# KAPITEL 2 Daten

Daten sind die wichtigste Grundlage für die erfolgreiche Untersuchung von Sequenzinformationen. Die Daten müssen die wichtigsten Eigenschaften für das gegebene Problem enthalten. Im Falle von Proteinsequenzdaten wird davon ausgegangen, daß die wesentliche Information für die Faltung, also die dreidimensionale Struktur, in der Reihenfolge der 20 Aminosäuren enthalten ist. Die dreidimensionalen Struktureigenschaften von Proteinen bestimmen die Funktion und Wirkungsweise der Proteine. Es ist nicht möglich, alle theoretisch möglichen Sequenzen zu untersuchen. Denn schon für ein Protein von 50 Aminosäuren gibt es $20^{50} \cong 10^{65}$ unterschiedliche Sequenzen. Daraus ergibt sich, daß nur ein kleiner Teil der möglichen Sequenzen untersucht werden kann. Dieser kleine Teil muß repräsentativ sein, das heißt, die wichtigen Informationen müssen zu etwa gleichen Teilen enthalten sein. Ist in einem gegebenen Datensatz ein Teil der Information überrepräsentiert, weil z.B. ein Analysesystem einen Bereich besser darstellt als einen anderen, so kann es sein, daß wichtige Aspekte des Problems nicht genügend berücksichtigt werden. Die Protein-Datenbank (PDB, [6]) enthält Kristall und NMR-Strukturen von Proteinen. Ein Modell, welches auf einem solchen Datensatz aufbaut, ist möglicherweise nicht in der Lage, die weniger gewichteten Eigenschaften korrekt zuzuordnen. Gibt es für ein gegebenes Problem eine Konsensussequenz, die die wesentliche Information beinhaltet, braucht man nur diese eine Sequenz, um ein adäquates Modell zu entwickeln. Leider ist diese Situtation nur dann sicher gegeben, wenn das Problem bereits verstanden ist.

Daraus folgt, daß nicht die Menge der Daten entscheidend ist, sondern einmal die Richtigkeit und zum anderen ihre Repräsentanz. Korrekte Daten sollten durch biochemische Experimente erhalten worden sein. Daten, die durch Anwendung von Modellen gewonnen wurden, geben nur die Informationen des Modells wieder.

Ein repräsentativer Datensatz besteht aus nicht-homologen Daten. Als Maß für eine Nicht-Homologie hat sich eine maximal 20%ige Sequenzidentität herausgestellt [19]. Zwei zu vergleichende Sequenzen sind dann nicht-homolog, wenn sie in einem bestimmten Sequenzfenster nicht mehr als 20% der Aminosäuren an gleichen Positionen besitzen.

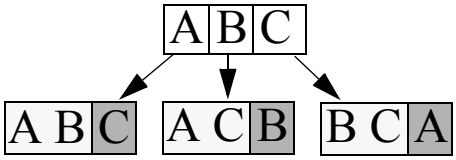


ABBILDUNG 6: Kreuzvalidierung: Der Datensatz wird in die Gruppen A, B und C aufgeteilt. Jeweils zwei Gruppen bilden die Daten zum Training (hell unterlegt); die verbleibende Gruppe dient als Test (dunkel unterlegt).

Wurde ein korrekter, möglichst repräsentativer Datensatz gefunden, sollten mit bestimmten Verfahren die gemeinsamen Merkmale identifiziert werden können. Die Methode sollte die Fähigkeit besitzen, neue, noch nicht bekannte Sequenzen, korrekt vorherzusagen. Diese Generalisierungsfähigkeit kann bestimmt werden, wenn nur ein Teil des Datensatzes zur Entwicklung (Training) der Methode benutzt wird. Mit den restlichen Daten wird die Methode auf ihre Generalisierungsfähigkeit hin überprüft (Test). Wenn der Datensatz in mehrere Teile aufgeteilt wird und diese in einem iterativen Prozeß zum Training und Test angewendet werden, wird das Kreuzvalidierung (Jack-Knife-Test) genannt. In der vorliegenden Arbeit wurde der Datensatz in drei Teile aufgeteilt und jeweils zwei Teile zum Training und ein Teil zum Test verwendet (Abbildung 6).

## 2.a Sequenzen der Schnittstellenregion humaner sekretorischer Proteine

Im Falle der Schnittstellensequenzen humaner sekretorischer Proteine ist es nicht üblich, die Schnittstellen experimentell zu bestimmen. Es werden vielmehr bereits entwickelte Modelle zur Vorhersage der Schnittstellen angewendet, um für eine neue Sequenz die passende Schnittstelle zu definieren. Es gibt in der SWISSPROT Datenbank (Version 34) [4] 4000 humane Proteine. 1141 davon besitzen eine gekennzeichnete Signalsequenz. 479 haben eine ausdrücklich putative Schnittstelle. Eine Literaturrecherche ergab, daß mindestens 76 von den 662 nicht-putativen Schnittstellen experimentell bestimmt wurden. Für die Bestätigung mußte sowohl die cDNA vorliegen als auch die N-terminale Sequenz des maturen Proteins durch Sequenzierung bestätigt worden sein. Daraus läßt sich direkt die Position der Schnittstelle ableiten. Im Anhang, Tabelle 29 auf Seite 127ff, sind die SWISSPROT-Einträge, die Schnittstellen und die Literaturstellen, mit denen die Bestätigung erfolgte, aufgelistet.

Drei mögliche Fehlerquellen für die Position der Schnittstellen bleiben trotz sorgfältiger Literaturrecherche:

1. Das Expressionssystem kann ein nicht-homologes System sein, bei dem sich die Peptidaseeigenschaften vom Menschen unterscheiden.

2. Präpropeptide können fälschlicherweise als Präpeptide identifiziert werden, da meist nur die N-terminale Sequenz des maturen Proteins bekannt ist, und weitere Modifikationen nach dem Schneiden durch die Signalpeptidase werden dann als Schnittstellen identifiziert. Einige der Präpropeptide sind bekannt, es gibt jedoch immer noch die Möglichkeit, daß einige nicht erkannt wurden.

3. Fehler in den Experimenten sind ebenfalls nicht auszuschließen.

Alle Fehlerquellen können einen nicht erwünschten Einfluß auf den Datensatz ausüben, der das darauf aufbauende Modell beeinflußt. Es ist somit zu erwarten, daß selbst mit einem optimalen Filtersystem nicht alle Sequenzen richtig erkannt werden.

Ein Sequenzfenster von 12 Aminosäuren wurde zur Generierung der Schnittstellendaten (positive Daten) verwendet. Es umfaßt, ausgehend von der Schnittstelle, die nächsten 10 Aminosäuren in Richtung des N-Terminus und die nächsten zwei Aminosäuren des maturen Proteins. Die folgenden 10 Sequenzfenster des maturen Proteins dienten als Nicht-Schnittstellendaten (negative Daten) (Abbildung 7).

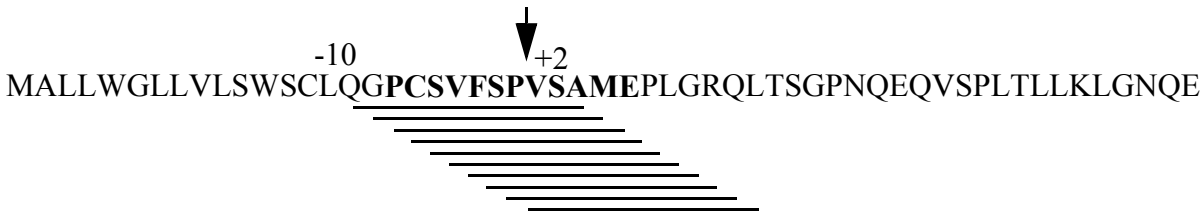


ABBILDUNG 7: Generierung der positiven und negativen Sequenzdaten. Die untersuchte Schnittstellenregion (positive Sequenz) ist fett dargestellt. Die nicht Schnittstellensequenzen sind mit einer Linie unterstrichen (negative Daten). Die Schnittstelle ist mit einem Pfeil markiert. Bei dem Beispielprotein handelt es sich um α-2-Anti-Plasmin Precursor (A2AP-HUMAN).

Die Datensätze wurden in jeweils drei Gruppen für die Kreuzvalidierung aufgeteilt (vergleiche Abbildung 6), aus denen die Datensätze für das Training und für den Test gebildet wurden. Die Gruppen sind in Tabelle 30 und Tabelle 31 im Anhang aufgelistet.

TABELLE 1. Anzahl der sekretorischen humanen Sequenzbeispiele

| Daten | Anzahl der Sequenzen |
|---|---|
| Schnittstellensequenzen | 76 |
| Nicht-Schnittstellensequenzen | 760 |

## 2.b Peptidylprolylsequenzen

Die Protein-Data-Bank (PDB) [6] beinhaltet alle veröffentlichten 3D-Strukturdaten von Proteinen und Nukleinsäuren. Es gibt zur Zeit 7.408 (Datum: 20. 7. 1998) solcher Datensätze. Viele Proteine liegen mit verschiedenen Auflösungen der Kristallstruktur oder mit leicht unterschiedlicher Sequenzzusammensetzung vor oder sind mit unterschiedlichen Liganden gebunden. Dadurch umfaßt diese Datenbank einen großen Teil redundanter Daten. Hobohm et al. haben aus dieser Datenbank den SELECT-Datensatz ausgewählt (Mai 1996)[38], der nach einer Homologieanalyse nur noch solche Sequenzen enthält, die in einem 80er Fenster nicht mehr als 25% Sequenzhomologie besitzen. Dieser Datensatz besteht aus 505 Sequenzen mit insgesamt 131.463 Aminosäuren.

Die Sequenzen der auswertbaren Daten dieses Datensatzes wurden mit der zugehörigen Konformation in einer neuen Datenbank zusammengefaßt. Auswertbar sind nur solche Daten, deren Koordinaten eine eindeutige Zuordnung der ω-Winkel zulassen. Aussortiert wurden solche Datensätze, bei denen nur Cα-Atome angegeben sind. Fehlt an einer Stelle eine Aminosäure, so wurde das Protein an dieser Stelle in zwei Teile geteilt. In dem neuen Datensatz befinden sich unter 126.123 Aminosäuren 5.987 Prolinreste. Diese teilen sich in einem 20er Fenster (±10 Aminosäuren vom zentralen Prolin) in 5.048 *trans*- und 327 *cis*- Konformere auf, wobei sich *cis* immer auf $\omega = (0° \pm 160.0°)$ bezieht. (Die restlichen 612 Prolinreste befinden sich in Sequenzabschnitten, die kein 20er Fenster ausfüllen können.) Damit liegen etwa 5,4% der Xaa-Pro Sequenzen in der *cis* Konformation vor. Die Daten wurden wieder in drei Gruppen für die Kreuzvalidierung aufgeteilt (Abbildung 6).

## 2.c Liganden-Docking

Für das Design antiviraler Peptide gegen das Rhinovirus HRV 14 wurden zwei verschiedene Methoden verwendet. Dazu wurden jeweils unterschiedliche Datensätze verwendet. Zum einen wurde eine Simulierte Molekulare Evolution (SME) [104] mit 68 experimentell getesteten Nona-Peptiden durchgeführt. Diese Sequenzen sind aus patentrechtlichen Gründen noch nicht zur Veröffentlichung freigegeben. Es lagen zu den Sequenzdaten auch die zellprotektiven Eigenschaften vor.

Die zweite Methode zur Vorhersage von bindenden Peptiden bei vorgegebener 3D-Struktur beruhte auf der Analyse von Enzym-Inhibitor-Komplexen der PDB. Dazu wurde in Zusammenarbeit mit Herrn Dr. Grunert (Institut für Infektionsmedizin, Universitätsklinikum Benjamin Franklin, Freie Universität Berlin) die PDB Datenbank auf verschiedene Enzym-Inhibitor-Komplexe durchsucht. Eine Liste der verwendeten kristallisierten Komplexe ist im Anhang aufgeführt. Die 3D-Struktur des HRV 14 befindet sich in der PDB unter 1HRV.