PresCont: Vorhersage von Protein-Protein Interaktionsflächen unter Verwendung struktureller und evolutionärer Eigenschaften

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.) DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN DER UNIVERSITÄT REGENSBURG

vorgelegt von

Hermann Josef Zellner

aus Straubing

im Jahr 2011

Das Promotionsgesuch wurde eingereicht am: 08. März 2011

Kolloquium fand statt am: 14. April 2011

Die Arbeit wurde angeleitet von: PD Dr. Rainer Merkl

Prüfungsausschuss:

Vorsitzender: Prof. Dr. Reinhard Wirth

Erstgutachter: PD Dr. Rainer Merkl

Zweitgutachter: apl. Prof. Dr. Wolfram Gronwald

Drittprüfer: Prof. Dr. Reinhard Sterner

Die vorliegende Arbeit wurde in der Zeit von April 2007 bis März 2011 am Lehrstuhl Biochemie II des Institutes für Biophysik und physikalische Biochemie der Fakultät für Biologie und vorklinische Medizin der Universität Regensburg unter Leitung von Herrn PD Dr. Rainer Merkl angefertigt.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungen

Å	Angström, 10^{-10} m
Ala	Alanin
Asn	Asparagin
Asp	Aspartat
Arg	Arginin
CRS	Gemeinsame reduzierte Oberfläche (Common Reduced Surface)
Cys	Cystein
DCLM	Dicubic Lattice Method
FPR	False Positive Rate
Gln	Glutamin
Glu	Glutamat
His	Histidin
HPA	Hydrophobic Patch Analyzer
Ile	Isoleuzin
Leu	Leuzin
LOR	Chancenquotient (Log Odds Ratio)
Lys	Lysin
MCC	Matthews Korrelationskoeffizient
Met	Methionin
MSA	Multiples Sequenzalignment
Phe	Phenylalanin
PIA	Protein Interface Analyzer
PIAS	Protein Interface Analyzer Schale
PPI	Protein Protein Interaktionen
PPK	Protein Protein Kontaktfläche
Pro	Prolin
PROC	Precision Recall Operating Characteristic
ROC	Receiver Operating Characteristic
RS	Reduzierte Oberfläche (Reduced Surface)
rSASA	relative Solvent Accessible Surface Area
SASA	Solvent Accessible Surface Area

Ser	Serin
SVM	Support Vektor Maschine
Thr	Threonin
TPR	True Positive Rate
Trp	Tryptophan
Tyr	Tyrosin
Val	Valin
vgl	vergleiche
VSO	Voronoi Shelling Order

1 Kurzfassung

Protein-Protein Interaktionen spielen eine essentielle Rolle für jeden lebenden Organismus. Sie sind bei der Aktivierung von Enzymen ebenso wichtig wie für die Signalübertragung und Transportvorgänge. Deswegen sind ca. 80% aller Proteine in größere Komplexe eingebunden. Für ein detailliertes Verständnis eines Protein-Protein Komplexes muss dessen 3D-Struktur bekannt sein. Experimentelle Methoden zur Bestimmung der Protein 3D-Struktur sind jedoch langwierig und aufwändig. Daher ist es sinnvoll, parallel oder alternativ Computerprogramme zu verwenden, um Strukturvorschläge zu generieren. Dazu gehört als wichtiger Teilaspekt die computergestützte Vorhersage von Protein-Protein Kontaktflächen (PPK).

In dieser Arbeit wurde die Software *PresCont* entwickelt, die anhand von 5 Merkmalen, basierend auf der 3D-Struktur des Monomers und evolutionärer Information aus einem Multiplen Sequenzalignment (MSA) homologer Proteinsequenzen, eine Vorhersage der PPK ableitet. Im Gegensatz zu anderen, etablierten Programmen benutzt *PresCont* lediglich solche Merkmale einer PPK, die einen hohen Beitrag zur Vorhersage leisten und ignoriert Merkmale, die im Vergleich zu anderen Eigenschaften wenig zusätzliche Information liefern. Die fünf, von *PresCont* verwendeten Merkmale sind Exponiertheit der Aminosäureseitenkette, Häufigkeiten von Aminosäurepaaren, Größe und Vorkommen hydrophober Patches, evolutionäre Konserviertheit und Konnektivität, die als Meta-Eigenschaft mehrere intermolekulare Scores zusammenfasst. Die ersten vier Merkmale wurden bereits häufiger zur Vorhersage von PPKs verwendet, die Eigenschaft Konnektivität wurde bisher nicht benutzt. In *PresCont* wird durch die Eigenschaft der Konnektivität ein Score für einzelne Positionen aus dem Vorkommen intermolekularer Kontaktpaare abgeleitet.

Die Klassifikationsleistung von *PresCont* konnte zusätzlich gesteigert werden durch die Mittelung der Signale über die lokale Nachbarschaft einzelner Positionen. Nach Normierung wurden die erwähnten Merkmale unter Verwendung einer *Support Vektor Maschine* (SVM) zu einer aussagekräftigen Vorhersage kombiniert. SVMs haben sich in der Bioinformatik als robuste Klassifikatoren bewährt. Ein wesentlicher Aspekt der Arbeit war es, einen robusten Ansatz zu entwickeln. Daher wurde bewusst die An-

zahl der Merkmale beschränkt und es wurden Signale gemittelt, um das Rauschen zu reduzieren.

Die Klassifikationsleistung von *PresCont* wurde mit der von *Sppider* und *ProMate* verglichen. *Sppider* ist ein Vertreter für Klassifikatoren obligater PPKs, *ProMate* wurde speziell für transiente PPKs entwickelt. Wie zu erwarten, übertrifft die Performanz von *Sppider* und *PresCont* gemessen an einem Datensatz obligater Homodimere diejenige von *ProMate*. Interessanterweise erreicht *PresCont* mit seinem wesentlich einfacheren Aufbau eine sehr ähnliche Vorhersagequalität wie *Sppider*. An einem Datensatz transienter Heterodimere hingegen übertrifft die Qualität der Vorhersage von *ProMate* diejenige von *PresCont* und *Sppider*. Es scheint folglich nicht möglich zu sein, einen Klassifikator zu entwickeln, der sowohl für obligate als auch für transiente Komplexe gleich hohe Klassifikationsleistung erreicht.

Mit dieser Arbeit wurde belegt, dass die Bewertung von fünf aussagekräftigen Merkmalen ausreicht, um mithilfe einer SVM einen leistungsfähigen Klassifikator zu entwickeln. Dieser steht anderen Verfahren, die ebenfalls den Stand der Technik repräsentieren, aber wesentlich mehr Eigenschaften bewerten und eine komplexere Software-Architektur besitzen, in der Klassifikationsleistung nicht nach.

2 Einleitung

Makromolekularen Interaktionen kommt bei der Organisation des Lebens eine bedeutende Rolle zu. Neben einer Vielzahl teils kleinerer organischer Moleküle sind Proteine an makromolekularen Interaktionen beteiligt. Alle natürlich vorkommenden Proteine sind aufgebaut aus 20 verschiedenen Arten von Aminosäuren. Eine Aminosäure wiederum besteht aus einer Aminogruppe (NH_3), die über ein C-Atom, das sogenannte C_α-Atom, mit einer Säuregruppe ($COOH$) kovalent verbunden ist. Daneben ist mit dem C_α-Atom eine weitere Gruppe, die als Rest oder auch als Seitenkette bezeichnet wird, kovalent verknüpft. Während alle 20 Aminosäurearten die Amino-, die Säuregruppe und ein C_α-Atom besitzen, unterscheiden sie sich lediglich anhand ihrer Seitenkette. Diese bestimmt alle physikalisch-chemischen Unterschiede verschiedener Arten von Aminosäuren wie Größe, Hydrophobizität und Ladung. Anhand dieser Unterschiede lassen sich Aminosäuren nach dem *Venn-Diagramm* (siehe Abbildung 2.1) in mehrere teilweise überlappende Gruppen einteilen.

Über die sogenannte *Peptidbindung* zwischen der Säuregruppe der einen und der Aminogruppe einer anderen Aminosäure können zwei Aminosäuren unter Abspaltung von Wasser kovalent miteinander verbunden werden. Auf diese Art lassen sich auch mehrere Aminosäuren zu einer längeren *Peptidkette* miteinander verknüpfen. Falls eine Peptidkette eine Länge von etwa 30 Aminosäuren überschreitet, so nennt man sie ein Protein. Die kovalent miteinander verbundenen Aminogruppen, C_α-Atome und Säuregruppen werden als seine Hauptkette bezeichnet. Anhand der Abfolge der Arten von Seitenketten innerhalb einer Peptidkette definiert sich die Struktur und Funktion eines Proteins eindeutig. Diese Ebene struktureller Organisation eines Proteins wird als die Primärstruktur bezeichnet. Längere Abschnitte einer Hauptkette können sich, aufgrund von Wasserstoffbrücken zwischen den Hauptkettenatomen, zu Sekundärstrukturelementen anordnen. Zu diesen zählt die α-Helix, in der sich die Hauptkette helixförmig windet, und das β-Faltblatt, bei dem zwei Hauptkettenabschnitte sich parallel bzw. antiparallel aneinander anlagern. Daneben existieren ungeordnete Schleifen oder Kehren, die die Richtung der Hauptkette ändern. Die Ebene dieser strukturellen Ordnung bezeichnet man als Sekundärstruktur. Innerhalb eines Proteins ordnen sich mehrere Sekundärstrukturelemente in einer ganz bestimmten Reihenfolge an und bilden verschlungene

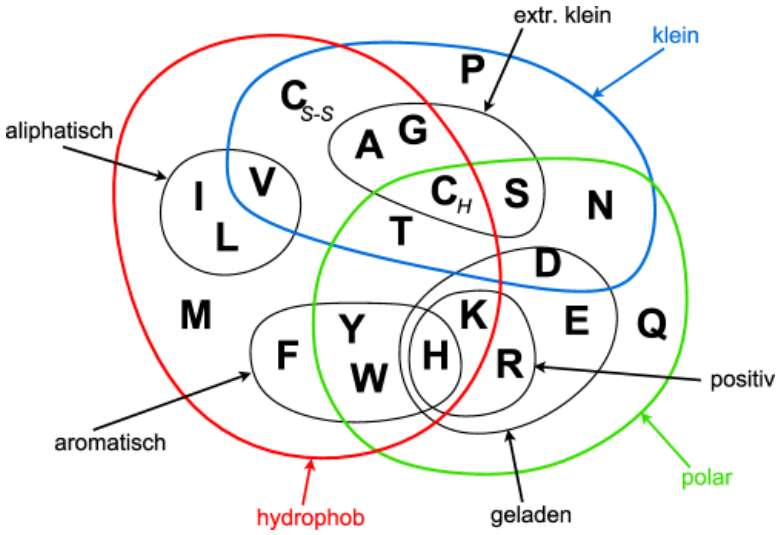


Abbildung 2.1: Venn-Diagramm der 20 natürlich vorkommenden Aminosäuren
Die Aminosäuren wurden anhand ihrer physikalisch-chemischen Eigenschaften gruppiert. Die Aminosäuren sind im Wesentlichen in zwei Gruppen (polar und hydrophob) eingeteilt. Eine dritte Gruppe umfasst die kleinen Aminosäuren. Abbildung aus [1].

teils symmetrische Strukturen. Die Anordnung mehrerer Sekundärstrukturelemente im Raum innerhalb einer Hauptkette bezeichnet man als Tertiärstruktur. Häufig interagieren mehrere Hauptketten miteinander über nicht-kovalente Wechselwirkungen und ordnen sich auf eine bestimmte Art zueinander im Raum an. Diese Anordnung zweier oder mehrerer interagierender Tertiärstrukturen im Raum wird als Quartärstruktur eines Proteins bezeichnet. Die einzelnen Hauptkettenabschnitte, die zu einer Tertiärstruktur gefaltet sind, werden dann Untereinheiten genannt. Bei physikalischen Interaktionen zwischen zwei oder mehreren Untereinheiten sprich man auch von Protein-Protein Interaktionen (PPI).

2.1 Bedeutung von Protein-Protein Interaktionen

PPIs kommt eine Schlüsselposition bei der Organisation des Lebens zu. Ohne PPIs könnten viele zelluläre Prozesse nicht ablaufen. Sie spielen bei der Regulation von enzymatischer Aktivität eine ebenso elementare Rolle wie bei der Funktion des Immunsystems, bei Signaltransduktion, Transport oder Zellbewegungen. Die Bedeutung von PPIs wurde auch im Zuge der jüngsten Genomprojekte deutlich. Die Sequenzierung der Genome höherer Arten wie Mensch, Affen, Maus und anderer Säugetiere hat gezeigt, dass sowohl die Anzahl an Basenpaaren von ca. 3×10^9 als auch die Anzahl der Gene,

die im Falle des menschlichen Genoms auf maximal 30 000 geschätzt wird, in derselben Größenordnung liegt wie die niedriger Arten. Die einzellige Hefe *Saccharomyces cerevisiae* besitzt beispielsweise ca. 5 800 Gene, die Proteine codieren. Die absolute Anzahl an Genen kann folglich nur geringe Aussagekraft über die Komplexität zellulärer Organisation besitzen. Aus diesem Grund wurde die frühere Lehrmeinung, dass ähnliche Moleküle in ähnlicher Weise funktionieren aufgegeben. Die neue Lehrmeinung besagt, dass höhere Arten sich aufgrund ihres komplexeren Netzwerkes von Interaktionen zwischen zellulären Bestandteilen von einfacheren Spezies unterscheiden [1]. So wurden beispielsweise für *S. cerevisiae* 18 000 – 30 000 binäre Interaktionen geschätzt [2], während sich für den Menschen die Anzahl der Interaktionen auf ca. 600 000 beläuft [3].

Diese Befunde belegen, wie wichtig das Verständnis von Protein-Protein Interaktionen ist um Proteinfunktionen zu erforschen und um biologische Systeme besser zu verstehen. Der Kenntnis der Struktur eines Komplexes auf Ebene der Aminosäuren oder gar Atome kommt dabei eine besondere Bedeutung zu. *Röntgenkristallographie* und *NMR* sind ohne Zweifel die präzisesten Techniken zur Bestimmung der 3D-Struktur von Proteinen. Da jedoch die experimentelle Strukturbestimmung von Makromolekülen immer noch teuer und aufwändig ist, ist die Anzahl an bekannten experimentell bestimmten 3D-Strukturen immer noch relativ gering. Folglich werden computergestützte Methoden benötigt, um die räumliche Struktur von Proteinen vorherzusagen.

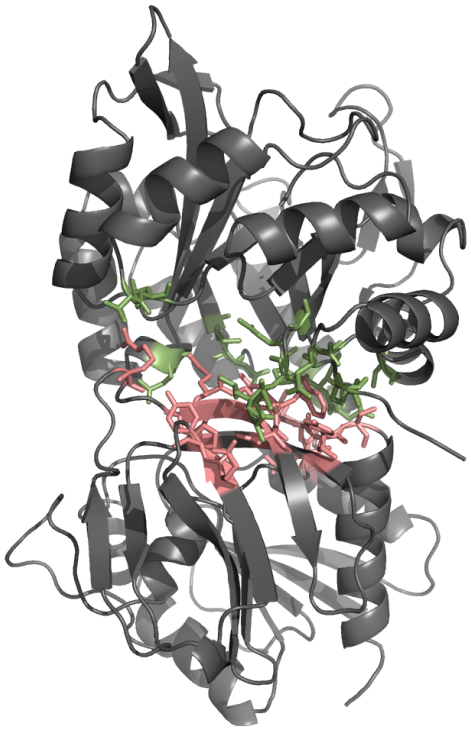


Abbildung 2.2: Der Dimer-Komplex aus *HisF* und *HisH*
Die Abbildung zeigt den enzymatischen Komplex aus *HisF* (oben) und *HisH* (unten). Die Reste beider Moleküle, die an der Interaktionsfläche liegen, sind in der Stäbchen-Darstellung grün bzw. rötlich eingefärbt.

Im Falle von Protein-Protein Komplexen stellt sich häufig das Problem, dass zwar die Struktur der einzelnen Untereinheiten eines Komplexes experimentell bestimmt und oftmals sogar der Interaktionspartner bekannt ist, jedoch keine Information über die Quartärstruktur des Komplexes vorliegt. In so einer Situation kön-

nen computergestützte Methoden helfen, die Interaktionsfläche zu finden, die wiederum Rückschlüsse auf Größe und Eigenschaften des Interaktionspartners erlaubt. Ist zusätzlich der Interaktionspartner bekannt, so können *Docking*-Verfahren genutzt werden um Vorschläge für die Struktur des Komplexes am Computer zu generieren. Da das Docking-Problem jedoch einer Suche in einem 6-dimensionalen Raum entspricht, liegt die Anzahl der zu überprüfenden Konformationen in der Größenordnung von 10^9 [4]. Auch wenn die meisten Implementierungen von Docking-Algorithmen verschiedene Techniken zur Beschleunigung der Suche und zur Verringerung des Rechenaufwandes verwenden, ist Protein-Protein Docking weiterhin ein schwieriges und rechenintensives Unterfangen. Die Kenntnis der Protein-Protein Kontaktfläche (PPK), die von den Aminosäuren gebildet wird, die mit Aminosäuren des Partnermoleküls physikalisch in Kontakt treten (siehe Abbildung 2.2), kann in diesem Fall sehr hilfreich sein. Akkurate computergestützte Vorhersagen von PPKs können daher die Güte von Docking Verfahren in hohem Maße verbessern und die benötigte Rechenzeit massiv verkürzen indem sie den Suchraum enorm einschränken.

Eine weitere Anwendung stellt sich bei der experimentellen Strukturbestimmung eines Komplexes durch *Röntgenkristallographie*. In Proteinkristallen ergeben sich nicht nur natürliche Kontakte zwischen den Untereinheiten, sondern zusätzliche Kristallkontakte. Um künstliche Kristallkontakte von *in vivo* Kontaktflächen unterscheiden zu können, können ebenfalls computergestützte Verfahren zur Vorhersage von PPKs genutzt werden.

2.2 Typen von Protein-Protein Komplexen

So verschieden die Aufgaben sind, die Protein-Protein Interaktionen in der Zelle erfüllen, so divers sind auch die dabei auftretenden Interaktionen. Zunächst lassen sich Protein-Protein Komplexe nach der Anzahl der beteiligten Ketten in binäre Komplexe, und höhere Oligomere, die aus mehr als zwei Untereinheiten aufgebaut sind, einteilen. Zu den binären Komplexen zählen Enzym-Inhibitor Komplexe und Antigen-Antikörperkomplexe. Multimere finden sich beispielsweise als Chaparone oder auch als Virushüllen. Die binären Komplexe können anhand der Sequenz ihrer Untereinheiten weiter unterteilt werden in Homodimere, die aus zwei identischen Untereinheiten aufgebaut sind, und Heterodimere, die aus zwei verschiedenen Untereinheiten bestehen. Weiter kann man Protein-Protein Komplexe unterteilen in obligate Komplexe, die in der Zelle nur im höheren Oligomerisierungszustand vorkommen und nicht-obligate Komplexe, deren Untereinheiten sowohl frei als auch in gebundener Form existieren. Daneben können PPIs anhand ihrer Lebenszeit in permanente und transiente Inter-

aktionen unterschieden werden. Anders als die permanente Interaktion, die sich durch ihre hohe Stabilität und Lebenszeit auszeichnet, sind transiente Interaktionen nicht dauerhaft, so dass Interaktionen zwischen Untereinheiten *in vivo* ständig gelöst und neue gebildet werden. Obligate Interaktionen zählen üblicherweise zu den permanenten Interaktionen, während nicht-obligate Interaktionen sich aufspalten in permanente und transiente Interaktionen [5].

Die Untereinheiten von Homodimeren bilden den Komplex meist bereits während des Faltungsvorganges aus und kommen als Monomer nicht stabil in der Zelle vor [6]. Daher zählen sie meist zu den obligaten Komplexen. Es existieren jedoch auch Homodimere mit schwacher Bindung, deren Untereinheiten in der Zelle frei nachgewiesen wurden [7] [8]. In einigen Fällen finden sich auch Paare orthologer Homooligomere mit unterschiedlichem Oligomerisierungszustand [9]. Daher ist es möglich über die Einführung destabilisierender Mutationen an der PPK manche Homooligomere in einen niedrigeren Oligomerisierungszustand zu zwingen.

Um Bindungseigenschaften von Homodimeren und Heterodimeren zu vergleichen, untersuchten *Noren* und *Thornton* in einer Studie die PPKs von Homodimeren und transienten Heterodimeren [8]. Dabei fanden sie, dass transiente Interaktionsflächen kleiner, planarer und polarer sind als diejenigen von Homodimeren. In einer anderen Arbeit wurde ein Maß für die Unebenheit zur Untersuchung von PPKs verwendet. Dabei stellte sich heraus, dass obligate und nicht-obligate PPKs zwar bezüglich der Größe und Häufigkeit der Unebenheiten ähnlich sind, jedoch bei obligaten Komplexen die Unebenheiten beider beteiligter PPKs stärker miteinander korreliert sind [10]. Außerdem gab es mehrere Arbeiten, in denen mit Hilfe maschineller Lernverfahren obligate von nicht-obligaten Komplexen unterschieden wurden [11] [12] [13] [14]. Dabei wurden Merkmale wie physikalisch-chemische Eigenschaften, atomare Kontaktvektoren, Oberflächenkomplementarität, Größe der PPKs, Aminosäurezusammensetzung und gewichtete Konserviertheit miteinander kombiniert und zu einer Vorhersage des Interaktionstyps verrechnet. In einem Übersichtsartikel [8] unterteilten *Noren* und *Thornton* transiente Komplexe weiter in starke und schwache Oligomere. Während starke Komplexe einen molekularen Trigger benötigen, damit sie binden, kommen schwache Komplexe in der Zelle in einem Gleichgewicht aus niedrigerem und höherem Oligomerisierungszustand vor, in dem kontinuierlich Kontakte auftrennen und neu entstehen. Viele Proteine interagieren auch mit mehr als einem Partner und sind so Teil eines komplexen Interaktionsnetzwerkes, in dem neben obligaten Interaktionen ständig transiente Interaktionen mit unterschiedlichen Interaktionspartnern getrennt und neu gebildet werden.

2.3 Energetische Betrachtungen von Protein-Protein Interaktionen

Trotz dieser Diversität von Protein-Protein Interaktionen sind die treibenden Kräfte stets die gleichen. Während einige Forschungsergebnisse besagen, dass die Aminosäurezusammensetzung der PPK vom Typ der Interaktion abhängt [15][16], fand sich während anderer Arbeiten, dass die Zusammensetzung verschiedener Arten von PPKs recht ähnlich ist [17][18][19][20]. So zeigte sich anhand von Datensätzen aus Homo- und Heterodimeren, dass die Kontakthäufigkeiten bei Homodimeren zwar extremer sind als bei Heterodimeren, die Signale jedoch in dieselbe Richtung tendieren [21].

Um im Einzelnen die Bedeutung einer einzelnen Seitenkette für die energetischen Stabilität einer PPI zu bestimmen, mutiert man die entsprechende Seitenkette zu einem Alanin, der zweitkleinsten Seitenkette, die nur aus einer CH_3-Gruppe besteht. Der Unterschied in der *Gibbsschen Energie* zwischen Wildtyp und Alaninmutante, stellt dann ein Maß dafür dar, wie stark die wildtypische Seitenkette an der Position zur Stabilisierung des Protein-Protein Komplexes beiträgt. Dem liegt jedoch die Annahme zugrunde, dass die gemessene Differenz der *Gibbsschen Energie* aus dem Fehlen der Effekte der wildtypischen Seitenkette resultiert und nicht aus einem Hohlraum, der eventuell durch die Alaninmutation an der Kontaktfläche entsteht. Dieses Risiko ungewollter Nebeneffekte wird dadurch verringert, dass man Alanin als die zweitkleinste Aminosäure wählt anstatt der kleinsten natürlich vorkommenden Aminosäure Glyzin, die den Hauptkettenverlauf extrem beeinflussen würde [22]. Um alle Seitenketten zu finden, die einen hohen Beitrag zur Stabilisierung des Komplexes leisten, wird dieses Vorgehen für jede Seitenkette an der PPK wiederholt. Derartige *Alaninscans* wurden für eine Vielzahl von Komplexen experimentell durchgeführt [23] [16] [24] [25] [26] [27]. Dabei stellte es sich heraus, dass die Stabilisierungsenergie eines Komplexes sehr ungleichmäßig über die PPK verteilt ist. Die meisten PPKs besitzen sogenannte *Hot Spots*, die für einen Großteil der Stabilisierungsenergie verantwortlich sind [28] [16] [29] [30] [31] [32] [33] [34] [35] [36].

Unterschiede in der Aminosäurekomposition wurden nicht nur zwischen verschiedenen Typen von PPKs untersucht, sondern auch zwischen verschiedenen Bereichen der PPK. So teilten *Chakrabarti* und *Janin* PPKs in einen Kern- und einen Randbereich ein und stellten fest, dass der Kernbereich zu einem höheren Anteil aus hydrophoben Aminosäuren besteht als der Rest der Oberfläche [37]. Der Randbereich dagegen besitzt eine ähnlich hydrophile Aminosäurezusammensetzung wie die restliche Oberfläche. Er hat die Aufgabe, wie ein O-Ring [23] dafür zu sorgen, dass kein Wassermolekül den hydrophoben Kernbereich erreichen kann. So kann der Kern von PPKs bei der Kom-

plexbildung vom Wasser abgeschirmt werden und als Ort der hydrophoben *Hot Spots* den Zustand des Komplexes durch den hydrophoben Effekt stabilisieren. Neben der Aminosäurezusammensetzung finden sich auch andere Merkmale, anhand derer sich der Zentralbereich von PPKs von ihrem Rand unterscheidet. So ist der Kernbereich einer PPK stärker konserviert als ihr Randbereich [38].

2.4 Computermethoden

Aus biochemischem Wissen über PPIs ist bekannt, dass Proteine ganz spezifisch an Kontaktflächen binden. Folglich muss es Merkmale geben, anhand derer sich PPKs von der restlichen Oberfläche unterscheiden. Da Bindungsstärke bei PPIs jedoch nur ein Parameter unter vielen ist, sind die Unterschiede zwischen PPKs und der restlichen Oberflächen nur gering und ihre Signale stark verrauscht.

Prinzipiell lassen sich zwei Arten von Programmen zur Vorhersage von PPKs unterscheiden. Zum einen gibt es Methoden, die allein sequenzbasierte Eigenschaften wie Konserviertheit oder die Aminosäurezusammensetzung berücksichtigen [39] [40] [41] [42]. In einer neueren Arbeit wurde gezeigt, dass die Performanz solch einfacher sequenzbasierter Verfahren zur Vorhersage von PPKs stark von der Art der Trainings- und Testdatensätze abhängt und aufgrund der Beschränkung auf sequenzbasierte Eigenschaften die Vorhersagequalität limitiert ist [43].

Zum anderen wird Information aus der Struktur zur Vorhersage von PPKs benutzt. Strukturinformationen werden z.T. auch mit Sequenzinformationen kombiniert um sie zu einer aussagekräftigen Vorhersage der PPK zu verrechnen [44] [45] [46] [47]. Es ist zu erwarten, dass es unter Berücksichtigung mehrerer nichtredundanter Eigenschaften möglich ist, trotz der stark verrauschten Signale eine bessere Vorhersage zu generieren als bei Verwendung nur eines Signals.

In vorliegender Arbeit wird eine Computermethode zur Vorhersage von PPKs entwickelt. Hierbei war es Ziel, mit einer geringen Anzahl an Merkmalen einen Klassifikator hoher Güte zu schaffen. Dabei wurden fünf positionsspezifische Merkmale ausgewählt. Diese Eigenschaften beschreiben die Exponiertheit an der Oberfläche, die Aminosäurezusammensetzung der intramolekularen Nachbarschaft an der Oberfläche, die Zugehörigkeit zu einem hydrophoben Patch, die evolutionäre Konserviertheit und die Anzahl günstiger Wechselwirkungen zum Interaktionspartner. Aus folgenden Gründen wurden diese Parameter ausgewählt:

1. Lösungsmittelzugänglichkeit (*Solvent accessible Surface Area, SASA*) ist eine der wichtigsten Eigenschaften bei der Vorhersage der Kontaktflächen von Homodimeren [15] [48]. *SASA* verbessert aber auch die Performanz von Methoden, die unter Einbeziehung von Information aus der Struktur PPKs von Heterodimeren vorhersagen [46] [47].

2. Die Aminosäurezusammensetzung der PPK unterscheidet sich im Durchschnitt signifikant von derjenigen der restlichen Oberfläche. Dieses Signal kann verstärkt werden, wenn nicht nur die Aminosäuretypen an den einzelnen Positionen, sondern auch an ihren Nachbarpositionen mit berücksichtigt werden [47] [49].

3. Hydrophopbe Positionen treten an der Kontaktfäche weitaus häufiger benachbart als sogenannte *hydrophobe Patches* auf als an der restlichen Oberfläche. So war in einer Studie in 90% der Fälle eines der beiden größten hydrophoben Oberflächenpatches an einer PPI beteiligt [50]. Daher wird die Auswertung hydrophober Patches auch zur Vorhersage von PPKs benutzt [47].

4. Konserviertheit von Positionen gemessen an einem multiplen Sequenzalignment (MSA) ist ein Hinweis auf die Bedeutung der Position für die Funktion das Proteins [51]. Daher können auch Positionen, die zur Stabilität eines Protein-Protein Komplexes beitragen, anhand ihrer höheren Konserviertheit identifiziert werden [52] [53].

5. Seitenketten, die über eine PPK hinweg in Kontakt treten mutieren häufig auf korrelierte Art [54]. Daher stellen korrelierte Mutationen zu mehreren Seitenketten an der Oberfläche des Interaktionspartners ein Signal für die Zugehörigkeit zur PPK dar.

In den folgenden Abschnitten wird auf diese Eigenschaften näher eingegangen.

2.4.1 Exponiertheit an der Oberfläche

Die Stärke der Bindung interagierender Proteine ergibt sich aus den Interaktionen der Aminosäureseitenketten. Die Wechselwirkungsenergie wird erhöht, wenn einzelne Reste aus der Kontaktfläche herausragen. An der restlichen Oberfläche hingegen hätten insbesondere die aromatischen und aliphatischen Seitenketten durch ihren überwiegend hydrophoben Charakter einen negativen Einfluss auf die Proteinstabilität infolge der Wechselwirkungen mit dem Wasser. Aufgrund dieses Beitrags zu Proteinstabilität ist zu

erwarten, dass stark exponierte hydrophobe Aminosäuren gehäuft an Kontaktflächen vorkommen.

Die Exponiertheit einer Aminosäure wird in dieser Arbeit anhand ihres Anteils an der Proteinoberfläche gemessen. Je größer dieser Anteil ist, desto weiter ragt die Seitenkette aus dem Protein heraus.

Meist definiert man die Oberfläche eines Proteins über ein hypothetisches sphärisches Lösungsmittelmolekül, das *in silico* über die Oberfläche des zu untersuchenden Makromoleküls rollt [56]. Während des Rollvorganges wird die Oberfläche des Probenmoleküls tangential zur *van der Waals*-Oberfläche des Makromoleküls gehalten [57]. Es wurden mehrere Verfahren entwickelt um sicherzustellen, dass das Probenmolekül an allen möglichen Kontaktpunkten zu liegen kommt. Dazu wird entweder der Ort, an dem sich die beiden Moleküle berühren (*Connolly Oberfläche*) oder der Mittelpunkt des Probenmoleküls (*SASA*) notiert und zur Oberfläche gezählt (siehe Abbildung 2.3). Die *van der Waals*-Oberfläche ergibt sich für den Grenzfall, bei dem der Radius des Probenmoleküls zu null gewählt wird. Je nach Algorithmus wird dabei analytisch oder numerisch die Größe der resultierenden Oberfläche in $Å^2$ bestimmt.

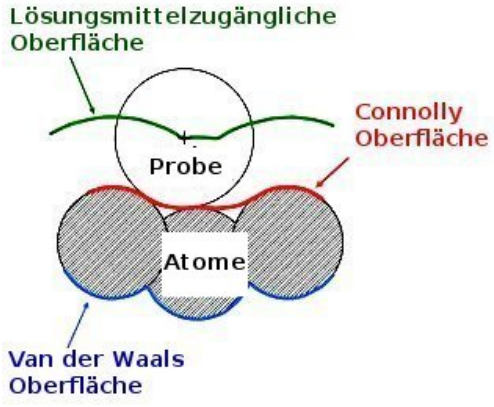


Abbildung 2.3: Definition der lösungsmittelzugänglichen Oberfläche
Ein kugelförmiges Probenmolekül rollt entlang der Van der Waals Oberfläche ab. Die Positionen, die sein Mittelpunkt dabei einnimmt definieren die lösungsmittelzugängliche Oberfläche (*SASA*) (Abbildung aus [55]).

Neben der Definition der Atomradien und dem gewählten Probenradius hängt die Größe der resultierenden Oberfläche davon ab, ob die *Connollyoberfläche* [58] oder die *lösungsmittelzugängliche Oberfläche (SASA)* [59] berechnet wird. Die *Connollyoberfläche* beschreibt die Punkte, auf der die Oberfläche der Probe zu liegen kommt, während die *SASA* durch diejenigen Punkte definiert ist, auf denen der Mittelpunkt des Probenmoleküls während des Rollvorganges wandert (siehe Abbildung 2.3).

In dieser Arbeit wird die lösungsmittelzugängliche Oberfläche (SASA) verwendet, wobei der Probenradius stets als $1,4$ Å gewählt wird, was in etwa dem Radius eines Wassermoleküls entspricht [59], das in der Zelle als Lösungsmittel fungiert. Das Zen-

trum dieses Probenmoleküls wird somit ungefähr die Fläche überstreichen, die für ein Wassermolekül zugänglich ist [60].

Bei festem Berechungsverfahren, Probenradius und *van der Waals* Radien beeinflusst nur der strukturelle Kontext eines Atoms bzw. einer Aminosäure im Makromolekül den Oberflächenanteil, der ihm zugewiesen wird. Eine Aminosäure, die im Proteininneren liegt, besitzt eine weit größere Oberfläche als eine Aminosäure, die sich an der Oberfläche in exponierter Lage befindet.

2.4.2 Aminosäurezusammensetzung von Protein-Protein Kontaktflächen

In mehreren Arbeiten wurde festgestellt, dass sich die Aminosäurezusammensetzung von PPKs signifikant von derjenigen der restlichen Oberfläche unterscheidet [14] [13] [61] [49] [28] [37]. Die genauen Werte der Häufigkeiten hängen in begrenztem Umfang davon ab, ob obligate oder transiente Komplexe, Homodimere oder Heterodimere betrachtet werden [62]. Generell sind jedoch neben den aliphatischen und aromatischen Aminosäuren auch Arginin, Histidin und Cystein an der PPK bevorzugt sind, während geladene und polare Aminosäuren an der restlichen Oberfläche häufiger gefunden werden [15] [49]. Diese Verteilung resultiert aus dem größeren Beitrag hydrophober Seitenketten zur Stabilität von Protein-Protein Interaktionen aufgrund des hydrophoben Effektes.

Aminosäurespezifische Information lässt sich anhand unterschiedlicher Scores quantifizieren um sie anschließend für die computergestützte Vorhersage von PPKs zu nutzen. *Porollo* und *Meller* [46] verwenden beispielsweise physikalisch-chemische Eigenschaften der Aminosäuren aus der *AAIndex-Datenbank* [63]. Diese Datenbank enthält numerische Indizes, die verschiedene Eigenschaften der Aminosäuren wie Größe und Hydrophobizität beschreiben. Damit lassen sich Informationen über Oberflächenaminosäuren von Proteinen erkennen und in maschinelle Lernverfahren mit integrieren.

Ähnliche Scores für einzelne Aminosäuren oder intermolekulare Kontaktpaare von Aminosäuren wurden bereits häufiger als wissensbasierte Potentiale aus Datensätzen von PPKs abgeleitet [21] [64]. Auch Scores, die neben dem Vorkommen einer einzelnen Aminosäure an der Oberfläche die Aminosäurezusammensetzung der räumlichen Nachbarschaft mit berücksichtigen, wurden bereits zur Vorhersage von PPKs [47] verwendet und haben sich als performanter erwiesen als Scores für einzelne Positionen [49]. Information über die Aminosäurezusammensetzung von PPKs und der restlichen Oberfläche ist sicherlich wichtig für eine qualitativ hochwertige Vorhersagen von Protein-Protein Interaktionsflächen.

2.4.3 Hydrophobe Patches

Daneben haben sich größere hydrophobe Bereiche an der Oberfläche eines Proteins als wichtiges Merkmal von PPKs herausgestellt. Im Folgenden wird zunächst die Eigenschaft der Hydrophobizität näher erläutert und ihre stabilisierende Wirkung auf PPIs begründet.

2.4.3.1 Definition der Hydrophobizität

Der Grund für die stabilisierende Wirkung hydrophober Wechselwirkungen ist der *hydrophobe Effekt*. Wasser, das in der Zelle als Lösungsmittel fungiert, hat die Tendenz hydrophobe Moleküle auszuschließen. Dieser Effekt ist nicht durch eine Abstoßung apolarer Moleküle durch Wasser bedingt. Vielmehr ergeben sich aufgrund der *Van der Waals* Interaktionen zwischen Wasser und einem apolaren Körper Anziehungskräfte. Das Verhalten des Wassers ist in der starken Neigung begründet zu anderen Wassermolekülen *Wasserstoffbrücken* auszubilden. Wassermoleküle sind im flüssigen Zustand tetraedrisch angeordnet, so dass sie durchschnittlich an 4 Wasserstoffbrücken beteiligt sind. Die Dynamik dieses Gitters erlaubt es einem Wassermolekül, seine Orientierung in relativ hohem Maße zu variieren ohne die Wasserstoffbrücken zu brechen. Aus diesem Grund ist die Entropie des Wassers im flüssigen Zustand sehr hoch. Die Situation ändert sich jedoch drastisch, wenn apolare Moleküle an ein Wassermolekül gelangen. Es bilden sich dann immer noch 4 Wasserstoffbrücken mit anderen Wassermolekülen aus, jedoch zahlt das Wassermolekül jetzt einen hohen entropischen Preis um die Brücken aufrechtzuerhalten. Der Winkelbereich, auf den seine 4 Wasserstoffbrücken verteilt werden, ist durch das apolare Molekül stark eingeschränkt, was sich negativ auf seine Bewegungsfreiheit auswirkt. In den meisten Fällen würde eine Veränderung der Orientierung die Wasserstoffbrücken brechen. Dies jedoch wäre ungünstig bezüglich der Enthalpie.

Für isobare und isotherme Bedingungen, wie sie in der Zelle vorherrschen, laufen Prozesse stets in diejenige Richtung ab, in der die *Gibbsche freie Energie*

$$G = H - T \cdot S \tag{2.1}$$

mit der *Enthalpie* H, der Temperatur T und der *Entropie* S minimiert wird. Gleichung (2.1) stellt den Zusammenhang zwischen den gegensätzlichen treibenden Kräfte der Enthalpie und Entropie dar. So führt zwar ein im Wasser gelöstes apolares Molekül

zu keinem enthalpischen Problem, da die Wasserstoffbrücken der Wassermoleküle nicht gebrochen werden und deshalb H auch nicht zunimmt. Über den Term $-TS$ bewirkt die sinkende Entropie jedoch einen Anstieg der *Gibbsschen freien Energie*. Ein optimaler Wert von G wird erreicht, indem die Anzahl der Kontakte zwischen Wassermolekülen und hydrophoben Molekülen minimiert wird.

2.4.3.2 Hydrophobizität an Protein-Protein Kontaktflächen

Seit den ersten Versuchen von Clothia und Janin [65] Protein-Protein Interaktionen vorherzusagen ist bekannt, dass der Eigenschaft der Hydrophobizität bei Protein-Protein Kontakten eine hohe Bedeutung zukommt. Hydrophobe Atome in Proteinen kommen meist geclustert als hydrophobe *Patches* vor. Während sie im Inneren des Proteins zur Proteinfaltung beitragen [66] [67], haben sie an der Proteinoberfläche oft die Aufgabe Liganden [68] [69] oder andere Proteine [65] [5] zu binden. Daher können hydrophobe Cluster an der Proteinoberfläche auch zur Vorhersage von Protein-Protein-Interaktionsflächen beitragen. Einen guten Überblick über die Natur hydrophober Patches findet man beispielsweise in [70].

2.4.4 Konserviertheit

Ein *Multiples Sequenzalignment* (MSA) ist eine zeilenweise Anordnung von Sequenzen homologer Proteine, so dass entsprechende Positionen zweier verschiedener Sequenzen untereinander zu liegen kommen. Abbildung 2.4 zeigt ein MSA, dessen Zeilen fünf Sequenzen (S1-S5) enthalten, während seine Spalten Aminosäurepositionen (P1-P6) der homologen Proteine entsprechen. Aus den Einträgen einer Spalte lassen sich positionsspezifische Informationen über die enthalten homologen Proteine ableiten. Unterschiede innerhalb einer Spalte repräsentieren Mutationen an einer Position des Proteins. Daher kann man aus einem MSA Informationen über evolutionären Druck, Mutationen und Rekombinationsereignisse ableiten. Sobald im Laufe der Evolution ein Protein lebenswichtig für eine Art geworden ist, können alle Mutationen in zwei Kategorien eingeordnet werden: Schädliche Mutationen und neutrale. Da sich schädliche Mutationen aufgrund evolutionären Drucks nicht durchsetzen, sind in einem MSA zu beobachtende Mutationen in der Regel neutral und repräsentieren keine Verbesserungen des Proteins. Anhand der Variabilität einer Position im MSA lässt sich daher die Toleranz das Protein gegenüber Mutationen an der entsprechenden Position ableiten. Falls die Ansprüche, die aus der Struktur oder Funktion eines Proteins erwachsen, nur von einer oder weniger Aminosäuren erfüllt werden können, so sind an dieser Position kaum Mutationen

erlaubt ohne die Funktion zu beeinträchtigen. Daher tauchen an solchen Positionen, wie P3 in Abbildung 2.4, kaum Mutationen auf und sie sind stark konserviert. P2 dagegen stellt ein Beispiel für eine Position dar, die zwar nicht strikt konserviert ist, an der jedoch nur die aliphatischen Aminosäuren *Leu*, *Ile* und *Val* erlaubt sind. Offensichtlich wird an dieser Position eine aliphatische Seitenkette benötigt. Regionen, die als Schleifen angeordnet sind besitzen dagegen in homologen Proteinen häufig unterschiedliche Längen und können, wie P4, anhand sehr vieler Lücken im MSA identifiziert werden. Positionen, an denen die Art der Seitenkette kaum eine Bedeutung für ein funktionsfähiges Protein besitzt, zeigen eine hohe Variabilität. Ein Beispiel für eine solche Position ist P6 in Abbildung 2.4, an der verschiedenste Arten von Seitenketten vorgefunden werden.

Daher hat sich die Konserviertheit einer Spalte in einem MSA als starker Indikator für funktional und strukturell wichtige Positionen eines Proteins erwiesen [51]. So ist es möglich über eine Konserviertheitsanalyse Positionen zu identifizieren, die für die Struktur wichtig sind [71] [72] [73], die an Ligandenbindung [74] [75] oder Protein-Protein Interaktionen [52] [53] beteiligt sind oder die die funktionale Spezifität von Proteinen bestimmen [76] [77] [78]. Konserviertheit wurde in vielen Anwendungen auch zusammen mit Information aus der Struktur verwendet [79] [80] um wichtige Positionen für die Funktion und Struktur des Proteins vorherzusagen.

	P1	P2	P3	P4	P5	P6
S1	D	L	W	-	R	S
S2	R	I	W	G	D	F
S3	D	L	W	-	R	V
S4	R	V	W	-	D	S
S5	R	L	W	G	D	G

Abbildung 2.4: Beispiel eines *Multiplen Sequenzalignments* (MSA)
Die Zeilen eines MSA beinhalten homologe Sequenzen (S1-S5), während seine Spalten (P1-P6) Aminosäurepositionen der homologen Proteine entsprechen. Anhand des Vorkommens bestimmter Aminosäuren kann man Information über die Aminosäurepositionen in den homologen Proteinen ableiten.

Während sich die Literatur darin einig ist, dass aktive Zentren und Ligandenbindestellen über viele verschiedene Proteinfamilien hinweg stark konserviert sind [81] [82], ist die Bedeutung von Konserviertheit an PPKs weniger klar. Die Konserviertheit an der PPK unterscheidet sich nicht signifikant von der im Proteininneren [81]. Die Kontaktfläche ist jedoch etwas stärker konserviert als die restliche Oberfläche [81] [83] [52]. Dieser Umstand für sich alleine betrachtet ist in den meisten Fällen allerdings nicht ausreichend um Protein-Protein Bindestellen zuverlässig vorherzusagen [52] [84]. Bei einer Vielzahl von Verfahren zur Vorhersage von PPKs wird Konserviertheit als zusätzliche orthogonale Information verwendet und verbessert dabei die Güte der Vorhersage [47] [74] [85].

2.4.5 Korrelierte Mutationen

Aber auch nicht strikt konservierte Positionen beinhalten ein Signal, das für die Vorhersage von PPKs genutzt werden kann. Da die Seitenketten einer PPK mit ihren Interaktionspartnern wechselwirken, sind nicht alle Mutationen zugelassen. Mutationen, die den Porteinkomplex schwächen, können nur dann überleben, wenn sie durch eine weitere kompensierende Mutation ausgeglichen werden. Solche Positionen eines Proteins, an denen derartige Mutationen auftauchen, lassen sich anhand von MSAs identifizieren. So erkennt man am MSA in Abbildung 2.4, dass die beiden Positione P1 und P5 auf korrelierte Art mutieren. Ein kleines negativ geladenes *Asp* an P1 bedingt ein großes positiv geladenes *Arg* an P5 und umgekehrt.

Korrelierte Mutationen werden sowohl intramolekular [86] [54] als auch intermolekular an Protein-Protein-Komplexen [54] [87] beobachtet. Intramolekular werden sie zur Vorhersage funktional wichtiger aber nicht strikt konservierter Positionen verwendet [88]. Eine weitere Anwendung korrelierter Mutationen ist die Vorhersage räumlicher Nachbarschaft von Positionen in der 3D-Struktur. In diesem Zusammenhang kann die Information aus korrelierten Mutationen mit weiterer orthogonaler Information kombiniert werden und dient der Strukturvorhersage eines monomeren Proteins [89].

Intermolekulare korrelierte Mutationen können zur Vorhersage von Protein-Protein Interaktionen [90] und zur Unterstützung von Dockingverfahren [91] verwendet werden. Allerdings befand eine andere Arbeit intermolekulare Korrelationen als nicht allgemein aussagekräftig für Protein-Protein Interaktionen [92]. In einer weiteren groß angelegten Studie wurde gezeigt, dass der Abstand korrelierter Positionen koevolvierender Proteine kleiner ist als der mittlere Abstand anderer Positionen [93]. In den Verfahren *Evolutionary Trace* [78], *ConSurf* [94] und *Phylogenetic Motifs* [95] werden Informationen zur Koevolution von Aminosäureresten mehrerer Ketten dazu verwendet, funktionale Bereiche von Proteinen zu identifizieren.

In dieser Arbeit werden intermolekulare korrelierte Mutationen als zusätzliche Information bei der Vorhersage von PPKs benutzt.

2.4.6 Maschinelle Lernverfahren

Eigenschaften, wie oben dargestellt, dienen dazu, einen Klassifikator zu entwickeln, der für eine Position an der Oberfläche des Proteins eine Vorhersage dafür generiert, ob sie sich an der PPK oder an der restlichen Oberfläche befindet. Maschinelle Lernverfahren

dienen dabei dem Zweck, sich möglicherweise widersprechende Eigenschaften zu einer optimalen Vorhersage zu kombinieren. Für den Fall der Mustererkennung von Datensätzen ohne weitere Information über die Klassenzugehörigkeit einzelner Datenpunkte können nicht-überwachte Lernverfahren wie *Principal Component Analysis*, *Independent Component Analysis* oder *Nonnegative Matrix Factorization* verwendet werden. Falls ein Datensatz aus korrekt klassifizierten Beispielen zur Verfügung steht, können überwachte Lernverfahren wie *neuronale Netze*, *Bayessche Netze* oder *Support Vektor Maschinen* (SVM) benutzt werden. Da überwachte Lernverfahren während des Trainings Informationen darüber sammeln, nach welchen Kriterien die Datenpunkte zu klassifizieren sind, ist ihre Performanz im Allgemeinen weit besser als die nicht-überwachter Lernverfahren. Falls ein korrekt klassifizierter Trainingsdatensatz vorhanden ist, ist es folglich erfolgversprechender, ein überwachtes Lernverfahren zu verwenden. Da aus Datenbanken Strukturdatensätze bekannter Protein-Protein Komplexe entnommen und zum Training eines Klassifikators verwendet werden können, werden zur Vorhersage von PPKs in der Regel überwachte Lernverfahren benutzt.

Neben *neuronalen Netzen* [46], *Conditional random fields* [96] und *Bayesschen Netzen* [97] haben sich vor allem *Support Vektor Maschinen (SVM)* [44] [45] in diesem Bereich etabliert. Ihr Vorteil besteht in ihrer Robustheit und geringen Anfälligkeit gegenüber *Overlearning*, dem auswendig lernen spezifischer Eigenschaften einzelner Beispiele anstatt dem Lernen generell gültiger Merkmale zur Unterscheidung der Klassen. Daneben bieten sie die Möglichkeit auf mathematisch fundierter Basis eine *a posteriori* Wahrscheinlichkeit für jede Vorhersage zu bestimmen. Für die weitere Einordnung einer Vorhersage ist dies weitaus nützlicher als eine binäre Klassifikation. Aus diesem Grund wird auch in dieser Arbeit eine *SVM* verwendet um die fünf positionsspezifischen Merkmale zu einer aussagekräftigen Vorhersage der PPKs zu verrechnen.

3 Materialien und Methoden

In diesem Kapitel werden die verwendeten Programme und Algorithmen detailliert beschrieben. Daneben werden die Datenquellen aufgeführt, die zur Herleitung charakteristischer Merkmale von Protein-Protein Kontaktflächen, sowie zum Training und zur Evaluation des Klassifikators benutzt wurden.

3.1 Strukturdatensätze von Protein-Protein Komplexen

Strukturen von Protein-Protein Komplexen können dem PDB-Archiv der RCSB Protein Datenbank [98] entnommen werden. Bei der Zusammenstellung eines Datensatzes ist zu beachten, dass bei Projekten der Strukturaufklärung bestimmte Proteine bevorzugt untersucht werden. Ist es das Ziel, einen möglichst repräsentativen Datensatz zu generieren, der alle bekannten Proteinstrukturen mit gleichem Gewicht repräsentiert, so sind Redundanzen aus dem Datensatz auszufiltern. In dieser Arbeit werden 2 nichtredundante Datensätze dreidimensionaler Strukturen von Protein-Protein Komplexen verwendet, die in den folgenden Abschnitten näher vorgestellt werden.

3.1.1 Der Datensatz von $\mathbf{Komp_{RN}}$

In der Arbeitsgruppe von *R. Nussinov* wurde der Datensatz $Komp_{RN}$ generiert, wobei redundante Strukturen von Protein-Protein Komplexen anhand ihrer strukturellen Ähnlichkeit erkannt und ausgefiltert wurden [99]. Deshalb ist davon auszugehen, dass dieser Datensatz ausreichend divers und redundanzfrei ist, um in dieser Arbeit zur Herleitung wissensbasierter Potentiale verwendet zu werden.

Der ursprüngliche Datensatz nach *Mintz et al.* beinhaltet sowohl Homodimere als auch Heterodimere. Bei der Auswahl verwandten die Authoren die folgenden Kriterien: Ausgeschlossen wurden modellierte Strukturen, Strukturen mit einer geringeren Auflösung als $3,5$ Å, Strukturen, die nur aus C_α-Atome bestehen und Strukturen, die weniger als

10 interagierende Reste in jeder Kette besitzen. Die Interaktionen zweier Aminosäuren wurde anhand des Kriteriums (3.8) aus Abschnitt 3.2 bestimmt. Die Strukturdaten wurden anschließend anhand räumlicher und physikalisch-chemischer Ähnlichkeit sowohl der Haupt- als auch der Seitenkettenatome unabhängig von der Sequenz der monomeren Proteinketten geclustert. Daraus resultieren 2582 Cluster von denen jeder eine Klasse zueinander ähnlicher Protein-Protein Komplexe repräsentiert. Die Komplexe innerhalb eines jeden Clusters wurden anschließend mithilfe von BLASTClust [100] aligniert und anhand ihrer Sequenzidentität miteinander verglichen. Falls eine Sequenzidentität von 50% von zwei Komplexen überschritten wurde, so wurde einer der beiden Komplexe entfernt. Zuletzt wurde aus jedem Cluster derjenige Komplex ausgewählt, der zum Rest des Clusters die höchste Sequenzidentität aufweist und dem in dieser Arbeit verwendeten Datensatz $Komp_{RN}$ hinzugefügt.

3.1.2 Der Datensatz $\mathbf{Komp_{trans}}$

Zur Evaluation der Performanz des in dieser Arbeit entwickelten Programms zur Vorhersage von Kontaktflächen anhand transienter Heterodimere wird der Datensatz *Benchmark 4.0* [101] benutzt. Ursprünglich wurde dieser Datensatz als Test der Performanz von Dockingverfahren erstellt. Er setzt sich aus den Strukturen nichtredundanter transienter Protein-Protein Komplexe der *PDB*-Datenbank zusammen, von denen sowohl die Struktur des niedrigeren Oligomerisierungszustandes als auch diejenige des höheren bekannt ist.

Für die monomeren Ketten gilt dabei eine Mindestlänge von 30 Resten. Außerdem muss der Komplex als Röntgenstruktur mit einer Auflösung von mindesten $3,25$ Å vorhanden sein. Für die Struktur des Monomers gilt dasselbe, jedoch werden hier auch NMR-Strukturen akzeptiert. Über einen Abgleich mit der *SCOP*-Datenbank [102] wurden anschließend Redundanzen auf Ebene der Proteinfamilie entfernt. Desweiteren wurden mittels Literaturrecherche obligate Komplexe entfernt.

Der Datensatz *Benchmark 4.0* besteht aus 121 “einfachen” Fällen, bei denen kaum Änderungen des Hauptkettenverlaufs durch die Komplexbildung auftreten, 31 “mittelschweren” und 25 “schwierigen” Fälle, die größere konformationelle Änderungen der Hauptkette bei der Komplexbildung erfahren. Da die in dieser Arbeit entwickelte Software keine Flexibilität der Hauptkette berücksichtigt, wird der Datensatz auf die 121 einfachen Fälle beschränkt, die bei der Komplexbildung kaum eine Deformationen der Hauptkette erfahren.

Viele interagierende Untereinheiten des Datensatzes haben jedoch nur eine geringe Sequenzlänge. Daher finden sich in Sequenzdatenbanken oft nur sehr wenige signifikante Treffer, so dass keine MSAs mit hinreichender Datengrundlage erstellt werden können. Deshalb wurde dieser Datensatz weiter eingeschränkt auf 35 Beispiele, für die MSAs mit mindestens 100 Sequenzen generiert werden konnten. Dieser Datensatz aus 35 transienten heterodimeren wird im Folgenden als $Komp_{trans}$ bezeichnet.

3.1.3 Kanonische Kontaktflächen

Wie im letzten Kapitel erläutert, sind Protein-Protein Interaktionen an einer Vielzahl verschiedener zellulärer Prozesse beteiligt und besitzen daher große Variabilität, was die Struktur des Komplexes anbelangt. Die allgemeine Vorstellung von *kanonischen* Protein-Protein Komplexen geht von zwei Untereinheiten aus, die sich berühren und dabei über ihre näherungsweise planare Interaktionsfläche miteinander wechselwirken. Viele Komplexe aus den vorgestellten Datensätzen zeigen jedoch ineinander verschlungene Strukturen und widersprechen somit dieser Vorstellung *kanonischer* Komplexe. Da *nicht-kanonische* Komplexe Probleme bei der Auswertung verursachen können, wurde im Rahmen dieser Arbeit ein Filterkriterium entwickelt, um PPKs, die zu stark von einer planaren Form abweichen, automatisiert aus einem Datensatz von Protein Komplexen zu entfernen.

Um zu bestimmen, ob die PPK einer Proteinkette *kanonisch* ist, soll zunächst eine approximierende Ebene in die PPK gelegt werden. Weicht die Kontaktfläche zu stark von dieser Ebene ab, so wird der zugehörige Komplex als Spezialfall verworfen.

Zur Bestimmung der approximierenden Ebene werden diejenigen 3 Kontaktaminosäuren R_1, R_2, R_3 bestimmt, die paarweise voneinander den größten Abstand besitzen

$$\overline{R_1R_2} + \overline{R_2R_3} + \overline{R_3R_1} = max. \tag{3.1}$$

Dabei wird der gegenseitige Abstand der Reste über den euklidischen Abstand ihrer C_α-Atome gemessen. Die drei Reste R_1, R_2 und R_3 spannen die approximierende Ebene auf, die in Abbildung 3.1 dargestellt ist. Da im nächsten Schritt Abstände bzgl. dieser Ebene berechnet werden sollen, ist außerdem die *Hessesche Normalenform* der zugehörigen Ebenengleichung nötig. Dazu wird ein Normalenvektor $\vec{n}$ auf die Ebene ermittelt. Man erhält $\vec{n}$ z.B. über das Kreuzprodukt der beiden Abstandsvektoren $\overrightarrow{R_1R_2}$ und $\overrightarrow{R_1R_3}$

$$\vec{n} = \overrightarrow{R_1R_2} \times \overrightarrow{R_1R_3}. \tag{3.2}$$

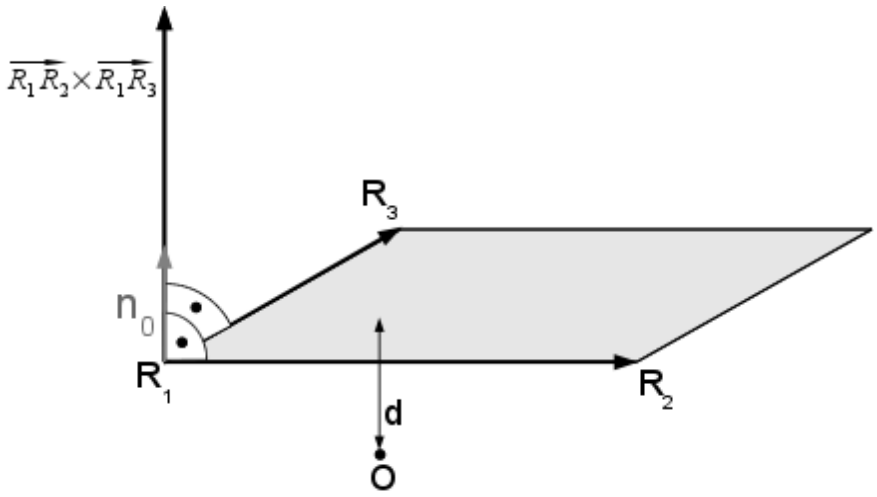


Abbildung 3.1: Die approximierende Ebene
Die Abbildung zeigt die Ebene, die in die Kontaktfläche gelegt wird. Aufhängepunkte sind drei Seitenketten, repräsentiert durch die Punkte R_1, R_2 und R_3. Die Ebene wird definiert über den Normalenvektor $\vec{n} = \overrightarrow{R_1R_2} \times \overrightarrow{R_1R_3}$, der zu dem Vektor $\overrightarrow{n_0}$ der Länge 1 normiert wird. d bezeichnet den Abstand der Ebene vom Ursprung O.

Alle Punkte X mit Ortsvektor $\vec{x}$, die auf der gesuchten Ebene liegen, erfüllen die *Normalenform* der Ebenengleichung

$$(\vec{x} - \overrightarrow{R_1}) \cdot \vec{n} = 0 \tag{3.3}$$

mit dem Aufhängepunkt $\overrightarrow{R_1}$ des Normalenvektors. Normiert man den Normalenvektor $\vec{n}$ auf die Länge 1

$$\overrightarrow{n_0} = \frac{\vec{n}}{\|\vec{n}\|} \tag{3.4}$$

so gilt

$$(\vec{x} - \overrightarrow{R_1}) \cdot \overrightarrow{n_0} = 0 \tag{3.5}$$

und man erhält mit $d = \overrightarrow{R_1} \cdot \overrightarrow{n_0} > 0$ die Hessesche Normalenform der Ebenengleichung

$$\vec{x} \cdot \overrightarrow{n_0} - d = 0. \tag{3.6}$$

Der Abstand d_p eines Punktes P mit Ortsvektor $\vec{p}$ von der Ebene lässt sich dann berechnen als

$$d_p = (\vec{p} - \overrightarrow{R_1}) \cdot \overrightarrow{n_0}. \tag{3.7}$$

Über (3.7) kann für jede Aminosäure an der Kontaktfläche der Abstand ihres C_α-Atoms zur approximierenden Hyperebene bestimmt werden. Nach Berechnung der Ebene sind nun Kriterien zu definieren, die eine Kontaktfläche zu erfüllen hat um als *kanonisch* eingeordnet zu werden. Als brauchbar um *kanonische* Kontaktflächen zu finden haben sich folgende beiden Kriterien erwiesen:

- **Kriterium 1:** Ist eine Aminosäure mehr als ein Schwellwert D von der approximierenden Hyperebene entfernt, so gilt die Aminosäure als Ausreißer.

- **Kriterium 2:** Besteht die Kontaktfläche zu einem größeren Bruchteil als b mit $0 < b < 1$ aus Ausreißern, so wird der dazu gehörende Protein-Protein Komplex als *nicht-kanonisch* verworfen.

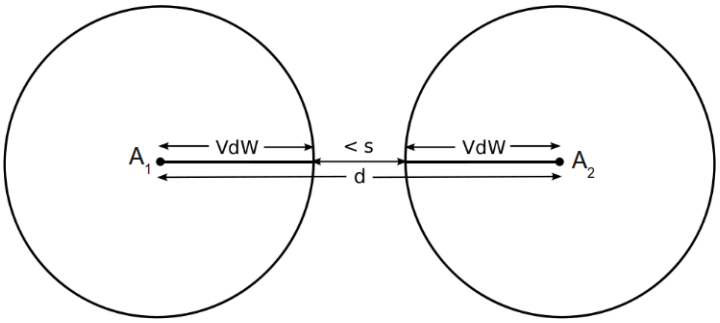
Die Schwellwerte D und b lassen sich dabei je nach Bedarf restriktiver oder weniger restriktiv anpassen. Sie wurden in dieser Arbeit als $D = 6$ Å und $b = 0,4$ bzw. $b = 0,6$ gewählt.

3.2 Definition der Protein-Protein Kontaktfläche

Die Kontaktfläche eines Proteinkomplexes wird in dieser Arbeit anhand der bekannten Komplexstruktur über ein Abstandskriterium definiert. Wie in Abbildung 3.2(a) dargestellt, werden zwei Aminosäuren an der Oberfläche der einzelnen Ketten als Kontakt gewertet, falls sich die Mittelpunkte mindestens zweier ihrer Atome näher sind als

$$d = s + v_1 + v_2. \tag{3.8}$$

Als Schwellwert s wird in dieser Arbeit meist $0,5$ Å verwendet. Die benutzten Van der Waals Radien v_1 und v_2 finden sich in Tabelle 3.2(b).



(a)

C	1,7
N	1,5
O	1,4
S	1,85

(b)

Abbildung 3.2: Abstandskriterium zur Bestimmung der Kontaktatome
(a) Zwei Atome müssen sich näher sein als die Summe ihrer Van der Waals Radien plus ein Schwellwert s. (b) Die atomspezifischen van der Waals Radien in Å nach [103].

3.3 Multiple Sequenzalignments

In dieser Arbeit wurden MSAs aus 2 verschiedenen Quellen benutzt. Zum einen fanden MSAs aus der *HSSP*-Datenbank [104] vom Stand des August 2009 Verwendung. Falls das zugehörige MSA aus der *HSSP*-Datenbank nicht alle geforderten Kriterien erfüllen konnte, so wurde ein MSA aus dem Ergebnis einer Abfrage von Sequenzdatenbanken generiert.

Bei der Analyse der Konserviertheit und korrelierter Mutationen fanden die MSAs aus der *HSSP*-Datenbank Verwendung. In diesem Datensatz wurden die Sequenzen interagierender Proteine in zwei separaten MSAs gehalten. Für die fehlerfreie Bestimmung der Koevolutionssignale ist es notwendig sicherzustellen, dass in beiden MSAs die gleich Sortierreihenfolge hinsichtlich der Herkunft aus phylogenetischen Arten eingehalten wird. Dies kann innerhalb der *HSSP*-Datenbank meist nur für Homodimere sicher erfüllt werden, da man im Falle von Homodimeren dasselbe MSA für beide interagierende Untereinheiten verwenden kann. Innerhalb der MSAs von Heterodimeren der *HSSP*-Datenbank finden sich meist nicht genügend Sequenzen für die die richtige Sortierreihenfolge sichergestellt werden kann.

Enthalten die MSAs der *HSSP*-Datenbank zu wenige Sequenzen so werden zunächst

mithilfe von *BLAST* [100] homologe Sequenzen aus der nichtredundanten Proteinsequenzdatenbank des NCBI [105] nach dem Stand vom 07.02.2011 abgeleitet. Anschließend wird mithilfe von *Muscle 3.8.31* [106] ein MSA generiert.

Jedes MSA wird vor jeglicher Auswertung mithilfe eines Ähnlichkeitsfilters prozessiert. Damit wird sichergestellt, dass alle Sequenzen im paarweisen Vergleich eine Sequenzidentität zwischen $id_{min} = 20\%$ und $id_{max} = 90\%$ besitzen. Somit wird eine hinreichende Diversität der Sequenzen im MSA sichergestellt und erreicht, dass die Datensätze frei von Redundanzen sind.

3.4 Konserviertheit

In der Literatur sind mehrere Ansätze beschrieben um die Konserviertheit von MSA-Spalten zu bewerten. Neben Scores, die auf Shannonscher Entropie basieren, gibt es Methoden, die auch physikalisch-chemisch Ähnlichkeiten der Aminosäuretypen berücksichtigen.

3.4.1 Shannonsche Entropie

1991 wurde erstmals die Shannonsche Entropie $H(i)$ einer Spalte i im MSA zur Bewertung der Konserviertheit benutzt [107]. Sie wird berechnet nach:

$$H(i) = \sum_{k=1}^{20} f_i(a_k) \log f_i(a_k). \tag{3.9}$$

Dabei bezeichnet $f_i(a_k),\ k = 1, \ldots, 20$ die Häufigkeit des Symbols a_k in der Spalte i des MSAs. Diese Werte $f_i(a)$ müssen aus den im MSA beobachteten Häufigkeiten $n_i(a)$ abgeschätzt werden. Enthält ein MSA nur eine geringe Anzahl von Sequenzen, so kommen möglicherweise einige Aminosäuren a_k in Spalte i nicht vor. Dann gilt für die geschätzte Häufigkeit dieser Aminosäure $n_i(a_k) = 0$ und der Wert von $\log(f_i(a_k))$ in (3.9) ist nicht bestimmbar. Dieses Problem wird umgangen, indem nach [108] die am MSA gemessenen Häufigkeiten $n(a_k)$ durch sogenannte *Pseudocounts* korrigiert werden gemäß

$$f_i(a_k) = \frac{n_i(a_k) + \lambda \sum_{l=1}^{20} \frac{n_i(a_l) B(a_k, a_l)}{\sqrt{n_i}}}{n_i + \lambda \sqrt{n_i}}. \tag{3.10}$$

In dieser Arbeit wird λ als $\lambda = 10^{-3}$ gewählt und als 20×20 Substitutionsmatrix der Aminosäurearten $B(a_k, a_l)$ die *BLOSUM62*-Matrix [109] verwendet. n_i bezeichnet die Anzahl aller Sequenzen im MSA, die an Position i keine Lücke (Gap) aufweisen.

Um der Tatsache gerecht zu werden, dass in wildtypischen Proteinen die Aminosäurearten mit unterschiedlichen relativen Häufigkeiten vorkommen, kann (3.9) zu einer relativen Shannonschen Entropie erweitert werden [110]:

$$H(i) = \sum_{k=1}^{20} f_i(a_k) \log \left(\frac{f_i(a_k)}{p_{BG}(a_k)} \right). \tag{3.11}$$

Dabei ist $P_{BG}(a_k)$ die Hintergrundhäufigkeit der Aminosäure a_k, die aus Statistiken von Sequenzdatenbanken entnommen werden kann. In dieser Arbeit wurden die Werte aus Tabelle 3.1 verwendet. Diese stammen aus der UniProt Sequenzdatenbank [111] vom Stand des 13.01.2011. Nach *Wang* wird durch diese Erweiterung um Hintergrundhäufigkeiten die Performanz der Shannonschen Entropie bei der Identifikation funktionaler Positionen signifikant verbessert [110].

A	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**
8,61	1,27	5,29	6,13	4,03	7,12	2,19	6,02	5,27	9,83
M	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
2,47	4,15	4,73	3,86	5,46	6,69	5,61	6,74	1,31	3,06

Tabelle 3.1: Relative Häufigkeiten der 20 Aminosäuretypen
Die Tabelle zeigt die relativen Häufigkeiten der 20 natürlich vorkommenden Aminosäurearten in der UniProt Sequenzdatenbank [111] vom Stand des 13.01.2011. Alle Werte sind in % angegeben.

3.4.2 Verbesserte Bewertung der Konserviertheit

Die meisten Konserviertheitsscores berücksichtigen entweder nur Häufigkeiten wie die im letzten Abschnitt vorgestellten Scores basierend auf Shannonscher Entropie, oder nur physikalisch-chemische Ähnlichkeit von Aminosäuren [112] [113]. Sogenannte *Sum of Pair - Scores* berechnen die Summe aller möglichen paarweisen Ähnlichkeiten der

Aminosäuren an der alignierten Position [94] [114] [115] [83].

In jüngerer Zeit wurde ein Score entwickelt, der sowohl Häufigkeiten in einer Spalte als auch physikalisch-chemische Ähnlichkeiten der Aminosäuren berücksichtigt [116]. Dieser Score wurde später reskaliert um doppelte Logarithmierung und die damit verbundenen Probleme zu vermeiden. Er erwies sich im Vergleich mit Shannonscher Entropie bei der Erkennung funktionaler Positionen von Proteinen als robuster [117]. Im Folgenden wird dieses Verfahren im Detail vorgestellt.

Der Grundgedanke bei dieser Methode ist es, dass der Beitrag eines Aminosäuretyps zur Konserviertheit einer Spalte davon abhängt, welcher Typ Aminosäure diese Spalte dominiert. Dabei ergibt sich die dominierende Aminosäureart zunächst als diejenige, die in der Spalte die größte Häufigkeit besitzt. Beispielsweise wird in einer *Aspartat*-dominierten Spalte die Aminosäure *Aspartat* den größten Beitrag zur Konserviertheit leisten. *Glutamat* wird einen größeren Beitrag zur Konserviertheit in einer *Aspartat*-dominierten Spalte leisten als *Phenylalanin*, da *Aspartat* und *Glutamat* beide klein und negativ geladen sind, während *Phenylalanin* groß und apolar ist.

Bei der folgenden Herleitung werden die vereinfachenden Annahmen benutzt, dass jeder Austausch relativ zur jeweils dominierenden Aminosäure geschieht und dass alle Austausche unabhängig voneinander erfolgen. Damit lässt sich die Stärke einer Mutation relativ zur dominierenden Aminosäure messen. Da die BLOSUM62-Matrix ein allgemein anerkanntes Maß für die Ähnlichkeit der physikalisch-chemischen Eigenschaften von Aminosäuren darstellt [109], wird diese Matrix zur Bewertung der Mutationen verwendet. Die BLOSUM62-Matrix wird so normiert, dass $B(a,a) = 10$ und $2 \leq B(a,b) \leq 9$ für $a \neq b$ (siehe Tabelle 3.2). Desweiteren wird analog zu anderen Arbeiten [94] [115] die Ähnlichkeit jeder Aminosäure zum Symbol für die Lücke mit null bewertet.

Im Folgenden wird o.B.d.A. angenommen, dass der Score für eine *Asp*-dominierte Spalte berechnet werden soll. Seien a_i und n_i mit $i = 1, \ldots, 20$ die Ähnlichkeitswerte aus der *Asp*-Zeile der Ähnlichkeitsmatrix B und die zugehörigen Häufigkeiten aus der Spalte des MSA. In diesem Fall entspricht a_i der vierten Zeilen in Tabelle 3.2. Dann wird der Konserviertheitsscore definiert als

$$C_{Sim\_Freq} = \sum_{i=1}^{20} n_i \cdot a_i. \tag{3.12}$$

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

Tabelle 3.2: Normierte BLOSUM62-Matrix
Die Einträge der Matrix wurden nach [117] so normiert, dass alle Werte zwischen 2 und 10 liegen. Die größten Werte befinden sich auf der Hauptdiagonale.

Angesichts der Tatsache, dass die Einträge der BLOSUM62-Matrix logarithmierte Wahrscheinlichkeiten sind, kann man diesen Score anschaulich interpretieren. Die Einträge der BLOSUM-Matrix B berechnen sich als $B_{ij} = c \cdot \log\left(\frac{p_{ij}}{p_i p_j}\right)$ wobei c eine Konstante ist, p_{ij} die beobachtete Austauschhäufigkeit der Aminosäure i mit Aminosäure j darstellt und p_i und p_j die Häufigkeiten repräsentieren mit denen die beiden Aminosäuren i und j einzeln erwartet werden. Schreibt man für eine *Aspartat*-dominierte Spalte a_i seiner Entstehung nach als Chancenquotienten für die Austauschhäufigkeit von *Aspartat* durch i

$$a_i = m \cdot \log\left(\frac{p_{Di}}{p_D p_i} + n\right) \tag{3.13}$$

mit den beiden Konstanten $m > 0$ und n, so stellt a_i ein Maß für die Ähnlichkeit der Aminosäuren *Aspartat* und i dar. Umgekehrt führt eine hohe Ähnlichkeit der Aminosäure i mit *Aspartat* dazu dass i häufig gemeinsam mit *Aspartat* in einer Spalte vorkommt. Da außerdem der Logarithmus eine streng monoton steigende Funktion ist, findet man

Aspartat und i umso häufiger gemeinsam in einer Spalte, je größer a_i. Der Wertebereich von C_{Sim_Freq} aus (3.12) liegt aufgrund der Normierung der BLOSUM62-Matrix (Tabelle 3.2) zwischen 0 und 10 mit einem Minimum von 0 falls sich nur Lücken in der Spalte befinden und einem Maximum bei 10 falls eine Spalte strikt konserviert ist.

Aufgrund mehrerer Ursachen (z.B. Forschungsschwerpunkte, Kultivierbarkeit von Mikroorganismen) stellen die Inhalte von Sequenzdatenbanken keine repräsentative redundanzfreie Stichprobe der in der Natur vorkommenden Proteine dar. Solche Redundanzen setzen sich in MSAs, deren Quelle stets Sequenzdatenbanken sind, fort. So kann es vorkommen, dass die Sequenzen in einem MSA aus vielen nicht verwandten Mikroben und einigen Säugetieren stammen, die eine eng verwandte Subgruppe bilden. Bei der Auswertung eines derart zusammengesetzten MSAs würden manche Spalten als stark konserviert erscheinen, die bei der Auswahl einer repräsentativeren Stichprobe eher variabel wären.

Eine Methode um zu hohe Redundanzen im MSA zu verringern besteht darin, neben zu unähnlichen Sequenzen auch sehr ähnliche Sequenzen auszufiltern. Trotzdem sind die Sequenzen innerhalb der Filtergrenzen nicht gleich verteilt. Aus diesem Grund wurden verschiedene Verfahren zum Gewichten einzelner Sequenzen bei der Auswertung von MSAs entwickelt. Üblicherweise werden dabei unterschiedliche Sequenzen, die mehr Variabilität und neue Information mitbringen stärker gewichtet als nahe miteinander verwandte Sequenzen.

So wird zur Verfeinerung der Scores eine positionsspezifische Gewichtung der Sequenzen verwendet [117], wie sie in [118] vorgeschlagen wurde. Dabei wird das Gewicht der l-ten Sequenz an Position x berechnet als $w_{lx} = \frac{1}{k_x n_{x_l}}$, wobei k_x die Anzahl verschiedener Arten von Aminosäuren in Spalte x und n_{x_l} die Häufigkeit der Aminosäure darstellt, die in der l-ten Sequenz in dieser Spalte x zu finden ist. Indem man über alle Positionen einer Sequenz mittelt bekommt man für jede Sequenz ein Gewicht w_l

$$w_l = \frac{1}{L} \sum_{x=1}^{L} w_{lx}. \tag{3.14}$$

Dabei ist die Länge des Alignments mit L bezeichnet. In dieser Metrik wird die l-te Sequenz mit w_l gewichtet. Seien die Gewichte der n_i Sequenzen, deren Symbol in der aktuellen Spalte die i-te Aminosäure ist, $w_{i1}, \cdots, w_{in_i}$. Dann lässt sich in (3.12) n_i substituieren durch $\sum_{j=1}^{n_i} w_{ij}$ und man erhält als Maß für die Konserviertheit bei Berücksichtigung der unterschiedlichen Gewichtung der Sequenzen

$$C_{Sim\_Freq} = \sum_{i=1}^{20} \left(\sum_{j=1}^{n_i} w_{ij} \right) a_i. \tag{3.15}$$

In diesem Fall ist die dominierende Aminosäure i konsequenterweise nicht über ihre absolute Häufigkeit, sondern über das Maximum von $m_i = \sum_{j=1}^{n_i} w_{ij}$ zu ermitteln.

3.5 Korrelierte Mutationen

Zur Bewertung korrelierter Mutationen wurden mehrere Arten von Scores entwickelt. In [119] und [120] wurden Korrelationskoeffizienten benutzt. Außerdem wurden *Transinformation* [121], [122], [123] und *Chi-Quadrat Tests* [120] verwendet. In dieser Arbeit wird sowohl ein Verfahren basierend auf *Pearson-Korrelation* [124] als auch eine Methode basierend auf normierter Transinformation [88] angewandt.

3.5.1 Pearson Korrelation

Häufig wird die Pearson Korrelation verwendet, um korrelierte Mutationen zu quantifizieren. Dabei werden zunächst für jede Position im MSA alle Mutationen erfasst. Dies geschieht für eine feste Position i indem für jedes Paar an Sequenzen k, l die Mutation an Position i anhand einer Ähnlichkeitsmatrix für Aminosäuren bewertet wird. In dieser Arbeit werden dafür die McLachlan-Substitutionsmatrix [125] und die BLOSUM50 [126] verwendet. Die so erhaltenen Werte füllen nun für jede Position i im MSA eine Distanzmatrix $D_i(k,l)$ der Dimension $N \times N$, wobei N die Anzahl der Sequenzen im MSA bezeichnet. Im Folgenden wird aus kombinatorischen Gründen nur das rechte obere Dreieck dieser Matrix ohne die Diagonale verwendet. Daraus lässt sich nun für jedes mögliche Paar an Positionen im MSA i, j die Pearson-Korrelation bestimmen als

$$r_{ij} = \frac{2}{N \cdot (N-1)} \cdot \frac{\sum_{k=1}^{N} \sum_{l=k+1}^{N} \left(D_i(k,l) \ - \overline{D_i} \right) \cdot \left(D_j(k,l) \ - \overline{D_j} \right)}{\sigma_i \cdot \sigma_j}. \tag{3.16}$$

Dabei bezeichnen $\overline{D_i}$ und σ_i den Mittelwert und die Standardabweichung der Einträge der Matrix D_i im rechten oberen Dreieck ohne die Werte auf der Hauptdiagonale. Für die Pearson-Korrelation gilt $-1 \leq r_{ij} \leq 1$. Falls die Spalten i und j völlig synchron mutieren gilt $r_{ij} = 1$, falls i und j völlig im Gegentakt mutieren gilt $r_{ij} = -1$.

Da strikt konservierte Positionen ebenso wenig ausgewertet werden können wie Positionen die hauptsächlich Lücken enthalten, wird die Korrelation nur für Spalten bestimmt, die zu weniger als 90% konserviert sind und die weniger als 25% Lücken enthalten. Außerdem werden bei der Berechnung des Korrelationskoeffizienten für ein Spaltenpaar i, j nur Sequenzen berücksichtigt, die weder an Position i noch an Position j eine Lücke besitzen. Alle anderen Einträge in $D_i(k,l)$ und $D_j(k,l)$ werden sowohl bei der Berechnung der Erwartungswerte $\overline{D_i}$ und $\overline{D_j}$ als auch bei der Summation in (3.16) übersprungen, womit sich auch der Normierungsfaktor entsprechend ändert.

Um ähnlichere Sequenzpaare bei der Summation in (3.16) geringer zu gewichten wurde vorgeschlagen [119], die Sequenzpaare mit dem Bruchteil ihrer Missmatches $0 < w_{kl} < 1$ gemessen über die gesamte Sequenzlänge zu gewichten. Damit erweitert sich (3.16) zu

$$r_{ij} = \frac{2}{N \cdot (N-1)} \cdot \frac{\sum_{k=1}^{N} \sum_{l=k+1}^{N} w_{kl} \cdot \left(D_i(k,l) \ - \overline{D_i}\right) \cdot \left(D_j(k,l) \ - \overline{D_j}\right)}{\sigma_i \cdot \sigma_j}. \tag{3.17}$$

3.5.2 Normierte Transinformation (*Mutual Information*)

Die Shannonsche Entropie für eine Position i in einem MSA ist ein Maß für die Variabilität. Sie definiert sich über die jeweiligen Wahrscheinlichkeiten $P_i(a_k),\ k = 1, \ldots, 20$ der auftretenden Aminosäuren als

$$H(i) = -\sum_{k=1}^{20} P_i(a_k) \cdot \log P_i(a_k). \tag{3.18}$$

Eine Erweiterung der Shannonschen Entropie auf Paare von Positionen stellt die gemeinsame Entropie zweier Spalten i und j dar

$$H(i,j) = -\sum_{k=1}^{20} \sum_{l=1}^{20} P_{ij}(a_k, a_l) \cdot \log P_{ij}(a_k, a_l) \tag{3.19}$$

die sich aus den gemeinsamen Wahrscheinlichkeiten $P_{ij}(a_k, a_l)$ für das Vorkommen der Aminosäuretypen a_k an Position i und a_l an Position j berechnet.

Die Transinformation $M(i,j) \geq 0$ (*Mutual Information*) zweier Spalten i und j im MSA wird berechnet als

$$M(i,j) = H(i) + H(j) - H(i,j) \tag{3.20}$$

und nimmt ihr Maximum an, falls die Spalten i und j völlig synchron mutieren. Nach Äquivalenzumformungen lässt sich $M(i,j)$ auch schreiben als

$$M(i,j) = -\sum_{k=1}^{20}\sum_{l=1}^{20} P_{ij}(a_k, a_l) \cdot \log\left(\frac{P_{ij}(a_k, a_l)}{P_{ij}(a_k) \cdot P_{ij}(a_l)}\right). \tag{3.21}$$

In dieser Form erkennt man die Verwandtschaft zu Chancenquotienten, die das logarithmierte Verhältnis der beobachteten Häufigkeit zur erwarteten Häufigkeit unter der Annahme statistischer Unabhängigkeit ausdrücken.

Rohe Transinformationswerte zeigen die Koevolution jedoch nur unpräzise an [123]. An synthetischen MSAs haben sich normierte M-Werte wie $\frac{M(i,j)}{H(i,j)}$ oder $\frac{M(i,j)}{H(i)+H(j)}$ performanter gezeigt. Analog zu [88] wird in dieser Arbeit $U(i,j)$ [127] verwendet, ein normiertes Maß, das berechnet wird nach

$$U(i,j) = 2\,\frac{H(i) + H(j) - H(i,j)}{H(i) + H(j)}. \tag{3.22}$$

Es gilt $0 \leq U(i,j) \leq 1$. Falls die beiden Spalten i und j vollkommen unabhängig voneinander sind, so gilt $H(i,j) = H(i) + H(j)$ und daher $U(i,j) = 0$. Wenn die beiden Spalten dagegen vollkommen abhängig voneinander sind, so gilt $H(i) = H(j) = H(i,j)$ und deshalb $U(i,j) = 1$ [127]. Ein hoher Wert $U(i,j)$ deutet somit auf ein starkes gemeinsames Vorkommen weniger Paare von Aminosäuren in den Spalten i und j hin, relativ zur Konserviertheit der einzelnen Spalten, die sich in den Einzelentropien $H(i)$ und $H(j)$ ausdrückt.

Die Wahrscheinlichkeiten $P_i(a_k)$, $P_j(a_l)$ und $P_{ij}(a_k, a_l)$ zur Berechnung der Entropien $H(i)$, $H(j)$ und $H(i,j)$ können direkt aus dem MSA als $f_i(a_k)$ bzw. $f_{ij}(a_k, a_l)$ abge-

schätzt werden. Aufgrund der begrenzten Größe des MSAs stellt sich hier das gleiche Problem wie in Abschnitt 3.4.1. An manchen Positionen bzw. Positionspaaren kommen manche Aminosäuretypen bzw. Paare von Aminosäuretypen nicht vor, so dass ihre geschätzte relative Häufigkeit $f_i(a_k)$ bzw. $f_{ij}(a_k, a_l)$ jeweils null ist. In diesen Fällen sind die Entropieterme (3.18) und (3.19) nicht berechenbar.

Um dieses Problem zu lösen, werden die geschätzten Einzelhäufigkeiten $f_i(a_k)$ durch *Pseudocounts* nach (3.10) korrigiert. Bei der Korrektur der paarweisen relativen Häufigkeiten $f_{ij}(a_k, a_l)$ werden in (3.23) die Häufigkeiten beider Spalten korrigiert:

$$f_{ij}(a_k, a_l) = \frac{n_{ij}(a_k, a_l) + \lambda \left(\sum\limits_{\substack{m=1 \\ m \neq k}}^{20} \frac{n_{ij}(a_m, a_l) \cdot B(a_m, a_l)}{\sqrt{n(i,j)}} + \sum\limits_{\substack{m=1 \\ m \neq l}}^{20} \frac{n_{ij}(a_k, a_m) \cdot B(a_k, a_m)}{\sqrt{n(i,j)}} \right)}{n_{ij}(a_k, a_l) + 2 \cdot \lambda \sqrt{n(i,j)}}. \quad (3.23)$$

Zu beachten ist, dass im Falle von $f_{ij}(a_k, a_l)$ 400 paarweise Häufigkeiten abgeschätzt werden müssen. Dies gelingt nur, wenn das MSA eine ausreichende Anzahl von Sequenzen enthält. Es wurde gezeigt, dass 125 Sequenzen notwendig sind um über *Transinformation* korrelierte Mutationen sinnvoll zu bestimmen [123]. Es gilt jedoch grundsätzlich: Je mehr Sequenzen ein MSA beinhaltet, umso besser können die Häufigkeiten abgeschätzt werden und umso besser ist die zu erwartende Performanz, solange man nicht die Qualität der MSAs auf Kosten ihrer Größe verringert.

3.6 Berechnung der Proteinoberfläche

Für mehrere der in den folgenden Absätzen beschriebenen Ansätze müssen die Aminosäuren eingeteilt werden in solche, die an der Oberfläche liegen und in diejenigen, die im Kern des Proteins vorkommen. Diese Einteilung erfolgt über die Berechnung der lösungsmittelzugänglichen Fläche (*Solvent Accessible Surface Area*, SASA) einer Aminosäure. Die SASA stellt ein Maß für den Anteil einer Aminosäure an der Oberfläche dar. Ihre Berechnung wird in den folgenden Abschnitten im Detail beschrieben.

3.6.1 Berechnung der SASA über DCLM

Numerische Verfahren zur Berechnung der Oberfläche können über die Art der diskretisierten Oberflächendarstellung eingeteilt werden. Dabei wurden bereits Scheiben zylindrischer Oberflächen, Würfel und Punktverteilungen auf kugelförmigen Atomen verwendet [128]. Die Implementation der BALL-Bibliothek [103], die in dieser Arbeit verwendet wird, verwendet ein doppelt kubisches Gitter [128] weshalb sie als *Double Cubic Lattice Method (DCLM)* bezeichnet wird. Sie ist ein schnelles und genaues Verfahren um *SASA*, *van der Waals*-Oberfläche und das Volumen und die Kompaktheit Molekularer Anordnungen zu bestimmen.

DCLM ist eine Variante eines Verfahrens nach Shrake und Rupley [129], das auf Punktverteilungen in atomaren Kugeln beruht. Der Grundgedanke dabei ist es, den Radius eines jeden Atoms durch den Radius eines kugelförmigen Probemoleküls zu erweitern und anschließend durch ein sphärisches Punktgitter, das über eine icosahedrische Tesselation eines kugelförmigen Atoms entsteht, die Oberfläche eines Atoms zu approximieren. Um das Gitter für ein Atom aus dem Molekül zu konstruieren, wird zunächst eine Menge gleich weit voneinander entfernter Punkte auf der Oberfläche des kugelförmigen Atoms erstellt. Da eine perfekt gleichförmige Verteilung dieser Punkte nicht existiert, werden diese näherungsweise gleich verteilt. Anschließend wird für jeden Punkt getestet, ob ihn Nachbaratome einhüllen. Ist dies der Fall, so wird der Punkt nicht mehr der lösungsmittelzugänglichen Oberfläche zugerechnet. Andernfalls ist der Punkt für das Lösungsmittelmolekül zugänglich. Die Anzahl der zugänglichen Punkte multipliziert mit der einem Punkt zugerechneten Oberfläche ergibt die *SASA* des zugehörigen Atoms.

Der größte Nachteil dieses Verfahrens ist, dass es viel Rechenzeit benötigt. In der DCLM-Variante ist es jedoch dadurch beschleunigt, dass zwei kubische Gitter verwendet werden – eines um die Mittelpunkte der Atome zu beschreiben und eines um benachbarte Oberflächenpunkte eines Atoms zu gruppieren. Dies führt zu einer drastisch kürzeren Rechenzeit bei moderatem Speicherbedarf [128]. Außerdem hängt der *DCLM*-Algorithmus nicht von der verwendeten sphärischen Punktverteilung ab. Im ersten kubischen Gitter werden benachbarte Atome im Molekül gruppiert. Ebenso werden im zweiten kubischen Gitter räumlich benachbarte Punkte der Kugeltesselation gruppiert. Auf diese Weise können Gruppen von Atomen und Punkte sehr schnell als nicht-überlappend erkannt werden. Um die Geschwindigkeit weiter zu steigern, wird bei der Prüfung auf Überlapp das Skalarprodukt anstatt des teuer zu berechnenden euklidischen Abstandes benutzt.

3.6.2 Relative SASA

Die Tatsache, dass sich die Seitenketten der Aminosäuren hinsichtlich der Größe und der Anzahl an schweren Atomen stark unterscheiden, muss bei Verwendung der *SASA* als Maß für die Oberflächenzugänglichkeit berücksichtigt werden. Andernfalls wäre beispielsweise ein kleines Glyzin, das so weit wie möglich an der Oberfläche platziert ist, weniger exponiert als ein großes Tryptophan, von dem nur 3 seiner 13 schweren Atome Zugang zum Lösungsmittel besitzen. Aus diesem Grund wird üblicherweise ein normiertes Maß $SASA_{rel}$ benutzt, das nach (3.24) berechnet wird, indem der über *DCLM* ermittelte Wert einer Aminosäure aa_k durch den maximal möglichen Wert für diesen Typ von Aminosäure $SASA_{ref}(aa_k)$ geteilt wird:

$$SASA_{rel}(aa_k) = \frac{SASA_{DCLM}(aa_k)}{SASA_{ref}(aa_k)}. \tag{3.24}$$

Als Referenzwerte wurden in dieser Arbeit die in [130] berechneten Werte verwendet. Wie man in Tabelle 3.3 sieht korrelieren diese Werte gut mit den Werten aus anderen Publikationen [130], [131], [132], [21].

3.6.3 Reduzierte Oberfläche

Eine weitere Möglichkeit zur Berechnung der *SASA* eines Atoms ist die Triangulation der RS. Da ihre Berechnung jedoch höheren Rechenaufwand erfordert als das *DCLM*-Verfahren, wird sie in dieser Arbeit nicht zur Berechnung der *SASA* benutzt. Dagegen eignet sich die Triangulation der RS gut um atomare Abstandkriterien zu definieren. Daher wird diese Triangulation bei der Berechnung von Kern- und Randbereich einer Kontaktfläche nach Abschnitt 3.7.1 benutzt.

3.6.3.1 Definitionen zur Reduzierten Oberfläche

Zur Definition der *reduzierten Oberfläche* eines Atoms benutzen *Sanner* und *Olson* [133] eine kugelförmige Probe mit Radius r_p, die über die Oberfläche der Atomkugeln abrollt. Wichtig ist bei diesem Verfahren das Konzept der stabilen Position. Die Probe sei in einer stabilen Position, sobald sie drei oder mehr Atomkugeln berührt wie in Abbildung 3.3 gezeigt. Verbindet man die Mittelpunkte der Atome 1-4 mit Kanten, so entstehen die sogenannten *RS-Flächen*. Für den Fall, dass die Probe in stabiler Lage

AA	Miller [130]	Bolser [131]	Chothia [132]	BALL [21]
Ala	113	117	115	118,692
Cys	140	147	135	147,997
Asp	151	163	150	158,295
Glu	183	200	190	190,091
Phe	218	222	210	216,866
Gly	85	95	75	87,201
His	194	208	195	195,977
Ile	182	180	175	190,263
Lys	211	214	200	224,539
Leu	180	185	170	194,365
Met	204	206	185	209,044
Asn	158	171	160	167,964
Pro	143	147	145	151,787
Gln	189	199	180	194,773
Arg	241	255	225	259,556
Ser	122	128	115	126,08
Thr	146	152	140	151,038
Val	160	164	155	165,203
Trp	259	271	255	255,299
Tyr	229	237	230	236,13

Tabelle 3.3: Aminosäurespezifische Referenzwerte der *SASA*
Die Angaben der aminosäurespezifischen maximal möglichen *SASA* nach *Miller*, *Bolser* und *Chothia* korrelieren gut mit den durch *BALL* berechneten Werten. Alle Werte sind in $Å^2$ angegeben.

3 Atomkugeln berührt, ist diese Fläche ein Dreieck wobei die drei Mittelpunkte der Atomkugeln $(c_i, 1 \leq i \leq 3)$ die Eckpunkte bilden.

Falls die Probe um eine Kante herumrollen kann ohne dabei eine dritte Atomkugel zu berühren, so wird diese als *freie Kante* bezeichnet, da diese Kante keiner *RS-Fläche* angehört (vgl. Abbildung 3.4 (b)). Analog dazu wird als freier Knoten die Sphäre eines Atoms a_x bezeichnet, auf der die Probe beliebig rotieren kann ohne auf eine weitere Atomkugel zu treffen. Für den Abstand ihres Mittelpunkts c_x zum nächsten Atommittelpunkt c_y gilt also $d(c_x, c_y) \geq r_x + 2 \cdot r_p + r_y$ mit den beiden Atomradien r_x und r_y.

Als *RS-Komponente* definiert man eine Menge K von *RS-Flächen*, die derart angeordnet sind, dass jede *RS-Fläche* mindestens über eine ihrer *RS-Kanten* mit einer anderen *RS-Fläche* aus K verbunden ist (siehe Abbildung 3.4 (a)). Für allgemein verteilte Atome ohne Spezialfälle wie in Abbildung 3.4 sind sämtliche *RS-Flächen* Dreiecke und es entsteht eine Oberflächentriangulation. Sonderfälle, für die über die beschriebene Me-

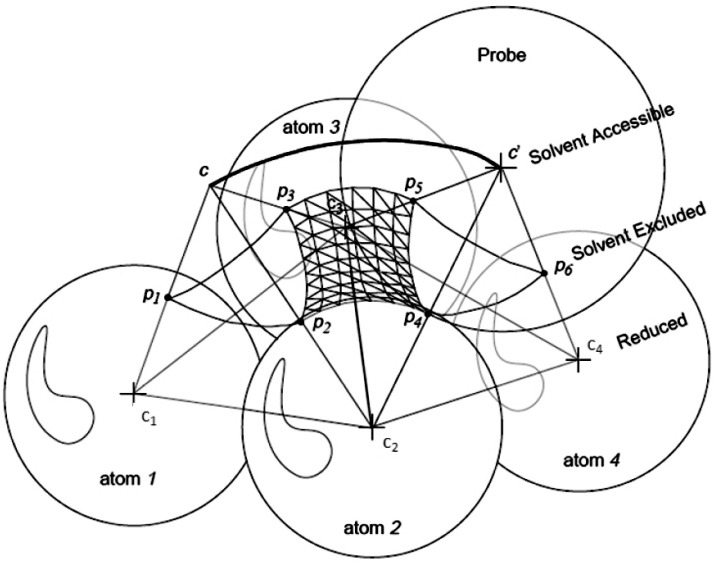


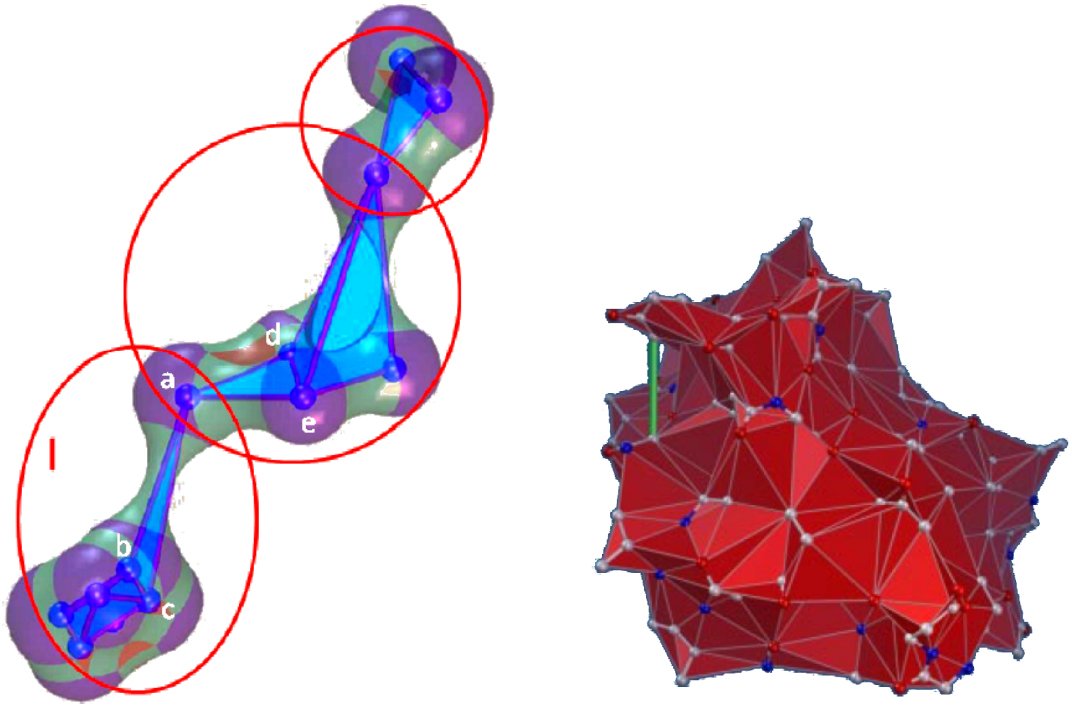
Abbildung 3.3: Berechnung der reduzierten Oberfläche – Abrollende Probe
Eine kugelförmige Probe rollt über die Atomkugeln vom Atom 2 und Atom 3, die über die Kante (c_2, c_3) miteinander verbunden sind und trifft dabei auf die Sphäre von Atom 4. Die eingezeichneten Punkte $p_{i,1\leq i\leq 5}$ sind die Berührungspunkte der Probenkugel mit den Atomkugeln in der stabilen Position, die bei der Berechnung der Connollyoberfläche von Bedeutung sind (Abbildung aus [133]).

thode keine vollständige Triangulation entsteht, müssen mit aufwändigeren Verfahren gesondert behandelt werden.

3.6.3.2 Berechnung der reduzierten Oberfläche

Die reduzierte Oberfläche wird wie folgt berechnet. Es seien n Atomkugeln in allgemeiner Lage gegeben und eine Probe mit Radius r_p. Bei der Initialisierung des Algorithmus wird aus den Mittelpunkten der Atome c_1, c_2 und c_3, zwischen denen der Probenball zu liegen kommt, ein Dreieck gebildet (siehe Abbildung 3.3 auf Seite 37). Anschließend muss die Probe zu einem weiteren Atom a_4 hin abrollen. Dabei verläuft die Bewegung entlang eines Kreisbogens. Als Atom a_4 mit Mittelpunkt c_4 wird unter den Nachbaratomen von a_2 und a_3 dasjenige Atom gewählt, zu dem ein Abrollen der Probe den minimalen Kreisbogen beschreibt. Dabei bildet das Dreieck mit den Eckpunkten c_2, c_3 und c_4 eine neue *RS-Fläche*.

Dieser Schritt 1 wird wiederholt bis alle *RS-Kanten* besucht wurden. Zum Schluss ist eine geschlossene Fläche als *RS-Komponente* entstanden. Anschließend wird in Schritt 2



(a) (b)

Abbildung 3.4: Konstellationen bei der Berechnung der reduzierten Oberfläche
(a) Eine *RS-Komponente* mit drei geschlossenen Oberflächen (rot umrandet). Die Triangulation der reduzierten Oberfläche ist blau eingefärbt. Die restlichen Bereiche geben Bestandteile der Connollyoberfläche an. (b) Das Probenmolekül kann um die grüne, freie Kante rotieren ohne ein weiteres Atom zu berühren. (Abbildungen aus [133]).

für jedes Atom a der *RS-Komponente* geprüft, ob ein Atom d existiert, das noch nicht über eine Kante an die *RS-Komponente* angeschlossen, jedoch nah genug ist um von der Probe beim Abrollen über a erreicht zu werden. Abbildung 3.4(a) stellt das Ergebnis dieses Schrittes 2 dar, der ausgehend von der Triangulation I durchgeführt wird. Dabei werden die Atome d und e gefunden und angeschlossen. Wird mindestens eine solche Fläche gefunden, so wird erneut über Schritt 1 iterativ nach Nachbaratomen mit minimalen Kreisbögen gesucht und diese an die *RS-Komponente* angeschlossen. Anschließend werden iterativ Schritt 1 und Schritt 2 für neu hinzugekommene Atome wiederholt.

Falls eine solche Fläche in Schritt 2 nicht gefunden werden kann, so sind 3 Fälle zu unterscheiden:

- **Fall 1:** Es wurde eine freie Kante gefunden um die die Probe ohne Berührung eines dritten Atoms rotiert. Dieser Fall muss gesondert behandelt werden.

- **Fall 2:** Die Probe umschließt stets ein weiteres Atom. In diesem Fall ist die Probe unter Berücksichtigung der eingeschlossenen Atome neu zu platzieren.

- **Fall 3:** Es befindet sich zu keinem Atom a mehr ein Nachbaratom nahe genug, um von der Probe erreicht zu werden. In diesem Fall ist die *RS-Komponente* fertiggestellt. Bisher nicht behandelte Atome bilden dann weitere *RS-Komponenten*, im Extremfall freie Knoten. Auch dieser Fall muss bei der Kern-Rand-Klassifikation gesondert betrachtet werden, da dann der Graph der RS nicht zusammenhängend ist.

3.7 Algorithmen zur Bestimmung von Kern und Rand

In diesem Abschnitt werden Methoden vorgestellt, mit denen auf geometrische Art eine Einteilung der Kontaktfläche nach Kern- und Randbereichen vorgenommen werden kann. Diese werden benötigt um verschiedene Bereiche von Kontaktflächen auf unterschiedliche Merkmale hin untersuchen zu können.

3.7.1 Protein Interface Analyzer (PIA)

Eine geometrische Methode zur Einteilung von PPKs in Kern- und Randbereich ist der *Protein Interface Analyzer (PIA)*, der von *Staudigel und Trenner* im Rahmen ihrer Diplomarbeit entwickelt wurde [134].

Dabei muss zu Beginn die Kontaktfläche nach Abschnitt 3.2 berechnet werden. Um den Rand der Kontaktfläche über die Nachbarschaft zum Außenbereich zu bestimmen soll im nächsten Schritt der Außenbereich definiert werden. Der intuitive Weg den Außenbereich als alle Nicht-Kontaktresiduen an der Oberfläche zu definieren erwies sich dabei als problematisch. Wie Abbildung 3.5 zeigt, kann eine

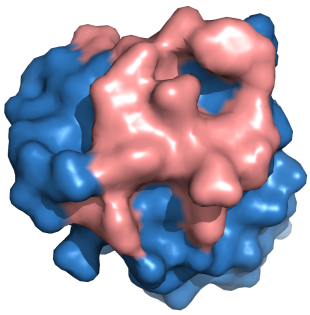


Abbildung 3.5: *1BRS* Kette *D* – Kontaktfläche
Kontaktfläche der Kette D aus $1BRS$ (Barnase-Barstar Komplex). Die Kontaktfläche (rot) enthält ein Loch, das damit zur Nicht-Kontaktfläche (blau) zählt.

Kontaktfläche, die nach Abschnitt 3.2 anhand eines Abstandskriteriums berechnet wurde, Löcher enthalten. Es widerspricht jedoch dem intuitiven Verständnis von “außen”, diese Löcher zum Außenbereich zu zählen.

In Abbildung 3.5 würde eine Definition des Außenbereichs als größte zusammenhängende Nicht-Kontaktfläche zum gewünschten Ergebnis führen. Es können jedoch bei nicht-kanonischen Protein-Protein Komplexen Konstellationen auftreten, in denen die Kontaktfläche größer ist als die restliche Oberfläche. Außerdem sind Fälle denkbar, in denen die restliche Oberfläche in zwei Teilbereiche unterteilt ist.

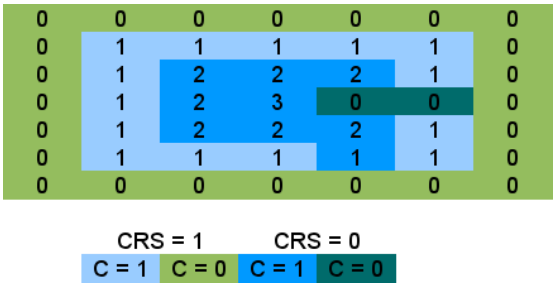


Abbildung 3.6: Einteilung in Kern und Rand nach PIA
Jedes Kästchen entspricht einer Aminosäure. $CRS = 1$ erhalten alle Reste auf der *CRS*. $C = 1$ gilt für Kontaktresiduen.

Um den Außenbereich eindeutig zu bestimmen wird deshalb die gemeinsame reduzierte Oberfläche (*Common Reduced Surface, CRS*) herangezogen, die definiert ist als die RS nach Abschnitt 3.6.3, die an der gebundenen Komplexstruktur berechnet wird. Damit lässt sich der Außenbereich definieren als die Menge aller Reste r, für die folgende Bedingungen erfüllt sind:

- r ist ein Nicht-Kontaktresiduum
- r besitzt mindestens ein Atom, das Teil der *CRS* ist

So ist sichergestellt, dass nur “Löcher” in der Kontaktfläche als einzige Nicht-Kontaktreste nicht zum Außenbereich gezählt werden. Abbildung 3.6 zeigt anhand eines Beispiel die Definition der Schalennummern. Alle Nicht-Kontaktreste, unabhängig von ihrer Eigenschaft als *CRS*-Reste erhalten die Nummer 0. Der Abstand in der Triangulation zum nächsten Nicht-Kontaktrest auf der *CRS* bestimmt dann die Schalennummern aller Kontaktreste.

3.7.2 Intervor

In [38] wird das Programm *Intervor* vorgestellt. Es teilt die PPK ähnlich wie PIA in zentralere Bereiche und Randbereiche ein. Allerdings berechnen sich in diesem Fall die

Schalennummern atomweise anhand einer Voronoi-Triangulation als sogenannte *Voronoi Shelling Order* (VSO). Zur Anwendung des Programmes wird die 64Bit Binärdatei benutzt, die unter [135] zur Verfügung gestellt wird. Aus Gründen der Konsistenz wurde auch bei Verwendung der VSO weiterhin die Kontaktfläche nach in Abschnitt 3.2 definiert, anstatt nach der Methode aus [38]. Den Berechnungen von Intervor wurde lediglich die VSO entnommen und über alle Atome an der PPK gemittelt, um ganze Aminosäuren mithilfe der aminosäurespezifischen *VSO* bewerten zu können.

3.8 Hydrophobe Patches

In diesem Abschnitt wird die von *Lijnzaad* entwickelte [136] und von *Bittkowski* reimplementierte [137] Methode *QUITE* vorgestellt, die hydrophobe Patches auf Ebene von Atomen definiert und berechnet. Da Hydrophobizität keine Eigenschaft der Aminosäuren sondern ihrer Atome ist, können auch Atome aus Seitenketten, die aufgrund einer polaren Gruppe gemeinhin als polar gelten, Teil hydrophober Patches sein. Daher ist eine Definition auf Atomebene einer Definition auf dem Level von Aminosäuren vorzuziehen.

3.8.1 Erzeugen einer zusammenhängenden Fläche

Ein hydrophober Patch soll zunächst als zusammenhängendes Stück der Proteinoberfläche aus C- und S-Atomen definiert werden. Zur Definition der Oberfläche wird die lösungsmittelzugängliche Oberfläche (SASA) benutzt, deren Berechnung in Abschnitt 3.6 genau beschrieben wird.

Eine sinnvolle formale Definition des Terms “zusammenhängend” gestaltet sich etwas schwieriger. Geht man in einem ersten Versuch davon aus, dass die SASA-Flächen zweier Atome A und B genau dann zusammenhängend sind, wenn sich ihre erweiterten Atomkugeln K_A und K_B, deren erweiterter Radius R_{ext} sich nach (3.25) als Summe aus *van der Waals*-Radius R_{vdW} und dem Radius des kugelförmigen Probenmolekül ($R_{Prob} = 1,4$ Å) zusammensetzt, so tritt häufig eine Situation wie in Abbildung 3.7 auf, in der die beiden Atomkugeln K_A und K_B von einer dritten Atomkugel K_C geschnitten werden.

$$R_{ext} = R_{vdW} + R_{Prob} \tag{3.25}$$

Die Bedingung, dass sich die Atomkugeln K_A und K_B schneiden ist folglich notwendig jedoch nicht hinreichend dafür dass die *SASA*-Flächen der beiden Atome A und B zusammenhängen. Als Ausgangspunkt für eine hinreichende Bedingung kann die Punktrepräsentation der *SASA* benutzt werden, die während des DCLM-Verfahrens aus Abschnitt 3.6.1 entsteht. Wie in [136] wird auf fundierte Art ein Schwellwert δ geschätzt, den die Länge einer Kante zwischen je einem nicht verdeckten Punkt der *SASA*-Repräsentation des Atoms A und des Atoms B unterschreiten muss, um den Zusammenhang der beiden *SASA*-Flächen zu garantieren.

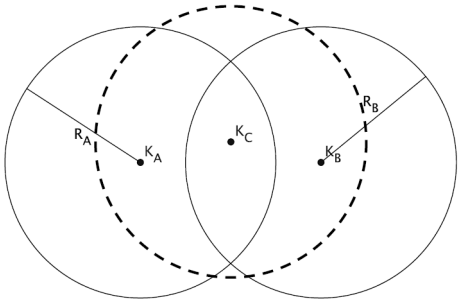
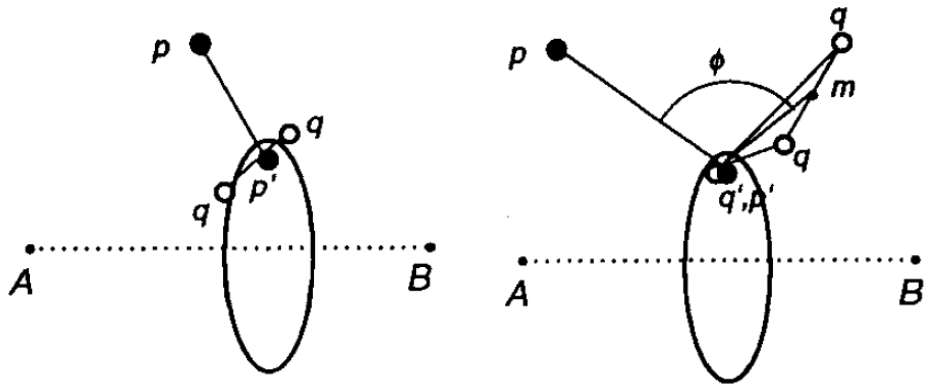


Abbildung 3.7: Nicht-zusammenhängende Oberfläche zweier sich schneidender Kugeln
Nicht-zusammenhängende Oberfläche zweier Kugeln, die sich schneiden (Abbildung aus [137]).

Nichtkontaktfläche. Vor allem im Falle von HisH erhält man einen starken Kontrast zwischen Kontaktfläche und Nichtkontaktfläche. Für HisF finden sich ebenfalls deutlich mehr und stärkere Signale an der Kontaktfläche als an der restlichen Oberfläche, auch wenn der Unterschied hier weniger deutlich ist. Der Grund für das vorhandene schwache Signal an der Nicht-Kontaktfläche von HisF liegt ist vermutlich die Ligandenbindestelle [138], die einen ähnlich hydrophoben Charakter besitzt wie Protein-Protein Kontaktflächen.

Der Schwellwert δ kann über diejenige Anordnung der beiden Kugeln K_A und K_B geschätzt werden, die zu einer maximalen Distanz der sich nächsten nicht verdeckten Punkte P_A und P_B aus den jeweiligen Punktrepräsentationen der *SASA* führt. Wie in Abbildung 3.8 sind dabei 2 Fälle zu unterscheiden. Sei in Fall I (Abbildung 3.8(a) die Strecke $a := pp'$ die maximale Kantenlänge in der Triangulation der benutzten Punktrepräsentation der Einheitskugel. Sei weiter die Strecke $b := qq$ die minimale Kantenlänge in dieser Triangulation und sei die Strecke $pq =: \delta$ die gesuchte maximale Entfernung zwischen zwei Punkten aus den Punktrepräsentationen der sich schneidenden Atome. So stellt Anordnung 3.8(a) den Fall dar, dass der nicht verdeckte Punkt p einen Nachbarpunkt p' der Punktrepräsentation von Atom A besitzt, der verdeckt wird. Außerdem halbiert p' die Strecke zwischen zwei benachbarten nicht verdeckten und äquivalenten Punkten q und q. Dann kann weder die Distanz von p zum Schnittkreis vergrößert werden (da pp als die maximale Kantenlänge gewählt wurde) noch die Distanz von q zum Schnittkreis verkleinert werden (da qq als die minimale Kantenlänge gewählt wurde).



(a) (b)

Abbildung 3.8: Herleitung des Schwellwertes δ
Die Abbildung stellt eine Skizze aus [136] zur Herleitung des Schwellwertes δ in schematischer perspektivischer Ansicht dar. Die Atommittelpunkte A und B sind durch kleine Punkte gekennzeichnet. Punkte aus der Punktrepräsentation von Atom A bzw. Atom B sind durch schwarz bzw. weiß gefüllte Punkte repräsentiert. Die Schnittkreise zwischen den Atomen mit erweitertem Radius sind perspektivisch verzerrt als Ellipsen skizziert.

Wird die Distanz von p zum Schnittkreis verringert, so verringert sich auch die Strecke pq. Dies geschieht auch, falls die Distanz von q zum Schnittkreis entlang der Achse pp' vergrößert wird. Daraus folgt, dass die Strecke pq maximal ist. Sie kann über den Satz von Pythagoras bestimmt werden als

$$\delta_I^2 = a^2 + \frac{1}{4}b^2. \tag{3.26}$$

Abbildung 3.8(b) ergibt sich für Fall II, in dem p und zwei äquivalente Punkte q und q zwei Nachbarpunkte p' und q' besitzen, die zusammenfallen und verdeckt sind. p und der Mittelpunkt m von qq befinden sich dann in maximaler Entfernung zum Schnittkreis.

Unter der Annahme aus [136], dass das Dreieck QPQ aufgrund seiner Lage in der triangulierten Einheitskugel gleichseitig ist, lässt sich $\delta_{II}^2 = pq$ mithilfe des Kosinussatzes nach Abbildung 3.9 berechnen durch

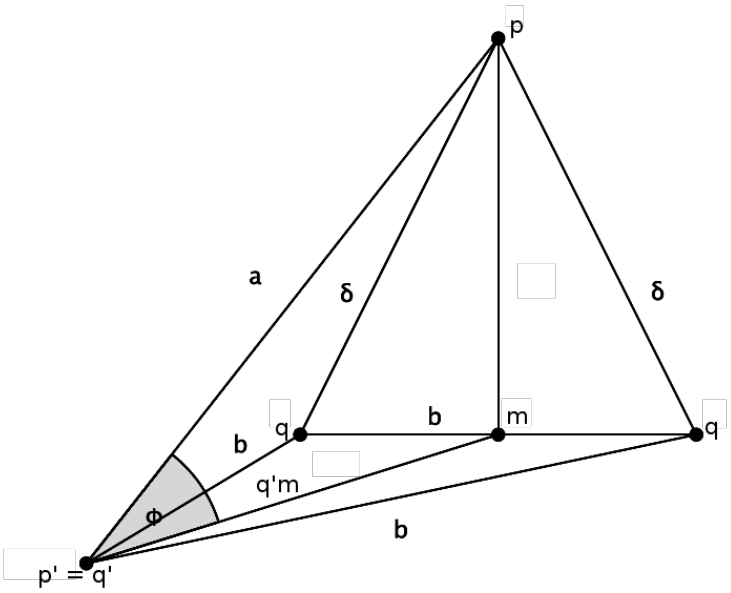


Abbildung 3.9: Dreiecke zur Herleitung von δ_{II}^2
δ_{II}^2 lässt sich über den Kosinussatz aus dem Dreieck QPQ berechnen. Die Abbildung stammt aus [137].

$$PM^2 = \delta_{II}^2 - \left(\frac{b}{2}\right)^2 \tag{3.27}$$

$$P'M = Q'M = \frac{b}{2}\sqrt{3} \tag{3.28}$$

$$PM^2 = a^2 + Q'M^2 - 2aQ'M\cos(\phi) \tag{3.29}$$

Wird (3.27) und (3.28) in (3.29) eingesetzt, so erhält man

$$\delta^2 - \left(\frac{b}{2}\right)^2 = a^2 + \left(\frac{b}{2}\sqrt{3}\right)^2 - 2a\left(\frac{b}{2}\sqrt{r}\right)\cos(\phi). \tag{3.30}$$

Umformung und Auflösen nach δ_{II}^2 ergibt

$$\delta_{II}^2 = a^2 + b^2 - \sqrt{3}ab\cos(\phi). \tag{3.31}$$

Der Winkel ϕ spannt einen Keil auf, der perfekt zwischen die beiden Kugeln passt und wird nach [136] aus der Summe zweier Winkel ψ und χ geschätzt (vgl. Abbildung 3.10)

$$\psi = \pi - \arccos\left(\frac{r_a^2 + r_b^2 - d_{ab}^2}{2r_a r_b}\right) \tag{3.32}$$

$$\chi = \frac{1}{2}\alpha + \arccos\left(\frac{\cos(\beta)}{\cos\left(\frac{1}{2}\beta\right)}\right). \tag{3.33}$$

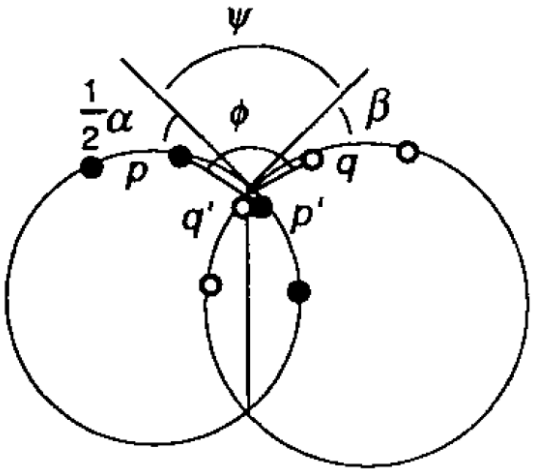


Abbildung 3.10: Skizze zu den Winkeln ϕ und χ
ϕ kann über die Winkel ψ und χ genähert werden (aus [136]).

Der Winkel ψ lässt sich wiederum über den Kosinussatz berechnen, während χ sich mithilfe sphärischer Trigonometrie bestimmen lässt. r_a und r_b in (3.32) bezeichnen die erweiterten Radii der Atome A und B und d_{ab} stellt den euklidischen Abstand zwischen den Mittelpunkten dieser beiden Atome dar. α und β aus (3.33) bezeichnen die mittlere Bogenlänge der kleinsten bzw. größten benutzten triangulierten Einheitskugel. Anschließend wird δ bestimmt als

$$\delta = \sqrt{\max(\delta_I^2, \delta_{II}^2)}. \tag{3.34}$$

3.8.2 Eliminierung zu kleiner Patches

Nachdem im letzten Abschnitt eine formale Definition für eine zusammenhängende Fläche gefunden wurde, kann nun für alle nach 3.6.1 nicht komplett verdeckten apolaren Atome nach anderen unverdeckten apolaren Atomen gesucht werden, mit denen sie eine zusammenhängende *SASA*-Fläche aufspannen. Die hydrophoben Cluster ergeben sich dann als diejenigen apolaren Atome, die (meist über mehrere andere apolare

Atome) zusammenhängen. Um allzu kleine hydrophobe Patches, denen keine physikalische Rolle zukommt, zu vermeiden, ist es sinnvoll für die hydrophoben Patches eine Mindestgröße festzulegen. Als vernünftiger Schwellwert hat sich eine SASA von mindestens $P_{min} = 500$ Å herausgestellt. Dieser Schwellwert entspricht in etwa der Größe von zwei vollkommen exponierten *Tryptophanen* an der Oberfläche. Besitzt ein hydrophober Patch weniger *SASA* als P_{min} angibt, so wird er verworfen.

3.8.3 Die polare Extension

Da alle Aminosäuren zu einem Großteil aus Kohlenstoff bestehen, ist auch der Anteil apolarer Atome an der Oberfläche von Proteinen relativ hoch (48–65% im Datensatz von *Lijnzaad* [136]). Aus diesem Grund würde in den allermeisten Fällen ein einziger großer hydrophober Patch pro Proteinkette gefunden werden [139]. In solch riesigen hydrophoben Patches, die sich über die ganze Kette erstrecken, werden großflächige Bereiche durch kleine *Kanäle* miteinander verbunden. Die Idee ist es nun, durch eine Erweiterung des Radius polarer Atome, die sogenannte *polare Extension* PE, diese Kanäle so weit zu verengen, dass sie verschwinden, womit die großflächigen Bereiche isoliert und somit als eigene Patches erkannt werden.

Durch die *polare Extension* entledigt man sich jedoch nicht nur der unerwünschten *Kanäle*. Es werden auch die großflächigen Bereiche, die als hydrophobe Patches übrig bleiben sollen, beschnitten. Daher werden in [136] alle Patches, die nach der *polaren Extension* übrig geblieben sind, um diejenigen hydrophoben Atome erweitert, die direkt mit einem der Atome aus dem Patch verbunden sind.

3.9 Häufigkeitsverteilungen von Aminosäuren

Chancenquotienten werden allgemein berechnet nach

$$L = \log\left(\frac{f_1}{f_2}\right). \tag{3.35}$$

Dabei stellen f_1 und f_2 die relativen Häufigkeiten bestimmter Ereignisse dar. Ist die Häufigkeit im Zähler von (3.35) f_1 größer als diejenige im Nenner f_2, so ist der Bruch > 1 und damit $L > 0$. Im umgekehrten Fall ist der Wert des Bruches < 1 und deshalb $L < 0$. In dieser Arbeit werden die beiden Chancenquotienten PW_{pair_inter} und

PW_{pair_intra} auf der Grundlage des Datensatzes von Protein-Protein Komplexen aus der Arbeitsgruppe von *R. Nussinov* [99] berechnet.

Zum Vergleich der Häufigkeiten eines Seitenkettentyps aa_i mit $i = 1, \ldots, 20$ an verschiedenen Regionen von Proteinen ersetzt man in (3.35) f_1 und f_2 durch die Häufigkeiten $f_1(aa_i)$ und $f_2(aa_i)$ der Seitenketten in den Regionen, die man vergleichen will. Die gefundenen Scores sind positiv falls der Seitenkettentyp i in Region 1 gegenüber Region 2 bevorzugt ist und negativ für den umgekehrten Fall.

Der intermolekulare Chancenquotient $PW_{pair\_inter}(aa_i, aa_j)$ soll die beobachtete Häufigkeit eines Kontaktpaares aus Aminosäuretyp aa_i mit Aminosäuretyp aa_j relativ zu der Häufigkeit vergleichend bewerten, die man unter der Annahme erwarten würde, dass die Aminosäuren unabhängig von ihrem Typ miteinander interagieren. Deshalb ersetzt man f_1 durch $f_{pair\_inter}(aa_i, aa_j)$, der beobachteten Häufigkeit eines Kontaktpaares. Die erwartete Häufigkeit f_2 wird als das Produkt der beiden Häufigkeiten, mit denen Aminosäuretyp aa_i und Aminosäuretyp aa_j an der Oberfläche auftreten, bestimmt. Dabei ist das Produkt der Einzelhäufigkeiten aufgrund der Alternativhypothese unabhängiger Interaktion zu wählen. Somit ergibt sich für alle $i, j = 1, \cdots, 20$

$$PW_{pair\_inter}(aa_i, aa_j) = \log\left(\frac{f^{cont}_{pair\_inter}(aa_i, aa_j)}{f^{surf}(aa_i) \cdot f^{surf}(aa_j)}\right). \tag{3.36}$$

In analoger Weise wird der zweite Chancenquotient PW_{pair_intra} definiert, der die Unterschiede in den Häufigkeiten intramolekularer Kontakte an der Kontaktfläche und an der restlichen Oberfläche bewertet. Hier lässt sich die Referenzhäufigkeit direkt an der restlichen Oberfläche der Proteinketten abschätzen und wird nicht als Produkt der Einzelhäufigkeiten bestimmt. Deshalb definiert man den intramolekularen paarweisen Häufigkeitsscore als

$$PW_{pair\_intra}(aa_i, aa_j) = \log\left(\frac{f^{cont}_{pair\_intra}(aa_i, aa_j)}{f^{surf}_{pair\_intra}(aa_i, aa_j)}\right). \tag{3.37}$$

Alle benötigten Häufigkeiten lassen sich aus einem Datensatz von Protein-Protein Kontaktflächen ableiten. Dazu müssen zuerst Kontaktpaare, Kontaktflächen und Oberfläche nach den Methoden aus den Abschnitten 3.2 und 3.6.2 bestimmt werden. Sowohl inter- als auch intramolekular ist dabei der maximale Abstand zweier Kontaktaminosäuren als Parameter $s^{(abl)}_{PW_{pair\_intra}}$ bzw. $s^{(abl)}_{PW_{pair\_inter}}$ zu wählen. Anschließend lassen

sich die absoluten inter- und intramolekularen Kontakthäufigkeiten $f^{cont}_{pair\_inter}(aa_i, aa_j)$, $f^{cont}_{pair\_intra}(aa_i, aa_j)$, $f^{surf}_{pair\_intra}(aa_i, aa_j)$ ebenso bestimmen wie die Einzelhäufigkeiten an der Oberfläche $f^{surf}(aa_i)$.

Durch Normieren nach der entsprechenden Gesamtzahl gezählter Kontakte bzw. Oberflächenaminosäuren N_{pair_inter}, N_{pair_intra} bzw. n^{surf} erhält man die relativen Häufigkeiten, die in (3.36) und (3.37) benötigt werden. Da bei der Berechnung von $f^{cont}_{pair\_inter}$ und $f^{cont}_{pair\_intra}$ die Reihenfolge der interagierenden Aminosäuren keine Rolle spielt, ist es angebracht diese relativen Häufigkeiten symmetrisch zu definieren:

$$f^{surf}(aa_i) = \frac{F^{surf}(aa_i)}{n^{surf}} \tag{3.38}$$

$$f^{cont}_{pair\_inter}(aa_i, aa_j) = \frac{F^{cont}_{pair\_inter}(aa_i, aa_j) + F^{cont}_{pair\_inter}(aa_j, aa_i)}{2 \cdot N_{pair\_inter}} \tag{3.39}$$

$$f^{cont}_{pair\_intra}(aa_i, aa_j) = \frac{F^{cont}_{pair\_intra}(aa_i, aa_j) + F^{cont}_{pair\_intra}(aa_j, aa_i)}{2 \cdot N_{pair\_intra}} \tag{3.40}$$

Die Auswertung der Scoringtabelle PW_{pair_inter} für Aminosäurekontaktpaare über das Interface hinweg, gestaltet sich trivial. Der intermolekulare Score zur Bewertung eines Kontaktes S_{pair_inter}, der nach der Methode aus Abschnitt 3.2 berechnet wurde, kann nach Bestimmung der Aminosäureart der beiden Kontaktpartner an der Matrix als $S_{pair\_inter} = PW_{pair\_inter}(aa_i, aa_j)$ abgelesen werden. Der Score PW_{pair_intra} dagegen wurde nicht zur Bewertung von Kontaktpaaren berechnet, sondern zur Bewertung einzelner Aminosäuren an der Oberfläche anhand ihrer intramolekularen Umgebung. Die Berechnung des Scores S_{pair_intra} einer Oberflächenaminosäure anhand der Chancenquotienten aus PW_{pair_intra} wird in den nächsten Absätzen näher erläutert.

Im Folgenden wird Information aus der Struktur und aus einem MSA miteinander kombiniert. Dabei bezeichnet der Begriff “Position” sowohl eine Aminosäure in der Struktur des Proteins als auch die zugehörige Spalte im MSA. Zur Berechnung von PW_{pair_intra} für eine Position p sind als erstes alle $n(p)$ räumlichen Nachbarpositionen q_k mit $k = 1, \ldots, n_p$ nach dem Abstandskriterium (3.8) mit $s^{(anw)}_{pair\_intra}$ als noch zu optimierendem Parameter innerhalb der Oberfläche derselben Proteinkette zu bestimmen. Anschließend lässt sich der Score $S_{pair\_intra}(p)$ der Position p durch Mittelung über alle gefundenen Kontaktpaare berechnen:

$$S_{pair\_intra}(p) = \frac{1}{n(p)} \sum_{k=1}^{n(p)} PW_{pair\_intra}(Type(p), Type(q_k)). \tag{3.41}$$

Dabei gibt die Funktion $Type(p)$ den Typ der Aminosäure an Position p aus.

Falls ein paarweise geordnetes MSA (siehe Abschnitt 3.3) zur Verfügung steht, so kann man sowohl bei der Bestimmung von S_{pair_inter} als auch von S_{pair_intra} den Score für jedes Positionenpaar p und q anhand einer paarweise positionsspezifischen 20×20 Häufigkeitsmatrix $P_{pq}(i,j)$ mit $i,j = 1,\ldots,20$, abgeleitet aus dem MSA, mitteln. Bei der Abschätzung der Häufigkeiten aus dem MSA sind nur diejenigen Sequenzen zu berücksichtigen, die weder an Position p noch an Position q eine Lücke aufweisen. Damit berechnen sich PW_{pair_inter} und PW_{pair_intra} als

$$S_{pair\_inter}(p,q) = \sum_{i=1}^{20}\sum_{j=1}^{20} P_{pq}(aa_i, aa_j) \cdot PW_{pair\_inter}(aa_i, aa_j) \tag{3.42}$$

$$S_{pair\_intra}(p) = \frac{1}{n_p} \sum_{k=1}^{n_p}\sum_{i=1}^{20}\sum_{j=1}^{20} P_{pq_k}(aa_i, aa_j) \cdot PW_{pair\_intra}(aa_i, aa_j) \tag{3.43}$$

3.10 Konnektivität

Bei der Bewertung einzelner Reste durch Scores für Paare $S^{pair}_{p_k q_l}$ von Positionen p_k aus dem Protein, für das eine Vorhersage generiert werden soll, und q_l aus dem Interaktionspartner müssen alle paarweisen Kontaktscores einer Position in einen Wert für die einzelne Position p_k verrechnet werden. Eine Möglichkeit dazu ist die Konnektivität. Man kann die Positionen beider Kontaktketten als Knoten in einem Netzwerk verstehen (siehe Abbildung 3.11), in dem alle Knotenpaare mit hohem Score $S^{pair}_{p_k q_l}$ durch eine Kante miteinander verbunden sind. Die Konnektivität ist dann für jede Position als die Anzahl an Kanten definiert, die diese Position erreichen. So ist es möglich, diejenigen Positionen zu identifizieren, die zum Interaktionspartner sehr viele Kopplungen besitzen. Eine hohe Anzahl an Kopplungen ist ein Indiz für eine physikalische Interaktion mit der anderen Untereinheit und damit für eine Lage an der PPK.

Um die Konnektivität in der Support Vektor Maschine als zusätzliches Feature verwenden zu können, muss Vergleichbarkeit der Werte für verschiedene Protein-Protein Komplexe sichergestellt sein. Unter der Annahme, dass für jedes Protein etwa derselbe

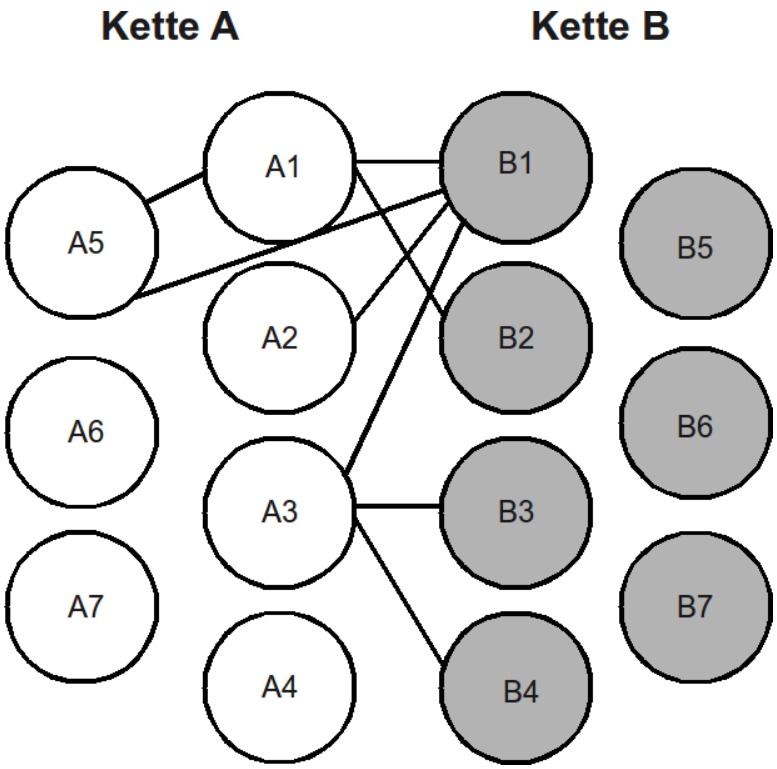


Abbildung 3.11: Konnektivität eines Netzwerks aus Aminosäuren
Die Abbildung zeigt den Kontaktbereich zweier interagierender Ketten A und B. Die Aminosäuren der Kette A (A1-A7) sind als weiße Kugeln dargestellt, die Aminosäuren der Kette B (B1-B7) als graue Kugeln. Scores zur Bewertung intermolekular interagierender Seitenketten werden für alle Aminosäurepaare aus untersschiedlichen Ketten berechnet. Die eingezeichneten Kanten repräsentieren einen hohen Score zwischen den Aminosäuren, die sie verbinden. Die Konnektivität $Conn^{abs}$ einer Aminosäure bestimmt sich als die Anzahl ihrer der Kanten.

feste Bruchteil aller $n_p \cdot n_q$ Positionspaare signifikante Kopplungen enthält, werden für die Berechnung der Konnektivität die $x \cdot n_p \cdot n_q$ größten Kopplungen berücksichtigt. Dabei stellt der Bruchteil x mit $0 < x < 1$ einen Parameter dar, der noch zu optimieren bleibt. Die Konnektivität $Conn_{p_k}^{abs}$ einer Position p_k im Protein, für das eine Vorhersage generiert werden soll, definiert sich folglich als die absolute Häufigkeit des Vorkommens dieser Position p_k unter den $x \cdot n_p \cdot n_q$ größten paarweisen Kopplungen.

Außerdem muss sichergestellt sein, dass die Konnektivität bei verschiedenen Protein-Protein Komplexen gleich skaliert. Nimmt man an, dass alle Kopplungen zufällig erfolgen, so wird der Wert der Konnektivität einer Position p_k mit der Anzahl der Positionen q_l wachsen, die der betrachteten Position p_k als Kopplungspartner zur Verfügung stehen. Um deshalb gleiche Skalierung bei unterschiedlichen Protein-Protein Komplexen sicherzustellen werden die Konnektivitätswerte gemäß der Sequenzlänge seq_2 der Se-

quenz des Partnerproteins normiert

$$Conn_{p_k}^{rel} = \frac{Conn_{p_k}^{abs}}{seq_2}. \tag{3.44}$$

3.11 Gewichtete Mittelung über die Nachbarschaft

Um bei der Berechnung einzelner Eigenschaften auch Informationen aus der Nachbarschaft der zu charakterisierenden Position mit einzubeziehen hat sich eine Methode von *Porollo und Meller* [46] als nützlich erwiesen. Dabei wird der Wert $P_k^{(e)}$ einer Eigenschaft e der Position k gewichtet gemittelt über intramolekulare Nachbarn an der Oberfläche des Proteins. Es hat sich dabei insbesondere eine Gewichtung direkt proportional der *relativen SASA* ($rSASA$) (3.45) und reziprok der Entfernung d_l vom betrachteten Nachbarn l (3.46) als sinnvoll herausgestellt:

$$P_{rSASA}^{(e)} = P_k + \sum_{l=1}^{N} w_{rSASA}^{(e)} \cdot P_l^{(e)} \cdot rSASA_l \tag{3.45}$$

$$P_{dist}^{(e)} = P_k + \sum_{l=1}^{N} w_{dist}^{(e)} \cdot \frac{P_k}{d_l} \tag{3.46}$$

Dabei ist N die Anzahl aller intramolekularen Nachbarn an der Oberfläche des Monomers im Umkreis eines Schwellwertes s. Dieser ist ebenso zu optimieren wie die Gewichte $w_{rSASA}^{(e)}$ bzw. $w_{dist}^{(e)}$.

3.12 Support Vektor Maschinen (SVM)

Support Vektor Maschinen (SVM) trennen einen Satz von Datenvektoren, die in zwei Klassen aufgeteilt sind, mithilfe einer Trennhyperebene derart, dass sich auf jeder Seite der Hyperebene nur Vektoren einer Klasse befinden. Die Hyperebene wird dabei so gewählt, dass der Abstand zu allen Datenpunkten möglichst groß ist. In dieser Arbeit wurde die Implementation einer SVM im Programmpaket *Libsvm* [140] mit einer *RBF*-Kernelfunktion verwendet.

3.12.1 C-Support Vector Classification (SVC)

Der Algorithmus der C-SVC [141] [142] löst bei gegebenen Trainingsvektoren $x_m \in \mathbb{R}, m = 1, \ldots, l$ mit bekannten Label $y_m \in \{-1, 1\}$ das Optimierungsproblem

$$\begin{aligned} \underset{\boldsymbol{w},b,\xi}{\text{minimiere}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{m=1}^{l}\xi_m \\ \text{unter der Nebenbedingung} \quad & y_m(\boldsymbol{w}^T\phi(x_m) + b) \geq 1 - \xi_m \\ & \xi_m \geq 0, m = 1, \ldots, l \end{aligned} \tag{3.47}$$

wobei $\phi(x_m)$ eine Funktion ist, die Datenvektoren x_m in einen hochdimensionalen Raum transformiert. $C \geq 0$ stellt einen Parameter dar, mit dem sich die Bedeutung der sogenannten *Slack Variablen* ξ_m regulieren lässt. *Slack Variable* ξ_m können Datenvektoren x_m, die z.B. aufgrund von verrauschten Signalen fehlerhaft klassifiziert werden, eine geringere Bedeutung beim Training geben. $\boldsymbol{w}$ beschreibt als Normalenvektor auf die Hyperebene die Richtung und zusammen mit b ihre genaue Lage im Raum.

Da $\boldsymbol{w}$ möglicherweise hochdimensional ist, wird anstatt dieses Problems das dazu duale Problem gelöst:

$$\begin{aligned} \underset{\alpha}{\textit{minimiere}} \quad & \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} - \boldsymbol{e}\boldsymbol{\alpha} \\ \text{unter der Nebenbedingung} \quad & \boldsymbol{y}^T\boldsymbol{\alpha} = 0 \\ & 0 \leq \alpha_m \leq C, m = 1, \ldots, l \end{aligned} \tag{3.48}$$

wobei $e = [1, \ldots, 1]^T$, Q eine $l \times l$ positiv semidefinite Matrix mit $Q_{mn} \equiv y_m y_n K(\boldsymbol{x}_m, \boldsymbol{x}_n)$ und $K(\boldsymbol{x}_m, \boldsymbol{x}_n) \equiv \phi(\boldsymbol{x}_m)^T\phi(\boldsymbol{x}_n)$ die Kernelfunktion ist. Anhand dieser Darstellung erkennt man, dass nicht die möglicherweise unendlichdimensionale Transformation $\phi(x_i)$ berechnet werden muss, sondern lediglich eine Kernelfunktion $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Sobald dieses Problem (3.49) gelöst ist, gilt für den Normalenvektor $\boldsymbol{w}$

$$\boldsymbol{w} = \sum_{m=1}^{l} y_m \alpha_m \phi(\boldsymbol{x}_m). \tag{3.49}$$

Nach diesem Trainingsschritt werden $y_m\alpha_m, m = 1, \ldots, l$, die Bezeichnungen der Label, die Supportvektoren und Kernelparameter gespeichert. Diese Daten lassen sich anschließend zum Testen nicht gelabelter Datenvektoren über folgende Entscheidungsfunktion nutzen:

$$\operatorname{sgn}(\boldsymbol{w}^T\phi(\boldsymbol{x}) + b) = \operatorname{sgn}\left(\sum_{m=1}^{l} y_m\alpha_m K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right). \tag{3.50}$$

Als Kernelfunktion fungiert eine radiale Basisfunktion

$$K(\boldsymbol{x}_m, \boldsymbol{x}_n) = \exp(-\gamma \cdot \|\boldsymbol{x}_m - \boldsymbol{x}_n\|^2) \tag{3.51}$$

mit dem zu optimierenden Parameter γ.

3.12.2 Vorverarbeitung der Daten

Da die Komponenten der Inputvektoren aus verschiedenen Parametern von Oberflächenaminosäuren gewonnen werden, sind sie unterschiedlich skaliert. Um eine optimale Performanz der SVM zu gewährleisten, müssen alle Inputdaten in einem Schritt der Vorverarbeitung reskaliert werden. Dazu werden am Trainigsdatensatz zunächst für jede Eigenschaft i mit $i = 1, \ldots, 5$ der maximale max_i und minimale Wert min_i bestimmt. Anschließend werden alle Einträge $val_{s,i},\ i = 1, \ldots, 5$ eines jeden Vektors vec_s nach (3.52) reskaliert, so dass für ihre skalierten Werte gilt $-1 \leq val_{s,i}^{scaled} \leq 1$. Im Detail berechnen sich die skalierten Werte durch:

$$val_{s,i}^{scaled} = \frac{val_{s,i} - min_i}{max_i - min_i}. \tag{3.52}$$

Testdaten müssen ebenfalls in einem Schritt der Vorverarbeitung nach (3.52) mit den gleichen Werten für min_i und max_i reskaliert werden, bevor mit der SVM eine Vorhersage generiert werden kann.

3.12.3 Training der SVM

Beim Training der SVM ist zu beachten, dass der Trainingsdatensatz aus jeder der beiden Klassen gleich viele Trainingsbeispiele enthalten muss. Der Grund für diese Forderung wird deutlich, wenn man annimmt, dass in einem Extremfall der Trainingsdatensatz zu 90% der Klasse 1 entstammt und nur zu 10% der Klasse -1. Falls die SVM z.B. aufgrund einer verrauschten Datengrundlage die Daten nur zu 80% richtig klassifizieren kann, so wird bei der Optimierung eine Hyperebene gefunden, die sämtliche Daten der Klasse 1 zuschlägt. Die Performanz dieser trivialen Klassifikation erreicht 90% Genauigkeit anstatt der 80% möglichen Genauigkeit bei der gewünschten nicht-trivialen Klassifikation.

Um ein solches Verhalten zu vermeiden, werden nicht alle Datenpunkte der größeren Klasse als Trainingsmenge verwendet. Bei der Zusammenstellung des Trainingsdatensatzes werden aus der Klasse mit der größeren Anzahl an Datenpunkten in zufälliger Auswahl genau so viele Trainingsbeispiele gezogen, wie die kleinere Klasse Datenpunkte besitzt.

Außerdem besteht bei SVMs prinzipiell, wie bei allen überwachten Lernverfahren, die Gefahr des sogenannten *Overlearnings*. Bei diesem unerwünschten Phänomen werden die Merkmale einzelner Trainingsbeispiele auswendig gelernt. Anschließend kann der Klassifikator jedoch die Klassenzugehörigkeit für neue Testbeispiele weitaus schlechter vorhersagen als für die bekannten Trainingsbeispiele. Aufgrund des Optimierungskriteriums, das besagt, dass die trennende Hyperebene den Abstand zu sämtlichen Trainingsbeispielen maximiert, ist eine *SVM* jedoch bei hinreichender Größe des Trainingsdatensatzes relativ robust gegenüber *Overlearning*.

3.12.4 Abschätzung der Wahrscheinlichkeit

Wie bisher beschrieben, liefert eine SVM bei der Bewertung unbekannter Objekte lediglich einen binären Wert, der den Datenvektor einer der beiden Klassen zuordnet. Es ist jedoch sinnvoll, die Vorhersage mit einer Wahrscheinlichkeit zu versehen, die angibt, wie zuverlässig die Klassifikation ist. In *Libsvm* [140] wird dazu der Abstand eines Datenpunktes von der Trennhyperebene in eine *a posteriori* Wahrscheinlichkeit verrechnet, mit der der Datenpunkt das Label 1 trägt [143].

3.13 Hierarchisches Clustern

Hierarchisches Clustern ist eine Methode um Datenpunkte anhand eines Abstandsmaßes baumförmig zu clustern, wie in Abbildung 3.12 schematisch dargestellt.

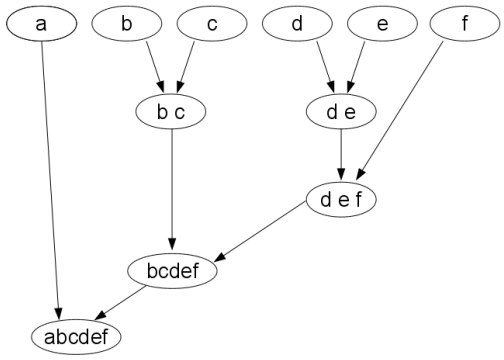


Abbildung 3.12: Schematische Darstellung des hierarchischen Clusterns In jedem Schritt werden die beiden Cluster mit dem geringsten Abstand vereinigt. Dies wird solange fortgesetzt, bis alle Objekte zu einem Cluster gehören.

Seien im Folgenden Datenpunkte $a - f$ gegeben. Im ersten Schritt werden sämtliche paarweisen euklidischen Abstände aller Datenpunkte $a - f$ bestimmt. In der Ausgangskonfiguration wird jedem Datenpunkt ein eigener Cluster zugewiesen. Anschließend werden in jeder Iteration des Clusteralgorithmus die beiden sich nächsten Cluster miteinander vereint. Dabei werden die Abstände zwischen zwei Clustern im Verfahren des *single linkage clustering*, das in dieser Arbeit verwendet wird, anhand der nächsten Mitglieder der beiden Cluster bestimmt. Jeder einzelne Schritt lässt sich durch die Entfernung d_{HC} der beiden neu vereinigten Cluster voneinander charakterisieren. Zum Schluss liegen alle Datenpunkte in einem einzigen großen Cluster vor.

Anschließend muss entscheiden werden, an welcher Stufe des Clusterbaumes abgebrochen werden soll. Dies geschieht in dieser Arbeit sobald d_{HC} einen festen Schwellwert überschreitet.

3.14 Bewertung der Klassifikationsleistung

Für die Evaluation eines Klassifikators wurden mehrere Verfahren entwickelt, die je nach Art des Datensatzes Vor- und Nachteile aufweisen. Im Folgenden werden die 3 in dieser Arbeit verwendeten Methoden zur Bewertung der Güte eines Klassifikators vorgestellt.

3.14.1 Receiver Operating Characteristic (ROC)

Viele Klassifikatoren für ein binäres Entscheidungsproblem liefern nicht nur einen binären Wert, sondern eine Sicherheit, mit der die Vorhersage eintritt. Eine Vorhersage kann dann erst nach Wahl einer Entscheidungswelle generiert werden, die die Sicherheit der Vorhersage überschreiten muss. Nach Wahl der Entscheidungsschwelle kann man zur Evaluation an einem Datensatz mit bekannter Klassenzugehörigkeit mehrere Werte bestimmen, die eine Aussage über die Qualität der Vorhersage treffen.

Wie in Abbildung 3.13(a) dargestellt, unterteilen sich die als positiv vorhergesagten Objekte ihrer wirklichen Klassenzugehörigkeit nach in *wahr positive* (*true positives, TP*) und *falsch positive* (*false positive, FP*) Vorhersagen. Analog dazu unterteilt man die negativ vorhergesagten Beispiele in *falsch negative* (*false negative*) und wahr negative (*true negative*) Vorhersagen.

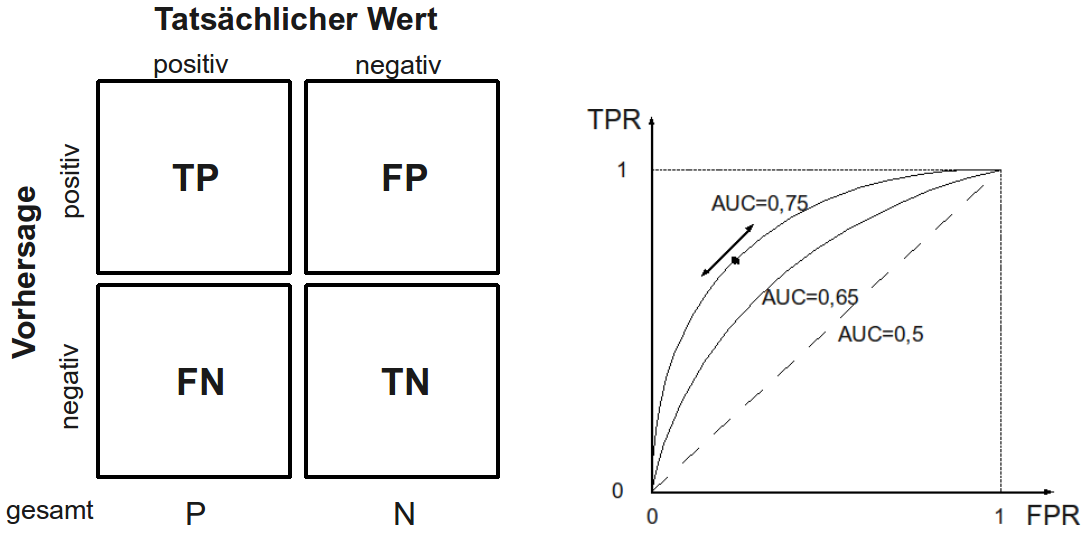


Abbildung 3.13: Receiver Operating Characteristic (ROC)
(a) Die Datenpunkte werden nach ihrem tatsächlichen Wert und der Vorhersage in *wahr Positive* (TP), *falsch Positive* (FP), *falsch Negative* (FN) und *wahr Negative* (TN) eingeteilt. (b) Bei Variation des Schwellwertes lässt sich eine *ROC*-Kurve als Plot der Rate der *wahr positiven* ($TPR = \frac{TP}{TP+FN}$) gegen die Rate der *falsch positiven* ($FPR = \frac{FP}{FP+TN}$) bestimmen.

Aussagekräftiger als absolute Werte sind jedoch Werte relativ zur Gesamtzahl aller getesteten Objekten. Daher definiert man die Rate der *wahr positiven (TPR)* und die Rate der *falsch positiven (FPR)* nach (3.53) und (3.54).

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{3.53}$$

$$FPR = \frac{FP}{P} = \frac{FP}{FP + TN} \tag{3.54}$$

Unter Variation der Entscheidungsschwelle lassen sich *TPR* und *FPR* für unterschiedlich restriktive Klassifikationen berechnen. Die *Receiver Operating Characteristic (ROC)* ist eine Grafik, in der die *TPR* gegen die *FPR* unter Variation der Entscheidungsschwelle aufgetragen ist. *TPR* und *FPR* fallen monoton zwischen dem kleinsten (am wenigsten restriktiven) Wert der Entscheidungsschwelle zum größten (restriktivsten) Wert von 1 auf 0. Falls der Klassifikator lediglich auf dem Niveau einer zufälligen Einteilung bewertet, so verläuft die *ROC*-Kurve als Gerade zwischen (0,0) und (1,1) (siehe Abbildung 3.13(b)). Je besser die Qualität des Klassifikators ist, umso mehr näher sich die *ROC*-Kurve dem Punkt (0,1) an. Daher stellt die *Area under the Curve (AUC)* ein Maß für die Vorhersagegenauigkeit eines Klassifikators dar. Ihr Wert liegt zwischen $0,5$ für einen zufälligen Klassifikator und 1 für einen optimalen Klassifikator der 100% *TPR* bei 0% *FPR* erreicht.

3.14.2 Precision Recall Operating Characteristic (PROC)

Zur Bewertung der Performanz eines Klassifikators werden neben *ROC*-Kurven häufig *Präzision-Trefferquote* (*Precision-Recall*) Kurven (*PROC*) benutzt [144]. In ihnen ist die *Precision* (3.55) gegen den *Recall*, die Äquivalent zur Rate der *wahr positiven* ist (3.53), aufgetragen. PROC-Kurven haben vor allem bei Datensätzen, in denen eine Klasse von weitaus mehr Beispielen repräsentiert wird als die andere, Vorteile gegenüber klassischen *ROC*-Kurven. Angenommen der Testdatensatz beinhaltet um einen Faktor 1000 mehr Negativbeispiele als Positivbeispiele, so sind die TPR und FPR wenig geeignet um die Qualität des Klassifikators zu bewerten. Auch wenn die absolute Anzahl an *falsch positiven* Vorhersagen diejenigen der *wahr positiven* Vorhersagen bei weitem übersteigt, kann daraus ein großer Wert der *AUC* einer *ROC*-Kurve resultieren. Die *Precision* dagegen kann auch bei stark asymmetrischen Datensätzen die Güte eines Klassifikators aussagekräftig bewerten:

$$Precision = \frac{TP}{TP + FP}. \tag{3.55}$$

3.14.3 Matthews Korrelationskoeffizient (MCC)

Ein weiteres, weit verbreitetes Maß zur Bewertung der Qualität eines Klassifikators ist der *Matthews Correlation Coefficient (MCC)* [145]. Der *MCC* (3.56) gibt an, wie stark die Vorhersage des Klassifikators mit der wirklichen Klasseneinteilung korreliert ist. Bei einem Wert von 1 stimmen vorhergesagte und wirkliche Klasseneinteilung vollkommen überein. Ein zufälliger Klassifikator würde einen Wert um 0 erreichen.

$$MCC = \frac{TP\ TN\ -\ FP\ FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \tag{3.56}$$

3.15 Implementation und verwendete Software

Zur Implementation aller strukturbasierten Methoden, wie die Berechnung der *SASA*, der Kontaktfläche und die Bestimmung hydrophober Patches wurde die *BALL-Biochemical Algorithms Library* (BALL) [103] verwendet. Die Verfahren zur Berechnung der Konserviertheit und korrelierten Mutationen wurden in C++ implementiert. Zur Berechnung eines globalen paarweisen Alignments mithilfe des *Needleman-Wunsch Algorithmus* wurde die Implementation der *SeqAn Library* [146] benutzt. Die *libsvm*-Software fand beim Training und Testen der Support Vektor Maschinen Verwendung [140], während hierarchisches Clustern in der Implementation der *C Clustering Library* [147] verwendet wurde. Die Visualisierungen der 3D Proteinstrukturen wurden mit *PyMol* [148] erstellt. Die Auswertung großer Datensätze wurde durch *MPI* parallelisiert. Beim Vergleich mit der Performanz von *ProMate* und *Sppider* wurden die entsprechenden Webinterfaces genutzt [149] [150].

4 Ergebnisse

In diesem Kapitel werden zunächst der innere Aufbau und die Arbeitsweise von *PresCont* erläutert. Anschließend werden alle optimierten Parameter und Trainingsbedingungen dargestellt, die Bedeutung der einzelnen Programmbausteine für die Qualität der Vorhersage untersucht und die Evaluierung der Performanz präsentiert. Zunächst werden jedoch die Datensätze vorgestellt, die dem Training und der Bewertung der Klassifikationsleistung dienten.

4.1 Datensätze und Datenaufbereitung

Für die Entwicklung eines Klassifikators und für die Bewertung seiner Performanz werden umfangreiche Datensätze benötigt. Diese müssen aus großen Datenbanken zusammengestellt werden, die Information über Proteinkomplexe beinhalten. Diese Datensätze müssen einerseits möglichst repräsentativ die Menge der in der Natur vorkommenden Proteine widerspiegeln, dürfen andererseits aber auch keine Redundanzen enthalten. In diesem Abschnitt wird beschrieben, wie die Datensätze ausgewählt wurden.

4.1.1 Bestimmen der Kontaktfläche

Ziel dieser Arbeit ist es, Kontaktaminosäuren anhand von Merkmalen zu indentifizieren, die aus der 2D-Struktur des nicht gebundenen Proteins abgeleitet werden können und solche, die aus zugehörigen MSAs stammen. Zur Bestimmung von Merkmalen, die Positionen an der PPK von Positionen an der restlichen Oberfläche unterscheiden, wird ein Datensatz benötigt, in dem Positionen, die zur PPK gehören, markiert sind. Dieser kann aus bekannten Strukturen von Protein-Protein Komplexen gewonnen werden. Um die PPK bekannter Strukturen zu bestimmen, wurden in der Literatur mehrere prinzipiell unterschiedliche Möglichkeiten vorgeschlagen, Kontaktaminosäuren zu definieren. Die erste Möglichkeit basiert auf einem Maß für die Reduktion der lösungsmittelzugängliche Oberfläche (Solvent Accessible Surface Area, SASA), die den Anteil einer

Aminosäure an der Oberfläche eines Proteins beschreibt. Die zweite Möglichkeit benutzt die Distanz zwischen Resten bzw. Atomen der beiden Moleküle in Kontakt.

Ersteres Kriterium rechnet eine Aminosäure der PPK zu, falls ihre *SASA* gemessen an der Struktur des Komplexes stärker als ein gewisser Schwellwert nach unten von der *SASA* der Monomerstruktur abweicht. Ein häufig verwendeter Schwellwert fordert, dass eine Aminosäure bei der Komplexbildung mindestens 0.1 $Å^2$ an *SASA* verlieren muss um als Kontaktaminosäure gewertet zu werden [28] [37] [151].

Daneben wurden geometrische Verfahren zur Identifizierung von Protein-Protein Kontaktflächen beschrieben, die *Voronoi-Delaunay Tesselation* oder *Alpha Shapes* benutzen [152] [151] [153] [154] [155] [38]. Der Vorteil dieser Verfahren ist es, dass sie keinen frei wählbaren Parameter, wie Schwellwerte für Atomdistanzen oder $\Delta SASA$ benötigen. Allerdings verliert man aber mit diesem Parameter auch Flexibilität bei der Definition der Kontaktfläche. Bei schwellwertbasierten Verfahren besteht die Möglichkeit, je nach Wahl des Schwellwertes die Kontaktfläche mehr oder weniger restriktiv zu definieren.

Das in dieser Arbeit verwendete Kontaktkriterium fordert, dass mindestens ein schweres Atom der Partnerkette innerhalb einer gewissen Entfernung zu mindestens einem schweren Atom der betrachteten Aminosäure liegen muss, um als Kontakt zwischen den Untereinheiten gewertet zu werden. Als Schwellwert für diese Distanz findet man in der Literatur maximal 6 Å gemessen an den Atommittelpunkten der C_β-Atome [18]. Meist wird jedoch ein geringerer Schwellwert verwendet [156] [157] [40] [151]. Häufig wird der Schwellwert auch spezifisch für jede Art und Größe der betrachteten Atome gewählt. Das in dieser Arbeit verwendete intermolekulare Kontaktkriterium definiert zwei Aminosäuren als Kontaktpaar, falls sich die Mittelpunkte mindestens zweier ihrer Atome näher als die Summe aus $0,5$ Å und ihrer Van der Waals Radien sind.

4.1.2 Identifizieren von Oberflächenaminosäuren

Zusätzlich zu den Kontaktaminosäuren muss für den Proteindatensatz bekannt sein, welche Aminosäuren an der Proteinoberfläche liegen. Als Klassifikationskriterium wird hier der relative Anteil der Lösungsmittelzugänglichkeit (rSASA) benutzt. Wie in [46] vorgeschlagen, gilt eine Aminosäure als an der Oberfläche liegend, wenn sie eine (rSASA) von mindestens 5% besitzt. Die Bewertung nach der *rSASA* ist notwendig, da ansonsten Aminosäuren mit großer Seitenkette aufgrund ihres höheren Wertes der maximalen *SASA* häufiger zur Oberfläche gezählt würden als kleinere Aminosäuren. Die SASA wird dabei nach der DCLM-Methode berechnet.

4.1.3 Der Datensatz $\mathbf{Komp_{kanon}}$

In der Arbeitsgruppe von R. Nussinov wurde ein redundanzfreier Strukturdatensatz von Proteinkomplexen zusammengestellt [99]. Dieser Datensatz an PDB-Einträgen enthält 2582 Proteinkomplexe und wird im Folgenden als $Komp_{RN}$ bezeichnet. Ein Teil der Komplexe aus $Komp_{RN}$ kann jedoch nicht für alle auszuführenden Analysen verwendet werden, da z.B. zu wenige homologe Sequenzen bekannt sind oder die Interfaces gewissen Anforderungen nicht genügen. Diese Einträge müssen eliminiert werden, so dass der Datensatz $Komp_{kanon}$ entsteht.

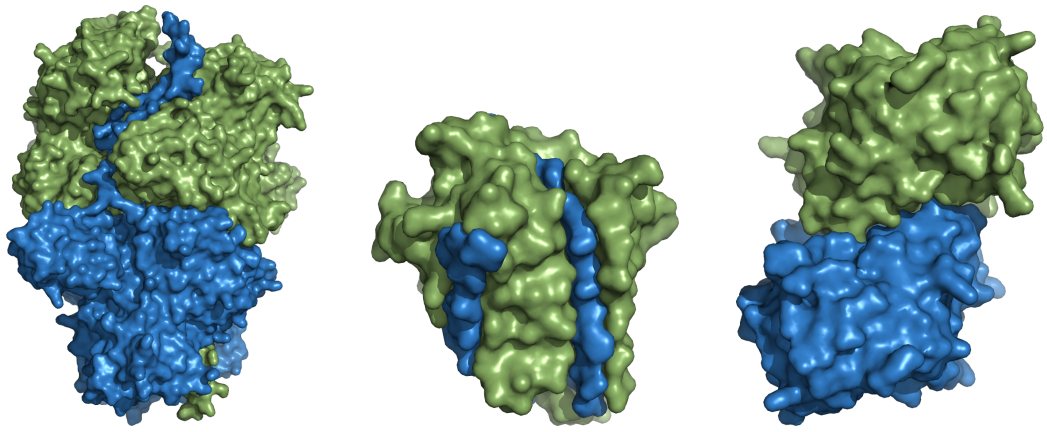
4.1.3.1 Filtern nach Typ der Komplexe und Datengrundlage

PresCont zielt darauf ab, PPKs von wasserlöslichen Proteinen zu identifizieren. Da Membranproteine in einem völlig anderem zellulären Millieue (der Zellmembran) vorkommen, besitzen sie eine anders beschaffene Oberfläche, die mit *PresCont* nicht untersucht werden kann. Daher werden alle offensichtlichen Membranproteine aus dem Datensatz entfernt. Außerdem werden Antigen-Antikörper Komplexe und Virushüllen, für die eine andere Beschaffenheit der Kontaktfläche erwartet wird, eliminiert.

Für die vollständige Analyse wird zu jedem Komplex ein korrespondierendes paarweises multiples Sequenzalignment (MSA) benötigt, das meist der *HSSP*-Datenbank [104] entnommen wird. Aus diesem Grund werden alle Komplexe entfernt, zu denen in der HSSP-Datenbank kein paarweises MSA verfügbar ist. Ein MSA-Paar ist hinreichend groß, wenn nach dem Vergleich der Sequenzen mit Schwellwerten von $id_{min} = 20\%$ und $id_{max} = 90\%$ Sequenzidentität (siehe Abschnitt 3.3) noch mindestens 100 Sequenzen übrig bleiben.

4.1.3.2 Bewerten der Kontaktfläche

Wie Abbildung 4.1 zeigt, sind PPKs zum Teil stark zerklüftet, oder die Proteinketten sind stark ineinander verschlungen. Um bei geometrischen Untersuchungen, die im Folgenden genauer ausgeführt werden, Artefakte zu vermeiden, müssen diese Komplexe aus dem Datensatz entfernt werden. PPKs, die diesen Bedingungen nicht genügen, werden *nicht-kanonische* Kontaktflächen genannt, alle anderen heißen *kanonische* Kontaktflächen.



(a) PDB: 1B0P (b) PDB: 1BQP (c) PDB: 1EQ9

Abbildung 4.1: Kanonische Kontaktflächen und Spezialfälle
Abbildungen (a) und (b) zeigen Spezialfälle nicht-kanonischer Protein-Protein Komplexe (PDB: 1B0P und 1BQP). Abbildung (c) zeigt einen typischen Vertreter eines kanonischen Protein-Protein Komplexes mit planarer Kontaktfläche (PDB: 1EQ9). Die Ketten A und B sind jeweils grün bzw. blau eingefärbt.

Kanonische Kontaktflächen müssen einen gewissen Grad an Planarität aufweisen. Um diesen zu bestimmen, wird zunächst für jede Kette K mit N_K Kontaktaminosäuren eines Protein-Protein Komplexes eine Ebene bestimmt, die die Kontaktfläche approximiert. Liegen mehr als $t \cdot N_K$ mit $0 < t < 1$ weiter als ein Abstandsschwellwert s von der approximierenden Ebene entfernt (siehe Abbildung 4.2), so wird der zugehörige Protein-Protein Komplex als *nicht-kanonisch* verworfen.

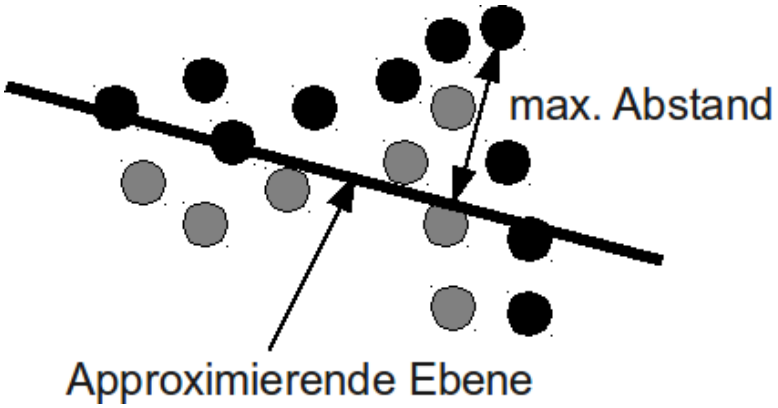


Abbildung 4.2: Skizze zur approximierenden Hyperebene
Der maximale Abstand einer Aminosäure an der PPK ist ein Kriterium für die Planarität der PPK.

Für das Zusammenstellen des Datensatzes $Komp_{kanon}$ wurden die Schwellwerte $s = 6$ Å und $t = 0,4$ verwendet. Da diese restriktiv gewählten Schwellwerte zu viele augenscheinlich kanonische Kontaktflächen verwerfen, wurden anschließend mehrere Komplexe, die

bei visueller Überprüfung als *kanonisch* befunden wurden, $Komp_{kanon}$ hinzugefügt.

$Komp_{kanon}$ besteht aus 64 Protein-Protein Komplexen. Aufgrund des Filterkriteriums korrespondierender MSA-Paare, sind diese 64 Komplexe durchweg Homodimere.

4.2 Kern-Rand Analyse

Die Hydrophobizität der Aminosäurereste ist, wie mehrfach gezeigt wurde [65] [5] bestimmend für die energetische Stabilität von Protein-Protein Interaktionen. Elektrostatische Interaktionen sind die zweite große treibende Kraft, die an Protein-Protein Interaktionen beteiligt ist [158] [159] [160] [161]. Elektrostatische Komplementarität von Proteinoberflächen lenkt die Formation der gebundenen Struktur [162] [163], indem sie die beteiligten Untereinheiten passend orientiert.

Von Chakrabarti wurde des Core-Rim Modell eingeführt [37]. Nach der gängigen Theorie sind Protein-Protein Kontaktflächen aufgebaut aus einem hydrophoben Kern, in dem die wenigen Reste platziert sind, die über hydrophobe Wechselwirkungen den größten Beitrag zur Stabilität des Komplexes leisten, die sogenannten *Hot Spots*. Der hydrophobe Effekt kann jedoch nur auftreten wenn sichergestellt ist, dass kein Wassermolekül in den Kern einzudringen vermag. Dafür sorgt ein O-Ring, der den hydrophoben Kern umgibt und eine ähnlich hydrophile Aminosäurezusammensetzung besitzt wie der Rest der Oberfläche.

4.2.1 Methoden zur Berechnung von Kern und Rand

Um zu bestimmen, welche Aminosäuren bevorzugt im Zentralbereich einer PPK vorkommen, muss zunächst eine sinnvolle, allgemein gültige Definition des Kernbereichs einer PPK gefunden werden. Dies ist aufgrund der vielen zerklüfteten und teils komplexen Formen, die PPKs im Raum einnehmen keine triviale Aufgabe. In früheren Arbeiten wurde der Begriff “Kern” meist über die Lösungsmittelzugänglichkeit definiert [164] [45] [52]. Jüngere Arbeiten hingegen verwenden Triangulationen um über Nachbarschaftskriterien das Innere der Kontaktfläche zu definieren.

Diesen Ansatz verwendet das Programm *Intervor* [38], das anhand einer Voronoi-Triangulation der Kontaktatome atomweise die *Voronoy Shelling Order (VSO)* als ein Maß dafür bestimmt, wie tief im Zentrum das entsprechende Atom liegt. Ein Programm mit

ähnlichem Ziel, *Protein Interface Analyzer (PIA)* basierend auf *Reduced Surface* (siehe Abschnitt 3.6.3) wurde kürzlich im Rahmen einer Diplomarbeit in der Arbeitsgruppe von R. Merkl entwickelt [134].

In den folgenden Abschnitten werden die drei genannten Arten der Berechnung von Kern und Rand genauer vorgestellt und resultierende Klassifikationen anhand der Struktur einer PPK miteinander verglichen.

4.2.1.1 Kern-Rand Klassifikationen basierend auf Lösungsmittelzugänglichkeit

In früheren Arbeiten wurden "Kern" und "Rand" meist über die Lösungsmittelzugänglichkeit definiert. Verbreitet ist die Definition des Kerns als die Menge derjenigen Aminosäuren, die mindestens ein vollkommen vergrabenes Atom besitzen [37] [165]. In [52] wird ein Kernresiduum als ein Rest definiert, dessen relative $SASA$ 7% in der Struktur des gebundenen Komplexes unterschreitet. Diese Methode würde jedoch keine Reste entdecken, die bereits im Monomer zum großen Teil vergraben sind und deshalb keinen großen Verlust an Lösungsmittelzugänglichkeit bei der Komplexbildung erleiden. Dies sind immerhin etwa 13% aller Kontaktreste [155]. Verstärkt wird dieses Problem durch die Tatsache, dass PPKs oft sehr zerklüftete Formen annehmen können. Außerdem entspricht diese Definition von Kern und Rand nicht immer der intuitiven geometrischen Vorstellung, wie man an Abbildung 4.3(a) auf Seite 66 erkennt. Dieses Beispiel belegt, dass ein solches, einfaches Kriterium nicht genügt, um PPKs zuverlässig in Kern und Rand aufzuteilen. Dies liegt daran, dass PPKs nur selten gleichmäßig konvex geformt sind. Nur unter dieser Voraussetzung würde die Lösungsmittelzugänglichkeit der Kontaktfläche gleichmäßig von außen nach innen abnehmen. Daher wurden Verfahren entwickelt, die auf geometrischen Algorithmen beruhen.

4.2.1.2 Intervor

Intervor wurde mit dem Ziel entwickelt, die Kontaktfläche von Proteinen anhand einer Voronoi-Triangulation der Atomkoordinaten zu bestimmen und die Kontaktatome über ihre Schalennummern nach ihrer geometrischen Lage in der PPK einzuteilen. Dabei gibt die *Voronoy Shelling Order (VSO)* als ein Maß auf Atomebene an, wie viele Atome sich an der Oberfläche des Monomers zwischen dem betrachteten Atom und dem nächsten Atom der restlichen Oberfläche in der Triangulation befinden. Je größer *VSO* umso weiter im Zentrum der Kontaktfläche ist das Atom platziert. Um damit die Lage eines Kontaktrestes in der Kontaktfläche zu bestimmen, muss über die Schalennummern sei-

ner Kontaktatome gemittelt werden. Die Authoren stellten fest, dass mit zunehmender *VSO* die Seitenketten weniger Kontakt zum Wasser haben. Außerdem korreliert die VSO mit typischen Merkmalen von PPKs wie Hydrophobizität und Konserviertheit [38]. Ein Beispiel für eine Klassifikation mit *Intervor* zeigt Abbildung 4.3(b) auf Seite 66.

4.2.1.3 Protein Interface Analyzer (PIA)

Ein ähnliches Programm, *PIA* genannt, wurde von *Staudigel* und *Trenner* [134] im Rahmen ihrer Diplomarbeit an der FU Hagen entwickelt. Es berechnet anhand der Triangulation der *Reduzierten Oberfläche* Nachbarschaften von Aminosäuren an der PPK und bestimmt damit für jede Kontaktaminosäure, wie weit im Zentrum der PPK sie lokalisiert ist.

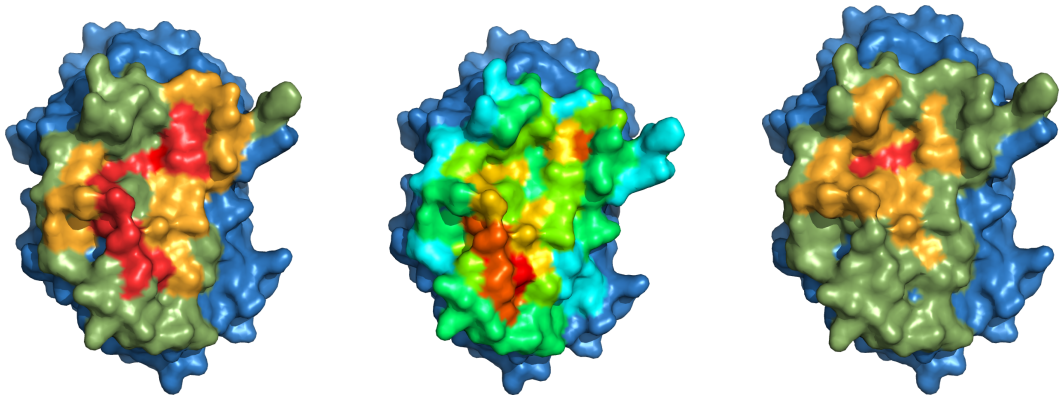
Die Berechnung der *Reduced Surface* wurde ursprünglich von *Sanner* und *Olson* für das Programm *MSMS* entwickelt [133], das über einen robusten Algorithmus die Berechnung von SASA und Connollyoberfläche ermöglicht (siehe Abschnitt 3.6). Die während der Berechnung der Oberfläche anfallende *Reduced Surface* (siehe Abschnitt 3.6.3) ist eine Triangulation der Oberfläche und kann als solche zur Definition der Nachbarschaft auf Atomebene benutzt werden.

Der äußere Rand der Kontaktfläche ergibt sich dann als die Menge derjenigen Kontaktreste, die direkte Nachbarn des Außenbereiches sind. Diese Reste werden der *PIA-Schale (PIAS) 1* zugerechnet. Anschließend lässt sich eine zweite Schale von Kontaktresten als direkte Nachbarn zum äußeren Rand bestimmen, die *PIA-Schale 2*. Analog wird eine PPK sukzessive abgeschält, bis der innerste Kern erreicht ist und alle Kontaktreste nach ihrer Lage in der PPK eingeteilt wurden. Die *PIAS* macht eine Aussage darüber, wie weit im Inneren der Kontaktfläche sich die entsprechende Kontaktaminosäure befindet. Je höher ihr Wert, umso mehr andere Kontaktaminosäuren befinden sich zwischen der betrachteten Position und dem Außenbereich und umso zentraler ist die Position in der PPK gelegen. Abbildung 4.3(c) auf Seite 66 zeigt ein Beispiel für eine Einteilung einer PPK nach *PIA*.

4.2.2 Vergleich der Methoden

In Abbildung 4.3 ist für eine PPK gezeigt, wie die drei beschriebenen Verfahren die Aminosäuren in Kern und Rand aufteilen. Man erkennt deutlich, dass eine Klassifika-

tion anhand der *SASA* nicht der intuitiven geometrischen Vorstellung entspricht. Wie erwähnt, klassifiziert *Intervor* einzelne Atome. Da alle folgenden Analysen jeweils eine Aminosäure als Einheit betrachten, müssen die von *Intervor* gelieferten atomspezifischen *VSO*-Werte gemittelt werden. Aus diesem Grund entspricht auch die resultierende Einteilung der Kontaktfläche nicht ganz der intuitiven Vorstellung von Kern und Rand einer PPK. Dies kommt zustande, da eine große am Rand der Kontaktfläche liegende Aminosäure meist auch Atome besitzt, die weiter ins Innere der Kontaktfläche hineinragen. Daher wird solchen Aminosäuren nach der Mittelung über ihre Kontaktatome eine höhere *VSO* zugeteilt. Aus diesem Grund liegen häufig Aminosäuren mit $VSO > 1$ am Rand der Kontaktfläche.



(a) (b) (c)

Abbildung 4.3: Definitionen von Kern und Rand einer Kontaktfläche
Diese Abbildung vergleicht 3 verschiedene Methoden zur Berechnung von Kern und Rand einer Kontaktfläche anhand des PDB-Eintrages 1O58. Die Farbtemperatur nimmt vom kalten Blau der Nichtkontaktfläche hin zu einem warmen Rot des innersten Kerns zu. (a) zeigt eine Klassifikation basierend auf $\Delta SASA$, (b) zeigt das Ergebnis von *Intervor* nach Mittelung der Werte einer Seitenkette über die *VSO* ihrer Kontaktatome (c) entstand durch Anwendung von *PIA*

Die Software *PIA* hingegen bestimmt die Schalennummer anhand eines intramolekularen Kontaktkriteriums zwischen Aminosäuren. Damit erhält man für sämtliche Positionen der Interaktionsfläche, die in der Triangulation benachbart zu Aminosäuren außerhalb der Kontaktfläche liegen $PIAS = 1$. Wie man an Abbildung 4.3(c) erkennt, entspricht diese Unterteilung der Kontaktfläche in Zentral- und Randbereich der intuitiven Erwartung. Das Zentrum liegt dabei jedoch nicht immer am Schwerpunkt der Kontaktfläche. Ein Grund dafür ist, dass wegen der unterschiedlichen Größe der Aminosäuren auch die Breite einer Schale nicht konstant ist. Der zweite Grund besteht darin, dass Kontaktflächen nur selten gleichmäßige konvexe Flächen sind. Häufig treten aufgrund der zerklüfteten Oberfläche Risse und Löcher in der Kontaktfläche auf. Risse in der Kontaktfläche schaffen zusätzliche Nachbarschaften zu Seitenketten außer-

halb der Kontaktfläche, die dazu führen, dass mehr Kontaktpositionen mit $PIAS = 1$ klassifiziert werden. Löcher hingegen nehmen Raum ein, der nicht genutzt wird um in der Triangulation den Abstand zum Außenbereich zu vergrößern. Dieser Effekt trägt ähnlich wie die unterschiedliche Größe der Aminosäuren zu einer unregelmäßigen Breite der Schalen bei. Details zur Behandlung von Löchern im PIA-Algorithmus sind in Abschnitt 3.7.1 dargestellt.

Der Hauptunterschied der beiden Methoden *Intervor* und *PIA* besteht darin, dass *Intervor* die PPK in Schalen von atomarer Dicke unterteilt, während PIA Schalen von der Dicke einer Aminosäure berechnet. Da jedoch in biologischer Hinsicht eine Aminosäure der kleinste mutierbare Baustein ist, arbeitet *PresCont* auf Aminosäureebene. Aus diesem Grund muss auch die Schalenordnung auf Aminosäureebene bestimmt werden.

Die Unterschiede zwischen *Intervor* und *PIA* führen dazu, dass *Intervor* gleiche PPKs in deutlich mehr Schalen einteilt als *PIA*. Um vergleichbare Definitionen von Kern und Rand einer Kontaktfläche zu erhalten wird bei der Kerndefinition durch PIA die Kontaktfläche auf Seitenketten mit einer $PIAS >= 2$ eingeschränkt, während bei *Intervor* ein Schwellwert von $VSO >= 3$ gewählt wird. So entstehen in grober Näherung etwa gleich große Zentralbereiche.

4.3 Eigenschaften zur Charakterisierung von Kontaktflächen

In der Literatur wurde eine Fülle von Parametern zur Klassifikation von PPKs beschrieben und angewendet [47] [46] [74]. Bei der Entwicklung von *PresCont* stand die Verwendung von Parametern im Vordergrund, die am MSA abgeleitet werden können. Die Performanz vieler Methoden der Bioinformatik, wie z.B. der Vorhersage der Proteinsekundärstruktur [166] [167] oder der Homologiemodellierung, konnte durch die Berücksichtigung dieser evolutionären Information drastisch gesteigert werden. Basierend auf diesen Befunden werden in *PresCont* Kenndaten aus den Protein 3D-Strukturen der zu klassifizierenden Proteine mit solchen aus MSAs kombiniert. Im Rahmen der Softwareentwicklung kristallisierte sich eine Kombination von 5 Eigenschaften als optimal für die Klassifikation heraus.

Aus der Protein 3D-Struktur wird die Exponiertheit einzelner Werte sowie die Verteilung hydrophober Patches abgeleitet. Aus den MSAs stammen Kenndaten zur Konserviertheit der Seitenketten, zur positionsspezifischen Häufigkeitsverteilung von Aminosäuren und zu korrelierten Mutationen.

Diese fünf Eigenschaften bilden die Eingabe für eine *Support Vektor Maschine* (SVM). Diese klassifiziert jede an der Proteinoberfläche liegende Aminosäure als zum Interface gehörend oder als zur Oberfläche gehörend. Zusätzlich wird jede Vorhersage mit einer Wahrscheinlichkeit versehen, mit der die Zuverlässigkeit der Vorhersage gekennzeichnet wird. In diesem Abschnitt werden sukzessive die Herleitung und das Berechnungsverfahren dieser fünf Merkmale vorgestellt.

4.3.1 Exponiertheit

Die Exponiertheit einer Aminosäure korreliert mit ihrer Lösungsmittelzugänglichkeit, die sich über die *SASA* anhand der Struktur des Monomers bestimmen lässt. In früheren Arbeiten wurde eine Korrelation zwischen Bindungsenergie eines Kontaktrestes und dem Verlust an Lösungsmittelzugänglichkeit durch die Komplexbildung gefunden [53]. Die Befunde aus [168] belegen, dass sich mit einer Kombination von Maßen der Lösungsmittelzugänglichkeit und der Konserviertheit energetische Hot Spots, die Seitenketten, die für einen Großteil der Bindungsenergie eines Protein-Protein Komplexes verantwortlich sind [169] [23], vorhersagen lassen. Wegen des großen Beitrages zur Klassifikationsleistung wird Lösungsmittelzugänglichkeit vielfach mit anderen Merkmalen kombiniert und zur Vorhersage von Protein-Protein Kontaktflächen verwendet (siehe z.B. [156] [170] [171] [46]).

In den Arbeiten von *Jones und Thornton* stellte sich Lösungsmittelzugänglichkeit als sehr aussagekräftige Information zur Vorhersage der Kontaktflächen von Homodimeren heraus [28] [15] [48]. *Zhou* und *Shan* verwendeten *SASA* erfolgreich bei der Kontaktflächenvorhersage transienter Heterodimere und erreichten unabhängig davon, ob sie die gebundene oder die ungebundene Struktur verwendeten eine sehr ähnliche Performanz [156]. Aufgrund dieser Befunde wird in *PresCont* die Exponiertheit der Aminosäureseitenkette als ein Merkmal bei der Klassifikation verwendet. Als Maß für die Exponiertheit wird die *rSASA* verwendet. Ihre Berechnung ist in Abschnitt 3.6 erläutert.

4.3.2 Häufigkeiten von einzelnen Seitenketten und Kontaktpaaren

Chancenquotienten dienen dazu, die Wahrscheinlichkeiten für zwei Hypothesen (*Likelihood*) in Abhängigkeit von einem Merkmal miteinander zu vergleichen. In der Regel wird das *log-likelihood* Verhältnis betrachtet, das sich als logarithmierter Quotient ergibt:

$$\log\left(\frac{f(a|H_a)}{f(a|H_0)}\right). \tag{4.1}$$

In der Bioinformatik wird als Nullhypothese H_0 meist ein zufälliges Auftreten einer bestimmten Aminosäure benutzt. Beispielsweise wurden bei der Berechnung der *BLOSUM*-Matrizen die beobachtete Austauschhäufigkeit von Aminosäuren in einem Datensatz von MSAs mit der Häufigkeit verglichen, die unter der Annahme der Unabhängigkeit der Austausche voneinander zu erwarten wäre [126].

In diesem Abschnitt werden Chancenquotienten (Log Odds Ratios, LORs) benutzt um auffällige Häufigkeitsverteilungen von Seitenketten und intramolekularen Kontaktpaaren von Seitenketten an PPKs vergleichend zu beurteilen. Nicht-kovalente Interaktionen zwischen Paaren von Aminosäuren, resultierend aus ihren spezifischen physikalisch-chemischen Eigenschaften, bilden die Basis für die Stabilität und Spezifität von Protein-Protein Interaktionen. Die Aminosäurezusammensetzung von PPKs [14] [13] [61] [49] [28] [37] und Häufigkeiten von Aminosäurepaaren [172] [18] [21] wurden bereits häufiger untersucht. Dabei wurde gezeigt, dass Kontaktflächen einen zentral gelegenen Kernbereich besitzen, deren Aminosäurezusammensetzung derjenigen des Proteininneren ähnelt und sich deutlich von der Zusammensetzung der restlichen Oberfläche unterscheidet. Somit ist anzunehmen, dass Scores, die diese Unterschiede quantifizieren, dazu beitragen können, Verfahren zur Vorhersage von Protein-Protein Kontaktflächen zu verbessern. In diesem Abschnitt werden derartige Scores als LORs berechnet um sie als zusätzliche Information bei der Vorhersage von Protein-Protein Kontaktflächen verwenden zu können.

Grundlage für die Berechnung ist der Datensatz $Komp_{RN}$ [99] aus Abschnitt 3.1.1. Dieser Datensatz bietet eine ausreichende Datengrundlage ohne Redundanzen um sicherzustellen, dass auch alle Werte mit hinreichender Genauigkeit abgeschätzt werden können.

4.3.2.1 Das Vorkommen von Aminosäuren in PPKs

Für die Berechnung der Chancenquotienten müssen für alle Aminosäuren aa_i mit $i = 1, \ldots, 20$ die relativen Häufigkeiten an der Kontaktfläche $f^{cont}(aa_i)$ und an der gesamten Oberfläche des Monomers $f^{surf}(aa_i)$ bestimmt werden. Anschließend lassen sich die LORs $L_0(aa_i)$ berechnen über

$$L_0(aa_i) = \log\left(\frac{f^{cont}(aa_i)}{f^{surf}(aa_i)}\right). \tag{4.2}$$

Positive Werte von $L_0(aa_i)$ weisen dabei auf eine Bevorzugung der Seitenkette a_i an der Kontaktfläche relativ zur gesamten Oberfläche hin.

Weiter kann man analog LORs berechnen, die die Häufigkeitsverteilungen im Zentralbereich einer Kontaktfläche nach *PIA* bzw. *Intervor* mit der Häufigkeitsverteilung an der gesamten Oberfläche vergleichen, indem man die relativen Häufigkeiten aller Aminosäurearten im Zentralbereich der Kontaktfläche nach *PIA* $f^{PIA_v}(aa_i)$ bzw. *Intervor* $f^{ITV_v}(aa_i)$ bestimmt. Die Schwellwerte $v \geq 2$ für die *PIA-Schale (PIAS)* bzw. $v \geq 2, \ldots, 8$ für die *Voronoy Shelling Order (VSO)* bei *Intervor* bestimmen, welche Reste bei der Berechnung der relativen Häufigkeiten zum Zentralbereich der PPK gerechnet werden. Ersetzt man in (4.2) $f^{cont}(aa_i)$ durch $f^{PIA_v}(aa_i)$ bzw. $f^{ITV_v}(aa_i)$ so erhält man analoge LORs für *PIA* und *Intervor*:

$$L_{PIA_v}(aa_i) = \log\left(\frac{f^{PIA_v}(aa_i)}{f^{surf}(aa_i)}\right) \tag{4.3}$$

$$L_{ITV_v}(aa_i) = \log\left(\frac{f^{ITV_v}(aa_i)}{f^{surf}(aa_i)}\right). \tag{4.4}$$

Vergleicht man die Scores für die Kontaktfläche mit denjenigen für den Kern der Kontaktfläche, so kann man eine Aussage darüber treffen, inwieweit sich die Aminosäurehäufigkeiten im Kern- und Randbereich von Protein-Protein Kontaktflächen unterscheiden.

Tabelle 4.1 auf Seite 71 belegt, dass sich die Häufigkeitsverteilungen der Aminosäuren für Kontaktflächen und die restliche Oberfläche unterscheiden. Alle aliphatischen Seitenketten (*Leu, Ile, Met, Val*) sind, wie ihre positiven Werte zeigen, an der Kontaktflä-

	Cont	$\mathbf{L_{PIA_2}}$	$\mathbf{L_{ITV_2}}$	$\mathbf{L_{ITV_3}}$	$\mathbf{L_{ITV_4}}$	$\mathbf{L_{ITV_5}}$	$\mathbf{L_{ITV_6}}$	$\mathbf{L_{ITV_7}}$	$\mathbf{L_{ITV_8}}$
Ala	-0,12	0,25	-0,06	0,13	0,25	0,37	0,32	0,27	0,34
Cys	0,19	0,43	0,29	0,58	0,74	0,78	0,91	0,84	0,96
Asp	-0,31	-0,47	-0,38	-0,68	-0,83	-0,86	-0,90	-0,53	-0,28
Glu	-0,27	-0,57	-0,43	-0,83	-1,10	-1.13	-1,08	-1,15	-1,56
Phe	0,48	0,46	0,57	0,76	0,83	0,83	0,75	0,54	0,39
GLy	-0,27	0,14	-0,23	-0,16	-0,03	0,05	0,02	-0,13	0,12
His	0,12	-0,06	0,14	0,09	-0,03	-0,07	-0,11	0,05	-0,11
Ile	0,29	0,43	0,37	0,58	0,66	0,68	0,66	0,74	0,62
Lys	-0,31	-0,79	-0,55	-1,09	1,45	-1,67	-1,80	-1,77	-1,29
Leu	0,30	0,39	0,36	0,53	0,59	0,56	0,51	0,38	0,17
Met	0,45	0,55	0,46	0,60	0,70	0,84	0,93	1,19	1,01
Asn	-0,17	-0,24	-0,19	-0,30	-0,41	-0,43	-0,49	-0,45	-0,74
Pro	-0,06	-0,06	-0,11	-0,26	-0,37	-0,36	-0,44	-0,58	-0,40
Gln	-0,03	-0,25	-0,09	-0,31	-0,42	-0,49	-0,31	-0,19	-0,16
Arg	0,12	-0,40	-0,01	-0,34	-0,71	-0,95	-0,99	-0,97	-0,57
Ser	-0,16	0,03	-0,13	-0,13	-0,06	-0,08	0,07	0,12	0,18
THr	-0,10	-0,03	-0,07	-0,08	-0,12	-0,11	-0,01	-0,09	-0,35
Val	0,15	0,31	0,22	0,35	0,44	0,47	0,53	0,56	0,50
Trp	0,41	0,08	0,52	0,68	0,68	0,63	0,64	0,71	0,67
Tyr	0,33	0,14	0,43	0,50	0,42	0,27	0,20	0,26	0,52

Tabelle 4.1: Chancenquotienten zur Häufigkeitsverteilung von Aminosäuren an der PPK
Diese Scores beschreiben Unterschiede in den Häufigkeitsverteilungen von Aminosäuren an verschiedenen Bereichen der Oberfläche eines Proteins. Positive Werte treten bei Bevorzugung der Aminosäure an der Kontaktfläche (L_0) bzw. dem Kern (L_{PIA_2} bzw. L_{ITV_v}) auf, negative Werte bei einer Unterrepräsentation. In der zweiten Spalte (L_0) werden die Scores gelistet, die sich ergeben, wenn die Häufigkeiten $f^{cont}(aa_i)$ aus den gesamten PPKs abgeleitet werden. Für die mit L_{PIA_2} überschriebene Spalte ergeben sich die Werte über die Häufigkeiten $f^{PIA_2}(aa_i)$ im *PIA*-Kern mit $PIAS \geq 2$. Die restlichen Spalten enthalten die Werte L_{ITV_v}, die sich aus $f^{ITV_v}(aa_i)$ ableiten, den Häufigkeiten am *Intervor*-Kern mit $VSO \geq v$ ($v = 2, \ldots, 8$).

che ebenso überrepräsentiert wie alle aromatischen (*Phe, Trp, Tyr*). Daneben sind auch *His* und *Arg* an der Kontaktfläche leicht bevorzugt. Dies spricht für die große Bedeutung hydrophober Interaktion für die Bindungsenergie von Protein-Protein Komplexen. Der positive Wert von *Cys* lässt sich durch intermolekulare Disulfidbrücken erklären. Hydrophile Seitenketten dagegen zeigen durchweg negative Werte. Die kleinsten Werte finden sich bei den geladenen Seitenketten *Asp, Glu, Lys*. Mit Ausnahme von *Arg* sind Ladungen an der Kontaktfläche ebenso unterrepräsentiert wie Seitenketten mit Dipolmoment (*Ser, Thr, Asn*).

Die Scores, die den Kernbereich (vgl. Werte in Spalte L_0 in Tabelle 4.1) mit der restlichen Oberfläche vergleichen, unterscheiden sich deutlich von denjenigen der gesamten

Kontaktfläche. Betrachtet man die aliphatischen Seitenketten (Ile, Leu, Met, Val) so zeigen ihre Scores unabhängig davon, ob der Kernbereich nach *PIA* oder nach *Intervor* definiert wurde, dieselbe Tendenz. Die Bevorzugung aliphatischer Aminosäuren ist im Kern stärker als an der gesamten Kontaktfläche. Da sich die gesamte Kontaktfläche additiv aus Kern- und Randbereich zusammensetzt, sind aliphatische Aminosäuren im Kernbereich gegenüber dem Randbereich bevorzugt. Dieser Effekt tritt sowohl bei der Kerndefinition nach *PIA* als auch nach *Intervor* auf. Beim Vergleich der aus *PIA* und *Intervor* resultierenden Scores muss die Spalte L^{ITV_3} herangezogen werden, da der Kern bei Definition nach *Intervor* mit einer $VSO \geq 3$ im Durchschnitt ähnlich viele Aminosäuren zum Zentralbereich der Kontaktfläche rechnet, wie *PIA*.

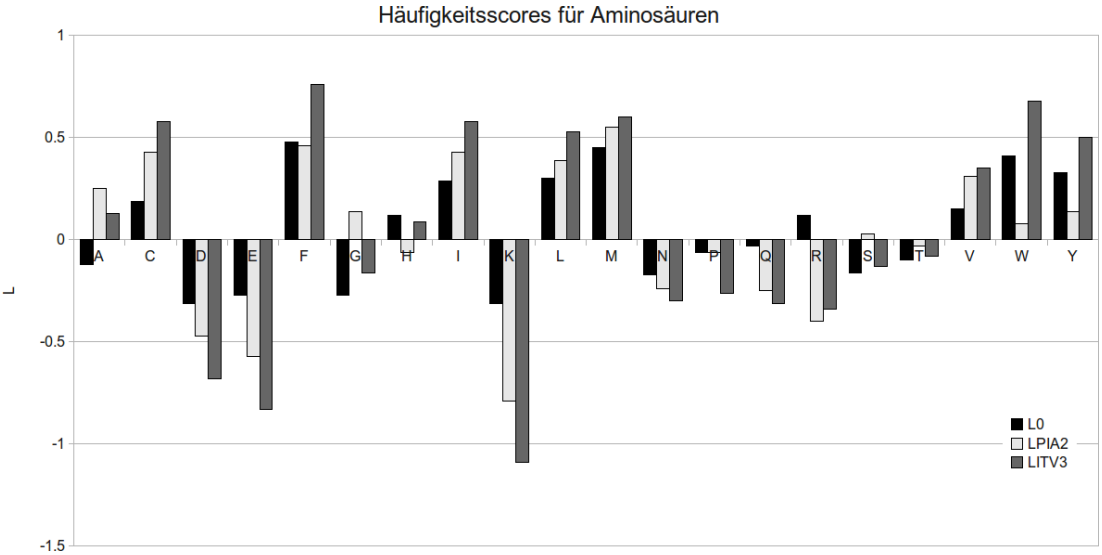


Abbildung 4.4: Häufigkeitsscores für Aminosäuren
Das Säulendiagramm stellt die Werte aus Tabelle 4.1 Seite 71 graphisch dar. Positive Werte Sprechen für eine Überrepräsentation der Aminosäure an der Kontaktfläche (L_0) bzw. im *PIA*-Kern (L_{PIA_2}) mit *PIA-Schale* ≥ 2 bzw. im *Intervor*-Kern (L_{ITV}) mit $VSO \geq 3$.

In Abbildung 4.4 sind die mit *Cont*, L_{PIA_2} und L_{ITV_3} überschriebenen Spalten aus Tabelle 4.1 Seite 71 graphisch dargestellt. Wie man erkennt, zeigen aromatische Seitenketten ein weniger eindeutiges Verhalten als aliphatische. Während im *PIA*-Kern die aromatischen Aminosäuren *Phe, Trp, Tyr* weniger stark bevorzugt sind als an der gesamten Kontaktfläche, sind sie bei Verwendung der Kern-Rand-Einteilung nach *Intervor* im Kern häufiger vertreten als am Rand.

An dieser Stelle erkennt man die in Abschnitt 4.2.2 festgestellten Unterschiede zwischen den Definitionen von Kern- und Randbereich einer PPK durch *PIA* und *Intervor*. Große Seitenketten, wie alle aromatischen Aminosäuren, liegen nach der Definition von *Intervor* weiter im Zentralbereich der PPK als nach *PIA*. Wie man sieht, hängen die

Aminosäurehäufigkeiten stark von der Art der Definition von Kern- und Randbereich einer Kontaktfläche ab.

4.3.2.2 Bewertung intermolekularer Kontaktpaare

Mit den oben eingeführten Scores können Unterschiede in der Zusammensetzung einzelner Oberflächenpatches unabhängig von den möglichen Wechselwirkungen mit einem Interaktionspartner bewertet werden. Es liegt jedoch nahe anzunehmen, dass die Häufigkeit von Paaren von Aminosäuren (aa_i^k, aa_j^l), die an der PPK benachbart liegen, aber zu den zwei verschiedenen Proteinen k bzw. l gehören, sich von der unterscheidet, die aufgrund zufälliger Kombination zu erwarten sind. Deswegen wurde der Score S_{pair_inter} berechnet. S_{pair_inter} ergibt sich aus

$$S_{pair\_inter} = \log\left(\frac{f^{cont}(aa_i, aa_j)}{f^{surf}(aa_i) \cdot f^{surf}(aa_j)}\right) \tag{4.5}$$

als LORs aus der beobachteten Häufigkeit eines intermolekularen Kontaktpaares der Aminosäuren aa_i und aa_j und dem Produkt der Häufigkeiten einzelner Aminosäuren an der Oberfläche, das die Annahme der Unabhängigkeit in der Nullhypothese ausdrückt.

Nach [13] und [14] unterscheiden sich Häufigkeitsverteilungen der Kontaktflächen von Homodimeren und Heterodimeren ebenso wie von obligaten und transienten Komplexen. So hat die PPK von obligaten Homodimeren einen hydrophoberen Charakter als diejenige von Heterodimeren unter denen sich viele schwache transiente Komplexe befinden. Diese Unterschiede werden auch in [21] deutlich, wenn die Chancenquotienten verglichen werden, die getrennt für einen Datensatz von Homodimeren und einen Datensatz von Heterodimeren berechnet wurden. Die Scores unterscheiden sich allerdings nur in ihrer Größe, nicht jedoch im Vorzeichen. Hydrophobe Wechselwirkungen von aromatischen und aliphatischen Seitenketten untereinander sind in beiden Fällen stärker bevorzugt als attraktive polare Wechselwirkungen, während gleichnamige Ladungen, die elektrostatische Abstoßungen bedingen, erwartungsgemäß unterrepräsentiert sind. Insgesamt lässt sich feststellen, dass Homodimere zwar extremere Signale zeigen, die Tendenz beider Tabellen von Chancenquotienten jedoch die gleiche ist. Daher werden im folgenden S_{pair_inter} Scores verwendet, die aus dem kompletten Datensatz $Komp_{RN}$ errechnet wurden.

	A	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
A	-0.23																			
C	-0.18	2.98																		
D	-1.08	-1.04	-0.78																	
E	-0.98	-0.67	-1.30	-0.55																
F	0.53	0.65	-0.05	0.18	2.32															
G	-0.90	-0.19	-0.99	-0.97	0.24	-0.47														
H	-0.24	0.00	0.26	0.34	0.93	-0.36	1.45													
I	-0.24	0.08	-0.75	-0.42	1.29	-0.56	0.35	1.15												
K	-1.08	-0.46	0.21	0.18	0.26	-0.82	-0.38	-0.53	-0.63											
L	-0.30	0.03	-0.86	-0.36	1.28	-0.58	0.20	0.55	-0.52	0.93										
M	0.12	0.71	-0.23	0.05	1.51	-0.12	0.83	0.94	-0.11	0.80	2.18									
N	-0.64	-0.30	-0.50	-0.57	0.46	-0.42	0.16	-0.25	-0.48	-0.29	0.08	0.45								
P	-0.60	-0.09	-0.69	-0.61	0.92	-0.59	0.26	-0.17	-0.79	-0.13	0.47	-0.38	0.11							
Q	-0.53	-0.05	-0.50	-0.50	0.72	-0.48	0.05	0.07	-0.40	0.16	0.54	0.07	-0.03	0.80						
R	-0.27	0.16	0.83	0.79	0.86	-0.15	0.31	0.16	-0.48	0.26	0.68	0.05	0.12	0.31	0.89					
S	-0.74	-0.27	-0.40	-0.34	0.44	-0.93	0.12	-0.43	-0.69	-0.58	0.08	-0.40	-0.59	-0.25	-0.13	-0.16				
T	-0.81	-0.25	-0.43	-0.44	0.33	-0.70	-0.05	-0.17	-0.65	-0.35	0.04	-0.30	-0.47	-0.07	-0.18	-0.50	0.11			
V	-0.34	0.10	-0.85	-0.54	0.96	-0.80	0.00,	0.30	-0.64	0.17	0.65	-0.49	-0.32	-0.14	-0.01	-0.60	-0.28	0.52		
W	0.46	0.99	0.39	0.41	1.91	0.42	1.14	1.13	0.46	1.08	1.54	0.47	1.09	0.71	1.17	0.40	0.43	0.70	2.68	
Y	0.21	0.48	0.33	0.45	1.59	0.19	1.04	0.93	0.33	0.88	1.21	0.43	0.85	0.62	0.90	0.17	0.19	0.64	1.37	1.56

Tabelle 4.2: Scores S_{pair_inter} für intermolekulare Aminosäurekontakte
Die Werte wurden aus dem Datensatz $Komp_{RN}$ abgeleitet. Sie bewerten, wie häufig zwei Aminosäuren über die Kontaktfläche hinweg intermolekular in Kontakt sind. Positive Werte sprechen für einen häufigen und günstigen Kontakt, während negative Werte auf einen seltenen und deshalb ungünstigen Kontakt hindeuten.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-0.04																			
C	-0.05	2.97																		
D	-0.97	-0.95	-0.86																	
E	-0.86	-0.59	-1.41	-0.83																
F	0.62	0.85	-0.02	0.13	2.29															
G	-0.73	-0.01	-0.79	-0.92	0.37	-0.30														
H	-0.12	0.17	0.16	0.17	0.88	-0.28	1.30													
I	-0.09	0.17	-0.79	-0.39	1.38	-0.38	0.45	1.21												
K	-0.99	-0.36	0.00	-0.19	0.23	-0.77	-0.48	-0.60	-0.92											
L	-0.14	0.08	-0.83	-0.40	1.37	-0.43	0.17	0.65	-0.58	1.01										
M	0.32	0.85	-0.09	0.01	1.62	0.05	0.94	1.07	-0.10	0.92	2.18									
N	-0.51	-0.11	-0.55	-0.73	0.49	-0.30	0.14	-0.24	-0.67	-0.22	0.18	0.35								
P	-0.48	0.13	-0.74	-0.67	0.94	-0.37	0.09	-0.15	-0.88	-0.06	0.55	-0.33	0.01							
Q	-0.42	0.02	-0.60	-0.70	0.75	-0.38	-0.12	0.01	-0.59	0.22	0.54	-0.07	-0.02	0.52						
R	-0.28	0.25	0.66	0.50	0.80	-0.06	0.21	0.14	-0.77	0.19	0.63	-0.02	0.07	0.17	0.49					
S	-0.54	-0.144	-0.34	-0.36	0.51	-0.78	0.18	-0.30	-0.76	-0.45	0.16	-0.36	-0.60	-0.23	-0.13	-0.12				
T	-0.64	-0.05	-0.50	-0.44	0.43	-0.57	-0.01	-0.07	-0.71	-0.29	0.14	-0.31	-0.40	-0.08	-0.24	-0.36	0.10			
V	-0.21	0.14	-0.84	-0.47	1.06	-0.70	0.03	0.43	-0.66	0.28	0.82	-0.45	-0.28	-0.22	-0.01	-0.60	-0.17	0.57		
W	0.62	1.05	0.27	0.29	1.93	0.50	1.04	1.14	0.38	1.08	1.52	0.46	1.07	0.65	1.06	0.45	0.56	0.80	2.39	
Y	0.34	0.60	0.25	0.36	1.59	0.27	1.02	1.03	0.21	0.91	1.27	0.38	0.86	0.59	0.75	0.25	0.22	0.68	1.31	1.50

Tabelle 4.3: Scores $S^{PIA}_{pair\_inter}$ für intermolekulare Aminosäurekontakte im Kernbereich von PPKs
Die Werte wurden aus dem Datensatz $Komp_{RN}$ abgeleitet. Sie bewerten, wie häufig zwei benachbarte Aminosäuren in PPKs vorkommen. Positive Werte sprechen für einen häufigen und günstigen Kontakt, während negative Werte auf einen seltenen und deshalb ungünstigen Kontakt hindeuten.

In Tabelle 4.2 finden sich die 210 paarweisen Werte $S_{pair\_inter}(aa_i^k, a_j^l)$. Ein großer positiver Wert resultiert aus einer hohen relativen Häufigkeit des entsprechenden Kontaktpaares, während ein negativer Wert besagt, dass das zugehörige Kontaktpaar seltener als erwartet auftritt und deshalb einen ungünstigen Kontakt darstellt. Der größte positive Wert findet sich bei $Cys-Cys$-Kontakten, was sich durch Disulfidbrücken erklären lässt. Ebenfalls bevorzugt sind Kontakte aromatischer und aliphatischer Aminosäuren untereinander. Darin spiegelt sich die große Bedeutung hydrophober Wechselwirkungen für die Stabilität von PPIs wider. Salzbrücken sind nur sehr gering bevorzugt. Die knapp positiven Werte für $Asp - Lys$ und $Glu - Lys$ lassen sich dadurch erklären, dass diese drei geladenen Seitenketten allgemein an der Kontaktfläche seltener auftreten als an der restlichen Oberfläche (siehe Tabelle 4.1). Häufiger sind dagegen Kontakte von $Asp - Arg$ und $Glu - Arg$, die ebenfalls in der Lage sind, Salzbrücken zu bilden. Elektrostatische Abstoßungen wie $Asp - Asp$, $Glu - Glu$ oder $Asp - Glu$, die zu einer starken Destabilisierung der Kontaktfläche führen würden, wurden erwartungsgemäß selten gefunden und zeigen deshalb deutlich negative Werte.

Die Werte $S^{PIA}_{pair\_inter}(aa_i^k, a_j^l)$ in Tabelle 4.3 wurden auf ähnliche Weise wie S_{pair_inter} berechnet. Einziger Unterschied ist, dass bei der Berechnung nach (4.5) die relative Häufigkeit $f^{cont}(aa_i^k, aa_j^l)$ für einen Kontakt zwischen den Seitenketten aa_i^k und aa_j^l ersetzt wird durch die Häufigkeit $f^{PIA}_{pair\_inter}(aa_i^k, aa_j^l)$ mit der ein Kontakt zwischen zwei Seitenkettenarten aa_i^k und aa_j^l auftritt, von denen sich mindestens eine der beiden Seitenketten im Kernbereich der Kontaktfläche nach *PIA* mit *PIA*-Schale ≥ 2 befindet.

Die Unterschiede in den Werten der Tabellen 4.2 und 4.3 sind nicht auffällig groß. Sie entsprechen den Unterschieden zwischen L_0 und L_{PIA_2} aus Tabelle 4.1, die für die gesamte Kontaktfläche und den *PIA*-Zentralbereich berechnet wurden (vgl. Tabelle 4.1). So sind die Kontakte großer aromatischer Seitenketten im Zentrum einer Kontaktfläche weniger bevorzugt als am Rand. Ansonsten verstärkt die Einschränkung auf den Zentralbereich der Kontaktfläche die Signale. So sind alle Kontakte aliphatischer Seitenketten untereinander durchwegs im Zentralbereich stärker bevorzugt als an der gesamten Kontaktfläche. Ungünstige elektrostatische Kollisionen zwischen gleichnamig geladenen Seitenketten dagegen (z.B. $Asp - Asp$, $Asp - Glu$, $Glu - Glu$, $Lys - Lys$ und $Lys - Arg$) sind im Kernbereich noch stärker benachteiligt als an der gesamten Kontaktfläche.

	A	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
A	-0,05																			
C	0,10	0,18																		
D	-0,25	0,02	-0,24																	
E	-0,29	0,06	-0,25	-0,20																
F	0,28	0,41	0,08	0,15	0,32															
G	-0,18	0,18	-0,41	-0,27	0,27	-0,14														
H	0,02	0,11	-0,17	-0,13	0,15	-0,09	0,05													
I	0,14	0,41	-0,06	-0,01	0,44	0,01	0,21	0,21												
K	-0,37	-0,12	-0,55	-0,51	-0,05	-0,47	-0,12	-0,13	-0,24											
L	0,13	0,34	-0,05	-0,01	0,43	0,03	0,11	0,35	-0,16	0,21										
M	0,43	0,41	0,05	0,17	0,54	0,21	0,31	0,51	-0,07	0,44	0,32									
N	-0,26	0,10	-0,41	-0,34	0,13	-0,30	0,02	0,02	-0,43	0,07	0,23	-0,13								
P	-0,16	0,10	-0,33	-0,34	0,20	-0,30	-0,21	0,04	-0,39	-0,03	0,21	-0,25	-0,10							
Q	-0,05	0,07	-0,30	-0,22	0,21	-0,18	-0,10	0,11	-0,26	0,13	0,31	-0,09	-0,18	-0,01						
R	-0,02	0,12	-0,23	-0,26	0,27	-0,12	0,06	0,09	-0,11	0,11	0,22	-0,10	-0,05	0,00	0,04					
S	-0,12	0,14	-0,28	-0,33	0,28	-0,20	0,02	0,09	-0,34	0,15	0,24	-0,21	-0,18	-0,14	0,02	0,01				
T	-0,07	0,18	-0,29	-0,28	0,19	-0,32	-0,06	0,14	-0,35	0,12	0,27	-0,28	-0,31	-0,11	-0,04	-0,12	-0,01			
V	0,06	0,27	-0,16	-0,11	0,43	-0,11	-0,03	0,28	-0,24	0,22	0,35	-0,10	-0,03	0,02	0,02	-0,04	-0,03	0,21		
W	0,19	0,39	0,03	-0,02	0,37	0,08	0,16	0,37	0,06	0,26	0,42	0,14	0,05	0,17	0,12	0,26	0,08	0,31	0,29	
Y	0,11	0,21	-0,07	-0,04	0,45	0,08	0,21	0,25	-0,09	0,34	0,39	0,09	0,03	0,19	0,10	0,19	0,15	0,22	0,32	0,27

Tabelle 4.4: Scores S_{pair_intra} für intramolekulare Aminosäurekontakte
Die Werte werden aus dem Datensatz $Komp_{RN}$ abgeleitet. Sie bewerten, wie häufig zwei benachbarte Aminosäuren in PPKs vorkommen. Beispiel: Das Paar Met-Phe kommt häufiger vor bezogen auf Met-Phe an der restlichen Oberfläche, da sein Score einen positiven Wert von 0,54 hat.

	A	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
A	0.19																			
C	0.57	0.33																		
D	-0.24	-0.11	-0.55																	
E	-0.34	0.13	-0.65	-0.49																
F	0.62	0.58	-0.17	0.04	0.50															
G	-0.59	-0.65	-1.90	-1.54	-0.61	-1.19														
H	0.16	0.06	-0.47	-0.43	0.30	-1.12	0.00													
I	0.59	0.83	-0.13	0.06	0.85	-0.57	0.26	0.55												
K	-0.68	-0.18	-1.24	-1.30	-0.29	-1.92	-0.54	-0.24	-0.70											
L	0.58	0.79	-0.08	0.00	0.88	-0.42	0.35	0.92	-0.36	0.53										
M	0.94	0.96	0.08	0.17	0.89	-0.37	0.54	0.95	-0.26	0.95	0.55									
N	-0.27	0.18	-0.65	-0.68	0.09	-1.67	0.02	0.10	-1.04	0.21	0.41	-0.30								
P	-0.09	-0.04	-0.80	-0.64	0.23	-1.33	-0.62	0.11	-1.04	0.08	0.25	-0.37	-0.24							
Q	-0.04	-0.07	-0.74	-0.73	0.19	-1.61	-0.38	0.24	-0.72	0.21	0.39	-0.36	-0.31	-0.18						
R	-0.18	0.04	-0.71	-0.76	0.06	-1.98	-0.34	0.04	-0.98	-0.08	0.03	-0.55	-0.59	-0.52	-0.31					
S	0.23	0.33	-0.36	-0.52	0.42	-1.06	0.06	0.43	-0.75	0.45	0.69	-0.17	-0.16	-0.14	-0.26	0.08				
T	0.19	0.39	-0.50	-0.39	0.33	-1.17	-0.05	0.41	-0.87	0.42	0.62	-0.44	-0.38	-0.29	-0.41	-0.11	0.09			
V	0.50	0.79	-0.17	-0.12	0.81	-1.00	0.22	0.86	-0.43	0.71	0.79	0.00	0.08	0.16	-0.06	0.21	0.27	0.58		
W	0.45	0.70	-0.22	-0.20	0.49	-0.99	-0.16	0.65	-0.49	0.42	0.59	-0.07	-0.10	-0.14	-0.32	0.28	0.19	0.55	0.23	
Y	0.29	0.32	-0.40	-0.32	0.55	-1.08	0.10	0.37	-0.44	0.55	0.49	-0.06	-0.10	0.03	-0.26	0.18	0.11	0.48	0.25	0.18

Tabelle 4.5: Scores $S^{PIA}_{pair\_intra}$ für intramolekulare Aminosäurekontakte im Zentralbereich der Kontaktfläche
Die Werte werden aus dem Datensatz $Komp_{RN}$ abgeleitet. Sie bewerten, wie häufig zwei benachbarte Aminosäuren in PPKs vorkommen. Beispiel: Das Paar Met-Phe kommt häufiger vor bezogen auf Met-Phe an der restlichen Oberfläche, da sein Score einen positiven Wert von 0,89 hat.

4.3.2.3 Intramolekulare Kontaktpaare

Mit den Scores L_0 kann bewertet werden, wie sehr ein Oberflächenpatch mit der Zusammensetzung von PPKs übereinstimmt. Ein Aufaddieren der L_0-Werte unterstellt jedoch, dass benachbarte Aminosäuren unabhängig voneinander in PPKs vorkommen. Diese naive Vorstellung gilt nicht, wie Befunde aus [18] belegen. Scores, die aus der Häufigkeit benachbarter Aminosäurereste berechnet wurden, klassifizieren PPKs besser als Scores für einzelne Aminosäuren. Daher wurde hier der Score S_{pair_intra} eingeführt. Für dessen Berechnung benötigt man die relative Häufigkeit für das Vorkommen aller Aminosäurepaare an der PPK $f^{cont}(aa_i, aa_j)$ und an der restlichen Oberfläche $f^{surf}(aa_i, aa_j)$. Anschließend lässt sich der Score S_{pair_intra} berechnen als

$$S_{pair\_intra}(aa_i, aa_j) = \log\left(\frac{f^{cont}(aa_i, aa_j)}{f^{surf}(aa_i, aa_j)}\right). \tag{4.6}$$

Tabelle 4.4 zeigt die 210 Werte, die dem Abstandsparameter $s^{rech}_{pair\_intra} = 1,0$ Å zur Bestimmung intramolekularer Kontakte nach (3.8) Seite 24 bestimmt wurden. Mit Hilfe dieser Werte lässt sich nun für eine Aminosäure an der Oberfläche eines Proteins bewerten, wie gut das Aminosäureprofil in der Umgebung dieser Position zu dem aus S_{pair_intra} abgeleiteten Profil passt. Dazu werden die Scores der zu untersuchenden Position mit den Nachbarn an der Oberfläche des Proteins im Umkreis von $s^{anw}_{PW_{pair\_intra}}$ addiert. Anschließend wird durch die Anzahl an Nachbarn geteilt um eine Verzerrung des Signals durch unterschiedliche Anzahl an Nachbarn zu vermeiden. Existiert für den zu untersuchenden Komplex ein paarweises MSA (vgl. Abschnitt 3.3), so wird der Mittelwert verwendet, der sich aus den zugehörigen Spalten ergibt. Im Detail wird dieses Vorgehen in Abschnitt 3.9 beschrieben.

4.3.2.4 Größe der Häufigkeitsunterschiede

Betrachtet man alle in diesem Abschnitt gezeigten Chancenquotienten, so ist trotz aller Varianz in den Daten festzustellen, dass die Unterschiede relativ gering ausfallen. Die größten Differenzen der Scores für einzelne Seitenketten liegen in der Größenordnung 2. Aufgrund der Tatsache, dass bei der Berechnung der Werte der natürliche Logarithmus benutzt wurde, unterscheiden sich die Häufigkeiten lediglich um einen Faktor $2,72^2 = 7,39$. Folglich ist davon auszugehen, dass derartige Signale zwar nützliche Information zur Bestimmung der Kontaktfläche beinhalten, jedoch stark verrauscht sind. Es ist nicht zu erwarten, dass sich Kontaktflächen allein anhand dieser Scores von der restlichen

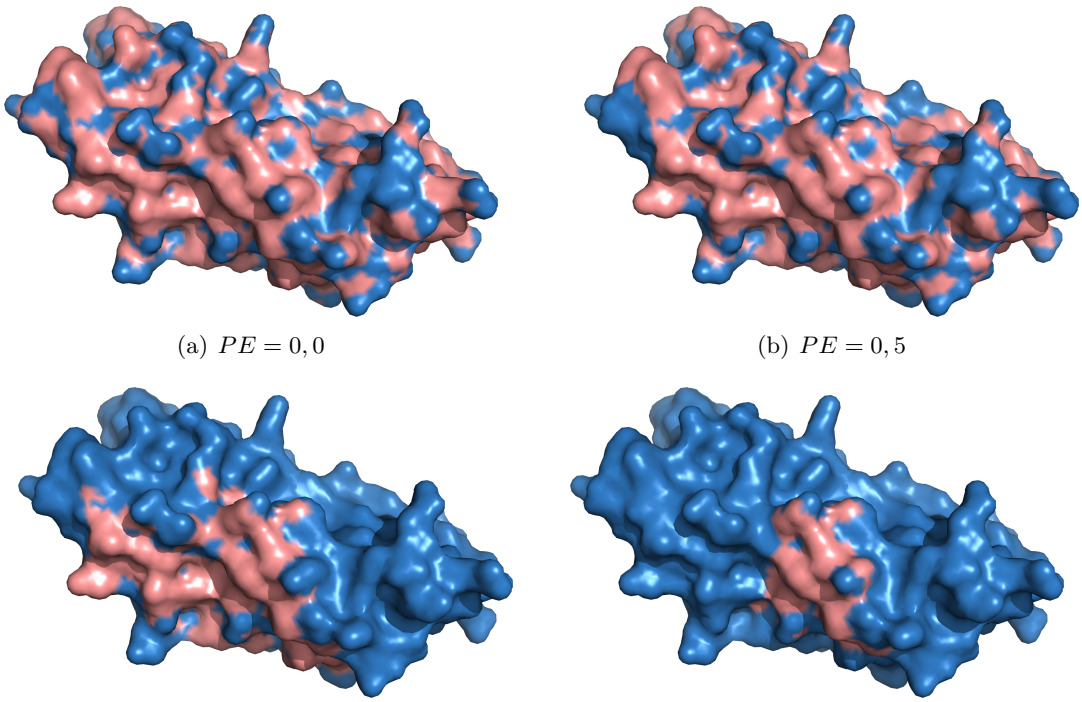
Oberfläche abheben. Die berechneten Scores können jedoch als zusätzliche Information zur Verbesserung der Vorhersage einer Kontaktfläche verwendet werden.

4.3.3 Hydrophobe Patches

Die energetisch wichtigste Rolle bei der Assoziation zweier Proteine zu einem Protein-Protein Komplex spielt der hydrophobe Effekt. Aus diesem Grund besitzen die Kontaktflächen zwischen den Untereinheiten eine ähnliche Aminosäurezusammensetzung wie der Kern eines Proteins, auch wenn der hydrophobe Charakter von PPKs marginal geringer ausfällt [173] [174]. Chothia und Janin haben gezeigt, dass Hydrophobizität die treibende Kraft bei der Assoziation eines obligaten Komplexes ist [65]. Etwas anders sieht die Situation bei *nicht-obligaten* Komplexen aus. Auch hier spielt Hydrophobizität eine Rolle, allerdings eine weniger dominierende [5] [175].

Deshalb finden sich in den Kontaktflächen interagierender Proteine hydrophobe Bereiche, die als Patches organisiert sind und die dazu tendieren aus der Oberfläche herauszuragen. Die Anzahl der Patches und ihre Größe kann zwischen 1 und 15 bzw. 200 und 400 $Å^2$ variieren [50]. Zur Bestimmung lokaler atomarer Details von Hydrophobizität an der Proteinoberfläche findet man in der Literatur nur wenige Methoden. In den meisten Fällen benutzen Kristallographen leicht ungenaue Definitionen von hydrophoben Patches um Proteinoberflächen zu beschreiben [176] [177] [178]. Ähnlich sind meist auch die hydrophoben Patches definiert, mit denen PPKs beschrieben werden [179] [15] [18] [45]. In diesen Arbeiten werden die Patches meist auf Ebene der Aminosäuren definiert. Vorteil dieser einfachen Definition ist es, dass die Patches sehr einfach und schnell bestimmt werden können. Die Hydrophobizität wird jedoch nicht exklusiv von der Aminosäureseitenkette bestimmt, sondern von der Verteilung hydrophober Atome. Als hydrophobe Atome gelten C und S während N und O hydrophilen Charakter zeigen. Desweiteren werden Reste aufgrund einer polaren Gruppe in der Seitenkette allgemein als hydrophil angesehen. Diese polaren Aminosäuren bestehen jedoch ebenfalls zu einem Großteil aus C-Atomen. Daher kommt es häufig vor, dass sich hydrophobe Patches aus hydrophoben Atomen (C und S) verschiedener Aminosäuren zusammensetzen, von denen einige gemeinhin als hydrophil gelten. Dieser Befund wurde jüngst in einer Abschlussarbeit der FU Hagen bestätigt [137].

Im Folgenden wird zur Bestimmung hydrophober Patches das Verfahren *QUILT* benutzt, das erstmals 1996 beschrieben [136] und das vor kurzem im Rahmen einer Masterarbeit der FU Hagen neu implementiert wurde [137]. Diese Implementation wird im Abschnitt 3.8 detailliert erklärt. *QUILT* besitzt den Parameter der polaren Extension



(a) $PE = 0,0$ (b) $PE = 0,5$

(c) $PE = 1,4$ (d) $PE = 2,5$

Abbildung 4.5: Hydrophobe Patches berechnet mit *QUILT*: Der Einfluss der polaren Extension
Hydrophobe Patches, die an der Kette A der HEAM-Bindestelle mit PDB-Code 1BCF berechnet wurden. Rot: hydrophobe Patches, Blau: restliche Oberfläche mit verschiedenen Werten der *polaren Extension* PE

(*PE*), der den Einfluss polarer Atome bei der Berechnung der hydrophoben Patches regelt. Über ihn lässt sich die Größe der hydrophoben Patches regulieren, wie man anhand des Beispiels in Abbildung 4.5 erkennen kann. Je größer der Wert umso besser werden enge "Kanäle" entfernt, umso leichter gehen jedoch auch großflächige hydrophobe Bereiche verloren. Die Abbildung macht deutlich, dass ein Wert von $1,4$ Å für die *polare Extension*, wie er in [136] vorgeschlagen wurde, eine sinnvolle Wahl ist um hydrophobe Patches von einer Größe und Form zu erhalten, wie sie in wildtypischen Proteinen vorkommen. In [137] wurde anhand des Datensatzes $Komp_{RN}$ (siehe Abschnitt 3.1.1) bestätigt, dass diese Wahl sinnvoll ist. Im Folgenden wird jedoch PE als freier Parameter variiert um eine optimale Abdeckung hydrophober Patches an der Kontaktfläche zu gewährleisten.

4.3.4 Bewertung der Konserviertheit einzelner MSA-Spalten

Positionen eines Proteins, die für die Struktur oder Funktion wichtig sind, unterliegen stärkeren Zwängen als weniger wichtige Positionen. Aus diesem Grund setzen sich Mutationen an wichtigen Positionen in der Evolution nicht durch, wenn sie die Fitness der Art vermindern. In einem MSA findet man deshalb fast ausschließlich neutrale Mutationen, die die Funktion des Proteins kaum beeinträchtigen [180]. Umgekehrt sind stark konservierte Positionen, an denen kaum Mutationen gefunden werden, intolerant gegenüber Mutationen und deshalb wichtig um die Funktion des Protein zu gewährleisten. Positionen an der Bindestelle zu einem anderen Protein übernehmen häufig eine wichtige Rolle für die Stabilität des Komplexes und leisten aus diesem Grund einen großen Beitrag für seine Funktion. Daher ist die Kontaktfläche eines Proteins durchschnittlich stärker konserviert als die restliche Oberfläche. Die Konserviertheit einer Spalte im MSA kann folglich als weiterer Hinweis für die Lage einer Position an der Kontaktfläche zu einem Interaktionspartner gewertet werden.

Zur Quantifizierung der Konserviertheit einer Spalte im MSAs wurden in den letzten Jahrzehnten viele verschiedene Verfahren entwickelt (siehe [180]). Es gibt keinen einfachen mathematischen Test für die Güte eines Konserviertheitsscores. Daher ist es nicht leicht, verschiedene Scores miteinander zu vergleichen. Seit den 90er Jahren wurden vor allem Verfahren verwendet, die auf *Shannonscher Entropie* [181] basieren. Während die *Shannon'sche Entropie* ausschließlich die Häufigkeiten der einzelnen Aminosäuren berücksichtigt, bewertet ein neuerer Score auch physikalisch chemische Ähnlichkeiten der Aminosäuren [116]. Dieses Verfahren zur Bewertung von Konserviertheit hat sich im Vergleich zu anderen Scores als robuster bei der Erkennung funktional wichtiger Position erwiesen [117].

PresCont erlaubt es, den Score zur Bewertung von Konserviertheit auszutauschen. So wurden im Rahmen dieser Arbeit Scores basierend auf Shannonscher Entropie und das Verfahren aus [117] vergleichend getestet. Beide Verfahren sind in Abschnitt 3.4 detailliert beschrieben.

4.3.5 Korrelierte Mutationen

Falls eine Seitenkette in einem Proteins mutiert, die im Kontext ihrer Nachbarschaft eine Rolle für die Funktion oder Struktur des Proteins übernimmt, so kommt es häufig vor, dass in ihrer Nähe eine kompensierende Mutation auftritt [119]. Bei Protein-Protein Komplexen finden sich solche korrelierten Mutationen nicht nur innerhalb derselben

Untereinheit [86] [54], sondern auch intermolekular über die PPK hinweg [54] [87].

In diesem Abschnitt soll untersucht werden, inwiefern das Signal intermolekularer korrelierter Mutationen dazu beitragen kann, intermolekulare Kontaktpaare von Aminosäuren vorherzusagen. Als Datengrundlage dazu dient der Datensatz $Komp_{RN}$ (siehe Abschnitt 3.1.1). Da zur Berechnung korrelierter Mutationen zwischen zwei Ketten ein paarweise geordnetes MSA zwingend erforderlich ist (vgl. Abschnitt 3.3) wurde dieser Datensatz weiter ausgefiltert.

Um zu gewährleisten, dass korrelierte Mutationen aus hinreichend großen paarweise sortierten MSAs berechnet werden, müssen zunächst die MSAs der beiden interagierenden Ketten aus der *HSSP*-Datenbank paarweise so aufeinander abgestimmt werden, dass die beiden Sequenzen in jeder Zeile derselben Spezies entstammen. Dies gewährleistet, dass die Proteine zu den Sequenzen einer Zeile im MSA, miteinander interagieren. Sequenzen, für die sich kein Partner im zweiten MSA finden lässt, werden eliminiert. Um eine redundanzfreie und repräsentative Datengrundlage zu gewährleisten, werden anschließend zu ähnliche und zu unähnliche Sequenzen verworfen. Dabei wird sichergestellt, dass für die paarweise Sequenzidentität I, gemessen jeweils am konkatenierten MSA, gilt: $20\% < I < 90\%$.

Sind in einem MSA-Paar nach dieser Filterprozedur noch mindestens 500 Sequenzen vorhanden, so verbleibt der zugehörige Protein-Protein Komplex im Datensatz. Dieser Schwellwert wird so restriktiv gesetzt um eine optimale Datengrundlage zu erhalten und damit das stets vorhandene Rauschen in den Daten so stark wie möglich zu reduzieren. Nach dieser Filterprozedur verbleiben aus dem Datensatz $Komp_{RN}$ 208 Komplexe, deren MSA aus der *HSSP*-Datenbank die Kriterien erfüllen. Dieser Datensatz wird im Folgenden $Komp_{MSA}$ genannt.

Da der $Komp_{MSA}$ sehr viele Homodimere enthält, ist nicht auszuschließen, dass die gemessenen Signale aus intramolekularen Wechselwirkungen anstatt intermolekularer Wechselwirkungen stammen. Wie man an Abbildung 4.6 sieht, weisen die Kontaktflächen von Homodimeren Spiegelsymmetrie auf. Aus diesem Grund interagiert eine zentral in der Kontaktfläche von Kette A gelegene Position $P_1^{(A)}$ nicht nur intermolekular mit einer Position $P_2^{(B)}$ in der identischen Partnerkette B, sondern auch intramolekular mit $P_2^{(A)}$.

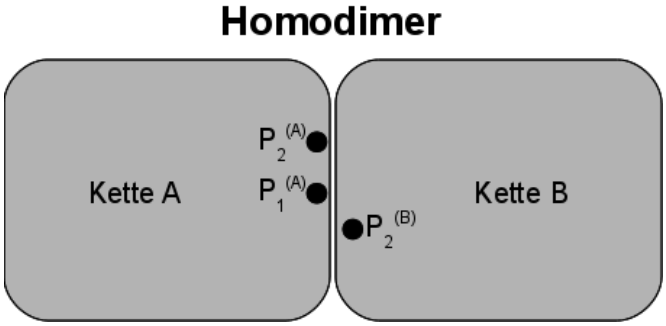


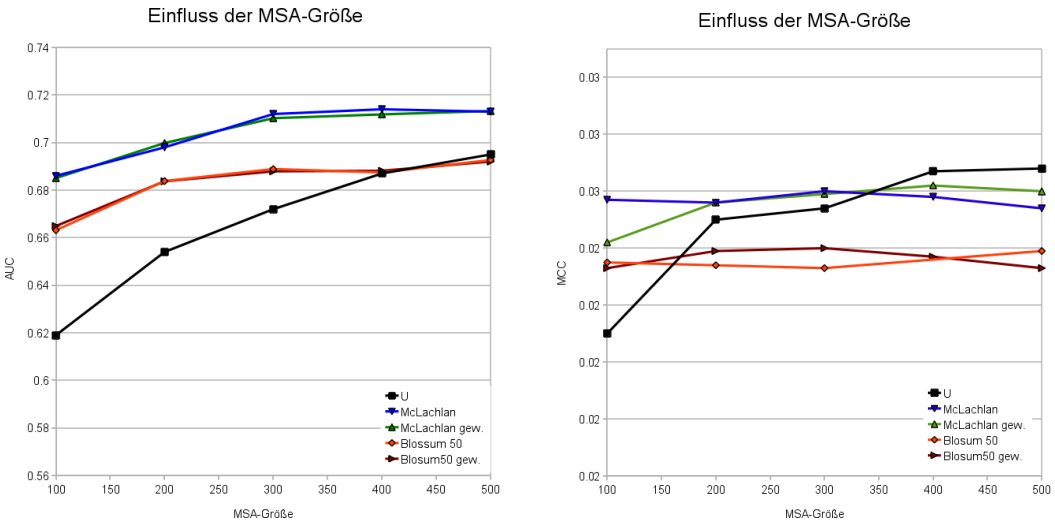
Abbildung 4.6: Spiegelsymmetrie der Kontaktfläche eines Homodimers
Da ein Homodimer aus zwei identischen Untereinheiten besteht, weisen seine Kontaktflächen Spiegelsymmetrie auf. Eine zentral in der Kontaktfläche gelegene Position P_1 interagiert mit einer Nachbarposition P_2 sowohl intra- als auch intermolekular.

4.3.5.1 Bewertung der Klassifikationsleistung

Im Folgenden wird anhand des *MSA*-Datensatzes getestet, inwiefern korrelierte Mutationen dazu benutzt werden können, intermolekulare Kontaktpaare von Seitenketten vorherzusagen. Dazu werden für jeden Komplex und jedes kombinatorisch mögliche Paar an Positionen aus unterschiedlichen Untereinheiten die Korrelationen der zugehörigen Spalten im MSA berechnet. Es wurde dabei sowohl *normierte Transinformation U* (3.22) als auch *Pearson*-Korrelation (3.16) verwendet. Zur Berechnung der *Pearson*-Korrelation ist ein Ähnlichkeitsmaß für Paare von Aminosäuren notwendig. Dazu wurden die *McLachlan*-Substitutionsmatrix [125] bzw. die *BLOSUM50-Matrix* [126] getestet. Daneben wurde auch der Einfluss der Anzahl an Sequenzen im MSAs auf die Vorhersagegenauigkeit untersucht.

Als Datensatz zur Evaluation der Performanz wird dabei der Datensatz $Komp_{RN}$ benutzt. Um zu testen, wie die Vorhersagequalität von der Anzahl an Sequenzen in den MSAs abhängt, werden so lange zufällig gewählte Sequenzen aus den MSAs entfernt, bis die gewünschte Anzahl an Sequenzen erreicht ist.

Zur Bewertung der Klassifikationsleistung werden *ROC*- und *PROC*-Kurven als Methoden zur Bewertung und Optimierung eines Klassifikators verwendet. In diesem Fall wird anhand eines Schwellwertes des *U*-Wertes bzw. der *Pearson*-Korrelation eines Paares entschieden, ob es als interagierendes Paar vorhergesagt wird. Unter Variation des Schwellwertes lassen sich verschiedene Werte der Raten der *wahr Positiven* (TPR) und *falsch Positiven* (FPR) sowie der *Präzision* (*Precision*) und der Trefferquote (*Recall*) bestimmen. Die *ROC*-Kurve ist eine graphische Darstellung der TPR gegen die FPR, während eine *PROC*-Kurve, die *Precision* in Abhängigkeit des *Recalls* unter Variation



(a) (b)

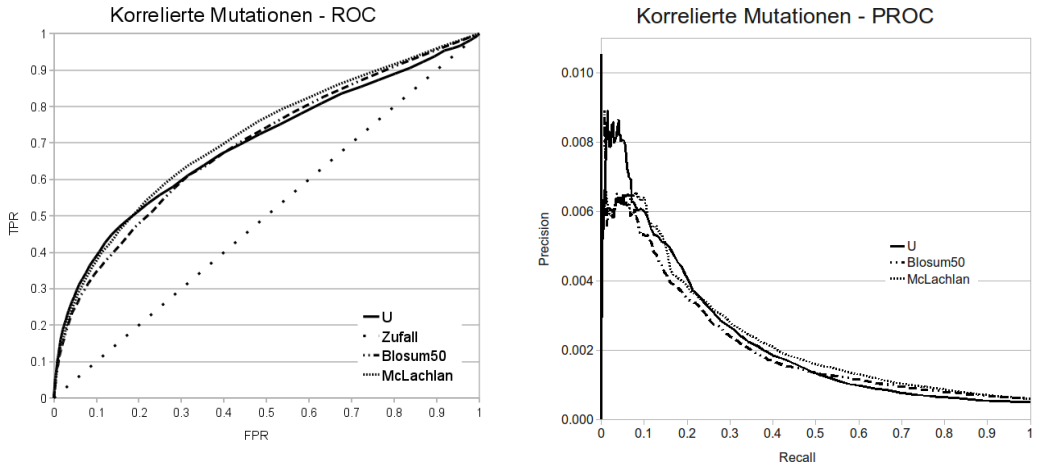
Abbildung 4.7: Qualität der Vorhersage in Abhängigkeit der MSA-Größe Diese Graphen zeigen die Abhängigkeit der Performanz von der MSA-Größe bei Verwendung von *normierter Transinformation* (U-Wert), und *Pearson*-Korrelation mit Hilfe der McLachlan-Matrix bzw. BLOSUM50-Matrix zur Bewertung der Ähnlichkeiten unter den Aminosäuren. Bei Verwendung der *Pearson*-Korrelation wurde auch eine unterschiedliche Gewichtung der Sequenzen im MSA getestet. (a) Abhängigkeit der AUC von der MSA-Größe bei Verwendung des U-Wertes, der *Pearson*-Korrelation mit McLachlan-Matrix und *Pearson*-Korrelation mit BLOSUM50-Matrix (gewichtet und ungewichtet) (b) Abhängigkeit der *MCC* von der MSA-Größe bei Verwendung des U-Wertes, der *McLachlan-Pearson*-Korrelation und der *BLOSUM50-Pearson*-Korrelation (Sequenzen jeweils gewichtet und ungewichtet).

der Schwelle zeigt. Ein weiteres Kriterium zur Bewertung der Klassifikationsleistung ist der *Metthews Correlation Coefficient (MCC)*, der angibt, wie stark die Vorhersage mit der realen Klasseneinteilung korreliert ist. Näheres dazu siehe Abschnitt 3.14.

Abbildung 4.7(a) zeigt den Einfluss der Anzahl der Sequenzen im MSA N_S auf die Performanz der Klassifikation gemessen anhand der *AUC* der zugehörigen *ROC*-Kurve. Man erkennt, dass die AUC wie erwartet für alle Methoden mit der N_S zunimmt. Während jedoch bei den Methoden basierend auf *Pearson*-Korrelation einem Wert für N_S von etwa 300 Sequenzen kaum mehr eine Verbesserung zu beobachten ist, steigt bei Benutzung der U-Werte auch für MSAs mit vielen Sequenzen die Güte der Vorhersage weiter mit N_S. Anhand des Kriteriums der *AUC* übertrifft die Performanz der Pearson-Korrelation unter Verwendung der *McLachlan* Matrix die Performanz der U-Werte für MSAs mit bis zu 500 Sequenzen. Es ist zu vermuten, dass für extrem große MSAs mit beiden Methoden dieselbe *AUC* erreicht werden kann, jedoch lässt sich dies aufgrund fehlender Datengrundlage nicht überprüfen. Vergleicht man die Performanz der

Pearson-Korrelation bei Verwendung unterschiedlicher Ähnlichkeitsmatrizen für Aminosäuren, so stellt man fest, dass die *McLachlan*-Matrix deutlich höhere Werte der *AUC* ermöglicht als die Verwendung der *BLOSUM50*-Matrix. Es wurde weiterhin getestet, ob bei Verwendung der Pearson-Korrelation eine unterschiedliche Gewichtung der Sequenzen im MSA (siehe Abschnitt 3.4.2) die Qualität der Vorhersage verbessern kann. Wie Abbildung (4.7(a)) zeigt, lassen sich jedoch keine signifikanten Unterschiede aufgrund unterschiedlichen Gewichtung der Sequenzen feststellen.

Nimmt man den *MCC* als Kriterium für die Performanz der Vorhersage, so ergibt sich ein ähnliches Bild. In Abbildung 4.7(b) erkennt man, dass bei Benutzung der U-Werte auch der *MCC* stetig mit N_S zunimmt. Bei korrelationsbasierten Methoden ist jedoch der *MCC* weitgehend unabhängig von N_S. Für MSA-Größen $N_S > 300$ Sequenzen übertrifft der *MCC* des U-Wertes denjenigen der Pearson-Korrelation. Auch gemessen am *MCC* wird bei Verwendung der *McLachlan*-Matrix eine signifikant höhere Performanz erreicht als mit der *BLOSUM50*-Matrix.



(a) (b)

Abbildung 4.8: Vergleich verschiedener Methoden zur Bewertung korrelierter Mutationen
Man erkennt die (a) *ROC*- bzw. (b) *PROC*-Kurven der Klassifikation von Kontaktpaaren durch Auswertung korrelierter Mutationen über U-Werte und Pearson-Korrelation unter Verwendung der McLachlan-Ähnlichkeitsmatrix für Aminosäuren und der BLOSUM50-Matrix.

Die *AUC* einer ROC-Kurve und der maximale *MCC* erlauben es, einen Klassifikator zu bewerten. Sie geben jedoch keinerlei Auskunft darüber, ob ein Klassifikator seine Stärken bei einem restriktiveren Schwellwert besitzt, wenn nur wenige Testbeispiele als *positiv* vorhergesagt werden und damit die *TPR* auf Kosten der *FPR* erhöht wird, oder bei einem weniger restriktiven Schwellwert. Dies erkennt man erst anhand der *ROC*-Kurven. In Abbildung 4.8(a) erkennt man, dass die höhere *AUC* der Methoden

basierend auf Pearson-Korrelation aus der besseren Klassifikation bei weniger restriktivem Schwellwert resultiert. Bei einem sehr hohen Schwellwert, der eine geringe *FPR* und *TPR* impliziert, übertrifft der U-Wert die Methoden der Pearson-Korrelation anhand der *AUC*.

Da die *TPR* und die *FPR* jeweils unabhängig von der Anzahl der negativen bzw. positiven Beispiele im Datensatz sind, kann bei einer stark asymmetrischen Aufteilung der Daten in die beiden Klassen die *ROC*-Kurve eine gute Klassifikationsleistung anzeigen, auch wenn die *Precision* der Vorhersage extrem gering ist. Daher besitzt bei derartigen Datensätzen, die *PROC*-Kurve meist eine größere Aussagekraft über die Qualität eines Klassifikators als die *ROC*-Kurve. Abbildung Abbildung 4.8(b) zeigt die *PROC*-Kurven dieses Klassifikationsproblems. Man sieht, dass anhand der *Precision* der U-Wert die Vorhersagequalität der Methoden basierend auf *Pearson*-Korrelation bei Wahl eines restriktiven Schwellwertes deutlich übertrifft. Während man über Pearson-Korrelation lediglich 0,65% *Precision* erreichen kann, erhält man bei Benutzung des U-Wertes 0,89%. Damit lässt sich für die Klassifikation über den U-Wert eine Anreicherung der Vorhersage mit *wahr Positiven* um einen Faktor 18 von $0,049\%$ im ursprünglichen Datensatz auf $0,89\%$ feststellen.

Im Falle eines derart asymmetrischen Datensatzes, bei dem die Anzahl der Negativbeispiele die Positivbeispiele um den Faktor 10^3 übertrifft, muss der Schwellwert für die Klassifikation sehr restriktiv gewählt werden. In solchen Fällen ist es sinnvoll, einige *falsch negative* Vorhersagen in Kauf zu nehmen, um den Anteil von *wahr positiven* Vorhersagen unter den Positiven hoch zu halten. Daher befindet sich der wichtige Bereich der Kurven in Abbildung 4.8 bei geringen Werten der *FPR* bzw. des *Recalls*. Bei restriktiver Wahl des Schwellwertes übertrifft die Klassifikation der *U*-Werte diejenige der *Pearson*-Korrelation. Insgesamt lässt sich daher sagen, dass anhand dieses Tests die *U*-Werte eine leicht bessere Klassifikationsleistung zeigen als die *Pearson*-Korrelation.

4.3.5.2 Signifikanzschwellen Korrelierter Mutationen

Um eine Schwelle zu bestimmen, über der U-Werte bzw. die *Pearson-Korrelation* auf eine signifikante Korrelation hinweisen, wurde untersucht, wie sich der Wertebereich der Korrelationsmaße anhand zufälliger MSAs verhält, deren Spalten unabhängig voneinander sind. Die zufälligen MSAs wurden aus den MSAs der *HSSP*-Datenbank zum gesamten Datensatz nach *Mintz* generiert, die nach der Entfernung zu ähnlicher und zu unähnlicher Sequenzen noch mindestens 500 Sequenzen enthielten. Da die Signifikanzschwelle von der Anzahl der Sequenzen im MSA abhängt, wurden MSAs der gewünschten Größe durch zufällige Auswahl der benötigten Anzahl an Sequenzen generiert. Anschließend wurde die Reihenfolge der Symbole in jeder Spalte unabhängig von allen anderen Spalten zufällig neu bestimmt. Damit ist sichergestellt, dass keine Kopplungen mehr zwischen den Spalten bestehen. Anschließend lässt sich abhängig von der Anzahl der Sequenzen im MSA eine Signifikanzschwelle durch das Kriterium festlegen, dass lediglich $0,01\%$ aller Werte aus zufälligen MSAs die Schwelle überschreiten dürfen.

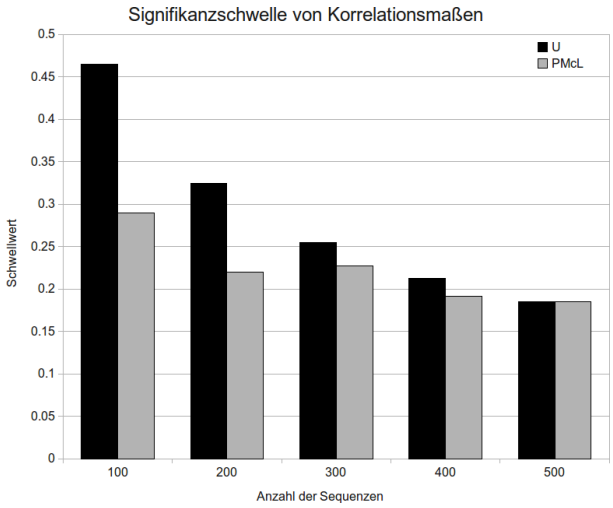


Abbildung 4.9: Signifikanzschwellen korrelierter Mutationen
Die Säulen geben den Schwellwert in Abhängigkeit der Anzahl an Sequenzen im MSA an, über dem ein U-Wert bzw. eine Pearson-Korrelation signifikant ist. Die Schwelle wurde anhand eines zufälligen MSAs der entsprechenden Größe als derjenige Wert bestimmt, über dem weniger als 0,01% aller Korrelationen liegen.

Anhand von Abbildung 4.9 erkennt man, dass die Signifikanzschwelle sowohl für U als auch für $PMcL$ erwartungsgemäß mit der Anzahl an Sequenzen in der zufälligen MSAs sinkt. Während jedoch ab einer Anzahl von 200 Sequenzen die Schwelle für $PMcL$ nur noch gering variiert, sinkt die Schwelle für U auch bei mehr als 300 Sequenzen im MSA noch weiter. Dieser Befund deckt sich mit dem Ergebnis aus dem letzten Abschnitt, dass der Aussagekraft von U im Gegensatz zu $PMcL$ auch bei großen MSAs mit mehr als 300 Sequenzen noch weiter mit der Anzahl der Sequenzen im MSA zunimmt.

4.3.5.3 Beispiele für korrelierte Mutationen

Die oben eingeführten Verfahren zur Identifizierung von korrelierten Mutationen bewerten das gemeinsame Vorkommen von Aminosäurepaaren auf abstrakte Weise. Um eine Idee zu bekommen, welche Paare von Aminosäuren an Positionen auftreten, die ein deutliches Korrelationssignal aufweisen, werden zwei Beispiele näher betrachtet. Dabei wird versucht, das Korrelationssignal anhand der physikalisch-chemischen Eigenschaften der Aminosäuren zu interpretieren.

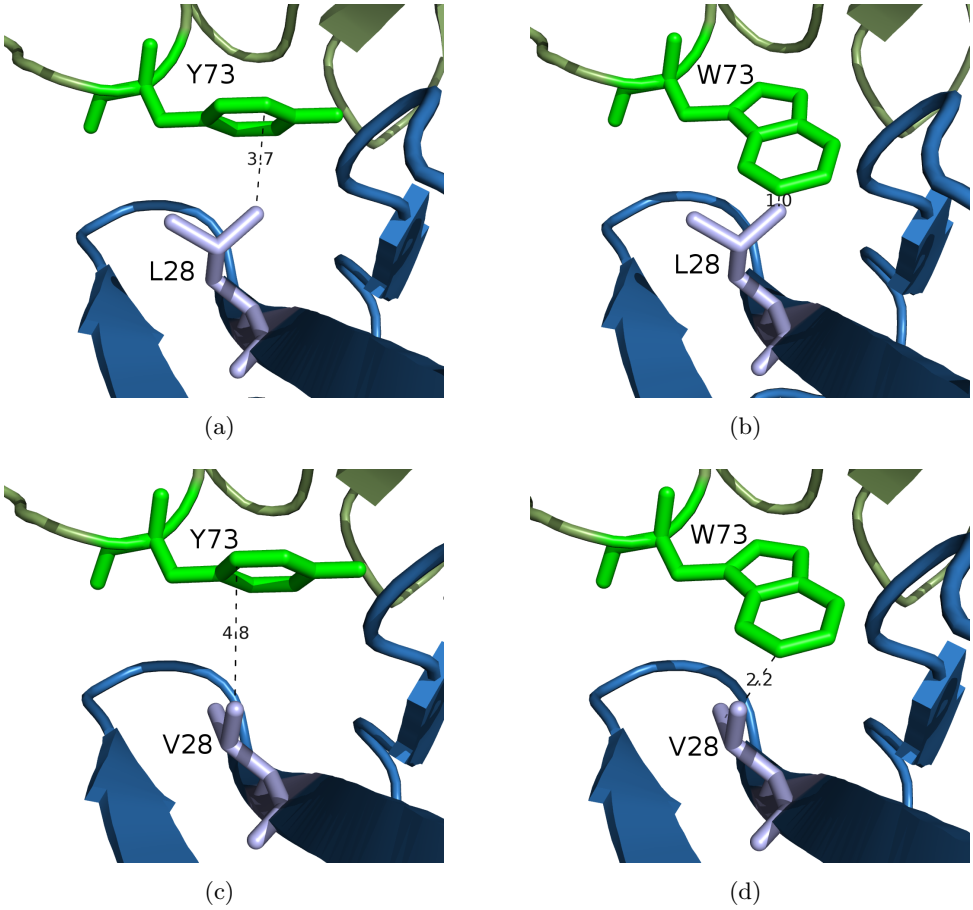


Abbildung 4.10: Korrelierte Mutation aufgrund der Aminosäuregröße
Die Abbildungen zeigen eine korrelierte Mutation zwischen den Positionen 73 und 28 der *Superoxid Dismutase* aus *Salmonella typhimurium* (PDB-ID 1EQW, Ketten A und C). In der Kristallstruktur (a) finden sich Y73 und L28 in einer Distanz von 3,7 Å. Einzelne Mutation Y73W bzw. L28V resultieren in einem Clash (b) bzw. einem Loch in der Kontaktfläche. Die gleichzeitige Mutation von Y73W und L28V (d) ergibt ohne Berücksichtigung von Backbonedeformationen einen Abstand von 2,2 Å zwischen den beiden Seitenketten. Dieser schwache Clash kann durch Backbonedeformationen ausgeglichen werden. Zur Einführung der *in silico* Mutationen und zur Erstellung der Grafiken wurde *PyMol* benutzt [148].

Im Homodimer der *Superoxid Dismutase* aus *Salmonella typhimurium* (PDB-ID: 1EQW) weisen die Positionen 73 und 28 eine deutliche Korrelation auf. U-Wert für dieses Spaltenpaar ist $0,480$ und die *Pearson*-Korrelation bei Verwendung der *McLachlan* Ähnlichkeitsmatrix (PMcL) liefert $0,487$. Für die Berechnung wurde ein MSA nach Abschnitt 3.3 unter Verwendung von *BLAST* [100] und *muscle* [106] generiert. Dieses enthält nach Vorverarbeitung durch den Ähnlichkeitsfilter noch 123 Sequenzen. An Abbildung 4.9 auf Seite 88 lässt sich ablesen, dass die Werte für U und $PMcL$ über dem Schwellwert für das Signifikanzniveau von $0,01\%$ liegen. Damit ist sichergestellt, dass eine signifikante Kopplung dieser beiden Positionen besteht.

Anhand der absoluten Werte für die paarweise Aminosäurehäufigkeiten in Tabelle 4.6 erkennt man, dass dieses Signal aus einer *kanonischen* korrelierten Mutation resultiert. Mit *kanonisch* ist in diesem Zusammenhang gemeint, dass genau zwei Paare von Aminosäuren (bzw. ähnlichen Aminosäuren) bevorzugt auftreten und alle anderen Paare nur geringe Häufigkeiten in den entsprechenden Spalten der MSAs zeigen. In diesem Fall sind dies die beiden Paare YL und WV, die mit ähnlicher Häufigkeit vorkommen. Alle anderen Kombinationen sind selten.

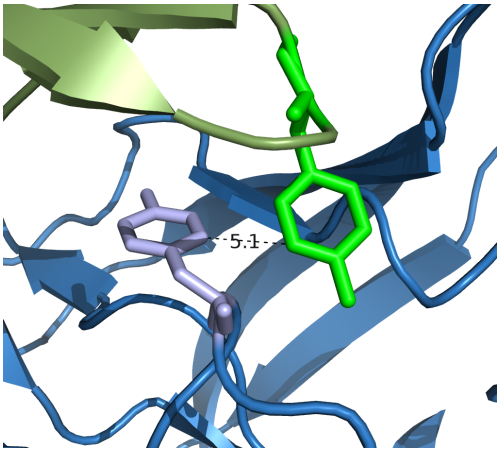


Abbildung 4.11: Korrelierte Mutation mit nicht-klassischem Charakter
Die Abbildung zeigt die beiden Kontaktreste Y28 und Y73, deren Spalten deutlich korreliert sind ($U = 0,480$ und $r = 0,487$).

Anhand der Struktur in Abbildung 4.10 erkennt man, dass sich die beiden Positionen an der Kontaktfläche gegenüberstehen. Die im MSA beobachteten Mutationen wurden *in silico* eingeführt. Aufgrund der unterschiedlichen Größe der wildtypischen und mutierten Seitenkette verursachen die beiden Einzelaustausche zu einer größeren (Y73W) bzw. kleineren (L28V) Seitenkette einen Zusammenstoß der Seitenketten bzw. einen Hohlraum an der Kontaktfläche. Beides wirkt sich energetisch äußerst ungünstig auf die Stabilität des Komplexes aus. Beim Doppelaustausch W73Y, L28V hingegen schafft das kleinere V an Position 28 Platz für das größere W an Position 73 und vermindert dabei den Zusammenstoß der beiden Seitenketten. Der Rest der verbleibenden Spannung wird vermutlich durch ein lokales Rearrangement aufgelöst.

Die Seitenkette Y73 ist in eine weitere korrelierte Mutation an der Kontaktfläche eingebunden. Die aromatischen Ringe des Paares TYR25 und TYR73 sind lediglich 5 Å

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-
C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
D	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
E	-	-	-	6	-	-	1	1	2	1	-	-	-	1	1	-	-	-	-	-
F	-	-	-	1	-	-	-	-	-	2	-	-	-	-	-	-	-	6	-	-
G	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
H	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
I	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
K	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
L	1	-	-	1	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-
M	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-
N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
P	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Q	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-
R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
S	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
T	-	-	-	-	-	-	-	2	-	1	-	-	-	1	-	-	-	-	-	-
V	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-
W	7	-	-	-	-	-	-	1	-	3	1	-	-	-	-	-	1	37	-	-
Y	-	-	-	1	-	-	-	2	-	27	-	-	-	3	-	-	-	2	-	-

Tabelle 4.6: Absolute Paarhäufigkeiten einer klassischen korrelierten Mutation
Die Tabelle enthält die absoluten Paarhäufigkeiten, die an den Positionen 73 und 28 der *Superoxid Dismutase* (PDB-ID: 1EQW) vorkommen. Zur Bestimmung der Häufigkeiten wurde ein MSA ausgewertet, das auf die Homodimer-Struktur des Enzyms aus *Salmonella typhimurium* projiziert wurde. Einträge des Wertes null sind durch - gekennzeichnet.

voneinander entfernt (siehe Abbildung 4.11) und zeigen ebenfalls einen auffälligen U-Wert von $U = 0,514$ sowie eine hohe *McLachlan-Pearson-Korrelation* von $r = 0,612$. Dieses Paar an Positionen ist somit unabhängig von der verwendeten Methode zur Bewertung korrelierter Mutationen deutlich miteinander korreliert. In Tabelle 4.7 finden sich die absoluten Werte der beobachteten Paare von Aminosäuren in den zugehörigen Spalten des MSAs. Wie man sieht, handelt es sich dabei um keine klassische Korrelation von zwei dominierenden Paaren von Seitenkettenpaaren wie im letzten Fall. Hier fordert ein Y25 entweder ein Y73 oder ein W73. Für den Fall, dass sich jedoch an Position 25 kein Y befindet besteht an Position 73 eine größere Freiheit bei der Wahl der Seitenkette.

Diese beiden Beispiele belegen, dass zwischen zwei Positionen komplexe Abhängigkeiten in der Besetzung mit Aminosäuren bestehen können. Im einfachsten Fall, den kanonischen Korrelationen treten überwiegend zwei disjunkte Paare von Aminosäuren auf. Bei nicht kanonischen Paaren sind mehrere Kombinationen möglich.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-
C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
D	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-
E	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
F	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	1
G	-	-	1	3	-	1	-	-	-	-	1	-	-	-	-	-	1	-	2	-
H	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-
I	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
K	-	-	1	7	-	-	-	-	-	1	-	-	-	1	-	-	1	-	-	-
L	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
N	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	1
P	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Q	1	-	-	-	7	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-
R	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-
S	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
V	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-
W	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
Y	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	40	30

Tabelle 4.7: Absolute Paarhäufigkeiten einer nicht-kanonischen korrelierten Mutation
Die Tabelle enthält die absoluten Paarhäufigkeiten, die an den Positionen 73 und 25 der *Superoxid Dismutase* (PDB-ID: 1EQW Ketten) vorkommen. Zur Bestimmung der Häufigkeiten wurde ein MSA ausgewertet, das auf die Homodimer-Struktur des Enzyms aus *Salmonella typhimurium* projiziert wurde. Einträge des Wertes null sind durch - gekennzeichnet.

4.3.6 Einbeziehung des Interaktionspartners – Konnektivität

Die meisten früheren Arbeiten zur Vorhersage von PPKs verwenden Scores, die Hydrophobizität, Aminosäurehäufigkeiten, Oberflächenbeschaffenheit, Exponiertheit oder evolutionäre Konserviertheit bewerten. D.h. zur Vorhersage der PPK werden nur Eigenschaften des betrachteten Proteins herangezogen. In der Regel wird kein Wissen über den Interaktionspartner verwendet. Dieser ist jedoch häufig bekannt. Da die PPKs der interagierenden Proteine eine gewisse Passgenauigkeit aufweisen müssen, ist zu vermuten, dass die Vorhersagegenauigkeit verbessert werden kann, wenn mögliche Interaktionen zwischen den Aminosäuren bewertet werden.

Informationen wie intermolekulare Chancenquotienten S_{pair_inter} und korrelierte Mutationen lassen sich als Potentiale, die nicht nur von der Position im Protein abhängen, die es zu bewerten gilt, sondern auch von eventuellen Kontaktpartnern in der anderen Proteinkette, nicht direkt in die SVM als Merkmal mit einbeziehen. Aus diesem Grund wird die Idee genutzt, dass die wirklichen Kontaktaminosäuren häufiger gute Scores bzw. intermolekularer Merkmale aufweisen als andere Oberflächenreste. Zum einen können die Signale der richtigen Kontaktreste im Interaktionspartner das Signal über das Hintergrundrauschen heben. Zum anderen existieren Signale korrelierter Mutationen auch

über weitere Distanzen als den in Abschnitt 3.2 definierten Abstandsschwellwert. So können auch Positionen, die im Komplex nur mittelbar benachbart liegen signifikante Korrelationen vorweisen.

Im Folgenden bezeichne der Begriff "Position" sowohl die Position einer Seitenkette in einem Protein als auch die zugehörige Spalte im MSA. Weiter sei das monomere Protein, für das die Kontaktfläche bestimmt werden soll das Query-Protein und seine Positionen p_k mit $k = 1, \ldots, n_p$. Der bekannte Interaktionspartner werde als Partnerprotein bezeichnet und seine Positionen seien als q_l mit $l = 1, \ldots, n_q$ gekennzeichnet. Ziel dieses Abschnittes ist es für jede Position p_k im Query-Protein aus den korrelierten Mutationen und den intermolekularen paarweisen Scores zu allen Positionen q_l des Partnerproteins einen Wert zu berechnen, anhand dessen sich abschätzen lässt, wie wahrscheinlich die Position p_k an der Kontaktfläche liegt.

Als erstes werden für alle Positionen p_k im Query-Protein und q_l im Partnerprotein mit $k = 1, \ldots, n_p$, $l = 1, \ldots, n_q$ Scores für korrelierte Mutationen $C(k, l)$ nach Abschnitt 3.5.1 oder 3.5.2 bestimmt und die intermolekularen paarweisen Chancenquotienten $PW_{pair\_inter}(aa_i, aa_j)$ als $S_{pair\_inter}(k, l) = PW_{pair\_inter}(Type(p_k), Type(q_l))$ abgelesen, wobei die Funktion $Type(p)$ den Aminosäuretyp an Position p ausgibt. Steht ein paarweises MSA zur Verfügung, so kann man die paarweisen Chancenquotienten über die entsprechenden Spalten im MSA mitteln. Daraus lässt sich dann für jedes Positionenpaar p_k und q_l mit $k = 1, \ldots, n_p$ und $l = 1, \ldots, n_q$ eine Linearkombination

$$L(k, l) = C(k, l) + w \cdot S_{pair\_inter}(k, l) \tag{4.7}$$

berechnen. Das positive Gewicht w bleibt als Parameter noch zu bestimmen. Wie in Abbildung 3.11 auf Seite 50 kann man diese Daten auch anhand eines Netzwerkes darstellen, dessen Knoten für die Positionen p_k und q_l aus beiden Ketten stehen und dessen Kanten zwischen den Knoten aus unterschiedlichen Proteinketten hohe Werte von L_{kl} repräsentieren. Aus diesem Grund lassen sich Konzepte aus der Netzwerkanalyse zur Bewertung der Ergebnisse verwenden. Ein allgemein genutzter Parameter zur Charakterisierung einzelner Knoten im Netzwerk ist die Konnektivität (siehe [182] und Referenzen darin). Zur Vorhersage funktional wichtiger Positionen anhand intramolekularer korrelierter Mutationen wurde die Konnektivität bereits in einer früheren Arbeit verwendet [88]. Die Konnektivität gibt die Anzahl an Kanten mit hohem Score an, die einen Knoten verbinden. Auf diese Art lassen sich Positionen im Query-Protein p_k identifizieren, die zu mehreren Positionen im Partnerprotein q_l einen hohen Score L_{kl} besitzen. Dies stellt dann einen Hinweis darauf dar, dass die entsprechende Position p_k an der PPK zum Partnerprotein liegt.

4.4 Der Klassifikator – Verrechnung der positionsspezifischen Eigenschaften

In den letzten Abschnitten wurden nichtredundante Eigenschaften von Oberflächenaminosäuren vorgestellt, anhand derer sich Kontaktpositionen von Positionen an der restlichen Oberfläche unterscheiden. Um daraus eine Vorhersage der Kontaktfläche zu generieren, müssen diese Eigenschaften miteinander kombiniert und verrechnet werden. Dazu wird eine Support Vektor Maschine (SVM) als überwachtes Lernverfahren verwendet. SVMs sind ein überwachtes Lernverfahren um Vektoren von Attributen anhand von Mustererkennung in zwei Klassen zu unterteilen. Aufgrund ihrer hohen Performanz und Robustheit insbesondere gegen das Phänomen des *Overlearnings* wurden sie in der Bioinformatik bereits häufiger bei der Analyse von Microarray-Datensätzen [183] oder wie in dieser Arbeit zur Vorhersage von Proteinbindestellen benutzt [46].

4.4.1 Training und Eingabedaten der SVM

Als überwachtes Lernverfahren benötigt die SVM zum Trainieren einen korrekt klassifizierten Datensatz aus Seitenketten an der PPK und an der restlichen Oberfläche einer Untereinheit. Diese können aus den in Abschnitt 3.1 vorgestellten Datensätzen von Protein-Protein Komplexen $Komp_{kanon}$ und $Komp_{trans}$ gewonnen werden.

Wie in Abschnitt 3.12.3 begründet, muss eine SVM anhand eines symmetrischen Datensatzes trainiert werden, d.h. ein Datensatz, der ebenso viele positive wie negative Beispiele beinhaltet. Da die hier benutzten Datensätze von Oberflächenaminosäuren eine deutliche Asymmetrie aufweisen, müssen beim Training der SVM ebenso viele Nicht-Kontaktreste zufällig ausgewählt werden, wie Kontaktreste vorhanden sind. Um die Diversität der Daten zu gewährleisten wird diese Auswahl für jeden Protein-Protein Komplex im Datensatz einzeln vorgenommen, so dass jeder Komplex ebenso viele Seitenketten an der PPK zum Datensatz beiträgt, wie an der restlichen Oberfläche.

In dieser Arbeit entspricht jeder Eingabevektor der SVM den Daten einer Oberflächenaminosäure. Seine Komponenten setzen sich zusammen aus den 5 Werten für Konserviertheit, Konnektivität, der relativen $SASA$, einem Mittelwert von S_{pair_intra} über alle intramolekularen Nachbarn an der Oberfläche und der Zugehörigkeit zu einem hydrophoben Patch. Die Zugehörigkeit zu einem hydrophoben Patch wird binär beschrieben. Die restlichen Eigenschaften werden mit reellen Zahlen dargestellt.

Die in dieser Arbeit verwendete C-SVM besitzt den Parameter C über den sich die Toleranz gegenüber fehlerhaft klassifizierten Datenpunkten durch das Konzept der *Soft Margin Hyperplane* justieren lässt [140] [184]. Außerdem wird als Kernelfunktion eine Gaußfunktion mit Parameter γ benutzt.

4.4.2 Optimierung der Parameter

Um eine optimale Klassifikation in Kontaktreste und Nicht-Kontaktreste zu erhalten, müssen alle freien Parameter optimiert werden. Zu den freien Parametern zählen sowohl Parameter der Berechnung der Chancenquotienten und der Input-Daten der SVM als auch Parameter der SVM selbst.

4.4.2.1 Optimierung der Chancenquotienten

Die Abstände $s^{anw}_{PW_{pair\_inter}}$ bzw. $s^{anw}_{PW_{pair\_intra}}$, die dazu dienen, benachbarte Positionen für die Berechnung der Chancenquotienten PW_{pair_intra} bzw. PW_{pair_inter} auszuwählen, sind als freier Parameter geeignet zu wählen. Dabei muss ein passender Kompromiss zwischen einer zu restriktiven und einer zu allgemeinen Wahl gefunden werden.

Wird im Falle von PW_{pair_inter} der Abstandsparameter $s^{anw}_{PW_{pair\_inter}}$ (3.8) zu groß gewählt, so werden auch Reste als intermolekularer Kontakt gezählt, die in der Struktur zu weit entfernt voneinander liegen um physikalisch miteinander interagieren zu können. Dadurch werden die gesuchten Signale stärker verrauscht und die Spezifität der Scores in PW_{pair_inter} sinkt. Wird dagegen $s^{anw}_{PW_{pair\_inter}}$ zu klein gewählt, so werden nicht alle physikalisch miteinander interagierenden Reste als intermolekulare Kontaktpaare identifiziert. Es besteht dann die Gefahr, dass selektiv Wechselwirkungen mit längerer Reichweite wie elektrostatische Wechselwirkungen ganz oder teilweise ausgefiltert werden und deshalb PW_{pair_inter} keine Information über derartige Kontaktpräferenzen beinhaltet.

Ähnlich verhält es sich mit dem Abstandsparameter $s^{anw}_{PW_{pair\_intra}}$ bei der Berechnung von PW_{pair_intra}. Wird $s^{anw}_{PW_{pair\_intra}}$ zu groß gewählt, so sind die Signale aufgrund des großzügigen Distanzkriteriums zu verrauscht, um optimal Abhängigkeiten vom Aminosäuretyp zu finden. Ist dagegen $s^{anw}_{PW_{pair\_intra}}$ zu klein, so werden nicht alle vorhandenen intramolekularen Nachbarn ausgezählt und es geht Information verloren.

4.4.2.2 Bewertung der Konserviertheit und korrelierter Mutationen

Ein weiterer Punkt, an dem die Charakterisierung von Kontaktflächen optimiert werden kann sind die Verfahren zur Berechnung von Konserviertheit und korrelierten Mutationen. Zur Berechnung der Konserviertheit einer Spalte im MSA haben sich Methoden basierend auf Shannonscher Entropie etabliert [107] [181]. Außerdem wurde ein Verfahren beschrieben [117], das zusätzlich physikalisch-chemische Ähnlichkeiten unter den Aminosäuren anhand einer Ähnlichkeitsmatrix berücksichtigt. Es sollte aufgrund der Verwertung dieser zusätzlichen Information robuster sein. Während der Parameteroptimierung werden diese Verfahren daraufhin getestet, inwiefern sie zur Vorhersage von Protein-Protein Kontaktflächen beitragen können.

Auch zur Bewertung korrelierter Mutationen existieren mehrere Verfahren. Daher werden auch verschiedene Methoden zur Bewertung korreliert mutierender Spalten in MSAs auf ihren Einfluss zur Vorhersage von Kontaktflächen hin untersucht.

4.4.2.3 Parameter bei der Berechnung der Eingabedaten

Bei der Berechnung der fünf Eigenschaften eines Oberflächenrestes gibt es eine Reihe von freien Parametern, deren Werte geeignet zu bestimmen sind.

- Bei der Auswertung der inter- sowie der intramolekularen Chancenquotienten entscheidet der Abstandsparameter $s^{anw}_{PW_{pair\_inter}}$ bzw. $s^{anw}_{PW_{pair\_intra}}$ über die Auswahl der Nachbarpositionen. Für seine Wahl gelten analoge Argumente wie bei der Berechnung der Scoring-Tabellen PW_{pair_inter} und PW_{pair_intra}.

- Die intermolekularen Chancenquotienten PW_{pair_inter} und der Konserviertheitsscore werden über eine Linearkombination miteinander verrechnet, bevor sie über die Konnektivität ausgewertet werden (siehe Abschnitt 4.3.6). Zur Berechnung der Linearkombination (4.7) muss das Gewicht $w > 0$ als freier Parameter so gewählt werden, dass PW_{pair_inter} und der Score für die korrelierten Mutationen optimal gewichtet werden.

- Bei der Berechnung der Konnektivität aus den N möglichen Werten der Linearkombination intermolekularer Scores, die durch paarweise Kombinationen aller Positionen aus beiden Untereinheiten entstehen, werden die höchsten $x \cdot N$ $(0 < x < 1)$ Werte benutzt. x muss so gewählt werden, dass sich Kontaktpaare möglichst gut vom Untergrundrauschen derjenigen Reste abheben, die im Komplex nicht benachbart liegen.

- Bei der Berechnung hydrophober Patches fällt der Parameter der *polaren Extension* an. Der ursprünglich gewählte Wert von $1,4$ Å [136] erzeugt hydrophobe Patches mit realistischer Größe und Verteilung, muss jedoch variiert werden, um mit dieser Methode hydrophobe Patches an Protein-Protein Kontaktflächen optimal abzudecken.

4.4.2.4 Parameter der SVM

Die verwendete SVM besitzt zwei freie Parameter:

- Als Kernel-Funktion wird eine RBF-Funktion nach (3.51) verwendet. Ihr Parameter γ ist so zu wählen, dass die zugehörige Transformation ϕ die Daten möglichst gut linear separabel in einen hochdimensionalen Raum abbildet.

- Die Algorithmen zum Trainieren und Testen der C-SVM beinhalten den Parameter C, der die Bedeutung fehlerhaft klassifizierter Punkte im Optimierungsproblem (3.47) variiert.

4.4.3 Bestimmen der Klassifikationsleistung

In diesem Abschnitt soll die Klassifikationsleistung der SVM bewertet werden. Dabei wird zunächst von einem optimierten Datensatz an Parametern ausgegangen und die maximale Klassifikationsleistung gemessen. Anschließend soll die Bedeutung der einzelnen Bausteine für die Performanz untersucht werden. Dazu werden ausgehend von den optimalen Einstellungen nacheinander verschiedene Parameter verändert und die daraus resultierende Verschlechterung der Klassifikationsleistung bestimmt.

4.4.3.1 Strategien zur Auswertung der Klassifikationsleistung

Zum Training der SVM benötigt man einen Trainingsdatensatz bestehend aus den Daten für die fünf genannten Merkmale von korrekt als Kontakt- und Nicht-Kontaktresten gekennzeichneten Positionen der Proteine. Für eine faire Evaluation der Performanz ist es wichtig, das Phänomen des *Overlearnings* ausschließen zu können. *Overlearning* tritt auf, falls der Klassifikator den Datensatz, an dem er trainiert wurde, besser vorhersagen kann, als unbekannte Testdaten. Dies würde bedeuten, dass der Klassifikator seine Entscheidung anhand von Unterschieden in den fünf Merkmalen vornimmt, die spezifisch

für einzelne Trainingsbeispiele sind, anstatt von Unterschieden, anhand derer sich die beiden Klassen generell unterscheiden. Wäre ersteres der Fall, so könnte der Klassifikator zwar seine Trainingsbeispiele richtig einordnen, würde jedoch aufgrund fehlender Generalisierung beim Testen unbekannter Beispiele scheitern. Um keine durch *Overlearning* künstlich erhöhte Performanz zu messen ist es daher wichtig, Trainings- und Testdatensätze stets getrennt zu halten.

Ist die Anzahl der Datensätze gering, so geschieht dies über eine Strategie, die als *Leave One out* bezeichnet wird. Im konkreten Fall wird dabei in jeder Runde ein Protein-Protein Komplex PPK_{sel} ausgewählt und aus dem Datensatz entfernt. Mit den Daten der Kontakt- und Oberflächenaminosäuren aller restlichen Komplexe ist anschließend die SVM zu trainieren. Im nächsten Schritt wird dann die trainierte SVM dazu verwendet, die Kontaktreste des ausgewählten Komplexes PPK_{sel} vorherzusagen. Dies wird so lange wiederholt, bis alle Protein-Protein Komplexe einmal getestet wurden. Da PPK_{sel} nicht zum Trainieren verwendet wurde, ist sichergestellt, dass die Performanz nicht künstlich durch *Overlearning* erhöht wurde.

Da die Trainingsbeispiele aus der restlichen Oberfläche zufällig ausgewählt werden, unterscheiden sich die resultierenden Hyperebenen der SVM zwischen zwei Trainingsläufen leicht. Daher werden in der folgenden Bewertungen der Performanz stets 10 SVMs mit jeweils zufällig ausgewählten Negativbeispielen trainiert und mit diesen anschließend der Testdatensatz ausgewertet. Ein Maß für die Performanz wird anschließend als Mittelung über die 10 Vorhersagen berechnet.

Die Bewertung der Klassifikationsleistung bei gegebenem Parametersatz erfolgt anhand einer *Receiver operating characteristic* (ROC) und dem damit verbundenen Maß der *Area Under the Curve* (AUC) (siehe Abschnitt 3.14.1). Die resultierende Fläche ist bei Mittelung über 10 Trainingsläufe bis auf die dritte Nachkommastelle reproduzierbar, was dem Rahmen der Messgenauigkeit entspricht.

4.4.3.2 Der optimale Parametersatz

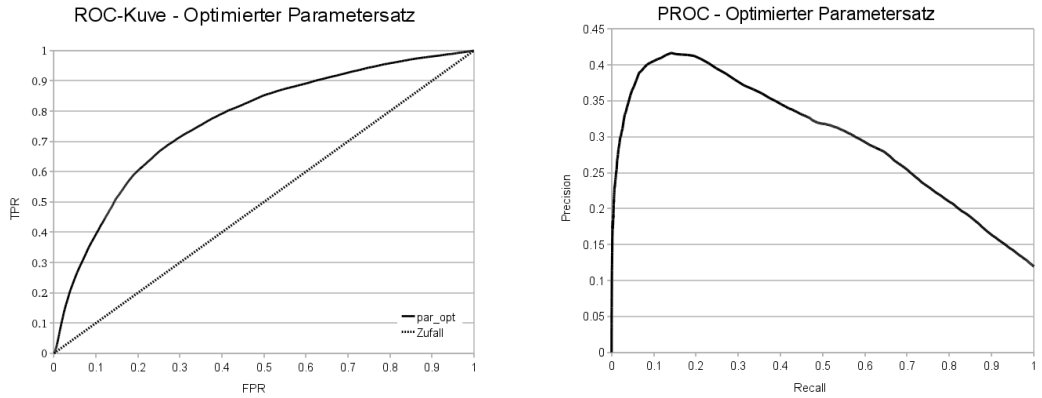
Für alle freien Parameter des Verfahrens müssen derart Werte gefunden werden, dass die Performanz der Klassifikation optimiert wird. Häufig legt man dazu für alle Parameter einen Satz an Werten fest, der getestet werden soll und testet alle Kombinationen dieser Werte durch um eine optimale Wahl der Parameter zu finden. Dieses als *Grid*-Suche bezeichnete Verfahren kann hier aufgrund der hohen Anzahl an freien Parametern nicht durchgeführt werden. Aufgrund der Art der Parameter ist jedoch zu vermuten, dass

sie unabhängig voneinander gewählt werden können. Daher können nacheinander einzelne bzw. wenige Parameter gemeinsam verändert werden, um einen Satz optimalen Parametern zu finden. Dieser Satz an Parametern wurde bestimmt, wobei der Datensatz $Komp_{kanon}$ verwendet wird. Dabei war es das Ziel, die Klassifikationsleistung zu maximieren.

Parameter	**Wert**
$s^{rech}_{pair\_intra}$	1,0 Å
$s^{rech}_{pair\_inter}$	0,5 Å
$s^{anw}_{pair\_intra}$	3,0 Å
$s^{anw}_{pair\_inter}$	0,5 Å
S_{Kons}	C_{Sim_Freq}
S_{Corr}	U
w_{Conn}	0,0
x_{Conn}	0,033
PE	1,7 Å
γ	0,0625
C	2,0

Tabelle 4.8: Parametersatz optimiert für den Datensatz $Komp_{kanon}$
Es wurden sowohl optimale Parameter zur Berechnung der Eingabe-Daten als auch optimale Parameter der *SVM* bestimmt. $s^{rech}_{pair\_intra}$ und $s^{rech}_{pair\_inter}$ sind die Abstandsparameter zur Definition der intra- bzw. intermolekularen Nachbarschaft bei der Berechnung von PW_{pair_intra} bzw. PW_{pair_inter}. $s^{anw}_{pair\_intra}$ und $s^{anw}_{pair\_inter}$ bezeichnen analoge Abstandsparameter, die zur Bewertung einer Position für optimal befunden wurden. S_{Kons} bezeichnet die Methode, mit der Konserviertheit bewertet wird, während S_{Korr} für die Methode zur Bewertung korrelierter Mutationen steht. w_{Konn} ist das Gewicht von PW_{pair_inter} in der Linearkombination mit dem Score für korrelierte Mutationen. x_{Conn} stellt den Bruchteil signifikanter Werte bei der Berechnung der Konnektivität dar. PE bezeichnet die polare Extension bei der Berechnung hydrophober Patches. γ und C sind die beiden Parameter der SVM.

Es resultierten die in Tabelle 4.8 aufgelisteten Parameter. Bei ihrer Verwendung resultiert die ROC-Kurve in Abbildung 4.12(a) mit $AUC = 0.7659$. Die *PROC*-Kurve in Abbildung 4.12(b) erreicht eine *Precision* von 42% bei einem *Recall* von 15%. Diese Leistung wird von einem Klassifikator erreicht, der vier Eigenschaften auswertet, die sich alleine aus der Besetzung der Proteinposition und seiner intramolekularen Nachbarschaft ergeben. Lediglich die fünfte Eigenschaft der Konnektivität ergibt sich aus der Berücksichtigung des Interaktionspartners. Nachbarschaft.



(a) (b)

Abbildung 4.12: *ROC*- und *PROC*-Kurven nach Optimierung der Eingabe-Parameter
(a) Die *ROC*-Kurve der *SVM* mit den optimierten Parametern C und γ. Die Eingabedaten wurden mit den in Tabelle 4.8 angegebenen optimierten Parametern berechnet. Ihre AUC beträgt $0,7670$. (b) Die zugehörige *PROC*-Kurve erreicht eine *Precision* von 42% bei einem *Recall* von 15%.

4.4.3.3 Der Einfluss der einzelnen Eigenschaften

Im Folgenden soll untersucht werden, welchen Einfluss die einzelnen Eigenschaften auf die Klassifikationsleistung haben. Nachdem in den letzten Abschnitten der optimale Parametersatz für alle positionsspezifischen Eigenschaften bestimmt wurde, kann nun die Bedeutung der einzelnen Merkmale für die Klassifikationsleistung untersucht werden. Dazu wird innerhalb der Prozedur der *Leave One out Kreuzvalidierung* nacheinander je eine der fünf Eigenschaften weggelassen. Anhand der Verschlechterung der Performanz lässt sich die Bedeutung dieser Eigenschaft für die Klassifikationsleistung abschätzen.

Eigenschaft	**AUC**	**$\mathbf{AUC_{full}}$−AUC**
SASA	0,6284	0,1377
Konnektivität	0,7612	0,0049
Konserviertheit	0,7452	0,0209
PW_{pair_intra}	0,7342	0,0319
HPA	0,7280	0,0381

Tabelle 4.9: Die Bedeutung der einzelnen Eigenschaften für die Qualität der Vorhersage
In dieser Tabelle sind die *AUC* ohne die in der ersten Spalte angegebene Eigenschaft und die Differenz zur *AUC* bei Verwendung aller 5 Eigenschaften gelistet. Der Wert AUC_{full} bei Verwendung aller fünf Eigenschaften beträgt $0,7661$

Tabelle 4.9 belegt, dass der relativen *SASA* die größte Bedeutung zukommt. Lässt

man sie weg, so kommt es zu einem dramatischen Abfall der AUC von $0,7661$ auf einen Wert von $0,6284$. Dies liegt zum Teil daran, dass im Datensatz $Komp_{kanon}$ die Monomerstrukturen durch Zerlegen der Dimerstruktur erzeugt werden. In der Tabelle folgen die beiden Scores, die vor allem den hydrophoben Charakter der Kontaktfläche beschreiben. Auch die Konserviertheit leistet einen deutlichen Beitrag zur Verbesserung der Performanz, während die Konnektivität den Wert der AUC lediglich um $0,0049$ erhöht, was jedoch einen deutlich messbarer und reproduzierbarer Effekt ist.

4.4.3.4 Der Einfluss des Abstandsschwellwertes $\mathbf{s^{anw}_{pair\_intra}}$

In diesem Abschnitt soll der Einfluss des Parameters $s^{anw}_{pair\_intra}$ auf die Klassifikationsleistung gemessen werden. Dazu werden alle anderen Parameter bei ihren optimalen Werten aus Tabelle 4.8 belassen und lediglich der Schwellwert für $s^{anw}_{pair\_intra}$ zwischen $0,5$ Å und 10 Å variiert. Die Änderung der Klassifikationsleistung der SVM wird anschließend über ROC-Kurven und deren AUC gemessen. Abbildung 4.13 zeigt die Änderung der AUC bei Variation von s_{pair_intra}. Man erkennt ein Maximum bei $s^{anw}_{pair\_intra} = 3.0$ Å, das von zwei Minima umgeben ist.

Der Grund für das Auftreten dieser Minima ist nur schwer zu erklären. Möglicherweise nimmt der optimale Abstand bestimmter funktionaler Gruppen von Aminosäureresten bei einer Distanz von 3.0 Å$+vdW_1+vdW_2$ ein energetisches Minimum an. Deshalb, sind die Rotamere günstiger Interaktionspartner, die diese Distanz leicht verfehlen würden, schwächer populiert.

Eine SVM, die ohne Auswertung der Chancenquotienten PW_{pair_intra} arbeitet, erreicht lediglich einen AUC-Wert von 0.7331. Dies ist deutlich weniger als wenn man den Abstandparameter $s^{anw}_{pair\_intra}$ extrem hoch setzt und damit sämtliche Positionen an der Oberfläche der monomeren Proteinkette als intramolekulare Kontakte wertet. Für $s^{anw}_{pair\_intra} = 100$ Å lässt sich beispielsweise $AUC = 0.7530$ messen.

Dies mag im ersten Moment verwundern, da es bei einem Abstandsschwellwert $s^{anw}_{pair\_intra}$ in der Größenordnung des Proteindurchmessers keinerlei Abhängigkeiten der Aminosäurearten der als "benachbart" gewerteten Positionen gibt. Weil also bei einem derart hohen Schwellwert $s^{anw}_{pair\_intra}$ sämtliche Positionen an der Oberfläche als Partnerpositionen gewertet werden, kann die Ursache der gemessenen Verbesserung der AUC nur im Typ der zu bewertenden Aminosäure selbst liegen. Die paarweisen Chancenquotienten PW_{pair_intra} beinhalten folglich implizit Information über die Bevorzugung bestimmter Typen von Aminosäuren an der Kontaktfläche. Dies erkennt man auch daran, dass die Werte in PW_{pair_intra} (Tabelle 4.4) für Aminosäuretypen, die nach Tabelle 4.1 an der

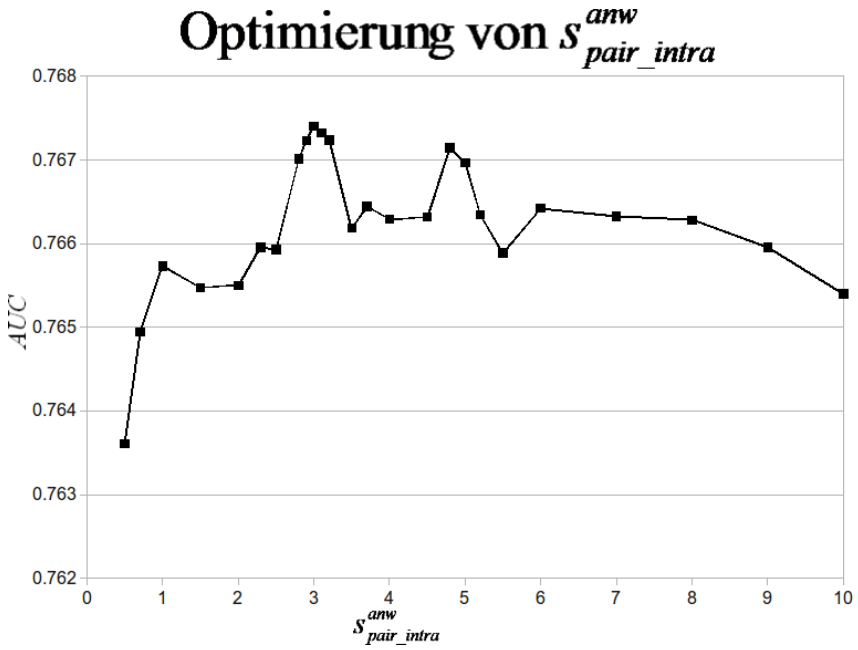


Abbildung 4.13: Einfluss des Abstandsschwellwertes $s^{anw}_{pair\_intra}$ auf die Klassifikationsleistung
Der Wert des Abstandsschwellwertes $s^{anw}_{pair\_intra}$ wurde zwischen 0,5 Å und 10 Å variiert und es wurde jedesmal der AUC-Wert gemessen. $s^{anw}_{pair\_intra}$ ist der Abstandsschwellwert zur Definition der Nachbarschaft bei der Anwendung von PW_{pair_intra}.

Kontaktfläche bevorzugt sind, im Durchschnitt größer sind als Werte für Aminosäuretypen, die an der restlichen Oberfläche bevorzugt vorkommen.

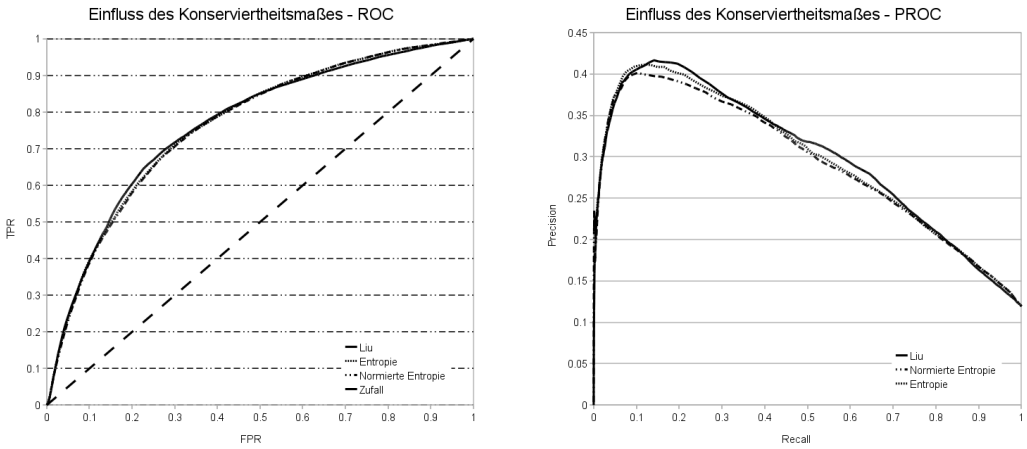
4.4.3.5 Bewertung der Konserviertheit

In der optimierten SVM wird die Konserviertheit einer Spalte im MSA durch die in [117] beschriebene Methode durch den Score C_{Sim_Freq} bewertet. Der Vorteil dieser Methode gegenüber *Shannonscher Entropie* besteht darin, dass sie nicht nur Häufigkeiten, sondern auch physikalisch-chemische Ähnlichkeiten unter den Aminosäuretypen berücksichtigt. Es stellt sich die Frage, wie die Methode zur Bewertung der Konserviertheit die Klassifikationsleistung beeinflusst. Daher wurde zusätzlich die Konserviertheit mit Hilfe

	AUC	**MCC**
C_{Sim_Freq}	0,7670	0,306
Entropie	0,7652	0,290
Norm. Entropie	0,7636	0,284

Tabelle 4.10: Einfluss des Konserviertheitsmaßes auf die Vorhersagequalität

der Entropie (siehe Abschnitt 3.4.1) und der normierten Entropie bestimmt. In Tabelle 4.10 sind die zugehörigen Kennwerte angegeben. Dabei zeigt sich ein ähnliches Bild von der Situation wie in den zugehörigen *ROC*- und *PROC*-Kurven in Abbildung 4.14. Der Score C_{Sim_Freq} eignet sich etwas besser zur Vorhersage von Protein-Protein Kontaktflächen als Shannonche Entropie. Eine Normierung der Entropie nach den Aminosäurehäufigkeiten vermindert die Güte der Vorhersage sogar marginal. Insgesamt sind die Unterschiede jedoch relativ gering wenn man berücksichtigt, dass die Klassifikationsleistung auf eine *AUC* von $0,7452$ sinkt, wenn die Konserviertheit nicht bewertet wird.



(a) (b)

Abbildung 4.14: Einfluss des Konserviertheitsmaßes auf die Vorhersagequalität
Die Abbildung zeigt den Einfluss des Konserviertheitsmaßes auf die (a) *ROC* bzw. (b)*PROC*-Kurven des Klassifikators zur Vorhersage der Kontaktfläche. In beiden Kurven zeigt der Score nach *Liu et al.* die beste Performanz.

4.4.3.6 Bewertung korrelierter Mutationen

Informationen über intermolekulare korrelierte Mutationen an der Kontaktfläche fließt über das Maß der Konnektivität in die Vorhersage der Kontaktfläche mit ein. Korrelierte Mutationen werden mit drei Verfahren bewertet. Der *U*-Wert ist ein Maß für die wechselseitige Information. Die beiden anderen Werte sind Korrelationskoeffizienten, die mit unterschiedlichen Scores (*McLachlan*, *BLOSUM50*) für die Aminosäureähnlichkeiten berechnet werden.

In Tabelle 4.15 wird die Performanz des optimalen Klassifikators unter Verwendung des *U*-Wertes verglichen mit Klassifikatoren, in denen korrelierte Mutationen über

	AUC	**MCC**
U	0,7670	0,304
Pearson McLachlan	0,7643	0,302
Pearson Blosum50	0,7645	0,301

Abbildung 4.15: Einfluss der Methode zur Berechnung korrelierter Mutationen

Pearson Korrelation quantifiziert werden. Dazu ist eine Ähnlichkeitsmatrix für Aminosäurepaare nötig. Als solche werden die *McLachlan*-Substitutionsmatrix [125] und die *BLOSUM50*-Matrix [126] getestet. Man erkennt, dass *normierte Transinformation* etwas bessere Vorhersagen liefert als Methoden basierend auf *Pearson-Korrelation*. Die Unterschiede sind jedoch auch hier marginal. Der geringe Einfluss der Methode zur Berechnung korrelierter Mutationen ist jedoch vor dem Hintergrund zu sehen, dass insgesamt Konnektivität, über die korrelierte Mutationen in die Vorhersage mit einfließen, eine untergeordnete Bedeutung besitzt. Wird der Parameter Konnektivität nicht berücksichtigt, so sinkt die AUC lediglich auf den Wert 0,7612.

4.4.3.7 Einfluss des Konnektivitätsparameters *x*

Die Konnektivität wird zur Bewertung von Informationen aus intermolekularen Scores herangezogen. Wichtig bei der Berechnung ist der Anteil x mit $0 < x < 1$ aller kombinatorisch möglichen Paarungen von Positionen der beiden Ketten, der zur Berechnung der Konnektivität herangezogen wird. Wird x zu klein gewählt, so werden zu viele sich kontaktierende Paare nicht bewertet. Wird x dagegen zu groß gewählt, so werden auch Aminosäurepaarungen mit kleineren, nicht signifikanten Scores bewertet, wodurch sich das Signal erniedrigt. Es ist daher notwendig, diesen Parameter zu bestimmen. In diesem Abschnitt wird untersucht, welchen Einfluss die Wahl von x auf die Klassifikationsleistung der SVM hat.

In Abbildung 4.16 ist die Abhängigkeit der AUC vom Konnektivitätsparameter x aufgetragen. Man erkennt zwei Maxima bei Werten von $x = 0,033$ und $x = 0,313$. Der Grund für die beiden Maxima ist unklar. Eigentlich würde man annehmen, dass es einen optimalen Schwellwert gibt und AUC monoton abnimmt, je weiter man sich von ihm entfernt. Da die AUC des ersten Maximums etwas größer ist, wird $x = 0,033$ als optimaler Wert gewählt. Ohne Berücksichtigung der Konnektivität erreicht die SVM eine AUC von $0,76117$. Man erkennt aus diesem geringen Abfall der AUC, dass die Bedeutung der Konnektivität für die Klassifikationsleistung nicht allzu groß ist.

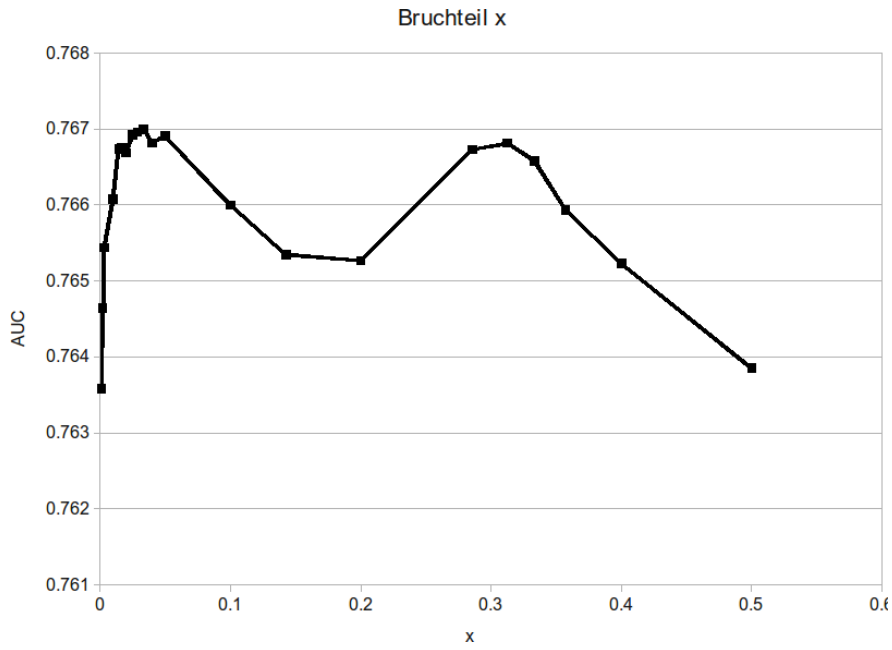


Abbildung 4.16: Der Einfluss des Konnektivitätsparameters x
Diese Abbildung zeigt die Performanz des Klassifikators gemessen anhand der AUC in Abhängigkeit des Bruchteils x aller paarweisen Werte, die zur Bestimmung der Konnektivität benutzt werden.

4.4.3.8 Die polare Extension

Es ist bekannt, dass Hydrophobizität ein wichtiger Parameter zur Bestimmung von PPKs ist [65]. Da anzunehmen ist, dass in PPKs größere hydrophobe Bereiche auftreten, werden hier hydrophobe Patches bewertet. Bei der Methode *QUILT* zur Bestimmung hydrophober Patches [136] bestimmt der Parameter der polaren Extension PE Größe und Aussehen der hydrophoben Patches. Der in der Literatur angegebene Wert von $1,4$ Å wird variiert um das Optimum für die Klassifikationsleistung von *PresCont* zu finden.

Abbildung 4.17 zeigt die Performance des Klassifikators gemessen anhand der AUC bei Variation der polaren Extension. Man erkennt deutlich das Maximum bei $PE = 1,7$ Å. Erhöhen bzw. Erniedrigen dieses Wertes vermindert die Klassifikationsleistung. Dieser Befund stimmt gut mit der Erwartung überein, dass für zu große Werte von PE die hydrophoben Patches an der Kontaktfläche nicht vollständig abgedeckt werden, während zu kleine Werte von PE auch hydrophile Bereiche an der Oberfläche zu den Patches zählt.

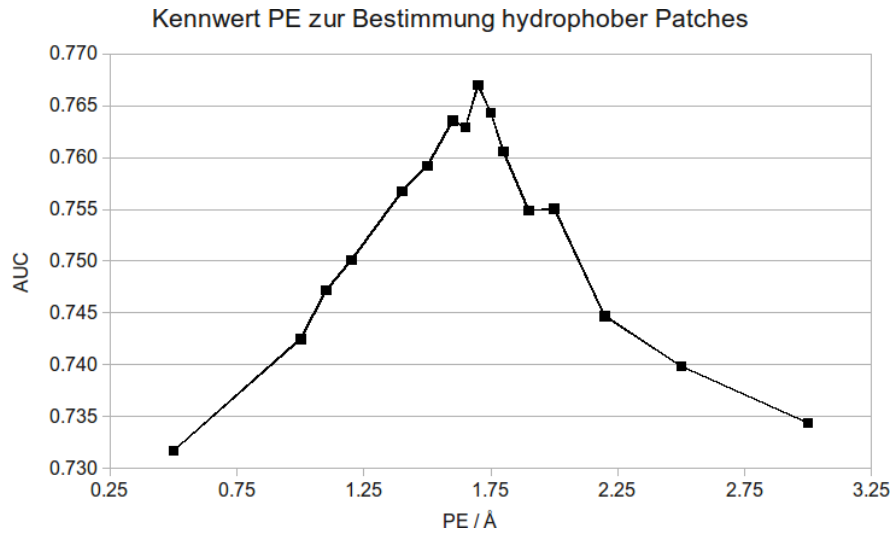


Abbildung 4.17: Der Einfluss der polaren Extension PE
Die Performanz des Klassifikators gemessen anhand der *AUC* in Abhängigkeit des Parameters der *polaren Extension PE* bei der Berechnung der hydrophoben Patches. Das Optimum liegt bei 1,70 Å.

4.4.3.9 Die Parameter der SVM

Da die SVM-Parameter γ und C nicht unabhängig voneinander sind, wurde ihre Optimierung anhand einer Grid-Suche durchgeführt. Um den Parameterraum effektiv absuchen zu können wurden exponentiell wachsende Werte aus dem Bereich $\{2^{-8}, 2^{-7}, \ldots, 2^{8}\}$ für γ und C kombiniert, wie es in [140] vorgeschlagen wird. Das Ergebnis dieser Grid-Suche ist in Abbildung 4.18 anhand einer Farbkarte dargestellt. Es ergab sich ein relativ breiter Bereich mit hoher Performanz gemessen anhand der AUC der zugehörigen ROC-Kurve, was für die Stabilität des Verfahrens spricht. Das globale Maximum lag bei $C = 2^8$, $\gamma = 2^{-6}$ mit $AUC = 0.7679$. Daneben gibt es ein lokales Maximum bei $C = 2^2$, $\gamma = 2^{-4}$ mit $AUC = 0.7670$. Da die Unterschiede in der AUC zwischen beiden Parametersätzen sehr gering sind und ein hoher Wert von C die beim Training benötigte Rechenzeit stark ansteigen lässt, wurden im Folgenden die Parameter des lokalen Maximums bei $C = 2^2$, $\gamma = 2^{-4}$ mit $AUC = 0.7670$ benutzt.

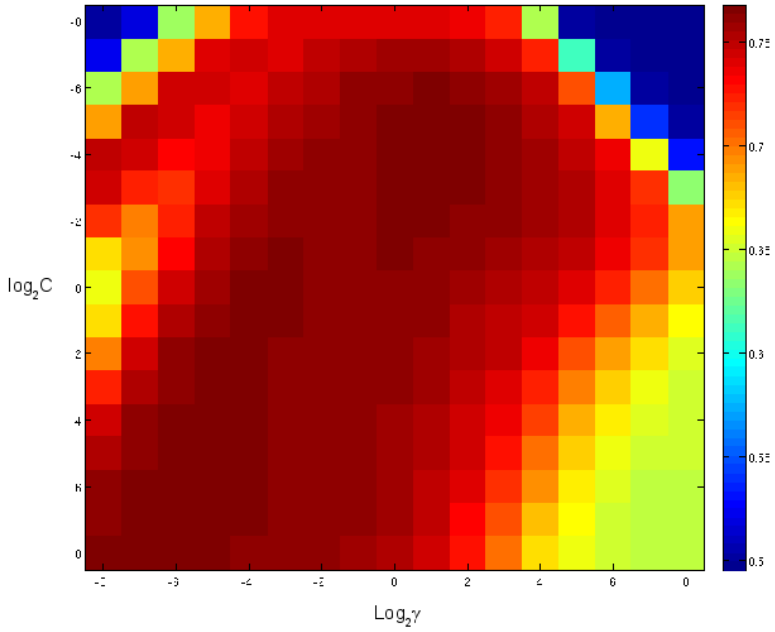
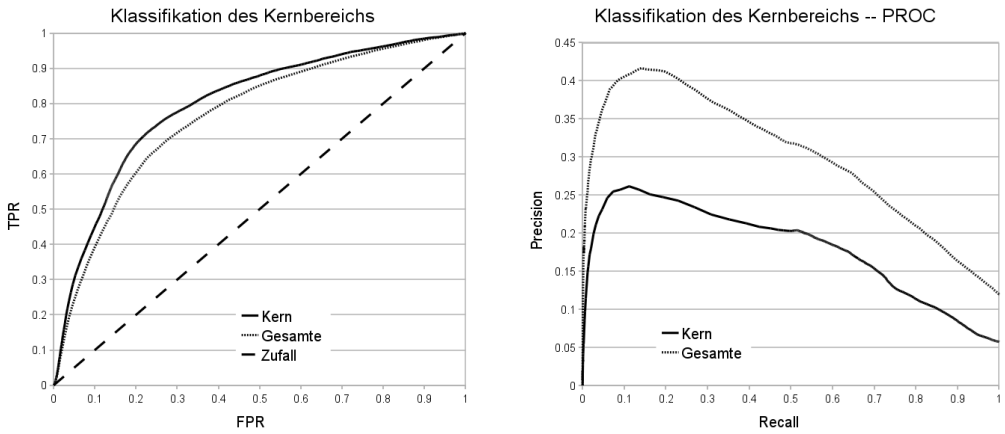


Abbildung 4.18: Performanz der SVM bei der Grid-Suche nach optimalen SVM-Parametern
Der Plot zeigt die Abhängigkeit der Performanz des Klassifikators gemessen über die *AUC* der *ROC*-Kurven bei Variation der SVM-Parameter γ und C während einer Grid-Suche. Zwei Optima fanden sich bei $C = 2^2, \gamma = 2^{-4}$ und $2^8, \gamma = 2^{-6}$

4.5 Klassifikationsleistung im Kernbereich von Kontaktflächen

Bei der Bewertung der Häufigkeiten von Seitenketten an der Kontaktfläche und im Zentralbereich der Kontaktfläche mit Hilfe von Chancenquotienten zeigte sich, dass sich die Häufigkeiten der Seitenketten an verschiedenen Bereichen der Kontaktfläche unterscheiden. Im Folgenden soll die Frage beantwortet werden, ob Seitenketten im Kernbereich der Kontaktfläche allgemein einfacher von Positionen an der restlichen Oberfläche unterschieden werden können als Seitenketten am Rand einer Kontaktfläche. Dazu wird untersucht, wie sich die Performanz von *PresCont* im Zentralbereich der Kontaktfläche verhält. Eine gute Datengrundlage zur Untersuchung des Kernbereiches von Kontaktflächen bietet der Datensatz $Komp_{RN}$. Er enthält keine Spezialfälle ineinander verschlungener Hauptketten, die eine Einteilung der Kontaktfläche in Kern- und Randbereich beeinflussen könnten.

Um die Vorhersagequalität von *PresCont* an verschiedenen Bereichen der Kontaktfläche bewerten zu können, wird analog zum vorhergehenden Abschnitt vorgegangen, wobei alle dort optimierten Parameter übernommen werden. Um ein *Overlearning* auszuschließen, wird wiederum *Leave One out Kreuzvalidierung* verwendet. Dabei wird die *SVM* mit allen Kontaktresten des Trainingsdatensatzes und ebenso vielen Positionen an der restlichen Oberfläche trainiert. Zum Testen werden wiederum alle Negativbeispiele der restlichen Oberfläche benutzt, als Positivbeispiele jedoch nur diejenigen, die sich im Zentralbereich der Kontaktfläche befinden. Diese sind als $PIAS \geq 2$ klassifiziert. Dieser Datensatz wird im Folgenden als $Komp_{core}$ bezeichnet.



(a) (b)

Abbildung 4.19: Performanz an verschiedenen Bereichen der Kontaktfläche
Es wurden jeweils die Kennwerte für die Analyse des gesamten Datensatzes, sowie für den Kernbereich aufgetragen. In (a) sind die *ROC*-Kurven angegeben, in (b) die *PROC*-Kurven. Bei der Analyse der Kernbereiche ergab sich ein größerer *AUC*-Wert von 0,799 verglichen mit 0,766 für die Analyse der gesamten Kontaktfläche.

Die resultierenden *ROC*- und *PROC*-Kurven in Abbildung 4.19 lassen keine eindeutige Aussage zu. Zwar findet man in Abbildung 4.19(a) eine größere *AUC* für das Zentrum der Kontaktfläche, jedoch zeigt die *PROC*-Kurve in Abbildung 4.19(b) für jeden Schwellwert eine höhere *Precision* bei der Klassifikation der gesamten Kontaktfläche.

Dieser scheinbare Widerspruch lässt sich wie folgt erklären: Im Datensatz $Komp_{core}$ ist der relative Anteil der Negativbeispiele TN größer als in $Komp_{kanon}$, da sich die beiden Datensätze nur um die im Randbereich liegenden Positivbeispiele unterscheiden. Somit bleibt für jeden Schwellwert $FPR = \frac{FP}{FP+TN}$ gleich. Wie Abbildung 4.19(a) zeigt, ist jedoch $TPR = \frac{TP}{TP+FN}$ des Zentralbereichs für jede FPR und daher auch für jeden Schwellwert größer als für die gesamte Kontaktfläche. Hieraus folgt, dass ein größerer Anteil der Kernbereiche korrekt klassifiziert wird. Da die TPR für die gesamte

Kontaktfläche niedriger ist, werden auch Residuen des Randbereichs seltener korrekt klassifiziert.

Die geringere *Precision* der *PROC*-Kurve in Abbildung 4.19(b) ist bedingt durch die größere Asymmetrie in den Anteilen von Positiv- und Negativbeispielen. Da die Positivbeispiele, die am Rand der Kontaktfläche liegen, aus dem Datensatz entfernt werden, sinkt das Verhältnis von Positiv- zu Negativbeispielen im Datensatz um etwa einen Faktor 2, wie am rechten Rand von Abbildung 4.19(b) bei einem *Recall* von 1 zu erkennen ist. Aufgrund der Definition der *Precision* (3.55) folgt, dass der Wert für die *Precision* sinkt.

4.6 Gewichtete Mittelung über die Nachbarschaft

Die im letzten Abschnitt vorgestellte SVM generiert Vorhersagen für die Zugehörigkeit einzelner Oberflächenaminosäuren der monomeren Untereinheit zu einer PPK. Da reale Kontaktaminosäuren nicht einzeln vorkommen, sondern eine geschlossene Fläche bilden, liegen bei hinreichend guter Qualität der Prognose die meisten als Kontakte vorhergesagten Aminosäuren räumlich benachbart. Falsch positive Vorhersagen dagegen werden gleichmäßiger über die restliche Oberfläche verteilt und daher einzeln erwartet. In diesem Abschnitt werden bei der Bestimmung positionsspezifischer Werte der fünf betrachteten Merkmale auch die Werte der Nachbarschaft mit berücksichtigt. Auf diese Weise sollen sich die Vorhersagen benachbarter Positionen gegenseitig verstärken und so die Güte der Vorhersage verbessern.

Im Detail wird zur Berechnung des Wertes eines Merkmals für eine Position im Protein wie folgt vorgegangen: Der Wert ergibt sich nun als gewichteter Mittelwert aus den Werten der intramolekularen Nachbarschaft (siehe Abschnitt 3.11). Durch dieses Vorgehen konnte von Porollo und Meller [46] der Informationsgehalt einiger Merkmale zur Vorhersage von PPKs deutlich erhöht werden. Dabei wird eine Nachbarposition k einer Position i im Protein entweder reziprok zu ihrem Abstand d_k von Position i oder proportional zu ihrer *relativen SASA* gewichtet. Die Gewichte $w^{(e)}_{rSASA}$ bzw. $w^{(e)}_{dist}$ aus (3.45) bzw. (3.46), die den Einfluss der Nachbarschaft bei der Berechnung des resultierenden gemittelten Wertes einer Eigenschaft e bestimmen, sind als Parameter ebenso zu optimieren, wie die Entfernung $s^{(e)}$ von der zu bewertenden Position i, in der intramolekulare Nachbarpositionen bei der Mittelung berücksichtigt werden.

4.6.1 Optimierung der Parameter

In diesem Abschnitt wird für jede untersuchte positionsspezifische Eigenschaft $e \in \{pair\_intra, hpa, rSASA, cons, conn\}$ der Einfluss der Gewichte $w^{(e)}_{rSASA}$ bzw. $w^{(e)}_{dist}$ und des Abstandsschwellwerts $s^{(e)}$ auf die Performanz während der Prozedur der *Leave One out Kreuzvalidierung* untersucht. Die Performanz des Klassifikators wird dabei wieder über die *AUC* einer ROC-Kurve bestimmt. Ausgangspunkt ist der Satz an optimierten Parametern der 5 betrachteten Eigenschaften ohne Berücksichtigung der Nachbarschaft und die SVM aus Abschnitt 4.4.3.2, die eine Performanz von $AUC = 0.7670$ erreicht.

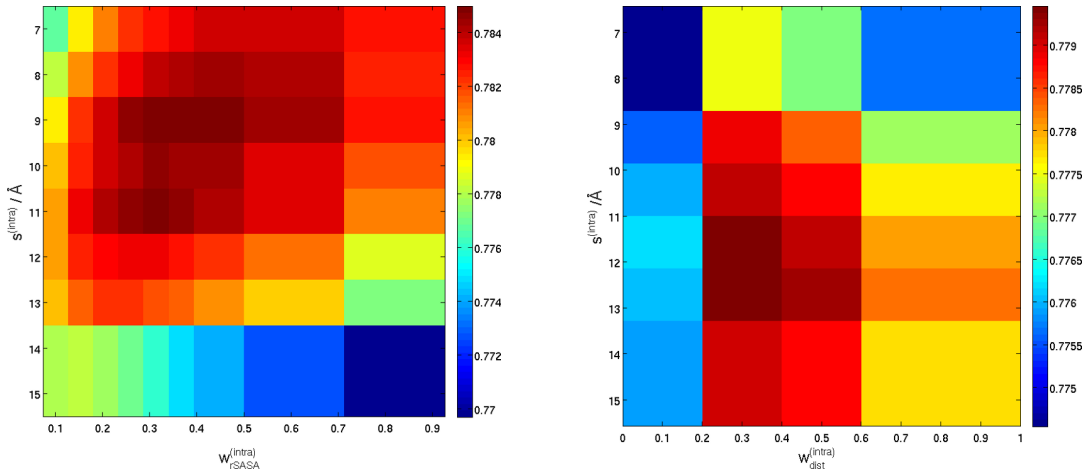
Da man davon ausgehen kann, dass die Mittelung einer einzelner Eigenschaft unabhängig von der Mittelung über andere Eigenschaften einen Gewinn an Performanz liefert, können die Parameter $w^{(e)}_{rSASA}$ bzw. $w^{(e)}_{dist}$ und $s^{(e)}$ für jedes e einzeln optimiert werden, während alle anderen Eigenschaften ohne Mittelung über die Nachbarschaft direkt verwendet werden. Im nächsten Schritt sollen dann alle Merkmale, die im vorhergehenden Schritt von der Mittelung über die Nachbarschaft profitierten, über die jeweils optimierten Parametern gleichzeitig gemittelt werden, bevor sie in der *SVM* zu einer Vorhersage verrechnet werden. Die Hoffnung dabei ist, dass sich der Gewinn an Performanz durch die Mittelung der einzelnen Eigenschaften addiert und so ein performanter Klassifikator entsteht, der von der Mittelung über mehrere Eigenschaften profitiert.

In den folgenden fünf Abschnitten werden nacheinander die je zwei, zweidimensionalen Parameterräume, die von $w^{(e)}_{rSASA}$ bzw. $w^{(e)}_{dist}$ und $s^{(e)}$ für $e \in \{pair\_intra, hpa, rSASA, cons, conn\}$ aufgespannt werden, anhand einer Grid-Suche nach optimalen Werten durchsucht.

4.6.2 Die intramolekulare Chancenquotienten $\mathbf{PW_{pair\_intra}}$

Abbildung 4.20 zeigt die Ergebnis der Grid-Suche nach optimalen Mittelungsparametern für die intramolekularen Chancenquotienten. Alle anderen Eigenschaften wurden dabei ohne Mittelung über die Nachbarschaft verwendet. Werden die Nachbarpositionen reziprok proportional zur jeweiligen Entfernung *dist* gewichtet, so ergibt sich ein Maximum bei den Werten $w^{(pair\_intra)}_{dist} = 0,3$ und $s^{(pair\_intra)} = 11$ Å mit einem Wert von 0.7795 (siehe Abbildung 4.20(b)). Da der Gewinn an Performanz bei Mittelung proportional zur jeweiligen *rSASA* jedoch deutlich höher ausfällt, wurde die rechenintensive Grid-Suche hier nicht mit höherer Auflösung durchgeführt.

Das Maximum im Parameterraum von $w_{rSASA}^{(pair\_intra)}$ und $s^{(pair\_intra)}$ dagegen wurde mit hoher Auflösung ermittelt, wie Abbildung 4.20(a) zeigt. Es befindet es sich bei $w_{rSASA}^{(pair\_intra)} = 0.35$ und $s^{(pair\_intra)} = 9$ Å und hat einen Wert von 0.7850. Die große Breite des Maximums im Parameterraum spricht für die Stabilität der Methode.



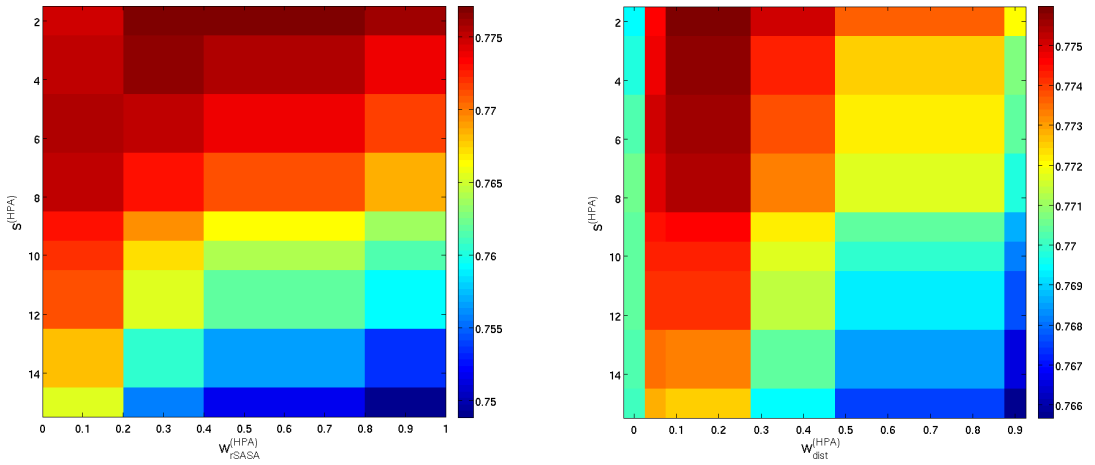
(a) Gewichtung nach *rSASA* (b) Gewichtung nach 1/dist

Abbildung 4.20: Gewichtete Mittelung von S_{pair_intra}
Der Farbplot zeigt das Ergebnis der Grid-Suche in dem durch (a) $w_{rSASA}^{(pair\_intra)}$ bzw. (b) $w_{dist}^{(pair\_intra)}$ und $s^{(pair\_intra)}$ aufgespannten Parameterraum. Die Farbtemperatur beschreibt die Performanz anhand der *AUC*-Werte und ist jeweils rechts in Form einer Skala angegeben.

4.6.3 Hydrophobe Patches

Als nächstes soll der Einfluss gewichteter Mittelung über die Nachbarschaft auf die Aussagekraft der Zugehörigkeit zu einem hydrophoben Patch untersucht werden. Wie Abbildung 4.21 zeigt, lässt sich die Klassifikationsleistung bei Mittelung über sehr nahe Nachbarn erhöhen. Die maximale *AUC* des Klassifikators erhält man bei einer Mittelung über Nachbarn im Umkreis von maximal 2 Å. Bei der Gewichtung nach der *rSASA* in Abbildung 4.21(a) wird die maximale *AUC* mit einem Gewicht von $w_{rSASA}^{(HPA)} = 0.3$ bei einem Wert der *AUC* von 0.7771 erreicht. Bei Gewichtung reziprok zur Entfernung ergibt sich in Abbildung 4.21(b) ein Maximum der *AUC* von 0,7760 bei $w_{dist}^{(HPA)} = 0.1$ mit einem *AUC*-Wert von 0.7760.

Obwohl der Unterschied zwischen den beiden Methoden der Gewichtung gering ausfällt, wird im Folgenden die Mittelung proportional zur jeweilige *rSASA* mit den oben bestimmten Parametern als Optimum benutzt.



(a) Gewichtung nach *rSASA*

(b) Gewichtung nach 1/dist

Abbildung 4.21: Gewichtete Mittelung der Zugehörigkeit zu einem hydrophoben Patch

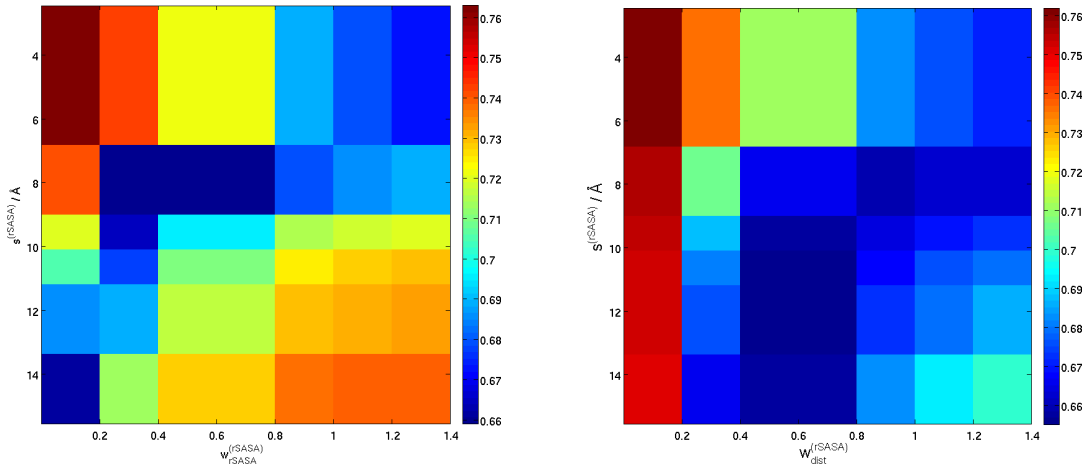
Der Farbplot zeigt das Ergebnis der Grid-Suche in dem durch (a) $w_{rSASA}^{(hpa)}$ bzw. (b) $w_{dist}^{(hpa)}$ und $s^{(hpa)}$ aufgespannten Parameterraum. Die Farbtemperatur beschreibt die Performanz anhand der *AUC*-Werte; diese sind in der Skala rechts angegeben.

4.6.4 Relative SASA

Da die relative *SASA* nicht nur als Gewicht für die Nachbarpositionen, sondern auch als Merkmal dient, wird diese Eigenschaft ebenfalls auf ihren Einfluss auf die Performanz des Klassifikators hin untersucht. Abbildung 4.22 zeigt wiederum das Ergebnis einer Grid-Suche der Mittelungsparameter. Wie man erkennt, verschlechtert eine Mittelung über die Nachbarschaft bei beiden Arten der Gewichtung die Performanz des Klassifikators. Der ursprüngliche *AUC*-Wert von 0.7650 wird für $s^{(rSASA)} \geq 0$ und $w_{rSASA}^{(rSASA)} \geq 0$ bzw. $w_{dist}^{(rSASA)} \geq 0$ stets unterboten. Folglich besitzt der Klassifikator die größte Vorhersagegenauigkeit, wenn die relative *SASA* ohne Mittelung über die Nachbarschaft direkt verwendet wird.

4.6.5 Konserviertheit

Analog wurde ein Mitteln der Konserviertheit über die Nachbarschaft analysiert. Abbildung 4.23(a) zeigt das Ergebnis der Grid-Suche nach optimalen Mittelungsparametern $w_{rSASA}^{(cons)}$ bzw. $w_{dist}^{(cons)}$ und $s^{(cons)}$ für Konserviertheit. Die Performanz gemessen an der *AUC* der *ROC*-Kurve konnte durch gewichtete Mittelung nur bei einem sehr kleinen Gewicht von $w_{rSASA}^{(cons)} = 0.1$ und einem kleinen Abstandsschwellwert von $s^{(cons)} = 3.0$ Å



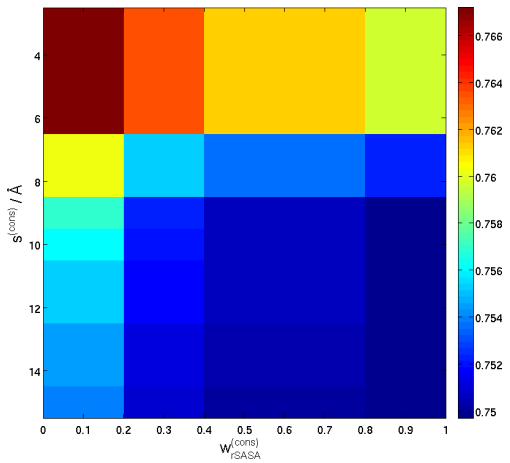
(a) Gewichtung nach *rSASA* (b) Gewichtung nach 1/dist

Abbildung 4.22: Gewichtete Mittelung der *rSASA*
Der Graph zeigt die Grid-Suche in dem durch (a) $w_{rSASA}^{(rSASA)}$ bzw. (b) $w_{dist}^{(rSASA)}$ und $s^{(rSASA)}$ aufgespannten Parameterraum. Die Farbtemperatur beschreibt die Performanz anhand der *AUC*-Werte; diese sind in der Skala rechts angegeben.

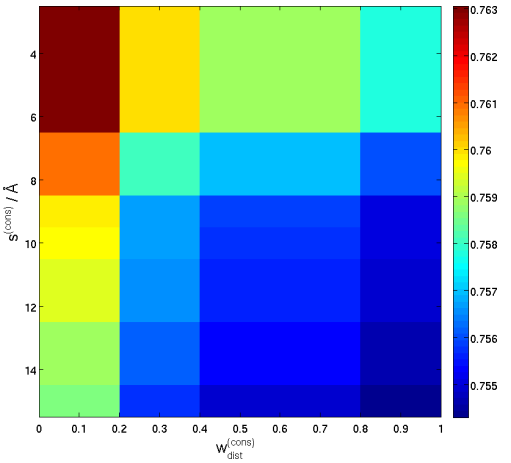
leicht verbessert werden. Es ergab sich eine Erhöhung des AUC-Wertes von 0,7670 auf 0,7672. Eine Gewichtung der Nachbarschaft reziprok proportional zu ihrer Entfernung dagegen brachte nur eine Verschlechterung der Performanz, wie man an Abbildung 4.23(b) sieht. Folglich kann die Performanz durch eine gewichtete Mittelung der Konserviertheit über die Nachbarschaft kaum verbessert werden. Aus diesem Ergebnis kann geschlossen werden, dasss in PPKs stark konservierte Positionen nicht geclustert vorkommen.

4.6.6 Konnektivität

Auch die Auswirkungen einer gewichteten Mittelung der Konnektivität auf die Performanz wurde anhand einer Grid-Suche des Parameterraumes geklärt. Wie Abbildung 4.24 zeigt konnte die Performanz sowohl bei einer Gewichtung proportional der *rSASA* als auch reziprok proportional dem Abstand d verbessert werden. Dabei fand sich der höchste Wert der *AUC* wieder bei einer Gewichtung proportional der *rSASA* mit den Parametern $w_{rSASA}^{(conn)} = 1.9$ und $s^{(conn)} = 2$ Å mit einem *AUC*-Wert von 0.7748.



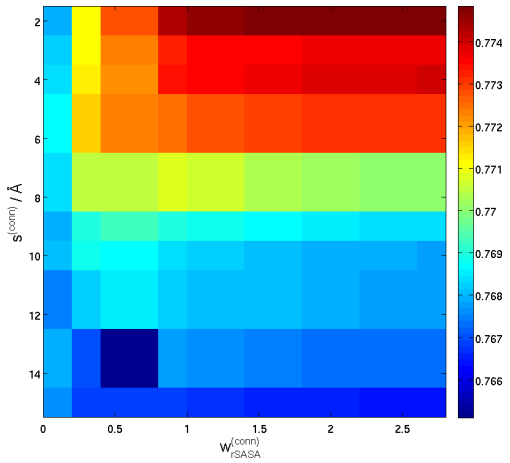
(a) Gewichtung nach *rSASA*



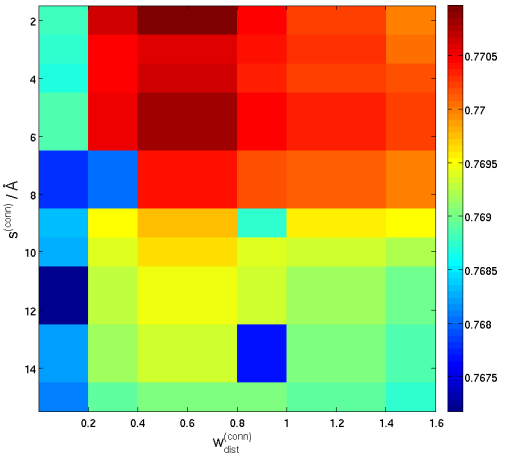
(b) Gewichtung nach 1/dist

Abbildung 4.23: Gewichtete Mittelung der Konserviertheit

Der Graph zeigt die Grid-Suche in dem durch (a) $w_{rSASA}^{(cons)}$ bzw. (b) $w_{dist}^{(cons)}$ und $s^{(cons)}$ aufgespannten Parameterraum. Die Farbtemperatur beschreibt die Performanz anhand der *AUC*-Werte; diese sind in der Skala rechts angegeben.



(a) Gewichtung nach *rSASA*



(b) Gewichtung nach 1/dist

Abbildung 4.24: Gewichtete Mittelung der Konserviertheit

Der Graph zeigt die Grid-Suche in dem durch (a) $w_{rSASA}^{(conn)}$ bzw. (b) $w_{dist}^{(conn)}$ und $s^{(conn)}$ aufgespannten Parameterraum. Die Farbtemperatur beschreibt die Performanz anhand der *AUC*-Werte; diese sind in der Skala rechts angegeben.

4.6.7 Kombination aller optimierten Parameter

In den letzten Abschnitten wurde jedes der fünf zur Klassifikation verwendeten Merkmale daraufhin untersucht, ob eine gewichtete Mittelung über die Nachbarschaft die Klassifikationsleistung erhöht. Für jede Eigenschaft wurde dabei die optimale Parameterkombination bestimmt. Diese sind in Tabelle 4.11 zusammengefasst. Eine Mittelung der *rSASA* über die Nachbarschaft brachte keine Verbesserung der Performanz. Auch eine Mittelung der Konserviertheit über die Nachbarschaft verbesserte die Performanz nur bei sehr kleinen Werten der Mittelungsparameter marginal. Dagegen konnte die Aussagekraft bei Konnektivität, der Zugehörigkeit zu einem hydrophoben Patch und vor allem bei intramolekularen Chancenquotienten in erheblichem Ausmaß durch eine Mittelung über die Nachbarschaft gesteigert werden. Bemerkenswert dabei ist, dass stets eine Gewichtung proportional der jeweiligen *rSASA* eine größere Verbesserung brachte als eine Gewichtung reziprok proportional der jeweiligen Entfernung.

Eigenschaft	**Gewichtung**	$\mathbf{w_M^{(e)}}$	$\mathbf{s^{(e)}}$ in Å
PW_{pair_intra}	rSASA	0,35	9
HPA	rSASA	0,3	2
rSASA	0	0	0
Konserviertheit	rSASA	0,1	3
Konnektivität	rSASA	1,9	2

Tabelle 4.11: Optimale Parameter der gewichteten Mittelung
Die Tabelle zeigt das Ergebnis der Grid-Suche nach optimalen Parametern für die gewichtete Mittelung über die Nachbarschaft der 5 Eigenschaften der Oberflächenaminosäuren

Mittelung	**AUC**	**ΔAUC**
keine	0.7670	0.0000
PW_{pair_intra}	0.7850	0.0180
HPA	0.7771	0.0101
rSASA	0.7670	0.0000
Konserviertheit	0.7672	0.0002
Konnektivität	0.7748	0.0078
Kombination	0.7945	0.0275

Abbildung 4.25: Einfluss der gewichteten Mittelung
Die Tabelle zeigt die Verbesserung der Performanz des Klassifikators durch gewichtete Mittelung über die Nachbarschaft einzeln für jede positionsspezifische Eigenschaft und für die kombinierte Mittelung aller Eigenschaften mit optimalen Parametern

Nun sollen alle Eigenschaften mit den jeweils optimalen Parametern über die Nachbarschaft gleichzeitig gemittelt und als Eingabedaten in den Klassifikator gegeben werden.

Es werden also die intramolekularen Chancenquotienten, die Zugehörigkeit zu einem hydrophoben Patch, die Konserviertheit und die Konnektivität jeweils proportional der *rSASA* über die Nachbarschaft mit den Parametern aus Tabelle 4.11 gemittelt, während die relative *SASA* als Merkmal ohne Mittelung benutzt wird. Damit erreicht die Vorhersagequalität der SVM die höchste Performanz, die mit den beschriebenen Methoden möglich ist. Der *AUC*-Wert ist dann $0,7945$, wie man an Tabelle 4.25 ablesen kann.

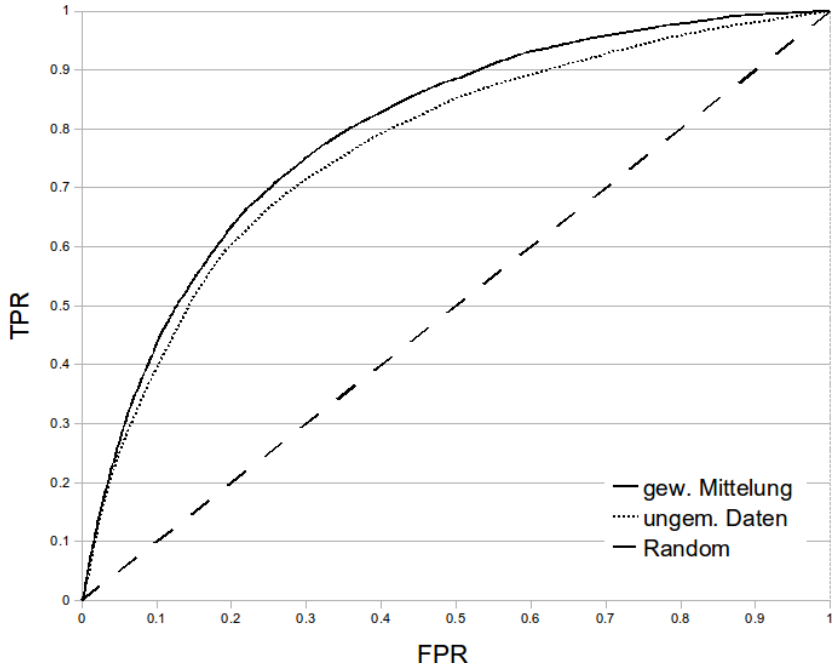


Abbildung 4.26: Einfluss der gewichteten Mittelung
Es sind die ROC-Kurven des Klassifikators angegeben, die sich bei einer Benutzung der optimalen Parameter für die gewichtete Mittelung über die Nachbarschaft (durchgezogene Linie) und aus den ungemittelten Daten (fein gestrichelte Linie) ergeben. Die gestrichelte Gerade repräsentiert einen zufälligen Klassifikator. Die *AUC*-Werte betragen 0,7945 bzw. 0,7670.

Mit der beschriebenen Strategie der *Leave One out Kreuzvalidierung* erhält man so Datensatz $Komp_{kanon}$ aus Abschnitt 4.1.3 die in Abbildung 4.26 gezeigte ROC-Kurve mit eine *AUC*-Wert von 0.7945. Da die hier behandelten Datensätze von Kontakt- und Nicht-Kontaktresten deutlich mehr Negativbeispiele als positive Fälle enthalten, ist in mancher Hinsicht eine *PROC*-Kurve, wie sie in Abbildung 4.27 gezeigt ist, aussagekräftiger. Man sieht, dass der Klassifikator eine *Precision* von 44% bei einem *Recall* von 20% ermöglicht.

Eine weitere Möglichkeit der Bewertung der Güte eines Klassifikators ist der *Matthews Korrelationskoeffizient* [145]. Er gibt an, inwieweit die vorhergesagte Klasseneinteilung

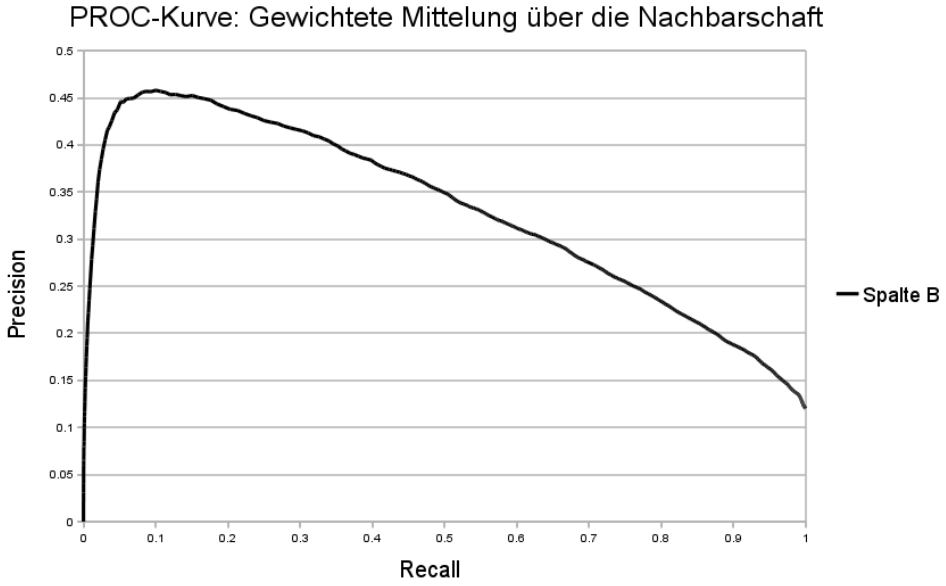


Abbildung 4.27: PROC-Kurve zur gewichteten Mittelung
Diese Abbildung zeigt die *PROC*-Kurve, die man bei gewichteter Mittelung über die Nachbarschaft mit optimierten Parametern erhält.

mit der richtigen Einteilung korreliert ist. Bei Verwendung aller optimierten Mittelungsparameter erreicht *PresCont* am Datensatz $Komp_{kanon}$ einen *MCC*-Wert von 0,323.

4.7 Nachbearbeitung der Ergebnisse

Clusterverfahren sind Methoden, um einen Satz an Datenpunkten derart in Teilmengen, sogenannte Cluster, zu unterteilen, dass die Daten innerhalb eines Clusters bezüglich eines bestimmten Abstandsmaßes eine stärkere Ähnlichkeit aufweisen. In dieser Arbeit wird *hierarchisches Clustern* dazu verwendet, *falsch positive* Vorhersagen von Kontaktaminosäuren, die einzeln an der Oberfläche der monomeren Proteinkette liegen, zu korrigieren.

4.7.1 Hierarchisches Clustern

Hierarchisches Clustern wurde in [185] zur Verminderung falscher Vorhersagen, die von mehreren richtigen Vorhersagen aus der anderen Klasse umgeben sind benutzt. Auf ähnliche Art wird hier *hierarchisches Clustern* verwendet um die Vorhersage der SVM

zu verbessern. In Abschnitt 4.4 wurde für alle Oberflächenaminosäuren eine Wahrscheinlichkeit für die Zugehörigkeit zu einer PPK berechnet. Anschließend wird ein Schwellwert festgelegt, der darüber entscheidet, welche Reste sich laut Prognose an einer Kontaktfläche befinden. Die so vorhergesagten Kontaktreste werden anschließend nach dem euklidischen Abstand ihrer C_α-Atome hierarchisch geclustert. Dabei wird ab einem Abstandsschwellwert von s_{hc} abgebrochen. Da Kontaktflächen unterschiedliche Größe und Form besitzen, wird der Abstand der Cluster voneinander nach dem *single linkage* Verfahren bestimmt.

Nach diesem ersten Schritt werden alle Positionen, die keinem Cluster mit einer Mindestgröße von N_{min} angehören, verworfen. Dadurch soll die Anzahl der *falsch positiven* Vorhersagen, die einzeln an der Oberfläche liegen, vermindert werden. Abschließend wird die Prognose für alle bisher als negativ vorhergesagten Positionen, die zu mehr als N_n positiv prognostizierten Positionen benachbart sind, geändert. Zur Definition dieser Nachbarschaft wird ein weiterer Abstandsschwellwert s_n verwendet. Mit diesem Schritt soll die Anzahl der *falsch negativen* Vorhersagen, die von mehreren *wahr positiven* Beispielen umgeben sind, vermindert werden.

Im letzten Schritt wird anschließend das *single linkage clustering* mit dem Abstandsschwellwert s_{hc} wiederholt und Cluster mit einer geringeren Größe als N_{min} verworfen.

4.7.2 Optimierung der Parameter

Für die beschriebene Clusterprozedur sind mehrere freie Parameter zu optimieren. Dazu zählen die Abstandsschwellwerte s_{hc} und s_n ebenso wie die Mindestgröße eines Clusters N_{min} und die Anzahl an positiv vorhergesagten Nachbarn N_n, die im letzten Schritt für einen Wechsel der Vorhersage von negativ nach positiv nötig ist. Diese Optimierung geschieht anhand einer *Grid*-Suche mit linearer Variation der Parameter im 4-dimensionalen Parameterraum.

$\mathbf{s_{hc}}$	$\mathbf{s_n}$	$\mathbf{N_{min}}$	$\mathbf{N_n}$
5	4	5	7

Tabelle 4.12: Parameter der Nachbearbeitung der Vorhersage ohne Verwendung der gewichtete Mittelung

Ausgangspunkt der Grid-Suche ist dabei die Vorhersage der *SVM* mit den optimalen Parametern aus Abschnitt 4.4 am Datensatz $Komp_{kanon}$ aus Abschnitt 4.1.3 ohne ge-

wichtete Mittelung über die Nachbarschaft. Der gewählte Satz an Parametern für die Clusterprozedur ist in Tabelle 4.12 gelistet. Als Optimierungskriterium wurde wieder die *AUC* der *ROC*-Kurve benutzt.

Bei der Aufnahme der ROC-Kurve wurde dabei folgendermaßen vorgegangen: Aus der Vorhersage der *SVM* wurde anhand eines variablen Schwellwertes eine primäre Vorhersage generiert. Diese wurde anschließend für einen bestimmten Satz an Cluster-Parametern nach dem im letzten Abschnitt beschriebenen Verfahren zu einer sekundären Vorhersage verrechnet. Anhand dieser sekundären Vorhersage können dann die Raten der *wahr positiven* und der *falsch positiven* Vorhersagen bestimmt werden. Unter Variation des Schwellwertes für die primäre Vorhersage, die sowohl Auswirkungen auf die primäre als auch auf die sekundäre Vorhersage hat, wird anschließend die *ROC*-Kurve der sekundären Vorhersage aufgenommen.

So konnte die primäre Performanz der SVM aus Abschnitt 4.4 mit einem *AUC*-Wert von 0.766 über die Nachverarbeitung durch hierarchisches Clustern mit den Parametern aus Tabelle 4.12 auf einen *AUC*-Wert von 0.773 gesteigert werden.

Weiterhin wird untersucht, in welchem Umfang die Performanz der *SVM* bei gewichteter Mittelung über die Nachbarschaft aus Abschnitt 4.6 durch hierarchisches Clustern erhöht werden kann. Die von diesem Klassifikator vorhergesagten Kontaktflächen weisen eine deutlich andere Form auf als die Kontaktfläche, die ohne gewichtete Mittelung über die Nachbarschaft vorhergesagt werden. Durch die gewichtete Mittelung über die Nachbarschaft treten weit weniger vereinzelte falsch positive und falsch negative Beispiele auf so dass sowohl die Kontaktfläche als auch die Nicht-Kontaktfläche geschlossenere Form mit wenigen einzelnen Löchern besitzen. Deshalb ist auch davon auszugehen, dass die Cluster-Parameter für diesen Klassifikator anders zu wählen sind. Daher wurde die Grid-Suche nach den Clusterparametern wiederholt, wobei sich die Werte in Tabelle 4.13 ergeben. Mit diesem Satz an Parametern konnte die Performanz allerdings lediglich marginal auf $AUC = 0,7958$ gesteigert werden.

s_{hc}	s_n	N_{min}	N_n
7	4	4	10

Tabelle 4.13: Parameter zur Nachverarbeitung der Vorhersage unter Verwendung der gewichteten Mittelung.

Dieses Ergebnis lässt sich mit einer gewissen Ähnlichkeit der beiden Strategien erklären. Da sowohl die gewichtete Mittelung über die Nachbarschaft als auch die Nachbearbeitung der Ergebnisse über hierarchisches Clustern eine Glättung der Eigenschaften bzw.

Vorhersage bewirkt, besitzen beide Verfahren eine gewisse Redundanz und können sich kaum gegenseitig verstärken. Daher ist es nicht verwunderlich, dass eine Nachverarbeitung der SVM-Ergebnisse, die aus gewichtet gemittelten Input-Daten resultieren, über hierarchisches Clustern die Performanz nicht mehr steigern kann.

Vergleicht man gewichtete Mittelung über die Nachbarschaft mit dem Verfahren des hierarchischen Clusterns anhand des Performanzgewinns, so stellt man fest, dass gewichtete Mittelung über die Nachbarschaft deutlich mehr Verbesserung bringt (*AUC*-Wert von $0,7945$) als hierarchisches Clustern (*AUC*-Wert von $0,773$). Da die Kombination beider Verfahren nur eine unwesentliche Verbesserung erzielt, ist es sinnvoll, sich auf die gewichtete Mittelung über die Nachbarschaft zu beschränken.

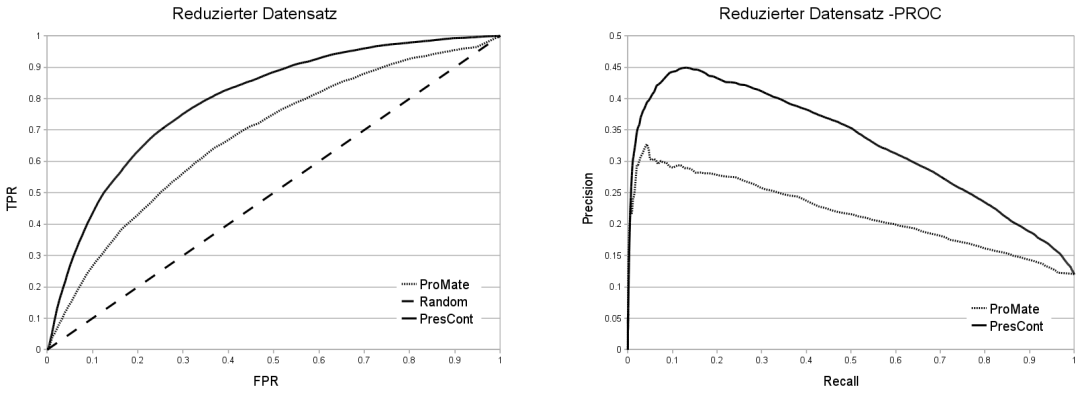
4.8 Vergleich mit anderen Verfahren

Um die Qualität der Vorhersage von *PresCont* beurteilen zu können, ist ein Vergleich mit etablierten Verfahren zur Vorhersage von Protein-Protein Kontaktflächen nötig. Um ein breiteres Spektrum an vorhandener Software abzudecken, wurde *ProMate* [47] als ein Programm herangezogen, das für die Vorhersage transienter PPKs entwickelt wurde. Daneben wird *PresCont* mit *Sppider* [46] verglichen, das sich an obligaten PPKs als sehr performant erwiesen hat. Beim Vergleich wird *PresCont* in der Form von Abschnitt 4.6.7 unter Verwendung gewichteter Mittelung über die Nachbarschaft ohne Nachbearbeitung der Ergebnisse durch *hierarchisches Clustern* über *Leave One Out Kreuzvalidierung* bewertet um *Overlearning* ausschließen zu können. Es wird die Performanz an verschiedenen Arten von Protein-Protein Kontaktflächen anhand der *AUC* einer *ROC*-Kurve, der *PROC*-Kurve und anhand des *MCC* miteinander verglichen und es werden die Stärken und Schwächen der einzelnen Methoden dargestellt.

In der Literatur angegebene Größen, mit der die Performanz einer Software zur Vorhersage von Protein-Protein Kontaktflächen bewertet wird, können höchst selten direkt zu einem fairen Vergleich genutzt werden. Einerseits hängt die Performanz stark vom verwendeten Datensatz, insbesondere vom Typ der Komplexe ab. Andererseits hat die Definition von Kontaktfläche und Oberfläche einen wesentlichen Einfluss auf Größen zur Bestimmung der Güte eines Klassifikators. Insbesondere hängt die Genauigkeit (*Accuracy*) stark vom Verhältnis der Anzahl positiver und negativer Beispiele ab. Aus diesem Grund werden die genannten Programme anhand derselben Datensätze von Protein-Protein Komplexen miteinander verglichen, wobei Kontaktfläche und Oberfläche der Proteine jeweils über die bisher benutzten Kriterien definiert werden.

4.8.1 ProMate

Das Programm *ProMate* wurde 2003 von der Gruppe um *G. Schreiber* zur Vorhersage von Protein-Protein Kontaktflächen entwickelt [47]. Es generiert zunächst an der Oberfläche der Monomerstruktur ein Punktgitter mit einer Dichte von 1 Å und berechnet für jeden Punkt mehrere Eigenschaften der Umgebung im Umkreis von 10 Å. Die Werte der verschiedenen Eigenschaften werden anschließend für jeden Punkt zu einem Score PM mit $0 \leq PM \leq 1$ kombiniert, der die Tendenz der Vorhersage angibt. Diese Scores werden anschließend über die Nachbarschaft gemittelt.



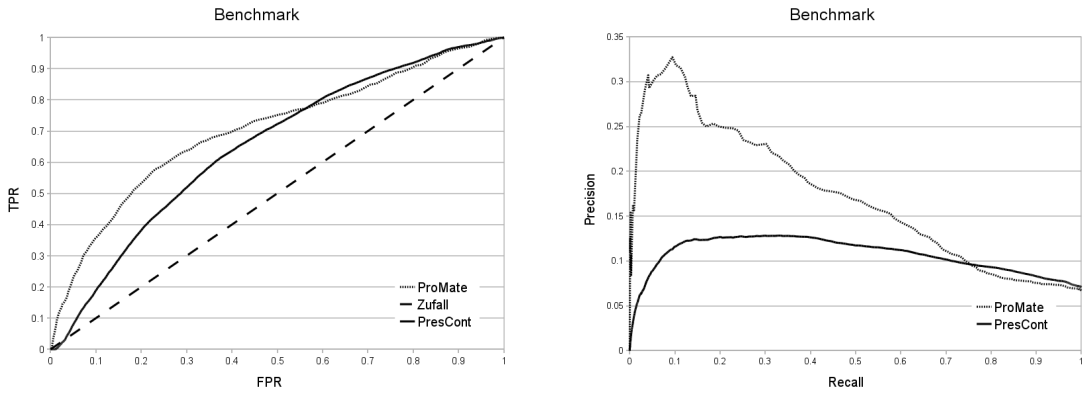
(a) (b)

Abbildung 4.28: ROC- und PROC-Kurven von ProMate und PresCont aufgenommen Datensatz $Komp_{kanon}$.
Unter Verwendung des Datensatzes $Komp_{kanon}$ wurde für *ProMate* und *PresCont* *ROC*- und *PROC*-Kurven aufgenommen. Es ergaben sich die folgenden *AUC*-Werte: 0.6779 (*ProMate*), 0,7943 (*PresCont*). In (a) sind die beiden *ROC*-Kurven dargestellt. (b) zeigt die *PROC*-Kurven. Die Kurve für *PresCont* liegt höher als die von *ProMate*.

Über eine Variation des Schwellwertes für den kombinierten und gemittelten Score lässt sich wiederum eine ROC-Kurve aufnehmen, die die Güte des Klassifikators beschreibt. Eine Bewertung anhand des Datensatzes $Komp_{kanon}$ ergibt die *ROC*-Kurve in Abbildung 4.28(a) mit einem *AUC*-Wert von $0,6779$. Damit übertrifft PresCont mit einem *AUC*-Wert von $0,7943$ die Performanz von ProMate anhand des Datensatzes $Komp_{kanon}$ erheblich. Die zugehörige *PROC*-Kurve in Abbildung 4.28(b) zeigt eine ähnliche Situation. Auch der Vergleich beider Verfahren anhand des *MCC* belegt die höhere Klassifikationsleistung von *PresCont*. Während der *MCC* von *ProMate* 0,1828 beträgt, erhält man für *PresCont* 0,323. Bei der Klassifikation des Datensatzes $Komp_{kanon}$ besitzt folglich *PresCont* eine weitaus höhere Performanz unabhängig davon, welches Verfahren zur Bestimmung der Güte eines Klassifikators dabei benutzt wird. Ein Grund für das schlechte Abschneiden von *ProMate* liegt darin, dass diese Software ursprünglich zur Vorhersage der Kontaktflächen transienter Heterodimere entwickelt wurde.

Als nächstes wird ein Vergleich der beiden Verfahren anhand des Datensatzes $Komp_{trans}$ angestellt, der aus transienten Komplexen besteht. Im Falle von *PresCont* stellte sich dabei das Problem, dass viele Sequenzen relativ geringe Länge besitzen. Daher wurden in der Sequenzdatenbank in den meisten Fällen zu wenige signifikante Treffer gefunden um paarweise geordnete MSAs mit hinreichend vielen Sequenzen generieren zu können. Daher wurde die Auswertung des Datensatzes $Komp_{trans}$ durch *PresCont* ohne Berücksichtigung der Konnektivität durchgeführt.

Wie man in Abbildung 4.29(a) erkennt, beträgt der *AUC*-Wert von *ProMate* $0,7032$ verglichen mit demjenigen von *PresCont* von 0.6510. Auch die zugehörige *PROC*-Kurve in Abbildung 4.29(b) zeigt ein ähnliches Ergebnis. Vergleicht man die beiden Werte für den *MCC*, so steht *ProMate* mit 0,203 einem *MCC*-Wert *PresConts* von 0,137 gegenüber. An diesem Datensatz aus transienten Heterodimeren übertrifft *ProMate* die Performanz von *PresCont* deutlich.



(a) (b)

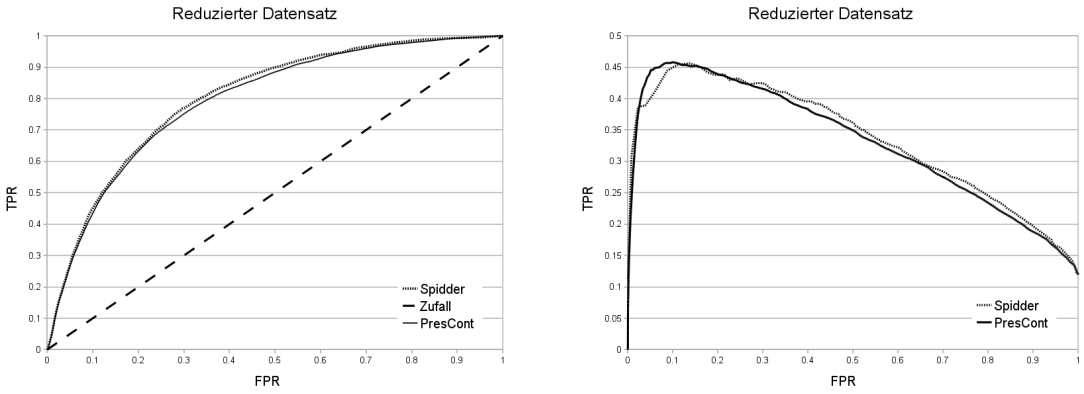
Abbildung 4.29: ROC- und PROC-Kurven von ProMate und PresCont aufgenommen am Datensatz $Komp_{trans}$.
Unter Verwendung des Datensatzes $Komp_{trans}$ wurden für *ProMate* und *PresCont* *ROC*- und *PROC*-Kurven aufgenommen. Es ergaben sich die folgenden *AUC*-Werte: ProMate: 0,7032, PresCont: 0,6510. In (a) sind die beiden *ROC*-Kurven dargestellt. (b) zeigt die *PROC*-Kurven.

4.8.2 Sppider

Porollo und Meller entwickelten die Software *Sppider* zur Vorhersage von Protein-Protein Kontaktflächen [46]. Dabei verwendeten sie eine Vielzahl positionsspezifischer Eigenschaften, insbesondere *rSASA* und den Unterschied zwischen der an der Struktur gemessenen *rSASA* und einem, durch die Software *Sable* sequenzbasiert vorhergesagten Wert. Diese Eigenschaften wurden, ähnlich wie in dieser Arbeit, einer gewichteten

Mittelung über die Nachbarschaft unterzogen und anschließend über *neuronale Netze*, *Diskriminanzanalyse* und *SVMs* zu einer Vorhersage kombiniert. Diese besteht für jede Oberflächenaminosäure aus einem Wert SP mit $0 \leq SP \leq 1$, der die Tendenz der Vorhersage angibt.

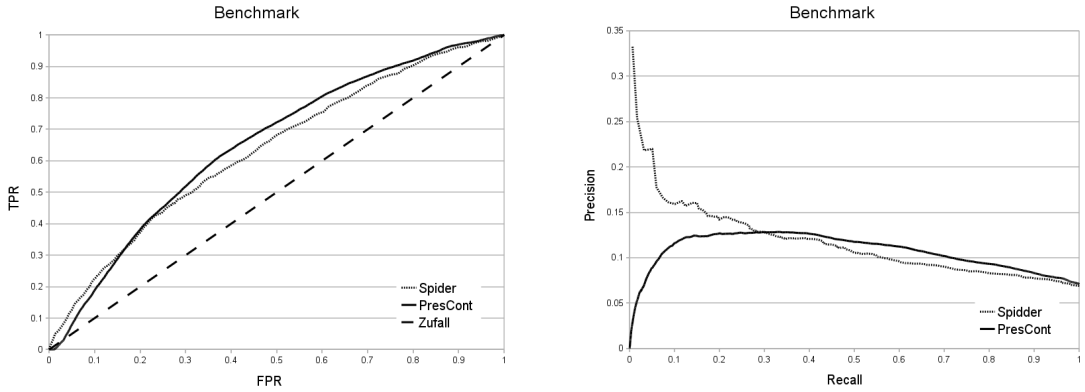
Über einen variablen Schwellwert lässt sich so wiederum für die Vorhersage an einem Datensatz bekannter Protein-Protein Komplexe eine ROC-Kurve und das zugehörige Maß der AUC ermitteln. Eine Evaluation anhand des Datensatzes $Komp_{kanon}$ ergibt die ROC-Kurve in Abbildung 4.30(a) mit $AUC^{rD}_{Sppider} = 0,8034$, die im selben Bereich liegt wie diejenige von PresCont mit $AUC^{rD}_{PresCont} = 0,7943$. Vergleicht man in diesem Fall die MCC beider Methoden miteinander, so stehen sich $MCC^{rD}_{Sppider} = 0,334$ und $MCC^{rD}_{PresCont} = 0,323$ gegenüber. Dies deutet ebenfalls auf eine sehr ähnliche Güte der beiden Klassifikatoren hin.



(a) (b)

Abbildung 4.30: ROC- und PROC-Kurven von Sppider und PresCont aufgenommen am Datendatz $Komp_{trans}$.
Diese Abbildung zeigt die *ROC*- bzw. *PROC*-Kurven, die die Performanz von Sppider und PresCont anhand des Datensatzes $Komp_{trans}$ charakterisieren. (a) *ROC*-Kurven von *Sppider* und *PresCont*. Die *AUC* von Sppider beträgt 0,8034, diejenige von *PresCont* 0,7943. (b) Die *PROC*-Kurven von *Sppider* und *PresCont*.

Um zu überprüfen, welchen Einfluss der Datensatz auf die Klassifikationsleistung hat, wurden die beiden Methoden anschließend anhand des Datensatzes $Komp_{trans}$ miteinander verglichen. Auch hier wurde *PresCont* wiederum ohne Berücksichtigung der Konnektivität ausgewertet. Wie Abbildung 4.31(a) zeigt, übertrifft die an *PresCont* gemessene *AUC* von 0,658 diejenige von *Sppider*, die 0,631 beträgt. Die *PROC*-Kurven in Abbildung 4.31(b) zeigen, dass *PresCont* im Bereich $Recall > 0,25$ Sppider leicht überlegen ist. Nur bei sehr kleinen Werten von $Recall < 0,15$ übertrifft die *Precision* von *Sppider* diejenige von *PresCont*.



(a) (b)

Abbildung 4.31: ROC- und PROC-Kurven von Sppider und PresCont aufgenommen am Datensatz $Komp_{trans}$.
(a) Vergleich der ROC-Kurven von *Sppider* und *PresCont*, die anhand des Datensatzes $Komp_{trans}$ transienter Homodimere aufgenommen wurden. Die *AUC* von *PresCont* beträgt 0,658, während *Sppider* 0,631 erreicht. (b) *PROC*-Kurven zur Bewertung der Performanz von Sppider und PresCont.

Ein Vergleich über den *MCC* ergibt für *Sppider* 0,114, während *PresCont* 0,137 erreicht. Ebenso wie anhand der *AUC* übertrifft *PresCont* die Performanz von *Sppider* auch anhand des *MCC*. Der *MCC*, der in der Publikation von Sppider [46] angegeben wurde von 0,4 konnte an keinem der hier verwendeten Datensätze erreicht werden.

Tabelle 4.14 fasst die Ergebnisse des Qualitätsvergleichs der in diesem Abschnitt getesteten Verfahren zur Vorhersage von Protein-Protein-Kontaktflächen zusammen. Während am Datensatz $Komp_{trans}$ aus obligaten Homodimeren *Sppider* und *PresCont* etwa gleiche Klassifikationsleistung zeigen und *ProMate* deutlich schlechter abschneidet, gewinnt *ProMate* den Vergleich anhand der transienten Heterodimere im Datensatz $Komp_{trans}$.

4.9 Sensitivität gegenüber Overlearning

Bei den bisherigen Untersuchungen zur Qualität der Vorhersagen von *PresCont* konnte aufgrund der Strategie der *Leave One out Kreuzvalidierung* der unerwünschte Effekt des *Overlearnings* ausgeschlossen werden. Um zu testen, inwieweit das Programm *PresCont* während des Trainings spezifische Merkmale einzelner Seitenketten anstatt genereller Merkmale von Protein-Protein Kontaktflächen lernt, wird in diesem Abschnitt mit dem gesamten Datensatz $Komp_{kanon}$ trainiert und dieser anschließend getestet. Dabei wer-

(a) Datensatz $Komp_{kanon}$

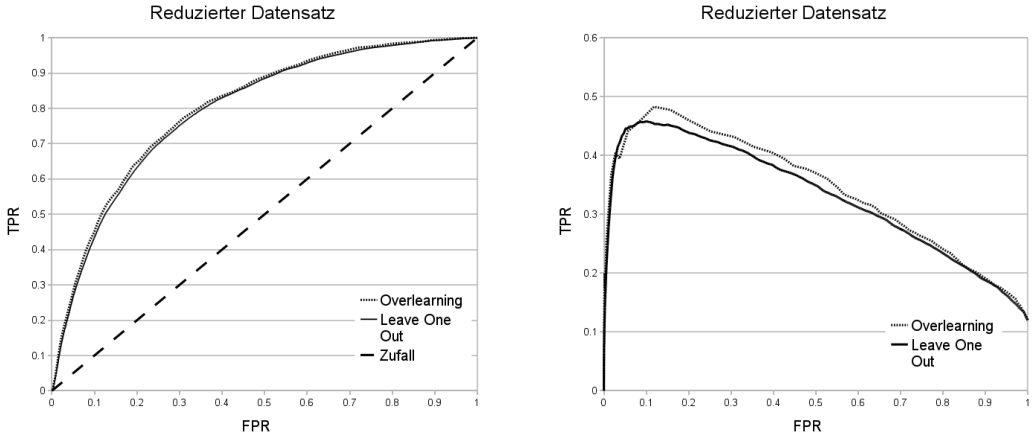
	PresCont	**ProMate**	**Sppider**
AUC	0,7943	0,6779	0,8034
MCC	0,323	0,1828	0,334

(b) Datensatz $Komp_{trans}$

	PresCont	**ProMate**	**Sppider**
AUC	0,6510	0,7032	0,631
MCC	0,137	0,203	0,114

Tabelle 4.14: Vergleich von *PresCont*, *ProMate* und *Sppider*.
(a) Die *AUC*- und *MCC*-Werte von *PresCont* und *Sppider* am Datensatz $Komp_{kanon}$
(b) Die *AUC*- und *MCC*-Werte von *PresCont* und *Sppider* am Datensatz $Komp_{trans}$.

den sämtliche in den letzten Abschnitten optimierten Parameter inklusive derjenigen der gewichteten Mittelung über die Nachbarschaft benutzt.



(a) (b)

Abbildung 4.32: Vergleich von *Leave One out Kreuzvalidierung* und *Overlearning*.
Diese Abbildungen vergleichen die Güte der Klassifikation anhand des Datensatzes $Komp_{kanon}$ unter Verwendung der gewichteten Mittelung über die Nachbarschaft mit und ohne die Strategie des *Leave One out Kreuzvalidierung*. (a) die *ROC*-Kurven mit 0,7945 (Kreuzvalidierung) 0,8021 (Overlearning) (b) *PROC*-Kurven mit und ohne Verwendung von *Leave One out Kreuzvalidierung*

Abbildung 4.32 vergleicht die *ROC*- und *PROC*-Kurven der resultierenden Klassifikation mit denjenigen, die im letzten Abschnitt unter Benutzung der *Leave One out Kreuzvalidierung* aufgenommen wurde. Anhand dieses Vergleichs erkennt man, dass ohne *Leave One out Kreuzvalidierung* nur eine marginal höhere Performanz erreicht wird. Die leicht geringere Güte der Klassifikation kann einerseits durch *Overlearning*

verursacht sein, andererseits aber auch durch die etwas geringere Größe des Traningsdatensatzes während der *Leave One out Kreuzvalidierung*-Prozedur. Folglich ist *PresCont* wenig empfindlich gegenüber dem Phänomen des *Overlearnings* und generalisiert gut. Dieser Befund ist einerseits der Tatsache geschuldet, dass *SVMs* allgemein robust gegenüber *Overlearning* sind. Andererseits jedoch besitzt *PresCont* aufgrund seines einfachen Aufbaus nur wenige freie Parameter, die auf einen Trainingsdatensatz abzustimmen sind. Dadurch, dass lediglich fünf Merkmale, die jeweils hohen und nichtredundanten Informationsgehalt besitzen, zur Charakterisierung von PPKs benutzt werden, steigt die Robustheit des Verfahrens gegenüber unerwünschten Einflüssen verschiedenster Art.

Im vorangegangenen Abschnitt wurde die Performanz von *PresCont* anhand des Datensatzes $Komp_{kanon}$ mit *ProMate* und *Sppider* verglichen. Da diese Methoden, ebenfalls auf überwachten Lernverfahren basieren, wurde anhand eines Datensatzes trainiert, bei dem nicht sichergestellt ist, dass er keine Redundanzen zum Datensatz $Komp_{kanon}$ oder zum Datensatz $Komp_{trans}$ aufweist. Daher ist auch im Vergleich des vorhergehenden Abschnittes *Overlearning* zwar im Falle von *PresCont* aufgrund der *Leave One out* ausgeschlossen, nicht jedoch im Falle von *Sppider* und *ProMate*. Vergleicht man den unter Inkaufnahme von *Overlearning* bestimmten Wert der AUC mit demjenigen von *Sppider* anhand des Datensatzes $Komp_{kanon}$, so ist der Unterschied in der AUC auf $\Delta AUC = 0,0013$ geschmolzen, was kaum den Rahmen der Messgenauigkeit übersteigt. Allerdings ist dieser Wert mit Vorsicht zu betrachten, da die Möglichkeit von *Overlearning* nicht ausgeschlossen wurde.

4.10 Beispiele

In den vorangegangenen Abschnitten wurde das Programm *PresCont* zur Vorhersage von Protein-Protein Kontaktflächen vorgestellt und seine Performanz validiert. Nun soll für einige Beispiele interagierender Proteine die Kontaktfläche vorhergesagt werde. Dabei wird die Vorhersage im Detail diskutiert und auf Stärken und Schwächen des Programms eingegangen.

4.10.1 *HisF-HisH*

Der HisF-HisH-Komplex ist ein enzymatisches Heterodimer bestehend aus der Glutaminase *HisH*, die durch eine Hydrolyse des Glutamins Ammoniak erzeugt. HisF setzt PRFAR mit dem durch HisH erzeugten Ammoniak zu ImGP und AICAR um [138]. Die Komplexbildung von HisF und HisH ist folglich Voraussetzung für die kontrollierte emzymatische Aktivität des Emzymkomplexes in der Zelle.

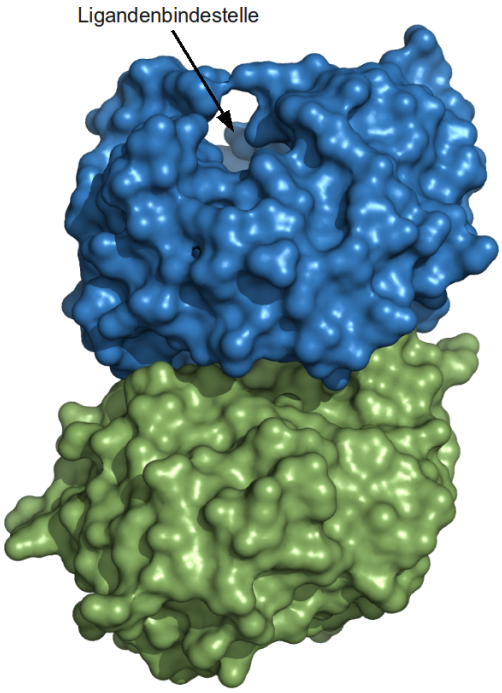


Abbildung 4.33: HisF-HisH aus *Thermothoga maritima*
Die Abbildung zeigt das enzymatische Heterodimer aus HisF (blau) und HisH (grün). An der Oberseite von HisF ist die Ligandenbindestelle zu erkennen.

In der *PDB*-Datenbank finden sich sowohl die Strukturen der einzelnen Monomere (PDB-ID: 1THF Kette D, 1K9V Kette F) als auch die des gebundenen Komplexes (PDB-ID 1GPW Ketten AB). Um einen möglichst fairen Test zu gewährleisten, sollen die Kontaktflächen von *HisF* und *HisH* anhand der Strukturen der Monomere vorhergesagt werden. Die Sequenzinformation wurde in Form von MSAs der *HSSP*-Datenbank entnommen. In diesem Fall war es möglich, genügend Spezies zu finden, von denen sowohl Sequenzen für *HisF* als auch für *HisH* vorhanden sind, was Voraussetzung zur Berechnung intermolekularer korrelierter Mutationen ist. Zur Vorhersage der Kontaktfläche wurde die SVM benutzt, die am Datensatz $Komp_{kanon}$ trainiert wurde.

Abbildung 4.34 zeigt das Ergebnis der Kontaktflächenvorhersage durch *PresCont*. Je tiefer der Rotton umso stärker wird die Position als an der PPK liegend vorhergesagt. Man erkennt in beiden Fällen, dass die Kontaktflächen in (a) und (c) deutlich stärker rot eingefärbt sind als die Nichtkontaktfläche. Vor allem im Falle von HisH erhält man einen starken Kontrast zwischen Kontaktfläche und Nichtkontaktfläche. Für HisF finden sich ebenfalls deutlich mehr und stärkere Signale an der Kontaktfläche als an der restlichen Oberfläche, auch wenn der Unterschied hier weniger deutlich ist. Der Grund für das vorhandene schwache Signal an der Nicht-Kontaktfläche von HisF liegt ist vermutlich die Ligandenbindestelle [138], die einen ähnlich hydrophoben Charakter besitzt wie Protein-Protein Kontaktflächen.

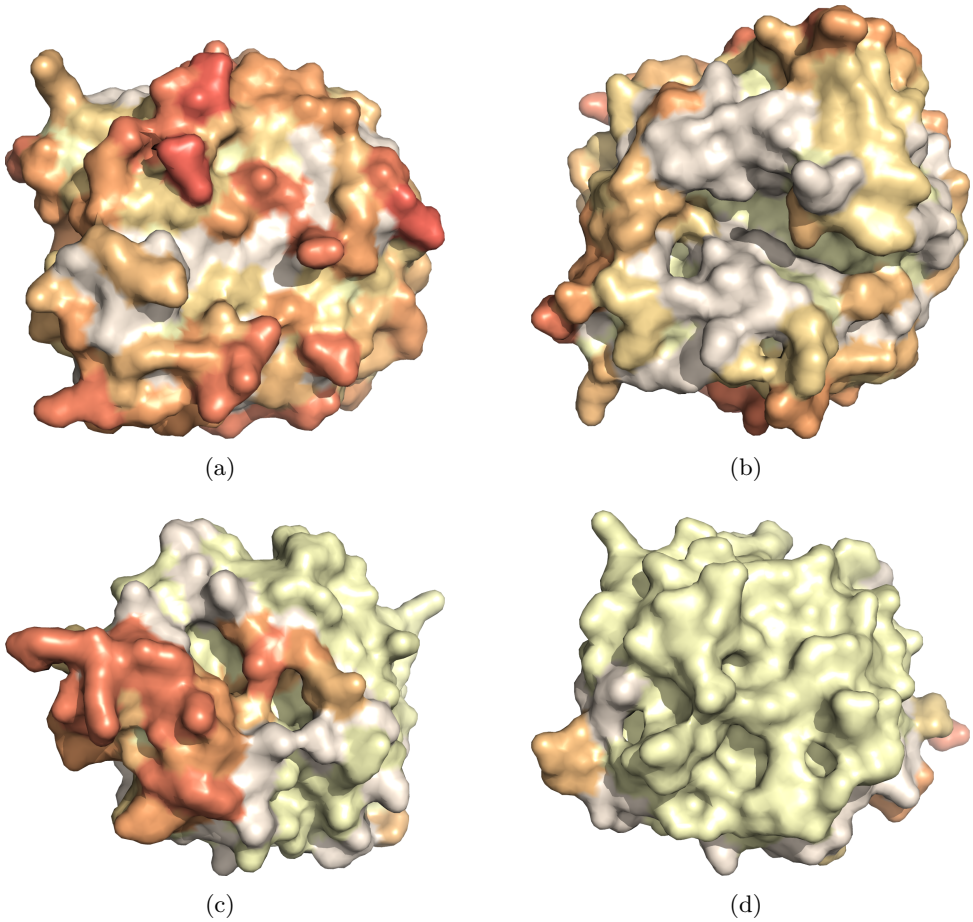


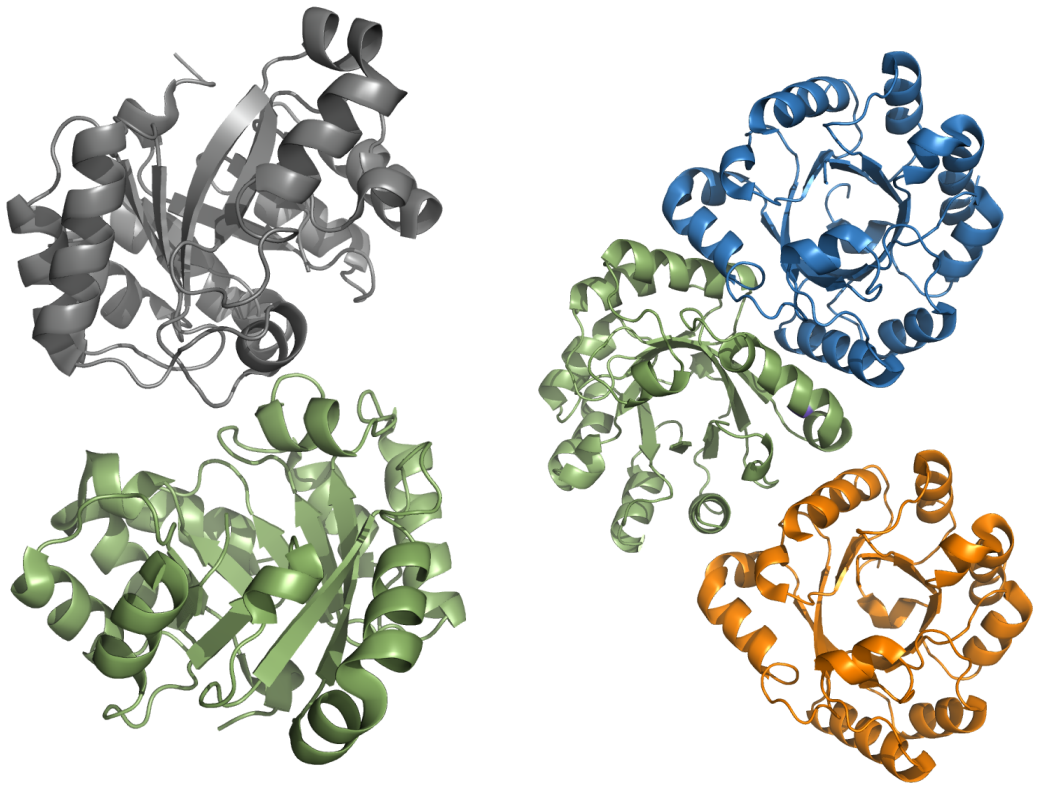
Abbildung 4.34: Vorhersage der Kontaktfläche am Komplex HisF-HisH. Die Abbildung zeigt anhand der Einfärbung der Oberfläche die Vorhersage der Kontaktfläche durch *PresCont*. Je intensiver der rote Farbton umso wahrscheinlicher liegt die entsprechende Position an der Kontaktfläche. (a) Die Kontaktfläche von HisF. (b) Die Gegenseite von HisF mit der Ligandenbindestelle. (c) Die Kontaktfläche von HisH. (d) Die der Kontaktfläche gegenüberliegende Seite von HisH. Gezeigt sind die Strukturen der Monomere mit PDB-IDs 1THF Kette D und 1K9V Kette F.

4.10.2 *PcrB* aus *B. subtilis*

PcrB aus *Bacillus subtilis* ist ein orthologes Protein zur Geranylgeranylglyceryl Phosphate Synthase (GGGPS) aus *Archaeoglobus fulgidus (Af)*. Das Homodimer *Af*GGGPS katalysiert die Kondensation von sn-Glycerin-1-Phosphat (G1P) mit Geranylgeranylpyrophosphat (GGPP) unter Abspaltung von Pyrophosphat [186]. Die Kristallstruktur von *PcrB*, das als Dimer vorliegt, wurde bereits 2005 im Rahmen eines Strukturgenomprojektes gelöst [187]. Aufgrund der Kristallsymmetrie existieren jedoch 3 mögliche Dimerisierungflächen. *PcrB* katalysiert dieselbe Reaktion wie *Af*GGGPS jedoch mit dem Unterschied, dass bevorzugt Geranylketten mit einer Länge von 35 anstatt von 20 umgesetzt werden, wie bei *Af*GGGPs. Für C_{20}-Ketten existiert lediglich *in vitro* wie *in vivo* eine promiskuitive Aktivität [188].

In der PDB-Datenbank ist die Komplexstruktur aus Abbildung 4.35(a) abgelegt, deren Kontaktfläche im Folgenden als PPK 1 bezeichnet wird. Aufgrund der sehr geringen und wenig komplementären Kontakte an dieser PPK erscheint es jedoch unwahrscheinlich, dass diese PPK *in vivo* als Dimerisierungsfläche fungiert. Im Proteinkristall finden sich zwei weitere Kontaktflächen des Dimers, die im Folgenden mit den Nummern 2 und 3 bezeichnet werden. Sie wurden durch Expansion der Kristallsymmetrie aus der PDB-Struktur 1VIZ *in silico* erzeugt und sind in Abbildung 4.35(b) dargestellt. Man erkennt, dass in beiden Fällen die Kontaktfläche deutlich größer und ihre Oberflächenkomplementarität deutlich höher ist als im Falle von PPK 1 in Abbildung 4.35(a). Überlagert man GGGPS mit PcrB, was mit einem sehr geringen RMSD von 1,1 Å möglich ist, so entspricht PPK 2 der Dimerisierungsfläche von *Af*GGGPS.

Aus biologischer Sicht hat das Aufklären der *in vivo* Kontaktfläche von *pcrB* weitreichende Relevanz. Das zu *PcrB* orthologe *Af*GGGPS besitzt an Position 99 ein Tryptophan, das aufgrund sterischer Hinderung dafür verantwortlich ist, dass Geranylketten länger als 20 C-Einheiten als Substrate gebunden und umgesetzt werden können. In *PcrB* entfällt dieser Substratlängenbegrenzer, da sich an seiner Position ein Alanin befindet (A100 in 1VIZ), das Platz für längere Substrate lässt. Für den Fall, dass Kontaktfläche Nr. 3 zwischen den grün und orange dargestellten Kette in Abbildung 4.35(b) *in vivo* auftritt, könnte eine Längenbegrenzung des Substrates bei einer Kettenlänge von 35 aufgrund sterischer Hinderung durch die jeweils andere Untereinheit erfolgen. Diese Funktion kann jedoch im Falle von PPK 1 und 2 von der grau dargestellten Untereinheit in Abbildung 4.35(a) bzw. der blau dargestellten Untereinheit in Abbildung 4.35(b) nicht erfüllt werden. Neuere moleküldynamische Untersuchungen weisen jedoch darauf hin, dass auch ohne sterische Hinderung durch die Dimerisierung C_{35}-Substrate die höchste Bindungsaffinität zeigen [188].

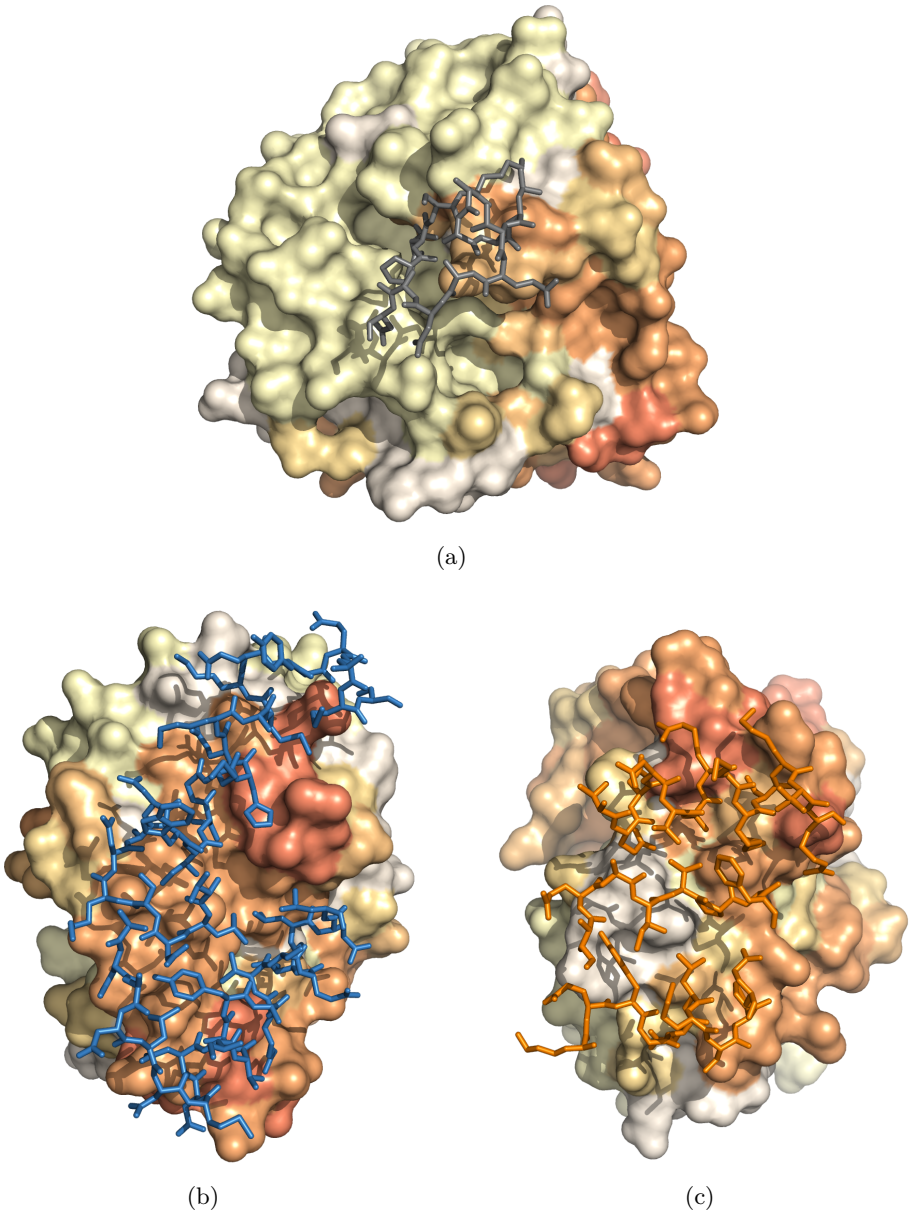


(a) (b)

Abbildung 4.35: Dimerstruktur des *PcrB* aus *Bacillus subtilis*
(a) In der PDB-Datenbank ist die abgebildete Struktur des Dimers *PcrB* abgelegt (PDB-ID 1VIZ, Ketten AB). Die beiden Ketten, die sich an der Kontaktfläche 1 berühren, sind grün und grau eingefärbt. (b) Im Proteinkristall sind auch die Kontaktflächen 2 und 3 zwischen der grün und blau bzw. der grün und orange dargestellten Kette vorhanden.

Um weitere Hinweise auf die Dimerisierungsfläche von *PcrB* zu erhalten, wurde mit Hilfe von *PresCont* die Kontaktfläche vorhergesagt. Die Vorhersage besteht für jede Aminosäure an der Oberfläche der monomeren Kette aus einem Wert zwischen 0 und 1, der die Wahrscheinlichkeit der Zugehörigkeit zu einer Protein-Protein-Kontaktfläche angibt. Abbildung 4.36 zeigt die Wahrscheinlichkeitswerte farblich codiert anhand der Proteinstruktur. Man erkennt, dass die PPK 1 die geringsten Werte unter allen drei möglichen PPKs besitzt. Dies stimmt mit dem vorher erwähnten Befund überein, dass an dieser PPK nur wenige Kontakte bei geringer Oberflächenkomplementarität existieren. Offensichtlich handelt es sich bei PPK 1, die in der PDB-Datei abgelegt ist, um einen künstlichen Kristallkontakt.

Es verbleibt die Frage, ob die Dimerisierung *in vivo* über Kontaktfläche 2 oder 3 stattfindet. Beide PPKs besitzen neben einer gewissen Oberflächenkomplementarität auch



(a)

(b)

(c)

Abbildung 4.36: Vorhersage der *in vivo* Kontaktfläche von *PcrB*
Die Oberfläche der Untereinheit in der Oberflächendarstellung ist nach dem vorhergesagten Wahrscheinlichkeitswert für die Zugehörigkeit zur Kontaktfläche eingefärbt. Je intensiver die Rotfärbung einer Position umso wahrscheinlicher ist sie Bestandteil der Kontaktfläche. Die grau, blau und orange in Stäbchendarstellung abgebildeten Aminosäuren, stellen die Kontaktaminosäuren der interagierenden Untereinheiten dar. (a) Ansicht 1: Die in der PDB-Datenbank abgelegte Dimerisierungsfläche PPK 1 (b) PPK 2 (c) PPK 3.

komplementäre Ladungsverteilungen. Anhand von Abbildung 4.36 (b) und (c) erkennt man jedoch, dass *PresCont* der Kontaktfläche 2 ein deutlich stärkeres Signal zuweist als Kontaktfläche 3. Dieses Signal spricht zwar einerseits dafür, dass Kontaktfläche 2 die

in vivo Dimerisierungsfläche ist und es sich bei Kontaktfläche 3 um einen künstlichen Kristallkontakt handelt. Andererseits zeigt jedoch auch Kontaktfläche 3 ein deutlicheres Signal als die restliche Oberfläche, so dass diese Kontaktfläche nicht komplett ausgeschlossen werden kann. In der Tat wurden *in vitro* Hinweise darauf gefunden, dass beide Kontaktflächen existieren und dass sogar ein Trimerzustand schwach populiert ist [189].

5 Diskussion

Im folgenden Abschnitt werden die Ergebnisse aus dem letzten Kapitel aufgegriffen und im Kontext bisheriger Forschung diskutiert. Dabei werden neben Merkmalen von PPKs auch Aspekte des Programmdesigns besprochen.

5.1 Der Aufbau von *PresCont*

Das Programm *PresCont* generiert durch Auswertung von fünf Merkmalen von Seitenketten an der Proteinoberfläche eine Vorhersage der Kontaktfläche. Bei der Berechnung der Scores für eine Seitenkette wird deren Nachbarschaft mit berücksichtigt. Im vorangegangenen Kapitel wurde die Qualität der Vorhersage bewertet und sichergestellt, dass jedes einzelne Merkmal einen Beitrag zur Güte der Vorhersage leistet. Im nächsten Abschnitt werden einzelne Programmteile mit den Elementen anderer Programme verglichen und es werden die Gemeinsamkeiten und Unterschiede diskutiert.

5.1.1 Relative SASA

Das Vorkommen an der Oberfläche eines Proteins ist notwendige Voraussetzung dafür, dass eine Seitenkette mit einer anderen Untereinheit interagieren kann. Je mehr eine Aminosäure an der Oberfläche eines Proteins exponiert ist, umso mehr Möglichkeiten hat sie, mit einer anderen Untereinheit zu interagieren. Daher sind Seitenketten an einer PPK im Durchschnitt stärker exponiert als an der restlichen Oberfläche [48]. Daneben hat sich dieses Signal in [46] ebenso wie in vorliegender Arbeit als starker Indikator für Protein-Protein Interaktionen erwiesen. So sinkt die Vorhersagegenauigkeit von *PresCont* ohne gewichtete Mittelung über die Nachbarschaft deutlich falls man die *rSASA* nicht berücksichtigt, was sich in einem Unterschied der *AUC*-Werte von $0,7661$ bzw. $0,6284$ ausdrückt. Damit hat die Exponiertheit einer Seitenkette an der Oberfläche mit Abstand den größten Einfluss auf die Klassifikationsleistung.

5.1.2 Intramolekulare Chancenquotienten

Neben der elektrostatischen Komplementarität [190], die für die korrekte Ausrichtung der Untereinheiten bei der Komplexbildung und die spezifische Orientierung verantwortlich ist [191], ist Hydrophobizität der bestimmende Faktor für die Stabilität von Protein-Protein Interaktionen [65]. Daher sind spezielle physikalisch-chemische Charakteristiken der Seitenketten an der Oberfläche ein starker Hinweis auf die Zugehörigkeit zu einer PPK. Um Informationen über Hydrophobizität und Elektrostatik mit einfließen zu lassen, benutzen viele Programme zur Vorhersage von PPKs mehrere physikalische oder wissensbasierte Potentiale [74] [192]. So werden in der Software *Sppider* [46] mehrere aminosäurespezifische Maße zur Bewertung der physikalisch-chemischen Eigenschaften der Datenbank AAIndex [63] entnommen und bei der Vorhersage berücksichtigt.

In dieser Arbeit dagegen wurde ein wissensbasiertes Potential für benachbarte intramolekulare Paare von Oberflächenseitenketten aus dem großen Datensatz von PPKs $Komp_{RN}$ [99] abgeleitet und bei der Vorhersage der PPKs ähnlich wie in [47] verwendet. Dieses Potential basiert auf Beobachtungen wildtypischer Kontaktflächen und beinhaltet Informationen über die Häufigkeitsunterschiede aller Aminosäurepaare. Daher enthält es Informationen über mehrere physikalisch-chemische und aminosäurespezifische Eigenschaften anhand derer sich PPKs von der restlichen Oberfläche unterscheiden. Durch die Verrechnung der Häufigkeitsunterschiede aller Aminosäuren beinhaltet dieses Potential Informationen über all die Kräfte, die für die Stabilität von Protein-Protein Komplexen verantwortlich sind. Merkmale wie Hydrophobizität, Größe und Polarität werden so zu einem einzigen aussagekräftigen Score für eine Oberflächenposition kombiniert. Durch die Verwendung eines Scores für Paare von Aminosäuren anstatt für einzelne Seitenketten wird die Nachbarschaft einer Position bei ihrer Bewertung mit berücksichtigt. Häufig ist nicht nur eine einzelne Seitenkette verantwortlich für die Eigenschaften eines Oberflächenbereichs. Oft treten hydrophobe und polare Seitenketten benachbart an der Oberfläche auf, so dass größere Oberflächenbereiche entstehen, die denselben physikalisch-chemischen Charakter zeigen [47]. Solche Signale können durch Bewertung benachbarter Paare weit besser berücksichtigt werden als nur bei Betrachten einer einzelnen Position.

5.1.3 Hydrophobe Patches

Die große Bedeutung des hydrophoben Effektes für die Stabilität von Protein-Protein Interaktionen durch das Vergraben großer apolarer Bereiche an der PPK bei der Kom-

plexbildung ist seit längerem bekannt [65] [18]. Auch wenn die Häufigkeit hydrophober Seitenketten in den PPKs von Homodimeren höher ist als in PPKs von Heterodimeren, zeigen auch diese PPKs eine deutlich höhere Hydrophobizität als die restliche Oberfläche [8]. Damit zwei Untereinheiten durch den hydrophoben Effekt eine stabile Interaktion über die Kontaktfläche ausbilden können, dürfen an die hydrophoben Bereiche der Kontaktfläche keine Wassermoleküle gelangen. Dies ist einfacher zu gewährleisten, falls die Kontaktfläche wenige großflächige hydrophobe Bereiche beinhaltet als wenn hydrophobe Positionen einzeln an der Kontaktfläche verstreut sind. Daher tauchen hydrophobe Stellen am Protein meist räumlich benachbart in Clustern auf [47].

Dieser Befund wurde in vorliegender Arbeit konsequent als weiteres Merkmal verwendet. Die Berechnung hydrophober Patches an der Oberfläche erfolgt dabei über ein Kriterium auf dem Niveau von Atomen mithilfe des Programms *QUILT* [136]. Dieses Verfahren hat sich als robust und aussagekräftig bei der Bestimmung hydrophober Patches an der Proteinoberfläche gezeigt und gilt weiterhin als beste Methode [193]. Ähnlich wie in [47] hat sich Information über die Zugehörigkeit zu einem hydrophoben Patch auch in vorliegender Arbeit als förderlich für die Vorhersage von PPKs erwiesen. So ist dieses Merkmal anhand des Klassifikators ohne gewichtete Mittelung über die Nachbarschaft für eine deutliche Steigerung der AUC von $AUC = 0,7280$ auf $AUC = 0,7661$ verantwortlich.

5.1.4 Konserviertheit

Ein weiteres positionsspezifisches Signal, das in dieser Arbeit benutzt wird um PPKs vorherzusagen, ist die evolutionäre Konserviertheit. Aufgrund der Tatsache, dass im Durchschnitt PPKs stärker konserviert sind als die restliche Oberfläche [81] [83] [52], erhöht auch dieses Merkmal die Klassifikationsleistung [47] [85] [156] [45] [44] [74]. Problematisch bei der Vorhersage von PPKs alleine anhand der Konserviertheit ist die Tatsache, dass außer einer PPK auch andere Bereiche der Proteinoberfläche stark konserviert sind. Viele Positionen an Ligandenbindestellen [74] [75] oder Positionen, die eine hohe Bedeutung für die Struktur des Proteins besitzen [71] [72] [73], jedoch nicht an der PPK liegen, sind ebenfalls stark konserviert. Daher würde man durch ein Entscheidungskriterium, das allein auf Konserviertheit basiert, zu viele falsch positive Vorhersagen generieren.

In dieser Arbeit wird Konserviertheit jedoch als zusätzliches Kriterium benutzt, das einen Beitrag zur Verbesserung der Klassifikation leistet. Wie ausgeführt, wurde festgestellt, dass die Qualität der Vorhersage ohne Verwendung des Merkmals Konserviertheit

deutlich abnimmt. Die *AUC* sinkt von $0,7661$ unter Verwendung von Konserviertheit ab auf $0,7452$ ohne Berücksichtigung der Konserviertheit. Als Maß der Konserviertheit hat sich der Score nach *Liu* [117] der Shannonschen Entropie auch nach Normierung gemäß der Aminosäurehäufigkeiten als überlegen erwiesen.

5.1.5 Korrelierte Mutationen

Neben der Konserviertheit haben sich korrelierte Mutationen in dieser Arbeit als zusätzliche, nichtredundante Sequenzinformationen nützlich zur Vorhersage von PPKs erwiesen. Während Konserviertheit vor allem für strikt konservierte Positionen ein starkes Signal liefert, stammen koevolutionäre Signale von Positionen, die nicht strikt konserviert sind, jedoch gewissen Zwängen bei der Wahl der Aminosäure unterliegen.

Die intermolekulare Koevolution von Positionen in PPKs wurden bereits studiert [91]. Die auf Koevolution zurückzuführenden Signale zweier Positionen werden jedoch durch eine Vielzahl anderer Signale überlagert, weshalb es schwierig ist, sie aufzuspüren [194]. So wurde in einer Arbeit [92] gefunden, dass intermolekulare korrelierte Mutationen nicht allgemein aussagekräftig zur Vorhersage von Seitenkettenkontakten zwischen verschiedenen Untereinheiten sind. Ein Grund für diese Aussage dürfte darin bestehen, dass es schwierig ist, Signale gemeinsamer Evolution aufgrund physikalischer Interaktion von korrelierten Mutationen zu unterscheiden, die aufgrund anderer Zwänge in Proteinen auftreten [195] [196].

Trotz dieser Ergebnisse konnte in vorliegender Arbeit gezeigt werden, dass intermolekulare korrelierte Mutationen von Homodimeren auf interagierende Paare von Seitenketten hinweisen. Auch wenn im Falle von Homodimeren nicht ausgeschlossen werden kann, dass das gemessene Signal der Koevolution aus intramolekularen, statt intermolekularen Interaktionen resultiert, konnte gezeigt werden, dass korrelierte Mutationen Informationen zur Vorhersage intermolekularer Kontaktpaare beinhalten. So konnte durch Filterung anhand dieses Signals der Anteil wirklicher intermolekularer Kontaktpaare in einem Datensatz aus allen kombinatorisch möglichen Aminosäurepaaren zweier interagierender Untereinheiten um einen Faktor 18 erhöht werden. Dies stimmt mit dem Befund einer Arbeit von *Mintseries* und *Weng* überein, in der gezeigt wurde, dass Protein-Protein Komplexe auf korrelierte Art evolvieren und dass das Signal korrelierter Mutationen bei obligaten Komplexen weitaus stärker ist als bei transienten [197]. Weiter besagt eine jüngere *in silico* Mutationsstudie von *Fromer* und *Linial* [195], dass transiente PPKs sogar mehr Möglichkeiten besitzen, korreliert zu evolvieren, als obligate PPKs, auch wenn wildtypische Sequenzen von transienten PPKs weniger korrelierte

Mutationen beinhalten als obligate PPKs.

Die bisherige Anwendungsmöglichkeit intermolekularer korrelierter Mutationen sind auf die Bewertung von Seitenkettenpaaren begrenzt. So werden in [91] korrelierte Mutationen dazu benutzt, die Qualität der Strukturvorschläge von Dockingverfahren zu bewerten. Die Software *PresCont* hingegen verrechnet den Score korrelierter Mutationen über das Maß der Konnektivität zu einer Bewertung einzelner Positionen einer interagierenden Untereinheit. Diese Möglichkeit, Scores von Aminosäurepaaren zu einem Score für eine einzelne Position zu kombinieren, wurde bereits zur Vorhersage wichtiger Positionen eines Proteins durch die Auswertung intramolekularer korrelierter Mutationen benutzt [88]. Dieses Verfahren hat sich als robust erwiesen, da es Signale mehrerer Positionspaare verknüpft und so den Einfluss von Rauschen stark unterdrückt.

Eine Schwierigkeit bei der Verwendung der Konnektivität als Signal für die Lage an einer PPK ist es, die wahren Kontaktpositionen aus der großen Anzahl der kombinatorisch möglichen Paare herauszufinden. Dies ist vermutlich der Grund, warum das Merkmal der Konnektivität keinen allzu großen positiven Einfluss auf die Vorhersage der PPK zeigt. Immerhin wurde aber die AUC bei Hinzuziehung korrelierter Mutationen ohne gewichtete Mittelung über die Nachbarschaft deutlich messbar von $0,7612$ auf $0,7661$ gesteigert.

Wie oben gezeigt, ist die Berechnung normierter Transinformation Methoden vorzuziehen, die auf *Pearson*-Korrelation basieren. Dies gilt insbesondere, wenn MSAs mit vielen Sequenzen verfügbar sind. So hat sich bei der Bewertung von Kontaktpaaren gezeigt, dass zwar die *Pearson*-Methode auch bei MSAs, die 500 Sequenzen beinhalten, die Werte der *normierten Transinformation* übertrifft, jedoch beschränkt sich der Vorteil auf den Bereich wenig restriktiv gewählter Schwellwerte. In dem Bereich der ROC- und PROC-Kurven, wo die TPR auf Kosten der FPR bzw. die *Precision* auf Kosten des *Recalls* erhöht wird, erwies sich normierte Transinformation korrelationsbasierten Methoden als deutlich überlegen.

5.1.6 Einbeziehung der Nachbarschaft

Zur Verbesserung der Performanz und Robustheit des Klassifikators wurden in dieser Arbeit zwei verschiedene Verfahren getestet, die es erlauben Information aus der Nachbarschaft einer Position mit einzubeziehen. Zum einen werden die fünf positionsspezifischen Merkmale gewichtet über die Nachbarschaft gemittelt. Zum anderen wurde hierarchisches Clustern benutzt, um räumlich isoliert liegende, *falsch positive* und *falsch*

negative Vorhersagen zu erkennen und zu ändern.

Die Tatsache, dass unter Verwendung einer gewichteten Mittelung über die Nachbarschaft die Werte des Merkmals in der räumlichen Umgebung mit einbezogen werden, macht das Verfahren wesentlich robuster und führt wie in [46] zu einer deutlichen Verbesserung der Klassifikationsleistung. Dabei kommt der *rSASA* innerhalb von *PresCont* neben ihrer Funktion als Merkmal für die Zugehörigkeit zu einer PPK auch eine Rolle als Gewicht bei der gewichteten Mittelung über die Nachbarschaft zu. Eine Gewichtung der Nachbarpositionen proportional der *rSASA* hat sich als deutlich performanter erwiesen als eine Gewichtung reziprok proportional der Entfernung. Der Grund für diesen Befund liegt vermutlich darin, dass die *rSASA* ein Maß für den Anteil einer Aminosäure an der Oberfläche ist. Daher wächst der Einfluss einer Aminosäure auf die physikalisch-chemischen Eigenschaften des untersuchten Oberflächenbereichs proportional mit ihrer *rSASA*. Durch eine gewichtete Mittelung über die Nachbarschaft der *rSASA* selbst lässt sich dagegen die Klassifikationsleistung nicht erhöhen.

Wie in den folgenden Abschnittel diskutiert, wurden bei Einbeziehung gewichteter Mittelung über die Nachbarschaft für die anderen vier Merkmale zum Teil deutliche Zugewinne der Performanz gemessen.

5.1.6.1 Gewichtete Mittelung der intramolekularen Chancenquotienten $\mathbf{PW_{pair\_intra}}$

Die Klassifikationsleistung bei gewichteter Mittelung über die Nachbarschaft der intramolekularen Chancenquotienten verbesserte sich deutlich wie der Anstieg der *AUC*-Werte von $0,7670$ auf $0,7850$ zeigt. Eine Gewichtung proportional der relativen *SASA* in einem großen Umkreis von 9 Å zeigte dabei die beste Performanz.

Die Werte der Scores und der Effekt der Mittelung lässt sich gut durch physikalisch-chemische Effekte erklären. Hydrophobe Bereiche dürfen nur an PPKs in exponierter Lage vorkommen ohne einen großen, negativen enthalpischen Effekt beim Kontakt mit dem Lösungsmittel Wasser zu verursachen. Die höhere Vorhersagegenauigkeit durch eine Gewichtung der Scores proportional zur relativen *SASA* anstatt reziprok zur Distanz lässt sich dadurch erklären, dass die Bedeutung bestimmter Seitenketten für die Eigenschaften der Proteinoberfläche mit ihrer Exponiertheit zunimmt. Seitenketten mit einem hohen *rSASA* Wert ragen weit aus dem Protein hervor, besitzen daher mehr Möglichkeiten für polare Interaktionen und erzeugen einen stärkeren hydrophoben Effekt, sobald sie an der Kontaktfläche bei der Komplexbildung vergraben werden. Vor allem

aromatische und aliphatische Aminosäuren würden bei einem hohen *rSASA*-Wert weit ins Lösungsmittel Wasser ragen und sich dabei ungünstig auf die Stabilität des Proteins auswirken, falls sie nicht durch die Protein-Protein Interaktion vom Wasser abgeschirmt würden.

5.1.6.2 Gewichtete Mittelung der Zugehörigkeit zu einem hydrophoben Patch

Ohne gewichtete Mittelung wird die Zugehörigkeit zu einem hydrophoben Patch als binärer Wert bei der Klassifikation berücksichtigt. Eine Mittelung über die Nachbarschaft hat den Effekt, die Ränder der großflächigen hydrophoben Patches zu glätten. Daher wird die Zugehörigkeit zu einem hydrophoben Patch am Rande des Patches weniger stark bewertet als in seiner Mitte. Durch einen flachen Übergang vom maximalen Wert im Zentralbereich eines großen hydrophoben Patches hin zum Wert 0 außerhalb des Patches lässt sich die Zugehörigkeit zur Kontaktfläche besser approximieren als durch einen Sprung vom Wert eins nach null am Patch-Rand.

Die Stärke der Abflachung wird durch den Abstandsschwellwert $s^{(hpa)}$ bestimmt, innerhalb dessen Nachbarpositionen bei der Mittelung berücksichtigt werden. Die optimale Steigung am Rande eines hydrophoben Patches wird bei geringen Abstandsschwellwerten $s^{(hpa)} = 2$ Å erreicht, was einem relativ steilen Übergang am Rand eines hydrophoben Patches bedeutet. Hydrophobe Patches approximieren folglich die Kontaktfläche bereits in ihrer ursprünglichen Form ohne Mittelung über die Nachbarschaft gut. Dies hängt damit zusammen, dass die Größe hydrophober Patches in Abschnitt 4.4.3.2 mit Hilfe des Parameters der polaren Extension $PE = 1,7$ Å optimal für die Anwendung bei Protein-Protein Kontaktflächen gewählt wurde.

5.1.6.3 Gewichtete Mittelung der Konserviertheit

Der Beitrag der Konserviertheit zur Klassifikation ist durch Mittelung kaum zu verbessern. Lediglich bei sehr geringen Werten von $w_{rSASA}^{(cons)}$ und $s^{(cons)}$ war eine unbedeutende Erhöhung der Performanz zu messen. Der Grund dieses Befundes mag darin liegen, dass zwar die Kontaktfläche grundsätzlich stärker konserviert ist als die restliche Oberfläche, jedoch an der restlichen Oberfläche auch andere funktional wichtige Positionen auftreten, die im zugehörigen MSA strikt konserviert sind [74] [75]. Durch die Übertragung dieses Signals auf die Nachbarpositionen würde es zu einer Erhöhung der falsch positiven Vorhersagen kommen, weshalb durch gewichtete Mittelung über die Nachbarschaft der Konserviertheit keine Verbesserung der Performanz erreicht werden konnte.

Ein weiterer Grund für diesen Befund könnte sein, dass konservierte Positionen bevorzugt einzeln liegen und daher nicht durch Signale aus der Nachbarschaft profitieren können.

5.1.6.4 Gewichtete Mittelung der Konnektivität

Der Beitrag der Konnektivität zur Klassifikationsleistung konnte durch eine gewichtete Mittelung über die Nachbarschaft kaum gesteigert werden. Bei der Parameteroptimierung wurde ein extrem großes Gewicht von $w_{rSASA}^{(conn)} = 1.9$ bei einem kleinen Wert für den Nachbarschaftsradius von $s^{(conn)} = 2$ Å als Optimum gefunden. Aus dem geringen Nachbarschaftsradius lässt sich folgern, dass Positionen mit hoher Konnektivität bevorzugt einzeln oder in sehr enger Nachbarschaft an der Kontaktfläche auftreten. Das extrem große Gewicht von $w_{rSASA}^{(conn)} = 1.9$ lässt sich durch die geringe Anzahl berücksichtigter Nachbarn aufgrund des geringen Nachbarschaftsradius erklären. Aufgrund der wenigen Nachbarn, über die gemittelt wird, behält die eigentlich zu bewertende Position auch bei einem extrem hohen Gewicht von $w_{rSASA}^{(conn)} = 1.9$ noch genügend Einfluss auf den resultierenden Wert. Würde dagegen bei einem derart hohen Gewicht ein Nachbarschaftsradius von 9 Å wie bei den intramolekularen Chancenquotienten benutzt, so würde der Einfluss der zu bewertenden Position auf den Wert der Konnektivität in der Mittelung über die Nachbarschaft untergehen.

5.1.6.5 Hierarchisches Clustern

Neben gewichteter Mittelung über die Nachbarschaft wurde hierarchisches Clustern verwendet um Informationen aus der Nachbarschaft zur Bewertung einer einzelnen Position mit einzubeziehen. Nach Verrechnung der Merkmale durch die SVM zu einer Prognose der Kontaktfläche werden die als positiv vorhergesagten Positionen durch hierarchisches Clustern anhand ihrer räumlichen Nachbarschaft zu Gruppen zusammengefasst. Dadurch ist es möglich *falsch positive* und *falsch negative* Vorhersagen, die bevorzugt einzeln bzw. umgeben von positiven Vorhersagen auftreten, zu erkennen und den Wert ihrer Vorhersage zu ändern. Um den Gewinn an Performanz durch hierarchisches Clustern zu messen, wurde die beschriebene Methode mit dem Klassifikator aus Abschnitt 4.4.3.2 und allen optimierten Parametern der Eingabedaten ohne gewichtete Mittelung über die Nachbarschaft, der am Datensatz $Komp_{kanon}$ einen AUC-Wert von $0,766$ aufweist, kombiniert. Durch diese Methode zur Nachbearbeitung der Ergebnisse war, ähnlich wie in [185], eine deutliche Steigerung der Klassifikationsleistung anhand des AUC-Wertes auf $0,773$ (siehe Abschnitt 4.7.1) zu erreichen.

Da gewichtete Mittelung über die Nachbarschaft die Vorhersage für eine Seitenkette auf ähnliche Art verbessert wie *hierarchisches Clustern*, besitzen beide Methoden eine gewisse Redundanz. Daher war es auch nicht möglich, eine weitere Verbesserung der Vorhersage durch Kombination von gewichteter Mittelung über die Nachbarschaft und *hierarchischem Clustern* zu erreichen. Die Güte der Vorhersage unter Einbeziehung gewichteter Mittelung von $AUC = 0,7945$ konnte durch *hierarchisches Clustern* der Vorhersagen lediglich auf $AUC = 0,7958$ gesteigert werden. Da dieser Unterschied die Messgenauigkeit nur gering übersteigt, wird in der fertigen Version von *PresCont* keine Nachbearbeitung der Ergebnisse durch hierarchisches Clustern verwendet.

5.2 Kern- und Randbereich von Kontaktflächen

Nach einer gängigen Theorie [37] enthalten typische PPKs einen Kernbereich mit überwiegend hydrophobem Charakter, der die energetischen *Hot Spots* enthält. Voraussetzung für die Stabilität der Bindung aufgrund des hydrophoben Effektes ist jedoch, dass kein Wasser in den hydrophoben Bereich der Kontaktfläche vordringen kann. Nach der erwähnten Theorie wird dies durch einen O-Ring mit hydrophilem Charakter am Randbereich der Kontaktfläche verhindert [23]. Gilt diese Theorie, so kommen in den PPKs die Aminosäuren mit unterschiedlichen Häufigkeiten vor. Dieser Befund wurde in [37] über eine Methode zur Bestimmung von Kern- und Randbereich, die auf *SASA* basiert, bestätigt.

Die neueren Methoden *Intervor* [38] und *PIA* [134] benutzen bei der Einteilung von PPKs in Zentral- und Randbereiche Triangulationen, um Nachbarschaften zwischen Atomen und Seitenketten zu definieren. Auf diese Weise erhält man eine Definition von Zentral- und Randbereich, die mehr der intuitiven, räumlichen Vorstellung von Zentrum und Rand einer Kontaktfläche entspricht, als bei Methoden basierend auf *SASA*. In vorliegender Arbeit wurden diese beiden Methoden miteinander verglichen. Dazu wurden wissensbasierte Potentiale für Aminosäurehäufigkeiten im Kernbereich nach beiden Verfahren berechnet und auf Unterschiede zu analogen Potentialen für die gesamte Kontaktfläche hin untersucht. Außerdem wurde die Vorhersagegenauigkeit von *PresCont* im Kernbereich verglichen mit derjenigen an der gesamten Kontaktfläche. Die Ergebnisse werden im nächsten Abschnitt diskutiert.

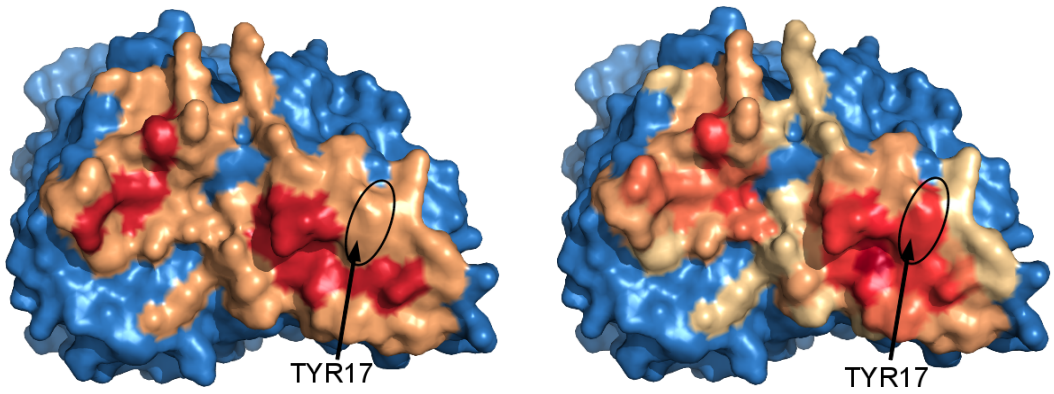
5.2.1 Aminosäurehäufigkeiten im Kern von Kontaktflächen

Aus dem Vergleich der Chancenquotienten für Häufigkeitsverteilungen einzelner Aminosäuren an der gesamten Kontaktfläche mit denjenigen des Kernbereichs, berechnet nach *PIA* bzw. *Intervor*, ergibt sich, dass im Zentrum einer Kontaktfläche aliphatische Aminosäuren deutlich stärker bevorzugt sind als an der gesamten Kontaktfläche und daher auch stärker als am Rand. Umgekehrt sind geladenen Aminosäuren an der gesamten Kontaktfläche stärker bevorzugt als im Kern. Diese Befunde sprechen dafür, dass der Kernbereich einer Kontaktfläche in Übereinstimmung mit [37] [38] einen stärker apolaren Charakter besitzt als der Randbereich.

Weniger eindeutig ist das Ergebnis im Hinblick auf aromatische Seitenketten. Während *Intervor* auch diese im Kernbereich stärker bevorzugt sieht als am Rand und erst ab einer extrem restriktiven Definition von Kern ($VSO \geq 7$) die Werte der Chancenquotienten sinken, sind im Kern nach *PIA* die Aromaten weniger stark bevorzugt als am Rand. Will man jedoch *Intervor* vergleichbar restriktiv justieren wie *PIA*, so müsste man $VSO \geq 3$ wählen um ungefähr gleich große Teilbereiche der Kontaktfläche als Kern zu definieren.

Dieser Befund kann nur von Unterschieden in den Konzepten der beiden Methoden *PIA* und *Intervor* verursacht sein. Während *PIA* die PPK ringförmig in Bereiche der Dicke einer Seitenkette einteilt, beträgt die Dicke einer Voronoi-Schale bei *Intervor* nur eine Atomlage. Bei größeren Seitenketten sind folglich größere Unterschiede zwischen beiden Definitionen zu erwarten als bei kleineren Seitenketten. Dies stimmt mit dem Befund überein, dass die größten Diskrepanzen der Chancenquotienten bei den großen aromatischen Seitenketten auftreten.

Dies gilt beispielsweise für die Seitenkette TYR17 in Abbildung 5.1, die sowohl Atome benachbart zum Außenbereich besitzt als auch Atome, die sich weiter im Zentrum der PPK befinden. Während *PIA* die Aminosäure als Ganzes zur äußersten ersten Schale rechnet, da einige ihrer Atome benachbart zum Außenbereich liegen, werden durch *Intervor* ihre Atome einzeln nach ihrer Lage in der Kontaktfläche in Schalen eingeteilt. Die aminosäurespezifische *VSO* ergibt sich dann durch Mittelung über ihre Atome und besitzt einen erheblich größeren Wert als bei der Einteilung durch PIA. Man erkennt also, dass es nicht trivial ist zu beurteilen, wann eine Seitenkette dem Kernbereich einer Kontaktfläche zuzuordnen ist. Vor allem für große Seitenketten, zu denen die Aromaten zählen, hängt die Unterteilung in Zentral- und Randbereich einer PPK stark vom Berechnungsverfahren ab.



(a) (b)

Abbildung 5.1: Einteilung großer Aminosäuren in Kern und Rand
Diese Abbildung zeigt am Beispiel der *Dihydropyrimidinase* (PDB-ID: 1GKP Kette A) aus *Thermus* Sp. die unterschiedliche Behandlung der aromatischen Seitenkette Tyrosin an Position 17 (TYR17). Die Nichtkontaktfläche der Kette ist blau eingefärbt. An der Kontaktfläche ändern sich die Farben zum Zentrum hin von hell- nach dunkelrot. (a) Aufgrund seiner Ausdehnung besitzt TYR17 bei der Berechnung der *PIA*-Schalennummer in der Triangulation Atome mit Kontakt zum Außenbereich und wird zum äußersten Rand der Kontaktfläche gerechnet. (b) Intervor berechnet für jedes Atom eine eigene Schalennummer (*VSO*). Die *VSO* einer ganzen Seitenkette erhält man durch Mittelung über ihre Atome. Dadurch erhält TYR17 hier die Schalennummer 5 zugewiesen, was auf eine zentrale Lage hindeutet.

Eine eindeutige Aussage darüber, welches der beiden Berechnungsverfahren besser ist, lässt sich nicht einfach treffen. Einerseits bietet Intervor durch die Einteilung in Schalen auf Atomniveau eine höhere Auflösung im Vergleich zu *PIA*. Andererseits jedoch stellt eine Aminosäure die kleinste Einheit dar, die durch eine Mutation verändert werden kann. Daher sind die meisten Signale, die in dieser Arbeit untersucht werden, auf Ebene von Aminosäuren und nicht von Atomen definiert. Um folglich Unterschiede in den Signalen an verschiedenen Bereichen der PPK messen zu können, wird eine Einteilung in Kern- und Randbereich auf Ebene der Aminosäuren benötigt. Während *PIA* nativ auf Ebene von Aminosäuren arbeitet, muss bei *Intervor* die *VSO* einer Aminosäure über die *VSO* ihrer Kontaktatome gemittelt werden.

5.2.2 Klassifikationsleistung im Kern von Kontaktflächen

Wie gezeigt, ist die Häufigkeitsverteilung der Aminosäureseitenketten im Randbereich einer Kontaktfläche derjenigen der restlichen Oberfläche ähnlicher als die Häufigkeitsverteilung im Kernbereich (siehe Abschnitt 4.3.2.1 und [37]). Daneben wurde gefunden, dass neben der Abschirmung von Wasser auch Eigenschaften wie Konserviertheit und apolarer Charakter mit der durch *Intervor* berechneten *VSO* korrelieren [38]. Im Pro-

gramm *PresCont* werden diese Merkmale entweder direkt, wie bei der Konserviertheit, oder indirekt wie im Falle des apolaren Charakters über die Zugehörigkeit zu einem hydrophoben Patch, als Eingabedaten des Klassifikators verrechnet. Daher ist zu erwarten, dass eine richtige Vorhersage für Seitenketten im Zentralbereich der Kontaktfläche leichter möglich ist als für den Randbereich.

Diese Vermutung hat sich in dieser Arbeit bestätigt. So konnte gezeigt werden, dass Seitenketten im Zentralbereich der Kontaktfläche zu einem deutlich höheren Bruchteil richtig als PPK eingeordnet werden als Positionen im Randbereich. Die *AUC* einer Vorhersage der gesamten PPKs im Datensatz $Komp_{kanon}$ steigt von $0,766$ auf $0,783$ bei der Vorhersage des *PIA*-Zentralbereichs.

5.3 Vergleich mit anderen Methoden zur Vorhersage von Kontaktflächen

Um die Klassifikationsleistung angemessen beurteilen zu können, wurde ein Vergleich von *PresCont* mit *ProMate* [47] und *Sppider* [46] angestellt, zwei etablierten Programmen zur Vorhersage von PPKs. Die Gemeinsamkeiten der Signale transienter und obligater PPKs sind hinreichend groß, um transiente Komplexe von Klassifikatoren, die anhand obligater Komplexe trainiert wurden, vorherzusagen [44]. Daher wurde die Klassifikationsleistung von *PresCont* sowohl anhand des Datensatzes $Komp_{kanon}$ als auch anhand des Datensatzes $Komp_{trans}$ ausgewertet. Da *Sppider* eine hohe Performanz am Datensatz $Komp_{kanon}$ gezeigt hat, während *ProMate* auf die Vorhersage transienter PPKs spezialisiert ist, wurden diese beiden Programme für den Vergleich herangezogen.

Angewandt auf den Datensatzes $Komp_{kanon}$ übertreffen *PresCont* und *Sppider* die Vorhersagegenauigkeit von *ProMate* deutlich, während *Sppider* und *PresCont* PPKs etwa gleich gut vorhersagen können. Vergleicht man die drei Programme jedoch anhand des Datensatzes $Komp_{trans}$, so übertrifft die Vorhersagegenauigkeit von *ProMate* deutlich diejenige von *Sppider* und *PresCont*.

Wie sich auch anderweitig gezeigt hat [198], hängt die Qualität der Vorhersage stark vom benutzten Datensatz und vor allem von der Art der Komplexe ab. Die beiden getesteten Datensätze unterscheiden sich hinsichtlich des Typs der enthaltenen Komplexe. Während der Datensatz $Komp_{kanon}$ durchwegs obligate Homodimere beinhaltet, besteht der Datensatz $Komp_{bench}$ aus transienten Heterodimeren. Da transiente Komple-

xe wesentlich schwächer binden als permanente Komplexe, sind ihre Interaktionsflächen kleiner, tendieren zu mehr polaren Seitenketten und besitzen eine geringere Oberflächenkomplementarität als die permanenten Komplexe [8] [7] [199]. Daher ist es schwieriger, transiente PPKs vorherzusagen als obligate [198]. Dies erklärt auch die allgemein geringere Vorhersagequalität am Datensatz $Komp_{bench}$ verglichen mit dem Datensatz $Komp_{kanon}$. Bemerkenswert ist jedoch, dass *ProMate* trotz aller Gemeinsamkeiten der Signale obligater und transienter Interaktionen, aus den schwächeren Signalen transienter PPKs eine bessere Vorhersage generiert als aus den stärkeren Signalen obligater Interaktionen. Der Grund dieses Befundes mag daran liegen, dass *ProMate* mehr strukturelle Informationen wie den *R-Faktor* auswertet. Derartige Informationen könnten bei transienten Komplexen, bei denen auch Komformationsänderungen der Hauptketten durch die Komplexbildung auftreten können, von größerer Bedeutung sein als bei obligaten Komplexen.

Insgesamt lässt sich das Fazit ziehen, dass *ProMate*, *Sppider* und *PresCont* im Durchschnitt ähnliche Vorhersagegenauigkeit zeigen. Während die Vorteile von *Sppider* und *PresCont* bei der Vorhersage obligater Kontaktflächen liegen, ist *ProMate* auf transiente Interaktionen spezialisiert.

Der Vergleich des Aufbaus von *Sppider*, *ProMate* und *PresCont* macht klar, dass sowohl *ProMate* als auch *Sppider* eine Vielzahl verschiedener Scores kombinieren, um daraus eine aussagekräftige Vorhersage zu generieren. *PresCont* dagegen berücksichtigt lediglich fünf Merkmale mit jeweils hohem Informationsgehalt mit Hilfe einer SVM, um daraus eine Vorhersage ähnlicher Qualität zu generieren. Wie in einem jüngeren Übersichtsartikel [198] festgestellt wird, scheint die Vorhersage von PPKs an einem Sättigungspunkt angelangt, an dem es unwahrscheinlich scheint, auf Grundlage der etablierten Merkmale zur Vorhersage von PPKs die Qualität der Vorhersagen noch deutlich verbessern zu können. Für weitere Fortschritte müssten neue Merkmale gefunden werden, die die Eigenschaften von PPKs mit deutlich höherer Genauigkeit beschreiben. Eine höhere Auflösung wirkt sich jedoch im Allgemeinen negativ auf die benötigte Rechenzeit aus. *PresCont* hingegen beschränkt sich auf eine minimale Anzahl an Eigenschaften, mit der eine Performanz erreicht wird, die derjenigen andrere Verfahren nicht nachsteht.

Der einfache Aufbau von *PresCont* trägt zur Robustheit des Verfahrens bei. Dies erkennt man auch in der geringen Empfindlichkeit von *PresCont* gegenüber Overlearning. Generell sind SVMs wenig empfindlich gegenüber *Overlearning* sind. So wird die trennende Hyperebene nicht aufgrund weniger Trainingsbeispiele optimiert, sondern anhand aller Trainingsbeispiele, die der SVM präsentiert werden. Andererseits fördert auch die geringe Anzahl an Eigenschaften und die damit verknüpfte geringe Anzahl an freien Parametern die Robustheit des Verfahrens. Wie in [200] gezeigt wurde, beeinflusst die

Verwendung extrem vieler Merkmale, die Redundanzen enthalten oder von denen einige irrelevant für die Klassifikation sind, die Performanz einer SVM negativ.

Generell kann gefolgert werden, dass *PresCont* für einen Großteil der Proteinkomplexe Vorhersagen hoher Güte erzeugt. Damit ist ein Werkzeug entstanden, das für eine Vielzahl von Aufgaben eingesetzt werden kann.

6 Ausblick

Im letzten Kapitel wurden bereits die Bestandteile von *PresCont* im Kontext bisheriger Forschungsergebnisse diskutiert. In diesem Abschnitt sollen Verbesserungsmöglichkeiten dargestellt und mögliche zukünftige Anwendungen der Software beschrieben werden.

6.1 Verbesserung des Merkmals der Konnektivität

In dieser Arbeit wurde das Merkmal der Konnektivität neu eingeführt um Information aus Aminosäurepaaren, die über die PPK hinweg in Kontakt sind, zu einem Merkmal für eine einzelne Position an der Oberfläche eines Proteins zu verrechnen und bei der Vorhersage der PPK zu berücksichtigen. Ein Grund der relativ geringen Bedeutung dieses Merkmals für die Qualität der Vorhersage ist, dass sich aufgrund der Einbeziehung des Interaktionspartners die wenigen Signale intermolekular interagierender Seitenketten aus der großen Menge aller kombinatorisch möglichen Paare abzeichnen müssen. Wie jedoch in dieser Arbeit gezeigt wurde, ist ein Signal vorhanden, das zur Vorhersage der PPK genutzt und ausgebaut werden kann.

Dieses Signal könnte durch Berücksichtigung der Tatsache verstärkt werden, dass die an intermolekularen Interaktionen beteiligten Positionen räumlich benachbart liegen müssen. Daher könnte es hilfreich sein, sich bei der Berechnung des Konnektivitätswertes auf Positionen des Interaktionspartners zu beschränken, die räumlich benachbart sind.

Ein anderer Ansatz, um Information der räumlichen Nachbarschaft von Kontaktpositionen mit einzubeziehen ist es, die signifikanten paarweisen Scores, die zur Berechnung der Konnektivität herangezogen werden, weiter einzuschränken auf Scores für Paare von Positionen, die sich bei räumlichem Clustern als benachbart herausstellen. Zum Clustern der Positionspaare könnten dabei die beiden 3-dimensionalen Koordinaten eines Seitenkettenpaares zu einer 6-dimensionalen Koordinate zusammengefasst werden,

anhand derer die räumlichen Abstände der Positionspaare als euklidischer Abstand im 6-dimensionalen Raum bestimmt werden. Durch die Prozedur des Clusterns würden zunächst eine oder mehrere ungenaue Vorhersagen der PPK generiert. Diese primäre Vorhersage würde anschließend zur Auswahl der Positionen genutzt, die bei der Berechnung der Konnektivität zu berücksichtigen sind.

Es ist zu erwarten, dass man durch diese zusätzliche Bedingung an signifikante intermolekulare Scores viele irrelevante Signale vor der Berechnung der Konnektivität ausfiltert und deshalb den Informationsgehalt der Konnektivität erhöht.

6.2 Weitere Merkmale

Der Aufbau von *PresCont* wurde auf den Datensatz $Komp_{kanon}$ abgestimmt, der aus obligaten Homodimeren besteht. Daher erreicht *PresCont* auch an diesem Datensatz näherungsweise die Vorhersagequalität von *Sppider*, dessen Stärken bei der Vorhersage obligater PPKs liegt. Bei der Vorhersage transienter PPKs erreicht *PresCont* einen etwas höheren Wert der *AUC* als *Sppider*. Anhand des deutlich höheren *AUC*-Wertes von *ProMate* bei $Komp_{trans}$ wird jedoch deutlich, dass es Möglichkeiten geben muss, mit etablierten Merkmalen, die Performanz von *PresCont* an transienten PPKs zu verbessern.

Wie aus dem Vergleich der drei Programme *Sppider*, *ProMate* und *PresCont* hervorgeht, scheint es insbesondere nicht möglich zu sein, mit Hilfe des gleichen Programms sowohl obligate Homodimere als auch transiente Heterodimere vorherzusagen. Die Performanz bei der Vorhersage transienter PPKs könnte daher einerseits durch eine Optimierung sämtlicher Parameter zur Berechnung der Eingabedaten und der SVM am Datensatz $Komp_{trans}$ verbessert werden. Andererseits ist es möglich, dass an transienten Kontaktflächen andere bzw. weitere Merkmale als die fünf in dieser Arbeit benutzten eine größere Rolle spielen. Daher könnte auch eine andere Auswahl der Merkmale die Qualität der Vorhersage transienter PPKs verbessern.

So könnte die Sekundärstruktur als zusätzliches Merkmal zur Vorhersage transienter PPKs, ähnlich wie bei *ProMate* [47], genutzt werden. Diese Vermutung ist darin begründet, dass sich die PPKs von Homodimeren und transienten Heterodimeren hinsichtlich der Häufigkeit von Sekundärstrukturelementen unterscheiden. So wurde an einem Datensatz transienter Heterodimere gefunden, dass β-Stränge an PPKs bevorzugt sind [47], während die PPKs in einem Datensatz, der zu $\frac{2}{3}$ aus Homodimeren und Antigen-Antikörper Komplexen besteht, eine gegenteilige Verteilung besitzen [201]. Ne-

ben Bereichen mit geordneter Sekundärstruktur spielen auch ungeordnete Schleifen an PPKs eine wichtige Rolle. So sind 40% aller Reste an PPKs in Schleifen positioniert [202].

Da sich in einer Vorauswahl der von *PresCont* benutzten Merkmale Oberflächenkrümmung als wenig aussagekräftig erwiesen hat, wird diese Art von Information in diesem Programm nicht berücksichtigt. Ein Grund für den geringen Informationsgehalt der Oberflächenkrümmung ist möglicherweise, dass häufig Wassermoleküle in die Kontaktfläche eingelagert sind, die über Wasserstoffbrücken Kontakte zwischen polaren Seitenketten vermitteln und so zur Stabilität des Komplexes beitragen [38]. Um den Informationsgehalt von geometrischen Merkmalen, wie der Oberflächenkrümmung, so weit zu erhöhen, dass sie die Vorhersagequalität von *PresCont* deutlich verbessern können, ist es daher ein erfolgversprechender Ansatz, Wassermoleküle mit zu berücksichtigen. Eine einfache Möglichkeit Wasser zu berücksichtigen ist es, wie in [47] die Anzahl der Wassermoleküle an Oberflächenpositionen miteinander zu vergleichen. Eine fortgeschrittener Ansatz wäre es, über Moleküldynamik Interaktionen des Wassers zu simulieren und so die Anzahl der Wassermoleküle, die mit der Oberfläche interagiert, besser abschätzen zu können. Die wichtige Rolle von Wassermolekülen an der PPK zeigte sich auch in [47]. Dort wurde gefunden, dass Wassermoleküle an der PPK der Monomerstruktur häufiger auftreten als an der restlichen Oberfläche. Unter Einbeziehung von Wassermolekülen ist anzunehmen, dass sich die Form der Oberfläche der PPK derjenigen nähert, die im Komplex vorliegt. Daher könnte unter Berücksichtigung von Wassermolekülen bei der Berechnung der Oberflächenkrümmung auch dieses Merkmal der Proteinoberfläche eine größere Rolle bei der Vorhersage der PPKs spielen.

6.3 Anwendungsmöglichkeiten

Wie an einem Beispiel gezeigt, kann *PresCont* dazu benutzt werden, *in vivo* Kontakte von künstlichen Kristallkontakten zu unterscheiden. Da die Vorhersage einer Kontaktfläche, im Gegensatz zum Training der SVM, wenig Rechenzeit in Anspruch nimmt, wäre durch *PresCont* eine Auswertung aller in der PDB vorhandenen Kontaktflächen zwischen zwei Untereinheiten möglich. Damit könnten frühere Annotationen relevanter Protein-Kontakte in der PDB [203] verbessert werden.

Daneben kann eine Vorhersage der PPK dazu genutzt werden, Verfahren des Protein-Protein Dockings zu unterstützen. Da es jedoch extrem viele Möglichkeiten gibt ($O(10^9)$) [4], zwei Proteine miteinander in Kontakt zu bringen, benötigen Docking Verfahren sehr viel Rechenzeit. Eine aussagekräftige Vorhersage der PPK kann die Anzahl mögli-

cher Konformationen des Komplexes stark reduzieren. Manche Programme für Protein-Protein Docking, wie z.B. *Haddock* [204] stellen Möglichkeiten zu Verfügung, Hinweise auf die Lage der PPK als zusätzliche Information mit einzubeziehen. Auf diese Weise kann die Qualität des Ergebnisses enorm erhöht und die benötigte Rechenzeit deutlich verringert werden.

Danksagung

Dieses Kapitel ist allen Leuten gewidmet, ohne die das Gelingen dieser Arbeit nicht möglich gewesen wäre.

Zuerst möchte ich meinem Doktorvater PD Dr. Rainer Merkl für die wunderbare Betreuung dieser Arbeit danken. Durch seine langjährige Erfahrung in der Bioinformatik konnte er mir stets mit Rat und Tat zur Seite stehen. Dabei ermöglichte er es mir, die Arbeit nach eigenen Vorstellungen zu gestalten. Vielen Dank für alle diversen Hilfestellungen und die Unterstützung während der letzten Jahre.

Großer Dank gilt Herrn Prof. Dr. Reinhard Sterner, dessen Forschungsinteressen im Bereich Protein-Protein Interaktionen die Basis dieser Arbeit gelegt haben. Durch seine Idee, die Untersuchung biochemischer Sachverhalte durch computergestütze Methoden voranzutreiben, habe ich als Physiker meine Aufgabe in der biochemischen Forschung gefunden.

Herrn PD Dr. W. Gronwald danke ich für die Übernahme der Funktion des Zweitgutachters dieser Arbeit und für die informative und lehrreiche Einführung in das Thema Protein-Protein Docking.

Besonderer Dank gilt Prof. Dr. Christian Icking und den Diplom- bzw. Masterstudenten Vincent Wolowski, Martin Staudigel, Thomas Trenner und Meik Bittkowski von der FernUniversität in Hagen, die maßgeblich an diesem Projekt beteiligt waren. Die Zusammenarbeit, insbesondere die fruchtbaren Diskussionen während unserer Skype-Konferenzen, hat sehr viel Spass gemacht.

Ich danke Prof. Dr. Jens Meiler von der Vanderbilt University in Nashville, der es mir durch einen Forschungsaufenthalt in seiner Arbeitsgruppe ermöglichte, tiefe Einblicke in ein großes Softwareprojekt und in eine hervorragende Arbeitsgruppe in den USA zu gewinnen. Ich danke ihm und dem DAAD für die Stipendien, die die Durchführung dieses Teilprojektes ermöglichten. Vielen Dank auch an alle Meilers für die freundliche Aufnahme, die Hilfestellungen und die schönen Ausflüge, die mir das Land näher brachten.

Großer Dank gilt allen aktuellen und ehemaligen Zimmerkollegen für die angenehme Atmosphäre und so manche Hilfe. Insbesondere danke ich André Fischer für die informative Einführung in viele Methoden der Bioinformatik. Dietmar Birzer danke ich sehr für die angenehme Zeit, die eindrucksvollen Ausflüge und die schönen Abende in unserer WG während des Forschungsaufenthaltes in den USA. Jan-Oliver Janda gebührt großer Dank für seine Hilfe bei der automatisierten Abfrage von Sequenzdatenbanken.

Allen aktuellen und ehemaligen Sternern danke ich für die geduldige Beantwortung biochemischer Fragen, die interessanten Einblicke in verschiedene Projekte und für das angenehme Arbeitsklima am Lehrstuhl. Außerdem danke ich für die lustigen Kickerpausen, die den Programmieralltag erheblich auflockerten. Ich danke Florian Busch, David Peterhoff und Monika Meier, die mir zur späten Abendstunde und am Wochenende solidarisch an der Uni sowie beim Schachtelwirt Gesellschaft leisteten.

Außerdem danke ich den Praktikanten/-innen der Bioinformatik Anke Behr, Andreas Geißner, Catrin Wartner, Felix Graßmann und Marion Huber für ihre Beiträge, ihr Engangement und für eine kurzweilige Zeit.

Ein weiteres Dankeschön gilt Dr. Marco Bocola für die wertvollen Ratschläge und Hilfen in vielen Belangen der Chemie und Strukturbiologie.

Klaus Tiefenbach danke ich für seine Hilfe bei der Hardwarebeschaffung und bei diverser Systemadministration. Daneben bedanke ich mich für die interessanten und unterhaltsamen Diskussionen rund um das Thema Computer.

Großer Dank gebührt Anke für ihren Zuspruch und ihre Unterstützung in allen Belangen. Ganz besonders danke ich meinen Eltern für ihr Vertrauen in mich während meines Studiums und meiner Promotion. Ohne ihre Unterstützung in all den Jahren wäre diese Arbeit nicht möglich gewesen.

Literaturverzeichnis

[1] Merkl, R. und S. Waack: *Bioinformatik Interaktiv.* Wiley-Blackwell, 2009.

[2] Yu, H., P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill und M. Vidal: *High-quality binary protein interaction map of the yeast interactome network.* Science, 322(5898):104–10, Oktober 2008.

[3] Stumpf, M. P. H., T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe und C. Wiuf: *Estimating the size of the human interactome.* Proc Natl Acad Sci U S A, 105(19):6959–64, Mai 2008.

[4] Ritchie, D. W.: *Recent progress and future directions in protein-protein docking.* Curr Protein Pept Sci, 9(1):1–15, Februar 2008.

[5] .Nooren, I. M. A und J. M. Thornton: *Diversity of protein-protein interactions.* EMBO J, 22(14):3486–92, Juli 2003.

[6] Bahadur, R. P. und M. Zacharias: *The interface of protein-protein complexes: analysis of contacts and prediction of interactions.* Cell Mol Life Sci, 65(7-8):1059–72, April 2008.

[7] Ofran, Y. und B. Rost: *Analysing six types of protein-protein interfaces.* J Mol Biol, 325(2):377–87, Januar 2003.

[8] Nooren, I. M. A. und J. M. Thornton: *Structural characterisation and functional significance of transient protein-protein interactions.* J Mol Biol, 325(5):991–1018, Januar 2003.

[9] Nishi, H. und M. Ota: *Amino acid substitutions at protein-protein interfaces that modulate the oligomeric state.* Proteins, 78(6):1563–74, Mai 2010.

[10] Bera, I. und S. Ray: *A study of interface roughness of heteromeric obligate and non-obligate protein-protein complexes.* Bioinformation, 4(5):210–5, 2009.

[11] Block, P., N. Weskamp, A. Wolf und G. Klebe: *Strategies to search and design stabilizers of protein-protein interactions: a feasibility study.* Proteins, 68(1):170–86, Juli 2007.

[12] Mintseris, J. und Z. Weng: *Atomic contact vectors in protein-protein recognition.* Proteins, 53(3):629–39, November 2003.

[13] De, S., O. Krishnadev, N. Srinivasan und N. Rekha: *Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different.* BMC Struct Biol, 5:15, 2005.

[14] Zhu, H., F. S. Domingues, I. Sommer und T. Lengauer: *NOXclass: prediction of protein-protein interaction types.* BMC Bioinformatics, 7:27, 2006.

[15] Jones, S. und J. M. Thornton: *Analysis of protein-protein interaction sites using surface patches.* J Mol Biol, 272(1):121–32, September 1997.

[16] Lo Conte, L., C. Chothia und J. Janin: *The atomic structure of protein-protein recognition sites.* J Mol Biol, 285(5):2177–98, Februar 1999.

[17] Keskin, O., I. Bahar, A. Y. Badretdinov, O. B. Ptitsyn und R. L. Jernigan: *Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions.* Protein Sci, 7(12):2578–86, Dezember 1998.

[18] Glaser, F., D. M. Steinberg, I. A. Vakser und N. Ben-Tal: *Residue frequencies and pairing preferences at protein-protein interfaces.* Proteins, 43(2):89–102, Mai 2001.

[19] Jones, S., A. Marin und J. M. Thornton: *Protein domain interfaces: characterization and comparison with oligomeric protein interfaces.* Protein Eng, 13(2):77–82, Februar 2000.

[20] Ansari, S. und V. Helms: *Statistical analysis of predominantly transient protein-protein interfaces.* Proteins, 61(2):344–55, November 2005.

[21] Wolowski, V.: *Computational analysis of protein-protein complexes related to knowledge-based predictions of interaction.* Diplomarbeit, Fernuniversität in Hagen, April 2008.

[22] Morrison, K. L. und G. A. Weiss: *Combinatorial alanine-scanning.* Curr Opin Chem Biol, 5(3):302–7, Juni 2001.

[23] Bogan, A. A. und K. S. Thorn: *Anatomy of hot spots in protein interfaces.* J Mol Biol, 280(1):1–9, Juli 1998.

[24] Hu, Z., B. Ma, H. Wolfson und R. Nussinov: *Conservation of polar residues as hot spots at protein interfaces.* Proteins, 39(4):331–42, Juni 2000.

[25] DeLano, W. L., M. H. Ultsch, A. M. de Vos und J. A. Wells: *Convergent solutions to binding at a protein-protein interface.* Science, 287(5456):1279–83, Februar 2000.

[26] DeLano, W. L.: *Unraveling hot spots in binding interfaces: progress and challenges.* Curr Opin Struct Biol, 12(1):14–20, Februar 2002.

[27] Thorn, K. S. und A. A. Bogan: *ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions.* Bioinformatics, 17(3):284–5, März 2001.

[28] Jones, S. und J. M. Thornton: *Principles of protein-protein interactions.* Proc Natl Acad Sci U S A, 93(1):13–20, Januar 1996.

[29] Wells, J. A.: *Systematic mutational analyses of protein-protein interfaces.* Methods Enzymol, 202:390–411, 1991.

[30] Wells, J. A. und A. M. de Vos: *Structure and function of human growth hormone: implications for the hematopoietins.* Annu Rev Biophys Biomol Struct, 22:329–51, 1993.

[31] Wells, J. A.: *Structural and functional basis for hormone binding and receptor oligomerization.* Curr Opin Cell Biol, 6(2):163–73, April 1994.

[32] Wells, J. A.: *Binding in the growth hormone receptor complex.* Proc Natl Acad Sci U S A, 93(1):1–6, Januar 1996.

[33] Thornton, J. M.: *The Hans Neurath Award lecture of The Protein Society: proteins– a testament to physics, chemistry, and evolution.* Protein Sci, 10(1):3–11, Januar 2001.

[34] Ma, B., H. J. Wolfson und R. Nussinov: *Protein functional epitopes: hot spots, dynamics and combinatorial libraries.* Curr Opin Struct Biol, 11(3):364–9, Juni 2001.

[35] Buckle, A. M., G. Schreiber und A. R. Fersht: *Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution.* Biochemistry, 33(30):8878–89, August 1994.

[36] Moreira, Irina S, Pedro A Fernandes und Maria J Ramos: *Hot spots–a review of the protein-protein interface determinant amino-acid residues.* Proteins, 68(4):803–12, September 2007.

[37] Chakrabarti, Pinak. und J. Janin: *Dissecting protein-protein recognition sites.* Proteins, 47(3):334–43, Mai 2002.

[38] Bouvier, B., R. Grünberg, M. Nilges und F. Cazals: *Shelling the Voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics, and composition.* Proteins, 76(3):677–92, August 2009.

[39] Yu, C.-Y., L.-C. Chou und D. T.-H. Chang: *Predicting protein-protein interactions in unbalanced data using the primary structure of proteins.* BMC Bioinformatics, 11:167, 2010.

[40] Park, J., M. Lappe und S. A. Teichmann: *Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast.* J Mol Biol, 307(3):929–38, März 2001.

[41] Shen, J., J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li und H. Jiang: *Predicting protein-protein interactions based only on sequences information.* Proc Natl Acad Sci U S A, 104(11):4337–41, März 2007.

[42] Guo, Y., M. Li, X. Pu, G. Li, X. Guang, W. Xiong und J. Li: *PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment.* BMC Res Notes, 3:145, 2010.

[43] Yu, J., M. Guo, C. J. Needham, Y. Huang, L. Cai und D. R. Westhead: *Simple sequence-based kernels do not predict protein-protein interactions.* Bioinformatics, 26(20):2610–4, Oktober 2010.

[44] Bradford, J. R. und D. R. Westhead: *Improved prediction of protein-protein binding sites using a support vector machines approach.* Bioinformatics, 21(8):1487–94, April 2005.

[45] Bordner, A. J. und R. Abagyan: *Statistical analysis and prediction of protein-protein interfaces.* Proteins, 60(3):353–66, August 2005.

[46] Porollo, A. und J. Meller: *Prediction-based fingerprints of protein-protein interactions.* Proteins, 66(3):630–45, Februar 2007.

[47] Neuvirth, H., R. R. und G. Schreiber: *ProMate: a structure based prediction program to identify the location of protein-protein binding sites.* J Mol Biol, 338(1):181–99, April 2004.

[48] Jones, S. und J. M. Thornton: *Prediction of protein-protein interaction sites using patch analysis.* J Mol Biol, 272(1):133–43, September 1997.

[49] Hamer, R., Q. Luo, J. P. Armitage, G. Reinert und Ch. M. Deane: *i-Patch: interprotein contact prediction using local network information.* Proteins, 78(13):2781–97, Oktober 2010.

[50] Lijnzaad, P. und P. Argos: *Hydrophobic patches on protein subunit interfaces: characteristics and prediction.* Proteins, 28(3):333–43, Juli 1997.

[51] Capra, J. A. und M. Singh: *Predicting functionally important residues from sequence conservation.* Bioinformatics, 23(15):1875–82, August 2007.

[52] Caffrey, D. R., S. Somaroo, J. D. Hughes, J. Mintseris und E. S. Huang: *Are protein-protein interfaces more conserved in sequence than the rest of the*

protein surface? Protein Sci, 13(1):190–202, Januar 2004.

[53] Guharoy, M. und P. Chakrabarti: *Conservation and relative importance of residues across protein-protein interfaces.* Proc Natl Acad Sci U S A, 102(43):15447–52, Oktober 2005.

[54] Thomas, J., N. Ramakrishnan und C. Bailey-Kellogg: *Graphical models of protein-protein interaction specificity from correlated mutations and interaction data.* Proteins, 76(4):911–29, September 2009.

[55] Düren, T.: *Calculating the accessible surface area*, März 2011. http://www.see.ed.ac.uk/~tduren/research/surface_area/.

[56] Richards, F. M.: *Areas, volumes, packing and protein structure.* Annu Rev Biophys Bioeng, 6:151–76, 1977.

[57] M. Gerstein, F. M. Richards: *Protein geometry: Volumes, areas and distances.* International Tables for Crystallography, 22:531–539, 2001.

[58] Connolly, M. L.: *Analytical molecular surface calculation.* J Appl Cryst, 16(4):548–558, 1983.

[59] Lee, B. und F. M. Richards: *The interpretation of protein structures: estimation of static accessibility.* J Mol Biol, 55(3):379–400, Februar 1971.

[60] Raychaudhuri: *Computational text analysis for functional genomics and bioinformatics.* Oxford University Press, Oxford, UK, 1, 2006.

[61] Saha, R. P., R. P. Bahadur und P. Chakrabarti: *Interresidue contacts in proteins and protein-protein interfaces and their use in characterizing the homodimeric interface.* J Proteome Res, 4(5):1600–9, 2005.

[62] Dong, Q., X. Wang, L. Lin und Y. Guan: *Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins.* BMC Bioinformatics, 8:147, 2007.

[63] Kawashima, S., H. Ogata und M. Kanehisa: *AAindex: Amino Acid Index Database.* Nucleic Acids Res, 27(1):368–9, Januar 1999.

[64] Liu, Q. und J. Li: *Propensity vectors of low-ASA residue pairs in the distinction of protein interactions.* Proteins, 78(3):589–602, Februar 2010.

[65] Chothia, C. und J. Janin: *Principles of protein-protein recognition.* Nature, 256(5520):705–8, August 1975.

[66] Heringa, J. und P. Argos: *Side-chain clusters in protein structures and their role in protein folding.* J Mol Biol, 220(1):151–71, Juli 1991.

[67] Tisi, L. C. und P. A. Evans: *Conserved structural features on protein surfaces: small exterior hydrophobic clusters.* J Mol Biol, 249(2):251–8, Juni 1995.

[68] Meador, W. E., A. R. Means und F. A. Quiocho: *Target enzyme recognition by calmodulin: 2.4 A structure of a calmodulin-peptide complex.* Science, 257(5074):1251–5, August 1992.

[69] Marina, A., P. M. Alzari, J. Bravo, M. Uriarte, B. Barcelona, I. Fita und V. Rubio: *Carbamate kinase: New structural machinery for making carbamoyl phosphate, the common precursor of pyrimidines and arginine.* Protein Sci, 8(4):934–40, April 1999.

[70] Lijnzaad, P.: *Hydrophobic patches on protein surface.* Doktorarbeit, Utrecht University, 2007.

[71] Karlin, S. und L. Brocchieri: *Evolutionary conservation of RecA genes in relation to protein structure and function.* J Bacteriol, 178(7):1881–94, April 1996.

[72] Schueler-Furman, O. und D. Baker: *Conserved residue clustering and protein structure prediction.* Proteins, 52(2):225–35, August 2003.

[73] Valdar, W. S. und J. M. Thornton: *Conservation helps to identify biologically relevant crystal contacts.* J Mol Biol, 313(2):399–416, Oktober 2001.

[74] Liang, S., C. Zhang, S. L. und Y. Zhou: *Protein binding site prediction using an empirical scoring function.* Nucleic Acids Res, 34(13):3698–707, 2006.

[75] Magliery, T. J. und L. Regan: *Sequence variation in ligand binding sites in proteins.* BMC Bioinformatics, 6:240, 2005.

[76] Hannenhalli, S. S. und R. B. Russell: *Analysis and prediction of functional sub-types from protein sequence alignments.* J Mol Biol, 303(1):61–76, Oktober 2000.

[77] Kalinina, O. V., A. A. Mironov, M. S. Gelfand und A. B. Rakhmaninova: *Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families.* Protein Sci, 13(2):443–56, Februar 2004.

[78] Lichtarge, O., H. R. Bourne und F. E. Cohen: *An evolutionary trace method defines binding surfaces common to protein families.* J Mol Biol, 257(2):342–58, März 1996.

[79] Landau, M., I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, T. Pupko und N. Ben-Tal: *ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures.* Nucleic Acids Res, 33(Web Server issue):W299–302, Juli 2005.

[80] Panchenko, A. R, F. Kondrashov und S. Bryant: *Prediction of functional sites by analysis of sequence and structure conservation.* Protein Sci, 13(4):884–

92, April 2004.

[81] Grishin, N. V. und M. A. Phillips: *The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences.* Protein Sci, 3(12):2455–8, Dezember 1994.

[82] Ouzounis, C., C. Pérez-Irratxeta, C. Sander und A. Valencia: *Are binding residues conserved?* Pac Symp Biocomput, Seiten 401–12, 1998.

[83] Valdar, W. S. und J. M. Thornton: *Protein-protein interfaces: analysis of amino acid conservation in homodimers.* Proteins, 42(1):108–24, Januar 2001.

[84] Bradford, J. R. und D. R. Westhead: *Asymmetric mutation rates at enzyme-inhibitor interfaces: implications for the protein-protein docking problem.* Protein Sci, 12(9):2099–103, September 2003.

[85] de Vries, S. J., A. D. J. van Dijk und A. M. J. J. Bonvin: *WHISCY: what information does surface conservation yield? Application to data-driven docking.* Proteins, 63(3):479–89, Mai 2006.

[86] Shackelford, G. und K. Karplus: *Contact prediction using mutual information and neural nets.* Proteins, 69 Suppl 8:159–64, 2007.

[87] Weigt, M., R. A. White, H. S., J. A. Hoch und T. Hwa: *Identification of direct residue contacts in protein-protein interaction by message passing.* Proc Natl Acad Sci U S A, 106(1):67–72, Januar 2009.

[88] Merkl, R. und M. Zwick: *H2r: identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments.* BMC Bioinformatics, 9:151, 2008.

[89] Fariselli, P., O. Olmea, A. Valencia und R. Casadio: *Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations.* Proteins, Suppl 5:157–62, 2001.

[90] Pazos, F. und A. Valencia: *In silico two-hybrid system for the selection of physically interacting protein pairs.* Proteins, 47(2):219–27, Mai 2002.

[91] Pazos, F., M. Helmer-Citterich, G. Ausiello und A. Valencia: *Correlated mutations contain information about protein-protein interaction.* J Mol Biol, 271(4):511–23, August 1997.

[92] Halperin, I., H. Wolfson und R. Nussinov: *Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families.* Proteins, 63(4):832–45, Juni 2006.

[93] Yeang, C.-H. und D. Haussler: *Detecting coevolution in and among protein domains.* PLoS Comput Biol, 3(11):e211, November 2007.

[94] Armon, A., D. Graur und N. Ben-Tal: *ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information.* J Mol Biol, 307(1):447–63, März 2001.

[95] La, D., B. Sutch und D. R. Livesay: *Predicting protein functional sites with phylogenetic motifs.* Proteins, 58(2):309–20, Februar 2005.

[96] Liu, B., X. Wang, L. Lin, B. T., Q. Dong und X. Wang: *Prediction of protein binding sites in protein structures using hidden Markov support vector machine.* BMC Bioinformatics, 10:381, 2009.

[97] Bradford, J. R., C. J. Needham, A. J. Bulpitt und D. R. Westhead: *Insights into protein-protein interfaces using a Bayesian network prediction method.* J Mol Biol, 362(2):365–86, September 2006.

[98] Rose, P. W., B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, J. D. Westbrook, J. Young, B. Yukich, C. Zardecki, H. M. Berman und P. E. Bourne: *The RCSB Protein Data Bank: redesigned web site and web services.* Nucleic Acids Res, 39(Database issue):D392–401, Januar 2011.

[99] Mintz, S., A. Shulman-Peleg, H. J. Wolfson und R. Nussinov: *Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions.* Proteins, 61(1):6–20, Oktober 2005.

[100] Altschul, S. F., W. Gish, W. Miller, E. W. Myers und D. J. Lipman: *Basic local alignment search tool.* J Mol Biol, 215(3):403–10, Oktober 1990.

[101] Hwang, H., T. Vreven, J. Janin und Z. Weng: *Protein-protein docking benchmark version 4.0.* Proteins, 78(15):3111–4, November 2010.

[102] Murzin, A. G., 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(4):536–40, April 1995.

[103] Hildebrandt, A., A. K. Dehof, A. Rurainski, A. Bertsch, M. Schumann, N. C. Toussaint, A. Moll, D. Stöckel, S. Nickels, S. C. Mueller, H.-P. Lenhof und O. Kohlbacher: *BALL–biochemical algorithms library 1.3.* BMC Bioinformatics, 11:531, 2010.

[104] Dodge, C., R. Schneider und C. Sander: *The HSSP database of protein structure-sequence alignments and family profiles.* Nucleic Acids Res, 26(1):313–5, Januar 1998.

[105] Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller und D. J. Lipman: *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 25(17):3389–402,

September 1997.

[106] Edgar, R. C.: *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 32(5):1792–7, 2004.

[107] Sander, C. und R. Schneider: *Database of homology-derived protein structures and the structural meaning of sequence alignment.* Proteins, 9(1):56–68, 1991.

[108] Kalinina, O. V., P. S. Novichkov, A. A. Mironov, M. S. Gelfand und A. B. Rakhmaninova: *SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins.* Nucleic Acids Res, 32(Web Server issue):W424–8, Juli 2004.

[109] Henikoff, S. und J. G. Henikoff: *Performance evaluation of amino acid substitution matrices.* Proteins, 17(1):49–61, September 1993.

[110] Wang, K. und R. Samudrala: *Incorporating background frequency improves entropy-based residue conservation measures.* BMC Bioinformatics, 7:385, 2006.

[111] Consortium, Uniprot: *The Universal Protein Resource (UniProt) in 2010.* Nucleic Acids Res, 38(Database issue):D142–8, Januar 2010.

[112] Taylor, W. R.: *The classification of amino acid conservation.* J Theor Biol, 119(2):205–18, März 1986.

[113] Zvelebil, M. J., G. J. Barton, W. R. Taylor und M. J. Sternberg: *Prediction of protein secondary structure and active sites using the alignment of homologous sequences.* J Mol Biol, 195(4):957–61, Juni 1987.

[114] Pilpel, Y. und D. Lancet: *The variable and conserved interfaces of modeled olfactory receptor proteins.* Protein Sci, 8(5):969–77, Mai 1999.

[115] Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin und D. G. Higgins: *The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.* Nucleic Acids Res, 25(24):4876–82, Dezember 1997.

[116] Liu, X., J. Li, W. Guo und W. Wang: *A new method for quantifying residue conservation and its applications to the protein folding nucleus.* Biochem Biophys Res Commun, 351(4):1031–6, Dezember 2006.

[117] Liu, X.-S. und W.-L. Guo: *Robustness of the residue conservation score reflecting both frequencies and physicochemistries.* Amino Acids, 34(4):643–52, Mai 2008.

[118] Henikoff, S. und J. G. Henikoff: *Position-based sequence weights.* J Mol Biol, 243(4):574–8, November 1994.

[119] Göbel, U., C. Sander, R. Schneider und A. Valencia: *Correlated mutations*

and residue contacts in proteins. Proteins, 18(4):309–17, April 1994.

[120] Larson, S. M., A. A. Di Nardo und A. R. Davidson: *Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions.* J Mol Biol, 303(3):433–46, Oktober 2000.

[121] Korber, B. T., R. M. Farber, D. H. Wolpert und A. S. Lapedes: *Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis.* Proc Natl Acad Sci U S A, 90(15):7176–80, August 1993.

[122] Atchley, W. R., K. R. Wollenberg, W. M. Fitch, W. Terhalle und A. W. Dress: *Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis.* Mol Biol Evol, 17(1):164–78, Januar 2000.

[123] Martin, L. C., G. B. Gloor, S. D. Dunn und L. M. Wahl: *Using information theory to search for co-evolving residues in proteins.* Bioinformatics, 21(22):4116–24, November 2005.

[124] Olmea, O. und A. Valencia: *Improving contact predictions by the combination of correlated mutations and other sources of sequence information.* Fold Des, 2(3):S25–32, 1997.

[125] McLachlan, A. D.: *Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551 .* J Mol Biol, 61(2):409–24, Oktober 1971.

[126] Henikoff, S. und J. G. Henikoff: *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 89(22):10915–9, November 1992.

[127] Press, W. H., S. A. Teukolsky, W. T. Vetterling und B. P. Flannery: *Numerical Recipies in C.* Cambridge University Press, 1992.

[128] Lijnzaad, P., P. Argos, C. Sander, M. Scharf und F. Eisenhaber: *The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies.* J Comput Chem, 16:273–284, 1995.

[129] Shrake, A. und J. A. Rupley: *Environment and exposure to solvent of protein atoms. Lysozyme and insulin.* J Mol Biol, 79(2):351–71, September 1973.

[130] Miller, S., J. Janin, A. M. Lesk und C. Chothia: *Interior and surface of monomeric proteins.* J Mol Biol, 196(3):641–56, August 1987.

[131] Bolser, D. M.: *The surfaces involved in the formation of protein complexes.* Doktorarbeit, University of Cambridge, 2007.

[132] Chothia, C.: *The nature of the accessible and buried surfaces in proteins.* J Mol Biol, 105(1):1–12, Juli 1976.

[133] Sanner, M. F., A. J. Olson und J. C. Spehner: *Reduced surface: an efficient way to compute molecular surfaces.* Biopolymers, 38(3):305–20, März 1996.

[134] Staudigel, M. und T. Trenner: *Analyse der Struktur von Proteinkontaktflächen mit Methoden der algorithmischen Geometrie.* Diplomarbeit, Fernuniversität in Hagen, November 2009.

[135] Cazals, F.: *Intervor*, August 2009. http://cgal.inria.fr/abs/Intervor/.

[136] Lijnzaad, P., H. J. Berendsen und P. Argos: *A method for detecting hydrophobic patches on protein surfaces.* Proteins, 26(2):192–203, Oktober 1996.

[137] Bittkowski, M.: *Bioinformatische Analyse hydrophober Patches auf der Proteinoberfläche.* Diplomarbeit, Fernuniversität in Hagen, April 2010.

[138] List, F.: *Die Imidazolglycerinphosphat-Synthase aus Thermotoga maritima: Struktur, Regulation und Evolution einer Glutaminamidotransferase.* Doktorarbeit, Universität Regensburg, Dezember 2009.

[139] Eisenhaber, F. und P. Argos: *Hydrophobic regions on protein surfaces: definition based on hydration shell structure and a quick method for their computation.* Protein Eng, 9(12):1121–33, Dezember 1996.

[140] Chang, C.-C. und C.-J. Lin: *LIBSVM: a library for support vector machines*, 2001. Software available at url http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[141] Boser, B. E., I. M. Guyon und V. N. Vapnik: *A training algorithm for optimal margin classifiers.* In: *Proceedings of the Workshop on Computational Learning Theory*, 1992.

[142] Cortes, C. und V. Vapnik: *Support-vector network.* Machine Learning, 20:273–297, 1995.

[143] T.-F. Wu, C.-J. Lin: *Probability Estimates for Multi-class Classification by Pairwise Coupling.* Journal of Machine Leaning Research, 5:975–1005, Mai 2004.

[144] Davis, J. und M. Goadrich: *The Relationship Between Precision-Recall and ROC Curves.* Proceedings of the 23rd International Conference on Machine Learning (ICML), 2006.

[145] Matthews, B. W.: *Comparison of the predicted and observed secondary structure of T4 phage lysozyme.* Biochim Biophys Acta, 405(2):442–51, Oktober 1975.

[146] Döring, A., D. Weese, T. Rausch und K. Reinert: *SeqAn an efficient, generic C++ library for sequence analysis.* BMC Bioinformatics, 9:11, 2008.

[147] de Hoon, M., S. Imoto und S. Miyano: *The C Clustering LIbrary.* Institute of Medical Science, University of Tokio, 2002.

[148] *The PyMol Molecular Graphics System, Version 1.1r2pre, Schrödinger, LLC.*

http://pymol.org.

[149] Schreiber, G.: *ProMate - Predicting the location of potential protein-protein binding sites for unbound proteins*, Januar 2011. http://bioinfo.weizmann.ac.il/promate/.

[150] Porollo, Aleksey und Jarosław Meller: *SPPIDER - Solvent accessibility based Protein-Protein Interface iDEntification and Recognition*, Januar 2011. http://sppider.cchmc.org/.

[151] Headd, J. J., Y. E. A. Ban, P. Brown, H. Edelsbrunner, M. Vaidya und J. Rudolph: *Protein-protein interfaces: properties, preferences, and projections.* J Proteome Res, 6(7):2576–86, Juli 2007.

[152] Y. A. Ban, H. Edelsbrunner J. Rudolph: *Interface surfaces for protein-protein complexe.* J ACM„ 53(3):361–378, 2006.

[153] Ray, N., X. Cavin, J.-C. Paul und B. Maigret: *Intersurf: dynamic interface between proteins.* J Mol Graph Model, 23(4):347–54, Januar 2005.

[154] Dupuis, F., J.-F. Sadoc, R. Jullien, B. Angelov und J.-P. Mornon: *Voro3D: 3D Voronoi tessellations applied to protein structures.* Bioinformatics, 21(8):1715–6, April 2005.

[155] Cazals, F., F. Proust, R. P. Bahadur und J. Janin: *Revisiting the Voronoi description of protein-protein interfaces.* Protein Sci, 15(9):2082–92, September 2006.

[156] Zhou, H. X. und Y. Shan: *Prediction of protein interaction sites from sequence profile and residue neighbor list.* Proteins, 44(3):336–43, August 2001.

[157] Fariselli, P., F. Pazos, A. Valencia und R. Casadio: *Prediction of protein–protein interaction sites in heterocomplexes with neural networks.* Eur J Biochem, 269(5):1356–61, März 2002.

[158] Stevens, J. M., R. N. Armstrong und H. W. Dirr: *Electrostatic interactions affecting the active site of class sigma glutathione S-transferase.* Biochem J, 347 Pt 1:193–7, April 2000.

[159] Sheinerman, F. B., R. Norel und B. Honig: *Electrostatic aspects of protein-protein interactions.* Curr Opin Struct Biol, 10(2):153–9, April 2000.

[160] Xu, D., S. L. Lin und R. Nussinov: *Protein binding versus protein folding: the role of hydrophilic bridges in protein associations.* J Mol Biol, 265(1):68–84, Januar 1997.

[161] Xu, D., C. J. Tsai und R. Nussinov: *Hydrogen bonds and salt bridges across protein-protein interfaces.* Protein Eng, 10(9):999–1012, September 1997.

[162] Vijayakumar, M., K. Y. Wong, G. Schreiber, A. R. Fersht, A. Szabo und H. X. Zhou: *Electrostatic enhancement of diffusion-controlled protein-protein association: comparison of theory and experiment on barnase and barstar.* J Mol Biol, 278(5):1015–24, Mai 1998.

[163] Camacho, C. J., Z. Weng, S. Vajda und C. DeLisi: *Free energy landscapes of encounter complexes in protein-protein association.* Biophys J, 76(3):1166–78, März 1999.

[164] Saha, R. P., R. P. Bahadur, A. Pal, S. Mandal und P. Chakrabarti: *ProFace: a server for the analysis of the physicochemical features of protein-protein interfaces.* BMC Struct Biol, 6:11, 2006.

[165] Bahadur, R. P., P. Chakrabarti, F. Rodier und J. Janin: *Dissecting subunit interfaces in homodimeric proteins.* Proteins, 53(3):708–19, November 2003.

[166] Rost, B und C Sander: *Prediction of protein secondary structure at better than 70% accuracy.* J Mol Biol, 232(2):584–99, Juli 1993.

[167] Rost, B: *PHD: predicting one-dimensional protein structure by profile-based neural networks.* Methods Enzymol, 266:525–39, 1996.

[168] Tuncbag, N., A. Gursoy und O. Keskin: *Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy.* Bioinformatics, 25(12):1513–20, Juni 2009.

[169] Clackson, T. und J. A. Wells: *A hot spot of binding energy in a hormone-receptor interface.* Science, 267(5196):383–6, Januar 1995.

[170] Cheng, J. und P. Baldi: *Improved residue contact prediction using support vector machines and a large feature set.* BMC Bioinformatics, 8:113, 2007.

[171] Hoskins, J., S. Lovell und T. L. Blundell: *An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements.* Protein Sci, 15(5):1017–29, Mai 2006.

[172] Bernauer, J., R. P. Bahadur, F. Rodier, J. Janin und A. Poupon: *DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions.* Bioinformatics, 24(5):652–8, März 2008.

[173] Janin, J., S. Miller und C. Chothia: *Surface, subunit interfaces and interior of oligomeric proteins.* J Mol Biol, 204(1):155–64, November 1988.

[174] Horton, N. und M. Lewis: *Calculation of the free energy of association for protein complexes.* Protein Sci, 1(1):169–81, Januar 1992.

[175] Fernandez-Recio, J., M. Totrov, C. Skorodumov und R. Abagyan: *Optimal docking area: a new method for predicting protein-protein interaction sites.* Proteins, 58(1):134–43, Januar 2005.

[176] Cutruzzolà, F., M. Arese, G. Ranghino, G. van Pouderoyen, G. Canters und M. Brunori: *Pseudomonas aeruginosa cytochrome C(551): probing the role of the hydrophobic patch in electron transfer.* J Inorg Biochem, 88(3-4):353–61, Februar 2002.

[177] Porter, S. W., Q. Xu und A. H. West: *Ssk1p response regulator binding surface on histidine-containing phosphotransfer protein Ypd1p.* Eukaryot Cell, 2(1):27–33, Februar 2003.

[178] Cheng, Z., Y. Liu, C. Wang, R. Parker und H. Song: *Crystal structure of Ski8p, a WD-repeat protein with dual roles in mRNA metabolism and meiotic recombination.* Protein Sci, 13(10):2673–84, Oktober 2004.

[179] Young, L., R. L. Jernigan und D. G. Covell: *A role for surface hydrophobicity in protein-protein recognition.* Protein Sci, 3(5):717–29, Mai 1994.

[180] Valdar, W. S. J.: *Scoring residue conservation.* Proteins, 48(2):227–41, August 2002.

[181] Shannon, C. E.: *A mathematical theory of communication.* Bell Szs Tech J, 27:623–656, 1948.

[182] Barabási, A.-L. und Z. N. Oltvai: *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 5(2):101–13, Februar 2004.

[183] Furey, T S, N Cristianini, N Duffy, D W Bednarski, M Schummer und D Haussler: *Support vector machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 16(10):906–14, Oktober 2000.

[184] Schölkopf, B. und A. J. Smola: *Learning with Kernels.* MIT Press, Cambridge, Massachusetts, 2002.

[185] Chung, J.-L., W. Wang und P. E. Bourne: *Exploiting sequence and structure homologs to identify protein-protein binding sites.* Proteins, 62(3):630–40, März 2006.

[186] Guldan, H.: *Charakterisierung der sn-Glycerin-1-Phosphat-abhängigen Enzyme AraM und PcrB aus Bacillus subtilis.* Diplomarbeit, Januar 2007.

[187] Badger, J., J. M. Sauder, J. M. Adams, S. Antonysamy, K. Bain, M. G. Bergseid, S. G. Buchanan, M. D. Buchanan, Y. Batiyenko, J. A. Christopher, S. Emtage, A. Eroshkina, I. Feil, E. B. Furlong, K. S. Gajiwala, X. Gao, D. He, J. Hendle, A. Huber, K. Hoda, P. Kearins, C. Kissinger, B. Laubert, H. A. Lewis, J. Lin, K. Loomis, D. Lorimer, G. Louie, M. Maletic, C. D. Marsh, I. Miller, J. Molinari, H. J. Muller-Dieckmann, J. M. Newman, B. W. Noland, B. Pagarigan, F. Park, T. S.

Peat, K. W. Post, S. Radojicic, A. Ramos, R. Romero, M. E. Rutter, W. E. Sanderson, K. D. Schwinn, J. Tresser, J. Winhoven, T. A. Wright, L. Wu, J. Xu und T. J. R. Harris: *Structural analysis of a set of proteins resulting from a bacterial genomics project.* Proteins, 60(4):787–96, September 2005.

[188] Guldan, H.: *DocGuldanNachweis Archaea-typischer Lipide in Bacteria über die Aufklärung der Funktion von AraM und PcrB aus Bacillis subtilis.* Doktorarbeit, Universität Regensburg, September 2010.

[189] Peterhoff, D.: *Einbau nichtnatürlicher Aminosäuren zur Untersuchung der Dimerisierung von PcrB.* Diplomarbeit, Universität Regensburg, Dezember 2009.

[190] McCoy, A. J., V. Chandana Epa und P. M. Colman: *Electrostatic complementarity at protein/protein interfaces.* J Mol Biol, 268(2):570–84, Mai 1997.

[191] Gabdoulline, R. R. und R. C. Wade: *On the protein-protein diffusional encounter complex.* J Mol Recognit, 12(4):226–34, 1999.

[192] Gruber, J., A. Zawaira, R. Saunders, C. P. Barrett und M. E. M. Noble: *Computational analyses of the surface properties of protein-protein interfaces.* Acta Crystallogr D Biol Crystallogr, 63(Pt 1):50–7, Januar 2007.

[193] Lijnzaad, P., K. Anton Feenstra, J. Heringa und F. C. P. Holstege: *On defining the dynamics of hydrophobic patches on protein surfaces.* Proteins, 72(1):105–14, Juli 2008.

[194] Dunn, S. D., L. M. Wahl und G. B. Gloor: *Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.* Bioinformatics, 24(3):333–40, Februar 2008.

[195] Fromer, M. und M. Linial: *Exposing the co-adaptive potential of protein-protein interfaces through computational sequence design.* Bioinformatics, 26(18):2266–72, September 2010.

[196] Pazos, F. und A. Valencia: *Protein co-evolution, co-adaptation and interactions.* EMBO J, 27(20):2648–55, Oktober 2008.

[197] Mintseris, J. und Z. Weng: *Structure, function, and evolution of transient and obligate protein-protein interactions.* Proc Natl Acad Sci U S A, 102(31):10930–5, August 2005.

[198] Ezkurdia, I., L. Bartoli, P. Fariselli, R. Casadio, A. Valencia und M. L. Tress: *Progress and challenges in predicting protein-protein interaction sites.* Brief Bioinform, 10(3):233–46, Mai 2009.

[199] Ma, B., T. Elkayam, H. Wolfson und R. Nussinov: *Protein-protein interactions: structurally conserved residues distinguish between binding sites and*

exposed protein surfaces. Proc Natl Acad Sci U S A, 100(10):5772–7, Mai 2003.

[200] Weston, J., S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio und V. Vapnik: *Feature selection for SVMs.* Seiten 668–674, 2000.

[201] Jones, S und J M Thornton: *Protein-protein interactions: a review of protein dimer structures.* Prog Biophys Mol Biol, 63(1):31–65, 1995.

[202] Miller, S: *The structure of interfaces between subunits of dimeric and tetrameric proteins.* Protein Eng, 3(2):77–83, November 1989.

[203] Bordner, Andrew J und Andrey A Gorin: *Comprehensive inventory of protein complexes in the Protein Data Bank from consistent classification of interfaces.* BMC Bioinformatics, 9:234, 2008.

[204] Dominguez, Cyril, Rolf Boelens und Alexandre M J J Bonvin: *HADDOCK: a protein-protein docking approach based on biochemical or biophysical information.* J Am Chem Soc, 125(7):1731–7, Februar 2003.