanalyse bereits berechneter Kodierungsvektoren bestimmen.

Auch eine Kombination von physikochemischen Eigenschaften, die der neuen adaptiven Kodierung am besten entspricht, läßt sich durch einen Vergleich der Basissysteme bestimmen. Die hierbei ermittelten Eigenschaften *Hydrophilizität*, *Hydrophobizität* und *polar oder Prolin* (Tabelle 11) sind in guter Übereinstimmung mit den Strukturuntersuchungen der Signalpeptidase von *e-coli*. Bei diesen Untersuchungen zeigte sich, daß Hydrophobizität eine wichtige Rolle für die effiziente Prozessierung spielt [85].

Die Analyse der entsprechenden Eigenvektoren unterstützt wiederum die (-1, -3)-Regel. Jedoch scheinen die Position -7 und -6 eine besondere Rolle für die menschlichen Schnittstellen zu spielen. Dies ging auch schon aus der Informationsanalyse hervor. Eine mögliche Erklärung mag hierfür die β-Faltblatt-Struktur sein, die für das native Protein im aktiven Zentrum vorhergesagt wird, die zwischen 5 bis 6 Aminosäuren lang sein soll [85].

An den Positionen -6 und -7 treten mit 60% beziehungsweise 67% die Aminosäuren Leucin, Valin, Alanin und Glycin auf. Diese Positionen können am besten mit der Eigenschaft *hydrophob* beschrieben werden (Tabelle 13). Hier könnte sich das C-terminale Ende der transmembranen Helix befinden, die von Matlack et al. [76] beschrieben wird. Dies wird auch durch die Arbeiten von Nielsen et al. [81] untermauert, die zeigen, daß sich die hydrophobe Region in Eukaryonten meistens in einer Region von -17 bis -6 relativ zur Schnittstelle befindet.

Das hier beschriebene ACN erkennt zwar zuverlässig die Schnittstellen in Signalpeptiden, jedoch werden darüber hinaus auch weitere möglicherweise falsche Schnittstellen im Signalpeptid erkannt. Es wurde deshalb ein weiterer Algorithmus vorgestellt, mit dem es möglich ist, solche Schnittstellen einfach herauszufinden. Eine Verbesserung dieses Algorithmus können möglicherweise Hidden Markov Modelle erzielen. Nielsen et al. konnten durch Einführen von Hidden Markov Modellen eine Verbesserung der Vorhersage von Signalpeptiden erreichen [82].

Wie in Abbildung 24 gezeigt wird, reicht die Anzahl der vorhandenen Daten nicht für ein völlig fehlerloses Vorhersagesystem aus. Hierzu müßten circa 200 experimentell gesicherte, nicht-homologe Sequenzdaten zur Verfügung stehen. So viele sind noch nicht experimentell untersucht worden. Mit der hier vorgestellten Methode lassen sich jedoch auch schon mit einer geringen Anzahl von Daten brauchbare und in der Praxis einsetzbare Vorhersagen machen.

Ein Datensatz mit 416 nicht explizit verifizierten Schnittstellendaten von Nielsen et al. [80] wurde zu Vergleichszwecken mit dem ACN untersucht. Dabei wurden 72% der Schnittstellen korrekt vorhergesagt. Nielsen et al. haben hier nur 68% erkannt. Bereits 3.8% dieser Daten haben geladene Aminosäuren an der Position -1. Dies ist im Widerspruch zur (-1, -3)-Regel und außerdem weisen die Ergebnisse der Strukturuntersuchung der Signalpeptidase von *e-coli* darauf hin, daß es sich hierbei nicht um Schnittstellen handelt [16, 85]. Es liegt deshalb nahe anzunehmen, daß ein nicht zu vernachlässigender Anteil der Schnittstellen humaner sekretorischer Proteine, die in der SWISSPROT Datenbank nicht explizit als putativ klassifiziert worden sind, falsch sind.

TABELLE 26. Schnittstellenzuordnung mit unterschiedlichen Programmen. Es sind die korrekt/falsch zugeordneten Schnittstellenposition angezeigt. SiganlP [80] Programm zur Vorhersage der Schnittstellen und Signalpeptide von Eukaryonten. 1C-adapt = ACN mit einem Kodierungsvektor, 2C-adapt = ACN mit 2 Kodierungsvektoren, 3C-adapt = ACN mit 3 Kodierungsvektoren. *: diese Netze sind mit den 74 Sequenzen trainiert worden, ohne geladene Aminosäuren an Position (-1).

	SignalP	**1C- adapt**	**2C- adapt**	**3C- adapt**	**2C-adapt ***	**3C-adapt ***
416 Daten (Nielsen et al.)	321/95	247/169	251/165	284/132	301/115	288/128
76 Daten (Jagla et al.)	65/11	50/26	61/15	72/4	71/5	73/3

Die Auswahl geeigneter Daten ist für die Erstellung eines zuverlässigen Vorhersagesystems sehr wichtig. Da die Einträge bezüglich der Schnittstellen der SWISSPROT-Datenbank nicht alle experimentell verifiziert sind, und dies nicht immer gekennzeichnet ist, wurden die angegebenen Literaturstellen auf experimentell gefundene Schnittstellen hin untersucht. Über die-

sen sehr zeitintensiven Prozeß wurden 76 Schnittstellen identifiziert. Dieser Datensatz ist zwar recht klein, sollte dafür aber sehr zuverlässig sein. Die Ergebnisse zeigen, daß sich selbst in diesem Datensatz mehrere Sequenzen befinden, die keine typischen Schnittstellen besitzen.

Es gibt zwei Sequenzen, die an der Position -1 eine geladene Aminosäure besitzen (FINC und CAP7). Desweiteren besitzt die Sequenz APVQALQQAG|IV (TRYA, mit | als Schnittstellenposition) sehr viele geladene Aminosäuren in der Nähe der Schnittstelle, was sich ungünstig auf die Vorhersageeigenschaften auswirkt. Alle drei Sequenzen werden von den meisten ACN nur sehr schlecht vorhergesagt. Sollten diese Sequenzen tatsächlich an diesen Positionen keine Schnittstellen besitzen, so sind in diesem Datensatz von experimentell verifizierten Daten circa 4% (3/76) falsch. Der Datensatz der SWISSPROT, der auch nicht experimentell verifizierte Daten enthält, enthält somit wahrscheinlich einen noch höheren Anteil an falschen Schnittstellenpositionen.

Die einfachen linearen Ansätze erlauben eine schnelle Einschätzung des Problems, inwieweit das gegebene Klassifikationsproblem lösbar erscheint. Von den Methoden zur Klassifikation/ Reklassifikation eignen sich die Adaptiv Kodierten Neuronalen Netze am besten. Sie zeichnen sich im Falle der Schnittstellen durch die beste Vorhersagequalität aus und bedürfen nur einer sehr geringen Anzahl an freien Parametern. Die Netze lassen sich einfach auf biologisch interessante Eigenschaften hin untersuchen. Selbst komplexe Zusammenhänge können hier in physikochemische und positionsspezifische Anteile aufgeteilt werden.

Wie können diese Ergebnisse in Bezug gesetzt werden zu den bereits bekannten Ergebnissen über den Translokationsapparat?

Der entscheidende initiale Schritt in der Biogenese vieler Proteine von eukaryontischen Zellen (insbesondere lösliche und Membranproteine des Endoplasmatischen Retikulums (ER), des Golgi-Apparats, der Lysosomen, der Plasmamembran und Membranproteine des Zellkerns und der Peroxisomen) ist deren Transport in das Endoplasmatische Retikulum [127]. Typischerweise bedarf der Proteintransport ins ER eines Signalpeptides am Aminoterminus des jeweiligen Precursorproteins und eine Transportmaschinerie mit löslichen- und Membranproteinen.

Die Vorhersage der Signale und die Zuordnung zu den einzelnen Kompartimenten ist hinreichend verstanden [14]. So gibt es bereits verschiedene Methoden für deren Vorhersage. Das am häufigsten angewendete Verfahren ist SignalP [81]. Für eine detaillierte Beschreibung dieser statistischen und bioinformatischen Verfahren sei auf [12, 14] und darin enthaltene Literatur verwiesen.

Die Signalpeptide von Proteinen, die ins ER transportiert werden und nicht membrangebunden bleiben, werden von der Signalpeptidase abgespalten. Somit ist die Funktion der Signalpeptidase *in vivo* die Freisetzung bestimmter Proteine, die an die ER-Membran gebunden sind.

Hier wird nur der Teil des Translokationsapparates betrachtet, der den Peptidasekomplex enthält und der das Signalpeptid vom naszierenden Protein abschneidet. Die Zusammensetzung von Signalpeptidasekomplexen verschiedener Systeme ist in Tabelle 27 zusammengestellt. Signalpeptidasen von Säugetieren sind aus dem Hundepankreas als Komplex aus fünf Untereinheiten gereinigt worden. Die cDNAs aller fünf Untereinheiten sind geklont und sequenziert worden [79]. Es bestehen große Unterschiede zwischen den Topologien dieser Proteine innerhalb der ER-Membran. Drei von ihnen, SPC18, SPC21 und das Glycoprotein SPC22/23 sind monotopische Membranproteine mit den Aminotermini in Richtung Cytosol. Der größte Bereich dieser Proteine befindet sich im Lumen des ER, und alle besitzen eine zweite hydrophobe Region in der Nähe ihres Carboxyterminus. SPC18 und SPC21 besitzen eine hohe Homologie. Darüber hinaus sind ihre Sequenzen mit der Leaderpeptidase verwandt, einem Enzym, das für die Signalabspaltung von sekretorischen Proteinen in Bakterien verantwortlich ist. Deshalb sollten diese Polypeptide als katalytische Untereinheiten funktionieren [79].

TABELLE 27. Beziehung zwischen mikrosomalen Signalpeptidaseuntereinheiten [68]. Proteine mit Peptidaseeigenschaften sind kursiv dargestellt.

Proteingruppen	**Hund**	**Huhn**	**Hefe**
25 kD	SPC25	nicht gefunden	20kD
Glykoprotein	*SPC22/23*	*GP23*	*25kD*
Sec11 Homologe	*SPC21* *SPC18*	*P19*	*Sec11p*
12kD	SPC12	nicht beobachtet	13kD

SPC12 und SPC25 besitzen bitopische Bereiche, wobei der Amino- und der Carboxyterminus jeweils ins Cytosol ragen. Fast keine Aminosäuren befinden sich im Lumen des ER. Die Beobachtung, daß diese Proteine im Huhn nicht gefunden werden, stützt die Annahme, daß die Funktion der Proteine nicht für die Peptidaseaktivität verantwortlich ist.

Die eukaryontischen und prokaryontischen Signalpeptidasen (SPase) gehören zu einer einzigen Peptidasefamilie und weisen vor allem im Bereich des katalytischen Zentrums die größte Übereinstimmung in der Sequenz auf. Da noch keine Kristallstruktur einer eukaryontischen Signalpeptidase bekannt ist, wird im folgenden die Kristallstruktur der Signalpeptidase aus

Escherichia coli als Komplex mit einem β-Lactamrezeptor betrachtet [85]. Die Domäne I, die alle wichtigen und konservierten katalytischen Elemente enthält, repräsentiert ein neues Proteasemotiv, das vom Bakterium bis hin zum Menschen konserviert zu sein scheint. Eine große, ungewöhnlicherweise exponierte hydrophobe Fläche erstreckt sich über die SPase und beinhaltet die Substratbindungsstelle und das katalytische Zentrum. Da Serin die Aminosäure ist, die den nucleophilen Angriff auf den Carboxykohlenstoff der Peptidbindung durchführt, handelt es sich um eine Serinprotease. Der Mechanismus des Spaltens der Peptidbindung verläuft über einen re-seitigen Angriff [62] einer Serinhydrolase. Diese Art von Angriff ist bisher unbekannt gewesen für Serinproteasen [85].

Das Substrat der Signalpeptidasen ist die C-terminale Region der Signalpeptide. Obwohl die Signalpeptide keine Sequenzhomologie aufweisen [122], besitzen sie doch ähnliche charakteristische Eigenschaften. Die Signalpeptide variieren in der Länge von ca. 13 bis 50 Aminosäuren [36]. Zwar gibt es keine Konsensussequenz der Signalpeptide, dafür haben alle eine drei-Domänenstruktur [34]. In der Nähe des N-Terminus befindet sich eine basische Aminosäure (N-Region). Danach folgt ein hydrophober Bereich aus mindestens acht apolaren Aminosäuren (H-Region), der in der Regel von einer Helix-brechenden Aminosäure (Ser, Gly oder Pro) von ungefähr 4-6 Aminosäuren vor der Schnittstelle abgeschlossen wird. Der dritte Bereich ist die Schnittstellenregion (C-Region), die sehr kleine Aminosäuren an der Position -1 (relativ zur Schnittstelle) benötigen (Ala, Gly, Ser oder Cys). An der Position -3 können zu den eben genannten vier Aminosäuren auch noch Ile, Val, Thr und Leu auftreten. Die Sequenz der Signalpeptide von Eukaryoten und Prokaryoten ist nicht sehr ähnlich. Die eubakteriellen Schnittstellen sind weniger variabel an den Positionen -1 und -3, wo Ala die dominierende Rolle spielt. Trotz dieses Unterschiedes sind die Signalsequenzen von eubakteriellen und eukaryotischen Signalsequenzen untereinander austauschbar [68].

Mutationen in der Schnittstellenregion von humanen sekretorischen Proteinen können bestimmte Krankheiten auslösen. Eine Reihe von Mutationen in unmittelbarer Umgebung der Schnittstelle wurden beschrieben [3, 10, 17, 123]. Die mutierten und nicht mutierten Proteine wurden mit dem ACN-Vorhersagesystem untersucht. Nur die Sequenzen der nicht mutierten Proteine von Parathyroid (PTHY_HUMAN) und Antithrombin (ANT3_HUMAN) waren im Trainingssatz des ACN vorhanden. Die Ergebnisse sind in Tabelle 28 zusammengefaßt.

TABELLE 28. Vorhersagen des Schnittstellenprediktors für die bekannten humanen Mutanten im Bereich der Schnittstelle. Unterstrichen: Vorhergesagte Signalpeptide, Fett: potentielle Schnittstellen, Leerzeichen: Schnittstelle laut Literaturangaben. Proteine mit einem * waren im Trainingssatz vorhanden. Die Mutationsstellen sind in der ersten Spalte angegeben. So entspricht V(30)->E einer Mutation des Valins an der Position 30 in eine Glutaminsäure.

Protein	SWISSPROT-ID	Sequenz/Vorhersage
Apolipoprotein	APB_HUMAN	MDPPRPALLALLAL**PA**LLLLLL**AG** **A**R**A** EEE
Antithrombin B	ANT3_HUMAN*	MYSNVIGTVT**S**GKRKVYLLSLLLI**G**FWDCV- **TC** HGSPVDICTAKPRDIPMNPMCIYRSPEK
V(30)->E		MYSNVIGTVT**S**GKRKVYLLSLLLI**G**FWDCE- TCHG SPVDICTAKPRDIPMNPMCIYRSPEK
Serum Albumin	ALBU_HUMAN	MKWVTFISLLFLFS**SA**Y**S** R**G**VFRRDAHKSE
R(23)->C		MKWVTFISLLFLFS**SA**Y**S**R**G**VFCRDAHKSE
Parathyroid Hormone	PTHY_HUMAN*	MIPAKDMAKVMIVML**A**I**C**FLTKSD**G** KSVKK
C(18)->R		MIPAKDMAKVMIVML**A**IRFLTKSD**G**KSVKK
Faktor X	FA10_HUMAN	MGRPLHLVLLSA**S**L**AG**LLLLGE**S** LFIRREQ
G(21)->R		MGRPLHLVLLSA**S**L**AG**LLLLRESLFIRREQ
Vasopressin	NEU2_HUMAN	MPDTMLPACFLGLL**A**F**S**S**A** CYFQNCPRGGK
A(18)->V		MPDTMLPACFLGLL**A**F**S**SV**C**YFQNCPRGGK

Ein Insertions/Deletions Polymorphismus ist im Signalpeptid des humanen Apolipoproteins B Gen bekannt. Dieses kodiert ein Polypeptid von 27 Aminosäuren und eines der Länge 24 [121]. Der hier vorgestellte Prediktor für Schnittstellen erkennt beide Schnittstellen korrekt.

Antithrombin Dublin ist eine Variante des Antithrombin, das durch eine Mutation an Position -3 hervorgeht, bei der Valin durch Glutaminsäure ersetzt wird [17]. Diese Mutation resultiert in einer Verschiebung der Schnittstelle um zwei Aminosäuren in Richtung C-Terminus. Die Schnittstelle des Antithrombin wird vom Vorhersagesystem erkannt. Es wird jedoch von den vorgeschlagenen Schnittstellen die falsche ausgewählt. Ein Algorithmus, der die H-Region identifiziert und danach erst nach einer Schnittstelle sucht, würde hier das richtige Ergebnis liefern. Die Schnittstelle wird in der Mutante nicht mehr erkannt, was den experimentellen Befunden entspricht.

Albumin Redhill ist eine Variante des humanen Serum Albumins, das zwei verschiedene Mutationen beinhaltet, wobei eine zur Verschiebung der Schnittstelle führen soll [10]. Eine Substitution an der vorletzten Position des Proproteins von Arginin durch Cystein, so wird vermutet, bewirkt, daß die Schnittstelle um fünf Position verschoben wird. Eine solche Schnittstelle kann von dem Prediktor nicht identifiziert werden. Die Schnittstelle des nicht mutierten

Präproalbumin wird erkannt, jedoch wird im Proprotein eine weitere Schnittstelle vorhergesagt.

Krankheiten, die durch Mutationen in Schnittstellen hervorgerufen werden, sind für das Präproparathyroid Hormon [3] und für den Koagulationsfaktor $X_{Santo\ Domingo}$ (FXsd; [123]) bekannt.

Die Substitution eines Cysteins durch ein Arginin innerhalb der H-Region des Präproparathyroid Hormon führt zu einer verminderten Translokation und/oder Prozessierung des naszierenden Proteins [3]. Die Schnittstelle des nicht mutierten Hormons wird richtig erkannt. Obwohl die Mutation im hier verwendeten Sequenzfenster der Schnittstellenregion liegt, hat sie keinen Einfluß auf die Vorhersage der Schnittstelle. Da die Mutation im Bereich der H-Region liegt, kann vermutet werden, daß die Translokationsrate beieinflußt wird und es sich somit nicht mehr um eine Schnittstellenregion handelt [90].

FXsd ist eine Mutation des humanen Koagulations Faktors X, bei der eine Punktmution zur Substitution von Glycin durch Arginin an der Position -3 führt [123]. Patienten mit dieser Mutation sind Bluter mit weniger als 1% FX-Aktivität und weniger als 5% zirkulierendem FX. Mutierte und nicht mutierte Schnittstellen werden korrekt erkannt.

Bei einer Mutation von Alanin nach Valin in der Position -1 des Arginin Vasopressin Hormons tritt ein autosomal dominanter Neurohypophysärer Diabetes insipidus auf, der auf eine vollständige Blockade der Schnittstelle zurückzuführen ist [94]. Die Schnittstelle des nicht mutierten Proteins wird korrekt erkannt und verschwindet nach der Mutation. Repaske et al. [94] vermuteten, daß durch die Mutation die Schnittstelle nicht mehr von der Signalpeptidase prozessiert wird. Mit Hilfe der vorliegenden Methode wird eine Verschiebung der Schnittstelle prognostiziert, die fatale Auswirkungen für die Funktion des Vasopressins zu haben scheint.

Die Ergebnisse der Vorhersagen zeigen, daß die Schnittstellen mit einer akzeptabelen Abweichung vorhergesagt werden. Eine noch größere Genauigkeit kann nur mit noch mehr nicht homologen Sequenzdaten erreicht werden (Abbildung 24). Außerdem ist eine weitere Methode wünschenswert, die die H-Region erkennt und somit eine Selektion der richtigen Schnittstellen ermöglicht. Die experimentellen Daten zeigen, daß mehrere Schnittstellen vorhanden sein können, die unterschiedlich gut prozessiert werden [1]. Auch dieses Ergebnis wird von der hier vorgestellten Methode wiedergegeben.

Die H-Region scheint die Schnittstellenregion in die Nähe des aktiven Zentrums der SPase zu bringen. Das würde auch erklären, warum die SPase keine transmembranen Helices von inte-

gralen Membranproteinen schneidet, oder auch Signalpeptide mit einer zu langen H-Region [63, 83]. Solche Helices erstrecken sich meistens noch über die Kopflipidgruppe hinaus, so daß nicht mehr die gestreckte Konformation präsentiert werden kann [37].

MHC-I Moleküle erkennen Peptide, die bis zu 33 Aminosäuren lang sind [125], was ungefähr der Länge der Signalsequenzen entspricht. Es liegt also nahe, daß wenigstens einige Signalsequenzen von den MHC-I Molekülen zur Zellerkennung verwendet werden [75]. Desweiteren gibt es Degradationsmechanismen im Endoplasmatischen Retikulum [8, 37, 53, 65]. Jedoch ist der Mechanismus des Transportes aus der ER-Membran ins Lumen noch ungeklärt. Ebensogut können die Sequenzen wieder ins Cytosol geschleust werden, um dort abgebaut zu werden [96].

5.b Konformationsvorhersage der Peptidylprolylbindungen

Der indirekte Weg zur SAR über die Vorhersage der Struktur von Proteinen muß nicht-lokale Sequenzbereiche einbeziehen. Mit den vorhandenen Daten ist keine Vorhersage aufgrund von lokalen Sequenzeigenschaften der Konformation von Peptidylprolylbindungen möglich, dafür lassen sich Major Histokompatibilitäts Komplexe (MHC-Moleküle) aufgrund von Prolinmotiven erkennen.

Die Bestimmung der Struktur eines Proteins aus der Aminosäuresequenz ist ein zentrales Problem bei der Funktionsbestimmung und dem *de novo* Design von Proteinen. Die Struktur kann bisher nur experimentell über die Röntgenstrukturanalyse oder die NMR-Spektroskopie bestimmt werden. Quantenmechanische Verfahren, die eine mathematische Beschreibung der Strukturen ermöglichen, scheitern zur Zeit noch an der Komplexität des Problems. Es gilt deshalb, einfache Näherungsverfahren zu entwickeln, die in der Lage sind, schnell möglichst genaue Strukturvorhersagen durchzuführen. Eine Möglichkeit dazu bietet die Suche nach Strukturhomologen durch Alignment oder Threading [116].

Da für die Analyse von Proteinstrukturen in Hinblick auf die Vorhersage der Struktur nichthomologer Proteine zu wenig Daten vorhanden sind, wird auch nach Möglichkeiten zur lokalen Strukturbestimmung gesucht. Die Tatsache, daß sich *trans*- von *cis*- Peptiden durch ihre Konformation unterscheiden und eine Umwandlung sehr langsam geschieht, legt die Überlegung nahe, zu ermitteln, unter welchen Voraussetzungen sich *cis*-Peptidgruppen bilden [42]. Die Ergebnisse dieser Arbeit zeigen unter anderem, daß es Unterschiede zwischen den beiden Isomeren auf Sequenzebene gibt. Mit den vorhandenen Daten ist es nicht möglich, eine lokale

Konformationsvorhersage der Peptidylprolylbindung zu machen. Die vorliegenden Analysen geben jedoch Hinweise auf Aspekte, die eine nähere Analyse sinnvoll erscheinen lassen.

Im folgenden werden zuerst die in dieser Arbeit gemachten Annahmen beschrieben. Anschließend werden die Ergebnisse dieser Arbeit im Hinblick auf eine mögliche Klassifikation untersucht.

Für die Untersuchungen mußten fünf Annahmen gemacht werden, die im folgenden aufgelistet werden.

(1.) Die Information über die *cis*- oder *trans*- Konformation ist lokal um das Prolin kodiert.

(2.) Die Information über die Konformation ist eindeutig. Kein Sequenzmuster kann sowohl in *cis*- als auch in *trans*-Konformation vorliegen.

(3.) Die Größe des Proteins darf keinen Einfluß auf die Isomerisierung haben. In den Proteinen werden alle Positionen gleich gut isomerisiert.

(4.) Xaa-Pro-Peptidgruppen werden durch Enzyme isomerisiert.

(5.) Die verwendeten Daten sind repräsentativ.

Die ersten vier Annahmen beruhen auf den Erkenntnissen zur enzymatischen Isomerisierung. Die letzte bezieht sich auf die verwendeten Daten und Methoden.

Die in dieser Arbeit benutzte Definition der *cis*- und *trans*-Konformation weicht von der allgemein üblichen Notation ab, weil hier der biologische Kontext mit einbezogen wurde (Annahme 4). Peptidgruppen werden entsprechend ihres ω-Winkels in *cis*- und *trans*-Isomere unterteilt. In der Literatur findet man Werte von 0° für *cis*- und 180° für *trans*-Isomere. In Peptidgruppen geht man von kleinen Abweichungen mit $\Delta\omega = -20°$ bis $+10°$ aus [15]. In der vorliegenden Arbeit wurden Peptidgruppen mit einem ω-Winkel im Bereich $\omega = 0° +/-160°$ als *cis*-Peptidgruppe klassifiziert. *Trans*-Peptidgruppen besitzen somit einen Winkelbereich von $\omega = 180° +/-20°$. Diese Definition läßt sich mit theoretischen Betrachtungen zur Rotationsbarriere und in vivo Isomerisierung begründen:

Die *in vivo* Synthese von Peptiden liefert stereospezifisch *trans*-Isomere. Da die Rotationsbarriere um die Peptidbindung mit 13 bis 20 kcal/mol sehr hoch ist [26, 89], bleiben alle ω-Winkel in einem kleinen Bereich um $\omega = 180°$. Für kleine Proteine wurde eine Standardabweichung von 7.0° des *trans*-Bindungswinkels (180°) mit *ab-initio*-Methoden berechnet [87]. Eine enzymatische Isomerisierung muß die partielle Doppelbindung schwächen, um eine Torsion zu ermöglichen. Wegen der stereospezifischen Bildung und hohen

Rotationsbarriere wird der große Winkelbereich für *cis*-Peptidgruppen gewählt. Es ist hingegen auch denkbar, daß Aminosäuren, die nicht unmittelbar in der Nähe vom Prolin sind, eine Änderung des dihedralen Winkels der Imidbindung bewirken und somit größere Abweichungen als 7° auch ohne Katalysator ermöglichen. So sind auch die größeren Abweichungen von $\Delta\omega$ -20° bis +10° zu erklären, die in der Literatur angegeben sind.

Die Folgen der Rotationsbarriere spiegeln sich in der Verteilung der ω-Winkel wider. Abbildung 32 zeigt die Verteilung der Xaa-Pro-Peptidgruppen in Abhängigkeit vom ω-Winkel. Xaa-Pro-Reste mit einem ω-Winkel zwischen $\omega = -160°$ und $\omega = 160°$ bilden ca. 7% aller Xaa-Pro-Peptidgruppen. Stewart et al. [117] haben in der Brookhaven Protein Data Bank 6.5% der Xaa-Pro-Reste in *cis*-Konformation gefunden (*cis*: -90° - +90°). Das Verhalten der hier gefundenen Kurve steht somit in Einklang mit den Überlegungen zur in vivo Synthese von Peptiden und der enzymatischen Isomerisierung. Das etwas asymmetrische Verhalten mit der Tendenz hin zu positiven ω-Winkeln tritt in allen bisherigen Untersuchungen auf. Die Ursachen dafür konnten noch nicht erklärt werden. Für die Vorhersage der Konformation ist dieser Befund nicht von Bedeutung, da der Effekt sowohl bei den *cis*- als auch bei den *trans*-Sequenzen auftritt.

Da angenommen wird, daß die Erkennung des *cis*-Prolins lokal kodiert ist, stellt sich die Frage nach der Größe des lokalen Sequenzbereiches. Eine Antwort geben Frömmel und Preissner [24], die ein 12er-Fenster für die Klassifizierung von *cis*- und *trans*-Prolinen verwenden. Wie aus Abbildung 34 und 35 hervorgeht, sind nicht zu vernachlässigende Abweichungen der relativen Häufigkeiten auch in Positionen, die wesentlich weiter entfernt sind, zu finden. Soweit nicht anders vermerkt, wurde hier ein Aminosäurefenster von 20 gewählt, um auch weiter entfernte Wechselwirkungen einzubeziehen und trotzdem die Anzahl der aus der Fenstergröße resultierenden freien Parameter möglichst klein zu halten.

Für eine Einschätzung des Datensatzes bezüglich der Verallgemeinerbarkeit auf nicht kristallisierte Proteine wurden die relativen Häufigkeiten der Aminosäuren aus der PDB-Datenbank, dem SELECT-Datensatz und den nach McCaldon und Argos [11] ermittelten Werten miteinander verglichen. Es treten keine wesentlichen Unterschiede zwischen den Strukturdatensätzen auf, außer für die Aminosäure Glycin. Letzere ermöglicht wesentlich flexiblere Strukturen als alle anderen Aminosäuren. Die Abweichungen, die mit zum Teil über 20% deutlich sind, müssen bei den weiteren Überlegungen berücksichtigt werden.

Die Häufigkeiten der ω-Winkel-Verteilung von Peptidbindungen in Tabelle 16 zeigt, daß Peptidylprolylbindungen von allen Peptidbindungen am häufigsten vorkommen. Die Werte liegen

im Bereich, der auch von anderen Arbeiten gefunden wurde [29, 124]. Eine getrennte Analyse der Xaa-Pro-Bindungen ist deshalb sinnvoll [103]. Hierzu sei auf die Arbeiten von Jabs et al. [41] verwiesen, die sich ausführlich mit nicht *cis*-Prolinbindungen beschäftigen.

Die Informationsanalyse der Xaa-Pro-Sequenzen zeigt, daß kein Unterschied zu einer zufälligen Verteilung der Aminosäuren an den einzelnen Positionen besteht. Wie bereits erwähnt, ist das ein sehr schlechtes Zeichen für eine mögliche Trennung der Daten, denn unter diesen Umständen müßten die Klassen vollständig durch Korrelationen zustandekommen, oder andere Einflüsse, die nicht in den Sequenzen berücksichtigt werden, haben einen entscheidenden Einfluß, wie z.B. nicht-lokale Wechselwirkungen.

Trotz dieser Ergebnisse zeigt die Analyse der relativen Häufigkeiten, daß es sehr große Abweichungen zwischen den Konformeren an verschiedenen Positionen des Fensters gibt (Tabelle 17, Abbildung 34 und 35). So treten die größten Abweichungen in einer Region um Position +/-30, 20 und +/-10 auf. Inwieweit diese Unterschiede auf strukturelle Besonderheiten der Sequenzen zurückzuführen sind, muß durch weitere Untersuchungen gekärt werden. Ein Grund hierfür könnte in Sekundärstrukturelementen liegen, die an diesen Positionen aufhören. So treten Cysteine an Position +/-7 häufiger auf. Sulfidbrücken könnten einen Loop der Länge 13 mit einem *cis*-Prolin in der Mitte bevorzugen. Eine Auswertung dieser Informationen läßt sich mit diesem Datensatz jedoch nicht für eine Klassifikation ausnutzen. Eine Bedeutung für spezielle Gebiete der Strukturvorhersage sind aber ausdrücklich nicht ausgeschlossen.

Der N-terminale Nachbar des Prolins ist von besonderer Bedeutung. Dieser Rest kann mit dem Prolinring in Wechselwirkung treten und eine bevorzugte gestaffelte Konformation annehmen, die in einer *cis*-Konformation begünstigt ist [93]. In der Tat finden sich hier Phenylalanin und Tyrosin häufiger in *cis*-Sequenzen (Abbildung 36). Auch Tryptophan kommt hier etwas häufiger vor.

Die Projektionen der *cis*- und *trans*-Daten auf die beiden größten Hauptkomponenten der *cis*-Daten zeigten, daß es in dieser Projektion fast keine Unterschiede zwischen den beiden Datenmengen gibt (Abbildung 38). Bei der Projektion auf die Hauptkomponenten der *trans*-Daten ist die Verteilung der *trans*-Sequenzen etwas ausgedehnter, aber die meisten Daten liegen in der Nähe des Koordinatenursprungs, wo auch die *cis*-Sequenzen hinprojeziert wurden. Es gibt hier keine Unterschiede zu einer zufälligen Verteilung der Aminosäuren. Dies zeigt sich auch in den einzelnen Eigenvektoren und Eigenwerten.

Ein etwas anderes Bild zeigt sich für die Projektion der Sequenzdaten auf die Hauptkomponenten der *cis*-Daten. Hier zeigen einige wenige *cis*-Sequenzen zum Teil erhebliche Abweichun-

gen. Es ist möglich, diese Proteinsequenzen bestimmten Proteinklassen zuzuordnen. Die MHC Moleküle zeigen hier die größten Unterschiede. Diese eigentlich sehr unterschiedlichen Sequenzen haben offensichtlich sehr eng begrenzte Sequenzbereiche, die eine hohe Homologie aufweisen. Die andere Proteinklasse, die sich unterscheidet, sind Hydrolasen, von denen einige ähnliche *cis*-Sequenzen aufweisen.

Bis auf diese Besonderheiten, die sich eventuell für die Strukturvorhersage bestimmter Proteinklassen ausnutzen lassen, gibt es auch hier keine Anzeichen für eine mögliche Trennung von *cis*- und *trans*- Sequenzen unter Verwendung dieses Datensatzes.

Von allen getesteten Methoden sind die adaptiv kodierten Neuronalen Netze am geeignetsten. Legt man die Erfahrungen an, die bei den Schnittstellensequenzen gewonnen wurden, so ist der Datensatz noch viel zu klein. Der Unterschied zwischen Training- und Testergebnis ist so groß, daß mindestens die zehn bis zwanzigfache Menge an Daten vorhanden sein müßte. Berücksichtigt man, daß es in einem 20iger Sequenzfenster 10^{26} mögliche Sequenzen gibt, von denen nur etwa der 10^{-23}ste Teil vorlag, ist sogar ein Korrelationskoeffizient von 0.2 für die Testdaten bemerkenswert. Jedoch sind mit diesem Datensatz keine zuverlässigen Vorhersagen möglich.

Eine Informationsanalyse von Sequenzen mit unterschiedlichem Oberflächenanteil des Prolins ergab keine Unterschiede zwischen den einzelnen Konformeren. *Cis*- und *trans*-Peptidylprolylgruppen kommen sowohl an der Proteinoberfläche als auch im Core von Proteinen vor. Es ist noch nicht geklärt, ob die Isomerasen vor allem an der Oberfläche von Proteinen befindliche Prolinreste isomerisieren. Für unzugängliche Proteine müßte jedoch die Faltung in einem sehr frühen Stadium aufgehalten werden, bis sich die *cis*-Konformation eingestellt hat. So wurde auch schon eine Chaperonfunktion für die Isomerasen festgestellt [2, 25].

Die Verteilung der Aminosäuren in der dreidimensionalen Nachbarschaft vom Prolin zeigt keine besonderen Unterschiede zwischen den beiden Klassen. Direkte Wechselwirkungen mit dem Prolin sind somit auch nicht für eine bestimmte Konformation verantwortlich.

Eine Analyse der PPIasen kann weitere Aufschlüsse über die Kodierung ergeben. Dazu wird zuerst die in der Literatur vorhandene Information über die PPIasen genauer untersucht.

Die PPIasen wurden zuerst von Fischer und Mitarbeitern 1984 [21] beschrieben. Erst später hat man festgestellt, daß Cyclophilin (CyP) und die von Fischer entdeckte PPIase dasselbe Molekül sind [22, 119].

Cyclophilin gehört zu den Immunsuppressiva-bindenden Substanzen [5]. Da alle PPIasen auch Immunsuppressiva binden, hat man die PPIasen entsprechend ihrer Eigenschaften gegenüber diesen Immunsuppressiva eingeteilt.

Die starken und klinisch wirksamen Immunsuppressiva Cyclosporin A (CsA), FK 506 und Rapamycin werden mit hoher Affinität von den PPIasen, die zu den Immunophilinen gehören, gebunden. Die unterschiedlichen Bindungseigenschaften erlauben eine Klassifizierung der PPIasen in die folgenden zwei Superfamilien: die Cyclophiline (CsA bindende Proteine) und die FKBP (FK 506/Rapamycin-bindende Proteine). Über eine kürzlich entdeckte dritte Familie ist zur Zeit noch sehr wenig bekannt [91, 101].

Immunsuppressiva unterdrücken die Immunantwort. Dazu gehören Azatiprine, die die Zellteilung verhindern; Steroide, die in den Reifungsprozeß eingreifen sowie Cyclosporin A, FK 506 und Rapamycin, die durch selektive Wirkung auf die Lymphozyten jeweils die Immunzellaktivierung blockieren [26]. Immunsuppressiva sind notwendig, um Autoimmunkrankheiten zu behandeln oder transplantierte Organe vor einer Abstoßung zu bewahren.

Die immunsuppressiven Wirkstoffe CsA und FK 506 sind ihrerseits in der Lage, das Zellwachstum von HIV- und Malaria infizierten Zellen zu verlangsamen [5, 49, 72].

Diverse posttranslationale Modifikationen wie die Glycosylierung, N-terminale Modifikationen und Phosphorylierung sind zusätzliche Funktionen der PPIasen. Auch Faltung, Zusammenlagerung und Transport von Proteinen werden durch die PPIasen reguliert. Diese Proteine haben zudem die Fähigkeit als Chaperone zu wirken [51]. Einige PPIasen sind koregulierende Untereinheiten molekularer Komplexe, etwa Hitzeschockproteine, Glucocorticoidrezeptoren und Ionenkanäle.

Über Wirkungsweise und Spezifizität der PPIasen ist noch nicht genügend bekannt, um schlüssige Vorhersagen treffen zu können, wann die *trans*-Prolinbindung in ihr Konformer überführt wird. Fischer et al. haben an Ribonuclease T1 zeigen können, daß von den insgesamt vier Prolinresten drei potentielle Substrate für die PPIase sind. Es sind aber nur zwei *cis*-Proline in der nativen Form vorhanden [64, 71, 78]. Dieses Beispiel zeigt, daß eine „overprediction“ bei den Vorhersagen nicht notwendigerweise schlecht sein muß.

Eine Aussage über die für eine Isomerisierung nötigen Substrateigenschaften können Enzym-Substrat-Komplexe liefern. Strukturell aufgeklärte Enzym-Substrat-Komplexe liegen schon vor, jedoch nur von Enzymen mit kleinen Peptiden. Kallen und Walkinshaw [47] haben die Struktur von humanem Cyclophilin A mit dem Tetrapeptid ac-Ala-Ala-Pro-amc (ac, Acetyl;

amc, Amidomethylcumarin) über die Röntgenstrukturanalyse aufklären können. Die Elementarzelle, die aus einem Dimer von Komplexen und 135 Wassermolekülen besteht, wurde bis zu einem kristallographischen R-Faktor von 17,7% für alle Daten in einem Bereich von 8 Å - 2,3 Å bestimmt. Das Tetrapeptid liegt in *cis*-Konformation vor. Aminosäurereste des Cyclophilin, die ein nicht-Wasserstoffatom in einem Bereich von 3,8 Å in der Nähe eines nicht-Wasserstoffatoms Tetrapeptids besitzen, wurden als aktives Zentrum bestimmt und sind: Arg-55, Ile-57, Phe-60, Gln-63, Ala-101, Asn-102, Gln-111, Phe 113, Leu-122, His-126 und Arg-148. Diese Aminosäuren bilden einen Kanal, der auf zwei antiparallelen β-Strängen sitzt (mit Kontakt von Phe-60, Gln-63, Gln111 und Phe-113). Loops, die aus der Oberfläche des Rumpfes herausragen, erzeugen eine spezifische Furche in der Proteinoberfläche. Ein Loop, bestehend aus den Resten 101 bis 110, beinhaltet die Reste Ala-101 und Asn-102 des aktiven Zentrums. Ein schmaler Durchgang trennt diesen Loop von einem zweiten (Reste 69 - 74). Eine weitere wichtige topologische Eigenschaft ist eine Wand auf einer Seite der Bindungsstelle, die von den Resten 118 - 126 gebildet wird und eine dichte helikale Form aufweist. Hier haben Leu-122 und His-126 Kontakt mit dem Peptidsubstrat. Das *cis*-Prolin sitzt in einer recht tiefen Tasche. Diese Tasche besteht aus den Aminosäuren Phe-60, Met-61, Phe 113 und Leu 122.

Die Guanidingruppe von Arg-55 bildet eine Wasserstoffbrücke zum Carbonyl-Sauerstoff des Pro-3 im Peptid (Abstand Arg-55 (NH1) - Pro-3(O): 2.8 Å und 3.0 Å für die beiden Monomeren). Es gibt weiterhin eine H-Brücke zum Gln-63, welche selbst wieder zu Gln-111 über eine H-Brücke verbunden ist. Der Rest von His-126 ist dicht benachbart zu der Alanyl-Prolyl-*cis*-imid-Bindung. Ke und Mitarbeiter haben mit dem Dipeptid Ala-Pro eine ähnliche Struktur entwickelt und beobachten auch hier eine Isomerisierung [50].

Die daraus abgeleiteten Rückschlüsse auf einen Mechanismus können nicht im Hinblick auf die Substratspezifizität ausgewertet werden, da noch zu wenig Information vorliegt.

Substitutionen im aktiven Zentrum lassen Rückschlüsse auf den Reaktionsmechanismus und Substrateigenschaften zu. Eine Substitution von Tyr-82 zu Leu im aktiven Zentrum von FKBP12 bewirkt eine starke Abschwächung der PPIase-Aktivität gegenüber allen getesteten Tetrapeptiden [9]. Die getesteten Sequenzen bestehen aus stark unpolaren Aminosäuren mit zwei Ausnahmen. Diese Ausnahmen, Ser-Leu-Pro-Phe und Suc-Arg-Leu-Pro-Phe-NA, weisen aber auch die geringste Aktivität gegenüber FKBP-12 auf. Die katalytischen Eigenschaften werden in diesen beiden Fällen nicht so sehr von der Mutation beeinflußt wie die vollständig hydrophoben Sequenzen. Es zeichnet sich hier also ab, daß FKBP eher Substrate katalysiert,

die an der Position -2 hydrophobe Reste besitzen. Die hier vorliegenden Ergebnisse zeigen keine Übereinstimmung mit diesen experimentellen Befunden zu FKBP.

Auch andere Mutationsuntersuchungen lassen vermuten, daß die hydrophobe Umgebung des aktiven Zentrums eine zentrale Rolle bei der Isomerisierung spielt [86]. Über die Substratspezifizität kann hier aber keine Angabe gemacht werden.

Die aus den Röntgenstrukturen und thermodynamischen Größen abgeleiteten Reaktionsmechanismen waren alle nicht eindeutig, weshalb sie in dieser Arbeit nicht weiter behandelt werden. Gute Übersichten zu diesem Thema finden sich in [50, 115].

Es hat sich gezeigt, daß die Reaktion in einem pH-Bereich von 5,5 bis 9 pH-unabhängig ist [29]. Außerdem zeigen Untersuchungen zur steady-state-Kinetik, daß es keine Unterschiede zwischen H_2O und D_2O als Solvens gibt [55], woraus man folgern kann, daß keine kovalenten Bindungen zu Wassermolekülen im geschwindigkeitsbestimmenden Schritt vorkommen. Auf eventuell auftretende lokale Änderungen des pH-Wertes kann hieraus nicht geschlossen werden.

In vivo ist bis auf die durch PPIasen katalysierte Isomerisierung keine andere Möglichkeit bekannt, wie eine *trans*-Peptidgruppe isomerisiert werden kann. Im Gegensatz dazu ist es *in vitro* möglich, durch extreme Bedingungen, wie zum Beispiel hohe Ammoniumsulfatkonzentrationen, die nicht katalysierte Isomerisierung zu beschleunigen [64]. Eine enzymatische Isomerisierung legt die Überlegung der lokalen Kodierung nahe. Die am Anfang gemachten Annahmen sollen nun in Bezug auf die Ergebnisse dieser Arbeit und anderer Arbeiten untersucht werden.

Die Statistik zeigt, daß die verwendeten Daten mit den Literaturwerten vergleichbar sind. Annahme 5 trifft also zu.

Da mit festen Fenstergrößen gearbeitet wurde, darf die Größe der Proteine ebenfalls nicht entscheidend für eine durch PPIasen katalysierte Isomerisierung sein. Es wird jedoch vermutet, daß kleinere Proteine besser mit PPIasen reagieren als große [66, 67]. Die Experimente zeigen, daß die Isomerisierungsgeschwindigkeit von der Größe abhängen kann, die Tatsache, daß isomerisiert werden kann, ist davon aber nicht unbedingt betroffen.

In einer NMR-Studie von 1993 wurde die nicht katalysierte *cis*-*trans*-Isomerisierung von Trp-$(Pro)_n$-Typ-OH (n = 1, ..., 5) untersucht [88]. Es zeigte sich hier, daß der Anteil an *cis*-Konformer mit zunehmender Länge des Peptids abnimmt. Der Anteil an *cis*- Konformer verringert sich von 0,62 für Trp-Pro-Tyr auf 0,42 für Trp-(Pro)5-Tyr.

In den vorliegenden Daten tritt an Positionen 349 - 352 von Alkohol Dehydrogenase (8ADH) die Sequenzfolge Val-Leu-Pro-Phe auf, mit einem ω-Winkel der Leu-Pro-Bindung von 177°. Diese Sequenz reagiert bei den Arbeiten von Bossard und Mitarbeitern mäßig mit FKBP und zeigt auch moderate Unterschiede in der Katalyse zwischen den Mutanten [9].

Diese Ergebnisse zeigen, daß sehr wohl eine Abhängigkeit der Isomerisierungsgeschwindigkeit von der Größe des Proteins besteht. Bei der Bestimmung der Konformation aus der Sequenz sollte die Stellung des Prolylrestes im Protein mit in die Überlegungen einbezogen werden. Prolylgruppen, die in schnell faltenden Regionen, wie α-Helices, zu finden sind, können aus zeitlichen Gründen eventuell nicht isomerisiert werden. Außerdem folgt daraus, daß größere Peptide nicht ohneweiteres mit kleinen Vierpeptiden vergleichbar sind.

Faltungsexperimente von Ribonuklease T1 (RNAse T1) zeigen, daß die schnelle Faltung von Proteinen die Isomerisierung verhindern kann [64]. Bei einer Klassifizierung kann man also eine „overprediction“ erwarten.

Soll die Information eindeutig sein, muß angenommen werden, daß die Spezifizität der PPIasen gleich ist, da nichts darüber bekannt ist, wie die einzelnen Imidbindungen isomerisiert werden. Es gibt sowohl Hinweise dafür, daß die Substratspezifizität ähnlich ist, als auch Belege für den individuellen Charakter der PPIasen.

Da die Bindungsstellen für Immunsuppressiva und Prolyl-Isomerisation unterschiedlich sind [114], ist es möglich daß die Substratspezifizität bezüglich der Isomerisierung zwischen den einzelnen Familien gleich, beziehungsweise ähnlich ist.

Untersuchungen der aktiven Zentren von FKBP und Cyp weisen große Ähnlichkeiten auf [18]. So lassen sich in den aktiven Zentren von FKBP und Cyclosporin folgende Aminosäuren überlagern (Cyp-FKBP): Trp121-Trp59, Leu122-Ile56, His126-Tyr82, Met100-Ile91, Ala101-Ile90, Phe113-Phe36, Arg55-Arg42, Phe60-Tyr26, His92-Phe99.

Experimente mit verschiedenen Substraten zeigen dagegen, daß die Reaktionen der einzelnen Isomerasen unterschiedlich katalysiert werden. Die Tetrapeptide Suc-Ala-Xaa-Pro-Phe-pNA (Suc, N-Carboxypropionyl; Xaa = Gly, Ala, Val, Leu, Phe, His, Lys, Glu; pNA, p-Nitroanilide) wurden mit Cyclophilin und FKBP umgesetzt, und es zeigte sich, daß FKBP selektiver katalysiert als Cyclophilin [29]. Es gibt sogar Unterschiede in der Reaktivität innerhalb der einzelnen Familien [32]. In Hefe kommen mindestens vier verschiedene Formen von Cyclophilin vor, die unterschiedliche hydrophobe und elektrostatische Eigenschaften haben. Auch die Substratspezifizität scheint unterschiedlich zu sein.

Wieviele Familien es gibt, ist zur Zeit auch unklar. Über eine kürzlich entdeckte neue Familie der PPIasen [91, 101], die auf Escherichia coli PpiC Protein zurückzuführen ist, liegen noch keine genauen Angaben vor. Trotzdem ist es nicht auszuschließen, daß bestimmte Eigenschaften und Aminosäuren in der näheren Umgebung von Prolin vorhanden sein müssen, um eine Isomerisierung überhaupt zu ermöglichen. So konnte gezeigt werden, daß das Substrat Gly-Pro-Ala etwa 16 mal schneller isomerisiert wird als Gly-Pro [67]. Bei der Faltung von RNase T1, bei der die Isomerisierungen von Pro-39 und Pro-55 geschwindigkeitsbestimmend sind, wird eine mögliche Beteiligung von Nachbargruppen der Prolinreste als Auswahlkriterium angenommen [52]. In RNase T1 kommen auch zwei *trans*-Prolinreste vor: Pro-60 und Pro-73.

Die Annahme der lokalen Kodierung ist die eigentliche Grundannahme. Diese Annahme basiert auf dem Wissen über die enzymatische Katalyse.

Es wird davon ausgegangen, daß alle *cis*-Proline durch PPIasen isomerisiert wurden. Andere *in vivo* Prozesse zur Isomerisierung sind nicht bekannt, jedoch auch nicht auszuschließen, da zum Beispiel hohe Ammoniumsulfatkonzentrationen die nicht katalysierte Isomerisierung beschleunigen können [64]. Des weiteren beschreiben Fox und Mitarbeiter zwei native Formen von Staphylococcal Nuclease, die beide in Lösung existieren und sich ineinander umwandeln [23].

Verschiedene Untersuchungen an kleineren Peptiden zeigen, daß lokale Unterschiede die enzymatische Isomerisierung beeinflussen können [29, 30, 67]. Anhand einer Gly6-Bradykininmutation konnte gezeigt werden, daß die durch PPIasen katalysierte *cis-trans*-Isomerisierung stark von der Sequenz abhängt [70]. Hier wurde die Aminosäure Serin des Nona-Peptids Bradykinin durch Glycin ausgetauscht und man beobachtete einen starken Rückgang der Isomerisierungsgeschwindigkeit. Dieser Sequenzunterschied deckt sich mit den Ergebnissen aus den statistischen Untersuchungen.

Die hohe Konservierung von Aminosäuren sowohl in Cyclophilin als auch in FKBP steht auch in Einklang mit der lokalen Kodierung [120].

Ziel dieser Analyse war die Vorhersage der Konformation von Peptidylprolylbindungen in Proteinen. Diese Situtation unterscheidet sich von der Vorhersage der Konformation von Peptiden. Das Protein wird immer diejenige Struktur annehmen, die in einem Energieminnimum liegt, das durch Konformationsänderungen aus der gestreckten Form erreichbar ist. In Proteinen wird die Struktur hauptsächlich durch nicht lokale Wechselwirkungen bedingt. Das ist einsichtig, da eine kompakte Struktur für die meisten Moleküle am günstigsten ist und hierbei die nicht lokalen Wechselwirkungen entscheidend sind. In seltenen Fällen wird durch eine *cis*

Konformation ein niedrigerer Energiezustand erreicht. In Peptiden ist der *cis* Zustand immer ungünstiger. Es gibt zwar Konstellationen in denen das Energieniveau des *cis* Zustandes etwas niedriger liegt als bei anderen, jedoch ist das Gleichgewicht immer auf der *trans*-Seite [93]. Die Rotationsbarriere hat auf die Struktur von Proteinen nur einen sehr geringen Einfluß. Nur wenn ein Faltungsweg in eine *cis* Konformation durch das schnelle Falten anderer Bereiche nicht mehr zugänglich wird, hat die Rotationsbarriere einen Einfluß auf die Faltung. Im allgemeinen jedoch wird ein lokales Minimum angesteuert, aus dem sich dann relativ langsam, entsprechend der Rotationsbarriere, die *cis* Konformation einstellt. Für den Fall der stabilen *cis* Bindung ist eine Rückreaktion sehr unwahrscheinlich. Im Falle von Proteinen mit nicht lokalen Wechselwirkungen ist der native Zustand mit dem *cis* Zustand verbunden. D.h. der *cis*-Zustand ist in Proteinen mit einer *cis* Bindung niedriger als der *trans* Zustand. Die Rückreaktion wird unwahrscheinlicher, da die Rotationsbarriere aus dem *cis* Zustand in den *trans* Zustand höher ist. Da, wie schon bemerkt, der *cis* Zustand im Peptid ungünstiger ist, muß der Energiegewinn, der durch das Einstellen der *cis* Konformation in Proteinen geschieht, die Energie kompensieren. Um den größtmöglichen Energiegewinn zu bekommen, sollte also die Aminosäure mit dem niedrigsten Energiezustand für die *cis* Konformation gewählt werden. Das ist in der Regel auch der Fall, da in den meisten Fällen ein Prolin beteiligt ist. Die relative Verteilung der *cis* Peptidylprolinbindungen in Proteinen stimmt mit der relativen Verteilung in Peptiden überein [93].

Die vorliegende Untersuchung zeigt, daß eine Konformationsvorhersage aus der lokalen Sequenz heraus nicht möglich ist. Es gibt zwar Präferenzen für bestimmte Peptidylprolylbindungen, jedoch lassen sich diese nicht für eine Vorhersage nutzen. Interessanterweise lassen sich aber die MHC-I Moleküle über die *cis*-Peptidylprolylbindung identifizieren.

5.c Ligandendesign

Bindungsstellen von Proteinen besitzen hoch selektive Erkennungsregionen für die Liganden. Das Ziel des rationalen Designs von Liganden wird auch als Docking-Problem bezeichnet. Die meisten Docking Methoden berechnen in einem komplexen und langwierigen Verfahren die Bindungsenergien zwischen Rezeptor und Ligand. Dazu muß ein Modell des Rezeptors bekannt sein und die aktive Stelle bestimmt werden. Die Konformation des Liganden ist ebenso unbekannt wie dessen Zusammensetzung [116]. Trotz dieser vielen Probleme gibt es Programme, die bei der Entwicklung von Liganden eine große Hilfe sind [7].

5.c.i Simulierte Molekulare Evolution

Ein Ansatz zur *de-novo* Entwicklung von Liganden, der ohne größeren Rechenaufwand auskommt und keine Struktur des Rezeptors benötigt, ist die Simulierte Molekulare Evolution [104, 108]. Sie verwendet die Erfahrungen, die aus bekannten Peptidinhibitoren gesammelt worden sind und hat sich bereits beim Design von Schnittstellen für die Signalpeptidase bewährt [110, 111]. Dieses Verfahren ist auch für das Design von Inhibitorproteinen des HRV 14 geeignet, wie in dieser Arbeit beschrieben. Die Inhibition wurde aufgrund von Eigenschaften nur weniger Peptide mit bekannter zellprotektiver Eigenschaft errechnet. Die Inhibition der besten designten Peptide liegt im Bereich der Ausgangsequenzen. Für diese wenigen Beispiele sind das beachtliche Zellprotektionen, zumal die Sequenzen zum Teil sehr wenig Ähnlichkeit zu den Originalsequenzen aufweisen. Dieses Ergebnis muß als erster Schritt in einem Designzyklus verstanden werden [104]. In diesem Designzyklus wird ein Modell aufgrund von Peptiden mit bekannter Funktion entwickelt. Bei dem Modell handelt es sich um ein Künstliches Neuronales Netz. Mit diesem KNN werden mittels der SME neue Sequenzen erzeugt, die anschließend in biologischen Tests auf ihre Funktion hin untersucht werden. Die Sequenzen werden im nächsten Zyklus für ein verbessertes Modell verwendet. So können die Peptide von Generation zu Generation optimiert werden. Der Sequenzraum wird mittels der Evolutionsstrategie gezielt durchsucht. Das KNN ist dabei die Fittnessfunktion.

Die vorliegenden Sequenzen wiesen keine besonders guten zellprotektiven Eigenschaften auf. Das Modell, das aufgrund dieser Beispiele entwickelt werden kann, wird nur in der lokalen Sequenzumgebung sinnvolle Ergebnisse liefern. Da die meisten Beispiele im Bereich von -0,2 der zellprotektiven Wirkung lagen, ist es bemerkenswert, daß mit dieser Methode Sequenzen mit einer hohen zellprotektiven Wirkung vorgeschlagen wurden (-45%, Tabelle 22). Es ist zu vermuten, daß in weiteren Schritten des Designzyklus wesentlich bessere Peptide entstehen.

5.c.ii Peptid-Docking

Für bekannte Strukturen ist es wichtig, neue Peptide zu entwickeln, die eine biologische Wirkung besitzen, das heißt eine feste Bindung mit der bekannten Struktur eingehen. Mit Hilfe der Simulierten Molekularen Evolution ist es möglich, neue funktionale Sequenzen zu generieren. Dabei werden die Ergebnisse, die von den künstlichen Neuronalen Netzen erhalten werden, verwendet. Dabei wird die Strukturinformation des Rezeptors nicht verwendet. Der hier vorgestellte Algorithmus zur Entwicklung neuer Ligandpeptide verwendet hingegen die Strukturinformation.

Die gängigen Verfahren des Docking sind sehr zeit- und kostenintensiv und beruhen meistens auf Kraftfeldberechnungen. In dem hier vorgestellten Ansatz hingegen wird die Information von bekannten 3D-Wechselwirkungen zwischen Proteinen ausgenutzt. Der Algorithmus, der in dieser Arbeit verwendet wurde, ist zum Test noch per Hand durchgeführt worden und ist noch nicht implementiert.

Er gliedert sich in vier Abschnitte:

1) Die Identifizierung der Bindungsregion am Rezeptor

Hierfür muß die aktive Stelle automatisch gefunden werden. Die hierzu bereits entwickelten Programme müssen so modifiziert werden, daß ein Raster der möglichen Bindungsregion ausgegeben wird.

2) Bestimmung der Eigenschaften des Rasters.

Zu diesem Raster, müssen anschließend die auf dem Protein/Rezeptor benachbarten Aminosäuren bestimmt werden.

3) Definition eines Startpunktes im Zentrum der Aktiven Region.

Nicht jeder Punkt ist gleich gut als Startpunkt für ein Inhibitorpeptid geeignet. Es könnte z.B. im Zentrum der aktiven Region begonnen werden, um dann sukzessive am C- und am N-Terminus weitere Aminosäuren anzuhängen, bis die gewünschte Länge erreicht ist.

4) Beurteilung und Auswahl der verschiedenen Möglichkeiten für die Lage der Aminosäuren.

Es muß derjenige Punkt auf dem Gitter ausgewählt werden, der die nächste Aminosäure repräsentiert. Dabei müssen die Kontaktfläche zwischen Peptid und Rezeptor, benachbarte Ladungen sowie Leerräume berücksichtigt werden.

Diese Berechnung der Energiefunktion wird beim manuellen Docking nicht so genau sein, jedoch ist in dieser Arbeit gezeigt worden, daß mit dieser Methode funktionale Peptide entwikkelt werden können. Eine Implementierung und Optimierung des vorgeschlagenen Algorithmus scheint nach dem bisherigen Erkenntnisstand lohnenswert.