Untersuchung eines wissensbasierten Potentials zur Bewertung von Protein-Protein-Docking-Studien

Inaugural - Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Vera Grimm
aus München

Köln 2002

Meinen Eltern

Als Julius Cäsar den Rubicon überschreitet,
weiß er nicht nur, daß er ein Sakrileg begeht;
er weiß auch, daß er, sobald er es einmal
begangen hat, nicht mehr zurückkann.

Umberto Eco

Über den Dank

Die Frage nach dem rechten Dank ist philosophischer Natur, denn wem sei an welcher Stelle zu welchem Zweck wie ausdrücklich gedankt. Auf der anderen Seite ist ein Dank zuviel besser als einer zuwenig. Das würde zu hemmungloser, seitenfüllender Dankerei führen. Reine Neugier läßt mich fragen wieviel Gedankenzeit bei Erstellen einer derartigen Arbeit im Mittel auf den Dank entfällt. Es bleibt zu hoffen, daß der Dank dabei nicht untergeht. In meiner Arbeit ist diese Seite die mit Abstand am häufigsten umgeschriebene und doch ist mir die rechte Gestaltung des Dankens nicht in den Sinn gekommen. Anstatt nun einen halbherzigen, einseitigen Dank an die Menschen, die mir aufgrund zeitlicher oder räumlicher Nähe am besten im Gedächtnis sind, auszusprechen, möchte ich mit der Tradition des klassischen Dankes brechen und hiermit einfach allen Personen von A bis Z, die an dieser Arbeit teilhatten oder die Fundamente dafür bauten oder für die Standhaftigkeit meiner Motivation sorgten oder mich bis hierhin freundschaftlich begleiteten auf herzlichste danken und es jedem dieser Menschen selber überlässen sich eine Reihenfolge der Dank-Empfangenden zu überlegen, allerdings mit der herzlichen Versicherung, daß ich lange über diesen Dank und die zu bedankenden Personen nachgesonnen habe. Vielen Dank !

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 der in A.8 angegebenen Teilpublikation – 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.

Vera Grimm

1. Referent: Prof. Dr. D. Schomburg
2. Referent: Priv.-Doz. Dr. T. Kraska
Eingereicht: Dezember 2002
Disputation: Februar 2003

Inhaltsverzeichnis

Begriffserklärungen und Abkürzungen

Gl.	Gleichung
vergl.	vergleiche
SAS	*solvent accessible surface*, Fläche, die vom Lösungsmittel zugänglich ist.
MSA	*molecular surface area* Einhüllende Fläche eines Proteins.
interface	Bindestelle zwischen zwei Makromolekülen.
buried surface	Während der Assoziation vergrabene Fläche in $Å^2$.
Ångstrøm	1 Å= 10^{-10}m.
ACE	*atomic contact energy*. Wissenbasiertes Potential basierend auf experminentellen Daten und strukturellen Informationen.
ASP	*atomic solvation parameter*. Potential zur Bestimmung der Desolvatationsenergie, basierend auf experimentellen Daten.
PCA	*principal component analysis*. Hauptkomponentenanalyse.
rmsd	*root mean square deviation*. Maß für die Ähnlichkeit zweier dreidimensionaler Strukturen.
FFT	Fast Fourier Transformation.
CAPRI	*Critical Assessment of Predicted Interactions*. Wettbewerb zum Testen von Docking-Programmen.
Docking	Simulation der Assoziation von Makromolekülen.
bound case	Docking von Proteinen, die aus einem Komplex ausgeschnitten wurden.
unbound case	Docking von Proteinen, die in ihrer nativen Form vorliegen.
backbone	Strukturgerüst eines Proteins, bestehend aus den C_α-, C-, N- und O-Atomen der Aminosäuren ohne die Seitenketten. Hauptkette.
decoy	Verdrehte oder verbogene Struktur.
pdb	Brookhaven Data Base (Berman *et al.* (2000)).
Strukturdatensatz	Eine Sammlung an Protein-Protein-Komplexen zur Erzeugung der Netto-Präferenzen.

Netto-Präferenz	Statistische Tendenz. Gibt die "Wahrscheinlichkeit" an, einen Atom-Atom-Kontakt in einem bestimmten Abstand zu finden.
Bindungs-Präferenz	Die Kombination der Netto-Präferenzen und der Häufigkeitsverteilung von Atom-Atom-Kontakten im zu bewertenden System (E_{gesamt}).
E^{nf}_{gesamt}	Bindungs-Präferenz **ohne** gesonderte Gewichtung von Atomen funktioneller Gruppen.
$E^{nf}_{spezifisch}$	Bindungs-Präfrerenz **ohne** die Berücksichtigung von Desolvatationseffekten, sowie ohne funktionelle Gruppen.

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

Abstract

Rigid body methods are very efficient for docking co-crystallized (bound) protein conformations using mainly geometric complementarity. When docking conformations crystallized individually (unbound) several thousands of false positive complexes with high scores can be generated. Therefore an atomic knowledge-based-potential has been developed which discriminates near-native conformations from non-native ones. The potential is based on an empirical energy-density-function and consists of two additive parts, namely a distance-dependent pair-potential of the interface atoms considering the specific atomic environment of forty different atom types (Melo & Feytmans (1997)), and a one-body-surface-potential to correct for missing entropic forces which are believed to be substantial for complex formation (Gohlke *et al.* (2000)). A trapeze function has been implemented to smooth the discrete distribution function. Hydrogen bonds and contacts between functional groups are taken into account as weighting factors. A repulsive part penalizes steric overlaps. The potential is derived from a database consisting mainly of the COMBASE (Vakser *et al.* (1999)), displaying the most complete collection of protein-protein-complexes applied in statistical approaches to date. In twelve of sixteen unbound docking simulations with the Fourier correlation program ckordo (Zimmermann (2002)) the potential ranked a near-native conformation within the top 1000. In eight of these cases the near-native was found with a rank even better than 200. The geometric ranking scored the best near-native conformation only in three of sixteen cases within the best 1000 and is therefor signifiant worse than the ranking with the presented potential.

Kurzzusammenfassung

Ein Hauptkriterium zur Bewertung von Strukturen beim Docken von Proteinen ist die geometrische Komplementarität. Vereinfachend werden die Strukturen als rigide Körper betrachtet. Für die Reproduktion der Proteinanordnung aus kokristallisierten Komplexen (*unbound*) ist diese Methode gut geignet, aber bei dem Versuch die Struktur aus den einzeln krisallisierten Proteinen zusammenbauen können viele tausend falsch-positive Lösungen erhalten werden. Ein wissens-

basiertes Kraftfeld auf atomarer Ebene soll eine Unterscheidung von nativen und nicht-nativen Anordnungen ermöglichen. Das Potential basiert auf einer empirischen Energiedichte-Funktion und besteht aus einem abstandsabhängigen Paarpotential der *interface*-Atome und einem Ein-Teilchen-Potential zur Repräsentation entropischer Kräfte (Gohlke *et al.* (2000)). Einem Vorschlag von Melo & Feytmans (1997) folgend werden vierzig Atomtypen verwendet. Eine Trapezfunktion berücksichtigt die Nachteile der Diskretisierung und glättet die Rohdaten. Wasserstoffbrücken und Kontakte zwischen Atomen funktioneller Gruppen werden als essentiell angesehen und mit einem eigenen Gewichtungsterm versehen. Ein repulsiver Term berücksichtigt sterische Überlappungen. Das Potential wird aus einem Strukturdatensatz erzeugt. Dieser besteht hauptsächlich aus der COMBASE (Vakser *et al.* (1999)), der umfangreichsten Datenbank an Protein-Protein-Strukturen, die bis dato in statistischen Ansätzen verwendet wurde. In sechzehn Docking Studien mit dem Fourier-Korrelations-Programm ckordo (Zimmermann (2002)) werden anhand des hier vorgestellten Potentials in zwölf der sechzehn Fälle nativ-ähnliche Anordnungen mit einem Rang kleiner 1000 gefunden. In acht von diesen Fällen wird der beste richtig-positive Komplex sogar unter den ersten 200 gefunden. Eine Bewertung anhand der geometrischen Korrelation fand eine nativ-ähnliche Strukur nur in drei von sechzehn Fällen unter den besten 1000 und damit deutlich schlechter, als die Bewertung mit dem hier vorgestellten Potential.

1 Einleitung

"Better understanding of the forces
that govern molecular recognition must
come from a realization that the process
takes place **under water**."

Ringe (1995)

1.1 Molekulare Erkennungsprozesse

Katalyse, Regulation, und Informationsübermittlung sind einige fundamentale Prozesse in lebenden Zellen. Makromoleküle wie Proteine, die sich in nicht-kovalenter Weise mit anderen Molekülen zu Komplexen verbinden können, spielen in allen diesen Vorgängen eine wichtige Rolle. Proteine sind aus niedermolekularen Einheiten, den Aminosäuren, zusammengesetzt und bilden kettenartige Polymere aus, die sich in einzigartige dreidimensionale Strukturen falten. Ein Beispiel sind die Enzyme, Proteine, die mit Substraten oder anderen Molekülen, wie Inhibitoren nicht-kovalente-Bindungen eingehen können. Enzyme katalysieren chemische Reaktionen und tragen entscheidend zur Regulation metabolischer Pfade bei. Weitere Beispiele für nicht-kovalente Verbindungen von Proteinen und anderen Molekülen sind Protein-DNA- oder Protein-RNA-Komplexe, die für die Transkription, Regulation und Reparatur der DNA wichtig sind. Weiterhin sind Antikörper-Antigen- ebenso wie MHC-Rezeptor-Komplexe als Beispiele anzuführen. Beide Komplexe sind integraler Bestandteil einer funktionstüchtigen Immunabwehr. Auch die Informationsweiterleitung kann auf Protein-Protein-Komplexen basieren, z.B. das Hormonsystem von Menschen.

Anhand dieser wenigen Beispiele wird bereits deutlich, daß die Kenntnis assoziativer Prozesse auf molekularer Ebene für das Verständnis biologischer Sachverhalte unumgänglich ist. Was sind die treibenden Kräfte für Assoziation, wie wird die "richtige" Bindungsstelle erkannt, wie spezifisch ist eine Bindung

und welche Funktion hat ein Komplex sind nur einige der entstehenden Fragen auf dem Weg zum vollständigen Verständnis der komplexen Interaktionsnetzwerke unter Beteiligung von Proteinen. Die vollständige Aufklärung des "Interaktoms", der Summe aller Protein-Interaktionen in einer Zelle nimmt in der postgenomen Ära an Bedeutung zu, da mit Aufklärung des Genoms sehr viele neue Informationen gewonnen werden können.

Die experimentelle Untersuchung von intermolekularen, nicht-kovalenten Wechselwirkungen in Protein-Protein-Komplexen ist oft schwierig, zeitintensiv oder gar nicht möglich. Mit Hybrid-Methoden (Tong *et al.* (2002)), dem Yeast-two-hybrid-Verfahren (Fields & Song (1989)), Protein-Chips (Zhu *et al.* (2001)), und neuerdings der Massenspektroskopie (Miranker (2000), Gavin *et al.* (2002)) lassen sich Interaktionen zwischen Proteinen feststellen. Kalorimetrische Studien (Pierce *et al.* (1999)) und Oberflächen-Plasmon-Resonanz-Spektroskopie (Fischer (2000)) können Informationen über Geschwindigkeitskonstanten und Bindungsstärken liefern. Zur Aufklärung der dreidimensionalen Struktur stehen die Röntgenstrukturanalyse und NMR-Techniken zur Verfügung. Für viele Proteine funktionieren sie hervorragend, bei Protein-Protein-Komplexen sind sie aber nur sehr bedingt anwendbar, da die Kristallisation von Protein-Protein-Strukturen nicht ohne weiteres möglich ist. Derzeit sind nur rund 100 Komplexstrukturen (Gardiner *et al.* (2001)) bekannt.

Eine *in silico* Methode den Bindungsmodus zweier Komponenten zu simulieren wird *Docking* genannt. Es wird zwischen dem Docking von Teilchen ähnlicher Größe (Protein-Protein) und dem Docking von Partikeln unterschiedlicher Größen (Protein-Ligand) unterschieden. Allgemein stellt das Docking eine Alternative und Ergänzung zur experimentellen Untersuchung dar. Für Komplexe von Proteinen mit kleinen Molekülen (Protein-Ligand) wird das Verfahren in der pharmazeutischen Industrie seit langem erfolgreich angewendet (*drug design*, *modelling*). Im Protein-Protein-Docking sind nur mäßige Erfolge zu erkennen. Das Hauptproblem hierbei ist eine probate Bewertungsfunktion, die eine native von einer nicht-nativen Anordnung unterscheiden kann.

1.1.1 Komplementarität

Assoziative Vorgänge in den Naturwissenschaften beruhen nach heutigem Kenntnisstand auf Komplementarität. Dieser Begriff wurde von Niels Bohr (1885-1965) im Zusammenhang mit der Quantenmechanik geprägt. Demnach heißt Komplementarität "die Zusammengehörigkeit verschiedener Möglichkeiten, dasselbe Objekt als verschiedenes zu erkennen. Komplementäre Erkenntnisse gehören zusammen, insofern sie Erkenntnisse desselben Objekts sind, sie schließen einander jedoch insofern aus, als sie nicht zugleich und für denselben Zeitpunkt erfolgen können" (Leibzig (2001)).

Der Begriff der Komplementarität wird seitdem in vielen verschiedenen Fachrichtungen verwendet. Im Docking wird z.B. von geometrischen Komplementäritäten gesprochen. Zwei Moleküle, die gegensätzliche und doch ineinander passende geometrische Form haben, sind zueinander komplementär wie das Ei zum Eierbecher. Auch andere Eigenschaften, wie Ladungsverteilungen oder Hydrophobizität können zueinander komplementär sein. Eine zentrale Frage im Docking ist die Vorhersagbarkeit von Komplementaritäten und damit möglicherweise von Bindestellen.

1.1.1.1 Geometrische Komplementarität

Die Substrat-Bindestelle eines Enzyms liegt nach der Schlüssel-Schloss-Hypothese (engl. *lock and key hypothesis*) von Emil Fischer (1894) vorgeformt vor. Die Spezifität des Enzyms (Schloß) für ein Substrat (Schlüssel) ergibt sich aus ihren geometrisch komplementären Formen. Diese Annahme ist noch heute die wichtigste Basis für Docking-Studien.

Die Weiterentwicklung des Modells stellt die *induced fit*-Hypothese dar, der zufolge die Substratbindestelle erst bei Bindung des Substrates ausgebildet wird. Die Oberflächengeometrie ist *a priori* nicht determiniert. Bei vielen anderen Bindungsprozessen wie dem Lehrbuchbeispiel Hexokinase mit Glukose finden subtile Konformationsänderungen an den Kontaktoberflächen statt, die manchmal sogar die Gesamtkonformation eines oder beider Bindungspartner nachhaltig

verändern. Solche Flexibilitäten können bis heute nur mit größtem Zeitaufwand, wenn überhaupt, berechnet werden.
Ein Maß für die Komplementarität ist nach Jones & Thornton (1996) der *gap index* oder nach Lawrence & Colman (1993) der *shape complementarity coefficient*. Enzym-Inhibitor- und permanente Komplexe weisen die höchste Komplementarität auf. Antikörper-Antigen- und nicht permanente Komplexe zeigen die geringste geometrische Komplementarität (Jones & Thornton (1996), Lawrence & Colman (1993)). Diese Tatsache liegt darin begründet, daß permanente Komplexe im Laufe der Evolution hoch optimierte Kontaktfläche mit guter Passform entwickelt haben. Temporäre Komplexe hingegen sind nur zeitlich begrenzt miteinander verbunden und müssen oft eine Vielzahl von Liganden binden können, wie eben z.B. Antikörper. Diese Diversität und zeitliche Knappheit stehen einer "guten Passform" im Wege. Die häufig trotzdem hohe Spezifität wird z.B. im Falle der Antikörper über intensive Kontakte der Seitenketten mit dem Antigen bewerkstelligt. Derartige Interaktionen sind mit theoretischen Methoden schwieriger zu modellieren, da die Stellung von Seitenketten nicht genau bekannt ist. Trotz dieser Einschränkungen ist die geometrische Komplementarität mit Abstand das wichtige Kriterium bei Docking.

1.1.1.2 Elektrostatische Komplementarität

Ein weiteres wichtiges Maß für die Unterscheidung nativer von nicht nativen Komplex-Anordnungen im Docking ist die elektrostatische Komplementarität. Die Ladungen auf der Oberfläche von Proteinen tragen zu einem charakteristischen elektrostatischem Feld bei. Die Felder zweier Proteine sind in vielen Fällen komplementär zueinander (Gabb *et al.* (1997), Sheinerman & Honig (2002)). Das bedeutet aber nicht zwangsläufig, daß auch die Ladungsverteilungen auf den Oberflächen komplementär sind. Untersuchungen von McCoy *et al.* (1997) und Xu *et al.* (1997) zeigen, daß es Beispiele sowohl für die gleichzeitige Komplementarität der elektrostatischen Felder und der Ladungsverteilungen, aber auch für den Fall der Komplementarität der elektrostatischen Felder und keiner Komplementarität der Ladungsverteilungen gibt.

Des weiteren bedingt die elektrostatische Komplementarität keine Unterteilung der Komplexe in verschiedene Gruppen, wie die geometrische Komplementarität. Sie ist zum Beispiel in Antikörper-Antigen-Komplexen genauso groß wie in Serin-Proteasen (McCoy *et al.* (1997)), während die geometrische Komplementarität in den beiden Fällen signifikant unterschiedlich ist. Weiterhin sind elektrostatische Wechselwirkungen nicht uniform über das gesamte *interface* verteilt, sondern nur auf einer Untermenge der Gesamtfläche lokalisiert (Novotny *et al.* (1989)), wohingegen das für die geometrische Komplementarität nicht allgemein der Fall ist.

1.1.1.3 Wasserstoffbrücken

Ein weiteres, oft benutztes Kriterium für die Bewertung gedockter Strukturen ist die Komplementarität von Wasserstoffbrücken und Netzwerken von Wasserstoffbrücken. Diese komplementäre Strukturen (Janin & Chothia (1990)) können ausgebildet werden, müssen es jedoch nicht. Ein Donor (oder Akzeptor) kann zu einem hohen energetischen Preis im Inneren der Kontaktfläche vergraben sein. 17,4 % der vollständig vergrabenen Donoren und Akzeptoren aus hochaufgelösten Kristallstrukturen bilden keine Wasserstoffbrücken aus (Xu *et al.* (1997)). Die Vorhersage von Wasserstoffbrücken-Bindungen zur Identifizierung der Bindestelle wurde von Meyer *et al.* (1996) und Krämer (2001) als möglicher Vorfilter in Docking-Simulationen untersucht.

1.1.2 Spezifität

Molekulare Wechselwirkungen beruhen auf zufälliger Kollision, wie in Kapitel 1.3 erläutert wird. Wegen dieser Zufallskomponente kann die Erkennung niemals perfekt sein. Selbst Reaktionen, die energetisch ungünstig sind, können manchmal vorkommen. Ebenso kann die Spezifität für ein Substrat niemals perfekt sein. Serin-Proteasen spalten proteolytisch eine Vielzahl von Peptidbindungen, sind demnach spezifisch in Bezug auf Peptidbindungen, aber sehr unspezifisch in Bezug auf die Art des Peptids. Antikörper hingegen bilden nur mit einem bestimmten Antigen einen Komplex. Komplexe haben demnach ein unterschiedlich spezifisches Bindungsverhalten ihrer Funktion entsprechend.

1.2 Molekulare Kontaktflächen

Molekulare Oberflächen mit komplementären Eigenschaften von Proteinen sind verantwortlich für die selektive Bindung von anderen Proteinen oder kleineren Liganden. Sie gelten als Ursache spezifischer Interaktionen und sind eine genauere Betrachtung wert. Die Kontaktstelle (engl. *interface*) wird für Enzym-Substrat-Komplexe auch aktives Zentrum oder allgemeiner Bindungsstelle (engl. *binding-site*) genannt. Daraus ergibt sich die vergrabene Fläche (engl. *buried surface area*) als der Bereich, der während der Komplexierung zweier Proteine vollständig verdeckt wird. Die Bindestelle zwischen zwei Makromolekülen kann auf zweierlei Arten definiert werden.Einerseits als aus den Aminosäuren bestehend, die auf unterschiedlichen Proteinketten innerhalb eines maximalen Abstandes (meist 6 Å) zu finden sind (Tsai *et al.* (1996)). Andererseits als die Summe der lösungsmittelzugänglichen Fläche (*solvent accessible surface*, SAS) der isolierten Komponenten abzüglich der SAS im ko-kristallisierten Komplex (Conte *et al.* (1999)). Die Anzahl der *interface*-Atome skaliert linear mit der *interface*-Fläche. Im Mittel befindet sich ein Atom auf 9.3 Å ^{2} Fläche. Die Kontaktfläche wird in Inneres und Peripherie unterteilt. Atome mit direktem Kontakt zu Atomen des anderen Proteins nehmen 3/4 der Fläche ein, die restlichen 1/4 verteilen sich auf die Peripherie (Conte *et al.* (1999)).

Die Vorhersage der exakten Lage von Bindestellen ohne biologische Informationen ist trotz zahlloser Untersuchungen bis dato nicht möglich. Allgemein werden Protein-Komplexe anhand ihres Interaktionspartners in verschieden Gruppen unterteilt, wie die Protein-Protein-, Protein-Ligand- und Protein-DNA-Komplexe. Im Folgenden werden ausschließlich Protein-Protein-Komplexe beschrieben, da nur sie Gegenstand dieser Arbeit sind.

1.2.1 Protein-Protein-Komplexe

Es werden zwei Arten von Protein-Protein-Komplexen definiert, solche deren Proteine sich als eigenständige Einheiten falten und jene, die einen gemeinsamen Faltungsprozess durchlaufen. Erstere sind nur temporär miteinander verbunden

und die Monomere in Lösung stabil. Zu ihnen gehören viele der gut untersuchten Protease-Inhibitor-, ebenso die Antikörper-Antigen-Komplexe und andere Heterooligomere. Komplexe, deren Einheiten einen gemeinsamen Faltungsprozess durchlaufen, heißen permanent oder obligatorisch und umfassen großteils Homooligomere. Die getrennten Ketten sind in Lösung selten stabil und haben oft keine Funktionalität. Viele Multienzym-Komplexe und Strukturproteine wie Collagen und Keratin gehören dieser Kategorie an, ebenso wie Komplexe, die bei der Proteinbiosynthese aus einer durchgängigen Kette bestanden und posttranslational verändert wurden.

Die Kontaktflächen der beiden Arten von Komplexen unterscheiden sich stark in ihrer chemischer Zusammensetzung, Größe und Form. Der in dieser Arbeit benutzte Dockingdatensatz besteht ausschließlich aus temporären Komplexen, wie Proteasen, Hydrolasen, Glykolasen und Antikörpern, wohingegen der Strukturdatensatz auch eine Vielzahl von permanenten Komplexen, wie Homooligomere und Membran-Proteinen enthält.

1.2.1.1 Temporäre Komplexe

Die chemischen Eigenschaften von temporären Komplexen variieren stark. Einige Komplexe zeigen im *interface* prozentual mehr hydrophobe Aminosäuren als auf der restlichen Oberfläche. Die beiden Protease-Inhibitor-Komplexe mit dem pdb-Kürzel 1ppf und 1abc haben z.B. ein zu etwa 70% nicht-polares *interface* bei einer mittleren Fläche von ca. 1100 $Å^2$. Andere Komplexe haben ähnliche Aminosäure-Zusammensetzungen auf der gesamten Oberfläche (im Mittel 56% nicht-polar). Eine simpler Zusammenhang ist nicht gegeben und die Analyse daher schwierig (Conte *et al.* (1999)).
Protease-Inhibitor-Komplexe haben im Mittel eine 1700 $Å^2$ große Substratbindestelle, der Thrombin-Hirudin-Komplex der Blut-Gerinnungs-Kaskade weist jedoch mehr als 3000 $Å^2$ auf. Für andere Enzym-Komplexe, als den Protease-Inhibitor-Systemen, werden durchschnittlich Flächen zwischen 1100 bis 3000 $Å^2$ gefunden (Chakrabarti & Janin (2002)). Antigen-Antikörper-Komplexe haben Flächen von 1200-2000 $Å^2$. Conte *et al.* (1999) zeigten, daß ein Standard-*interface*

mit einer Fläche von ca. 1600 $Å^2$ ausreichend zur Stabilität und Spezifität eines beliebigen Komplexes beiträgt. Eine Vergrößerung der Fläche bringt keinen signifikanten Stabilitätsgewinn. Die *interfaces* sind meist stark gekrümmt und weisen Spalten und Furchen auf. In wenigen Fällen werden planare Flächen gefunden. Die Existenz und Relevanz von *patches*, zusammenhängenden Oberflächenstücken ähnlicher Eigenschaften (Jones & Thornton (1997)) wird kontrovers diskutiert. Sie könnten der Vorhersage von Bindestellen dienen und das Docking erleichtern. Chakrabarti & Janin (2002) fanden *recognition patches*, Flächenstücke mit mittleren Größen von 1560 ± 340 $Å^2$ in allen von ihnen untersuchten Komplexen. Sie dienen der spezifischen Erkennung.

1.2.1.2 Permanente Komplexe

Proteine mit Bindestellen für ihre eigene Oberfläche können sich zu Dimeren oder Oligomeren, zu geschlossenen Ringen, Kugelschalen oder helikalen Polymeren zusammenlagern. Sie sind selbstkomplementär, d.h. die Kontaktfläche ist flach oder konvex an der einen und konkav an der anderen Seite. In permanenten Komplexen ist die chemische Zusammensetzung der Kontaktfläche verglichen mit gesamten Oberfläche sehr ähnlich. Der vergrabene Bereich ist anders als bei temporären Komplexen deutlich weniger polar. Im Mittel sind 20% polare, 67% nicht-polare und 13% geladene Aminosäuren im *interface* zu finden (Miller (1989)). Dieses Verhältnis kann aber zwischen verschiedenen Oligomeren stark variieren (Janin *et al.* (1988), Miller *et al.* (1987a), Miller *et al.* (1987b)). Homooligomer-*interfaces* haben weniger herausragende Schleifen, tiefe Taschen und ihnen wird häufig ein hydrophober Kern aufgrund des größeren Anteils hydrophober Aminosäuren unterstellt. Larsen *et al.* (1998) zeigte, daß nur eine Minderheit der von ihm untersuchten Oligomere einen derartigen Kern haben. Die meisten werden durch eine Kombination von kleinen hydrophoben Stückchen (engl. *patches*), polaren Interaktionen und Wasser-Molekülen stabilisiert. Bei der Oligomerisierung spielen hydrophobe *patches* eine wichtigere Rolle (Lijnzaad *et al.* (1996)) als bei temporären Komplexen.

1.2.2 Kristallkontakte

Kristallkontakte sind Artefakte des Kristallisationsprozeßes, die in der Zelle nicht vorkommen. Sie sind deutlich kleiner als Oligomer- und Protein-Protein-*interfaces* und zeigen eine höhere Neigung zur Paarung polarer Residuen im Gegensatz zu Oligomeren, die eine Tendenz zur Kombination hydrophober Aminosäuren, wie Met-Met, Met-Leu, Leu-Leu und Leu-Ile haben. Polare Interaktionen zwischen Lys-Glu und Asp-Arg sind gleichermaßen in Oligomer-*interfaces*, wie unter Kristallkontakten zu finden. Kristallkontakte gehören zu den instabilen Kontaktflächen, genau wie kleine *interface*-Flächen von Multimeren, die nur einen kleinen Beitrag zur Gesamtstabilität beitragen. Bisher ist es nicht möglich instabile Interaktionsflächen von stabilen, wie Oligomer-*interfaces* zu unterscheiden. Aus diesem Grunde werden Kristallkontakte aus dem Strukturdatensatz ausgeschlossen.

Es sei angemerkt, daß trotz der Unterschiedlichkeit von Oligomer-Bindestellen und Kristallkontakten, nach Dasgupta *et al.* (1997) ähnliche Paarpotentiale für beide erhalten werden.

1.3 Assoziation von Proteinen in Lösung

Allgemein wird der hydrophobe Effekt als die treibende Kraft der Assoziation und Faltung verstanden (Honig & Nicholls (1995)). Weitreichende elektrostatische Kräfte spielen allerdings in einigen Fällen ebenfalls eine bedeutsame Rolle. Direkte, lokale Wechselwirkungen sind im Übergangszustand nicht zu finden, da die Komplexpartner zu weit voneinander entfernt sind.
Für Protein-Protein-Komplexe **ohne** weitreichende elektrostatische Interaktionen wie Elastase/OMTKY (1ppf) und α-Chymotrypsin/OMTKY (1cho) ist die Desolvatation der beiden Proteine die treibende Kraft für die Reaktion zum Komplex *AB* (Camacho *et al.* (1999)). Die Desolvatation ist um Größenordnungen schneller als die Diffusion. Derartige Komplexe zeigen ein hohes Maß an Oberflächenkomplementarität und sind mit entsprechenden *in silico* Methoden relativ gut simulierbar (Camacho & Vajda (2002)). Komplexe **mit** weitreichenden elektrosta-

tischen Interaktionen zeigen deutlich beschleunigte Reaktionsgeschwindigkeiten. Für den Barnase/Barstar-Komplex zeigten Gabdoulline & Wade (1997), daß alleine die elektrostatischen Kräfte eine sehr hohe Geschwindigkeitskonstante von $10^9 M^{-1}s^{-1}$ herbeiführen (vergl. Kapitel 1.3.1). Derartige Komplexe weisen oft lange Arginin oder Lysin-Seitenketten auf, deren Anordnung im ungebundenen Zustand deutlich von dem im gebundenen abweicht. Einige dieser Aminosäuren nehmen erst im Komplex gemäß der *induced-fit*-Hypothese ihre native Konformation ein. Die Anordnung ist *a priori* nicht bestimmbar (Kimura *et al.* (2001)). Die *in silico*-Modellierung solcher Systeme ist schwierig. Sollen Eigenschaften wie Stabilität und Spezifität untersucht werden, müssen thermodynamische Größen ermittelt werden.

1.3.1 Kinetik

Die Geschwindigkeitskonstante von Protein-Protein-Assoziationen liegt für langsame Reaktionen im Bereich von $10^5 - 10^6 M^{-1}s^{-1}$ für schnelle jedoch bei etwa $10^9 M^{-1}s^{-1}$. Für die bimolekulare Reaktion zweier Makromoleküle A und B in wässriger Lösung mit dem Übergangszustand $A:B$ gilt:

Der erste Schritt der Assoziation besteht aus der diffussionskontrollierten, zufälligen Kollision von A und B. Das sich bildende Konglomerat kann nur dann als zweiten Schritt einen stabilen Komplex formen, wenn die beiden Reaktionspartner korrekt zueinander angeordnet sind. Ist dies nicht der Fall, zerfällt der instabile Übergangskomplex (Janin (1997)). Die zufällige Kollision der beiden Teilchen wird von der Brownschen Molekularbewegung diktiert. Die Geschwindigkeitsrate für diese Reaktion wird von der Smolochowski-Einstein-Gleichung beschrieben und erreicht Größen von maximal $10^9 M^{-1}s^{-1}$ (Camacho *et al.* (2000)).

1.3.2 Thermodynamik

Das chemische Potential μ_{AB} für Reaktion 1.3.1 im thermodynamischen Gleichgewicht ergibt sich zu:

$$\mu_A + \mu_B = \mu_{AB} \tag{1.1}$$

Die Standard molare Gibbsche freie Energie G ist demnach:

$$\Delta G = \mu_{AB} - (\mu_A + \mu_B) = -RT \ln K \tag{1.2}$$

mit K als Assoziationskonstante. Zwei Beispiele für die Größenordnung der dimensionslosen Konstante K sind Biotin mit Streptavidin ($K = 10^{14}$) und N-C-Acetyltryptophanamid mit α-Chymotrypsin ($K = 5 * 10^3$) (Miyamoto & Kollman (1993)). Der Zusammenhang mit der Enthalpie H und der Entropie S bei einer definierten Temperatur T ist gegeben durch:

$$\Delta G(T) = \Delta H(T) - T\Delta S(T) \tag{1.3}$$

Die freie Energie $\Delta G_B(X)$ für die Bindung eine Makromoleküls B an die spezifische Stelle X eines anderen Makromoleküls hat nach Ben-Naim *et al.* (1990) den niedrigsten Wert, verglichen mit allen anderen Bindungsstellen:

$$\Delta G_B(X) = \min_i \Delta G_B(i) \tag{1.4}$$

$\Delta G_B(A)$ wird auch als effektive Energie oder engl. *potential of mean force* bezeichnet (Lazaridis & Karplus (2000)). $\Delta G_B(i)$ setzt sich zusammen aus direkter intramolekularer Wechselwirkungsenergie im Vakuum $\Delta U_B(i)$ und einem indirekten, solvensabhängigen Anteil $\delta G(i)$:

$$\Delta G_B(i) = \Delta U_B(i) + \delta G(i) \tag{1.5}$$

Der lösungsmittelabhängige Beitrag stellt eine Mittelung über alle Konfigurationen aller Solvens-Moleküle dar. Die exakte Beschreibung aller Lösungsmittelmoleküle für jede Konformation ist derzeit nicht möglich. Die Aufstellung einer expliziten Energiefunktion für Proteine oder Protein-Protein-Komplexe in Lösung ist ebenso wenig möglich (Lazaridis & Karplus (2000)). Da die Desolvatation eine wichtige treibende Kraft der makromolekularen Assoziation ist (Kauzmann *et al.* (1974),Horton & Lewis (1992), Jayaram *et al.* (2002)), sind zahlreiche implizite Solvens-Modelle erstellt worden. Es werden im wesentlichen drei Ansätze unterschieden:

- Kontinuumsmodelle
- empirische Modelle
- wissenbasierte Modelle.

Einen direkten Weg den elektrostatischen Beitrag zur Desolvatationsenergie zu berechnen, liefern die Kontinuumsmodelle, in denen weitreichende und lokale Wechselwirkungen über die Poisson-Boltzmann-Gleichung berechnet werden (Gilson & Honig (1988), Honig & Nicholls (1995), Dennis & C. J. Camacho (2000), Holst *et al.* (1994)).
Das Problem der Kontinuumsmodelle ist die Sensitivität bei der Positionierung der Dielektrika sowie kleine strukturelle Störungen (Jackson & Sternberg (1995), Vorobjev *et al.* (1998)). Die Berechnung des elektrostatischen Feldes für jede mögliche Konfiguration ist zeitintensiv. Empirische Solvatationsmodelle nähern daher die Solvatationsenergie mit der Änderung der *solvent accessible surface*

(Eisenberg & McLachlan (1986), Ooi *et al.* (1987), Vajda *et al.* (1995), Eisenhaber & Argos (1996)), dem Volumen der Hydratationsschale (Kang *et al.* (1987), von Freyberg *et al.* (1993), Eisenhaber (1996)), Vila *et al.* (1991), Augspurger & Scheraga (1995)) oder dem Gruppenvolumen (Lazaridis & Karplus (1999)) an.

Die Kombination der Änderung der Oberflächenzugänglichkeit und den meist experimentell bestimmten Transferenergien kleiner organischer Moleküle von Oktanol zu Wasser (Wolfenden *et al.* (1981)) und später von Gas zu Wasser (Wesson & Eisenberg (1992)) geht auf Eisenberg & McLachlan (1986) zurück. Ein Faktor σ verknüpft die beiden Größen und wird atomarer Solvatationsparameter (engl. *atomic solvation paramater*, ASP) genannt. Ein Vergleich von acht unterschiedlichen atomaren Solvatationsparamter-Sets (Juffer *et al.* (1995)) zeigt Inkonsistenzen in den Sets auf. Die Ergebnisse variieren in den Absolutwerten und deren Vorzeichen, weswegen teilweise gegenteilige Aussagen gemacht werden. Bei einem Vergleich eines ASP-Sets mit einem strukturellen Solvensmodell für die Diskriminierung nativer von nicht nativen Faltungen stellten Gatchell *et al.* (2000) jedoch die Überlegenheit der ASP-Methode fest. ASP-Sets werden erfolgreich für das Protein-Protein-Docking verwendet (Cummings *et al.* (1995), Horton & Lewis (1992)).

Das ACE-Modell (*atomic contact energies*, ACE, (Zhang *et al.* (1997))) ist ein strukturbasiertes Kontaktpotential. Ein Atom-Wasser-Kontakt wird durch ein Atom-Atom-Kontakt ersetzt um den Faltungs- oder Assoziationsprozeß zu simulieren. Es findet zahlreiche Einsatzgebiete im Konformationsscreening (Vasmatzis *et al.* (1996)), im Protein-Protein-Docking (Chen & Weng (2002)) sowie in der Brownschen Dynamik (Camacho *et al.* (2000)).

Nachteil aller vorgestellter Modelle ist die Tatsache, daß die Änderung der ASA nicht direkt proportional zur Desolvatationsenergie ist, sondern vielmehr die Desolvatation**entropie** wiedergibt (Vajda *et al.* (1994a)). Der Zusammenhang zwischen der ASA und der Desolvatationsenergie ist im Einklang mit dem Klathrat-Modell (Kauzmann *et al.* (1974)). Das Lösen eines Moleküls in Wasser reduziert die Freiheitsgrade der Translation und Rotation der nahe gelegenen

Wassermoleküle und damit die Entropie. Dieser Effekt hängt nicht von geladenen Gruppen ab, sondern lediglich von der ASA. Elektrostatik und Desolvatationsenergie hängen primär von Ladungsträgern an der Oberfläche ab. Im einfachsten Fall ist die Zahl geladener Aminosäuren, die während der Komplexierung vergraben werden proportional zur Desolvatationsenergie. Bei perfekter Ladungskomplementarität wäre auch die elektrostatische Energie proportional dazu.

Aus diesem Grund wird in vorliegender Arbeit ein aus der statistischen Thermodynamik stammender Ansatz verwendet. Er beschreibt die Desolvatationsenergie anhand eines wissensbasierten Ein-Teilchen-Potentials (Jones *et al.* (1992), Jones (1994), Kocher *et al.* (1994), Miyazawa & Jerningan (1985)). Es wird angenommen, daß dieses Potential alle oben genannten Energiebeiträge implizit enthält, so daß eine schwierige explizite Beschreibung einzelner Kräfte entfällt. Das Verfahren wird im Detail in Kapitel 2.1.7 dargestellt.

1.3.3 Stabilität von Komplexen in Lösung

Die Stabilität eines Komplexes in Lösung wird von der Dissoziationskonstante beschrieben. Je stärker die Bindung der Moleküle im Komplex, desto geringer die Geschwindigkeit ihrer Dissoziation (vergl. Kapitel 1.3.1). Im einen Extrem ist die Energie der ausgebildeten Bindung gegenüber thermischen Bewegungen zu vernachlässigen, und die Moleküle dissozieren ebenso schnell, wie sie zusammengekommen sind. Im anderen Extrem haben die Bindungen eine derart hohe Energie, daß Dissoziation fast nicht vorkommt. Die Bindungsstärke zwischen zwei Molekülen ist ein Maß für die Spezifität des Erkennungsvorganges. Unter physiologischen Bedingungen sind die Mehrzahl nativer globulärer Proteine in der Nähe des globalen Minimums der freien Energie stabil (Anfinsen (1973)). Ob diese Annahme auch für Komplexe gültig ist, gilt als umstritten. Die einzige derartige Aussage stammt von Ben-Naim *et al.* (1990) und wird als Grundlage dieser Arbeit verwendet (vergl. Kapitel 1.3.2). Eine entsprechend kritische Betrachtung erfolgt in der Diskussion.

1.4 Das Docking Problem

Docking bedeutet die "korrekte" Orientierung zweier Makromoleküle zueinander mittels theoretischer Methoden zu finden. Die "richtige" Anordnung ist die, die im nativen, ko-kristallisierten Komplex beobachtet wird. Es werden das Docking von Proteinen mit kleinen Molekülen (Protein-Ligand), mit ähnlich dimensionierten (Protein-Protein) und mit der DNA (Protein-DNA) unterschieden. Die Docking-Verfahren in den drei Gebieten sind angepasst an die physiko-chemischen Eigenschaften der Komplexe und daher unterschiedlich. Komplementaritäten (geometrische, elektrostatische, Wasserstoffbrücken) sind die Hauptauswahlkriterien für die Unterscheidung nativer von nicht-nativen Anordnungen.

Alle Docking-Ansätze basieren auf einer probaten Oberflächenrepräsentation, einer Suchstrategie und einer effizienten Bewertungsfunktion. Oberflächen können anhand zahlreicher mathematischer Modelle wie z.B. geometrischen Deskriptoren, *spherical harmonics* (Ritchie & Kemp (2000)), *sparse crititcal points* (Lin *et al.* (1994)), Graphen- (Gardiner *et al.* (2001)) oder Gitter-basiert (Katchalski-Katzir *et al.* (1992), Meyer *et al.* (1996)) beschrieben werden (Halperin *et al.* (2002)). 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 exponentiell mit der Größe steigt. Eine Beschleunigung kann anhand voriger Kenntnis der Bindungsstelle erfolgen, da dann der Suchraum erheblich eingeschränkt wird. Für unbekannte Proteine ist die Bindestelle jedoch nicht verfügbar.

Soll ein Protein gegen eine ganze Datenbank von Proteinen gedockt werden, wird von 1:n-Docking gesprochen. Im Vergleich zum "normalen" Docking zweier Proteine, ist der 1:n-Ansatz wesentlich zeitkritischer. Die zugrunde liegenden Algorithmen müssen besser besser werden und sehr zeitaufwändige Berechnungen sind nicht möglich. Das 1:n-Docking ist dann interessant, wenn *in silico* nach noch unbekannten Interaktionspartnern eines Proteins gesucht wird. Das

" normale" Docking versucht lediglich den Bindungsmodus zweier Makromoleküle vorherzusagen. Problematisch ist auch die Modellierung von Flexibilitäten. Seitenketten-, aber auch Domänen- und *loop*-Bewegungen müssen berücksichtigt werden. Für einen aktuellen Überblick über die Problematik sei Halperin *et al.* (2002) empfohlen.

1.4.1 Protein-Protein-Docking

Im Protein-Protein-Docking werden beide Komponenten als rigide Körper genähert, Flexibilitäten von Haupt- und Seitenketten werden meist vernachlässigt (engl. *rigid body docking*). In einigen Komplexen sind kaum Verschiebungen der Haupt- und Seitenketten von ungebundener (engl. *unbound*) zu gebundener (engl. *bound*) Struktur zu erkennen. In diesen Fällen ist die Annahme starrer Körper gerechtfertigt. In vielen anderen Proteinen kommt es jedoch zu starken Veränderungen der Hauptkettengeometrie und die Vorhersage des Bindungsmodus wird damit erheblich schwieriger. Flexibilitäten der Seitenketten sind ebensowenig zufriedenstellend modellierbar, obwohl hierfür eine Vielzahl von Methoden verfügbar ist (vergl. 1.4.1.2).
Es wird zwischen *bound*- und *unbound*-Docking unterschieden. Im *bound*-Docking wird versucht den Komplex aus beiden Proteinen, wie sie im ko-kristallisierten Komplex vorkommen, zu rekonstruieren. Hierzu werden die beiden *bound*-Strukturen direkt aus dem Komplex genommen. Im *unbound*-Docking hingegen wird versucht den Komplex aus den beiden ungebundenen Strukturen zusammenzusetzen.

1.4.1.1 Verschiedene Ansätze

Ein weithin benutztes *rigid body*-Verfahren ist der DOCK-Algorithmus. Das aktive Zentrum eines Proteins wird mit überlappenden Kugeln gefüllt. Die Zentren der Kugeln eines Clusters wird mit ähnlichen Clustern des Liganden verglichen (Shoichet & Kuntz (1991), Meng *et al.* (1991)). Eine ganze Reihe von Programmen basiert auf der Fourier-Korrelations-Technik (Katchalski-Katzir *et al.* (1992), Meyer *et al.* (1996), FTDOCK von Gabb *et al.* (1997), Vakser & Aflalo (1994), Vakser (1996a), Mandell *et al.* (2001), ZDock von Chen & Weng (2002)) und der

Weiterentwicklung zu den *sperical harmonics* (hex von Ritchie & Kemp (2000)). Im Fourier-Docking werden die beiden Proteine auf Gitter projeziert, das größere Protein wird starr gehalten, während das kleinere rotiert wird. Die globale Suche findet im Fourier-Raum statt, die geometrische Komplementarität wird durch eine Korrelationsfunktion bestimmt. Der große Vorteil der Fourier Transformation ist die Translations-Invarianz, die eine globale Suche zu akzeptablen Laufzeiten überhaupt erst möglich macht. Anstatt N^6 Berechnungen sind nur noch N^3 durchzuführen. Elektrostatische Wechselwirkungen wurden nachträglich eingebaut und konnten die Ergebnisse erheblich verbessern. (Heifetz *et al.* (2002)). Kenntnis vom Ort der Bindestelle ist nicht nötig. Andere Ansätze basieren auf den *sparse critical points* und der Verwendung von *geometric hashing* (Nussinov & Wolfson (1991), Lin *et al.* (1994), Norel *et al.* (1995), Fischer *et al.* (1995)).

1.4.1.2 Flexibilität

In fast allen zitierten und nicht zitierten Ansätzen zum Protein-Protein-Docking ist versucht worden zumindest die Seitenketten flexibel zu gestalten. Beispiel hierfür ist das *low resolution docking* (Vakser (1996b)), in dem versucht wird durch eine "ungenaue" Oberfläche allgemeine Flexibilität einzuführen. Das erste *soft docking* Verfahren von Jiang & Kim (1991) arbeitet mit einer Gitterdarstellung der Oberflächen. Die Zellen werden verglichen und damit die Oberflächenkomplementarität bestimmt. Mit einer bitweisen Darstellung der Proteine und der Kombination mit NMR-Techniken versuchen Palma *et al.* (2000) und Morelli *et al.* (2001) Flexibilität zu modellieren. Genetische Algorithmen mit einer Gradientenminimierung sind von Taylor & Burnett (2000) verwendet worden. Eine Kombination aus Brownscher Dynamik und Monte Carlo Optimierung von Fernandez-Recio *et al.* (2002) erbrachte erst kürzlich gute Ergebnis im *unbound*-Docking. Die von Camacho *et al.* (2000) entwickelten Methode konvergiert von anfänglichen 10 Å *rmsd* bis auf 2 Å *rmsd*, was für strukturelle Untersuchungen ausreichend ist. Vielfach verwendet werden Rotamer-Bibliotheken, die nach den energetisch günstigsten Rotameren aus einer Datenbank suchen. Derartige Datenbanken oder Bibliotheken sind umso genauer, je besser die Auflösung der Strukturen ist, aus denen die Rotamere extrahiert wurden (Xiang & Honig (2001)). Viele weitere

Ansätze bemühen sich dem Problem des flexiblen Dockings nachzugehen, eine allgemeine Lösung ist jedoch noch nicht in Sicht.

1.4.1.3 CAPRI (*Critical Assessment of Predicted Interactions*)

Die qualitative und quantitative Bewertung der zahlreichen und sehr unterschiedlichen Methoden ist schwierig. Daher wurde 2001 der Wettbewerb *Critical Assessment of Predicted Interactions* (CAPRI) ins Leben gerufen (*Conference on Modeling of Protein Interactions* Vajda *et al.* (2002)). Dieser Wettstreit soll es möglich machen, Protein-Protein-Komplexe zu docken bevor deren ko-kristallisierte Struktur bekannt ist und anschließend mit der experimentellen Struktur zu vergleichen. Ähnliches wurde 1994 von John Moult für die Protein-Struktur-Vorhersage ins Leben gerufen (*Critical Assessment of Structure Prediction*, CASP). Die CASP-Wettbewerbe sind so erfolgreich, daß mittlerweile der fünfte davon stattgefunden hat (Schonbrun *et al.* (2002)). Im CAPRI-Wettbewerb stehen momentan drei Targets zur Verfügung. Die Ergebnisse stehen noch aus. Eine Teilnahme an diesem Wettbewerb wird angestrebt.

1.4.2 Protein-Ligand-Docking

Das Docking kleiner Moleküle unterscheidet sich vom Protein-Protein-Docking signifikant. Die Annahme rigider Körper ist hier noch weniger gerechtfertigt als bei Protein-Protein-Komplexen, da die für eine gute geometrische Bewertung vorhandenen Kontaktflächen deutlich kleiner sind. Die Verwendung einer einzigen Liganden-Anordnung ist somit nicht erfolgreich, während dies für Protein-Protein-Komplexe oftmals ausreichend ist. Die Wertung elektrostatische Wechselwirkungen ist damit für Protein-Ligand-Komplexe wesentlich wichtiger. Die meisten Methoden verwenden rigide Rezeptoren und flexible, sequenziell in der Bindetasche aufgebaute Liganden (FlexX von Rarey *et al.* (1996), Rarey *et al.* (1999), GOLD Jones *et al.* (1995), AutoDock Morris *et al.* (1998)). Die Verwendung eines Ensembles an Rezeptoren ermöglicht vollständige Flexibilität für beide Komponenten (FlexE von Claussen *et al.* (2001)).

1.4.3 Bewertungsfunktionen

Die Bewertung der vielen Komplex-Möglichkeiten stellt das eigentliche Problem des Dockings dar. Eine exakte geometrische und energetische Beschreibung der Komplex-Anordnung in wässriger Lösung ist die einzige Möglichkeit den nativen Komplex zu identifizieren. Die vollständige Berechnung der freien Energie jeder erzeugten Variante ist zeitkritisch und exakt nicht möglich. Es werden daher empirische Funktionen entwickelt, die zusammen betrachtet begrenzt diskriminatorische Kraft haben (Norel *et al.* (1999)). Die praktische Anwendbarkeit dieser Funktionen ist im Mittel beschränkt. Es werden integrierte Berwertungsfunktionen von solchen zur nachfolgenden Filterung unterschieden. Alle Protein-Protein- und Protein-Ligand-Docking- Verfahren haben eine geometrische Bewertung und Strafe für sterische Überlappung integriert. Beim *unbound*-Docking entstehen hier bereits die ersten Schwierigkeiten. Es muß ein Gleichgewicht zwischen "tolerieren" von Überlappung aufgrund von Konformationsänderungen und "verbieten" gefunden werden. Elektrostatische Komplementarität wird bei fast allen Methoden berücksichtigt. Trotzdem bleiben mitunter tausende Möglichkeiten von ursprünglich etlichen Millionen übrig.
Der Bedarf an einer probaten Bewertungsfunktion, die schnell und sicher eine native Konformation erkennt und möglichst alle in Kapitel 1.3 vorgestellten energetischen Beiträge berücksichtigt, ist groß. Im folgenden Kapitel soll ein Überblick über einige Ansätze gegeben werden.

1.5 Vorhersage thermischer Größen aus der Literatur

Es werden im wesentlichen drei Ansätze zur Berechnung von Bindungsaffinitäten unterschieden. Diese sollen auf den folgenden Seiten in Kürze erläutert werden, für detailliertere Beschreibungen sei auf die zitierte Literatur verwiesen.

1.5.1 Störungsrechnungen und thermodynamische Integration

Die Freie-Energie-Störungsrechnung stellt die einzig korrekte Methoden zur Vorhersage relativer freier Energien von Makromolekülen dar. Ebenbürtig ist die

thermodynamische Integration mit expliziter Berücksichtigung von Solvensmolekülen und Flexibilitäten (Murcko (1995)). Grundlage für die Störungsrechnung ist der Zusammenhang der freien Energie eines Systems mit dem Ensemble-Mittelwert einer Energiefunktion, die das System beschreibt. Die Konfigurationsensemble können mit einer Monte-Carlo(MC)- oder Molekulardynamik(MD)-Simulation gewonnen werden. Problematisch ist die Durchsuchung des Phasenraumes und die Genauigkeit des Hamiltonian (Kraftfeld), so daß eine allgemeine Anwendbarkeit nicht gegeben ist. Die langen Rechenzeiten sowie die Beschränkung auf kleine Änderungen verhindern die Anwendung auf Protein-Protein-Komplexe (Kasper *et al.* (2000), Leach (2001)). Dennoch sind sie standardmäßig in den meisten Kraftfeldern integriert (Pearlman *et al.* (1991)).

1.5.2 Regressionsbasierte Methoden

Empirische Funktionen basieren auf der Anpassung von Gewichtungstermen an strukturelle Daten. Experimentelle und berechnete Bindungsdaten werden mittels Regressionsverfahren für einen Strukturdatensatz optimiert. Obwohl die Zerlegung der Bindungsenergie in Einzelbeträge unter theoretischen Gesichtspunkten zweifelhaft ist (Horovitz (1987), Dill *et al.* (1997)) werden die Wichtungsfaktoren als freie Energie-Beiträge verschiedener Kräfte interpretiert. Das Hauptproblem dieser Methoden ist die Abhängigkeit von Größe und Qualität des Trainingsdatensatzes, sowie die Mehrfachbewertung von Interaktionen, die als unabhängig angenommen werden, es aber letztendlich nicht sind.

1.5.3 Wissensbasierte Methoden

Wissensbasierte (engl. *knowledge-based*) Systeme beruhen auf der Annahme, daß aus einer ausreichend großen Sammlung von Informationen Regeln abgeleitet und auf neue Problemstellungen angewendet werden können. Diese Methoden lassen sich auf molekularbiologische Sachverhalte übertragen, wie beispielsweise der Gesamtheit an bekannten dreidimensionalen Kristallstrukturen von Proteinen. Aus einer Datenbank dieser Strukturen kann eine bestimmte Eigenschaft, wie der Abstand zweier Partikel extrahiert werden. Eine Verteilungsfunktion des

Merkmals wird erstellt. Sie beschreibt die Wahrscheinlichkeit zwei Partikel in einem bestimmten Abstand zu finden.

Soll eine neue Proteinstruktur hinsichtlich des gewählten Merkmales bewertet werden, so kann die Verteilung im Protein mit der Verteilung der gesamten Datenbank verglichen werden. Eigenschaften in der unbekannten Proteinstruktur, die in der Nähe von Maxima der Verteilung aus der Datenbank liegen, werden als besonders "gut" bewertet. Werden die Häufgkeitsverteilungen interatomarer Wechselwirkungen in Proteinen untersucht, wird meist das *inverse* Boltzmann-Gesetz verwendet um aus den Merkmalsverteilungen "freie-Energie-Beiträge" (*potentials of mean force*) zu berechnen (Sippl (1990),Miyazawa & Jerningan (1985), Miyazawa & Jernigan (1996), Reva *et al.* (1997)). Dem Boltzmannschen Verteilungsgesetz zufolge korrespondieren stark besetzte Zustände mit einem geringem Energieinhalt. Die native Struktur eines Proteins oder Protein-Protein-Komplexes liegt in der Nähe des thermodynamischen Minimums (Ben-Naim *et al.* (1990), Janin (1996), Maiorov & Crippen (1992)).

Die physikalisch sinnvolle Verwendung der Boltzmann-Statistik für Proteine ist umstritten (Ben-Naim (1987), Zhang & Skolnick (1998), Finkelstein & Gutin (1995)). Proteine formen eher einzigartige dreidimensionale Strukturen als ergodische Ensemble von unabhängigen Aminosäuren oder Atomen. Es ist fraglich ob sich die Proteine einer Datenbank im thermodynamischen Gleichgewicht befinden, noch ist sicher, daß sich ein Protein oder gar alle Proteine im eigenen und im gemeinsamen thermodynamischen Minimum befindet. Die Proteinstrukturen sind unter verschiedenen experimentellen Bedingungen gewonnen worden, von gemeinsamer Temperatur oder Druck kann nicht ausgegangen werden. Das Boltzmannsche Gesetz ist auf ein Ensemble von Teilchen im thermodynamischen Gleichgewicht anwendbar, nicht aber auf jedes Teilchen alleine. Nach Thomas *et al.* (1996) wird der Ensemble-Mittelwert effektiv ersetzt durch einen Mittelwert über einen Satz an repräsentativen Proteinstrukturen. Die Korrelation ist gut, doch die Absolutwerte unpräzise.

Das Boltzmannsche Gesetz beschreibt das Verhältnis zweier Besetzungszustände, also zweier Verteilungen. Die Wahl des daher erforderlichen "Referenzzustandes" ist kritisch für die Qualität des Potentials (Betancourt & Thirumalai (1999)).

Der Referenzzustand kann als zufällige Mischung von Aminosäuren angenom-

men werden (Miyazawa & Jerningan (1985)). Das Potential ist dann ein Maß für die Abweichung der nicht-zufallsgesteuerten Natur von Aminosäure-Kontakt-Verteilungen in Protein Strukturen (Gan *et al.* (2001)). Diese quasichemische Näherung wurde von Skolnick *et al.* (1997) und Godzik *et al.* (1995) untersucht. Es werden ungefaltete Proteinketten, sog. *random coils* oder kompakte aber zufällige Proteine (Godzik (1996)) verwendet. In einigen Fällen wird ein globaler Referenzzustand verwendet (Robert & Janin (1998)). Anschaulich entspricht das der mittleren Aminosäure-Kontakt-Verteilung.
Finkelstein & Gutin (1995) postulierten für ein Modell von zufälligen Heteropolymeren eine Boltzmann-ähnliche Verteilung. Thomas *et al.* (1996) hingegen stehen der Boltzmann-ähnlichen Verteilung von Strukturen kritisch gegenüber. Mohanty *et al.* (1999) fand gute Korrelation für die Entfaltung von GCN4 Leucin Zipper zwischen einem *knowledge-based-potential* und einer detaillierten atomaren Beschreibung mit MD-Simulation.
Es wurde von Sippl (1993a) und von Miyazawa & Jernigan (1999) angemerkt, daß Lösungsmitteleffekte durch diesen Formalismus nicht ausreichend berücksichtigt werden. Die zuvor erläuterten empirischen Solvatationsmodelle können als Korrekturterme verwendet werden.

1.6 Problemstellung

Es soll ein Potential entwickelt werden, daß eine native Protein-Protein-Anordnung von einer nicht-nativen unterscheiden kann. Die Funktion muß möglichst alle wichtigen Energiebeiträge der Assoziation aus den letzten Kapiteln berücksichtigen und soll gleichzeitig ein so schnelles Laufzeitverhalten aufweisen, daß es als Post-Filter von Docking-Studien verwendet werden kann. Eine Einbindung in das Fourier-Docking-Program ckordo als integrale Bewertungsfunktion soll ebenfalls möglich sein. Das Potential soll für die Anwendung im reinen *unbound*-Docking validiert und optimiert werden. Aus den zahlreichen vorgestellten Ansätzen, ist ein wissenbasiertes Potential mit impliziten Solvensmodell am ehesten in der Lage den Anforderungen gerecht zu werden.

2 Theorie und Methoden

The distribution of distances
for any given pair was determined by evolution,
or by god, not by Boltzmann !

Ben-Naim (1987)

2.1 Begründung des Ansatzes und zu erwartende Vorhersagequalität

Zur Abschätzung von Protein-Protein-Interaktionen wird ein abstandsabhängiges, atombasiertes Potential ausgehend von dem von Sippl (1990) und Gohlke *et al.* (2000) entwickelten Formalismus verwendet. Der Ansatz ist als heuristisch motiviert zu betrachten.
Der Vorteil dieser Methode ist die Tatsache, daß alle intermolekularen Kräfte, ob bisher verstanden oder nicht, **implizit** berücksichtigt werden. Die Aufstellung einer **expliziten** Energiefunktion für Proteine in Lösung ist schwierig. Die Form eines Potentials und die Art der betrachteten Wechselwirkungen wird nach Godzik *et al.* (1995) nur durch die Größe von r_{max}, also dem definierten Maximalabstand, bestimmt. Das hier entwickelte Potential wird nur für kurze Abstände bis acht Å bestimmt und beinhaltet daher keine Solvatationseffekte. Die entropischen Kräfte werden durch ein lösungsmittelabhängiges Ein-Teilchen-Potential beschrieben. Die energetisch wichtigen Wasserstoffbrücken und Kontakte funktioneller Gruppen werden besonders gewichtet, um deren höheren Beitrag herauszuarbeiten. Des weiteren wird ein repulsiver Korrekturterm eingeführt, um sterische Überlappungen als Artefakte aus dem Docking-Prozess zu bestrafen. Eine in den folgenden Kapiteln erläuterte Trapez-Gewichtung führt zu einer "Glättung" der Verteilungen. Daraus resultieren weichere Wahrscheinlichkeitsdichteverteilungen. Diese Ungenauigkeit oder Oberflächlichkeit in den Verteilungen ist wichtig für die Modellierung von Flexibilitäten und wirkt einer

Überbewertung von Details entgegen. Eine weiterer Vorteil der Glättung ist eine gegen kleine Änderungen, wie nicht korrekte Geometrien sehr robuste Funktion. Die teils erstaunlichen Ergebnisse mit wissensbasierten Methoden in der Diskriminierung nativer von nicht nativen Strukturen (Sippl (1993b), Skolnick *et al.* (2000), Jones *et al.* (1992), Novotny *et al.* (1989), Reva *et al.* (1997)), im *Threading* (Miyazawa & Jernigan (1996)), dem *de novo* Proteindesign (Jones (1994)), bei der Vorhersage von Thermostabilitäten von Proteinen (Hoppe (2002), Gilis & Rooman (1997), Guerois *et al.* (2002)) , im Bereich des Protein-Ligand-Dockings (Gohlke *et al.* (2000), Nobeli *et al.* (2000), Mitchell *et al.* (1999a), Mitchell *et al.* (1999b)), sowie in der Untersuchung von Protein-Protein-Interaktionen (Jiang *et al.* (2002), Glaser *et al.* (2001), Moont *et al.* (1999)) stellen eine Rechtfertigung hinsichtlich guter Ergebnisse und Reproduzierbarkeit dar. Dennoch muß nochmal betont werden, daß die vorgestellte Methode mit einigen physikalischen Prinzipien nur schwer in Einklang zu bringen ist und die trotzdem guten Vorhersagen streng genommen nicht mit entsprechenden naturwissenschaftlichen Theorien erklärt werden können. Eine genauere Betrachtung erfolgt in der Diskussion.

2.1.1 Atombasiert *vs* Aminosäurebasiert

Für die Beschreibung von Protein-Protein-Interaktionen werden wegen der höheren Genauigkeit atombasierte, distanzabhängige radiale Verteilungsfunktionen verwendet. Die *coarse grained*-Methoden nähern eine Aminosäure als einen Punkt im Raum an und entsprechen keineswegs einer detaillierten atomaren Beschreibung. Meist werden nur vier bis fünf Atomtypen definiert, so fallen alle aliphatischen Kohlenstoffatome in eine Kategorie. Zhang *et al.* (1997) untersuchten die Verwendung von achtzehn Atomtypen. Die Kontaktenergien variieren über die gesamte Länge der Verteilung. Es wird vorgeschlagen eine genauere Unterteilung zu verwenden. Die zwanzig Standard Aminosäuren enthalten weit über hundert verschiedene Atome. Jedes Atom befindet sich in einer spezifischen chemischen Umgebung und müsste daher gesondert betrachtet.
Für die vorliegende Arbeit werden einem Ansatz von Melo & Feytmans (1997) folgend vierzig Atomtypen zur Beschreibung aller Atome ausgewählt und als eigenständige Entitäten behandelt (vergl. Abb. 2.1).

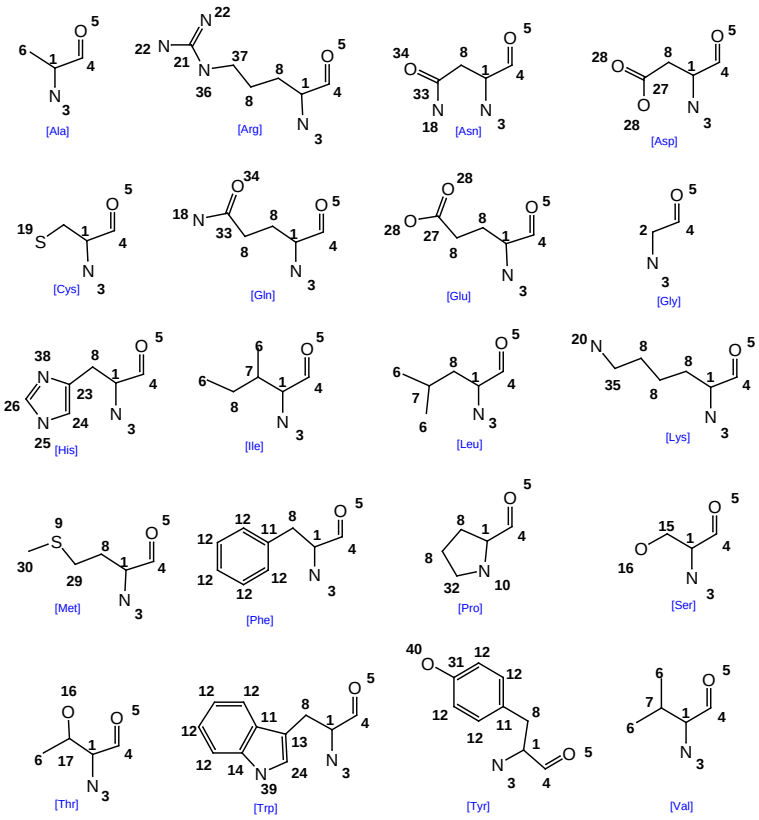


Abbildung 2.1: Einteilung der Atome in vierzig verschiedene Typen.

Den Atomtypen werden numerische Werte von eins bis vierzig zugeordnet. Es ergeben sich $N(N+1)/2 = 820$ Kombinationsmöglichkeiten mit $N = 40$. Jede Paarungsmöglichkeit erhält einen eindeutigen numerischen Wert UID:

$$UID = (x-1) * 40 - \left[(x-1) * \frac{x}{2}\right] + y \tag{2.1}$$

mit x als dem numerischen Wert des ersten Atoms (mit kleinerem numerischem Wert) und y dem Wert des zweiten Atoms (mit größerem numerischen Wert).

2.1.2 Funktionelle Gruppen

Die chemische Umgebung von Hauptkettenatomen ändert sich von Protein zu Protein nur marginal. Die Atome funktioneller Gruppen von Aminosäuren hingegen bilden den Hauptanteil der Wechselwirkungsenergie und ihre Umgebung ändert sich stärker. Als einer funktionellen Gruppe zugehörig wird jedes Atom definiert, das nicht Teil der Proteinhauptkette ist (vergl. Tabelle A.2, Anhang). Es wird angenommen, daß die Hauptkette mehr zum Protein hin gerichtet ist, als die Seitenketten. Daher werden als erste Näherung nur Seitenkettenatome betrachtet und von diesen auch fast ausschließlich besonders stark polarisierten Atome, wie Stickstoff und Sauerstoff. Alle Kombinationsmöglichkeiten der ausgewählten funktionellen Atome untereinander werden anhand eines speziellen Gewichtungsfaktors höher bewertet als andere Atomkombinationen (vergl. Kapitel 2.1.6).

2.1.3 Wasserstoffbrücken

Die Anzahl an Wasserstoffbrücken-Donoren und -Akzeptoren wird durch Auszählen aller Stickstoff-, Sauerstoff- und Schwefelatome mit einem Abstand r_d mit $2,8\ \text{Å} < r_d < 3,0\ \text{Å}$ bestimmt. Eine Direktionalität wird außer Acht gelassen. Alle entsprechenden Atom-Atom-Kombinationen werden gesondert gewichtet.

2.1.4 Paarkorrelationsfunktionen

Die physiko-chemische Grundlage des *potential of mean force*-Ansatzes geht auf die statistisch-mechanische Theorie fluider Systeme zurück (Ben-Naim (1987), Lucas (1986), Hill (1986)). Danach kann eine n-Teilchen-Korrelationsfunktion $g_n(\mathbf{r}_1\omega_1, ..., \mathbf{r}_n\omega_n)$ mit den Orts- und Orientierungskoordinaten $\mathbf{r}$ und ω in ein *potential of mean force* $W_n(\mathbf{r}_1\omega_1, ..., \mathbf{r}_n\omega_n)$ umgewandelt werden:

$$W_n(\mathbf{r}_1\omega_1, ..., \mathbf{r}_n\omega_n) = -RT \ln g_n(\mathbf{r}_1\omega_1, ..., \mathbf{r}_n\omega_n) \tag{2.2}$$

Für ideale Gase, in denen keine intermolekularen Wechselwirkungen vorliegen, die Moleküle vollkommen unabhängig sind, ist $g = 1$ Die Abweichung von eins im allgemeinen Fall ist damit ein Maß für den Einfluss der intermolekularen Kräfte auf die Korrelation der Einzelmolekülpositionen - und Orientierungen untereinander.

Für zwei wechselwirkende Teilchen mit sphärischer Symmetrie wird $g_{AB}(\mathbf{r_{AB}})$ als radiale Paarkorrelationsfunktion oder auch radiale Verteilungsfunktion bezeichnet. Sie gibt das Verhalten des Systems vollständig wieder. Sämtliche thermodynamische Größen können aus Integralen über Korrelationsfunktionen berechnet werden, wenn exakte paarweise Additivität der Wechselwirkungen (Superpositionsprinzip) angenommen wird (Hill (1986), Findenegg (1985)). Für Proteine in Lösung ist die exakte paarweise Additivität umstritten (Ben-Naim (1987)). Die Verwendung empirischer Paarpotentiale für derartige Systeme geht auf eine Näherung des Superpositionsprinzip zurück (Kirkwood (1935)). Radiale Verteilungsfunktionen lassen sich theoretisch auch aus wechselwirkenden Atom-Paaren kristallisierter Protein-Protein-Strukturen erhalten. Dieser Ansatz ist aus physikalischen Sicht kritisch, da die radiale Verteilungsfunktion druck-, temperatur- und systemabhängig ist (Ben-Naim (1987), Thomas & Dill (1996)).

Die Korrelationsfunktion g_{AB} ist eine Funktion des skalaren Abstandes $\mathbf{r_{AB}} \equiv |\mathbf{r}_A - \mathbf{r}_B|$ der Teilchen A und B und ist proportional zur Gesamtwahrscheinlichkeit $W_{AB}(\mathbf{r}_A, \mathbf{r}_B)dr_A dr_B$, die das Auffinden des Teilchen A an der Position r_A und B bei r_B innerhalb des Volumenelementes dr_A und dr_B beschreibt, verglichen mit der erwarteten Wahrscheinlichkeit für eine vollständig zufällige Verteilung. Die winkelabhängige Paarkorrelationsfunktion $g_{AB}(\mathbf{r_{AB}}\omega_{AB})$ beschreibt zusätzlich die relative Orientierung ω_{AB} des einen Moleküls gegenüber dem anderen. In dieser Arbeit wird eine Direktionalität jedoch außer Acht gelassen.

In einem isotropen, homogenen Fluid ist die Wahrscheinlichkeit ein Molekül mit den Koordinaten $(\mathbf{r}_1\omega_1)$ ohne Spezifikation der Koordinaten der anderen Moleküle zu finden unabhängig von $\mathbf{r}_1$ und ω_1. Daraus ergibt sich die Anzahl $N_{AB}(\mathbf{r_d})$ der Paare A, B in einem Abstand r_d (Leach (2001)):

$$N_{AB}(\mathbf{r_{AB}}) = N_{B|A}(\mathbf{r_d})N_A = N_A\rho_B g_{AB}(\mathbf{r_{AB}})4\pi r^2 dr \tag{2.3}$$

$N_{B|A}(\mathbf{r_d})$ ist die Anzahl an Partikeln des Types B im Kugelvolumen $4\pi r^2 dr$ mit einem Abstand r vom Zentral-Teilchen A, ρ_B ist die Dichte von B und N_A die Teilchenzahl der Atome A.

Die radiale Verteilungsfunktion kann durch Auszählen der Kontakthäufigkeiten der Atome A und B im Intervall $[r_{AB}, r_{AB} + dr)$ bestimmt werden. Es wird angenommen (Gohlke *et al.* (2000)), daß

- alle Distanzen paarweise unabhängig sind
- die Verteilung interatomarer Distanzen eines Atompaares in verschiedenen Umgebungen ähnlich ist
- die Verteilung der Distanzen hinreichend scharf und voneinander getrennt sind

Mehrkörperwechselwirkungen werden vernachlässigt. Nach Gan *et al.* (2001) erbringen sie keine signifikante Verbesserung. Die Additivität der Zwei-Körper-Wechselwirkungen wird vorausgesetzt.

2.1.5 Theorie der *inversen* Boltzmann Gleichung

Die Theorie der *inversen Boltzmann Gleichung* wird von mehreren Autoren klar und übersichtlich dargestellt (Sippl (1990), Sippl (1995), Koppensteiner & Sippl (1998), Godzik *et al.* (1995), Godzik (1996), Gohlke *et al.* (2000)), die folgenden Abschnitte sollen diese Ausführungen kurz zusammenfassen und Erweiterungen aufzeigen.
Die Verteilung von Molekülen in Mikrozuständen wird mit dem Boltzmannschen Gesetz beschrieben. Demnach korrespondiert die Energie $E_{ij}(r_{ij})$ eines Subsystems mit der Wahrscheinlichkeitsdichtefunktion $\rho_{ij}(r_{ij})$. Für zwei Atome i und j mit einem Abstand r_{ij} gilt:

$$\rho_{ij}(r_{ij}) = \frac{exp(-E_{ij}(r_{ij})/kT)}{Z_{ij}} \tag{2.4}$$

$Z_{ij} = \sum_{ij} exp(-E_{ij}(r_{ij})/kT)$ ist die Zustandssumme.
Sind alle $E_{ij}(r_{ij})$ bekannt, so läßt sich daraus die Verteilung der Zustände bestimmen. Wenn wiederum die Wahrscheinlichkeitsdichtefunktion $\rho_{ij}(r_{ij})$ bekannt ist, kann die Energie eines Mikrozustandes mit dem inversen Gesetz berechnet werden:

$$E_{ij}(r_{ij}) = -kTln(\rho_{ij}(r_{ij})) - kT \ln Z_{ij} \tag{2.5}$$

Ein Vergleich der Energien der Mikrozustände mit einem Referenzsystem $E(r)$ ergibt das **Netto**-Potential $\Delta E_{ij}(r_{ij})$:

$$\Delta E_{ij}(r_{ij}) = E_{ij}(r_{ij}) - E(r) = -kT \ln \left[\frac{\rho_{ij}(r_{ij})}{\rho(r)}\right] - kT \ln \left[\frac{Z_{ij}}{Z}\right] \tag{2.6}$$

Beide Verteilungsdichtefunktionen $\rho_{ij}(r)$ und $\rho(r)$ müssen normiert sein (Bahar & Jernigan (1997)) um eine schnelle Konvergenz gegen 0 für große Distanzen zu gewährleisten: $\rho(r) = \sum_{ij} \rho_{ij}(r)$ mit $\sum_{ij} \rho(r) = \sum_{ij} \rho_{ij}(r) = 1$.

Als Näherung gilt $Z_{ij} \approx Z$, da Z und Z_{ij} nicht aus den Wahrscheinlichkeitsdichtefunktionen bestimmt werden können nicht von der Zustandsvariablen r abhängen. Daraus folgt: $-kT \ln \left[\frac{Z_{ij}}{Z}\right] \approx 0$.
Die normalisierte radiale Verteilungsfunktion ergibt sich nach Kapitel 2.1.4 zu:

$$g_{ij}(r_d) = \frac{N_{ij}(r_d)/4\pi r_d^2}{\sum(N_{ij}(r_d)/4\pi r_d^2)} \tag{2.7}$$

$N_{ij}(r_d)$ ist die Kontakthäufigkeit der Atome i und j im Abstand r_d, wenn man den kontinuierlichen Abstand $d_{ij} =\mid i - j \mid$ auf ein Intervall $[r_{min}, r_{max})$ beschränkt. Bei festgelegter Intervallweite dr ergibt sich der dazugehörige Wert r_d:

$$r_d = r_{min} + \llcorner \frac{(d_{ij} - r_{min})}{dr} \lrcorner dr \tag{2.8}$$

mit $r_d \leq d_{ij} < (r_d + dr)$. Die Summation läuft über alle Intervalle d im Bereich $[r_{min}, r_{max})$. $\llcorner \lrcorner$ bedeutet, daß die größte ganze Zahl zu nehmen ist, die kleiner als der errechnete Wert ist.
$g_{ij}(r_d)$ konvergiert nach dem Gesetz der großen Zahlen mit der Wahrscheinlichkeitsdichte:

$$\rho_{ij}(r_d) : \lim_{n->\infty} g_{ij}(r_d) \equiv \rho_{ij}(r_d)$$

.
Die Wechselwirkungen zwischen zwei Atomen i und j wird als symmetrisch angenommen, obwohl aufgrund der Richtungsabhängigkeit der Proteinkette $(i, j) \neq (j, i)$ gilt.

2.1.6 Der Referenzzustand

Die Frage nach einem passenden Referenzzustand ist kontrovers diskutiert. Man könnte beispielsweise die vollständige Separierung von beiden Proteinen als gutes System annehmen. Dagegen spricht die Schwierigkeit ein derartiges Modell in den verwendeten Formalismus ohne explizites Lösungsmittel einzubauen.
Außerdem wäre ein Referenzzustand wünschenswert, der auch nicht-spezifische, nicht-native Geometrien erfasst. Der Argumentation von Gohlke *et al.* (2000) folgend sind die benutzen Komplexstrukturen mehr oder minder kompakt und die darin enthaltenen Informationen nach einer Mittelung nicht mehr spezifisch. Ein Referenzsystem läßt sich demnach aus einer Datenbank kristallographischer Strukturen mit Atomen willkürlichen Typs (entspricht einer Mittelung über alle Atomtypen) erzeugen. Der Referenzzustand ist also die *a*

priori Verteilung zweier Atome A und B in einem Abstand r_d in einem nativen oder nicht nativen Komplex.
Sippl (1993a) schlug ein Referenzsystem aus dem Mittelwert über ein Set an Subsystemen bei festgelegtem i und j und der unabhängigen Variablen r_d vor.
Die **gemittelte**, normalisierte radiale Paarverteilungsfunktion $\widetilde{g}_{r_d}$ ergibt sich demnach als Mittelwert über alle normalisierten radialen Paarverteilungsfunktionen:

$$\widetilde{g}(r_d) = \frac{\sum_i \sum_j N_{ij}(r_d)/4\pi r_d^2}{\sum_i \sum_j \sum_d N_{ij}(r_d)/4\pi r_d^2} \tag{2.9}$$

Dieser Referenzzustand ist abhängig von der Anzahl häufig auftretender Fälle N_{ij} in einzelnen Paarverteilungen. Wird über die Anzahl an Kombinationsmöglichkeiten von Atomen der Typs j mit denen des Typs i gemittelt, wird dieses Problem umgangen:

$$g(r_d) = \frac{\sum_i \sum_j g_{ij}(r_d)}{||j|| * ||i||} \tag{2.10}$$

In der Literatur ist die Form des besten Referenzsystems umstritten. Eine andere Annahme, als die oben diskutierte "Zufallsverteilung", ist die Idee des energetischen "Nullzustandes" (Thomas & Dill (1996)). Gesucht ist ein System, in dem die Interaktionsenergie null ist, also ein wechselwirkungsfreies System. Es gibt jedoch - nicht unerwartet - keine Daten über ein Protein-System dessen Interaktionsenergie null ist. Daraus ergibt sich die Schwierigkeit ein solches zu beschreiben. Eine übliche Annahme für derartige "nicht-interagierende" Systeme ist das Produkt der Molfraktionen (Godzik *et al.* (1995)), gleichwohl dieser Ansatz wie jeder andere seine Schwachpunkte hat.
Aus diesem Grund wird die "These der Gleichverteilung" ins Leben gerufen und der daraus resultierende Referenzzustand gegen den aus Gleichung 2.10 getestet. Eine Beantwortung der Frage ob überhaupt und wenn, inwieweit

die im Folgenden vorgestellte These in physikalisch sinnvoller Weise mit dem Boltzmann-Ansatz kombiniert werden kann, erfolgt in der Diskussion.

Es wird angenommen, daß alle interatomaren Abstände eines Sytems mit der gleichen Wahrscheinlichkeit auftreten. Daraus wird eine monoton steigende Verteilungsfunktion $P(x)$ in den Grenzen (a,b) mit $P(a) = 0$ und $P(b) = 0$ erhalten.

$$P(x) = \begin{cases} 1 & \text{für } (0 \leq x \geq 1) \\ 0 & \text{sonst} \end{cases} \tag{2.11}$$

Die Wahrscheinlichkeitsdichte $p(x)$ ergibt sich als Differentialquotient der Verteilungsfunktion:

$$p(x) = \frac{dP(x)}{dx} \tag{2.12}$$

Für eine Gleichverteilung gilt $p(x) = 1/(b-a)$. Die Wahrscheinlichkeitsdichte $p(x)$ wird auf die Anzahl der Atomtypen normiert, woraus sich eine skalare Verteilung mit dem Wert $g(r_d) = 0.025$ ergibt.

Zur Unterscheidung der beiden Ansätze, soll die skalare Verteilung $g(r_d) = 0.025$ als **Profil 1** und die aus gemittelten, normierten, radialen Verteilungsfunktionen erhaltene Verteilung gemäß Gleichung 2.10 als **Profil 2** bezeichnet werden.

Die Größe $\Delta E_{ij}(r_d)$ (Gleichung 2.6) soll wegen der fragwürdigen Anwendbarkeit des Boltzmann-Gesetzes (vergl. Kapitel 1.5.3 im Folgenden einem Vorschlag von Gohlke *et al.* (2000) entsprechend als statistische **Netto-Präferenz** $\Delta W_{ij}(r_d)$ bezeichnet werden.
Für die statistische Netto-Präferenz $\Delta W_{ij}(r_d)$ als Differenz des betrachteten Subsystems $W_{ij}(r_d)$ und des Referenzzustandes $W(r_d)$ ergibt sich analog zu Gleichung 2.6:

$$\Delta W_{ij}(r_d) = -kT \ln \frac{g_{ij}(r_d)}{g(r_d)} \tag{2.13}$$

Die Anzahl der Atompaare $N_{ij}(r_d)$ mit den Atomen i und j ergibt sich durch Auszählen der Häufigkeit des Auftretens, summiert über alle Atome beider Proteine P und Q eines Komplexes für alle nativen Komplexe K:

$$N_{ij}(r_d) = \sum_{k \in K_n} \sum_{p \in P} \sum_{q \in Q} \delta(d_{ij}, r_d) \tag{2.14}$$

mit $\delta(d_{ij}, r_d) = 1$ wenn $d_{ij} \in [r_d, r_d + dr)$, sonst 0.
Unter Berücksichtigung der funktionellen Gruppen und Wasserstoffbrücken-bildenden Atome ergibt sich die Netto-Präferenz:

$$\Delta W_{ij}^{Paar}(r_d) = \sum_{p \in P} \sum_{q \in Q} \left[\Delta W_{ij}^{f}(r_d) + \Delta W_{ij}^{nf}(r_d) + \Delta W_{ij}^{wb}(r_d) \right] \tag{2.15}$$

mit f = funktioneller Gruppe angehörig, nf = keiner funktionellen Gruppe angehörig, wb = Wasserstoffbrücken-bildend und α, β, γ als Gewichtungsfaktoren und

$$\Delta W_{ij}^{f}(r_d) = \alpha \Delta W_{ij}(r_d, f) \tag{2.16}$$

$$\Delta W_{ij}^{nf}(r_d) = \beta \Delta W_{ij}(r_d, nf) \tag{2.17}$$

$$\Delta W_{ij}^{wb}(r_d) = \gamma \Delta W_{ij}(r_d, wb) \tag{2.18}$$

2.1.7 Das Oberflächenpotential

Ein an die Oberflächenzugänglichkeit des Proteins angepasstes Ein-Teilchen-Potential gleicht den niedrig gewählten r_{max} aus und soll somit entropische Faktoren repräsentieren. Die bisher bekannten Ein-Teilchen-Potentiale dienen der Evaluierung von Protein-Ligand-Komplexen und müssen daher an die Gegebenheiten von Protein-Protein-Komplexen angepasst werden.
Die Potentiale ergeben sich als negative Logarithmen des Verhältnisses zweier lösungsmittelabhängiger, normierter Verteilungsfunktionen, die Wahrscheinlichkeit angeben mit der ein Proteinatom von P oder Q eines Typs im komplexierten bzw. im freien Zustand bei einer gegebenen Solvenszugänglichkeit (SAS) angetroffen wird. Abweichend von den Paarpotentialen gibt es hier keinen gemeinsamen Referenzzustand, die Präferenzen sind also unabhängig.
Nach Gohlke *et al.* (2000) ergibt sich für getrennt betrachtete Atome x der Proteine P und Q eine statistische Präferenz von $\Delta W_x^{P/Q}(S_i^{gb}, S_j^{ug})$ als Funktion der solvenzugänglichen Fläche im gebundenen Zustand SAS_x^{gb} innerhalb eines Intervalles i mit den Grenzen $[S_i^{gb}, S_i^{gb} + dS)$ und der entsprechenden Fläche im ungebundenen Zustand SAS_x^{ug} innerhalb des Intervalles j mit den Grenzen $[S_j^{ug}, S_j^{ug} + dS)$:

$$\Delta W_x^{P/Q}(S_i^{gb}, S_j^{ug}) = W_x^{P/Q}(S_i^{gb}) - E_x^{P/Q}(S_j^{ug}) = -ln\frac{g_x^{P/Q}(S_i^{gb})}{g_x^{P/Q}(S_j^{ug})} \tag{2.19}$$

mit gb = gebunden und ug = ungebunden. Die $SAS_x^{ug/gb}$ ist auf das Intervall $[S_{min}, S_{max})$ beschränkt. Bei festgelegtem dS ergibt sich der Wert $S_{i,j}^{ug/gb}$:

$$S_{i,j}^{ug/gb} = S_{min} + \left[\frac{SAS_x^{ug/gb} - S_{min}}{dS}\right] dS \tag{2.20}$$

$g_x(S_i^{gb})$ ist die normalisierte Verteilungsfunktion des Atoms x bezüglich der Oberfläche im komplexierten Zustand:

$$g_x^{P/Q}(S_i^{gb}) = \frac{N_x^{P/Q}(S_i^{gb})}{\sum_i N_x^{P/Q}(S_i^{gb})} \tag{2.21}$$

Die Summation läuft über alle Intervalle i der diskreten Verteilung. $g_x^{P/Q}(S_i^{gb})$ beschreibt die Wahrscheinlichkeit ein Atom x mit $SAS > 0$ im Komplex zu finden. Die analoge Gleichung für $g_x^{P/Q}(S_j^{ug})$ beschreibt die Wahrscheinlichkeit das gleiche Atom des Typs x mit gleicher SAS im ungebundenen Zustand zu finden. $g_x^{P/Q}(S_j^{ug})$ ist keine über alle Atomtypen gemittelte Funktion, sondern bezieht sich auf einen speziellen Atomtyp. $\Delta W_x^{P/Q}$ beschreibt nur den Beitrag, der durch Unterschiede in der SAS in komplexierten und freien Protein zustande kommen.

Analog zu den Paarpotentialen kann die Anzahl der Atome x im Intervall $[S_i^{gb/ug}, S_i^{gb/ug} + dS)$ durch Auszählen der Auftrittshäufigkeiten über alle Atome der Protein P_p und Q_q eines Komplexes k für alle nativen Komplexe K_n im Strukturdatensatz K ermittelt werden:

$$N_x^{P/Q}(S_i^{gb/ug}) = \sum_{k \in K_n} \sum_{x \in P_k/Q_k} \delta(SAS_x^{gb/ug}, S_i^{gb/ug}) \tag{2.22}$$

Im Gegensatz zu der von Gohlke *et al.* (2000) verwendeten getrennten Betrachtung der Präferenzen von Protein und Ligand wird hier die Summe der Präferenzen für beide Proteine verwendet. Im Protein-Ligand-Komplexen können die Verteilungen der Solvenszugänglichkeiten signifikant unterschiedlich sein. In Protein-Protein-Komplexen hingegen sind die Komplexpartner von vergleichbarer Größe. Der auf die wissensbasierten Ein-Teilchen-Potentiale zurückgehende Anteil an Wechselwirkungen für eine Konfiguration von P und Q ergibt sich somit zu:

$$\Delta W_x^{einzel}(S_i^{gb}, S_j^{ug}) = \sum_{p\in P}\sum_{q\in Q} \Delta W_p^P(S_i^{gb}, S_j^{ug}) + \Delta W_q^Q(S_i^{gb}, S_j^{ug}) \tag{2.23}$$

2.1.8 Das Gesamt-Netto-Potential

Zur Berechnung der Gesamt-Netto-Präferenz ergibt sich folgende Gleichung:

$$\begin{aligned}\Delta W^{ges} = &\sum_{p\in P}\sum_{q\in Q}\left[\Delta W_{ij}^f(r_d) + \Delta W_{ij}^{nf}(r_d) + \Delta W_{ij}^{wb}(r_d)\right] \\ &+\delta\left[\sum_{p\in P}\sum_{q\in Q}\Delta W_p^P(S_i^{gb}, S_j^{ug}) + \Delta W_q^Q(S_i^{gb}, S_j^{ug})\right]\end{aligned} \tag{2.24}$$

mit f = funktioneller Gruppe zugehörend, nf = keiner funktionellen Gruppe zugehörend und wb = Wasserstoffbrücken-Donor-Akzeptor-Paar.

2.2 Glättung und Gewichtung der Häufigkeitsverteilungen

2.2.1 Trapezfunktion

Um den Effekt der Diskretisierung der Distanzverteilung zu berücksichtigen und der Tatsache, daß Atompositionen in Kristallstrukturen mit einer Unsicherheit in den Ortskoordinaten von etwa 1/6 der maximalen Auflösung (bei 2,5 Å also 0,4 Å Ortsunschärfe) Wlodawer *et al.* (1987)) behaftet sind, Rechnung zu tragen, muss eine Glättung der Rohdaten vorgenommen werden. Des weiteren sind Atome keine Punkte im Raum, sondern haben eine räumliche Ausdehnung. Damit liegt ein Atom B in einer Kugelschale der Dicke dr mit einem Abstand r_d vom Atom A meist nicht vollständig in dieser Kugelschale. Oft liegt das Atom "gerade noch" innerhalb der Kugel, manchmal vollständig darin, manchmal zu zweidrittel. Zur Glättung wird häufig eine Gaußfunktion benutzt. Die Standardabweichung

entspricht dann der Breite der Kurve auf halber Höhe. Die Gaußfunktion geht unter bestimmten Bedingungen in eine Dreiecksfunktion über, ebenso wie die Trapezfunktionen. Der Methode von Gohlke *et al.* (2000) folgend, kann zur Glättung der Daten eine Dreiecksfunktion verwendet werden. Die Einführung einer Trapezfunktion (Nauck *et al.* (1994), Kruse *et al.* (1997)) ist die stringente Fortführung dieses Ansatzes.

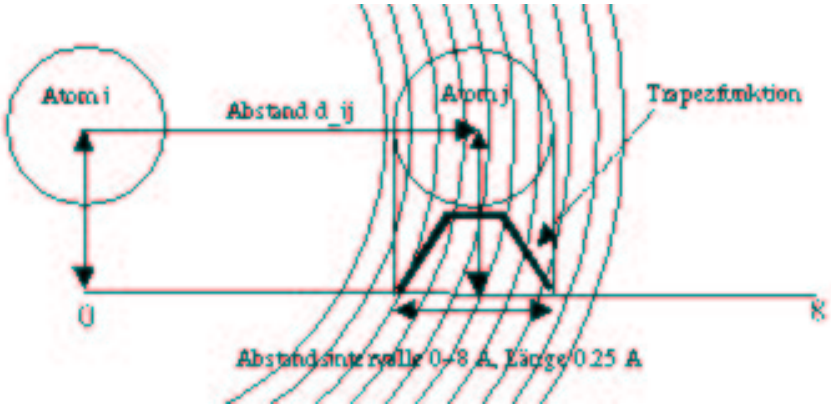


Abbildung 2.2: Veranschaulichung der Trapezfunktion zur Glättung von diskreditierten Atom-Atom-Abständen.

Dreiecksfunktionen sind degenerierte Trapezfunktionen. Die Vorteile der Dreiecksfunktion, der lineare Abfall und die leichte Berechenbarkeit der Funktionswerte über Fallunterscheidungen der Sprungfunktion werden bei Verwendung eines Trapezes beibehalten und gleichzeitig die realen Gegebenheiten besser beschrieben. Das Flächenstück unter dem Trapezmittelpunkt ist im Mittel größer als bei Verwendung eines Dreiecks. Dadurch erhält das Intervall, daß den Mittelpunkt des Trapezes bestimmt ein deutlich größeres Gewicht, wie Abbildung 2.2 zeigt. Das hier verwendete Trapez hat eine Gesamtbreite von 0.75 Å und eine Höhe von zwei bei festgelegter Fläche von eins. Die Breite ist so gewählt, daß bei eine Intervallweite von 0.25 Å bis zu drei Intervalle berücksichtigt werden.
Das Zentrum des Trapezes befindet sich jeweils bei einem Abstand $d_{i,j}$ eines betrachteten Atom-Atom-Kontaktes. Dieser Kontakt wird nun über das umschließende Intervall $[r_i, r_i + dr)$ und die angrenzenden Nachbarintervalle verteilt, wie es dem von den Intervallen überstrichenen Flächeninhalt der Trapezfunktion ent-

spricht. Die Funktionswerte lassen sich über eine Fallunterscheidung berechnen (vergl. auch Abb. 2.3):

$$\mu_A(x) = \begin{cases} 0 & x < a_1 \\ \frac{x-a_1}{a_2-a_1}h & a_1 \leq x \leq a_2 \\ 1 & a_2 \leq x \leq a_3 \\ \frac{x-a_4}{a_4-a_3}h & a_3 \leq x \leq a_4 \\ 0 & x > a_4 \end{cases} \tag{2.25}$$

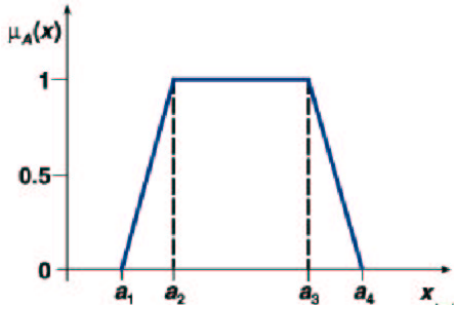


Abbildung 2.3: Darstellung der Trapezfunktion.

Zur Glättung der diskreten Verteilung der Lösungsmittelzugänglichkeiten wird kein so aufwändiges Verfahren benötigt, da das Verhältnis zweier unabhängiger, diskreter Verteilungen betrachtet wird. Des weiteren ist die Intervallänge mit 2 Å großzügig bemessen, ganz im Gegensatz zum engmaschigen Sampling von 0.25 Å bei den Paarpotentialen. Eine Glättung der Verteilungen auf die zwei umliegenden Intervalle ist ausreichend. Jede Beobachtung wird über die beiden angrenzenden Bins verteilt. Das mittige Intervall erhält ein Gewicht von $\frac{1}{2}$, die zwei benachbarten jeweils $\frac{1}{4}$.

2.2.2 Repulsiver Korrekturterm

In künstlich generierten Komplexen werden ungünstige Konformationen erzeugt. Es kann zu *clashes* und *overlaps* kommen. Um den hohen Abstoßungskräften in Distanzen kleiner der Summe zweier Van-der-Waals-Radien Rechnung zu tragen, wird ein Abstoßungsterm eingeführt. Der halbe maximal Wert jeder einzelnen Netto-Präferenz-Verteilung wird als höchster Abstoßungsterm im kleinsten Intervall festgelegt. Dies bedeutet, daß das Maximum ein jeder Verteilung als Ausgangswert für den repulsiven Term im kleinsten Abstandsintervall genommen wird. In den beiden folgen Intervallen wird jeweils $\frac{1}{4}$ und $\frac{1}{8}$ des Maximums addiert, so daß sich ein stark abstoßendes, fast exponentiell abfallendes Potential für Distanzen unter 2.5 Å ergibt.

2.2.3 Gewichtung der Häufigkeitsverteilungen

Generell zwischen der Wahrscheinlichkeitsdichte $\rho_{ij}(r_d)$, wie sie im Boltzmann Gesetz verwendet wird und der radialen Verteilungsfunktion $g_{ij}(r_d)$ unterschieden werden. In Kapitel 2.1.5 wird $\lim_{n->\infty} g_{ij}(r_d) \equiv \rho_{ij}(r_d)$ gesetzt. Diese Zuordnung gilt nur für eine unendliche Zahl an Beobachtungen. Je mehr Daten zur Verfügung stehen, desto präziser gibt die Verteilungsfunktion das betrachtete Merkmal wieder. Bei der Auszählung der Kontakthäufigkeiten von Atom-Atom-Paaren kann sich sehr schnell das Problem geringer Datenmengen ergeben. Dies ist abhängig von der Größe des eingesetzten Strukturdatensatzes, im vorliegenden Fall etwa 600 Protein-Protein-Komplexe. Werden für diese Komplexe etwa 1.200.000 Atome angenommen, so ist die Näherung der Wahrscheinlichkeitsdichte mit der Verteilungsfunktion gerechtfertigt. In einem Einzelfall, bei selten auftretenden Atom-Atom-Kontakten ist die zu erwartende Häufigkeit für das Paar A, B gerade mal $1.200.000/820 \approx 1463$. In diesem Fall ist $g_{ij}(r_d)$ nur eine schwache Näherung für $\rho_{ij}(r_d)$.

Sippl (1990) schlug ein von der Anzahl der Beobachtungen in jeder Verteilung abhängiges Mischen der normalisierten Verteilungsfunktionen für ein spezielles Atompaar ij vor:

$$g'_{ij}(r) = \frac{1}{1 + m_{ij}\sigma} g(r) \frac{m_{ij}\sigma}{1 + m_{ij}\sigma} g_{ij}(r) \tag{2.26}$$

m_{ij} ist die Anzahl aller Kontakte zwischen den Atomen i und j, $g_{ij}(r)$ ist der Anteil dieser Kontakte in dem Abstand r, und $g(r)$ der Referenzzustand. r bezieht sich auf die Länge der Intervalle.
Die von Gohlke *et al.* (2000) weiterentwickelte Form beinhaltet die konstante Größe κ, die eine Reduzierung "lokaler Unsicherheit" einführt:

$$g''_{ij}(r) = \frac{1}{1 + m_{ij}\sigma} (g(r) + \kappa) \frac{m_{ij}\sigma}{1 + m_{ij}\sigma} (g_{ij}(r) + \kappa) \tag{2.27}$$

Abweichend von diesem Formalismus, wird in dieser Arbeit nicht die Anzahl aller Kontakte zwischen zwei Atomen i und j für m_{ij}, verwendet, sondern die die bereits normierte Anzahl dieses Kontaktes benutzt. Die Verwendung der absoluten Häufigkeiten resultiert in unrealistisch großen Netto-Präferenzen und ist daher weniger sinnvoll, als die Benutzung der relativen Häufigkeiten.
Problematisch sind Atom-Atom-Kombinationen AB, die in einem bestimmten Abstand r_d gar nicht vorkommen. In diesem Fall ist die Datenmenge nicht ausreichend, denn es kann *a priori* nicht davon ausgegangen werden, daß die Atome AB bei r_d grundsätzlich keine Wechselwirkungen haben. Die Kombination ist so selten, daß der Strukturdatensatz diesen Fall nicht repräsentiert. Die Boltzmann Statistik ist für Wahrscheinlichkeitsdichten nahe null nicht definiert, geringe Datenmenge erfordern die Definition von Randbedingungen:

$$\Delta E_{ij}(r_d) = \infty : f_{ij}(r_d) = 0 \tag{2.28}$$

$$\Delta E_{ij}(r_d) = kT f_{ij}(r_d) : f(r_d) = 0 \wedge f_{ij}(r_d) \neq 0 \tag{2.29}$$

Um derartige "Nullkontakte" von vornherein zu vermeiden, wird zu jedem Häufigkeitswert eine eins addiert.

2.3 Generierung der Gesamt-Netto-Präferenzen

Das *interface* jedes Komplexes des Strukturdatensatzes wird anhand folgender Definition bestimmt. Als dem *interface* zugehörig gelten Atome, die auf unterschiedlichen Proteinketten lokalisiert sind und einen Abstand r_{ij} mit $r_{min} < r_{ij} \leq r_{max}$ haben. Dabei wird $r_{min} = 2,5$ Å und $r_{max} = 8$ Å gesetzt. Der Abstand zweier Atome r_{ij} wird nach Gleichung 2.8 einem Abstands-Intervall zugeordnet.

Die Intervallbreite dr ist $0,25$ Å . Alle Atompaare mit $r_d > r_{max}$ werden als "nicht-interagierend" angenommen. Alle nicht-kovalent gebundenen Atompaare mit $r_d < r_{min}$ werden als artifizielle sterische Überlappung angesehen.

Die ausgezählten Häufigkeiten werden mittels der Trapezfunktion aus Kapitel 2.2.1 geglättet und nach Gleichung 2.7 in normalisierte radiale Verteilungsfunktionen überführt. Profil 1 wird auf 0.025 gesetzt, das Profil 2 nach Gleichung 2.10 berechnet. Zur Gewichtung der Verteilungen wird Gleichung 2.26 verwendet. Die Netto-Präferenzen werden nach Gleichung 2.15 unter Berücksichtigung des repulsiven Korrekturterms berechnet. Die SAS für jedes Atom im kristallisierten Protein wird mit dem Programm *naccess* (Hubbard & Thornton (1993)) berechnet. Dieses enthält eine Implementierung des Algorithmus von Lee & Richards (1971). Die Oberflächenzugänglichkeit im gebundenen Zustand wird ebenfalls bestimmt. Die Oberflächenzugänglichkeits-Werte von 1-70 Å^2 werden in 36 Intervalle für jeden der 40 Atomtypen unterteilt. Die Intervallweite dS ist 2 Å^2 und $S_{min} = 0$. Die Solvenszugänglichkeit wird nach Gleichung 2.20 bestimmt und deren Häufigkeiten in Abhängigkeit vom Atomtyp ausgezählt (Gleichung 2.22). Die Rohdaten werden wie in Kapitel 2.2.1 beschrieben, geglättet. Gleichung 2.21 überführt die Häufigkeitswerte in normalisierte Verteilungsfunktionen. Die maximale Solvenszugänglichkeit in Å^2 abhängig vom Atomradius wird aus dem Strukturdatensatz extrahiert und als Maximum definiert.

Nach Gleichung 2.23 wird die Oberflächen-Präferenz bestimmt und letztendlich mittels Gleichung 2.24 die Gesamt-Netto-Präferenz.

2.4 Protein-Protein-Docking

Für die Docking-Studien wird die Fourier-Korrelationstechnik benutzt. Die neu implementierte und weiterentwickelte Methode von kordo (Meyer *et al.* (1996)) wird für alle Simulationen verwendet (ckordo, Zimmermann (2002)). Das Verfahren ist für das 1:*n*-Docking entwickelt worden und benötigt keine zusätzlichen biologischen Informationen. Die ungebundenen Proteine werden mit dem kokristallisierten Komplex superpositioniert. Hierfür wird das Programm Bragi Reichelt & Schomburg (1996) verwendet. Wasserstoffatome und Wasser werden nicht berücksichtigt. Das Docking wird mit der nativen Orientierung der ungebunden Proteine, so wie sie im Komplex zu finden wären, begonnen. Dies entspricht nicht der Realität, denn die ungebundenen Moleküle sind "irgendwo" im Lösungsmittel. Das Docking randomisierter Proteinstrukturen ist schwieriger. Da hier die Qualität der Bewertungsfunktion evaluiert werden soll, wird diese Vereinfachung in Kauf genommen.
Der Konformationsraum wird global abgesucht. Dazu wird das kleinere Protein in bestimmten Inkrementen um alle drei kartesischen Winkel um das größere rotiert. In dieser Arbeit werde Inkremente von 15, 20 und 50 Grad verwendet. Für jede Rotation wird der Translationsraum anhand einer Fast-Fourier-Transformation berechnet und hierbei die geometrische Komplementarität über eine Korrelationsfunktion bestimmt. Die fünf besten Translationen jeder Rotation werden ausgegeben.
Es werden keine weiteren Auswahlkriterien wie Elektrostatik oder Desolvatation für diese Arbeit verwendet. Für alle generierten Konformationen wird die Abweichung vom superpositionierten Komplex berechnet. Die Ähnlichkeit zweier Strukturen gleicher Länge wird mit dem *root mean square distances* (rmsd) D bestimmt:

$$D(\mathbf{r}_i\mathbf{r}_i^{'}) = \left(\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_i^{'})^2\right)^{1/2}$$

Hierbei ist zu beachten, daß unterschiedliche rmsd-Werte berechnet werden können, je nachdem, ob nur CA-Atome oder alle Atome zweier Strukturen betrachtet werden. In dieser Arbeit werden die *interface*-rmsd- und Gesamt-rmsd-Werte für alle Atome und nur für die CA-Atome ermittelt. Die Berechnung dieser rmsd-

Werte wird durch die Programme ckordo (Zimmermann (2002)) und BRAGI (Reichelt & Schomburg (1996)) vorgenommen. Für die Docking-Studien werden vorgegebene Standard-Parameter verwendet. Lediglich die Schichtdicke wird auf 1.75 Å und die Gitterweite auf 1.5 Å gesetzt. Für eine detaillierte Erklärung der genauen Docking-Methode und der Bedeutung einzelner Parameter sei auf Zimmermann (2002) verwiesen.

2.4.1 Bewertung

Für das zu bewertende System werden die Häufigkeiten von Atom-Atom-Kontakten und die Verteilung der solvensabhängigen Fläche im ungebundenen Zustand mittels *naccess* bestimmt. Die lösungsmittelzugängliche Fläche im gebundenen Zustand wird näherungsweise auf null gesetzt, um Rechenkapazität zu sparen. Die Häufigkeitsverteilungen werden dafür künstlich erzeugt. Hierzu wird die Summe der Häufigkeiten im ungebundenen Protein auf die ersten Intervalle der Verteilungen vom gebundenen Protein verteilt. Die Funktion ist exponentiell fallend in den angrenzenden Intervallen. Anschaulich entspricht dies der Annahme, daß fast alle *interface*-Atome im komplexierten Zustand vergraben (nicht zugänglich für Lösungsmittel) sind. Einige wenige sind nicht vollständig, sondern nur teilweise vergraben.

Die so bestimmten Häufigkeiten werden nun mit den Gesamt-Netto-Präferenzen aus dem Strukturdatensatz zu "Bindungsenergien" kombiniert. Diese "Bindungsenergien" von Komplexen sollen im Folgenden als Bindungs-Präferenzen bezeichnet werden. Zur Berechnung des Anteils an spezifischen Interaktionen (ohne solvensabhängigen Teil) an der Bindungs-Präferenz wird die Netto-Präferenz eines Atom-Atom-Paares mit der soeben gefundenen Anzahl dieses Paares multipliziert. Dieser Teil des Potentials gibt hauptsächlich spezifische Interaktionen wieder, weil die Atom-Atom-Abstände in kleinen Intervallen bis zu r_{max} gesammelt werden. Der Anteil $E_{spezifisch}$ an der Bindungs-Präferenz E_{gesamt} ergibt sich als Summe über alle Paarkombinationen ij und alle Intervalle r_d:
mit n^{ij} als der Anzahl an Kontakten der Atome i und j im Abstand r_d und $\Delta W_{ij}^{Paar}(r_d)$ als der Netto-Präferenz des Atompaares ij im Abstand r_d.

$$E_{spezifisch} = \sum_{ij} \sum_{r_d} n_{ij}(r_d) * \Delta W_{ij}^{Paar}(r_d) \tag{2.30}$$

Der durch Solvensänderungen bedingte entropische Beitrag zur Bindungs-Präferenz wird nicht durch das Paarpotential repräsentiert. Er ergibt sich analog als Summe über alle Intervalle i und alle Atomtypen x:

$$E_{entrop} = \sum_{i} \sum_{x} \frac{n_x^{gb}}{m_x^{ug}} * \Delta W_x^{einzel}(S_i^{gb}, S_j^{ug}) \tag{2.31}$$

Mit $\Delta W_x^{einzel}(S_i^{gb}, S_j^{ug})$ als der Oberflächen-Präferenz für beide Proteine und $\frac{n_x^{gb}}{m_x^{ug}}$ als dem Verhältnis der Anzahl von Oberflächenzugänglichkeiten des Atoms x vom gebundenen zum ungebundenen Zustand. Das Gesamtpotential setzt sich additiv aus den beiden Termen zusammen.

$$E_{gesamt} = E_{spezifisch} + E_{entrop} \tag{2.32}$$

Für jede aus der Docking-Studie erhaltene Konformation wird nach 2.32 die Bindungs-Präferenz dieser einen Konformation berechnet und mit allen anderen verglichen. Nativ-ähnliche Anordnungen sollten einen möglich negativen Wert aufweisen.
Zur Untersuchung des Einflusses von funktionellen Gruppen, wird der Term E_{gesamt}^{nf} eingeführt. Er ist äquivalent zu Gleichung 2.32, vernachlässigt die Gewichtung von besonderen Atomen nach Gleichung 2.16 sowie Gleichung 2.17 jedoch vollständig. $E_{spezifisch}^{nf}$ ist äquivalent zu E_{gesamt}^{nf} definiert für die hydrophobe Netto-Präferenz.

2.4.2 Kombination der Kriterien

Zur Klärung der Frage ob die geometrische Korrelation und die statistische Präferenz die gleichen Anordnungen äquivalent gut bewerten, werden alle falsch-

positiven Strukturen genau betrachtet. Wenn die beiden Lösungsräume eine gemeinsame Schnittmenge besitzen, so würde eine Kombination dieser und weiterer Kriterien wie elektrostatische und hydrophobe Komplementarität die Unterscheidung falsch-positiver von richtig-positiven erleichtern. Die gefundenen Ränge werden entsprechend der beiden Maßstäbe addiert und erneut numerisch sortiert. Anschließend wird untersucht, ob nativ-ähnliche Anordnungen, die zuvor mäßig gut bewertet wurden durch die Kombination der Kriterien einen höheren Rang erhalten.

2.5 Datensätze

2.5.1 Strukturdatensatz

Größe und Zusammensetzung des Strukturdatensatzes sind entscheidend für die Qualität des Potentials (Shimada *et al.* (2000)). Furuichi & Koehl (1998) zeigte, daß wissensbasierte Potentiale eine Art "Gedächtnis" haben. Sie erinnern sich an die Zusammensetzung des Strukturdatensatzes. Besteht dieser hauptsächlich aus Serin-Proteasen, lassen sich gute Vorhersagen für Serin-Proteasen erzielen, nicht aber für Immunoglobuline.
Der verwendete Datensatz besteht aus 584 nicht redundanten ko-kristallisierten Protein-Protein-Komplexen, extrahiert aus einem von Glaser *et al.* (2001) beschriebenen Datensatz. Dieser umfasst 621 Einträge und stellt eine Untermenge der von I.A. Vakser und A. Sali gesammelten Komplexe dar (http://guitar.rockefeller.edu/sub-pages/combase.html). 37 Komplexe sind für die vorliegende Arbeit ausgemustert worden. Sie zeigten Probleme mit dem Programm *naccess* zur Berechnung der SAS. Die Strukturen sind der *Protein Data Bank* (Bernstein *et al.* (1977), Berman *et al.* (2000)) entnommen. Ketten mit weniger als 30 % Sequenzidentität gehören verschiedenen Familien an. Aus jeder Familie wird ein repräsentativer Komplex ausgewählt. Kriterien sind die Auflösung der Kristallstruktur und eine möglichst große *interface*-Fläche.
Oligomere, Enzym-Inhibitor-Komplexe werden ebenso berücksichtigt wie Membran-Proteine und verschiedene Ketten des selben Komplexes, um eine möglichst große Zahl an Daten zu erzeugen. Es ist zu beachten, daß der Datensatz

sehr viele permanente Komplexe beinhaltet, obwohl das Potential zur Bewertung von *unbound*-Strukturen und damit fast ausschließlich für temporäre Komplexe entwickelt wird. Bei der alleinigen Verwendung von temporären Komplexen, wäre die Datenbasis zu dünn für eine aussagekräftige Statistik. In Tabelle A.3 ist der verwendete Trainings-Datensatz aufgelistet. Die Komplexe sind mit ihrem dreistelligen PDB-Bezeichner und Ketten-Bezeichner angegeben.

2.5.2 Dockingdatensatz

Einundzwanzig Komplexe werden für reine *unbound* Docking-Studien verwendet. Die Strukturen der ungebundenen Proteine und der ko-kristallisierten Komplexe sind bekannt und werden ebenfalls der *Protein Data Bank* entnommen. In Tabelle 3.4 im Ergebnisteil sind die verwendeten Daten aufgelistet. Elf der Komplexe gehören zu den Proteasen. Von dieser Klasse gibt es mit Abstand die meisten Strukturen. Außerdem sind Endonuklease-, Glykolase-, Dehydrognese- und Hydrolase- Strukturen vertreten. Zwei Immunoglobulin-Komplexe werden ausgewählt. Sie sind wegen des kleinen *interfaces* schwierig zu docken. Sämtliche Untersuchungen bezüglich Immunoglobulin-Komplexen zeigten mäßige Ergebnisse, in den meisten Fällen wurde gar keine native Konformation anhand geometrischer Korrelation gefunden (vergl. Kapitel 3.7).

2.5.3 Decoydatensatz

Zur Untersuchung der diskriminatorischen Kraft des Potentials werden frei verfügbare so genannte *decoy*-Strukturen verwendet. Darunter wird das Verdrehen oder Verschieben eines Komplexpartners gegen den anderen verstanden. Das Ausmaß der artifiziell herbeigeführten Störung ist bestimmbar und kann in definierten Inkrementen variiert werden. Der rmsd zum nativen Komplex dient als Maß für die Störung. Der verwendete Datensatz ungeordneter Komplexstrukturen ist aus zwei verschiedenen frei zugänglichen Sets zusammengestellt. Chymotrypsin/Inhibitor (1cgi), Glykolase/Inhibitor (1ugh), Kalikrein/Inhibitor (2kai), Acetylcholinesterase/Fascicullin II (1fss) und Subtilisin BPN/Inhibitor (2sic) stammen aus dem Test-Systemen von FTDock

(http://www.bmm.icnet.uk/docking/systems.html). Jeder Set besteht aus einhundert Strukturen. Eine davon entspricht dem ko-kristallisierten Komplex. Drei weitere sind sehr ähnliche Modelle, die aus Docking-Studien mit einem *interface*-Cα-rmsd kleiner 7 Å von der experimentell bestimmten Struktur erhalten werden. Diese sehr guten Modelle beinhalten mindestens 25% der nativen Aminosäurekontakte. Die restlichen 96 Strukturen sind verdrehte oder verschobene Anordnungen aus den Docking-Simulationen mit Abweichungen von 9 bis 41 Å. Chymotrypsion/OMTKY (1cho) gehört zu einem von Vakser (1996b) entwickelten Datensatz (http://engpub1.bu.edu/bioinfo/MERL/databases.html). Dieser Set besteht auch aus einhundert Strukturen mit einem nativen, fünf nativähnlichen und vielen falsch-positiven Anordnungen. Diese falschen Komplexe stellen die fünf besten Lösungen einer *unbound*-Docking-Studie mit dem Programm GRAMM (Vakser & Aflalo (1994)) dar. Die native Struktur entspricht der Superposition der einzeln kristallisierten Proteine mit dem ko-kristallisierten Komplex. Eine detaillierte Beschreibung und Analyse ist von Camacho *et al.* (2000) vorgenommen worden.

2.6 Statistische Analyse

2.6.1 Korrelationsanalyse

Als Maß für den Zusammenhang zweier metrischer Merkmale dient der Korrelationskoeffizient von Bravais und Bravais-Pearson (Helge Toutenburg (1998)), der die Abstände zwischen den Beobachtungen zweier Merkmale und deren arithmetischem Mittel zueinander in Beziehung setzt. Für zwei Paarverteilungen X und Y ergibt sich die Korrelation r_{corr}:

$$r_{corr} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} \tag{2.33}$$

mit dem arithmetischen Mittel $\overline{x} = \frac{\sum x_i}{n}$
Für $\overline{y}$ gilt entsprechendes, n ist die Anzahl der Beobachtungen.

Der Korrelationskoeffizient ist ein dimensionsloses Maß, das beide Merkmale symmetrisch behandelt, d. h. es gilt $r_{corr}(X,Y) = r_{corr}(Y,X)$. r_{corr} liegt in den Grenzen (-1,1), ist $r_{corr} \pm 1$, liegt ein exakt linearer Zusammenhang vor.
Der Nachteil dieser Methode ist, daß r_{corr} nur die Form, nicht aber die Größe von Verteilungen vergleicht. Zwei Verteilungen mit ähnlicher Form, aber vollständig anderen Werten, würden in diesem Fall einen hohen r_{corr}-Wert erhalten. Das ist nicht gewollt, denn beim Aufsummieren für die Bindungs-Präferenzen über alle Verteilungen macht die Größe der Werte einen signifikanten Unterschied. Unter diesem Aspekt macht es Sinn, Verteilungen auch hinsichtlich ihrer Größe zu vergleichen. Von Hodgin und Richards (1987) wurde ein Maß vorgeschlagen, daß den Nachteil umgeht (Hodgkin & Richards (1987)). Für den Korrelationskoeffizienten im vorliegenden Fall ergibt sich:

$$r_{corr} = \frac{2\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2 + \sum (y_i - \overline{y})^2} \tag{2.34}$$

Die Korrelation der Verteilung der Netto-Präferenzen innerhalb des Strukturdatensatzes wird nach Gleichung 2.34 bestimmt.

2.7 Umsetzung und Laufzeitverhalten

Die beschriebene Methode ist für zeitkritische Routinen in C/C++, sowie in Python für die statistische Auswertung realisiert. Das externe Programm *naccess* (Hubbard & Thornton (1993)) zur Berechnung der *solvent accessible surface* wird eingebunden. Alle weiteren verwendeten programmiertechnischen Hilfsmittel und Programme sind im Anhang aufgelistet. Das Hauptprogramm zur Berechnung aller Abstände, der Bestimmung des *interfaces* und der Glättung aller Daten braucht pro Komplex ohne *naccess* im Mittel etwa 0.2 s auf einem Linux Rechner mit 1700MHz AMD-Prozessor. Das in Python realisierte Auswertungsprogramm zur Berechnung der Präferenzen braucht pro Komplex etwa 0.5 s auf dem selben Computer. Jeder Aufruf von *naccess* kostet in Abhängigkeit von der Größe des Proteins etwa 1s.

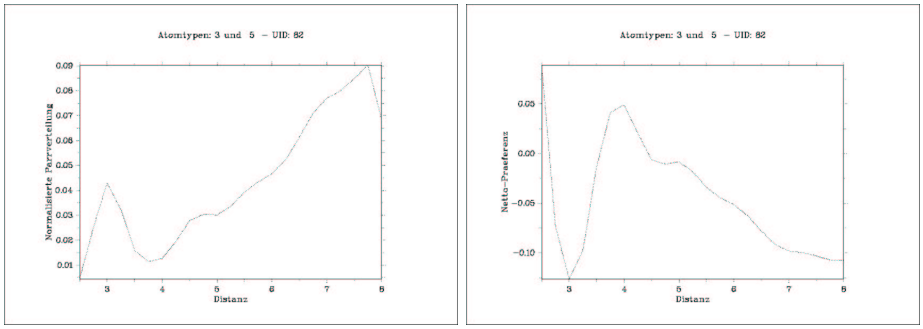
3 Ergebnisse

"... the structure of water is less well understood
than the structure of proteins ..."

Wesson & Eisenberg (1992)

3.1 Radiale Verteilungsfunktionen

Abb. 3.1 (a) gibt die normalisierte radiale Paarverteilungsfunktion exemplarisch für den Atom-Atom-Kontakt UID=82 (Carbonylsauerstoff mit Hauptketten-Stickstoff) und die daraus generierte Netto-Präferenz gemäß Gleichung 2.6 wieder. Erwartungsgemäß nimmt die Häufigkeit von Atom-Atom-Kontakten mit steigendem Abstand zu. Bei Entfernungen unter 3 Å sind nur wenige Sauerstoff- und Stickstoffatome zu finden, bei etwa 3 Å sind es entsprechend der möglichen Ausbildung von Wasserstoffbrücken wesentlich mehr.



(a) normalisierte Paarverteilung **(b)** Netto-Präferenz

Abbildung 3.1: UID = 82 Carbonylsauerstoff mit Hauptketten-Stickstoff a) Normalisierte Paarverteilung, b) Netto-Präferenz. x-Achse: Distanzen in Å .

Im weiteren Verlauf wird bei etwa 4 Å ein Minimum durchlaufen und anschlie-

ßend ein Anstieg der Anzahl von Kontakten gefunden, da die Wahrscheinlichkeit für diese Atom-Atom-Interaktion mit der Entfernung steigt, insofern keine spezifischen Wechselwirkungen vorliegen.
Die daraus erzeugte "Energie" oder Netto-Präferenz spiegelt dieses Verhalten deutlich wieder.
Die ersten 400 Kombinationsmöglichkeiten von Atomen (UID 1 - UID 400, vergl. Abb. 3.2) sind deutlich stärker bevölkert als die Kombinationen von UID 401 - UID 820. Unter diesen ersten 400 befinden sich die ausgesprochen häufig vorkommenden Atome der Hauptkette, sowie Methyl- und Methylen-Gruppen (vergleiche Abbildung 2.1). In den Klassen mit UIDs größer 400 sind viele Atome und Kombinationen von selteneren Aminosäuren enthalten. Die Besetzungshäufigkeiten sind dementsprechend geringer, wie in Abb. 3.2 zu sehen ist. Weniger als 50% der 820 Klassen sind mit mehr als 10.000 Beobachtungen vertreten.

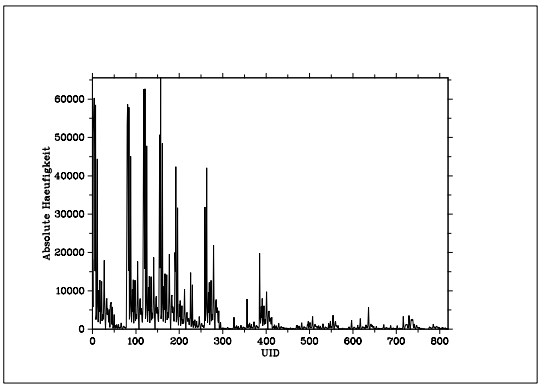


Abbildung 3.2: Absolute Häufigkeiten (y-Achse) in Abhängigkeit von der Art des Atom-Atom-Kontaktes (UID) (x-Achse).

3.1.1 Der Referenzzustand

Profil 2 wird nach Gl. 2.10 berechnet und ergibt sich als gemittelte, normalisierte Paarverteilungsfunktion, summiert über alle Atom-Atom-Kombinationsmöglichkeiten. Dadurch wird eine möglichst zufällige Verteilung erhalten, die in Abbildung 3.3 dargestellt ist. Kurvenverlauf und Größe der Werte für $g(r_d)$ stimmen gut mit denen von Gohlke *et al.* (2000) erzeugten Referenzwerten überein.

Das alternativ hierzu getestete Profil 1 entspricht einer “Gleichverteilung” mit dem Wert 0.025.

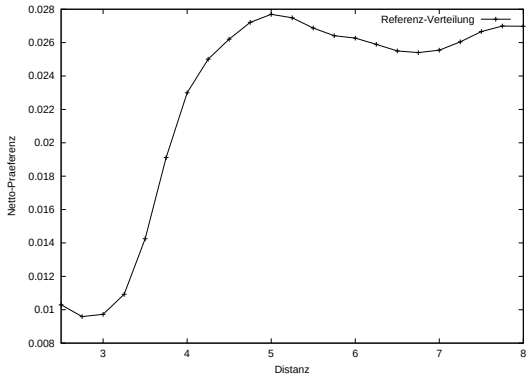
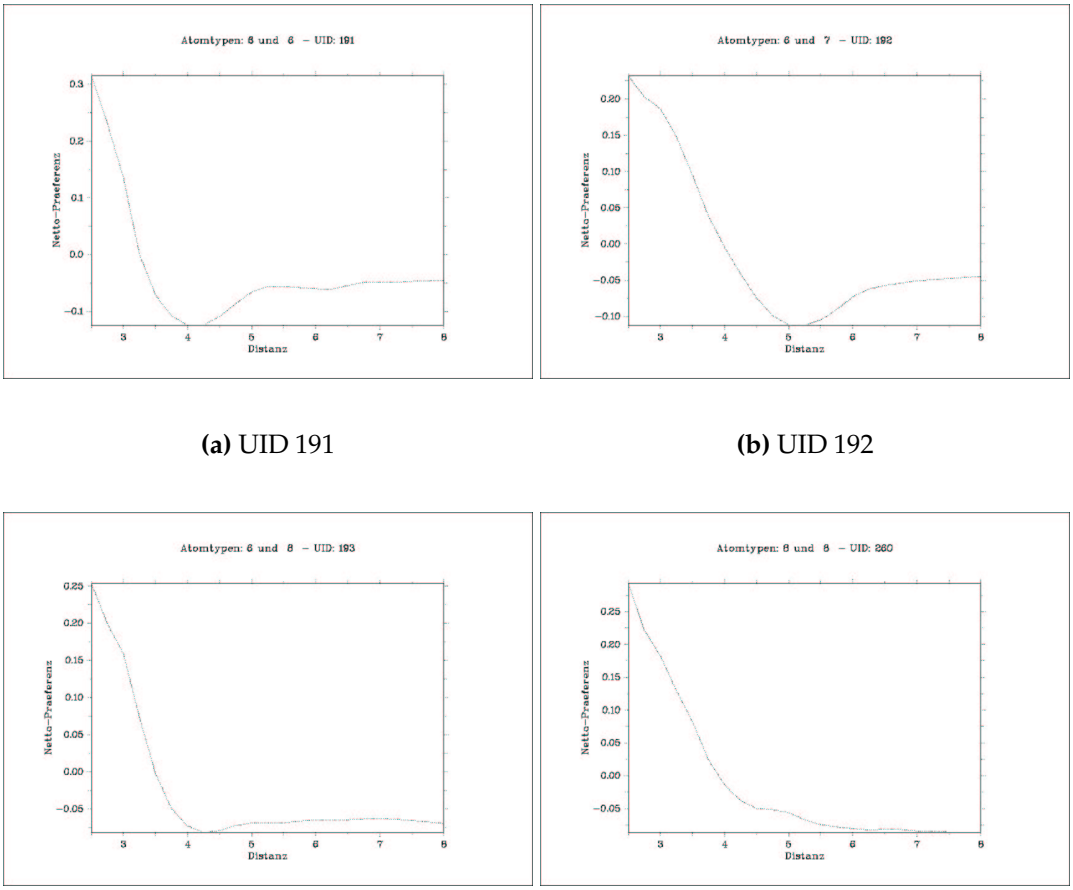


Abbildung 3.3: Profil 2 als Mittelung über die normalisierten Verteilungsfunktionen. x-Achse: Distanzen in Å , y-Achse: Netto-Präferenz.

3.2 Distanzabhängige Paarpotentiale

Im Folgenden werden einige exemplarische Netto-Präferenzen für 820 mögliche vorgestellt. Wird nur Profil 2 betrachtet, wobei die Kurvenverläufe der Paarpotentiale bei Verwendung des Profils 1 denen bei Benutzung von Profil 2 tendenziell entsprechen, läßt sich eine leichte Verschiebung der Extrema beobachten, was an einem Beispiel in der Diskussion erläutert wird.
Der Verlauf für hydrophobe Kontakte in Abb. 3.4 ist im allgemeinen flach und ohne signifikante Extrema. Bei geringen Abständen zeigen sich tendenziell positive Netto-Präferenzen, die bis etwa 4-5 Å abnehmen, um dann gegen -0.05, zu konvergieren. Dieser Wert ergibt sich aus der Gewichtung der Rohdaten.
Hydrophobe Kontakte sind demnach ab einer Distanz von ca. 4 Å günstig (kleiner Null), allerdings auch weniger strukturiert. Das entspricht der Tatsache, daß diese Wechselwirkungen nahezu keiner geometrischer Beschränkung unterliegen. Diese Aussage trifft jedoch nicht auf aromatische Systeme zu. Die Interaktionen von Benzyl-, Phenyl-, Imidazol- und Indol-Derivaten, wie in den zwanzig Standard-Aminosäuren enthalten, ebenso wie die der Stickstoff-Basen der DNA sind stark richtungsabhängig. Eine “Stapelung” wie aus der DNA bekannt, be-

günstigt π-π-Wechselwirkungen, jede andere Konformation schwächt sie ab. Die hier aufgetragenen Verteilungen aliphatischer Atome zeigen Minima über 4 Å. Dies deutet daraufhin, daß nicht nur direkte Wechselwirkungen zwischen zwei Atomen sondern auch mit nächsten Nachbarn involviert sind.



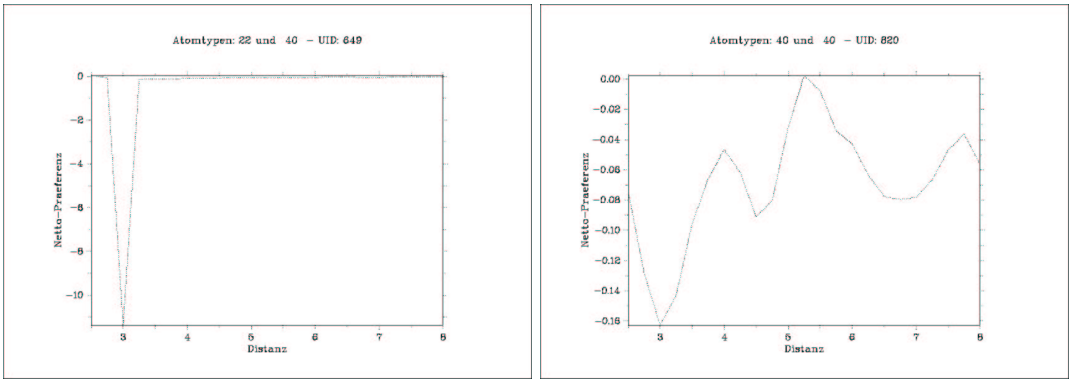
(a) UID 191 **(b)** UID 192

(c) UID 193 **(d)** UID 260

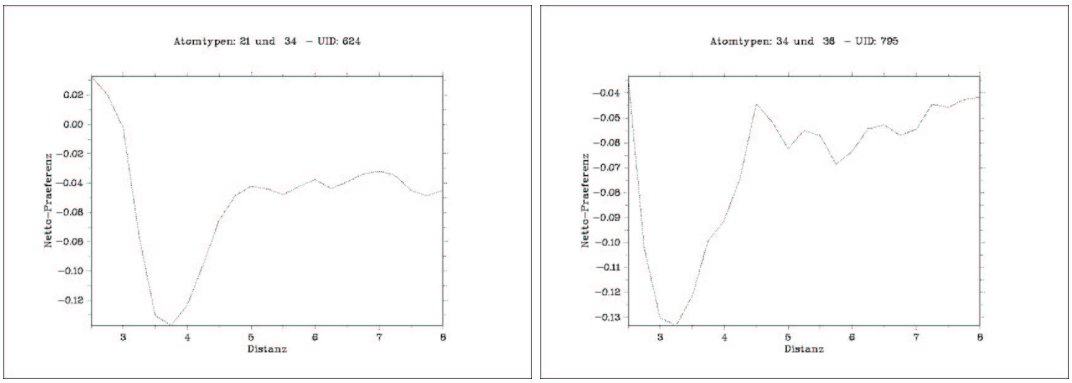
Abbildung 3.4: Netto-Präferenzen für hydrophobe Kontakte. (a) CB/CG/RCD (VAL, LEU, ILE, THR u.a.) mit der gleichen Art. (b) CB/CG/CD (VAL, LEU, THR u.a.) mit CB/CG (VAL, LEU, ILE). (c) CB/CG/CD (VAL, LEU, ILE, THR u.a.) mit CB/CG (MET, ILE, PHE, ASN, TYR u.a.), (d) CB/CG (MET, ILE,PHE, ASN, TYR u.a.) mit der selben Art. Ordinate: Distanzen in Å , Abszisse: Netto-Präferenzen.

Kurvenverlauf und Lage der Minima zeigen gute Übereinstimmungen mit den

von Mitchell *et al.* (1999b) ermittelten Daten.
In Abb. 3.5 sind die Netto-Präferenzen für einige polare Kontakte angegeben.



(a) UID 649 **(b)** UID 820



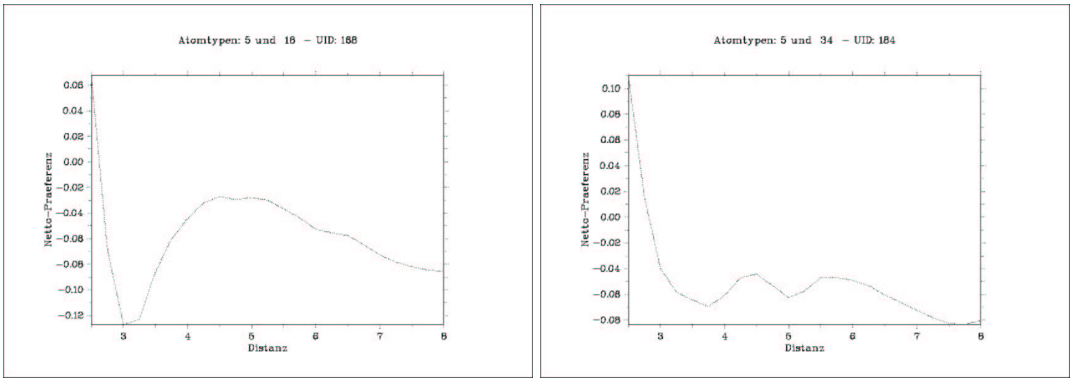
(c) UID 624 **(d)** UID 795

Abbildung 3.5: Netto-Präferenzen für polare Kontakte. (a) O (TYR) mit N aus Amidgruppe (ARG). (b) O (TYR) mit dem selben Typ. (c) C der Amidgruppe (ARG) mit O (GLN). (d) N (ARG) mit O (GLN). Ordinate: Distanzen in Å . Abszisse: Netto-Präferenzen.

Die Kurven sind gekennzeichnet durch tiefe Minima zwischen 3-3.7 Å. Hydroxyl-Sauerstoff (TYR) und Amid-Stickstoff (ARG) als Wasserstoffbrückenbildner weisen den eindeutigsten Extremwert auf. Die Verteilung scheint nur aus wenigen Punkten zu bestehen. In Abb. 3.5 (b) und (d) sind zusätzliche Minima zwischen

5-6 Å erkennbar, die nach Ben-Naim (1992) die zweite Koordinationssphäre des jeweiligen Interaktionspartners um das betrachtete Zentralatom wiedergeben und charakteristisch für kondensierte Systeme sind.

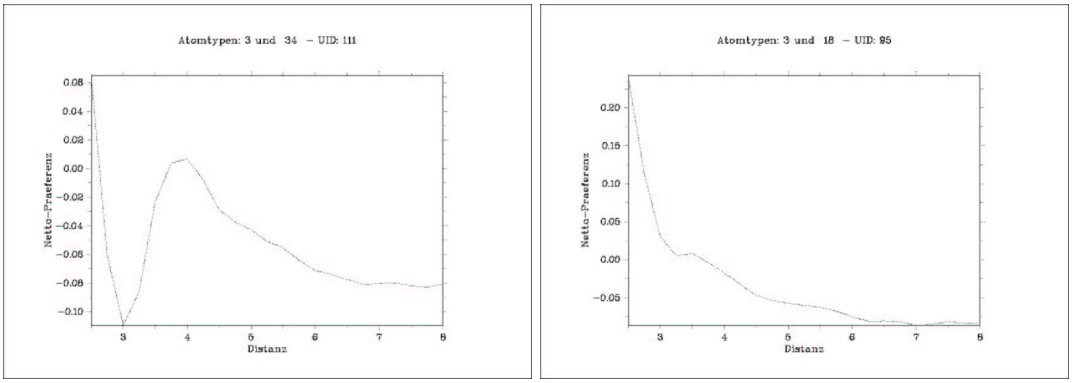
Der Gleichgewichtsabstand zwischen Amid-Stickstoff und Carbonyl-Sauerstoff beträgt etwa 2.9 Å (Sippl (1996)). Die Netto-Präferenzen für Peptid-Stickstoff mit Sauerstoff (GLN) und Peptid-Sauerstoff mit Amid-Stickstoff (GLN) sind in den Abbildungen 3.6 (a) und 3.7 (a) aufgetragen. Sie zeigen vergleichbar tiefe Minima bei 3 Å. Die Ausbildung von Wasserstoffbrücken ist sehr wahrscheinlich.



(a) UID 168 **(b)** UID 184

Abbildung 3.6: Netto-Präferenzen für (a) Peptid O mit N (GLN) und (b) Peptid O mit O (GLN). Ordinate: Distanzen in Å , Abszisse: Netto-Präferenzen.

Die Kurvenverläufe zeigen gute Übereinstimmung mit den von Sippl (1996) gefundenen. Wesentlich weniger stark ausgeprägt sind die Minima bei der Betrachtung von Peptid-Sauerstoff mit Sauerstoff von Glutamin (Abb. 3.6) und Peptid-Stickstoff mit Stickstoff von Glutamin (Abb. 3.7). Eine erkennbare Schulter liegt zwischen 3-3.5 Å und deutet wieder auf die begünstigte Ausbildung von Wasserstoffbrücken hin. Im Anhang (vergl. A.1) sind die Interaktionen von NZ (LYS) mit Atomen von Aspartat (CB, CG und O) dargestellt.



(a) UID 111 **(b)** UID 95

Abbildung 3.7: Netto-Präferenzen für (a) Peptid N mit O (GLN) und (b) Peptid N mit N (GLN). Ordinate: Distanzen in Å , Abszisse: Netto-Präferenzen.

Die NZ-CB Interaktion ist hydrophober Natur, zeigt einen relativ flachen Verlauf und ein schwaches Minima bei $r_{min} = 5$ Å. Der NZ-CG-Kontakt ist deutlich ausgeprägter mit $r_{min} = 3.5$ Å. Die starke Anziehung zwischen NZ-O aufgrund der Fähigkeit zur Ausbildung von Wasserstoffbrücken wird bei $r_{min} = 3$ Å deutlich. Die Bindung von NZ zu Lysin ist demnach stark attraktiv.

Die Ergebnisse sind in guter Übereinstimmung mit den Werten von DeBolt & Skolnick (1996), ebenso wie die Untersuchung von Wechselwirkungen von CD (LEU) mit Atomen von Lysin (CB/CG/CD, CE und NZ) (Abb. A.2 im Anhang). Interaktionen von CD (LEU) mit Lysin sind weniger begünstigt. Die Kurvenverläufe sind flach und pendeln um -0.05, dem nach unten skalierten Nullwert. Mehrere Minima deuten auf die bevorzugte Häufung von vergrabenen und exponierten CD-Lysin-Kontakten hin.

Die hier aus Platzgründen nicht vorgestellten Verteilungen von Netto-Präferenzen sind analog den hier exemplarisch diskutierten sinnvoll zu deuten und zeigen darüberhinaus gute Übereinstimmung mit der Literatur-bekannten Ergebnissen (DeBolt & Skolnick (1996), Sippl *et al.* (1996), Gohlke *et al.* (2000),

Mitchell *et al.* (1999b)).

3.3 Korrelationsanalyse

Die Korrelationswerte r_{corr} aller Netto-Präferenzen zueinander werden mit Gleichung 2.34 berechnet. In Abb. 3.8 sind die Netto-Präferenzen für einige Arten von CA-Kontakten abgebildet und in Tabelle 3.1 die entsprechenden Korrelationswerte angegeben. Für jede Verteilung werden 22 Datenpunkte zugrunde gelegt. Besonders die hydrophoberen Wechselwirkungen von CA-CA, CA-CB und CA-C weisen hohe Korrelation und damit große Ähnlichkeit auf. Deutlich abweichend verhält sich die stärker polare Wechselwirkung von CA und Carbonylsauerstoff ($r_{corr} = 0.6$).

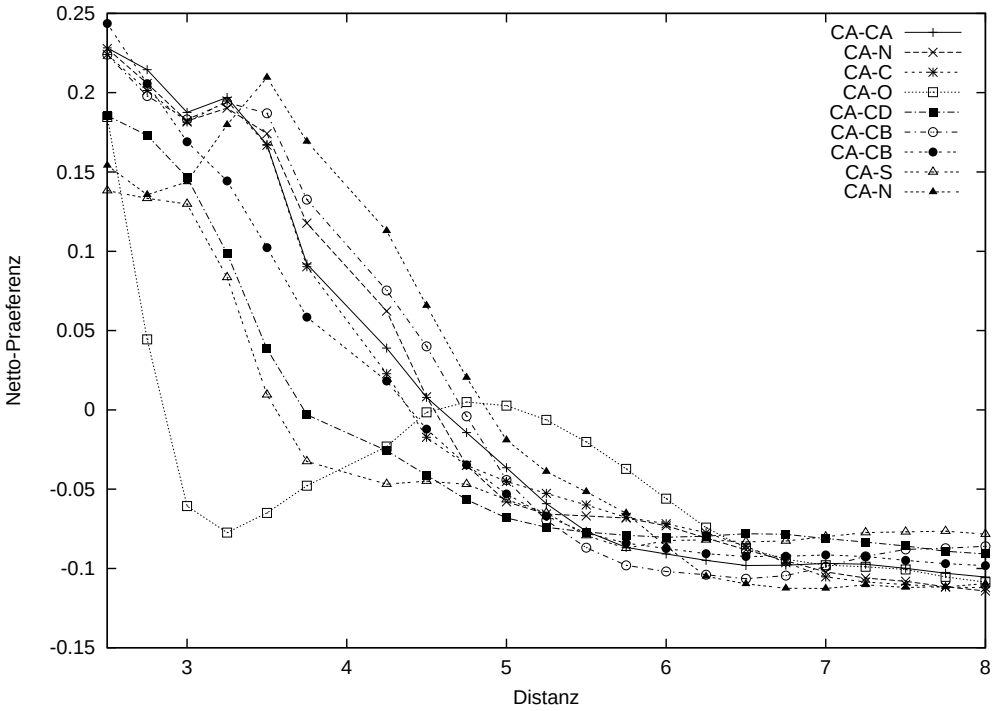
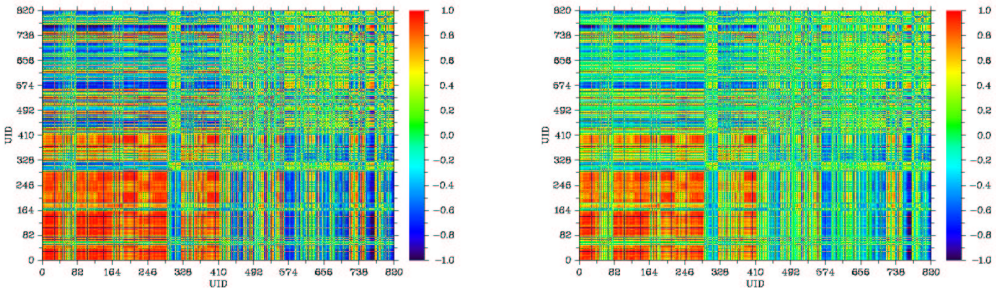


Abbildung 3.8: Netto-Präferenzen für Kontakte von CA mit: CA, Peptid-N, Carbonyl-C, Carbonyl-O, CD, verschiedene CB, Schwefel und N (PRO). Ordinate: Distanzen in Å, Abszisse: Netto-Präferenzen.

CA-CA zu	r_{corr}
CA-CA	1.000
CA-N (Peptid)	0.995
CA-C (C=O)	0.996
CA-O (C=O)	0.622
CA-CD	0.935
CA-CB	0.992
CA-CB	0.988
CA-S	0.866
CA-N(PRO)	0.947

Tabelle 3.1: Ähnlichkeiten von neun Verteilungen der Netto-Präferenzen mit der von CA-CA. r_{corr} wird nach Gl. 2.34 berechnet.

Eine Untersuchung des Korrelationskoeffizienten r_{corr} für alle Verteilungen ist in Abb. 3.9 gezeigt. Das Ausmaß der Ähnlichkeit nach Pearson (Gl. 2.33) ist deutlich höher als bei einer Berechnung von r_{corr} nach Hodgkin. Insbesondere hydrophobe Wechselwirkungen führen zu einem ähnlichen Verlauf der Netto-Präferenzen, wie im vorangegangenen Abschnitt erklärt. Die Größe der Absolutwerte kann aber sehr unterschiedlich sein.



(a) nach Pearson **(b)** nach Hodgkin

Abbildung 3.9: Korrelationen der Netto-Präferenzen. Abszisse und Ordinate: UIDs aller 820 möglicher Interaktionen. r_{corr} farbkodiert von -1 bis +1. Die Korrelationsmatrix ist symmetrisch um die Diagonale. (a) und (b) sind mit Profil 1 bestimmt.

Ein Vergleich der Korrelationen der Netto-Präferenzen bei Verwendung der beiden Profile zeigt deutlich größere Ähnlichkeiten für Profil 2, wie in Tabelle 3.2 dargestellt.

	Profil 1		Profil 2	
	Hodgkin	Pearson	Hodgkin	Pearson
$r_{corr} > 0.9$	4 %	10 %	19 %	35 %
$r_{corr} > 0.5$	38 %	57 %	76 %	85 %

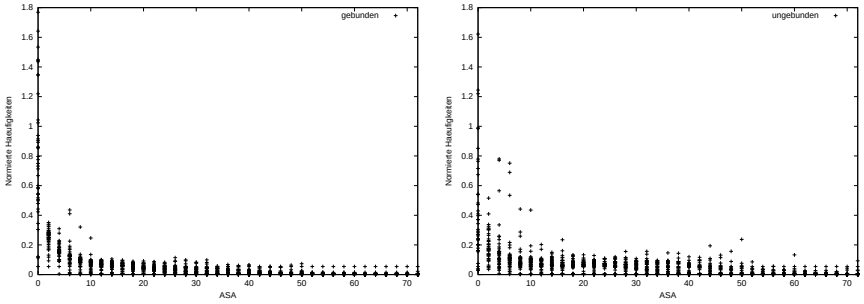
Tabelle 3.2: Korrelation der Netto-Präferenzen in den beiden Profilen mit unterschiedlichen Maßstäben berechnet.

Die Verwendung eines Maßes, daß die Größe der Werte berücksichtigt ist nach Abbildung 3.9 und Tabelle 3.2 sinnvoller. Die Verteilungen sind weniger stark korreliert, eine Unterscheidung von verschiedenen Interaktionen einfacher. Als Fazit ist aus dieser Korrelationsanalyse zu schließen, daß die Verteilungsfunktionen bei Benutzung von Profil 1 "mehr Informationen" beinhalten, da sie weniger ähnlich sind, als bei Profil 2.

3.4 Von der lösungsmittelzugänglichen Fläche abhängige Ein-Teilchen-Potentiale

Atome der Kontaktfläche mit mehr als 50 Å ^{2} sind selten, wie in Abb. 3.10 zu erkennen ist. Die auch im unkomplexierten Zustand stark vertretenen Bereiche um 10 Å ^{2} erklären sich aus der *interface*-Definition. Wie eingangs beschrieben wird das *interface* definiert als aus den Atomen bestehende Fläche, die einen bestimmten r_{max} voneinander auf verschiedenen Proteinketten einnehmen und nicht durch die Differenz der lösungsmittelzugänglichen Flächen von Komplex und einzelnen Proteine. Dadurch ergibt sich zwangsläufig, daß nicht nur Oberflächenatome auf dem anderen Protein betrachtet werden, sondern auch die erste innere Schicht. Das wiederum führt dazu, daß selbst im ungebundenen Zustand

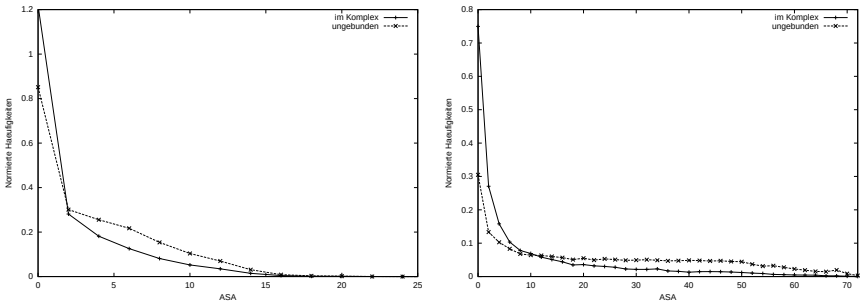
eine relativ hohe Besetzungszahl für kleine Lösungsmittelzugänglichkeiten erhalten werden (Abbildung 3.10).



(a) gebunden **(b)** ungebunden

Abbildung 3.10: Häufigkeitsverteilung der lösungsmittelzugänglichen Fläche im (a) komplexierten und (b) ungebundenen Zustand für alle vierzig Atomtypen. Ordinate: Fläche in Å ^{2}, Abszisse: normierte Häufigkeiten.

Im Folgenden werden einige exemplarische Häufigkeitsverteilungen der Solvenszugänglichkeit für 40 der möglichen erläutert.

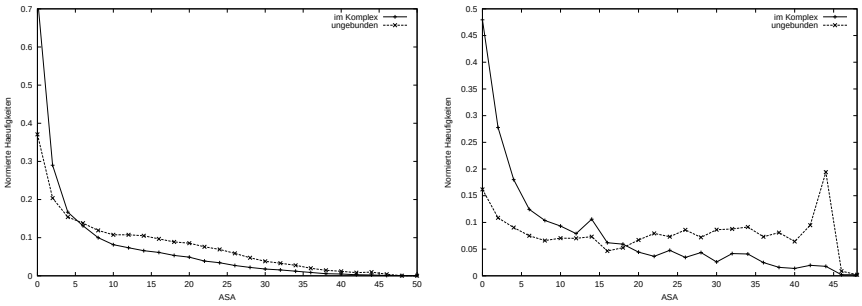


(a) CA

(b) aliphatisches C (ALA, VAL, LEU, ILE u.a)

Abbildung 3.11: Häufigkeitsverteilung der lösungsmittelzugänglichen Fläche für den ungebunden (-x-) und gebundenen Zustand (-+-). (a) Atomtyp 1. (b) Atomtyp 6. Ordinate: Fläche in Å ^{2}, Abszisse: normierte Häufigkeiten.

Atome von relativ hydrophoben Aminosäuren wie verschiedene aliphatische Kohlenstoffe zeigen flache Kurven. Solvenszugänglichkeit im komplexierten und freien Zustand unterscheiden sich marginal.

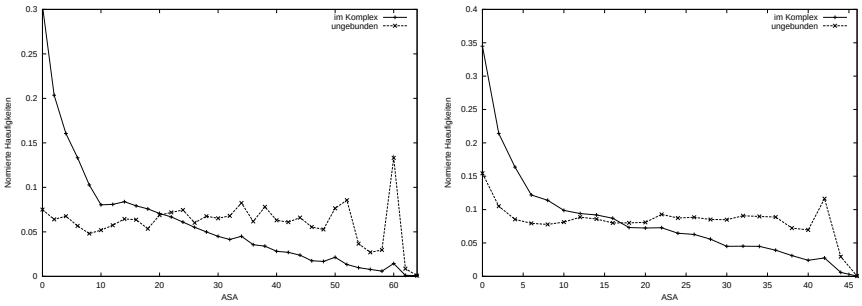


(a) aliphatisches C (LYS, MET, ARG, ASN u.a)

(b) O (TYR)

Abbildung 3.12: Häufigkeitsverteilung der lösungsmittelzugänglichen Fläche für den ungebunden (-x-) und gebundenen (-+-) Zustand. (a) Atomtyp 8. (b) Atomtyp 40. Ordinate: Fläche in Å ^{2}, Abszisse: normierte Häufigkeiten.

Die Kurven schneiden sich zwischen fünf und zehn Å ^{2} (vergl. Abb. 3.11 (a), (b) und 3.12 (a)).



(a) N (ARG)

(b) O (GLU, ASP)

Abbildung 3.13: Häufigkeitsverteilung der lösungsmittelzugänglichen Fläche für den ungebunden (-x-) und gebundenen (-+-) Zustand. (a) Atomtyp 22. (b) Atomtyp 28. Ordinate: Fläche in Å ^{2}, Abszisse: normierte Häufigkeiten.

Abbildung 3.12 (b) und 3.13 (a), (b) zeigen SAS-abhängige Verteilungsfunktionen für einige polare Atome.
Für Atome des Typs O (TYR, GLU, ASP) und N (ARG) ist der Zustand weitgehendster Lösungsmittelexposition im freien Zustand deutlich häufiger als für hydrophobe Atome. Im komplexgebundenen Fall werden SAS-Werte zwischen 0 und 30 Å ^{2} gefunden. Die vollständige Vergrabung des Sauerstoffes vom Tyrosin im Komplex beruht wahrscheinlich auf der Bildung von Wasserstoffbrücken im *interface*. Die ebenfalls zahlreich vorhandenen mittleren SAS-Werte für Sauerstoff (TYR) im proteingebundenen Zustand ergeben sich aus "teilweiser" Vergrabung. Eine vollständige Vergrabung in der Bindestelle ohne die Ausbildung einer Wasserstoffbrücke ist möglich, aber energetisch "teuer".

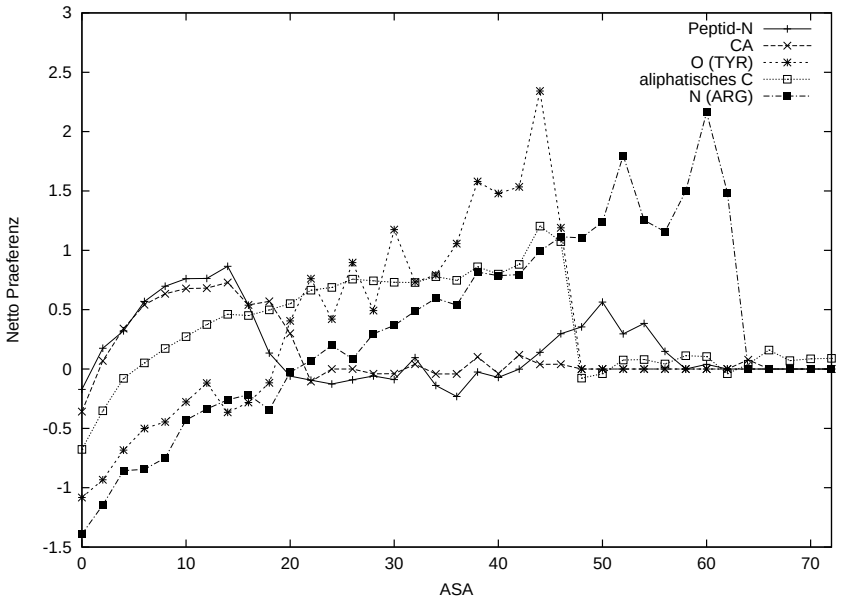


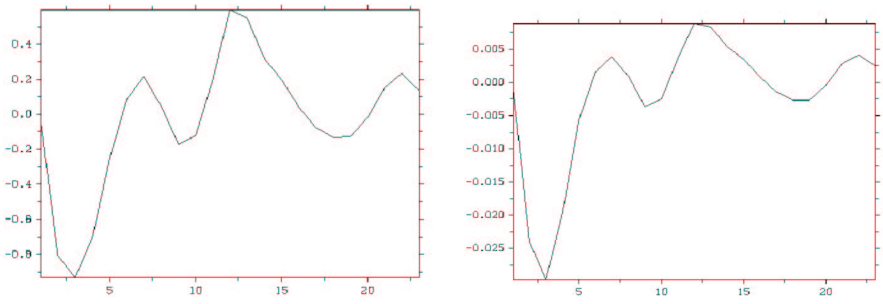
Abbildung 3.14: Oberflächen-Präferenzen für Atomtypen 1, 3, 8, 22, 40. Ordinate: Fläche in Å ^{2} skaliert auf die maximal zugängliche Fläche des jeweiligen Atomtyps (vergl. Anhang A.1), Abszisse: Netto-Präferenzen.

Es ist wahrscheinlich, daß ein Großteil Hydroxylgruppen in entsprechenden Bindungen zu finden ist oder eine Position vom *interface*-Inneren weg einnehmen wird. Abb. 3.14 zeigt einige aus den Häufigkeitsverteilungen gewonnenen Oberflächen-Präferenzen. Allgemein ist eine Vergrabung zu stark begünstigt

(negative Werte in den Oberflächen-Präferenzen), was auf die Einbeziehung der ersten inneren Atomschicht des anderen Proteins beruht. Tendenziell zeigen aliphatische Atome eine begünstigte Vergrabung und einen neutralen Verlauf für Exposition größerer Bereiche. Polare Atome zeigen teilweise Vergrabung. Eine große SAS im proteingebundenen Zustand ist ungünstig, da Wechselwirkungen z.B. mit Carboxylatgruppen für die Protein-Protein-Bindung bevorzugt sind.

3.5 Gewichtung

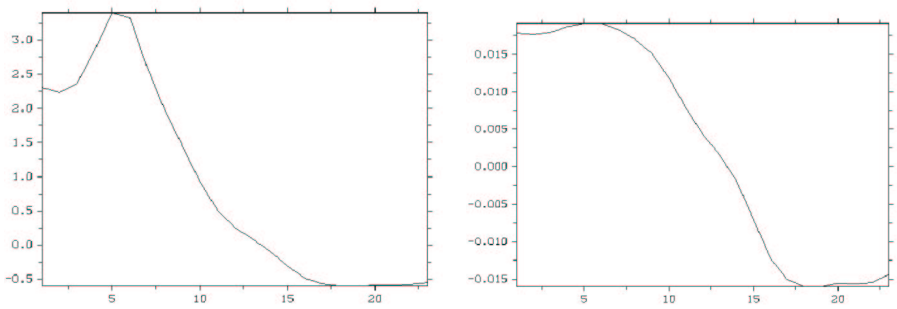
Die Abbildungen 3.15 bis 3.17 zeigen die Netto-Präferenzen zweier Atom-Atom-Kontakte im ungewichteten Zustand, gewichteten Zustand und mit repulsivem Korrekturterm. Für die ungewichteten Verteilungen werden repulsive Kräfte bei Abständen kleiner 2.5 Å wenig berücksichtigt. Der Kontakt von C_α mit Stickstoff der Hauptkette (UID 10) zeigt ohne Gewichtung keinerlei Abstoßung. Nach Einführung der Gewichtung und der damit geringeren Gefahr der Überbewertung von Einzelereignissen ist auch eine Abstoßung bei kleinen Distanzen erkennbar, welche aber erst nach Verwendung des entsprechenden Korrekturterms realistische Größenordnungen zeigt, was sich im Verhältnis der Funktionswerte bei y=0 zu x=0 zeigt.



(a) UID 820 ungewichtet **(b)** UID 820 gewichtet

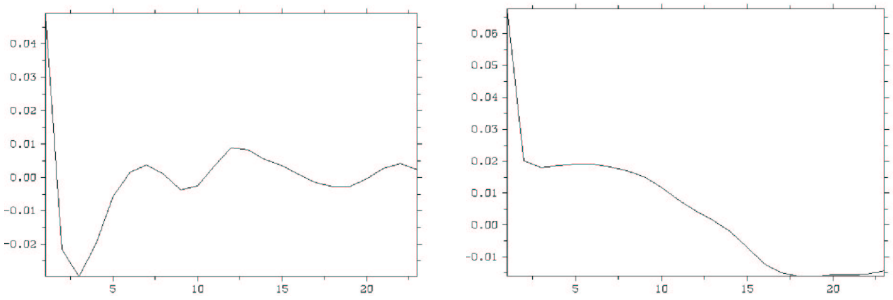
Abbildung 3.15: Statistische Netto-Präferenzen, gewichtet sind (a), ungewichtet (b). UID = 820 (Atomtyp 40 mit 40).

Dieses Verhältnis ist für den ungewichteten Zustand 0.025, im gewichteten 0.4 und nach Einführung des repulsiven Terms 0.5 (vergl. Abbildung 3.15, 3.16 und 3.17). Die eingeführte Gewichtung verändert den allgemeinen Verlauf der Verteilungsfunktionen nur geringfügig.



(a) UID 10 ungewichtet **(b)** UID 10 gewichtet

Abbildung 3.16: Statistische Netto-Präferenzen, gewichtet sind (a), ungewichtet (b). UID=10 (Atomtyp 1 mit 10).



(a) UID 820 gewichtet mit repulsivem Anteil

(b) UID 10 gewichtet mit repulsivem Anteil

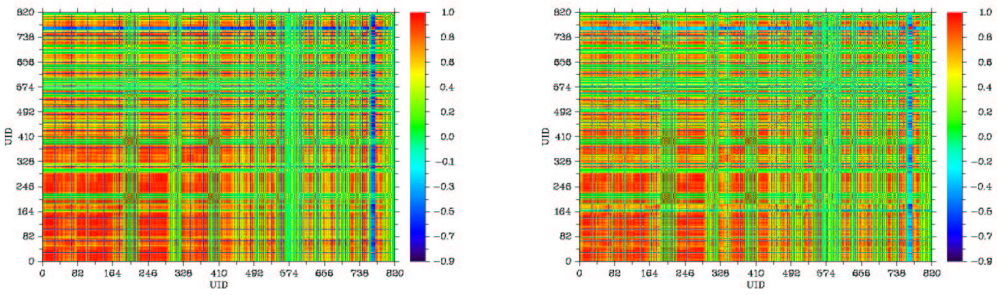
Abbildung 3.17: Statistische Netto-Präferenzen, gewichtet sind (a), ungewichtet (b), UID = 820 und UID = 10, beide mit repulsivem Korrekturterm.

Minima werden deutlicher herausgearbeitet und repulsive Kräfte bei kleinen

Abständen berücksichtigt. Verschiedene σ- und κ-Werte werden hinsichtlich der Bewertungen von Docking-Studien getestet. Ein kleiner Trainingsdatendatz, extrahiert aus dem im Folgenden Kapitel dargestellten Dockingdatensatz diente zum iterativen Optimieren dieser Werte. Für σ erweist sich ein Wert von 0.02 am sinnvollsten, für κ einer von 0.001. Dies entspricht weitgehend den Erkenntnissen von Gohlke *et al.* (2000). Weitere hier nicht gezeigte Untersuchungen von verschiedenen Gewichtungsarten brachten keine einheitliche Verbesserung und wurden daher nicht weiter verfolgt.

3.6 Hauptkomponentenanalyse

Eine Hauptkomponentenanalyse (PCA) wird durchgeführt, um Abhängigkeiten von Daten zu untersuchen. Abb. 3.18 zeigt einerseits in (a) r_{corr} für die Netto-Präferenzen (nach Profil 1) und andererseits in (b) r_{corr} für dieselben Präferenzen, die jedoch nach einer PCA und Multiplikation mit der Faktor-Matrix (auf der Heyde (1990)) unabhängig voneinander sind.



(a) Statistische Präferenzen

(b) Aus Faktormatrix generierte Präferenzen

Abbildung 3.18: Korrelationen nach Hodgkin. (a) Korrelationen der Präferenzen für Profil 1 (b) Korrelationen der Präferenzen unter Benutzung der Faktormatrix aus der PCA.

Es wird deutlich, daß die Korrelationen zwischen den Verteilungen erwartungsgemäß abnehmen. Vorhersagen für einige Docking-Systeme dieser Arbeit haben jedoch gezeigt, daß die Bewertung bei Verwendung der Faktor-Matrix anstatt der Netto-Präferenzen signifikant schlechter werden, obwohl bei einer PCA alle Ausgangsdaten zu 100% repräsentiert sein sollten. Im vorliegenden Fall kommt es dennoch zum Verlust von Daten, da bei der Berechnung der Faktormatrix Vorzeicheninformationen verloren werden (auf der Heyde (1990)). Für die sinnvolle Verwendung der Bindungs-Präferenzen sind deren Vorzeichen jedoch von zentraler Bedeutung. Deren Verlust führt zu einer deutlichen Veränderung und/oder Verschlechterung der Vorhersage. Eine weitere Fehlerquelle sind die extrem kleinen Zahlen in der Größenordnung von 10^{-20}, die zu deutlichen Rundungsfehlern führen.
Die Frage nach einer Reduktion der Dimensionalität konnte nicht abschließend geklärt werden. Tabelle 3.3 zeigt einige Eigenwerte des nach auf der Heyde (1990) berechneten Eigenwertsystems und deren Varianz bezüglich der Summe aller Eigenwerte. Es zeigt sich, daß bereits ein einziger Eigenwert und der dazugehörige Eigenvektor über 60% der Ausgangsdaten repräsentiert. Die ersten vier Eigenwerte decken sogar über 90% ab.

Eigenwert	Varianz %
0.001733246497	0.9
0.002399456694	1.2
0.003026322309	1.5
0.005300276229	2.7
0.010862319988	5.5
0.019154883995	9.6
0.028330486608	14.3
0.122671794966	61.7

Tabelle 3.3: Angegeben sind einige Eigenwerte und die entsrechende Varianz dieser Werte bezogen auf alle Eigenwerte.

Aufgrund der komplexen Wechselbeziehungen der Potential-Parameter können einige dieser Variablen nicht einfach weggelassen oder als Linearkombination von anderen dargestellt werden. Eine genauere Betrachtung dieser nicht-trivialen Verknüpfungen soll in der Diskussion erfolgen.

3.7 Untersuchung einiger *unbound*-Docking-Systeme

Die in Tabelle 3.4 beschriebenen einundzwanzig Systeme werden in einem *unbound*-Docking mit dem Fourier-Korrelations-Program ckordo (Zimmermann (2002), nur geometrische Korrelation) untersucht. Für alle Fälle sind die dreidimensionalen Strukturen der einzelnen Proteine sowie des ko-kristallisierten Komplexes bekannt. Einige der Komplexe oder der Proteine haben eine Auflösung größer 2.5 Å. Die Elektronenverteilungen der Strukturen sind demnach nicht vollständig lokalisierbar und derartige Proteine sollten aussortiert werden. Für das *unbound*-Docking stehen jedoch, wie eingangs erklärt, nur etwa dreißig Systeme insgesamt zur Verfügung, so daß kein Auswahlspielraum hinsichtlich Auflösung und anderer Eigenschaften besteht.
Sechzehn von diesen einundzwanzig Komplexen werden den Enzym-Inhibitor-Systemen zugeschrieben. Diese Protein-Klasse ist sehr gut untersucht und mit zahlreichen Strukturen in der pdb vertreten. Im verwendeten Datensatz sind hauptsächlich Proteasen (Chymotrypsin, Subtilisin, Trypsin, Kallikrein, Elastase), eine Glykolase, eine Dehydrogenase und das Barnase/Barstar-System (Endonuklease) vertreten. Des weiteren sind zwei unterschiedliche Acetycholinesterasen und eine Peroxidase vertreten. Für die Klasse der Antikörper sind zwei Fab-Fragment/Lysozym-Systeme ausgewählt worden. Insgesamt sind über 50% aller untersuchten Fälle Proteasen. Die beiden Antikörper/Antigen-Komplexe werden im Folgenden vernachlässigt, da das Docking-Programm ckordo in keiner Simulation eine einzige "plausible" Lösung findet. Als nativ-ähnlich oder "plausibel" werden Protein-Protein-Anordnungen bezeichnet, die einen Gesamt-rmsd kleiner als 3 Å vom superpositionierten Komplex haben. Alle Konformationen mit einem Gesamt-rmsd größer 3 Å werden als falsch angesehen. Als sehr gute Lösung werden Strukturen bezeichnet, deren Gesamt-rmsd weniger als 2 Å beträgt. Von dieser Definition kann in Einzelfällen abgesehen werden, wenn der rmsd des *interfaces* kleiner als 3 Å ist. Das würde bedeuten, daß die Bindestelle eine hohe Ähnlichkeit zum superpositionierten Komplex hat, der Rest des oder der Proteine jedoch stärker abweicht. In einem derartigen Fall ist der *interface*-rmsd ausschlaggebend, weil er die Übereinstimmung der Struktur in der Bindestelle angibt. Welcher rmsd verwendet werden sollte, ist *a priori* nicht bekannt.

Beschreibung	PDB	Res.	PDB	Res.	rmsd	PDB	Res.	rmsd
Enzym/Inhibitor								
$a-$Chymoptrypsin/ Ovomuciod OMTKY	1cho	1.8	5cha	1.67	0.62[1]	2ovo	1.5	
$a-$Chymotrypsin/ pankreatischer Trypsin Inhibitor	1cgi	2.3	1chg	2.5	1.53[1]	1hpt	2.3	
Barnase/ Barstar	1brs	2.0	1a2p 1bao	1.5 2.2	0.464 0.424	1a19 1bta	2.8 NMR	0.466 0.881
Subtilisin BPN/ Inhibitor	2sic	1.8	1sup 2st1	1.6 1.8	0.246 0.314	3ssi 3ssi	2.3 2.3	0.542 0.542
Subtilisin Novo/ Inhibitor	2sni	2.1	1sbc 1sup	2.8 1.6	0.582 0.582	2ci2 2ci2	2.0 2.0	0.459 0.459
Kalikrein A/ pankreatischer Trypsin Inhibitor	2kai	2.5	2pka	2.1	0.542	4pti	1.5	0.510
β-Trypsin/ pankreatischer Trypsin Inhibitor	2ptc	1.9	2ptn	1.5	0.335	5pti	1.0	0.397
Elastase/ Ovomucoid OMTKY	1ppf	1.8	1ppg	2.3	0.342	2ovo	1.5	0.499
Trypsin/ APPI	1brc	2.5	1bra	2.2	0.403	1aap	1.5	0.426
Uracil-DNA Glykolase/ Inhibitor	1ugh	1.9	1akz	1.57	0.389	1ugi	1.55	0.571
Trypsin/ pankreatischer Trypsin Inhibitor	1brb	2.1	1bra	2.2	0.361	1bpi	1.1	0.290
Hydrolase/ Inhibitor	1bvn	2.5	1pif	2.3	0.422	2ait	NMR	0.837
Dehydrogenase/ Inhibitor	1mda	2.5	2bbk	1.75	0.529	1aan	2.0	0.887
Andere								
Acetylcholinesterase/ FasciculinII	1fss	3.0	2ace	2.5	0.76[1]	1fsc	1.9	
Peroxidase/ Cytochrome C	2pcc	2.3	1ccp	2.2	0.39[1]	1ycc	1.23	
Acetylcholinesterase/ FasciculinII	1mah	3.2	1maa	2.9	0.60[1]	1fsc	1.9	
Antikörper-Antigen								
Fab Fragment/ Lysozym	mlc	2.1	1mlb	2.1	0.953	1lza	1.6	0.627
Fv Fragment Lysozym	vfb	1.8	1vfa	1.8	0.325	1lza	1.6	1.066

Tabelle 3.4: Verwendete Komplexe für die *unbound*-Docking-Studien. Angegeben ist die pdb-Kennung für: den ko-kristallisierten Komplex, das ungebundene größere Protein, das ungebundene kleinere Protein. Res. ist die Auflösung in Å aus den entsprechenden Einträgen in der pdb. Alle rmsd-Werte in Å der superpositionierten Strukturen werden mit BRAGI (Reichelt & Schomburg (1996)) bestimmt. [1] ist aus Chen & Weng (2002) entnommen.

3.7.1 Analyse der Bindestellen

Tabelle 3.5 stellt einige statistische Erhebungen der Komplexe dar. Die *buried surface*-Werte sind in guter Übereinstimmung mit den Berechnungen verschiedener Arbeitsgruppen. Für einige Komplexe sind keine Einträge in der Literatur gefunden worden, so daß ein Vergleich nicht möglich ist. Ebenfalls angegeben ist die Anzahl der Atome der beiden Proteine. Einige der Komplexe bestehen aus sehr großen Komponenten, wie z.B. die beiden Acetylcholinesterasen 1mah und 1fss.

PDB	Atome R	Atome L	*bur. surface*	*bur. surface*	# H-Brücken
1cho	1735	418	1466	1420[1]	10
1cgi	1800	440	2052	1993[3]	8
1brs	846	696	1555	1560[2]	13
2sic	1939	765	1616	1620[2]	10
2sni	1934	521	1627	1601[1]	10
2kai	1799	453	1421	1406[1]	10
2ptc	1629	464	1429	1400[1]	12
1ppf	1637	419	1324	1320[2]	6
1brc	1644	412	1316		10
1ugh	1808	649	2192		12
1brb	1651	391	1398		9
1bvn	3907	537	2221		11
1mda	2583	790	2765		7
1fss	4226	464	1966	1970[2]	7
2pcc	2371	847	1140	1140[2]	1
1mah	4116	461	2145		10

Tabelle 3.5: Anzahl der Atome beider Proteine (Spalte zwei und drei) in den gedockten Komplexen (Spalte eins), *buried surface* in $Å^2$, aus dieser Arbeit (Spalte vier) und aus der Literatur (Spalte fünf), sowie Anzahl der Wasserstoffbrücken, bestimmt mit dem Programm HBPLUS (McDonald & Thornton (1994)) (Spalte sechs). [1]Wallqvist *et al.* (1995), [2]Chakrabarti & Janin (2002), [3]Krämer (2001) [4]Conte *et al.* (1999).

Die mit dem Programm HBPLUS (McDonald & Thornton (1994)) bestimmte Anzahl an Wasserstoffbrücken ist ebenfalls angegeben. In guter Übereinstimmung mit den Ergebnissen von Conte *et al.* (1999) werden im Mittel etwa neun Wasserstoffbrücken in der Bindestelle pro Komplex gefunden. Tabelle 3.6 zeigt darüberhinaus die Anzahl an gefundenen *interface*-Atomen. Diese Zahl differiert

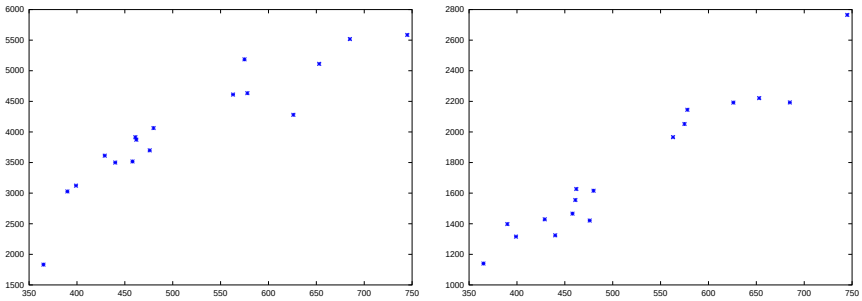
stark von der daneben angegeben Anzahl aus der Literatur. Der Grund hierfür könnte in der verwendeten Definition der Bindestelle liegen. Nach dieser werden nicht nur die tatsächlich an der Oberfläche liegenden Atome berücksichtigt, sondern auch solche, die in der ersten inneren Schale unter der Oberfläche liegen. Es ist zu erwarten, daß dann die Anzahl an "*interface*-Atomen" wesentlich höher ist, als wenn lediglich Atome an der Oberfläche betrachtet würden. Es ist davon auszugehen, daß über die Hälfte der als Bindestelle zugehörig definierten Atome nicht an der Oberfläche liegen. Das vorgestellte Potential reicht demnach ins Innere des Proteins hinein und berücksichtigt mehr Wechselwirkungen (Tabelle 3.6).

Komplex	# Atome	# *int*.-Atome	# *int*.-Atome[1]	# Kontakte	*buried surface*	$\frac{interface-Fl.}{\#Atome}$	$\frac{\#interface-Atome}{\#Atome}$
1cho	2153	458	166	3518	1466	0.7	0.2
1cgi	2240	575		5187	2052	0.9	0.3
1brs	1542	461	177	3915	1555	1.0	0.3
2sic	2704	480	180	4064	1616	0.6	0.2
2sni	2455	462	181	3873	1627	0.7	0.2
2kai	2252	476	150	3699	1421	0.6	0.2
2ptc	2093	429	167	3613	1429	0.7	0.2
1ppf	2056	440	151	3500	1324	0.6	0.2
1brc	2056	399		3123	1316	0.6	0.2
1ugh	2457	626		4280	2192	0.9	0.3
1brb	2042	390		3028	1398	0.7	0.2
1bvn	4444	653		5113	2221	0.5	0.1
1mda	3373	745		5585	2765	0.8	0.2
1fss	4971	563	220	4611	1966	0.4	0.1
2pcc	3218	365	106	1832	1140	0.4	0.1
1mah	4577	578		4634	2145	0.5	0.1

Tabelle 3.6: Angegeben sind von links nach rechts der Komplex mit pdb-Kürzel, die Anzahl der Atome des ganzen Komplexes, die Anzahl der *interface*-Atome, die Anzahl der *interface*-Atome aus der Literatur, die Anzahl an gefundenen Atom-Atom-Kontakten in der Bindestelle , die *buried surface* in $Å^2$, das Verhältnis von *interface*-Fläche zur Anzahl aller Atome und das Verhältnis der Anzahl an *interface*-Atomen zur Anzahl aller Atome.[1] entnommen aus Chakrabarti & Janin (2002).

Eine weitere logische Konsequenz der größeren Anzahl an *interface*-Atomen ist auch eine höhere Anzahl an Atom-Atom-Kontakten in der Bindestelle. Sie beträgt im Durchschnitt 4064 $\pm$ 980 Kontakte bei einer mittleren Anzahl von 516 $\pm$

111 *interface*-Atomen. Des weiteren ist in Tabelle 3.6 das Verhältnis von *interface*-Fläche bzw. von der Anzahl an *interface*-Atomen zur Gesamtanzahl an Atomen angegeben. Abbildung 3.19 zeigt eine Auftragung dieser Werte. In beiden Fällen ist, wie erwartet, eine lineare Korrelationen zu erkennen. Anschaulich bedeutet dies, daß die Anzahl an *interface*-Atomen mit der Anzahl der gefundenen Atom-Atom-Kontakte, sowie mit der vergrabenen Fläche korreliert.



(a) **(b)**

Abbildung 3.19: Auftragung von Anzahl der *interface*-Atome (x-Achse) zur Anzahl an gefundenen Atom-Atom-Kontakten (y-Achse) (a) und zur *interface*-Fläche in $Å^2$ (y-Achse) (b), vergl. Tabelle 3.6.

3.7.2 Ergebnisse der Docking-Simulationen

Tabelle 3.7 und 3.8 zeigen die Bewertungsergebnisse für einundzwanzig Protein-Protein-Docking-Studien. Für jedes System werden maximal 22006 Anordnungen erhalten, weil der Raum mit einer Schrittweite von 15 Grad abgesucht und aus jeder Rotation die besten fünf Translationen behalten werden. Die dargestellte Reihenfolge der Komplexe, sowie die Gesamtzahl der erhaltenen Lösungen ist von der Gradzahl des Samplings abhängig. Wird der Raum mit einem Winkel von zwanzig Grad abgesucht und nur eine Translation pro Rotation ausgegeben, ist die Anzahl an Möglichkeiten wesentlich geringer, als wenn der Raum mit fünfzehn Grad abgesucht und fünf Translationen pro Rotation behalten würden.

Komplex	Protein	Ligand	rmsd int.	rmsd int. C α	rmsd CA	rmsd	Rang geom.	Rang für E^{nf}_{gesamt}		Rang für E_{gesamt}	
								Profil 1	Profil 2	Profil 1	Profil 2
Enzym/Inhibitor											
1cho	5cha	2ovo	1,308	1,371	1,089	1,097	3239	601	628	**153**	716
			1,119	1,352	1,205	1,249	781	2897	2850	3540	3193
			1,215	1,384	1,639	1,649	989	505	718	185	819
			1,523	1,655	1,652	1,654	4757	1269	1529	473	1715
			2,060	2,183	1,752	1,775	2685	850	869	10321	736
1cgi	1chg	1hpt	2,434	2,759	1,881	2,03	10797	1530	1809	**227**	2013
			3,459	3,797	2,980	3,126	11965	5768	5930	2396	6336
1brs	1a2p	1a19	2,184	2,164	2,151	2,147	8000	2300	1920	3667	1982
			2,645	2,550	2,183	2,187	4265	9469	9637	1985	9986
			3,190	3,053	2,961	2,978	5435	3821	3175	**962**	3537
			3,442	3,267	3,317	3,340	6265	3494	2718	**528**	3032
1brs	1bao	1bta	2,095	2,446	2,42	2,671	19576	1609	20556	2333	19621
			2,119	2,365	2,204	2,696	13281	2211	13396	**832**	14032
			1,923	2,176	2,499	3,001	3152	3885	7767	1033	8578
2sic	2st1	3ssi	1,072	1,146	1,082	1,088	6030	2687	3047	1646	3156
			2,166	2,315	2,518	2,484	8882	7101	7361	3969	7531
			1,158	1,335	2,762	2,765	2817	751	1104	**196**	1273
			2,275	2,708	3,370	3,421	879	2171	2795	5237	2755
2sni	1sbc	2ci2	2,828	2,802	2,709	2,873	8858	13104	14344	8968	14787
			2,070	2,507	2,843	3,025	18773	2550	2792	12082	2626
			1,959	2,434	2,924	3,107	8406	4368	5044	12787	4862
			2,184	2,762	3,718	3,961	15870	1595	1818	**1311**	2099
2kai	2pka	4pti	3,203	3,360	3,362	2,409	15996	10254	10784	16360	10815
			2,868	3,049	2,349	2,441	13344	1888	1988	**597**	2449
			3,163	3,382	3,358	2,484	3260	3119	3218	8323	4359
			3,254	3,421	2,675	2,707	6137	7175	7072	5351	7899
			3,049	3,300	3,493	3,529	1269	1068	1188	**323**	1571
			2,313	2,634	4,043	4,033	16531	1439	1224	471	1605
2ptc	2ptn	5pti	0,676	0,908	0,644	0,849	790	133	134	149	188
			1,281	2,065	1,699	2,334	2817	480	654	208	786
			1,053	1,632	1,981	2,75	4271	87	116	**31**	157
1ppf	1ppg	2ovo	0,524	0,563	0,466	0,465	7188	846	1235	355	1481
			0,991	1,264	1,545	1,526	11784	58	94	**54**	116
			2,611	3,147	3,333	3,324	19159	6226	7959	6795	7290
1brc	1bra	1aap	4,278	4,278	2,729	2,771	9813	9301	9071	6169	9290
			4,273	4,388	3,175	3,279	1319	9585	9473	6939	9643
			2,378	2,499	3,578	3,675	3468	4718	4608	3705	4821
			2,520	2,680	4,474	4,589	7499	1745	1517	6176	1503
			2,987	3,465	5,804	6,040	5216	2180	1679	8110	1668
1ugh	1akz	1ugi	0,796	0,793	0,596	0,580	7465	123	113	6847	94
			2,055	2,128	1,544	1,573	**1**	8	8	16	10
			2,027	2,040	1,577	1,595	3570	74	**1**	190	**1**
			1,772	1,823	1,600	1,628	320	**1**	75	35	73
			2,809	2,884	2,191	2,207	6104	106	117	966	109
			3,029	3,017	2,232	2,208	1994	153	115	52	127
1brb	1bra	1bpi	0,536	0,560	0,499	0,511	131	**74**	114	**75**	115
			0,855	0,983	1,588	1,643	2819	146	225	101	290
			0,886	1,039	1,638	1,688	1671	165	269	304	315
			1,511	1,652	1,882	1,944	7454	695	798	1252	779
			1,387	1,606	1,801	1,944	12339	1378	1773	1393	1307
1bvn	1pif	2ait	2,750	3,578	1,767	2,415	2881	13992	12463	10090	13447
			3,034	3,943	1,854	2,556	18301	13768	12226	12658	12579
			3,458	4,554	2,121	2,908	9542	17901	17100	15313	17334
			3,502	4,612	2,143	2,971	13148	17617	17094	14865	17344
			3,321	4,267	2,240	3,056	12982	15760	1044	15714	865

Tabelle 3.7: Ergebnisse von 13 *unbound*-Docking-Studien. Profil 1 entspricht $g_{ij} = 0.025$ mit $\alpha = 30$, $\beta = 1$, $\gamma = 100$ und $\delta = 1$. Profil 2 entspricht Gleichung 2.10 mit $\alpha = 0.1$, $\beta = 0.9$, $\gamma = 90$ und $\delta = 0.1$. rmsd-werte in Å : *interface* (int.), CA-Atome des *interfaces*, alle Atome und nur CA-Atome.

Für eine detaillierte Beschreibung der Zusammenhänge von Schrittweite der Rotation und der daraus resultierenden Anzahl an möglichen Komplexen sei auf Meyer *et al.* (1996), Katchalski-Katzir *et al.* (1992) und für genaue Angaben zum verwendeten Programm ckordo auf (Zimmermann (2002)) verwiesen.
In den ersten drei Spalten beider Tabellen sind die pdb-Bezeichner der einzeln kristallisierten Proteine und des ko-kristallisierten Komplexes aufgelistet. In den nächsten vier Spalten sind verschiedene rmsd-Werte angegeben und in den letzten fünf die Ränge der jeweiligen nativ-ähnlichen Struktur, bewertet mit verschiedenen Methoden (geometrische Korrelation, Bindungs-Präferenzen mit und ohne die Verwendung funktioneller Gruppen, zwei verschiedene Profile) angegeben. Die Anzahl der nativ-ähnlichen Strukturen wird auf die ersten fünf beschränkt, obwohl für einige Systeme (Trypsin/Inhibitor (1brb), Acetylcholinesterase/Fascicullin II (1mah, 1fss), u.a.) mehr als zehn Anordnungen mit einem Gesamt-rmsd kleiner als 3 Å gefunden werden.

Komplex	Protein	Ligand	rmsd int.	rmsd int. C α	rmsd CA	rmsd	Rang ckordo	Rang für E^{nf}_{gesamt}		Rang für E_{gesamt}	
								Profil 1	Profil 2	Profil 1	Profil 2
Andere											
1fss	2ace	1fsc	1,518	1,502	0,838	0,825	7718	1049	1030	846	1226
			1,412	1,453	1,020	1,010	2362	340	337	128	425
			3,443	1,537	1,153	1,139	18179	18	16	**9**	23
			4,434	1,514	1,185	1,159	15455	1075	866	343	1053
			3,644	1,563	1,262	1,210	853	755	734	202	888
2pcc	1ccp	1ycc	3,824	4,351	2,251	2,251	20517	21339	18328	18758	19448
			2,352	2,523	2,715	2,715	16661	21359	20485	21058	20342
			3,824	3,885	2,968	2,94	10676	13428	9592	17909	13109
1mah	1maa	1fsc	0,365	0,374	0,205	0,203	4573	71	74	18	101
			1,423	1,450	1,033	0,993	15797	351	321	81	408
			1,593	1,633	1,055	1,033	1228	56	54	**9**	73
			2,066	2,192	1,103	1,125	12813	5833	6285	3112	6819
			1,539	1,628	1,191	1,154	843	304	315	93	396
2sni	1up	2ci2	keine native Anordnung gefunden								
1mda	2bbk	1aan	keine native Anordnung gefunden								
1mlc	1mlb	1lza	keine native Anordnung gefunden								
1vfb	1vfa	1lza	keine native Anordnung gefunden								
2sic	1sup	3ssi	keine native Anordnung gefunden								

Tabelle 3.8: Ergebnisse von 8 *unbound* Docking Studien. Beschriftung siehe Tabelle 3.7

Die Tabelle ist nach dem Gesamt-rmsd sortiert und die jeweils erste Lösung entspricht der besten, gefundenen Struktur. Aufgrund genereller methodischer

Schwierigkeiten im Docking-Verfahren (Mapping auf ein Gitter), wird keine Struktur mit einem idealen rmsd von null gefunden, da selbst die Anordnung, mit der das Docking begonnen wird, eine Abweichung "von sich selbst" nach einer Projektion auf das dreidimensional Gitter zeigt.
Für die in den folgenden Kapiteln durchgeführten statistischen Berechnungen von Mittelwerten und Standardabweichungen sind alle richtig-positiven Komplexe aus den vorgestellten einundzwanzig Docking-Studien zugrunde gelegt. Dies gilt auch für hier nicht aufgelistete Daten.

In Tabelle 3.9 sind die entsprechend strukturierten Ergebnisse für weitere sechs Docking-Tests angegeben. Die ersten drei Protein-Protein-Komplexe wurden mit einer Schrittweite von sechzig Grad (das entspricht 72 Anordnungen beim Behalten einer Translation), die letzten drei Systeme wurden mit zwanzig Grad (9335 Möglichkeiten bei Verwendung von fünf Translationen) gedockt. Angegeben ist der Gesamt-rmsd.

Komplex	Protein	Ligand	rmsd	Rang ckordo	Rang für E^{nf}_{gesamt}		Rang für E_{gesamt}	
					Ref. 1	Ref 2	Ref. 1	Ref 2
Enzym/Inhibitor								
1cho	5cha(A)	2ovo	1,570	20	1	13	1	14
	5cha(B)	2ovo	1,100	3	1	3	2	6
2sic	1sup	3ssi	2,437	6	1	3	3	1
1brb	1bra	1bpi	1,098	-	4	-	4	-
1brs	1a2p	1a19	2,147	3381	1140	965	1709	965
			3,035	1472	1602	1563	593	1742
1brs	1bao	1bta	3,642	5948	1815	1873	417	2078

Tabelle 3.9: Ergebnisse aus 6 *unbound*-Docking-Studien mit einem Winkel von zwanzig und sechzig Grad. Sonstige Beschriftung wie in Tabelle 3.8.

3.7.3 Vergleich der geometrischer Korrelation mit der Bewertung durch die Bindungs-Präferenzen

In zwölf von den sechzehn erfolgreichen Docking-Studien aus den Tabellen 3.7 und 3.8 wird mindestens eine nativ-ähnliche Anordnungen mit einem Rang kleiner 1000 von dem vorgestellten Potential gefunden. In Tabelle 3.10 ist der jeweils beste Rang für diese sechzehn Komplexe, sowie verschiedene rmsd-Werte aus

den obigen Tabellen zusammengetragen. Es wird deutlich, daß in acht von diesen sechzehn Fällen die beste Lösung sogar innerhalb der ersten 200 Möglichkeiten lokalisiert ist. Im Vergleich dazu, wird ein Platz unter den ersten 1000 Lösungen von ckordo nur in drei von sechzehn Fällen gefunden (nicht gezeigt).

Komplex	Rang unter 1000 nach E_{gesamt}	C_α-rmsd gesamt	rmsd gesamt	C_α-rmsd des *interface*
1mah	18	0,205	0,203	0,365
1brb	75	0,499	0,511	0,536
2ptc	149	0,644	0,849	0,676
1cho	153	1,089	1,097	1,308
1fss	9	1,153	1,139	1,443
1ppf	54	1,545	1,526	1,308
1ugh	8	1,544	1,573	2,055
1cgi	227	1,881	2,030	2,434
2kai	597	2,349	2,441	2,868
2sic (2st1)	196	2,762	2,765	1,158
1brs (1a2p)	962	2,961	2,978	3,190
1brs (1bao)	832	2,204	2,696	2,119
1bvn	—			
2sni	—			
1brc	—			
2pcc	—			

Tabelle 3.10: Ergebnisse von 16 *unbound*-Docking-Studien. Gezeigt sind nur die am besten bewerteten nativen Anordnungen aus den Tabellen 3.8 und 3.7. rmsd-Werte in Å.

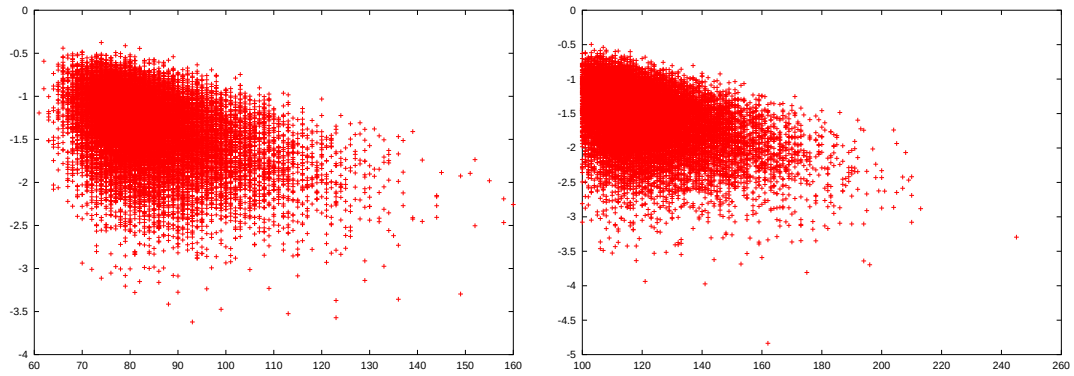
Des weiteren verbessert die Einführung der Bindungs-Präferenz E_{gesamt} zusammen mit Profil 1 die Vorhersage um durchschnittlich 4345 Plätze. In 82 % aller Systeme aus den Tabellen 3.7 und 3.8 ist die Vorhersage mit der Bindungs-Präferenz der mit rein geometrischer Korrelation (ckordo) deutlich überlegen.
Bei einer Schrittweite von sechzig Grad (vergl. Tabelle 3.9) beobachtet man den gleichen Sachverhalt. Aufgrund der kleineren Anzahl an Rotationen, wird der beste nativ-ähnliche Komplex eher auf dem ersten Platz gefunden.

	Rang bei	
Komplex	15 Grad	20 Grad
1brs/1bao/1bta	832/22006	417/9335
1brs/1a2p/1a19	528/22006	593/9335
1brb/1bra/1bpi	75/22006	4/9335

Tabelle 3.11: Vergleich der Docking-Studien von Barnase/Barstar bei fünfzehn und zwanzig Grad. Vergleiche Tabellen 3.7 und 3.9.

Es sei daran erinnert, daß das Docking mit der "richtigen" Anordnung begonnen

wird und daher das “Wiederfinden” nicht sonderlich schwierig ist, insbesondere wenn der Raum mit einem so groben Maß von sechzig Grad abgesucht wird. Wird die Schrittweite auf zwanzig Grad verkleinert, so werden 9335 Möglichkeiten (bei fünf Translationen) erhalten. Die beiden Barnase/Barstar-Systeme sind auf mittleren Rängen zu finden, wohingegen das 1brb/1bra/1bpi-System sehr gut diskriminiert wird (vergl. Tabelle 3.11).
Ein Vergleich der geometrischen Korrelation mit der Bindungs-Präferenz, wie in Abbildung 3.20 für zwei Beispiele dargestellt, zeigt keine Abhängigkeiten der beiden Eigenschaften. Wäre eine lineare Korrelation vorhanden, ließe sich ein Kriterium durch das jeweils andere darstellen. Die Verwendung zweier Bewertungskriterien würde keinen Informationsgewinn bringen. In keinem der untersuchten Systeme ist eine lineare Abhängigkeit zu erkennen.



(a) 1cho **(b)** 2kai

Abbildung 3.20: Vergleich der geometrischen Korrelation (x-Achse) mit der Bindungs-Präferenz (y-Achse) für (a) Chymotrypsin/OMTKY (1cho) und (b) Kalikrein/Inhibitor (2kai).

3.7.4 Kombination von Geometrie und Präferenzen

In Tabelle 3.12 sind die Ergebnisse des Versuches einer einfachen Kombination von geometrischer Korrelation und den Bindungs-Präferenzen nach der in Kapitel 2.4.2 beschriebenen Methode dargestellt. In fast allen Fällen ist die Bewertung

anhand geometrischer Komplementarität wesentlich schlechter als die bei Verwendung der Bindungs-Präferenzen. Daraus folgt fast zwangsläufig, daß eine kombinierte Bewertung der Strukturen zwar zu einer Aufwertung der geometrischen Korrelation, aber auch zu eine Abwertung der Bindungs-Präferenzen führt.

	Rang		
PDB	geom.	E_{gesamt}	geom/. E_{gesamt}
1cho	3239	153	677
	2897	3540	953
	989	185	157
	4757	473	1275
	2685	10321	4723
1cgi	10797	227	367
	11965	2396	5375
1brs	8000	3667	3876
(1a2p)	4265	198	1509
	5435	962	1557
	6265	528	1710
1brs	19576	2333	10808
(1bao)	13281	832	5295
	3152	1033	914
2sic	6030	1646	3181
	8882	3969	7333
	2817	196	719
	879	5237	2169
2sni	8858	8968	7540
	18773	12082	17572
	8406	12787	10198
	15870	1311	7099
2kai	15996	16360	18404
	13344	597	5123
	3260	8323	3873
	6137	5351	3823
	1269	323	1268
	16531	471	16531
2ptc	790	149	155
	2817	208	811
	4271	31	1313
1ppf	7188	355	1523
	11784	54	3441
	19159	6795	14936

	Rang		
PDB	geom.	E_{gesamt}	geom./ E_{gesamt}
1brc	9813	6169	9203
	1319	6939	3200
	3468	3705	2538
	7499	6176	7532
	5216	8110	7260
1ugh	7465	6847	7300
	1	16	2
	3570	190	906
	320	35	31
	6104	966	2275
	1994	52	387
1brb	131	75	19
	2819	101	704
	1671	304	426
	7454	1252	2830
	12339	1393	5245
1bvn	2881	10090	4637
	18301	12658	17378
	9542	15313	12851
	13148	14865	15373
	12982	15714	15833
1fss	7718	846	2485
	2362	128	445
	18179	9	7700
	15455	343	6203
	853	202	136
2pcc	20517	18758	21260
	16661	21058	20716
	10676	17909	15883
1mah	4573	18	1080
	15797	81	6337
	1228	9	176
	12813	3112	6373
	843	93	112

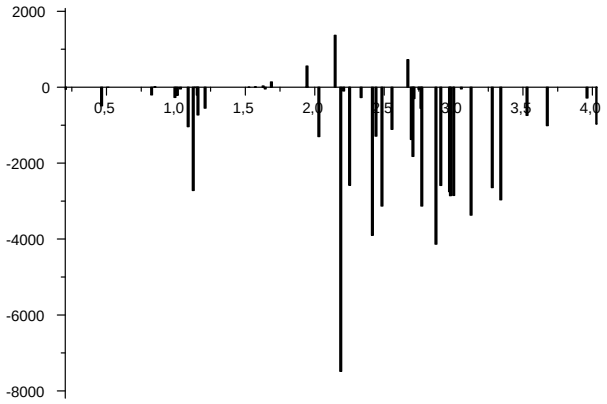
Tabelle 3.12: Kombiniertes Sortieren von geometrischen Korrelation und Bindungs-Präferenz. Vergleich mit den Ergebnissen der Einzelwertungen.

Die Abstände zwischen den beiden Bewertungen sind zu groß, als das eine gemeinsame Sortierung mit so einfachen Mitteln sinnvoll wäre. In Einzelfällen sind die Schnittmengen so klein, daß es zu einer Verbesserung beider Ranglisten kommt, wie z.B. für einen Komplex des Barnase/Barstar(1brs)-Systems, einer Anordnung der Acetylcholinesterase/Fascicullin II (1fss) und eine Struktur bei Kalikrein/Inhibitor (2kai).

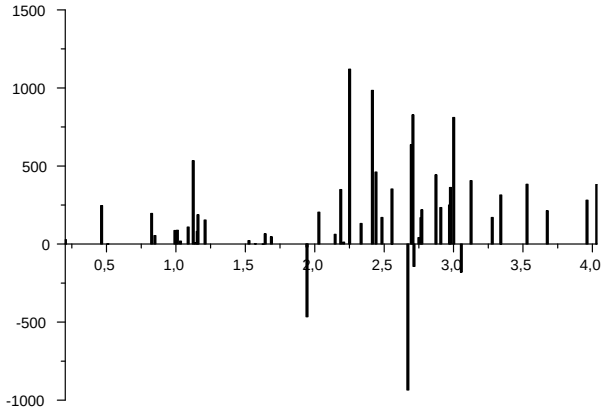
3.7.5 Vergleich von E^{nf}_{gesamt} und E_{gesamt}

Die Verwendung von der Bindungs-Präferenz **mit** funktionellen Gruppen E_{gesamt} anstatt von E^{nf}_{gesamt}, also der Bindungs-Präfrenenz **ohne** funktionelle Gruppen bringt eine deutliche Verbesserung der Positionierung nativ-ähnlicher Konformationen für Profil 1.
Für die Erzeugung der Bindungs-Präferenzen siehe Gleichung 2.32 auf Seite 44, Gleichung 2.17 und Gleichung 2.16 auf Seite 33. Im Mittel wird bei Profil 1 für E_{gesamt} ein um etwa 1100 $\pm$1600 Plätze höherer Rang erreicht. Diese Werte sind jedoch nur als grobe Richtlinie zu verstehen, da eine Statistik über 10-15 Datenpunkte nicht aussagekräftig ist. Für Profil 2 ergibt sich eine Verschlechterung der Platzierung um durchschnittlich ca. 200 nach Einführung der funktionellen Gruppen. Auch in diesem Fall sind die Schwankungen um den Mittelwert enorm. Für einige Strukturen ist durchaus eine Verbesserung des Ranges zu vermerken, was bei der Mittelwertsbildung jedoch verloren geht.
In Abbildung 3.21 ist der Rang von E_{gesamt} minus dem Rang von E^{nf}_{gesamt} gegen den Gesamt-rmsd für Profil 1 (a) und Profil 2 (b) aufgetragen. Negative Werte bedeuten eine bessere Bewertung bei Verwendung von E_{gesamt}. Für Profil 1 ist die Verwendung von E^{nf}_{gesamt} nur in 17.6% aller Fälle günstiger, für Profil 2 hingegen in 88.2%.
In Abbildung 3.22 ist die Platzierung der Strukturen bei Verwendung der beiden Bindungs-Präferenzen E^{nf}_{gesamt} und E_{gesamt} für beide Profile gegeneinander aufgetragen. Es ist eine lineare Korrelation ($r_{corr} = 0.968$ und $r_{corr} = 0.999$) zwischen den Platzierungen der Komplexe bei Verwendung von E_{gesamt} und von E^{nf}_{gesamt} zu erkennen. Die Einführung von funktionellen Gruppen verändert die Bindungs-Präferenzen nicht, sondern sorgt für eine Verschiebung der Platzierung. Im Fall von Profil 1 ist eine Verschiebung hin zu kleineren und damit besseren Rängen erkennbar, wohingegen bei Profil 2 eine allgemeine Verschlechterung eintritt.

Als Fazit läßt sich festhalten, daß für Profil 1 die Einführung von funktionellen Gruppen eine deutliche Verbesserung bringt, wohingegen sie bei Profil 2 zu einer leichten Verschlechterung führt.

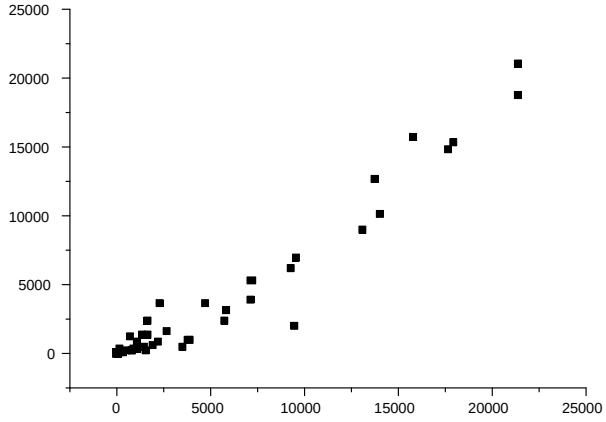


(a) Profil 1

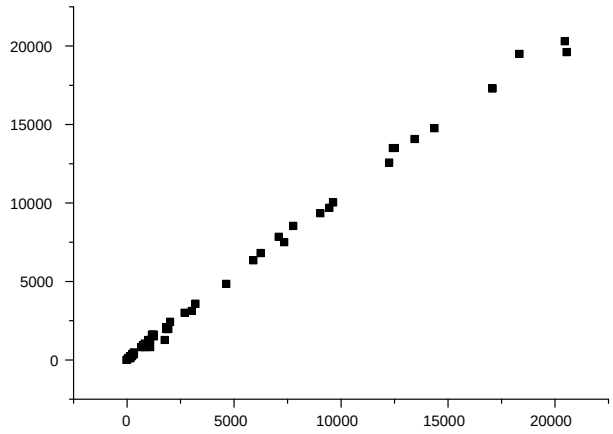


(b) Profil 2

Abbildung 3.21: Abszisse: Rang der nativ-ähnlichen Lösungen berechnet von E_{gesamt} minus entsprechendem Rang von E^{nf}_{gesamt}. Ordinate: Gesamt-rmsd in Å .



(a)

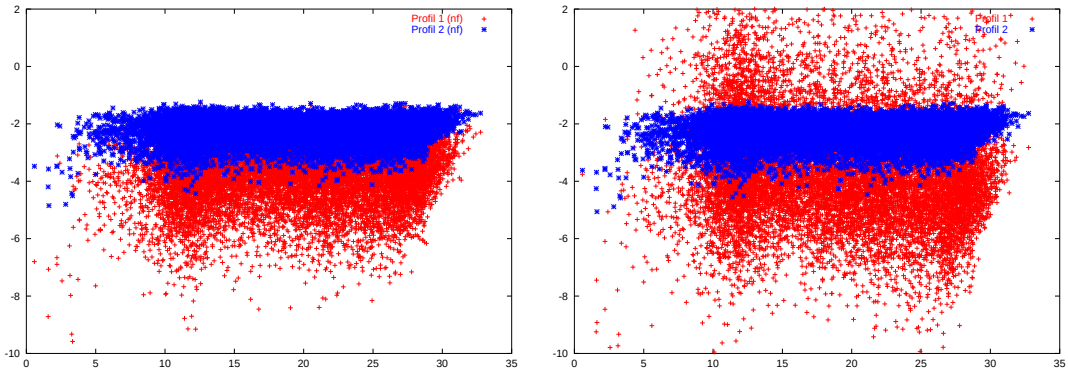


(b)

Abbildung 3.22: Auftragung der Reihenfolge der richtig-positiven Lösungen für E^{nf}_{gesamt} gegen E_{gesamt}, Profil 1 mit y = 1.02x, r_{corr} = 0.968 (a) und Profil 2 mit y = 1.02x, r_{corr} = 0.999 (b).

3.7.5.1 Vergleich der Performance von Profil 1 und Profil 2

Ein Vergleich der Bewertungsqualität von Profil 1 und Profil 2, zeigt eine deutliche Verbesserung in dreizehn von sechzehn Docking-Studien bei Benutzung von Profil 1. Der Anstieg in den Platzierung von nativ-ähnlichen Strukturen läßt sich noch genauer separieren, bei Betrachtung von E^{nf}_{gesamt} und E_{gesamt}.



(a) E^{nf}_{gesamt} **(b)** E_{gesamt}

Abbildung 3.23: E^{nf}_{gesamt} und E_{gesamt} von 1ugh für beide Profile. Abszisse: Bindungs-Präferenz, Ordinate: Gesamt-rmsd. Helles Grau (Rot) jeweils Profil 1, Schwarz (Blau) entspricht Profil 2.

Im Mittel steigen von Profil 2 zu Profil 1 alle Ränge der richtig-positiven Komplexe bei Verwendung von E^{nf}_{gesamt} um 260 und bei Benutzung von E_{gesamt} um 1600 Plätze. Dies ist im Einklang mit den in Kapitel 3.7.5 gewonnenen Erkenntnis, daß die Einführung funktioneller Gruppen eine deutliche Verbesserung für Profil 1 erbringt.
Abbildung 3.23 verdeutlicht die Unterschiede in der Diskriminierungskraft von Profil 1 zu Profil 2 für E^{nf}_{gesamt} und E_{gesamt}. Profil 1 zeigt deutlich häufigeres Auftreten von Anordnungen mit kleinem rmsd und kleiner Bindungs-Präferenz (helles Grau in Abb. 3.23). Dies führt zu einer Entzerrung der "Energie"-Werte, während sie für Profil 2 dicht zusammen liegen. Der Effekt steigt von E^{nf}_{gesamt} zu E_{gesamt} ((a) zu (b)).

3.7.6 Vergleich von $E^{nf}_{spezifisch}$ mit E^{nf}_{gesamt}

Tabelle 3.13 zeigt die Änderung der Bindungs-Präferenz bei Verwendung des solvensabhängigen Terms E^{nf}_{gesamt} und ohne diesen $E^{nf}_{spezifisch}$.
Es ist keine derart drastische Steigerung zu erkennen, wie bei Einführung eines anderen Profils oder der funktionellen Gruppen. In nur wenigen Fällen ist die Differenz Rang(E^{nf}_{gesamt})-Rang($E^{nf}_{spezifisch}$) positiv, d.h. das Potential schlechter als vorher. Auffälligerweise gilt dies für Chymotrypsin/Inhibitor (1cgi), ein Komplex mit extrem schlechter geometrischer Komplementarität, ebenso wie für Peroxidase/Cytochrom C (2pcc), bei dem alle untersuchten Kriterien nicht zu einer Identifizierung des nativen Komplexes führen.

PDB	Rangdifferenz			PDB	Rangdifferenz		
1cho	−153.7	±	120.7	1brc	−17.4	±	22.0
1cgi	34.0	±	125.9	1ugh	1.3	±	2.8
1brs (1bao)	−91.0	±	59.5	1brb	−5.7	±	20.4
1brs (1a2p)	−133.5	±	86.5	1ppf	6.3	±	4.0
2sic	−3.2	±	8.1	1bvn	−171.3	±	78.5
2sni	36.5	±	19.8	1fss	−31.1	±	121.5
2kai	−10.3	±	40.5	2pcc	106.7	±	130.2
2ptc	0.3	±	3.5	1mah	−65.9	±	123.6

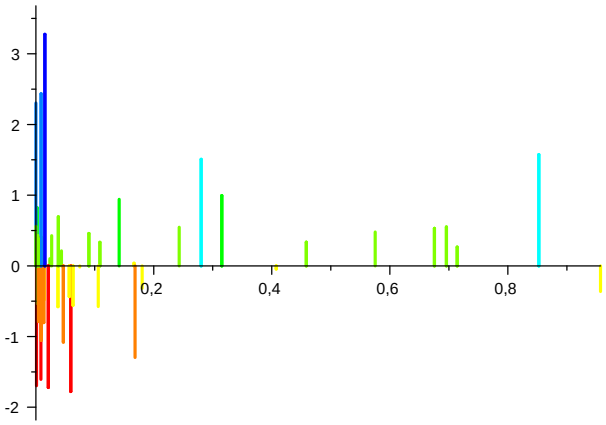
Tabelle 3.13: Vergleich der Präferenzen mit dem solvensabhängigem Term (E^{nf}_{gesamt}) und ohne diesen ($E^{nf}_{spezifisch}$) für die sechzehn erfolgreichen Docking-Studien aus Kapitel 3.7. Angegeben ist das pdb-Kürzel des Komplexes, und die Rangdifferenz Rang(E^{nf}_{gesamt})-Rang($E^{nf}_{spezifisch}$).

Auch bei anderen Komplexen (Elastase/OMTKY (1ppf), Glykolase/Inhibitor (1ugh), Subtilisin Novo/Inhibitor (2sni), Trypsin/Inhibitor (2ptc)) bringt die Einführung des solvensabhängigen Terms in dieser Form im Mittel wenig oder sogar schlechtere Ergebnisse. Auf der anderen Seite stehen Komplexe wie Chymotrypsin/OMTKY (1cho), Barnase/Barstar (1brs) und Hydrolase/Inhibitor (1bvn) für die das Gegenteil gilt.

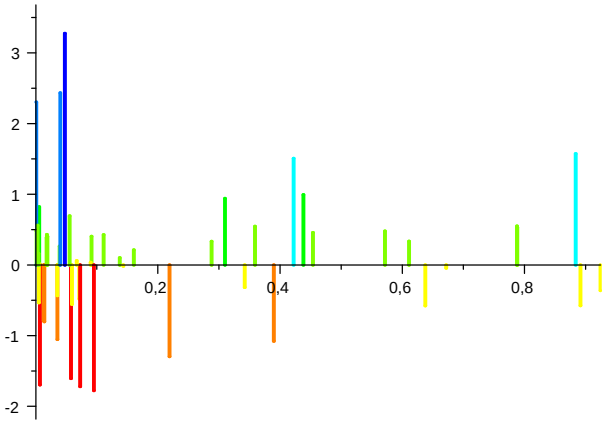
3.7.7 Betrachtung unterschiedlicher rmsd-Werte

Um den *interface*-rmsd mit dem Gesamt-rmsd zu vergleichen, zeigt Abbildung 3.24 das Verhältnis vom Rang des besten richtig-positiven Komplexes zur Ge-

samtzahl $\frac{\text{Rang(richtig-positiver Komplex)}}{\text{Anzahl aller Komplexe}}$ (x-Achse) aufgetragen gegen die Differenz von *interface*-rmsd und Gesamt-rmsd (y-Achse) für beide Profile.



(a) Profil 1



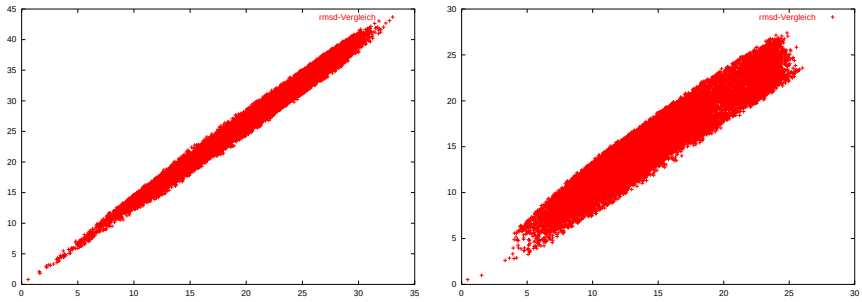
(b) Profil 2

Abbildung 3.24: x-Achse: Verhältnis des Ranges für den besten richtig-positiven zur Anzahl aller möglichen Ränge $\frac{\text{Rang(richtig-positiver Komplex)}}{\text{Anzahl aller Komplexe}}$. Skalierung von 0-1. y-Achse: Differenz *interface*-rmsd -Gesamt-rmsd in Å.

Ein positiver Wert (grün-blauer Balken, Farbabstufung je nach Absolutwert) bedeutet, daß der *interface*-rmsd größer als der Gesamt-rmsd ist. In diesen Fällen ist die "allgemeine" Superpositionierung unter Berücksichtigung aller Atome recht gut, das *interface* hingegen "passt deutlich weniger gut". Dieser Trend ist umso stärker ausgeprägt, je mehr das Verhältnis des besten richtig-positiven Komplexes zu allen aus dem Docking-Prozess vorgeschlagenen Strukturen gegen eins tendiert (x-Achse). Im Umkehrschluss bedeutet dies aber auch, daß in Docking-Systemen mit relativ schlechter Bewertung der *interface*- dem Gesamt-rmsd vorzuziehen ist.

Als "relativ schlechte" Bewertung wird ein Verhältnis von richtig-positiver Struktur zu allen möglichen größer 0.2 (x-Achse) angesehen. Anschaulich entspricht das einem höherem Rang für den besten nativ-ähnlichen Komplex als etwa 4400 bei 22006 Anordnungen.

In den meisten Systemen korrelieren *interface*- und Gesamt-rmsd sehr gut (Abb. 3.25, (a), (b)), in anderen Fällen weichen sie deutlicher voneinander ab (Abb. 3.26, (c), (d)). Für die Komplexe 1ugh und 1ppf liegt beispielsweise eine gute, lineare Korrelation vor. Hier kann der Gesamt-rmsd als sinnvolles Maß zum Vergleichen der nativ-ähnlichen Lösungen mit dem superpositionierten Komplex verwendet werden.



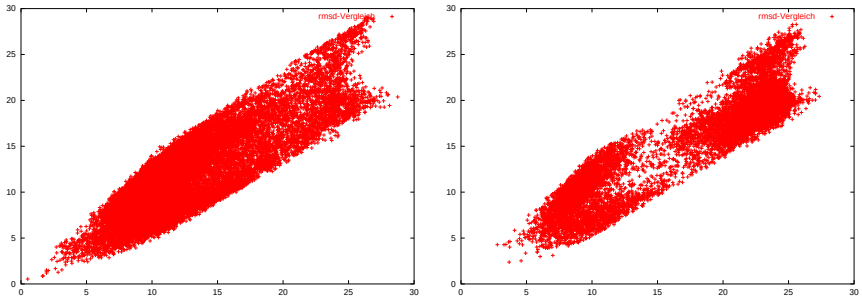
(a) 1ugh **(b)** 1ppf

Abbildung 3.25: Vergleich von Gesamt-rmsd (x-Achse) und *interface*-rmsd (y-Achse) aus den erfolgreichen sechzehn *unbound*-Docking-Studien aus Kapitel 3.7. rmsd-Werte in Å.

Im System Trypsin/Inhibitor (1brb) sind einige *interface*-rmsd-Werte kleiner als der Gesamt-rmsd. Das ist nicht unerwartet, denn für eine nativ-ähnliche Anordnung ist es sinnvoll, daß die Bindestelle deutlich besser "passt", als der gesamte Komplex. Eine derartige Kombination ist die optimale Voraussetzung für eine gute Vorhersage.

Anders ist der Sachverhalt bei Systemen wie Trypsin/APPI (1brc). Die meisten Anordnungen haben einen kleineren Gesamt-rmsd als *interface*-rmsd. Die allgemeine Ähnlichkeit der Struktur zum superpositionierten Komplex ist also gut, der Kontaktbereich weicht jedoch deutlich ab. In diesem Fall ist die Bewertungsqualität schlechter als bei Systeme wie Glykolase/Inhibitor (1ugh). Hier sollte der *interface*-rmsd verwendet werden, da er ein besseres Ähnlichkeitsmaß darstellt.

Allgemein betrachtet ist vor der Bewertung von Docking-Ergebnisse nicht bekannt, welcher rmsd-Wert am sinnvollsten ist. Nach Möglichkeit sollte daher der *interface*-rmsd dem Gesamt-rmsd vorgezogen werden.



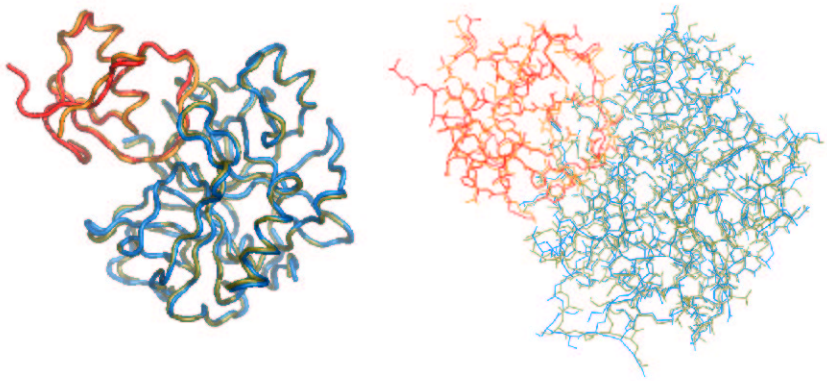
(a) 1brb **(b)** 1brc

Abbildung 3.26: Vergleich von Gesamt-rmsd (x-Achse) und *interface*-rmsd (y-Achse) aus den erfolgreichen sechzehn *unbound*-Docking-Studien aus Kapitel 3.7. rmsd-Werte in Å.

3.8 Ein detailliertes Beispiel - α-Chymotrypsin/OMTKY

3.8.1 Die Superpositionierung

Chymotrypsin gehört zu den gut untersuchten Endopeptidasen. Sie dienen der Spaltung von Peptidbindungen. Die Struktur wurde mit einer Auflösung von 1.8 Å bestimmt. Allerdings sind siebzehn Aminosäuren so schlecht aufgelöst, daß die Position einiger Atome nicht genau angegeben werden kann. Der Komplex besteht aus zwei Ketten, dem 1735 Atome großem Enzym und dem 418 Atome langen Inhibitor OMTKY. Der Gesamt-CA-rmsd der einzeln kristallisierten Proteine 5cha und 2ovo zum nativen Komplex Chymotrypsin/OMTKY (1cho) beträgt 0.62 Å. Diese Abweichung ist verglichen mit den rmsd-Werten anderen superpositionierter Komplexe hoch (vergl. Tabelle 3.4 auf Seite 67). Trotzdem werden so gut wie keine größeren Bewegungen der Hauptkette beobachtet, was eine günstige Ausgangsposition für Docking-Untersuchungen ist.

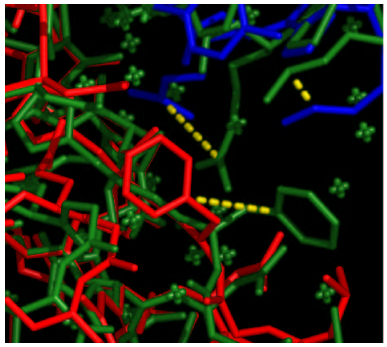


(a) **(b)**

Abbildung 3.27: (a) und (b): 1cho_i (nativer Komplex-kleineres Protein) in orange, 1cho_e (nativer Komplex-größeres Protein) in oliv, 2ovo in rot und 5cha in blau.

In Abbildung 3.27 (a) ist die Superpositionierung der Hauptkette dargestellt. Der

allgemeine Verlauf der Ketten stimmt mit dem im nativen Komplex überein. Bei Berücksichtigung der Seitenketten (Abb.3.27 (b)) fallen auf Anhieb einige Atome mit größeren Abweichungen auf. Abbildung 3.27 (c) zeigt diese für Atome dreier *interface*-Aminosäuren. Die Distanz zwischen den beiden CG-Atomen der Seitenkette PHE39 vom Chymotrypsin/OMTKY-Komplex (1cho) und Chymotrypsin (5cha) beträgt 3.63 Å. Die CZ-Atome von ARG21 von Chymotrypsin/OMTKY (1cho) und dem Liganden OMTKY (2ovo) haben einen Abstand von 3.36 Å und die CE von LYS55 immerhin 1.68 Å. Die Aminosäure LYS 55 vom Liganden gehört zu den anfangs erwähnten, schlecht aufgelösten. Die Positionen der Atome hinter CG sind nicht genau bestimmbar.



(a)

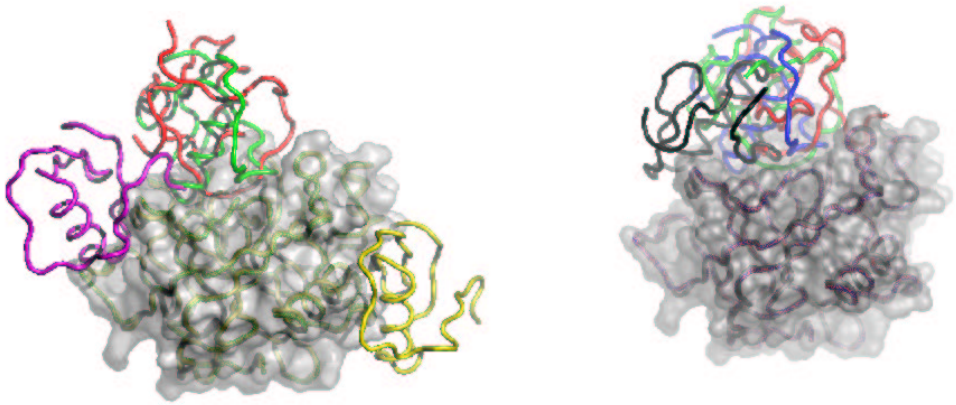
Abbildung 3.28: Abstände in Å: 3.36, 1.68 und 3.63. In grün ist die native Komplexstruktur (1cho) gekennzeichnet, in rot das größere Protein (5cha) und in blau das kleinere Protein (2ovo)

Die CZ-Atome von ARG21 von Chymotrypsin/OMTKY (1cho) und dem Liganden OMTKY (2ovo) haben einen Abstand von 3.36 Å und die CE von LYS55 immerhin 1.68 Å. Nach optischer Begutachtung sind keine langen, falsch orientierten Seitenketten in der Bindestelle vorhanden. Systeme wie 2kai, in denen Arginine oder andere Aminosäuren im Protein starke Abweichungen von der Konformation im nativen Komplex haben, bewirken oftmals eine schlechte geometrische Korrelation und erschweren damit das Docking (Kimura *et al.* (2001)). Im vorliegenden Fall des 1cho sind zwar große Abweichungen einzelner

Aminosäuren zu erkennen, keine jedoch inmitten des *interfaces*.

3.8.2 Docking-Ergebnisse

Beim Docking mit unterschiedlichen Gradzahlen schneidet das System Chymotrypsin/OMTKY (1cho) sehr gut ab. Bei fünfzehn Grad werden fünf nativähnliche Lösungen bei Verwendung der hier berechneten Bindungs-Präferenz E_{gesamt} mit Profil 1 unter den ersten 2000 Rängen gefunden. Die hierunter am besten gewertete Anordnung ist auf Platz 135 von 22006 möglichen mit einem Gesamt-rmsd von 1.097 Å zu finden (vergl. Tabelle 3.7).



(a) **(b)**

Abbildung 3.29: Grün: 1cho (nativer Komplex); rot: bester falsch-positiver (Potential, Profil 1), Rang 3641 (Geometrie); pink: bester falsch-postiver (Geometrie), Platz 556 (Potential, Profil 1). (a) Gelb: Platz zwei (Potential, Profil 2), Rang 139 (Geometrie). (b) Blau: zweiter Platz (Potential, Profil 1), Rang 14221 Geometrie); schwarz: Rang 4 (Potential, Profil 1), Rang 2745 (geometrische Korrelation).

Zur Untersuchung der Frage ob die Anordnungen, die von der geometrischen Korrelation gut abschneiden auch eine gute "energetische" Bewertung erhalten, sind einige Komplexe analysiert worden. In Abbildung 3.29 ist der native Komplex (1cho_i in grün und 1cho_e als Oberflächendarstellung), sowie einige

gefundenen Lösungen aufgetragen. In rot ist die beste falsch-positive Struktur nach Bewertung durch die Bindungs-Präferenz dargestellt. Dieser liegt sehr nahe an der richtigen Anordnungen. Der Ligand ist lediglich um einige Grad verdreht, aber in der Bindetasche lokalisiert. Gleiches kann von dem besten falsch-positiven Komplex nach der geometrischen Beurteilung nicht gesagt werden (pink). Diese Struktur liegt deutlich neben der Kontaktstelle (Gesamt-rmsd = 5.355 Å). Vom hier vorgestellten Potential hingegen wird die Struktur auf Rang 556 gefunden. Dieser Platz ist relativ zur Gesamtanzahl aller Möglichkeiten sehr hoch, also schlecht. Ähnliches gilt auch umgekehrt, wie am Beispiel des vom Potential (Profil 2) als zweit-besten bewerteten falsch-positiven Komplex verdeutlicht wird (gelbe Struktur des Liganden in Abb. 3.29). Nach geometrischer Sortierung erhält die Struktur den ebenfalls sehr hohen Rang von 139, obwohl sie falsch ist. Warum in beiden Fällen für die falsch-positive Struktur eine so gute Bewertung der geometrischen Korrelation und der Bindung-Präferenz erhalten wird, ist nicht klar. Die Struktur mit der geometrischen Bewertung von Platz 14221 erhält bei Benutzung der Bindungs-Präferenzen einen Rang von drei. Eine optische Begutachtung zeigt, daß auch diese Struktur fast in der Bindetasche liegt (blau in Abb: 3.29 (b)). Für die Anordnung 2756 (geometrische Korrelation) liegt ein Rang von 4 bei "energetischer" Betrachtung vor. Der Ligand ist auch in diesem Fall in der Nähe der Kontaktstelle zu finden (schwarz in der selben Abbildung).

Rang			
geometrisch	Profil 1	Profil 2	*interface*-rmsd
1	552/556	505/591	17.221
3641	1/1	1/1	8.377
139	3/274	2/2	19.907
14221	6/2	9/13	6.104
4956	12/3	17/18	12.862
2756	14/4	11/11	8,395

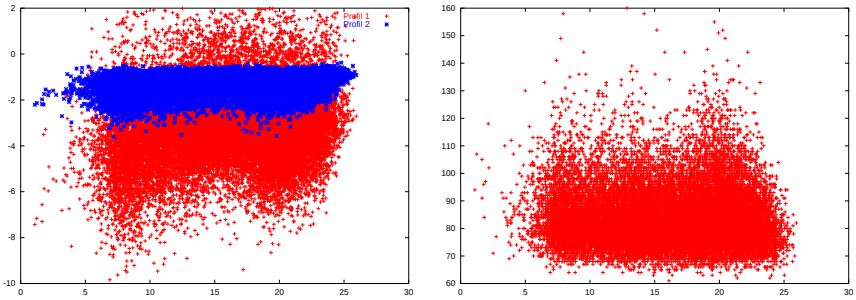
Tabelle 3.14: Gegenüberstellung der geometrischen Korrelation (Spalte eins) mit den Platzierungen für E^{nf}_{gesamt} und E_{gesamt} bei beiden Profilen (Spalten zwei und drei), sowie der *interface*-rmsd.

Tabelle 3.14 faßt die grafisch dargestellten Ergebnisse zusammen. Es wird deutlich, daß die vom Potential als gut bewerteten Lösungen in den untersuchten

Fällen nahe an der Bindestelle liegen. Gleichermaßen wird klar, daß viele falsche Strukturen von beiden Kriterien als zu gut berwertet werden.

3.8.3 Vergleich mit der geometrischen Korrelation

Abbildung 3.30 (a) zeigt die Verteilung der Bindungs-Präferenzen aller Anordnungen für 1cho bei fünfzehn Grad im Verhältnis zum Gesamt-rmsd für E^{nf}_{gesamt} (schwarz, bzw. blau) und E_{gesamt} (grau, bzw. rot). Daneben ist die geometrische Korrelation gegen den Gesamt-rmsd aufgetragen. In beiden Fällen ist ein dichtes Hauptfeld zwischen fünf und fünfundzwanzig Å zu erkennen. Die geometrische Komplementarität ist demnach an vielen auch von der Bindestelle entfernten Stellen stark ausgeprägt. Die Strukturen mit der höchsten geometrischen Bewertung sind sehr weit von der nativen Komplex-Struktur entfernt, wie auch schon exemplarisch im vorigen Kapitel gezeigt wurde. Anordnungen mit einem rmsd kleiner 5 Å finden sich im Mittelfeld der Bewertung wieder. Im Falle der Bindungs-Präferenzen kommt es zu einer etwas besseren Unterscheidung nativähnlicher von nicht-nativen Komplexen.

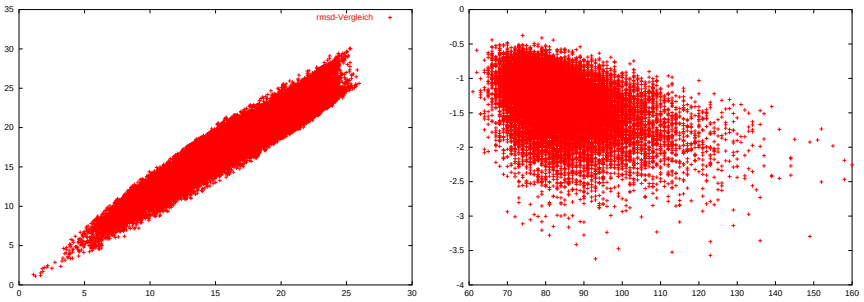


(a) **(b)**

Abbildung 3.30: (a) E^{nf}_{gesamt} (schwarz, bzw. blau) E_{gesamt} (grau, bzw. rot), (y-Achse). (b) geometrische Korrelation (y-Achse). Beide jeweils gegen den Gesamt-rmsd in Å (x-Achse).

In Abbildung 3.31 sind wieder *interface*- gegen Gesamt-rmsd (a) und geometrische Korrelation gegen die Bindungs-Präferenz (b) aufgetragen. In diesem

System sind keine großen Abweichungen der unterschiedlichen rmsd-Werte zu erkennen. Der Gesamt-rmsd, sowie der *interface*-rmsd auch, kann hier als sinnvolles Ähnlichkeitsmaß auch der Bindetasche verwendet werden.
Die beiden in dieser Arbeit verwendeten Kriterien der geometrischen Komplementarität und der Bindungs-Präferenz sind nicht offensichtlich korreliert. Eine allgemeine Tendenz von kleiner geometrischer Korrelation und schlechter Bindungs-Präferenz hin zu großer Komplementarität und guter "energetischer" Bewertung ist erkennbar. Die Strukturen mit der höchsten geometrischen Passform haben jedoch nicht zwangsläufig eine hohe statistische Präferenz und *vice versa*.



(a) **(b)**

Abbildung 3.31: (a): Korrelation von *interface*- (y-Achse) und Gesamt-rmsd (x-Achse) jeweils in Å.(b): Vergleich der geometrischen Korrelation (x-Achse) mit der Bindungs-Präferenz (y-Achse) für Profil 2.

3.9 Decoy-Bewertung

Das vorgestellte Potential unterscheidet ungeordnete Strukturen im Mittel nicht hinreichend gut vom nativen Komplex. In Abbildung 3.32 sind die hier berechneten Bindungs-Präferenzen für Acetylcholinesterase/Fasciculin II (1fss) mit und ohne Gewichtung von Atomen funktioneller Gruppen für beide Profile aufgetragen.

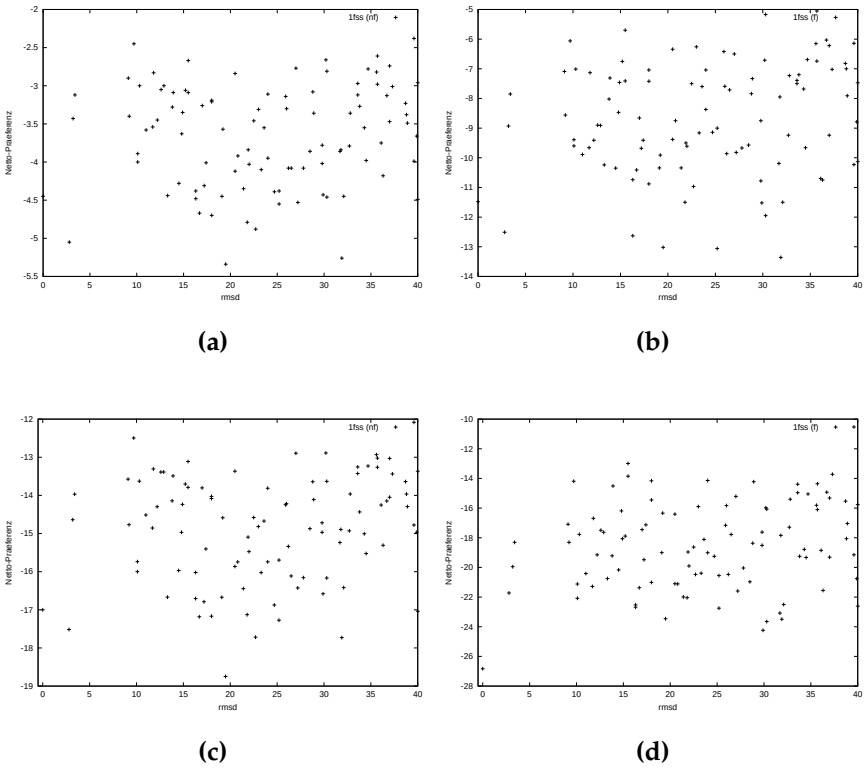


Abbildung 3.32: 1fss-decoy: rmsd in Å (*interface*-CA). (a) und (b) Profil 2. (c) und (d) Profil 1. nf = ohne Gewichtung von Atom-Atom-Kontakten funktioneller Gruppen (a), (c). f = mit Gewichtung derselben (b), (d).

Die Verwendung von Profil 1 und Berechnung von E_{gesamt} platzieren den nativen Komplex auf den ersten Rang. Profil 2 hingegen findet ihn auf Rang elf und ist damit bedeutend schlechter. Wie bereits für die Docking-Studien gezeigt, verbessert die Einführung funktioneller Gruppen auch in diesem Fall die Bewertung.

Für das System Chymotrypsin/Inhibitor (1cgi) in Abbildung 3.33 wird in allen vier Fällen die richtige Struktur auf dem ersten Platz gefunden.

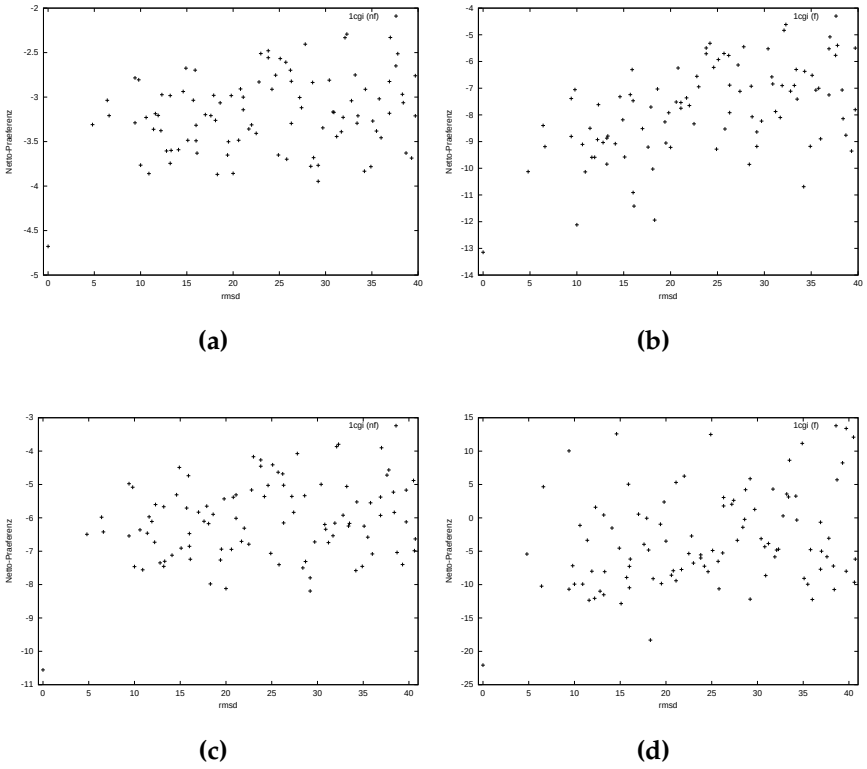


Abbildung 3.33: 1cgi-decoy: rmsd in Å (*interface*-CA). (a) und (b) Profil 2. (c) und (d) Profil 1. nf = ohne Gewichtung von Atom-Atom-Kontakten funktioneller Gruppen (a), (c). f = mit Gewichtung derselben (b), (d).

Dieses Ergebnis ist erstaunlich, da die Docking-Simulationen mit diesem Komplex sehr schlechte Bewertungen ergeben (vergl. Kapitel 3.7).
Werden die beiden Proteine mit dem ko-kristallisierten Komplex superpositioniert, so stehen einige Lysin-Ketten *interface* sterisch ungünstig fast senkrecht von der Oberfläche ab. Diese verschlechtern die geometrische Komplementarität und führen somit zu einer Verschlechterung der Docking-Ergebnisse. Für die verwendeten Chymotrypsin/Inhibitor(1cgi)-decoy-Struturen ist nicht bekannt, ob die Stellungen dieser Seitenketten für die Auswahl des Datensatzes besonders

betrachtet wurden. Dies könnte der Grund für die deutlichen Unterschiede der Bewertung von Chymotrypsin/Inhibitor(1cgi)-decoys und selbst durchgeführten Chymotrypsin/Inhibitor(1cgi)-Docking-Studien sein.
Im Vergleich zum System Acetylcholinesterase/Fascicullin II (1fss), sind für alle vier Parameterkombinationen ähnliche Punktverteilungen zu erkennen. Die Diskriminierung vom nativen Komplex, sowie Komplexen mit nur geringer Störung (kleinem rmsd) ist in den vier Fällen vergleichbar.

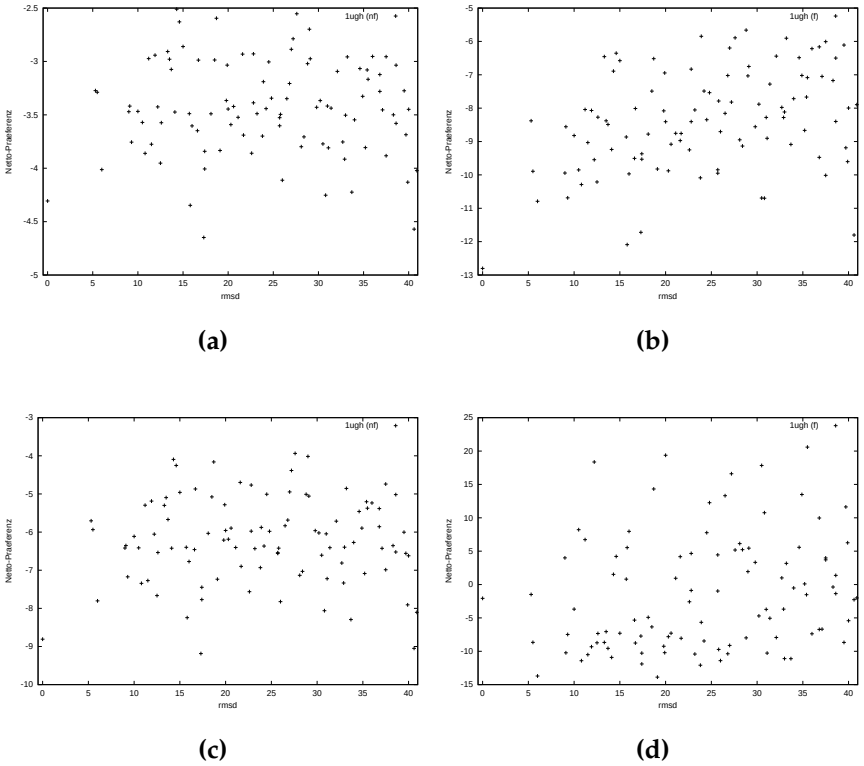


Abbildung 3.34: 1ugh-decoy: rmsd in Å (*interface*-CA). (a) und (b) Profil 2. (c) und (d) Profil 1. nf = ohne Gewichtung von Atom-Atom-Kontakten funktioneller Gruppen (a), (c). f = mit Gewichtung derselben (b), (d).

Für das in Abbildung 3.34 wiedergegebene System 1ugh ist der Sachverhalt anders. Hier zeigt Profil 2 die besten Ergebnisse. Die Einführung funktioneller Gruppen ist wie im Fall von 1fss positiv für die Bewertung für Profil 2, sehr

schlecht hingegen für Profil 1. Die Form der Punktwolken für E_{gesamt} zeigt eine leichte Tendenz von kleinem rmsd und niedrigen Präferenzen zu hohem rmsd und hoher Präferenz. Für E^{nf}_{gesamt} ist dieser Verlauf schwächer ausgeprägt. In den Abbildungen A.3 und A.4 im Anhang sind die entsprechenden Größen für die Systeme Subtilisin BPN/Inhibitor (2sic) und Kalikrein/Inhibitor (2kai) aufgetragen. Für Subtilisin BPN/Inhibitor (2sic) ist die Kombination Profil 1 mit E_{gesamt} die erfolgreichste um native Strukturen zu erkennen. Während mit Profil 2 der beste zu erreichende Rang bei neunzehn liegt, wird der "richtige" Komplex mit Profil 1 auf Platz drei gefunden. Für Kalikrein/Inhibitor (2kai) bringt die Einführung funktioneller Gruppen eine deutliche Verschlechterung.
Tabelle 3.15 faßt die Ergebnisse für E^{nf}_{gesamt}, $E^{nf}_{spezifiscj}$ und E_{gesamt} nochmals zusammen.

Decoy	Rang für E^{nf}_{gesamt}		Rang für $E^{nf}_{spezifisch}$		Rang für E_{gesamt}	
	Profil 1	Profil 2	Profil 1	Profil 2	Profil. 1	Profil 2
1cgi	1	1	1	1	1	1
1ugh	3	4	3	4	53	1
2kai	8	18	10	18	79	16
1fss	10	15	10	15	1	11
2sic	18	19	18	19	3	25
5cha	2	-	2	-	4	-

Tabelle 3.15: $E^{nf}_{gesamt} = \sum_i \sum_j \Delta W^{nf}_{ij}(r_d)$,$E^{nf}_{spezifisch}$ analog, $\alpha = 30$, $\beta = 1$, $\gamma = 100$, $\delta = 1$. Profil 1 entspricht 0.025 und Profil 2 folgt Gleichung 2.10.

3.10 Korrelation mit experimentellen Bindungsdaten

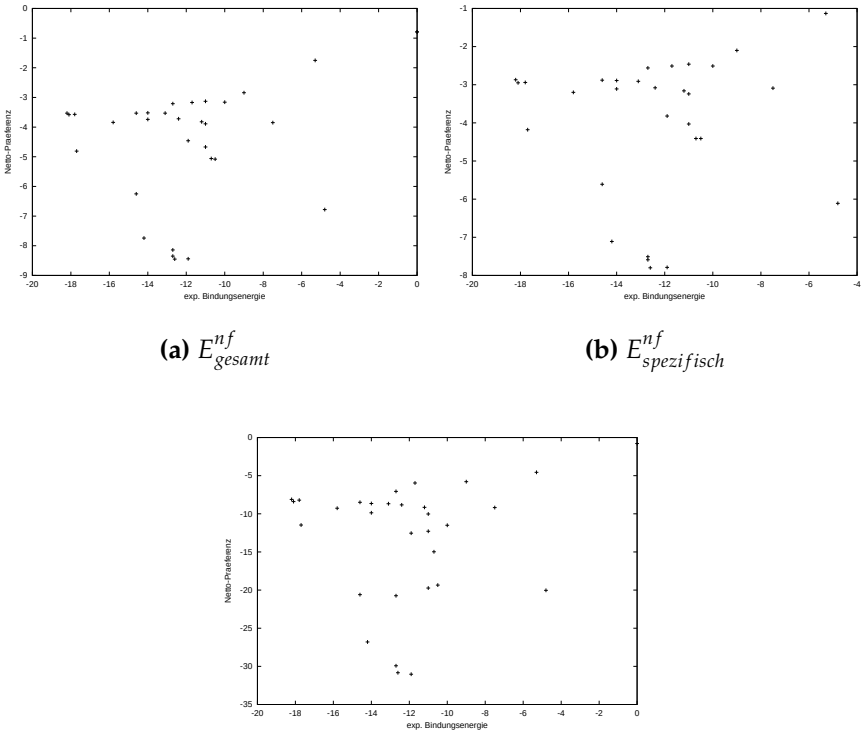
In Tabelle 3.16 sind die experimentellen Bindungsdaten von 38 Komplexen aus der Literatur zusammengetragen. Die fünf in Tabelle 3.16 angegebenen Autoren berufen sich hierbei jeweils auf Original-Veröffentlichungen. Die experimentellen Bindungsenergien wurden von verschiedenen Arbeitsgruppen mittels kalorimetrischer Messungen bestimmt. Die genauen Methoden der Messung und der daraus resultierenden Messungenauigkeiten werden von den Autoren nicht erwähnt, obwohl sie die Daten verwenden.
Starke Schwankungen um Werte von 0.1 - 0.2 kcal/mol bei Insulin (4ins),

Chymotrypsion/OMTKY (1cho), Subtilisin Carlsberg/Inhibitor (1cse), Kalikrein/Inhibitor (2kai) und anderen bis hin zu über 10 kcal/mol für Trypsinogen/Inhibtor (2tgp) sind zu erkennen. Keine Arbeitsgruppe hat alle erhältlichen experimentellen Daten verwendet, die Arbeit von Jiang *et al.* (2002) stellt die umfassendste Untersuchung dar.

Komplex	Quelle				
	Horton[1]	Weng[5]	Wallqvist[4]	Vajda[3]	Jiang[2]
1cho	-15.7	-14.4	-14.6	-14.4	-14.6
1cse	-13.1	-13.1	-	-	-15.2
1cna	-	-9.7	-	-	-
1ppf	-	-13.5	-	-	-
1tec	-	14.0	-	-14.0	-14.2
2hfl	-14.2	-14.2	-	-	-14.2
2kai	-12.4	-12.5	-12.4	-12.5	-
2ptc	-	-18.1	-18.1	-18.1	-18.1
2sec	-13.1	-14.0	-13.1	-14.0	-13.1
2sni	-	-15.8	-15.0	-15.8	-15.8
3sgb	-14.7	-12.7	-	-12.7	-12.7
4sgb	-	-11.7		-11.7	-11.7
2tpi	-	-	-	-	-5.9
3tpi	-18.1	-17.3	-	-	-
4tpi	-	-17.3	- 17.7	-	-17.7
1tpa	-18.1	-	-17.8	-	-18.1
4cpa	-10.0	-	-	-	-
3cpa	5.3	-	-	-	-
4ins	-7.4	-	-	-	-7.5
1hbs	-4.8	-	-	-	-
2tgp	-	-	-18.2	-7.9	-
1axi	-	-	-	-	-10.7
1mel	-	-	-	-	-11.0
1nmb	-	-	-	-	-11.0
1fdl	-	-	-	-	-11.9
7hvp	-	-	-	-	-12.7
4htc	-	-	-	-	-14.6
3hfm	-	-	-	-	-12.6
1igc	-	-	-	-	-12.7
1abi	-	-	-	-	-11.9
1dvf	-	-	-	-	-11.0
1vfb	-	-	-	-	-10.5
1dkz	-	-	-	-	-9.0
1fle	-	-	-	-	-11.2
1cgi	-	-	-	-	-14.8

Tabelle 3.16: Experimentelle Bindungsenergien in kcal/mol. Die Daten sind [1] Horton & Lewis (1992), [2] Jiang *et al.* (2002), [3] Vajda *et al.* (1994b), [4] Wallqvist *et al.* (1995), [5] Weng *et al.* (1997) entnommen.

In Abbildung 3.35 sind die experimentellen Werte aus Tabelle 3.16 gegen E^{nf}_{gesamt}, $E^{nf}_{spezifisch}$ und E_{gesamt} aufgetragen. In keinem der drei Fälle ist eine linear Zusammenhang mit den experimentellen Werten bei Berücksichtigung aller Datenpunkte zu erkennen.



(a) E^{nf}_{gesamt} **(b)** $E^{nf}_{spezifisch}$

(c) E_{gesamt}

Abbildung 3.35: Korrelation experimenteller Bindungsdaten mit berechneten Präferenzen. Daten sind [1] Horton & Lewis (1992), [2] Jiang *et al.* (2002), [3] Vajda *et al.* (1994b), [4] Wallqvist *et al.* (1995), [5] Weng *et al.* (1996) entnommen.

4 Diskussion

"Now the time has arrived for the artist to come out
from behind his protective coloring of adopted
abstractions and indirections."

Buckminster Fuller (1947)

4.1 Kritische Betrachtung des Boltzmann-Ansatzes

Die Verwendung der normierten, gemittelten Paarverteilungsfunktion als Referenzzustand (Profil 2), wie sie von Gohlke *et al.* (2000) und anderen vorgeschlagen wurde, ist im vorliegenden Fall nicht die günstigste Wahl. In fast allen Docking-Studien erbrachte die Benutzung von Profil 1 eine deutliche Verbesserung gegenüber Profil 2.

Eine Interpretation und Verwendung von Profil 1 im Sinne des Boltzmannschen Verteilungssatzes ist nicht sinnvoll, wie im Folgenden gezeigt werden soll. Bei Benutzung der Boltzmann-Statistik läge dem Profil 1 die Annahme zugrunde, daß **jeder** Abstand zwischen zwei beliebigen Atomen gleich wahrscheinlich wäre. Der "Referenzzustand" entspräche dann einer Gleichverteilung aller Distanzen. Atomabstände von null Å sind jedoch unrealistisch, weil es zu einer Durchdringung der Atomhüllen kommt. Demnach müssten für eine Verwendung dieses Ansatzes die Atome als Punkte im Raum approximiert werden. Diese Annahme ist zu stark vereinfachend und widerstrebt den realen Gesetzmäßigkeiten. Bei Vernachlässigung von Normierungsfaktoren würde die Benutzung der Gleichverteilung im Boltzmann-Gesetz zum "Weglassen" des klassischen Referenzzustandes führen: $\Delta W = -\ln \frac{Paarverteilung}{Referenzzustand}$. Dies wiederum bedeutet, daß nur der negative natürliche Logarithmus der Paarverteilung dividiert durch einen konstanten Faktor (eins ohne Normierung, 1/40 mit Normierung auf die Anzahl an Atomtypen oder 1/8 wenn man den abgesuchten Raum von acht Å auf

eins normiert) verwendet wird. Ist hier ein Widerspruch zur Definition der Boltzmann-Statistik vorhanden und darf diese überhaupt verwendet werden? Die Antworteten sind erstens ein klares ja und zweitens ein klares nein.

Diese Art der Statistik ist keinesfalls für die hier verwendete Methode definiert. Wie an mehreren Stellen bereits erwähnt, liegt kein Ensemble gleicher Teilchen im thermodynamischen Gleichgewicht vor, die Statistik ist für dünn populierte Zustände nicht gültig, eine Boltzmann-ähnliche-Verteilung für Proteine ist umstritten, die Gültigkeit der Gruppenadditivität zwischen Kooperativität und Anti-Kooperativität ist noch strittiger und die Anwendbarkeit der radialen Verteilungsfunktionen als Wahrscheinlichkeitsdichten ist fraglich. Die Einführung eines "Referenzzustandes" kann in die gleiche Kategorie eingeordnet werden.

Ein solcher Zustand wird benötigt, weil der Boltzmannsche Verteilungssatz das Besetzungsverhältnis zweier Energieniveaus angibt. Wenn also zwei Zustände eines Systems bekannt wären, könnte man dieses Verhältnis bilden. Ist jedoch nur einer bekannt, wie im vorliegenden Fall im Strukturdatensatz, so muß ein weiterer Zustand "konstruiert" werden. Wie soll dieser "zu erwartende" Zustand aussehen, wie ist er beschreibbar? Als erste Annahme wurde die "Zufallsmischung" postuliert (Miyazawa & Jerningan (1985)). Diese Hypothese ging davon aus, daß Aminosäuren und Lösungsmittel einheitlich im zur Verfügung stehendem Volumen verteilt sind. Die Anzahl an Kontakten ist damit nur von der Konzentration der einzelnen Teilchen abhängig. Natürlich werden Alanin-Alanin-Kontakte häufiger gefunden werden, als Methionin-Methionin-Kombinationen. Dieses Modell mag für ein stark verdünntes Gas akzeptabel sein, aber kann damit ein Protein beschrieben werden? Sippl (1990) entwickelte ebenfalls ein Referenzsystem. Danach ist die fragliche Dichte eines bestimmten Atom-Paares in einem definierten Abstand abhängig von der Anzahl aller gefundener Paare. In einer Entfernung von 10 Å befinden sich mehr Kontakte, als bei 80 Å. Eine Mittelung über alle gefundenen Kontakte in jedem Abstandsintervall wird auch in vorliegender Arbeit im Profil 2 verwendet. Dieses Verfahren entspricht anschaulich einem mehrfachen Mitteln, bis von einer Zufallsverteilung ausgegangen werden kann.

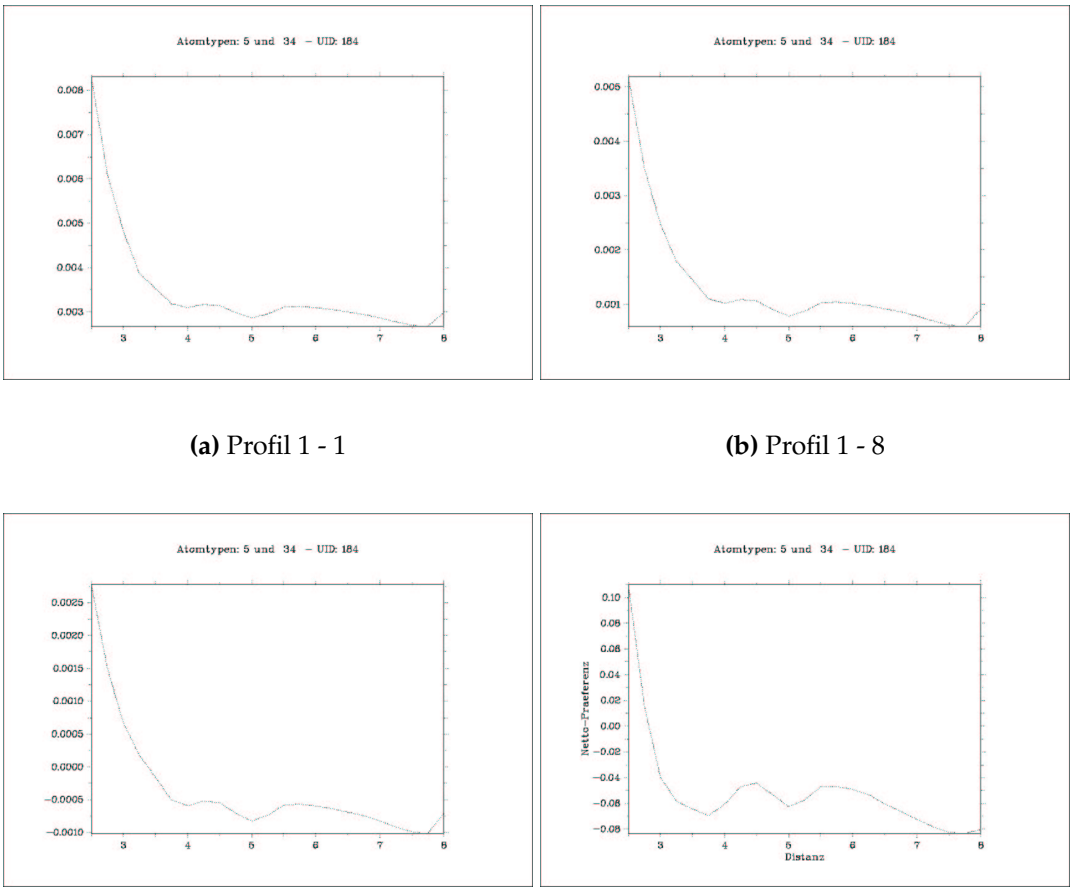
Ein über die Annahme einer Zufallsmischung hinaus gehendes Konzept stellt den Referenzzustand als willkürliche Sammlung von beliebigen Teilchen

ohne jedwede Interaktion dar. Es wird ein Zustand gesucht, in dem "die Paarverteilungsdichte mit einer Null-Energie korrespondiert" (Thomas & Dill (1996)). Die Gleichverteilung aller Abstände könnte dann unter großem Vorbehalt hinsichtlich der aufgeführten physikalischen Unsinnigkeiten (Durchdringung der Atomhüllen) ein System im energetischen "Nullzustand" beschreiben. Es sei angemerkt, daß es sehr unwahrscheinlich ist, so dichte Systeme wie Proteine überhaupt in einem Zustand ohne Wechselwirkungen anzutreffen.

Aus den soeben angeführten Gründen ist die Annahme, bei Profil 1 könnte von einem Referenzzustand gesprochen werden, nicht sinnvoll. Vielmehr wird deutlich, daß für die ausreichende Beschreibung der Verteilungen von interatomaren Abständen im Strukturdatensatz kein Referenzzustand benötigt wird. Daraus folgt aber auch, daß der gesamte Boltzmann-Ansatz nicht haltbar ist. Eine Exponentialfunktion, gewichtet mit einem konstanten Faktor, hier 0.025, hat sich nach den vorgestellten Ergebnissen als signifikant besser erwiesen, als die Verwendung der gemittelten Paarverteilung (Profil 2) im Boltzmann-Ansatz. Die Einführung einer empirischen Energiedichtefunktion, und nichts anderes ist hier bei Profil 1 geschehen, auf der Basis von additiven, gewichteten Termen ist nicht neu und bereits an anderer Stelle diskutiert (Dengler (1998)).

Im Folgenden soll die eingeführte Terminologie der Netto-Präferenzen und der Profile aus Gründen der Konsequenz beibehalten werden, obwohl für Profil 1 eine empirische Energiedichtefunktion und für Profil 2 der Boltzmann-Ansatz mit gemittelter Paarverteilung als Referenzsystem angenommen wird.
Es stellt sich die Frage wie die beiden unterschiedlichen Ansätze auf den Verlauf der Netto-Präferenzen einwirken. In Abbildung 4.1 ist dieser in Abhängigkeit vom Profil aufgetragen. Es wird deutlich, daß sich der allgemeine Kurvenverlauf von Profil 1 zu Profil 2 nicht stark ändert. Die Absolutwerte schwanken erwartungsgemäß, die Lage der Minima verschiebt sich jedoch nur beim Übergang von Profil 1 zu Profil 2, nicht bei Verwendung verschiedener Normierungsfaktoren innerhalb von Profil 1. Die insgesamt geringen Änderungen der Kurven werden verständlich, wenn nochmals der gemittelte Referenzzustand 2 (Abb. 3.3) betrachtet wird. Die Extrema liegen etwa 0.02 Einheiten auseinander (vergl. Abb.

3.3). In der Arbeit von Gohlke *et al.* (2000) wird eine sehr ähnlicher Referenzzustand mit einer Maximal-Differenz von etwa 0.035 erhalten.



(a) Profil 1 - 1

(b) Profil 1 - 8

(c) Profil 1 - 40

(d) Profil 2

Abbildung 4.1: Vergleich des Einflußes verschiedener Profile auf die Netto-Präferenz. (a) bis (c) entspricht Profil 1, erstens ohne Normierung, zweitens mit Normierung des Abstandes (8 Å) und drittens Normierung auf die Anzahl der Atomtypen (40). Ordinate: Distanz in Å , Abszisse: Netto-Präferenz.

Dieses äquivalente Ergebnis, obwohl aus zwei völlig verschiedenen Datensätzen gewonnen (Protein-Protein-Komplexe und Protein-Ligand-Komplexe) bestätigt die Annahme, daß eine mehrfache Mittelung zu einem Zufallszustand führt (Gohlke *et al.* (2000), Finkelstein & Gutin (1995)).

Die Distanz zwischen den Extrema der Verteilung des Referenzsystems liegt in der gleichen Größenordnung wie der Wert von Profil 1 (0.025). Bei Benutzung von Profil 2 als Referenzzstand wird durch einen Wert zwischen 0.01 und 0.028 geteilt, bei Profil 1 immer durch 0.025. Ähnliche Kurvenverläufe für Profil 1 und Profil 2 sind zu erwarten und werden auch beobachtet.

4.2 Docking-Bewertung

Die Grundannahme für eine Bewertung von Docking-Ergebnissen ist die Annahme, daß sich der native Komplex Nähe des thermodynamischen Minimums befindet (Ben-Naim *et al.* (1990)). Zunächst stellt sich die Frage nach einer Rechtfertigung dieser Annahme, denn aus experimenteller Sicht gibt es hierfür keinerlei Beweise.

Theoretisch scheint es sinnvoll, daß eine native Komplexstruktur in Abhängigkeit von der Art des Komplexes so stabil ist, daß sie eine bestimmte Zeitspanne lang nicht auseinander diffundiert. Permanente Komplexe als Extrembeispiel haben häufig sehr kleine Dissoziationskonstanten, weil die Einzelproteine in Lösung meist nicht stabil sind. Daraus läßt sich folgern, daß ein stabiler Komplex keine hohe, positive freie Gibbs-Energie haben sollte. Der Absolutwert ist von der Art des Komplexes und auch vom Einzelfall abhängig. Wird davon ausgegangen, daß der native Komplex in der Nähe des thermodynamischen Minimums zu finden ist, folgt für diese Arbeit, daß Strukturen mit hohen, positiven Werten als nicht-native, falsch-positive Lösung angesehen werden können. Der native Komplex sollte dann einen möglichst kleinen Wert haben. Diese Überlegung kann nicht als allgemeingültig aufgefasst werden, stellt also nur eine grobe Richtlinie für die Bewertung von gedockten Komplexen dar. Es ist zwar sehr wahrscheinlich, daß sich ein Protein-Protein-Komplex im thermodynamischen Minimum befindet wenn keine kinetische Hinderung vorhanden ist, aber der endgültige Beweis von experimenteller Seite steht wie schon erwähnt noch aus. Im Zweifel muß jeder Komplex individuell begutachtet und obige Annahme in keinem Fall als allgemeines Naturgesetz aufgefasst werden.

4.2.1 Vergleich mit Ergebnissen aus der Literatur

Der Vergleich mit den Ergebnissen anderer Arbeitsgruppen ist schwierig, weil zum einen in fast allen Fällen keine Einzelterme (geometrische Korrelation, Elektrostatik, Bindungsenergie und andere) gegeneinander aufgeführt werden und zum anderen die Docking- und Bewertungs-Methoden sehr unterschiedlich sind. Die Angabe von verschiedenen rmsd-Werten stellt das erste Problem bei einem Vergleich dar. Es werden Gesamt-rmsd-, *interface*-rmsd-, so genannten Basis-rmsd- und etliche andere rmsd-Werte berechnet. Des weiteren werden unterschiedliche Strukturen, wie ko-kristallisierte Komplexe oder die superpositionierte Struktur als Referenz bei der rmsd-Berechnung verwendet. Der direkte Vergleich von Docking-Ergebnissen ist daher schwierig. Die Einigung auf gewisse Standards soll der momentan laufende CAPRI-Wettbewerb erbringen. Die folgenden Vergleiche mit den Arbeiten anderer Gruppen sind qualitativer Natur.

Weiterhin ist anzumerken, daß in dieser Arbeit nur die geometrische Korrelation zum internen Vergleich verwendet werden kann. Die benutzte Programmversion von ckordo stellte zum Zeitpunkt der Arbeit keine weiteren Funktionalitäten zur Verfügung. In der Zwischenzeit ist nicht nur die geometrische Korrelation um mindestens eine Größenordnung verbessert worden, es sind auch weitere Kriterien, wie Elektrostatik und Hydrophobizität berücksichtigt worden. Eine sinnvolle und komplexere Kombination der einzelnen Bewertungsmaßstäbe, als die in Kapitel 3.7.4 vorgestellte wird derzeit mittels *support vector machines* untersucht (Zimmermann (2002)). Es ist zu erwarten, daß durch die Einbeziehung des hier vorgestellten Potentials die Gesamt-Bewertung der Docking-Ergebnisse erheblich verbessert wird.

Der Übersicht halber faßt Tabelle 4.1 einige literaturbekannte Potentiale und deren Parameter zusammen. Eine Vielzahl mehr sind bekannt, eine Auflistung würde den Rahmen dieser Arbeit sprengen. Leider sind auch die Potentiale aus der Literatur nur schwer direkt vergleichbar. Sie sind für unterschiedliche Fragestellungen entworfen worden, stammen aus dem Protein-Ligand-Docking

oder sind auf einen linearen Zusammenhang mit experimentellen Bindungsdaten optimiert. Eines der wenigen Potentiale, das ähnlich wie in dieser Arbeit Protein-Protein-Komplexe bewertet, stammt von Moont *et al.* (1999).

Art des Potentials	Struktur-datensatz	Sonstiges	Literatur
Atome	30	HIV-Inhibitoren Desolvatation, Konformationsänderungen	Verkhivker & Rejto (1996)
Atome	38	Kalibrierung anhand experimenteller Daten	Wallqvist *et al.* (1995)
Aminosäuren	17/109	Protein-Ligand-Komplexe	DeWitte & Shaknovich (1996)
Atome (40)	90	Protein-Ligand-Komplexe	Mitchell *et al.* (1999b)
Atome (5)	1376	Protein-Ligand-Komplexe	Gohlke *et al.* (2000)
Aminosäuren	11/23/385	Bewertung von gedockten Strukturen; naccess	Moont *et al.* (1999)
Atome (40)	147	Untersuchung von Protein-Protein-Interaktionen	Melo & Feytmans (1997)
Atome (2)	126	Protein-Protein-*interface*-Analyse Verbesserung von Moonts Methode (Moont *et al.* (1999))	Robert & Janin (1998)
Atome (4)	191	Protein-Protein-Komplexe Wasserstoffbrücken berücksichtigt	Jiang *et al.* (2002)
Atome (4)	191	Protein-Protein-Komplexe	Lu & Skolnick (2001)

Tabelle 4.1: Vergleich der Parameter einiger Potentiale aus der Literatur.

Tabelle 4.2 zeigt einen Vergleich einiger Ergebnisse aus dieser Arbeitsgruppe und der vorliegenden Arbeit. Das Docking wurde ebenfalls nach der Fourier-Korrelations-Methode (ftdock) durchgeführt. Die Bewertungskriterien von ftdock sind nicht einzeln aufgeschlüsselt. Sie bestehen aus geometrischer und elektrostatischer Komplementarität und beinhalten biologische Informationen zur Filterung (Gabb *et al.* (1997)).

Moont et al. testeten mehrere Paar-Potentiale und Strukturdatensätze. Sie kommen zu dem Schluss, daß ein Aminosäure-Aminosäure-Potential, erstellt aus einem Domänendatensatz (385 Domänen aus SCOP Murzin *et al.* (1995)) die beste Wahl darstellt. Erstaunlicherweise zeigt dieses auf **intra**molekularen, anstatt auf intermolekularen Paarungen basierendem Potential wesentlich bessere Ergebnisse, als die ebenfalls erzeugten intermolekularen Potentiale aus Homo-(11) und Heterodimeren (23). Von den Autoren wird die These aufgestellt, daß die benutzten Strukturdatensätze von elf bzw. dreiundzwanzig Komplexen für die intermolekularen Paarwechselwirkungen zu gering sind.

Komplex	Rank geom	Rank E_{gesamt}	ftdock	Pair-Potential
2ptc	790/9327	149/9327	12/205	14/205
1cho	3239/22006	153 /22006	11/85	6/85
2kai	1269/21155	323/21155	128/349	13/349
2sni	8858/21322	8968/21322	8/26	1/26

Tabelle 4.2: Vergleich der Bewertung einiger Systeme mit den Ergebnissen von ftdock (Spalte vier und fünf). (Orginaldaten und Beschriftung sind Moont *et al.* (1999) entnommen).

Anhand der Docking-Ergebnisse ist zu erkennen, daß für diese Arbeit und die Arbeit von Moont et al. die Einführung eines Potentials eine deutliche Verbesserung der Vorhersage bewirkt. In einem Fall, dem System Trypsin/Inhibitor (2ptc) führt die Benutzung eines Paar-Potentials bei ftdock jedoch zu einer Verschlechterung der Reihenfolge. Für die hier präsentierte Methode gilt dies nicht. Als Gegenbeispiel ist der Komplex Subtilisin Novo/Inhibitor (2sni) anzuführen, der in den hier gemachten Vorhersagen schlechter abschneidet. Die Unterschiede in der Anzahl möglicher Komplexe in Tabelle 4.2 läßt sich daraus zurückführen, daß die Komplexe von Moont et al. bereits gefiltert und vorbewertet sind. Daraus ergibt sich eine kleinere Anzahl an zu bewertenden Möglichkeiten (jeweils zweiter Wert in der Tabelle). In dieser Arbeit ist dies nicht der Fall und die Gesamtanzahl ist entsprechend größer. Ein quantitativer Vergleich des Ausmaßes der Verbesserung ist nicht möglich.

Einen anderen Ansatz zur Beschreibung von Interaktionen verfolgen Chen & Weng (2002). Sie haben die meisten der sechzehn in dieser Arbeit benutzten Komplexe ebenfalls anhand einer Fourier-Docking-Technik (Zdock) untersucht. In Tabelle 4.3 sind die jeweils gefundenen Ränge der nativ-ähnlichen Strukturen beider Arbeiten gegenübergestellt. Chen et al. verwenden die von Zhang *et al.* (1997) entwickelten *atomic contact energies*, basierend auf achtzehn Atomtypen. Die *atomic contact energies* geben die Desolvatationsenergie oder auch freie Energie wieder, die benötigt wird einen Atom-Wasser-Kontakt durch einen Atom-Atom-Kontakt zu ersetzen. Es werden also fast ausschließlich Desolvatationseffekte berücksichtigt, ein Paapotential wird nicht verwendet. Der maximale

Abstand von *interface*-Atomen in der Arbeit von Chen et al. beträgt 6 Å und der Raum wird im Docking mit 15 Grad abgesucht. Eine Lösung ist eine Struktur, deren *interface*-Cα-rmsd kleiner als 2.5 Å ist. Die verwendeten Parameter des Dockings sind demnach gut vergleichbar.
Bei Betrachtung von Tabelle 4.3 wird deutlich, daß mehr nativ-ähnliche Komplexe von ckordo im Vergleich zu Zdock bei alleiniger Betrachtung der geometrischen Korrelation gefunden werden. In einigen Fällen wird übereinstimmend gar keine Lösung unter den ersten 2000 Plätzen von beiden Programmen gefunden. Dazu gehört z.B. der Komplex Peroxidase/Cytochrome C (2pcc).

Komplex	Anzahl an Hits unter den ersten 1000			
	$corr_{geo}$	E_{gesamt}	S_{SC}	$\alpha S_{SC} + S_{DS}$
1mah	4	5	0	16
1brb	3	7	–	–
2ptc	1	3	0	21
1cho	2	4	1	22
1fss	1	5	0	0
1ppf	0	2	–	–
1ugh	4	6	10	3
1cgi	0	1	1	55
2kai	1	3	0	20
2sic (2st1)	1	1	0	112
1brs (1a2p)	0	2	13	0
1brs (1bao)	0	1	–	–
1bvn	0	0	–	–
2sni	0	0	0	43
1brc	0	0	–	–
2pcc	0	0	0	0

Tabelle 4.3: Vergleich der gefundenen richtig-positiven Strukturen bei Verwendung verschiedener Auswahlkriterien. Spalte eins ist die geometrische Komplementarität (ckordo), Spalte zwei die hier berechnete Bindungs-Präferenz (Profil 1), Spalte drei die geometrische Korrelation und Spalte vier gibt die geometrische Komplementarität plus der Desolvatationsenergie an (Originaldaten und Beschriftung für Spalte drei und vier übernommen aus Chen & Weng (2002)).

Bei Bewertung mit der Bindungs-Präferenz wird in zwölf von sechzehn Versuchen mindestens ein nativ-ähnlicher Komplex gefunden. Die Sortierung von Chen & Weng (2002) anhand der Desolvatationsenergien plus der geometrischen Komplementarität identifiziert jedoch etliche Komplexe mehr. Hier könnte ein Hinweis dafür liegen, daß der solvensabhängige Term des vorgestellten Potentials die Desolvatationsenergie nur mäßig beschreibt, wohingegen der entsprechende Term in der Arbeit von Chen et al. ausschließlich auf Solvenseffekte ab-

gestimmt ist. Weiterhin ist der Vergleich der reinen Netto-Präferenz dieser Arbeit mit einer kombinierten Wertung von Geometrie und Desolvatation von Chen et al. nur qualitativ möglich. Ein Vergleich der geometrische Korrelation von ckordo und Zdock zeigt, daß ckordo in acht von sechzehn Fällen eine Struktur findet, Zdock jedoch nur in drei von neun aufgeführten. Die von Chen et al. ebenfalls integrierte elektrostatische Korrelation konnte zum Zeitpunkt der Arbeit noch nicht getestet werden. Diese Ergebnisse deuten abermals darauf hin, daß eine sinnvolle Kombination der Bindungs-Präferenzen mit weiteren Kriterien zu einer deutlichen Verbesserung der Bewertung führen wird.

4.2.2 Das Ein-Teilchen-Potential

In Kapitel 3.4 auf Seite 58 sind die Verteilungen der vom lösungsmittelabhängigen Ein-Teilchen-Potentiale dargestellt und an einigen Beispiele diskutiert.
Daraus wird deutlich, daß die Wahl der Definition des *interfaces* für diese Potential-Art nicht geeignet ist. Durch den maximalen Abstand von 8 Å werden Atome der inneren Schichten berücksichtigt. Diese haben jedoch keinen Anteil an der Desolvatation. Es resultiert die in allen Abbildungen des Kapitels 3.4 präsente Überbewertung des vergrabenen Zustandes. Selbst polare Atome wie Stickstoff von Arginin und Sauerstoff von Glutamin und Asparaginsäure zeigen eine zu starke Begünstigung beim Übergang vom exponiertem zum vollständig verdecktem Zustand. Es wäre günstiger für diesen Teil des Potentials das *interface* als die Differenz von der Oberfläche des gesamten Komplexes und der Einzelproteine zu berechnen. Daraus ergibt sich jedoch die Schwierigkeit, daß die *interface*-Atome für den kontaktabhängigen Teil anders definiert wären, als die für den solvensabhängigen. Dies zu vermeiden wurde eingangs (vergl. Kapitel 2.3 eine Definition für beide Potentialterme verwendet, da die Auswirkung sowohl von Paarwechselwirkungen als auch Desolvatation von ein und dem selben Atom betrachtet werden sollten. Als Mittelweg und einfache Näherung könnte man für jedes gefundene *interface*-Atom prüfen, ob es wirklich an der Oberfläche liegt und es sonst ignorieren. Die Ergebnisse aus den Docking-Studien in Kapitel 3.7 zeigen, daß die Einführung eines Solvens-Terms eine allgemeine Verbesserung der Vorhersagekraft des Potentials bewirkt. Die Größe dieses Effektes ist jedoch

nicht so hoch wie erwartet. Das läßt darauf schließen, daß die Beschreibung von entropischen Kräften durch die Verwendung des entwickelten Ein-Teilchen-Potential noch nicht ausreichend ist.

4.2.3 Verbesserungen durch Einführung funktioneller Gruppen

In Kapitel 3.7.5 auf Seite 77 wurde die Auswirkungen der Einführung funktioneller Gruppen auf die Vorhersage-Ergebnisse im Docking dargestellt und diskutiert. Kapitel 3.9 auf Seite 91 stellt die Resultate entsprechend für die Decoys vor und interpretiert sie. Zusammenfassend verbessert die Einführung von funktionellen Gruppen und die gesonderte Gewichtung der Atome die Vorhersagen in beiden Fällen signifikant. Aus der Bewertung der Docking-Studien hat sich aber auch gezeigt, daß die Verbesserung mit der Verwendung eines bestimmten Profils einhergeht. Profil 1 führt zu einer Entzerrung der Bindungs-Präferenzen, während die entsprechenden Werte bei Profil 2 dicht beieinander liegen (vergl. Abb. 3.23 auf Seite 80). Für das Profil 2 ist im Mittel sogar eine Verschlechterung bei Verwendung funktioneller Gruppen zu verzeichnen. Dieser Widerspruch könnte darin begründet sein, daß die durchgeführte iterative Parameteroptimierung für Profil 2 nicht optimal ist. Für beide Profile finden sich sehr unterschiedliche Absolutwerte für die drei Gewichtungsfaktoren.

Chemisch betrachtet ist die Verwendung der funktionellen Gruppen sinnvoll. Die meisten Reaktionen laufen unter Beteiligung stärker polarisierter oder polarisierbarer Atome ab, wie aus zahllosen Namensreaktionen bekannt ist. Häufig bilden wenige Interaktionen zwischen zwei Proteinen den Hauptbeitrag zur Bindungsenergie. Seit langem bekannt ist das Beispiel der Sichelzellenanämie. Eine einzige Mutation reicht aus, daß sich Hämoglobin-Moleküle zusammenlagern und den Blutfluß behindern. Bogan & Thorn (1998) zeigten an zweiundzwanzig Alanin Mutanten, daß wenige Aminosäuren, so genannte *hot spots* im *interface* einen signifikanten Anteil an der Bindungsenergie haben. Diese *hot spots* sind in der Mitte der Interaktionsfläche lokalisiert. Ein "Dichtungsring"-ähnlicher Kreis von energetisch unwichtigen Aminosäuren schirmt sie hermetisch gegen Lösungsmittel ab. Die *hot spots* sind komplementär zueinander auf den beiden Proteinen angeordnet. Es ist naheliegend, daß die Atome der funktionellen

Gruppen dieser *hot spots* hauptsächlich zu einer Bindung beitragen. Ob die in Tabelle A.2 auf Seite 122 angegebenen Atome die optimale Auswahl darstellen, ist nicht untersucht worden. So fehlen beispielsweise alle Hauptkettenatome, wie Carbonylkohlenstoff und Stickstoff, obwohl auch reaktiv sind und entscheidend zu einer Bindung beitragen können. Eine weitergehende Untersuchung wäre sinnvoll.
Hu (2000) bestätigte die strukturelle Konservierung einiger weniger polarer Aminosäuren, die von einem Ring hydrophober Aminosäuren umgeben sind. Interessanterweise zeigen Antikörper-Antigen-Komplexe bei einem Sequenzvergleich weniger strukturell konservierte Residuen, als Homodimere, Heterodimere und Enzym-Inhibitor-Komplexe. Die Ergebnisse einer detaillierten Analyse von Serin-Protease-Inhibitor- und Antikörper-Antigen-Komplexen weisen in die selbe Richtung (Jackson (1999)). Wechselwirkungen zwischen Atomen der Proteinhauptkette bilden den größten Beitrag in Serin-Proteasen. Sie zeigen ein klares Bindungsmotiv in der Hauptkette. In Antikörper-Antigen-Komplexen hingegen überwiegen die Seitenketten-Kontakte. Die notwendige Diversität in Antikörpern verbietet eine Erkennung über die Hauptketten-Geometrie. Der "Muster"-Antikörper wäre zu unflexibel für die Erkennung hoch divergenter Antigene.

Diesen Ausführungen zufolge, ist eine deutliche Verbesserung der Bewertung von Serin-Proteasen bei Verwendung funktioneller Gruppen nicht zu erwarten. Dieser scheinbare Widerspruch wird aufgelöst, wenn in Erinnerung gerufen wird, daß lediglich eine Gewichtung funktioneller Gruppen vorgenommen wird. Natürlich werden die Netto-Präferenzen über alle Atom-Atom-Kontakte berechnet, einige bestimmte Kombinationen jedoch hervorgehoben. Werden diese Kontakte im zu bewertenden Komplex gar nicht beobachtet, so ist die höhere Netto-Präferenz dieses Atom-Atom-Kontaktes unerheblich. Weiterhin wird die Bindungsenergie in Serin-Proteasen vermutlich nicht ausschließlich über Kontakte der Hauptkette bestimmt, sondern auch durch Interaktionen polarer Atome. In diesem Fall würde die Verwendung funktioneller Gruppen zu einem besseren Ergebnis führen. Es ist zu untersuchen, ob es insbesondere für Proteasen zu einer weiteren Verbesserung der Vorhersage kommt, wenn auch

Atome der Hauptkette berücksichtigt werden.

Des weiteren sind Antikörper-Antigen-Komplexe, die nach obigen Ausführungen besonders gut vorherzusagen sein sollten, gar nicht berechnet worden, weil das Programm ckordo in allen untersuchten Fällen keine native Lösung gefunden hat. Die geometrische Komplementarität dieser Komplexe ist, wie eingangs beschrieben, zu gering. Dies steht im Einklang mit der Annahme, daß die Erkennung hauptsächlich über Seitenketten gestaltet wird, die im Docking-Prozess naturgemäß schwierig zu modellieren sind. Weiterhin ist völlig ungewiss, ob die Bewertung dieser Komplexe nicht besser wäre als bei Proteasen, so wie es nach obiger Argumentation anzunehmen wäre.
Zusammenfassend kann die Verwendung funktioneller Gruppen im vorgestellten Potential als Akzentuierung verstanden werden. Die schon mehrfach angesprochene Kombination von geometrischer Komplementarität und statistischen Netto-Präferenzen ist unter diesem theoretischen Gesichtspunkt ebenfalls sinnvoll. Die beiden Kriterien können sich ergänzen, nicht aber ersetzen. Ein Indiz für die Richtigkeit dieser Theorie sind die Auftragungen von Bindungs-Präferenzen und geometrischer Korrelation im Ergebnisteil (vergl. Kapital 3.7.4 auf Seite 75). Sie zeigen keine lineare Abhängigkeit. In manchen sehr drastischen Fällen, wie dem Paradebeispiel 1cgi sieht man die großen Unterschiede in der Bewertung durch die geometrische Komplementarität und der Bindungs-Präferenz. Die geometrische Komplementarität ist für diesen Komplex extrem schlecht, die Bindungs-Präferenzen zeigen hingegen Cluster um den nativen Komplex herum (O. Zimmermann, persönliche Mitteilung, 2002).

4.2.4 Einführung von Wasserstoffbrücken

Wasserstoffbrücken wirken stabilisierend auf Protein-Protein-Komplexe und können einen großen Anteil an der Bindungsenergie haben (Wallqvist *et al.* (1995)). In den sechzehn hier untersuchten Komplex-Systemen sind im Mittel neun solcher Bindungen im *interface* zu finden. Es ist nicht gewährleistet, daß diese Art von Bindung im Strukturdatensatz adäquat repräsentiert wird. Da im Mittel um die neun Wasserstoffbrücken pro Komplex gefunden werden und

der Datensatz nur 584 Komplexe enthält, werden gerade mal ca. 5256 solcher Wechselwirkungen gefunden. Diese Anzahl ist nicht ausreichend jede Art von möglichem Donor-Akzeptor-Paar in zahlenmäßig ausreichender Weise darzustellen. Es kann zu einer Unterbewertung einzelner Wasserstoffbrücken kommen. Die Einführung eines besonderen Gewichtungsterm ist daher sinnvoll.
Wasser-vermittelte Interaktionen im *interface* machen einen Anteil von 25% an der Bindungsenergie von Komplexen aus und nehmen Einfluss auf die spezifische Erkennung. Im Mittel sind achtzehn Wasser-Moleküle im *interface* zu finden und sind demnach zahlreicher als reine Wasserstoff-Brücken (Covell & Wallqvist (1997)). Im Protein-Protein-Docking wird Wasser in der Bindestelle jedoch vernachlässigt, weil *a priori* weder die Bindestelle selbst noch die Anzahl in ihr enthaltener Wassermoleküle bekannt ist. Dies kann in einigen Fällen zu deutlich schlechteren "Energien", oder Bindungs-Präferenzen einer Komplex-Struktur führen.

4.2.5 Einfluss von Gewichtung und Glättung

In Kapitel 3.5 ist der Einfluß der Gewichtung von Häufigkeiten auf die Netto-Präferenzen dargestellt. Die Gewichtung wurde eingeführt um statistisch unterrepräsentierte Beobachtungen auszugleichen. Der allgemeine Verlauf der Verteilung ändert sich nicht, jedoch die Akzentuierung. Extrema werden abgeschwächt und somit eine Glättung herbeigeführt. Dies entspricht der Annahme möglichst weiche, oberflächlichen Netto-Präferenzen zu entwerfen und Details erst später einzubringen, wie beispielsweise die Einführung funktioneller Gruppen, die erst nach Erzeugung aller Netto-Präferenzen als Gewichtungsfaktor berücksichtigt wird. Die Glättung der Häufigkeiten über eine Trapezfunktion gehört in die gleiche Kategorie wie die Gewichtung der Rohdaten. Sie gleicht den Nachteil diskreter und ungenauer Daten aus. Flexibilitäten werden dadurch besser modelliert, da ungenaue Atompositionen nicht zu einer sofortigen Verschlechterung des Potentials führen. Diese Eigenschaft ist besonders bezüglich flexibler Seitenketten in der Bindestelle sehr wichtig. Lange senkrecht zur Proteinoberfläche ins Lösungsmittel stehende Aminosäuren bereiten bei der geometrischen Korrelation, wie mehrfach erwähnt, enorme Probleme. Das vorgestellte Potential hingegen

ist aufgrund sehr weicher Netto-Präferenz-Verteilungen robust gegen sterische Ungenauigkeiten. In einer Auftragung dieser Eigenschaft als Hyperfläche, würde die Extrema umgeben von relativ flachen Hängen zeigen. Die Steigungen wären entsprechend klein. Dies gilt nicht für die geometrische Korrelation, bei der oft Verschiebungen um wenig Å ausreichend sind aus der schmalen Zone guter Korrelation herauszukommen (O. Zimmermann, persönliche Mitteilung, 2002).
Ein Vergleich der Trapez-Funktion mit der alternativ zur Verfügung stehenden Dreiecks-Funktion zeigt nur geringe Unterschiede zwischen beiden (Ergebnisse nicht gezeigt), so daß eine Implementierung der Dreiecks-Funktion nach Gohlke *et al.* (2000) ebenfalls zur Verfügung stehen würde.

4.2.6 Die optimale Atomtypen-Anzahl

In den Kapiteln 3.3 und 3.6 sind die Korrelationen der erzeugten Netto-Präferenzen untereinander anhand des Korrelationskoeffizienten nach Pearson und nach Hodgkin dargestellt. Es wird deutlich, daß einige lineare Korrelationen zwischen einzelnen Verteilungen vorhanden sind. Besonders die Netto-Präferenzen sehr häufig vorkommender Atome, wie Kohlenstoff und Sauerstoff der Hauptkette zeigen Abhängigkeiten voneinander. Trotzdem sind diese Korrelationen so komplex und verwoben, daß nicht einfach einige Atomtypen weniger verwendet werden können. Die Atome und damit das Auftreten ihrer Kontakte sind nicht paarweise abhängig oder unabhängig voneinander. Sind z.B. die Verteilungen von C und O (beide Hauptkette) und C und N (beide Hauptkette) sehr ähnlich, so kann dies nicht dazu führen beide mit einem Term zu beschreiben. Der Zusammenhang zu anderen Kontakten, wie N (Hauptkette) und N (Prolin) wäre nicht mehr korrekt wiedergegeben. Wegen dieser komplexen Abhängigkeiten ist eine Reduktion der Dimensionalität nicht so einfach möglich. Die Evaluierungen mit der Hauptkomponentenanalyse zeigen zwar die Abhängigkeiten, weitere Rechnungen mit der so genannten Faktormatrix nach auf der Heyde (1990) bringen jedoch in allen Versuchen eine Verschlechterung der Docking-Bewertungen. Die Gründe hierfür sind unklar, könnten aber auch in der Wahl der Methode zur Untersuchung von Abhängigkeiten liegen. Einer Untersuchung der optimalen Atomtypen-Anzahl sollte die Analyse aller möglichen Methoden und Algorith-

men zur Reduktion von Dimensionalitäten, hinsichtlich ihrer Anwendbarkeit und zu erwartender Ergebnisse vorausgehen. Gleiches gilt für die Suche eines guten Verfahrens zur Optimierung der 18040 (820 Atomtypenkombinationen bei jeweils 22 Abstandsintervallen) justierbarer Parameter, wobei funktionelle Gruppen und Wasserstoffbrücken noch nicht berücksichtigt sind.

4.3 Decoy-Bewertung und Korrelation mit experimentellen Bindungsdaten

Es gibt zwei klassische Methoden ein neu entwickeltes Potential zu testen. Die erste Möglichkeit ist die Bewertung von decoy-Strukturen hinsichtlich der Unterscheidung nativer von nicht-nativer Anordnungen. Das zweite Verfahren ist die Korrelation der Bindungs-Präferenzen von Komplexen mit den experimentellen Bindungsdaten. Beide Methoden sind verwendet und die Ergebnisse in den Kapiteln 3.9 und 3.10 dargestellt worden. Zusammenfassend kann gesagt werden, daß diese Ergebnisse deutlich schlechter ausfallen, als nach den sehr guten Vorhersagen im *unbound*-Docking zu erwarten gewesen wäre. Es stellt sich die Frage nach dem Warum. Die beiden Methoden werden daher im Folgenden näher betrachtet.
Ein kritischer Punkt in der Bewertung von decoy-Anordnungen ist das Verfahren mit der diese Strukturen erzeugt werden. Ein optimaler Datensatz sollte sehr genau die zu klärende Problemstellung beschreiben, also möglichst gute falsch-positive Strukturen aus Docking-Versuchen beinhalten. Es sei daran erinnert, daß ein decoy-Datensatz im Grunde genommen nichts anderes ist, als ein Docking, bei dem allerdings nur wenige der vom Programm vorgeschlagenen Komplexe ausgewählt werden, wohingegen in den hier gemachten Docking-Experimenten mit fünfzehn Grad bis zu 22006 Lösungen gefunden werden. Die verwendeten decoy-Strukturen stammen alle aus derartigen Simulationen, sind jedoch mit verschiedenen Programmen erzeugt worden, was allerdings aufgrund der universellen Anwendbarkeit des entwickelten Potentials unerheblich sein sollte. Der einzige Unterschied zu den hier gemachten Docking-Untersuchungen ist die Tatsache, daß die decoys nicht selber erstellt wurden. Sie genügen offensichtlich

nicht den Anforderungen an einen guten Test- bzw. Dockingdatensatz.
Abschließend sei darauf hingewiesen, daß die Ergebnisse in der decoy-Bewertung nur im Vergleich mit den guten Resultaten der *unbound*-Docking-Studien aus Kapitel 3.7 schlechter sind als erwartet, relativ zu anderen Arbeitsgruppen jedoch nicht. Camacho *et al.* (2000) untersuchten einige decoy-Datensätze und erhielten insbesondere für nicht-minimierte Strukturen ähnlich diffuse Verteilungen wie in dieser Arbeit auch gefunden werden (vergl. Abb. 3.33). Im Gegensatz zu den Arbeiten von Camacho et al. wird mit dem hier entwickelten Potential der richtig-positive Komplex trotzdem auf dem ersten Platz gefunden. Nach einer vorherigen Minimierung der Strukturen wird die nativ-ähnliche Anordnung auch von Camacho et al. auf dem ersten Platz gefunden. In der vorliegenden Arbeit ist jedoch nicht mit minimierten Strukturen gearbeitet worden. An diesen Beispielen wird die geringe Vergleichbarkeit und Aussagekraft der Bewertung von decoy-Strukturen deutlich, insbesondere wenn Herkunft und Erzeugungsmethode nicht sehr genau beachtet werden.

Die Ergebnisse für die zweite gängige Validierungs-Methode, dem Vergleich mit experimentellen Daten, bleiben ebenfalls hinter den Erwartungen zurück. Es kann keine lineare Korrelation zwischen experimentellen und berechneten Werten festgestellt werden (vergl. Kapitel 3.10). In Tabelle 3.16 sind die Bindungsenergien in kcal/mol aus verschiedenen Literatur-Quellen zusammengetragen. In keinem der Fälle haben die Autoren den experimentellen Fehler oder die genaue Messmethode angegeben. Es wird vermutet, daß eine kalorimetrische Methode wie die isothermale Titrationskalometrie (Pierce *et al.* (1999)) verwendet worden ist. Die Autoren berufen sich auf Primärliteratur oder aufeinander. Damit ist schwer zu beurteilen, wie stark die Werte schon aufgrund von Messungenauigkeiten schwanken. Eine genaue Fehlerabschätzung ist nur bei Kenntnis der Messfehler möglich. Weiterhin fällt auf, daß die Bindungsdaten aus der Literatur untereinander Abweichungen von 0.1 - 0.2 kcal/mol aufweisen. Diese vermeintlich kleine Schwankung wird sehr groß gegenüber den viel kleineren Abständen in den Bindungsenergien zweier nur geringfügig verschiedener Strukturen. In dieser Arbeit sind die Unterschiede in den Netto-Präferenzen von einem Komplex zu einem anderen sehr ähnlichen mitunter erst in der zweiten

Nachkommastelle zu sehen. Die von den zitierten Arbeitsgruppen gefundenen linearen Korrelationen müssen wegen fehlender Messungenauigkeiten als kritisch betrachtet werden. Abgesehen davon wird in vier von fünf Fällen nicht einmal die Hälfte aller vorhandener Bindungsdaten verwendet. Die Gründe hierfür sind unklar. Jiang *et al.* (2002) haben mit Abstand die meisten Komplexe untersucht und finden eine lineare Korrelation mit den von ihnen berechneten Potentialwerten. Eine Anwendung auf Docking-Studien ist leider nicht vorgenommen worden. In der hier vorgestellten Arbeit wurde der Hauptakzent auf die Docking-Simulationen gelegt. Die Korrelation mit den experimentellen Daten ist - wie erwähnt - nicht sehr hoch. Es ist zwar denkbar das Potential anhand der zu justierenden Paramter auf eine lineare Korrelation mit den wenigen Messwerten zu optimieren, ist aber für hier vorliegende und diskutierte Problemstellung nicht sinnvoll (vergl. Kapitel 1.6). Die Beweiskraft dieser beiden Validierungsmethoden wird aus oben angeführten Gründen als kritisch betrachtet. Der beste Test - wenn auch zeitintensiver - stellt die Erprobung am "Ernstfall" dar, also die Betrachtung von *unbound*-Protein-Protein-Docking-Studien.

4.4 Ausblick

Die im Verlauf der gesamten Arbeit immer wieder angemerkte Nicht-Anwendbarkeit der Boltzmann-Statistik auf eine Sammlung von dreidimensionalen Proteinstrukturen führte in der Diskussion schließlich zu einer Abkehr von der physikalisch äußerst schwach untermauerten Sicherheit dieses Verteilungsgesetzes. Es stellt sich daher generell die Frage wie eine Funktion beschaffen sein muß, um die Bindungseigenschaften von Proteinen anhand von Häufigkeitsverteilungen in einer Datenbank zu beschreiben und diese Information in geeigneter Weise auf eine neue Problemstellung zu projizieren. Es bleibt zu untersuchen, ob die in dieser Arbeit entwickelten Exponential-Funktion bereits das Optimum darstellt oder mittels probater Verfahren eine Verbesserung möglich ist.

Die Verwendung der *accessible solvent area* ist nach neueren Erkenntnissen nicht

die beste Wahl (Rank & Baker (1997), Shimizu & Chan (2002)). Simulationen und Vergleiche des solvensabhängigen Teils des *potentials of mean force* (PMF) bei Verwendung der ASA und der MSA (engl. *molecular surface area*) zeigen deutlich, daß die ASA nicht in der Lage ist die Desolvatations-Barriere wiederzugeben. Eine Verwendung der MSA anstatt der ASA würde demzufolge den solvensabhängigen Teil des Potentials genauer beschreiben und könnte auch bessere Vorhersagen zumindest in dem Bereich um 5 Å bewirken (Rank & Baker (1997)). Systematische Untersuchungen des Unterschiedes von ASA und MSA könnten die Verwendbarkeit für die Bewertung von Docking-Ergebnissen zeigen.

Eine weitere lohnenswerte Untersuchung ist die Verwendung einer anderen oder der Kombination mehrerer *interface*-Definitionen, damit nur an der Oberfläche liegende Atome berücksichtigt werden. Es ist zu testen, ob die gleiche Definition auch für den kontaktabhängigen Teil des Potentials verwendet werden kann. Die Betrachtung reiner Oberflächenatome läßt eine akzentuiertere Verteilung der Oberflächen-Präferenzen vermuten. Die Überbewertung der Vergrabung im komplexgebundenen Zustand wäre damit behoben. Alternativ ist zu testen, ob nicht eine einfache Überprüfung der Oberflächenzugänglichkeit der Atome bei der momentan verwendeten Definition ausreichend wäre, nur Atome an der Oberfläche zu berücksichtigen.

Ein andere Ansatz zur Verbesserung des Potentials ist die Erzeugung spezieller Strukturdatensätze. Soll beispielsweise eine Serin-Protease gedockt und bewertet werden, so ist die alleinige Verwendung von Serin-Proteasen als Strukturdatensatz aufgrund des in Kapitel 2.5.1 beschriebenen Gedächtnisses sinnvoll. Die Entscheidung welcher Strukturdatensatz zu verwenden sei, ließe sich zumindest für die Serin-Proteasen automatisieren, da sie leicht an ihrer katalytischen Triade von jedem Motivsuch-Programm erkannt werden können, gleiches gilt für die Antikörpern. Sollen zwei unbekannte Proteine gedockt werden, könnte eine schnelle Motivsuche bestimmen welche Art von Komplex vorliegt und so den passenden Strukturdatensatz für die Erzeugung der Netto-Präferenzen verwenden. Es wird erwartet, daß die Bewertung mit der Einführung spezieller Strukturdaten deutlich verbessert werden kann.

Zur Evaluierung des Potentials könnte eine Korrelation mit berechneten Energien aus physikalisch fundamentierten Methoden, wie Kraftfeldern, versucht werden. Es wäre interessant zu sehen, ob hier eine lineare Korrelation zwischen den, mit beiden Methoden berechneten, "Energien" vorliegt.

Als letzten Ausblick sei auf die *interface*-Analysen von Tsai *et al.* (1997) hingewiesen. Sie zeigen, daß der Anteil hydrophober Aminosäuren im Innerem gegenüber dem Anteil in der Bindestelle bei Oligomeren überwiegt. Sie kommen zu dem Schluss, daß der Beitrag des hydrophoben Effektes zur Protein-Protein-Assoziation nicht so stark ist, wie bei der Faltung. Als logische Konsequenz wären Paarpotentiale, die auf Oligomeren basiere, nicht besonders geeignet assoziative Vorgänge von temporären Komplexen zu beschreiben. Dies wiederum würde bedeuten, daß eine genaue Untersuchung der Zusammensetzung des Strukturdatensatzes auch in diese Hinsicht sinnvoll wäre.

5 Zusammenfassung

Protein-Protein-Wechselwirkungen spielen eine zentrale Rolle in biologischen Prozessen. Da die experimentelle Untersuchung solcher Komplexe schwierig und zeitaufwendig ist, wird versucht den Bindungsmodi zweier Proteine *in silico* mit der Methode des Dockings vorherzusagen. Hierbei können viele tausende falsch-positive Anordnungen erhalten werden, so daß der Bedarf an einer Bewertungsfunktion zur Unterscheidung nativ-ähnlicher und nicht-nativer Anordnungen groß ist. In der hier vorgestellten Arbeit ist daher ein wissenbasierten Potential zur Bewertung von Protein-Protein-Docking-Ergebnissen entwickelt worden. Es besteht aus zwei additiven Termen, einem kontaktabhängigen Paarpotential und einem von der lösungsmittelzugänglichen Fläche abhängigen Ein-Teilchen-Potential. Das Potential ist atombasiert, da die Beschreibung anhand von Aminosäuren als zu ungenau betrachtet wird. Es werden 40 Atomtypen (Melo & Feytmans (1997) zur Beschreibung aller Atome verwendet.

Des weiteren werden zwei spezielle Gewichtungsfaktoren eingeführt, die Wasserstoffbrücken und die Atomen funktioneller Gruppen höher gewichten sollen. Wasserstoffbrücken können einen signifikanten Anteil an der Bindungsenergie haben und sind gleichzeitig in dem hier verwendeten Datensatz zur Erzeugung der Potentiale unterrepräsentiert. Gleiches gilt für die Atome funktioneller Gruppen. Außerdem wird angenommen, daß die meisten Interaktionen, wie aus chemischen Reaktionen bekannt, polarisierte oder polarisierbare Atome beinhalten. Es wird die Annahme gemacht, daß Atome funktioneller Gruppen allgemein einen größeren Anteil an der Bindungsenergie zweier Proteine haben. Einzelfälle und Ausnahmen werden nicht berücksichtigt. Die Potentiale werden aus den Häufigkeitsverteilungen der interatomaren Abständen der 40 Atomtypen unter Verwendung der *inversen* Boltzmann-Gleichung aus einem Strukturdatensatz gewonnen. Dieser beinhaltet 584 Komplex-Strukturen und besteht hauptsächlich aus der COMBASE (Vakser *et al.* (1999)), der momentan

größten Sammlungen von Protein-Protein-Komplexen. Die generierten, diskreten Häufigkeitsverteilungen werden mit einer Trapezfunktion geglättet. Zuletzt wird wird ein repulsiver Korrekturterm eingeführt, der sterische Überlappungen als Artefakte aus dem Docking-Prozess bestrafen soll.

In insgesamt einundzwanzig *unbound*-Docking-Studien mit dem Fourier-Korrelations-Programm ckordo (Zimmermann (2002)) wird das Potential gegen die Bewertung mit reiner geometrischen Korrelation getestet. In fünf der einundzwanzig Fälle findet das Programm ckordo keine nativ-ähnliche Lösung, so daß diese Komplexe mit diesem Potential nicht bewertet werden können. Zu den fünf Systemen gehören auch die beiden einzigen hier getesteten Antikörper-Antigen-Komplexe 1mlc und 1vfb. In zwölf von den erfolgreichen sechzehn Docking-Studien findet das vorgestellte Potential eine nativ-ähnliche Lösung unter den ersten 1000 Plätzen. ckordo hingegen kann nur in drei der sechzehn Fällen eine passende Lösung unter den besten 1000 finden. Die Bewertung mit Gewichtung der Atome funktioneller Gruppen bringt eine signifikante Verbesserung im Vergleich zu den Vorhersagen ohne funktionelle Gruppen, so daß die eingangs gemachten Annahmen als bestätigt angesehen werden können. Die Einführung eines solvensabhängigen Terms erbringt im Mittel keine deutliche Verbesserung. Gründe hierfür könnten eine unpassende *interface*-Definition und die Verwendung der SAS anstatt der MSA sein.

Die jedoch mit Abstand größte Verbesserung der Bewertung wird durch die Abkehr vom Boltzmannschen Verteilungsgesetz hin zu einer empirischen Energiedichtefunktion erlangt. Die Ergebnisse der lediglich mit einem konstanten Faktor (hier 0.025) gewichteten normierten radialen Paarverteilungsfunktionen sind signifikant besser als eine Verwendung der *inversen* Boltzmann-Gleichung mit der gemittelten, normierten, radialen Paarverteilungsfunktion als Referenzzustand.

Die Untersuchung zur Unterscheidung von nativen-und nicht-nativen-decoy-Strukturen mit dem entwickelten Potential zeigt gute Ergebnisse, die allerdings nicht so überzeugend sind, wie nach den Docking-Studien zu erwarten gewesen

wäre. Gründe hierfür könnten vor allem die nicht bis ins Detail bekannte Methode zur Erzeugung der decoy-Datensätze sein.

Ein Vergleich der hier berechneten Bindungs-Präferenzen einzelner Komplexe mit experimentellen Bindungsdaten aus der Literatur zeigt keine linearen Abhängigkeiten. Es ist zu bemerkten, daß die Werte in der Literatur bereits um 0.1-0.2 kcal/mol von Autor zu Autor schwanken. Die hier zitierten Autoren verwenden diese Daten zwar, aber geben keinen experimentellen Fehler an, so daß keine Fehlerabschätzung vorgenommen werden kann. Es wird angenommen, daß es möglich wäre das Potential auf einen linearen Zusammenhang mit den wenigen experimentellen Bindungsdaten aus der Literatur zu optimieren. Dies war nicht Ziel dieser Arbeit. Weiterhin wird die Bewertung des "Ernstfalles", also die Untersuchung von *unbound*-Protein-Protein-Docking-Ergebnisse als die sinnvollste Methode zur Validierung dieses Potentials betrachtet.

Abschließend kann gesagt werden, daß das vorgestellte Potential sehr gut geeignet ist, die Ergebnisse von Protein-Protein-Docking-Studien hinsichtlich der Diskriminierung nativer von nicht-nativen Komplex-Strukturen zu bewerten. Gleichzeitig deuten die in der Diskussion und dem Ausblick vorgestellten Ideen weitere Verbesserungen an.

A Anhang

A.1 Atomeinteilung und maximale atomare ASA

Ala		
N	3	72
CA	1	79
CB	6	75
C	4	53
O	5	58

Arg		
N	3	72
CB	8	76
CG	8	76
CD	37	74
NE	36	58
NH1	22	62
NH2	22	62
CZ	21	72
CA	1	79
C	4	53
O	5	58

Asn		
N	3	72
CB	8	70
CG	33	61
OD1	34	43
ND2	18	63
CA	1	79
C	4	53
O	5	58

Asp		
N	3	72
CB	8	76
OD1	28	45
OD2	28	45
CG	27	67
CA	1	79
C	4	53
O	5	58

Cys		
N	3	72
CB	29	70
SG	19	77
CA	1	79
C	4	53
O	5	58

Gln		
N	3	72
CB	8	76
CG	8	76
CD	33	61
OE1	34	43
NE2	18	63
CA	1	79
C	4	53
O	5	58

Glu		
N	3	72
CB	8	76
CG	8	76
CD	27	67
OE1	28	45
OE2	28	45
CA	1	79
C	4	53
O	5	58

His		
N	3	72
CB	8	72
CG	23	5
ND1	38	29
CD2	24	40
NE2	25	33
CE1	26	46
CA	1	79
C	4	53
O	5	58

Ile		
N	3	72
CB	7	68
CG1	8	76
CG2	6	75
CD1	6	75
CA	1	79
C	4	53
O	5	58

Leu		
N	3	72
CB	8	72
CG	7	68
CD1	6	75
CD2	6	75
CA	1	79
C	4	53
O	5	58

Lys		
N	3	72
CB	8	70
CG	8	77
CD	8	88
CE	35	53
NZ	20	58
CA	1	79
C	4	53
O	5	58

Met		
N	3	72
CB	8	70
CG	29	70
SD	9	50
CE	30	79
CA	1	79
C	4	53
O	5	58

Phe		
N	3	72
CB	8	76
CG	11	15
CD1	12	38
CD2	12	38
CE1	12	38
CE2	12	38
CZ	12	38
CA	1	79
C	4	53
O	5	58

Ser		
N	3	72
CB	15	68
OG	16	43
CA	1	79
C	4	53
O	5	58

Thr		
N	3	72
CB	17	68
OG1	16	43
CG2	6	75
CA	1	79
C	4	53
O	5	58

Trp		
N	3	72
CB	8	76
CG	13	23
CD1	24	40
CD2	11	15
NE1	39	30
CE2	14	7
CZ2	12	38
CE3	12	38
CZ3	12	38
CA	1	79
C	4	53
O	5	58

Tyr		
N	3	72
CB	8	76
CG	11	15
CD1	12	38
CD2	12	38
CE1	12	38
CE2	12	38
CZ	31	
OH	40	45
CA	1	79
C	4	53
O	5	58

Val		
N	3	72
CB	7	68
CG1	6	75
CG2	6	75
CA	1	79
C	4	53
O	5	58

Gly		
N	3	72
CA	2	69
C	4	53
O	5	58

Pro		
N	10	26
CB	18	63
CG	8	75
CD	32	50
CA	1	79
C	4	53
O	5	58

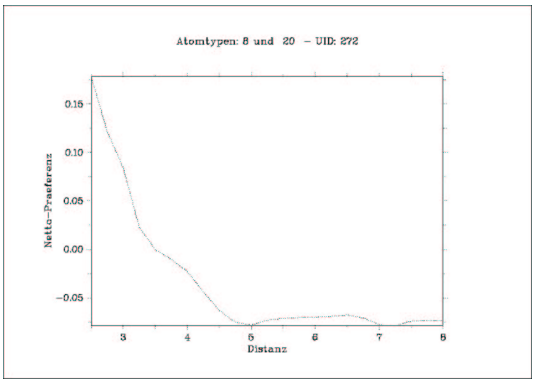
Tabelle A.1: Für jede Aminosäure sind die Atome in der ersten Spalte, den ihnen zugewiesenen numerischen Werte in der zweiten Spalte und die maximale dem Lösungsmittel zugängliche Fläche in $Å^2$ (extrahiert aus dem Strukturdatensatz) in der dritten Spalte angegeben.

A.2 Atome funktioneller Gruppen

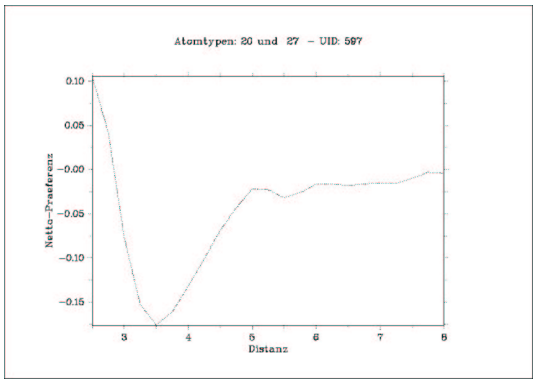
Atomtyp	Numerischer Wert
C	6
S	9
C (arom)	12
N (Gln, Asn)	18
O (Ser)	16
S (Cys)	19
N (Lys)	20
N (Arg)	22
N (His)	25
C (His)	26
O (Asp)	28
C (Met)	30
O (Gln)	34
N (Arg)	36
N (His)	38
N (Trp)	39
O (Tyr)	40

Tabelle A.2: Atome funktioneller Gruppen mit dem ihnen zugeordneten numerischen Wert.

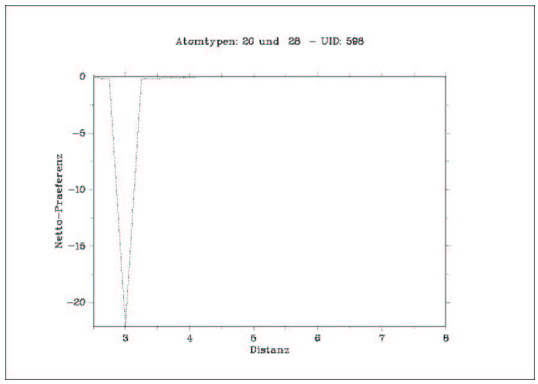
A.3 Distanzabhängige Paarpotentiale



(a) UID 272

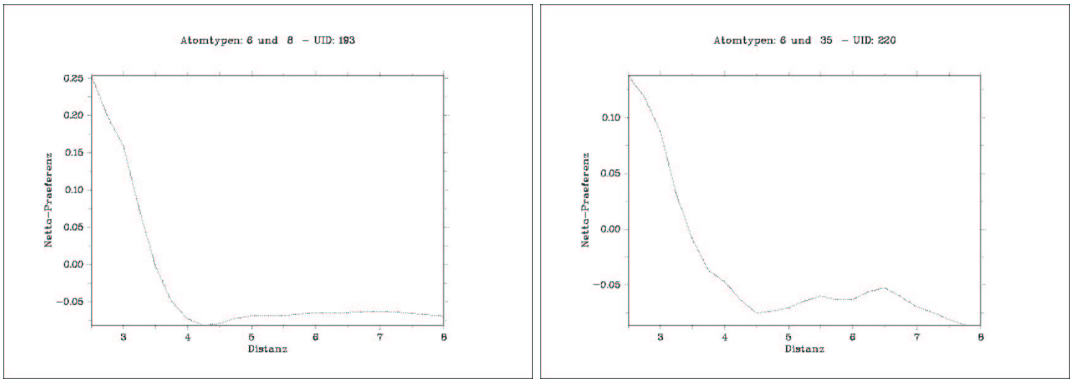


(b) UID 597

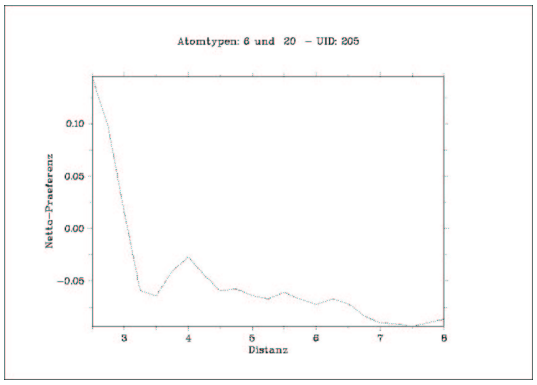


(c) UID 598

Abbildung A.1: Netto-Präferenzen für NZ (LYS) mit (a) CB (ASP), (b) CG (ASP) und (c) O (ASP).Ordinate: Distanzen in Å , Abszisse: Netto-Präferenzen.



(a) UID 193 **(b)** UID 220



(c) UID 205

Abbildung A.2: Netto-Präferenzen für CD (LEU) mit (a) CB/CG/CD (LYS), (b) CE (LYS) und (c) NZ (LYS). Ordinate: Distanzen in Å, Abszisse: Netto-Präferenzen.

A.4 Strukturdatensatz

104l-AB	1aap-AB	1aar-AB	1aaz-AB	1abr-AB	1acb-EI	1ade-AB	1adu-AB	1aer-AB
1ahs-BC	1aiz-AB	1aks-AB	1alk-AB	1all-AB	1anw-AB	1aor-AB	1aoz-AB	1apx-AB
1apy-AC	1apy-AB	1apy-BD	1asy-AB	1atn-AD	1avd-AB	1bar-AB	1bbb-AB	1bbh-AB
1bbp-BD	1bbr-EK	1bbt-14	1bbt-24	1bbt-23	1bcf-AB	1bcm-AB	1bdm-AB	1bgs-FG
1bhm-AB	1bin-AB	1bjm-AB	1blb-AB	1bmf-AG	1bmf-CD	1bmf-DF	1bmf-DG	1bmt-AB
1bmv-12	1bnc-AB	1bnd-AB	1bov-BC	1bpl-AB	1bql-LY	1bro-AB	1brs-CF	1bsr-AB
1bun-AB	1bvp-23	1c2r-AB	1cax-AC	1cax-CF	1cbi-AB	1cdk-AB	1cdl-AC	1cea-AB
1cgj-EI	1cgl-AB	1chk-AB	1chm-AB	1cho-EI	1chr-AB	1cki-AB	1cle-AB	1clx-AB
1cmc-AB	1cns-AB	1col-AB	1cpc-AB	1cpc-AK	1csg-AB	1csk-AD	1csm-AB	1cud-AB
1cwe-AC	1cwp-AB	1cyd-CD	1d66-AB	1daa-AB	1dbq-AB	1dcp-GH	1dea-AB	1dek-AB
1dfj-EI	1dfn-AB	1dif-AB	1dir-AB	1dkt-AB	1dky-AB	1dlh-BE	1dmx-AB	1dnp-AB
1dok-AB	1dpg-AB	1dpp-AC	1dpr-AB	1dsb-AB	1dth-AB	1dut-AB	1dvf-BD	1dvr-AB
1dyn-AB	1ebd-BC	1ece-AB	1ecf-AB	1ecm-AB	1ecp-BD	1ecz-AB	1edh-AB	1edm-BC
1efn-AB	1efn-BD	1efu-AC	1efu-AB	1efu-BD	1epa-AB	1ept-AB	1ept-AC	1ept-BC
1esf-AB	1etp-AB	1ext-AB	1fat-AB	1fba-BC	1fbi-QY	1fc1-AB	1fc2-CD	1fcb-AB
1fcc-AC	1fcd-AB	1fcd-BD	1fia-AB	1fie-AB	1fin-AB	1fjl-AB	1fjm-AB	1fki-AB
1fle-EI	1fod-12	1fod-13	1fos-GH	1frp-AB	1frt-AC	1frt-BC	1frv-AC	1frv-CD
1fss-AB	1fug-AB	1fuj-AB	1fuq-AB	1fvp-AB	1fxi-AD	1fxr-AB	1gad-OP	1gam-AB
1gar-AB	1gdh-AB	1gdt-AB	1ges-AB	1gff-12	1gfl-AB	1ggg-AB	1ghs-AB	1gif-BC
1gla-FG	1glq-AB	1glu-AB	1got-AB	1got-BG	1gp1-AB	1gpm-BD	1gri-AB	1gse-AB
1gto-BC	1gtp-BI	1gtq-AB	1gua-AB	1gyl-AB	1hav-AB	1hbh-CD	1hcg-AB	1hde-AB
1hge-AC	1hge-CD	1hge-DF	1hiw-AR	1hjr-BD	1hle-AB	1hlp-AB	1hmp-AB	1hng-AB
1hpc-AB	1hpl-AB	1hrd-BC	1hrh-AB	1hro-AB	1hsa-AD	1hsb-AB	1hsl-AB	1hst-AB
1htm-DF	1htt-AB	1huc-AB	1hul-AB	1hxp-AB	1hyh-AB	1hyl-AB	1ice-AB	1ids-AC
1ies-BE	1ihf-AB	1ilr-12	1inh-AB	1isu-AB	1ith-AB	1jst-AC	1kba-AB	1kir-BC
1kny-AB	1kob-AB	1kpb-AB	1kpt-AB	1lcp-AB	1leh-AB	1lgb-AC	1lmb-34	1lmk-EG
1lmw-BD	1lpb-AB	1lti-AG	1luc-AB	1lwi-AB	1lya-AB	1hur-AB	1lya-BD	1lyl-AC
1lyn-AB	1mac-AB	1mas-AB	1mdp-12	1mdt-AB	1mdy-AB	1mec-14	1mee-AI	1mhl-AB
1mhl-AC	1mhl-CD	1mka-AB	1mld-AB	1mmo-BC	1mmo-CH	1mmo-CE	1mmo-DE	1mmo-EH
1mol-AB	1mpm-BC	1msa-AD	1msp-AB	1mtn-BF	1mtn-FH	1mtn-GH	1myk-AB	1nal-23
1nba-AB	1nci-AB	1nco-AB	1nfk-AB	1noy-AB	1npo-AC	1nsc-AB	1nsn-HS	8ruc-KL
9ldt-AB	1nsn-LS	1oac-AB	1obp-AB	1occ-NO	1occ-NP	1occ-NQ	1occ-NS	1occ-GN
1occ-NU	1occ-NV	1occ-NW	1hcn-AB	1occ-NX	1occ-NY	1occ-NZ	1occ-OQ	1occ-OR
1occ-OU	1occ-OV	1occ-OX	1occ-PS	1occ-PT	1occ-PU	1occ-PW	1occ-QR	1occ-QS
1occ-QV	1occ-QX	1occ-QZ	1occ-RS	1occ-RV	1occ-FS	1occ-ST	1occ-SW	8atc-AB
8cat-AB	8ruc-EK	1occ-TU	1occ-HU	1occ-WY	1occ-YZ	1one-AB	1onr-AB	1ord-AB
1oro-AB	1ort-BF	1osj-AB	1otf-BE	1otg-BC	1ova-AB	1ova-CD	1ovo-AB	1pag-AB
1pam-AB	1pbw-AB	1pdg-AB	1pfx-CL	1pge-AB	1pio-AB	1pky-AC	1pml-BC	1pnk-AB
1pov-03	1pox-AB	1poy-12	1ppf-EI	1prc-CL	1prc-CM	1prc-CH	1prc-LM	1prc-HM
1pre-AB	1prt-AB	1prt-AE	1prt-AF	1prt-EF	1prt-HJ	1prt-HL	1prt-JK	1psa-AB
1psd-AB	1pvc-12	1pvc-13	1pvc-24	1pvc-34	1pvd-AB	1pvu-AB	1pxt-AB	1pya-CE
1pya-CD	1pya-DF	1pyi-AB	1pyt-BD	6q21-AB	1pyt-AB	1pyt-AC	1qap-AB	1qas-AB
1qbe-BC	1qor-AB	1qpa-AB	1qrd-AB	1rah-BD	1rba-AB	1rcm-AB	1rco-RV	1rcp-AB
1rdl-12	1reg-XY	1rfb-AB	1rgf-AB	1rhg-AC	1rlb-AF	1rn1-AC	1rth-AB	1rtm-12
1rtp-23	1rva-AB	1sac-CD	1sce-BD	1sch-AB	1scm-AB	1scm-BC	1scu-DE	1scu-BE
1seb-AB	1seb-EH	1sei-AB	1sem-AB	1set-AB	1sft-AB	1sgp-EI	1slt-AB	1slu-AB
1smn-AB	1smp-AI	1spb-PS	1sph-AB	1sri-AB	1seb-EH	1stf-EI	1stm-BC	1tab-EI
1taf-AB	1tah-AC	1tbr-KS	1tcb-AB	1tco-AC	1tco-AB	1tco-BC	1tcr-AB	1tgx-AB
1the-AB	1thj-BC	1tht-AB	1tii-AC	1tii-EF	1tlf-AB	1tmc-AB	1tme-12	1tme-13
1tme-23	1tmf-13	1tmf-14	1tmf-23	1tmf-24	1tmf-34	1tnd-AC	1tnf-AB	1tnr-AR
1tph-12	1trk-AB	1tro-AC	1tsd-AB	1tsr-AB	1tta-AB	1tvx-BD	1ubs-AB	1ucy-HK
1udi-EI	1umu-AB	1una-AB	1urn-AB	1vcp-BC	1vfb-AB	1vfb-AC	1vhi-AB	1vmo-AB
1vok-AB	1vol-AB	1vrt-AB	1vsc-AB	1vsg-AB	1wap-BC	1wdc-AC	1wfb-AB	1wgt-AB
1wht-AB	1wtl-AB	1xik-AB	1xim-AC	1xso-AB	1xva-AB	1xxa-DF	1xyp-AB	1ycb-AB
1ygp-AB	1yha-AB	1ypp-AB	1ypt-AB	1yrn-AB	1ytf-AD	1ytf-BD	1ytt-AB	1zop-AB
256b-AB	2abx-AB	2ach-AB	2adm-AB	2afn-BC	2bbk-HJ	2bbk-HL	2bbv-BC	2bpa-13
2btf-AP	2ccy-AB	2cht-DE	2cst-AB	2dhf-AB	2dld-AB	2drp-AD	2eip-AB	2hhm-AB
2hip-AB	2hmq-CD	2hnt-CF	2hpp-HP	2kai-AI	2kai-BI	2kau-AB	2kau-AC	2kau-BC
2lig-AB	2ltn-AC	2ltn-AB	2mev-12	2mev-23	2mev-24	2mev-34	2mta-AC	2nac-AB
2pcc-AB	2pcd-BC	2pcd-BN	2pcd-MP	2pel-BC	2phl-BC	2pka-AB	2pka-BY	2plv-14
2pol-AB	2psp-AB	2ptc-EI	2rbi-AB	2rmc-EG	2rsl-AB	2rsp-AB	2scp-AB	2spc-AB
2tbv-AB	2tmd-AB	2trx-AB	2utg-AB	2zta-AB	3bto-BC	3cro-LR	3hhr-AB	3hhr-BC
3ink-CD	3ins-BD	3lad-AB	3mde-AB	3mon-CE	3mon-CD	3mon-BD	3pga-24	3pmg-AB
3sdh-AB	3sic-EI	4aah-AC	4aah-CD	4ake-AB	4cha-AB	4cts-AB	4dfr-AB	4htc-HI
4kbp-BC	4sbv-AB	4sgb-EI	5cna-AB	6chy-AB	6gsv-AB	6pfk-CD		

Tabelle A.3: PDB- und Kettenbezeichner der Komplexe des Strukturdatensatzes.

A.5 Decoy Bewertung

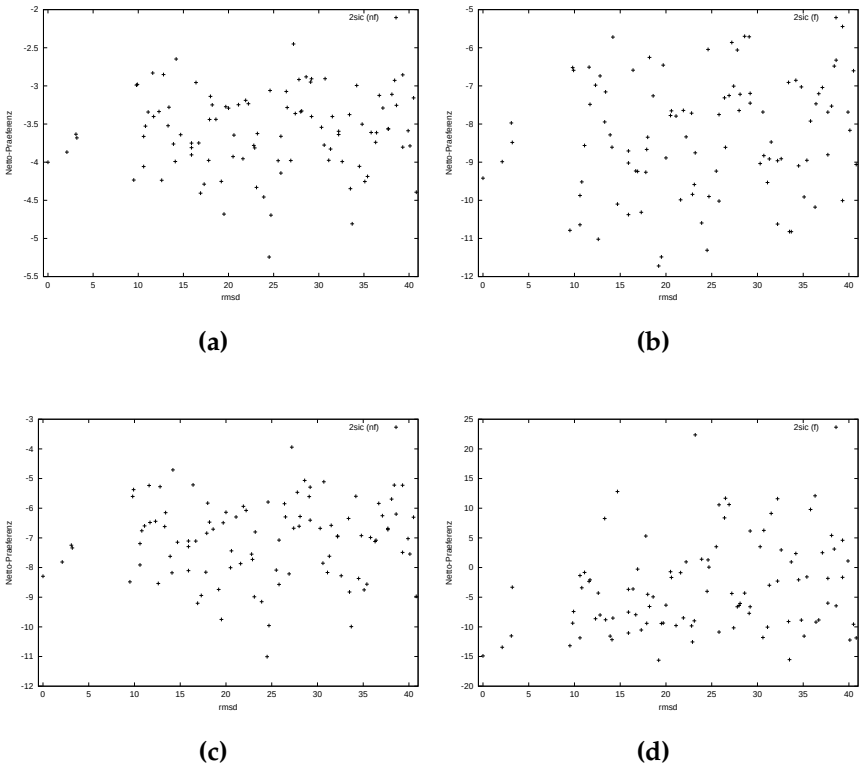


Abbildung A.3: 2sic-decoy: rmsd in Å (*interface*-CA). (a) und (b) Profil 2. (c) und (d) Profil 1. nf = ohne Gewichtung von Atom-Atom-Kontakten funktioneller Gruppen. f = mit Gewichtung derselben.

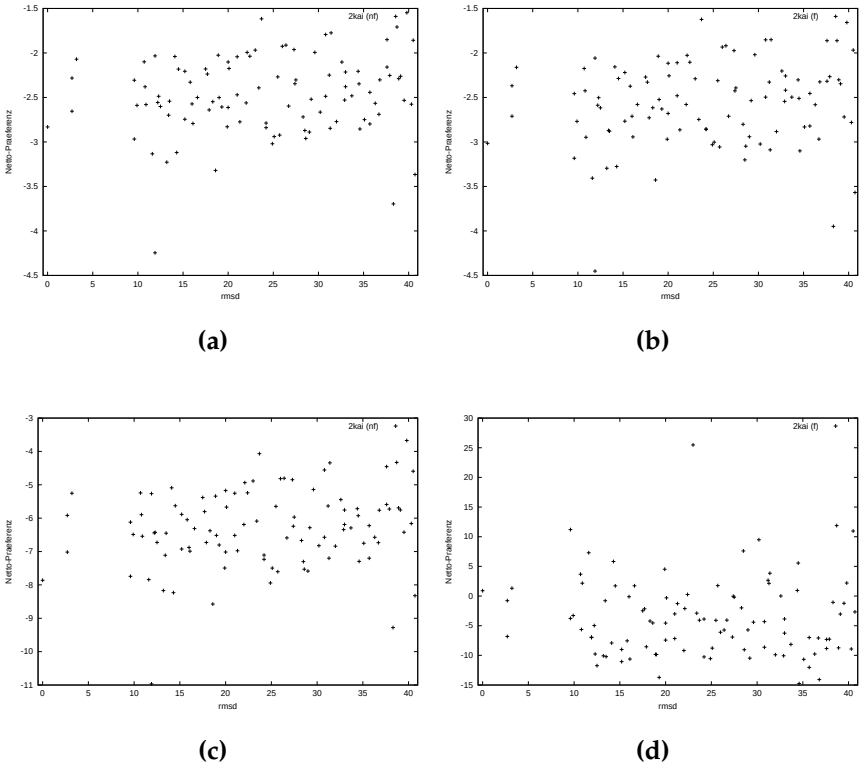


Abbildung A.4: 2kai-decoy: rmsd in Å (*interface*-CA). (a) und (b) Profil 2. (c) und (d) Profil 1. nf = ohne Gewichtung von Atom-Atom-Kontakten funktioneller Gruppen. f = mit Gewichtung derselben.

A.6 Hilfsmittel

Visualisierung und Auswertung	
Visual Molecular Dynamics (VMD v1.7)	http://www.ks.uiuc.edu/Research/vmd/
RasMol v2.7.1	http://www.umass.edu/microbio/rasmol/
PyMol v0.83	http://www.pymol.org
BRAGI	http://www.uni-koeln.de/math-nat-fak/biochemie/ds/dsbrag_e.htm
dislin v7.6	http://www.linmpi.mpg.de/dislin/
gnuplot v3.7.1	http://www.gnuplot.info/
gnumeric v0.70	http://www.gnome.org/projects/gnumeric/
Excel	http://www.microsoft.com/office/excel/default.asp
Origin	http://www.originlab.com/
mathlab	http://mathlab.com/
Programmiersprachen und Module	
Python v2.1.1	http://www.python.org/
gcc3.0-3.0.1	http://gcc.gnu.org/
Numeric-21.3	http://sourceforge.net/projects/numpy
stats	http://www.nmr.mgh.harvard.edu/Neural_Systems_Group/gary/python.html
scientific	http://starship.python.net/crew/hinsen/scientific.html
Sonstige Programme	
naccess v2.1.1	http://sjh.bi.umist.ac.uk/naccess.html
HBPLUS v3.0	http://www.biochem.ucl.ac.uk/ mcdonald/hbplus/home.html
Betriebssysteme	
Linux 2.2.18 (Mandrake 7.1)	http://www.mandrake.com/
Linux 2.4.10 (Suse 8.1)	http://www.de.kernel.org/
SGI Irix 6.5	http://www.sgi.com/
SOLARIS	
Verwendete Rechner	
Sunfire 880V, 6 Ultra Sparc 750 MHz Prozessoren	
1700 AMD Athlon	
800 AMD Athlon	

Tabelle A.4: Liste der verwendeten Programme, Programmiersprachen und Betriebssysteme.

A.7 Lebenslauf

6. Oktober 1973	geboren in München, Staatsangehörigkeit deutsch
1981–1982	Städt. Kath. Grundschule Bensberg
1983–1984	Städt. Gemeinschaftsgrundschule Bensberg
1984–1993	Otto-Hahn-Gymnasium Bensberg
Mai 1993	Abitur
Oktober 1993	Beginn des Chemiestudiums an der Universität zu Köln
Oktober 1994	Wechsel an die Ruprecht-Karl-Universität zu Heidelberg
Februar 1996	Vordiplomprüfung in Heidelberg
April 1996	Wechsel an die Universität zu Köln
Dezember 1998	Diplomprüfung
Dezember 1998- Oktober 1999	Diplomarbeit bei Herrn Prof. Dr. D. Schomburg an der Universität zu Köln Thema: „Atomare Strukturmotive in Proteinen“
seit März 2000	Doktorarbeit bei Herrn Prof. Dr. D. Schomburg an der Universität zu Köln

A.8 Vorabveröffentlichungen

- Grimm, V., Schomburg D. (2002). Development and Evaluation of a Knowledge-Based-Potential for the Scoring of Protein-Protein-Docking Results. Posterpräsentation, European Conference on Computational Biology, Saarbrücken.

Literaturverzeichnis

Anfinsen, C. B. (1973) Principles that govern the folding of protein chains. *Science* **181**, 223–30.

auf der Heyde, T. E. A. (1990) Analyzing Chemical Data in more than two Dimensions. *J Chem Educat* **67**, 461–469.

Augspurger, J. & Scheraga, H. (1995) An Efficient, Differentiable Hydration Potential for Peptides and Proteins. *J Comp Chem* **17**, 1549–1558.

Bahar, I. & Jernigan, R. L. (1997) Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J Mol Biol* **266**, 195–214.

Ben-Naim, A. (1987) *Solvation Thermodynamics*. Plenum Press, Ney York.

Ben-Naim, A. (1992) *Statistical Thermodynamics for Chemists and Biochemists*. Plenum Press, New York.

Ben-Naim, A., Ting, K. L. & Jernigan, R. L. (1990) Solvent effect on binding thermodynamics of biopolymers. *Biopolymers* **29**, 901–19.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res* **28**, 235–42.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **112**, 535–42.

Betancourt, M. R. & Thirumalai, D. (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci* **8**, 361–9.

Bogan, A. A. & Thorn, K. S. (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol* **280**, 1–9.

Camacho, C. J. & Vajda, S. (2002) Protein-protein association kinetics and protein docking. *Curr Opin Struct Biol* **12**, 36–40.

Camacho, C. J., Weng, Z., Vajda, S. & DeLisi, C. (1999) Free energy landscapes of encounter complexes in protein-protein association. *Biophys J* **76**, 1166–78.

Camacho, C. J., Gatchell, D. W., Kimura, S. R. & Vajda, S. (2000) Scoring docked conformations generated by rigid-body protein-protein docking. *Proteins* **40**, 525–37.

Chakrabarti, P. & Janin, J. (2002) Dissecting protein-protein recognition sites. *Proteins* **47**, 334–43.

Chen, R. & Weng, Z. (2002) Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins* **47**, 281–94.

Claussen, H., Buning, C., Rarey, M. & Lengauer, T. (2001) FlexE: efficient molecular docking considering protein structure variations. *J Mol Biol* **308**, 377–95.

Conte, L. L., Chothia, C. & Janin, J. (1999) The atomic structure of protein-protein recognition sites. *J Mol Biol* **285**, 2177–98.

Covell, D. G. & Wallqvist, A. (1997) Analysis of protein-protein interactions and the effects of amino acid mutations on their energetics. The importance of water molecules in the binding epitope. *J Mol Biol* **269**, 281–97.

Cummings, M. D., Hart, T. N. & Read, R. J. (1995) Atomic solvation parameters in the analysis of protein-protein docking results. *Protein Sci* **4**, 2087–99.

Dasgupta, S., Iyer, G. H., Bryant, S. H., Lawrence, C. E. & Bell, J. A. (1997) Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins* **28**, 494–514.

DeBolt, S. E. & Skolnick, J. (1996) Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. *Protein Eng* **9**, 637–55.

Dengler, U. (1998) *Kristallstruktur der D2-Hydroxyisocaproat-Dehydrogenase aus Lactobacillus casei. Verfeinerung, Interpretation und Anwendung in einem Verfahren zur Erkennung der Proteinfaltung*. Dissertation, Technische Universität Carolo-Wilhelmina zu Braunschweig.

Dennis, S. & C. J. Camacho, S. V. (2000) Exploring potential solvation sites of proteins by multistart local minimization. *Optimization in Computational Chemistry and Molecular Biology* 1–2.

DeWitte, R. S. & Shaknovich, E. I. (1996) SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J Am Chem Soc* **118**, 11733–11744.

Dill, K. A., Phillips, A. T. & Rosen, J. B. (1997) Protein structure and energy landscape dependence on sequence using continous engery function. *J Comput Biol* **4**, 227–39.

Eisenberg, D. & McLachlan, A. D. (1986) Solvation energy in protein folding and binding. *Nature* **319**, 199–203.

Eisenhaber, F. (1996) Hydrophobic regions on protein surfaces. Derivation of the solvation energy from their area distribution in crystallographic protein structures. *Protein Sci* **5**, 1676–86.

Eisenhaber, F. & Argos, P. (1996) Hydrophobic regions on protein surfaces: definition based on hydration shell structure and a quick method for their computation. *Protein Eng* **9**, 1121–33.

Fernandez-Recio, J., Totrov, M. & Abagyan, R. (2002) Soft protein-protein docking in internal coordinates. *Protein Sci* **11**, 280–91.

Fields, S. & Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature* **20**, 245–6.

Findenegg, G. H. (1985) *Statistische Thermodynamik*. Steinkopf Verlag.

Finkelstein, A. V. & Gutin, A. M. (1995) Why do protein architectures have Boltzmann-like statistics? *Proteins* **23**, 142–50.

Fischer, D., Lin, S. L., Wolfson, H. L. & Nussinov, R. (1995) A geometry-based suite of molecular docking processes. *J Mol Biol* **248**, 459–77.

Fischer, T. (2000) *Oberflächen-Plasmon-Resonanz spektroskopische Studie an verschiedene Biosensoroberflächen. Charakterisierung von Protein-Peptid und Protein-Lipid Wechselwirkungen*. Forschungszentrum Jülich.

Furuichi, E. & Koehl, P. (1998) Influence of protein structure databases on the predictive power of statistical pair potentials. *Proteins* **31**, 139–49.

Gabb, H. A., Jackson, R. M. & Sternberg, M. J. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* **272**, 106–20.

Gabdoulline, R. R. & Wade, R. C. (1997) Simulation of the diffusional association of barnase and barstar. *Biophys J* **72**, 1917–29.

Gan, H. H., Tropsha, A. & Schlick, T. (2001) Lattice protein folding with two and four-body statistical potentials. *Proteins* **43**, 161–74.

Gardiner, E., Willett, P. & Artymiuk, P. (2001) Protein docking using a genetic algorithm. *Proteins* **44**, 44–56.

Gatchell, D. W., Dennis, S. & Vajda, S. (2000) Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins* **41**, 518–34.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. & Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–7.

Gilis, D. & Rooman, M. (1997) Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J Mol Biol* **272**, 276–90.

Gilson, M. K. & Honig, B. (1988) Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins* **4**, 7–18.

Glaser, F., Steinberg, D., Vakser, I. & Ben-Tal, N. (2001) Residue Frequencies and Pairing Preferences at Protein-Protein Interfaces. *Proteins* **43**, 89–102.

Godzik, A. (1996) Knowledge-based potentials for protein folding: what can we learn from known protein structures? *Structure* **4**, 363–6.

Godzik, A., Kolinski, A. & Skolnick, J. (1995) Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci* **4**, 2107–17.

Gohlke, H., Hendlich, M. & Klebe, G. (2000) Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* **295**, 337–56.

Guerois, R., Nielsen, J. E. & Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* **320**, 369–87.

Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409–43.

Heifetz, A., Katchalski-Katzir, E. & Eisenstein, M. (2002) Electrostatics in protein-protein docking. *Protein Sci* **11**, 571–87.

Helge Toutenburg, Andreas Fieger, C. K. (1998) *Deskriptive Statistik*. Prentice Hall.

Hill, T. L. (1986) *An Introduction to Statistical Thermodynamics*. Dover Publications Inc.

Hodgkin, E. E. & Richards, W. G. (1987) Molecular Similarity Based on Electrostatic Potential and Electric Field. *Int I Quant Chem: Quant Biol Symp* **14**, 105–110.

Holst, M., Kozack, R. E., Saied, F. & Subramaniam, S. (1994) Treatment of electrostatic effects in proteins: multigrid-based Newton iterative method for solution of the full nonlinear Poisson-Boltzmann equation. *Proteins* **18**, 231–45.

Honig, B. & Nicholls, A. (1995) Classical electrostatics in biology and chemistry. *Science* **268**, 1144–9.

Hoppe, C. (2002) *Entwicklung einer richtungs- und abstandsabhängigen wissensbasierten Bewertungsfunktion für die Vorhersage der Thermostabilität von Proteinen*. Dissertation, Universität zu Köln.

Horovitz, A. (1987) Non-additivity in protein-protein interactions. *J Mol Biol* **196**, 733–5.

Horton, N. & Lewis, M. (1992) Calculation of the free energy of association for protein complexes. *Protein Sci* **1**, 169–81.

Hu, J. C. (2000) A guided tour in protein interaction space: coiled coils from the yeast proteome. *Proc Natl Acad Sci U S A* **97**, 12935–6.

Hubbard, S. J. & Thornton, J. M. (1993) *NACCESS*. University College London.

Jackson, R. M. (1999) Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem. *Protein Sci* **8**, 603–13.

Jackson, R. M. & Sternberg, M. J. (1995) A continuum model for protein-protein interactions: application to the docking problem. *J Mol Biol* **250**, 258–75.

Janin, J. (1996) Quantifying biological specificity: the statistical mechanics of molecular recognition. *Proteins* **25**, 438–45.

Janin, J. (1997) The kinetics of protein-protein recognition. *Proteins* **28**, 153–61.

Janin, J. & Chothia, C. (1990) The structure of protein-protein recognition sites. *J Biol Chem* **265**, 16027–30.

Janin, J., Miller, S. & Chothia, C. (1988) Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol* **204**, 155–64.

Jayaram, B., McConnell, K., Dixit, S., Das, A. & Beveridge, D. (2002) Free-Energy Component Analysis of 40 Protein-DNA Complexes: A Consensus View on the Thermodynamics of Binding at the Molecular Level. *J Comp Chem* **23**, 1–14.

Jiang, F. & Kim, S. H. (1991) Soft docking: matching of molecular surface cubes. *J Mol Biol* **219**, 79–102.

Jiang, F., Lin, W. & Rao, Z. (2002) SOFTDOCK: understanding of molecular recognition through a systematic docking study. *Protein Eng* **15**, 257–63.

Jones, D. T. (1994) De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci* **3**, 567–74.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86–9.

Jones, G., Willett, P. & Glen, R. C. (1995) Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J Mol Biol* **245**, 43–53.

Jones, S. & Thornton, J. (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**, 13–20.

Jones, S. & Thornton, J. M. (1997) Analysis of protein-protein interaction sites using surface patches. *J Mol Biol* **272**, 121–32.

Juffer, A. H., Eisenhaber, F., Hubbard, S. J., Walther, D. & Argos, P. (1995) Comparison of atomic solvation parametric sets: applicability and limitations in protein folding and binding. *Protein Sci* **4**, 2499–509.

Kang, Y. K., Nemethy, G. & Scheraga, H. A. (1987) Free Enegery of Hydration of Solute Molecules. 1. Improvement of the Hydration Shell Model by Exact Computations of Overlapping Volumes. *J Phys Chem* **91**, 4105–4109.

Kasper, P., Christen, P. & Gehring, H. (2000) Empirical calculation of the relative free energies of peptide binding to the molecular chaperone DnaK. *Proteins* **40**, 185–92.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C. & Vakser, I. A. (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A* **89**, 2195–9.

Kauzmann, W., Moore, K. & Schultz, D. (1974) Protein densities from X-ray crystallographic coordinates. *Nature* **248**, 447–9.

Kimura, S. R., Brower, R. C., Vajda, S. & Camacho, C. J. (2001) Dynamical view of the positions of key side chains in protein-protein recognition. *Biophys J* **80**, 635–42.

Kirkwood, J. G. (1935) Statistical mechanics of fluids mixtures. *J Chem Phys* **3**, 300–13.

Kocher, J. P., Rooman, M. J. & Wodak, S. J. (1994) Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol* **235**, 1598–613.

Koppensteiner, W. A. & Sippl, M. J. (1998) Knowledge-based potentials–back to the roots. *Biochemistry (Mosc)* **63**, 247–52.

Krämer, P. (2001) *Ermittlung, Charakterisierung und effiziente Verarbeitung von Oberflächenparametern für die Simulation von molekularen Wechselwirkungen der Proteine.* Dissertation, Universität zu Köln.

Kruse, R., Nauck, D. & Klawonn, F. (1997) Neuronale Fuzzy-Systeme. *Spektrum der Wissenschaft Dossier: Kopf und Computer* .

Larsen, T. A., Olson, A. J. & Goodsell, D. S. (1998) Morphology of protein-protein interfaces. *Structure* **6**, 421–7.

Lawrence, M. C. & Colman, P. M. (1993) Shape complementarity at protein/protein interfaces. *J Mol Biol* **234**, 946–50.

Lazaridis, T. & Karplus, M. (1999) Effective energy function for proteins in solution. *Proteins* **35**, 133–52.

Lazaridis, T. & Karplus, M. (2000) Effective energy functions for protein structure prediction. *Curr Opin Struct Biol* **10**, 139–45.

Leach, A. R. (2001) *Molecular Modelling Principles and Applications*. Pearson Education Limited.

Lee, B. & Richards, F. M. (1971) The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**, 379–400.

Leibzig, U. (2001) Werner Heisenberg Eine Ausstellung zum 100. Geburtstag. http://www.uni-leipzig.de/archiv/heisenberg/intro.htm.

Lijnzaad, P., Berendsen, H. J. & Argos, P. (1996) A method for detecting hydrophobic patches on protein surfaces. *Proteins* **26**, 192–203.

Lin, S. L., Nussinov, R., Fischer, D. & Wolfson, H. J. (1994) Molecular surface representations by sparse critical points. *Proteins* **18**, 94–101.

Lu, H. & Skolnick, J. (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* **44**, 223–32.

Lucas, K. (1986) *Angewandte Statistische Thermodynamik*. Springer Verlag.

Maiorov, V. N. & Crippen, G. M. (1992) Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol* **227**, 876–88.

Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I. & Ten Eyck, L. F. (2001) Protein docking using continuum electrostatics and geometric fit. *Protein Eng* **14**, 105–13.

McCoy, A. J., Chandana Epa, V. & Colman, P. M. (1997) Electrostatic complementarity at protein/protein interfaces. *J Mol Biol* **268**, 570–84.

McDonald, I. K. & Thornton, J. M. (1994) Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**, 777–93.

Melo, F. & Feytmans, E. (1997) Novel knowledge-based mean force potential at atomic level. *J Mol Biol* **267**, 207–22.

Meng, E. C., Shoichet, B. K. & Kuntz, I. D. (1991) Automated Docking with Grid-Based Energy Evaluation. *J Comp Chem* **13**, 505–524.

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

Miller, S. (1989) The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng* **3**, 77–83.

Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987a) Interior and surface of monomeric proteins. *J Mol Biol* **196**, 641–56.

Miller, S., Lesk, A. M., Janin, J. & Chothia, C. (1987b) The accessible surface area and stability of oligomeric proteins. *Nature* **328**, 834–836.

Miranker, A. D. (2000) Protein complexes and analysis of their assembly by mass spectrometry. *Curr Opin Struct Biol* **10**, 601–6.

Mitchell, J. B. O., Laskowski, R. A., Alex, A., Forster, J. & Thornton, J. M. (1999a) BLEEP - Potential of Mean Force Describing Protein-Ligand Interactions: II. Calculation of Binding Energies and Comparison with Experimental Data. *J Comp Chem* **20**, 1177–1185.

Mitchell, J. B. O., Laskowski, R. A., Alex, A. & Thornton, J. M. (1999b) BLEEP - Potential of Mean Force Describing Protein-Ligand Interactions: I. Generating Potential. *J Comp Chem* **20**, 1165–1176.

Miyamoto, S. & Kollman, P. A. (1993) What determines the strength of noncovalent association of ligands to proteins in aqueous solution? *Proc Natl Acad Sci U S A* **90**, 8402–6.

Miyazawa, S. & Jernigan, R. L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* **256**, 623–44.

Miyazawa, S. & Jernigan, R. L. (1999) Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins* **34**, 49–68.

Miyazawa, S. & Jerningan, R. L. (1985) Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules* **18**, 534–552.

Mohanty, D., Dominy, B. N., Kolinski, A., & Skolnick, J. (1999) Correlation between knowledge-based and detailed atomic potentials: application to the unfolding of the GCN4 leucine zipper. *Proteins* **35**, 447–52.

Moont, G., Gabb, H. A. & Sternberg, M. J. (1999) Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins* **35**, 364–73.

Morelli, X. J., Palma, P. N., Guerlesquin, F. & Rigby, A. C. (2001) A novel approach for assessing macromolecular complexes combining soft-docking calculations with NMR data. *Protein Sci* **10**, 2131–7.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K. & Olson, A. J. (1998) Automated Docking usind a Lamarckian Genetic algorithm and an Empirical Binding Free Energy Function. *J Comp Chem* **19**, 1639–1662.

Murcko, M. A. (1995) Computational methods to predict binding free energy in ligand-receptor complexes. *J Med Chem* **38**, 4953–67.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536–40.

Nauck, D., Klawonn, F. & Kruse, R. (1994) *Neuronale Netze und Fuzzy-Systeme. Grundlagen des Konnektivismus neuronaler Fuzzy-Systeme und der Kopplung mit wissenbasierten Methoden.* Vieweg Verlag Braunschweig Wiesbaden.

Nobeli, I., Mitchell, J. B. O., Alex, A. & Thornton, J. M. (2000) Evaluation of a Knowledg-Based Potential of Mean Force for Scoring Docked Protein-Ligand Complexes. *J Comp Chem* **22**, 673–688.

Norel, R., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1995) Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. *J Mol Biol* **252**, 263–73.

Norel, R., Petrey, D., Wolfson, H. & Nussinov, R. (1999) Examination of Shape Complementarity in Docking of Unbound Proteins. *Proteins* **36**, 307–317.

Novotny, J., Bruccoleri, R. E. & Saul, F. A. (1989) On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. *Biochemistry* **28**, 4735–49.

Nussinov, R. & Wolfson, H. J. (1991) Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci U S A* **88**, 10495–9.

Ooi, T., Oobatake, M., Nemethy, G. & Scheraga, H. A. (1987) Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc Natl Acad Sci U S A* **84**, 3086–90.

Palma, P. N., Krippahl, L., Wampler, J. E. & Moura, J. J. (2000) BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins* **39**, 372–84.

Pearlman, D. A., Case, D., Caldwell, J. C., Seiberl, G. L., Singh, U. C., Weiner, P. & Kollman, P. (1991) *AMBER 4.0*. Univeristiy of California.

Pierce, M. M., Raman, C. R. & Nall, B. T. (1999) Isothermal Titration Calometry of Protein-Protein-Interactions. *Methods* **19**, 213–221.

Rank, J. A. & Baker, D. (1997) A desolvation barrier to hydrophobic cluster formation may contribute to the rate-limiting step in protein folding. *Protein Sci* **6**, 347–54.

Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **261**, 470–89.

Rarey, M., Kramer, B. & Lengauer, T. (1999) Docking of hydrophobic ligands with interaction-based matching algorithms. *Bioinformatics* **15**, 243–50.

Reichelt, J. & Schomburg, D. (1996) BRAGI. *http://www.uni-koeln.de/math-nat-fak/biochemie/ds/dsbrag.htm* .

Reva, B. A., Finkelstein, A. V., Sanner, M. F. & Olson, A. J. (1997) Residue-residue mean-force potentials for protein structure recognition. *Protein Eng* **10**, 865–76.

Ringe, D. (1995) What makes a binding site a binding site? *Curr Opin Struct Biol* **5**, 825–9.

Ritchie, D. W. & Kemp, G. J. (2000) Protein docking using spherical polar Fourier correlations. *Proteins* **39**, 178–94.

Robert, C. & Janin, J. (1998) A soft, mean-field potential derived from crystal contacts for predicting protein-protein interactions. *J Mol Biol* **283**, 1037–47.

Schonbrun, J., Wedemeyer, W. J. & Baker, D. (2002) Protein structure prediction in 2002. *Curr Opin Struct Biol* **12**, 348–54.

Sheinerman, F. B. & Honig, B. (2002) On the role of electrostatic interactions in the design of protein-protein interfaces. *J Mol Biol* **318**, 161–77.

Shimada, J., Ishchenko, A. V. & Shakhnovich, E. I. (2000) Analysis of knowledge-based protein-ligand potentials using a self-consistent method. *Protein Sci* **9**, 765–75.

Shimizu, S. & Chan, H. S. (2002) Anti-cooperativity and cooperativity in hydrophobic interactions: Three-body free energy landscapes and comparison with implicit-solvent potential functions for proteins. *Proteins* **48**, 15–30.

Shoichet, B. K. & Kuntz, I. D. (1991) Protein docking and complementarity. *J Mol Biol* **221**, 327–46.

Sippl, M. J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213**, 859–83.

Sippl, M. J. (1993a) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des* **7**, 473–501.

Sippl, M. J. (1993b) Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355–62.

Sippl, M. J. (1995) Knowledge-based potentials for proteins. *Curr Opin Struct Biol* **5**, 229–35.

Sippl, M. J. (1996) Helmholtz free energy of peptide hydrogen bonds in proteins. *J Mol Biol* **260**, 644–8.

Sippl, M. J., Ortner, M., Jaritz, M., Lackner, P. & Flockner, H. (1996) Helmholtz free energies of atom pair interactions in proteins. *Fold Des* **1**, 289–98.

Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. (1997) Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci* **6**, 676–88.

Skolnick, J., Kolinski, A. & Ortiz, A. (2000) Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins* **38**, 3–16.

Taylor, J. S. & Burnett, R. M. (2000) DARWIN: a program for docking flexible molecules. *Proteins* **41**, 173–91.

Thomas, C., DeVries, P., Hardin, J. & White, J. (1996) Four-dimensional imaging: computer visualization of 3D movements in living specimens. *Science* **273**, 603–7.

Thomas, P. D. & Dill, K. A. (1996) Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol* **257**, 457–69.

Tong, A. H. Y., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C. W. V., Fields, S., Boone, C. & Cesareni, G. (2002) A Combined Expermental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules. *Science* **295**, 321–324.

Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1996) Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. *Crit Rev Biochem Mol Biol* **31**, 127–52.

Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1997) Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci* **6**, 53–64.

Vajda, S., Weng, Z., Rosenfeld, R. & DeLisi, C. (1994a) Effect of conformational flexibility and solvation on receptor-ligand binding free energies. *Biochemistry* **33**, 13977–88.

Vajda, S., Weng, Z., Rosenfeld, R. & DeLisi, C. (1994b) Effect of conformational flexibility and solvation on receptor-ligand binding free energies. *Biochemistry* **33**, 13977–88.

Vajda, S., Weng, Z. & DeLisi, C. (1995) Extracting hydrophobicity parameters from solute partition and protein mutation/unfolding experiments. *Protein Eng* **8**, 1081–92.

Vajda, S., Vakser, I. A., Sternberg, M. J. & Janin, J. (2002) Modeling of protein interactions in genomes. *Proteins* **47**, 444–6.

Vakser, I. A. (1996a) Long-distance potentials: an approach to the multiple-minima problem in ligand-receptor interaction. *Protein Eng* **9**, 37–41.

Vakser, I. A. (1996b) Low-resolution docking: prediction of complexes for under-determined structures. *Biopolymers* **39**, 455–64.

Vakser, I. A. & Aflalo, C. (1994) Hydrophobic docking: a proposed enhancement to molecular recognition techniques. *Proteins* **20**, 320–9.

Vakser, I. A., Matar, O. G. & Lam, C. F. (1999) A systematic study of low-resolution recognition in protein-protein complexes. *Proc Natl Acad Sci U S A* **96**, 8477–82.

Vasmatzis, G., Zhang, C., Cornette, J. & D.DeLisi (1996) Computational determination of side chain specificity for pockets in class I MHC molecules. *Mol Immunol* **33**, 1231–9.

Verkhivker, G. M. & Rejto, P. A. (1996) A mean field model of ligand-protein interactions: implications for the structural assessment of human immunodeficiency virus type 1 protease complexes and receptor-specific binding. *Proc Natl Acad Sci U S A* **93**, 60–4.

Vila, J., Williams, R. L., Vasquez, M. & Scheraga, H. A. (1991) Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. *Proteins* **10**, 199–218.

von Freyberg, B., Richmond, T. J. & Braun, W. (1993) Surface area included in energy refinement of proteins. A comparative study on atomic solvation parameters. *J Mol Biol* **233**, 275–92.

Vorobjev, Y. N., Almagro, J. C. & Hermans, J. (1998) Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. *Proteins* **32**, 399–413.

Wallqvist, A., Jernigan, R. L. & Covell, D. G. (1995) A preference-based free-energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. *Protein Sci* **4**, 1881–903.

Weng, Z., Vajda, S. & Delisi, C. (1996) Prediction of protein complexes using empirical free energy functions. *Protein Sci* **5**, 614–26.

Weng, Z., Delisi, C. & Vajda, S. (1997) Empirical free energy calculation: comparison to calorimetric data. *Protein Sci* **6**, 1976–84.

Wesson, L. & Eisenberg, D. (1992) Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci* **1**, 227–35.

Wlodawer, A., Nachman, J., Gilliland, G. L., Gallagher, W. & Woodward, C. (1987) Structure of form III crystals of bovine pancreatic trypsin inhibitor. *J Mol Biol* **198**, 469–80.

Wolfenden, R., Andersson, L., Cullis, P. M. & Southgate, C. C. (1981) Affinities of amino acid side chains for solvent water. *Biochemistry* **20**, 849–55.

Xiang, Z. & Honig, B. (2001) Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol* **311**, 421–30.

Xu, D., Tsai, C. J. & Nussinov, R. (1997) Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng* **10**, 999–1012.

Zhang, C., Vasmatzis, G., Cornette, J. L. & DeLisi, C. (1997) Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol* **267**, 707–26.

Zhang, L. & Skolnick, J. (1998) How do potentials derived from structural databases relate to truepotentials ? *Protein Sci* **7**, 112–22.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Mitchell, T., Miller, P., Dean, R. A., Gerstein, M. & Snyder, M. (2001) Global Analysis of Protein Activities Using Proteome Chips. *Science* **293**, 2101–2105.

Zimmermann, O. (2002) *Untersuchungen zur Vorhersage der nativen Orientierung von Protein-Komplexen mit Fourier Korrelationsmethoden.* Dissertation, Universität zu Köln.