Universität zu Köln
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Biochemie
Lehrstuhl Prof. Dr. D. Schomburg

Vorhersage der SCOP-Klassifizierung aus Alignmentdaten unter Verwendung neuronaler Netzwerke

Diplomarbeit

vorgelegt von
Thorsten Fallisch
aus Düsseldorf

im Oktober 2001

1. Referent: Prof. Dr. D. Schomburg
2. Referent: Dr. S. Waffenschmidt

Die vorliegende Diplomarbeit wurde angefertigt in der Zeit vom
11. Januar 2001 bis zum 9. Oktober 2001
am Institut für Biochemie der Universität zu Köln,
Lehrstuhl Prof. Dr. D. Schomburg

Ich versichere, dass ich diese Diplomarbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Köln, den 9. Oktober 2001

Meinen Eltern und meinem Bruder gewidmet

Herrn Prof. Dr. D. Schomburg möchte ich für die Möglichkeit diese Arbeit durchzuführen, eine interessante Themenstellung, anregende Diskussionen und Hilfestellungen danken.

Bei allen Mitgliedern der Arbeitsgruppe, Freunden und sonstigen Personen, die mich durch Ideen, Ratschläge und Hilfe (sowohl fachlicher als auch moralischer Art) unterstützt haben, möchte ich mich bedanken.
Besonderen Dank möchte ich hier Prof. Dr. Arne Elofsson vom *Stockholm Bioinformatics Center* aussprechen.

Besonderen Dank auch an meine Eltern, ohne deren Unterstützung mir diese Arbeit so nicht möglich gewesen wäre.

Inhaltsverzeichnis

Begriffserklärungen und Abkürzungen

Begriffserklärungen

Alignment, alignieren	„Zur Deckung bringen“ zweier Sequenzen oder Strukturen, auch als Abgleich, abgleichen oder Vergleich, vergleichen bezeichnet
Ångstrøm	$1\,\text{Å} = 10^{-10}\,\text{m}$
Backbone	Strukturgerüst eines Proteins, bestehend aus den C_α-, C-, N- und O-Atomen der Aminosäuren ohne die Seitenketten
Cutoff	Scheidepunkt zur Aufteilung von Daten, alle Werte unter dem C. werden als ein Wert a interpretiert, alle anderen als ein Wert b
Input	Eingabe(daten)
Output	Ausgabe(daten)
Performance	Leistung, Ausführung
Ranking	Einteilung nach dem Platz in einer vorher festgelegten Reihenfolge
Score	Punktzahl, Bewertung
Task	Aufgabe, hier häufig: Proteinsequenz unbekannter Struktur
Template	Schablone, hier häufig: für die Struktur der Task-Sequenz, teilweise auch als Modell bezeichnet

Abkürzungen

CATH	*C*lass, *A*rchitecture, *T*opology, *H*omologous Superfamily
FSSP	*F*old Classification based on *S*tructure-*S*tructure Alignmnet of *P*roteins (Databank)
PDB	*P*rotein *D*ata *B*ank
RMSD	*R*oot *M*ean *S*quare *D*eviation
SCOP	*S*tructural *C*lassification *o*f *P*roteins
Sek.Str.	*Sek*undär*str*uktur

1 Einleitung und Zielsetzung

Die fortlaufenden Genomprojekte produzieren eine ständig wachsende Zahl an Proteinsequenzen mit unbekannter Struktur und Funktion der Proteine. Gerade die Funktion ist für viele Zweige der Wissenschaft von Interesse. Das Protein Hämoglobin ist für den Sauerstofftransport im Blut verantwortlich. Die Keratine bilden den Hauptbestandteil von Haaren, Nägeln und Federn. Die Festigkeit der Knochen beruht auf den Eigenschaften des Faserproteins Kollagen. Nicht zuletzt sind die Moleküle des Immunsystems Proteine. Um zu verstehen wie der Körper auf Eindringlinge reagiert, ist es notwendig die Funktionsweise dieser Proteine zu kennen.

Die Funktion eines Proteins hängt direkt von seiner Struktur ab, welche wiederum allein von der Aminosäuresequenz bestimmt wird, wie Anfinsen schon 1973 gezeigt hat [1]. Die experimentelle Aufklärung dieser Strukturen ist zwar möglich, aber kompliziert und sehr zeitaufwändig. Die Bestimmung der Aminosäuresequenz hingegen ist aufgrund des geringen Aufwandes sogar schon automatisiert worden. Dementsprechend ist der Nutzen, den man aus computergestützten Vorhersagen der Proteinstruktur aus der Sequenz ziehen kann, groß.

Das Problem ein dreidimensionales Bild der Struktur vorherzusagen, ist in den vergangenen Jahren zu einem der Schwerpunkte der Bioinformatik geworden, der Wissenschaft, welche die Aufgabenstellungen der Biochemie und Genetik von einer rechnerunterstützten Seite angeht [2]. Die Versuche reichen von *ab initio* Verfahren, welche vollständig auf physikalischen und chemischen Prinzipien beruhen und deren alleiniger Ausgangspunkt die Aminosäuresequenz ist [3, 4], bis hin zum *Homology Modelling*, welches die Information zu Hilfe nimmt, die in Sequenz- und Strukturdatenbanken verfügbar ist [5].

Der Prozess des Modelling läuft normalerweise nach dem in Abbildung 1 dargestellten Schema ab. Für eine Sequenz mit unbekannter Struktur - der „Task-Sequenz“ - wird nach geeigneten Proteinen gesucht, die als Templates - „Strukturschablonen“ - verwendet werden können. Dies geschieht z.B. durch Abgleichen der Task-Sequenz mit allen Proteinsequenzen die in einer

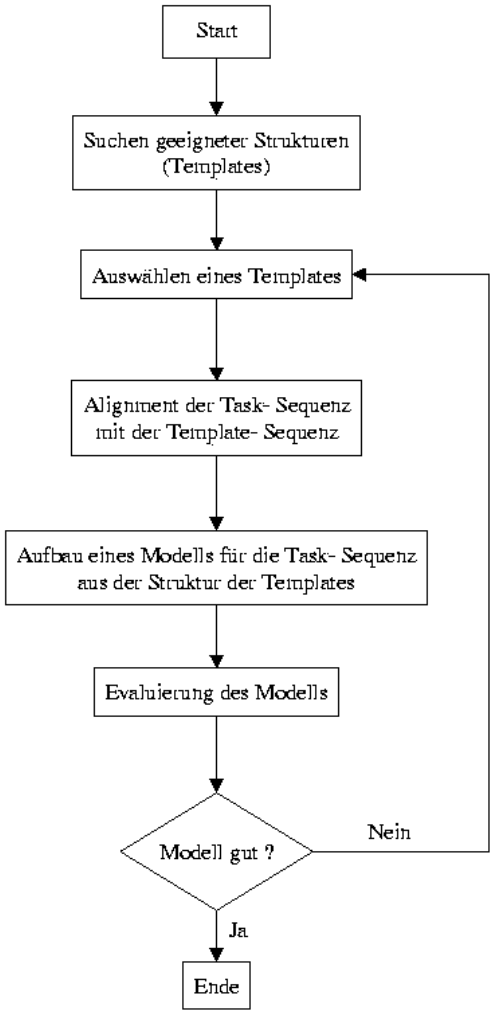


Abbildung 1: Vorgehensweise beim Protein-Modelling

Datenbank verfügbar sind. Dabei wird eine Wertung für die Qualität dieses Alignments berechnet. Als Modell für die Struktur wird das Protein verwendet, welches die beste Wertung erzielt hat.
Ist ein Modell ausgewählt wird seine Basisstruktur auf die Sequenz übertragen, deren Bild man erhalten möchte. Energetische Berechnungen erlauben eine Evaluierung des Modells als letzten Schritt.

Der schwedische Student Jesper Lundström hat in seiner Diplomarbeit [6] die Aufgabe angegriffen, die Qualität eines Modells vorherzusagen. Dazu hat er sechs verschiedene Server verwendet, die „online“ verfügbar sind. Namentlich verwendete er PDB-BLAST [7], FFAS [8], 3D-PSSM [9], GenTHREADER [10], Sam-T98 [11] und Inbgu [12]. Diese Server lieferten ihm jeweils zehn gegen die Test-Sequenz abgeglichene Templates. Zusammen mit diesen wird eine Bewertung der Anpassung angegeben. Basierend auf dieser Wertung wurden drei Gruppen aufgestellt. Die erste Gruppe enthielt alle Templates. Gruppe Zwei beinhaltete nur die Templates, welche innerhalb der Web-Server die beste Benotung erhielten. Die letzte Gruppe bestand aus denjenigen Proteinen, deren Bewertung über einem festgelegten Wert lag. Für jede dieser Gruppen bildete er Modelle für die Struktur der Task-Proteine. Der nächste Schritt war, pro unbekanntem Protein, die sechs Modelle der verschiedenen Server auf Ähnlichkeit zu untersuchen und jedem die Anzahl ähnlicher Modelle zuzuordnen. Das Gleiche wurde mit den Strukturen der Templates vor dem eigentlichen Modelling durchgeführt. Für verschiedene Kombinationen dieser Zahlen und mit dem zusätzlichen Gebrauch der auf ein einheitliches Maß gebrachten Bewertungen der Alignments durch die Web-Server wurde dann versucht, die Qualität eines Modells vorherzusagen.
Die grundlegende Idee der vorliegenden Arbeit ist, eine ähnliche Strategie zu verwenden. Zum einen soll der Einsatz von Web-Servern wegfallen, da ihre Dienste je nach aktueller Belastung viel Zeit in Anspruch nehmen können. Zum anderen ist der Vergleich der Modelle und Templates sehr zeitaufwändig. Es wird demnach ein Ansatz gesucht, bei dem auch dieser Schritt entfällt. Benötigt wird eine existierende Einteilung der Proteine, die den Schritt des Modellings ersetzt und einen Hinweis darauf gibt, welche Proteine gute Templates sein können. Aus den möglichen Datenbanken, wie z.B. FSSP [13], CATH [14] oder SCOP [15] (s. Abschnitt 2.2) wird die Einteilung der SCOP-Datenbank verwendet. In ihr werden Proteine manuell in unterschiedliche Strukturklassen auf vier verschiedenen Leveln sortiert. Die enthaltene Information sollte eine fundierte Grundlage für die gestellten Anforderungen

liefern.

Ziel dieser Arbeit ist es, die SCOP-Klassifizierung von Proteinen vorauszusagen. Die Eindeutigkeit dieser Vorhersage soll als Maß für die Qualität eines Modells verwendet werden. Zu diesem Zweck werden neuronale Netzwerke mit aus Alignments gewonnenen Daten und einer auf der SCOP-Einteilung beruhenden Wertung trainiert. Diese Wertung besagt, ob Task und Template die gleiche Einteilung besitzen. Es werden Vorhersagen gemacht, um die Qualität des Trainings zu bestimmen. Abschließend wird eine Prognose für ein Set von Proteinen durchgeführt, welche noch nicht in der SCOP-Datenbank klassifiziert sind. Die Qualität dieser Voraussage wird durch ein Modelling und dessen Bewertung überprüft.

2 Grundlagen

2.1 Proteinstruktur

2.1.1 Aufbau von Proteinen

Proteine sind aus Aminosäuren aufgebaut, deren grundlegende Struktur in Abbildung 2 gezeigt wird. In natürlichen Proteinen kommen 20 verschiedene Aminosäuren vor. Sie unterscheiden sich nur im Aufbau der Seitenkette, die im Allgemeinen mit R bezeichnet wird. Verknüpft werden diese Bausteine durch Ausbilden einer Bindung zwischen der Aminogruppe und der Carboxylgruppe. Die entstehende Stickstoff-Kohlenstoffbindung heißt Peptidbindung. Dementsprechend werden derartig aufgebaute Ketten auch als Polypeptide[1] bezeichnet. Die Aminosäuren werden sowohl mit einem Drei-Buchstaben-Code wie auch mit einem Ein-Buchstaben-Code bezeichnet, um das Ausdrücken langer Sequenzen zu vereinfachen (vgl. Tab. 1). Die Länge der Polypeptidketten in Proteinen reicht von nur 40 bis zu mehr als 4000 Aminosäureresten. Typische Proteine mit einer Kettenlänge von 250 Aminosäuren haben so eine Variationsvielfalt von $20^{250} = 1.8{*}10^{325}$ möglichen, unterschiedlichen Sequenzen. Obwohl nur ein kleiner Bruchteil aller vorstellbaren Proteine auch wirklich existiert, gibt es noch immer eine immense

[1] poly [gr.] = viel

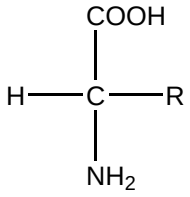


Abbildung 2: Allgemeine Strukturformel einer α-Aminosäure. Der zentrale Kohlenstoff ist mit einer Aminogruppe (NH_2), einem Carbonsäurerest (COOH), einem Wasserstoffatom (H) und einem variierenden Rest R verbunden. Es gibt 20 verschiedene proteinogene R-Gruppen (Tab. 1).

Tabelle 1: Drei- und Ein-Buchstaben-Code der proteinogenen Aminosäuren.

Aminosäure	Drei-Buchstaben-Code	Ein-Buchstaben-Code
Alanin	Ala	A
Arginin	Arg	R
Asparagin	Asn	N
Asparaginsäure	Asp	D
Cystein	Cys	C
Glutamin	Gln	Q
Glutaminsäure	Glu	E
Glycin	Gly	G
Histidin	His	H
Isoleucin	Ile	I
Leucin	Leu	L
Lysin	Lys	K
Methionin	Met	M
Phenylalanin	Phe	F
Prolin	Pro	P
Serin	Ser	S
Threonin	Thr	T
Tryptophan	Trp	W
Tyrosin	Tyr	Y
Valin	Val	V

Anzahl von verschiedenen Aminosäureketten.

2.1.2 Proteinfaltung

Da die dreidimensionale Struktur von nativen Proteinen nur von ihrer Peptidsequenz abhängt [1], gibt es ebensoviele verschiedene Strukturen wie Sequenzen. Mehrere Faktoren bestimmen zusammen die Faltung eines Proteins [16]. Dabei spielen Wechselwirkungen mit der natürlichen Umgebung ebenso eine Rolle wie attraktive und repulsive Kräfte innerhalb des Moleküls.
Es hat sich gezeigt, dass die Hydrophobizität eine der wichtigsten Kräfte zwischen Molekül und Umgebung ist. Ihr Einfluss auf die Faltung beruht auf den ungünstigen Wechselwirkungen zwischen polaren Wassermolekülen und zum Teil unpolaren Aminosäureseitenketten. Proteine, die im Allgemeinen polare sowie unpolare Reste enthalten, tendieren dazu Strukturen anzunehmen, in denen hauptsächlich polare Seitenketten in Kontakt mit einer wässrigen Umgebung treten. Membranproteine zeigen umgekehrte Vorlieben, da der innere Bereich einer Membran eine unpolare Umgebung darstellt.
Innerhalb des Proteins spielt neben den schwach attraktiven Van-der-Waals Kräften [17] die Ausbildung von Wasserstoffbrücken eine große Rolle. In ihnen besteht eine attraktive Wechselwirkung zwischen einem Elektronendonatoratom (D) und einem Elektronenakzeptoratom (A) :

$$\text{A–H}\ldots|\text{D}$$

In einem korrekt gefalteten Protein werden alle möglichen Wasserstoffbrücken auch wirklich ausgebildet. Weitere Kräfte, welche die dreidimensionale Struktur von nativen Proteinen bestimmen, sind ionische Interaktionen und die Ausbildung von Disulfidbrücken zwischen zwei Cystein-Resten (s.a. [18]).

Die Struktur eines nativen Proteins wird auf vier verschiedenen Ebenen angegeben. Die *Primärstruktur* beschreibt die Abfolge der Aminosäuren in der Polypeptidkette. Die *Sekundärstruktur* eines Polymers stellt die lokale Gerüstkonformation dar. Bei Proteinen bezeichnet sie Faltmuster wie He-

lices, Faltblattstrukturen und Windungen. Die *Tertiärstruktur* ist die dreidimensionale Anordnung der Sekundärstrukturelemente, sowie die Plazierung der Aminosäureseitenketten im Raum. Proteine können aus mehr als einer Aminosäurekette bestehen. Zusammengehalten werden solche Assoziate grundsätzlich durch alle attraktiven Kräfte, die keine echte chemische Bindung beinhalten[2]. Die räumliche Anordnung der Polypeptid-*Untereinheiten* wird durch die *Quartärstruktur* beschrieben.
Weiterhin falten sich Polypeptidketten, die aus mehr als 200 Aminosäureresten bestehen, gewöhnlich in mehrere globuläre Gruppen. Diese Gebilde sind hierarchisch zwischen der Sekundär- und Tertiärstruktur einzuordnen. Sie verleihen den entsprechenden Proteinen eine zwei- oder mehrlappige Struktur und werden als *Domänen* bezeichnet.

2.1.3 Sequenz-Alignments

Der Begriff Sequenz-*Alignment* bedeutet ein „zur Deckung bringen“ von zwei oder mehr Sequenzen. Solche Alignments können z.B. verwendet werden um evolutionäre Verwandtschaften zwischen Proteinen aufzufinden. Alle Alignment-Methoden verwenden ein System zur Bewertung des Abgleiches der Sequenzen. Wird eine Task-Sequenz gegen eine ganze Datenbank abgeglichen, so wird das am nächsten verwandte Protein wahrscheinlich die beste Note erhalten.
Die einfachste Methode eine Bewertung vorzunehmen ist das *Identity Scoring*. Nach Vergleich der Sequenzen wird ein Aminosäurenpaar entweder mit einer Eins bewertet, wenn es aus zwei identischen Säuren besteht, oder mit einer Null, sofern sie sich unterscheiden. Da die Proteinsequenzen der Evolution unterliegen, was bedeutet, dass Substitutionen, Insertionen oder auch Deletionen vorkommen, scheint ein fortgeschritteneres Wertungssystem sinnvoll.
Die evolutionären Substitutionen sind keineswegs völlig zufällig, sondern fol-

[2] Ausnahme von dieser Aussage bilden z.B. Disulfidbrücken zwischen zwei Cystein-Resten [18].

gen einem Muster. Es ist wahrscheinlicher, dass eine Aminosäure durch eine chemisch ähnliche ersetzt wird, als durch eine völlig andere Eigenschaften aufweisende. So ist z.B. ein Austausch eines Valin-Restes gegen ein Isoleucin wesentlich wahrscheinlicher als ein Austausch eines Valins gegen ein Tryptophan (s. Abb. 3). Entsprechend weiterenwickelte Bewertungs-Schemata beno-

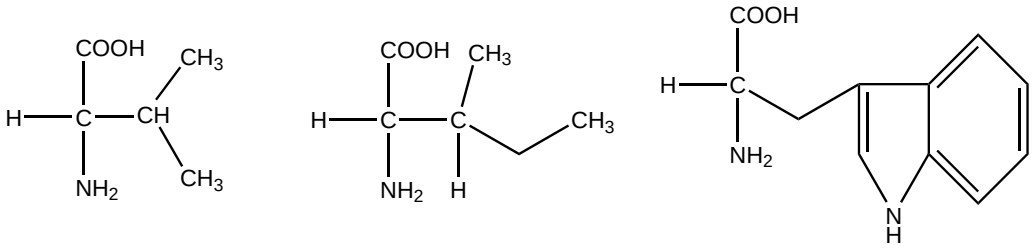


Abbildung 3: Die Aminosäuren Valin (I), Isoleucin (II) und Tryptophan (III). Ein evolutionärer Austausch von Valin gegen Isoleucin erscheint warscheinlicher als gegen Tryptophan.

ten Aminosäurenpaare nach der Wahrscheinlichkeit eines Austausches, üblicherweise repräsentiert durch eine 20 x 20 Matrix [19]. Um Insertionen und Deletionen gerecht zu werden, wird einem Alignment erlaubt Lücken aufzuweisen (s. Abb. 4). Um zu verhindern, dass ein Alignment z.B. nur aufgrund eines genauen Abgleichs der Enden der Sequenzen gut bewertet wird, bestimmt man Strafen für das Eröffnen und Erweitern einer Lücke.
Der nächste Schritt ist, den Abgleich mit der höchsten möglichen Benotung zu finden, um das bestmögliche Alignment zwischen zwei Sequenzen zu er-

Task-Sequenz :	`-RHYPGDFSPA-`
Template-Sequenz :	`ARF-PADFT-AE`

Abbildung 4: Beispiel eines paarweisen Sequenz-Alignments. Die Länge der alignierten Sequenzen beträgt 10 Aminosäurereste; „ - “ stellt eine Lücke dar. Bzgl. der Kodierung der Aminosäuren siehe Tabelle 1 auf Seite 8.

halten. Naiv wäre hier die Vorgehensweise, alle möglichen Alignments durchzuführen und dann das Beste auszuwählen. Eine Anpassung von zwei Sequenzen aus nur jeweils 100 Aminosäuren erlaubt über 10^{75} Möglichkeiten, was diesen Ansatz schon im Keim erstickt. Erfreulicherweise gibt es eine Gruppe von Algorithmen, die das am besten punktende Alignment mit einer Anzahl von Schritten berechnen können, welche in der Größenordnung von mn liegt. Dabei sind m und n die Längen der abzugleichenden Sequenzen.

Ein Weg die Qualität eines Alignments zu verbessern ist Informationen über vorhergesagte Sekundärstrukturelemente zu verwenden. Diese Vorhersage ist mit über 75%iger Genauigkeit möglich [20]. Statt die Aminosäuren nur aufgrund der Verwandschaft ihrer Reste miteinander in Beziehung zu setzen, wird ebenfalls berücksichtigt, ob die alignierten Säuren in gleichen Sekundärstrukturelementen liegen.
Um auch Proteine zu detektieren, die nicht evolutionär verwandt sind (i.e. weniger als 30 % Sequenzidentität besitzen), aber dennoch ähnliche Strukturen aufweisen, werden multiple Sequenz-Alignments verwendet. Die Task-Sequenz wird gegen alle Proteine in einer Datenbank abgeglichen. Die Treffer, welche eine Bewertung über einem zuvor festgelegten Schwellenwert erzielen, werden verwendet um ein *Profil* aufzustellen. Basierend auf der Häufigkeit, mit der jede einzelne Aminosäure gegen alle 20 möglichen abgeglichen wird, baut man eine neue n x 20 Matrix zur Bewertung des Alignments auf; n bezeichnet die Länge der Task-Sequenz. Unter Verwendung dieser neuen Matrix wird nochmals die Sequenz gegen die gesamte Datenbank aligniert. Der Vorgang kann wiederholt werden, um weitere mögliche Templates zu finden. Die Detektion wird also prinzipiell ermöglicht, wenn zwei Proteine zu einem dritten verwandt sind. Diese Methode liefert, neben der Möglichkeit auch entfernte Verwandte eines Proteins aufzuspüren, eine signifikante Verbesserung der Qualität des Alignments für Sequenzidentitäten, bei denen paarweise Alignments bekanntermaßen nicht mehr funktionieren [5].

2.1.4 Protein Modelling

Beim Protein Modelling wird ein Protein bekannter Struktur als Schablone für die unbekannte Struktur eines zweiten Proteins verwendet. Der erste Schritt im Protein Modelling ist demnach das Auffinden eines geeigneten Templates (s.a. Abb. 1, S. 4). Dies geschieht, wie in Abschnitt 1 erwähnt, z.B. durch Verwendung von (multiplen) Sequenz-Alignments. Aus allen möglichen Templates, deren Benotung einen bestimmten Wert überschreitet, wird eines ausgewählt. Dies kann nach dem einfachen Kriterium des besten Alignments erfolgen, oder nach anderen Maßstäben, wie z.B. einer möglichst engen evolutionären Verwandtschaft. Wurde das Schablonenprotein im ersten Schritt nicht gegen die Sequenz unbekannter Struktur aligniert, so wird dies nun getan.
Im nächsten Schritt wird die Struktur des Backbones des Templates als Modell für die Backbonefaltung des Task-Proteins verwendet. Mit Hilfe des Sequenz-Alignments können die Seitenketten des zu modellierenden Proteins an das Rohmodell gehängt werden. Energetische Berechnungen erlauben die Optimierung ihrer Positionen und leichte Anpassungen der Struktur des Peptid-Rückgrates. Je nach verwendeter Methode und benötigter Qualität können einzelne Bereiche der Struktur von Hand modelliert werden. Das Modell wird im letzten Schritt evaluiert (s. nächsten Pkt. (2.1.5)).

2.1.5 Strukturvergleiche

Es gibt eine Reihe von Methoden um die Ähnlichkeit zwischen Proteinen zu messen [21]. Im Allgemeinen gilt dabei: je ähnlicher sich Modell und korrekte Struktur sind, umso besser ist die Qualität des Modells.
Die gebräuchlichste Methode die Ähnlichkeit zu messen ist die Berechnung der *Root Mean Square Deviation (RMSD)* nach möglichst exakter Überlagerung der Strukturen.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(q_i - q_i')^2}{n}} \tag{1}$$

Dabei entsprechen q_i und q_i' den Positionen der alignierten Atome im Backbone von Modell und wahrer Struktur; n ist die Anzahl der alignierten Backbone-Atome pro Molekül. Eine hohe RMSD bedeutet also eine geringe strukturelle Übereinstimmung zwischen Modell und wirklichem Protein. Der Nachteil der RMSD ist ihre globale Eigenschaft. Ein gutes Modell, dessen Struktur nur an einer Stelle (z.B. an einem Ende) stark von der Schablone abweicht, bekommt eine hohe RMSD. Andere Methoden, wie der in dieser Arbeit verwendete LGScore [22], vermeiden dieses Problem, indem sie die Bewertung auf das signifikanteste gemeinsame Segment beziehen. Dies sollte idealerweise möglichst lang und ähnlich sein. Details werden unter Punkt 3.7.3 beschrieben.

2.2 Strukturdatenbanken

Es gibt mehrere Datenbanken, die Informationen über die Struktur von Proteinen zur Verfügung stellen. Die bekannteste dürfte die *Brookhaven Protein Data Bank* (PDB) sein [23, 24]. Sie enthält fast alle aufgeklärten Proteinstrukturen, in Form der Koordinaten aller Atome in der nativen Faltung der Proteine. Andere Datenbanken verwenden die von der PDB zur Verfügung gestellten Daten, um eine Einteilung der Proteinstrukturen in verschiedene Klassen vorzunehmen.

2.2.1 FSSP

Die FSSP (F*old classification based on* S*tructure-*S*tructure alignment of* P*roteins*) Datenbank [13, 25] verwendet alle in der PDB beschriebenen Strukturen, deren Peptidketten mehr als 30 Aminosäurereste enthalten. Die Ketten werden in ein *repräsentatives Set* und *Sequenzhomologe* der Strukturen des repräsentativen Sets aufgeteilt. Letztere haben mehr als 25 % Sequenzidentität und das repräsentative Set enthält keine Paare solcher Sequenzhomologen. Ein Strukturvergleich aller Proteine gegen alle anderen wurde im repräsentativen Set durchgeführt. Die resultierenden Alignments sind in den

FSSP-Einträgen aufgeführt. Zusätzlich sind auch die Struktur-Alignments der Proteine mit ihren Sequenzhomologen angegeben.

2.2.2 CATH

CATH [14, 26, 27] ist eine hierarchische Klassifikation von Proteinstrukturen, die Proteine auf vier Hauptebenen gruppiert: *Class* (C), *Architecture* (A), *Topology* (T) und *Homologous Superfamily* (H). Die Klasse (C) wird aus der Sekundärstruktur für mehr als 90 % der Proteinstrukturen automatisch abgeleitet. Die Architektur (A) beschreibt die grundlegende Orientierung der Sekundärstrukturen, unabhängig von Verbindungen. Sie wird von Hand festgelegt. Auf dem Topologielevel (T) werden Strukturen nach den topologischen[3] Verbindungen und der Anzahl der Sekundärstrukturelemente zusammengefasst. Die homologen „Superfamilien“ (H) beinhalten Proteine mit stark ähnlichen Strukturen und Funktionen. Die Zuteilungen für die Topologie und die homologen Superfamilien werden durch Sequenz- und Strukturvergleiche bestimmt. Für Details siehe [26] und [28].

2.2.3 SCOP

SCOP ist ein Akronym für *Structural Classification of Proteins*. Die SCOP-Datenbank [15] an der Universität von Cambridge enthält zur Zeit (1. März 2001) 13 220 PDB-Einträge, die auf vier verschiedenen Ebenen klassifiziert werden [29, 30]. Die Einteilung wird nach visueller Inspektion und Vergleich der Strukturen vorgenommen. Klassifikationseinheit ist die Protein-Domäne (s. Pkt. 2.1.2, S. 10). Kleine Proteine und die meisten mittlerer Größe haben nur eine Domäne und werden demnach als Ganzes betrachtet. Die Domänen großer Proteine werden üblicherweise einzeln eingeteilt. Die hierarchischen Ebenen der SCOP-Klassifizierung sind folgende:

Family. Proteine werden in gleiche Familien gruppiert, wenn sie entweder

[3] [gr.] die Struktur und Anordnung von geometrischen Figuren im Raum betreffend

Sequenzidentitäten von 30 % oder mehr aufweisen oder weniger Identität in der Sequenz zeigen, aber sehr ähnliche Strukturen und Funktionen haben; z.B. Globine mit Sequenzidentitäten von 15 %.

Superfamily. Familien deren Proteine geringe Sequenzidentitäten zeigen, aber deren Strukturen und häufig auch funktionelle Eigenschaften eine gemeinsame evolutionäre Herkunft wahrscheinlich erscheinen lassen, sind zusammen in Superfamilien platziert; bspw. die variablen und konstanten Domänen der Immunglobuline.

Common Fold. Superfamilien und Familien besitzen eine gemeinsame Faltung, wenn ihre Proteine die gleichen wichtigen Sekundärstrukturen in gleicher Anordnung und mit den gleichen topologischen Verknüpfungen aufweisen.

Class. Die verschiedenen Faltungen sind in Klassen gruppiert. Die meisten Faltungen sind in eine der folgenden fünf strukturellen Klassen eingeteilt:

1. All-α, Strukturen hauptsächlich aus α-Helices aufgebaut.
2. All-β, Strukturen hauptsächlich aus β-Faltblättern aufgebaut.
3. α/β, Strukturen enthalten sowohl α-Helices als auch β-Faltblätter.
4. $\alpha+\beta$, Strukturen in denen α-Helices und β-Faltblätter weitgehend getrennt sind.
5. *multi-domain*, Strukturen mit Domänen verschiedener Faltung, für die momentan keine Homologen bekannt sind.

Für Oligopeptide, kleine Proteine, theoretische Modelle, Nucleinsäuren und Kohlenhydrate sind andere Klassen festgelegt.

Neben den bereits zuvor erwähnten Datenbanken FSSP und CATH gibt es noch weitere wie z.B. Entrez [31] und DDBASE [32]. Die Unterscheidung zwischen evolutionären Verwandtschaften und solchen, die auf der Physik und Chemie der Proteine beruhen, ist jedoch einzig bei SCOP realisiert.

2.3 Neuronale Netzwerke

Wenn Computer verwendet werden, um Probleme zu lösen, kann meist explizit beschrieben werden, wie das gesuchte Ergebnis aus den vorhandenen Daten gewonnen wird. Es gibt aber auch Beispiele, bei denen dieser einfache Ansatz nicht möglich oder zu aufwändig ist. Das Wiedererkennen von Gesichtern auf diese Weise würde erfordern, dass der Computer zuvor alle Gesichter in jeglichen Gefühlslagen und Grimassen gesehen und gespeichert hat. Selbst wenn dies geschehen ist, würde eine einfache Schnittwunde oder eine aufgeplatzte Lippe eine Wiedererkennung verhindern.
In einem anderen Ansatz wird dem Computer der Zusammenhang zwischen den Daten beigebracht, indem man Input/Output-Paare zur Verfügung stellt. Durch bestimmte Algorithmen [33] kann der Rechner die Funktionalität lernen. Das verwendete Konstrukt für derartiges Lernen ist häufig ein *neuronales Netzwerk*.
Trotzdem die hier zugrundeliegende verknüpfende Idee zum ersten Mal schon von Aristoteles in seiner Vorstellung mentaler Assoziationen formuliert wurde, gewann die Forschung auf diesem Gebiet erst in den achtziger Jahren des letzten Jahrhunderts an Bedeutung [34]. Die Theorie neuronaler Netzwerke basiert auf der Physiologie des menschlichen Gehirns. Die Bausteine werden *Neuronen* oder *Knoten* genannt. Diese sind durch *Synapsen* verbunden, denen eine Gewichtung zugeordnet ist, um Signale zu verstärken oder zu unterdrücken. Die Knoten sind in *Schichten* angeordnet und jeder Knoten erhält nur Daten aller mit ihm verbundenen Knoten aus der übergeordenten Schicht und gibt nur Daten an Knoten der nächsten Schicht weiter. Ein Beispiel eines zweischichtigen Netzwerkes ist in Abbildung 5 gezeigt.

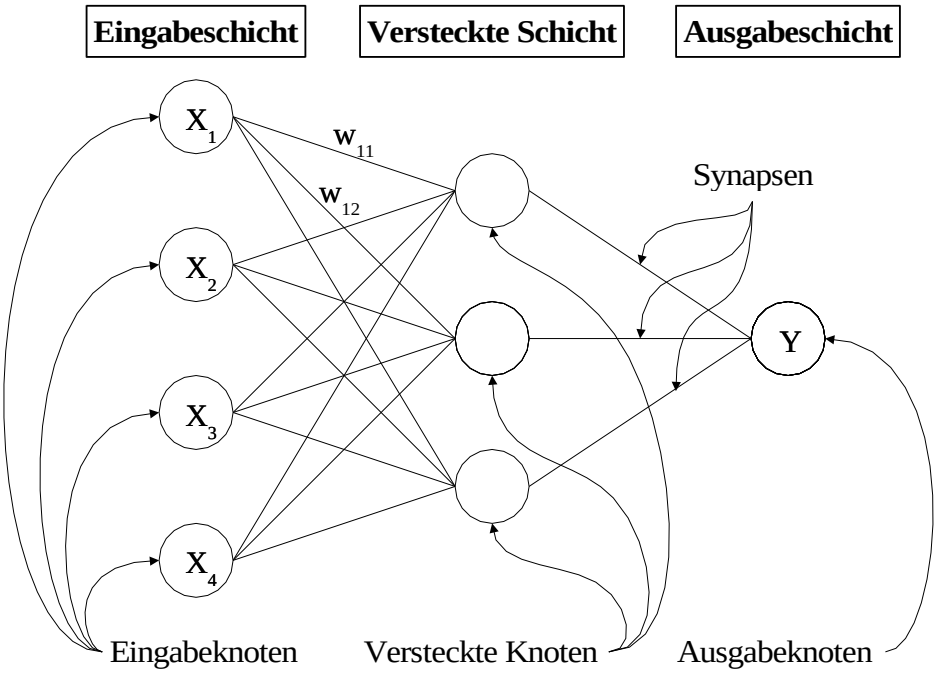


Abbildung 5: Schematischer Aufbau eines zweischichtigen neuronalen Netzwerkes mit vier Eingabeknoten ($X_{1\text{-}4}$), drei versteckten und einem Ausgabeknoten (Y). Der Input für einen Knoten wird aufsummiert und ein Tendenzparameter addiert. Dieser Wert wird dann durch eine Funktion geleitet, die üblicherweise linear oder logistisch ist.

Die anfänglichen Daten (X_i) werden entlang der Synapsen mit einer Gewichtung (w_{ji}) multipliziert. In den Neuronen der versteckten Schicht werden diese Werte aufsummiert und ein Tendenzparamter wird addiert:

$$a_j = \sum_{i=1}^{d} (w_{ji} * X_i) + w_{j0} \tag{2}$$

Der so erhaltene Wert a_j wird durch eine Funktion skaliert. Hier wird meist eine lineare (3) oder logistische (4) Funktion verwendet.

$$y = m * x + b \tag{3}$$

$$y = \frac{1}{1 + e^{-x}} \tag{4}$$

Das Gleiche geschieht beim Übergang zur nächsten Schicht, die in einem zweilagigen Netzwerk schon der Ausgabeschicht entspricht.
Beim Trainieren eines solchen Konstruktes werden die Gewichtungen und Tendenzparameter durch wiederholtes Durchlaufen lang entwickelter Algorithmen [33] den gegebenen Daten angepasst. Dabei wird der Input durch das Netzwerk geleitet und die Differenz zwischen der Netzwerk-Vorhersage und dem benötigten realen Output minimiert.

2.3.1 Trainings- und Testset

Um die Qualität der Vorhersage eines neuronalen Netzes zu bestimmen, wird der vorhandene Datensatz in zwei Teile aufgeteilt. Einer wird zum Trainieren des Netzes verwendet, das Trainingsset. Mit dem anderen Teil der Daten - dem Testset - wird eine Vorhersage gemacht. Da das korrekte Ergebnis bekannt ist, kann so die Performance bestimmt werden. Es sollte darauf geachtet werden, dass Trainings- und Testset beide repräsentativ für den ursprünglichen Datensatz sind.

2.3.2 Übertraining

Wird ein neuronales Netzwerk für eine Aufgabe entwickelt und trainiert, gibt es zwei grundsätzliche Dinge, die definiert werden müssen: der Aufbau des Netzes und die Dauer des Lernvorganges [35, 36]. Die Anzahl der Eingabe- und Ausgabeknoten wird durch die verwendeten Daten festgelegt. In einem zweischichtigen Netz bleibt somit, die Anzahl der Neuronen der versteckten Schicht[4] zu bestimmen. Die Dauer des Lernvorganges wird definiert durch die Anzahl der Zyklen, die der Trainingsalgorithmus durchlaufen wird. Die Justierung dieser beiden Parameter ist aus einem bestimmten Grund sehr wichtig.
Wird einer dieser beiden Werte überdimensioniert, passiert etwas, das als *Übertraining* bezeichnet wird [33, 37]. Das Netz lernt das Hintergrundrauschen im verwendeten Datensatz. Eine Vorhersage der Funktionswerte desselben Datensatzes würde extrem gute Ergebnisse zeigen. Aber das Netzwerk hat seine Fähigkeit zu generalisieren verloren (s.a. Abb. 6). Eine Probe mit einem unabhängigen Testsatz wird keine befriedigenden Ergebnisse mehr zeigen. Die richtige Anzahl der versteckten Knoten und Trainingszyklen hängt stark von der Menge der Trainingsdaten ab [38]. So kann bei vorkommendem Übertraining eine Optimierung dieser Werte ebenso helfen wie die Verwendung eines größeren Datensatzes.

[4] Normalerweise verlaufen die Synapsen von allen Neuronen einer Schicht zu jeweils allen Neuronen der nächsten Schicht.

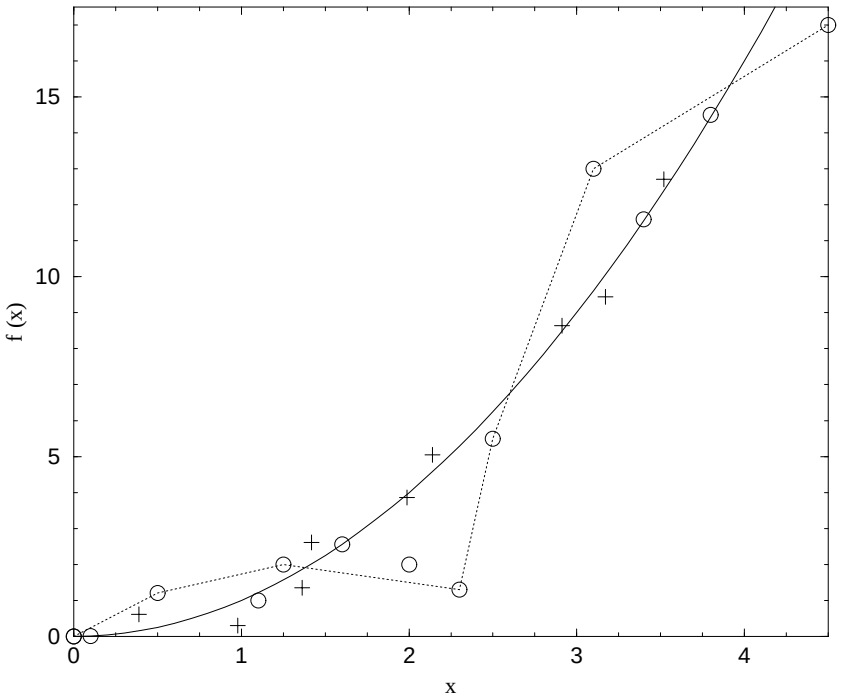


Abbildung 6: Die Vorhersagen eines optimal trainierten und eines übertrainierten neuronalen Netzwerkes. Die Kreise symbolisieren das Trainingsset, die Pluszeichen das Testset. Die durchgezogene Linie repräsentiert ein optimal trainiertes Netzwerk und die gepunktete Linie zeigt die Vorhersage eines übertrainierten Netzwerkes.

2.3.3 Optimierung der Netzwerkparameter

Üblicherweise wird zur Optimierung das neuronale Netz für eine stetig ansteigende Zahl von Trainingszyklen trainiert und die Performance gegen den Testdatensatz gemessen. Wenn die Qualität der Vorhersage schlechter wird, beginnt das Übertraining. An diesem Punkt wird das Training gestoppt. Das beschriebene Verfahren kann für eine variierende Anzahl an versteckten Knoten durchgeführt werden. Allerdings löst diese Methode häufig Diskussionen darüber aus, ob die Performance am Ende der tatsächlichen Qualität entspricht, da das Netzwerk auf den Testdatensatz hin optimiert wurde.
Nicht vergessen werden darf hier der Aspekt der benötigten Rechenzeit. Werden die Netzwerke zu kompliziert, enthalten sie z.B. mehr als zwei Schichten von Synapsen oder wird ein sehr großer Datensatz verwendet, so kann die benötigte Rechenzeit zum optimalen Training leicht ineffiziente Ausmaße annehmen. Es bleibt am Ende dem Wissenschaftler überlassen zu entscheiden welche Parameter für ihn am zweckdienlichsten sind.

3 Daten und Methoden

3.1 Verwendete Proteinsätze

3.1.1 Haupt-Proteinsatz

In dieser Arbeit wurde von einem Satz von 12 805 Proteindomänen aus der SCOP-Datenbank [15] (Version 1.39) ausgegangen (s. S. 15 ff.). Alle unvollständigen und fehlerbehafteten Proteindateien wurden ausgefiltert. So blieben 10 306 verwendbare Domänen. Daraus wurde ein repräsentatives Set ausgewählt, dessen Proteine weniger als 40 % Sequenzidentität aufwiesen. Dieses Set bestand aus 1058 Proteinen. Aufgrund leichter Fehler in den Dateien konnten 31 nicht als Templates verwendet werden. Der verwendete Proteinsatz bestand aus 1058 Task- und 1027 Template-Proteinen.

3.1.2 Unabhängiger Proteinsatz

Um die Qualität der gewonnenen Ergebnisse zu überprüfen, wurde ein Satz unabhängiger, zu diesem Zeitpunkt (Juli 2001) noch nicht SCOP-klassifizierter Proteine verwendet (zu SCOP s. Pkt. 2.2.3, S. 15). Ausgewählt wurden die Task-Proteine von LiveBench-2 [39, 40]. Jede Woche werden dort, mittels neu aufgeklärter Proteinstrukturen, Strukturvorhersagen getestet. In dieser Arbeit wurden die Proteinsequenzen benutzt, welche sich von August 2000 bis inkl. Dezember 2000 bei LiveBench-2 akkumulierten. Das verwendete unabhängige Set enthielt 201 Task- und die gleichen 1027 Template-Proteine wie der Haupt-Proteinsatz.

3.2 Methodik durchgeführter Alignments

Alle für diese Arbeit benötigten Alignments wurden mit dem Programm *Palign* durchgeführt. Es verwendet standardisierte „dynamic programming“ Algorithmen und ist verfügbar unter:
http://www.sbc.su.se/~arne/pscan/palign.tar.gz .
Für weitere Informationen siehe [41] und [42]. Es wurden alle Task-Sequenzen

aus beiden Proteinsätzen gegen alle Template-Sequenzen aligniert. Dafür wurden acht verschiedene Methoden verwendet, die in Tabelle 2 aufgelistet sind.

Tabelle 2: Die verschiedenen verwendeten Alignment-Methoden. Aufgelistet ist die Art des Alignments und die Strafen für das Eröffnen einer Lücke (gap opening) und das Erweitern einer Lücke (gap extension).

Alignment Methode	Beschreibung
1	Lokales Sequenz-Profil Alignment
	gap opening: go = -15
	gap extension: ge = -1
2	Globales Sequenz-Profil Alignment
	gap opening: go = -15
	gap extension: ge = -1
3	Lokales Sequenz-Profil Alignment
	gap opening: go = -10
	gap extension: ge = -1
4	Globales Sequenz-Profil Alignment
	gap opening: go = -10
	gap extension: ge = -1
5	Lokales Seq. + Sek.Str.-Profil + Sek.Str. Alignment
	gap opening: go = -15
	gap extension: ge = -1
6	Globales Seq. + Sek.Str.-Profil + Sek.Str. Alignment
	gap opening: go = -15
	gap extension: ge = -1
7	Lokales Seq. + Sek.Str.-Profil + Sek.Str. Alignment
	gap opening: go = -10
	gap extension: ge = -1
8	Globales Seq. + Sek.Str.-Profil + Sek.Str. Alignment
	gap opening: go = -10
	gap extension: ge = -1

3.3 Basisdaten

Gleichzeitig mit dem reinen Sequenz-Alignment wurden die Koordinaten der Cα-Atome und der Seitenkettenschwerpunkte des bekannten Proteins auf die jeweiligen Alignmentpartner übertragen. Dies ermöglicht Berechnungen simpler Kontaktpotentiale, die attraktive und repulsive Wechselwirkungen berücksichtigen [43, 44]. Die daraus und aus den durchgeführten Alignments gewonnenen Daten, die in dieser Arbeit verwendet wurden, werden im Folgenden erläutert.

3.3.1 $E_{C\alpha\text{-}C\alpha}$

Zur Berechnung der auf den Cα-Atomen beruhenden Wechselwirkungsenergie wurden, wie oben erwähnt, die Koordinaten der entsprechenden Atome auf die Partner im Sequenz-Alignment übertragen. Die daraus bestimmbare Energie ist ein distanzabhängiges Kontaktpotential, ähnlich der Van-der-Waals Funktion [17]. Alle α-Kohlenstoffe werden als energetisch gleichwertig betrachtet. Die Cα-Cα-Energie einer Konformation wird nach der, im nächsten Abschnitt beschriebenen, Funktion (5) berechnet. In diesem Fall wird nur der erste Term der Gleichung betrachtet. Zu beachten ist, dass die in dieser Arbeit verwendete Cα-Cα-Energie eine heuristische Größe mit willkürlichen Einheiten ist. Weitere Details unter [44].

3.3.2 E_{total}

Um die totale Energie kalkulieren zu können, wurden die Koordinaten sowohl der Cα-Atome als auch der Seitenkettenschwerpunkte, gemäß den Sequenz-Alignments, vom Template auf die Task-Sequenz übertragen. Im Gegensatz zu den Cα-Atomen werden die Schwerpunkte der Aminosäurereste als energetisch verschieden betrachtet. In [44] wurde die folgende Formel entwickelt:

$$
\begin{aligned}
E &= \sum_{(1\leq i\leq N)} \sum_{(i+4\leq j\leq N)} \left(\frac{A_{ij}}{r_{ij}^8} - \frac{B_{ij}}{r_{i\alpha_j}^4}\right) \\
&+ \sum_{(1\leq i\leq N)} \sum_{(i+4\leq j\leq N)} \left(\frac{A_{\alpha\alpha}}{r_{\alpha\alpha}^8} - \frac{B_{\alpha\alpha}}{r_{\alpha\alpha}^4}\right) \\
&+ \sum_{(1\leq i\leq N)} \sum_{((1\leq j\leq i-3)\cup(i+3\leq j\leq N))} \left(\frac{A_{i\alpha}}{r_{i\alpha_j}^8} - \frac{B_{i\alpha}}{r_{i\alpha_j}^4}\right)
\end{aligned} \tag{5}
$$

mit:

$$
\begin{aligned}
A_{ij} &= -\epsilon_{ij}(R_{ij}^a)^8 \\
B_{ij} &= -2\epsilon_{ij}(R_{ij}^a)^4
\end{aligned} \tag{6}
$$

Sie beschreibt ein distanzabhängiges Kontaktpotential, in welchem die Kontaktenergien durch die ϵ_{ij}-Werte und die Kontaktabstände durch die R_{ij}-Werte repräsentiert werden (s. Gl. (6)). Diese sind in [44] tabelliert und in [45] abgeleitet. Die r_{ij}-Werte der Gleichung (5) entsprechen den geometrischen Abständen zwischen zwei Seitenketten und werden unter Verwendung kartesischer Koordinaten berechnet.
Die drei Terme der Gleichung (5) beschreiben die Interaktionen zwischen den α-Kohlenstoffen, den Seitenketten und zwischen α-Kohlenstoffen und Seitenketten. Auch die totale Energie stellt in dieser Arbeit eine heuristische Größe dar, deren Werte ohne Einheiten angegeben werden.

3.3.3 Alignment-Score

Die Bewertung des Alignments ist eine dimensionslose Zahl, deren Berechnung auf einer Ähnlichkeitsmatrix beruht. Diese wird bei der Erstellung des Profiles aufgebaut. Das hier verwendete Profil wurde von Altschul et al [46] übernommen. Zur Entwicklung von Profil-Alignments siehe [47].

3.3.4 Alignment-Länge

Der Wert der Alignment-Länge bezeichnet die Anzahl der alignierten Aminosäuren. In der verwendeten Form enthält er ebenfalls die Lücken in den Alignments.

3.3.5 TOP 5 und TOP 80

Zur Berechnung des von uns TOP 5 genannten Wertes wurden alle Alignments einer Task-Sequenz nach dem oben beschriebenen Alignment-Score sortiert. Für jedes Alignment wurde gezählt, wie oft die SCOP-Werte *Class* und *Fold* (vgl. Pkt. 2.2.3) des Templates unter den fünf - in dieser Reihenfolge - besten Templates für das gleiche Task-Protein, vorkamen. Diese Zahl wird im Folgenden als „TOP 5“ bezeichnet.

Der als TOP 80 bezeichnete Wert wurde folgendermaßen berechnet: es wurden die alignierten Proteinpaare für jedes Task-Protein absteigend nach dem Alignment-Score sortiert. Aus allen acht Alignment-Methoden (s. Tab. 2, S. 25) stellten wir die jeweils zehn besten Paare zusammen in eine TOP 80-Liste. Dann wurde für jedes Task/Template-Paar gezählt, wie oft die SCOP-Werte (*Class* und *Fold*) des Template-Proteins bei den Template-Proteinen der Liste (der entsp. Taskseq.) vorkamen. Diese Zahl wird im Folgenden als „TOP 80“ bezeichnet.

3.3.6 SCOP-Score

Die SCOP-Einteilungen (s. Pkt. 2.2.3) der miteinander alignierten Proteine wurden verglichen. Dabei betrachteten wir nur die Einteilungen bezüglich *Fold* und *Class*. Eine neue Bewertung wurde jedem Paar zugeteilt. Besaßen die abgeglichenen Proteine die gleiche Einteilung bezüglich *Fold* und *Class*, so wurde dem Paar eine Eins gegeben. Unterschieden sie sich in der Einteilung mindestens eines Levels der beiden betrachteten, so wurde ihnen eine Null zugeteilt. Diese Bewertung wird im Folgenden als „SCOP-Score“ bezeichnet.

Wird in dieser Arbeit in Zusammenhang mit der SCOP-Datenbank von gleicher Faltung oder Einteilung zweier Proteine gesprochen, so ist im Allgemeinen die Einteilung bezüglich Fold *und* Class gemeint. Es soll damit eine SCOP-Score von Eins verdeutlicht werden.

3.4 Durchgeführte Ansätze

Um die neuronalen Netze zu trainieren, wurden aus den über eine Million durchgeführten Alignments (s. Pkt. 3.1.1) repräsentative Paare ausgewählt. Dabei wurden drei Ansätze gemacht.

3.4.1 Aufbau des ersten Ansatzes

Der erste Ansatz enthielt pro Datenset (s. Abschnitt 3.5) 1000 Proteinpaare. Diese wurden aus allen alignierten Paaren zufällig ausgewählt. Fünfhundert der verwendeten Alignments besaßen einen SCOP-Score (s. Pkt. 3.3.6) von Eins und die anderen 500 wurden mit einer Null bewertet.

3.4.2 Aufbau des zweiten Ansatzes

Die im zweiten Ansatz aufgebauten Datensätze enthielten die Daten aller 19 451 alignierten Paare, deren Proteine die gleiche Einteilung in der SCOP-Datenbank (s. Pkt. 2.2.3) bezüglich Faltung und Klasse besaßen. Weiterhin wurde die gleiche Anzahl an Paaren mit sich unterscheidender Einteilung verwendet. Dabei haben wir die abgeglichenen Proteine nach dem Kriterium des höchstmöglichen Alignment-Score (vgl. Pkt. 3.3.3) ausgewählt.

3.4.3 Aufbau des dritten Ansatzes

Der dritte Ansatz bestand aus den gleichen 19 451 Alignments, welche mit einem SCOP-Score (Pkt. 3.3.6) von Eins bewertet wurden, von denen schon im zweiten Ansatz Gebrauch gemacht wurde. Repräsentativ für die Alignments von Proteinen, die nicht die gleiche Faltung besitzen, wurden hier

die Daten doppelt so vieler Proteinpaare verwendet. Diese 38 902 mit einem SCOP-Score von Null bewerteten Paare wurden zufällig aus allen möglichen (vgl. Pkt. 3.1.1) ausgewählt.

3.5 Aufbau der Datensätze

Durch Kombination der verfügbaren Basisdaten (s. Abschnitt 3.3) und der acht verschiedenen Alignment-Methoden (Tab. 2, S. 25) wurden, in allen Ansätzen gleichermaßen, die im Folgenden beschriebenen Datensets erstellt.

3.5.1 Set 1 bis Set 5

Die Datensätze 1 bis 5 wurden aus der, im vorigen Abschnitt beschriebenen, Auswahl der verfügbaren Daten aufgebaut. Redundante Alignments und Alignments zwischen identischen Proteinen wurden ausgefiltert. Die verwendeten Basisdaten sind in Abschnitt 3.3 beschrieben. Der Aufbau der Sets verlief folgendermaßen:

Set 1. Das erste Datenset enthielt nur die Spalten Alignment-Länge und Alignment-Score als Daten für die Eingabeknoten des neuronalen Netzes. Die Daten für den Ausgabeknoten waren die entsprechenden SCOP-Scores.

Set 2. Das zweite Set bestand aus vier Basisspalten und dem SCOP-Score. Die verwendeten Basisdaten waren $E_{C\alpha\text{-}C\alpha}$, E_{total}, Alignment-Score und die Alignment-Länge.

Set 3. Beim dritten Set wurden den Basisdaten des zweiten Sets als weitere Spalte die TOP 5-Wertung hinzugefügt. Es bestand also aus $E_{C\alpha\text{-}C\alpha}$, E_{total}, Alignment-Score, Alignment-Länge, TOP 5 und dem SCOP-Score der ausgewählten Alignments.

Set 4. Das vierte Datenset stellte eine Kombination aus den Basisdaten aller acht verschiedenen Alignment-Methoden dar (s. Abschnitt 3.2). Damit

enhielt es 8*4 + 1 = 33 Eingabedaten, wobei neben den Energien, der Alignment-Länge und dem Alignment-Score, der bei allen Methoden gleiche SCOP-Score nur einmal als Referenz für das Ausgabeneuron verwendet wurde.

Set 5. Im fünften Set wurden die Sets 3 und 4 kombiniert. Es enthielt, repräsentativ für die Eingabeknoten des neuronales Netzes, die vier Basisdaten: $E_{C\alpha\text{-}C\alpha}$, E_{total}, Score und Länge des Alignments, aus allen acht Alignment-Methoden. Im Gegensatz zu Set 3 wurde hier aber die TOP 80-Wertung als weitere Datenspalte verwendet, nicht TOP 5. Repräsentativ für den Ausgabeknoten des Netzes wurde wieder der, bei allen Alignment-Methoden gleiche, SCOP-Score verwendet.

Die Datensets, die keine Kombination der verschiedenen Alignment-Methoden verwendeten, wurden für alle acht Methoden einzeln aufgebaut und getestet.

3.5.2 Set 1.1 bis Set 4.1

Die Datensätze 1.1 bis 4.1 wurden nahezu aufgebaut wie die Sets 1 bis 4. Sie enhielten aber nicht nur die Daten, der in Abschnitt 3.4 ausgewählten Alignments des jeweiligen Ansatzes, sondern die Daten aller 1 086 566 durchgeführten Alignments (s. Pkt. 3.2).
Der Aufbau des Sets 5.1 war aus computertechnischen Begrenzungen leider nicht möglich. Es mangelte an Speicher und Rechenzeit.

3.5.3 Set 1.2 bis Set 5.2

Die Datensätze 1.2 bis 5.2 wurden prinzipiell aufgebaut wie die Datensätze 1 bis 5. Jedoch wurden hier die Informationen aus dem unabhängigen Proteinsatz verwendet (s. Abschnitt 3.1.2). Dabei wurde keine Auswahl getroffen, sondern alle 206 427 dort durchgeführten Alignments einbezogen.

3.6 Neuronale Netzwerke

Alle in dieser Arbeit verwendeten neuronalen Netzwerke waren zweilagig. Sie wurden für die Datensätze 1 bis 5 optimiert und getestet. Die Sets 1.1 bis 4.1 und 1.2 bis 5.2 verwendeten zur Vorhersage die von Set 1 bis 5 trainierten und danach gespeicherten Netzwerke.
Die Software für die neuronalen Netzwerke wurde vom Paket NETLAB [48] für MATLAB [49] zur Verfügung gestellt.

3.6.1 Aktivierungsfunktionen

Die Aktivierungsfunktionen der versteckten Knoten der Netze waren als logistische Funktionen von NETLAB festgelegt. Als Aktivierungsfunktion des Ausgabeknotens konnte eine lineare oder eine logistische Funktion gewählt werden (s. Gl. (3) u. (4), S. 19). Ein Unterschied besteht darin, dass eine lineare Aktivierungsfunktion auch Vorhersagen grösser eins und kleiner null zulässt, während eine logistische Funktion nur Vorhersagen im Intervall von null bis eins, inklusive der Ränder, ergibt.

3.6.2 Optimierung der verwendeten Netzwerkparameter

Sämtliche, für die vorliegende Arbeit benötigten Netzwerke wurden in ihrem Aufbau, bezüglich der Anzahl der Trainingszyklen und hinsichtlich der Aktivierungsfunktion des Ausgabeneurons optimiert. Dabei wurde nicht die unter Abschnitt 2.3.3 beschriebene Methode verwendet, sondern es wurde im Abstand von 25 Trainingszyklen der *Matthews Koeffizient* (s. Gl. (7), S. 34) berechnet. Die maximale Anzahl Trainingszyklen betrug 600 für die Datensets 1 bis 3 und 1000 für die Sets 4 und 5. Um die Anzahl der versteckten Knoten für Set 1 zu optimieren wurde diese Prozedur für zwei, fünf und acht Knoten angewendet. Bei Set 2 und 3 wurde für drei, sechs und neun Knoten optimiert und die beiden letzten Sets wurden für 5, 10, 15, 20 und 25 Knoten getestet. Zum Festlegen der Aktivierungsfunktion des Outputneurons führten wir die oben beschriebene Vorgehensweise jeweils für eine lineare und

eine logistische Funktion durch (Gl. (3) u. (4), S. 19).
Für die Datensätze 1 bis 3 wurden sämtliche Daten für alle acht verschiedenen Alignment-Methoden (Tab. 2, S. 25) ausgewertet. Jede Optimierung wurde im *Cross Validation* Test vorgenommen. Die Methode ist in Abschnitt 3.6.4 beschrieben. Die Parameter, welche die besten Matthews Koeffizienten erreichten, wurden im Weiteren verwendet.

3.6.3 Aufbau der optimierten Netze

Die zur Vorhersage der Sets 1.1 bis 4.1 und 1.2 bis 5.2 verwendeten neuronalen Netze sind in ihrem Aufbau in Tabelle 3 zusammengefasst.

Tabelle 3: Die für die angegebenen Datensätze optimierten Netzwerkparameter.

Set	Eingabeknoten	Versteckte Knoten	Ausgabeknoten	Trainingszyklen
1	2	5	1	600
2	4	6	1	600
3	5	6	1	600
4	32	25	1	1000
5	33	25	1	1000

3.6.4 Cross Validation

Die Performance der verwendeten Netze wurde in einem *Cross Validation* Test überprüft. Die Datensets wurden dazu in fünf gleich große Segmente aufgeteilt. Von diesen wurden reihum jeweils vier als Trainingsset zusammengefasst und das fünfte als Testset verwendet. So können Instabilitäten in der Qualität der Vorhersage schnell erkannt werden. Meist deuten diese Instabilitäten auf ein Übertraining oder eine ungleichmäßige Verteilung der Daten im Set hin. Zur Theorie siehe Punkt 2.3.1 bzw. 2.3.2.

3.6.5 Evaluierung der Performance

In dieser Arbeit wurden drei Maße verwendet, um die Qualität der Voraussagen zu bestimmen. Alle beruhen auf der Einteilung der vorhergesagten Werte in *wahr positiv (tp), wahr negativ (tn), falsch positiv (fn)* und *falsch negativ (fn)*[5]. Dazu wird für die vorhergesagten Werte ein Cutoff bestimmt. Alle Werte die über ihm liegen werden als Eins interpretiert, alle Werte die darunter liegen als Null. Der in dieser Arbeit verwendete Cutoff war 0.5. Eine Vorhersage von genau 0.5 wurde als Eins interpretiert.
Um die Einteilung in tp, tn, fp und fn vorzunehmen, wird der vorhergesagte Wert nach Zuordnung zu Null oder Eins mit dem korrekten Wert verglichen. Stimmt der zugeordnete Wert mit dem wirklichen Wert überein, so ist die Einteilung wahr, ansonsten unwahr. Ist der vorhergesagte Wert eine Eins, so wird er als positiv bezeichnet, andernfalls als negativ. Abbildung 7 verdeutlicht dies.

Die verwendeten Maße zur Evaluierung waren:

$$M_c = \frac{(tp * tn) - (fp * fn)}{\sqrt{(tn + fn)(tn + fp)(tp + fn)(tp + fp)}} \tag{7}$$

Der Matthews Koeffizient M_c variiert zwischen -1 und +1, wobei +1 einer perfekten Vorhersage entspricht, 0 einer zufälligen Vorhersage und -1 einer gegenteiligen Vorhersage.
Die ebenfalls verwendeten Parameter Spezifität und Sensitivität sind definiert als:

$$Spezifität = \frac{tp}{tp + fp} \tag{8}$$

$$Sensitivität = \frac{tp}{tp + fn} \tag{9}$$

[5][engl.] t steht für true (wahr), f für false (falsch), p für positive (positiv) und n für negative (negativ).

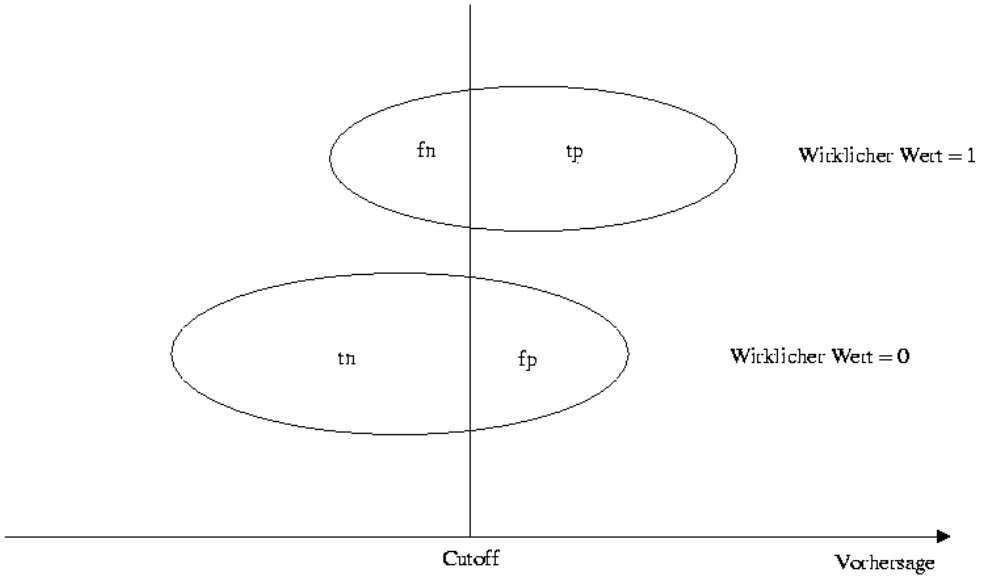


Abbildung 7: Die Graphik verdeutlicht die Terme tp, tn, fp und fn. Alle Vorhersagen die mit dem wirklichen Wert übereinstimmen sind wahr (t), die anderen falsch (f). Ist die Vorhersage eine Eins, ist sie positiv (p), ist sie Null wird sie als negativ (n) bezeichnet.

Die Sensitivität ist die Fraktion der positiv vorausgesagten Werte, die korrekt prognostiziert wurden. Die Spezifität ist der Bruchteil aller möglichen positiven Werte, der erkannt wurde.

3.7 Modellauswahl

Für alle in den Sets 1.1 bis 4.1 und 1.2 bis 5.2 enthaltenen Proteinpaare wurde eine Vorhersage durchgeführt. Dazu wurden, wie auf Seite 32 erwähnt, die von den Sets 1 bis 5 trainierten und optimierten Netzwerke verwendet. Hier wurden die gesamten Daten dieser Sets zum Training genutzt und nicht nur 4/5 wie in der Cross Validation. Für Details siehe Abschnitt 3.5 bzw. 3.6.4. Von allen Templates für jeweils eine Task-Sequenz wurde dasjenige mit der höchsten Prognose ausgewählt. Dieses wurde als bestmögliches Modell für die unbekannte Struktur interpretiert.

3.7.1 Evaluierung für Set 1.1 bis 4.1

Um die oben beschriebene Auswahlmethode des besten Templates zu evaluieren, wurde für die Sets 1.1 bis 4.1 verglichen, ob bzw. wie häufig Task-Protein und ausgewähltes Template in der SCOP-Einteilung bezüglich Class und Fold übereinstimmten (bzgl. SCOP s. Pkt. 2.2.3).

3.7.2 Evaluierung für Set 1.2 bis 5.2

Für die Sets 1.2 bis 5.2 war die obige Methode nicht möglich, da noch keine SCOP-Einteilung der Task-Proteine vorgenommen war (vgl. Pkt. 3.1.2 u. 3.5.3). Es wurde ein einfaches Modell dieser Proteine erstellt, in dem die Koordinaten der C_α-Atome und der Seitenkettenschwerpunkte des Template-Proteins auf die Task-Sequenz übertragen wurden. Die so erhaltene Struktur wurde mit der wahren Struktur verglichen. Zur Evaluierung wurde der im Folgenden erläuterte *LGScore* verwendet.

3.7.3 LGScore

Vor kurzem führten Michael Levitt und Mark Gerstein den nach ihnen benannten LGScore ein [22]. Er dient zur Berechnung der Signifikanz der Ähnlichkeit zwischen zwei Strukturen und beruht auf einer lokalen Betrachtungsweise. Mittels eines heuristischen Algorithmus wird das, die höchste Wertung erzielende, Fragment der alignierten Proteinsequenzen gesucht. Jenes sollte möglichst lang und ähnlich (zwischen den Strukturen) sein. Der LGScore beruht auf der folgenden Wertung:

$$S_{str} = M \left(\sum \frac{1}{1 + (\frac{d_{ij}}{d_0})^2} - \frac{N_{gap}}{2} \right) \tag{10}$$

wobei M gleich 20 ist; d_{ij} ist der Abstand zwischen den Resten i und j, d_0 ist gleich 5 Å und N_{gap} ist die Anzahl der Lücken im Alignment. Um die statistische Signifikanz dieser Wertung zu bestimmen, verwendeten Levitt und Gerstein ein Set von strukturellen Alignments nicht verwandter Proteine. Sie berechneten eine Verteilung von S_{str} in Abhängigkeit von der Alignmentlänge l. Aus dieser Verteilung wurde ein Wert P berechnet, der von S_{str} und l abhing. Dieser Wert gibt die Wahrscheinlichkeit an, dass eine bessere Bewertung zufällig auftritt. Der LGScore ist der negative dekadische Logarithmus des P-Wertes.

$$LGScore = -\log_{10}(P) \tag{11}$$

Ein höherer LGScore bedeutet demnach eine geringere Wahrscheinlichkeit, dass die Ähnlichkeit zwischen zwei Fragmenten zufällig ist.
In dieser Arbeit wurde der LGScore mit dem gleichnamigen Programm LGSCORE berechnet. Auf früheren Erfahrungen beruhend wurde ein LGScore größer als drei als Indikator für eine zufriedenstellende Qualität des Modells verwendet. Für weitere Informationen siehe [21, 22].

4 Ergebnisse

Der Aufbau der durchgeführten Ansätze ist in Abschnitt 3.4 beschrieben, die Berechnung der dazu benötigten Daten in Abschnitt 3.3. Die Ergebnisse der Vorhersagen für eine lineare und eine logistische Aktivierungsfunktion (vgl. Gl. (3) u. (4), S. 19) unterscheiden sich nur minimal. Die am besten prognostizierbaren Sets erreichen mit einer linearen Funktion leicht höhere Matthews Koeffizienten (s. Gl. (7), S. 34) als mit einer logistischen. Der Unterschied liegt maximal in der Größenordnung von 10^{-3}. Aus diesem Grund werden zur besseren Vergleichbarkeit im Folgenden nur Werte für eine lineare Aktivierungsfunktion angegeben.

4.1 Ergebnisse des ersten Ansatzes

Die Berechnung der Matthews Koeffizienten zeigte im Cross Validation Test (Pkt. 3.6.4) Schwankungen von bis zu $\pm$ 0.3. Dies deutet sehr stark auf ein Übertraining des neuronalen Netzes hin. Aus diesem Grund wurde der Ansatz nicht weiter verfolgt. (Bzgl. Übertraining siehe Abschnitt 2.3.2 auf Seite 20.)

4.2 Ergebnisse des zweiten Ansatzes

Die Matthews Koeffizienten der Vorhersagen zeigten durchgängig Werte von $\geq$ 0.98 - unabhängig von dem verwendeten Datensatz. Abbildung 8 zeigt eine Auftragung der vorhergesagten Werte über dem Alignment-Score für das Set 2. Man sieht deutlich, dass Vorhersagen kleiner als 0.5 hauptsächlich im Bereich von 450 bis 700 entlang der Abszisse auftreten.

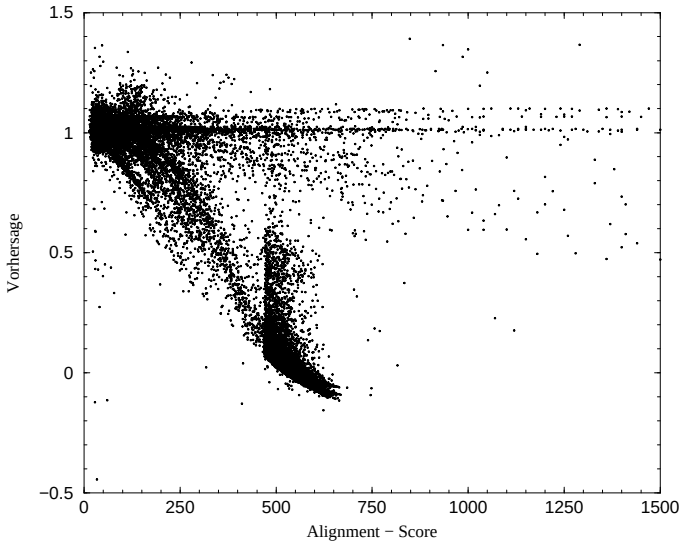


Abbildung 8: Vorhergesagter Wert gegen den Alignment-Score für ein lokales Alignment (Methode 1, Set 2). Im Trainingsset wurden die Zeilen mit dem höchsten Alignment-Score für den Fall unterschiedlicher SCOP-Werte gewählt.

4.3 Ergebnisse des dritten Ansatzes

4.3.1 Set 1 bis Set 5

Eine der Abbildung 8 entsprechende Graphik wird in Abbildung 9 gezeigt. Man sieht dass ein Alignment-Score kleiner als null zu einer Vorhersage von Null führt, während ein Alignment-Score größer als etwa 300 mit einer Voraussage von Eins einhergeht.

Eine Einteilung der vorhergesagten Werte in tp, tn, fp und fn, wie in Abschnitt 3.6.5 beschrieben, eröffnet die Möglichkeit, die im gleichen Abschnitt definierten Matthews Koeffizienten, Sensitivität und Spezifität zu betrachten. Abbildung 10 zeigt die Berechnung des Matthews Koeffizienten für einen steigenden Cutoff. Es ist zu erkennen, dass ein Cutoff nahe der Grenzen null und eins eine schlechtere Performance zeigt als um 0.5. Für die Sets, welche einen „TOP-Score“ beinhalten (Set 3 und Set 5), steigt der Matthews Koeffizient

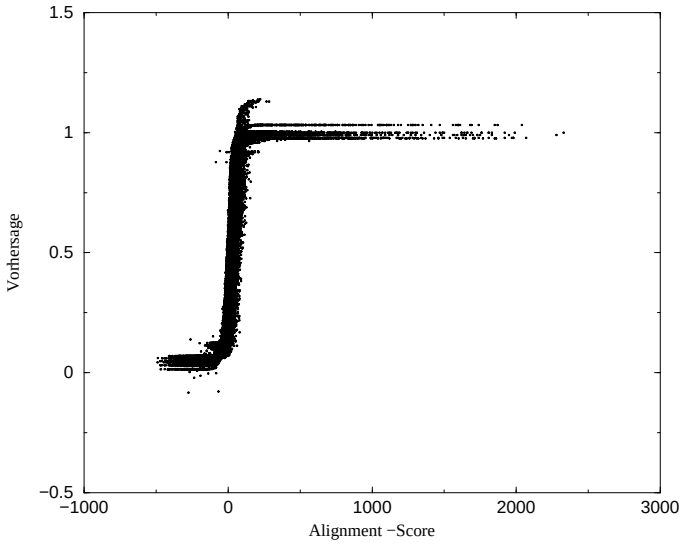


Abbildung 9: Die Graphik zeigt den vorhergesagten Wert über dem Alignment-Score für ein globales Alignment (Methode 2, Set 2). Im Trainingsset wurden die Zeilen mit einem SCOP-Score von Null zufällig gewählt. Das Verhältnis der Zeilen mit einem SCOP-Score von Eins ggü. einem SCOP-Score von Null war 1:2 .

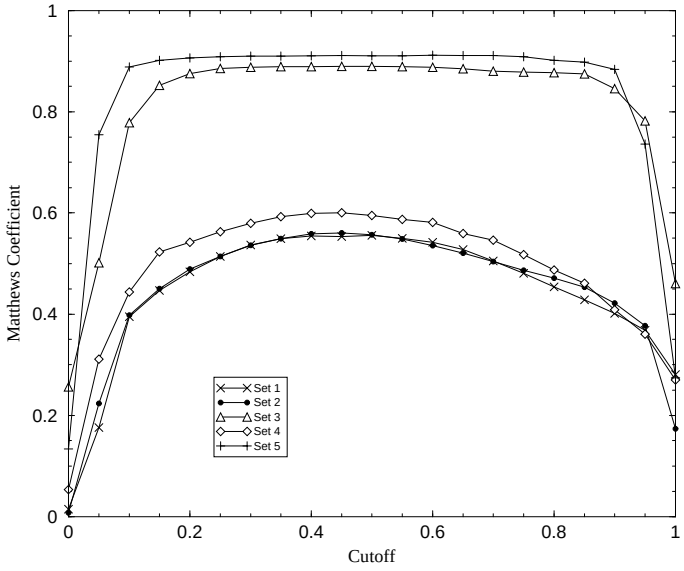


Abbildung 10: Matthews Koeffizient vs. Cutoff für die Sets 1 bis 5. Der Cutoff steigt von null bis eins in Stufen von 0.05. Für die Sets 1 bis 3 ist die bei einem Cutoff von 0.5 am besten abschneidende Alignment-Methode gezeigt.

an den Rändern steiler an als bei den anderen Sets und bleibt bei einem Cutoff zwischen 0.2 und 0.8 nahezu konstant. Insgesamt zeigen Set 3 und 5 eine durchgängig höhere Performance als die Datensätze 1, 2 und 4.
Die Werte des Koeffizienten, berechnet für den in dieser Arbeit verwendeten Cutoff von 0.5 für alle Sets und Alignment-Methoden (vgl. Pkt. 3.5.1 u. Tab. 2, S. 25), sind in Tabelle 4 zusammengefasst. Ein Vergleich der Zahlen zeigt, dass Set 1 und 2 nahezu gleich abschneiden. Die Verwendung der Energien in Set 2 bewirkt bei der besten Alignment-Methode sogar eine leichte Verschlechterung um etwa $2.5*10^{-3}$. Die Kombination der Basisdaten: $E_{C\alpha\text{-}C\alpha}$, E_{total}, Alignment-Score und Alignment-Länge aller acht Methoden in Set 4 ergibt einen Matthews Koeffizienten, der um $4*10^{-2}$ höher liegt als in den ersten beiden Sets. Eine deutliche Verbesserung zeigt sich bei Verwendung des TOP 5-Scores in Set 3. Die mit höchstem Koeffizienten prognostizierende Methode wurde beim Aufbau von Set 5 verwendet. Insgesamt lässt sich durch Vergleich der jeweils besten Werte pro Set die Reihenfolge :

$$\text{Set } 2 \approx \text{Set } 1 < \text{Set } 4 < \text{Set } 3 < \text{Set } 5$$

aufstellen.

Eine Auftragung der Spezifität über der Sensitivität ist in Abbildung 11 zu sehen. Die Parameter werden auf Seite 34 definiert. Auch in dieser Darstellung ist erkennbar, dass die Sets 1 und 2 eine sehr ähnliche Performance zeigen. Set 4 liegt leicht darüber. Wieder schneiden die einen „TOP-Score“ verwendenden Sets als Beste ab. Der Wert der Spezifität fällt bei Set 5, für sinkenden Cutoff, etwas früher ab als für Set 3.
Das bei einem Cutoff von Null keine Sensitivität von Eins erzielt wird, liegt an den - durch die lineare Aktivierungsfunktion möglichen - Voraussagen unter null.

Tabelle 4: Die Matthews Koeffizienten aller Sets und Alignment-Methoden. Die besten Werte pro Set sind *kursiv* gedruckt.

Set	Alignment-Methode	Matthews Koeffizient
1	1	0.34836
1	2	0.43949
1	3	0.35906
1	4	0.45355
1	5	0.40915
1	6	0.54329
1	7	0.41041
1	8	*0.55926*
2	1	0.34854
2	2	0.44532
2	3	0.36037
2	4	0.45214
2	5	0.41923
2	6	0.54484
2	7	0.41646
2	8	*0.55656*
3	1	0.79110
3	2	0.86071
3	3	0.81504
3	4	0.86283
3	5	0.85636
3	6	0.88773
3	7	0.85567
3	8	*0.89099*
4	Kombination	*0.59596*
5	Kombination	*0.91308*

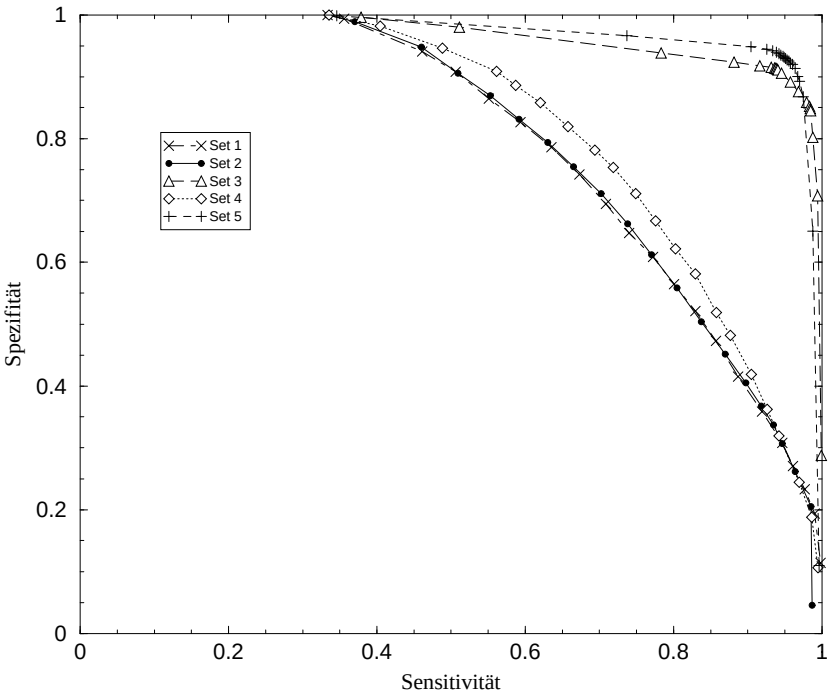


Abbildung 11: Spezifität vs. Sensitivität für einen von null bis eins, in Schritten von 0.05, steigenden Cutoff. Für die Sets 1 bis 3 ist die Alignment-Methode gezeigt, welche bei einem Cutoff von 0.5 den höchsten Matthews Koeffizienten erzielt.

4.3.2 Analyse der Eingabedatensätze

Die im Folgenden häufig verwendeten Terme: $E_{C\alpha\text{-}C\alpha}$, E_{total}, Alignment-Score, Alignment-Länge, TOP 5 bzw. TOP 80 und SCOP-Score, sind in Abschnitt 3.3 beschrieben. Dort wurde schon darauf hingewiesen, dass die in der vorliegenden Arbeit verwendeten Energien heuristische Größen sind, welche ohne Einheiten angegeben werden. Bezüglich des Aufbaus der hier analysierten Datensätze siehe Punkt 3.5.1.
In Set 1 bis 3 wurden dieselben Daten der Alignments von Proteinen mit gleicher SCOP-Einteilung (s. Pkt. 2.2.3) verwendet. Die Alignments mit einem SCOP-Score von Null wurden zufällig ausgewählt. Da von einer statistischen Verteilung ausgegangen wird, werden die Daten aus Set 3 als repräsentativ für die Sets 1 und 2 angesehen.
Wie schon unter Punkt 3.3.6 angemerkt, schließt die in den histographischen Darstellungen verwendete Angabe „SCOP (Fold) gleich“ bzw. „SCOP (Fold) nicht gleich“ die Einteilung bezüglich der SCOP-Klasse ein. Die in diesem Abschnitt beschriebenen Graphiken sind auf den Seiten 46 bis 52 zu finden.

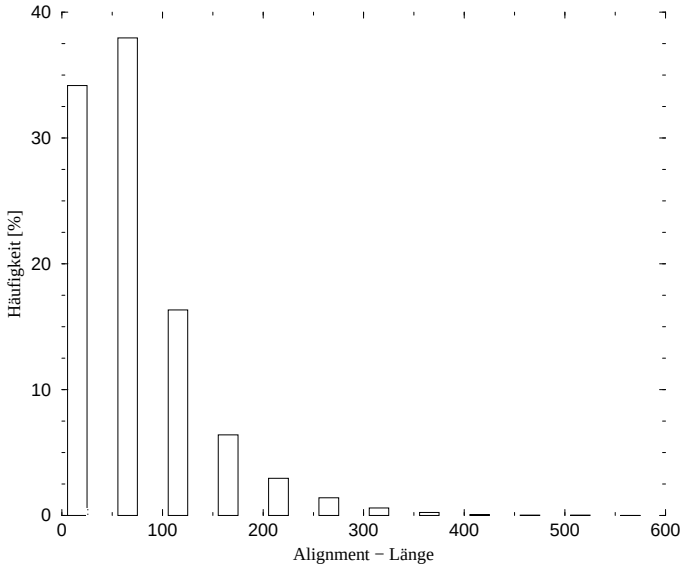
Abbildung 12 zeigt eine histographische Auftragung der vorkommenden Alignment-Längen in Set 3 für ein lokales Alignment. Es ist zu sehen, dass längere Alignments häufiger für einen SCOP-Score von Eins auftreten. Die gleiche Darstellung für ein globales Alignment zeigt Abbildung 13. Da eine globale Eigenschaft einen Abgleich über die gesamte Länge der Peptidkette erzwingt, ist die Häufigkeit der verschiedenen Längen für einen SCOP-Score von Null in etwa gleich derjenigen eines SCOP-Scores von Eins.
Die Betrachtung des Alignment-Scores zeigt ebenfalls Unterschiede zwischen einem globalen und einem lokalen Alignment. Die entsprechenden Auftragungen sind die Abbildungen 14 und 15. Durch die Globalität sind auch negative Bewertungen möglich. In beiden Abbildungen wird deutlich, dass höhere Alignment-Scores häufiger bei einem SCOP-Score von Eins auftreten und niedrigere eher auf einen SCOP-Score von Null hindeuten.
Die Betrachtung der $C\alpha$-$C\alpha$-Energien in Abbildung 16 und 17 signalisiert

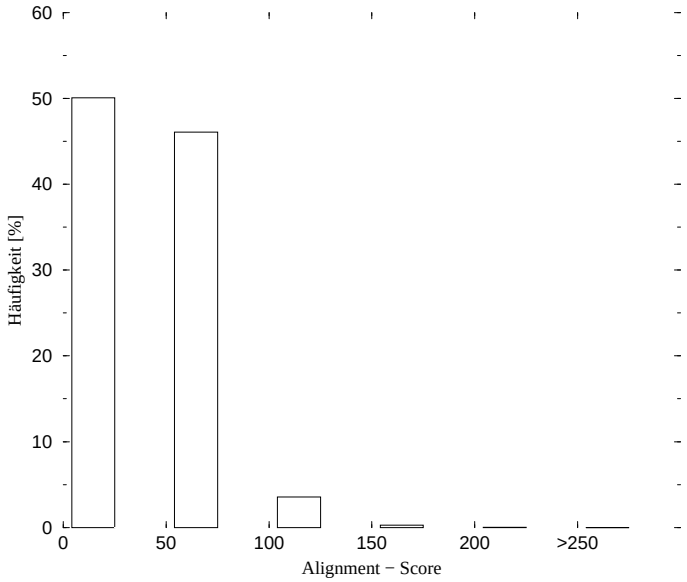
zum einen keine bemerkenswerten Unterschiede zwischen lokalem und globalem Alignment. Zum anderen ist auch keine deutlich verschiedene Verteilung der Häufigkeiten für einen SCOP-Score von Eins bzw. Null zu sehen. Gleiches gilt für die, in den Abbildungen 18 und 19 gezeigten, Histogramme der totalen Energien.

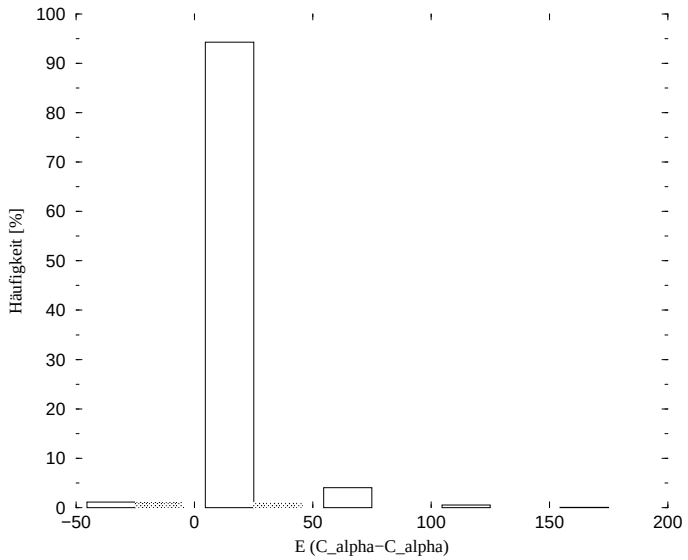
Eine augenfällige Verbesserung in den Matthews Koeffizienten (s. Pkt. 3.6.5) zeigte die Verwendung des TOP 5-Scores (vgl. Tab. 4, S. 42). Die Abbildungen 20 und 21 erklären dies. Für ein lokales wie auch für ein globales Alignment erkennt man die Verknüpfung einer höheren TOP 5-Wertung mit einem SCOP-Score von Eins.

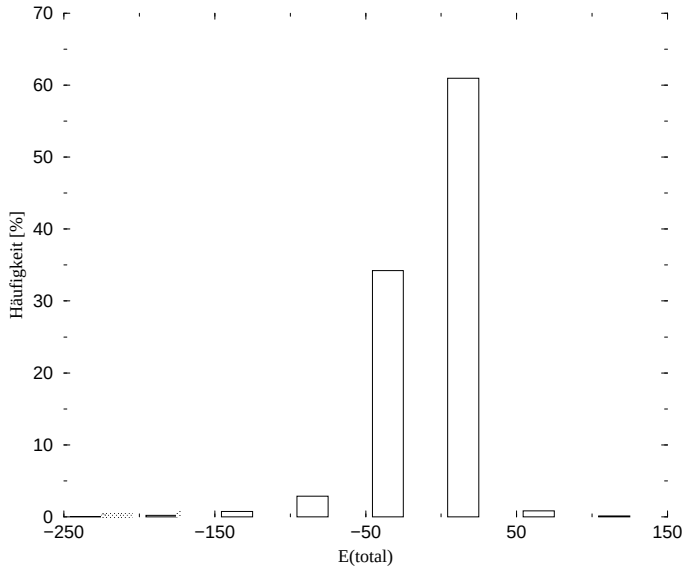
Im Gegensatz zur Verwendung des TOP 5-Scores ergab die Kombination der Daten aller acht Alignment-Methoden nur eine geringe Erhöhung der Matthews Koeffizienten. Abbildung 22 versucht diese zu erklären. Hierfür wurden die im jeweiligen Datensatz ausgewählten Alignments nach absteigendem Alignment-Score sortiert. Für Set 2 wurden die Daten der globalen Methode 8 verwendet, für Set 4 der Durchschnittswert aller acht in diesem Set kombinierten Methoden. Die Abszisse der Graphik zeigt die sich aufsummierende Anzahl der vorkommenden Alignments mit einem SCOP-Score von Null, die Ordinate entsprechend die Alignments mit einem SCOP-Score von Eins. Der Schnittpunkt der zwei Kurven liegt bei einem Alignment-Score von Null. Damit lässt sich erkennen, dass positive Alignment-Scores bei Set 4 etwas stärker auf einen SCOP-Score von Eins hindeuten, als bei Set 2.

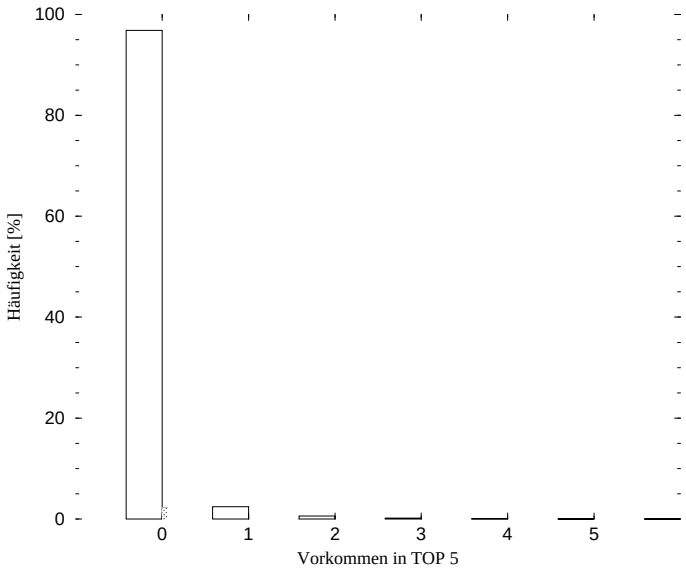
Im fünften Set wurden sowohl alle acht Alignment-Methoden verwendet, als auch ein TOP-Score. Eine histographische Auftragung letzteren ist in Abbildung 23 gezeigt. In dem vergrößerten Ausschnitt ist sichtbar, dass eine TOP 80-Wertung größer als 24 eindeutig auf einen SCOP-Score von Eins hinweist. Aber auch schon Wertungen von fünf und größer deuten eher auf eine gleiche als auf eine unterschiedliche Faltung hin.


40
30
20
10
0
Häufigkeit [%]
0
100
200
300
400
500
600
Alignment − Länge


60
50
40
30
20
10
0
Häufigkeit [%]
0
50
100
150
200
>250
Alignment − Score


Häufigkeit [%]
100
90
80
70
60
50
40
30
20
10
0
−50
0
50
100
150
200
E (C_alpha−C_alpha)


70
60
50
40
30
20
10
0
Häufigkeit [%]
−250
−150
−50
50
150
E(total)


100
80
60
40
20
0
Häufigkeit [%]
0
1
2
3
4
5
Vorkommen in TOP 5

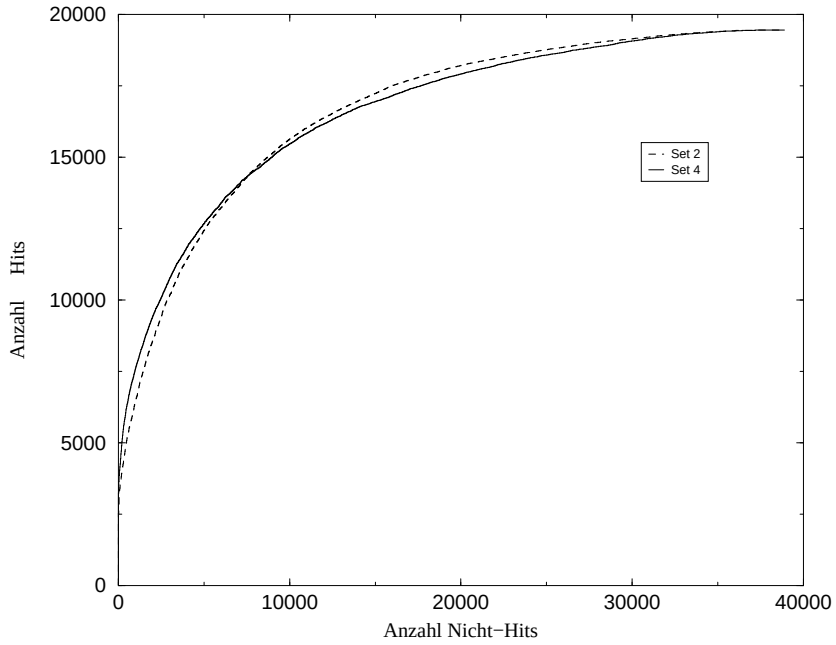
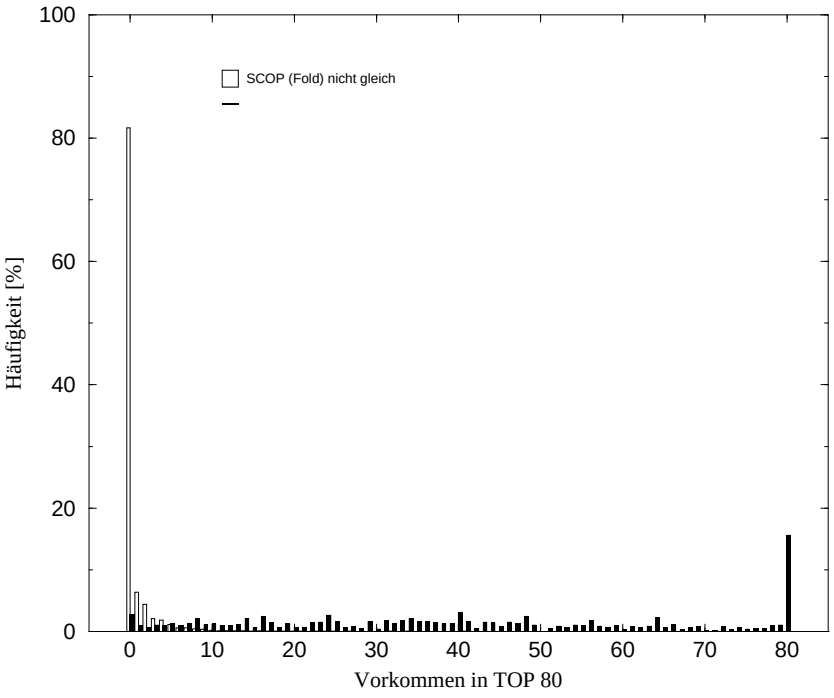


Abbildung 22: Für einen sinkenden Alignment-Score ist die Anzahl der Hits (SCOP-Score = 1) über der Anzahl der Nicht-Hits (SCOP-Score = 0) aufgetragen. Der Graph ist für Set 2 und 4 gezeigt. Im ersten Fall wurden die Werte der globalen Methode 8 verwendet. Im Fall von Set 4 wurde die durchschnittliche Bewertung aller acht Alignment-Methoden als Cutoff verwendet.


SCOP (Fold) nicht gleich
Häufigkeit [%]
Vorkommen in TOP 80
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80

4.3.3 Set 1.1 bis Set 4.1

Der Aufbau der Sets, deren Ergebnisse hier beschrieben werden, ist unter Punkt 3.5.2 erläutert.

Wie unter Punkt 3.7.1 beschrieben, wurde für die Sets 1.1 bis 4.1 eine Vorhersage des SCOP-Scores aller Alignments durchgeführt. Pro Task-Protein wurde das Template mit dem höchsten vorhergesagten Wert als bestmöglichstes interpretiert. Tabelle 5 auf Seite 55 listet die Häufigkeit, mit der das ausgewählte Template die gleichen SCOP-Werte wie das Task-Protein besitzt, auf. Die Ergebnisse zeigen eine sehr schwache Verbesserung von Set 1.1 zu Set 3.1 hin. Set 4.1, welches nach dem Matthews Koeffizienten zwischen Set 1 und 3 liegt (vgl. S. 41), schneidet hier am schlechtesten ab. Zur Darstellung wie sie in Abbildung 24 auf Seite 55 gegeben ist wurden sämtliche Alignments nach sinkendem Score sortiert. Das Template mit der höchsten Vorhersage wurde nach seinem Platz in dieser Ordnung eingeteilt. Die Abszisse zeigt das Ranking, während die Ordinate die Häufigkeit der entsprechenden Einteilungen zeigt. Es ist zu sehen, dass die Templates mit den höchsten Vorhersagen nicht immer auch diejenigen mit dem höchsten Alignment-Score sind. Besonders deutlich wird dies für Set 4.1.

4.3.4 Set 1.2 bis Set 5.2

Um die Vorhersagen der Sets 1.2 bis 5.2 zu evaluieren, wurde pro Task-Protein für die Templates mit der höchsten Vorhersage der in Abschnitt 3.7.3 beschriebene LGScore berechnet. Tabelle 6 auf Seite 56 zeigt die Summe der berechneten LGScores für die, nach dieser Wertung, besten Alignment-Methoden, sowie die Häufigkeit mit der ein LGScore > 3 erhalten wurde. Zum Vergleich sind die von PDB-Blast [7] erzielten Werte mit angegeben. Es handelt sich dabei um einen Web-Server zum Auffinden homologer Sequenzen in einer spezifischen Datenbank. Es sind auch für die Sets 4.2 und 5.2 Alignment-Methoden angegeben, da zur Berechnung des LGScores eines der acht durchgeführten Alignments ausgewählt werden musste.

Auffällig ist, dass die laut den Matthews Koeffizienten besten Sets 3 und 5 hier als Schlechteste einzustufen sind. Besonders unerwartet ist die Performance von Set 5.2. Keiner der getesteten Datensätze hält einem Vergleich mit PDB-Blast stand. Eine der Abbildung 24 entsprechende Graphik ist in Abbildung 25 auf Seite 56 gezeigt. Die Ergebnisse sind ähnlich denen der Sets 1.1 bis 4.1 (vgl. vorigen Pkt. (4.3.3)). Hier ist für Set 5.2 am deutlichsten zu sehen, dass das vermeintlich beste Template häufig nicht mit dem höchsten Score abgeglichen wurde.

Tabelle 5: Anzahl der – nach vorhergesagtem SCOP-Score – besten Alignments mit einem wirklichen SCOP-Score von 1 und diese Angabe in Prozent der max. möglichen 1058 Treffer. Für Set 1.1 bis 3.1 sind die Werte der in dieser Hinsicht besten Alignment-Methoden aufgelistet.

Set	Alignment-Methode	Anzahl SCOP-Score = 1	in Prozent
1.1	6	810	76.56
2.1	8	816	77.13
3.1	8	825	77.98
4.1	Kombination	727	68.71

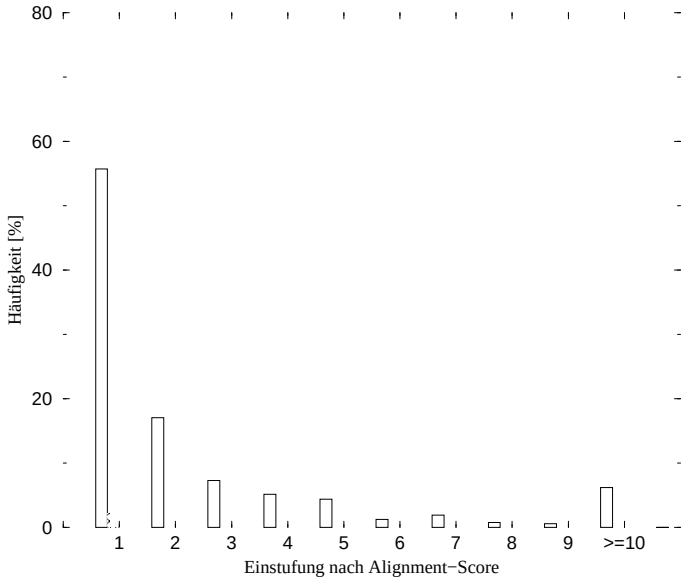
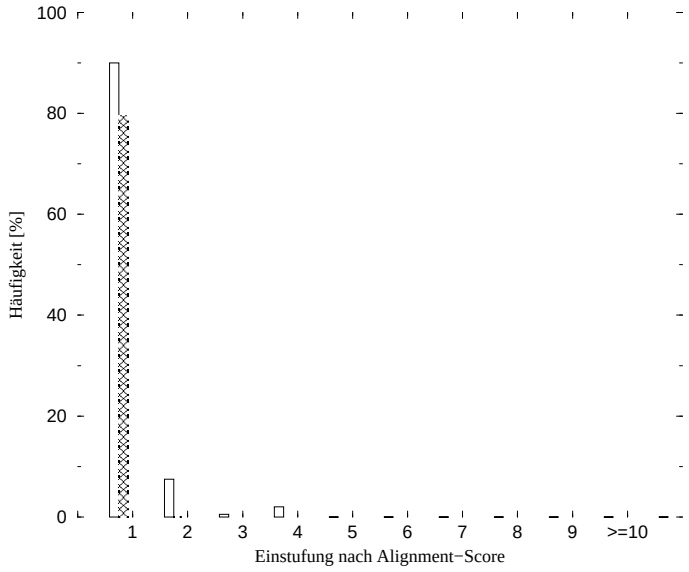


Tabelle 6: Summe des LGScores, Anzahl der – nach vorhergesagtem SCOP-Score – besten Alignments mit einem LGScore > 3 und diese Angabe in Prozent der max. möglichen 201 Treffer. Es sind die Werte für die besten Alignment-Methoden aufgelistet.

Set	Alignment-Methode	$\sum$(LGScore)	Anzahl LGScore >3	in Prozent
1.2	6	279.43791	37	18.41
2.2	6	266.87012	35	17.41
3.2	6	246.35981	26	12.94
4.2	6	234.60891	30	14.93
5.2	6	87.44137	8	3.98
PDB-Blast	—	364.3000	59	29.35



5 Diskussion

Mit einem Satz von knapp über 1000 Proteinen haben wir, jeweils über eine Million, Alignments nach acht verschiedenen Methoden (vgl. Tab. 2, S. 25) durchgeführt. Mit deren Hilfe wurden die in Abschnitt 3.3 beschriebenen Daten berechnet. Wir versuchten, durch verschiedene Kombinationen dieser Daten, neuronale Netzwerke zu trainieren. Dabei sollten sie lernen vorherzusagen, ob die beiden Proteine, aus deren Alignment die gegebenen Daten stammten, die gleiche Einteilung in der SCOP-Datenbank [15] (s. a. Pkt. 2.2.3) besaßen. Repräsentiert wird dieser Fall durch einen SCOP-Score von Eins, während ein SCOP-Score von Null einer unterschiedlichen Einteilung entspricht. Für Details vergleiche Punkt 3.3.6.
Wie auf Seite 22 erläutert, musste auch der Aspekt der benötigten Rechenzeit beachtet werden. Dies stellte uns vor die Aufgabe, aus allen Alignments eine repräsentative Auswahl zu treffen. Dabei wurden drei Ansätze gemacht (vgl. Abschnitt 3.4 ab S. 29), deren Vorgehen und Ergebnisse im Folgenden diskutiert werden.

5.1 Diskussion des ersten Ansatzes

Der erste Ansatz sollte eine möglichst geringe Rechenzeit beanspruchen. Aus diesem Grund wählten wir nur 1000 aller möglichen Alignments aus. Um den enstehenden Datensatz repräsentativ zu gestalten, geschah dies zufällig. Einschränkend wurde nur beachtet, dass die Anzahl der Alignments mit einem SCOP-Score von Null, derjenigen der Alignments mit einem SCOP-Score von Eins glich (bzgl. SCOP-Score s. S. 28). Dies sollte dem Netzwerk die Möglichkeit zu geben, die Zusammenhänge zwischen den Basisdaten (Abschnitt 3.3) und beiden SCOP-Scores gleichermaßen lernen zu können. Die Ergebnisse zeigten hier sehr schnell, dass zwar eine geringe Rechenzeit erreicht wurde, die Voraussagen jedoch nicht zu gebrauchen waren.

Der Matthews Koeffizient (vgl. S. 34) kann zwischen -1 und +1 liegen. Ein

Unterschied zwischen zwei Koeffizienten von 0.3 , wie er in den Ergebnissen dieses Ansatzes gefunden wurde (s. S. 38), entspricht somit 15 % der möglichen Bandbreite. In einem Cross Validation Test (s. Pkt. 3.6.4) werden sehr ähnliche Werte erwartet. Die hier beobachtete Schwankung ist eindeutig zu groß, um ignoriert zu werden. Da solch starke Abweichungen meist Folge eines Übertrainings sind (vgl. Pkt. 2.3.2), wurde der erste Ansatz abgebrochen und direkt zum zweiten Ansatz übergegangen.

5.2 Diskussion des zweiten Ansatzes

Um dem Problem des Übertrainings, wie es im ersten Ansatz auftauchte, entgegenzutreten, haben wir hier wesentlich mehr Daten verwendet. Wie unter Punkt 3.4.2 erwähnt, existierten nur 19 451 Alignments mit gleicher Faltung der entsprechenden Proteine, deren Daten alle verwendet wurden. Analog zum vorigen Ansatz zogen wir die Daten gleich vieler Alignments mit einem SCOP-Score von Null (vgl. Pkt. 3.3.6) hinzu. Auch hier war der Hauptgedanke, dem neuronalen Netz die gleiche Möglichkeit zu bieten, die Zusammenhänge zwischen Eingabedaten und einem SCOP-Score von Eins bzw. Null zu lernen.
Die Auswahl der mit einem SCOP-Score von Null bewerteten Abgleiche wurde nach dem Kriterium eines möglichst hohen Alignment-Scores (vgl. Pkt. 3.4.2) getroffen. Dieses sollte eher auf eine ähnliche Faltung der alignierten Proteine hindeuten. Der Gedanke war also, einen möglichst schwierig prognostizierbaren Datensatz zu erstellen. Nach dem Training eines neuronalen Netzes mit diesen Daten sollte der SCOP-Score realer einfacher Alignments sehr gut und jener schwierigerer Alignments zumindest zufriedenstellend voraussagbar sein. Die in Abschnitt 4.2 dargestellten Ergebnisse führten jedoch zu einem unerwarteten Problem.

Die berechneten Matthews Koeffizienten (Gl. (7), S. 34) im zweiten Ansatz waren ausnahmslos ungewöhnlich hoch. Weiterhin konnte kein signifikanter Unterschied zwischen den einzelnen Sets beobachtet werden. Diese beiden

Feststellungen führten zu einer genaueren Betrachtung der Ergebnisse. In der Auftragung der Prognosen über dem Alignment-Score (Abb. 8, S. 39) ist zu sehen, dass Vorhersagen < 0.5 nur über einem kleinen Bereich der Abszisse vorkommen. Das ist leicht verständlich, wenn man das Auswahlkriterium für die Abgleiche mit einem SCOP-Score von Null bedenkt (s. o.). So beschränkt sich der Alignment-Score in diesen Fällen auf den Ausschnitt zwischen etwa 450 und 700. Dieses ist auch der einzige Bereich in dem das neuronale Netzwerk falsche Voraussagen machen kann, welche meist als *false negative* (s. Pkt. 3.6.5) auftreten. Alignment-Scores außerhalb dieses Bereiches deuten eindeutig auf einen SCOP-Score von Eins hin.
Die in diesem Ansatz verwendete Methode war offensichtlich zu weit von einer realistischen Verteilung (vgl. a. Abb. 14 u. 15 auf S. 47) entfernt, um noch brauchbare Ergebnisse zu liefern. Die Matthews Koeffizienten waren nur deshalb so hoch, weil die Verteilung der Daten in den Testsets während der Cross Validation derjenigen der Trainingssets entsprach (bzgl. Cross Validation s. Pkt 3.6.4). Das Problem soll verdeutlicht werden: zwei Proteine, die keinerlei Ähnlichkeit in der Faltung aufweisen, würden höchstwahrscheinlich mit sehr niedriger Score aligniert werden. Nach dem Training mit den Daten dieses Ansatzes, würde ein neuronales Netzwerk jenes Proteinpaar aber mit einem SCOP-Score von Eins bewerten, da es eine Kausalität zwischen einem niedrigen Alignment-Score und einem SCOP-Score von Eins gelernt hat. Dies wird durch Abbildung 8 auf S. 39 bestätigt. Prognosen mit Proteinsätzen, deren Daten einer realen Verteilung entsprächen, würden also nur mit unbefriedigender Qualität ablaufen. Aus diesem Grund wurde auch der zweite Ansatz abgebrochen.

5.3 Diskussion des dritten Ansatzes

5.3.1 Set 1 bis Set 5

In diesem Ansatz sollten die zuvor aufgetretenen Probleme behoben werden. Aus diesem Grund wurden, analog dem zweiten Ansatz, alle 19451 Align-

ments mit einem SCOP-Score (s. Pkt. 3.3.6) von Eins verwendet. Weiterhin benutzten wir zum Aufbau der Sets 1 bis 5 (vgl. Pkt. 3.5.1) die Daten der doppelten Anzahl zufällig ausgewählter Alignments von Proteinen mit verschiedener SCOP-Einteilung.
Die Auswahl wurde zufällig getroffen, um dem im vorigen Abschnitt beschriebenen Problem entgegenzuwirken. Wir erhöhten die Menge der Daten, um dem Verhältnis von Alignments, mit einem SCOP-Score von Null zu einem solchen von Eins, wie es im Haupt-Proteinsatz (s. Pkt. 3.1.1) vorlag[6], näherzukommen. Dabei musste erstens noch immer der Aspekt der Rechenzeit bedacht werden (vgl. Pkt. 2.3.3) und zweitens durfte das Verhältnis in den Sets nicht zu stark auf Seite des SCOP-Scores von Null liegen. Dies hätte zu einer Unterdrückung der Zusammenhänge, welche auf gleiche Proteinfaltung hinweisen, während des Lernvorganges führen können. Anders ausgedrückt: bei verhältnismäßig zu wenigen Datenzeilen mit einem SCOP-Score von Eins, könnten diese vom Netzwerk als Hintergrundrauschen interpretiert werden.
Die unter Punkt 4.3.1 gezeigten Ergebnisse unterstützen die Richtigkeit dieser Überlegungen.

Die einführende Abbildung 9 auf Seite 40 zeigt ein typisches Ergebnis der Prognosen des dritten Ansatzes. Die Auftragung entspricht der Erwartung, nach der ein hoher Alignment-Score (vgl. Pkt. 3.3.3) eher auf ähnliche Proteine hindeuten sollte und vice versa. Es ist offenbar gelungen die Fehler des zweiten Ansatzes (vgl. hier Abb. 8, S. 39) zu korrigieren. Trotzdem muss klargestellt werden, dass die Verbindung „hoher Alignment-Score ⇒ SCOP-Score von Eins“ bzw. „niedriger Alignment-Score ⇒ SCOP-Score von Null“ kein absolut kausaler Zusammenhang ist.

Der Vergleich der Matthews Koeffizienten (Gl. (7), S. 34) in Abbildung 10 auf Seite 40 und in Tabelle 4 auf Seite 42, sowie der Spezifitäten und Sensitivitäten in Abbildung 11, Seite 43 begründet die unter Punkt 4.3.1 aufgestellte

[6]Dort galt: $\frac{(SCOP-Score=0)}{(SCOP-Score=1)} \approx \frac{55}{1}$

Abfolge:

$$\text{Set } 2 \approx \text{Set } 1 < \text{Set } 4 < \text{Set } 3 < \text{Set } 5$$

der Sets 1 bis 5 nach der Performancequalität. In beiden Abbildungen und der Tabelle ist klar erkennbar, dass die Verwendung eines TOP-Scores einen deutlich größeren Einfluss als die alleinige Kombination der Alignment-Methoden hat. Um die hier wirkenden Ursächlichkeiten besser zu verstehen, analysierten wir die Eingabedatensätze.

5.3.2 Diskussion der Eingabedatensätze

Die im Folgenden besprochenen Werte der Energien, des Scores und der Länge eines Alignments, der TOP 5- und TOP 80-Wertung sowie des SCOP-Scores sind ab Seite 26 in Abschnitt 3.3 erklärt.

Die analytische Betrachtung der Eingabedaten der Sets 1 bis 5 (s. a. Pkt. 3.5.1) verifiziert, dass die neuronalen Netzwerke die Zusammenhänge zwischen den Input- und Output-Daten gelernt haben.
Die Abbildungen 14 und 15 auf der Seite 47 verdeutlichen den Bezug zwischen einem hohen Alignment-Score und einem SCOP-Score von Eins. Gleiches gilt für die Länge eines lokalen Alignments, aufgetragen in Abbildung 12 auf Seite 46. Dass die Länge eines globalen Abgleiches zweier Proteine keinen Hinweis auf die Ähnlichkeit ihrer Faltungen gibt, liegt daran, dass die globale Eigenschaft ein Alignment über die gesamte Länge der Sequenzen bedeutet. Auch die Betrachtung der Energien in den Abbildungen 16 bis 19 (S. 48 u. 49) geht konform mit der sehr ähnlichen Performance von Set 1 und 2. Ihre Verwendung kann die Anzahl der korrekten Voraussagen nur unbedeutend erhöhen, da keine ausgeprägte Kausalität beobachtbar ist.
Die merklich besseren Vorhersagen der Datensätze 3 und 5 werden durch die Abbildungen 20, 21 und 23 auf den Seiten 50 und 52 erklärt. Das neuronale Netz lernt die starke Assoziation zwischen einem hohen TOP-Score und der gleichen Faltung der Proteine. Offenbar werden, gemäß der Eindeutigkeit dieser Verflechtung, die entsprechenden Synapsen im Netzwerk stärker

stimuliert.
Die reine Kombination der Alignmentdaten aller acht Methoden, ohne die Verwendung eines TOP-Scores, wurde in Set 4 getestet (vgl. 3.5.1). Die Ergebnisse haben gezeigt, dass diese Vorgehensweise eine zwar merkbare, aber dennoch geringe Verbesserung im Matthews Koeffizienten brachte. Vergleiche hierzu Tabelle 4 auf Seite 42. Die Abbildung 22 (S. 51) soll die in diesem Fall wirkenden Zusammenhänge verdeutlichen. Wie schon bei der Präsentation der Ergebnisse (vgl. hier S. 45) aufgezeigt wurde, liegt der Schnittpunkt der zwei dargestellten Kurven bei einem Alignment-Score von Null. Letzterer sinkt entlang der Graphen von links nach rechts. Mit diesen Überlegungen ist erkennbar, dass ein positiver, mittlerer Alignment-Score in Set 4 etwas stärker mit einem SCOP-Score von Eins verbunden ist, als eine positive Bewertung der Alignments in Set 2.
Man bedenke, dass die angewendete Darstellungsform die Interaktion der Synapsen und Neuronen im Netzwerk nur unzureichend darzustellen vermag. Die Annahme, das neuronale Netz verwende den Mittelwert der Alignment-Scores aller kombinierten Methoden, ist grob vereinfachend und wird hier nur gemacht, um eine graphische Darstellung zu ermöglichen.

Die Analyse der Ergebnisse und Eingabedaten zeigt, dass die verwendeten Netzwerke, nach der genannten Reihenfolge (s. S. 41 o. 61), einen größer werdenden Prozentsatz der Daten richtig voraussagen. Vergleicht man aber die Entwicklung der Sets 1 bis 5 mit jener, der Sets 1.1 bis 4.1 und besonders 1.2 bis 5.2, wird klar, dass dieses Ergebnis kritisch zu betrachten ist.

5.3.3 Set 1.1 bis 4.1

Mit Hilfe der Datensätze 1.1 bis 4.1 sollten die Vorhersagen, der mit den Sets 1 bis 5 trainierten Netzwerke, überprüft werden. Zu diesem Zweck wurden die gespeicherten Netze verwendet, um die SCOP-Scores der, aus allen 1 086 566 möglichen Alignments aufgebauten, Sets 1.1 bis 4.1 vorauszusagen (s. a. Pkt. 3.5.2). Aus diesen Prognosen wurde pro Task-Protein das Template

mit dem höchsten vorhergesagten Wert als bestmögliche Strukturvorlage interpretiert (vgl. Abschnitt 3.7, S. 36). Es wurde gezählt, wie häufig dieses Kriterium ein Protein als Modell auswählte, welches die gleiche Einteilung in der SCOP-Datenbank (s. Pkt. 2.2.3) besaß wie das Task-Protein.

In Tabelle 5 auf Seite 55 sind die Ergebnisse der besten Alignment-Methoden zusammengefasst. Stellt man diese in eine aufsteigende Reihe, so erhält man:

$$\text{Set } 4.1 \ll \text{Set } 1.1 < \text{Set } 2.1 < \text{Set } 3.1.$$

Auffällig ist die Stellung des kombinatorischen Sets 4 als schlechtest prognostizierbarer Datensatz. Auch ist aus der Tabelle ersichtlich, dass der Vorsprung von Set 3 nicht so groß ist, wie aufgrund der Ergebnisse der Sets 1 bis 5 erwartet. Dort stieg die Qualität der Performance bei Verwendung des TOP 5-Score in Set 3 sprunghaft an (vgl. Pkt. 4.3.1, insb. Tab. 4, S. 42). Diese Resultate weisen schon darauf hin, dass eine Übertragung, der auf dem Matthews Koeffizienten beruhenden Reihenfolge der Sets 1 bis 5 auf die Ergebnisse nach Anwendung des oben beschriebenen Auswahlkriteriums für das beste Modell, nicht ohne weiteres möglich ist. Diesbezüglich siehe auch Punkt 5.3.5.

5.3.4 Set 1.2 bis 5.2

Die Vorhersagen der Sets 1.2 bis 5.2 stellten den Test gegen einen unabhängigen, zu dem Zeitpunkt (Juli 2001) nicht SCOP-klassifizierten Datensatz dar. Zur Auswahl der Proteine siehe Punkt 3.1.2. Wie auch für die Sets 1.1 bis 4.1 wurde pro Task-Sequenz das Template mit dem höchsten vorhergesagten Wert als bestmögliches Modell angesehen. Ein dem Vorgehen bei diesen Sets vergleichbarer Test der Richtigkeit der Prognosen konnte, aufgrund der fehlenden SCOP-Einteilung der Proteine, nicht durchgeführt werden. Aus diesem Grund wurden die Ergebnisse der Sets 1.2 bis 5.2 mittels des LGScore (s. Pkt. 3.7.3, S. 37) evaluiert. Er ist ein Maß für die Ähnlichkeit zweier Proteinstrukturen.

In Tabelle 6 auf Seite 56 sind die Resultate der besten Methoden dargestellt. Dabei wurde, wie auf Seite 37 gesagt auf früheren Erfahrungen beruhend, ein LGScore > 3 als Indikator für eine zufriedenstellende Qualität des Modells betrachtet. Das erlaubt, hier folgende qualitative Reihe aufzustellen:

$$\text{Set } 5.2 \ll \text{Set } 3.2 < \text{Set } 4.2 < \text{Set } 2.2 \approx \text{Set } 1.2.$$

Diese Abfolge ist entgegengesetzt derjenigen der Sets 1 bis 5 ! Set 5.2 zeigt eine extrem schlechte Performance und keine der getesteten Methoden erreicht die Qualität von PDB-Blast [7]. Auch dies demonstriert die Unvereinbarkeit zwischen den Ergebnissen der Sets 1 bis 5 und der angewendeten Auswahlmethode des vermeintlich besten Modells (vgl. Pkt. 5.3.3).

5.3.5 Vergleich aller Ergebnisse

Vergleicht man die Ergebnisse der Sets 1 bis 5 mit jenen der Sets 1.1 bis 4.1 und besonders 1.2 bis 5.2, so muss festgestellt werden, dass die Abfolge der Datensätze 1 bis 5 nicht das gewünschte Resultat beschreibt.
Die histographischen Abbildungen 24 und 25 (S. 55 u. 56) demonstrieren, dass die Netzwerke gelernt haben nicht nur den Alignment-Score zu beachten, sondern auch die anderen verwendeten Daten in die Voraussagen einzubeziehen. Die Matthews Koeffizienten (Tab. 4, S. 42; Def. Gl. (7), S. 34) validieren, dass die Netze gelernt haben entsprechend der Zusammenhänge in den Eingabedatensätzen (s. Pkt. 5.3.2) zu prognostizieren. Folglich muss der offensichtlich vorhandene Fehler im Aufbau der Datensätze bzw. beim Auswahlkriterium des bestmöglichen Modells liegen (s. Pkt. 3.5 bzw. 3.7).

Laut den Matthews Koeffizienten haben die verwendeten neuronalen Netze, von Set 1 zu Set 5 hin, den Prozentsatz der korrekten Vorhersagen erhöht. Anders ausgedrückt: die mit den „besseren“ Methoden durchgeführten Vorhersagen sind mit größerer Warscheinlichkeit richtig. Es ist also gelungen die Qualität der Vorhersage, ob *ein* Task- und *ein* Template-Protein die gleiche SCOP-Einteilung bzgl. Class und Fold besitzen, zu verbessern! (Bzgl. SCOP

s. Pkt. 2.2.3)
Auf die Auswahl eines Modells für ein Protein unbekannter Struktur bezogen bedeutet dies: es werden mehr mögliche Modelle erkannt. Die unter Punkt 3.6.5 beschriebenen Parameter Spezifität und Sensitivität erreichen also höhere Werte. Doch die zusätzlich detektierten Modelle beinhalten auch Polypeptide, die trotz gleicher SCOP-Einteilung keine hinreichend ähnliche Struktur aufweisen.
Die SCOP-Klassifizierung eines unbekannten Proteins vorherzusagen bedeutet einen Hinweis darauf zu erhalten, welches Protein das *beste* Modell ist, aber nicht möglichst viele *mögliche* Modelle zu finden!
Eine verbesserte Detektion des bestmöglichen Modells konnte in dieser Arbeit nicht erreicht werden. Damit wurde auch die gestellte Aufgabe, die SCOP-Klassifizierung eines Proteins vorauszusagen und die Qualität dieser Prognose als Qualität der möglichen Modelle zu interpretieren, nicht erfüllt.

Zusammengefasst repräsentiert die Steigerung der Performance keine Verbesserung in der Vorhersage der SCOP-Klassifizierung eines Task-Proteins, sondern nur eine mit höherer Wahrscheinlichkeit richtige Vorhersage bezüglich *eines alignten Proteinpaares*. Aus diesem Grund lässt sich die beobachtete qualitative Reihenfolge der Sets 1 bis 5 nicht zwischen den Sets 1.1 bis 4.1 bzw. 1.2 bis 5.2 wiederfinden. So ist zwar anzunehmen, dass das beste Modell die gleiche Einteilung in der SCOP-Datenbank besitzt wie das Task-Protein, aber es ist nicht anzunehmen, dass es auch als mit höchster Wahrscheinlichkeit die gleiche SCOP-Einteilung besitzend vorausgesagt wird.

5.4 Weiterführende Ansätze

Wie in der Einleitung gesagt, sollte diese Arbeit eine Umgestaltung und Weiterführung der Arbeit von Jesper Lundström [6] darstellen. Ein sehr ähnlicher Ansatz wurde bei der Entwicklung des Web-Servers GenTHREADER [10] gemacht. Die dort von Jones verfolgten Ideen wurden auch in dieser Arbeit eingeflochten.

Der Name GenTHREADER stammt von der in [50] entwickelten Methode ab. Es wurde mit einem Sequenz-Alignment begonnen. Zusätzlich zu den daraus gewonnenen Daten wurde das Programm THREADER verwendet, welches der Autor schon früher entwickelte [51]. Es berechnet einfache Potentiale durch die Verbindung des Sequenz-Alignments mit dem implizierten Strukturmodell. Dieses Verbinden wird im Englischen als „threading“ bezeichnet. Die Kombination der Potentiale mit den Daten aus dem Alignment wurde verwendet, um ein neuronales Netz zu trainieren. Die drei Ebenen *Class*, *Architecture* und *Topology*, der unter Punkt 2.2.2 beschriebenen Datenbank *CATH*, wurden von Jones als Kriterium struktureller Ähnlickeit verwendet. Das vorangestellte „Gen“ stammt von der Überlegung, dass die gesamte Methode allein von genetischen Proteine kodierenden Sequenzen ausgehend arbeiten kann.
Mit diesem Ansatz gelang der arbeitenden Gruppe bei einem Test, mit einem unabhängigen von Fischer *et al.* vorgeschlagenem Set [52], eine zu 73,5 % richtige Vorhersage. Das beste Ergebnis im gleichen Test, mit einer etwas anderen Methode, erzielten Fischer und Eisenberg [53] mit 76.5 %.

In Anbetracht der von Jones verwendeten Methode und der recht guten Resultate die er erreichte, sollte die Vorgehensweise der hier vorliegenden Arbeit überdacht werden.
Zu Beginn des Vergleichs aller Ergebnisse (S. 64) wurde dargelegt, dass es gelungen ist, den neuronalen Netzwerken die Zusammenhänge zwischen den Eingabe- und Ausgabedaten beizubringen. Dementsprechend erscheint der Gedanke, mit anderen Eingabedaten als den in Abschnitt 3.3 beschriebenen Basisdaten zu arbeiten, nicht mehr Erfolg versprechend.
Sinnvoller sollte eine Änderung des Kriteriums für ein gutes Modell sein. Hier wäre denkbar, die Informationen einer anderen Datenbank (z.B. CATH) zu verwenden. Es könnten auch die anderen Ebenen der SCOP-Datenbank als Kriterium für einen positiven SCOP-Score mit einbezogen werden. Namentlich wären dies die Familien und Superfamilien. Eigentlich sollten gleiche

Einteilungen auf diesen Ebenen aufgrund der hierarchischen Struktur der Datenbank (vgl. Pkt. 2.2.3) durch gleiche Faltung und Klasse impliziert sein. Durchgeführte Stichproben zeigten aber, dass dies nicht immer zutrifft.
Vielversprechend ist die Überlegung den LGScore als Ähnlichkeitskriterium einzusetzen. Er evaluiert klar die strukturelle Gleichartigkeit zwischen Proteinen. Nachteilig wäre in diesem Fall aber die nötige Berechnung dieses Scores für alle alignierten Proteinpaare. Sie kann je nach Größe des Datensatzes sehr zeitaufwändig werden. Spielt der Zeitfaktor keine Rolle, kann natürlich ein komplizierteres, automatisches Modelling durchgeführt werden. Damit ließen sich auch andere Bewertungen der Ähnlichkeit berechnen und als Referenz für das Ausgabeneuron der Netzwerke verwenden. Eine Möglichkeit wäre die auf Seite 13 erläuterte *Root Mean Square Deviation*.

Auf anderer Ebene muss der Matthews Koeffizient (s. S. 34) als Hilfe zur Evaluierung der Performance überdacht werden, denn die nach ihm aufgestellte Reihenfolge, in der sich die Qualität der Vorhersagen verbesserte (vgl. S. 61), konnte nicht beibehalten werden. Grundsätzlich beruht der Matthews Koeffizient auf dem Verhältnis richtiger zu falschen Prognosen. Es stellte sich aber in dieser Arbeit heraus, dass dies kein hinreichendes Kriterium ist (s. z.B. S. 65). Es ist nicht einmal sicher, ob es sich um ein notwendiges Kriterium handelt, denn prinzipiell würde die korrekte Vorhersage des besten Modells ausreichend sein. Dementsprechend sollte versucht werden, eine Bewertung zu optimieren, welche sich auf die richtig erkannten *besten* Modelle bezieht. Möglich ist hier z.B. die Berechnung der Sensitivität *nach* deren Auswahl.

6 Zusammenfassung und Ausblick

Ziel der Arbeit war es, Methoden zu entwickeln um die SCOP-Klassifizierung von Proteinen vorherzusagen. Weiterhin sollte die Eindeutigkeit dieser Vorhersagen als Basis für die Auswahl eines Templates zum Modellieren der Proteinstruktur verwendet werden.
Es wurden alle Proteine eines ausgewählten, unter SCOP (s. Pkt. 2.2.3) klassifizierten Proteinsatzes nach acht unterschiedlichen Methoden (Tab. 2, S. 25) gegeneinander aligniert. Aus diesen Alignments wurden verschiedene Basisdaten gewonnen, namentlich die Länge des Alignments, eine Bewertung, eine auf den C_α-Atomen beruhende Energie, eine totale Energie und zwei TOP-Scores. Als vorherzusagenden Wert verwendeten wir den SCOP-Score, welcher angibt ob zwei Proteine in der SCOP Datenbank bzgl. Faltung und Klasse die gleiche Einteilung besitzen. Diese Daten sind detailliert in Abschnitt 3.3 erläutert.
Es wurden fünf verschiedene Methoden angewendet die Eingabedaten zu kombinieren (vgl. Abschnitt 3.5). Als Werzeug zur Vorhersage bedienten wir uns zweilagiger neuronaler Netzwerke, die in ihrem Aufbau den jeweiligen Datensets angepasst und optimiert wurden (s. Pkt. 3.6.2). Als vorwiegendes Mittel zur Evaluierung der Vorhersagen wurde der Matthews Koeffizient (Gl. (7), S. 34) verwendet.

Es zeigte sich eine klare Reihenfolge, in welcher sich die Qualität der Vohersage-Performance laut dem Matthews Koeffizienten verbesserte. Der Test gegen einen unabhängigen und zu diesem Zeitpunkt (Juli 2001) noch nicht SCOP-klassifizierten Proteinsatz sollte dieses Ergebnis validieren (vgl. Pkt. 3.1.2). Dies ergab jedoch eine der zuvor beobachteten Reihenfolge entgegengesetzte Abfolge. Unter Zuhilfenahme einer Analyse der Eingabedaten konnte aber auch bestätigt werden, dass die neuronalen Netze sowohl alle verfügbaren Trainingsdaten verwendeten als auch den Zusammenhang zwischen Input- und Output-Daten gelernt hatten. Aus diesem scheinbaren Widerspruch folgende Überlegungen führten zu dem Fazit, dass es uns gelungen war die

Vorhersage, ob zwei abgeglichene Proteine die gleiche Faltung besitzen, zu verbessern. Es war damit aber nicht gelungen eine bessere Voraussage der SCOP-Klassifizierung eines Task-Proteins zu erreichen. Dies würde verlangen, die Qualität der Auswahl des besten Template-Proteins zu optimieren. Die Verwendung der Eindeutigkeit der hier durchgeführten Prognosen als Auswahlkriterium ließ sich nicht befriedigend resümieren.
Insgesamt muss unerfreulicherweise gesagt werden, dass es nicht gelungen ist das gestellte Ziel zu erreichen.

Auf den bisher gewonnenen Erkenntnissen beruhende Weiterentwicklungen der verwendeten Methoden wurden, durch die begrenzte Zeit, in der diese Arbeit erstellt werden musste, unterbunden. Denkbar wäre hier der Austausch des SCOP-Scores durch eine Bewertung der Qualität des Templates als Modell. Um kein zu zeitaufwändiges Modelling durchführen zu müssen, könnte der LGScore (s. Pkt. 3.7.3) als eine mögliche Variante in Betracht gezogen werden. Ebenfalls ist zu überdenken, ob der Matthews Koeffizient durch ein anderes Evaluierungsmittel ersetzt wird, da seine Aussage sich prinzipiell auf das Verhältnis richtiger zu falschen Vorhersagen bezieht. Wünschenswert wäre hier eine Bewertung, welche direkt mit der Vorhersage der besten Templates zusammenhängt.

Literatur

[1] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223 – 230, Juli 1973.

[2] P. Baldi und S. Brunak. *Bioinformatics,the Machine Learning Approach.* The MIT Press, 2 edition, 1998.

[3] D. J. Osguthorpe. Ab initio protein folding. *Current Opinion in Structural Biology*, 10:146 – 152, 2000.

[4] D. J. Osguthorpe. Improved ab initio predictions with a simplified, flexible geometry model. *Proteins: Structure, Function, and Genetics Suppl*, 3:186 – 193, 1999.

[5] B. Al-Lazikani, J. Jung, Z. Xiang und B. Honig. Protein structure prediction. *Current Opinion in Structural Biology*, 5:51 – 56, Februar 2001.

[6] Jesper Lundström. Pcons: A consensus approach to protein fold recognition, Januar 2001. Diplomarbeit.

[7] PDB-Blast Website. http://bioinformatics.ljcrf.edu/pdb_blast/.

[8] FFAS Website. http://bioinformatics.burnham-inst.org/FFAS/.

[9] 3D-PSSM Website. http://www.bmm.icnet.uk/servers/3dpssm/.

[10] GenTHREADER Website. http://insulin.brunel.ac.uk/psipred/.

[11] Sam-T98 Website. http://www.cse.ucsc.edu/research/compbio/HMM-apps/T98-query.html.

[12] Inbgu Website. http://www.cs.bgu.ac.il/~bioinbgu/form.html.

[13] FSSP Website. http://www.ebi.ac.uk/dali/fssp/fssp.html.

[14] CATH Website. http://www.biochem.ucl.ac.uk/bsm/cath_new/.

[15] SCOP Website. http://scop.mrc-lmb.cam.ac.uk/scop/.

[16] Ken A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133 – 7155, August 1990.

[17] Peter W. Atkins. *Physikalische Chemie*. VCH Verlagsgesellschaft mbH, 2nd edition, 1996. S. 720 ff.

[18] D. Voet und J. G. Voet. *Biochemie*. VCH Verlagsgesellschaft mbH, 1992. S. 173 ff.

[19] S. Henikoff und J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89(22):10915 – 10919, November 1992.

[20] T. N. Petersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G. P. Gippert und O. Lund. Prediction of protein secondary structure at 80 % accuracy. *Proteins*, 41(1), Oktober 2000.

[21] S. Cristóbal, A. Zemla, D. Fischer, L. Rychlewski und A. Elofsson. A study of quality measures for protein threading models. *BMC Bioinformatics 2001*, 2(5), August 2001.

[22] Michael Levitt und Mark Gerstein. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. USA*, 95:5913 – 5920, Mai 1998.

[23] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov und P. E. Bourne. The protein data bank. *Nucleid Acid Research*, 28(1):235 – 242, 2000.

[24] PDB Website. http://www.rcsb.org/pdb/.

[25] L. Holm und C. Sander. Mapping the protein universe. *Science*, 273:595 – 602, 1996.

[26] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells und J. M. Thornton. Cath- a hierarchic classification of protein domain structures. *Structure*, 5(8):1093 – 1108, 1997.

[27] F. M. G. Pearl, D. Lee, J. E. Bray, I. Sillitoe, A. E. Todd, A. P. Harrison, J. M. Thornton und C. A. Orengo. Assigning genomic sequences to cath. *Nucleic Acids Research*, 28(1):277 – 282, 2000.

[28] CATH-Info Website. http://www.biochem.ucl.ac.uk/bsm/cath_new/cath_info.html.

[29] A. G. Murzin, S. E. Brenner, T. Hubbard und C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536 – 540, 1995.

[30] L. Lo Conte, A. Ailey, T. J. P. Hubbard, S. E. Brenner, A. G. Murzin und C. Chothia. Scop: a structural classification of proteins database. *Nucleic Acids Research*, 28(1):257 – 259, 2000.

[31] C. W. Hogue, H. Ohkawa und S. H. Bryant. A dynamic look at structures: www-entrez and the molecular modeling database. *Trends Biochem. Sci.*, 21(6):226 – 229, Juni 1996.

[32] R. Sowdhamini, S. D. Rufino und T. L. Blundell. A database of globular protein structural domains: clustering of representative family members into similar folds. *Fold Des.*, 1(3):209 – 220, 1996.

[33] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. OXFORD University Press, 1995.

[34] David A. Medler. A brief history of connectionism. *Neural Computing Surveys*, 1:61 – 101, 1998. http://www.icsi.berkeley.edu/~jagota/NCS.

[35] H. White. Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3(5):535 – 549, 1990.

[36] A. R. Gallant und H. White. On learning the derivatives of an unknown mapping with multilayer feedforward networks. *Neural Networks*, 5(1):129 – 138, 1992.

[37] D. J. Livingstone, D. T. Manallack und I. V. Tetko. Data modelling with neural networks: Advantages and limitations. *Journal of Computer-Aided Moleular Design*, 11:135 – 142, 1997.

[38] Steve Lawrence, C. Lee Giles und A.C. Tsoi. Lessons in neural network training: Overfitting may be harder than expected. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI-97*, pages 540 – 545. AAAI Press, Menlo Park, California, 1997.

[39] LiveBench-2 Website. http://bioinfo.pl/livebench/2/.

[40] J. M. Bujnicki, A. Elofsson, D. Fischer und L. Rychlewski. Livebench-2: large-scale automated evaluation of protein structure prediction servers. *Eingereicht*, 2001.

[41] Arne Elofsson. A study on how to best align protein sequences. *Eingereicht*, Juli 2000.

[42] Kontakt für weitere Informationen zu *Palign*: Arne Elofsson, *Stockholm Bioinformatics Center*. Email: arne@sbc.su.se.

[43] B. H. Park, E. S. Huang und M. Levitt. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.*, 266:831 – 846, 1997.

[44] Britt Park und Michael Levitt. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.*, 266:831 – 846, 1997.

[45] D. A. Hinds und M. Levitt. A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci.*, 89:2536 – 2540, 1992.

[46] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller und D. J. Lipman. Gapped blast and psi-blast: a new generation of protein databse search programs. *Nucleic Acids Res.*, 25(17):3389 – 3402, September 1997.

[47] M. Gribskov, R. Luthy und D. Eisenberg. Profile analysis. *Methods Enzymol.*, 183:146 – 159, 1990.

[48] Verfügbar unter: http://www.ncrg.aston.ac.uk/netlab/index.html .

[49] Informationen bei: http://www.mathworks.com .

[50] David T. Jones. Genthreader: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, 287:797 – 815, 1999.

[51] D. T. Jones, W. R. Taylor und J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358:86 – 89, 1992.

[52] D. Fischer und D. Eisenberg. Protein fold recognition using sequence-derived predictions. *Protein Science*, 5:947 – 955, 1996.

[53] D. Fischer und D. Eisenberg. Assigning folds to the proteins encoded by the genome of mycoplasma genitalium. *Proc. Natl. Acad. Sci. USA*, 94:11929 – 11934, 1997.