Optimierte atom- und aminosäurespezifische Gewichtungsfaktoren für Protein-Docking Methoden

I n a u g u r a l - D i s s e r t a t i o n

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Philipp Heuser

aus Münster

Köln, 2006

Berichterstatter:	Prof. Dr. D. Schomburg
	Prof. Dr. R. Schrader
Tag der mündlichen Prüfung:	12. Dezember 2006

I Abstract

Protein-protein docking is an *in silico* method to predict the three-dimensional structure of protein complexes. The complex structure is predicted from the known subunits structures. The detailed structural knowledge about the structure is of great scientific interest, because it is often the only way to discover the function of a complex. The existing experimental methods to determine protein structures are usually complicated and time consuming, thus a strong need for *in silico* docking-methods exists.

Present protein-protein docking methods are able to predict a near native structure for most complexes. Unfortunately this correct prediction is one among thousands. Therefore it was the aim of this work to develop a method to identify the near native structures. The commonly used representation of proteins for docking as a 3-D grid is extended by weighting factors being specific for residues and atoms represented by each cell of the grid. The values for these weighting factors are mathematically optimised for three different classes of complexes (enzyme-inhibitor, antibody-antigen and others).

The obtained optimised parameters comply with previously described properties of protein-protein interfaces. High weighting factors are optimised for aromatic, hydrophobic and rigid residues, whilst amino acids with long flexible side-chains obtain low factors. Since the number of freely rotatable bonds is correlated with the obtained weighting factors for enzyme-inhibitor complexes, the method is a computational rather cheap way to take the flexibility of amino acid side-chains on the surface of proteins at least partially into account, which is one of the major problems of unbound protein docking methods.

Sorting all proposed structures by the score obtained from the optimised weighting factors and by the comprehensive scoring function from Martin increases the portion of complexes for which a near native structure can be found among the top 10 structures by a factor of 10.

For more than 80% of all enzyme-inhibitor and antibody-antigen complexes a near-native structure can be found within the first percent of the prediction output. The top 8% of the prediction for ‘other’ complexes have to be searched until a near native solution for 80% of the evaluated complexes is found

II Kurzzusammenfassung

Protein-Protein *Docking* ist eine bioinformatische Methode, bei der aus zwei Protein Strukturen, von denen bekannt ist, dass sie *in vivo* aneinander binden, die dreidimensional Struktur des Komplexes mit atomarer Auflösung vorhergesagt wird. Die Kenntnis der Struktur ist von großem wissenschaftlichem Interesse, da oft nur so die genaue Funktion der jeweiligen Interaktion studiert werden kann. Komplexstrukturen können zwar meistens auch experimentell ermittelt werden, allerdings sind die bekannten Methoden oftmals extrem zeitaufwändig, so dass es einen beträchtlichen Bedarf an *in silico* Methoden zur Vorhersage von Proteinkomplexstrukturen gibt.

Aktuelle *Docking* Methoden können zwar für die meisten Komplexe eine der nativen Struktur ähnliche Struktur berechnen, allerdings ist diese richtige Vorhersage in der Regel eine von mehreren tausend als möglich vorhergesagten Strukturen. Das Ziel dieser Arbeit war daher die Entwicklung einer Methode zur Identifikation der richtigen, nahe nativen Struktur aus allen als möglich vorhergesagten Strukturen. Um dieses Ziel zu erreichen, wurde die - in den meisten *Docking* Programmen benutzte - Darstellung der Proteine als 3D-Gitter um Gewichtungsfaktoren erweitert, welche spezifisch für die jeweils repräsentierten Atome und Aminosäuren durch mathematische Methoden optimiert wurden. Die Optimierungen wurden für drei Komplexklassen (Enzym-Inhibitor, Antikörper-Antigen und ‚Andere') durchgeführt.

Die erhaltenen optimierten atom- und aminosäurespezifischen Gewichtungsfaktoren stehen z.T. in Einklang mit zuvor beobachteten Eigenschaften von Protein-Protein Bindungsstellen. So wurden z.B. hohe Gewichtungsfaktoren für aromatische, hydrophobe und rigide Residuen optimiert, während die Optimierung für Aminosäuren mit langen flexiblen Seitenketten eher niedrige Faktoren ergeben hat. Da die erhaltenen Gewichtungsfaktoren für Enzym-Inhibitor Komplexe in Relation zu der Anzahl frei drehbarer Seitenketten stehen, ist die in dieser Arbeit vorgestellte Methode eine äußerst effiziente Möglichkeit Flexibilität beim Docking zu berücksichtigen.

Durch eine Sortierung der potentiellen Komplexstrukturen nach der mit den Gewichtungsfaktoren berechneten geometrischen Korrelation in Kombination mit Martins umfassender Scoringfunktion kann der Anteil an Komplexen, für die eine nahe native Struktur unter den ersten 10 vorhergesagten Strukturen liegt, verzehnfacht werden.

Für 80% der Enzym-Inhibitor und der Antikörper-Antigen Komplexe wird nach Anwendung der Gewichtungsfaktoren mindestens eine nahe native Struktur im ersten Prozent der sortierten Vorhersage gefunden. Um für 80% der ‚Anderen' Komplexe mindestens eine nahe native Struktur zu finden, müssen die ersten 8% der Vorhersage durchsucht werden.

III Abkürzungen

Abkürzungen

2D/3D	Zwei- bzw. dreidimensional
Å	1 Ångstrøm= 10^{-10} m
AA	Antikörper-Antigen [Komplexe] (vgl.: Kapitel 3.1.1.2)
ACE	*Atomic Contact Energies* (ein Atom-Atom Paar Potential)
AS	Aminosäurespezifische Scores
AS*ATM	Produkt aus aminosäure- und atomspezifischen Scores
AS+ATM	Summe aus aminosäure- und atomspezifischen Scores
ATM	Atomspezifische Scores
CAPRI	*Critical Assessment of PRedicted Interactions*
CK	Ckordo (vgl.: Kapitel 4.1)
EI	Enzym-Inhibitor/Substrat [Komplexe] (vgl.: Kapitel 3.1.1.2)
FFT	Fast Fourier Transformation
I1	Inneres des größeren (ersten) Proteins bei Gitterrepräsentation (vgl.: Kapitel 4.1)
kDa	Kilo Dalton
MD	Molekulare Dynamik
NMR	kernmagnetische Resonanzspektroskopie (vgl.: Kapitel 3.1.4.1.2)
OTH	*Others* = ‚Andere' [Komplexe] (vgl.: Kapitel 3.1.1.2)
PDB	*Protein Data Bank* (www.pdb.org); zentrale Datenbank für Proteinstruktur Daten
RMS/RMSD	*Root mean square deviation* (vgl.: Kapitel 4.8); Maß für Ähnlichkeit zwischen (Protein-)Strukturen
RMSiCα	RMSD der Interface Cα Atome
SVM	Support Vector Machine
UUPPDD	*Unbound-Unbound Protein-Protein Docking Dataset*

Häufig genutzte englische Begriffe

Backbone	Strukturgerüst eines Proteins, bestehend aus den Cα-, C-, N- und O-Atomen der Aminosäuren ohne die Seitenketten. Hauptkette.

Bound	*Docking* von Proteinen, die aus einem Komplex ausgeschnitten wurden
Clash	Unnatürliche Kollision zwischen Atomen, die aus der Behandlung von Proteinstrukturen als starre Körper entstehen kann
Docking	Simulation der Assoziation von Makromolekülen.
Interface	Bindestelle zwischen zwei Makromolekülen.
Propensity	Neigung/Tendenz
Refinement	Verfeinerung von Komplexstrukturen (vgl.: Kapitel 3.1.3.2.2 und 3.1.4.2.1.4)
Rigid-body	Behandlung von Proteinen als rigide Körper im Gegensatz zu ihrer flexiblen Natur
Score	Wert, der die Qualität einer potentiellen Komplexstruktur beschreibt, berechnet durch Bewertungsfunktionen
Scoring/Ranking	Sortierung von möglichen Komplexstrukturen nach verschiedenen Bewertungsfunktionen
Target (CAPRI)	Zielstruktur: Vorherzusagende Struktur bei *CAPRI*
Unbound	*Docking* von Proteinen, deren Struktur in einem ungebundenen Zustand gelöst wurde

Aminosäuren

Alanin	ALA	A
Cystein	CYS	C
Aspartat	ASP	D
Glutamat	GLU	E
Phenylalanin	PHE	F
Glycin	GLY	G
Histidin	HIS	H
Isoleucin	ILE	I
Lysin	LYS	K
Leucin	LEU	L
Methionin	MET	M
Asparagin	ASN	N
Prolin	PRO	P
Glutamin	GLN	Q
Arginin	ARG	R
Serin	SER	S
Threonin	THR	T
Valin	VAL	V
Tryptophan	TRP	W
Tyrosin	TYR	Y

IV Inhaltsverzeichnis

1 Einleitung

Proteine sind einer der essentiellen Bausteine des Lebens. Nahezu jeder biologische Prozess ist von diesen abhängig. Die Funktionen von Proteinen sind vielfältig, so werden z.B. die meisten Reaktionen des Metabolismus durch Enzyme katalysiert, das Immunsystem basiert auf Proteinen, die Weiterleitung von biochemischen Signalen erfolgt oft durch Protein-Rezeptoren und komplexe Mechanismen werden durch Proteine reguliert, wie z.B. Transkription und Translation. Die Erfüllung dieser diversen Aufgaben ist häufig nur möglich, wenn verschiedene Proteine miteinander interagieren.

Für den Menschen sind z.B. 70.000 Proteininteraktionen von 6.200 Proteinen vorhergesagt worden[52]. Für einen großen Teil der vorhergesagten Interaktionen ist bis jetzt keine Funktion bekannt. Oftmals kann sich die Funktion eines Proteins bzw. eines Proteinkomplexes nur durch genaue Kenntnis der Struktur des Proteins vollständig aufklären lassen. Allerdings ist die experimentelle Strukturaufklärung von Proteinen ein komplizierter und zeitaufwändiger Prozess, so dass den 2.965.756 bekannten Proteinsequenzen (in UniProtKB/TrEMBL[3], Juni 2006) nur 37.269 bekannte Strukturen (PDB[8], Juni 2006) gegenüberstehen. Um trotz der Komplikationen bei der experimentellen Strukturaufklärung die dreidimensionalen Strukturen von Proteinen zu erhalten, wurden in den letzten Jahren verschiedene Methoden zur computergestützten Vorhersage von Proteinstrukturen entwickelt, mit welchen inzwischen erfolgreich Modelle mit nahezu atomarer Auflösung erstellt werden können.

Mit zunehmendem Erfolg der Strukturvorhersagemethoden nimmt auch das Bedürfnis zu, die Interaktionen der vorhergesagten Proteinstrukturen mit anderen Proteinen auf struktureller Ebene studieren zu können. Die Vorhersage von Komplexstrukturen, also wie zwei oder mehr Proteine miteinander interagieren, wird als Protein-Protein Docking bezeichnet. Die Entwicklung von Dockingmethoden ist natürlich nicht nur für Strukturmodelle von Bedeutung, sondern auch für Interaktionen von experimentell gelösten Strukturen. Diese sind z.T. als Monomere aber auch in anderen Komplexen gelöst worden. Da geschätzt wird, dass jedes Protein mit

durchschnittlich drei anderen Proteinen interagiert, können die im Komplex gelösten Strukturen auch in anderen vorherzusagenden Komplexen vorkommen.

Die Entwicklung von Dockingmethoden ist bislang primär von wissenschaftlichem Interesse, aber auch die pharmazeutischen Industrie entdeckt mehr und mehr das Potential dieser Methoden. Für das medizinische Interesse an Proteininteraktionen gibt es mindestens zwei verschiedene Gründe. Einerseits stehen viele Krankheiten in Verbindung mit Proteininteraktionen, so dass eine genaue Kenntnis der jeweiligen Interaktionen hilfreich sein kann, um die Ursachen für Krankheiten zu finden. Andererseits spielen Proteininteraktionen auch für den Einsatz und die Entwicklung von Medikamenten eine herausragende Rolle. Zum einen können Proteine selbst Medikamente sein und z.B. bestimmte Enzyme inhibieren und zum anderen kann es das Ziel einer medikamentösen Behandlung sein, spezifische Proteininteraktionen zu verhindern. Mit therapeutischen Proteinen und Antikörpern konnten bereits klinische Erfolge erzielt werden, allerdings primär bei extrazellulären Proteinen. Da Proteine in der Regel relativ groß sind, können sie nicht ohne weiteres in Zellen eindringen und sind ferner auch nicht oral als Medikament verabreichbar, da sie verdaut würden. Um intrazelluläre Proteinkomplexe inhibieren zu können, bieten sich kleine organische Moleküle an. Allerdings ist es schwierig, geeignete Bindungsstellen für kleine Moleküle zu identifizieren, da die zur Protein-Protein Interaktion genutzten Interface Regionen selten tiefe Spalten (*cavities*) anbieten, wie sie von kleinen Molekülen zur Bindung bevorzugt werden. Um dennoch erfolgreiche Angriffspunkte für Medikamente zu identifizieren, ist es zwingend erforderlich, die Interaktion zweier Proteine auf atomarem Level studieren zu können.[4,82]

Seit Beginn der 1990er Jahre wird weltweit intensiv an der Entwicklung von computergestützten Dockingmethoden gearbeitet, wobei insbesondere zu Beginn der Entwicklung nur wenige Datensätze zur Entwicklung und Evaluation der Methoden zur Verfügung standen. Selbst heute gibt es lediglich 83 nicht redundante Komplexe[66], für die sowohl die Strukturen der Untereinheiten im ungebundenen Zustand als auch im gebundenen Zustand bekannt sind.

Für die Entwicklung der ersten Dockingmethoden wurden bekannte Komplexstrukturen in ihre Untereinheiten getrennt und anschließend wurde versucht,

die Struktur des Komplexes aus diesen Untereinheiten wieder vorherzusagen. Diesen Vorgang bezeichnet man als *Bound Docking*. Das Problem des *Bound Docking* lies sich relativ einfach lösen, indem man nach der Orientierung der beiden Untereinheiten zueinander gesucht hat, die am besten aneinander passen. Die Orientierung mit der größten geometrischen Passgenauigkeit entspricht in der Regel der nativen Struktur. Bezieht man allerdings die zu dockenden Untereinheiten nicht aus der Komplexstruktur, sondern löst die Strukturen unabhängig voneinander oder benutzt Modelle von Proteinstrukturen (*Unbound Docking*), dann reicht die geometrische Passgenauigkeit als alleiniges Kriterium nicht aus, um die Komplexstruktur vorherzusagen. Die Schwierigkeiten des *Unbound Dockings* sind in erster Linie durch die Flexibilität von Proteinen zu begründen. Proteine sind keine starren Körper, sondern ihre Oberflächenstruktur passt sich in vielen Fällen der Struktur des Bindungspartners an. Hierfür kommen einerseits sehr kleine Bewegungen der Seitenketten der verschiedenen Aminosäuren aber auch Bewegungen kleinerer (Loops) oder größerer Proteinbereiche (Domänen) in Frage.

Da die Simulation und Vorhersage der Flexibilität von Proteinstrukturen äußerst komplex und im Hinblick auf die benötigte Rechenzeit sehr aufwändig ist, werden die Proteine bei den meisten Dockingmethoden nach wie vor als starre Körper behandelt (*rigid-body Docking*) und es wird versucht, die Komplexstrukturen, die der nativen Struktur ähnlich sind, an Hand physikochemischer Eigenschaften zu identifizieren.

In der hier vorliegenden Arbeit wird durch die Einführung von optimierten aminosäurespezifischen Gewichtungsfaktoren, die unterschiedliche Bedeutung der 20 verschiedenen Aminosäuren bei der Proteininteraktion berücksichtigt. Durch die Optimierung dieser Faktoren können verschiedene physikochemische Eigenschaften sowie die Flexibilität der Aminosäuren in je einem Parameter zusammengefasst werden. Des Weiteren werden auch atomtypspezifische Gewichtungsfaktoren optimiert und deren Effektivität evaluiert.

In den folgenden Kapiteln werden zunächst die physikochemischen Eigenschaften von Proteinkomplexen und deren Bindungsstellen, die verschiedenen Dockingmethoden und die Schritte zur Entwicklung der Gewichtungsfaktoren

beschrieben. Daran anschließend werden die durch die Gewichtungsfaktoren erzielten Ergebnisse vorgestellt.

2 Ziel der Arbeit

Ziel der Arbeit ist es, durch Einführung und Optimierung von atom- und aminosäurespezifischen Gewichtungsfaktoren eine schnelle und effiziente Methode zu entwickeln, mit der sich die Vorhersagequalität von Dockingstudien erheblich steigern lässt. Mit diesen Gewichtungsfaktoren wird eine gewichtete geometrische Korrelation berechnet. Durch eine Optimierung der Gewichtungsfaktoren sollen alle für Proteininteraktionen relevanten Eigenschaften der Aminosäuren bzw. der Atome in jeweils einem Parameter zusammengefasst werden. Die Optimierung der Parameter wird mit *unbound* Docking Beispielen durchgeführt, so dass durch die Gewichtungsfaktoren auch die Flexibilität der Aminosäuren indirekt berücksichtigt werden kann. Für die Entwicklung und Validierung dieser Gewichtungsfaktoren wird das gitterbasierte FFT-Docking Programm *ckordo* benutzt, welches in der Arbeitsgruppe entwickelt wurde.

Die Verbesserung der Vorhersagequalität soll sowohl in der Anzahl an Strukturen, die der nativen Struktur ähnlich sind, auf den vorderen Rängen einer Vorhersage als auch in Form des Ranges, auf dem die erste nahe native Struktur für jeden Komplex gefunden wird, sichtbar werden.

3 Biochemischer und bioinformatischer Hintergrund

3.1 Proteinkomplexe

3.1.1 Klassifizierungen von Proteinkomplexen

Protein-Protein Komplexe unterscheiden sich zum Teil erheblich in Bezug auf ihre Zusammensetzung, Funktion, Stabilität, Lebensdauer, evolutionäre Geschichte und ihren Bindungsmodus. Um Protein-Protein Interaktionen zu studieren bzw. vorherzusagen, ist eine Unterteilung in Komplexklassen hilfreich. Hierfür wurden diverse sich z.T. überlappende Klassifizierungsschemata entwickelt. Nooren und Thornton[71] haben verschiedene Klassifizierungseigenschaften in Bezug auf den Bindungsmodus beschrieben:

3.1.1.1 Klassifizierung nach Bindungsmodus

3.1.1.1.1 Homo- und Heterooligomere

Proteinkomplexe können sowohl aus identischen oder homologen Proteinen als auch aus verschiedenen Untereinheiten gebildet werden. Komplexe aus gleichen Untereinheiten werden als Homooligomere bezeichnet, während man bei verschiedenen Untereinheiten von Heterokomplexen spricht. Die Interaktion gleicher Proteine kann entweder isolog oder heterolog sein, wobei ersteres die Interaktion mit identischen Interfaces und letzteres die Interaktion mit verschiedenen Interfaces beschreibt[71].

3.1.1.1.2 Obligate und nicht obligate Komplexe

Ferner können Proteinkomplexe in obligate und nicht obligate Komplexe unterteilt werden, abhängig davon, ob die jeweiligen Interaktionspartner auch ohne den Bindungspartner stabil sind. Die meisten Heterokomplexe sind nicht obligate Komplexe, das heißt sie sind meistens auch unabhängig voneinander stabil, da sie

z.T. an verschiedenen Orten des Organismus synthetisiert werden. Beispiele für nicht obligate Heterokomplexe sind die Antikörper-Antigen, Enzym-Inhibitor oder die Rezeptor-Ligand Wechselwirkungen[71].

3.1.1.1.3 Transiente und permanente Komplexe

Die Unterteilung in transiente und permanente Komplexe beschreibt die Lebensdauer der jeweiligen Komplexe. Die Bindungen der transienten Komplexe sind *in vivo* häufig nur von kurzer Dauer und können mehrmals getrennt und neu gebildet werden. In der Regel sind obligate Komplexe auch permanent, da die Untereinheiten alleine nicht stabil sind und daher permanent aneinander gebunden sein müssen. Die nicht obligaten Komplexe hingegen können sowohl transient als auch permanent sein[71].

Die Interfaces der obligaten, permanenten Komplexe sind tendenziell größer, weniger planar und dichter gepackt als die der transienten Komplexe. Die transienten Komplexe zeichnen sich durch eine eher hydrophile Oberfläche aus, da diese auch alleine im wässrigen Medium existieren müssen[47].

3.1.1.2 Funktionelle Klassifizierung

Eine andere Art der Klassifizierung, die insbesondere für die Entwicklung von Dockingmethoden genutzt wird, bezieht sich auf die Funktion der Komplexe und basiert auf den evolutionären Unterschieden der Interaktionen. Die Komplexe werden hierbei in die Gruppen Enzym-Inhibitor/Substrat (EI), Antikörper-Antigen (AA) und ‚Andere' Komplexe (OTH)[18,66] eingeordnet. Die meisten der für *Unbound Docking* Studien genutzten Beispiele sind transiente, nicht obligate Komplexe. Diese Klassifizierung kann historisch begründet werden, da es lange Zeit nur einige Enzym-Inhibitor Komplexe, wenige Antikörper-Antigen und wenige andere Komplexe gab, für die sowohl die Komplexstruktur als auch die Strukturen der Untereinheiten im ungebundenen Zustand bekannt waren (vgl. Kapitel 4.5).

Allerdings ist diese Klassifizierung auch weiterhin sinnvoll, da diese verschiedenen Klassen sich im Hinblick auf ihre Bindung und evolutionäre Geschichte

unterscheiden. Während die Interaktionen von Enzymen mit ihrem Substrat bzw. Inhibitor oft hoch spezifisch sind, sich im Laufe der Evolution gemeinsam entwickelt und ihre Bindung dabei optimiert haben, müssen Antikörper die entsprechenden Antigene spontan erkennen. Antikörper müssen auch unbekannte Antigene binden können, um eine erfolgreiche Immunabwehr zu gewähren. Dies führt dazu, dass die Interfaces von Antikörper-Antigen Komplexen zu den am wenigsten dicht gepackten gehören[47] und sich außerdem von den Enzym-Inhibitor Komplexen darin unterscheiden, dass die Konserviertheit der Interfaceresiduen weitaus geringer ist[80].

Die ‚Anderen' Komplexe sind eine relativ heterogene Gruppe, bestehend aus den restlichen bekannten Komplexstrukturen, zu denen auch die Strukturen der ungebundenen Untereinheiten bekannt sind. Eine weitere Unterteilung der ‚Anderen' Komplexe kann sinnvoll sein, allerdings ist die Anzahl an Beispielen nach wie vor zu gering, um diese weitere Einteilung vorzunehmen, was aber in Zukunft möglich sein könnte.

3.1.2 Bindungsstellen (Interfaces)

Die Bindungsstellen, mit denen Proteine an andere Proteine binden, unterscheiden sich in ihren physikochemischen Eigenschaften von der restlichen Proteinoberfläche. Insbesondere die Häufigkeit, mit der manche Aminosäuren im Interface vorkommen, ist unterschiedlich.

Chakrabarti und Janin[16] haben die Bindungsstellen von Protein-Protein Komplexen durch geometrisches Clustering in verschiedene so genannte *Patches* unterteilt, wobei die Interfaces solcher Komplexe, bei denen das Interface eine Fläche von weniger als 2.000 Å² einnimmt, in der Regel aus nur einem *Patch* bestehen, während bei Proteinkomplexen mit größeren Bindungsstellen diese häufig aus mehreren *Patches* zusammengesetzt sind. Bei *Multi-Patch* Interfaces hat mindestens eines der *Patches* jedoch die Größe eines *Single-Patch* Interfaces. Bei jedem Interface *Patch* kann zwischen „*Core*" und „*Rim*", also Zentrum und Rand, unterschieden werden. Als *Core* wird der Teil der Bindungsstelle bezeichnet, der nach der Komplexbildung völlig verdeckt ist, während die Residuen, die auch nach der Komplexbildung noch

teilweise dem umgebenden Lösungsmittel zugänglich sind, als *Rim* beschrieben werden.

Im Schnitt verdeckt ein *Patch* 1.320 ± 520 Å² der Proteinoberfläche beider Proteine zusammen. Die *Patches* in *Single-Patch* Interfaces sind mit 1.560 ± 340 Å² durchschnittlich etwas größer als in *Multi-Patch* Interfaces. Insgesamt sind 170 ± 39 Atome an einem *Single-Patch* Interface beteiligt, bzw. 85 Atome je *Patch*. Ein Protein-Protein Interface setzt sich aus 57 ± 22 Residuen zusammen. Dem *Core* werden 53% und dem *Rim* 47% der Interface Residuen zugeordnet, während allerdings 72% der Interface Fläche von *Core* Residuen eingenommen werden und nur 28% von *Rim* Residuen[16].

Ferner wurde in den Kristallstrukturen der ungebundenen Proteine in den Interface Regionen im Schnitt ein niedrigerer B-Faktor (experimentelles Maß für die Flexibilität) gemessen[69,99] als auf dem Rest der Oberfläche und außerdem waren mehr Wassermoleküle an die Interface Residuen gebunden. Allerdings kann dieses Phänomen auch auf Kristallkontakte zurückzuführen sein[69].

3.1.2.1 Sekundärstruktur

Im Hinblick auf strukturelle Eigenschaften der Interface Regionen haben Neuvirth *et al.*[69] bei einer Analyse von transienten Komplexen festgestellt, dass in Interfaceregionen β-Faltblätter und lange Loops bevorzugt werden, während α-Helices eher auf der restlichen Oberfläche zu finden sind. Das erhöhte Auffinden von β-Faltblättern wird bei Neuvirth *et al.*[69] damit erklärt, dass β-Faltblätter eher eng gepackte Strukturen bilden können, wenn zwei Faltblätter direkt aneinander gelegt werden. In einer weiteren Studie mit 28 Homodimeren waren allerdings 53% der Kontaktflächen α-helical, 22% β-Faltblätter, 12% αβ, und der Rest Windungen ("*coils*")[93].

3.1.2.2 Komplementarität

Die Komplementarität zwischen zwei interagierenden Proteinen ist insbesondere für Computermethoden zur Vorhersage von Proteinkomplexstrukturen eine der

wichtigsten und wirksamsten Eigenschaften von Interaktionsflächen. Die Komplementarität zwischen den Bindungspartnern ist sowohl geometrischer als auch chemischer Natur. Insbesondere bei Enzym-Inhibitor Komplexen und stabilen Heterokomplexen findet sich eine hohe geometrische Oberflächenpassform (*fitting surface shape*), so dass im gebundenen Zustand das Volumen zwischen den beiden Bindungspartnern sehr gering ist, diese also sehr eng gepackt sind (*packing density*). Die Antikörper-Antigen Komplexe sind etwas weniger dicht gepackt, als die oben genannten Komplexe, so dass die geometrische Komplementarität etwas geringer ist[50]. Der geringere Grad an Komplementarität bei Antikörper-Antigen Komplexen kann dadurch erklärt werden, dass diese Bindungsstellen nicht über lange Zeiträume hinweg auf die jeweiligen Bindungspartner optimiert wurden, sondern auch auf neue unbekannte Antigene reagieren müssen[47,50].

Zwischen den Bindungspartnern besteht ferner auch eine chemische Komplementarität, im Hinblick auf geladene und ungeladene, auf polare und apolare Gruppen.

In einigen Fällen wird die Komplementarität zwischen Rezeptor und Ligand allerdings erst durch den Bindungsprozess hergestellt. Da insbesondere die chemische Komplementarität häufig von den geladenen Atomen in den langen Seitenketten von Arg, Lys oder Glu abhängig ist, können diese Atome im ungebundenen Zustand bis zu 10 Å von der Position entfernt sein, die sie im gebundenen Zustand einnehmen[44,47].

3.1.2.3 Aminosäurezusammensetzung

In der Literatur finden sich mehrere Studien, die die Aminosäurenzusammensetzung von Protein-Protein Bindungsstellen analysieren und mit der Aminosäurenzusammensetzung der restlichen Oberfläche vergleichen[16,37,46,56]. Die verschiedenen Untersuchungen unterscheiden sich z.T. bezüglich der angewandten Methodik und des untersuchten Datensatzes. Einer der Hauptunterschiede liegt darin, dass z.B. Huang und Schroeder[37] nur die Anzahl der Aminosäuren im Interface ausgewertet haben, während Jones[46], Chakrabarti[16] und Lo Conte[56] jeweils die

Flächen, die von der jeweiligen Aminosäure eingenommen werden, analysiert haben. Chakrabarti *et al*. haben zudem die Interface Regionen in *Core* und *Rim* unterteilt.

	Huang, Schröder 2005[37]		Chakrabarti 2002[16]		Lo Conte 1999[56]	Jones & Thornton. 1997[46]
	***Propensities* für Trypsin-like serine proteasen**	**Propensities für Antikörper**	**Zentrum (*Core*)**	**Rand (*Rim*)**	***Propensities* nach Fläche**	***Propensities* nach Fläche**
ALA	0,53	1,16	-0,40	-0,26	-0,43	-0,17
ARG	0,72	0,92	0,13	0,11	0,13	0,27
ASN	0,51	0,51	-0,14	0,03	-0,12	0,12
ASP	0,95	1,13	-0,46	-0,07	-0,31	-0,38
CYS	8,99	0,00	1,00	0,62	0,76	0,43
GLN	0,64	0,84	-0,34	-0,36	-0,36	-0,11
GLU	0,55	0,84	-0,80	0,02	-0,47	-0,13
GLY	0,97	0,69	-0,08	0,35	0,02	-0,07
HIS	2,07	1,01	0,84	0,23	0,64	0,41
ILE	0,88	0,86	0,71	0,38	0,56	0,44
LEU	1,05	0,77	0,34	0,25	0,29	0,40
LYS	0,52	0,75	-0,82	-0,20	-0,57	-0,36
MET	0,92	1,27	1,13	0,51	0,98	0,66
PHE	1,94	2,73	1,01	-0,60	0,79	0,82
PRO	0,69	1,32	-0,38	-0,22	-0,25	-0,25
SER	1,25	0,75	-0,56	-0,14	-0,42	-0,33
THR	0,94	0,61	-0,44	-0,21	-0,35	-0,18
TRP	4,18	1,98	1,41	0,21	1,25	0,83
TYR	1,50	3,55	1,22	0,50	1,04	0,66
VAL	1,07	0,91	0,08	0,11	0,09	0,27

Tabelle 3.1: Interface Propensities berechnet durch Huang, Chakrabarti, Lo Conte und Jones

Aus den ermittelten Häufigkeiten bzw. aus der jeweils eingenommenen Fläche wurden *Interface-Propensities* berechnet. Die *Interface-Propensity* beschreibt die Neigung einer bestimmten Aminosäure im Interface zu liegen, also wie hoch die Wahrscheinlichkeit ist, dass eine bestimmte Aminosäure im Interface ist, unter Berücksichtigung der generellen Häufigkeit dieser Aminosäure auf der Proteinoberfläche. Allen *Propensity* Skalen (s. Tabelle 3.1) ist gemeinsam, dass hohe

Werte für die aromatischen Residuen PHE, TYR und TRP sowie für MET berechnet wurden.

Im Hinblick auf die Häufigkeit hydrophober Aminosäuren im Interface finden sich in der Literatur unterschiedliche Angaben in Abhängigkeit davon, welche Art von Komplexen untersucht wurde. Während von Neuvirth *et al.* bei transienten Heterokomplexen keine erhöhte Häufigkeit für hydrophobe Residuen festgestellt wurde, so bilden die hydrophoben Residuen aber dennoch *Patches*[69]. Bordner *et al.* beschreiben hingegen eine Anreicherung großer, hydrophober und ungeladener, polarer Residuen im Interface, während geladene Aminosäuren eher selten zu finden sind[12] ohne hierbei Unterschiede zwischen Hetero- und Homodimeren festzustellen (abgesehen von Asparaginsäure und Glycin).

Die Aminosäuren Zusammensetzung der *Rim* Region ist der Zusammensetzung der restlichen Oberfläche des Proteins ähnlich, während bei der Zusammensetzung des Interface-*Cores* erhebliche Unterschiede zum Rest der Oberfläche festgestellt werden konnten. Der *Core* des Interfaces entspricht eher dem Inneren eines Proteins und zeichnet sich durch relativ viele aromatische und wenige geladene Residuen aus[16].

Neuvirth *et al.*[69] haben zusätzlich zur Analyse der Präferenz ganzer Residuen Teil des Interfaces zu sein auch die Präferenz einzelner Atome untersucht. Dabei haben nur die Atome der aromatischen Ringe von TRP, PHE und TYR eine deutlich höhere Neigung im Interface zu liegen gezeigt. Hingegen haben die Cα Atome von ALA, LYS, ASN, und SER eine höhere *Propensity* sich auf der restlichen Oberfläche zu befinden als im Interface. Dieses Ergebnis unterstreicht die Bedeutung der aromatischen Residuen für die Bindung zweier Proteine.

3.1.2.4 Bindungskräfte

Bereits 1940 haben Pauling und Delbrück postuliert, dass intermolekulare Interaktionen durch die Kräfte von van-der-Waals Anziehung und Abstoßung, durch Elektrostatik und durch die Bildung von Wasserstoffbrücken zu Stande kommen, sowie dass für stabile Interaktionen die Komplementarität der Oberflächen und der

aktiven Gruppen nötig ist[74]. Diese mehr als 65 Jahre zurück liegende Aufzählung der Kräfte ist nach wie vor gültig und wurde nur um den Hydrophoben Effekt ergänzt. Letzterer beschreibt den Energiegewinn, der durch das Zusammenbringen von nicht polaren Residuen in bzw. aus wässriger Umgebung erreicht wird. Dies wird als eine der treibenden Kräfte für die Stabilisierung von Proteinkomplexen beschrieben. In der Literatur finden sich unterschiedliche Angaben bezüglich der Quantifizierung dieser Kraft. So werden Werte von 25 über 47 bis zu 72 Kalorien je Å² beschrieben[39,47,88].

Elektrostatische Wechselwirkungen beschreiben die Interaktionen geladener Atome, wie sie in den Seitenketten einiger Aminosäuren vorkommen. Wenn sich in einem Proteinkomplex gegensätzlich geladenen Atome gegenüber stehen, können Salzbrücken (auch Ionenbindung genannt) ausgebildet werden. Salzbrücken sind generell nicht zwingend notwendig für die Interaktion zweier Proteine, können aber die Stabilität von Interaktionen erheblich vergrössern[47]. So kommen manche Komplexe ohne Salzbrücken aus, während in anderen Komplexen bis zu 5 intermolekulare Salzbrücken gefunden werden[93].

Wenn zwei Moleküle über ein H-Atom miteinander interagieren, so spricht man von Wasserstoffbrücken. Die Bindungsenergie von Wasserstoffbrücken liegt deutlich unter der von kovalenten Bindungen und von ionischen Wechselwirkungen. In Protein-Protein Interfaces findet sich je 170 Å² durchschnittlich eine Wasserstoffbrücke, also ca. 10 ± 5 H-Brücken je Interface[93]. Allerdings variiert die Anzahl von 0 - z.B. in *Uteroglobin* - bis zu 46 im *Variant Surface Glycoprotein*[47].

Ein weiterer Beitrag zur Interaktion zwischen zwei Proteinen wird durch die van-der-Waals Kräfte geleistet. Van-der-Waals Kräfte tauchen zwar auch zwischen Proteinen und umgebendem Medium auf, da allerdings aufgrund der größeren Packungsdichte in Protein-Protein Interfaces, zwischen zwei Proteinen mehr van-der-Waals Interaktionen stattfinden als zwischen Protein und Wasser, tragen diese auch zur Bindungsenergie bei[47,93].

Sehr selten kommt es auch zu Disulfidbrücken zwischen interagierenden Proteinen. Wenn sich allerdings eine solche kovalente Bindung ausbilden kann, dann spielt diese meistens auch eine wichtige Rolle bei der Interaktion[47].

Die Gesamtenergie einer Protein-Protein Bindung setzt sich aus den oben beschriebenen Kräften zusammen, wobei sich allerdings keine verallgemeinernde Regel bezüglich der Zusammensetzung der verschiedenen Kräfte ableiten lässt. Die Stabilität einer Interaktion wird durch die individuelle Kombination der unterschiedlichen Kräfte bestimmt.

3.1.3 Strukturaufklärung und –vorhersage von Proteinkomplexen

Für ein genaues Verständnis der Proteine und insbesondere der Interaktionen von Proteinen mit Proteinen und anderen Molekülen ist die Kenntnis der Struktur derselben in atomarer Auflösung von erheblichem Nutzen.

Zur Ermittlung der Strukturen gibt es mehrere experimentelle Methoden, die verschiedene Genauigkeiten erreichen, sich aber auch im Hinblick auf Komplexität, Zeitaufwand und Schwierigkeit unterscheiden. Einerseits gibt es die hochauflösenden Methoden, wie die unten beschriebene Röntgenstrahlen-kristallographie und die kernmagnetische Resonanzspektroskopie (NMR) und andererseits Methoden wie Elektronenmikroskopie und -tomographie, die die Strukturen nur in niedriger Auflösung abbilden können.

Da es insbesondere mit den hochauflösenden Methoden meistens extrem zeitaufwändig und z.T. sogar unmöglich ist, die Struktur eines Proteins bzw. eines Proteinkomplexes zu lösen, wird zur Zeit intensiv an bioinformatischen Methoden gearbeitet, die die Struktur eines Komplexes an Hand der Strukturen der Untereinheiten oder an Hand von Modellen der Untereinheiten vorhersagen können.

Ferner gibt es Ansätze experimentelle Methoden und *in silico* Methoden miteinander zu kombinieren, um so die Strukturen der Komplexe zu ermitteln. Einen detaillierten Überblick über die Methoden zur Strukturaufklärung gibt ein Review Artikel von Russell *et al.*[81] und eine Einführung in *data-driven* Docking, also Docking Methoden, welche auch experimentelle Ergebnisse berücksichtigen, ein Artikel von van Dijk *et al.*[96].

Im Folgenden werden die oben genannten experimentellen Methoden kurz vorgestellt und ein Überblick über die unterschiedlichen Dockingmethoden gegeben. Zu Beginn des Methoden Teils (Kapitel 4.1) werden die Fast Fourier Transformations (FFT) Methoden detaillierter betrachtet, da das für diese Arbeit verwendetet Docking-Programm *ckordo* auf dieser Methode basiert.

3.1.3.1 Experimentelle Methoden

3.1.3.1.1 Röntgenstrahlkristallographie (X-ray)

Die am häufigsten benutzte Methode zur experimentellen Bestimmung von Proteinstrukturen ist die Röntgenstrahlkristallographie. 85% der Strukturen in der PDB[8] wurden mit dieser Methode ermittelt. Die Röntgenstrahlkristallographie ermöglicht es, an Hand von Röntgenstrahlenbeugungsmustern Proteinstrukturen mit nahezu atomarer Auflösung zu messen.

Für die Röntgenstrahlkristallographie muss in einem ersten Schritt von dem zu untersuchenden Protein ein Kristall gezüchtet werden. Der Kristall wird dann mit Röntgenstrahlen beschossen, welche von den Elektronen des Proteins in einem spezifischen Muster abgelenkt werden. Anhand des Ablenkungsmusters kann in einem nachgeschalteten Schritt die genaue Position der Atome berechnet werden.

Mit dieser Methode können zwar die am höchsten aufgelösten Strukturen von Proteinen ermittelt werden, allerdings ist insbesondere der Kristallisationsschritt oftmals sehr zeitaufwändig und manchmal sogar unmöglich (z.B. bei sehr großen und bei hydrophoben Proteinen wie Membranproteinen). Ein weiterer Nachteil von Röntgenstrukturen ist, dass die Struktur nur eine Momentaufnahme darstellt, so dass die Flexibilität der Proteine nicht berücksichtigt werden kann.

3.1.3.1.2 NMR - kernmagnetische Resonanzspektroskopie

Die NMR-Spektroskopie beruht auf dem Prinzip, dass alle Atome einen Kernspin besitzen, welcher abhängig von dem umgebenden Magnetfeld ist. Der Kernspin der Atome kann durch einen Radiofrequenzimpuls beeinflusst werden. Die Änderung der

Kernspins kann gemessen werden und erlaubt es, die Struktur der Proteine zu berechnen.

Der große Vorteil der NMR-Spektroskopie liegt darin, dass Proteinstrukturen in Lösung ermittelt werden können und keine Kristalle gezüchtet werden müssen. Dies erlaubt ferner auch die Strukturen von schwachen (transienten) Komplexen zu messen und durch eine Abfolge von ‚Aufnahmen' ist es möglich, die Bewegungen der Proteine zu untersuchen. Lange Zeit konnten allerdings mittels NMR-Spektroskopie nur die Strukturen kleiner Proteine mit einer Länge bis zu 300 Aminosäuren ermittelt werden.

Mit neueren Varianten der NMR-Spektroskopie, wie *Transverse Relaxation Optimised Spectroscopy* (TROSY) oder *Chemical Shift Perturbations* (CSP), können inzwischen Strukturen von Proteine mit einem Gewicht von bis zu 50 kDa (ca. 420 Aminosäuren) ermittelt werden[10].

3.1.3.1.3 Elektronenmikroskopie und –tomographie

Für Moleküle mit einem Gewicht von mehr als 200-500 kDa kann mit *single-particle cryo*-Elektronenmikroskopie eine zweidimensionale Aufnahme mit einer Auflösung von bis zu 5 Å gemacht werden. Die dreidimensionale Struktur lässt sich dann aus mehreren 2D Aufnahmen rekonstruieren. Aktuelle Entwicklungen im Bereich der Elektronentomographie erlauben inzwischen Aufnahmen mit einer Auflösung von bis zu 2-5 nm[81].

Bilder in dieser Auflösung können bereits Aufschlüsse über die grobe Form von Proteinkomplexen geben und somit eindeutige Hinweise auf die relative Orientierung der bindenden Proteine zueinander geben und so die Identifikation der richtigen Struktur mit Docking-Programmen vereinfachen.

3.1.3.2 Protein-Protein Docking

Unter Protein-Protein Docking versteht man die Vorhersage der Interaktion zweier Proteine. In der hier vorgestellten Arbeit wird das so genannte 1:1 Docking

behandelt, bei welchem vorhergesagt wird, *wie* zwei Proteine miteinander interagieren, also wie sich beide Proteine im dreidimensionalen Raum (3D) zueinander ausrichten und aneinander binden. Im Gegensatz dazu gibt es noch das 1:n Docking, welches versucht, für ein Protein mögliche Bindungspartner zu finden.

Alle Docking-Ansätze basieren auf einer probaten Oberflächenrepräsentation, einer Suchstrategie und einer effizienten Bewertungsfunktion.

Zwei Verfahren beim Docking sind besonders zeitkritisch. Erstens die globalen Suchmethoden, da sechs Freiheitsgrade (drei der Rotation und drei der Translation) berücksichtigt werden müssen und zweitens die Größe der Reaktanden, da die Anzahl an möglichen Komplex-Konformationen mit der Größe steigt. Eine Beschleunigung kann an Hand vorheriger Kenntnis der Bindungsstelle erfolgen, da dann der Suchraum erheblich eingeschränkt wird. Für unbekannte Proteine ist die Bindungsstelle jedoch nicht verfügbar.

Einen allgemeinen Überblick über den aktuellen Stand der Forschung im Bereich Docking ist einem der Review-Artikel zu entnehmen, die in letzter Zeit erschienen sind[23,32,43,63,90]. Insbesondere sind hierbei die Artikel von Halperin[32] und von Eisenstein[23] hervorzuheben. Der Artikel von Halperin ist der wohl umfassendste Bericht über Docking, und Eisensteins Artikel ist auf das FFT-Docking im Besonderen fokussiert und somit für diese Arbeit von großem Interesse.

3.1.3.2.1 Allgemeiner Ablauf

Die meisten Dockingmethoden lassen sich in die im Folgenden beschriebenen und in Abbildung 3.1 dargestellten Schritte unterteilen. Zu Beginn muss eine Darstellungsform der Proteine gewählt werden, die eine effiziente Berechnung der Komplementarität der beiden Bindungspartner erlaubt. Die Berechnung der Komplementarität ist im Bezug auf die Rechenzeit oftmals der aufwändigste Schritt der Docking Prozedur, bei dem mehrere tausend bis mehrere zehntausend mögliche Strukturen generiert werden.

Da insbesondere beim *unbound* Docking die geometrische Passgenauigkeit als alleiniges Kriterium für die Detektion der nativen Struktur nicht ausreicht, wird in einem nachgeschalteten Schritt mit so genannten Postfiltern und Rerankingfunktionen versucht, diese an Hand von physikochemischen und biochemischen Eigenschaften zu identifizieren.

Da die Flexibilität der Proteine gar nicht oder nur geringfügig in die oben genannten Schritte einfließt, muss der Dockingprozedur ein Refinement-Schritt folgen, bei dem mit einem Energieminimierungsprogramm oder mit den Methoden der molekularen Dynamik mögliche Kollisionen entfernt werden und die Seitenkettenatome in die günstigste Konformation gebracht werden.

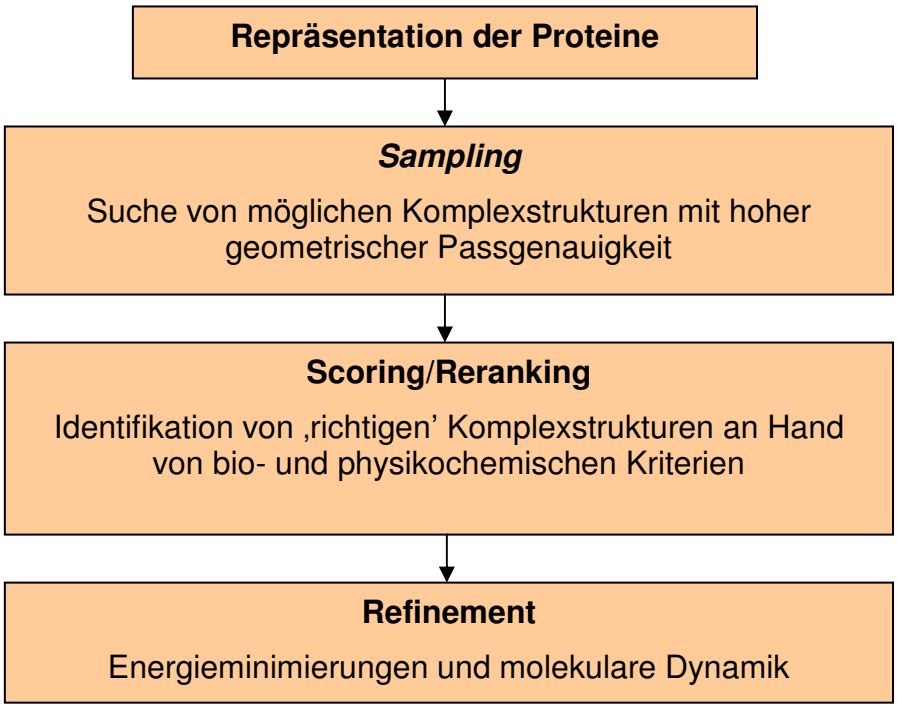

Abbildung 2.1: Allgemeiner Ablauf von Docking Prozeduren

3.1.3.2.1.1 Repräsentation der Proteinstrukturen

Die Wahl der Repräsentationsform der Proteinstrukturen stellt immer einen Kompromiss dar zwischen Auflösung und einer möglichst einfachen - für Computer effizient zu behandelnden - Darstellung. Der Rechenaufwand für das Absuchen des kompletten dreidimensionalen Raumes nach den Orientierungen der beiden

Strukturen zueinander, die die höchste geometrische Komplementarität zeigen, steigt in der Regel mit der Anzahl an Punkten, die für die Beschreibung der Geometrie der Proteinoberfläche genutzt werden.

Die detaillierteste Darstellungsform ist die atomare Darstellung, wie sie in den PDB-Dateien benutzt wird. Diese Darstellungsform wird ob der großen Anzahl an Datenpunkten allerdings nur vereinzelt[28,92] genutzt. In der Regel wird versucht, die Geometrie durch andere geometrische Deskriptoren zu abstrahieren.

Die für Docking am weitesten verbreitete Abstraktion der Strukturen ist die Abbildung auf 3D-Gitter[7,15,17,20,38,45,48,65]. Hierbei können durch verschiedene Gitterzellengrößen unterschiedliche Abstraktionslevel erreicht werden. Die Gitterdarstellung ist in Kapitel 4.1.1. detailliert beschrieben.

Die Grundlage für weitere alternative Proteinoberflächendarstellungen wurde durch Connolly[21] gelegt, welcher die Oberfläche mit einem Netzwerk von konvexen, konkaven und sattelförmigen Oberflächenfragmenten beschreibt und somit für die Repräsentation der Struktur die kleinen Lücken und Löcher zwischen den Atomen vernachlässigt. Basierend auf dieser *Connolly-Surface* wurden verschiedene für das Docking genutzte Oberfächenrepräsentationen entwickelt[5,54,55].

3.1.3.2.1.2 Berechnung der geometrischen Korrelation

Im so genannten Sampling Schritt, mit welchem die meisten Docking Protokolle beginnen, wird der gesamte Konformationsraum nach möglichen relativen Orientierungen von Rezeptor und Ligand zueinander durchsucht. Hierbei muss möglichst effizient eine sehr große Anzahl denkbarer Konformationen erstellt und zuverlässig bewertet werden, so dass sämtliche Konformationen ausgeschlossen werden, die unmöglich sind, z.B. solche mit großflächigen Überlappungen oder ohne jeglichen Kontakt. Das Bewertungskriterium für den Sampling Schritt ist in der Regel die geometrische Passgenauigkeit, also wie gut die Oberflächen von Rezeptor und Ligand miteinander korrelieren.

Um das Absuchen des gesamten Raums so effizient wie möglich zu gestalten, wird hierfür meistens eine vereinfachte Repräsentation der Proteine gewählt und möglichst einfache Bewertungsfunktionen für die zu evaluierenden Strukturen.

Für die Evaluierung aller möglichen Konformationen werden verschiedene Algorithmen eingesetzt. Nach wie vor ist die Benutzung der Fast Fourier Transformation[17,23,25,48,65,101] die am häufigsten angewandte Methode für den Sampling Schritt. Zusammen mit der *Geometric Hashing*[72,98] Methode und der Polaren Fourier Transformation gehören die FFT-basierten Algorithmen mit zur Gruppe der schnellsten und somit effizientesten Methoden relative Rezeptor-Ligand Konformationen zu evaluieren. Alternative Methoden, wie z.B. das Monte Carlo Sampling[28,29] (Bonvin, Gray und Bates), oder die systematische Suche mit Polar Koordinaten (Bates, Zacharias, Poupon), konnten bei CAPRI 5 (vgl. Kapitel 4.9) nicht überzeugen[64]. Neuere Methoden, die in CAPRI 5 eingeführt wurden, umfassen unter anderem „*Conformational Space Annealing*“[51] und „*Molecular Interaction Fields*“ (Fano).[64]

Um die Auswahl an weiter zu evaluierenden Konformationen so gering wie möglich zu halten, schließt der Sampling Schritt bei den meisten Gruppen einen abschließenden Clustering Schritt mit ein, bei dem ähnliche Konformationen im weiteren Docking Prozess durch einen Repräsentanten vertreten sind.[64]

3.1.3.2.1.3 Scoringfunktionen

Die meisten Dockingmethoden sind in der Lage, allein durch Berechnung der geometrischen Passgenauigkeit mindestens eine nahe native Struktur zu finden, allerdings ist diese beim *unbound* Docking meistens nicht die Struktur, die die höchste geometrische Korrelation aufweist. Da es nicht möglich ist, mehrere tausend Strukturen per Hand auszuwerten, um die nahe native Struktur zu finden, sind in den letzten Jahren zahlreiche Bewertungsfunktionen für potentielle Proteinkomplexe entwickelt worden. Diese Bewertungsfunktionen versuchen, an Hand von biochemischen und physikochemischen Eigenschaften die nahe nativen Strukturen zu identifizieren und im Ranking der potentiellen Strukturen nach oben zu bringen, so dass nur wenige Strukturen per Hand ausgewertet werden müssen.

Die unterschiedlichen Bewertungskonzepte umfassen Elektrostatik[13,17,25,33,40,51,60,73], Hydrophobizität[7,26,30,40,60,94], Wasserstoffbrücken[26,28,65], Desolvatationsenergie[13,17,51,73,100], Paarpotentiale[30,67,73], geometrische Packungsdichte der Interfaces[27,37], aber auch biologische Kriterien, wie die Konserviertheit der Residuen[53,58] oder Analysen der Interaktionen verwandter Proteine[36].

3.1.3.2.1.4 Refinement

Insbesondere bei den *rigid-body* Dockingprogrammen kann die nahe native Lösung Kollisionen zwischen beiden Bindungspartnern aufweisen. Diese Kollisionen können in einem nachgeschalteten Verfeinerungsschritt zum Teil entfernt werden. Hierfür kommen Methoden zur Energieminimierung und molekularen Dynamik zum Einsatz. Um Kollisionen zu entfernen, die durch die flexiblen Seitenketten entstehen, werden auch Rotamer Libraries genutzt (vgl.: 3.1.4.2.2).

3.1.4 Flexibilität im Interface

Die Strukturen der Proteinuntereinheiten eines Proteinkomplexes können sich zwischen gebundenem und ungebundenem Zustand z.T. erheblich voneinander unterscheiden. Die Ursache für diesen Unterschied liegt in der flexiblen Natur der Proteinstrukturen.

Während des Bindungsprozesses tauchen verschiedene Arten von Flexibilität auf, die unterschiedliche Behandlungen erfordern:

- Seitenkettenflexibilität
- Backbonebewegungen[56]
    - Bewegungen flexibler Loops
    - Bewegungen ganzer Domänen oder Bereiche durch *hinge* oder *shear* Bewegungen (dt.: Gelenk- oder Scherbewegungen)
    - *Disorder-to-order* Bewegungen (dt.: Ungeordnet-zu-Geordnet)

Die Seitenketten der Aminosäure haben bis zu vier frei drehbare Bindungen und können daher unterschiedliche Konformationen einnehmen, sofern sie eine hohe Solventzugänglichkeit haben und nicht z.B. über Wasserstoffbrückenbindungen stabilisiert sind. Die Röntgenstrukturen sind nur eine Momentaufnahme der Proteine, so dass die Seitenketten häufig in einer von mehreren möglichen Konformationen abgebildet sind. Zum Teil können sich aber auch ganze Loops (also Regionen der Proteine ohne reguläre Sekundärstruktur) oder ganze Proteindomänen bewegen, um einen Interaktionspartner zu binden. Während bei Proteinkomplexen mit einem Interface in Standardgröße (entspricht einem *Single-Patch* Interface (vgl. 3.1.2) in der Regel nur geringfügige Bewegungen bei der Bindung stattfinden, sind für die Assoziation großer Proteinkomplexe (Interface > 2.000 Å^2) auch häufig große strukturelle Änderungen nötig[56].

Abbildung 3.2-A zeigt am Beispiel von 1ACB (Komplex aus Chymotrypsin und Eglin C) exemplarisch einen durch Seitenkettenflexibilität verursachten *‚Clash'*, welcher durch gitterbasierte Dockingprogrammen wie *ckordo* bei der Berechnung der geometrischen Korrelation zu einer 'Bestrafung' führen würde, also zu einem negativen Beitrag zum Gesamtscore (s. Kapitel 4.1). In Abbildung 3.2-B wird an

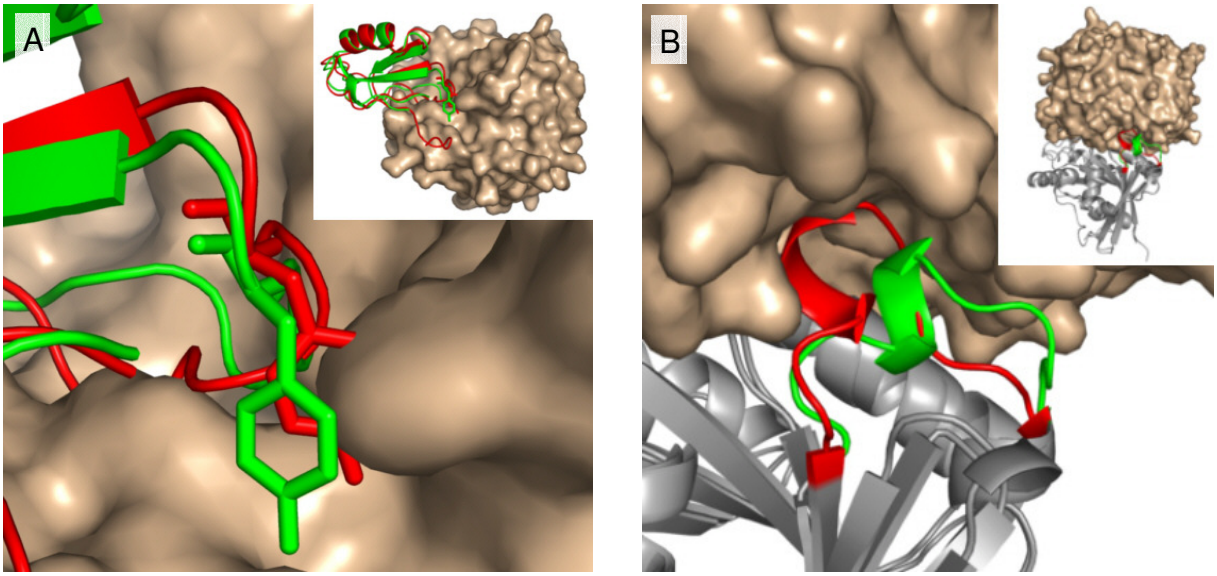

Abbildung 3.2: Auswirkungen von Flexibilität auf geometrische Komplementarität von Komplexen: (**A**) Durch Seitenkettenflexibilität verursachte Kollisionen bei 1ACB. In rot ist die Struktur des Liganden im ungebundenen Zustand gezeigt, während die Struktur im Komplex in grün dargestellt ist. (**B**) Kollisionen zwischen Ligand im ungebundenen Zustand (rot) mit Rezeptor durch Bewegungen eines Loops bei 1I2M. Die Struktur des Liganden im gebundenen Zustand ist in grün gezeigt.

Hand von 1I2M (Komplex aus Ran GTPase mit RCC1) gezeigt, welche Auswirkungen ein flexibler Loop auf die geometrische Komplementarität des Interfaces haben kann. Der in rot gezeigte Loop des Liganden im ungebundenen Zustand kollidiert über mehrere Residuen hinweg mit dem Rezeptor, während sich im gebundenen Zustand der in grün gezeigte Loop an den Rezeptor anlegt.

Alleine die Seitenkettenbewegungen können allerdings ausreichen, um die Oberflächenstruktur eines Proteins so zu verändern, dass *rigid-body* Dockingprogramme nicht mehr in der Lage sind, alleine an Hand der geometrischen Korrelation die nahe native Struktur zu finden. Daher spielt die Behandlung von Seitenkettenflexibilität eine entscheidende Rolle bei *in silico* Methoden zur Vorhersage von Proteinkomplexstrukturen.

Najmanovich *et al.*[68] haben 1333 Protein-Bindungsstellen im Bezug auf die Flexibilität der beteiligten Aminosäuren untersucht und daraus eine Wahrscheinlichkeitsskala entwickelt, die angibt, wie wahrscheinlich eine Konformationsänderung bei der Bindung ist. Nach dieser Skala ergibt sich die folgende Flexibilitätsreihenfolge: LYS > ARG, GLN, MET > GLU, ILE, LEU > ASN, THR, VAL, TYR, SER, HIS, ASP > CYS, TRP, PHE. Ferner haben Najmanovich *et al.* festgestellt, dass nur bei einem geringen Anteil der Interface-Residuen überhaupt Konformationsänderungen bei Bindung stattfinden. Bei 85% der Interfaces haben nur drei oder weniger Residuen im gebundenen und ungebundenen Zustand verschiedene Konformationen. Zusätzlich ändert sich chi-1 in 94,4% und chi-2 in 95,7% der Interfaceresiduen gar nicht (chi-1/chi-2: Torsionswinkel der Seitenketten). Dennoch sind die Strukturänderungen am häufigsten, die durch flexible Seitenketten entstehen. Bei den von Najmanovich *et al.* untersuchten Beispielen fand nur in 12% der Fälle eine Backbonebewegung von mehr als 2 Å statt und 75% aller Strukturen zeigten keine Backbonebewegung von mehr als 1 Å. Insgesamt sind die Interface Regionen dennoch flexibler als die restliche Proteinoberfläche[9]. Der größte Anteil an der erhöhten Flexibilität von Bindungsstellen liegt bei den Residuen, die am Rand des Interfaces liegen, während das Zentrum der Interfaces als eher rigide bezeichnet werden kann[91].

3.1.4.1 Bioinformatische Methoden zur Simulation von Flexibilität

Prinzipiell stehen Methoden zur Verfügung, die versuchen die Flexibilität von Proteinstrukturen *in silico* zu simulieren, zum Beispiel gibt es Programme zur Berechnung der molekularen Dynamik wie Amber[75] oder Gromacs[95]. Allerdings sind diese Methoden sehr rechenintensiv.

Insbesondere beim Docken von Proteinstrukturen steigt diese Komplexität erheblich an, da der ‚richtige' Zustand eines jeden Proteins auch von der Struktur des anderen Proteins abhängt. Die Simulation der Flexibilität müsste für jede Struktur durchgeführt werden, die im Laufe des Dockingprozesses evaluiert wird. Alleine die Flexibilitätssimulation aller potentiellen Komplexstrukturen, die eine gewisse geometrische Komplementarität zeigen, würde Simulationen für mehrere 10.000 Proteinstrukturen gleichkommen, da bei jeder Struktur der Bindungspartner in einer anderen Orientierung angelagert ist, so dass das Umfeld verschieden ist. Zudem muss für eine zuverlässige Flexibilitätssimulation von Proteinstrukturen auch das umgebende Wasser berücksichtigt werden, was die Komplexität des Flexibilitätsproblems erheblich steigert.

Um Flexibilität beim Protein-Protein Docking zu berücksichtigen, müssen also andere Methoden entwickelt werden als die direkte Simulation der Bewegungen.

3.1.4.2 Flexibilität und Docking

Viele Dockingmethoden basieren auf der Suche nach der besten geometrischen Passform zwischen den beiden Bindungspartnern. Insbesondere in den frühen Versionen der Dockingprogramme wurden die Proteine als *rigid-bodies*, also rigide, nicht flexible Körper, behandelt. Aufgrund der oben beschriebenen Konformationsänderungen, die im Laufe des Bindungsprozesses auftauchen können, kann allerdings so die richtige gebundene Struktur oftmals nicht gefunden werden. Daher wird in den meisten Dockingprogrammen inzwischen versucht, die Flexibilität mit einzubeziehen. Bonvin gibt in einem kürzlich erschienen Review Artikel einen guten Überblick über aktuelle Entwicklungen aus dem Bereich Flexibilität und Docking[11].

Im Folgenden werden verschiedene Ansätze, Flexibilität von Proteinstrukturen beim Docking einzubeziehen, kurz beschrieben. Die Flexibilität kann zu zwei verschiedenen Zeitpunkten in den Docking Prozess einfließen, einerseits vor bzw. während der Suche nach der Komplexstruktur mit der höchsten geometrischen Komplementarität oder in einem nachgeschalteten Refinement Schritt:

3.1.4.2.1 Flexibilität vor oder während des Dockings

‚Soft Penetration'

Durch asymmetrische Behandlung beider Proteine bei Programmen, die mit Gitterrepräsentationen arbeiten, werden leichte Kollisionen zwischen beiden Proteinen zugelassen, bzw. nicht 'bestraft', so dass z.B. kollidierende Seitenketten keinen negativen Beitrag zum Gesamtscore leisten[23,73].

Multi Copy Ansätze/Ensemble Docking

Ein anderer Ansatz um den verschiedenen möglichen Konformationen der Proteine Rechnung zu tragen, ist der *‚multi-copy'* Ansatz bei dem der Dockingprozess mit mehreren möglichen Konformationen durchgeführt wird. Die unterschiedlichen Konformationen werden vor der eigentlichen Dockingprozedur durch kurze molekulare Dynamiksimulationen (MD) generiert oder aus NMR Strukturen entnommen. Dieser Ansatz wurde sowohl für ganze Proteine[31,59,97] als auch für einzelne Bereiche der Proteine, wie z.B. bestimmte Loops[6], entwickelt.

Ein großer Nachteil dieses Ansatzes ist, dass logischerweise die Rechenzeit mit jeder zusätzlichen gedockten Konformation erheblich zunimmt. Allerdings ist dies die bislang einzige Möglichkeit, auch Backbone Bewegungen beim Docking mit den weit verbreiteten FFT-Dockingmethoden (vgl.: Kapitel 4.1) zu berücksichtigen.

Gelenkbewegungen

Um die Bewegungen ganzer Domänen zueinander in den Dockingprozess einzubinden, wird das Protein oft in rigide Untereinheiten geteilt, die unabhängig voneinander gedockt und anschließend wieder aneinandergesetzt werden. Hierfür müssen allerdings die Gelenkregionen vorher bekannt sein[84].

Abgeschnittene Seitenketten

Da insbesondere die langen und sehr flexiblen Seitenketten von Lysin, Arginin und Gluatmin häufig *Clashes* verursachen, haben Heifetz und Eisenstein[34] versucht, durch ‚Abschneiden' dieser Seitenketten mehr native Konformationen zu finden. Palma *et al.*[73] erlauben für Arginin, Lysin, Asparaginsäure, Glutaminsäure und Methionin eine Penetration des gebundenen Proteins, was einem Abschneiden der Seitenketten gleich kommt.

3.1.4.2.2 Flexibilität im Refinement Schritt

Rotamerbibliotheken

Da die Berechnung der korrekten Konformation der Seitenketten während bzw. vor der Dockingprozedur zu rechenintensiv ist, erfolgt die Suche nach der richtigen Seitenkettenkonformation in der Regel in einem dem eigentlichen Docking nachgeschalteten Schritt, dem so genannten ‚Refinement Schritt'. Da es für verschiedene Konformationen der Aminosäuren unterschiedliche Präferenzen gibt, bestimmte Winkelkombinationen *in vivo* einzunehmen, diese also deutlich häufiger auftauchen als andere, wurden Bibliotheken von möglichen Seitenkettenkonformationen (‚*Rotamerlibraries*') erstellt, welche in die zu untersuchenden Strukturen eingesetzt werden. Die wahrscheinlichste und energetisch günstigste Seitenkettenkonformation ersetzt dann die möglicherweise falsche Konformation der ungebundenen Struktur[15,41].

Nachgeschaltete MDs und Energieminimierung

Den meisten Dockingprogrammen wird eine kurze Simulation molekularer Dynamik mit anschließender Energieminimierung nachgeschaltet. Hierdurch werden in erster Linie Kollisionen der beiden Proteine miteinander entfernt. Es gibt verschiedene Ansätze, die Atome zu definieren, welche sich während der MD frei bewegen dürfen. Einige Dockingprotokolle erlauben hier nur Bewegungen der Seitenkettenatome, andere erlauben auch Bewegungen des Backbones[41,97].

4 Methoden

4.1 Dockingprogramm ckordo

Für diese Arbeit wurde das Dockingprogramm *ckordo*[65,101] verwendet. *Ckordo* ist ein *rigid-body* (d.h. beide Proteine werden als starre nicht flexible Körper behandelt) Dockingprogramm, basierend auf der FFT-Methode (Fast Fourier Transformation), welche 1992 von Katchalski-Katzir[48] zuerst auf das Docking Problem angewandt wurde.

Durch den Einsatz der FFT wird der Raum für die Suche nach der optimalen geometrischen Korrelation von sechs auf vier Dimensionen reduziert, was zu einer deutlichen Verringerung der benötigten Rechenzeit für diesen aufwändigen Schritt führt.

4.1.1 Repräsentation der Proteinstrukturen als Gitterzellen

Für das FFT-Docking werden die Proteinstrukturen als Gitter repräsentiert (Abbildung 4.1).

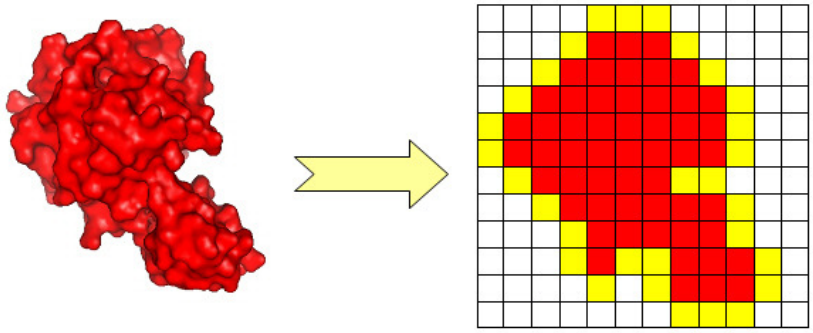

Abbildung 4.1: Projektion des Proteins auf ein Gitter. Das größere der beiden Proteine wird komplett als Proteininneres dargestellt (rot) und mit einer Randschicht (gelb) umgeben.

Es gibt zwei Arten von Gitterzellen, die das Protein repräsentieren. Zum einen gibt es Zellen, die das Innere darstellen, und zum anderen solche, die die Oberfläche des Proteins markieren. Die verschiedenen Gitterzellen unterscheiden sich in den ihnen zugewiesenen Zahlenwerten. Für das größere der beiden Proteine, den Rezeptor, wird das komplette Protein durch Zellen, die den Zahlenwert für ‚Innen' tragen, dargestellt und eine Schicht definierbarer Breite von Oberflächenzellen außen herum gelegt. Bei dem kleineren Protein, dem Liganden, werden die Oberflächenzellen abgezogen.

Zur Berechnung der geometrischen Passgenauigkeit beider Proteine wird der komplette Raum nach einer optimalen Kombination von Translationen und Rotationen für das kleinere Protein abgesucht, so dass eine maximale Anzahl von Oberflächenzellen beider Gitter überlappt, ohne dass es zu Überlappungen (*Clashes*) von Zellen des Inneren kommt. Durch den unterschiedlichen Charakter der Oberflächenzellen beider Proteine schmiegen sich beide Proteine bei einer maximalen Anzahl an Oberflächenzellenüberlappungen aneinander an, wie an Hand der gepunkteten Linie in Abbildung 4.2 deutlich wird.

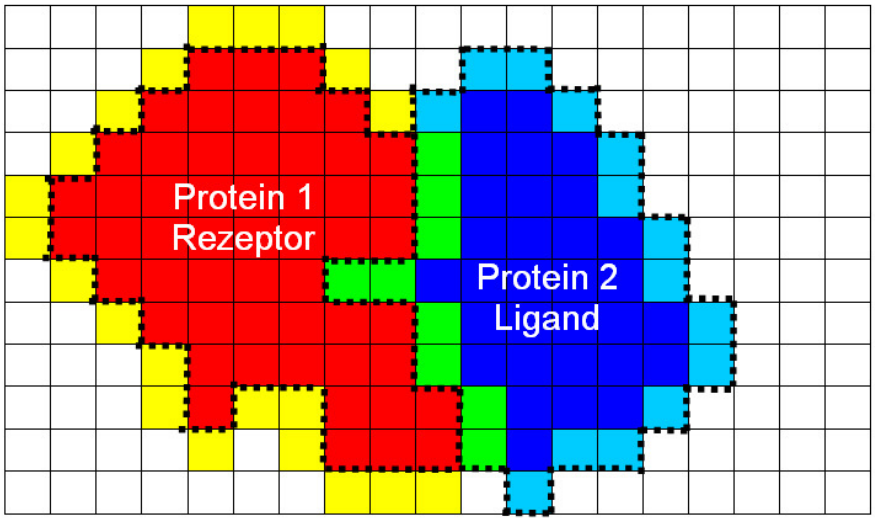

Abbildung 4.2: Bei maximaler Anzahl von sich überschneidenden Oberflächenzellen (Protein 1: gelb, Protein 2: hellblau, Überschneidend: grün) zeigen beide Proteine eine optimale geometrische Passgenauigkeit. Durch die gepunktete dicke Linie werden die Grenzen der beiden Proteine angezeigt

4.1.2 Berechnung der geometrischen Korrelation im Fourier-Raum

Zur Berechnung der geometrischen Korrelation werden beide Gitter gegeneinander verschoben und die Zahlenwerte miteinander multipliziert, die in den jeweils überlappenden Gitterzellen stehen. Da eine maximale Anzahl von überlappenden Oberflächenzellen zu einem hohen Wert führen soll, müssen die den Oberflächenzellen zugewiesenen Werte so gewählt sein, dass Oberfläche mal Oberfläche einen hohen Wert ergibt, während Inneres mal Inneres einen niedrigen Wert ergibt, da dies *in vivo* nicht auftauchenden Überlappungen beider Proteine entspräche. Die Schwierigkeit der Auswahl der Zahlenwerte liegt darin, dass Überlappungen zwischen der Oberfläche des einen Proteins mit dem Inneren des anderen Proteins zwar nicht gewünscht sind, also 'bestraft' werden müssen, aber dennoch auch bei der nativen Konformation auftauchen können. Der Grund hierfür liegt in der Flexibilität der Proteine. Leichte Überlappungen zwischen Oberfläche und Innerem können beim *rigid-body* Docking z.B. durch lange flexible Seitenketten der Aminosäuren entstehen, welche sich *in vivo* während des Bindungsprozesses umlagern. Ein solcher Clash müsste zur Berechnung der geometrischen Korrelation akzeptiert werden und zu einem späteren Zeitpunkt des Dockingprozesses durch Energieminimierung oder Molekulare Dynamik Programme korrigiert werden.

Bei den meisten bisher veröffentlichten Dockingprogrammen, die auf dieser Gitterrepräsentation beruhen, werden die Zahlenwerte so verteilt, dass dem Inneren des Rezeptors ein niedriger negativer Wert zugewiesen wird und alle anderen Zellen mit dem Wert 1 versehen werden. Alle Zellen, die nicht das Protein darstellen, tragen den Wert 0.

Protein 2 / Protein 1	Oberfläche (1)	Inneres (1)	Leer (0)
Oberfläche (1)	**1**	**1**	0
Inneres (-6)	**- 6**	**- 6**	0
Leer (0)	0	0	0

Tabelle 4.1: Verteilung der Zahlenwerte für verschiedene Zellen bei *ckordo* und die daraus resultierenden Ergebnisse bei Überlappungen;

Durch diese Auswahl der Zahlenwerte – in Tabelle 4.1 am Beispiel der in *ckordo* benutzten Zahlenwerte gezeigt – kommt es dazu, dass Oberfläche/Inneres Kontakte je nach Richtung unterschiedlich gewertet werden. Diese Zahlenwerte sind bislang willkürlich gewählt, bzw. durch ‚Ausprobieren' gesetzt worden. Verschiedene Dockingprogramme benutzen unterschiedliche Werte für das Innere von Protein 1, so benutzt MolFit von Katchalski-Katzir[48] z.B. -12, während diesen Zellen bei *ckordo* -6 zugewiesen wird.

Für die Suche nach der Orientierung der beiden Proteine zueinander, die die höchste geometrische Komplementarität aufweist, muss das gesamte Spektrum an potentiellen Orientierungen möglichst umfassend abgesucht werden. Hierfür werden zunächst mit definierbarem Winkelinkrement verschiedene Rotationen des kleineren Proteins durchgeführt. Übliche Werte für das Winkelinkrement sind 12° oder 15°. Anschließend werden die Gitter mit der Fast Fourier Transformation in den Fourier Raum überführt und dort für alle Translationen die geometrische Korrelation berechnet. Eine detaillierte Beschreibung des *ckordo*-Algorithmus ist in der Original Arbeit von Zimmermann[101] gegeben.

Wenn die Suche nach den Komplexstrukturen mit der größten geometrischen Korrelation im direkten Raum durchgeführt wird, sind N^3 Multiplikationen und Additionen nötig, für jede der N^3 möglichen Kombinationen der Rotationswinkel. Dies führt zu N^6 nötigen Rechenschritten (Gittergröße: N *x* N *x* N). Bei einer Berechung der geometrischen Korrelation mit Hilfe der Fast Fourier Transformation reduziert sich der Rechenbedarf auf $N^3 \cdot \ln(N^3)$ Schritte, da die Berechnung der geometrische Korrelation für alle Translationen je Rotation in einem Rechenschritt durchgeführt werden kann[23].

4.1.3 SVM optimierte Bewertungsfunktionen

Die bislang erfolgreichste Bewertungsfunktion wurde von Oliver Martin[61] entwickelt. Martin hat mehrere unterschiedliche Bewertungsfunktionen zusammengefasst und eine Support Vector Machine (SVM) dahingehend trainiert, die verschiedenen

Kriterien im richtigen Verhältnis zueinander anzuwenden, so dass nahe native Strukturen erkannt und falsche Lösungen ausgeschlossen werden[61].

Diese von Martin verwendeten Bewertungsschemata beinhalten spezialisierte Energiefunktionen molekularer Fragmente, evolutionäre Verwandtschaft, komplexklassen-spezifische Wahrscheinlichkeitsverteilungen von Residuen, Lückenvolumen, die Größe der verborgenen Oberfläche, empirische Paarpotentiale auf atomarer und Aminosäureebene sowie ein Maß für die Festigkeit der Bindung. Die entwickelten Bewertungsfunktionen sind hochspezifisch für die einzelnen Kategorien von Proteinkomplexen und in der Lage, nahe native Lösungen mit hoher Sensitivität aus einer großen Anzahl potentieller Komplexanordnungen heraus zu filtern. Eine Sortierung der Lösungsvorschläge durch Anwendung der Bewertungsfunktionen führt zu einer signifikanten Anreicherung von nahe nativen Komplexen in den oberen Rängen.

4.2 Methode zur Optimierung von Gewichtungsfaktoren

4.2.1 Allgemeiner Überblick

In Kapitel 2 wurde gezeigt, dass es eine große Anzahl verschiedener Eigenschaften gibt, mit denen die Interaktionsflächen von Proteinen charakterisiert werden können. Viele dieser Eigenschaften sind aminosäure- oder atomspezifisch. Die zentrale Idee dieser Arbeit besteht darin, dass diese Eigenschaften direkt in die Gitterdarstellung der Proteine einfließen und somit entweder bei der Berechnung der geometrischen Korrelation oder als Postfilter angewandt werden können. Bei der herkömmlichen Gitterdarstellung (vgl. Abbildung 4.3 A-C) gibt es nur drei verschiedenen Zelltypen (leer/Oberfläche/Inneres).

Durch diese Darstellung werden Informationen über die zu dockenden Proteine vernachlässigt, die zum einen vorhanden sind und zum anderen hilfreiche Hinweise für die Identifikation der nativen Bindung geben können. Jede Zelle kann nicht nur das Innere oder die Oberfläche der Proteine repräsentieren, sondern auch die

dahinter stehenden Aminosäuren (Abb. 4.3 D) und Atome. Jeder Zelle wird daher zusätzlich zu der Eigenschaft das Innere oder die Oberfläche des Proteins zu repräsentieren, in Abhängigkeit von der jeweilige Aminosäure bzw. dem jeweiligen Atomtyp ein Gewichtungsfaktor zugewiesen (Abb. 4.3 E).

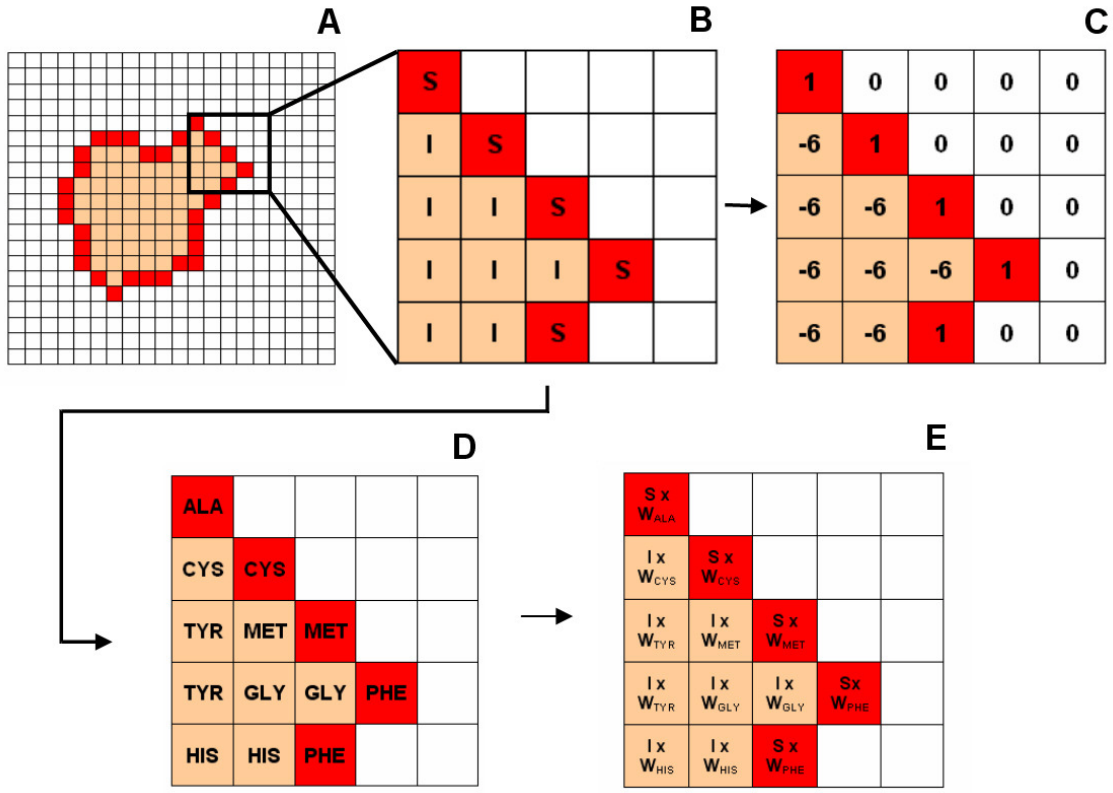

Abbildung 4.3: Erweiterung der Gitterdarstellung um aminosäurespezifische Gewichtungsfaktoren

Die Gewichtungsfaktoren können unterschiedliche Gegebenheiten des Docking Prozesses abbilden: Zum einen spielen bestimmte Aminosäuren bei der Interaktion von Proteinen eine größere Rolle als andere. So sind z.B. die aromatischen Residuen deutlich häufiger im Interface anzutreffen als auf dem Rest der Oberfläche. Wenn diesen Aminosäuren ein höherer Wert zugeordnet würde als den anderen Aminosäuren, so würden mögliche Komplexstrukturen, bei denen aromatische Residuen in die Berechnung der geometrischen Korrelation einfließen, einen höheren Score erhalten als andere. Zum anderen können Aminosäuren, die lange und flexible Seitenketten haben, auch bei nahe nativen Strukturen Kollisionen zwischen den beiden Proteinen hervorrufen, da sich diese *in vivo* umlagern würden. Diesen Aminosäuren könnten niedrigere Gewichtungsfaktoren zugewiesen werden,

so dass Kollisionen, die durch sehr flexible Seitenkettenatome entstehen, nur zu einer geringen 'Bestrafung' führen.

Für einzelne Aminosäuren und Atome ist an Hand der bekannten Eigenschaften nicht klar ersichtlich, ob diese eher höhere oder eher niedrigere Gewichtungsfaktoren erhalten sollten. So ist z.B. für Methionin in der Literatur[16] eine hohe Interface-*Propensity* beschrieben, während Methionin gleichzeitig eine relativ lange und flexible Seitenkette hat. Daher werden die Gewichtungsfaktoren in dieser Arbeit durch mathematische Methoden so optimiert, dass durch den Einsatz derselben nahe native Strukturen höhere Scores erhalten als falsche.

Für den Optimierungsprozess werden zunächst für alle von *ckordo* vorgeschlagenen Strukturen Kontaktmatrizen erstellt. Diesen Kontaktmatrizen ist zu entnehmen, welche Aminosäure, welches Atom wie häufig mit welcher Aminosäure bzw. mit welchem Atom interagiert. Für jede Struktur ist die Qualität in Form des RMSiCα (s. Kapitel 4.8) bekannt, welcher die jeweilige Ähnlichkeit zum nativen Komplex darstellt. Mit Hilfe der Kontaktmatrizen konnte die Berechnung der jeweiligen Komplementarität im direkten Raum (also nicht im Fourier Raum) und ohne die Gitterdarstellung durchgeführt werden (vgl. Formel 1) und die Formel zu Berechnung der Komplementarität um die jeweiligen Gewichtungsfaktoren erweitert werden (Formel 2).

F1

$$geoscore = \sum (P1xP2)$$

Formel 1: Formel für die Berechnung der geometrischen Korrelation im direkten Raum. Die Werte der jeweils überlappenden Zellen des größeren Proteins (P1) werden mit denen des kleineren (P2) multipliziert und aufsummiert.

F2

$$weighted\_geo\_score = \sum ((W_{AA}xP1)x(W_{AA}xP2))$$

Formel 2: Formel für die Berechnung des geometrischen Scores mit Gewichtungsfaktoren. Jeder Zelle wird jetzt ein atom- oder aminosäurespezifischer Gewichtungsfaktor (W_{AA}) zugewiesen und mit den jeweiligen Werten für P1 bzw. P2 multipliziert.

Als Datengrundlage für die Optimierungen wird der komplette Benchmark 2.0 (vgl. Kapitel 4.5) mit einem Winkelinkrement von 15° gedockt und für jede Rotation die 5 Strukturen (Translationen) mit der höchsten geometrischen Korrelation weiter verwendet. Dies führt zu 22.000 potentiellen Strukturen je Komplex[101]. Das Docking wird mit einer Randschichtdicke (Oberflächenschicht) von 2 Å und mit einer Gitterzellengröße von 1 Å durchgeführt.

Für einige Komplexe findet *ckordo* nur sehr wenige nahe native Strukturen, daher werden für alle Komplexe künstlich weitere Strukturen mit niedrigem RMS-Wert generiert. Hierfür wird ein *ckordo* Lauf vorgenommen, bei dem 1.000 zufällig gewählte Rotationen aus dem Bereich zwischen plus 10° und minus 10° um die native Struktur untersucht werden. Hieraus werden jeweils alle Strukturen mit einem RMSiCα ≤ 5 Å (bis zu 4700) den Datensätzen zugefügt. Stichprobenartige Vortests haben ergeben, dass die Gewichtungsfaktoren weitaus besser funktionieren, die mit den künstlich generierten nahe nativen Lösungen optimiert werden.

Für die jeweiligen Optimierungen werden 5 zufällige Untergruppen der Komplexe gebildet und die Optimierungen 5 mal durchgeführt, wobei jedes Mal eine der 5 Untergruppen nicht mit optimiert wird (*5-fold-cross-validation*), so dass gezeigt werden kann, dass die errechneten Gewichtungsfaktoren mit einem beliebigen Datensatz berechnet werden könnten.

Da die Optimierungen selbst, abhängig von der Anzahl der zu optimierenden Parameter und von der Anzahl der Strukturen, sehr viel Arbeitsspeicher beanspruchen, können nicht alle 22.000 Strukturen je Komplex genutzt werden. Daher wird für jeden Komplex eine bestimmte Anzahl an Strukturen ausgewählt, die für die Optimierung genutzt wird. Dieses Subset besteht für die Optimierung der aminosäurespezifischen Faktoren aus 10.000 Strukturen je Komplex und für die atomspezifischen Faktoren aus 4.000 Strukturen. Die Anzahl für die atomspezifische Optimierung ist ob der größeren Anzahl an Gewichtungsfaktoren (40 Atomtypen + I1 = 41) niedriger. Die genutzten Strukturen müssen alle Strukturen mit einem RMSiCα ≤ 5 Å beinhalten, jedoch muss mindestens die Hälfte der Strukturen einen schlechteren RMSD aufweisen. In den Fällen, in denen es mehr nahe native

Strukturen gibt, werden so viele Strukturen zufällig ausgewählt, bis maximal 50% des Subsets einen RMSiCα ≤ 5 Å haben.

Die jeweiligen Gewichtungsfaktoren werden unabhängig voneinander für die drei Komplexklassen Enzym-Inhibitor/Substrat, Antikörper-Antigen und ‚Andere' Komplexe optimiert.

Mit Hilfe der nicht linearen Minimierung (nlm()-Funktion) des R-Paketes[77] werden die jeweiligen Gewichtugsfaktoren so optimiert, dass die Komplementaritätsscores für jede Struktur möglichst nah an die in Abb. 4.4 gezeigte Funktion herankommen oder anders formuliert, dass der Fehler zwischen tatsächlichem Score und gewünschtem Score möglichst minimal ist.

Als Zielfunktion wurden in diversen Vortests verschiedene Funktionen getestet, wie z.B. durchgängig lineare oder logarithmische Funktionen. Die besten Ergebnisse wurden für die in Abbildung 4.4 gezeigte Zielfunktion erreicht. Nahe nativen Strukturen (bis RMSiCα 5 Å) soll ein Wert von 10.000 und falschen (RMSiCα > 10 Å)

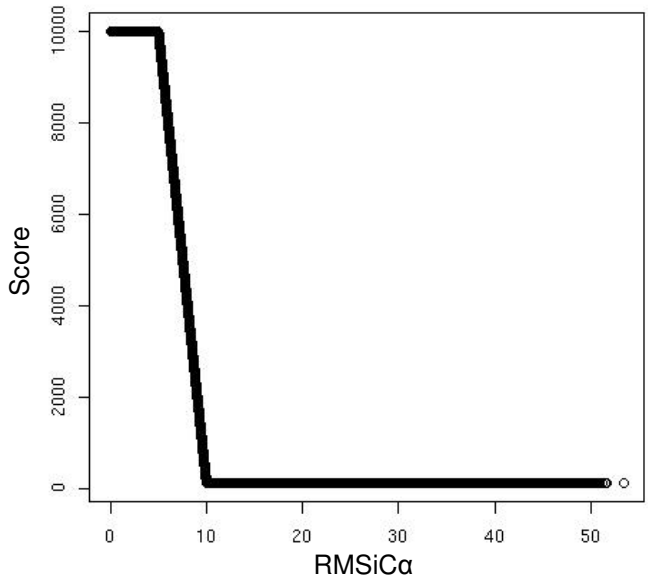

Abbildung 4.4: Zielfunktion für die Optimierung. Nahe nativen Strukturen (bis RMSiCα 5Å) soll ein Wert von 10.000 und falschen (RMSiCα > 10Å) ein Wert von 100 zugewiesen werden. Die Zielwerte für Strukturen mit einem RMSiCα zwischen 5 Å und 10 Å werden durch eine lineare Funktion, die die beiden Levels miteinander verbindet, ermittelt.

ein Wert von 100 zugewiesen werden (s. auch Abbildung 4.7). Die Zielwerte für Strukturen mit einem RMSiCα zwischen 5 Å und 10 Å werden durch eine lineare Funktion ermittelt, die die beiden Stufen miteinander verbindet.

Der Wert für die nahe nativen Strukturen wird so gewählt, dass er 100-mal größer ist als der Wert für Lösungen mit hohem RMS-Wert, welcher die gleiche Größenordnung wie die bisherigen Werte für die geometrische Korrelation haben sollte. Da es nicht möglich ist, eine klare Grenze zwischen richtigen und falschen Lösungen zu ziehen, wird der fließende Übergang in Form einer linearen Funktion zwischen den beiden Stufen gewählt.

4.2.1.1 Optimierung von aminosäurespezifischen Gewichtungsfaktoren

Der eleganteste Einsatz der optimierten Gewichtungsfaktoren ist es, diese direkt bei der ersten Berechnung der geometrischen Komplementarität einzusetzen. Dadurch könnte ohne zusätzliche Rechenzeit ein besseres Ergebnis erzielt werden.

Bei stichprobenartigen Vortests wurde auch der generelle Wert für das Innere des größeren Proteins (I1) optimiert. Diese Optimierungen haben gezeigt, dass z.T. für I1 auch positive Werte optimiert werden. Würde man allerdings positive Werte für I1 in *ckordo* einsetzen und zur Berechnung der geometrischen Korrelation nutzen, so

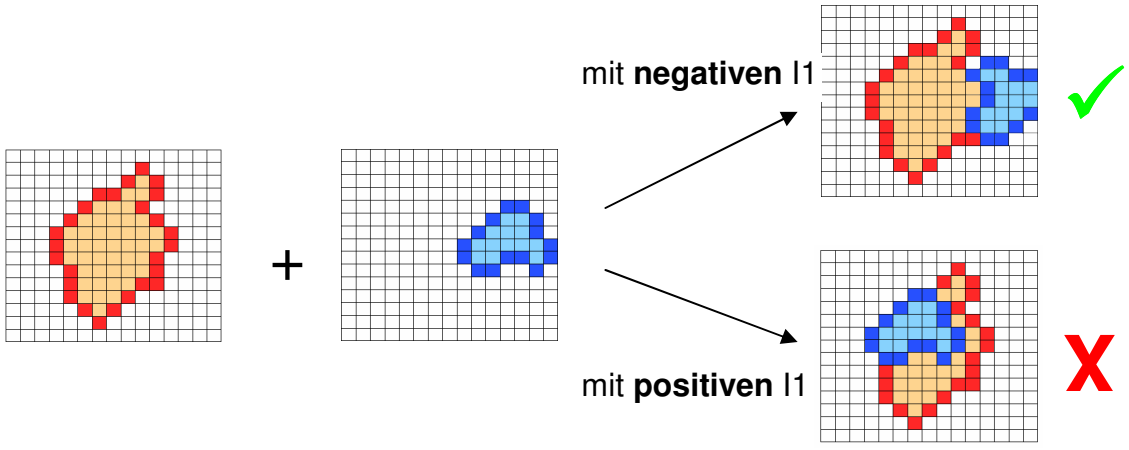

Abbildung 4.5: Würde man positive Werte für I1 bei der Berechnung der geometrischen Korrelation verwenden, so würden Strukturen, bei denen sich beide Proteine überlappen, die höchsten Scores erhalten

würden die höchsten Scores für Strukturen errechnet, die sich weitgehend überlappen (vgl. Abb. 4.5).

Da aber die Anwendung der positiven I1 Werte als Reranking Kriterium möglich ist, werden Optimierungen vorgenommen, bei denen einerseits I1 ein optimierbarer Parameter ist und andererseits mit einem fixen Wert von I1 = -6.

Die Gewichtungsfaktoren, die gleichzeitig mit dem Wert für I1 optimiert werden, werden als Rerankingfunktion evaluiert, also im Hinblick darauf, wie gut diese aus einem mit *ckordo* generierten Set von möglichen Komplexstrukturen die nahe nativen herausfiltern können. Hierbei ist es nicht notwendig, dass I1 negativ ist, da die Sets mit den ursprünglichen Werten (also I1 = -6) erstellt werden, so dass die zu bewertenden Strukturen nicht überlappend sind, sondern höchstens geringfügige Kollisionen einzelner Aminosäuren bzw. Atome enthalten.

4.2.1.2 Optimierung von atomspezifischen Gewichtungsfaktoren

Da die meisten Kollisionen beider Proteine durch besonders flexible Bereiche der Aminosäuren zustande kommen und z.B. der Backbone eher selten beteiligt ist, könnten atomspezifische Gewichtungsfaktoren eher in der Lage sein, nahe native Strukturen zu identifizieren als aminosäurespezifische. Ferner wurden von Neuvirth *et al.* [69] Atome identifiziert, die häufiger im Interface von Proteinen vorkommen als auf der restlichen Oberfläche.

Um möglichst spezifische Gewichtungsfaktoren für die verschiedenen Atomtypen zu erhalten, wird die Atomtypenklassifizierung nach Melo *et al.*[62] gewählt, so dass 40 verschiedene Gewichtungsfaktoren für 40 verschiedene Atomtypen optimiert werden. Abbildung 4.6 zeigt welche Atome welchem Atomtyp zugeordnet sind.

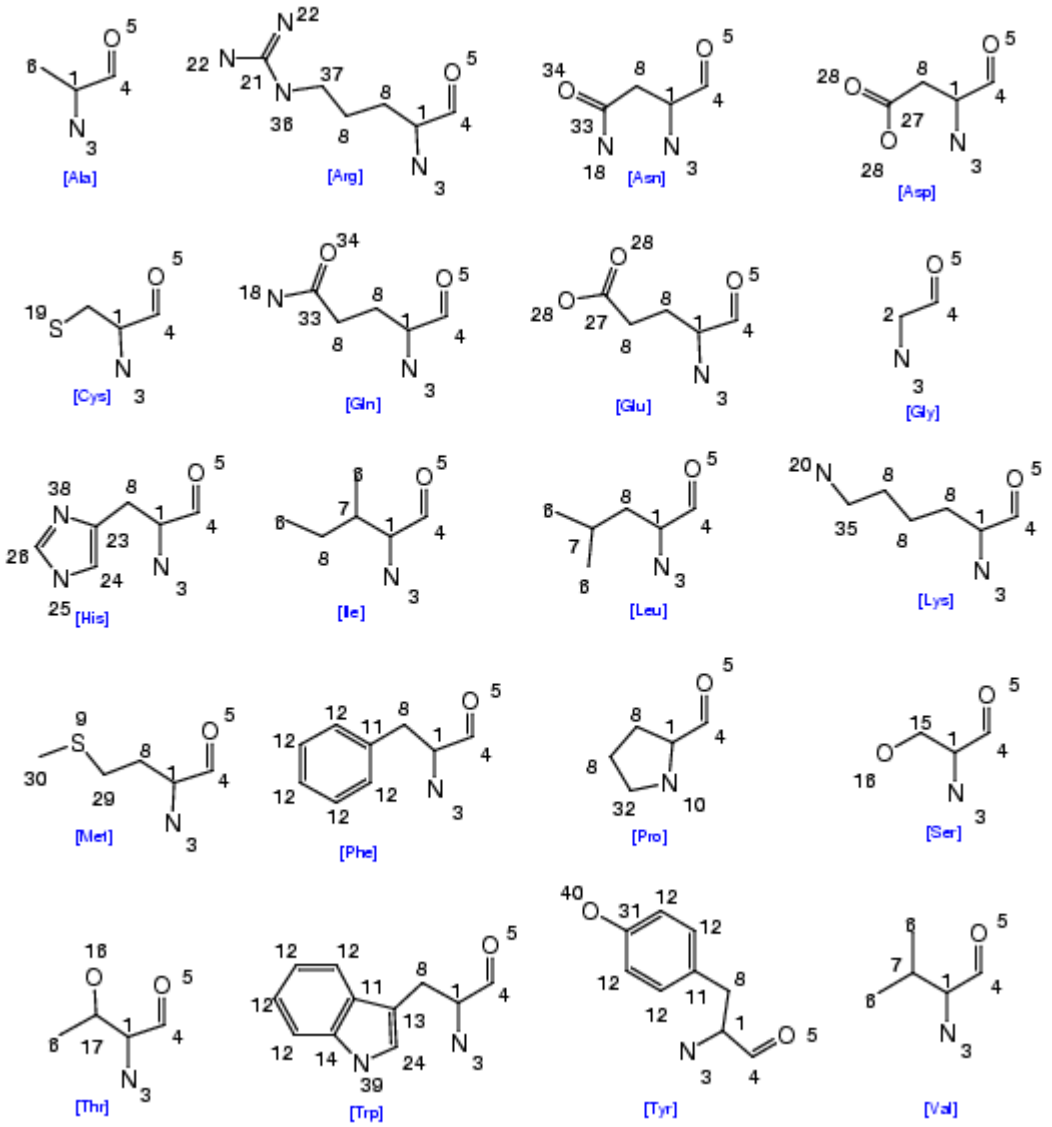

Abbildung 4.6: Atomtypen nach Melo *et al.*[62]

4.3 Postfilter

Für die Optimierung und Evaluation der Gewichtungsfaktoren wird ein eigenständiges Programm entwickelt, welches für die von *ckordo* generierten Strukturen die geometrische Korrelation und die gewichtete Passgenauigkeit neu berechnet.

Dieses Programm projiziert genauso wie *ckordo* die rotierten und translatierten Strukturen auf ein Gitter. Die Korrelationsberechnungen finden im direkten Raum statt. Da die zu evaluierenden Strukturen nur eine relativ kleine Untergruppe der

Strukturen sind, die beim Absuchen des kompletten Raumes berücksichtigt werden, ist die Anwendung der Gewichtungsfaktoren hier etwa dreimal schneller als bei einem *ckordo* Lauf.

Für die finale Anwendung ist die Einbindung des gewichteten Rerankings in *ckordo* allerdings deutlich schneller, da dann die Proteine nicht erneut auf ein Gitter projiziert werden müssen, sondern die für die Berechnung der geometrischen Korrelation erstellten Gitter genutzt werden können. Die von *ckordo* unabhängige Version der Rerankingfunktion mit Gewichtungsfaktoren kann aber auch für die Bewertung von potentiellen Strukturen genutzt werden, die mit anderen Dockingprogrammen erstellt werden.

4.4 Kombination der Filter

Es wird ferner untersucht, ob eine Kombination der beiden Scoringfunktionen eine weitere Verbesserung der Vorhersagequalität erbringt. Hierfür werden die mit den aminosäurespezifischen und den atomspezifischen Gewichtungsfaktoren berechneten Scores einerseits miteinander addiert und andererseits multipliziert.

Für die in *ckordo* genutzten aminosäurespezifischen Gewichtungsfaktoren wird die Kombination zusätzlich zwischen dem von *ckordo* berechneten gewichteten Score und der Rerankingfunktion berechnet.

Die oben beschriebenen (Kapitel 4.1.3) Scoringfunktion von Martin[61] berücksichtigt die geometrische Passgenauigkeit der beiden interagierenden Proteine nur in der Form, dass die Ergebnisse zuerst nach dem SVM-optimierten umfassenden Scoringschema und anschließend nach der von *ckordo* berechneten geometrischen Korrelation sortiert werden. Daher wird auch die Kombination zwischen Martins Scoringfunktion und den hier vorgestellten Gewichtungsfaktoren evaluiert. Hierfür werden die Strukturen zuerst nach Martins Werten und anschließend nach den Werten sortiert, die die Multiplikation aus atom- und aminosäurespezifischen Scores ergeben. Des Weiteren wird eine Multiplikation aus Martins Scores und den mit den Gewichtungsfaktoren berechneten Scores evaluiert.

Martins Scoringfunktion selbst ist leider zurzeit nicht verfügbar, sondern nur die vorberechneten Ergebnisse für den mit 12° Winkelinkrement gedockten UUPPDD Datensatz (vgl.: Kapitel 4.5.1) und den mit 12° Winkelinkrement gedockten Benchmark 2.0 (vgl.: Kapitel 4.5.2). Die Kombination der hier vorgestellten Gewichtungsfaktoren mit Martins Filtern kann daher auch nur mit diesen Daten durchgeführt werden und nicht mit den für die Optimierung genutzten mit 15° gedockten Benchmark 2.0 Strukturen.

4.5 Datensätze

Für die Optimierung und Evaluation der Gewichtungsfaktoren werden zwei verschiedene Datensätze verwendet. Zum einen eine 34 Komplexe umfassende Sammlung zuvor in der Literatur[14,17,18,25,26,32,33,57,60,73] benutzter Beispiele, die 2003 von Martin und mir unter dem Titel *Unbound-Unbound Protein-Protein Docking Dataset* (UUPPDD) gesammelt und im Internet verfügbar gemacht wurden[35]. Der andere Datensatz wurde 2005 von Mintseris *et al.* veröffentlicht[66] und besteht aus 83 Beispielen.

Insgesamt gibt es relativ wenige nicht redundante Testfälle für die Entwicklung von *unbound* Docking Methoden, da für ein adäquates Testset sowohl die Strukturen des jeweiligen Komplexes als auch die unabhängig voneinander gemessenen Strukturen der Untereinheiten verfügbar sein müssen. Dies ist nötig, um eine realistische Vorhersage simulieren und evaluieren zu können.

4.5.1 UUPPDD

Der UUPPDD-Literaturdatensatz setzt sich aus 21 Enzym-Inhibitor, 4 Antikörper-Antigen und 4 ‚Anderen' Komplexen zusammen. Zusätzlich gibt es 4 als schwierig klassifizierte Komplexe, bei denen es im Laufe der Bindung zu größeren konformationellen Änderungen kommt, und die daher schwierig vorherzusagen sind (s. Tabelle 4.2).

Die Komplexe weisen untereinander höchstens eine Sequenzähnlichkeit von 75% (*positive hits* in BLAST[1]) auf, und die ungebundenen Proteine haben mindestens eine Sequenzähnlichkeit von 75% zu den jeweiligen gebundenen Untereinheiten.

Art der Interaktion[a]	Komplex			Rezeptor ungebunden			Ligand ungebunden		
	PDB-Code	Rezeptor Kette	Ligand Kette	PDB-Code	Kette	Anzahl an Aminosäuren	PDB-Code	Kette	Anzahl an Aminosäuren
Enzym-Inhibitor/Substrat Komplexe									
HYDROLASE(SERINE PROTEASE)	**1ACB**	E	I	**5CHA**	A	237	**1CSE**	I	63
PROTEINASE/INHIBITOR	**1AVW**	A	B	**2PTN**	-	223	**1BA7**	A	165
PROTEINASE/INHIBITOR	**1BRC**	E	I	**1BRA**	-	223	**1AAP**	A	56
ENDONUCLEASE	**1BRS**	A	D	**1A2P**	B	108	**1A19**	A	89
HYDROLASE/HYDROLASE INHIBITOR	**1BVN**	P	T	**1PIF**	-	495	**2AIT**	-	74
SERINE PROTEASE/INHIBITOR	**1CGI**	E	I	**1CHG**	-	230	**1HPT**	-	56
SERINE PROTEINASE	**1CHO**	E	I	**5CHA**	A	237	**2OVO**	-	56
SERINE PROTEASE/INHIBITOR	**1CSE**	E	I	**1SCD**	-	274	**1ACB**	I	63
ENDONUCLEASE/INHIBITOR	**1DFJ**	I	E	**2BNH**	-	456	**7RSA**	-	124
SERINE ESTERASE/TOXIN	**1FSS**	A	B	**2ACE**	-	527	**1FSC**	-	61
PHOSPHOTRANSFERASE	**1GLA**	G	F	**1BU6**	Y	499	**1F3G**	-	150
HYDROLASE/TOXIN	**1MAH**	A	F	**1MAA**	B	536	**1FSC**	-	61
SERINE PROTEINASE	**1PPF**	E	I	**1PPG**	E	218	**2OVO**	-	56
PROTEINASE/INHIBITOR	**1TGS**	Z	I	**2PTN**	-	223	**1HPT**	-	56
GLYCOSYLASE	**1UGH**	E	I	**1AKZ**	-	223	**1UGI**	A	83
PROTEINASE/INHIBITOR	**2KAI**	AB	I	**2PKA**	XY	232	**6PTI**	-	57
ELECTRON TRANSPORT	**2MTA**	LH	A	**2BBK**	LH	480	**1AAN**	-	105
OXIDOREDUCTASE/ELECTRON TRANSPORT	**2PCB**	A	B	**1CCP**	-	293	**1HRC**	-	105
OXIDOREDUCTASE/ELECTRON TRANSPORT	**2PCC**	A	B	**1CCA**	-	291	**1YCC**	-	107
PROTEINASE/INHIBITOR	**2PTC**	E	I	**2PTN**	-	223	**6PTI**	-	57
PROTEINASE/INHIBITOR	**2SNI**	E	I	**1SUP**	-	275	**2CI2**	I	65
Antikörper-Antigen Komplexe									
IMMUNOGLOBULIN/TISSUE FACTOR	**1AHW**	DE	F	**1FGN**	LH	428	**1BOY**	-	211
IMMUNE SYSTEM/HYDROLASE	**1DQJ**	AB	C	**1DQQ**	AB	424	**3LZT**	-	129
IMMUNOGLOBULIN/HYDROLASE(O-GLYCOSYL)	**1VFB**	AB	C	**1VFA**	AB	224	**1LZA**	-	129
ANTIKÖRPER/ELECTRON TRANSPORT	**1WEJ**	LH	F	**1QBL**	LH	433	**1HRC**	-	105
‚Andere' Komplexe									
MYRISTYLATION/TRANSFERASE	**1AVZ**	B	C	**1AVV**	-	99	**1SHF**	A	59
CHEMOTAXIS/TRANSFERASE	**1BDJ**	A	B	**3CHY**	-	128	**2A0B**	-	118
IMMUNE SYSTEM	**1L0Y**	A	B	**1BEC**	-	238	**1B1Z**	A	218
GTP-BINDING/GTPASE ACTIVATION	**1WQ1**	G	R	**1WER**	-	324	**5P21**	-	166
Schwierige Testfälle									
SERINE PROTEASE/INHIBITOR	**1BTH**	LH	P	**2HNT**	LCEF	292	**6PTI**	-	57
TRANSFERASE/CYCLIN	**1FIN**	A	B	**1HCL**	-	294	**1VIN**	-	252
HYDROLASE/TRANSFERASE	**1FQ1**	B	A	**1B39**	A	290	**1FPZ**	F	178
GTP-BINDING/TRANSDUCER	**1GOT**	BG	A	**1TBG**	AE	408	**1TAG**	-	314

[a] Art der Interaktion, wie im Header der jeweiligen PDB Datei angegeben

Tabelle 4.2: Unbound-Unbound Protein-Protein Docking Dataset[35]

4.5.2 Benchmark 2.0

Mintseris *et al.*[66] haben 2005 einen neuen Testdatensatz veröffentlicht, für den die PDB nach geeigneten Testfällen durchsucht wurde. Alle potentiellen transienten Komplexe der PDB wurden unter anderem so gefiltert, dass sie strukturell nicht redundant sind. Als redundant wurden die Strukturen angesehen, die zur gleichen SCOP-Familie[2] (Structural Classification of Proteins) gehören. Allen Komplexen wurden die ungebundenen Untereinheiten zugeordnet mit der höchsten Sequenzidentität, der niedrigsten Auflösung und den wenigsten fehlenden Residuen. Alle Komplexe wurden nach der biochemischen Funktionen und der Vorhersageschwierigkeit klassifiziert. Die Vorhersageschwierigkeit wurde durch ein FFT Docking ermittelt, bei dem die jeweilige Anzahl an gefundenen nahe nativen Strukturen als Schwierigkeitsmaß herangezogen wurde. Für 11 Antikörper-Antigen Komplexe konnten keine ungebundenen Strukturen für die jeweiligen Antikörper gefunden werden, so dass die Strukturen aus dem jeweiligen Komplex genutzt werden (*crossbound*).

Der Datensatz setzt sich aus 23 Enzym-Inhibitor/Substrat (rigid-body: 21, mittlere Schwierigkeit: 2), 22 Antikörper-Antigen (rigid-body: 9, crossbound: 11, crossbound mit mittlerer Schwierigkeit: 1, schwierig: 1) und 38 ‚Anderen' Komplexen (rigid-body:22, mittlere Schwierigkeit: 9, schwer: 7) zusammen (Tabelle 4.3).

Art der Interaktion[a]	Komplex			Rezeptor ungebunden			Ligand ungebunden		
	PDB-Code	Rezeptor Kette	Ligand Kette	**PDB-Code**	Kette	Anzahl an Aminosäuren	**PDB-Code**	Kette	Anzahl an Aminosäuren
Enzym-Inhibitor/Substrat Komplexe									
PROTEINASE/INHIBITOR	**1AVX**	A	B	**1QQU**	A	223	**1BA7**	B	169
ENZYME/INHIBITOR	**1AY7**	A	B	**1RGH**	B	96	**1A19**	B	89
HYDROLASE/HYDROLASE INHIBITOR	**1BVN**	P	T	**1PIG**	-	495	**1HOE**	-	74
SERINE PROTEASE/INHIBITOR COMPLEX	**1CGI**	E	I	**2CGA**	B	245	**1HPT**	-	56
HYDROLASE	**1D6R**	A	I	**2TGT**	-	223	**1K9B**	A	58
ENDONUCLEASE/INHIBITOR	**1DFJ**	I	E	**2BNH**	-	456	**9RSA**	B	124
OXIDOREDUCTASE	**1E6E**	A	B	**1E1N**	A	455	**1CJE**	D	107
SERINE PROTEASE/INHIBITOR	**1EAW**	A	B	**1EAX**	A	241	**9PTI**	-	58
OXIDOREDUCTASE	**1EWY**	A	C	**1GJR**	A	295	**1CZP**	A	98
HYDROLASE/INHIBITOR	**1EZU**	AB	C	**1ECZ**	AB	284	**1TRM**	A	223
HYDROLASE/HYDROLASE INHIBITOR	**1F34**	A	B	**4PEP**	-	326	**1F32**	A	127
PROTEASE/INHIBITOR	**1HIA**	AB	I	**2PKA**	XY	232	**1BX8**	-	49

HYDROLASE/TOXIN	**1MAH**	A	F	**1J06**	B	533	**1FSC**	-	61
HYDROLASE(SERINE PROTEINASE)	**1PPE**	E	I	**1BTP**	-	223	**1LU0**	A	29
HYDROLASE/HYDROLASE INHIBITOR	**1TMQ**	A	B	**1JAE**	-	470	**1B1U**	A	117
HYDROLASE/INHIBITOR	**1UDI**	E	I	**1UDH**	-	228	**2UGI**	B	83
ELECTRON TRANSPORT	**2MTA**	HL	A	**2BBK**	JM	480	**2RAC**	A	105
OXIDOREDUCTASE/ELECTRON TRANSPORT	**2PCC**	A	B	**1CCP**	-	293	**1YCC**	-	107
PROTEINASE/INHIBITOR	**2SIC**	E	I	**1SUP**	-	275	**3SSI**	-	108
PROTEINASE/INHIBITOR	**2SNI**	E	I	**1UBN**	A	274	**2CI2**	I	65
IMMUNE SYSTEM	**7CEI**	B	A	**1M08**	B	131	**1UNK**	D	87
Antikörper-Antigen Komplexe									
IMMUNOGLOBULIN/TISSUE FACTOR	**1AHW**	AB	C	**1FGN**	LH	428	**1TFH**	A	202
HUMANIZED ANTIKÖRPER/HYDROLASE	**1BVK**	DE	F	**1BVL**	BA	224	**3LZT**	-	129
IMMUNE SYSTEM/HYDROLASE	**1DQJ**	AB	C	**1DQQ**	CD	424	**3LZT**	-	129
HIV CAPSID PROTEIN (P24)	**1E6J**	HL	P	**1E6O**	HL	429	**1A43**	-	72
IMMUNE SYSTEM	**1JPS**	HL	T	**1JPT**	HL	425	**1TFH**	B	182
ANTIKÖRPER/ANTIGEN	**1MLC**	AB	E	**1MLB**	AB	432	**3LZT**	-	129
IMMUNOGLOBULIN/HYDROLASE(O-GLYCOSYL)	**1VFB**	AB	C	**1VFA**	AB	224	**8LYZ**	-	129
ANTIKÖRPER/ELECTRON TRANSPORT	**1WEJ**	HL	F	**1QBL**	HL	433	**1HRC**	-	105
HEMAGGLUTININ/IMMMUNOGLOBULIN	**2VIS**	AB	C	**1GIG**	LH	431	**2VIU**	ACE	960
Antikörper-Antigen Crossbund Komplexe									
ANTIKÖRPER/ANTIGEN	**1BJ1**	HL	VW	**1BJ1**	HL	431	**2VPF**	GH	189
IMMUNE SYSTEM	**1FSK**	BC	A	**1FSK**	BC	434	**1BV1**	-	159
CYTOKINE/IMMUNE SYSTEM	**1I9R**	HL	ABC	**1I9R**	HL	434	**1ALY**	ABC	438
IMMUNE SYSTEM/BLOOD CLOTTING	**1IQD**	AB	C	**1IQD**	AB	408	**1D7P**	M	159
MEMBRANE PROTEIN	**1K4C**	AB	C	**1K4C**	AB	431	**1JVM**	ABC D	394
HYDROLASE, IMMUNE SYSTEM	**1KXQ**	A	H	**1PPI**	-	496	**1KXQ**	H	120
HYDROLASE(O-GLYCOSYL)	**1NCA**	HL	N	**1NCA**	HL	435	**7NN9**	-	388
IMMUNOGLOBULIN/HYDROLASE	**1NSN**	HL	S	**1NSN**	HL	427	**1KDC**	-	137
IMMUNE SYSTEM	**1QFW**	HL	AB	**1QFW**	HL	224	**1HRP**	AB	196
IMMUNE SYSTEM	**1QFW**	IM	AB	**1QFW**	IM	229	**1HRP**	AB	196
ANTIKÖRPER/ANTIGEN	**2JEL**	HL	P	**2JEL**	HL	435	**1POH**	-	85
‚Andere' Komplexe									
TRANSPORT/NUCLEAR PROTEIN	**1A2K**	AB	C	**1OUN**	AB	246	**1QG4**	A	202
CAPSID PROTEIN/CYCLOSPORIN	**1AK4**	A	D	**2CPL**	-	164	**1E6J**	P	210
MHC I/PEPTIDE/CD8	**1AKJ**	AB	DE	**2CLR**	DE	375	**1CD8**	AB	228
ISOMERASE/PROTEIN KINASE	**1B6C**	B	A	**1IAS**	A	330	**1D6O**	A	107
TRANSFERASE	**1BUH**	A	B	**1HCL**	-	294	**1DKS**	A	76
SIGNALLING COMPLEX	**1E96**	B	A	**1HH8**	A	192	**1MH1**	-	183
TRANSFERASE	**1F51**	AB	E	**1IXM**	AB	343	**1SRR**	C	121
IMMUNOGLOBULIN	**1FC2**	D	C	**1FC1**	AB	414	**1BDD**	-	60
SIGNALING PROTEIN	**1FQJ**	A	B	**1TND**	C	316	**1FQI**	A	133
SIGNALING PROTEIN/SIGNALING PROTEIN	**1GCQ**	C	B	**1GCP**	B	67	**1GRI**	B	211
IMMUNE SYSTEM/VIRUS RECEPTOR	**1GHQ**	A	B	**1C3D**	-	294	**1LY2**	A	130
SIGNALLING COMPLEX	**1HE1**	C	A	**1MH1**	-	183	**1HE9**	A	131
SIGNALING PROTEIN	**1I4D**	AB	D	**1I49**	AB	402	**1MH1**	-	183
VIRUS/VIRAL PROTEIN/RECEPTOR	**1KAC**	A	B	**1NOB**	F	185	**1F5W**	B	121
IMMUNE SYSTEM/TOXIN	**1KLU**	AB	D	**1H15**	AB	369	**1STE**	-	238
CYTOKINE/CYTOKINE RECEPTOR	**1KTZ**	B	A	**1M9Z**	A	105	**1TGK**	-	112
CONTRACTILE PROTEIN/PROTEIN BINDING	**1KXP**	D	A	**1KW2**	B	453	**1IJJ**	B	371
IMMUNE SYSTEM	**1ML0**	AB	D	**1MKF**	AB	742	**1DOL**	-	70
IMMUNE SYSTEM	**1QA9**	A	B	**1HNF**	-	179	**1CCZ**	A	171
PROTEIN/PROTEIN	**1RLB**	ABCD	E	**2PAB**	ABCD	456	**1HBP**	-	175
IMMUNE SYSTEM	**1SBB**	B	A	**1SE4**	-	239	**1BEC**	-	238
ACETYLATION AND ACTIN-BINDING	**2BTF**	A	P	**1IJJ**	B	371	**1PNE**	-	140
Enzym-Inhibitor (mittlere Schwierigkeit)									
HYDROLASE(SERINE PROTEASE)	**1ACB**	E	I	**2CGA**	B	245	**1EGL**	-	70
TRANSFERASE,HYDROLASE/TRANSPORT PROTEIN	**1KKL**	ABC	H	**1JB1**	ABC	471	**2HPR**	-	86
Antikörper-Antigen Crossbund Komplexe (mittlere Schwierigkeit)									
POLYMERASE/INHIBITOR	**1BGX**	T	HL	**1CMW**	A	817	**1AY1**	HL	423
‚Andere' Komplexe (mittlere Schwierigkeit)									
COMPLEX (GTP-BINDING/TRANSDUCER)	**1GP2**	BG	A	**1TBG**	DH	405	**1GIA**	-	309
GENE REGULATION	**1GRN**	B	A	**1RGP**	-	189	**1A4R**	A	190
COMPLEX (PHOSPHOINOSITIDE KINASE/RAS)	**1HE8**	A	B	**1E8Z**	A	839	**821P**	-	166
CELL CYCLE	**1I2M**	B	A	**1A12**	A	401	**1QG4**	A	202

SIGNALING PROTEIN/TRANSFERASE	**1IB1**	AB	E	**1QJB**	AB	460	**1KUY**	A	166
BLOOD CLOTTING/TOXIN	**1IJK**	BC	A	**1FVU**	AB	254	**1AUQ**	-	208
SIGNALING PROTEIN/SIGNALING ACTIVATOR	**1K5D**	AB	C	**1RRP**	AB	338	**1YRG**	B	343
BLOOD CLOTTING	**1M10**	B	A	**1M0Z**	B	266	**1AUQ**	-	208
COMPLEX (GTP-BINDING/GTPASE ACTIVATION)	**1WQ1**	G	R	**1WER**	-	324	**6Q21**	D	171
‚Andere' Komplexe (schwierig)									
ENDODEOXYRIBONUCLEASE	**1ATN**	A	D	**1IJJ**	B	371	**3DNI**	-	258
METAL TRANSPORT INHIBITOR/RECEPTOR	**1DE4**	CF	AB	**1CX8**	AB	1278	**1A6Z**	AB	371
CYTOKINE/RECEPTOR	**1EER**	BC	A	**1ERN**	AB	414	**1BUY**	A	166
BLOOD CLOTTING	**1FAK**	HL	T	**1QFK**	HL	348	**1TFH**	B	182
HYDROLASE/TRANSFERASE	**1FQ1**	B	A	**1B39**	A	290	**1FPZ**	F	178
SMALL GTPASE	**1IBR**	B	A	**1F59**	A	440	**1QG4**	A	184
RT/DNA/FAB	**2HMI**	AB	CD	**1S6P**	AB	979	**2HMI**	CD	434
Antikörper-Antigen (schwierig)									
ACTIN-BINDING	**1H1V**	A	G	**1IJJ**	B	371	**1D0N**	B	729

[a] Art der Interaktion, wie im Header der jeweiligen PDB Datei angegeben

Tabelle 4.3: Benchmark 2.0[66]

4.6 Minimierung des quadratischen Fehlers mit dem R-Paket

Zur Optimierung der Gewichtungsfaktoren wird das R-Paket[77], ein open-source Statistik Paket, verwendet. Das Programm wird genutzt, da es die Bedingungen freier Verfügbarkeit, ausführlicher Dokumentation und Anpassungsfähigkeit an neue Fragestellungen erfüllt. Die Fähigkeit, große Datenmengen für die Optimierung von bis zu 41 Parametern zu verarbeiten, hat bei der Suche nach einem geeigneten, bereits implementierten Optimierungsverfahren die größte Schwierigkeit dargestellt, welche aber von der nlm()-Funktion des R-Paketes erfüllt wird.

4.6.1 *nlm()* Funktion

Die Optimierung entspricht einer Minimierung des quadratischen Fehlers zwischen tatsächlichem Score und gewünschtem Score (vgl. Kapitel 4.2.1). Für diese Fehlerminimierung wird die nichtlineare Minimierungsfunktion (nlm())[22,83] des R-Paketes genutzt, welcher eine Newton-Raphsonsche Methode[70,78] zu Grunde liegt. Die nlm()-Funktion versucht durch numerische Berechnungen der Ableitung das Minimum einer Funktion zu finden.

Die *nlm()*-Funktion kann mit verschiedenen Startparametern gestartet werden. Unterschiedliche Sets an Startparametern sind in Vortests ausprobiert worden, führten allerdings immer zu gleichen Ergebnissen. Die finalen Optimierungen werden daher mit 1 als Startwert für alle Gewichtungsfaktoren und - soweit mitoptimiert - mit -6 für I1 durchgeführt.

Die Optimierungen werden mit einer Genauigkeit von 2 Nachkommastellen (*ndigits*=2) durchgeführt. Die maximale Anzahl an Iterationen wird auf 1.000 gesetzt, welche aber bei keiner Optimierung erreicht wurden.

Als Stoppkriterien für die Optimierung werden die Standardparameter gewählt, so dass die Optimierung abbricht, wenn das Verhältnis der Änderung des Scores zu der Änderung der Parameter $1x10^{-6}$ unterschreitet.

4.7 Evaluation und Validierung

Die im UUPPDD gesammelten Komplexe werden zur Validierung der optimierten Parameter genutzt. Diese Beispiele werden nicht für die Optimierung genutzt, so dass diese für Dockingvorhersagen mit den optimierten Parametern unbekannten Komplexen entsprechen. Einige Komplexe kommen zwar in beiden Datensätzen (Benchmark 2.0 und UUPPDD) vor, allerdings jeweils mit verschiedenen Strukturen der ungebundenen Proteine. Daher spiegeln die Ergebnisse, die für diesen Datensatz erreicht werden, ein realistisches Maß für die erwartete Vorhersagequalität neuer Komplexe wieder.

Für die Validierung der Parameter als Postfilter werden mit *ckordo* für jeden Komplex 43.080 mögliche Strukturen generiert. Dies wird durch einen Dockinglauf mit einem Winkelinkrement von 12° und einem Behalten der 5 Translationen mit der höchsten geometrischen Komplementarität erreicht. Die Vorhersagequalität wird im Hinblick auf die Anreicherung der Anzahl nahe nativer Strukturen in den oberen Rängen und im Hinblick auf die Anzahl an Komplexen, die mindestens eine nahe native Struktur auf den oberen Rängen haben, durchgeführt. Ferner wird untersucht, ob sich durch den Einsatz der optimierten Gewichtungsfaktoren der Rang und der RMS-Wert der

ersten nahe nativen Struktur verändert hat. Die Evaluation der Vorhersagequalität mit optimierten Parametern in *ckordo* selbst umfasst ferner den RMS-Wert der besten Struktur und die Anzahl an nahe nativen Strukturen, die gefunden werden.

Der Nachweis, dass die jeweiligen optimierten Parameter allgemeine Gültigkeit haben - unabhängig von dem für die Optimierung benutzten Datensatz - wird durch eine bereits oben beschriebene 5-fache Crossvalidierung (Kapitel 4.2.1) erbracht.

Für die Evaluierung des Effektes der Faktoren auf die Vorhersagequalität wird der Durchschnittswert der jeweils 5 Optimierungen benutzt.

Die Verbesserung der Vorhersagequalität wird zum Abschluss mit der Vorhersagequalität von 6 anderen Scoringfunktionen verglichen. Für diesen Vergleich wird die Anreicherung nahe nativer Strukturen auf den niedrigen Rängen durch Atomic Contact Energies (ACE)[100], durch ein Residuen-Residuen Potential[67], durch ein Atom-Atom Potential[30], durch die komplexklassen-spezifische Residuen Interface-*Propensity* basierte Scoringfunktion von Huang *et al.*[37], durch die Berechnung der Packungsdichte[27,37] sowie durch die umfassende Scoringfunktion von Martin[61] für den UUPPDD Datensatz berechnet.

4.8 Das Qualitätskriterium – der RMSD-Wert

Als Qualitätskriterium für die verschiedenen Konformationen dient der RMSD-Wert (*root mean square deviation*) der Interface C_α-Atome zwischen der jeweils vorhergesagten Orientierung der ungebundenen Proteine im Raum und den auf den Komplex gelegten (*gefitteten*) ungebundenen Strukturen.

Als Interfaceatome werden die Atome definiert, die mindestens ein Atom des Bindungspartners in einem Abstand von 6 Å haben. Für die qualitative Bewertung der vorhergesagten Komplexe werden nur die Interface Atome herangezogen, da die korrekte Vorhersage der relativen Position dieser Atome wesentlich interessanter ist als die Position von Atomen, die weit entfernt vom Interface liegen. Außerdem können z.B. kleinere Rotationen um das Interface zu enormen Distanzen bei

Residuen führen, die weit vom Interface entfernt sind, obwohl das Interface selbst richtig vorhergesagt wurde.

Auf Grund der bereits beschriebenen relativ häufigen Konformationsänderung der Seitenketten wird der RMS-Wert nur für die C_α-Atome berechnet.

Als Referenz für die RMS-Berechnung dienen die auf den nativen Komplex gelegten ungebundenen Proteine. Das strukturelle *Alignement* zwischen Komplex und ungebundenen Proteinen wurde mit dem Programm *CE*[89] durchgeführt. Da in *ckordo* keinerlei Flexibilität explizit behandelt wird, ist die strukturelle Ähnlichkeit zum auf den Komplex gelegten ungebundenen Protein die maximal erreichbare Qualität, so dass eine RMS-Berechnung relativ zum gebundenen Komplex nicht sinnvoll ist.

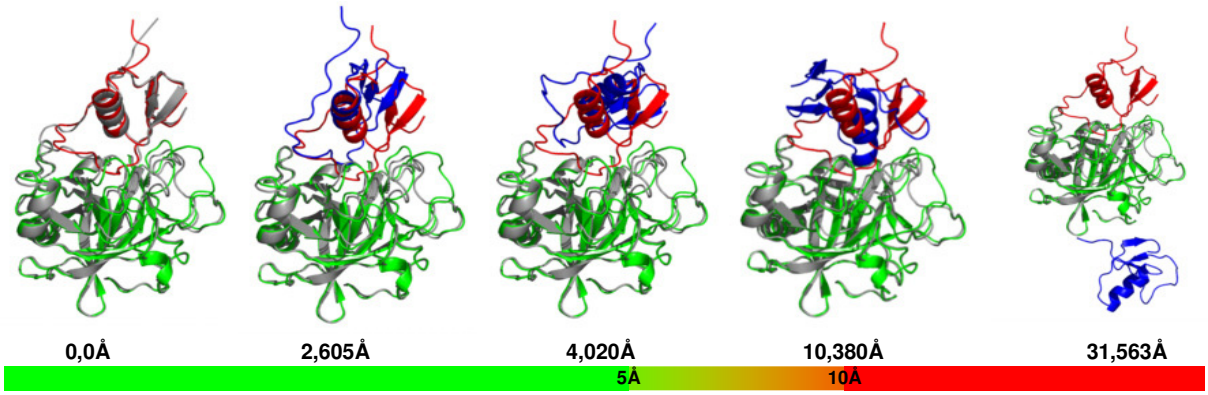

Abbildung 4.7: Strukturen von 1CGI mit unterschiedlichem RMSiCα. Die Struktur ganz links zeigt die ungebundenen Proteine (grün/rot) auf den nativen Komplex (grau) gelegt. In den folgenden Strukturen sind die jeweiligen Orientierungen des Liganden in blau gezeigt. Unter den Strukturen ist der jeweilige RMSiCα angegeben.

Als nahezu native Komplexe werden solche Konformationen behandelt, die einen Interface C_α RMS-Wert von bis zu 5 Å aufweisen. Abbildung 4.7 zeigt verschiedene von *ckordo* vorgeschlagene Komplexstrukturen für 1CGI mit unterschiedlichen RMS-Werten und dient der Visualisierung der RMSD-Werte.

4.9 CAPRI

CAPRI steht für “Critical Assessment of PRedicted Interactions”. CAPRI ist ein Experiment, welches es allen Arbeitsgruppen, die an der Entwicklung von Protein-

Protein Dockingmethoden arbeiten, ermöglicht, ihre Methoden mit denen der anderen Arbeitsgruppen zu vergleichen. Dafür werden unbekannte Komplexstrukturen, deren Untereinheiten bereits bekannt sind oder modelliert werden müssen, von experimentell arbeitenden Wissenschaftlern vor der Veröffentlichung zur Verfügung gestellt. Die Teilnehmer haben 2-3 Wochen Zeit, ihre Strukturvorschläge zur Auswertung an das European Bioinformatics Institute (EBI) in Cambridge zu übermitteln. Alle vorgeschlagenen Strukturen werden zentral evaluiert.

Es gibt 4 verschiedene Qualitätskategorien, in welche die Strukturvorschläge eingeteilt werden: *High-Quality, Medium, Akzeptabel* und *Falsch*. Für die Evaluation werden folgende Kriterien genutzt: Anzahl der nativen und der nicht nativen Residuum-Residuum Kontakte geteilt durch die gesamte Anzahl an Kontakten (f_{nat}, $f_{non\text{-}nat}$), der Anteil an nativen Kontakten im vorhergesagten Interface (f_{IR}), RMSD des Liganden (L_rms), Interface RMSD (I_rms) und die Rotation und Translation, die nötig sind, um den vorhergesagten Liganden auf den nativen Liganden zu platzieren, wenn der vorhergesagte Rezeptor perfekt auf den nativen Rezeptor gelegt wurde (Θ_L, d_L). Die Einteilung in die 4 Klassen erfolgt nach den in Tabelle 4.4 gezeigten Grenzwerten.[64]

Um die hier vorgestellte Methode und die SVM-basierte umfassende Scoringfunktion von Martin[61] auszuwerten, wurden Vorhersagen für die Targets 24-26 durchgeführt.

Kategorie	Kriterien
High-Quality	$f_{nat} \geq 0{,}5$ UND (L_rms ≤ 1,0 ODER I_rms ≤ 1,0)
Medium	($f_{nat} \geq 0{,}3$ UND $f_{nat} \leq 0{,}5$) UND (L_rms ≤ 5,0 ODER I_rms ≤ 2,0) ODER $f_{nat} \geq 0{,}5$ UND L_rms > 1,0 UND I_rms > 1,0
Akzeptabel	($f_{nat} \geq 0{,}1$ UND $f_{nat} \leq 0{,}3$) UND (L_rms ≤ 10,0 ODER I_rms ≤ 4,0) ODER $f_{nat} \geq 0{,}3$ UND L_rms > 5,0 UND I_rms > 2,0
Falsch	$f_{nat} < 0{,}1$ ODER (L_rms > 10,0 ODER I_rms > 4,0)

Tabelle 4.4: Evaluationskriterien für CAPRI-Experiment [64]

4.9.1 Targets 24, 25 und 26

Der Proteinkomplex, der in Runde 9 von CAPRI vorhergesagt werden sollte, war für die beiden Targets 24 und 25 identisch. Beide Male sollte der Komplex aus ADP-Ribosilierungs Faktor 1 (Arf1-GTP) und der Arf Bindungs Domäne (ArfBD) von ArhGAP10 (Rho GTPase aktivierendes Protein 10) vorhergesagt werden.

Für Target 24 sollte 1O3Y Kette A (Arf1-GTP) gegen ein selbst zu erstellendes Strukturmodell von ArfBD gedockt werden. Für die Erstellung des Modells war als Information gegeben, dass ArfBD homolog ist zu der PH-Domäne von Beta-Spektrin (1BTN). Die Erstellung des Modells wurde mit dem vollautomatischen Strukturvorhersageserver SWISS-MODEL[86,87] durchgeführt.

Für Target 25 wurde die zu modellierende Struktur von ArfBD durch die entsprechende Kristallstruktur aus dem vorherzusagenden Komplex ersetzt, die von Menetrey (Institut Curie, Paris) gelöst und zur Verfügung gestellt wurde.

Target 26 ist ein Komplex zwischen den *E. coli* Proteinen TolB (1C5K) und Pal (1OAP). Die vorherzusagende Komplexstruktur wurde von Kleanthous (University of York, UK) gelöst und zur Verfügung gestellt. In der Literatur finden sich Hinweise aus Experimenten auf die Bindungsstellen beider Proteine. So sollen die Residuen 89-104 und 126-130 von PAL[19,79] und der C-Terminale Teil von TolB[79] an der Interaktion beteiligt sein.

Das Docking wurde mit einem Winkelinkrement von 12° mit *ckordo* durchgeführt und die dadurch von *ckordo* vorgeschlagenen Strukturen anschließend mit der umfassenden Scoringfunktion von Martin[61] und den in dieser Arbeit beschriebenen aminosäurespezifischen Gewichtungsfaktoren bewertet. Da beide Methoden zum Zeitpunkt der Teilnahme am CAPRI-Wettbewerb noch nicht vollständig fertig entwickelt und evaluiert waren, erfolgte die Auswertung per Hand. Dabei wurden die 10 Strukturen, die zur Auswertung geschickt werden dürfen, so ausgewählt, dass diese von beiden Filtern eine hohe Bewertung erhalten und zudem möglichst divers sind. Die vorgeschlagenen Strukturen für Target 26 wurden des Weiteren so

ausgewählt, dass die Residuen Teil des Interfaces sind, deren Beteiligung an der Interaktion vorher experimentell nachgewiesen wurde.

5 Ergebnisse

5.1 Ckordo Ergebnisse / Datengrundlage

Bei einem Winkelinkrement von 15° findet *ckordo* nur anhand der geometrischen Korrelation im Schnitt 33 nahe native Strukturen für Enzym-Inhibitor Komplexe, 10 für Antikörper-Antigen Komplexe und 10 für die ‚Anderen'. Bei den Enzym-Inhibitor Komplexen wird im Schnitt die erste nahe native Struktur mit einem RMSiCα ≤ 5 Å auf Rang 1.222 von 22.000 gefunden. Bei den Antikörper-Antigen Komplexen ist der durchschnittliche Rang der ersten nahe nativen Lösung Rang 3.119 und bei den ‚Anderen' Komplexen Rang 3.088. Der Mittelwert des jeweils niedrigsten RMS-Wertes ist bei den EI-Komplexen 2,34 Å, bei den Antikörper-Antigen 2,98 Å und bei den ‚Anderen' 3,28 Å.

Für einige Komplexe kann *ckordo* keine nahe native Lösung finden. Die beste Struktur, die für 1EZU (ein EI Komplex) gefunden wird, hat einen RMSiCα von 6.57 Å. Für die drei Antikörper-Antigen Komplexe 1BGX, 1H1V und 1K4C liegt der niedrigste RMS-Wert jeweils bei 9,66 Å, 13,33 Å bzw. 5,57 Å. Auch für 7 der ‚Anderen' Komplexe kann *ckordo* bei einem Winkelinkrement von 15° keine Lösung mit einem RMSiCα ≤ 5 Å finden (1ATN (6,493 Å), 1H1V (13,33 Å), 1FAK (6,91 Å), 1I4D (5,88 Å), 1IB1 (6,04 Å), 1IBR (6,85 Å), 1M10 (5,38 Å)).

Tabelle 5.1 gibt einen Überblick, für welchen Komplex wie viele gute Lösungen gefunden werden, auf welchem Rang die Struktur mit dem niedrigsten RMS-Wert gefunden wird und auf welchen Rang die bestplatzierte Struktur liegt.

	Erste nahe native Struktur		Beste Struktur		Anzahl nahe nativer Strukturen
	Rang	RMS	Rang	RMS	
Enzym-Inhibitor					
1ACB	653	4,87	15781	2,52	15
1AVX	203	4,8	19748	1,52	41
1AY7	7	2,59	177	1,73	20
1BVN	100	2,31	1894	2,18	23
1CGI	107	4,53	12142	2,6	58
1D6R	621	3,69	16882	2,71	38
1DFJ	3815	4,9	8633	2,37	3
1E6E	1074	4,89	2086	3,24	10
1EAW	101	3,24	13305	2,38	44
1EWY	479	4,92	1912	2,56	34
1EZU	-	-	*14300*	*6,57*	*0*
1F34	7448	3,85	8233	2,63	5
1HIA	75	4,81	2629	3,25	43
1KKL	2027	3,52	5898	0,46	14
1MAH	864	1,6	8450	1,45	12
1PPE	6	4,4	13238	1,12	214
1TMQ	28	1,05	28	1,05	15
1UDI	249	4,25	15189	1,86	16
2MTA	838	3,53	18881	1,36	18
2PCC	7589	4,66	7856	3,86	5
2SIC	15	4,82	7288	1,84	40
2SNI	419	2,53	10529	2,12	44
7CEI	167	4,17	6653	2,34	14
Antikörper-Antigen					
1AHW	487	2,34	1890	0,91	10
1BGX	-	-	*1616*	*9,67*	*0*
1BJ1	18489	3,96	18489	3,96	1
1BVK	2140	4,57	12258	1,49	16
1DQJ	5318	3,61	5318	3,61	8
1E6J	146	4,05	19139	1,51	23
1FSK	1206	1,78	3213	1,75	16
1H1V	-	-	*4999*	*13,33*	*0*
1I9R	81	3,84	891	1,44	11
1IQD	1765	3,41	21710	1,38	4
1JPS	613	1,7	8021	1,46	14
1K4C	-	-	*8782*	*5,57*	*0*
1KXQ	6	2,34	862	1,42	17
1MLC	6322	1,19	6322	1,19	3
1NCA	718	3,08	3134	1,15	7
1NSN	1655	2,89	8606	1,21	7
1QFW a	2575	1,96	20349	1,01	9
1QFW b	3228	4,32	4072	0,74	14
1VFB	2430	4,65	11259	1,59	15
1WEJ	6348	4,69	12087	3,69	8
2JEL	2610	4,42	2610	4,42	2
2VIS	-	-	4643	6,99	0
Andere					
1A2K	338	4,77	14123	4,08	13
1AK4	3589	4,75	10904	3,61	24
1AKJ	76	3,5	632	2,61	16
1ATN	-	-	*20146*	*6,49*	*0*
1B6C	467	2,42	467	2,42	9
1BUH	1425	4,71	20926	2,56	4
1DE4	17410	2,31	17410	2,31	1
1E96	787	4,03	14339	2,88	11
1EER	3261	4	3261	4	3
1FAK	-	-	*10111*	*6,91*	*0*
1F51	819	1,92	5299	0,43	26
1FC2	1122	4,75	3627	2,72	15
1FQ1	3747	3,65	15922	2,84	5
1FQJ	231	3,98	19952	2,62	9
1GCQ	1653	2,11	12014	1,92	14
1GHQ	3547	4,17	11109	1,36	14
1GP2	3917	1,71	10924	1,53	10
1GRN	250	4,5	2210	1,42	15
1H1V	-	-	*4999*	*13,33*	*0*
1HE1	5691	3,16	8685	1,39	7
1HE8	106	2,08	7352	0,99	10

1I4D	-	-	*167*	*5,88*	*0*
1I2M	131	3,14	131	3,14	7
1IJK	597	4,99	15441	2,09	10
1K5D	9264	4,94	15180	4,12	2
1KAC	345	2,85	2585	2,35	15
1KLU	14603	3,94	14739	2,91	4
1KTZ	9816	3,37	13686	1,1	6
1KXP	37	1,29	37	1,29	7
1M10	-	-	*2277*	*5,38*	*0*
1ML0	1109	4,62	1449	2,15	6
1QA9	6120	3,45	9438	1,74	7
1RLB	205	4,81	5120	3,5	8
1SBB	1804	4,24	9182	2,19	12
1WQ1	165	3,23	5396	2,36	10

Tabelle 5.1: ckordo Ergebnisse für Docking Benchmark 2.0

5.2 Gewichtungsfaktoren

5.2.1 Aminosäurespezifische Gewichtungsfaktoren (Reranking)

Tabelle 5.2 zeigt für die drei Komplexklassen die optimierten Gewichtungsfaktoren und die Werte, die dem Inneren des größeren Proteins (I1) zugewiesen werden. Die Werte, die als 0,00 dargestellt sind, sind nicht exakt Null, sondern liegen im Bereich zwischen $1 \cdot 10^{-4}$ und $8 \cdot 10^{-4}$.

Es gibt zwei Gruppen von Aminosäuren, bei denen die optimierten Werte für alle drei Gruppen vergleichbar sind. Zum einen werden für die aromatischen und hydrophoben Aminosäuren (TRP, TYR, PHE, ILE, VAL) für alle Komplexe hohe Werte berechnet. Zum anderen hat die Optimierung für die Aminosäuren, mit langen flexiblen Seitenketten (ARG, GLN, GLU, LYS) und für die Aminosäuren mit einer geringen Neigung (*Propensity*) im Interface zu liegen (ASP, SER, PRO), sehr niedrige Werte ergeben.

Die optimierten Werte für die anderen Aminosäuren unterscheiden sich bei den drei Komplexklassen. Während ALA, ASN und HIS bei der Identifikation nahe nativer Antikörper-Antigen Strukturen eine wichtige Rolle spielen, hat die Optimierung für MET einen extrem hohen Wert bei den Enzym-Inhibitor Komplexen ergeben. Für LEU und THR wurden für die EI und die ‚Anderen' Komplexe Werte im mittleren

Bereich errechnet, während diese Werte für Antikörper-Antigen Komplexe bei Null liegen.

Aminosäure	Antikörper/ Antigen	Enzym/Inhibitor	‚Andere'
ALA	2,32	0,68	0,74
ARG	0,00*	1,70	1,09
ASN	5,23	0,49	0,92
ASP	0,00*	0,00*	0,00*
CYS	0,14	1,11	0,18
GLN	0,00*	0,00*	1,06
GLU	0,15	0,00*	0,00*
GLY	2,49	1,47	7,24
HIS	8,74	3,17	0,00*
ILE	6,83	2,30	4,34
LEU	0,02	3,24	4,83
LYS	0,67	0,59	1,29
MET	1,46	9,63	3,90
PHE	4,46	3,14	7,89
PRO	1,98	0,00*	1,13
SER	0,00*	0,00*	0,00*
THR	0,25	2,67	3,08
TRP	4,99	5,19	2,56
TYR	11,64	4,94	8,83
VAL	3,10	2,78	3,02
I1	-3,45	8,34	0,15

Tabelle 5.2: Atomspezifische Gewichtungsfaktoren für Scoring-Funktion.
*Werte sind nicht exakt Null; zwischen $1 \cdot 10^{-4}$ und $8 \cdot 10^{-4}$

Insbesondere die Werte für das Innere des Rezeptors (I1) unterscheiden sich massiv. Dieser Wert ist bei den Enzym-Inhibitor Komplexen 8,34, für die Antikörper-Antigen Komplexe -3,45 und für die ‚Anderen' Komplexe 0,15.

5.2.2 Aminosäurespezifische Gewichtungsfaktoren (ckordo)

Um die Gewichtungsfaktoren nicht nur als Rerankingfunktion zu nutzen, sondern diese auch direkt bei der Berechnung der geometrischen Korrelation anwenden zu können, wurde die Optimierung der Gewichtungsfaktoren wiederholt, ohne dabei den

Wert für das Innere des Rezeptors (I1) mitzuoptimieren. Die daraus erhaltenen Gewichtungsfaktoren sind in Tabelle 5.3 gezeigt.

Die aminosäurespezifischen Gewichtungsfaktoren korrelieren stark mit den Gewichtungsfaktoren für die Rerankingfunktion (Korrelationskoeffizienten Antikörper-Antigen: 0,97, Enzym-Inhibitor: 0,71, ‚Andere': 0,79), so dass die oben gegebene Beschreibung der Parameter auch hier zutrifft.

Aminosäure	Antikörper/ Antigen	Enzym/Inhibitor	Andere
ALA	2,69	0,72	1,64
ARG	0,01	0,77	1,55
ASN	5,27	1,35	2,69
ASP	0,01	0,98	1,09
CYS	0,14	4,77	0,02
GLN	0,00*	0,45	1,29
GLU	1,09	0,28	0,75
GLY	3,06	1,15	4,53
HIS	8,82	7,34	0,00*
ILE	7,96	3,58	0,97
LEU	0,09	2,82	7,13
LYS	1,01	0,00*	1,51
MET	3,73	6,17	3,51
PHE	2,92	0,88	5,25
PRO	0,74	2,53	0,66
SER	0,00*	0,84	0,29
THR	0,60	5,50	2,64
TRP	4,88	7,97	3,04
TYR	12,06	6,67	11,23
VAL	4,44	5,88	7,85

Tabelle 5.3: Aminosäurespezifische Gewichtungsfaktoren für ckordo.
*Werte sind nicht exakt Null; zwischen $1 \cdot 10^{-3}$ und $1 \cdot 10^{-4}$

5.2.3 Atomspezifische Gewichtungsfaktoren (Reranking)

Die optimierten Parameter für die 40 verschiedenen Atomtypen nach Melo *et al.*[62] sind in Tabelle 5.4 für die drei verschiedenen Komplexklassen gezeigt. Zu

Visualisierungszwecken sind die erhaltenen Werte für die Abbildung 4.1 - 4.3 in 4 Klassen eingeteilt worden (Niedrige Werte: 0 - 1, Mittlere Werte: 1 - 5, Hohe Werte: 5 - 10 und sehr hohe Werte: > 10) und in verschiedenen Farben auf die Strukturformeln der Aminosäuren gelegt worden (Abbildungen 5.1-5.3).

Auffällig ist zunächst, dass die Atome des Backbones für alle drei Komplexklassen einen sehr niedrigen gegen Null gehenden Wert zugewiesen bekommen (Atomklassen: 1, 3, 4). Das Sauerstoffatom des Backbones (Klasse 5) spielt bei der Berechnung der gewichteten geometrischen Korrelation nur bei den Antikörper-Antigen Komplexen eine Rolle. Das Cα-Atom von Glycin bildet eine eigene Atomklasse (2) und hat für alle drei Komplexklassen einen mittleren bis sehr hohen Wert erhalten.

Übereinstimmend bei allen drei Komplexklassen ist außerdem, dass für die Kohlenstoffatome der aromatischen Ringe von TRP, TYR, HIS und PHE hohe bis sehr hohe Werte optimiert werden.

Die Atome der eher kurzen hydrophoben Seitenketten von ALA, ILE, LEU und VAL (Atomtypen 6 und 7) erhalten bei allen drei Klassen mittlere bis hohe Werte (6-9).

Die Methylengruppen in den Seitenketten (Atomtyp 8) erhalten nach der Optimierung nur bei den Enzym-Inhibitor Komplexen einen mittelhohen Wert von 2,6, während die Werte bei AA- und OTH-Komplexen gegen Null gehen.

Bei den Enzym-Inhibitor Komplexen erhalten die endständigen Atome von Methionin sehr hohe Werte, während bei den Antikörper-Antigen Komplexen nur das Schwefel Atom und bei den ‚Anderen' Komplexen nur das Cε einen sehr hohen Wert erhält.

Die Seitenkettenatome der langen und flexiblen Seitenketten von ARG, LYS, ASP, und GLU erhalten bis auf wenige Ausnahmen eher niedrige Werte. Auffällig ist, dass bei Enzym-Inhibitor Komplexen für ARG Cζ und für Nε hohe Werte optimiert wurden.

Atomtyp	Antikörper-Antigen	Standard-abweichung	Enzym-Inhibitor	Standard-abweichung	Andere	Standard-abweichung
1	**0,00**	0,00	**0,01**	0,01	**0,00**	0,00
2	**8,62**	3,45	**3,30**	1,78	**14,88**	2,65
3	**0,00**	0,00	**0,04**	0,04	**0,00**	0,00
4	**0,01**	0,00	**0,01**	0,01	**0,00**	0,00
5	**1,28**	0,85	**0,07**	0,09	**0,00**	0,00
6	**7,50**	0,39	**6,18**	0,95	**3,75**	0,83
7	**6,52**	1,37	**6,20**	2,96	**9,52**	4,01
8	**0,18**	0,28	**2,61**	0,32	**0,00**	0,00
9	**10,08**	8,49	**25,77**	8,78	**5,59**	6,05
10	**4,78**	2,12	**0,07**	0,03	**11,03**	9,92
11	**9,21**	9,07	**19,61**	7,27	**17,64**	12,79
12	**10,00**	1,48	**8,81**	1,12	**8,76**	2,20
13	**3,51**	4,17	**2,08**	3,28	**0,02**	0,03
14	**1,93**	2,96	**6,81**	7,19	**11,03**	11,03
15	**0,00**	0,00	**0,00**	0,01	**3,14**	2,71
16	**0,00**	0,00	**1,83**	1,64	**6,88**	1,77
17	**0,00**	0,00	**6,29**	1,18	**2,90**	2,81
18	**5,31**	3,16	**2,28**	1,19	**0,00**	0,00
19	**0,34**	0,33	**5,28**	5,37	**5,72**	7,51
20	**0,24**	0,38	**0,67**	0,75	**0,00**	0,00
21	**0,01**	0,00	**11,61**	5,29	**0,01**	0,02
22	**1,18**	1,86	**0,01**	0,00	**6,29**	2,45
23	**11,29**	5,60	**25,18**	13,60	**1,68**	2,57
24	**19,60**	5,97	**14,26**	5,46	**0,00**	0,00
25	**0,73**	0,93	**1,56**	2,19	**2,90**	3,88
26	**0,03**	0,03	**2,79**	1,98	**0,01**	0,01
27	**0,26**	0,40	**0,01**	0,01	**0,00**	0,00
28	**3,31**	1,37	**0,02**	0,01	**0,00**	0,00
29	**0,05**	0,05	**5,18**	4,22	**0,00**	0,00
30	**1,84**	2,00	**16,88**	4,50	**11,23**	5,73
31	**0,18**	0,24	**30,84**	5,48	**8,21**	6,79
32	**4,91**	2,02	**0,05**	0,03	**0,11**	0,22
33	**5,66**	0,99	**1,32**	1,51	**6,80**	2,73
34	**0,95**	1,49	**0,01**	0,01	**9,00**	5,60
35	**1,22**	1,33	**0,01**	0,00	**9,92**	2,99
36	**0,04**	0,05	**10,91**	4,36	**1,56**	1,53
37	**0,03**	0,04	**0,30**	0,33	**0,00**	0,00
38	**20,75**	7,43	**1,61**	2,39	**3,15**	4,17
39	**8,46**	5,35	**0,03**	0,03	**0,00**	0,00
40	**23,18**	1,49	**5,46**	1,39	**18,08**	7,25
I1	**-0,87**	0,29	**0,67**	0,57	**0,70**	0,28

Tabelle 5.4: Durchschnittliche optimierte Gewichtungsfaktoren nach Atomtypen und deren Standardabweichungen (aus 5-facher Crossvalidierung) für die drei Komplexklassen

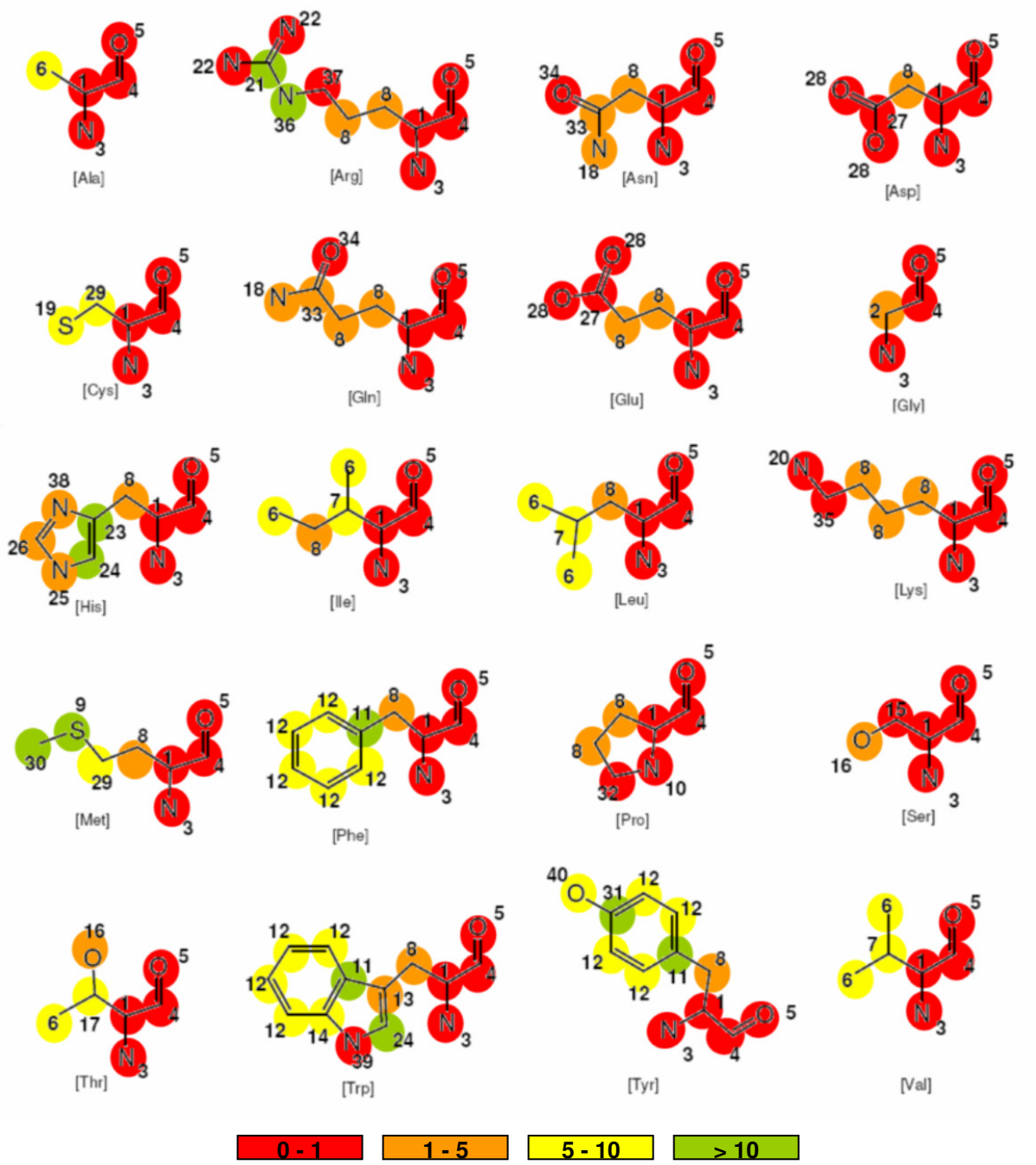

Abbildung 5.1: Visualisierung der atomtypenspezifischen Gewichtungsfaktoren für **Enzym-Inhibitor Komplexe**. Die verschiedenen Farben zeigen verschiedene Gewichtungsfaktoren an.

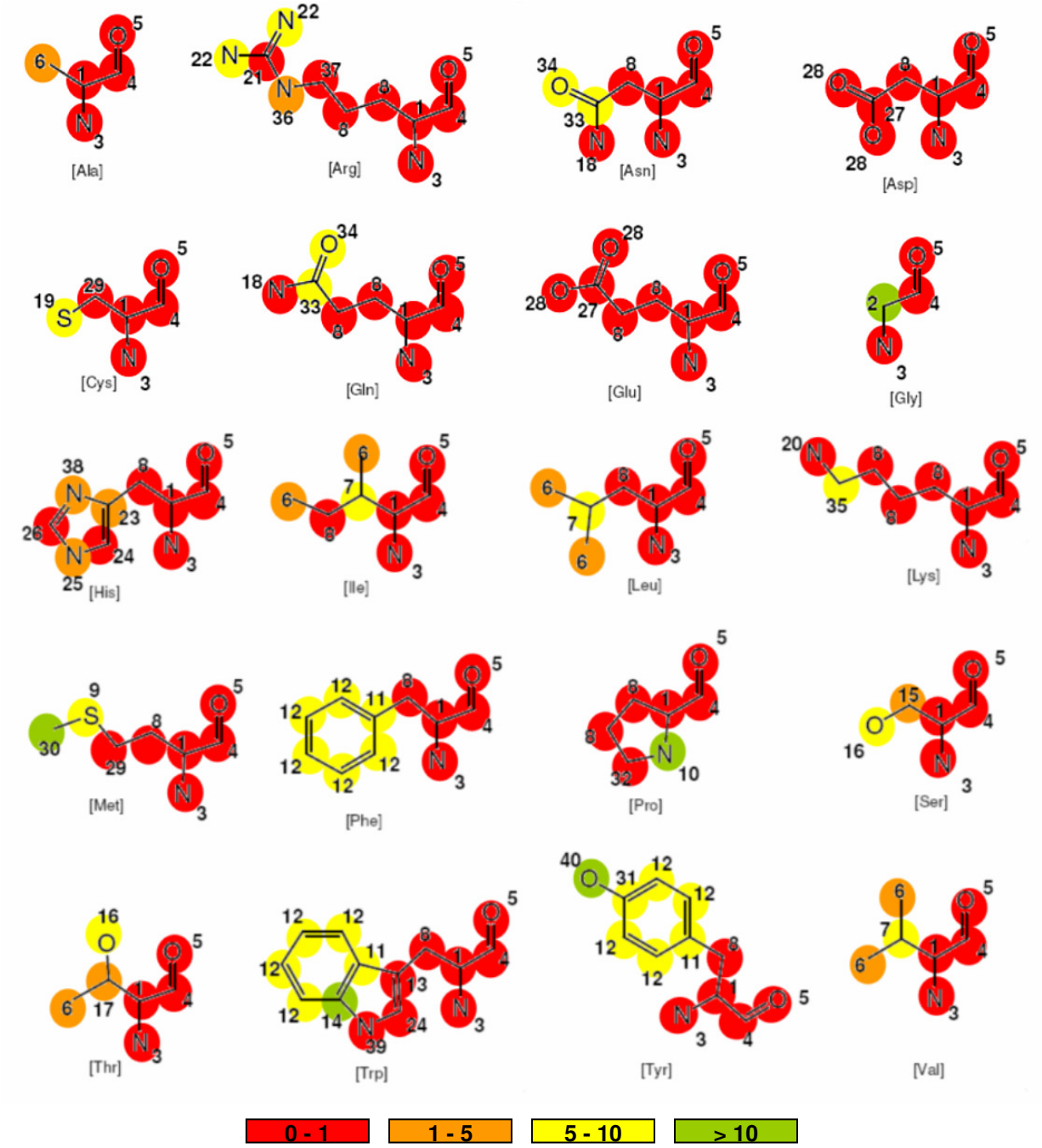

Abbildung 5.2: Visualisierung der atomtypenspezifischen Gewichtungsfaktoren für **Andere Komplexe**. Die verschiedenen Farben zeigen verschiedene Gewichtungsfaktoren an.

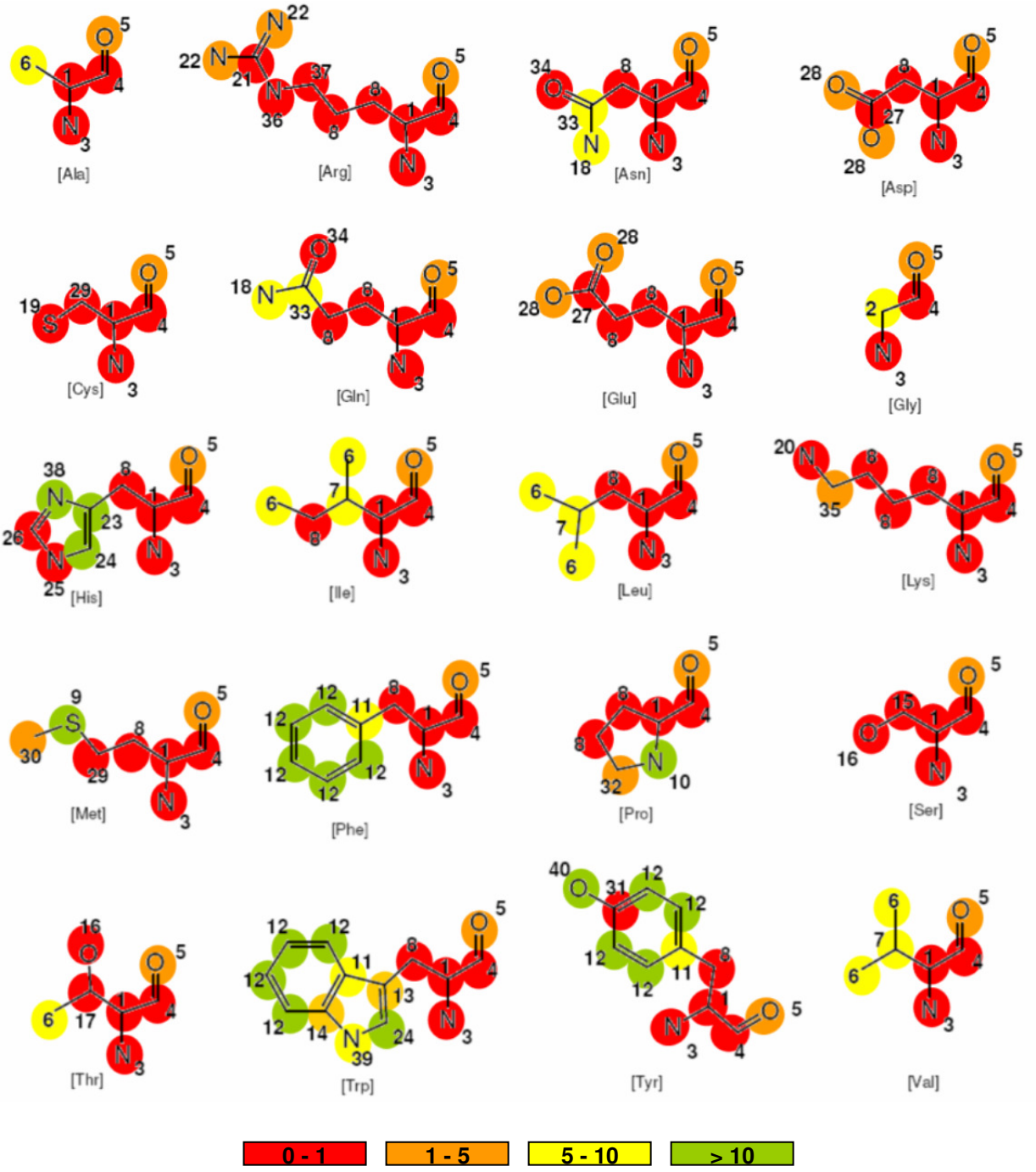

Abbildung 5.3: Visualisierung der atomtypenspezifischen Gewichtungsfaktoren für **Antikörper-Antigen** Komplexe. Die verschiedenen Farben zeigen verschiedene Gewichtungsfaktoren an.

5.3 Reranking mit Gewichtungsfaktoren

5.3.1 Enzym-Inhibitor/Substrat Komplexe

Wie in Abbildung 5.4 zu sehen ist, sind sowohl die für Atomtypen als auch die für Aminosäuren spezifischen Gewichtungsfaktoren in der Lage, die Anzahl der Strukturen, die auf den vorderen Rängen gefunden werden, erheblich zu steigern. In der Abbildung ist gezeigt, wieviel Prozent aller von *ckordo* vorgeschlagenen

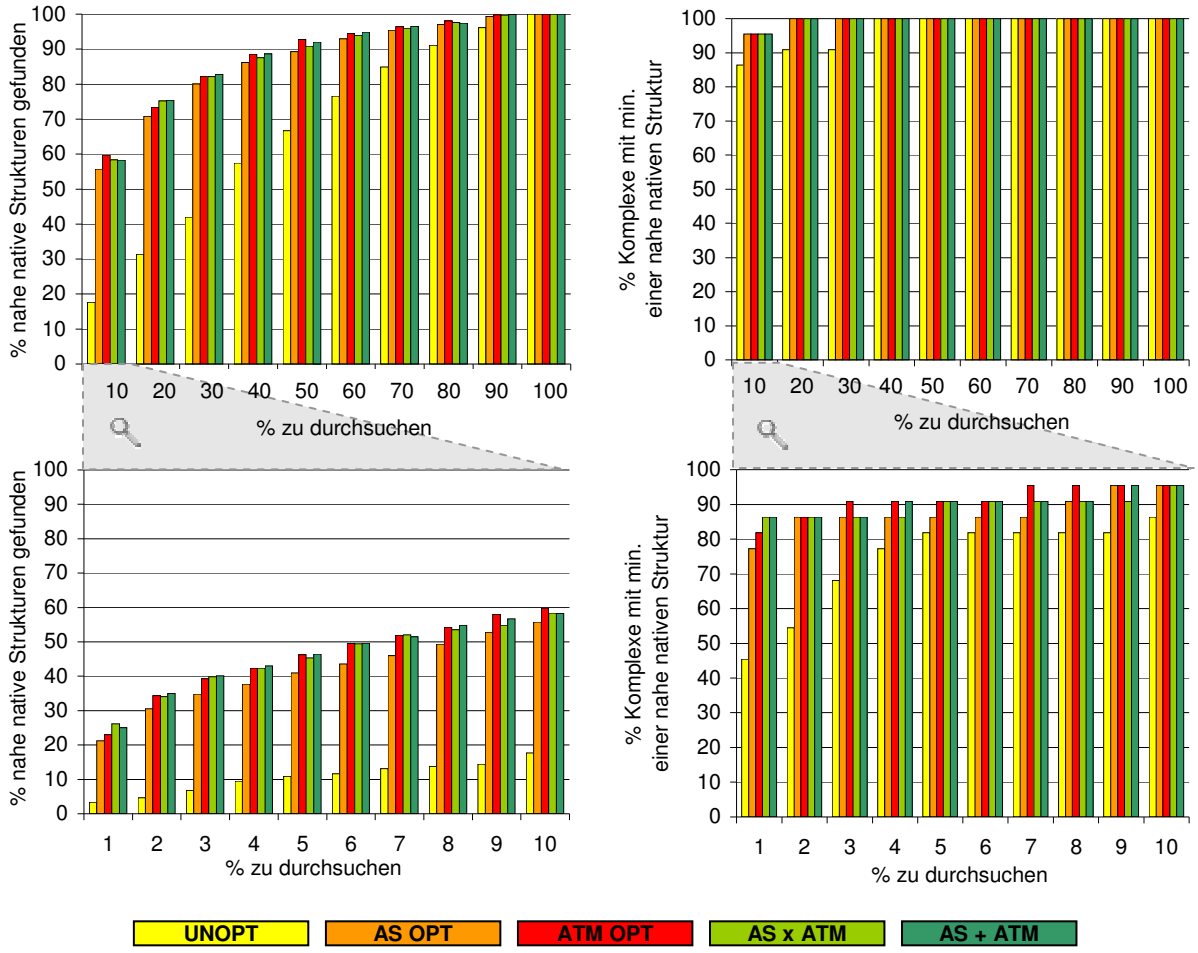

Abbildung 5.4: Verbesserung der Vorhersagequalität von **Enzym-Inhibitor Komplexen** durch Anwendung der Gewichtungsfaktoren. Es ist gezeigt, wieviel Prozent aller von *ckordo* vorgeschlagenen Strukturen sortiert nach der geometrischen Korrelation bzw. nach den gewichteten Scores durchsucht werden müssen, bis wieviel Prozent aller nahe nativer Strukturen gefunden werden (links) bzw. bis wieviel Prozent aller Komplexe mindestens eine nahe native Struktur haben. Anzahl aller nahe nativer Strukturen: 726; Anzahl aller Komplexe: 22)

Strukturen sortiert nach der geometrischen Korrelation bzw. sortiert nach dem gewichteten Score, zu durchsuchen sind (X-Achse), bis wieviel Prozent aller nahe nativer Strukturen gefunden werden (Y-Achse Abbildung 5.4 linke Seite). Die Y-Achse der beiden rechten Diagramme in Abbildung 5.4 zeigt an, für wieviel Prozent aller Komplexe mindestens eine nahe native Struktur gefunden wird.

Innerhalb des ersten Prozent der vorgeschlagenen Strukturen kann nur mit dem Wert der geometrischen Korrelation für 45% der Komplexe eine nahe native Struktur gefunden werden, was 3% aller nahe nativen Enzym-Inhibitor Strukturen entspricht. Durch die Anwendung der optimierten Gewichtungsfaktoren für Aminosäuren erhöht sich der Anteil an Komplexen mit einer nahe nativen Struktur im ersten Prozent der Vorhersage auf 77% und durch Anwendung der atomspezifischen Faktoren auf 82%. Der Anteil nahe nativer Strukturen im ersten Prozent steigt durch Anwendung der aminosäurespezifischen Gewichtungsfaktoren auf 21% und auf 23% durch die atomspezifischen Faktoren.

Während ohne die Gewichtungsfaktoren mindestens 40% aller potentiellen Strukturen durchsucht werden müssen, bis für alle Komplexe mindestens eine nahe native Struktur gefunden wird, halbiert sich dieser Anteil durch Anwendung der Gewichtungsfaktoren auf 20%.

Durch die Kombination der atom- und der aminosäurespezifischen Gewichtungsfaktoren durch Addition bzw. Multiplikation kann der Anteil an nahe nativen Strukturen sowie der Anteil an Komplexen mit mindesten einer nahe nativen Struktur im ersten Prozent der Vorhersage weiter gesteigert werden. Durch Multiplikation können 26% aller nahe nativen Strukturen im ersten Prozent gefunden werden, so dass 86% aller Komplexe mindestens eine nahe native Struktur im ersten Prozent aufweisen.

Wenn eine nahe native Struktur als eine solche mit einem RMSiCα ≤ 2,5 Å definiert wird, dann werden zwar nur für 14 EI-Komplexe überhaupt nahe native Strukturen berechnet, und im Schnitt auch nur 2,6 je Komplex, aber das Produkt aus aminosäure- und atomspezifischen Scores kann diese sehr guten Strukturen besonders gut erkennen und auf den niedrigen Rängen platzieren. Dies führt dazu,

dass bereits im ersten Prozent der Ausgabe 62% aller Strukturen mit einem RMSiCα ≤ 2,5 Å platziert werden.

Tabelle 5.5 zeigt, auf welchem Rang die erste nahe native Struktur je nach angewandtem Filter zu finden ist und welchen RMS-Wert diese Struktur hat. Für zwei Komplexe (1D6R und 7CEI) kann keiner der Filter eine Verbesserung erreichen. Insgesamt wird das beste Ergebnis durch eine Multiplikation der Werte erzielt, die mit aminosäurespezifischen und atomspezifischen Gewichtungsfaktoren berechnet werden. Für 14 der 22 Enzym-Inhibitor Komplexe kann mit dieser Kombination der vergleichsweise niedrigste Rang erzeugt werden, so dass 3 Komplexe auf dem

	Unoptimiert		**Aminosäure-spezifisch**		**Atom-spezifisch**		**Aminosäure-spezifisch x Atomspezifisch**		**Aminosäure-spezifisch + Atomspezifisch**	
	Rang	**RMS**	**Rang**	**RMS**	**Rang**	**RMS**	**Rang**	**RMS**	**Rang**	**RMS**
1ACB	653	4,87	63	4,13	72	4,13	**37**	4,13	43	4,13
1AVX	203	4,8	76	3,39	**14**	3,39	26	3,39	27	3,39
1AY7	7	2,59	23	2,37	15	2,37	**5**	2,37	7	2,37
1BVN	100	2,31	2	2,92	**1**	2,92	**1**	2,92	**1**	2,92
1CGI	107	4,53	**1**	3,85	**1**	3,85	**1**	3,85	**1**	3,85
1D6R	**621**	3,69	3330	2,71	1449	3,88	2304	3,88	1963	3,88
1DFJ	3815	4,9	79	2,75	79	2,75	**36**	2,75	41	2,75
1E6E	1074	4,89	5	3,9	4	3,9	**3**	3,9	**3**	3,9
1EAW	101	3,24	203	4,7	**64**	4,93	95	4,93	77	4,93
1EWY	479	4,92	269	3,96	269	3,07	**192**	3,07	217	3,07
1F34	7448	3,85	16	3,72	4	3,85	**3**	3,72	**3**	3,72
1HIA	75	4,81	31	4,22	**29**	4,76	51	4,22	41	4,76
1KKL	2027	3,52	310	0,46	**135**	4,79	155	4,79	147	4,79
1MAH	864	1,6	15	1,45	24	3,96	**13**	1,45	14	1,45
1PPE	**6**	4,4	13	4,99	16	3,6	**6**	3,6	7	3,6
1TMQ	28	1,05	2	3,49	**1**	2,78	4	3,49	4	2,78
1UDI	249	4,25	3	1,86	2	1,86	**1**	1,86	**1**	1,86
2MTA	838	3,53	195	4,31	194	1,36	**161**	4,31	194	4,31
2PCC	7589	4,66	**1886**	3,86	2325	4,42	2170	4,42	2255	4,42
2SIC	15	4,82	26	1,88	12	1,84	**14**	1,84	17	1,84
2SNI	419	2,53	12	2,53	14	2,53	**9**	2,53	11	2,53
7CEI	**167**	4,17	1572	2,74	599	2,74	934	2,74	827	2,74
Mittelwert	1222	3,8	369,6	**3,2**	**242,0**	3,3	282,8	3,4	268,2	3,4
Median	334,0	4,2	28,5	**3,4**	**20,0**	3,5	**20,0**	3,5	22,0	3,5
Anzahl = 1	0		1		**3**		**3**		**3**	
Anzahl ≤ 10	2		5		6		**9**		8	
Anzahl ≤ 100	6		15		**16**		**16**		**16**	
Anzahl Beste	3		2		7		**14**		5	

Tabelle 5.5: Platzierung der ersten nahe nativen Struktur je nach angewandtem Filter
In Fett sind für jede Zeile die jeweils besten Werte hervorgehoben. Die statistische Auswertung umfasst den Mittelwert, den Median, die Anzahl an Komplexen mit einer nahe nativen Lösung unter den ersten 1, 10, 100 Rängen und für wie viele Komplexe der jeweilige Filter das beste Ergebnis erzielen konnte.

ersten Platz, 9 Komplexe unter den ersten 10 Plätzen und 16 Komplexe unter den ersten 100 Plätzen der sortierten Ausgabe eine nahe native Lösung haben. Bis auf die beiden oben genannten Komplexe kann diese Rerankingfunktion den Rang der ersten nahe nativen Struktur für alle Komplexe erheblich verbessern.

5.3.2 Antikörper-Antigen Komplexe

Die Vorhersagequalitätsverbesserung für die Antikörper-Antigen Komplexe ist in Abbildung 5.5 gezeigt. Auch bei den Antikörper-Antigen Komplexen sind sowohl die atom- als auch die aminosäurespezifischen Gewichtungsfaktoren in der Lage, den Anteil an nahe nativen Strukturen auf den vorderen Rängen und den Anteil der Komplexe, die eine solche auf den vorderen Rängen haben, außerordentlich zu steigern.

Durch den Einsatz der atomspezifischen Gewichtungsfaktoren findet sich für mehr als 90% der Antikörper-Antigen Komplexe eine nahe native Struktur unter den ersten 2% der vorgeschlagenen Strukturen. Der Anteil an nahe nativen Strukturen, die im ersten Prozent liegen, verfünfzehntfacht sich durch Neusortierung der Strukturen mit den aminosäurespezifischen Gewichtungsfaktoren von 1,4% auf 21%.

Im Allgemeinen schneidet der mit den atomspezifischen Gewichtungsfaktoren berechnete Score leicht besser ab als der aminosäurespezifische.

Die Kombination aus mit atomspezifischen und mit aminosäurespezifischen Gewichtungsfaktoren berechneten Scores durch Addition bzw. Multiplikation kann den Anteil an gut vorhersagbaren Komplexen weiter steigern. Der Anteil an Komplexen mit mindestens einer nahe nativen Struktur im ersten Prozent der Ränge erreicht 89%, und es finden sich bis zu 31% der nativen Strukturen im ersten Prozent.

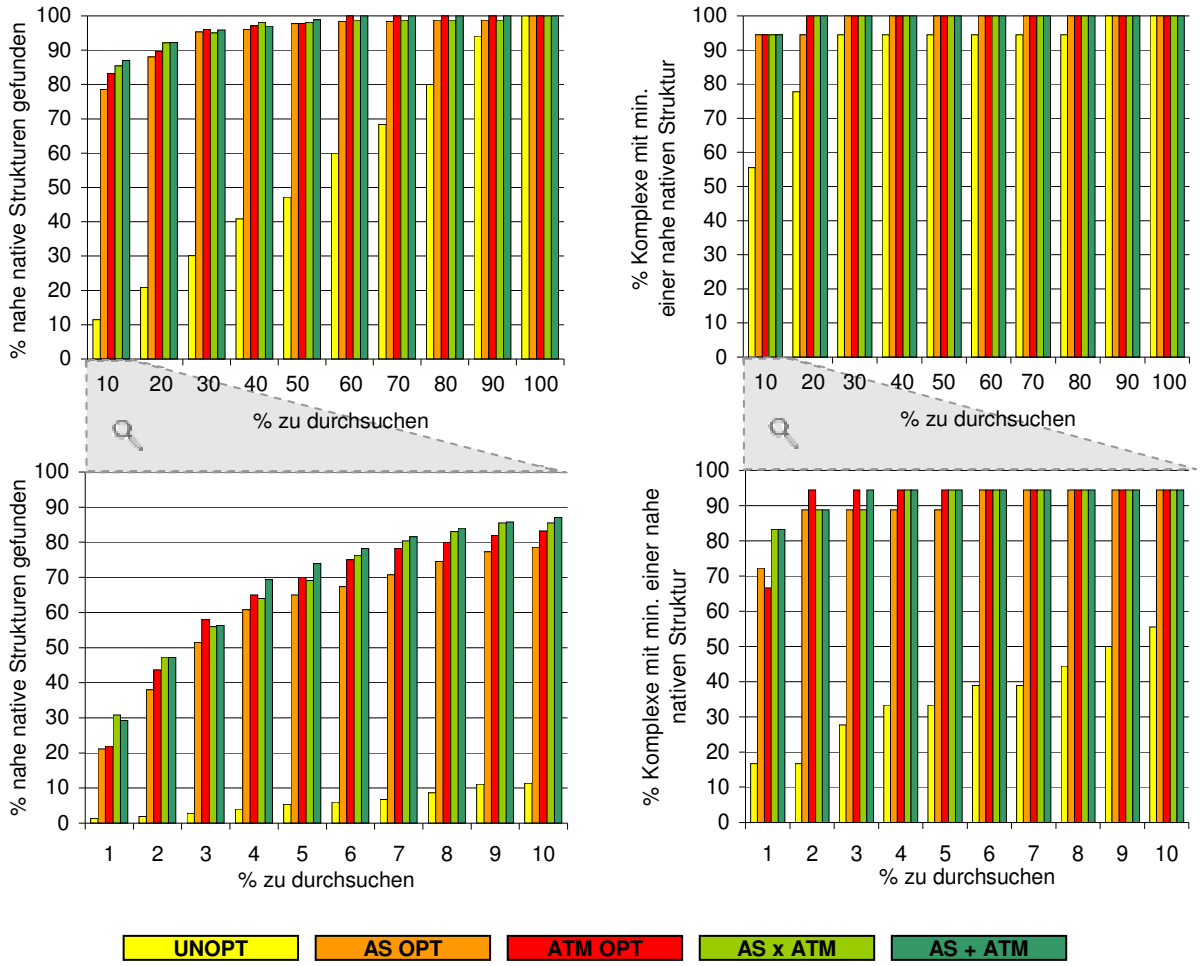

Abbildung 5.5: Verbesserung der Vorhersagequalität von **Antikörper-Antigen Komplexen** durch Anwendung der Gewichtungsfaktoren. Es ist gezeigt, wieviel Prozent aller von ckordo vorgeschlagenen Strukturen sortiert nach der geometrischen Korrelation bzw. nach den gewichteten Scores durchsucht werden müssen, bis wieviel Prozent aller nahe nativer Strukturen gefunden werden (links) bzw. bis wieviel Prozent aller Komplexe mindestens eine nahe native Struktur haben. (Anzahl aller nahe nativer Strukturen: 185; Anzahl aller Komplexe: 18)

Auch bei den Antikörper-Antigen Komplexen kann das Produkt aus aminosäure- und atomspezifischen Scores besonders gut solche Strukturen mit einem RMSiCα ≤ 2,5 Å erkennen. Auch hier werden nur für 14 Komplexe überhaupt solche Strukturen vorhergesagt, aber nach dem Reranking finden sich 43% aller sehr nahe nativen Strukturen im ersten Prozent der Ausgabe und mehr als 90% in den ersten 8%.

Tabelle 5.6 zeigt, dass sowohl beide Scoringfunktionen als auch deren Kombinationen den Rang der ersten richtig vorhergesagten Struktur für alle Antikörper-Antigen Komplexe verbessern.

Die verschiedenen Filter haben bei der Neusortierung der vorgeschlagenen Strukturen für Antikörper-Antigen Komplexe unterschiedliche Stärken, so dass der niedrigste Rang für verschiedene Komplexe durch jeweils andere Filter erreicht wird. Der atomtypspezifische Filter kann für 8 Komplexe einer nahe nativen Struktur den niedrigsten Rang zuweisen. Dieser Filter, wie auch das Produkt mit dem aminosäurespezifischen Filter können jeweils für 8 Komplexe eine richtige Lösungen unter die ersten 100, 3 unter die ersten 10 und für einen auf den ersten Rang bringen. Nur mit dem aminosäurespezifischen Filter wird für 9 Komplexe eine nahe

	Unoptimiert		**Aminosäure-spezifisch**		**Atom-spezifisch**		**Aminosäure-spezifisch x Atomspezifisch**		**Aminosäure-spezifisch + Atomspezifisch**	
	Rang	RMS	Rang	RMS	Rang	RMS	Rang	RMS	Rang	RMS
1AHW	487	2,34	380	2,34	260	2,34	**255**	2,34	260	2,34
1BJ1	18489	3,96	**126**	3,96	338	3,96	163	3,96	184	3,96
1BVK	2140	4,57	**80**	4,38	244	4,57	127	4,57	151	4,57
1DQJ	5318	3,61	**62**	4,97	155	3,78	136	4,97	153	4,97
1E6J	146	4,05	**1**	2,96	2	2,96	**1**	2,96	**1**	2,96
1FSK	1206	1,78	278	1,78	**183**	1,78	192	1,78	197	1,78
1I9R	81	3,84	80	2,1	**38**	3,62	42	3,62	46	3,62
1IQD	1765	3,41	18	3,41	**1**	3,41	3	3,41	**1**	3,41
1JPS	613	1,7	77	1,46	54	1,46	**48**	1,46	49	1,46
1KXQ	6	2,34	49	2,34	**2**	2,34	4	2,34	4	2,34
1MLC	6322	1,19	4929	1,19	**2319**	1,19	3049	1,19	2800	1,19
1NCA	718	3,08	**78**	1,15	196	1,15	85	1,15	96	1,15
1NSN	1655	2,89	182	2,52	**47**	2,89	119	2,52	112	4,96
1QFW$_a$	2575	1,96	1164	1,96	**413**	4,34	663	1,96	643	1,96
1QFW$_b$	3228	4,32	**143**	2,02	263	2,02	126	2,02	150	2,02
1VFB	2430	4,65	212	4,65	157	4,51	154	4,51	**155**	4,51
1WEJ	6348	4,69	**43**	3,69	97	3,69	53	3,69	60	3,69
2JEL	2610	4,42	413	4,42	**31**	4,42	89	4,42	50	4,42
Mittelwert	3118,7	3,3	461,9	**2,9**	**266,7**	3,0	294,9	2,9	284,0	3,1
Median	1952,5	3,5	**103,0**	**2,4**	156,0	3,2	122,5	2,7	131,0	3,2
Anzahl =1	0		1		1		1		**2**	
Anzahl ≤10	1		1		**3**		**3**		**3**	
Anzahl ≤100	2		**9**		8		8		8	
Anzahl Beste	0		6		**8**		4		2	

Tabelle 5.6: Platzierung der ersten nahe nativen Struktur je nach angewandtem Filter
In Fett sind für jede Zeile die jeweils besten Werte hervorgehoben. Die statistische Auswertung umfasst den Mittelwert, den Median, die Anzahl an Komplexen mit einer nahe nativen Lösung unter den ersten 1, 10, 100 Rängen und für wie viele Komplexe der jeweilige Filter das beste Ergebnis erzielen konnte.

native Struktur unter den ersten 100 gefunden.

Vergleicht man die verschiedenen Filter direkt miteinander, ist das Produkt der beiden Gewichtungsfaktoren jeweils bei mehr Komplexen besser als die anderen Scoringfunktionen.

5.3.3 ‚Andere' Komplexe

Bei den ‚Anderen' Komplexen können beide Rerankingfunktionen sowohl den Anteil an nahe nativen Strukturen als auch den Anteil an Komplexen mit mindestens einer nahe nativen Struktur auf den vorderen Rängen steigern (vgl. Abbildung 5.6).

Durch eine Neusortierung der vorgeschlagenen Strukturen nach dem mit den aminosäurespezifischen Gewichtungsfaktoren berechneten Scores kann der Anteil nahe nativer Strukturen in den ersten 10% der Vorhersage von 14% auf 39% und durch Anwendung der atomspezifischen Gewichtungsfaktoren sogar auf 41% gesteigert werden, so dass der Anteil an Komplexen mit einer nahe nativen Struktur in den Top 10% von 63% auf 83% steigt.

Die Kombination beider Filter führt bei den ‚Anderen' Komplexen zu keiner weiteren Verbesserung der Ergebnisse. Je nachdem welchen prozentualen Abschnitt man betrachtet, sind die Ergebnisse minimal besser oder sogar etwas schlechter als nur mit den atom- oder aminosäurespezifischen Gewichtungsfaktoren.

Für 17 der ‚Anderen' Komplexe konnten im Schnitt 1,8 Strukturen mit einem RMSiCα ≤ 2,5 Å gefunden werden. Durch das Reranking mit dem Produkt aus aminosäure- und atomspezifischen Scores werden mehr als 30% dieser sehr nahe nativen Strukturen in den Top 2% der Vorhersage gefunden.

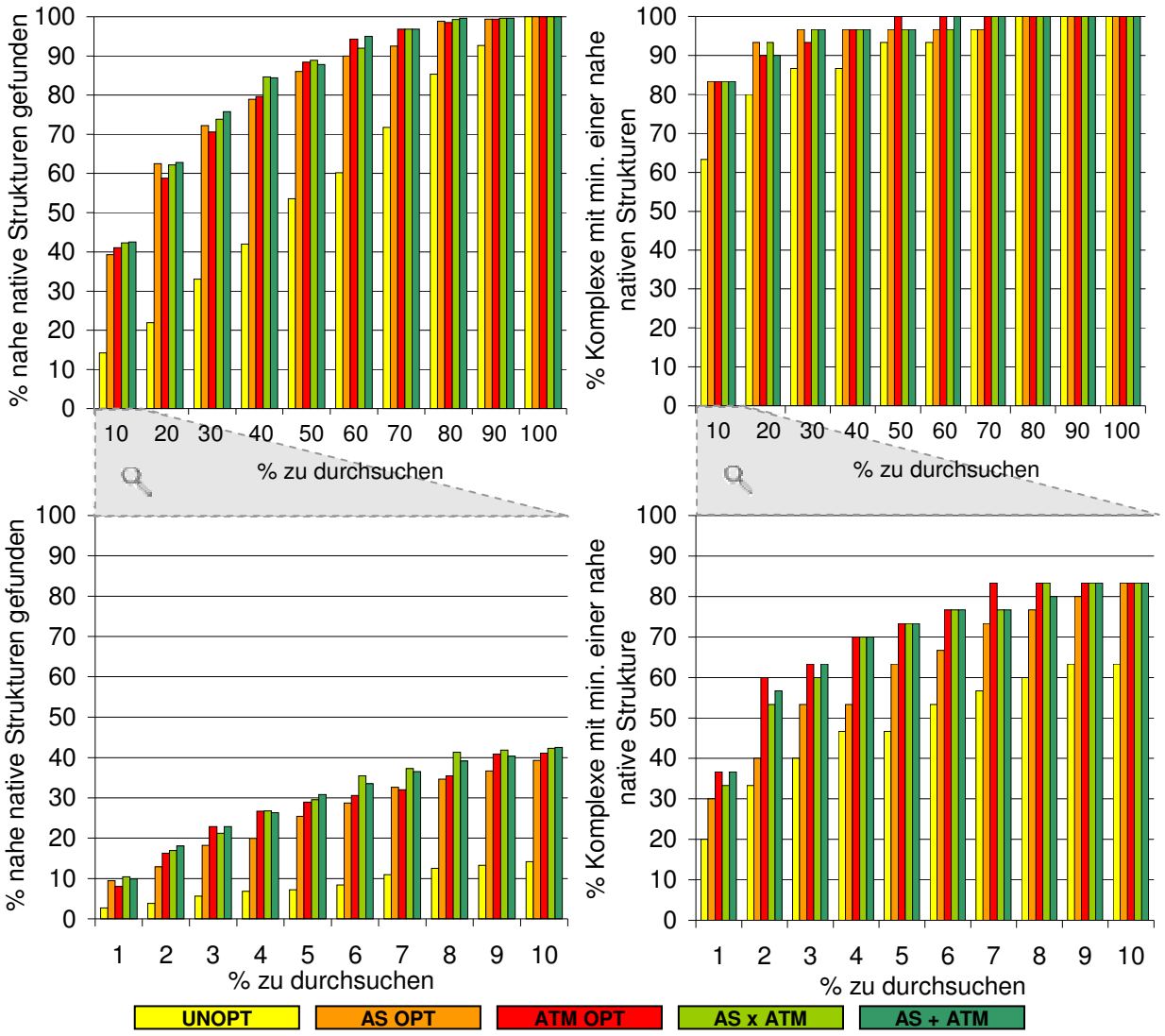

Abbildung 5.6: Verbesserung der Vorhersagequalität von **‚Anderen' Komplexen** durch Anwendung der Gewichtungsfaktoren. Es ist gezeigt, wieviel Prozent aller von ckordo vorgeschlagenen Strukturen sortiert nach der geometrischen Korrelation bzw. nach den gewichteten Scores durchsucht werden müssen, bis wieviel Prozent aller nahe nativer Strukturen gefunden werden (links) bzw. bis wieviel Prozent aller Komplexe mindestens eine nahe native Struktur haben. (Anzahl aller nahe nativer Strukturen: 300; Anzahl aller Komplexe: 30)

Betrachtet man den Mittelwert der Ränge, auf denen die erste nahe native Lösung zu finden ist (Tabelle 5.7), so wird das beste Ergebnis durch eine Reranking mit den atomspezifischen Scores erreicht. Für 12 der 30 Komplexe kann mit dieser Rerankingfunktion der jeweils niedrigste Rang für eine nahe native Struktur berechnet werden, so dass für 8 der Komplexe eine nahe native Struktur innerhalb der ersten 100 Ränge zu finden ist.

	Unoptimiert		Aminosäure-spezifisch		Atomspezifisch		Aminosäure-spezifisch x Atomspezifisch		Aminosäure-spezifisch + Atomspezifisch	
	Rang	RMS	Rang	RMS	Rang	RMS	Rang	RMS	Rang	RMS
1A2K	338	4,77	**66**	4,34	198	4,84	106	4,34	116	4,34
1AK4	3589	4,75	**136**	4,99	363	4,99	124	4,99	150	4,99
1AKJ	76	3,5	623	3,49	**33**	3,65	271	3,65	104	3,65
1B6C	467	2,42	**9**	2,87	42	3,65	14	3,65	19	3,65
1BUH	1425	4,71	1057	2,56	**507**	4,71	607	2,56	633	2,56
1DE4	17410	2,31	15458	2,31	**10376**	2,31	13244	2,31	12392	2,31
1E96	787	4,03	**21**	3,05	334	3,05	55	3,05	71	3,05
1EER	**3261**	4	5843	4,88	4588	4,88	4761	4,88	4879	4,88
1F51	819	1,92	146	0,43	**38**	0,43	40	0,43	39	0,43
1FC2	1122	4,75	83	3,6	**74**	4,31	**74**	3,6	77	3,6
1FQ1	3747	3,65	1229	3,65	**233**	3,65	498	3,65	392	3,65
1FQJ	**231**	3,98	450	3,78	233	3,78	224	3,78	227	3,78
1GCQ	1653	2,11	2487	1,92	**877**	1,92	1318	1,92	1238	1,92
1GHQ	3547	4,17	**2038**	4,63	4391	4,92	3571	4,92	3883	4,92
1GP2	3917	1,71	291	3,67	**64**	1,53	134	3,58	118	3,58
1GRN	**250**	4,5	1662	2,78	363	2,27	731	2,27	633	2,27
1HE1	5691	3,16	1391	4,91	**1145**	5	1619	5	1549	5
1HE8	**106**	2,08	598	0,99	222	0,99	286	0,99	269	0,99
1I2M	**131**	3,14	1045	3,14	168	4,21	325	3,14	258	3,14
1IJK	597	4,99	448	4,99	**115**	4,8	397	4,8	297	4,8
1K5D	9264	4,94	**3007**	4,94	6947	4,94	4157	4,94	4981	4,94
1KAC	**345**	2,85	1890	4,78	1495	4,54	1721	4,78	1920	4,78
1KLU	14603	3,94	3640	3,65	**2450**	2,91	2636	2,91	2673	2,91
1KTZ	9816	3,37	**401**	3,06	1525	3,37	785	3,06	861	3,06
1KXP	37	1,29	8	3,49	8	3,49	**5**	3,49	**5**	3,49
1ML0	1109	4,62	329	2,15	361	3,94	**265**	2,15	300	2,15
1QA9	6120	3,45	1397	1,74	1041	1,74	**982**	1,74	1011	1,74
1RLB	205	4,81	45	4,85	81	3,5	**28**	3,5	40	3,5
1SBB	1804	4,24	896	3,95	787	3,95	**672**	3,95	715	3,95
1WQ1	165	3,23	120	4,37	**31**	2,36	46	4,37	53	4,37
Mittelwert	3087,73	3,58	1560,47	3,47	**1303**	3,49	1323,2	3,41	1330,1	3,41
Median	1115,5	3,795	610,5	3,625	**347,5**	3,65	361	3,59	298,5	3,59
Anzahl ≤ 100	2		6		**8**		7		7	
Anzahl ≤ 10	0		**2**		1		1		1	

Tabelle 5.7: Platzierung der ersten nahe nativen Struktur je nach angewandtem Filter
In Fett sind für jede Zeile die jeweils besten Werte hervorgehoben. Die statistische Auswertung umfasst den Mittelwert, den Median und die Anzahl an Komplexen mit einer nahe nativen Lösung unter den ersten 10 bzw. 100 Rängen.

5.3.4 Spezifität

Abbildung 5.7 zeigt, dass die für die jeweilige Komplexklasse optimierten aminosäurespezifischen Gewichtungsfaktoren auch für die jeweils selbe Komplexklasse die meisten nahe nativen Strukturen auf den niedrigen Rängen anreichern, also spezifisch für die jeweilige Klasse sind.

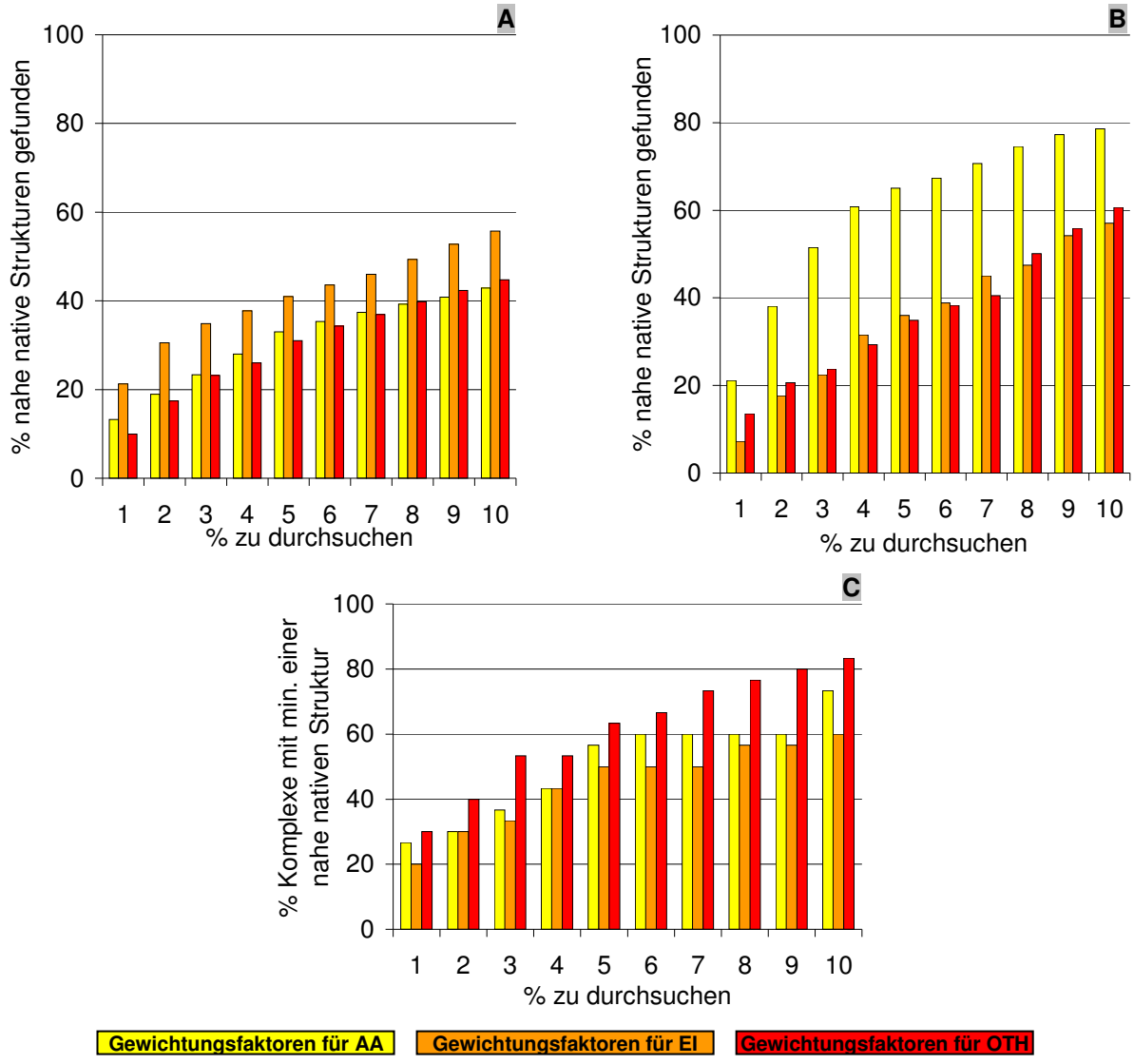

Abbildung 5.7: Spezifität der aminosäurespezifischen Gewichtungsfaktoren; (A) bei Enzym-Inhibitor Komplexen, (B) Antikörper-Antigen Komplexen und (C) den anderen Komplexen

Insbesondere die Faktoren für die Antikörper-Antigen Komplexe sind hochspezifisch. Mit den für die AA-Komplexe optimierten Gewichtungsfaktoren können mehr als 50% aller nahe nativen Strukturen unter die ersten 3% der jeweiligen Vorhersage gebracht werden, während die für EI bzw. ‚Andere' Komplexe optimierten

Gewichtungsfaktoren nur etwas mehr als 20% aller nahe nativen AA-Strukturen unter die ersten 3% bringen (Abbildung 5.7 B).

Die für die ‚Anderen' Komplexe optimierten Faktoren sind nicht so spezifisch, wie die AA-Faktoren, da diese nur 18% aller nahe nativen OTH-Strukturen unter die Top 3% bringen, während sich nach einem Reranking mit den für EI- und AA-Komplexe optimierten Faktoren 11% bzw. 13% der nahe nativen Strukturen in den ersten 3% finden. Betrachtet man für die ‚Anderen' Komplexe allerdings den Anteil an Komplexen, die eine nahe native Struktur unter den ersten 3% der Ausgabe haben,

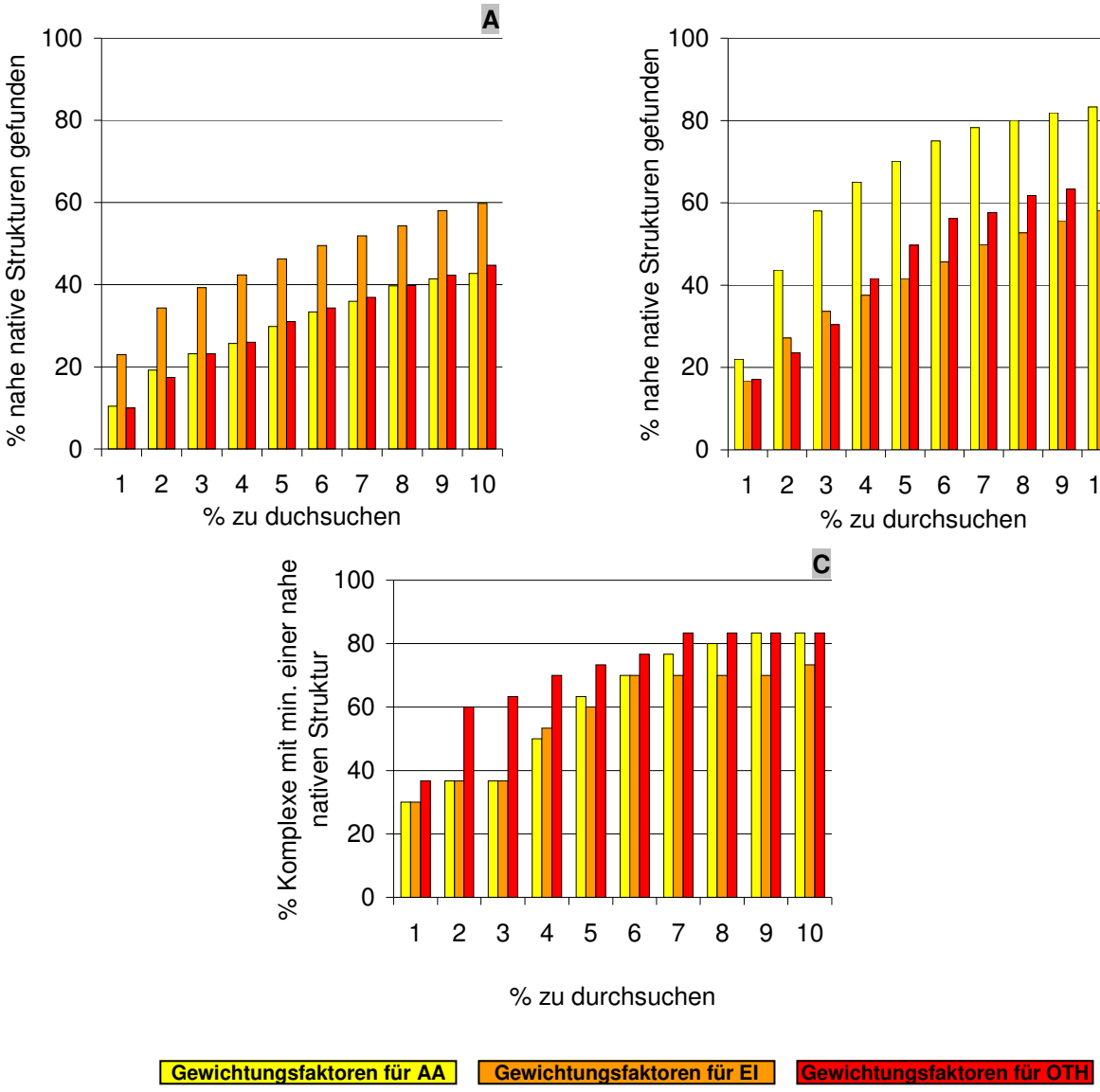

Abbildung 5.8: Spezifität der atomspezifischen Gewichtungsfaktoren; (A) bei Enzym-Inhibitor Komplexen, (B) Antikörper-Antigen Komplexen und (C) den anderen Komplexen

wird deutlich, dass auch die für die OTH-Komplexe optimierten Faktoren für diese hochspezifisch sind (Abbildung 5.7 C).

Die Spezifität für die atomspezifischen Gewichtungsfaktoren ist in Abbildung 5.8 gezeigt.

Die für Enzym-Inhibitor Komplexe optimierten atomspezifischen Gewichtungsfaktoren können 16% mehr nahe native Strukturen unter den Top 3% der Vorhersage platzieren als die anderen Faktoren (Abbildung 5.8 B). Bei den Antikörper-Antigen Komplexen beträgt die Verbesserung für den Anteil der nahe nativen Strukturen unter den Top 3% mit den spezifisch für dieselben optimierten Faktoren sogar bis zu 28% (Abbildung 5.8 A).

Bei den ‚Anderen' Komplexen ist der Anteil nahe nativer Strukturen unter den Top 3% zwar nur 8% höher durch Verwendung der für OTH-Komplexe optimierten atomspezifischen Faktoren, allerdings liegt der Anteil an Komplexen mit mindestens einer nahe nativen Struktur unter den ersten 3% bei 63%, was 27% mehr sind als mit den anderen Gewichtungsfaktoren erreicht werden könnte (Abbildung 5.8 C).

Es lässt sich zusammenfassen, dass sowohl die atom- als auch die aminosäurespezifischen Gewichtungsfaktoren jeweils für die Komplexklasse, für die sie optimiert wurden, spezifisch sind.

5.3.5 Validierung (UUPPDD)

5.3.5.1 Enzym-Inhibitor/Substrat Komplexe

Auch bei den Komplexen, die nicht Teil des Optimierungsprozesses sind, sind die Gewichtungsfaktoren in der Lage, nahe native Strukturen auf den vorderen Rängen anzureichern (vgl. Abbildung 5.9 links) und somit die Vorhersagequalität zu verbessern.

	Unoptimiert		Aminosäure-spezifisch		Atomspezifisch		Aminosäure-spezifisch x Atomspezifisch		Aminosäure-spezifisch + Atomspezifisch	
	Rang	RMS	Rang	RMS	Rang	RMS	Rang	RMS	Rang	RMS
1ACB	48	0,93	3	2,64	**1**	2,64	**1**	2,64	**1**	2,64
1AVW	1434	3,32	200	4,86	**8**	1,64	55	1,64	33	1,64
1BRC	**3**	4,44	15	4,56	14	5	34	4,56	33	5
1BRS	483	4,21	5	1,97	2	1,97	**1**	1,97	**1**	1,97
1BVN	33	3,98	2332	4,42	**29**	3,28	211	3,91	437	3,91
1CGI	**21**	4,55	31	4,29	25	3,5	35	4,29	51	3,5
1CHO	33	1,03	**1**	2,78	2	3,15	**1**	4,64	**1**	4,64
1CSE	**427**	2,27	4334	4,79	1947	4,89	2924	4,89	2717	4,89
1DFJ	3598	4,14	532	4,14	**171**	4,14	241	4,14	226	4,14
1FSS	2713	1,62	**149**	4,1	165	1,09	170	4,76	191	4,76
1GLA	4475	4,57	**3508**	4,63	6294	4,79	5169	4,05	6032	4,05
1MAH	471	4,51	94	2,44	**44**	3,25	54	3,25	48	3,25
1PPF	162	4,59	9	4,35	3	4,35	**1**	4,35	**1**	4,35
1TGS	55	4,22	2	1,97	3	1,97	**2**	1,97	**2**	1,97
1UGH	129	2,72	13	2,67	42	2,8	**14**	2,02	21	2,02
2KAI	**1**	4,74	18	3,96	21	4,9	15	3,96	18	3,96
2MTA	466	4,74	374	4,64	**351**	4,78	581	4,64	633	4,78
2PCB	1357	4,48	881	4,42	1671	4,42	**825**	4,42	1037	4,42
2PCC	**1843**	4,91	2548	4,12	3212	4,12	2341	4,12	2717	4,12
2PTC	174	3,76	20	4,57	5	4,57	**4**	4,57	**4**	4,57
2SNI	1361	4,57	59	3,73	**32**	4,98	38	4,66	43	4,66
Mittelwert	918,43	3,73	720,38	3,81	668,67	**3,63**	**605,57**	3,78	678,43	3,77
Median	427	4,22	59	4,14	**29**	4,12	38	4,14	43	4,12
Anzahl = 1	1		1		1		**4**		**4**	
Anzahl ≤ 10	2		5		**7**		6		6	
Anzahl ≤ 100	7		12		**14**		13		13	

Tabelle 5.8: Ränge der ersten nahe nativen Strukturen und deren RMSD je nach angewandten Gewichtungsfaktoren

Der Anteil der nahe nativen Strukturen, die im ersten Prozent der Vorhersage liegen, kann mit den atomspezifischen Gewichtungsfaktoren von 2% auf 16% gesteigert werden und der Anteil der Komplexe mit nahe nativer Struktur im ersten Prozent erhöht sich von 55% auf 82%.

Die Kombination der beiden Gewichtungsfaktoren bringt prozentual gesehen keine weitere Steigerung der Vorhersagequalität, allerdings wird sowohl mit der Multiplikation als auch mit der Addition für 4 Komplexe eine nahe native Lösung auf dem ersten Platz der Vorhersage gefunden und für 13 Komplexe unter den ersten 100. Durchschnittlich wird der niedrigste Rang durch die Multiplikation der Scores von aminosäure- und atomspezifischen Gewichtungsfaktoren erreicht (Tabelle 5.8).

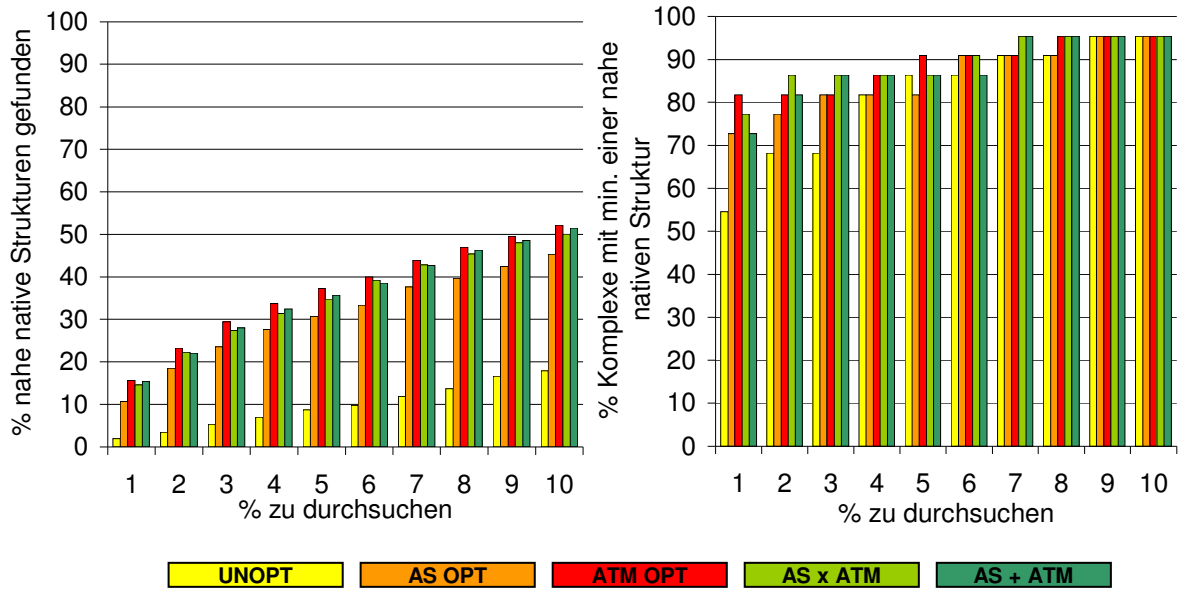

Abbildung 5.9: Anreicherung nahe nativer Strukturen auf den niedrigen Rängen für die Enzym-Inhibitor Komplexe des UUPPDD.

5.3.5.2 Antikörper-Antigen Komplexe

Für alle vier Antikörper-Antigen Komplexe lässt sich durch die Anwendung der atomspezifischen Gewichtungsfaktoren mindestens eine nahe native Struktur unter die Top 2 Prozent der Vorhersage bringen. Die mit den aminosäurespezifischen Gewichtungsfaktoren berechneten Scores können nur für 3 der 4 Komplexe eine nahe native Struktur unter die vorderen 2% bringen, jedoch auch für alle vier unter die ersten 3%.

In Abbildung 5.10 wird deutlich, dass sich der Anteil an nahe nativen Strukturen, die auf den vorderen Rängen liegen, durch Anwendung beider Gewichtungsfaktoren und deren Kombination massiv steigern lässt. Der Anteil an nahe nativen Strukturen unter den Top 2% der Vorhersage verzehnfacht sich, wenn alle potentiellen Strukturen nach den gewichteten Scores sortiert werden.

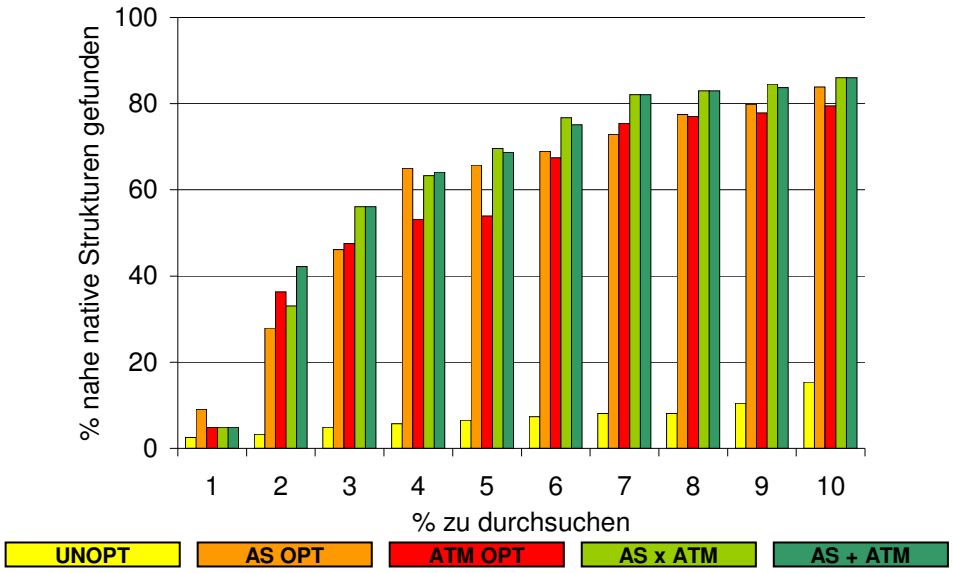

Abbildung 5.10: Anreicherung nahe nativer Strukturen für die Antikörper-Antigen Komplexe des UUPPDD.

	Unoptimiert		Aminosäure-spezifisch		Atomspezifisch		Aminosäure-spezifisch x Atomspezifisch		Aminosäure-spezifisch + Atomspezifisch	
	Rang	RMS	Rang	RMS	Rang	RMS	Rang	RMS	Rang	RMS
1AHW_1	**20**	1,62	580	4,77	508	1,07	469	1,07	498	1,07
1DQJ_1	34411	4,45	1066	3,97	**657**	3,97	723	3,97	717	3,97
1VFB_1	855	4,89	323	4,89	400	4,46	**316**	4,89	368	4,89
1WEJ_1	4263	4,03	304	4,03	430	4,03	**260**	4,03	293	4,03
Mittelwert	9887,25	**3,7475**	568,25	4,415	498,75	3,3825	**442**	3,49	469	3,49

Tabelle 5.9: Rang und RMS der ersten nahe nativen Struktur für die Antikörper-Antigen Komplexe des UUPPDD

Im Schnitt kann der jeweils niedrigste Rang einer nahe nativen Struktur für die Antikörper-Antigen Komplexe durch die Multiplikation der aminosäurespezifischen mit den atomspezifischen Scores erreicht werden. Für 1AHW allerdings wird der niedrigste Rang einer nahe nativen Struktur durch Sortierung nur nach der geometrischen Korrelation erzielt (Tabelle 5.9).

5.3.5.3 'Andere' Komplexe

Die Anreicherung nahe nativer Strukturen auf den vorderen Rängen funktioniert für die vier untersuchten ‚Anderen' Komplexe des UUPPDD nur eingeschränkt. Zwar ist der Anteil an nahe nativen Strukturen nach Anwendung der aminosäurespezifischen Gewichtungsfaktoren unter den ersten 10% etwa 3% höher als nur nach der geometrischen Korrelation, und nach Anwendung der atomspezifischen Faktoren unter den ersten 30% um 15% höher (Abbildung 5.11), allerdings können unter den

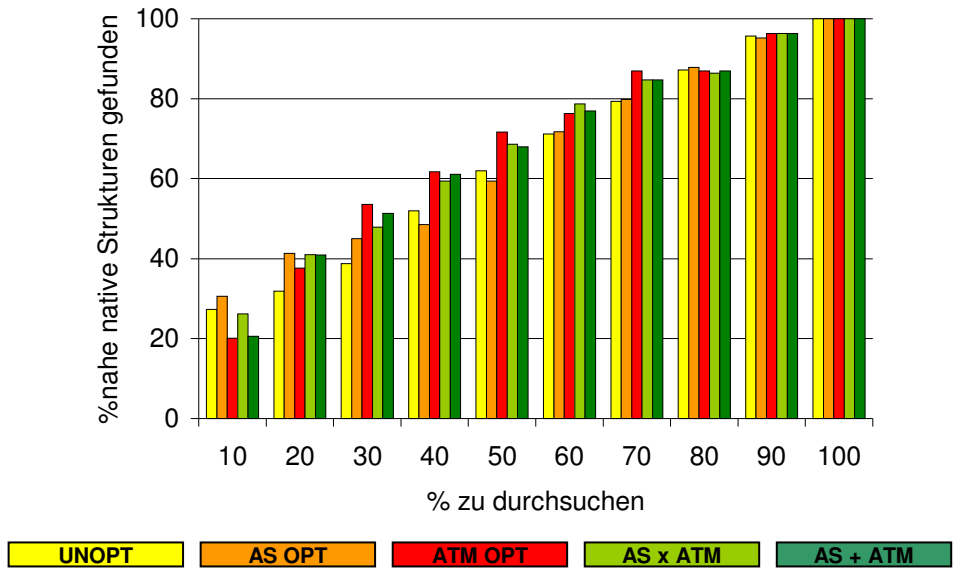

Abbildung 5.11: Anreicherung nahe nativer Strukturen für die **‚Anderen' Komplexe** des UUPPDD.

	Unoptimiert		Aminosäure-spezifisch		Atomspezifisch		Aminosäure-spezifisch x Atomspezifisch		Aminosäure-spezifisch + Atomspezifisch	
	Rang	**RMS**	**Rang**	**RMS**	**Rang**	**RMS**	**Rang**	**RMS**	**Rang**	**RMS**
1AVZ	13281	4,5	19685	3,96	**8999**	3,89	13117	3,89	12855	3,89
1BDJ	**184**	2,8	2216	4,55	2561	4,93	2156	4,55	2500	4,55
1L0Y	**507**	2,64	839	2,27	2265	4,83	1123	2,27	1444	2,27
1WQ1	**1**	3,4	305	2,76	225	4,75	356	2,76	404	4,75
Mittelwert	**3493,25**	3,335	5761,25	3,385	3512,5	4,6	4188	3,3675	4300,75	3,865
Median	**345,5**	3,1	1527,5	3,36	2413	4,79	1639,5	3,325	1972	4,22
Anzahl = 1	**1**		0		0		0		0	
Anzahl ≤ 10	**1**		0		0		0		0	
Anzahl ≤ 100	**1**		0		0		0		0	

Tabelle 5.10: Rang und RMS der ersten nahe nativen Struktur für die ‚Anderen' Komplexe des UUPPDD

ersten 1-5% der Vorhersage alleine durch die geometrische Korrelation die meisten nahe nativen Strukturen auf die vorderen Ränge gebracht werden.

Der Rang der ersten nahe nativen Struktur kann nur für 1AVZ von 13281 auf 8999 gesteigert werden, bei den anderen 3 Komplexen verschlechtert sich der Rang der ersten nahe nativen Struktur deutlich (Tabelle 5.10).

5.4 ckordo mit Gewichtungsfaktoren

5.4.1 Enzym-Inhibitor/Substrat Komplexe

Tabelle 5.11 vergleicht die Vorhersagequalität von *ckordo* für Enzym-Inhibitor Komplexe basierend auf der geometrischen Korrelation mit der Vorhersagequalität mit aminosäurespezifisch gewichteten geometrischen Scores.

Sowohl der niedrigste Rang, auf dem eine nahe native Struktur gefunden wird, als auch der Rang der besten Struktur ist für 67% der Komplexe durch den Einsatz der Gewichtungsfaktoren niedriger als ohne die Gewichtungsfaktoren.

Für 83% der Komplexe konnten durch den Einsatz der Gewichtungsfaktoren deutlich mehr nahe native Strukturen gefunden werden und für 67% der Komplexe ist der niedrigste RMS kleiner.

Insgesamt lässt sich die Vorhersagequalität durch die Benutzung der aminosäurespezifischen Gewichtungsfaktoren also leicht steigern.

Eine leichte zusätzliche Steigerung der Anzahl nahe nativer Strukturen auf den niedrigen Rängen wird durch die Anwendung der atomspezifischen Gewichtungsfaktoren als Rerankingfunktion erreicht (Abbildung 5.12). Die Neusortierung der Strukturen durch Anwendung der aminosäurespezifischen Gewichtungsfaktoren führt zu einem schlechteren Ergebnis, wie auch die

Kombination aus aminosäure- und atomspezifischem Reranking keine Verbesserung gegenüber dem ausschließlich atomspezifischen Reranking darstellt.

	Ckordo					**Ckordo mit Gewichtungsfaktoren**				
	Bester Rang	RMS	Rang der besten Struktur	RMS	Anzahl nahe native Lösungen	Bester Rang	RMS	Rang der besten Struktur	RMS	Anzahl nahe native Lösungen
1ACB	653	4,87	15781	2,52	15	**61**	**3,84**	**293**	**1,89**	**64**
1AVX	203	4,8	19748	1,52	41	**141**	**4,39**	**11865**	**0,75**	**73**
1AY7	**7**	**2,59**	**177**	**1,73**	20	85	4,35	10717	**0,56**	**40**
1BVN	100	**2,31**	1894	2,18	23	**5**	3,35	**161**	**1,53**	**58**
1CGI	**107**	4,53	12142	2,6	58	129	**4,38**	**7214**	**2,43**	**144**
1D6R	**621**	3,69	16882	2,71	**38**	5645	**3,22**	**13910**	**1,98**	11
1DFJ	3815	4,9	8633	**2,37**	3	**32**	**4,16**	**289**	2,44	**30**
1E6E	1074	4,89	**2086**	3,24	10	**58**	**4,14**	4212	**2,85**	**21**
1EAW	101	3,24	13305	2,38	44	**2**	**1,74**	**3**	**1,18**	**154**
1EWY	**479**	4,92	**1912**	**2,56**	**34**	13929	**4,49**	15592	3,51	5
1HIA	75	4,81	2629	3,25	43	**3**	**3,85**	**1378**	**1,71**	**265**
1KKL	**2027**	**3,52**	**5898**	**0,46**	**14**	11196	4,91	20991	3,14	2
1MAH	864	1,6	8450	1,45	12	**5**	**1,48**	**105**	**0,39**	**35**
1PPE	**6**	4,4	13238	1,12	214	51	**2,65**	**140**	**0,94**	**411**
1TMQ	28	**1,05**	**28**	**1,05**	15	**2**	2,34	29	1,88	**49**
1UDI	249	**4,25**	15189	**1,86**	16	**2**	4,71	**1453**	2,19	**31**
2MTA	**838**	**3,53**	18881	**1,36**	**18**	1318	4,58	**6089**	2,51	15
2PCC	**7589**	**4,66**	**7856**	**3,86**	**5**	14931	4,98	19484	4,93	3
2SIC	**15**	4,82	**7288**	**1,84**	40	1371	**4,6**	8630	1,86	**75**
2SNI	419	**2,53**	**10529**	2,12	**44**	**240**	4,42	11474	**1,69**	36
7CEI	167	**4,17**	**6653**	**2,34**	14	**163**	4,76	9927	2,58	**24**
Mittelwert	**925,57**	**3,81**	9009,48	2,12	34,33	2350,90	3,87	**6855,05**	**2,04**	**73,62**
Median	249,00	**4,25**	8450,00	2,18	20,00	**85,00**	4,35	**6089,00**	**1,89**	**36,00**
Anzahl besser	9	9	9	9	6	**12**	**12**	**12**	**12**	**15**

Tabelle 5.11: Vergleich der Vorhersagequalität von ckordo ohne und mit optimierten aminosäurespezifischen Gewichtungsfaktoren. Die jeweils besten Werte sind fett hervorgehoben.

Das beste Ergebnis wird durch eine Multiplikation der Scores der atomspezifischen Rerankingfunktion mit den aminosäurespezifischen *ckordo* Scores erreicht.

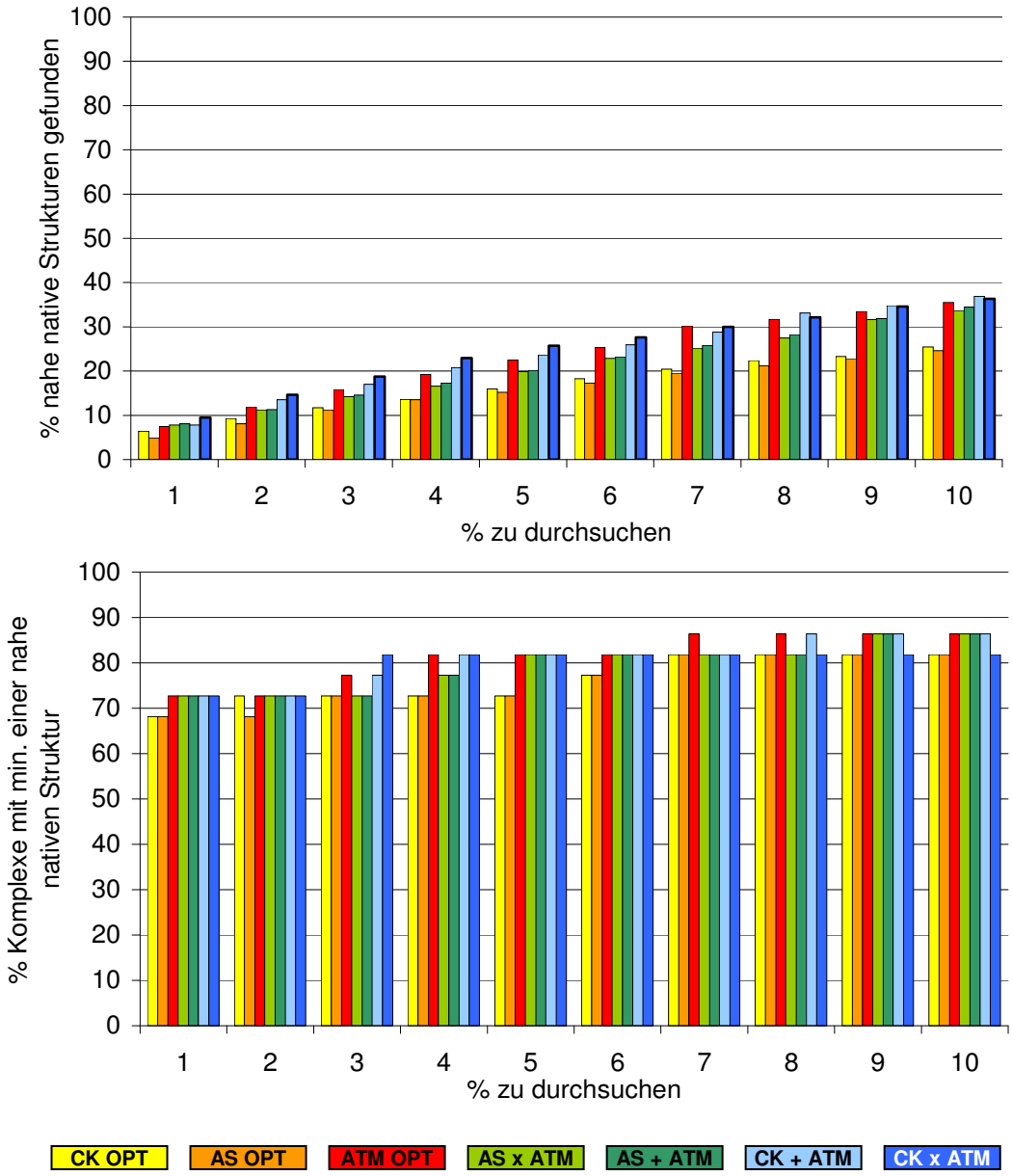

Abbildung 5.12: Verbesserung der Vorhersagequalität bei Enzym-Inhibitor Komplexen mit aminosäurespezifischen Gewichtungsfaktoren in ckordo und anschließendem Reranking in verschiedenen Kombinationen. (CK: ckordo mit aminosäurespezifischen Faktoren; AS/ATM: Reranking Funktionen mit Gewichtungsfaktoren)

Ein Vergleich des Anteils nahe nativer Strukturen, die durch einen gewichteten *ckordo* Lauf und anschließendes Reranking auf den vorderen Rängen gefunden werden, mit dem Anteil, der durch einen ungewichteten *ckordo* Lauf mit anschließendem atom- und aminosäurespezifischen Reranking erreicht wird, ist in Abbildung 5.13 gezeigt.

Nur durch den gewichteten *ckordo* Lauf können zwar mehr nahe native Strukturen auf den vorderen Rängen vorhergesagt werden als mit einem normalen *ckordo* Lauf,

allerdings ist der Anteil nahe nativer Strukturen, welche durch einen normalen *ckordo* Lauf mit anschließendem Reranking in den Top 10% der Vorhersage gefunden werden, etwa anderthalb mal so groß.

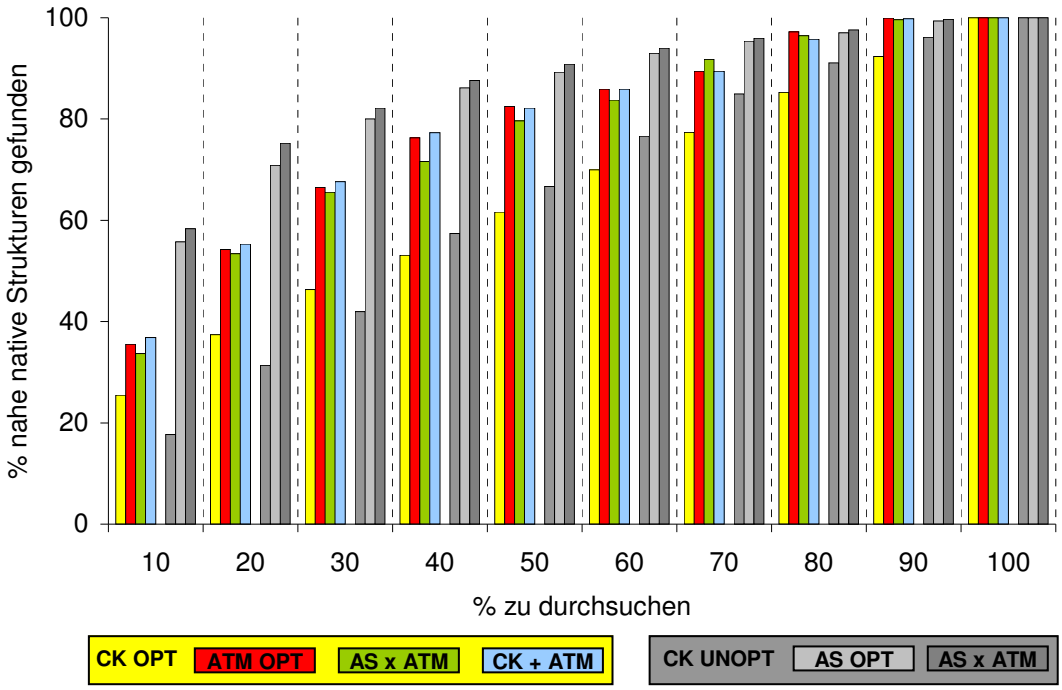

Abbildung 5.13: Vergleich Anteil nahe native Lösungen auf niederen Rängen zwischen ckordo-Lauf mit optimierten aminosäurespezifischen Gewichtungsfaktoren und anschließendem Reranking und normalem ckordo Lauf und anschließendem Reranking

5.4.2 Antikörper-Antigen Komplexe

Durch einen Dockinglauf mit *ckordo*, bei dem die geometrische Korrelation mit den aminosäurespezifisch optimierten Gewichtungsfaktoren berechnet wird, können für zwei Komplexe (1K4C und 2VIS), für die ein normaler *ckordo* Lauf keine nahe native Lösung findet, 4 bzw. 12 Strukturen mit einem RMSiCα ≤ 5 Å vorhergesagt werden. Allerdings kann auch mit den optimierten Gewichtungsfaktoren für die beiden anderen Kandidaten, für die ein normaler *ckordo* Lauf keine nahe native Lösungen findet, keine solche berechnet werden.

Insgesamt werden jedoch durch Benutzung der Gewichtungsfaktoren etwa 2,9 mal so viele nahe native Strukturen gefunden. Im Durchschnitt ist mit den

Gewichtungsfaktoren der niedrigste RMS-Wert etwas niedriger, allerdings ist der Median der RMS-Werte ohne die Gewichtungsfaktoren etwas niedriger.

	Ckordo					**Ckordo mit Gewichtungsfaktoren**				
	Bester Rang	RMS	Rang der besten Struktur	RMS	Anzahl nahe native Lösungen	Bester Rang	RMS	Rang der besten Struktur	RMS	Anzahl nahe native Lösungen
1AHW	**487**	**2,34**	**1890**	**0,91**	**10**	4272	4,56	4580	2,36	8
1BJ1	18489	**3,96**	**18489**	3,96	1	**371**	4,84	21471	**2,15**	**47**
1BVK	2140	**4,57**	12258	**1,49**	16	**512**	4,8	**9928**	2,29	**40**
1DQJ	5318	**3,61**	5318	3,61	8	**87**	4,72	**999**	**0,5**	**36**
1E6J	**146**	4,05	19139	**1,51**	23	1967	**2,96**	**5822**	2,35	**38**
1FSK	**1206**	**1,78**	**3213**	**1,75**	16	1990	2,81	10898	1,78	**21**
1I9R	**81**	3,84	**891**	**1,44**	11	1661	**3,4**	10001	1,82	**21**
1IQD	1765	**3,41**	21710	**1,38**	4	**1642**	3,94	**5751**	2,42	**25**
1JPS	**613**	1,7	8021	1,46	14	1689	**0,65**	**1689**	**0,65**	**12**
1K4C	-	-	8782	5,57	0	**2802**	**4,43**	**6178**	**2,4**	**4**
1KXQ	**6**	2,34	862	1,42	17	98	**0,27**	**98**	**0,27**	**19**
1MLC	**6322**	**1,19**	**6322**	**1,19**	3	13738	4,9	15446	3,64	3
1NCA	**718**	**3,08**	**3134**	1,15	7	1768	4,6	4126	**0,73**	**15**
1NSN	**1655**	**2,89**	8606	**1,21**	7	2112	3,53	**7253**	2,35	**26**
1QFW$_b$	**3228**	**4,32**	**4072**	0,74	14	3656	4,75	9661	0,74	**16**
1VFB	2430	**4,65**	**11259**	**1,59**	15	**1829**	4,49	12477	2,68	**40**
1WEJ	6348	4,69	**12087**	3,69	8	**2726**	**4,21**	19918	**2,14**	**37**
2JEL	2610	4,42	**2610**	4,42	2	**20**	**3,15**	**5524**	**0,5**	**84**
2VIS	-	-	**4643**	6,99	0	**9107**	**4,92**	15451	**1,04**	**12**
Mittelw.	3150,71	**3,34**	**8068,74**	2,39	9,26	**2739,32**	3,79	8803,74	**1,73**	**26,53**
Median	**1765,00**	**3,61**	**6322,00**	**1,49**	8,00	1829,00	4,43	7253,00	2,14	**21,00**
Anzahl besser	**10**	**11**	**11**	9	1	9	8	9	9	**17**

Tabelle 5.12: Vergleich der Vorhersagequalität von ckordo ohne und mit optimierten aminosäurespezifischen Gewichtungsfaktoren für **Antikörper-Antigen Komplexe**. Die jeweils besten Werte sind fett hervorgehoben.

Ferner ist die Anzahl an Komplexen höher, für die ohne die Gewichtungsfaktoren eine nahe native Struktur auf einem niedrigeren Rang gefunden wird (vgl. Tabelle 5.12).

Die weitere Anwendung spezifischer Gewichtungsfaktoren als Rerankingfunktion kann den Anteil nahe nativer Strukturen auf den vorderen Rängen der Vorhersage leicht steigern. Das beste Ergebnis wird hierbei durch die atomspezifischen Gewichtungsfaktoren erzielt (s. Abbildung 5.14).

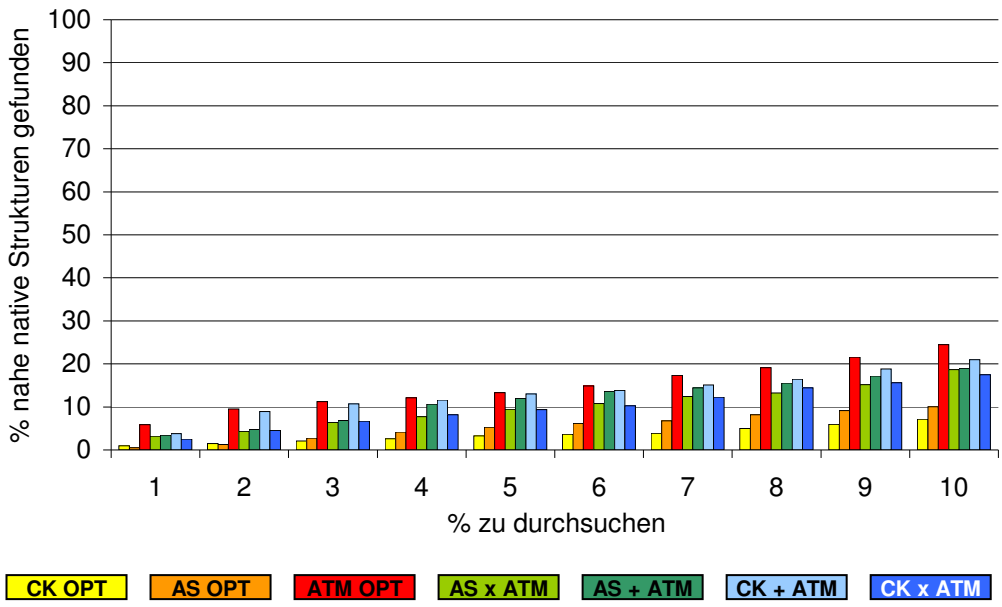

Abbildung 5.14: Verbesserung der Vorhersagequalität bei **Antikörper-Antigen Komplexen** mit aminosäurespezifischen Gewichtungsfaktoren in *ckordo* und anschließendem Reranking in verschiedenen Kombinationen (CK: ckordo mit aminosäurespezifischen Faktoren; AS/ATM: Reranking Funktionen mit Gewichtungsfaktoren)

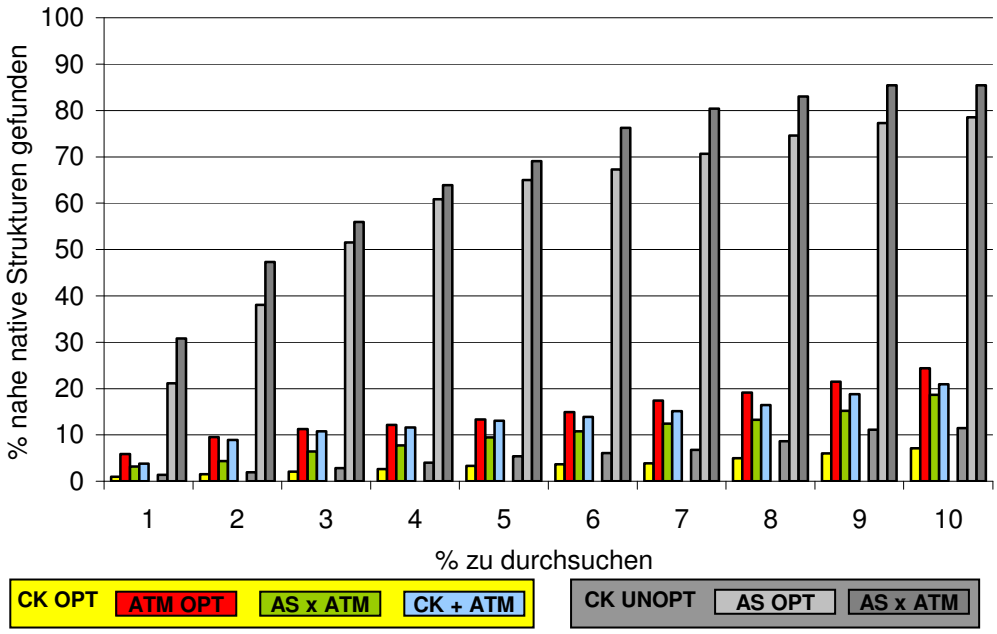

Abbildung 5.15: Vergleich Anteil nahe native Lösungen auf niederen Rängen zwischen *ckordo*-Lauf mit optimierten aminosäurespezifischen Gewichtungsfaktoren und anschließendem Reranking und normalem *ckordo* Lauf und anschließendem Reranking

In Abbildung 5.15 wird indes deutlich, dass verglichen mit einem normalen *ckordo* Lauf mit anschließendem Reranking, die Berechnung der geometrischen Korrelation mit Gewichtungsfaktoren in *ckordo* nicht erfolgreich ist.

Während durch einen gewichteten *ckordo* Lauf mit anschließendem Reranking mit atomspezifischen Gewichtungsfaktoren nur 5,9% aller nahe nativen Strukturen im ersten Prozent der Vorhersage gefunden werden, so werden durch einen normalen *ckordo* Lauf mit anschließendem Ranking mehr als 30% aller nahe nativen Strukturen im ersten Prozent der Vorhersage gefunden (vgl. Abbildung 5.15).

5.4.3 ‚Andere' Komplexe

Benutzt man die aminosäurespezifischen optimierten Gewichtungsfaktoren (ohne I1 optimiert) für die Vorhersage mit *ckordo*, so kann für 17 der ‚Anderen' Komplexe keine Struktur mit einem RMSiCα ≤ 5 Å gefunden werden. Für nur einen Komplex (1I4D) der sieben, für die *ckordo* ohne Gewichtungsfaktoren keine nahe native Lösung findet, können mit den Faktoren vier nahe native Strukturen gefunden werden. Auch im Hinblick auf den Rang der ersten nahe nativen Struktur sind die Gewichtungsfaktoren in *ckordo* nicht erfolgreich. In nur sechs Fällen ist der Rang der ersten nahe nativen Struktur mit Gewichtungsfaktoren niedriger als ohne. Die Anzahl an nahe nativen Strukturen, die gefunden werden - wenn überhaupt welche gefunden werden - ist für 11 der 20 Komplexe mit Gewichtungsfaktoren höher. Tabelle 5.13 zeigt den jeweils besten Rang einer nahe nativen Struktur, ihr RMS-Wert und die Anzahl an nahe nativen Strukturen (für die Komplexe, für welche keine Struktur mit einem RMSiCα ≤ 5 Å gefunden wird, ist jeweils der RMS-Wert und Rang der besten Struktur angegeben).

	Ckordo			**Ckordo mit Gewichtungsfaktoren**		
	Bester Rang	RMS	Anzahl nahe native Lösungen	Bester Rang	RMS	Anzahl nahe native Lösungen
1A2K	338	4,77	13	**12**	**4,76**	**128**
1AK4	3589	4,75	**24**	*9431*	*5,78*	*0*
1AKJ	**76**	**3,5**	**16**	5962	4,54	3
1B6C	467	**2,42**	9	**1**	3,14	**41**
1BUH	**1425**	**4,71**	**4**	*18367*	*5,56*	*0*
1DE4	**17410**	**2,31**	**1**	*11441*	*14,15*	*0*
1E96	787	**4,03**	11	**732**	4,79	**12**

1EER	**3261**	**4**	**3**	*8150*	*7,31*	*0*
1F51	**819**	**1,92**	**26**	2412	3,79	19
1FC2	**1122**	**4,75**	**15**	*9843*	*5,41*	*0*
1FQ1	**3747**	**3,65**	**5**	*16111*	*7,05*	*0*
1FQJ	**231**	**3,98**	**9**	*12054*	*8,03*	*0*
1GCQ	**1653**	**2,11**	14	3249	4,24	**19**
1GHQ	**3547**	4,17	**14**	16038	**4,13**	3
1GP2	**3917**	**1,71**	**10**	4269	4,55	6
1GRN	**250**	**4,5**	**15**	11944	5	3
1HE1	5691	**3,16**	**7**	**2487**	4,81	6
1HE8	**106**	**2,08**	**10**	*18106*	*10,87*	*0*
1I2M	**131**	**3,14**	7	514	4,46	**10**
1I4D	167	5,88	0	**8909**	**4,65**	**4**
1IJK	**597**	4,99	10	9201	**4,73**	**14**
1K5D	**9264**	**4,94**	**2**	*19684*	*5,24*	*0*
1KAC	**345**	**2,85**	**15**	7097	3,6	4
1KLU	14603	3,94	4	**7573**	**3,74**	2
1KTZ	**9816**	**3,37**	**6**	*16540*	*7,35*	*0*
1KXP	**37**	**1,29**	7	113	3	**22**
1ML0	**1109**	**4,62**	6	2148	3,07	**16**
1QA9	**6120**	**3,45**	**7**	*3788*	*8,86*	*0*
1RLB	**205**	4,81	8	255	**4,62**	**52**
1SBB	1804	**4,24**	12	**118**	4,8	**29**
1WQ1	**165**	**3,23**	**10**	18587	3,54	1
Mittelwert	**3087,73**	**3,58**	10,00	7907,61	5,47	**12,71**
Median	**1115,50**	**3,80**	**9,50**	7573,00	4,76	3,00
Anzahl besser	**22**	**23**	**19**	7	6	11

Tabelle 5.13: Bester Rang einer nahe nativen Struktur, ihr RMS und die Anzahl an nahe nativen Strukturen nach Anwendung der optimierten Gewichtungsfaktoren in *ckordo* für **‚Andere' Komplexe** (für die Komplexe, für welche keine Struktur mit einem RMSiCα ≤ 5 Å gefunden wird, ist jeweils der RMS und Rang der besten Struktur angegeben).

5.5 Kombination aus Gewichtungsfaktoren und SVM-Scoringfunktion

5.5.1 Enzym-Inhibitor/Substrat Komplexe

Die von Martin entwickelte umfassende Scoringfunktion[61] liefert im Hinblick auf den Anteil nahe nativer Strukturen auf den niederen Rängen ein deutlich besseres Ergebnis als die hier vorgestellten optimierten Gewichtungsfaktoren (s. Abbildung 5.16). Betrachtet man allerdings die Anzahl an Komplexen, für die eine nahe native

Struktur auf dem ersten Platz der sortierten Vorhersage für die EI-Komplexe des UUPPDD gefunden wird, dann ist das Ergebnis nach den multiplizierten atom- und aminosäurespezifischen Scores besser. Während die SVM Scoringfunktion nur für einen Komplex eine richtige Lösung auf den ersten Platz bringt, gelingt das mit den Gewichtungsfaktoren für 4 Komplexe (Tabelle 5.14).

	Unoptimiert		**Aminosäure-spezifisch x Atomspezifisch**		**SVM Scoring & Geo Score Sortierung**		**SVM Scoring & AS x ATM Sortierung**	
	Rang	**RMS**	**Rang**	**RMS**	**Rang**	**RMS**	**Rang**	**RMS**
1ACB	48	0,93	**1**	2,64	5	0,93	**1**	2,64
1AVW	1434	3,32	55	1,64	16	3,32	**5**	1,64
1BRC	**3**	4,44	34	4,56	10	2,59	7	4,35
1BRS	483	4,21	**1**	1,97	77	2,24	122	2,24
1BVN	33	3,98	211	3,91	**14**	4,62	17	4,62
1CGI	21	4,55	35	4,29	11	4,81	**9**	4,66
1CHO	33	1,03	**1**	4,64	2	1,03	**1**	4,64
1CSE	427	2,27	2924	4,89	**19**	3,53	38	3,53
1DFJ	3598	4,14	241	4,14	192	4,14	**53**	4,14
1FSS	2713	1,62	**170**	4,76	386	4,41	345	4,41
1MAH	471	4,51	54	3,25	42	2,44	**5**	3,25
1PPF	162	4,59	**1**	4,35	5	3,47	**1**	2,46
1TGS	55	4,22	2	1,97	24	4,53	**1**	1,97
1UGH	129	2,72	14	2,02	3	2,72	**1**	2,02
2KAI	**1**	4,74	15	3,96	**1**	4,74	3	3,96
2MTA	466	4,74	581	4,64	**33**	4,26	65	4,64
2PCB	1357	4,48	825	4,42	**266**	4,42	324	4,89
2PCC	1843	4,91	2341	4,12	**68**	4,91	156	4,12
2PTC	174	3,76	4	4,57	27	4,21	**2**	4,57
2SIC	18	2,97	8	4,11	**2**	3,82	**2**	4,11
2SNI	1361	4,57	38	4,66	13	4,83	**7**	4,83
Mittelwert	706	3,65	360	3,79	58	**3,62**	**55**	3,70
Median	174	4,21	35	4,14	16	4,14	**7**	**4,12**
Anzahl 1	1		4		1		**5**	
Anzahl <10	2		7		7		**13**	
Anzahl <100	8		14		**18**		17	

Tabelle 5.14: Ränge und RMS-Werte der ersten nahe nativen Struktur bei Sortierung nach verschiedenen Scoring Schemata (vgl. Text)

Die SVM Scoringfunktion berechnet für relativ viele Strukturen den höchsten möglichen Wert von 1, so dass alle Strukturen, die somit als nahe nativ klassifiziert

werden, nach einem weiteren Kriterium sortiert werden können. Bei Martin[61] erfolgt diese Sortierung nach der von *ckordo* berechneten geometrischen Korrelation.

Wenn dieser Sortierungsschritt für alle als nahe nativ klassifizierten Strukturen mit dem durch die Multiplikation von aminosäure- und atomspezifischen Scores berechneten Wert durchgeführt wird, kann die Vorhersagequalität weiter gesteigert werden. Die Anzahl an Komplexen, für die eine nahe native Struktur auf dem ersten Platz der Vorhersage zu finden ist, erhöht sich auf 5 und für 13 Komplexe findet sich eine nahe native Struktur unter den ersten 10 Lösungen (vgl. Tabelle 5.14).

Betrachtet man die Ergebnisse, die für den mit 12° gedockten Benchmark 2.0 durch diese Kombination von Martins Scoringfunktion und den Gewichtungsfaktoren erreicht werden, so kann der Anteil an Komplexen mit mindestens einer nahe nativen Struktur unter den ersten 50 Strukturen von 82% (nur SVM-Scoring) auf 86% gesteigert werden. Eine Multiplikation der beiden Scoring Schemata führt zu keiner Verbesserung der Endergebnisse.

5.5.2 Antikörper-Antigen Komplexe

Für das Reranking der potentiellen Antikörper-Antigen Strukturen liefert die Kombination aus der umfassenden SVM basierten Scoringfunktion und der

	Unoptimiert		**Aminosäure-spezifisch x Atomspezifisch**		**SVM Scoring & Geo Score Sortierung**		**SVM Scoring & AS x ATM Sortierung**	
	Rang	**RMS**	**Rang**	**RMS**	**Rang**	**RMS**	**Rang**	**RMS**
1AHW	20	1,62	469	1,07	**1**	1,62	39	1,07
1DQJ	34411	4,45	723	3,97	937	3,97	**225**	3,97
1VFB	855	4,89	316	4,89	46	4,76	**21**	4,76
1WEJ	4263	4,03	260	4,03	213	1,66	**176**	1,66
Mittelwert	9887,25	3,7475	442	3,49	299,25	3,0025	**115,25**	2,865
Median	2559	4,24	392,5	4	129,5	2,815	**107,5**	2,815

Tabelle 5.15: Ränge und RMS-Werte der ersten nahe nativen Struktur der UUPPDD Antikörper-Antigen Komplexe bei Sortierung nach verschiedenen Scoring Schemata (vgl. Text)

Multiplikation aus aminosäure- und atomspezifischen Scores im Hinblick auf den Anteil nahe nativer Strukturen im ersten Prozent der Vorhersage das beste Ergebnis (Abbildung 5.17). Für 1AHW kann mit einem Reranking mit Martins Funktion eine nahe native Struktur auf den ersten Platz gebracht werden, was mit den anderen in Tabelle 5.15 aufgeführten Scoringschemata nicht möglich ist. Bei den drei anderen AA-Komplexen allerdings wird der niedrigste Rang durch ein Reranking mit der Kombination aus der SVM-basierten und dem Produkt aus aminosäure- und atomspezifischen Scores erreicht.

Eine nahe native Lösung unter den ersten x Plätzen	Ckordo unoptimiert	SVM-Scoring	Aminosäure-spezifisch x Atomspezifisch	Aminosäure-spezifisch x Atomspezifisch x SVM	Sortierung nach SVM und anschl. nach AS X ATM
1	0	0	1	1	1
10	1	3	2	3	4
25	1	6	3	3	6
50	1	7	3	4	10
100	1	9	7	8	12

Tabelle 5.16: Anzahl Komplexe mit einer nahe nativen Lösung unter den ersten Plätzen für die 19 **Antikörper-Antigen** Komplexe des Benchmark 2.0 (12°). Kombination aus gewichteten Scores und SVM-Scoring

Um einen besseren Eindruck von der Leistungsfähigkeit der kombinierten Scoringfunktionen zu erhalten, müssen hier die Ergebnisse für den mit 12° Winkelinkrement gedockten Benchmark 2.0 Datensatz betrachtet werden (Tabelle 5.16). Das beste Ergebnis wird für die Antikörper-Antigen Komplexe dadurch erreicht, dass die vorgeschlagenen Strukturen erst nach den SVM-Scores und anschließend nach dem Produkt aus aminosäure- und atomspezifischen Gewichtungsfaktoren sortiert werden. Somit kann für 12 der 19 evaluierten AA-Strukturen eine nahe native Lösung unter den ersten 100 Plätzen gefunden werden.

5.5.3 ‚Andere' Komplexe

Keine der untersuchten Rerankingfunktionen oder deren Kombinationen ist in der Lage, den Anteil der nahe nativen Strukturen auf den niedrigen Rängen im Vergleich zur geometrischen Korrelation für die 4 ‚Anderen' Komplexe des UUPPDD

anzureichern oder den besten Rang einer nahe nativen Struktur zu verbessern. Nur die Scoringfunktion von Martin kann den besten nahe nativen Rang für 1AVZ von 13.281 auf 5.699 anheben.

Da die Anzahl der ‚Anderen' Komplexe im UUPPDD so gering ist, ist für diese Komplexe eine zuverlässige Evaluation der Kombination aus SVM- und gewichtetem Scoring nur mit dem mit 12° gedockten Benchmark 2.0 möglich.

Tabelle 5.17 zeigt die Anzahl Komplexe mit einer nahe nativen Lösung unter den ersten Plätzen für verschiedene Kombinationen. Multipliziert man SVM-Scores und gewichtete Scores, so wird für 6 von 29 Komplexen eine nahe native Lösung unter den ersten 10 und für 12 Komplexe unter den ersten 100 Plätzen gefunden.

Eine nahe native Lösung unter den ersten x Plätzen	Ckordo unoptimiert	SVM-Scoring	Aminosäure-spezifisch x Atomspezifisch	Aminosäure-spezifisch x Atomspezifisch x SVM	Sortierung nach SVM und anschl. nach AS X ATM
1	0	0	0	**1**	0
10	1	0	4	**6**	0
25	2	2	5	**7**	2
50	3	5	**7**	**7**	5
100	3	7	9	**12**	7

Tabelle 5.17: Anzahl Komplexe (von 29) mit einer nahe nativen Lösung unter den ersten Plätzen für die 29 **‚Anderen'** Komplexe des Benchmark 2.0 (12°). Kombination aus gewichteten Scores und SVM-Scoring

Das Produkt der aminosäure- und atomspezifischen Scores alleine erbringt ein besseres Ergebnis als das unoptimierte und als das nur mit den SVM-Scores erstellte Ranking (vgl. Tabelle 5.17). Das Ergebnis, welches durch Sortierung nach SVM-Scores und anschließendem Sortieren nach gewichteten Scores erreicht wird, ist dem nur mit den SVM-Scores erzielten Ergebnis gleich.

5.6 Vergleich mit anderen Scoringfunktionen

Im Vergleich mit einigen vorher veröffentlichten Scoringfunktionen ist die hier vorgestellte Methode, bei der die durch aminosäure- und atomspezifischen Gewichtungsfaktoren berechneten Werte multipliziert werden und alle potentiellen Komplexstrukturen danach bewertet und neu sortiert werden relativ erfolgreich. Die einzige Scoringfunktion, die für Enzym-Inhibitor Komplexe ein besseres Ergebnis erbringt, ist die umfassende Scoringfunktion von Martin. Der Anteil an nahe nativen Strukturen im ersten Prozent der Vorhersage kann allerdings durch die Kombination der hier vorgestellten Methode mit Martins Scoringfunktion noch minimal gesteigert werden.

Abbildung 5.16 zeigt für EI-Komplexe den Anteil nahe nativer Strukturen in dem ersten Prozent der Vorhersage für Atomic Contact Energies (ACE)[100], ein Residuen-Residuen Potential[67], ein Atom-Atom Potential[30], die auf komplexklassen-

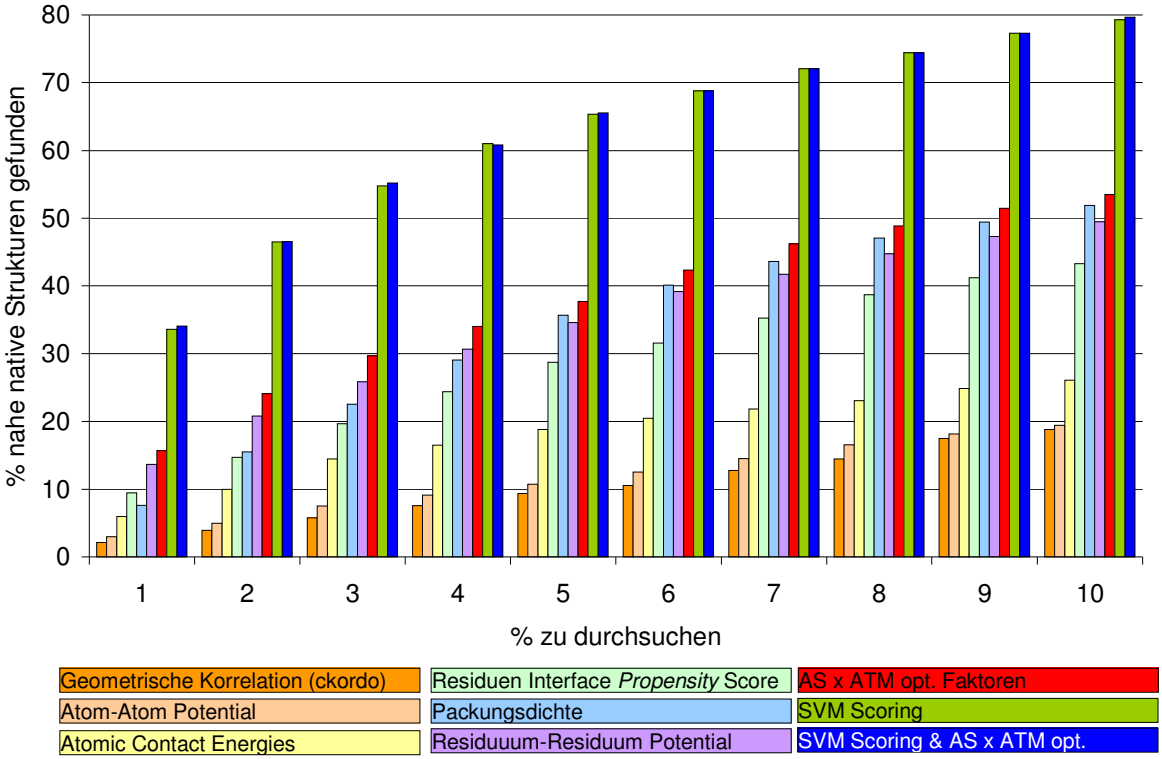

Abbildung 5.16: Vergleich von Vorhersagequalität mit Gewichtungsfaktoren für **Enzym-Inhibitor** Komplexe mit 6 anderen Scoringfunktionen.

spezifischen Residuen Interface-*Propensity* basierte Scoringfunktion von Huang *et al.*[37], die Berechnung der Packungsdichte[27,37] sowie für die umfassende Scoringfunktion von Martin[61] im Vergleich mit den hier vorgestellten Scoringfunktionen.

Für die Antikörper-Antigen Komplexe kann ein Reranking nach der Packungsdichte, nach dem Residuen-Residuen Potential oder mit Martins Scoringfunktion deutlich mehr nahe native Strukturen in das erste Prozent der Vorhersage bringen als dies nur durch Anwendung des Produktes aus aminosäure- und atomspezifischen Faktoren möglich ist. Die Kombination dieses Produktes mit Martins Scoringfunktion kann hingegen für alle 4 Antikörper-Antigen Komplexe eine nahe native Struktur unter das erste Prozent der Vorhersage bringen und führt dazu, dass 50% aller nahe nativen AA-Strukturen im ersten Prozent der Vorhersage liegen (vgl. Abbildung 5.17).

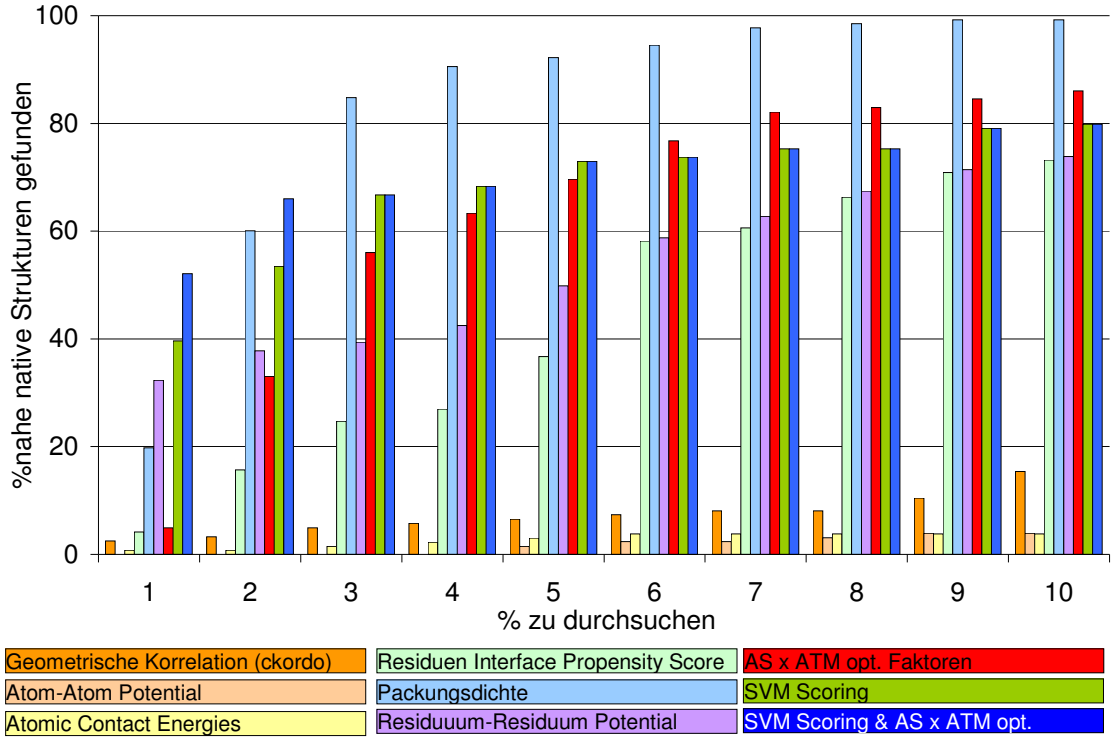

Abbildung 5.17: Vergleich der Vorhersagequalität durch gewichtete geometrische Korrelation für **Antikörper-Antigen Komplexe** mit 6 anderen Scoringfunktionen.

Für die ‚Anderen' Komplexe kann nur die SVM-basierte Scoringfunktion eine leichte Anreicherung nahe nativer Strukturen im ersten Prozent der Vorhersage bewirken. Der Anteil nahe nativer Strukturen steigt von 7,6% auf 11,3%. Alle anderen

Funktionen sind schlechter als das Ranking, welches von *ckordo* nur mit der geometrischen Korrelation erzeugt wird.

5.7 CAPRI

Insgesamt kann die Teilnahme an den CAPRI Runden 9 und 10 als erfolgreich bewertet werden. Da die Kristallstrukturen für alle Targets bislang (August 2006) noch nicht veröffentlicht sind, können hier nur die Ergebnisse relativ zu den Ergebnissen der anderen teilnehmenden Gruppen gezeigt werden, jedoch keine Abbildungen der Strukturen oder andere detaillierte Strukturinformationen.

5.7.1 Target 24

Insgesamt wurden für Target 24 von 36 teilnehmenden Gruppen 321 Strukturen erstellt. Darunter befanden sich keine *High Quality* und keine *Medium* Strukturen. Lediglich vier Strukturen, die von den Arbeitsgruppen um Camacho, Weng und um Totrov vorhergesagt wurden, konnten als *Acceptable* klassifiziert werden.[24]

Die von uns (Heuser, Martin, Schomburg) erzeugten Strukturen sind leider falsch, da dem mit SWISS-Modell erstellten Strukturmodell eine komplette Helix fehlt, die an der Interaktion beteiligt ist.

5.7.2 Target 25

Für Target 25 haben 36 Gruppen insgesamt 336 Strukturen vorhergesagt. Die Arbeitsgruppe von Eisenstein konnte für dieses Target als einzige eine *High Quality* Struktur mit einem Interface RMS von 0,904 Å vorhersagen. Dreizehn Arbeitsgruppen haben eine Struktur der Kategorie *Medium Quality* erstellt (Schomburg, PatchDock, Gramm-X-Server, Vajda, Fernandez-Recio, Totrov, Facemeyer, SKE-DOCK, Takeda-Shitaka, Weng, Smith, Smoothdock, Negi). Ferner gab es 20 Strukturen der Kategorie *Acceptable*.[24]

Die von uns erstellte *Medium Quality* Struktur, ist im Hinblick auf den Backbone Interface RMS mit 1,06 Å die zweitbeste nach der Eisenstein Struktur. Von den 109 Interface Residuen des nativen Komplexes konnten wir 105 richtig vorhersagen.

5.7.3 Target 26

Für Target 26 wurden von 39 Arbeitsgruppen 351 Strukturen vorhergesagt. Allerdings konnte keine Arbeitsgruppe eine *High Quality* Struktur berechnen, aber acht Arbeitsgruppen mindestens eine *Medium Quality* Struktur und weitere 8 Arbeitsgruppen eine *Acceptable* Struktur vorhersagen.[24]

Mit *ckordo*, der umfassenden Scoringfunktion von Martin und den aminosäurespezifischen Gewichtungsfaktoren wurde eine *Acceptable* Struktur mit einem Backbone Interface RMS von 2,32 Å erstellt. Die vorhergesagte Struktur enthält 63% der nativen Interface Residuen.

6 Diskussion

6.1 Zusammenfassung der Ergebnisse

Ziel dieser Arbeit war die Einführung und Optimierung von atom- und aminosäurespezifischen Gewichtungsfaktoren in Dockingprogramme wie *ckordo*, um damit die Vorhersagequalität von Proteinkomplexstrukturen zu steigern.

Dieses Ziel wurde mit den hier vorgestellten aminosäure- und atomspezifischen Gewichtungsfaktoren erfolgreich erreicht. Nach der Anwendung der komplexklassen-spezifischen Gewichtungsfaktoren ist der Anteil nahe nativer Strukturen im ersten Prozent der Vorhersagen 5-15 mal größer. Für bis zu 72% der Komplexe findet sich eine nahe native Struktur auf den ersten 100 Plätzen der Vorhersage.

Insbesondere die Multiplikation der aminosäure- und atomspezifischen Scores kann nahe native Strukturen auf den vorderen Rängen anreichern. In Kombination mit der von Martin entwickelten SVM-basierten, umfassenden Scoringfunktion kann für 62% der Enzym-Inhibitor Komplexe eine nahe native Struktur unter die ersten 10 Plätze der Vorhersage gebracht werden.

Im Folgenden werden zunächst grundlegende Probleme von Protein-Protein Docking diskutiert und im Anschluss daran die hier vorgestellten Rerankingfunktionen analysiert. Die Diskussion der Filter umfasst die Interpretation der optimierten Gewichtungsfaktoren und deren Einordnung in den biochemischen Kontext. Abschließend erfolgt ein Ausblick auf die weiteren Schritte in der Entwicklung von *ckordo*.

6.2 Protein-Protein Docking

Programme zur Strukturvorhersage von Komplexen nähern sich immer mehr der Anwendungsreife. In den letzten CAPRI-Runden konnten für jeden Komplex von

mehreren Gruppen mindestens eine ‚*Medium'* Struktur vorhergesagt werden (mit Ausnahme von Target 24), allerdings jeweils von verschiedenen Gruppen, so dass die Entwicklung von Dockingmethoden noch nicht als erfolgreich abgeschlossen gelten kann.

Bei der Entwicklung von Methoden im Bereich des Protein-Protein Dockings ist die Datengrundlage nach wie vor eines der Hauptprobleme. Selbst der größte nicht redundante Datensatz für Docking umfasst lediglich 83 Komplexe. In Relation zu den eingangs erwähnten geschätzten 70.000 Protein-Protein Interaktionen ist dies ein verschwindend geringer Anteil (0,1%). Die hohe Spezifität der hier vorgestellten Filter zeigt, dass es sinnvoll ist, die Gesamtheit der Proteinkomplexe für die Vorhersage in funktionelle Gruppen einzuteilen. Ob der geringen Anzahl an bekannten Komplexstrukturen ist es allerdings nicht möglich die verschiedenen Komplexe sinnvoll in feinere als die hier benutzten Gruppen einzuteilen. Insbesondere für die Gruppe der ‚Anderen' Komplexe wäre eine weitere Unterteilung möglicherweise sinnvoll.

Allerdings erschwert nicht nur die pure Anzahl an bekannten Strukturen, sondern auch deren Qualität die Entwicklung neuer *unbound* Docking Methoden. Zum Teil stammen z.B. ungebundene und gebundene Strukturen aus verschiedenen Organismen, so dass diese zwar oft sehr ähnlich sind, aber z.T. auch strukturelle und sequentielle Unterschiede aufweisen, und es auch unbekannt ist, ob diese *in vivo* überhaupt aneinander binden.

Erhebliche Schwierigkeiten für das Docking werden auch durch fehlende Residuen in den ungebundenen Proteinen verursacht. In der ungebundenen Struktur des Rezeptors von 1GP2 fehlt z.B. eine komplette Helix, die Teil des Interfaces sein müsste. Aber auch bei 1IB1 fehlen in der ungebundenen Struktur offensichtlich einige im Interface liegende Residuen des N-Terminus.

Ein weiteres Problem stellen solche Testfälle dar, bei denen eine der beiden Untereinheiten ein Homomultimer ist. Diese Multimere müssen erst per Hand erstellt werden, da *ckordo* nicht in der Lage ist, solche Homomere mit C_n-Symetrie vorherzusagen. Es gibt allerdings Programme, die sich auf die Vorhersage von

solchen Homomultimeren spezialisiert haben (z.B. M-ZDOCK[76], SYMMDOCK[85]). Dockt man solche Multimere gegen ein anderes ungebundenes Protein, so gibt es theoretisch mehr als eine richtige Lösung, welche aber mit den in *ckordo* integrierten Methoden zur RMS-Berechnung nicht berücksichtigt werden, so dass nur der 1/n-te Teil der richtigen Strukturen auch als solche erkannt wird. Im benutzten Datensatz wurden für 8 Komplexe künstlich Multimere erstellt (2VIS, 1AKJ, 1ML0, 1RLB, 1I9R, 1K4C, 1KKL, 1EER).

Für das hier benutzte geometriebasierte *rigid-body* Dockingprogramm stellen noch zwei weitere Aspekte ein erhebliches Problem dar: Einerseits kann die Identifikation nahe nativer Strukturen durch die Flexibilität der Proteine erschwert bzw. verhindert werden und andererseits sinkt die Wahrscheinlichkeit basierend auf der geometrischen Korrelation nahe native Lösungen zu finden, wenn das Interface im Verhältnis zur Gesamtoberfläche des Proteins relativ klein ist.

Das Auftreten größerer konformationeller Änderungen zwischen gebundenem und ungebundenem Zustand stellt für *ckordo* das größte Problem dar. Bei fast allen Lösungen, bei denen *ckordo* keine nahe native Lösung gefunden hat, treten während der Bindung mindestens Bewegungen eines Interface Loops auf. Bei einzelnen Komplexen, wie z.B. bei 1H1V, gibt es bei den auf den Komplex gelegten ungebundenen Strukturen sogar massive *Clashes* von Sekundärstrukturelementen.

6.3 Gewichtete geometrische Korrelation

Die mit optimierten aminosäurespezifischen und atomspezifischen Faktoren gewichtete geometrische Korrelation kann als Rerankingfunktion für von einem *rigid-body* Dockingprogramm vorgeschlagene mögliche Komplexstrukturen überzeugen. Durch eine Neusortierung der Komplexstrukturen nach den mit den Gewichtungsfaktoren berechneten Scores können nahe native Strukturen und insbesondere solche mit einem sehr niedrigen RMS auf den vorderen Rängen für alle drei Komplexklassen angereichert werden. Allerdings ist es nicht möglich, die Gewichtungsfaktoren so zu optimieren, dass ein Einsatz direkt bei der Berechnung der geometrischen Korrelation in *ckordo* als erfolgreich zu bewerten wäre.

Es werden durch den direkten Einsatz in *ckordo* für viele Komplexe mehr nahe native Strukturen mit niedrigerem RMS-Wert und auf einem niedrigeren Rang gefunden, so dass die optimierten Gewichtungsfaktoren prinzipiell auch in diesem Fall funktionieren. Die Tatsache, dass sich der Anteil an nicht vorhersagbaren Komplexen, also solche für die keine nahe native Struktur gefunden wird, mehr als verdoppelt hat, ist allerdings inakzeptabel. Das insgesamt unbefriedigende Abschneiden der Gewichtungsfaktoren für die Anwendung in ckordo lässt sich dadurch erklären, dass die Variabilität der zu bewertenden Strukturen im Sampling Schritt weitaus größer ist als während des Reranking Schrittes, während die Gewichtungsfaktoren nur mit den Daten optimiert wurden, die beim Reranking vorliegen. Die Nutzung aller potentiellen Strukturen für die Optimierung hätte die vorhandenen Rechenressourcen überschritten.

Ferner ist bei den Komplexen, bei denen die aminosäurespezifischen Gewichtungsfaktoren bereits in *ckordo* genutzt wurden, ein weiteres Reranking mit den hier vorgestellten Gewichtungsfaktoren nicht erfolgreich. Da alle so berechneten Strukturen bereits einer optimalen Zusammensetzung der Aminosäuren im Interface entsprechen, können die Rerankingfunktionen ihr Wirkungsspektrum nicht mehr entfalten. Daher liefert ein Reranking nach einem klassischen *ckordo* Lauf auch in den Fällen bessere Ergebnisse, wo durch den Einsatz der Gewichtungsfaktoren in *ckordo* mehr nahe native Strukturen auf niedrigeren Rängen gefunden wurden. Als weiteres Argument kann auch hier angeführt werden, dass die Filter die Unterscheidung zwischen richtigen und falschen Strukturen an Hand eines anderen Spektrums an diversen Strukturen *gelernt* haben als nach einem gewichteten *ckordo* Lauf vorhanden ist. Eine erneute spezifische Optimierung an Hand dieser Strukturen ist allerdings nicht sinnvoll, da der Einsatz der Gewichtungsfaktoren angesichts der großen Anzahl an nicht vorhersagbaren Komplexen ohnehin nicht in Frage kommt.

Insgesamt lässt sich feststellen, dass die Rerankingfunktion umso erfolgreicher ist, je homogener die Gruppe der Komplexe ist, für die sie optimiert wurden. Im Hinblick auf den Anteil nahe nativer Strukturen auf den vorderen Rängen wird das beste Ergebnis für die Antikörper-Antigen Komplexe erreicht, gefolgt von den Enzym-Inhibitor Komplexen, während die Ergebnisse für die sehr heterogene Gruppe der ‚Anderen'

Komplexe nicht ganz so gut sind. Das etwas schlechtere Abschneiden der Ergebnisse für die ‚Anderen' Komplexe in Kombination mit der Tatsache, dass die optimierten Faktoren jeweils spezifisch für die Gruppe an Komplexen ist, für die sie optimiert wurden, spricht für eine weitere funktionelle Unterteilung der ‚Anderen' Komplexe, so dass auch für diese Komplexe spezifischere Faktoren optimiert werden können.

In einigen Fällen war es nicht möglich - trotz Einsatzes der Gewichtungsfaktoren - eine nahe native Struktur unter die ersten 1.000 Ränge zu bringen. Diese Komplexe durchlaufen entweder größere konformationelle Änderungen und sind daher auch als schwierig oder mittel schwierig klassifiziert (z.B. 1DE4, 1EER), oder diese Komplexe haben ein sehr kleines Interface (z.B. 1GHQ: 800 $Å^2$ ΔASA) oder ein sehr locker gepacktes Interface (z.B. 2PCC) und sind daher mit geometriebasiertem Docking schwierig vorherzusagen. Allerdings ist kein klarer Zusammenhang zwischen der Wirksamkeit der Rerankingfunktion und der jeweiligen Schwierigkeit erkennbar. Sofern von *ckordo* selbst überhaupt nahe native Strukturen gefunden werden, können diese auch bei den meisten Komplexen von den Scoringfunktionen erkannt werden.

Durch die Validierung der optimierten Gewichtungsfaktoren konnte gezeigt werden, dass diese in der Lage sind, auch die Vorhersage von unbekannten Komplexen zu verbessern. Allerdings war es nicht möglich, an den vier ‚Anderen' Komplexen des UUPPDD das Funktionieren für ‚Andere' Komplexe nachzuweisen. Da aber die Ergebnisse für den mit 12° Winkelinkrement berechneten Benchmark 2.0 überzeugen, ist es wahrscheinlich, dass auch die für ‚Andere' Komplexe optimierten Gewichtungsfaktoren die Vorhersage neuer bislang unbekannter Komplexe deutlich verbessern.

6.3.1 Aminosäurespezifische Gewichtungsfaktoren

Die optimierten aminosäurespezifischen Gewichtungsfaktoren stehen zum Teil in Einklang mit den bekannten Eigenschaften von Protein-Protein Bindungsstellen und lassen sich somit in den biochemischen Kontext einordnen.

Für solche Aminosäuren, die wegen ihrer langen und flexiblen Seitenketten wahrscheinlich Kollisionen mit dem Bindungspartner auslösen, und somit *rigid-body unbound* Dockingmethoden in die Irre führen können, wurden sehr niedrige Gewichtungsfaktoren optimiert. Für ARG, ASP, GLN, GLU und LYS, also für die flexiblen und polaren Aminosäuren, die zudem eine sehr niedrige Interface-*Propensity* aufweisen, wurden die niedrigsten Werte errechnet.

Dem gegenüber stehen sehr hohe Werte für die aromatischen Residuen, die eine sehr hohe Interface-*Propensity* aufweisen und mit ihren Ringsystemen so genannte aromatische *π-stacks* (*π-π-Stapel-Wechselwirkungen)* ausbilden können. Des Weiteren lassen sich die hohen Faktoren für die Aromaten durch die Rigidität der Ringsysteme erklären.

Gerade für die Enzym-Inhibitor Komplexe korrespondieren die Gewichtungsfaktoren mit den Interface Propensities[16,56], der Anzahl frei drehbarer Bindungen in den Seitenketten und der Hydrophobizität[49] der Aminosäuren.

Der sehr hohe Faktor für Methionin steht in Einklang mit der hohen Wahrscheinlichkeit MET im Interface zu finden und der sehr geringen Wahrscheinlichkeit Methionin überhaupt auf der Proteinoberfläche anzutreffen[16,56], so dass ein auf der Oberfläche befindliches MET ein deutlicher Hinweis auf die Interface-Region sein kann.

Auch die für die Antikörper-Antigen Komplexe optimierten Parameter lassen sich mit den oben genannten Argumenten erklären, allerdings korrelieren die Faktoren nicht so stark mit diesen Eigenschaften, wie die Faktoren für die EI-Komplexe. Die relativ hohen Werte für ASN und TYR entsprechen der größeren Bedeutung von

Wasserstoffbrückenbindungen, bzw. der größeren Bedeutung der elektrostatischen Wechselwirkungen bei Antikörper-Antigen Bindungen[42].

Die Gruppe der ‚Anderen' Komplexe ist so heterogen, dass eine eindeutige Interpretation der Parameter schwierig ist. Dennoch wurden auch hier die höchsten Gewichtungsfaktoren für die apolaren, hydrophoben und rigiden Residuen optimiert, während die Werte für die Aminosäuren mit langen flexiblen Seitenketten eher gering ausfallen.

Da insbesondere bei den Enzym-Inhibitor Komplexen, aber auch bei den ‚Anderen' Komplexen die optimierten Werte mit der Anzahl an frei drehbaren Bindungen in der Seitenkette zusammenhängen, können die Gewichtungsfaktoren als eine nicht besonders rechenintensive Methode angesehen werden, um implizit Flexibilität von Seitenketten beim Docking zu berücksichtigen. Die Methode kann nicht zur richtigen Orientierung der Seitenketten führen, aber dennoch helfen den Fehler zu relativieren, der durch ‚falsch' liegende Seitenketten hervorgerufen wird, die auch bei nahe nativen Strukturen Kollisionen verursachen. Die richtige Konformation der Seitenketten muss mit den Methoden des Refinements berechnet werden.

Auch die Werte, die die Optimierung für das Innere des Rezeptors (I1) ergeben hat, machen im Hinblick auf die unterschiedliche Art der Interaktion Sinn. Der positive Wert, der für das Innere der Enzyme optimiert wurde, irritiert auf den ersten Blick, da in einer Standard Fourier Docking Prozedur dieser Wert immer negativ ist. Da allerdings in der Optimierung und auch während des Rerankings lediglich solche Komplexe vorkommen, die primär Oberflächen-Oberflächen Kontakte aufweisen und somit keine größeren Kollisionen zwischen den Bindungspartnern aussortiert werden müssen, können in diesem Schritt durchaus positive Werte für I1 herauskommen. Gerade bei EI-Komplexen, die ein sehr eng gepacktes Interface aufweisen, können auch die nahe nativen Strukturen auf Grund der Flexibilität der Proteine kleinere Kollisionen aufweisen. Der erhaltene hohe positive Wert deutet darauf hin, dass gerade bei den nahe nativen Strukturen Kollisionen vorkommen. Die Strukturen, die im Reranking Schritt noch Kollisionen aufweisen, trotz der 'Bestrafung' von Kollisionen während der Berechnung der geometrischen Korrelation in *ckordo*, müssen eine so hohe geometrische Passgenauigkeit aufweisen, dass sie nicht

aussortiert werden. Diese Komplexstrukturen haben also eine exzellente Passgenauigkeit, abgesehen von einigen kleineren vermutlich durch flexible Seitenketten ausgelösten Kollisionen und erhalten durch eine Neusortierung mit einem positiven I1 Wert einen niedrigeren Rang.

Die Antikörper-Antigen Komplexe hingegen haben eine weitaus geringere Packungsdichte, so dass hier auftretende Kollisionen keinen Hinweis auf das native Interface geben können und daher auch während des Rerankings durch einen negativen I1 Wert 'bestraft' werden.

Für die sehr heterogene Gruppe der ‚Anderen' Komplexe wurde ein Wert nahe Null optimiert. Dies deutet darauf hin, dass Kollisionen hier weder durchgängig negativ noch positiv zu bewerten sind.

6.3.2 Atomspezifische Gewichtungsfaktoren

Die sehr niedrigen Werte für die Backbone Atome, die bei der Optimierung der atomspezifischen Gewichtungsfaktoren berechnet wurden, liegen wahrscheinlich darin begründet, dass sich die Art und Anzahl der Kontakte des Backbones bei nahe nativen und falschen Strukturen nicht wesentlich unterscheiden und die Gewichtungsfaktoren für diese Atome somit keine Rolle spielen bei der Differenzierung zwischen nahe nativen und falschen Strukturen. Ferner decken sich die niedrigen Backbone Werte auch mit der bei Neuvirth *et al.*[69] beschriebenen niedrigen Interface-*Propensity* für einige Cα-Atome.

Wie aus der Interpretation der aminosäurespezifischen Faktoren zu erwarten, wurden für die Atome, die in den Ringen der aromatischen Residuen liegen, hohe bis sehr hohe Werte errechnet, so dass auch bei den atomspezifischen Parametern den hohen Interface Propensities und der Fähigkeit, aromatische *π-stacks* auszubilden, Rechnung getragen wird.

Insbesondere bei den Antikörper-Antigen und bei den ‚Anderen' Komplexen wurden hohe und mittlere Werte für die Seitenkettenatome optimiert, die häufig für die

Bildung von Wasserstoffbrücken benötigt werden. Die Seitenkettenatome, die nur bei den hydrophoben Residuen mit kurzer und wenig flexibler Seitenkette, wie z.B. SER oder LEU vorkommen, erhielten bei allen drei Komplexgruppen mittlere Werte durch die Optimierung.

Die Methylengruppen, die bei mehreren - auch langen - Seitenketten vorkommen, haben bei den AA- und bei den OTH-Komplexen gegen Null gehende Werte erhalten, spielen aber bei den EI-Komplexen eine größere Rolle.

Während sich also abgesehen von den Kohlenstoffatomen der kurzen hydrophoben Seitenketten die höheren Werte bei den Antikörper-Antigen und den ‚Anderen' Komplexen nur auf die funktionellen Gruppen, welche an Wasserstoffbrücken und polaren Wechselwirkungen beteiligt sind, und die Atome der aromatischen Ringe konzentrieren, haben bei den EI-Komplexen die Seitenkettenatome der Residuen mit einer hohen Interface-*Propensity* (ILE, TRP, TYR, MET, PHE, CYS, HIS) durchgängig höhere Gewichtungsfaktoren erhalten. Zusammenfassend kann also gesagt werden, dass auch die atomspezifischen Gewichtungsfaktoren der jeweiligen biochemischen Rolle der Atome und deren Bedeutung für die Interaktion zweier Proteine entsprechen.

Für die Atome, die in mehreren unterschiedlichen Residuen vorkommen (wie z.B. die Backboneatome und die Kohlenstoffatome der Gruppe 8), wäre eine weitere Unterteilung in Untergruppen möglicherweise sinnvoll, abhängig von den Residuen, in denen sie vorkommen, oder abhängig von der Distanz zum Backbone. Allerdings würde eine weitere Unterteilung zu einem höheren Anspruch an Speicherbedarf und Rechenzeit führen, der die zur Verfügung stehenden Ressourcen überstiegen hätte.

Die für das Innere des Rezeptors (I1) optimierten Werte verhalten sich bei der atomspezifischen Optimierung ähnlich wie bei der aminosäurespezifischen, so dass auch hier die oben gegebenen Erklärungen zutreffen. Die jeweiligen Werte sind zwar niedriger (zwischen 1 und -1), was aber eher an der Abhängigkeit von den anderen optimierten Faktoren liegt, als eine biochemische Ursache hat.

6.3.3 Kombination mit SVM-Scoring

Die Kombination aus der mit den spezifischen Gewichtungsfaktoren berechneten geometrischen Korrelation und der umfassenden SVM-Scoringfunktion von Martin kann die Vorhersagequalität weiter steigern.

Im direkten Vergleich von SVM-basiertem Scoring und der gewichteten geometrischen Korrelation ist die Anreicherung nahe nativer Strukturen auf den vorderen Rängen durch die SVM-Scoringfunktion erfolgreicher. Allerdings ist der niedrigste Rang auf dem eine nahe native Struktur gefunden wird nach dem Reranking mit der gewichteten geometrischen Korrelation oftmals niedriger.

Da bislang die geometrische Korrelation in Martins Scoringschema nur in Form der von *ckordo* berechneten Passgenauigkeit einfließt, ist eine weitere Steigerung durch die optimierte gewichtete Korrelation zu erwarten gewesen.

Die erfolgreichste Art der Kombination unterscheidet sich jedoch nach den verschiedenen Komplexklassen. Während für die Enzym-Inhibitor und für die Antikörper-Antigen Komplexe das beste Ergebnis dadurch erreicht wird, die potentiellen Komplexstrukturen zuerst nach Martins Scoringfunktion zu sortieren, und anschließend alle Strukturen, die identische SVM-Scores erhalten haben, nach der gewichteten Korrelation zu sortieren, wird für die meisten ‚Anderen' Komplexe eine nahe native Struktur auf den vorderen Rängen durch das Produkt aus aminosäurespezifischen, atomspezifischen und SVM-Scores gefunden.

Dies kann darin begründet liegen, dass die Spezifität von Martins Scoringfunktion für die ‚Anderen' Komplexe deutlich schlechter ist als für AA- und EI-Komplexe, so dass es relativ viele ‚Falsch Negative' und ‚Falsch Positive' gibt[61]. Ferner ist die Anzahl der Strukturen, die durch die SVM-Scoringfunktion den Wert 1 zugewiesen bekommen, relativ gering, so dass eine sekundäre Sortierung nach der gewichteten geometrischen Korrelation keine weitere Steigerung des Endergebnisses bringt.

7 Ausblick

Mit dem erfolgreichen Abschluss der Entwicklung der Rerankingfunktionen mit aminosäure- und atomspezifischen Gewichtungsfaktoren und der umfassenden Scoringfunktion von Martin ist der Rahmen für Fortschritte im Scoring Schritt der Dockingprozedur weitgehend erschöpft.

Damit *ckordo* allerdings auch in Zukunft ein konkurrenzfähiges Dockingprogramm ist, sollte insbesondere an drei Stellen der Dockingprozedur weiterhin Arbeit investiert werden. Der wohl wichtigste Schritt ist die Geschwindigkeitsoptimierung des Samplingschrittes, die unter Berücksichtigung der neuesten Methoden zur Parallelisierung von Rechenaufgaben erfolgen sollte, so dass es gegebenenfalls möglich ist, die Flexibilität der Proteine durch Ensemble Docking in realistischer Zeit zu berücksichtigen.

Die Entwicklung neuer Algorithmen zur Verbesserung der Vorhersagequalität könnte sich auf zwei Stellen konzentrieren, nämlich zum einen auf die Analyse der Proteine vor dem eigentlichen Docking und auf das Refinement der nahe nativen Strukturen.

Fortschritte im Bereich des Refinements können durch Energieminimierungen und durch gezielte Behandlungen der flexiblen Bereiche der Proteine zu niedrigeren RMSD-Werten führen und somit die Qualität der vorhergesagten Strukturen erheblich steigern.

Die Analyse der Proteine vor dem eigentlichen Docking sollte zwei Ziele verfolgen. Zum einen kann versucht werden, die Interface Regionen vorherzusagen, um somit das Absuchen des geometrischen Raumes effizienter gestalten zu können. Der andere, möglicherweise wichtigere Teil der Voranalyse kann sich auf die Identifikation flexibler Bereiche der Proteine beziehen, um die Komplexität der expliziten Behandlung flexibler Bereiche einzuschränken.

Insgesamt bietet der aktuelle Entwicklungsstand von *ckordo* allerdings eine gute Grundlage für ein erfolgreiches Dockingprogramm.

8 Literatur

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. **Basic local alignment search tool**. J Mol Biol 1990; 215(3): 403-410.

2. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. **SCOP database in 2004: refinements integrate structure and sequence family data**. Nucleic Acids Res 2004; 32(Database issue): D226-229.

3. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. **UniProt: the Universal Protein knowledgebase**. Nucleic Acids Res 2004; 32 Database issue: D115-119.

4. Arkin M. **Protein-protein interactions and cancer: small molecules going in for the kill**. Curr Opin Chem Biol 2005; 9(3): 317-324.

5. Ausiello G, Cesareni G, Helmer-Citterich M. **ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure**. Proteins 1997; 28(4): 556-567.

6. Bastard K, Prevost C, Zacharias M. **Accounting for loop flexibility during protein-protein docking**. Proteins 2006; 62(4): 956-969.

7. Berchanski A, Shapira B, Eisenstein M. **Hydrophobic complementarity in protein-protein docking**. Proteins 2004; 56(1): 130-142.

8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. **The Protein Data Bank**. Nucleic Acids Res 2000; 28(1): 235-242.

9. Betts MJ, Sternberg MJ. **An analysis of conformational changes on protein-protein association: implications for predictive docking**. Protein Eng 1999; 12(4): 271-283.

10. Bonvin AM, Boelens R, Kaptein R. **NMR analysis of protein interactions**. Curr Opin Chem Biol 2005; 9(5): 501-508.

11. Bonvin AM. **Flexible protein-protein docking**. Curr Opin Struct Biol 2006; 16(2): 194-200.

12. Bordner AJ, Abagyan R. **Statistical analysis and prediction of protein-protein interfaces**. Proteins 2005; 60(3): 353-366.

13. Camacho CJ, Gatchell DW, Kimura SR, Vajda S. **Scoring docked conformations generated by rigid-body protein-protein docking**. Proteins 2000; 40(3): 525-537.

14. Camacho CJ, Vajda S. **Protein docking along smooth association pathways**. Proc Natl Acad Sci U S A 2001; 98(19): 10636-10641.

15. Carter P, Lesk VI, Islam SA, Sternberg MJ. **Protein-protein docking using 3D-Dock in rounds 3, 4, and 5 of CAPRI**. Proteins 2005; 60(2): 281-288.

16. Chakrabarti P, Janin J. **Dissecting protein-protein recognition sites**. Proteins 2002; 47(3): 334-343.

17. Chen R, Weng Z. **Docking unbound proteins using shape complementarity, desolvation, and electrostatics**. Proteins 2002; 47(3): 281-294.

18. Chen R, Mintseris J, Janin J, Weng Z. **A protein-protein docking benchmark**. Proteins 2003; 52(1): 88-91.

19. Clavel T, Germon P, Vianney A, Portalier R, Lazzaroni JC. **TolB protein of Escherichia coli K-12 interacts with the outer membrane peptidoglycan-associated proteins Pal, Lpp and OmpA**. Mol Microbiol 1998; 29(1): 359-367.

20. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. **ClusPro: a fully automated algorithm for protein-protein docking**. Nucleic Acids Res 2004; 32(Web Server issue): W96-99.

21. Connolly ML. **Solvent-accessible surfaces of proteins and nucleic acids**. Science 1983; 221(4612): 709-713.

22. Dennis JE, Schnabel RB. **Numerical Methods for Unconstrained Optimization and Nonlinear Equations**. Prentice-Hall, Englewood Cliffs, NJ 1983.

23. Eisenstein M, Katchalski-Katzir E. **On proteins, grids, correlations, and docking**. C R Biol 2004; 327(5): 409-420.

24. EMBL/EBI-MSD Group. **CAPRI-Home Critical Assessment of PRediction of Interactions.** 2006 [http://capri.ebi.ac.uk/]

25. Gabb HA, Jackson RM, Sternberg MJ. **Modelling protein docking using shape complementarity, electrostatics and biochemical information**. J Mol Biol 1997; 272(1): 106-120.

26. Gardiner EJ, Willett P, Artymiuk PJ. **Protein docking using a genetic algorithm**. Proteins 2001; 44(1): 44-56.

27. Gottschalk KE, Neuvirth H, Schreiber G. **A novel method for scoring of docked protein complexes using predicted protein-protein binding sites**. Protein Eng Des Sel 2004; 17(2): 183-189.

28. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. **Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations**. J Mol Biol 2003; 331(1): 281-299.

29. Gray JJ. **High-resolution protein-protein docking**. Curr Opin Struct Biol 2006; 16(2): 183-193.

30. Grimm V. **Untersuchung eines wissensbasierten Potentials zur Bewertung von Protein-Protein-Docking-Studien.** Köln: Universität zu Köln; 2003.

31. Grunberg R, Leckner J, Nilges M. **Complementarity of structure ensembles in protein-protein binding**. Structure 2004; 12(12): 2125-2136.

32. Halperin I, Ma B, Wolfson H, Nussinov R. **Principles of docking: An overview of search algorithms and a guide to scoring functions**. Proteins 2002; 47(4): 409-443.

33. Heifetz A, Katchalski-Katzir E, Eisenstein M. **Electrostatics in protein-protein docking**. Protein Sci 2002; 11(3): 571-587.

34. Heifetz A, Eisenstein M. **Effect of local shape modifications of molecular surfaces on rigid-body protein-protein docking**. Protein Eng 2003; 16(3): 179-185.

35. Heuser P, Martin O. **UUPPDD - Unbound Unbound Protein Protein Docking Dataset.** 2003 [http://biotool.uni-koeln.de:8080/uuppdd/]

36. Heuser P, Bau D, Benkert P, Schomburg D. **Refinement of unbound protein docking studies using biological knowledge**. Proteins 2005; 61(4): 1059-1067.

37. Huang B, Schroeder M. **Using residue propensities and tightness of fit to improve rigid-body protein-protein docking.** In: Matthias R, Andrew T, Kurtz S, Willhoeft U, editors. Proceedings of the German Conference on Bioinformatics (GCB2005), Hamburg, Germany, October 5-7, 2005. Volume 71, LNI: GI; 2005. S. 159-173.

38. Huang PS, Love JJ, Mayo SL. **Adaptation of a fast Fourier transform-based docking algorithm for protein design**. J Comput Chem 2005; 26(12): 1222-1232.

39. Jackson RM, Sternberg MJ. **Application of scaled particle theory to model the hydrophobic effect: implications for molecular association and protein stability**. Protein Eng 1994; 7(3): 371-383.

40. Jackson RM, Sternberg MJ. **A continuum model for protein-protein interactions: application to the docking problem**. J Mol Biol 1995; 250(2): 258-275.

41. Jackson RM, Gabb HA, Sternberg MJ. **Rapid refinement of protein interfaces incorporating solvation: application to the docking problem**. J Mol Biol 1998; 276(1): 265-285.

42. Jackson RM. **Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem**. Protein Sci 1999; 8(3): 603-613.

43. Janin J, Chothia C. **The structure of protein-protein recognition sites**. J Biol Chem 1990; 265(27): 16027-16030.

44. Janin J. **Kinetics and thermodynmaics of protein-protein interactions**. In: Kleanthous C, editor. Protein-Protein Recognition, Frontiers in molecular biology 31. Oxford: Oxford University Press; 2000. S. 1-32.

45. Jiang F, Kim SH. **"Soft docking": matching of molecular surface cubes**. J Mol Biol 1991; 219(1): 79-102.

46. Jones S, Thornton JM. **Analysis of protein-protein interaction sites using surface patches**. J Mol Biol 1997; 272(1): 121-132.

47. Jones S, Thornton JM. **Analysis and classification of protein-protein interactions from a structural perspective**. In: Kleanthous C, editor. Protein-Protein Recognition, Frontiers in molecular biology 31. Oxford: Oxford University Press; 2000. S. 33-59.

48. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. **Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques**. Proc Natl Acad Sci U S A 1992; 89(6): 2195-2199.

49. Kyte J, Doolittle RF. **A simple method for displaying the hydropathic character of a protein**. J Mol Biol 1982; 157(1): 105-132.

50. Lawrence MC, Colman PM. **Shape complementarity at protein/protein interfaces**. J Mol Biol 1993; 234(4): 946-950.

51. Lee K, Sim J, Lee J. **Study of protein-protein interaction using conformational space annealing**. Proteins 2005; 60(2): 257-262.

52. Lehner B, Fraser AG. **A first-draft human protein-interaction map**. Genome Biol 2004; 5(9): R63.

53. Lichtarge O, Bourne HR, Cohen FE. **An evolutionary trace method defines binding surfaces common to protein families**. J Mol Biol 1996; 257(2): 342-358.

54. Lin SL, Nussinov R, Fischer D, Wolfson HJ. **Molecular surface representations by sparse critical points**. Proteins 1994; 18(1): 94-101.

55. Lin SL, Nussinov R. **Molecular recognition via face center representation of a molecular surface**. J Mol Graph 1996; 14(2): 78-90, 95-77.

56. Lo Conte L, Chothia C, Janin J. **The atomic structure of protein-protein recognition sites**. J Mol Biol 1999; 285(5): 2177-2198.

57. Lorber DM, Udo MK, Shoichet BK. **Protein-protein docking with multiple residue conformations and residue substitutions**. Protein Sci 2002; 11(6): 1393-1408.

58. Ma B, Elkayam T, Wolfson H, Nussinov R. **Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces**. Proc Natl Acad Sci U S A 2003; 100(10): 5772-5777.

59. Ma XH, Li CH, Shen LZ, Gong XQ, Chen WZ, Wang CX. **Biologically enhanced sampling geometric docking and backbone flexibility treatment with multiconformational superposition**. Proteins 2005; 60(2): 319-323.

60. Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF. **Protein docking using continuum electrostatics and geometric fit**. Protein Eng 2001; 14(2): 105-113.

61. Martin O. **Effcient comprehensive scoring of docked protein complexes - a machine learning approach**. Cologne: University of Cologne; 2006.

62. Melo F, Feytmans E. **Novel knowledge-based mean force potential at atomic level**. J Mol Biol 1997; 267(1): 207-222.

63. Mendez R, Leplae R, De Maria L, Wodak SJ. **Assessment of blind predictions of protein-protein interactions: current status of docking methods**. Proteins 2003; 52(1): 51-67.

64. Mendez R, Leplae R, Lensink MF, Wodak SJ. **Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures**. Proteins 2005; 60(2): 150-169.

65. Meyer M, Wilson P, Schomburg D. **Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking**. J Mol Biol 1996; 264(1): 199-210.

66. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z. **Protein-Protein Docking Benchmark 2.0: an update**. Proteins 2005; 60(2): 214-216.

67. Moont G, Gabb HA, Sternberg MJ. **Use of pair potentials across protein interfaces in screening predicted docked complexes**. Proteins 1999; 35(3): 364-373.

68. Najmanovich R, Kuttner J, Sobolev V, Edelman M. **Side-chain flexibility in proteins upon ligand binding**. Proteins 2000; 39(3): 261-268.

69. Neuvirth H, Raz R, Schreiber G. **ProMate: a structure based prediction program to identify the location of protein-protein binding sites**. J Mol Biol 2004; 338(1): 181-199.

70. Newton I. **Methodus fluxionum et serierum infinitarum**. 1664-1671.

71. Nooren IM, Thornton JM. **Diversity of protein-protein interactions**. Embo J 2003; 22(14): 3486-3492.

72. Norel R, Petrey D, Wolfson HJ, Nussinov R. **Examination of shape complementarity in docking of unbound proteins**. Proteins 1999; 36(3): 307-317.

73. Palma PN, Krippahl L, Wampler JE, Moura JJ. **BiGGER: a new (soft) docking algorithm for predicting protein interactions**. Proteins 2000; 39(4): 372-384.

74. Pauling L, Delbrück M. **The nature of the intermolecular forces operative in biological processes**. Science 1940; 92: 77-79.

75. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham III TE, DeBolt S, Ferguson D, Seibel G, Kollman P. **AMBER, a package of computer programs for applying molecular meechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the**

structural and energetic properties of molecules. Comp Phys Commun 1995; (91): 1-41.

76. Pierce B, Tong W, Weng Z. **M-ZDOCK: a grid-based approach for Cn symmetric multimer docking**. Bioinformatics 2005; 21(8): 1472-1478.

77. R Development Core Team. **R: A language and environment for statistical computing**. 2005.

78. Raphson J. **Analysis Aequationum universalis**. London; 1690.

79. Ray MC, Germon P, Vianney A, Portalier R, Lazzaroni JC. **Identification by genetic suppression of Escherichia coli TolB residues important for TolB-Pal interaction**. J Bacteriol 2000; 182(3): 821-824.

80. Reddy BV, Kaznessis YN. **A quantitative analysis of interfacial amino acid conservation in protein-protein hetero complexes**. J Bioinform Comput Biol 2005; 3(5): 1137-1150.

81. Russell RB, Alber F, Aloy P, Davis FP, Korkin D, Pichaud M, Topf M, Sali A. **A structural perspective on protein-protein interactions**. Curr Opin Struct Biol 2004; 14(3): 313-324.

82. Ryan DP, Matthews JM. **Protein-protein interactions in human disease**. Curr Opin Struct Biol 2005; 15(4): 441-446.

83. Schnabel RB, Koontz, J. E. and Weiss, B. E. **A modular system of algorithms for unconstrained minimization.** ACM Trans Math Software 1985; 11: 419-440.

84. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. **Geometry-based flexible and symmetric protein docking**. Proteins 2005; 60(2): 224-231.

85. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. **PatchDock and SymmDock: servers for rigid and symmetric docking**. Nucleic Acids Res 2005; 33(Web Server issue): W363-367.

86. Schwede T, Kopp J, Guex N, Peitsch MC. **SWISS-MODEL.** [http://swissmodel.expasy.org/SWISS-MODEL.html]

87. Schwede T, Kopp J, Guex N, Peitsch MC. **SWISS-MODEL: An automated protein homology-modeling server**. Nucleic Acids Res 2003; 31(13): 3381-3385.

88. Sharp KA, Nicholls A, Friedman R, Honig B. **Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models**. Biochemistry 1991; 30(40): 9686-9697.

89. Shindyalov IN, Bourne PE. **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path**. Protein Eng 1998; 11(9): 739-747.

90. Smith GR, Sternberg MJ. **Evaluation of the 3D-Dock protein docking suite in rounds 1 and 2 of the CAPRI blind trial**. Proteins 2003; 52(1): 74-79.

91. Smith GR, Sternberg MJ, Bates PA. **The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking**. J Mol Biol 2005; 347(5): 1077-1101.

92. Taylor JS, Burnett RM. **DARWIN: a program for docking flexible molecules**. Proteins 2000; 41(2): 173-191.

93. Uetz P, Pohl E. **Protein-Protein und Protein-DNA-Interaktionen**. In: Wink M, editor. Molekulare Biotechnologie. Weinheim: Wiley-VCH; 2004. S. 385-407.

94. Vakser IA, Aflalo C. **Hydrophobic docking: a proposed enhancement to molecular recognition techniques**. Proteins 1994; 20(4): 320-329.

95. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. **GROMACS: fast, flexible, and free**. J Comput Chem 2005; 26(16): 1701-1718.

96. van Dijk AD, Boelens R, Bonvin AM. **Data-driven docking for the study of biomolecular complexes**. Febs J 2005; 272(2): 293-312.

97. van Dijk AD, de Vries SJ, Dominguez C, Chen H, Zhou HX, Bonvin AM. **Data-driven docking: HADDOCK's adventures in CAPRI**. Proteins 2005; 60(2): 232-238.

98. Wolfson HJ. **Geometric Hashing: An Overview**. IEEE Computational Science & Engineering 1997; 4(4): 10-21.

99. Yuan Z, Zhao J, Wang ZX. **Flexibility analysis of enzyme active sites by crystallographic temperature factors**. Protein Eng 2003; 16(2): 109-114.

100. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. **Determination of atomic desolvation energies from the structures of crystallized proteins**. J Mol Biol 1997; 267(3): 707-726.

101. Zimmermann O. **Untersuchungen zur Vorhersage der nativen Orientierung von Protein-Komplexen mit Fourier-Korrelationsmethoden**. Cologne: University of Cologne; 2002.

Danksagung

Ich möchte mich an dieser Stelle bei allen bedanken, die zum Gelingen dieser Arbeit beigetragen haben und mich während meiner Promotionszeit unterstützt haben.

Insbesondere danke ich Prof. Schomburg für das interessante und herausfordernde Projekt, das mir stets entgegengebrachte Vertrauen, dass er mir immer ein offenes Ohr schenkte und mir Freiheiten bei Planung und Durchführung des Projektes einräumte.

Oliver Martin gebührt Dank für die enge Zusammenarbeit während der gesamten Entstehungszeit dieser Arbeit und zahllose konstruktive Diskussionen.

Zusammenfassend möchte ich mich bei allen Korrekturlesern, Freunden und Kollegen des Cologne University Bioinformatics Center für die Unterstützung bei meiner Arbeit und für ein angenehmes Arbeitsklima bedanken.

Vielen Dank

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. Dietmar Schomburg betreut worden.

Philipp Heuser

Teilpublikationen

Paper

Heuser P, Schomburg D. **Optimised amino acid specific weighting factors for unbound protein docking.** BMC Bioinformatics 2006, 7:344.

Poster bei Konferenzen

Martin O, Heuser P, Schomburg D. **Protein-Protein Docking@CUBIC.** Bioperspectives05 (Wiesbaden)

Heuser P, Schomburg D. **Amino Acid Dependent Weighting Factors for FFT-based Unbound Protein Docking. ECCB05 (Madrid)**

Lebenslauf

Dipl. Biologe
Philipp Heinrich Heuser
Geb.: 16. März 1976 in Münster
Staatsangehörigkeit: deutsch

Mauenheimer Str.3
D-50733 Köln
Email: contact@philipp-heuser.de
Web: www.philipp-heuser.de

seit **2002**	**Wissenschaftlicher Mitarbeiter** Cologne University Bioinformatics Center (CUBIC) / Universität zu Köln (Arbeitsgruppe Prof. Schomburg) und **Promotion** zum Thema Protein-Protein Docking
2002	**Diplomarbeit** zur Vorhersage von Strukturveränderungen in Proteinen durch Insertionen und Deletionen (Looppredict ion) Arbeitsgruppe Prof. Schomburg – Universität zu Köln
1999/2000	**Erasmussemster** an der University of Manchester (UK)
1996-2001	Studium der **Biologie** an der Universität zu Köln (Schwerpunkte: Biochemie, Entwicklungsbiologie und Genetik)
1995/1996	**Zivildienst** bei Jugendhilfezentrum Michaelshoven (Köln)
1986-1995	**Abitur** am Bodelschwingh Gymnasium Herchen der ev. Kirche im Rheinland
1982-1986	Gemeinschaftsgrundschule Windeck-Obernau