Untersuchungen zur Vorhersage der nativen Orientierung von Protein-Komplexen mit Fourier-Korrelationsmethoden

Inaugural - Dissertation

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

vorgelegt von
Olav Zimmermann
aus Remscheid

Köln 2003

Danksagung

Diese Arbeit wurde von Mai 1997 bis Oktober 2002 bei Herrn Prof. Dr. Dietmar Schomburg am Institut für Biochemie der Universität zu Köln angefertigt.
Mein erster Dank gilt allen Mitgliedern der Arbeitsgruppe Schomburg, insbesondere Christian aus dem Spring, Oliver Martin, Holger Klein und Jörn Behre, deren Praktikumsergebnisse zu dieser Arbeit beigetragen haben. Für die vielen anregenden Diskussionen, die hoffentlich noch lange weitergeführt werden gilt mein nächster Dank den Mitgründern der "Science Factory", Sebastian Schneckener, Oliver Leven und Fuad Abdallah, sowie Gerd Wohlfahrt. Sie sorgten für reichlich Inspiration, Motivation und manchen wissenschaftlichen Diskurs. Allen die mich in der Endphase unterstützt und mir den Rücken freigehalten haben möchte ich besonders herzlich danken, allen voran Vera Grimm, Christian Hoppe, Oliver Hofmann und Alexander Schliep.
Mein besonderer Dank gilt meinem Doktorvater Herrn Prof. Dr. Dietmar Schomburg. Er vertraute mir dieses interessante Thema an, gab mir viel Freiraum und war immer zu einem fachlichen Dialog bereit.

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 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.

Olav Zimmermann

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

Tag der mündlichen Prüfung: 12. Februar 2003

Kurzzusammenfassung

Diese Arbeit beschreibt die Entwicklung des Dockingprogramms CKORDO und analysiert seine Leistung bei der Vorhersage von Protein-Komplexen sowohl aus Teilen von Kokristallstrukturen als auch aus unabhängigen Einzelstrukturen. Das Verfahren basiert auf einer von Katchalski-Katzir *et al.* (1992) eingeführten Methode, bei der die Korrelation zweier diskretisierter Proteinoberflächen im Fourierraum um ein Vielfaches effizienter erfolgt als bei direkter Berechnung. Eins der Proteine wird in regelmäßigen Winkelabständen rotiert bis der gesamte Rotationsraum abgesucht ist. CKORDO berechnet für jede Orientierung Korrelationen für die Oberflächengeometrie, sowie für ein elektrostatisches und ein hydrophobes Potential. Für die Maxima der geometrischen Korrelation wird außerdem der Abstand der Interfaceflächen und der Wert eines Atom-Atom Kontaktpotentials berechnet. Die ermittelten Werte werden zur Filterung und zur Klassifizierung durch eine *Support Vector Machine* eingesetzt. Praktisch alle Kokristallstrukturen werden innerhalb der ersten 100 Lösungen korrekt vorhergesagt. Auch Teilstrukturen eines Transmembranproteins können richtig gedockt werden. Von zwölf aus unabhängigen Einzelstrukturen bestehenden Testfällen werden für acht Komplexe richtige Orientierungen unter den ersten 1000 Vorhersagen ermittelt. Die Diskussion untersucht die Gründe für die Schwierigkeiten mit den anderen vier Komplexen, vergleicht die Leistungsfähigkeit mit weiteren Ansätzen und zeigt im Ausblick Möglichkeiten zur Weiterentwicklung des Programms CKORDO auf.

Abstract

This work describes the development of the protein docking program CKORDO analyzing it's performance to predict protein structures from both parts of co-crystallized structures and from individually crystallized protein structures. The approach is based on a method introduced by Katchalski-Katzir *et al.* (1992) which accelerates the calculation of the correlation between two protein surfaces by transforming the discretized grid representation into Fourier space prior to the actual calculation of the correlation. By rotating one of the proteins at fixed angle

increments a full scan of the rotational space is performed. CKORDO calculates correlation values for the surface geometry, an electrostatic and a hydrophobic potential. For each maximum of the geometric correlation the gap width of the interface and the value of an atom-atom contact potential is calculated. The values are used for filtering and for classification by a support vector machine classifier. Almost all cocrystallized structures show near native orientations within the top 100 predictions. Also parts of a transmembrane protein complex can be successfully predicted. Starting from individual structures eight out of twelve unbound test cases have correct orientations within the top 1000 predictions. In the discussion the difficulties in predicting the other four complexes are analyzed. The comparison of the performance with other approaches leads to an outlook discussing possible enhancements of CKORDO.

Inhaltsverzeichnis

Abkürzungen und Begriffe

Ångstrøm	Längeneinheit: 1 Å= 10^{-10}m
Backbone	Hauptkette eines Polypeptids: C- und N-Atome der Peptidbindung, C_α-, und Carbonyl O-Atome der Aminosäuren ohne die Seitenketten
Bound-Docking	Docking von Komplexteilen, die aus einem Komplex ausgeschnitten wurden
Unbound-Docking	Docking von Proteinen, die in ihrer nativen Form vorliegen
Buried Surface	Während der Assoziation vergrabene Fläche in $Å^2$, auch als bur. surface abgekürzt
Docking	Assoziation von Makromolekülen
Interface	Kontaktfläche zwischen zwei Makromolekülen (Proteinen)
NMR	*Nuclear Magnetic Resonance*
FFT	Fast Fourier Transformation
RMSD	*root mean square deviation*
RNA	*ribo nucleic acid* = Ribonukleinsäure
Sampling	Probennahme
SVM	*Support Vector Machines*
DSSP	Dictionary of protein secondary structure
ITC	*Isothermal Titration Calorimetry*
CAPRI	*Critical Assessment of Predicted Interactions*
DNA	*desoxyribo nucleic acid*=Desoxyribonukleinsäure
ASA	*accessible surface area*
AS	Aminosäure
tp	*true positive* = richtig Positiver
tn	*true negative* = richtie Negativer
fp	*false positive* = falsch Positiver
tn	*true negative* = richtig Falscher
PPV	*positive prediction value*
Sens.	Sensitivität

Spez.	Spezifität
Res.	*resolution* = Auflösung
pankr.-Inh.	pankreatischer Inhibitor
vdW	van der Waals
geom.	geometrisch
z.B.	zum Beispiel
z.T.	zum Teil
z.Z.	zur Zeit
d.h.	das heißt

Aminosäuren

Alanin	ALA	A
Cystein	CYS	C
Asparaginsäure	ASP	D
Glutaminsäure	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

1 Einleitung

Lebende Systeme unterscheiden sich von nicht lebenden Systemen durch eine Faktorenkombination aus Selbstreproduktion, Stoffwechsel und Reaktivität. Der Begriff System impliziert dabei ein koordiniertes Verhalten. Die Fähigkeit zu einer solchen Koordination innerhalb lebender Systeme beruht auf den besonderen Eigenschaften der biologischen Makromoleküle. Insbesondere die Variabilität der Proteine bezüglich 3D-Struktur und physiko-chemischer Eigenschaften ist die entscheidende Voraussetzung für die Entwicklung von Spezifität. Die dreidimensionale Gestalt der Proteine und ihre dadurch definierte Bindungsspezifität verhindert, dass in einer Zelle mit mehreren tausend unterschiedlichen Molekülen beliebige Interaktionen stattfinden und so zu einem Chaos führen. Stattdessen herrscht im komplexen System Zelle eine fein regulierte Ordnung, so dass in den 40er und 50er Jahren auch eine Reihe von Physikern begann sich für die Molekularbiologie zu interessieren und die Zelle aus den neuen Blickwinkeln von Chemie und Physik zu studieren (Schrödinger, 1946; Delbrück, 1963).

Alle wichtigen Vorgänge in der Zelle basieren auf Spezifität: Stoffwechsel, Signalverarbeitung, genetische Regulation; und fast immer stehen spezifische Wechselwirkungen von Proteinen im Mittelpunkt. Zur Durchführung ihrer Aufgaben als Rezeptoren, Signalüberträger, Enzyme, Regulatoren oder Transporter interagieren Proteine mit DNA, RNA und niedermolekularen Substanzen, vor allem aber – mit anderen Proteinen. Spezifische Protein-Protein Wechselwirkungen sind daher die zentrale intra- und interzelluläre Kommunikationsplattform und damit ein wichtiges Forschungsfeld in der Molekularbiologie. Die Vorhersage solcher Interaktionen hilft Forschern ihre experimentellen Ergebnisse zu interpretieren und ist ein wichtiger Schritt auf dem Weg zur Simulation zellulärer Systeme. Computerprogramme, die Vorhersagen über die Orientierung der Untereinheiten in einem Protein-Komplex liefern werden *Docking*-Programme genannt. Die Weiterentwicklung eines solchen *Docking-Programms*, seine Parametrierung, Einsatzmöglichkeiten und Grenzen sollen in dieser Arbeit beschrieben werden.

1.1 Protein-Protein Komplexe

Protein-Protein Interaktionen können nach verschiedenen Gesichtspunkten kategorisiert werden, unter anderem nach ihrer zeitlichen Dauer, der Ähnlichkeit der Bindungspartner oder nach der Funktion der Komplexe. Allerdings lassen sich beispielsweise große multifunktionelle Protein-Komplexe nur schwer in ein solch starres Schema einordnen (s. Tabelle 1.1).

Bindungspartner	dynamisch	statisch
identisch	dynamische Zellstrukturen (Mikrotubuli)	statische Zellstrukturen (Keratin)
unterschiedlich	Kontaktkomplexe regulatorische Komplexe Signaltransduktionskomplexe (Rhodopsin-Transducin)	heteromere Komplexe (ATP-Synthase)

Tabelle 1.1: Beispiele zur Einteilung von Protein-Protein-Komplexen

Protein-Protein-Interaktionen haben in der Zelle ganz unterschiedliche Funktionen. Die Interaktionen polymerer Gerüstproteine verleihen biologischen Strukturen Stabilität, sie existieren oft während der ganzen Lebensdauer der Zelle als Komplex. Andere Komplexinteraktionen sind nur von kurzer Dauer, insbesondere wenn die Bindung eine Signalweitergabe darstellt, wie bei der Phosphorylierung von Proteinen durch spezifische Proteinkinasen. Serin-Proteasen wie Kallikrein und Thrombin interagieren in einer proteolytischen Kaskade um die Blutgerinnung zu veranlassen. Bestimmte Proteininteraktionen dienen sogar als Waffen. So verwenden Giftschlangen wie die Mamba (*D*endroaspis sp.) Proteine zur Inhibition der Acetylcholinesterase. Dies führt dazu, daß der Neurotransmitter Acetylcholin nicht mehr abgebaut werden kann und verursacht dadurch bei dem Opfer Nervenlähmungen. Als Teil des körpereigenen Abwehrarsenals fungieren Antikörper, die eine Vielzahl unterschiedlicher Bindungsspezifitäten haben können, bei der Abwehr von körperfremden Substanzen wie Krankheitskeimen.

1.2 Spezifität und Evolution

Spezifität besteht aus zwei unterschiedlichen Komponenten: der Möglichkeit einer stabilen Bindung zwischen den Bindungspartnern und der geringeren Wahrscheinlichkeit zur Ausbildung anderer stabiler Bindungen. Die Spezifität von Interfaces wird nicht durch großräumige Strukturunterschiede erreicht, sondern bereits einzelne Mutationen reichen aus um eine Bindungsspezifität aufzuheben, zu verändern oder neu zu erzeugen. Dramatische Folgen hat dies z.B. beim Hämoglobin, wo die Mutation einer einzigen Aminosäure ausreicht um die Bindungsspezifität so zu erweitern, daß sich Fäden aus Hämoglobinmolekülen bilden und zum Krankheitsbild der Sichelzellenanämie führen (Ingram, 1956). Dimerisierungverlust kann gleichfalls durch eine einzige Mutation ausgelöst werden (Bennett *et al.*, 1994, 1995; Dong *et al.*, 1999). Beim Lac-Repressor aus *E.coli* reicht eine Mutation von Aspartat 278 nach Leucin in einem Monomer aus, um die Dimerisierung nur noch mit einem anderen Lac-Repressor Monomer mit der gleichen Mutation zu ermöglichen, die Bindung an ein Wildtyp Monomer aber vollständig zu verhindern (Spott *et al.*, 2000). Die Fähigkeit eine neue Spezifität in einem Schritt ohne die Zwischenstufe eines unspezifischen Interfaces erreichen zu können, ist eine wichtige Voraussetzung zur Evolution von Interfaces (Xu *et al.*, 1998). Auch die spontane Neuentwicklung eines Dimerinterfaces mit hoher Stabilität aus einem monomeren Protein durch eine einzige Mutation ist schon beschrieben worden (Green *et al.*, 1995).

1.3 Protein-Protein Interfaces

1.3.1 Größe

Typische Interfacegrößen bei Protease-Inhibitor Komplexen und Antikörper-Antigen Komplexen liegen bei 1600 ± 400 Å^2 (Janin & Chothia, 1990; Janin, 2001). Diese Standardgröße ist in bemerkenswert guter Übereinstimmung mit den Interfacegrößen, die bei Zufallsassoziationen von Proteinen in Computersimulationen bestimmt wurden (Janin & Rodier, 1995). Auch in Proteinkristallen gibt es solche zufallsbedingten nicht-nativen Interfaces zwischen den einzelnen Molekülen.

Verschiedene Arbeiten haben diese Kristallkontakte mit nativen Interfaces verglichen (Janin & Rodier, 1995; Müller, 1996; Carugo & Argos, 1997; Dasgupta *et al.*, 1997). Die Gesamtfläche der Kristallkontakte zwischen zwei Proteinen unterscheidet sich mit $1100 - 4400$ Å^2 nicht erkennbar von der Größe funktioneller Kontaktflächen. Durch die starke Fragmentierung der Kristallkontaktflächen sind die einzelnen Oberflächenabschnitte im Durchschnitt deutlich kleiner als bei funktionellen Proteininterfaces. 45% haben eine Größe von weniger als 100 Å^2 (Carugo & Argos, 1997).

1.3.2 Aufbau

Lijnzaad und Argos fanden, daß der größte oder zweitgrößte hydrophobe Oberflächenabschnitt (Patch) in 90% der Fälle mit dem Interface überlappt (Lijnzaad & Argos, 1997). Der Anteil solcher hydrophober Patches unterscheidet sich jedoch je nach Art der Proteinkomplexe. So finden sich große hydrophobe Bereiche fast nur bei Homodimeren, also Proteinen die im allgemeinen während ihrer gesamten Lebensdauer gebunden vorkommen. Genau aus diesem Grund existieren 3D-Strukturen nur für die Dimere, nicht aber für die isolierten Monomere.

1.3.3 Form

Große hydrophobe Interfaces wie die der angesprochenen Homodimere sind eher planar, während sie bei Heterodimeren oft gekrümmt sind. Dies führt bei einigen Proteinen zu tief verzahnten Interfaces, die nur in der exakten Orientierung zueinander eine hohe geometrische Komplementarität aufweisen (s. Kapitel 1.6.1), bei nur geringfügig falscher Rotation aber bereits zu sterischen Inkompatibilitäten führt (s.a. Kapitel 4.1.2 auf Seite 52).

1.3.4 Vorhersage von Interfaceoberflächen

Wäre einem Stück Proteinoberfläche anzusehen, ob es als Interaktionsfläche in Frage kommt und wäre außerdem der Interaktionspartner bekannt, so könnten die Vorhersagen von Proteinkomplexen erheblich beschleunigt werden. Darüber hinaus könnte vorab geklärt werden, ob ein Protein überhaupt einen Komplex

mit irgendeinem anderen Protein eingeht. Diese Information ist selbst für strukturell aufgeklärte Proteine nicht immer bekannt. Die Analyse von hydrophoben *Patches* (Lijnzaad & Argos, 1997) und weiteren Eigenschaftsprofilen von Oberflächen haben zu Ergebnissen geführt, die bei der Vorhersage von Interfaces eingesetzt werden können (Jones & Thornton, 1997a,b).

1.4 Experimentelle Techniken

Trotz einer Reihe experimenteller Techniken zur Untersuchung von Protein-Protein Wechselwirkungen ist es gegenwärtig noch nicht möglich, das vollständige Netzwerk aller Interaktionen einer Zelle zu bestimmen. Diese Techniken unterscheiden sich bezüglich ihrer Zielsetzung, sowie in der Art und dem Detailgrad ihrer Resultate.
Neben einer Reihe klassischer Verfahren, darunter *phage display*, chemisches Crosslinking, analytische Ultrazentrifugation und Retardierungsgele sind in den letzten Jahren einige neue Techniken in den Mittelpunkt gerückt um, zunehmend automatisiert, qualitative und quantitative Daten über Protein-Protein Wechselwirkungen zu liefern.

Technik	Information über Interaktion
ITC	Bindungsstärke, Kinetik
2-Hybrid	Bindungspartner
Affinitätschromatographie	Bindungspartner
Plasmon Resonanz	Bindungsstärke, Kinetik
Kraftmessung	Bindungsstärke
Röntgenstrukturanalyse	3D Feinstruktur
Kernspinresonanz	3D Feinstruktur und Dynamik

Tabelle 1.2: Neuere experimentelle Techniken zur Protein-Protein Interaktion

- ITC

  Die Isotherme Titrations-Kalorimetrie bestimmt die Reaktionsenthalpie bei der Bindung eines Liganden an einen Rezeptor. Aus der Sättigungskinetik dieser Reaktion lassen sich die Bindungskonstanten bei schwach bindenden

Proteinkomplexen und die Stöchiometrie des Komplexes bestimmen. Bindungskonstanten oberhalb von 10^9 M, wie sie typisch für Enzym-Inhibitor Komplexe oder die Bindung hochaffiner Antikörper an ihre Epitope sind, lassen sich damit nicht bestimmen (Pierce *et al.*, 1999).

- *Yeast Two Hybrid System*

  Das verbreitetste experimentelle System zur qualitativen Untersuchung von Protein-Protein Interaktionen ist das sogenannte *yeast two hybrid system* (Fields & Song, 1989). Bestimmte Transkriptionsfaktoren aus Hefe lassen sich in eine DNA-bindende Domäne und eine aktivierende Domäne trennen, und sind nur dann aktiv, wenn beide Domänen sich in unmittelbarer Nähe zueinander befinden. Das Gemisch der beiden unverbundenen Domänen ist weitgehend inaktiv. Wird nun Protein X, dessen Interaktionspartner gesucht wird, mit der einen Domäne und die potentiellen Bindungspartner Y mit der anderen Domäne fusioniert, so führt nur die Bildung eines Proteinkomplexes X-Y zur Rekonstitution eines funktionsfähigen Transkriptionsfaktors. Als Folge wird ein Reportergen transkribiert, welches hinter einen zum Transkriptionsfaktor passenden Promoter kloniert wurde. Je nach Reportersystem kann das von dem Reportergen kodierte Protein durch Enzymaktivität, Lichtemission oder Zellwachstum detektiert werden.

- Affinitätschromatographie

  Die Firma Cellzome in Heidelberg hat ein als TAP (*Tandem Affinity Purification*) bezeichnetes Verfahren entwickelt, bei dem ein Protein durch gezielte PCR-Mutagenese am 3'-Ende des Gens mit einem zusätzlichen Sequenzstück versehen wird. Dieses Zusatzstück ermöglicht die Aufreinigung an einer passenden Affinitätssäule. (Rigaut *et al.*, 1999). Unter milden Bedingungen können dadurch eventuell an das Protein gebundene Komplexpartner mit aufgereinigt werden. Mit Hilfe dieses Ansatzes wurden 232 Komplexe in der Bäckerhefe *Saccharomyces cerevisiae* identifiziert (Gavin *et al.*, 2002).

- Oberflächen-Plasmon-Resonanz

  Die Oberflächen-Plasmon-Resonanz (engl.: *Surface Plasmon Resonance*, SPR) mißt den Totalreflektionswinkel an der Grenzfläche zwischen einer goldbeschichteten Glasfläche und einem optisch weniger dichten Medium wie Luft oder Wasser. Durch Resonanz der von den Photonen erzeugten Elektronendichtewellen mit immobilisierten Makromolekülen auf der goldbeschichteten Seite wird der Totalreflektionswinkel beeinflußt. So lassen sich Interaktionen dieser Makromoleküle mit hoher Empfindlichkeit messen und über zeitaufgelöste Messungen auch kinetische Konstanten bestimmen (Leatherbarrow & Edwards, 1999).

- Kraftmessung

  Durch die Atom-Kraft-Mikroskopie (AFM) ist es möglich, mit einer ultrafeinen Feder die Kräfte zwischen zwei Makromolekülen direkt zu messen (Clausen-Schaumann *et al.*, 2000). Eine weiterentwickelte Technik mit dem Namen C-FIT (*Congruent Force Intermolecular Test*) hat die Firma Nanotype aus München zum Patent angemeldet. Dieses Verfahren mißt die Kraft mit einer Referenzinteraktion statt mit einer ultrafeinen Feder, d.h. es wird festgestellt ob der Komplex stabiler oder unstabiler als eine bekannte Interaktion ist. Als Referenzen dienen dabei leicht synthetisierbare molekulare Komplexe bekannter Stärke, z.B. DNA-Duplexstränge bestimmter Sequenz (H. Clausen-Schaumann, pers. Mitteilung).

- Röntgenkristallografie

  Röntgenstrukturen bieten die z.Z. genauesten strukturellen Daten über Protein-Protein Interaktionen. Die Technik nutzt die Röntgenbeugung an den Atomen regelmäßig aufgebauter Proteinkristalle. Aus den Beugungsmustern kann die dreidimensionale Struktur der Proteine und ihrer Seitenketten z.T. in atomarer Auflösung berechnet werden. Die erfolgreiche Züchtung der für die Röntgenkristallografie benötigten Protein-Einkristalle ist das zentrale Problem für die Gewinnung einer größeren Anzahl solch hochaufgelöster Strukturen. Große Proteinkomplexe, insbesondere Membranproteinkomplexe lassen sich nur schwer kristallisieren.

- *Nuclear Magnetic Resonance* (NMR)

  Die Kernspinresonanz erlaubt die Messung interatomarer Abstände <500 pm (Voet & Voet, 1992). Derartige Untersuchungen können Aufschluss über die Dynamik von Seitenketten in Protein-Protein Interfaces geben. Im Gegensatz zur Röntgenkristallografie befinden sich die Proteine bei der NMR-Messung frei in Lösung, so daß die Strukturen der Seitenketten nicht durch Kristallkontakte beeinflußt werden. Mit den gegenwärtigen Methoden ist es allerdings nur möglich Strukturen bis zu einer Größe von maximal 300 Aminosäuren durch Kernspinresonanz aufzuklären (Wüthrich, 1989).

1.5 Sequenzbasierte Methoden zur Vorhersage von Komplexen

1.5.1 Homologie

Viele Eigenschaften von Proteinen lassen sich vorhersagen, wenn diese für sequenzhomologe Proteine bekannt sind. In begrenztem Maße gilt dies auch für die Identifizierung von Interfaces. Dies ist insbesondere dort möglich, wo sich in der Natur bestimmte spezialisierte Bindungsdomänen herausselektiert haben, die in verschiedenen Proteinen vorkommen und eine bestimmte Protein-Bindungseigenschaft vermitteln. Beispiele für solche in verschiedenen Kontexten verwendeten Protein-Bindungsdomänen sind die SH3-Domänen, die mit kurzen, Prolin-reichen Motiven anderer Proteine interagieren. Darüber hinaus herrscht ein höherer Selektionsdruck auf funktionelle Oberflächen als auf den Rest der Proteinoberfläche. Dies führt zu stärkerer Konservierung von Aminosäuren, die zum Interface gehören. Durch Vergleich der Sequenzen eines Proteins in verschiedenen Spezies lassen sich sogenannte evolutionäre Spuren auffinden und zur Vorhersage von Interfaces einsetzen (Lichtarge *et al.*, 1996).

1.5.2 Korrelierte Mutationen

Bei Interfaces zwischen zwei unterschiedlichen Proteinen muß eine eventuelle Mutation in dem einen Bindungspartner meistens durch eine Mutation in dem anderen Bindungspartner kompensiert werden um die Erhaltung der Interaktion zu gewährleisten (s.a. Kapitel 1.2). Einige Forschergruppen versuchen solche korrelierten Mutationen durch Sequenzvergleiche über verschiedene Spezies hinweg zu detektieren und für sequenzbasierte Vorhersagen von Bindungspartnern und den am Interface beteiligten Aminosäuren auszunutzen (Pazos *et al.*, 1997).

1.6 Strukturbasierte Methoden zur Vorhersage von Komplexen

1.6.1 Komplementarität

Praktisch alle Methoden zur Vorhersage von Komplexen, die von den Atomkoordinaten der einzelnen Proteine ausgehen, beruhen auf der Annahme, daß die Komplexteile zueinander komplementär sind. Dies gilt sowohl für die Vorhersage von Protein-Protein als auch von Protein-Ligand Komplexen. Neben der Frage welche Eigenschaften komplementär sind, ist entscheidend ob diese Komplementarität bereits vor der Komplexbildung ausgebildet ist und somit für die Vorhersage ausgenutzt werden kann oder ob sie erst während der Assoziation der Komplexteile entsteht.

1.6.1.1 Geometrie

Es wird angenommen, daß der Energiegewinn bei der Komplexbildung hauptsächlich aus dem hydrophoben Effekt stammt (Honig & Nicholls, 1995). Demzufolge sollten sich die Komplexpartner in unmittelbarer Nähe zueinander befinden, da hydrophobe Wechselwirkungen kurzreichweitiger Natur sind. Weiterhin sollte das Interface wenig Lücken aufweisen, was zur Folge hat, daß die beiden Komplexteile auf der Makroskala einen ähnlichen Krümmungsradius und auf der Mikroskala zueinander passende Seitenkettenanordnungen haben. Nur dann

stehen auf beiden Detailebenen Konvexitäten in Teil A Konkavitäten in Teil B gegenüber und umgekehrt.

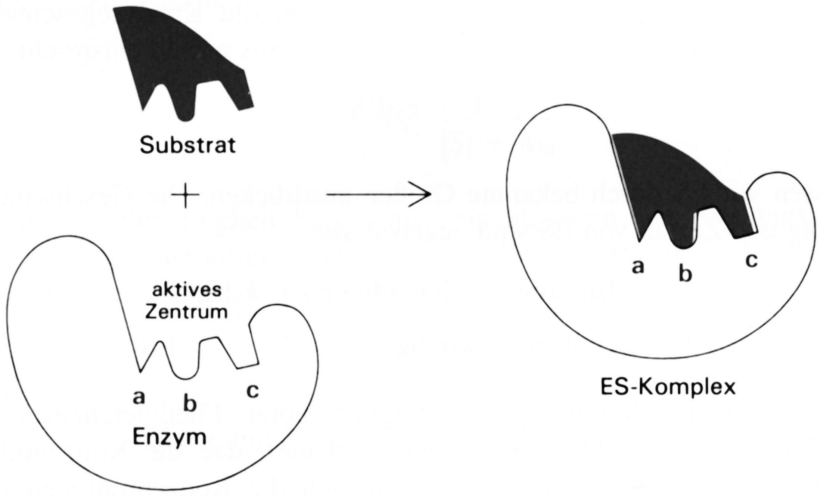


Abbildung 1.1: Emil Fischers Schlüssel-Schloß Modell. Abbildung entnommen aus: (Stryer, 1995)

Diese genaue Passform der beiden Bindungspartner läßt sich an vielen strukturell aufgeklärten Protein-Komplexen auch im unkomplexierten Zustand beobachten (Betts & Sternberg, 1999). Die Abbildung des Bindungspartners ist demnach also bereits vorgeformt. Dies zeigt, daß die vor mehr als Hundert Jahren von Emil Fischer für die Erklärung der Enzym-Substrat Bindung aufgestellte Schlüssel-Schloß Hypothese z.T. auch für Protein-Protein Komplexe Gültigkeit hat (Fischer, 1894) (s. Abbildung 1.1). Für eine Reihe von Proteinen-Komplexen ist aber bekannt, daß Teile des Interfaces beim Kontakt mit dem Bindungspartner ihre Form ändern. Diese folgen dem *induced fit* Modell und die Proteine sind im unkomplexierten Zustand nur eingeschränkt komplementär zueinander (s. Abbildung 1.2). Zur Quantifizierung geometrischer Komplementarität wurden verschiedene Ansätze entwickelt. Bereits 1986 hat Michael Connolly eine Methode beschrieben, die an den topographischen Extremwerten, d.h. den Gipfeln und Tälern der Oberflächenlandschaft die Raumwinkel (*solid angles*) berechnet. In jedem dieser Extremwerte sollten die Raumwinkel beider Teile zusammengenommen annähernd dem Raumwinkel einer Kugel entsprechen (Connolly, 1986).

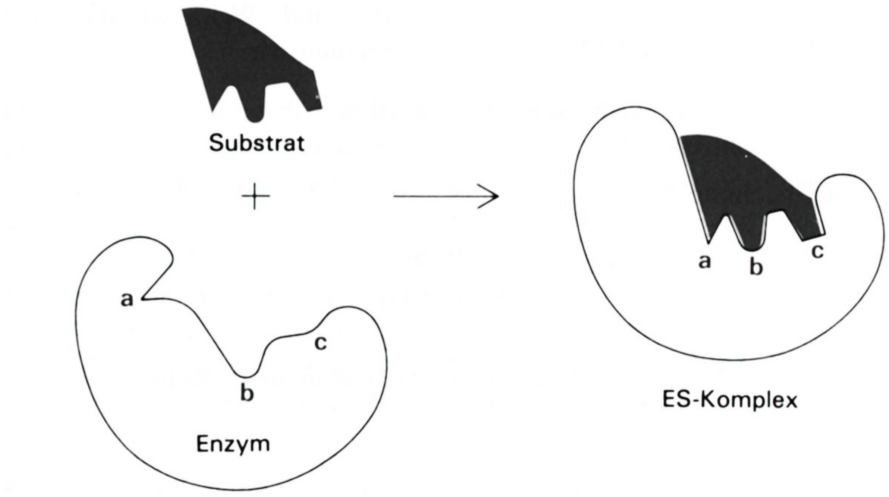


Abbildung 1.2: Das *induced fit* Modell. Abbildung entnommen aus: (Stryer, 1995)

Andere Verfahren klassifizieren die Form einzelner Oberflächenteile (Connolly, 1992), suchen komplementäre Dreiecke (Lin *et al.*, 1994) oder beschreiben die Abweichungen der Oberfläche zur Konvexen Hülle (Badel-Chagnon *et al.*, 1994) bzw. deren Verallgemeinerung, den sogenannten Alpha-Formen (Edelsbrunner & Mucke, 1994; Peters *et al.*, 1996). Schließlich kann statt einzelner kritischer Oberflächenpunkte auch die gesamte Oberfläche auf Komplementarität überprüft werden. Hierzu werden die entsprechenden Oberflächen als mehrdimensionale Funktionen dargestellt und miteinander korreliert. Wie in Abschnitt 2.3.3 dargestellt, sind Fourier-Methoden in besonderem Maße für solche Korrelationsberechnungen geeignet. Anders als auf der bisher besprochenen Makroskala sind auf der Mikroskala Komplementaritäten physiko-chemischer Eigenschaften wichtig. Einige von diesen werden in den folgenden Kapiteln erläutert.

1.6.1.2 Wasserstoffbrückenbindungen

Insbesondere geladene Interfaces in bestimmten Protein/Protein-Inhibitor Komplexen zeigen ein Muster an Wasserstoffbrückenbindungen. Ausgehend von der Annahme, daß sich die entsprechenden komplementären Wasserstoffdonor/-akzeptor Muster auch auf den Oberflächen der Einzelproteine vor der Kom-

plexierung finden lassen, wurde von Peter Wilson ein entsprechendes Suchprogramm entwickelt. Dieses Programm, das als Vorfilter für das Dockingprogramm KORDO eingesetzt wurde, sucht nach Orientierungen mit einer maximalen Anzahl komplementärer Paarkombinationen von Wasserstoffdonoren und -akzeptoren (Meyer *et al.*, 1996). Peter Krämer hat diese Methode um die Suche nach komplementären Tripelkombinationen erweitert (Krämer, 2001). Es sind allerdings Komplexe bekannt, die keine Wasserstoffbrückenbindungen im Interface aufweisen (Xu *et al.*, 1997) für die Wasserstoffbrückenvorfilter nur falsche Vorhersagen liefern können.

1.6.1.3 Elektrostatik

Alle Ladungen in einem Protein tragen zu einem charakteristischen Ladungsfeld um die Oberfläche bei. Solche Ladungsmuster in Protein-Komplexen können komplementär zueinander sein, da unkompensierte Ladungen energetisch ungünstig sind (Gabb *et al.*, 1997; Sheinerman & Honig, 2002). Dies gilt insbesondere wegen der geringen Dielektrizitätskonstante des Interfaces, die mit der im Proteininneren vergleichbaren ist. Untersuchungen der Gruppe von Peter Colman zeigen allerdings, daß hierbei nicht allein von den lokalen Ladungen ausgegangen werden darf, sondern das Ladungsfeld des gesamten Proteins betrachten werden muß (McCoy *et al.*, 1997). Die Berechnung solcher Felder z.B. durch Poisson-Boltzmann Methoden ist jedoch sehr rechenzeitaufwendig.

1.6.1.4 Aminosäuremuster

Schließlich lassen sich insbesondere Ladungskomplementaritäten auch in Form von Aminosäuremustern beschreiben. Dies hat zu einer Vielzahl von empirischen Potentialen geführt, die durch Bestimmung von Häufigkeiten und Abständen von Interaktionspaaren d.h. Paaren von Aminosäuren, Atomgruppen oder einzelnen Atomen versuchen die Bindungsenergie abzuschätzen. Meist geschieht dies durch die Annahme einer Boltzmann-Beziehung zwischen Häufigkeit und Energie (Sippl, 1990), sowie der Additivität einzelner Energiebeiträge. Die resultierenden Pseudoenergiewerte können dann zur Bewertung von Interfacehypothesen eingesetzt werden (Moont *et al.*, 1999; Gohlke *et al.*, 2000; Grimm, 2002).

1.6.2 Protein-Protein Docking

1.6.2.1 Anwendung

Die häufigste Frage, die mit Hilfe von Docking-Programmen zu klären versucht wird, ist die nach der Orientierung, in der zwei (oder mehr) Proteinmonomere unter natürlichen Bedingungen aneinander binden, d.h. nach der Lage des Bindungsinterfaces. Diese Vorhersagen können dabei helfen die Positionen in den entsprechenden Proteinen zu finden, die zur Beeinflussung der Bindungsstärke oder Spezifität auf DNA-Ebene mutiert werden müssen. Ebenso lassen sich auf diese Weise Hypothesen für den Entwurf inhibitorisch wirkender Peptide, Peptidomimetika oder niedermolekularer Liganden für Diagnose und Therapie entwickeln (Protein-Ligand-Docking, Drug Design).

1.6.2.2 Tests und Vergleiche von Docking-Vorhersagen

Zum Testen der Vorhersage von Komplexen können zwei Problemklassen unterschieden werden.

- *bound*-Docking

  Die weitaus größte Zahl der Veröffentlichungen zum Thema Protein-Protein Docking testet ihre Verfahren und Algorithmen mit den Koordinaten der Einzelteile der kokristallisierten Komplexstruktur (*bound*-Docking).

- *unbound*-Docking

  Erst seit wenigen Jahren wird versucht, Komplexe aus den Koordinaten getrennt kristallisierter Strukturen vorherzusagen. Von diesen Proteinen muß bekannt sein, das sie interagieren und auf welche Weise. Die Schwierigkeit beim *unbound*-Docking besteht darin, daß die Einzelproteine nicht genau die Form haben, die nach Ausbildung des Komplexes vorliegt. Die Strukturen können sich also durch eine gegenseitige Adaptation der Orientierungen der Komplexpartner bei der Interaktion ändern: Polare Seitenketten können sich ausrichten um Salz- oder Wasserstoffbrücken auszubilden, hydrophobe Aminosäurereste werden z.T. erst exponiert, wenn hydrophobe Gruppen in unmittelbarer Nähe vorhanden sind. Schließlich kann sich auch

das Rückgrat der Proteine verbiegen oder knicken. *Unbound*-Docking Tests sind auf Grund der Strukturänderungen bei der Komplexbildung zwar erheblich schwieriger, stellen aber den realistischen Anwendungsfall dar.

- Vergleichender Wettbewerb

  Da die Ansätze zum Protein-Protein Docking z.T. sehr unterschiedlich, die Anzahl der verfügbaren Testfälle aber immer noch gering ist, besteht eine gewisse Gefahr, daß die Methoden zu sehr auf die wenigen Testfälle zugeschnitten sind. Um einen Überblick über den Status Quo allgemein im Bereich Docking zu bekommen wurde daher bereits 1996 ein erster Blindtest publiziert (Strynadka *et al.*, 1996), bei der eine noch nicht publizierte Komplexstruktur aus den separat kristallisierten Einzelproteinen vorhergesagt werden sollte. Im darauf folgenden Jahr gab es bereits eine eigene Docking Rubrik im CASP-Wettbewerb (*Critical Assessment of Structure Prediction*), der eigentlich ein Vergleichstest der Vorhersagemethoden für Proteinstrukturen aus ihren Aminosäuresequenzen ist (Dixon, 1997). Durch einen Mangel an Testfällen mußte diese Rubrik in den folgenden Jahren ausfallen. Im Sommer 2001 wurde schließlich auf einer von Sandor Vajda und Ilya Vakser organisierten Dockingkonferenz ein neuer vergleichender Wettbewerb für Docking Programme ins Leben gerufen. Für diesen CAPRI (*Critical Assessment of Predicted Interactions*) genannten Vergleichstest wurden im September 2001 erste Einsendungen entgegengenommen, eine detaillierte Auswertung steht jedoch noch aus (Vajda *et al.*, 2002).

1.6.2.3 Verfahren

Ein großer Teil der in der Literatur beschriebenen Dockingverfahren betrachten die Proteine als starre Körper. Diese Verfahren werden als *rigid body* Docking bezeichnet. Kokristallisierte Komplexstrukturen aus ihren Einzelteilen wieder zusammenzusetzen (*bound* Docking) ist mit *rigid body* Docking weitgehend unproblematisch (s. Kapitel 4.1), und bei einer Reihe von Proteinkomplexen sind die beschriebenen Konformationsänderungen so gering, daß sie bei der Vorhersage vernachlässigt werden können (Betts & Sternberg, 1999). In den letzten Jahren werden jedoch zunehmend *unbound* Fälle zum Test eingesetzt, so

daß Dockingprogramme zunehmend auch Konformationsänderungen bei der Komplexbildung berücksichtigen müssen. Die aktuellen Veröffentlichungen zeigen erste Lösungsansätze, sowohl für die Berücksichtigung der Flexibilität von Seitenketten, als auch für die Behandlung von Knicken des Proteinrückgrates zwischen verschiedenen Proteindomänen (Sandak *et al.*, 1998). Eine Beschreibung verschiedener aktueller Ansätze zum Protein-Docking sei der Diskussion vorbehalten (s. Kap. 5.5 auf Seite 77). Eine sehr gute Zusammenfassung zum Thema Protein-Docking findet sich auch bei Halperin *et al.* (2002). In den letzten Jahren haben *rigid body* Dockingverfahren, die auf Fourier-Korrelation basieren stark an Beachtung gewonnen. Die Gründe hierfür liegen vor allem bei der hohen Geschwindigkeit der Methode, die es daher erlaubt, den gesamten 6-dimensionalen Konformationsraum für die Interaktion zweier starrer Strukturen abzusuchen. Darüber hinaus lassen sich mit dieser einfach zu implementierenden Methode sowohl geometrische als auch physiko-chemische Eigenschaften korrelieren. Mit besonderen Techniken, wie der Koeffizientenfilterung, läßt sich auch die Toleranz der Fourier-Korrelation gegenüber Seitenkettenumlagerungen erweitern (3.1.2 auf Seite 38). Auch das in dieser Arbeit vorgestellte Verfahren nutzt die Vorteile dieser Methode als Basistechnologie.

1.6.2.4 Die Arbeit von Katchalski-Katzir

1992 publizierte Ephraim Katchalski-Katzir eine neue Methode zum Protein-Protein Docking auf Basis von Fourier-Korrelation. Er zeigte für einige *bound* Docking Fälle, daß die native Komplexkonformation ein lokales und häufig auch globales Korrelationsmaximum darstellt, sowie daß die Methode geeignet ist, bei hinreichender Dichte des Rotationssamplings dieses Maximum zu finden.

1.6.2.5 Erweiterungen des Fourier Dockings

Anwendungen der Fourier-Korrelation für das Protein-Protein Docking wurden von mehreren Gruppen entwickelt und erweitert (Gabb *et al.*, 1997; Ritchie & Kemp, 2000; Chen & Weng, 2002). Ilya Vakser versuchte durch die Verwendung sehr niedriger Auflösungen das Problem der Seitenkettenflexibilität zu umgehen, indem er nur eine sehr grobe Repräsentation der Proteine einsetzte (Vakser,

1995, 1996). Eine Ergänzung der Methode um Berechnung der Korrelation von Hydrophobizitäten erfolgte ebenfalls (Vakser & Aflalo, 1994). Henry Gabb aus der Gruppe von Michael Sternberg integrierte eine Elektrostatikberechnung der Einzelproteine und schloß Orientierungen mit ungünstiger elektrostatischer Korrelation aus (Gabb *et al.*, 1997).

1.6.2.6 KORDO von Michael Meyer

Das Programm KORDO von Michael Meyer ist eine Fortran Implementation des Algorithmus von Katchalski-Katzir. Neben der Grundfunktionalität zur Korrelationsberechnung besitzt KORDO eine Routine zum gleichmäßigen Sampling des Rotationsraums, eine Verfeinerungsstufe die mit höherer Gitter- und Winkelauflösung in der Umgebung von hohen Korrelationswerten nach Maxima sucht und eine dritte Stufe, die mit einem *simulated annealing* Ansatz diese Lösungen optimiert. Darüber hinaus benutzt KORDO beim Gittermapping verschiedene Gewichtungen in Abhängigkeit von der Besetzung und dem Temperaturfaktor der Atome. Schließlich lassen sich auch vordefinierte Winkellisten anstelle des regelmäßigen Samplings des Rotationsraums abarbeiten.

1.7 Ziel der Arbeit

Ziel dieser Arbeit ist es, die Methoden des Dockings mit Fourier-Korrelationsmethoden und die Bewertung solcher Korrelationen soweit zu verbessern, daß damit aus der Struktur separat kristallisierter Komplexteile die Struktur des nativen Komplexes vorhergesagt werden kann. Dies beinhaltet die geeignete Parametrisierung der Methode, die Analyse verschiedener Fehlerarten und eine entsprechende programmiertechnische Lösung zur Vermeidung solcher Fehler. Das in dieser Arbeit entwickelte Programm CKORDO basiert auf dem Programmcode von Michael Meyer für das Fourier-Docking-Programm KORDO (Meyer *et al.*, 1996).

2 Material und Methoden

2.1 Komplexdaten

Unter den mehr als 11.000 öffentlich zugänglichen Proteinstrukturen erfüllen nur relativ wenige Strukturen alle Bedingungen, die an relevante Testfälle für Dockingalgorithmen zu stellen sind (Berman *et al.*, 2000, 2002):

- Es sollte sich um Heteromultimerkomplexe handeln.
- Alle Bestandteile des Komplexes sollten sich auch als unkomplexierte Einzelstrukturen in der Datenbank befinden.
- Es sollte sich nicht um Modellstrukturen handeln.
- Das Interface sollte vollständig aufgelöst sein.

Die Suche nach geeigneten Proteinen entsprechend obiger Kriterien gestaltet sich in der PDB äußerst schwierig, da die Daten als einzelne Dateien mit einem für die Auswahl unzureichendem Maß an Attributen und ohne Beziehungen zueinander vorliegt. Zwar lassen sich einige Randbedingungen (keine Modellstrukturen, Mindestauflösung etc.) relativ leicht selektieren, Verbindungen zu sequenzidentischen oder -ähnlichen Strukturen sind aber für die Auswahl weitaus wichtiger und stehen innerhalb der PDB-Datenbank nicht zur Verfügung. Positiver Nebeneffekt der geringen Anzahl an Testfällen ist, daß dadurch viele Arbeiten zum Protein-Docking die gleichen Testfälle verwenden und so eine, wenn auch begrenzte, Vergleichbarkeit der Ergebnisse gegeben ist. Eine Liste von 26 bekannten Testfällen, welche die genannten Kriterien erfüllen und daher von verschiedenen Autoren verwendet werden zeigt Tabelle 2.1 auf der nächsten Seite.

Nr.	Komplex	Res.	Typ	Part 1	Res	N.Res	Part 2	Res.	N.Res	Ref.
1	1brb	2.1	PI	1bra	2.2	223	1bpi	1.1	58	[1]
2	1cgi	2.5	PI	1chg	2.5	245	1hpt	2.3	56	[1, 2, 3, 4]
3	2kai	2.5	PI	2pka	2.1	232	1bpi	1.1	58	[1, 3, 4]
4	2kai	2.5	PI	2pka	2.1	232	6pti	1.7	57	[2]
5	2ptc	1.9	PI	3ptn	1.7	223	4pti	1.5	58	[1]
6	2ptc	1.9	PI	2ptn	1.55	229	6pti	1.7	57	[2]
7	2ptc	1.9	PI	2ptn	1.55	229	4pti	1.5	58	[5, 3, 4, 6]
8	1smf	2.2	PI	2ptn	1.55	229	1pi2	2.5	61	[5]
9	2sni	2.1	PI	1sup	1.6	275	2ci2	2.0	65	[2, 3, 4]
10	2sic	1.8	PI	1sup	1.6	275	2ssi	2.6	107	[1, 3, 2]
11	1tec	2.2	PI	1thm	1.37	279	2sec	1.8	62	[5]
12	1cho	1.8	PI	5cha	1.7	245	2ovo	NMR	56	[2, 3, 4, 6]
13	1cho	1.8	PI	5cha	1.7	245	1ovo	1.9	56	[5]
14	1mah	3.2	I	1maa	2.9	545	1fsc	2.0	61	[2, 7]
15	2sec	1.8	I	1scd	2.3	275	1tec	2.2	62	[5]
16	1avw	1.75	I	1ept	1.8	223	1avu	2.3	181	[5]
17	1bth	2.3	EI	4htc	2.3	360	4pti	1.5	58	[5]
18	1brs	2.0	EI	1a2p	1.5	107	1a19	2.8	89	[1, 2, 4]
19	1fss	3.0	EI	2ace	2.5	531	1fsc	2.0	61	[2, 4, 5]
20	1brs	2.0	EI	1bni	2.1	107	1bta	NMR	89	[5]
21	1mlc	2.1	AB	1mlb	2.1	432	1lza	1.6	129	[1, 4, 2, 3]
22	1fdl	2.5	AB	1vfa	1.8	223	1lza	1.6	129	[3, 4]
23	1vfb	1.8	AB	1vfa	1.8	223	1lza	1.6	129	[1, 2, 5]
24	1vfb	1.8	AB	1vfa	1.8	223	1hel	1.7	129	[5]
25	1mda	2.5	OT	2bbk	1.6	489	1aan	2.0	103	[1]
26	2pcc	2.3	OT	1ycc	1.2	108	1ccp	2.2	293	[2, 7, 4]

Tabelle 2.1: Zum Docking verwendete Protein Komplex Systeme mit Angabe von Arbeiten aus der Literatur. Gardiner, *bound/unbound* Fälle. Heifetz02: Verwendung von modellierten Seitenketten für 1avu, 2ace, 1bni.[1] Gardiner *et al.* (2001), [2] Chen & Weng (2002), [3] Gabb *et al.* (1997), [4] Palma *et al.* (2000), [5] Heifetz *et al.* (2002), [6] Lorber *et al.* (2002), [7]Mandell *et al.* (2001)

2.2 Grundlagen

2.2.1 Begriffsbestimmungen

Dockingprogramme besitzen häufig ein mehrstufiges Ablaufschema. Zuerst wird eine große Anzahl von Orientierungen erzeugt; danach werden diese mit unterschiedlichen Verfahren bewertet und gefiltert. Die am Besten bewerteten Strukturen werden dann ggf. abgewandelt und erneut bewertet. Am Schluß steht meistens eine Minimierung, die wegen des hohen Rechenaufwands nur noch an wenigen Kandidaten durchgeführt werden kann. Vorausgesetzt, daß sich bei der Erzeugung der Orientierungen auch immer die richtige d.h. eine der

Kokristallstruktur ähnliche Anordnung befindet, können zwei Arten von Fehlern für die Vorhersage unterschieden werden:

- die richtige Lösung wird als falsch interpretiert und z.B. in einem Filterschritt entfernt. Diese Art Fehler wird in dieser Arbeit als „falsch negativ“ (*false negative*, fn) bezeichnet.
- die falsche Lösung wird als richtig interpretiert, d.h. eine falsche Lösung wird z.B. besser bewertet als die richtige Lösung. Diese Art Fehler wird in dieser Arbeit als „falsch positiv“ (*false positive*, fp) bezeichnet.

Entsprechend werden die korrekten Identifizierungen von richtigen und falschen Orientierungen als „richtig positiv“ (*true positive*, tp) und „richtig negativ“ (*true negative*, tn) bezeichnet. Mehrere Indikatoren ermöglichen eine vergleichende Beurteilung der Qualität von Vorhersagen: Die Spezifität gibt an wie hoch der Anteil negativer Lösungen ist, die korrekt identifiziert wurden:

$$Spez. = \frac{tn}{(tn + fp)}$$

Die Sensitivität dagegen gibt dagegen den Anteil an positiven Lösungen an, die korrekt identifiziert wurden:

$$Sens. = \frac{tp}{(tp + fn)}$$

Spezifizität und Sensitivität sollten im Idealfall beide eins sein. Zur vergleichenden Bewertungen unterschiedlicher Algorithmen werden Spezifität und Sensitivität häufig gegeneinander aufgetragen. Welcher Anteil der positiven Vorhersagen korrekt ist, das heißt wie hoch die Verläßlichkeit einer positiven Klassifizierung ist, wird durch den *positive prediction value* (PPV) bestimmt:

$$PPV = \frac{tp}{(tp + fp)}$$

2.2.2 Eulerwinkel

Die Nomenklatur der für das Rotationssampling (s. Kapitel 2.3.8 auf Seite 28) verwendeten Eulerwinkel orientiert sich an der Arbeit von (Rossmann & Blow,

1962). Sie soll im Folgenden kurz erläutert werden, da ihre Benennung in der Literatur uneinheitlich ist.
Eine Rotation besteht aus:

- einer Rotation von Θ_1 um die X_3-Achse,
- einer Rotation von Θ_2 um die neue Lage der X_1-Achse und
- einer Rotation von Θ_3 um die neue Lage der X_3-Achse.

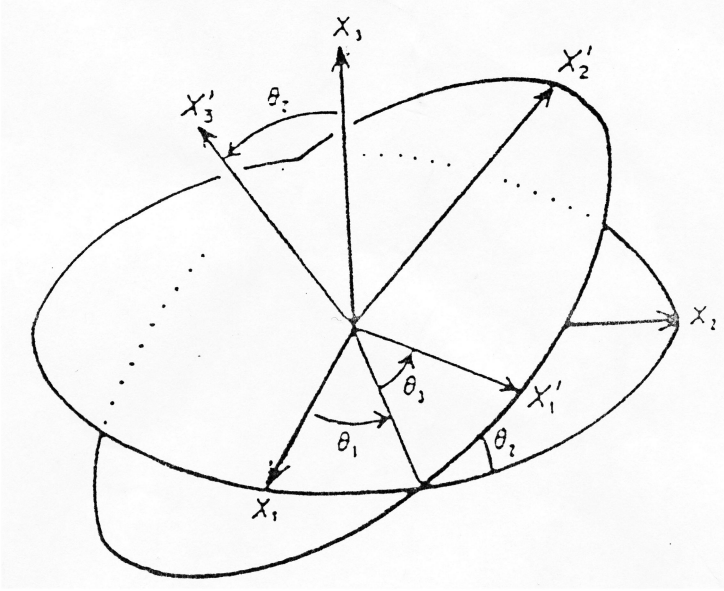


Abbildung 2.1: Definition der Eulerwinkel aus: (Rossmann & Blow, 1962)

Die Winkel bilden ein rechtshändiges System, d.h. positive Winkel verlaufen, in Richtung der Achsen betrachtet, im Uhrzeigersinn. In vielen Publikationen finden sich auch die Bezeichnungen α für Θ_1, β für Θ_2 und γ für Θ_3. In Abbildung 2.1 ist das Verhältnis der ursprünglichen kartesischen Achsen X_1, X_2, X_3 zu den neuen Achsen $X_1^{'}, X_2^{'}, X_3^{'}$ durch eine Rotation um $\Theta_1, \Theta_2, \Theta_3$ dargestellt.

2.3 Methoden des Dockingprogramms CKORDO

2.3.1 Aufbau von CKORDO

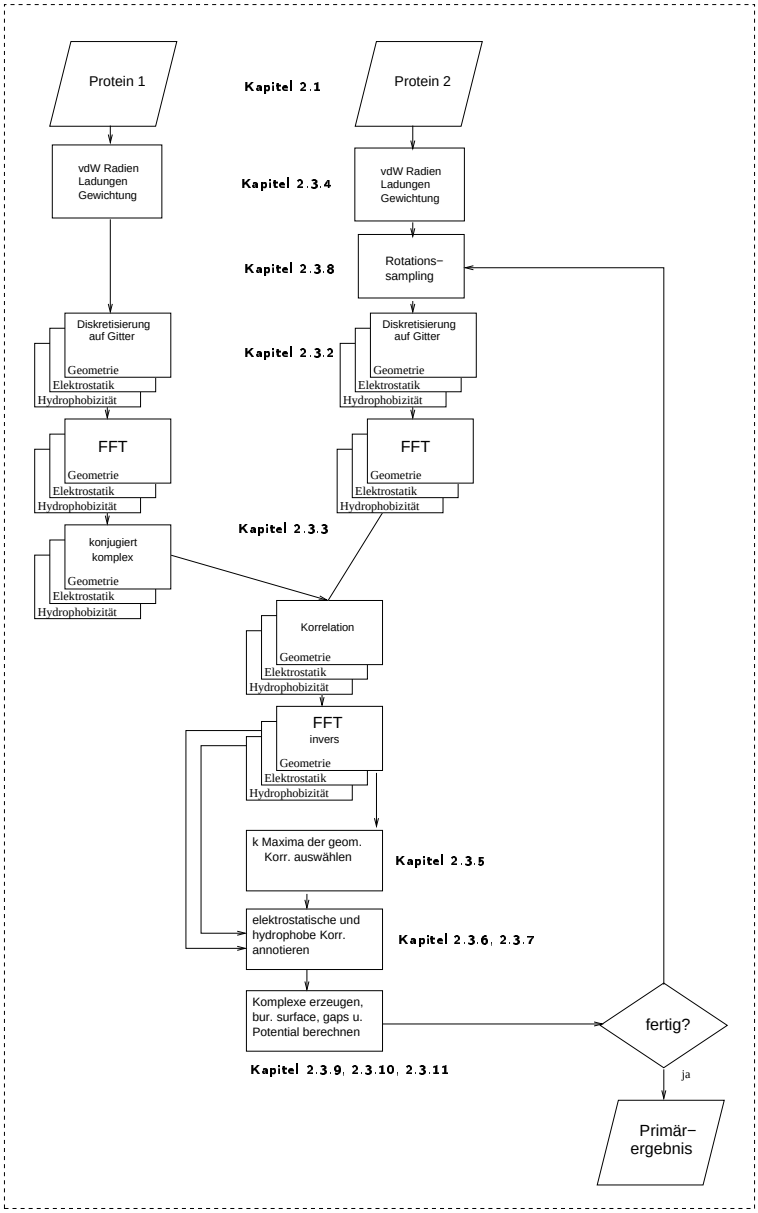


Abbildung 2.2: Schema von CKORDO

In Abbildung 2.2 ist der allgemeine Ablaufplan des entwickelten Docking-Programms CKORDO abgebildet. Die einzelnen Punkte werden in den folgenden Kapiteln entsprechend der Kennzeichnung im Schema erläutert.
Den Atomen der beiden zu dockenden Proteine werden zu Beginn die entsprechenden van der Waals Radien und Ladungen zugeordnet, sowie ein von mehreren Parametern abhängiger Gewichtungswert. Danach erfolgt die Abbildung der beiden Proteine auf je drei dreidimensionale Gitter, welche die Repräsentationen der Geometrie, des elektrostatischen Potentials und des hydrophoben Potentials enthalten. Nach FFT-Transformation der Gitterdarstellungen wird im Fourierraum die Korrelationsberechnung für jedes der drei Gitterpaare vorgenommen und nach Rücktransformation ein Clustering der geometrischen Korrelationen durchgeführt. Die Maxima der ersten k Cluster stellen die vorhergesagten Orientierungen dar. Allen diesen Orientierungen wird die Korrelation der Hydrophobizität und Elektrostatik als Attribut mitgegeben. Weitere Attribute erhalten die Lösungen durch Berechnung der Größe der Interfaceflächen, der Lücken (*gaps*) zwischen diesen Flächen und der Bestimmung eines Pseudoenergiewertes aus einem Atom-Atom Kontaktpotential. Die Berechnungen werden nach Rotation des zweiten Proteins um ein festgelegtes Winkelinkrement solange wiederholt, bis der gesamte Rotationsraum abgesucht ist. Die Ergebnisse werden von CKORDO sortiert nach der geometrischen Korrelation ausgegeben und in weiteren Klassifizierungs- und Filterschritten auf wenige Vorhersagen reduziert.

2.3.2 Komplementarität als diskrete Korrelation

Komplementaritäten zwischen zwei Proteinen lassen sich bei geeigneter Repräsentation der Proteine in einem dreidimensionalen Gitter als diskrete Korrelationsfunktion berechnen, wie im folgenden Kapitel beschrieben werden soll. Die Darstellung der Proteine auf dem Gitter muß sicherstellen, daß nur zwei möglichst große, nah benachbarte Oberflächenteile hohe Korrelationswerte ergeben und eine starke Überlappung der Strukturen zu negativen Korrelationswerten führt. Eine solche Diskretisierung für Proteine wurde in der Arbeit von Katchalski-Katzir erstmals vorgestellt (Katchalski-Katzir *et al.*, 1992) und wird mit Abwandlungen in allen auf Fourier-Korrelation basierenden Docking-

Programmen implementiert (Meyer *et al.*, 1996; Gabb *et al.*, 1997; Chen & Weng, 2002). Die in dieser Arbeit verwendete Methode ist an Gabb *et al.* (1997) angelehnt. Für die Repräsentation zweier Proteine A und B auf einem Gitter werden den Gitterknoten folgende diskrete Werte zugewiesen: Das starre Protein A erhält

$$f_{A_{l,m,n}} = \begin{cases} 1: & \text{Randschicht um das Protein} \\ \rho: & \text{Proteininneres} \\ 0: & \text{außerhalb} \end{cases} \qquad f_{B_{l,m,n}} = \begin{cases} 1: & \text{Proteininneres} \\ 0: & \text{außerhalb} \end{cases}$$

eine dreifache Unterteilung. Für das rotierte Protein B ist die Definition einer Randschicht nicht erforderlich. Zur Abwertung von Orientierungen, die zu einer Überlappung der beiden Proteine führen, wird den Gitterzellen des ersten Proteins ein negativer Wert ρ zugeordnet. In dieser Arbeit wird mit $\rho = -6$ gearbeitet. Aus Laufzeitgründen wird in der Regel das kleinere Protein rotiert.

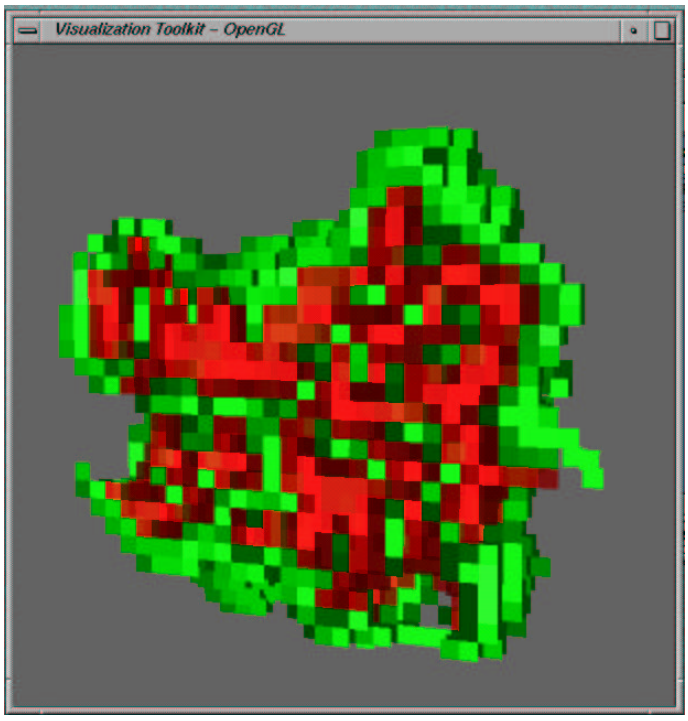


Abbildung 2.3: Schnitt durch ein Protein bei 1.5 Å Gitterauflösung

Abbildung 2.3 zeigt einen Schnitt durch die Gitterrepräsentation eines Proteins bei einer Gitterauflösung von 1.5 Å. Grüne Gitterzellen haben den Wert 1 und rote Gitterzellen den Wert -6 (Abweichende Schattierungen sind Artefakte der Beleuchtungsberechnung für diese 3D-Darstellung).

2.3.3 Fourier-Korrelation

Eine besonders effiziente Berechnung der Komplementarität zwischen den diskreten Funktionen der Proteingeometrie, sowie des elektrostatischen und des hydrophoben Potentials kann durch Transformation in den Fourier-Raum vorgenommen werden. Die Korrelationsfunktion $f_{C_{\alpha,\beta,\gamma}}$ gibt die Korrelation zwischen zwei dreidimensionalen diskreten Funktionen an und ist definiert als:

$$f_{C_{\alpha,\beta,\gamma}} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} f_{A_{l,m,n}} \times f_{B_{l+\alpha,m+\beta,n+\gamma}} \tag{2.1}$$

wobei N die Größe des kubischen Gitters in x-, y- und z-Richtung mit den Indices l, m, n und α, β und γ die jeweiligen Komponenten des Translationsvektors von B in Gitterzelleneinheiten darstellen.
Da die Korrelationsberechnung für die geometrische Komplementarität die Multiplikation der Besetzungswerte beinhaltet, erzeugt eine Überlappung des Proteins B mit der Randschicht von Protein A positive Summanden, die Überlappung von Protein B mit Protein A negative Summanden und Orientierungen ohne Überlappungen Nullwerte. Zur Berechnung dieser Korrelationen auf direktem Weg sind N^3 Multiplikationen für jeden Translationsvektor α, β, γ notwendig d.h. insgesamt N^6. Werden die diskreten Funktionen A und B in den Fourierraum transformiert, so reduziert sich die Anzahl der Multiplikationen von N^6 auf N^3. Der Preis hierfür ist der Aufwand für die Fourier-Transformation des Gitters, auf dem die Proteine repräsentiert werden. Bei diskreten Funktionen läßt sich mit Hilfe der sogenannten *Fast Fourier Transformation* (FFT) (Cooley & Tukey, 1965) die Komplexität für diese Operation auf $N^3 log(N^3)$ mit N als Länge des Gitters in Gitterpunkten begrenzen, während sonst N^6 Operationen erforderlich wären. Die Korrelation im Fourierraum berechnet sich nach Gleichung 2.2 bis 2.5.

$$F_A = FFT(f_A) \tag{2.2}$$

$$F_B = FFT(f_B) \tag{2.3}$$

$$F_C = (F_A^*)(F_B) \tag{2.4}$$

$$f_C = FFT^{-1}(F_C) \tag{2.5}$$

Insgesamt werden also 3 diskrete Fourier-Transformationen durchgeführt, zwei Hintransformationen für die Proteinrepräsentationen (notiert als FFT für *F*ast *F*ourier *T*ransformation) und die Rücktransformation der Korrelationsfunktion (notiert als FFT^{-1}). Die Korrelationsfunktion f_C ist im Fourier-Raum definiert als das Produkt des konjugiert Komplexen der Fourier-Darstellung von f_A (notiert als F_A^*) mit der Fourier-Darstellung von f_B. Sie benötigt im Gegensatz zur direkten Korrelationsberechnung nur N^3 Multiplikationen, so daß der Gesamtalgorithmus eine Laufzeitkomplexität der Größenordnung O($N^3 log(N^3)$) statt O(N^6) hat.

2.3.4 Gewichtung

Die Gitterrepräsentation der Proteine wird anhand des Temperaturfaktors gewichtet. Dieser ist ein Maß für die Beweglichkeit einzelner Atome und damit auch für die Flexibilität von Seitenketten. Dabei erhalten alle Atome mit Temperaturfaktoren, die mehr als drei Standardabweichungen über dem Durchschnitt liegen, Null als Gewichtungswert zugeordnet, so daß diese Atome auch bei Penetration des anderen Proteins keine negativen Terme bei der Korrelationsberechnung verursachen. Andere Gewichtungen, z.B. für bestimmte Aminosäuretypen, wenn deren Seitenketten stark solvensexponiert sind, befinden sich in Vorbereitung. Hierbei können auch rationale Zahlen als Gewichtungsfaktoren verwendet werden.

2.3.5 Clustern der Korrelationen in der Translationsebene

Alle Korrelationswerte für eine Rotation befinden sich nach der Korrelationsberechnung und der Rücktransformation des Gitters aus dem Fourier-Raum in den einzelnen Gitterzellen. Dabei entspricht die Lage der Gitterzelle der Translation des rotierten Proteins. Für jede Rotation ergibt sich also ein Gitter mit Korrelationswerten. Nicht in jedem Fall entspricht die höchste Korrelation innerhalb einer Rotation auch der nativen Orientierung. Manchmal findet sich die richtige Anordnung erst als zweit- oder drittgrößter Korrelationswert. Um diese Orientierungen nicht zu verlieren, müssen evtl. mehrere Translationen für die genauere Untersuchung aufgehoben werden. Da die Gitterzelle mit der nächstniedrigeren Korrelation sich häufig unmittelbar neben dem Korrelationsmaximum befindet, also eine Flanke des Hauptpeaks darstellt, wurde ein einfaches Clusterverfahren entwickelt, daß in der Lage ist, lokale Maxima von Flanken eines Korrelationsmaximums zu unterscheiden. Bei der Entwicklung mußte berücksichtigt werden, daß das Gitter sich randlos im Raum fortsetzt, d.h. daß in einem Gitter A der Größe $N \times N \times N$ in allen Achsrichtungen $k \in x, y, z$ die Achskoordinate k_{N+1} der Achskoordinate k_1 entspricht. Für das Clustering werden alle Korrelationswerte oberhalb eines Schwellwerts absteigend sortiert. Der erste Wert bildet den Kern eines Clusters. Für alle nachfolgenden Werte wird überprüft ob sich die Gitterzelle näher als ein Grenzwert x zu einem bereits geclusterten Wert befindet. Ist das nicht der Fall, so bildet die Gitterzelle den Kern eines neuen Clusters, wenn die festgelegte Anzahl an auszugebenden Werten noch nicht erreicht ist. Die Zellen mit den Korrelationsmaxima der Cluster stellen die Translationen dar, die für jede Rotation ausgegeben werden.

2.3.6 Berechnung der Elektrostatik

Für die Berechnung der elektrostatischen Komplementarität über eine diskrete Korrelation muß das elektrostatische Potential geeignet auf eine Gitterstruktur abgebildet werden. Zur Repräsentation der Elektrostatik wurden zwei verschiedene Ansätze untersucht: Beim Ansatz von Gabb und Sternberg wird die Ladung jedes Atoms des rotierten Proteins auf die acht Gitterpunkte aufgeteilt, die zum jeweiligen Atommittelpunkt den kürzesten Abstand haben (Edmonds *et al.*, 1984)

und diese Ladungsverteilung mit dem Coulomb-Feld des nicht rotierten Proteins korreliert (Gabb *et al.*, 1997). Alternativ dazu wurde ein Coulomb-Feld für jedes der beiden Proteine berechnet und die beiden Felder miteinander korreliert. Um den Rechenaufwand zu begrenzen ist die Reichweite des Coulomb-Feldes für die Berechnung der Elektrostatik im Programm auf 7 Å voreingestellt. Der Verlauf des entsprechenden Pseudo-Coulombschen Potentials ist in Abbildung 2.4 dargestellt.

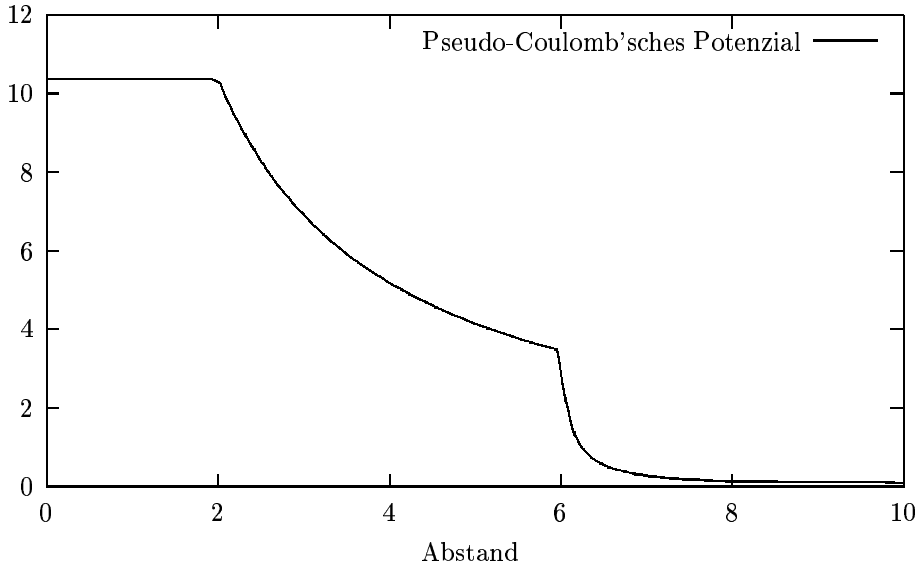


Abbildung 2.4: Pseudo-Coulmbsches Potential. Aufgetragen sind das Potential (y-Achse) gegen den Abstand (x-Achse) in [Å]

Die Berechnung der Elektrostatik mit zwei korrelierten Coulomb-Feldern ähnelt dem Ansatz von Heifetz *et al.*. Allerdings verwenden diese Autoren eine Poisson-Boltzmann Näherung zur Berechnung der elektrostatischen Felder und korrelieren nur die Vorzeichen der Feldstärken (Heifetz *et al.*, 2002).

2.3.7 Berechnung der Hydrophobizität

Für die Berechnung der atombasierten Hydrophobizität wurden die Absolutwerte der Atompartialladungen aus dem Kraftfeldprogramm AMBER 5 gewählt. Es wurden die unterschiedlichen Ladungszustände des Histidins und explizite

Wasserstoffatome vernachlässigt. Bei der Hydrophobizitätsberechnung wurden die Absolutwerte der Atomladungen beider Proteine mit Hilfe eines Coulomb-Feldes auf die Gitterpunkte verteilt. Da hydrophobe Wechselwirkungen kurzreichweitig sind, wurde die Reichweite des Feldes auf 4 Å voreingestellt. Dieser Maximalabstand wird als besonders günstig für hydrophobe und insbesondere π-π-Wechselwirkungen angenommen (Hunter *et al.*, 1991).

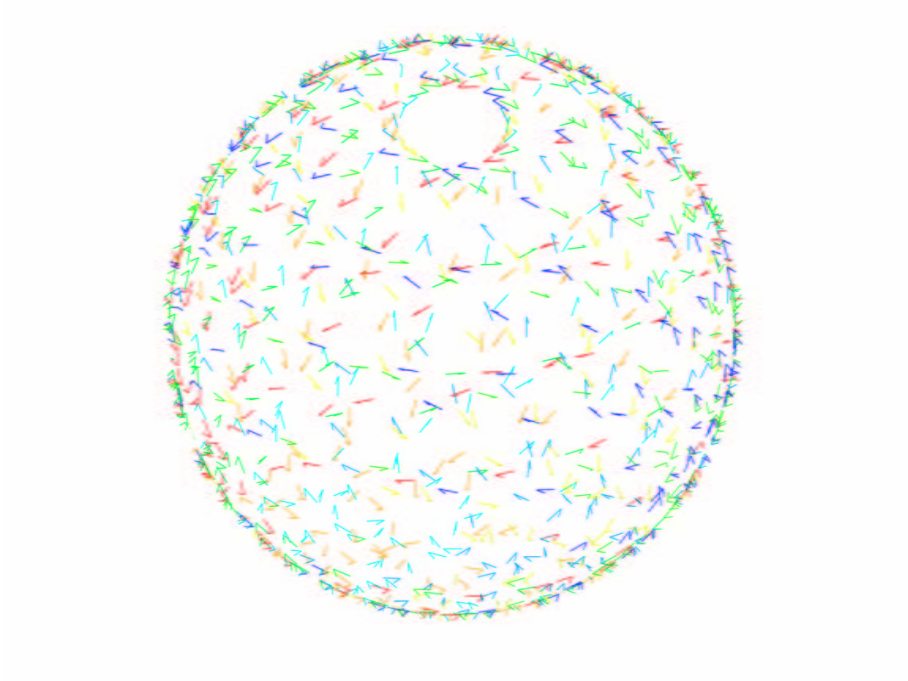
2.3.8 Rotationssampling

Die beschriebenen Korrelationsfunktionen für Geometrie, Elektrostatik und Hydrophobizität werden für den gesamten 6-dimensionalen Konformationsraum bestimmt. Während die Korrelationsberechnung für alle Translationen durch die Fourier-Korrelation des Gitters in einem Schritt erfolgt, müssen die drei Rotationswinkel durch ein geeignetes Sampling abgesucht werden. Der Rotationsraum in Eulerwinkeln $(\Theta_1, \Theta_2, \Theta_3)$ beträgt 360^o x 180^o x 360^o. Bei unabhängiger Unterteilung dieser Winkel wären viele Orientierungen redundant wie in Abbildung 2.5, (b) an der ungleichmäßigen Verteilung insbesondere in Äquatornähe zu sehen ist. Ein gleichmäßigeres Sampling des Winkelraums kann durch eine Rotationssuche mit einem von Θ_2 abhängigen Winkelinkrement erreicht werden (Lattman, 1972). Dabei ergeben sich für die beiden anderen Eulerwinkel:

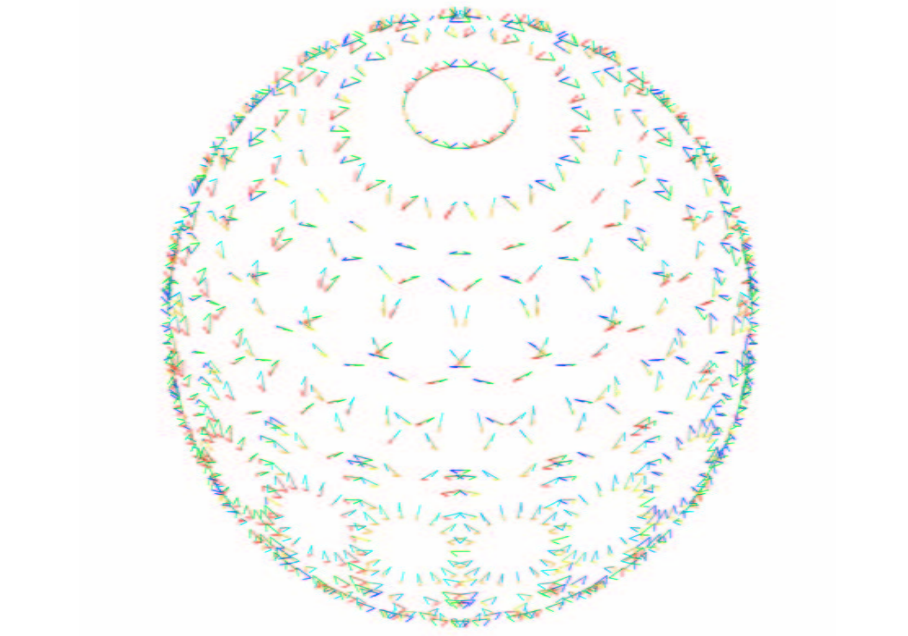
$$\Delta\Theta_1 = 0.5\frac{\Delta\Theta_2}{\cos(\Theta_2/2)} + 0.5\frac{\Delta\Theta_2}{\sin(\Theta_2/2)} \tag{2.6}$$

$$\Delta\Theta_3 = 0.5\frac{\Delta\Theta_2}{\cos(\Theta_2/2)} - 0.5\frac{\Delta\Theta_2}{\sin(\Theta_2/2)} \tag{2.7}$$

In Abbildung 2.5 auf der nächsten Seite ist zu erkennen, daß die Samplingdichte der Winkel, mit Ausnahme des Pols selbst, bei diesem regelmäßigem Sampling überall gleich groß ist (s. Abb. 2.5 auf der nächsten Seite, (a)).



(a) Sampling durch Eulerwinkel nach Lattman



(b) Sampling durch Kugelwinkel

Abbildung 2.5: Sampling des Rotationsraums: Farbe der Pfeile $= \Theta_3$

Bei Verwendung des Lattman Samplings muß keine Filterung redundanter Winkelkombinationen vorgenommen werden, wie von einer Reihe anderer Autoren beschrieben (Gabb *et al.*, 1997). Die Anzahl der Winkelkombinationen zum Erreichen eines bestimmten Raumwinkelabstands liegt um etwa ein Drittel niedriger als die von Gabb angegebene Zahl nach Filterung redundanter Winkelkombinationen (s. Tabelle 2.2).

	Anzahl der Winkelkombinationen		
Winkeldistanz	Lattman	unabhängig	gefiltert[1]
22.5°	1307	2048	
20.0°	1857	2916	
18.0°	2578	4000	
15.0°	4401	6912	6385
12.0°	8616	13500	
10.0°	14877	23328	22105

Tabelle 2.2: Anzahl der benötigten Samples bei gegebenem Winkelinkrement.
[1] Gabb *et al.* (1997)

2.3.9 Verdeckte Oberfläche (*buried surface*)

Für die Maxima der geometrischen Korrelation ist es möglich aus den Orientierungen die Kandidatenkomplexe zu erzeugen um geometrische und physikochemische Parameter zur Bewertung der Komplexstrukturen zu berechnen. Die folgenden Abschnitte beschreiben die Berechnung dieser Größen.
Wenn zwei Proteine miteinander in Kontakt treten, wird ein Teil der Oberflächen der Zugänglichkeit für Wassermoleküle entzogen. Diese Fläche wird *buried surface* genannt und errechnet sich als Summe der lösungsmittelzugänglichen Oberflächen der Komplexteile abzüglich der lösungsmittelzugänglichen Oberfläche des Komplexes. In Verbindung mit dem Lückenvolumen kann so ein Bild von der Packungsdichte des Interfaces gewonnen werden. Die Berechnung der einzelnen lösungsmittelzugänglichen Oberflächen (*Accessible Surface Area* (ASA)) wurde mit dem Programm DSSP (Kabsch & Sander, 1983) durchgeführt.

2.3.10 Interfacelücken (*gaps*)

Die Packungsdichte von Interfaces ist ein wichtiges Maß für die Güte der geometrischen Komplementarität. Natürliche Interfaces haben eine Dichte, die nur unwesentlich geringer ist als im Inneren eines Proteins. Da es schwierig ist, die lokale Dichte im Proteininterface zu berechnen, wird stattdessen ein Programm zur näherungsweisen Berechnung des Lückenvolumens verwendet. Das Programm SURFNET (Laskowski, 1995) bestimmt das Lückenvolumen, indem die Zwischenräume im Interface mit Kugeln verschiedenen Durchmessers welche beide Proteine berühren, gefüllt werden.

2.3.10.1 Lückenweite

Der Abstand der Interfaceflächen der beiden Proteine, in dieser Arbeit Lückenweite genannt, kann zur Bewertung von Protein-Komplexen verwendet werden. Unter dem Begriff *gap index* wird bei Janet Thorntons Protein Interaction Server (Thornton, 1998) der Quotient aus Lückenvolumen und der *buried surface*, also der Summe der Interfaceflächen beider Proteine verstanden. Für den in dieser Arbeit verwendeten Begriff der Lückenweite wird als Divisor stattdessen das arithmetische Mittel aus den beiden Interfaceflächen verwendet, so daß die hier ermittelten Werte etwa doppelt so groß sind, wie die entsprechenden *gap index*-Werte des Protein Interaction Servers. Dadurch läßt sich die Lückenweite anschaulich als durchschnittlicher Abstand zwischen den beiden Proteinoberflächen der *buried surface* interpretieren.

2.3.11 Statistisches Potential

Zur Bewertung der Wahrscheinlichkeit, ob die Kontakte im Interface des Kandidatenkomplexes eine native Zusammensetzung haben, wird ein statistisches Potential eingesetzt, welches auf der abstandsabhängigen Bewertung von Atom-Atom Kontakten basiert (Grimm, 2002). Dazu werden die Atome in vierzig verschiedene Typen und die interatomaren Abstände bis zu einer Entfernung $r_{max} = 8$ Å in 23 Intervalle eingeteilt. Zur Glättung der Intervallzuordnungen wird eine Trapezfunktion eingesetzt. Die so gewonnenen Häufigkeitsverteilun-

gen werden mittels einer empirischen Funktion (Profil 1 aus Grimm (2002)) in Pseudoenergien umgerechnet.

2.3.12 RMSD Berechnung

Das gebräuchlichste Maß zum Vergleich dreidimensionaler Strukturen ist der RMSD. Er ist definiert als die Wurzel der Summe der Abweichungsquadrate der einander entsprechenden Atome. Leider sind viele verschiedene Abwandlungen dieses Maßes in Gebrauch, die sich bezüglich der berücksichtigten Atome und der Wahl der Referenzstruktur unterscheiden. Sofern nicht besonders gekennzeichnet, ist in dieser Arbeit der RMSD zwischen den Cα-Atomen der nativen Interfaces einer Struktur zur Referenzstruktur gemeint. Dabei ist die Referenzstruktur nicht die Kristallstruktur des Komplexes, sondern die sich aus der optimalen Superpositionierung der Einzelproteine ergebende Position des rotierten Proteins. Als Interfaceatome werden diejenigen Atome definiert, die maximal 10 Å von irgendeinem Atom des Komplexpartners entfernt sind. Diese Definition wird in vielen Publikationen über Protein-Protein Docking verwendet und daher zur besseren Vergleichbarkeit mit den Ergebnissen anderer Autoren auch in dieser Arbeit eingesetzt. Die Superpositionierung erfolgte mit den entsprechenden Routinen des Molecular Modelling Programms BRAGI (Schomburg & Reichelt, 1988; Lessel & Schomburg, 1994) und mit dem Programm CE (Shindyalov & Bourne, 1998).

2.3.13 Klassifizierung durch *Support Vector Machines* (SVM)

Die aus der *algorithmischen Lerntheorie* stammenden *Support Vector Machines* (SVM) stellen einen methodischen Rahmen dar um einen Klassifizierungsmechanismus aus annotierten Daten abzuleiten (sog. *supervised learning*). SVMs haben in der Bioinformatik in der letzten Zeit zunehmende Beachtung bei Klassifizierungsproblemen gefunden, u.a. bei der Auswertung von Expressionsexperimenten. Ein Anwendungsfall dabei ist die Klassifikation von Gewebeproben anhand der Aktivitäten einer Reihe von Genen darauf, ob es sich um Tumorgewebe handelt oder nicht. Diese Aufgabe ist sehr anspruchsvoll, da Expressionsdaten in

hohem Grade fehlerbehaftet sind und die Aktivität der meisten getesteten Gene irrelevant in Hinblick auf die Tumordetektion ist.

Dem Erfolg des SVM-Ansatzes liegen zwei wesentliche Prinzipien zu Grunde (Boser *et al.*, 1992). Beim idealisierten Klassifizierungsproblem bei zwei gegebenen, durch eine Hyperebene, also linear, trennbaren Klassen von Daten, ist es einsichtig, daß es keine eindeutige trennende Hyperebene gibt. Das erste Kernprinzip sagt nun, daß diejenige trennende Hyperebene gewählt werden soll, die den sogenannten *Margin*, also den Abstand von der Hyperebene zu den nächsten Datenpunkten beider Klassen, diese werden als *Supportvektoren* bezeichnet, maximiert. Dieses Optimierungsproblem ist ein quadratisches Programm, welches effizient *global* gelöst werden kann.

Im Allgemeinen sind Klassen nun nicht linear zu trennen. Das zweite Kernprinzip sieht vor, die Daten geeignet in einen höherdimensionalen Raum, den *Feature-Raum*, abzubilden, in dem die beiden Klassen wieder mit einer Hyperebene getrennt werden können. Praktikabel wird dieser Ansatz nun dadurch, daß diese Abbildung nicht tatsächlich vorgenommen werden muß. Für die Berechnung der trennenden Hyperebene wird nur das Skalarprodukt zweier Vektoren im Feature-Raum benötigt, welches durch die sogenannten *Kernel-Funktionen* — ausgehend direkt von Eingabedaten — gegeben wird. Erweiterungen von SVMs für nicht trennbare Klassen oder Mehr-Klassen-Probleme sind in der Literatur beschrieben.

SVMs stellen im Gegensatz zu neuronalen Netzen keine vollständige Black Box dar, optimieren im Gegensatz zu diesen die Diskriminationsfähigkeit und sind darüberhinaus global statt nur lokal zu optimieren. SVMs werden in dieser Arbeit eingesetzt, um native Orientierungen von Protein-Interfaces von nicht-nativen zu unterscheiden. Als Eingangsdaten werden die unterschiedlichen Korrelationen, die geometrischen Daten und das statistische Potential verwendet. Das benutzte Programm LIBSVM beinhaltet eine Reihe verschiedener Teilprogramme zur Skalierung und Klassifizierung von Daten (Chang & Lin, 2001). Drei verschiedene Kernelfunktionen (Linearer-, Exponentieller- und RBF-Kernel) werden untersucht.

2.3.14 Vorbereitung von Proteinen

Einige Seitenkettenorientierungen führen beim Docking zu sterischen Hinderungen und schlechten geometrischen Korrelationswerten. Daher wurden Experimente durchgeführt, welche die Seitenketten vor dem Docking in eine der Orientierung im Komplex angenäherte Position zu bringen suchen. Zwei unterschiedliche Programme wurden verwendet:

- SCWRL (Bower *et al.*, 1997)

  SCWRL wählt aus einer Rotamerbibliothek unterstützt von einer Potentialfunktion Seitenkettenorientierungen aus.

- XPLOR/CNS (Brünger *et al.*, 1987)

  Mit CNS wurde versucht in einer hydrophoben Umgebung eine Minimierung durchzuführen. Dabei wurde das Proteinrückgrat unverändert gelassen.

Beide Ansätze zeigten keine Verbesserung der Resultate (s. Kap. 4.2.1 und 3.5) und wurden daher bislang nicht weiter verfolgt.

2.3.15 Analyse von *induced fit* Effekten

Um die Veränderung der Proteinstrukturen durch den Dockingvorgang zu analysieren wurden Superpositionen von Komplexstrukturen mit einzeln kristallisierten Proteinen durchgeführt. Um dies für eine große Anzahl von Kombinationen automatisch durchführen zu können, wurde die Superpositionierungsroutine aus dem Molecular Modelling Programm BRAGI, dessen Programmcode vorlag, ausgebaut und zu einem eigenständig lauffähigen Programm umgeschrieben. Nach Berechnung der verschiedenen Superpositionen wurden die Atom-Atom-Abstände, sowie die Änderung der lösungsmittelzugänglichen Oberfläche zwischen den Atomen von Komplex und Einzelprotein bestimmt.

2.3.16 Intermolekulare Atomkontakte

Zur Analyse von Interfaces ist es häufig notwendig, Abstände zwischen den Atomen der beiden verschiedenen Bindungspartner festzustellen. Zu diesem

Zweck wurde ein einfaches C-Programm mit Namen CCONTACT entwickelt, daß die intermolekularen Atompaare unterhalb eines einstellbaren Schwellwertes ausgibt.

2.3.17 Wasserstoffbrücken, Salzbrücken

Wasserstoffbrücken und Salzbrücken wurden durch DSSP (Kabsch & Sander, 1983) und HBPLUS (McDonald & Thornton, 1994) bestimmt.

3 Entwicklung von CKORDO

CKORDO besteht aus etwa 5000 Zeilen Programmcode in der Programmiersprache C++. Es ist lauffähig unter Unix und kann von der Kommandozeile aus vielfältig parametriert werden. Die Laufzeiten betragen ohne Berechnung der Lückenweite zwischen 40 und 400 Minuten. In den folgenden Abschnitten werden die einzelnen in CKORDO implementierten Techniken detailliert beschrieben.

3.1 Parameterauswahl

Die Fourier-Korrelation besitzt viele Parameter, deren optimale Einstellung darin besteht, für möglichst unterschiedliche Komplexe gute Ergebnisse zu erzielen. Tests können aufgrund der Anzahl der möglichen Kombinationen nur eine Annäherung an dieses Optimum erreichen. In der Regel wird daher die Änderung für einen oder zwei Parameter gleichzeitig untersucht.

3.1.1 Winkelauflösung

Um die Rechenzeit in Grenzen zu halten wird normalerweise mit einem Winkelinkrement von 20^o entsprechend 1867 nichtredundanten Rotationen gearbeitet. In den meisten Fällen konnte dabei zumindest eine Lösung mit großer Ähnlichkeit zur nativen Orientierung erhalten werden. Um mehrere solcher Lösungen zu bekommen, wird für einen umfassenden Test des Dockings von *unbound*-Fällen mit einer Winkelauflösung von 12^o gearbeitet, was zu insgesamt 8616 nichtredundanten Rotationen führt (s.a. Kap. 2.2 auf Seite 30). Katchalski-Katzir gibt an, daß die Winkeltoleranz der diskreten Fourier-Korrelation von der verwendeten Randschichtdicke abhängt (Katchalski-Katzir *et al.*, 1992). Die Optimierung von Randschichtdicke und Gitterauflösung stand deshalb im Mittelpunkt der ersten Versuche zur Parametrierung der geometrischen Fourier-Korrelation.

3.1.2 FFT-Auflösung (Koeffizientenfilterung)

Die Genauigkeit mit der Funktionen durch Fourier-Reihen approximiert werden, hängt von der Anzahl der Terme in dieser Reihe ab. Werden nur wenige Terme der Fourier-Entwicklung berücksichtigt, wird eine gewisse Unschärfe der Darstellung der Proteinstruktur im Fourier-Raum erreicht. Insbesondere beim *unbound*-Docking lassen sich die Effekte nutzen um Abwertungen bei der Berechnung der geometrischen Korrelation aufgrund einander durchdringender Seitenketten abzumildern. Die Vergleiche wurden mit einer Reihe von künstlichen Trypsin/Trypsin-Inhibitor Komplexen durchgeführt. Diese Strukturen sind künstlich erzeugte Übergänge von der Proteinstruktur im kokristallisierten Komplex 1tgs zu den Einzelstrukturen 1tgn bzw. 5pti. Dazu wurden einzelne Reste aus dem Interface der Kokristallstruktur in die Einzelstrukturen modelliert (für Details s. Kap. 3.5 auf Seite 47). Die Koeffizientenfilterung führt zu einer geringeren Auflösung der Strukturen. Daher wird das Maß der Filterung hier in Ångstrøm angegeben.

Test	Geänderte Seitenketten	Rang (geometrische Korrelation)		
Nr.	1tgn-5pti	Ohne Filterung	Filterung (5.0 Å)	Differenz
80	39,192-15,17	80	67	-13
82	39,192-15,39	71	3	-68
87	39,192-15,17,39	85	3	-82
95	39,195-15,17	260	267	+7
97	39,195-15,39	175	15	-160
102	39,195-15,17,39	389	31	-358
140	192,195-15,17	37	122	+85
142	192,195-15,39	26	4	-22
147	192,195-15,17,39	65	15	-50
185	39,192,195-15,17	27	65	+38
187	39,192,195-15,39	7	4	-3
189	39,192,195-17,39	182	19	-163
192	39,192,195-15,17,39	31	11	-20
225	39,99,192,195-15,17,18,39	1	1	0

Tabelle 3.1: Einfluß der FFT-Koeffizientenfilterung beim *unbound* Docking, Winkelinkrement 20^o

In der Mehrzahl der Fälle wird durch die FFT-Koeffizientenfilterung eine deutliche Verbesserung des Rangs erreicht (s. Tabelle 3.1). Dies ist in Übereinstimmung mit den Untersuchungen von Vakser zum *Low Resolution* Docking (Vakser, 1996;

Vakser *et al.*, 1999). Alle folgenden *unbound* Dockings wurden daher mit FFT-Koeffizientenfilterung entsprechend einer Auflösung von 5.0 Å durchgeführt.

3.1.3 Gitterauflösung und Randschichtdicke

Komplex	R.-dicke	Gitterauflösung		
		1.0 Å	1.2 Å	1.5 Å
4cha-B/2ovo	1.0 Å	1520		
	1.2 Å	1453	95	
	1.4 Å		29	
	1.5 Å	245	42	615
	1.6 Å	223	45	
	1.8 Å		352	
	2.0 Å		615	380
1bra/1bpi	1.1 Å	–		
	1.2 Å	9	9	
	1.3 Å	8		
	1.4 Å	4	6	
	1.5 Å	3	4	4
	1.6 Å	3	3	4
	1.7 Å	3	2	2
	1.8 Å	3	2	2
	1.9 Å	4	4	2
	2.0 Å	5	4	2
	2.1 Å			7
	2.2 Å			10
	2.5 Å			13

Tabelle 3.2: Rang der "pseudonativen" Orientierungen (n=1867): 4cha-B/2ovo (α-Chymotrypsin/Ovomucoid Third Domain) und 1bra/1bpi (Trypsin/pankreatischer Inhibitor), *unbound* Docking, FFT-Auflösung 5.0 Å (s. Kapitel 3.1.2 auf der vorherigen Seite), – = nicht gefunden

Tabelle 3.2 zeigt den Rang der nativen Orientierung für das *unbound* Docking des Komplexes aus α-Chymotrypsin und seinem Inhibitor bei unterschiedlichen Einstellungen für die Dicke der Randschicht um das nicht rotierte Protein und für die Gitterauflösung. Die Ergebnisse zeigen, daß eine höhere Gitterauflösung nicht notwendigerweise auch zu einer Verbesserung der Bewertung des nativen Komplexes gegenüber den anderen Orientierungen führt. Vielmehr zeigt sich, daß das Verhältnis aus Randschichtdicke und Gitterauflösung bei 1 bis 1.2 ein Optimum hat. Da sich bei einer Verdopplung der Auflösung die Anzahl der Zellen verachtfacht und die Rechenzeit fast linear mit der Anzahl der Gitterzellen steigt, ist eine höhere Gitterauflösung als 1.2 Å für das komplette Absuchen des

Rotationsraumes nicht sinnvoll. Die untersuchten *unbound* Testfälle scheinen zudem unterschiedlich empfindlich gegenüber den Einstellungen zu sein. Für den Chymotrypsin/Ovomucoid-Komplex (4cha/2ovo) führt sowohl eine zu hohe (1.0 Å) als auch eine zu niedrige (1.5 Å) Auflösung dazu, daß die richtige Orientierung nicht mehr innerhalb der 500 Korrelationsmaxima liegt. Die beste getestete Parameter-Kombination liegt hier bei einer Randschichtdicke von 1.4 Å und einer Gitterauflösung von 1.2 Å. Für den Testfall eines Trypsin/Trypsin-Inhibitor Komplexes ergaben sich bei 1.5 Å Gitterauflösung für die Randschichtdicke in einem Bereich von 1.7 bis 2.0 Å keine erkennbaren Unterschiede. Eine Reihe ähnlicher Tests führte für alle nachfolgend beschriebenen *unbound* Dockings zu einer Parameterwahl von 1.5 Å für die Gitterkonstante und 1.75 Å für die Randschichtdicke.

3.1.4 Werte für Proteininneres

Der Strafwert für eine geometrische Überlappung hängt von dem Wert ab, den das Gitter für proteinbesetzte Zellen im Inneren des nichtrotierten Proteins aufweist. Während Gabb *et al.* den Wert -15.0 von Katchalski-Katzir et al. übernommen haben, ist der Parameter im Code von KORDO auf -6.0 eingestellt. Mehrere Experimente mit Werten zwischen -3.0 und -10.0 wurden durchgeführt und zeigten, daß mit dem Wert -6.0 die besten Ergebnisse erzielt werden konnten.

3.2 Cluster im Rotationsraum

Eine naheliegende Hypothese bei einem systematischen Sampling ist es, daß es in der Nähe der nativen Orientierung zu einer Häufung von Orientierungen kommt, die eine hohe Korrelation aufweisen und deren Interfaces von der Geometrie her einem Komplexinterface ähneln. Während das Clustering von Lösungen mit der gleichen Rotation bereits in Kapitel 2.3.5 auf Seite 26 besprochen wurde, wurden Visualisierungen vorgenommen um zu überprüfen ob es eine solche Häufung identifizierbarer Lösungen gibt. Für die Visualisierung wurde die gleiche Kugeldarstellung verwendet, wie in der Visualisierung des Rotationssamplings in Kapitel 2.3.8. In dieser Darstellung werden die Pfeile zusätzlich

entsprechend translatiert und der Winkel Θ_3 ist nur noch an der Orientierung der asymmetrischen Pfeilspitze relativ zur Pfeilachse ablesbar, da die Farbe für den entsprechenden Eigenschaftswert (Korrelation bzw. Potential) verwendet wird.

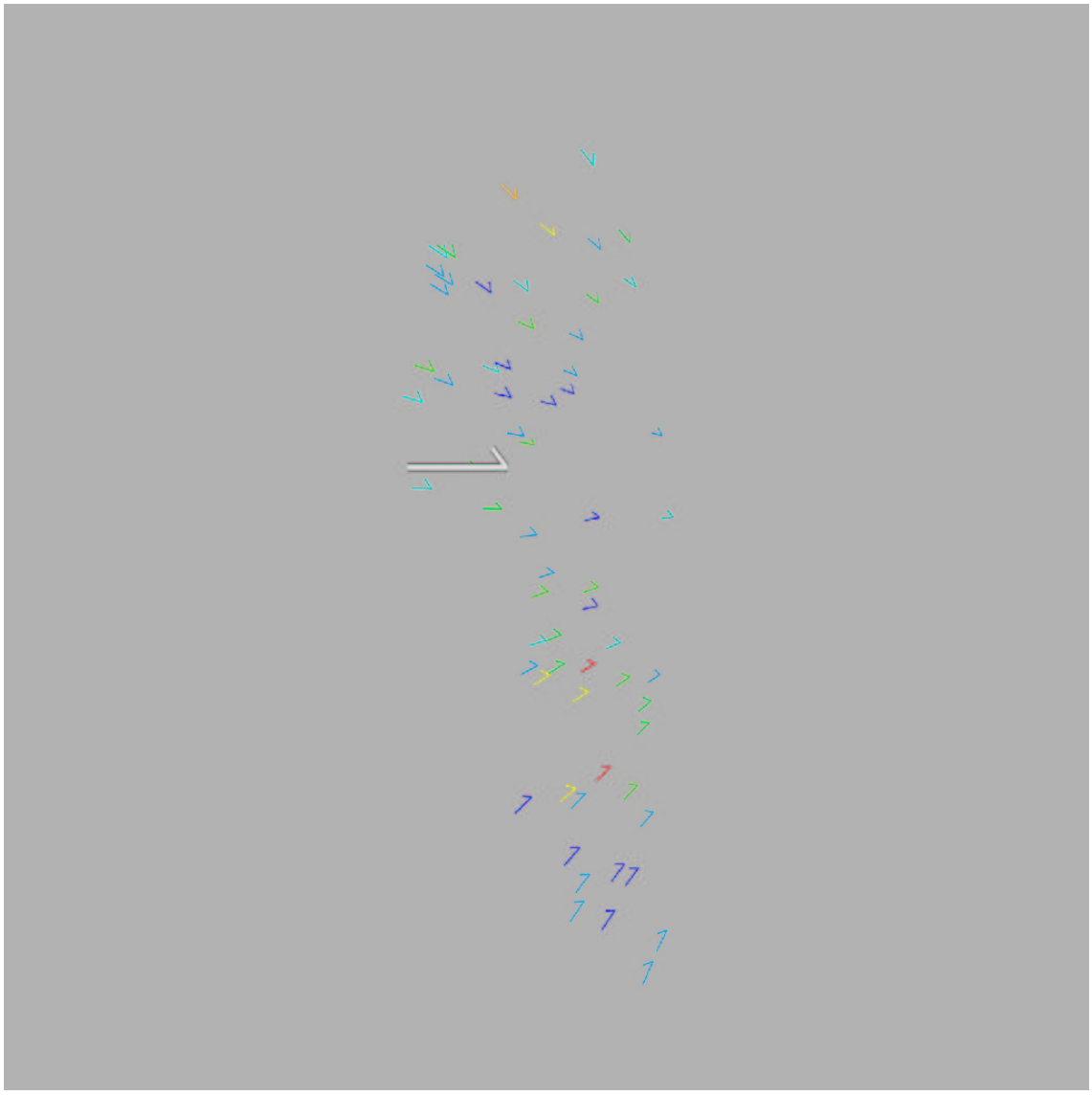


Abbildung 3.1: Geometrische Korrelation für 2kai

Die Farbskala verläuft für die geometrische Korrelation von rot (geringe Korrelation) nach blau (hohe Korrelation) und für das Potential auf Grund des

negativen Vorzeichens von blau (ungünstiger Potentialwert) nach rot (günstiger Potentialwert). Der große Pfeil gibt jeweils die native Orientierung an. Da die Darstellung von mehr als 40000 Ergebnissen etwas unübersichtlich ist, sollen hier nur die Ergebnisse gezeigt werden, die sich unmittelbar in der Nähe der nativen Orientierung befinden, d.h. einen RMSD von <3.5 Å zur optimal superpositionierten Struktur haben.

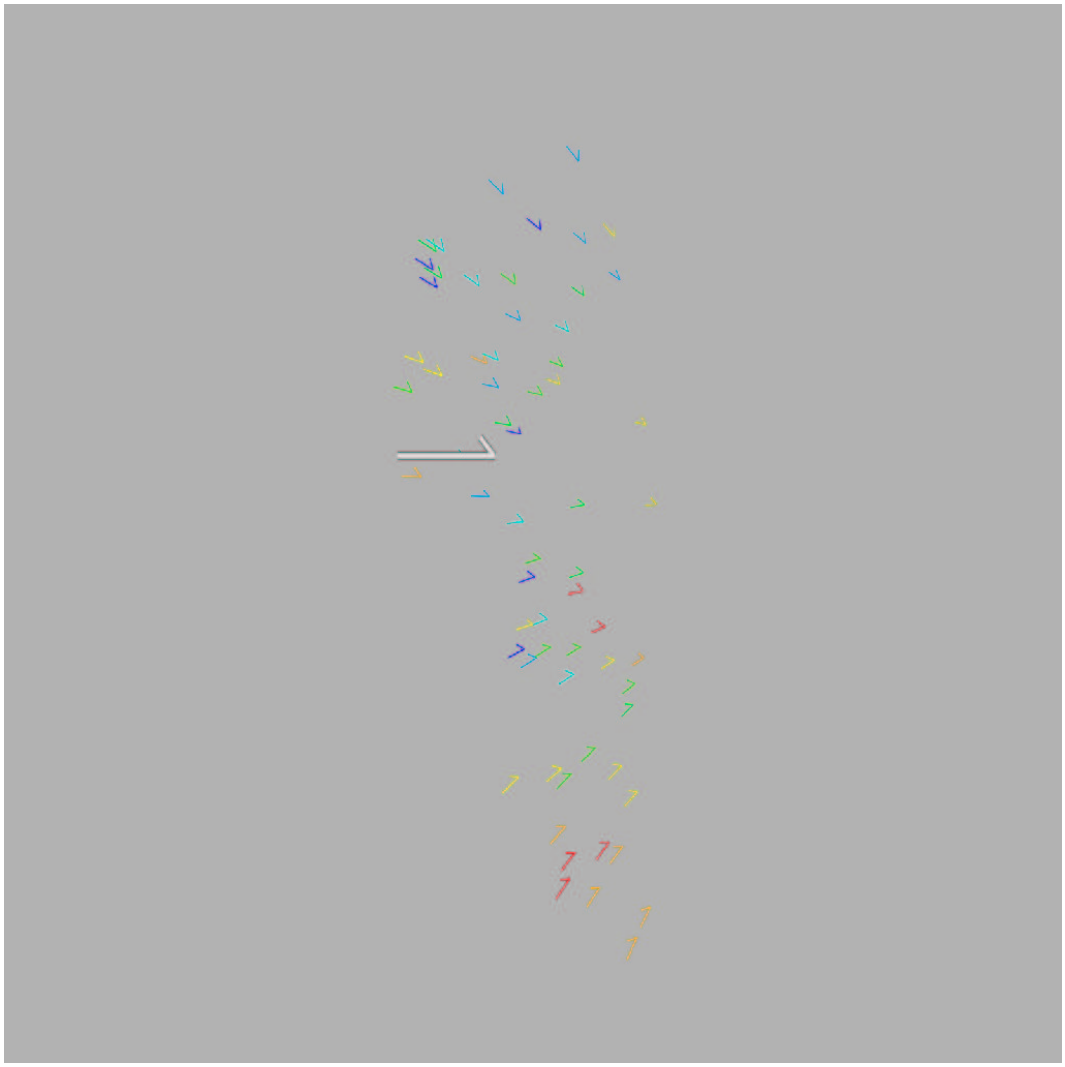


Abbildung 3.2: Statistische Präferenz für 2kai

Abbildung 3.1 zeigt eine entsprechende Auftragung der geometrischen Korrela-

tion für das System Kallikrein/Inhibitor (2kai). Deutlich sind zwei Komplexe mit sehr hoher geometrischer Korrelation zu erkennen (dunkelblau im Bild), davon einer sehr nah an der korrekten Orientierung und einer am unteren Bildrand. Es ist jedoch auch ein deutlicher ungünstiger Cluster zu erkennen. Die Grafik macht deutlich, daß bei dieser Lösungsdichte ein Clustering von Lösungen möglich ist: blaue Pfeile sind gegenüber orangen und roten Pfeilen in einem Abstand von 3.5 Å klar in der Mehrheit, zeigen also eine überwiegend positive geometrische Korrelation.

Die Ergebnisse für denselben Komplex bei Betrachtung des statistischen Potentials gestalten sich ähnlich. Deutlich ist in Abbildung 3.2 der Cluster am unteren Bildrand zu sehen, der vom Kontaktpotential positiv bewertet wird. Wenn auch etwas weniger deutlich als bei der geometrischen Korrelation, so zeigt sich doch auch hier eine Mehrzahl von positiven Bewertungen (orange, rot) gegenüber negativen Bewertungen (blau, hellblau).

3.3 Hydrophobe Korrelation

Im Rahmen eines Praktikums wurden aminosäurebasierte Hydrophobizitätsskalen von Kyte-Doolittle, Cornette und Boyko getestet (s. Anhang A.2), um eine größere Toleranz gegenüber der Seitenkettenflexibilität zu erreichen. Die Ergebnisse zeigten jedoch, daß sich mit diesen Skalen für Protein-Protein Interfaces keine brauchbaren Korrelationswerte erzielen lassen. Als zweiter Ansatz wurde daher als Hydrophobizitätswert jedem Atom der Absolutbetrag der Ladungswerte aus den AMBER 5 Kraftfeldparametern zugeordnet und mit einer bei 4.0 Å abgeschnittenen Coulomb-Funktion auf dem Gitter verteilt.

Auffällig ist, daß bei praktisch allen Testfällen die hydrophobe Korrelation der nativ-ähnlichen Orientierungen nur im schwach positiven Bereich liegt (s. Abbildung 3.3 für ein Beispiel (5cha) und die Abbildungen A.1 bis A.12 im Anhang). Zu hohe Korrelationswerte weisen damit auf nicht-native Orientierungen hin, die offensichtlich größere hydrophobe Oberflächen beinhalten. Besonders auffällig ist dies bei den Antikörperkomplexen (s. Abbildung 3.4). Eine große Anzahl falsch positiver Lösungen findet sich hier in einer flach-konkaven Tasche des An-

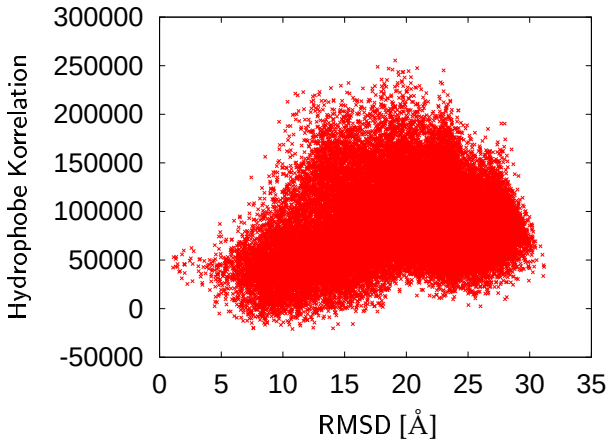


Abbildung 3.3: Hydrophobe Korrelation beim Docking des α-Chymotrypsin-Ovomucoid-Komplexes 1cho aus den Koordinaten der Einzelstrukturen 5cha und 2ovo

tikörpers. Durch die große Oberfläche wird hier sowohl eine hohe geometrische als auch eine sehr hohe hydrophobe Korrelation erzielt. Da extrem hohe hydrophobe Korrelationswerte im Gegensatz zu hohen geometrischen Korrelationen offensichtlich auf falsch positive Lösungen hinweisen, lassen sich diese Werte gut als Filterkriterium verwenden.

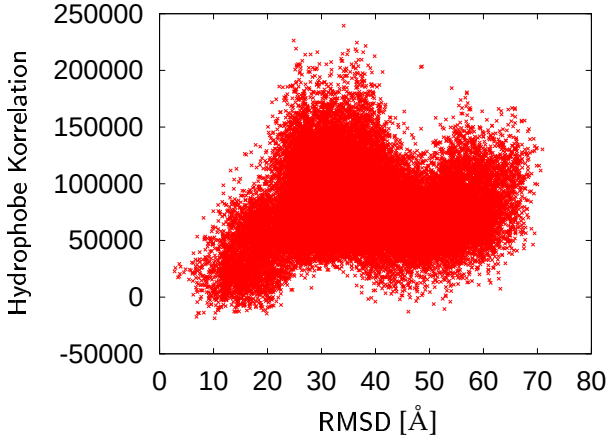
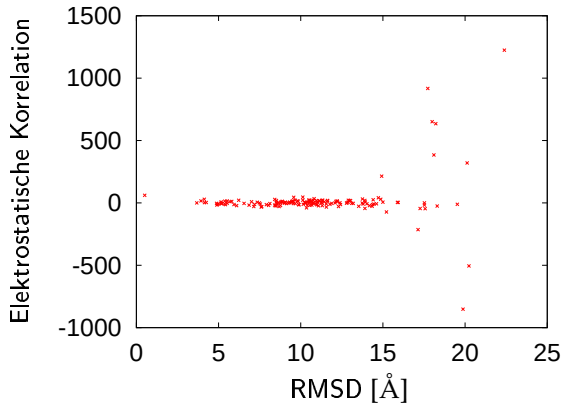


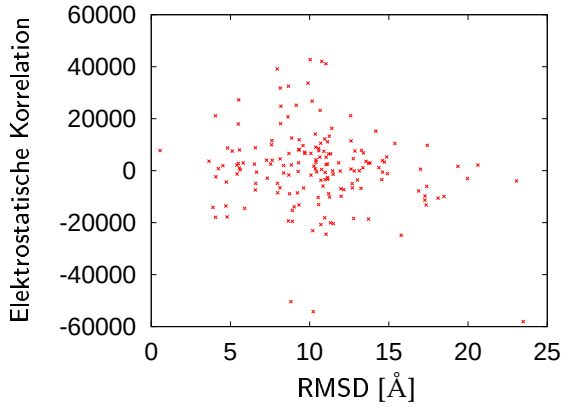
Abbildung 3.4: Hydrophobe Korrelation beim Docking des Antikörper-Lysozym-Komplexes 1mlc aus den Koordinaten der Einzelstrukturen 1mlb und 1lza

3.4 Elektrostatische Korrelation

Die elektrostatische Korrelation nach dem Verfahren von Sternberg (Gabb *et al.*, 1997; Edmonds *et al.*, 1984) ergab keine Unterscheidungskriterien, die sich zum Filtern oder zur direkten Bewertung eignen (s. Abbildung 3.5(a)).



(a) Ansatz von Gabb *et al.* (1997)



(b) Korrelation zweier Coulomb-Felder

Abbildung 3.5: Elektrostatische Korrelation beim Docking des Trypsin/Trypsin-Inhibitor Komplexes 1brb aus den Koordinaten der Einzelstrukturen 1bra und 1bpi

Als zweite Möglichkeit wurde die Korrelation zwischen den Coulomb-Feldern untersucht. Ein ähnlicher Ansatz wird von Heifetz *et al.* beschrieben (Heifetz *et al.*, 2002). Allerdings berechnen diese Autoren die elektrostatischen Felder durch Lösung der linearisierten Poisson-Boltzmann Gleichung mit dem Programm DELPHI. Auf Grund der großen Dynamik des Wertebereichs verwenden Heifetz *et al.* für ihre Dockingexperimente nur die Vorzeichen der Felder. Die gleichen Orientierungen wie im vorigen Test (s. Abbildung 3.5(a)) zeigen mit diesem Ansatz zur Ladungsverteilung eine größere Spreizung der elektrostatischen Korrelationswerte, wie in Abbildung 3.5(b) zu sehen ist.

Als weiterer Test wurde der Acetylcholinesterase-Fasciculin II-Komplex (1fss) gewählt, dessen unabhängig voneinander kristallisierten Einzelteile in den PDB-Einträgen 2ace und 1fsc vorhanden sind. Sowohl in der Arbeit von Heifetz *et al.*, als auch in der Arbeit von Chen & Weng zeigten sich Verbesserungen in der Vorhersage dieses Komplexes durch Einbeziehung der Elektrostatik in die Korrelationsberechnung (Heifetz *et al.*, 2002; Chen & Weng, 2002), wie sie auch Abbildung 3.6 zeigt.

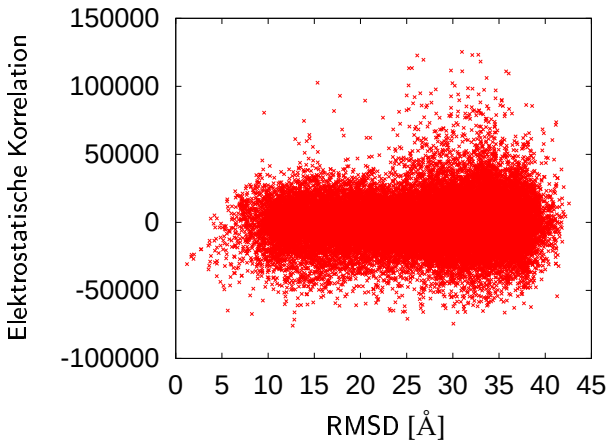


Abbildung 3.6: Elektrostatische Korrelation beim Docking des Acetylcholinesterase-Fasciculin II-Komplexes 1fss aus den Koordinaten der Einzelstrukturen 2ace und 1fsc

Es ist zu erkennen, daß die Werte unterhalb eines RMSD von 4 Å gegen einen Wert von ca. -30000 konvergieren und daß unterhalb eines RMSD von 10 Å die

Absolutwerte im Durchschnitt geringer werden. Insbesondere die Anzahl großer positiver Korrelationswerte nimmt ab. Da die elektrostatische Korrelationen so berechnet worden sind, daß negative Werte eine gute Korrelation darstellen, ist die Elektrostatik für den Komplex 1fss leicht vorteilhaft. Es ist jedoch erkennbar, daß sich die Werte alleine nicht zur Diskriminierung nativer Lösungen eignen. Da die Verjüngung der Verteilung auch bei allen anderen Testkomplexen beobachtet werden konnte (s. Kapitel 4.3), eignet sich die Elektrostatik aber zum nachträglichen ausfiltern vieler nicht-nativer Lösungen. Diese Verwendung ist auch bei Gabb *et al.* beschrieben (Gabb *et al.*, 1997).

3.5 Behandlung der langen Aminosäuren

Schon seit langem ist aus vielen Dockingexperimenten mit Serin-Proteasen bekannt, daß Seitenketten von oberflächennahen Argininen und Lysinen in Lösung problematisch sind. Eine Möglichkeit mit diesem Problem umzugehen, ist es die Aminosäureseitenketten abzuschneiden. Gabb *et al.* führen diesen Schritt für Reste mit hoher Solvenszugänglichkeit durch. Da die betreffenden Aminosäuren geladen sind, wird der Einfluß auf die Elektrostatikberechnung dadurch verringert, daß die entsprechende Ladung auf das CG-Atom der Aminosäure verlagert wird. Um den Einfluss verschiedener Behandlungsarten langer, exponierter Seitenketten zu untersuchen, wurden Dockingexperimente mit der Trypsinstruktur 1tgn und der Trypsininhibitorstruktur 5pti durchgeführt. Eine Analyse der entsprechenden Komplexstruktur 2tgp führte zu folgenden für die Ausbildung des Interface relevanten Aminosäuren (s. Tabelle 3.3 und 3.4):

AS	Kurzb.	ASA 2tgp	ASA 1tgn
Tyr 39	Y39	74.4%	77.2%
Leu 99	L99	31.2%	23.4%
Gln 192	Q192	87.2%	125.5%
Ser 195	S195	26.9%	32.0%

Tabelle 3.3: Relevante Aminosäure (AS) und solvenszugängliche Fläche (ASA) im Trypsin (1tgn)

AS	Kurzb.	ASA 2tgp	ASA 5pti
Lys 15	K15	128.6%	117.0%
Arg 17	R17	90.9%	96.9%
Ile 18	I18	46.4%	42.0%
Arg 39	R39	90.9%	92.9%

Tabelle 3.4: Relevante Aminosäuren (AS) und solvenszugängliche Fläche (ASA) im Trypsin Inhibitor (5pti)

Die Aminosäuren wurden in einem ersten Experiment durch Alanin ersetzt, was einem Abschneiden der Seitenkette hinter dem CB-Atom gleichkommt. In einem zweiten Experiment wurden die Solvensstrukturen der Aminosäure durch die Komplexstrukturen ersetzt, also für die jeweilige Aminosäure die richtige Orientierung aus dem Komplex eingesetzt. Im dritten Ansatz wurde mit dem Programm SCWRL eine Konformation über das dort implementierte Seitenkettenpotential gewählt (s. Tabelle 3.5).

Versuch	1tgn Mutation	5pti Mutation	Rang		
			Alanin	korrekt	*SCWRL*
1	Y39/Q192	K15/R17	–	67	294
2	Y39/Q192	K15/R39	–	3	177
3	Y39/Q192	R17/R39	1657	65	5
4	Y39/Q192	K15/R17/R39	–	3	118
5	Y39/S195	K15/R17	1095	267	
6	Y39/S195	K15/R39	878	15	
7	Y39/S195	R17/R39	708	–	
8	Y39/S195	K15/R17/R39	983	31	
9	Q192/S195	K15/R17	503	122	
10	Q192/S195	K15/R39	247	4	
11	Q192/S195	R17/R39	403	27	
12	Q192/S195	K15/R17/R39	876	15	
13	Y39/Q192/S195	K15/R17	1374	65	
14	Y39/Q192/S195	K15/R39	933	4	
15	Y39/Q192/S195	R17/R39	1140	19	
16	Y39/Q192/S195	K15/R17/R39	–	11	
17	Y39/L99/Q192/S195	K15/R17/I18/R39	–	1	

Tabelle 3.5: Einfluß der Konformation der langen Aminosäuren auf die geometrische Korrelation im Interface

3.6 Support Vector Machines zur Vorhersage

Zur Wahl der entsprechenden Kernelfunktion und der Parameter wurde zunächst ein Teildatensatz ausgewertet. Dieser bestand aus sechs Dockingergebnissen mit insgesamt 29406 Dockingkandidaten. Von diesen hatten 180 einen RMSD < 4.0 Å und wurden als positiv (+1) klassifiziert. Die restlichen 29226 hatten einen RMSD >= 4.0 Å und wurden als negativ (-1) klassifiziert. Jeder Dockingkandidat stellt einen Eigenschaftenvektor aus vier Elementen dar: geometrische, elektrostatische und hydrophobe Korrelation, sowie das abstandsabhängige Atom-Atom Potential. Die Werte wurden auf Standardabweichungen normiert, indem die Abweichung vom Mittelwert berechnet wurde:

$$Z_j = k_j - m_j / s_j$$

mit m_j = Mittelwert aller k_j, k_j = Messwert und s_j = Standardabweichung aller k_j. Je nach Wahl der Kernelfunktion und der Gewichtung der beiden Ausgangsklassen wurden sehr unterschiedliche Ergebnisse erhalten. Insgesamt standen in dem verwendeten Programm LIBSVM vier verschiedene Kernelfunktionen zur Verfügung. Da der sigmoidale Kernel auf dem verwendeten Datensatz nicht konvergierte, konnten nur drei Kernelfunktionen getestet werden: Linearer Kernel, Exponentieller Kernel und RBF-Kernel (=Radial Basis Function).
Mit einer Sensitivität von 51.7% bei nur ca. 10 falsch Positiven je richtig positiver Klassifizierung war die Leistung der Radialen Basis Funktion als Kernelfunktion bei einer Gewichtung von 4000 für positive Testfälle und 120 für negative Testfälle am höchsten (s. Tabelle 3.6). Diese Parameter bildeten die Grundlage für das Training von SVMs auf größeren Datensätzen, wurden jedoch in der Art der Skalierung und der Gewichtung im weiteren Verlauf der Arbeit verändert. Um für das Training eine klarere Trennung zwischen positiven und negativen Lösungen zu erreichen, wurden als positive Lösungen nur solche mit einem RMSD von weniger als 3.5 Å und als negative Lösungen solche mit einem RMSD von mindestens 10 Å eingesetzt. Das gleiche gilt auch für die Testdatensätze. Ein Problem ergibt sich aufgrund der insgesamt sehr geringen Anzahl an positiven Lösungen. Zwar lagen für alle Testfälle die Ergebnisse zweier unabhängiger Dockingläufe vor, um ein Übertraining des Klassifizierers zu verhindern, wurden

Gew. +1/-1	nSV	nBSV	tp	fp	Spez.	Sens.	Prob. Rang
Linearer Kernel							
100/1	16123	16092	65	4100	0.043	0.361	63
100/3	6208	6126	0	0			
Exponentieller Kernel							
100/1	16893	16882	47	1728	0.101	0.261	37
100/3	6556	5896	3	15	10	0.017	
RBF Kernel							
100/1	11994	11865	132	4553	0.038	0.733	34
100/3	5733	5532	47	551	0.301	0.261	12
300/3	11154	11004	136	4393	0.040	0.756	32
500/5	10807	10619	145	4366	0.040	0.806	30
600/18	5399	5167	67	710	0.232	0.372	11
1000/30	5298	5044	74	719	0.227	0.411	10
2000/60	5161	4893	80	794	0.206	0.444	10
4000/120	5035	4751	93	902	0.181	0.517	10
4000/240	2992	2637	47	160	0.870	0.261	3
6000/240	4002	3713	68	394	0.390	0.378	6

Tabelle 3.6: Testläufe mit verschiedenen SVM-Kerneln und Parametern. Gew. = Gewichtung, nSV = Anzahl Support Vektoren, nBSV = Anzahl beschränkter Support Vektoren, tp = Anzahl richtig positiver Lösungen, fp = Anzahl falsch positiver Lösungen, Spez. = Spezifität, Sens. = Sensitivität, Prob. Rang = Wahrscheinlicher Rang

jedoch nur die Ergebnisse zweier Serinprotease-Testfälle (2ptc, 2kai) eingesetzt. Der Trainingsdatensatz hatte 78011 Einzelergebnisse mit jeweils 7 Eigenschaften. Zusätzlich zu den Korrelationen und dem Potential sind in diesen Eigenschaftenvektoren die *buried surface*, das Lückenvolumen und die Lückenweite enthalten. Die 117 positiven Lösungen wurden beim Training mit einem Faktor von 4000, die 77894 negativen Lösungen mit einem Faktor von 100 gewichtet. Der trainierte Klassifizierer besteht aus 4678 Support Vektoren, von denen 4665 d.h. 99.7% beschränkt sind. Dies weist darauf hin, daß keine Übertrainierung stattgefunden hat.

4 Ergebnisse

4.1 *Bound*-Docking

4.1.1 Reproduktion der Meyer-Ergebnisse

Als erster Test wurden einige Dockingexperimente, die M. Meyer mit dem Programm KORDO durchgeführt hatte mit CKORDO wiederholt. Da die mit den KORDO-Experimenten verwendeten Parameter nicht zur Verfügung standen, wurden die Standardparameter von KORDO (Randschichtdicke 2.0 Å, Gitterauflösung 1.5 Å) verwendet (s.a Meyer *et al.* (1996)).

PDB-Code	Anzahl Atome		Auflsg.	CKORDO				KORDO
	Protein 1	Protein 2		Rang	Korrelation	Korr. Rang 2	Performance	Rang
2mhb	1178	1113	2.0 Å	1	372	283	131 %	1
4phv2	760	760	2.1 Å	1	679	283	240 %	1
2sec	1923	522	1.8 Å	1	336	329	102 %	1
1tec	2007	522	2.2 Å	27	263	(344)	76 %	1
4tpi	1644	456	2.2 Å	1	370	293	126 %	1
2ptc	1629	454	1.9 Å	1	329	282	117 %	1
1tgs	1652	416	1.8 Å	1	336	271	124 %	1
1cho	1750	400	1.8 Å	1	331	297	111 %	1
4sgb	1310	380	2.1 Å	3	306	(316)	97 %	1
4cpa	2438	290	2.5 Å	13	237	(273)	87 %	1
2cpk	2665	75	2.7 Å	1	286	196	146 %	1
2igf	3378	58	2.8 Å	1	226	193	117 %	1
1fdl	3308	1001	2.5 Å	74	263	(344)	76 %	33/3
3hfm	3295	1001	3.0 Å	1	207	194	107 %	>75/1

Tabelle 4.1: *Bound* Docking Ergebnisse. Korrelationen, Randschichtdicke 2.0 Å, Gitterkonstante 1.5 Å, 1867 Anordnungen

Die native Orientierung hat bei diesen *bound* Docking Versuchen in fast allen Fällen die höchste geometrische Korrelation (s. 4.1). Ausnahmen sind die Enzym-Inhibitor-Komplexe von Thermitase (1tec) und C-terminaler Peptidase (4cpa). Bei feinerer Gitterauflösung hat die native Orientierung auch bei 1tec ein globales Maximum in der geometrischen Korrelation (s. Kap. 3.1.3). Der Antikörper-Komplex 1fdl weist sowohl bei der Auswertung mit CKORDO, als auch mit KORDO eine Reihe von falsch positiven Orientierungen auf. Dies ändert sich auch

nicht bei anderer Gitterauflösung oder Randschichtdicke (s. Kap. 3.1.3) und wird deshalb in Kap. 4.1.3 näher untersucht.

PDB-Code	R.-dicke	Gitter-auflsg.	CKORDO				KORDO
			Rang	Korrelation	Korr. Rang 2	Performance	Rang
1scd-E/1cse-I	2.0 Å	1.5 Å	1	346	316	109 %	3
2gch-E/1acb-I	2.0 Å	1.5 Å	1	1393	373	373 %	7
2gch-E/1cho-I	2.0 Å	1.5 Å	1	1395	357	391 %	9
3est/2est-I	2.0 Å	1.5 Å	145	149	(186)	80 %	1
3est/2est-I	1.2 Å	1.2 Å	350	105	(140)	75 %	(1)
5cpa/4cpa-I	2.0 Å	1.5 Å	318	195	(265)	74 %	-
5cpa/4cpa-I	1.2 Å	1.2 Å	22	99	(122)	81 %	-
3cpa/4cpa-I	2.0 Å	1.5 Å	13	247	(273)	90 %	69
3cpa/4cpa-I	1.5 Å	1.2 Å	44	154	(192)	80 %	(69)
3cpa/4cpa-I	1.2 Å	1.2 Å	26	104	(122)	85 %	(69)

Tabelle 4.2: Vergleich von CKORDO und den KORDO-Ergebnissen von Meyer *et al.*: *cross bound* und *bound-unbound*-Docking, R.-dicke = Randschichtdicke

Die Werte in Tabelle 4.2 zeigen, daß mit CKORDO auch die Ergebnisse für *cross-bound* und *bound-unbound*-Dockings von Meyer *et al.* weitgehend reproduziert werden können. Eine mögliche Erklärung für das schlechtere Abschneiden der Elastase (3est/2est-I) könnte darin liegen, daß die KORDO-Versuche evtl. mit den Wassermolekülen in den PDB-Daten durchgeführt wurden. 3est enthält 129 Wassermoleküle, die zu einem scharfen Maximum der geometrischen Korrelationsfunktion in unmittelbarer Nähe der nativen Korrelation führen würden, da sie den Inhibitor aussparen und so ein größeres Interface erzeugen. Mit Wassermolekülen ist bei CKORDO die native Orientierung für 3est/2est-I auf Rang eins.

4.1.2 RAS/RAF-Komplex 1gua

Einige Proteine bereiten auch beim *bound*-Docking bei der Vorhersage besondere Schwierigkeiten. Der Komplex aus den beiden Protoonkogenen RAS und RAF hat ein stark verzahntes Interface (s. Abb. 4.1 auf der nächsten Seite). Dies führt dazu, daß die Korrelationsfunktion ein steiles Gefälle hat. Die native Orientierung stellt zwar das globale Maximum der geometrischen Korrelation dar, aber bereits bei einer Rotation um $\Theta_2 = 4°$ sinkt die geometrische Korrelation auf 77%

des Maximalwerts. Bei einer Rotation um 4° bezüglich aller drei Eulerwinkel ist der Translationsvektor, welcher der nativen Orientierung am nächsten kommt, schon nicht mehr das Korrelationsmaximum. Bei zufallsrotierter Ausgangsorientierung wurde die native Anordnung bei einem Winkelinkrement von 20° nicht gefunden.

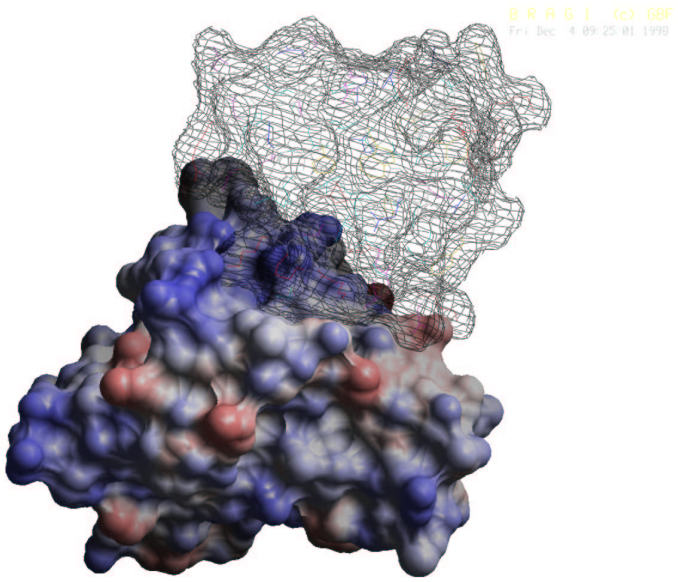


Abbildung 4.1: RAS/RAF Komplex 1gua, dargestellt mit elektrostatischem Potential in Gitterdarstellung

4.1.3 Immunglobulinkomplexe

Die Immunoglobuline stellen eine eigene Problemklasse für das Protein-Protein Docking dar. Bereits in den Arbeiten von Michael Meyer zeigten sich hier auch beim *bound*-Docking Schwierigkeiten (Meyer *et al.* (1996) und Tabelle 4.1). Die Vorhersagbarkeit verschiedener Immunglobulin-Lysozym-Komplexe ist jedoch höchst unterschiedlich. Um diese Differenzen zu analysieren, wurden die PDB-Daten der beiden Antikörper-Lysozym-Komplexe 3hfm und 1fdl untersucht. Während der Komplex 3hfm ein klares geometrisches Korrelationsmaximum in der nativen Orientierung zeigt, finden sich beim Docken der Einzelteile von 1fdl

eine Reihe falsch positiver Lösungen. Charakteristisch für diese falsch positiven Komplexstrukturen ist die große Kontaktfläche, die sich durch die Lage des konvexen Lysozyms in der Konkavität zwischen VL und VH-Domäne ergibt. Die native Kontaktfläche zwischen Immunglobulin und dem Antigen Lysozym ist dagegen weitaus kleiner und erzielt daher niedrigere geometrische Korrelationswerte.

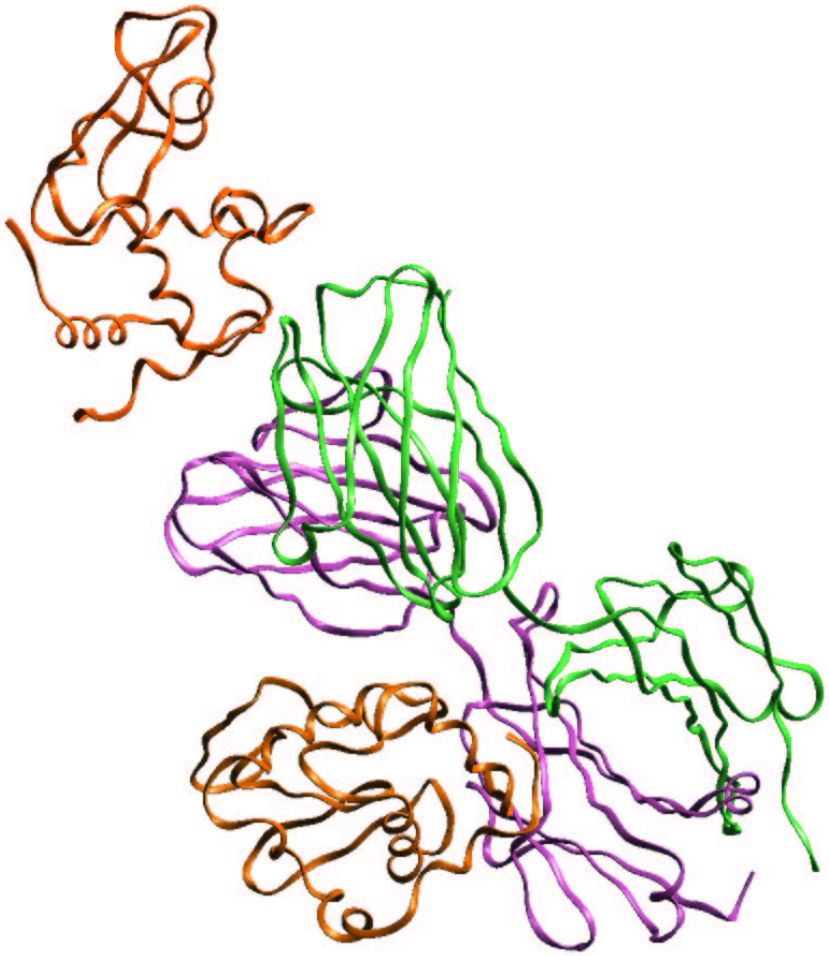


Abbildung 4.2: Immunglobulin mit Lysozym (1fdl): native Orientierung des Lysozyms (orange) im Komplex (oben) und falsch positive Orientierung (mitte) in der Konkavität des Immunglobulins (blau/violett)

Um zu untersuchen worin sich die einzelnen Interfaces unterscheiden, wurde

die beste falsch-positive Lösung ($\Theta_1 = 285.6^o$, $\Theta_2 = 60.0^o$, $\Theta_3 = 245.6^o$, $dx = 34.6$ Å, $dy = 39.5$ Å, $dz = -30.6$ Å), sowie die native Orientierung des 1fdl Komplexes und die native Lösung des 3hfm Komplexes mit Hilfe von Janet Thorntons Protein Interaction Server untersucht (Thornton, 1998). Die Ergebnisse sind in Tabelle 4.3 zusammengetragen.
Die Analyse zeigt, daß der 3hfm Komplex ein etwa 30% größeres Interface besitzt als der 1fdl Komplex. 843 Å^2 des Antikörpers und 864 Å^2 des Liganden bei 3hfm stehen 647 Å^2 und 674 Å^2 bei 1fdl gegenüber. Das Interface der falsch positiven Lösung ist durch die Lage in einer konkaven Tasche recht groß (s. Abbildung 4.2). Es bedeckt 891 Å^2 der Antikörperoberfläche und 868 Å^2 der Lysozymoberfläche und ist damit sogar noch etwas größer als das Interface des 3hfm Komplexes. Die Daten weisen darauf hin, daß die Packungsdichte, obwohl nicht explizit bestimmt, im falsch positiven Interface geringer ist: Das Lückenvolumen ist mit 6300-8600 Å^3 fast doppelt so groß wie in den nativen Interfaces mit ca 2600-3400 Å^3. Des weiteren ist bei der nicht nativen Orientierung der größere Interfaceteil zwischen der L-Kette des Antikörpers und dem Lysozym sehr viel stärker gekrümmt als bei den beiden korrekten Interfaces. Schließlich sind im falsch positiven Interface nur zwei Wasserstoffbrückenbindungen vorhanden, während die natürlichen Interfaces elf bzw. zwölf aufweisen. Insgesamt gibt die Analyse genug Anhaltspunkte um das nicht-native Interface als solches zu identifizieren.

4.2 Docking von Membranproteinen und Substrukturen

Eine interessante Fragestellung, der man mit Hilfe von Docking-Programmen nachgehen kann, ist ob die Oberflächen von Sekundärstrukturelementen im Inneren von Proteinen ähnlich spezifische Interfaces aufweisen, wie die Oberflächen von Untereinheiten. Besondere Bedeutung hat diese Fragestellung bei Transmembranproteinen. Da sich die membrandurchspannenden Helices in einer hydrophoben Umgebung befinden, könnten hier andere Verhältnisse vorliegen als in Lösung. Im Rahmen eines Praktikumsprojekts wurden aus Transmembranproteinkomplexen Teilstücke, meist einzelne Helices mit Hilfe von CKORDO

Komplex		3hfm		1fdl (true positive)		1fdl (false positive)	
	Ketten	L/H	Y	L/H	Y	L/H	Y
dASA [Å^2]	H-Y	474	511	376	355	353	347
	L-H	369	353	271	319	538	521
H-Brücken (#)	H-Y	7		8		1	
	L-Y	4		4		1	
Lückenvolumen [Å^3]	H-Y	2589		2764		6310	
	L-Y	3429		3021		8573	
Planarität [Å]	H-Y	1.96	2.2	1.5	1.16	1.59	1.86
	L-H	0.94	0.97	1.2	1.57	4.01	3.73
Polarität [%]	H-Y	37	56	52	47	40	43
	L-H	48	49	45	57	47	419

Tabelle 4.3: Ergebnisse einer Analyse von drei Antigen/Antikörper Komplexen: 3hfm, 1fdl (true positive und false positive), dASA = Änderung der lösungsmittelzugänglichen Fläche

gedockt. Dabei wurde außerdem eine Anzahl weiterer Eigenschaften der erzeugten Orientierungen berechnet um Möglichkeiten für Postfilter zu untersuchen. Ein geeigneter Testfall für das Docken von Transmembranstrukturen ist die Cytochrom-C Oxidase aus Rinderherzmitochondrien.

4.2.1 Cytochrom-C Oxidase aus Rinderherzmitochondrien

Der 1995 von Tsukihara *et al.* aufgeklärte Cytochrom-C Oxidase-Komplex mit der PDB-Bezeichnung 1occ hat senkrecht zur Membran betrachtet die Form einer Schiffsschraube, deren zwei Flügel identische Subkomplexe mit je dreizehn Untereinheiten darstellen (s. Abb. 4.3 auf der nächsten Seite nach: Tsukihara *et al.* (1996)). Mit insgesamt etwa 3000 Aminosäuren und 28 Transmembranhelices stellt 1occ einen der größten strukturaufgeklärten Transmembrankomplexe dar. Er katalysiert die abschließende Oxidation der Zellatmung, ein Prozeß bei dem molekularer Sauerstoff durch Elektronen des Cytochrom-C zu Wasser reduziert wird. Gleichzeitig werden Protonen von der Matrixseite der inneren Mitochondrienmembran zur cytosolischen Seite in den Intermembranraum gepumpt. Die dreiflügelige Untereinheit I (gelb) flankiert von den ebenfalls mitochondrial kodierten Untereinheiten II (blau) und III (grün) bildet das Zentrum des Komplexes

mit insgesamt 21 Transmembranhelices. Neben den sieben kerncodierten Untereinheiten (weiße römische Ziffern) mit jeweils einer Transmembranhelix liegen noch mindestens acht Phospholipidmoleküle im Komplex.
Tabelle 4.4 gibt einen Überblick über verwendete Helices und Parameter der durchgeführten Dockings.

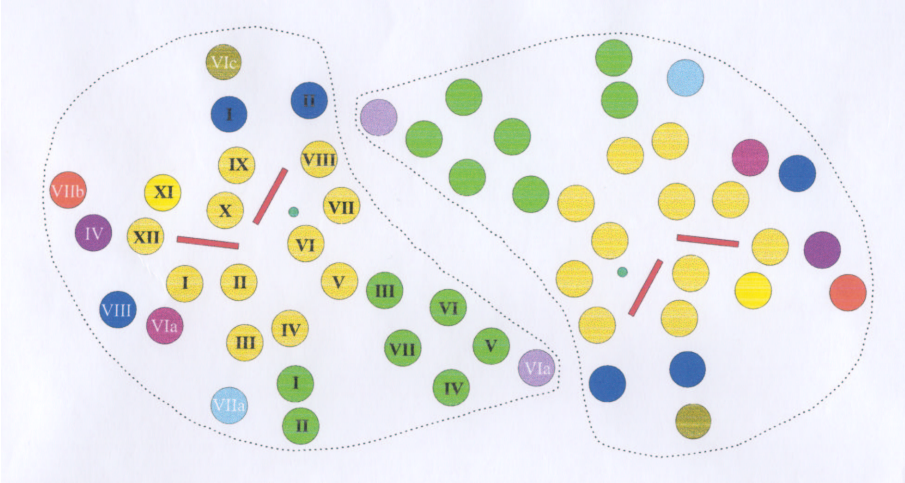


Abbildung 4.3: Transmembranhelices von Cytochrom-C Oxidase

Nr.	Helices	Winkel	geom. Korr.	Gapvol.	Buried Surf.	Gapindex	Intermol. Kontakte	H-Brücken	Salz-brücken
		[°]	[Rang]	[Å ^{3}]	[Å ^{2}]	[Å ^{2}]			
1	I(III)/II(III)	20	1	759/842	1084/1022	0.70/0.82	n.d/n.d.	n.d./n.d.	n.d./n.d.
2	I(VIa)/IV(III)	20	1	972/890	1078/1097	0.90/0.81	78/134	0/0	0/0
3	I(VIc)/I(II)	20	1	1279/1283	1590/1583	0.80/0.81	190/172	1/4	4/3
4	I(VIc*)/I(II)	20	9	1711/1283	1575/1583	1.09/0.81	202/180	2/4	0/3
5	I(VIc)/I(II*)	20	56	1885/1283	1119/1583	1.68/0.81	110/180	1/4	0/3
6	I(VIc*)/I(II*)	20	266	4845/1283	1020/1583	4.75/0.81	110/180	0/4	0/3
7	I(III)/III(I)+IV(I)	20	1	2580/2600	1159/1152	2.23/2.26	102/86	1/1	1/1
8	I(VIIa)/I(III)+II(III)	20	1	1632/1592	1439/1448	1.13/1.10	68/74	3/2	2/2
9	I(IV)/XI(I)+XII(I)	20	109	3780/3023	1127/1191	3.35/2.54	68/74	1/4	0/1
10	I(VIII)/I(I)+XII(I)	50	13	4560/5294	1305/888	3.49/5.96	140/46	0/0	0/0
11	I(VIII)/I(IV)+XII(I)	50	1	4482/4452	1260/1212	3.56/3.67	186/108	2/2	0/0
12	I(VIIc)/I(I)+XII(I)	50	1	4203/4389	1482/1409	2.84/3.11	210/164	2/3	0/0

Tabelle 4.4: Ergebnisse des Dockings von Transmembranhelices der Cytochrom-C Oxidase. Jeweils für gedockte und native Struktur

Die Ergebnisse in Tabelle 4.4 zeigen, daß parallel liegende Helices mit einem Winkel von maximal 20° problemlos einzeln gedockt werden können (Nr. 1-3). Auch bei größerem Winkel der Helices zueinander lassen sich noch Substruktu-

ren docken (Nr. 12), solange die Interfaces eine Mindestanzahl an Atom-Atom Kontakten nicht unterschreiten (Nr. 9+10). Die Seitenketten der mit * bezeichneten Helices sind mit dem Programm SCWRL positioniert worden, um zu testen wie robust die verschiedenen Messgrößen bzgl. Seitenkettenumlagerungen sind (Nr. 4-6). Dabei werden Umlagerungen in einer Helix noch toleriert (Nr. 4+5). Werden hingegen beide Helices mit SCWRL behandelt, so sinkt die Qualität des Interfaces deutlich (Nr. 6). Dieser Fall entspricht der Vorbereitung zweier unabhängig voneinander kristallisierten Strukturen mit SCWRL.

4.3 *Unbound*-Docking

Um sicherzustellen, daß mehrere Orientierungen, die dem nativen Komplex ähneln gefunden werden, wurden 12 Dockingläufe mit einem sehr feinen Winkelinkrement von 12° durchgeführt. Die Auswahl der Testfälle erfolgte unter dem Gesichtspunkt der Vergleichbarkeit. Daher wurden nur solche Proteinkombinationen ausgewählt, die mindestens von zwei verschiedenen Gruppen vorher untersucht worden sind (s. Tabelle 2.1 auf Seite 18). Details der letztendlich verwendeten Testfälle sind in Tabelle 4.5 zusammengetragen.

Komplex			Einzelproteine							
PDB	Res	Beschreibung	PDB	Res	AS	RMSD	PDB	Res	AS	RMSD
2kai	2.5	Kallikrein A/pankr.Tryp.-Inhib.	2pka	2.1	232	0.542	1bpi	1.1	58	0.539
2ptc	1.9	β-Trypsin/pankr.Tryp.-Inhib.	3ptn	1.7	223	0.322	4pti	1.5	58	0.457
1cho	1.8	α-Chymotrypsin/OMTKY	5cha	1.7	245	0.620[1]	2ovo	NMR	56	
2sni	2.1	Subtilisin Novo/ Chymotrypsin-Inhib.	1sup	1.6	275	0.582	2ci2	2.0	65	0.459
1brs	2.0	Barnase/Barstar	1bni	2.1	107	0.770	1bta	NMR	89	1.44
1fss	3.0	Acetylcholinesterase/Fasciculin II	2ace	2.5	531	0.760[1]	1fsc	2.0	61	
1mah	3.2	Acetylcholinesterase/Fasciculin II	1maa	2.9	545	0.600[1]	1fsc	2.0	61	
1mda	2.5	Methylamin-Dehydrogenase/ Amicyanin	2bbk	1.6	489	0.529	1aan	2.0	103	0.887
2pcc	2.3	Peroxidase/Cytochrom-C	1ccp	2.2	293	0.930[1]	1ycc	1.2	108	
1mlc	2.1	Fab D44.1 (Ketten A,B)/Lysozym	1mlb	2.1	432	0.953	1lza	1.6	129	0.627
1vfb	1.8	Fv D1.3 (Ketten A,B)/Lysozym	1vfa	1.8	223	–	1hel	1.7	129	
1vfb	1.8	Fv D1.3 (Ketten A,B)/Lysozym	1vfa	1.8	223	0.325	1lza	1.6	129	1.066

Tabelle 4.5: Verwendete Testfälle für das *unbound*-Docking. [1] entnommen aus Chen & Weng (2002)

Für das Training der *Support Vector Machines* wurden von den Tests unabhängige Dockingläufe verwendet. Diese wurden ebenfalls mit einem Winkelinkrement

von 12° durchgeführt, um Datensätze zu produzieren, die genug richtige Orientierungen enthalten. Für jede Rotation wurden die fünf Orientierungen mit der höchsten geometrischen Korrelation ausgegeben. Für alle Anordnungen wurden folgende Größen bestimmt:

- geometrische Korrelation
- elektrostatische Korrelation
- hydrophobe Korrelation
- verdeckte Oberfläche (*buried surface*)
- Lückenvolumen
- Lückenweite
- Wert des statistischen Potentials
- RMSD

Die graphischen Auftragungen dieser Eigenschaften für alle zwölf Docking-Fälle sind im Anhang in den Abbildungen A.1 bis A.12 zusammengetragen. Durch das schlechte Laufzeitverhalten des Programms SURFNET, welches von CKORDO zur Berechnung des Lückenvolumens eingesetzt wird, dauerten die einzelnen Dockings je nach Größe des Proteins zwischen 16 und 70 Stunden auf einem PC mit 1.4 GHz Athlon-Prozessor. Davon entfielen fast 90% auf die Laufzeit von SURFNET, so daß bei Abschalten der Berechnung des Lückenvolumens Rechenzeiten zwischen 1.5 und 8 Stunden erreicht werden.
Die Ergebnisse für die zwölf Dockings sind in Tabelle 4.6 zusammengetragen und zeigen, daß das Winkelinkrement von 12° fein genug gewählt wurde um in allen Fällen, mit Ausnahme des Cytochrom-C Oxidase-Komplexes 2pcc, mehrere Orientierungen mit einem RMSD von weniger als 3.5 Å zur superpositionierten Ligandenstruktur (=positive Lösungen) zu erhalten. Aus jeweils $1.9 \times 10^9 - 6.3 \times 10^9$ untersuchten Orientierungen sind insgesamt maximal 43080 Lösungen mit hoher geometrischer Korrelation ausgegeben worden. Besonders viele korrekte Strukturen wurden bei den Serin-Protease-Dockings erzeugt. Mit

Komplex	Kleiner 3.5 Å			Bester RMSD	A-1000	Kontakte <2.5 Å
	Gesamt-Anzahl	Anzahl unter 1000	Höchster Rang			
2kai:2pka-1bpi	56	7	82	1.27	1.58	40
2ptc:2ptn-4pti	55	6	2	1.70	2.34	32
1cho:5cha-2ovo	31	3	144	1.10	1.14	13
2sni:1sup-2ci2	10	0	15037	2.66	4.30	31
1brs:1bni-1bta	23	0	1736	1.33	3.99	24
1fss:2ace-1fsc	7	1	295	1.22	1.22	25
1mah:1maa-1fsc	27	6	29	0.87	0.87	12
1mda:2bbk-1aan	7	0	13590	0.80	6.92	10
2pcc:1ccp-1ycc	0	0	–	3.815	7.29	12
1mlc:1mlb-1lza	3	0	7829	2.79	9.339	38
1vfb:1vfa-1hel	6	0	3354	1.96	4.637	4
1vfb:1vfa-1lza	8	0	2556	2.32	7.88	3

Tabelle 4.6: Ergebnisse der geometrischen Korrelation bei fünf Korrelationsmaxima je Rotation. A-1000 = bester RMSD innerhalb der ersten 1000 Lösungen

Ausnahme des Subtilisin-Komplexes 2sni ist für alle Protease-Komplexe eine Anordnung unter 2.5 Å im Ergebnisdatensatz enthalten und in vier von sechs Fällen wird eine solche Lösung unter den 1000 höchsten geometrischen Korrelationen gefunden. Außer den Protease-Komplexen finden sich auch für die beiden Acetylcholinesterase-Komplexe positive Lösungen unter den 1000 besten geometrischen Korrelationen. Für die drei Antikörper-Testfälle gibt es nur wenige positive Lösungen. Sie werden in keinem der Fälle unter den ersten 1000 Korrelationsmaxima gefunden. Das gleiche gilt für den Methylamin-Dehydrogenase-Komplex. Die Anzahl der Kontakte mit zu niedrigem interatomaren Abstand steht nicht im direkten Verhältnis mit der erreichbaren geometrischen Komplementarität. Die Tabellen 4.7 und 4.8 fassen die Ergebnisse des Rankings bei Verwendung verschiedener Filter in den zwölf *unbound*-Docking-Fällen zusammen. Sie zeigen jeweils den Rang der ersten Lösung mit einem RMSD unter 3.5 Å. Die Kennzeichnung des jeweiligen Filters ist der Beschriftung zu entnehmen. Tabelle 4.7 auf der nächsten Seite zeigt die Ergebnisse ohne Einsatz des SVM-Klassifikators. Der Vergleich der ersten beiden Spalten zeigt, daß sich nur bei beim Methylamin-Dehydrogenase-Komplex die Erhöhung der Anzahl

von Lösungen, die je Rotation behalten werden, positiv auf das Ranking auswirkt. In allen anderen Fällen verschlechtert sich der Rang zum Teil erheblich. Die Anwendung des Elektrostatik-Filters (Spalte 3) zeigt nur bei den Protease-Komplexen und dem Barnase-Komplex nennenswerte Wirkung, der Hydrophobizitätsfilter (Spalte 4) nur beim Antikörper-Komplex 1mlc. Der Lückenfilter (Spalte 5) begrenzt vor allem die in Kap. 4.1.3 auf Seite 53 besprochenen falsch positiven Lösungen bei Antikörper-Komplexen. Diese starke Filterwirkung führt allerdings auch zum Ausfiltern positiver Lösungen, wie man am schlechteren Ranking des Acetylcholin-Esterase-Komplexes 1fss und dem Verschwinden aller positiven Lösungen beim Subtilisin-Komplex 2sni erkennen kann. Der effizienteste Filter ist der Potentialfilter. Das verwendete Atom-Atom Kontaktpotential zeigt über alle Testfälle hinweg starke Filterwirkung. Insgesamt finden sich für acht Komplexe richtige Lösungen unter den ersten tausend Vorhersagen, in der Hälfte der Fälle sogar unter den ersten hundert. Nur in zwei Fällen werden alle positiven Lösungen durch die Filter ausgesondert.

Komplex	Rang bei verschiedenen Filtern					
	G5	G1	G1+E	G1+E +H	G1+E +H+L	G1+E +H+L+P
2kai:2pka-1bpi	82	81	46	34	34	20
2ptc:2ptn-4pti	2	2	2	2	2	2
1cho:5cha-2ovo	144	143	109	80	80	23
2sni:1sup-2ci2	15037	8484	7866	6328	-	-
1brs:1bni-1bta	1736	1574	1072	1003	951	137
1fss:2ace-1fsc	295	292	278	234	250	141
1mah:1maa-1fsc	29	29	28	27	27	20
1mda:2bbk-1aan	13590	-	-	-	-	-
2pcc:1ccp-1ycc	-	-	-	-	-	-
1mlc:1mlb-1lza	7829	5128	5006	4561	2114	1102
1vfb:1vfa-1hel	3354	2755	2643	2423	1810	711
1vfb:1vfa-1lza	2556	2198	2140	1983	1551	684

Tabelle 4.7: Ergebnisse der geom. Korrelation bei Anwendung verschiedener Filter: Gx : x geom. Korrelationsmaxima je Rotation, E: Elektrostatische Korrelation auf -30000. . .20000 beschränkt, H: Hydrophobizität auf 0. . .150000 beschränkt, L: Lückenweite auf 1.0. . .6.0 Å beschränkt, P: statistisches Potential auf max. -4.0 beschränkt

Tabelle 4.8 auf der nächsten Seite zeigt die Ergebnisse für die gleiche Abfolge

an Filtern wie Tabelle 4.7 auf der vorherigen Seite aber unter Einbeziehung des SVM-Klassifikators. Mit Ausnahme des Antikörper-Komplexes 1mlc zeigt der Klassifikator eine ausgezeichnete Filterwirkung (Spalte 2). Im Falle des Subtilisin-Komplexes 2sni wird allerdings keine Lösung als positiv klassifiziert, beim β-Trypsin-Komplex 2ptc wird die geometrisch zweitbeste Lösung vom SVM-Klassifikator falsch klassifiziert, was das Ranking für diesen Komplex stark verschlechtert. Der Elektrostatikfilter zeigt seine Wirkung bei den stark geladenen Interfaces des Barnase- und des Kallikrein-Komplexes, der nachgeschaltete Hydrophbizitäts-Filter entfaltet praktisch keine Wirkung. Der große Methylamin-Dehydrogenase-Komplex 1mda und der Antikörperkomplex mit dem großen FAB-Fragment 1mlc werden durch den Lückenweitenfilter von vielen falsch positiven Lösungen befreit, die durch ein zwar großes aber nur ungenau passendes Interface einen hohen geometrischen Korrelationswert erhalten haben. Der Potentialfilter hat in Kombination mit dem SVM-Klassifikator so gut wie keine Wirkung.

Komplex	Rang bei verschiedenen Filtern					
	G5	G5+S	G5+S+E	G5+S+E +H	G5+S+E +H+L	G5+S+E +H+L+P
2kai:2pka-1bpi	82	11	5	5	5	5
2ptc:2ptn-4pti	2	212	194	188	178	168
1cho:5cha-2ovo	144	7	7	7	7	6
2sni:1sup-2ci2	15037	-	-	-	-	-
1brs:1bni-1bta	1736	28	19	19	19	19
1fss:2ace-1fsc	295	115	110	106	95	86
1mah:1maa-1fsc	29	18	18	17	17	16
1mda:2bbk-1aan	13590	3310	2913	2543	1315	1095
2pcc:1ccp-1ycc	-	-	-	-	-	-
1mlc:1mlb-1lza	7829	5035	4889	4348	1267	1369
1vfb:1vfa-1hel	3354	622	591	528	427	390
1vfb:1vfa-1lza	2556	513	498	458	381	354

Tabelle 4.8: Ergebnisse der geom. Korrelation bei Anwendung der SVM-Klassifizierung und verschiedener Filter: Gx: x geom. Korrelationsmaxima je Rotation, S: SVM-Klassifikator, E: Elektrostatische Korrelation auf -30000. . .20000 beschränkt, H: Hydrophobizität auf 0. . .150000 beschränkt, L: Lückenweite auf 1.0. . .6.0 Å beschränkt, P: statistisches Potential auf max. -4.0 beschränkt

4.3.1 Ergebnisse des *Suport Vector Machine* Klassifikators

Von insgesamt 248 positiven Lösungen in allen zwölf Dockings wurden 94 gefunden, dies sind 38% (s. Tabelle 4.9). Allerdings sind die Ergebnisse für die einzelnen Testfälle sehr unterschiedlich. So werden je positiver Lösung beim Kallikrein-Testfall 2kai nur neun falsch positive Lösungen über 10 Å RMSD erzeugt, während beim Antikörper-Testfall 1mlc mehr als 8000 falsch positive Lösungen mit hohem RMSD erzeugt werden. Insgesamt zeigen die Werte gute Diskriminierung für Protease-Inhibitor Komplexe, die ja auch die Trainingsmenge darstellten. Ausnahmen in dieser Klasse sind Subtilisin (2sni), wo keine der 10 positiven Lösungen gefunden wurde und der β-Trypsin Komplex 2ptc, dessen Ergebnisse aus einem anderen Docking 50% der Trainingsmenge stellt.

Test Komplex	tp	tn	fp	fn	Spez.	Sens.	fp/tp	PPV
1cho_rrot	9	40463	143	22	0.996	0.290	15.9	0.0592
1cgi_unrot	4	23954	38	9	0.998	0.308	9.5	0.0952
1MLC_rrot	2	25181	16785	1	0.600	0.667	8392.5	0.0001
1VFB_1hel_rrot	5	39513	3200	1	0.925	0.833	640.0	0.0016
1VFB_1lza_rrot	5	39552	3034	3	0.929	0.625	606.8	0.0016
1brs_rrot	1	42071	58	22	0.999	0.043	58.0	0.0169
1fss_rrot	4	36964	4912	3	0.883	0.571	1228.0	0.0008
2sni_rrot	0	36706	176	10	0.995	0.0		0.0
1mah_rrot	21	36338	5556	6	0.867	0.778	264.6	0.0038
2pcc_rrot	0	40207	2553	0	0.940	0.0		0.0
1MDA_rrot	4	35440	7176	3	0.832	0.571	1794.0	0.0006
2kai_rrot	38	38876	347	19	0.991	0.667	9.1	0.0987
2ptc_rrot	1	38673	91	55	0.998	0.018	91.0	0.0109

Tabelle 4.9: Ergebnisse der SVM Untersuchungen. rrot = randomisierte Proteinkoordinaten, unrot = nicht randomisierte Proteinkoordinaten, tp = Anzahl richtig positiver Lösungen, fp = Anzahl falsch positiver Lösungen, tn = Anzahl richtig negativer Lösungen, fn = Anzahl falsch negativer Lösungen, Spez. = Spezifität, Sens. = Sensitivität, PPV = *positive prediction value*. Zur Erklärung s. Kapitel 2.2.1 auf Seite 18.

Für die restlichen Enzym-Komplexe (Acetylcholinesterase, Methylamin-Dehydrogenase) und die Antikörper-Komplexe wird zwar die Mehrzahl der Lösungen gefunden, die Anzahl der falsch positiven Lösungen ist jedoch relativ gesehen zu hoch, d.h. eine positive Klassifizierung ist nur in 0.1 - 3.7

‰ der Fälle korrekt. Um die Spezifität für Antikörper-Komplexe zu erhöhen, wurde nachfolgend versucht, einen Klassifizierer für Antikörper-Dockings zu trainieren. Als Trainingsmenge wurden Ergebnisse der beiden Dockingläufe mit dem Referenzkomplex 1vfb verwendet. Es konnte jedoch kein Klassifizierer gefunden werden, der in der Lage ist auch nur eine der positiven Lösungen des Antikörper-Komplexes 1mlb korrekt zu klassifizieren.

5 Diskussion

Bei der Entwicklung von CKORDO stand im Vordergrund, die sehr guten Leistungen des Vorläuferprogramms KORDO bezüglich der Komplexvorhersage mit geometrischer Korrelation in einer Neuimplementation durch die Berechnung weiterer Eigenschaften so zu erweitern, daß beliebige Komplexe auch aus den Strukturdaten einzeln kristallisierter Proteinen vorhergesagt werden können. Als Erweiterungen wurden im wesentlichen fünf Neuerungen eingeführt: die elektrostatische Korrelation, die hydrophobe Korrelation, die Berechnung geometrischer Eigenschaften des Interfaces insbesondere der Lückenweite, die Bewertung der Atom-Atom Kontakte durch ein statistische Potential und die Verwendung von *Support Vector Machines* zur Entwicklung eines Klassifikators, der alle berechneten Werte in geeigneter Weise kombiniert.

5.1 Qualität der Dockingvorhersagen mit CKORDO

5.1.1 Ergebnisse der *bound*-Docking Testfälle

Bound Docking ist seit mehreren Jahren ein weitgehend gelöstes Problem. CKORDO ist, genau wie andere Protein-Docking-Programme, in den meisten Fällen in der Lage, aus den getrennten Koordinaten der beiden Teile des kokristallisierten Komplexes, eine der nativen Orientierung sehr ähnliche Struktur zu rekonstituieren (s. Tabelle 4.1 auf Seite 51). Schwierigkeiten ergeben sich durch steile Korrelationsfunktionen bei stark verzahnten Interfaces und durch große flachkonkave Oberflächen. Letztere führen bei Dockingexperimenten mit Antikörpern zu vielen falsch positiven Lösungen.
Diese Schwierigkeiten sind nicht spezifisch für das *bound* Docking. Beim Docking der Antikörper-Komplexe aus Einzelstrukturen sind die falsch-positiven Lösungen in gleicher Weise zu beobachten. Dem verzahnten Interface entsprechen beim *unbound* Docking stark vorspringende Seitenketten im Interface von Einzelproteinen, die sich in der Komplexstruktur in eine andere Position bewegt haben.

Im direkten Vergleich mit CKORDO scheint die Berechnung der geometrischen Komplementarität des Vorläuferprogramms KORDO in einigen Fällen andere, z.T. bessere Ergebnisse zu liefern als der Nachfolger (Heuser, 2002). Untersuchungen zur Ursache der Differenzen sind in Arbeit. Die Tatsache, daß CKORDO im Gegensatz zu KORDO Heteroatome noch ignoriert, könnte die Unterschiede in einigen Fällenerklären. Desweiteren sind die verwendeten van der Waals Radien und das Gewichtungsschema für Temperaturfaktor und Besetzung bei den beiden Programmen nicht identisch.
Interessante Ergebnisse ergaben sich bei einem bisher nur selten untersuchten Testfall für das *bound* Docking, der Zusammensetzung von Transmembranproteinkomplexen aus Substrukturen, im Extremfall einzelnen Helices. Das Docking von Substrukturen wurde an Hand des Beispiels des Cytochrom-C Oxidase Komplexes 2occ untersucht. Ein wichtiges Ergebnis ist, daß in Transmembranproteinen d.h. in hydrophober Umgebung offensichtlich eine ebenso starke Notwendigkeit zur Ausbildung geometrischer Komplementarität herrscht wie in löslichen Proteinen. Darüberhinaus wurden verschiedene Filtermethoden getestet, die je nach Erfolg in CKORDO eingebaut wurden. Die Analyse der Ergebnisse läßt sich wie folgt zusammenfassen:

- Ähnlich wie bei anderen *bound* Dockings ist nahe der nativen Orientierung eine hohe geometrische Komplementarität zu beobachten.

- Interfaces müssen eine bestimmte Mindestgröße von etwa 900 Å haben, um von dem Algorithmus detektiert werdem zu können.

- Bezüglich einer abschließenden Filterung sind die Anzahl der Atomkontakte, Salzbrücken und Wasserstoffbrücken nicht robust gegenüber Seitenkettenumlagerungen, kleineren Translations- oder Rotationsänderungen. Nur die Lückenweite erscheint aus diesem Grund für einen Postfilter geeignet. Offensichtlich ist die Anzahl von Salzbrücken und Wasserstoffbrücken in der nativen Seitenkettenanordnung jedoch deutlich höher als in den leicht abweichenden Orientierungen, so daß sich diese Werte für die Entwicklung von speziellen Seitenkettenpotentialen für die Minimierung von Komplexstrukturen einsetzen lassen.

- Der SURFNET-Algorithmus berechnet als Lückenvolumen offensichtlich auch Teilvolumina, die außerhalb des eigentlichen Interfaces liegen. Hierdurch werden bei stark gegeneinander verkippten Helices zu große Lückenvolumina bestimmt.

- Wie die Versuche zur Vorbehandlung der Seitenketten durch statistische Potentiale zeigen, werden Seitenketten durch das Programm SCWRL nicht in eine Position gebracht, die derjenigen im Komplex ähnelt.

Der letzte Punkt ist sicher teilweise dadurch erklärbar, daß die von SCWRL zur Berechnung eingesetzten Potentiale nicht auf Protein-Protein Interfaces hin optimiert sind, sondern auf die bei der Vorhersage von Proteinfaltung betrachteten Interfaces innerhalb von Proteinen. Eine Arbeit aus der Arbeitsgruppe von Barry Honig konnte zeigen, daß die Lage der Seitenketten z.T. stark von Kristallwechselwirkungen bestimmt ist (Jacobson *et al.*, 2002).

5.2 *Unbound* Docking Resultate

Die in den letzten Kapiteln dargestellten Ergebnisse zeigen, daß CKORDO in Verbindung mit einer Bewertung der Primärergebnisse durch *Support Vector Machines* und empirische Filter in der Lage ist, sinnvolle Vorhersagen für *bound* Docking-Fälle, sowie für *unbound* Docking-Fälle von Protease-Inhibitor-Komplexen zu treffen. Die Vorhersagequalität bei *unbound* Testfällen von Antikörper-Komplexen und Enzym-Komplexen mit Ausnahme der Protease-Inhibitor-Komplexe sind sehr unterschiedlich und z.T. für praktische Anwendungen noch nicht ausreichend.
Die geometrische Korrelation von CKORDO alleine betrachtet ist nicht in der Lage *unbound* Testfälle mit ausreichender Genauigkeit vorherzusagen (s. Tab. 5.1 auf der nächsten Seite). Im Fall der beiden Acetylcholinesterase-Komplexe, sowie der Serinprotease-Komplexe mit Ausnahme von 2sni ist die geometrische Korrelation hoch genug um ohne weitere Kriterien ein vergleichbar gutes Ranking zu erzielen. Für 2pcc findet sich überhaupt keine Struktur mit einem RMSD unter 3.5 Å in der Lösungsmenge.

Insgesamt bleibt die geometrische Korrelation aber weiterhin die wichtigste Größe zur Auswahl von Lösungen. Die Verwendung von einem (G1) bzw. fünf (G5) Korrelationsmaxima je Rotation aus den 216000 bis 729000 Korrelationswerten die pro Rotation erzeugt werden, reduziert die Menge der Lösungen um fünf Größenordnungen und behält dabei dennoch bis zu 56 nativ-ähnliche Orientierungen (2kai). Alle Sortierungen zur Bestimmung des Rangs werden anhand der geometrischen Korrelation durchgeführt. Die Verwendung anderer Sortierkriterien, wie des statistischen Potentials wurden getestet, ergaben aber schlechtere Rankings. Die Ergebnisse aus den zwölf Dockingexperimenten machen deutlich, daß es bei einem Winkelinkrement von 12° nicht notwendig ist mehrere Korrelationen pro Rotation zu speichern. Fast alle Rankings werden durch die zusätzlichen Orientierungen schlechter.

Komplex	Filter			Ranking aus der Literatur					
	G1	G1 +Filter	G5+svm +Filter	[1]	[2]	[3]	[4]	[5]	[6]
2kai	81	20	5	4	-	33	n.f.	-	-
2ptc	2	2	168	-	-	9	52	63	-
1cho	143	23	6	-	72	8	6	-	-
2sni	8484	n.f.	n.f.	-	2	2	16	-	-
1brs	1574	137	19	-	-	-	-	1	-
1fss	292	141	86	-	210	-	11	3	-
1mah	29	20	16	-	5	-	-	-	82
1mda	n.f.	n.f.	1095	1	-	-	-	-	-
2pcc	n.f.	n.f	n.f.	-	n.f.	-	50	-	266
1mlc	5128	1102	1369	55	18	1	n.f.	-	-
1vfb_1hel	2755	711	390	-	-	-	-	894	-
1vfb_1lza	2198	684	354	n.f.	n.f.	-	-	4673	-

Tabelle 5.1: Ergebnisse von CKORDO und anderen Docking-Programmen, [1] Gardiner *et al.* (2001), [2] Chen & Weng (2002), [3] Gabb *et al.* (1997), [4] Palma *et al.* (2000), [5] Heifetz *et al.* (2002), [6] Mandell *et al.* (2001)

Das Ranking der unterschiedlichen Testfälle durch CKORDO entspricht, mit einigen Ausnahmen, den mit anderen aktuellen Docking-Programmen erreichten Leistungen, wie Tabelle 5.1 zeigt. Die Ergebnisse sollten nur als eine Zusammenstellung betrachtet werden, nicht jedoch als direkter Vergleich, da die Vorraussetzungen zum Zustandekommen der einzelnen Resultate höchst unter-

schiedlich sind. So sind die RMSD-Werte für den Methylamin-Dehydrogenase-Komplex 2mda und den Kallikrein-Komplex 2kai bei Gardiner *et al.* ([1]) mit 5.05 bzw. 4.39 Å bereits weit außerhalb dessen, was für CKORDO als positive Lösung gewertet wurde. Bei den Antikörper-Testfällen 1vfb und 1mlc sind in allen anderen Arbeiten, entweder bei der Repräsentation der Proteine für das Dockingprogramm oder bei der Bewertung der Ergebnisse, Kenntnisse über die Lage der Antigen-Bindungsstelle eingeflossen. In der Arbeit von Gabb *et al.* wird in den Filterschritten auch bei den Serinproteasen die Kenntnis über die Lage der Bindungsstelle verwendet. Die Einbeziehung solcher Informationen ergibt selbstverständlich bessere Rankings, als alleine durch Verwendung der PDB-Strukturdaten zu erhalten ist, wie in vorliegender Arbeit geschehen.

5.2.1 Leistung des SVM-Klassifizierers

Werden die Einzelergebnisse analysiert, so fällt auf, daß die Klassifizierungsqualität durch die *Support Vector Machine* insgesamt gut ist, auch wenn innerhalb der Serinprotease-Inhibitor-Komplexe höchst unterschiedliche Bewertungsergebnisse erzielt werden. So sind die Klassifizierungen für 2ptc deutlich schlechter als diejenigen für 2kai (s.a. 4.9 auf Seite 63) obwohl das Training der SVM mit zwei unabhängig erzeugten Dockings dieser beiden Komplexe durchgeführt wurde und die Trainingsmenge von beiden Proteinen eine ähnliche Anzahl positiver und negativer Lösungen beinhaltete. Die guten Klassifizierungsleistungen für den Barnase-Barstar-Komplex, die beiden Acetylcholinesterase-Komplexe und die 1vfb Antikörper-Lysozym-Testfälle belegen, daß die SVM keine für das Interface von Serinproteasen spezifischen Eigenschaften gelernt hat. Die Filterleistung des SVM-Klassifizierers ist in acht von elf Fällen höher als die für alle anderen Filter zusammen, d.h die alleinige Anwendung der SVM-Klassifizierung ergibt einen höheren Rang als die Filterung durch alle anderen Kriterien.

5.2.2 Leistungen der Einzelfilter

- Die Verwendung der Elektrostatikfilters wirkt sich besonders positiv beim Barnase-Barstar-Komplex aus. Hier werden etwa ein Drittel der falsch po-

sitiven Lösungen vor der ersten richtig positiven Lösung eliminiert. Das Interface des Barnase-Barstar-Komplexes ist stark geladen und enthält mehrere Salzbrücken. Die Anwesenheit vieler geladener Reste an der Oberfläche führt zu einer hohen Wahrscheinlichkeit von falsch positiven Lösungen, bei denen sich Ladungsfelder mit gleichem Vorzeichen überlagern. Daraus resultieren stark negative elektrostatische Korrelationen. Andererseits ist das Potential empfindlich bei unphysiologisch nahen Überlappungen entgegengesetzter Ladungen. Beide Phänomene werden durch die Elektrostatik gemessen und Lösungen bei entsprechender Häufung durch den Filter als unphysiologisch aussortiert.

- Die hydrophobe Korrelation ist sowohl in seiner Implementierung als auch in seiner Wirkung der elektrostatischen sehr ähnlich. Hauptunterschiede sind, daß sich Überlagerungen gleicher Ladungen und Überlagerungen ungleicher Ladungen nicht gegenseitig kompensieren und daß nur kurzreichweitige Felder verwendet werden. Die von elektrostatischem und hydrophobem Filter ausgesonderten Komplexstrukturen überschneiden sich, so daß die Wirkung des jeweils hinteren Filters gering ist. Nennenswerte Filterwirkung wird durch den der Elektrostatik nachgeschalteten Hydrophobizitätsfilter daher nur bei den Antikörper-Komplexen und beim Subtilisin erreicht.

- Die Entwicklung des Lückenweitefilters ist auf die Analyse der falsch positiven Lösungen beim *bound*-Docking des Antikörper-Lysozym Komplexes 1fdl zurückzuführen. Seine Aufgabe besteht im Wesentlichen darin kleine Interfaces mit guter Paßform größeren Interfaces mit mäßiger Paßform vorzuziehen und somit das Unvermögen der geometrischen Korrelation zwischen beiden zu unterscheiden zu kompensieren. Eine hohe Filterwirkung erzielt er daher auch eher bei großen Komplexen wie dem FAB-Fragment-Komplex 1mlc und dem Methylamin-Dehydrogenase-Komplex 1mda, bei denen große Interfaces mit höherer Wahrscheinlichkeit vorkommen und die kleineren nativen Interfaces mit vielen falsch positiven Lösungen mit geometrischer Korrelation alleine undetektierbar machen. Beim Subtilisin-Komplex 2sni führt die Anwendung des Lückenfilters zum Verwerfen der

letzten Orientierung mit einem RMSD unter 3.5 Å . Dieser Komplex hat eine Lückenweite von 6.7 Å, würde aber mit einem Potentialwert von -2.1 auch von dem nachfolgenden Potentialfilter verworfen.

- Ohne Anwendung des SVM-Klassifikators ist das statistische Potential von Grimm das wirksamste Filterkriterium zur Reduktion falsch positiver Lösungen. Besonders groß ist die Wirkung bei den Antikörper-Testfällen und dem Barnase-Barstar Komplex, bei dem der Potentialfilter die verbliebenen falsch positiven Lösungen um 85% reduziert. Die stark reduzierte Wirkung dieses Filters bei Einsatz der SVM weist darauf hin, daß das Potential im Klassifikator ein hohes Gewicht hat (s. Tabelle 4.8 auf Seite 62 und 4.7 auf Seite 61).

Da alle als Einzelfilter verwendeten Kriterien beim Training in den SVM-Klassifikator eingeflossen sind, ist der Effekt dieser Filter nach der Klassifizierung durch die SVM deutlich geringer als ohne die Klassifizierung. Eine Ausnahme stellt der Lückenweitenfilter bei den Komplexen 1mlc (Antikörper/Lysozym-Komplex) und 1mda (Methylamin-Dehydrogenase/Inhibitor-Komplex) dar. Für beide Testfälle hat der in dieser Arbeit verwendete SVM-Klassifikator nur geringe Spezifität, so daß hier die falsch positiven Lösungen durch den Lückenweitenfilter (nach Anwendung der Filter für Elektrostatik und Hydrophobizität) um weitere 50% (1mda) bzw. 70% (1mlc) reduziert werden können.

5.3 Analyse der Problemfälle

Von den zwölf Komplexen, die als Testfälle für das *unbound* Docking verwendet werden, konnten zwei gar nicht und zwei weitere nur mit mangelhafter Qualität vorhergesagt werden. Es ist daher für die Weiterentwicklung von CKORDO interessant zu klären, warum diese Komplexe schwierig zu docken sind und welche Ansätze anderer Gruppen in diesen Fällen erfolgreich waren.

- **Cytochrom-C Oxidase/Cytochrom-C-Komplex 2pcc**

  Der Testfall 2pcc mit den Einzelstrukturen 1ccp und 2ycc hat mehrere Eigenarten, die ihn für CKORDO zu einer schwierigen Herausforderung

machen. Einerseits hat der superpositionierte Komplex aus 1ccp und 1ycc eine starke Überlappung des Rückgrats von Alanin 193 der Cytochrom-C Oxidase und der Seitenkette des Glutamin 16 des Cytochrom-C mit Atomabständen bis unter 1.0 Å. Andererseits, und dies dürfte der Hauptgrund des Scheiterns sein, befindet sich im Interface des Cytochrom-C eine Häm-Gruppe, die von CKORDO ignoriert wird. Damit entsteht ein Loch im Interface. Zusammen mit einer insgesamt geringen Anzahl von Kontakten im Interface führt dies bei allen Orientierungen mit hinreichender Ähnlichkeit zu der superpositionierten Struktur zu einer sehr geringen geometrischen Komplementarität, die nicht unter das erste bzw. die ersten fünf Korrelationsmaxima fällt. Selbst wenn das geometrische Docking noch positive Lösungen erkennen würde, könnten die nachfolgenden Filter kaum entsprechend reagieren. Das Programm DSSP ignoriert die Häm-Gruppe, wodurch die Werte für den Lückenweitenfilter verfälscht werden. Das statistische Potential hat keine Typzuordnungen für die Atome des Häms (Grimm, 2002) und dürfte daher ebenfalls versagen.

Einige Gruppen haben 2pcc erfolgreich aus den beiden Einzelstrukturen vorhergesagt (Palma *et al.*, 2000; Mandell *et al.*, 2001) während anderen dies nicht gelang (Chen & Weng, 2002). Die Veröffentlichung von Mandell *et al.* beschreibt einen gitterbasierten Fourier-Korrelationsansatz, allerdings unter Verwendung einer aufwendigen Energieberechnung mit Kontinuum-Elektrostatik. Da die Interaktion der Cytochrom-C Oxidase mit Cytochrom-C durch langreichweitige elektrostatische Wechselwirkungen dominiert ist, läßt sich durch diese Maßnahme in Verbindung mit einem sehr feinen Rotationssampling mit 54000 Rotationen in ca. 130 Prozessorstunden eine Lösung finden. Das Verfahren von Palma *et al.* berechnet die Korrelation zweier Gitterrepräsentationen ohne Fourier-Transformation und verwendet dazu schnelle Bitoperationen. Die Filterstufe des Programms besteht aus einem trainierten neuronalen Netz. Hier liegt wahrscheinlich der Hauptgrund für den Erfolg des Ansatzes. Die Trainingsmenge für dieses neuronale Netz umfaßt den Cytochrom-C Oxidase/Cytochrom-C-Komplex des

Pferdes, der sich nur sehr geringfügig von dem hier untersuchten Komplex aus Hefe unterscheidet. Es kann daher nicht beurteilt werden, ob die Methode übertrainiert worden ist und nur deshalb in der Lage war 2pcc zu docken.

- **Subtilisin/Chymotrypsin-Inhibitor-Komplex 2sni**

  Mit 31 Kontakten unter 2.5 Å, die bis zu 0.8 Å erreichen, stellt die überlagerte Struktur von 1sup mit 2ci2 das Programm CKORDO vor ein Problem. Insbesondere Histidin 64 des Subtilisins und die Reste Threonin 58, Methionin 59 und Glutamat 60 des Chymotrypsin-Inhibitors sind für die zu nahen Kontakte verantwortlich, die zu einer gegenseitigen Durchdringung der van der Waals Radien und damit zu hohen Strafwerten bei der geometrischen Korrelation führen. Andere Gruppen konnten mit diesem Maß alleine ebenfalls keine nativen Orientierungen mit akzeptablem Ranking finden (Gabb *et al.*, 1997; Palma *et al.*, 2000; Chen & Weng, 2002).

Rang	Korrelation			L	RMSD	Pot.	T-Max
	G	E	H				
20664	125	4776	**168199**	5.0	2.655	**-3.103**	**3**
40955	109	1961	101972	**9.1**	2.719	**-2.098**	**3**
32586	117	-855	140594	**6.7**	3.010	**-2.116**	1
22938	123	475	**236036**	4.3	3.119	**-3.895**	**3**
26136	121	-1011	**220314**	4.2	3.210	**-3.386**	**2**
15037	130	-7563	120374	5.0	3.252	**-2.908**	**5**
22770	124	1828	89033	**10.5**	3.397	**-2.031**	**3**
31914	117	-5767	**251218**	**6.7**	3.433	-4.030	**4**
27545	120	16839	**345108**	5.0	3.442	-4.867	**5**
16014	129	-3532	**208867**	4.4	3.459	-4.281	**3**

Tabelle 5.2: Positive Lösungen von CKORDO für Subtilisin (1sup)/-Chymotrypsin-Inhibitor (2ci2), fett gedruckte Werte stellen ein Ausschlußkriterium dar, G = geometrische Korrelation, E = elektrostatische Korrelation, H = hydrophobe Korrelation, L = Lückenweite, T-Max = Translationsmaximum, L und RMSD in Å

Der Einsatz eines Desolvatationsterms bei Chen & Weng, bzw. der Elektrostatik bei Gabb *et al.* führte bei diesen Autoren jedoch immer zu detektierbaren positiven Lösungen. Daher ist interessant zu untersuchen, warum keine

der zehn Lösungen, die CKORDO bei Speicherung von fünf Korrelationsmaxima je Rotation gefunden hat, alle Filter passiert. Die positiven Lösungen (Cα-RMSD < 3.5 Å) und ihre filterrelevanten Werte sind in Tab. 5.2 auf der vorherigen Seite dargestellt. Demnach ist nur eine gefundene Lösung innerhalb einer Rotation das globale geometrische Korrelationsmaximum. Dennoch würden auch die zusätzlichen neun Lösungen bei Verwendung des Filters G5 nicht alle Stufen passieren. Der Elektrostatikfilter filtert keine positiven Lösungen aus und der Lückenfilter würde nur vier positive Lösungen verwerfen. Hydrophobizitäts- und Potentialfilter lassen jedoch zusammengenommen keine Lösung passieren: Alle drei vom Potential akzeptablen Lösungen haben einen sehr hohen hydrophoben Korrelationswert. Dies weist darauf hin, daß das statistische Potential von Grimm hydrophobe Interaktionen besonders günstig bewertet. Die starke Korrelation zwischen Hydrophobizität und statistischem Potential bestätigt diese Annahme (s. Abb. 5.1).

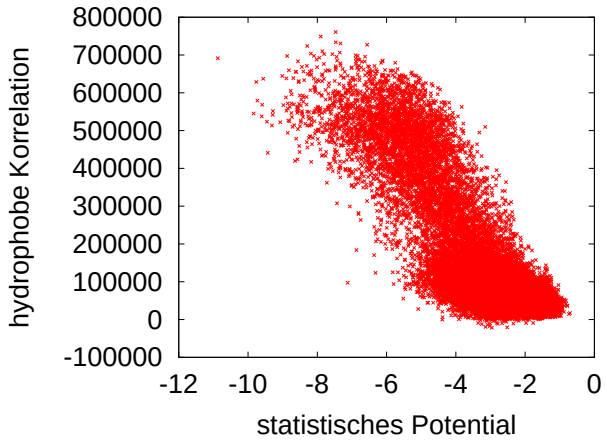


Abbildung 5.1: Korrelation von Hydrophobizität und statistischem Potential beim Docking von Subtilisin (1sup) mit Chymotrypsin-Inhibitor (2ci2)

Eine weniger restriktive Einstellung des Hydrophobozitätsfilters würde zu vielen falsch positiven Lösungen bei Antikörperkomplexen führen. Es erscheint daher sinnvoller erst die Entstehung der außerordentlich hohen

Hydrophobizitätstswerte in der Nähe der superpositionierten Orientierung von 1sup und 2ci2 näher zu untersuchen.

- **Methylamin-Dehydrogenase-Komplex 1mda**

  Der Methylamin-Dehydrogenase-Komplex ist ein Beispiel für die möglicherweise auftretenden Probleme beim Docking von Proteinen, die bei der Komplexbildung größere strukturelle Anpassungen erfahren. Bei einem Interface-RMSD der superpositionierten Strukturen 2bbk und 1aan von 2.69 Å zur Komplexstruktur ist es für ein Programm wie CKORDO, das die Struktur der Proteine als rigide betrachtet, nur schwer möglich, dieses offensichtlich nicht vorgeformte Interface zu detektieren. Um so bemerkenswerter ist daher, daß CKORDO ohne a priori Annahmen über die Lage des Interfaces zu verwenden, eine Orientierung findet, die unter den ersten 1100 Lösungen liegt und mit einem Interface Cα-RMSD von 1.82 Å sehr nah an der nativen Orientierung ist. Palma *et al.* konnten hingegen nur unter Vorgabe der Lage der Bindungsregion auf dem Enzym eine nativ-ähnliche Orientierung finden.

- **Antikörper/Lysozym-Komplex 1mlc**

  Die Problematik den Komplex 1mlc aus den Einzelstrukturen 1mlb und 1lza zusammenzusetzen wird in Palma *et al.* (2000) ausführlich diskutiert. Das wesentliche Problem stellt eine stark veränderte Lage des Arginin 45 im Lysozym gegenüber der Kokristallstruktur dar, so daß es zu einer fast vollständigen Überlappung mit der Seitenkette von Tryptophan 94 im Antikörper kommt. Mit einer minimalen Entfernung von nur 0.85 Å führt dies zu starker Abwertung in der geometrischen Korrelation. Palma *et al.* lösen das Problem, indem sie die Strafwerte für die Seitenketten jenseits von CB für Arginin, Lysin, Aspartat, Glutamat und Methionin auf Null setzen. Diese fünf Aminosäuren sind am häufigsten Ursache für die Abweichungen zwischen Einzelstrukturen und kokristallisierten Strukturen. Ein ähnlicher Ansatz für CKORDO ist in Vorbereitung.

5.4 Stärken und Grenzen von Fourier-Korrelationsmethoden

FFT-Korrelationsverfahren haben eine Reihe von wichtigen Eigenschaften, die sie zu den meistverwendeten Verfahren für die Vorhersage von Proteinkonformationen gemacht haben (Katchalski-Katzir *et al.*, 1992; Vakser & Aflalo, 1994; Meyer *et al.*, 1996; Gabb *et al.*, 1997; Vakser *et al.*, 1999; Mandell *et al.*, 2001). Dazu zählt vor allem die hohe Geschwindigkeit, die es erlaubt auch ohne a priori Informationen über die Lage der Bindungsregionen oder die Zugehörigkeit zu einer bestimmten Proteinklasse gute Ergebnisse erzielen zu können. Die Vielseitigkeit, über die Geometrie hinaus die Korrelationsmaxima anderer Eigenschaften wie Hydrophobizität oder elektrostatisches Potential zur Vorhersage zu verwenden, kennzeichnet die Methode. Die Grenzen des Verfahrens zeigen sich insbesondere bei solchen Fällen, in denen mindestens einer der Bindungspartner größere Konformationsänderungen bei der Komplexbildung erfährt (s. Kap. 5.3 auf Seite 71). Wie bei allen gitterbasierten Verfahren ist die direkte Modellierung von Flexibilität praktisch nicht möglich. Zwar hat die FFT-Korrelation z.B. durch Anwendung der Koeffizientenfilterung (Meyer *et al.*, 1996) oder die spezielle Repräsentation flexibler Aminosäuren (Gabb *et al.*, 1997) eine gewisse Robustheit bezüglich der Konformationsänderungen von einzelnen Seitenketten, Domänenflexibilität ist jedoch außerhalb der Reichweite dieser Maßnahmen. Zudem hat die Stärke, durch Translationsunabhängigkeit ein vollständiges Sampling in kurzer Zeit durchführen zu können als Kehrseite den Nachteil, daß durch Kenntnis der Bindungsstelle auf nur einem Bindungspartner der Suchraum nicht eingeschränkt werden kann. Schließlich ergibt sich durch die Gitterrepräsentation noch das Problem von Samplingfehlern. Zwar scheint in manchen Fällen eine abstrakte, ungenaue Repräsentation der Proteine zur Kompensation vorteilhaft zu sein (Vakser, 1996), die Fehler bei der Gitterrepräsentation führen aber bei unterschiedlichen Rotationen zu unterschiedlichen Erscheinungsbildern derselben Struktur und sind daher nachteilig. Versuche zur Kompensation dieses Effekts für CKORDO sind in Vorbereitung, die Verwendung rationaler Zahlen zwischen Null und Eins je nach Füllgrad einer Zelle führte jedoch zu einer Verschlechte-

rung der Ergebnisse. Für einige der angesprochenen Probleme gibt es vielversprechende Lösungsansätze. Eine besonders interessante Entwicklung ist sicher die Verwendung von Fourier-Korrelation für andere Proteinrepräsentationen. Ritchie & Kemp verwenden eine sogenannte *spherical harmonics*-Repräsentation, die im Gegensatz zur Gitterrepräsentation rotationsinvariant ist und die Verwendung von a priori Information über die Bindungsstelle von nur einem Protein zur Einschränkung des Suchraums verwenden kann (Ritchie & Kemp, 2000). Die Rechenzeiten liegen, obwohl kein der FFT entsprechendes Verfahren für die *spherical harmonics*-Repräsentation verfügbar ist, im Bereich der gitterbasierten FFT-Korrelation. Nachteile dieses Verfahrens sind z.Z. noch die aufwendige Berechnung der Überlappungsintegrale, Schwierigkeiten bei der Repräsentation von Flexibilität und eine zunehmende Ungenauigkeit, je größer die zu dockenden Strukturen sind.

5.5 Alternativen zu Fourier-Korrelationsmethoden

Trotz ihrer Verbreitung sind Fourier-Korrelationsmethoden nur ein Ansatz unter vielen um Protein-Komplexe vorherzusagen. Eine ebenfalls gitterbasierte Methode ohne Verwendung von FFT beschreiben Palma *et al.*. Dabei enthalten die Gitterzellen nur den Wert 0 oder 1 und können somit durch einzelne *bits* im Speicher eines Rechners repräsentiert werden. Über schnelle logische Bitoperationen kann die Randschicht erzeugt werden und das Verfahren erreicht durch intelligente Heuristiken eine Laufzeit von 2-8 Stunden auf einem normalen PC. Einen Schritt weiter geht ein Verfahren das von Fernandez-Recio *et al.* vorgestellt wurde. Zunächst wird eine gitterbasierte Grobsuche durchgeführt, in der als Sampling-Methode ein Monte-Carlo-Verfahren verwendet wird. Im Anschluß erfolgt eine Minimierung der Seitenketten des Liganden-Proteins. Dies führt auch bei solchen Proteinen, deren native Orientierung aufgrund sterischer Hinderung von *rigid body*-Dockingmethoden kaum erfolgreich vorhergesagt werden können zu brauchbaren Ergebnissen. Der Preis dafür ist die Laufzeit. Zwar ist die Grobsuche in vier bis zehn Stunden auf einem PC möglich, die Minimierung benötigt aber je Struktur zehn bis dreißig Minuten. Bei den angegebenen 300 Strukturen

in der Minimierungsstufe bedeutet dies eine Laufzeit von bis zu 150 Stunden.
Ruth Nussinov und Haim Wolfson entwickelten ein Verfahren zum Suchen und Vergleichen von Dreiecken aus sogenannten *critical points* (Lin *et al.*, 1994) . Diese *critical points* entsprechen beim geometrischen Docking den Nullstellen der ersten Ableitung einer Proteinoberfläche, also den Apices der Ein- und Ausbuchtungen, es gibt jedoch auch die Möglichkeit den *critical points* physikochemische Parameter zuzuordnen. Die Vorteile dieser Methode liegen in der hohen Geschwindigkeit (wenige Minuten) und darin, daß es kein Sampling gibt, wodurch eine Lösung mit einer guten Korrelation aber sehr steilen Korrelationsflanken übersehen werden könnte. Die Herausforderung dieses Verfahrens besteht darin, die einzelnen Lösungen geeignet zu clustern, wodurch die Laufzeit eine ähnliche Größe erreicht wie FFT-basierte Methoden. Erweiterungen des ursprünglichen Verfahrens sind in der Lage auch scharnierartige Domänenflexibilität mit einzubeziehen (Sandak *et al.*, 1998).
Den Einsatz genetischer Algorithmen unter Verwendung der *critical points* Repräsentation beschreiben Gardiner *et al.*. Genetische Algorithmen optimieren Lösungen durch ein Verfahren das aus Mutation, Selektion und Rekombination von Lösungsvektoren (Chromosomen) besteht, die in diesem Fall die Orientierung der Proteine zueinander kodieren. Als Terme der Bewertungsfunktion für das Selektionsverfahren werden die Antiparallelität der Oberflächennormalen, die Komplementarität der Oberflächenform (*critical points*), und ein Wasserstoffbrücken-Potential eingesetzt. Die Laufzeit des Verfahrens beträgt zwischen 3.5 und 33.5 Stunden auf einer Workstation.

5.6 Ausblick

5.6.1 Weiterentwicklung von CKORDO

CKORDO besitzt neben den in dieser Arbeit beschriebenen Elementen eine Reihe von weiteren Eigenschaften, die bei der Vorhersage nützlich sind. So lassen sich Rotationen vorgeben um beispielsweise die Korrelationslandschaft in der Umgebung der nativen Orientierung zu untersuchen, sowie die Möglichkeit fast alle besprochenen Parameter einzustellen ohne den Programmcode ändern zu

müssen. Die Möglichkeiten zur Weiterentwicklung von CKORDO sind vielfältig. Dies betrifft sowohl die bereits im Programm implementierten Methoden als auch die Ergänzung des Programms zur Verfeinerung der Ergebnisstrukturen.

5.6.2 Proteinrepräsentation

Noch immer gibt es Probleme beim Docking von Einzelproteinen, die eine Überlappung von Aminosäuren zeigen. Mit einem geeigneten Torsionswinkelpotential ließen sich die Seitenketten in eine Position bringen, die derjenigen im Komplex näher kommt als die Ausgangsstruktur. Einige Veröffentlichungen zeigen in diesem Gebiet große Fortschritte (Xiang & Honig, 2001). Zusätzlich sollte sich durch Analyse von möglichst vielen Docking-Testfällen eine Heuristik finden lassen, die die Erkennung kritischer Seitenketten und ihre Gewichtung im Inneren des Proteins (Palma *et al.*, 2000) oder in der Randschicht geeignet verringert. Der Ansatz von Palma *et al.* (2000) behält die Korrelationsschicht aller Aminosäureatome bei, setzt aber jenseits von CB den Wert für die inneren Gitterzellen der Aminosäurereste Arginin, Lysin, Aspartat, Glutamat und Methionin auf Null, statt auf einen negativen Wert. Diese Aminosäuren weisen die größten Differenzen zwischen und unkomplexierter Struktur auf und können so nicht mehr zur Abwertung führen, wenn sie in ungünstiger Orientierung das andere Protein penetrieren. Ein zweites Problemfeld der Proteinrepräsentation liegt in der Vernachlässigung von allen Atomen die nicht zu den Polypetidketten gehören wie Kofaktoren oder Zuckerreste. Hier müßten entsprechende Erweiterungen der Einleseroutinen erfolgen und auch die Werte für die van der Waals Radien und Ladungen ergänzt werden. Schließlich wurden erste Versuche zur Kompensation des Samplingfehlers unternommen.

5.6.3 Elektrostatik

Den bisherigen Erkenntnissen zufolge werden bei einigen Testfällen bessere Rankings unter Einbeziehung der Elektrostatik erhalten. Während der Entstehungszeit dieser Arbeit haben fast alle auf dem Gebiet des Protein-Docking arbeitenden Gruppen die Elektrostatik als Erweiterung bestehender geometrischer

Methoden in die Bewertung von Docking-Lösungen eingebaut. Dabei wurde von allen Gruppen, die mit gitterbasierten FFT-Korrelationsmethoden arbeiten, die Elektrostatik direkt auf dem Gitter als Korrelationswert berechnet, ebenso wie in dieser Arbeit geschehen. Die einzelnen Implementationen unterscheiden sich jedoch sowohl bezüglich der Berechnung der Elektrostatikwerte für das Gitter, als auch in der Art, wie diese Werte in die Bewertung einbezogen werden.
Sowohl bei Gabb *et al.* als auch in den Arbeit von Chen & Weng wird die elektrostatische Korrelation durch ein separates Gitter mit FFT-Korrelation berechnet, während in der Arbeit von Heifetz *et al.* die Elektrostatik im Imaginärteil einer komplexen Zahl kodiert wird, deren Realteil die geometrische Repräsentation darstellt. In dieser Arbeit wird der Ansatz von Gabb *et al.* verfolgt, da die Methode von Heifetz *et al.* durch die Multiplikation in der Korrelationsberechnung zu Summationstermen führt, die Geometrie und Elektrostatik gemischt enthalten. Daher muß die Gewichtung zwischen Geometrie und Elektrostatik vor der Fourier-Korrelation vorgenommen werden. Während der Entwicklung von CKORDO erschien es sinnvoller, die Elektrostatikberechnung von der Geometrieberechnung zu trennen. Dies hat den Vorteil verschiedene Gewichtungen ohne Neuberechnung der Korrelation austesten zu können. CKORDO verwendet eine einfache Coulomb-Funktion zur Berechnung des Feldes des stationären und des rotierten Proteins. Zumindest für das stationäre Protein ließe sich eine aufwendigere Berechnung wie eine Lösung der linearisierten Poisson-Boltzmann-Gleichung durchführen ohne die Gesamtrechenzeit des Programms wesentlich zu beeinflussen. Eine besonders elegante Methode zur Implementation um auch für das rotierte Protein das elektrostatische Potential nur einmal berechnen zu müssen beschreiben Heifetz *et al.*. Dazu wird das mit dem Programm DELPHI berechnete Potential aus dem Gitter in Kugelform diskretisiert. Diese Potentialkugeln werden dann analog zu den Atomkoordinaten der Proteine rotiert und auf das in der Regel etwas gröbere Gitter für die Fourier-Korrelation abgebildet.

5.6.4 Hydrophobizität

Die Hydrophobizität wird auf ähnlichem Weg berechnet wie die Elektrostatik. Die Ergebnisse unterscheiden sich jedoch deutlich. Wichtigste Gemeinsamkeit ist,

daß große absolute Werte auf nicht-native Orientierungen hinweisen. Diese Beobachtung ist mit Ausnahme des bereits besprochenen Subtilisin/Chymotrypsin-Inhibitor Komplexes (2sni) bei allen Testfällen gemacht worden. Sie steht damit in Widerspruch zu Vaksers Ansatz des „hydrophoben Dockings“(Vakser & Aflalo, 1994). Die Autoren dieser Arbeit gehen davon aus, daß es im wesentlichen hydrophile Kontakte sind, die zu falsch positiven Lösungen führen. Die beschriebene Methode verwendet keine atomaren Hydrophobizitäten, sondern ordnet in der Gitterrepräsentation nur dem Teil der Randschicht einen von Null verschiedenen Wert zu, der durch Abbildung einer hydrophoben Aminosäure zustande kommt. Insofern stehen die Beobachtungen dieser Arbeit und der von Vakser et al. nicht im Widerspruch. Ein *bound-unbound*-Testfall des β-Trypsins aus dem Komplex 2ptc mit der Trypsin-Inhibitor-Struktur 4pti wird in der zitierten Arbeit untersucht. Sie erreicht einen Rang von 5394 bei Verwendung diverser Filter. Bei Betrachtung der reinen geometrischen Korrelation wird von ihnen jedoch keine nativ-ähnliche Struktur unter den ersten 7000 Lösungen gefunden (Vakser & Aflalo, 1994). Mit CKORDO hingegen wird bei dem entsprechenden *unbound*-Testfall (2ptn/4pti) nur mit geometrischer Korrelation bereits Rang zwei erreicht. Das bei der Diskussion von 2sni angesprochene Problem (s. Kap. 5.3 auf Seite 71) der Wechselwirkung zwischen Potentialfilter und Hydrophobizitäts-Filter könnte darauf hinweisen, daß der Potentialverlauf zu steil für eine Korrelationsberechnung ist, bei der die Potentialwerte beider Proteine multipliziert werden.

5.6.5 Geometrische Eigenschaften von Interfaces

Das Kriterium der Lückenweite zeigt zwar eine sehr gute Filterwirkung, die Berechnung des Lückenvolumens durch das Programm SURFNET verzehnfacht jedoch die Laufzeit des Programms auf bis zu 70 Stunden. SURFNET berechnet weitaus mehr als nur das Lückenvolumen zwischen den beiden Komplexteilen (Laskowski, 1995). Diese Berechnungen werden für CKORDO nicht benötigt. Es erscheint daher sinnvoll die Berechnung des Lückenvolumens neu zu implementieren. Auch für die Flächenberechnung ließe sich statt des Programms DSSP (Kabsch & Sander, 1983) eine Routine einsetzen, die Atome aus Kofaktoren und

Zuckern berücksichtigt.

5.6.6 Statistisches Kontaktpotential

Die bereits bei der Hydrophobizität in Abschnitt 5.6.4 auf Seite 80 diskutierten Probleme bei der Anwendung des Potentials auf hydrophobe Interfaces haben ihre Ursache möglicherweise darin, daß das verwendete Potential nur eine von zwei additiven Komponenten aus der Originalarbeit darstellt und daher Desolvatationseffekte nicht berücksichtigt (Grimm, 2002). Interessant wäre ein Vergleich der Filterleistungen mit anderen Kontaktpotentialen, die explizite Desolvatationsterme enthalten wie die von Chen & Weng eingesetzten Atom-Kontakt-Energien (ACE) von Zhang & Skolnick (1998).

5.6.7 *Support Vector Machines*

Die Anwendung von Methoden des algorithmischen Lernens für Klassifizierungsaufgaben ist naheliegend. Bisher sind davon für das Protein-Docking ausschließlich Neuronale Netze zum Einsatz gekommen. Diese Arbeit zeigt erstmals die Verwendung von *Support Vector Machines* zur Unterscheidung von nativen und nicht-nativen Protein-Komplexen. Die Möglichkeiten in diesem Bereich können durch die vorliegende Arbeit nur angedeutet werden. Ein wesentlicher Schwachpunkt, der Entwicklung guter Klassifikatoren betrifft das Training. Zur Zeit gibt es nur wenige Proteinkomplexe, deren Einzelteile als getrennte Strukturen zur Verfügung stehen. Dabei sind zwei Proteinklassen innerhalb dieser Testfälle dominierend: Serinprotease-Inhibitor-Komplexe und Antikörper-Antigen-Komplexe. Für die Entwicklung eines allgemein anwendbaren Klassifikators ist dies nicht ausreichend. Nähere Untersuchungen müssen zeigen ob es sinnvoller ist Klassifikatoren speziell für solche einzelnen Proteinklassen zu entwickeln. Auch die Frage welche Parameter für das Training eines Protein-Komplex-Klassifikators herangezogen werden sollten, bleibt weiteren Untersuchungen vorbehalten. Schließlich gibt es auch noch eine Reihe weiterer Parameter wie die Wahl einer geeigneten Skalierung der Daten, sowie der Kernelfunktion.
Die SVM-Klassifikation und die verschiedenen Filter sollten dann in das eigent-

liche Dockingprogramm CKORDO integriert werden.

5.6.8 Verfeinerung und Minimierung

Das Vorläuferprogramm KORDO enthält einen dreistufigen Prozeß zur Vorhersage der Komplex-Strukturen. Außer der in dieser Arbeit von CKORDO verwendeten Stufe eines regelmäßigen Samplings, wird bei KORDO mit den besten Orientierungen in einer zweiten Runde in der Umgebung der ursprünglichen Rotationswinkel nach Lösungen gesucht. Die letzte Stufe verfeinert diese Suche durch ein *simulated annealing*-Verfahren, um das jeweilige lokale Maximum der Korrelationsfunktion zu finden. Während die zweite Stufe in einer vorläufigen Version bereit in CKORDO implementiert wurde, steht die dritte Stufe noch aus. Die Entwicklung einer Minimierungsstufe schließlich, welche die Seitenkettenpositionen verändert, würde zusammen mit der Entwicklung einer entsprechend optimierten Potentialfunktion ein eigenes Dissertations-Projekt darstellen.

5.7 Zukünftige Entwicklungen

In den letzten Jahren werden weltweit Anstrengungen unternommen, die Bestimmung dreidimensionaler Strukturen von Proteinen durch Einsatz von Robotern von der Proteinreinigung bis zum Einspannen des Proteinkristalls in die Röntgenapparatur um ein Vielfaches zu beschleunigen. Dennoch steigt die Anzahl der bekannten Proteinsequenzen sehr viel schneller als die Anzahl der Strukturen. Ein großes Interesse besteht daher einerseits an Verfahren, die Proteinstrukturen aus verwandten Strukturen mit hoher Genauigkeit vorhersagen können, andererseits an Docking-Programmen, die gegenüber den verbleibenden strukturellen Ungenauigkeiten tolerant genug sind, um auch mit Modellstrukturen verläßliche Komplexvorhersagen treffen zu können.
Eine weitere Entwicklung der letzten Jahre betrifft den Bereich Datenintegration, der versucht, Informationen über Proteine miteinander zu vernetzen. Für die Eignung von Dockingprogrammen für die praktische Anwendung wird es von Bedeutung sein, ob sich experimentelle Befunde für die Beurteilung von Protein-Komplex-Vorhersagen einsetzen lassen. Die Art dieser experimentellen Daten ist

sehr heterogen und reicht von der Sequenzverwandtschaft bis zu Ergebnissen von Mutationsanalysen zur Identifizierung einzelner Interfacereste. Zur Vorhersage der Struktur großer Funktionskomplexe (Ribosom, Oxidoreduktasekomplexe, Komplementrezeptor), sowie von Mikrotubuli und anderen Gerüstelementen der Zelle werden in den entsprechenden Dockingprogrammen bereits jetzt elektronenmikroskopische Aufnahmen als Zusatzwissen verwendet (Chacon & Wriggers, 2002).
Protein-Docking wird vielleicht in Zukunft helfen die vielen Daten zu filtern, die durch die in Abschnitt 1.4 auf Seite 5 beschriebenen experimentellen Verfahren erhoben worden sind und viele falsch positive Interaktionen enthalten. Dem Ziel einer vollständigen Beschreibung des Protein-Interaktionsnetzwerks einer Zelle wäre die Wissenschaft damit einen großen Schritt näher.

6 Zusammenfassung

Diese Arbeit beschreibt die Entwicklung des Programms CKORDO zur Vorhersage von Protein-Protein-Komplexen aus den Strukturdaten der einzelnen Komplexbestandteile. Dieses unter dem Begriff Docking bekannte Verfahren unterscheidet zwei Klassen von Testfällen: *bound* Docking, bei dem die zusammenzusetzenden Strukturen aus den Daten eines kokristallisierten Komplexes stammen und das schwierigere *unbound* Docking, bei dem die Strukturen der Einzelproteine unabhängig voneinander bestimmt worden sind.

CKORDO beinhaltet eine Neuimplementation von Teilen des Protein-Docking-Programms KORDO, das Michael Meyer in der Arbeitsgruppe von Prof. D. Schomburg entwickelt hat (Meyer *et al.*, 1996). Beide Programme basieren auf Arbeiten zur Fourier-Korrelationen diskretisierter Proteinstrukturen (Katchalski-Katzir *et al.*, 1992). Bei diesem Verfahren wird das erste Protein so auf ein Gitter abgebildet, daß das Proteininnere mit negativen Werten und die Randschicht mit positiven Werten belegt wird. Das zweite Protein wird mit positiven Werten besetzt und besitzt keine Randschicht. Bei der Korrelation beider Gitterdarstellungen entstehen positive Werte, wenn das zweite Protein mit der Randschicht überlappt, negative Werte, wenn das zweite Protein mit dem Inneren des ersten Proteins überlappt und ansonsten Nullwerte. Der Vorteil des von Katchalski-Katzir eingeführten Verfahrens liegt darin, daß die Berechnung der Korrelationsfunktion aus den Fouriertransformationen der beiden Proteine um ein vielfaches schneller ist als die direkte Berechnung. Die Transformation der diskreten Proteindarstellungen in den Fourier-Raum läßt sich durch eine *Fast Fourier Transformation* (FFT) genannte Methode ebenfalls sehr effizient durchführen, so daß sich der Gesamtrechenaufwand für die Korrelation bei einer Gittergröße von $N \times N \times N$ Zellen von N^6 auf $N^3 log(N^3)$ reduziert, wobei N in der Regel zwischen sechzig und neunzig liegt. Das zweite Protein wird in regelmäßigen Winkelintervallen rotiert, abgebildet und jeweils ein bis fünf Maxima der Korrelationsfunktion für jede Rotation ausgegeben. Bei einem Winkelinkrement zwischen zwölf und zwanzig Grad werden so zwischen 2000 und 40000 Orientierungen erhalten

und absteigend nach ihren Korrelationswerten sortiert. Die Lösungen mit den höchsten Korrelationswerten entsprechen häufig den nativen Orientierungen der Proteine im kokristallisierten Komplex. Beim *unbound* Docking ist die geometrische Korrelation der aus Einzelstrukturen zusammengesetzten Komplexe in der nativen Orientierung durch einander sterisch behindernde Seitenketten oft nicht ausreichend, so daß diese richtigen Lösungen ein schlechteres Ranking haben als andere sogenannte „falsch positive“ Lösungen, die keine Ähnlichkeit mit dem gesuchten Komplex haben aber eine gute geometrische Korrelation aufweisen. Anhand einer Reihe von *bound* Docking-Tests wurden die Parameter der geometrischen Korrelation optimiert. Eine Gitterauflösung von 1.5 Å und eine Randschichtdicke von 1.75 Å erweisen sich als geeignete Parameter. Der Wert für die Korrelationsschicht und das zweite Protein wird auf 1.0, das Innere des ersten Proteins auf -6.0 eingestellt und für das *unbound* Docking sorgt eine Filterung der FFT-Koeffizienten für eine Glättung der Proteindarstellung im Fourier-Raum. In einer Anwendungsstudie an Cytochrom-C Oxidase (2occ) wird das Docking von Substrukturen bis hin zu einzelnen Helices untersucht. Das Docking gelingt wenn eine Mindestinterfacegröße von ca. 900 Å vorhanden ist. Durch künstliches Verändern der Lage von einzelnen Seitenketten zeigt sich allerdings, daß die Toleranz gegenüber Strukturabweichungen sehr gering ist. CKORDO implementiert zur Verbesserung der Vorhersageleistung insbesondere beim *unbound* Docking eine Reihe von weiteren Verfahren:

- die Korrelation eines elektrostatischen Potentials
- die Korrelation eines hydrophoben Potentials
- die Bestimmung des mittleren Abstands der beiden Proteine an ihrer Kontaktfläche (Lückenweite)
- die Bestimmung einer Pseudoenergie über ein abstandsabhängiges Atom-Atom-Kontaktpotential

Die mit diesen Verfahren bestimmten Werte werden verwendet, um eine Reihe von Ausschlußkriterien zum Filtern von Lösungen zu bestimmen und um einen *Support Vector Machines* genannten Klassifikationsalgorithmus zu trainieren. Der

Test von CKORDO beim *unbound* Docking wird anhand von zwölf Komplexen durchgeführt: vier Protease/Inhibitor-Komplexe, drei Antikörper/Lysozym-Komplexe, zwei Acetylcholin-Esterase/Fasciculin-Komplexe, die Komplexe von Barnase/Barstar und Methylamin-Dehydrogenase mit ihren Inhibitoren, sowie der Komplex aus Cytochrom-C Oxidase mit Cytochom-C. Der Test zeigt, daß CKORDO in der Lage ist, für die Mehrzahl der Komplexe Vorhersagen unterhalb eines RMSD von 3.5 Å mit guten bis zufriedenstellenden Rankings zu treffen: Bei Einsatz aller Filter erreichen ohne SVM-Klassifikator vier Komplexe einen Rang unter hundert, vier weitere unter tausend. Drei Komplexe können nicht vorhergesagt werden. Bei Einsatz des SVM-Klassifikators werden etwas bessere Werte erreicht: Für fünf Komplexe ist der Rang unter hundert, für weitere drei unter tausend und nur zwei Proteine können nicht vorhergesagt werden. Im Vergleich mit anderen Arbeiten zum Protein-Docking schneidet CKORDO unter Berücksichtigung der Tatsache, daß keinerlei a priori Annahmen über die Lage der Bindungsstelle verwendet werden, sehr gut ab.
Stärken und Grenzen der Fourierkorrelation, sowie alternative Ansätze zum Protein-Docking werden in der Diskussion dargestellt. Die Gründe für die Schwierigkeiten beim Docking von vier Komplexen werden detailliert analysiert. Aus der Analyse werden Möglichkeiten zur weiteren Verbesserung von CKORDO entwickelt und die Verbindungen zwischen Protein-Docking, Proteinstrukturvorhersage und Datenintegration als Vorraussetzung für praktische Anwendungen von Docking-Programmen aufgezeigt.

A Anhang

A.1 Programmiertechnische Hilfsmittel

Softwareentwicklung

Aus verschiedenen Gründen sollte das bestehende Programm KORDO in eine modernere Programmiersprache umgeschrieben werden. Insbesondere der Wunsch das Programm auf freien Betriebssystemen entwickeln und laufen lassen zu können und die schlechte Unterstützung für Fortran auf diesen Systemen führte zur Wahl von C/C++ als neuer Implementierungssprache.

Betriebssysteme

- **Linux:** Als freies Betriebssystem wurde Linux von der Firma RedHat (Version 5.0, 5.1, 5.2, 6.0), davon abgeleitete Distributionen der Firma Lehmann (Helloween II + III), sowie die Version 7.0 der Firma SuSe GmbH auf handelsüblichen PCs eingesetzt. Die PCs waren mit 256 MByte Hauptspeicher ausgerüstet.

- **IRIX:** Die Workstations der Firma Silicon Graphics (jetzt SGI) laufen unter dem Betriebssystem IRIX. Es wurden die Versionen 6.2, 6.3, 6.4 und 6.5 auf Indigo II, O2 und Octane Workstations sowie einer Power Challenge eingesetzt. Der Hauptspeicher dieser Workstations war auf 512 bis 768 MByte ausgebaut.

Programmiersprachen

- **Fortran:** Zur Kompilierung von KORDO und den späteren Ausgliederungen der FFT- Routinen unter IRIX wurde der f77 Compiler von Silicon Graphics eingesetzt. Seit Ende 2000 war auch eine Version des g77 Fortran Compilers der GNU Foundation in der Lage die FFT-Routinen unter Linux zu kompi-

lieren. Der g77 Compiler Version 0.5 wurde für alle neueren Versionen von CKORDO eingesetzt.

- **C/C++:** Da die FFT-Routinen als Fortran Code gelinkt werden mußten, wurde bis Ende 2000 nur mit dem Compiler CC unter IRIX gearbeitet. Später wurden auch Tests mit den freien C/C++ Compilern (g++, gcc Versionen 2.72, 2.95, egcs 1.0, 1.1) unter IRIX und Linux durchgeführt.

- **python:** Python ist eine Skriptsprache, die auch objektorientierte Konzepte enthält. Skriptsprachen finden insbesondere bei kleineren Programmen, die keine aufwendigen Rechnungen durchführen Verwendung. Python wurde im Rahmen dieser Arbeit zur Vor- und Nachbearbeitung von Daten eingesetzt, insbesondere auch zur Vorbereitung der Visualisierung von Ergebnissen.

- **perl:** Zur Bearbeitung von Textlisten, z.B. zur Zusammenstellung von Testdatensätzen oder zur Vorbereitung von Input- oder Nachbereitung von Outputlisten wurde anfänglich die Skriptsprache perl in der Version 5 eingesetzt. Darüber hinaus sind fast alle im Rahmen von Etagenpraktika erstellten Skripte zur Analyse von Interfaces in dieser Programmiersprache erstellt, da sie auf einfache Weise die Einbindung von anderen Programmen ermöglicht. Später wurde zu diesen Zwecken verstärkt Python eingesetzt.

Softwareentwicklungswerkzeuge

Für die verschiedenen Aufgabenstellungen beim Entwickeln und Testen von CKORDO wurden Standardwerkzeuge der Softwareentwicklung eingesetzt: ftncheck zur Analyse des Programmablaufs in KORDO, f2c um die Fortranroutinen aus KORDO automatisch in C zu übersetzen. Zur Fehleranalyse von Fortran und C/C++-Programmcode unter IRIX wurde der Debugger cvd von Silicon Graphics eingesetzt, unter Linux der gnu Debugger gdb der Free Software Foundation. Weitere Werkzeuge waren xemacs als Editor für den Programmcode, kdevelop als integrierte Programmierumgebung, CVS als Versionskontrollsystem, sowie das Datenbanksystem Access Version 97 und 2000 von Microsoft unter Windows und das für Linux frei erhältliche Datenbanksystem mysql.

Visualisierung

- **gnuplot:** gnuplot ist ein von Thomas Williams et al. geschriebenes Programm zum Anzeigen von Funktionen und Meßwerten. Alle Diagramme dieser Arbeit wurden mit gnuplot 3.7 erstellt.
- **vtk:** Der Visualization Tool Kit ist eine komplexe Bibliothek mit Visualisierungsroutinen, die für verschiedene Programmiersprachen vorliegen. vtk wurde für alle Gitterabbildungen eingesetzt.
- **BRAGI** BRAGI ist ein an der GBF Braunschweig entwickeltes Molecular Modelling Programm. Neben der grafischen Darstellung von Proteinstrukturen wurde das Programm auch zum Superpositionieren von Strukturen eingesetzt.

A.2 Hydrophobizitätsdaten

Die Daten für die aminosäurebasierte Hydrophobizitätsberechnung wurde unterschiedlichen Skalen entnommen:

Kyte-Doolittle Skala

Die Kyte-Doolittle Skala (Kyte & Doolittle, 1982) ist die bekannteste Skala von Hydrophobizitätswerten für Aminosäuren (s. Tabelle A.1). Sie basiert wie viele andere Skalen auf der Abschätzung der freien Energie bei der Überführung einer Aminosäure von einer wässrigen in eine hydrophobe Umgebung. Die 1982 publizierte Skala wurde aus experimentell bestimmten Löslichkeiten von Aminosäuren in Ethanol gewonnen (Tanford, 1970).

Aminosäure	Wert	Aminosäure	Wert
A	1.8	M	1.9
C	2.5	N	-3.5
D	-3.5	P	-1.6
E	-3.5	Q	-3.5
F	2.8	R	-4.5
G	-0.4	S	-0.8
H	-3.2	T	-0.7
I	4.5	V	-4.2
K	-3.9	W	-0.9
L	3.8	Y	-1.3

Tabelle A.1: Die Kyte-Doolittle Skala

Cornette Skala

Die Cornette-Skala (Cornette et al., 1987) ist als gewichteter Mittelwert aus mehreren anderen Skalen ermittelt worden und in Tabelle A.2 zu erkennen.

Aminosäure	Wert	Aminosäure	Wert
A	0.2	M	4.2
C	4.1	N	-0.5
D	-3.1	P	-2.2
E	-1.8	Q	-2.8
F	4.4	R	1.4
G	0.0	S	-0.5
H	0.5	T	-1.9
I	4.8	V	4.7
K	-3.1	W	1.0
L	5.7	Y	3.2

Tabelle A.2: Die Cornette Skala

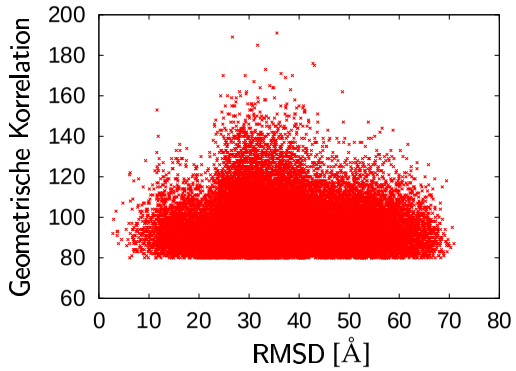
Boyko Skala

Die Boyko Skala wurde von einem PepTool Autor (R. Boyko) auf Basis von Proteinstrukturdaten aus der PDB berechnet. Sie wurde zur Identifikation von coil Regionen optimiert und sollte nicht auf Transmembranhelices ansprechen. Sie hat gegenüber den Skalen von Kyte und Cornette ein umgekehrtes Vorzeichen, stellt also eher eine “Hydrophilizitätsskala” dar.

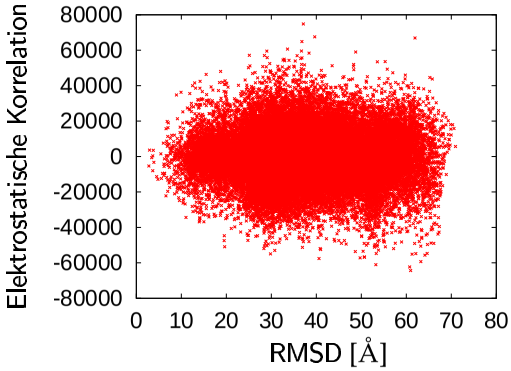
Aminosäure	Wert	Aminosäure	Wert
A	-0.4	M	-8.5
C	-0.5	N	1.2
D	6.2	P	3.0
E	3.6	Q	6.4
F	-8.8	R	-1.8
G	0.0	S	0.9
H	2.4	T	3.8
I	-9.5	V	-9.7
K	6.1	W	-2.0
L	-10.6	Y	-7.4

Tabelle A.3: Die Boyko Skala

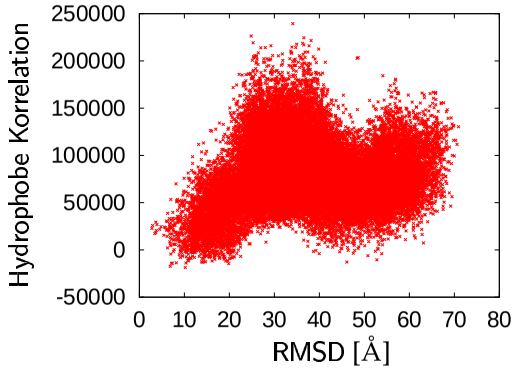
A.3 Ergebnisse für 12 *unbound*-Dockings



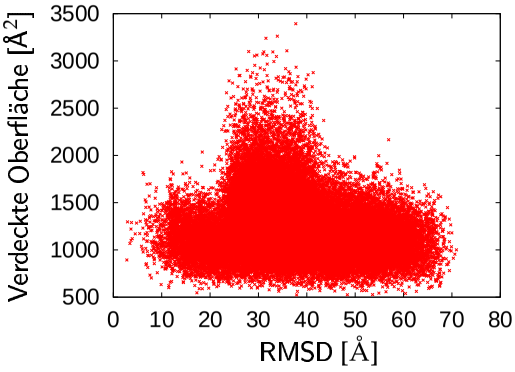
(a) geometrische Korrelation



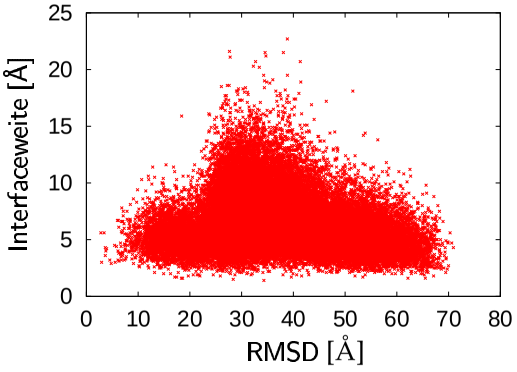
(b) elektrostatische Korrelation



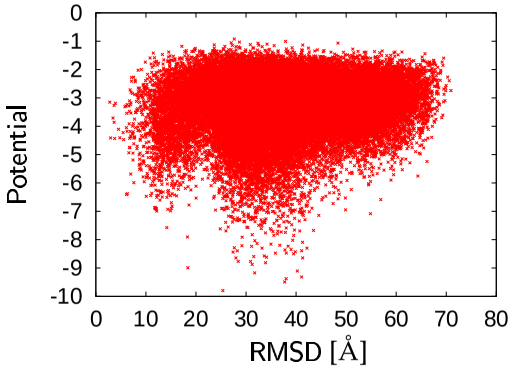
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche

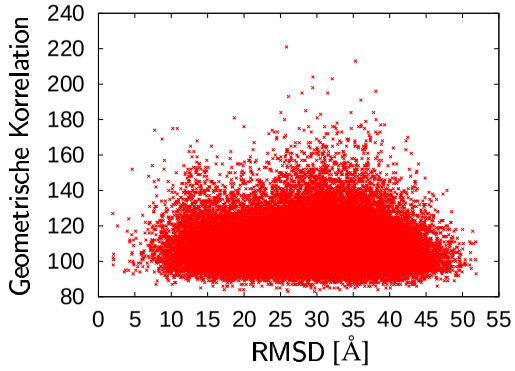


(e) Lückenweite

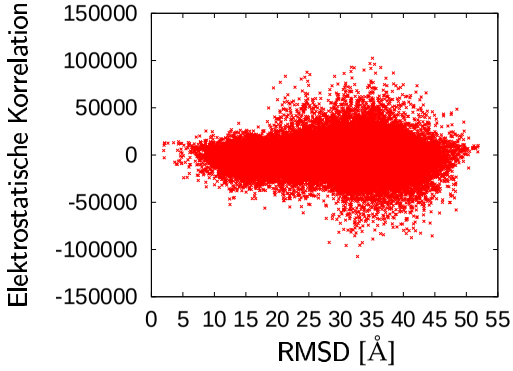


(f) statistisches Potential

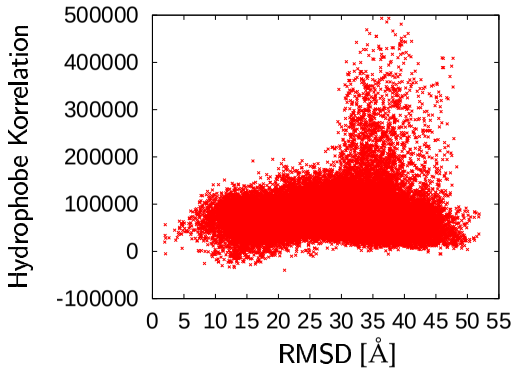
Abbildung A.1: FAB Fragment des Antikörpers D44.1 (1mlb) im Komplex mit Hühnerei-Lysozym (1lza), Referenzkomplex 1mlc verlegt wird.



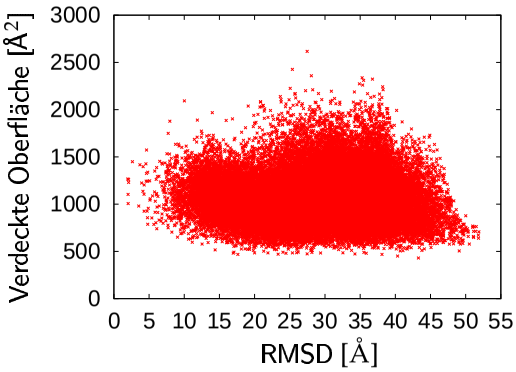
(a) geometrische Korrelation



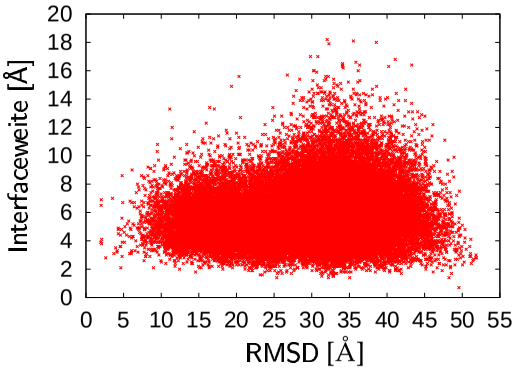
(b) elektrostatische Korrelation



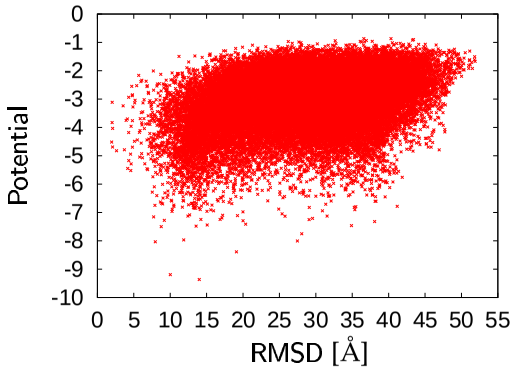
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche

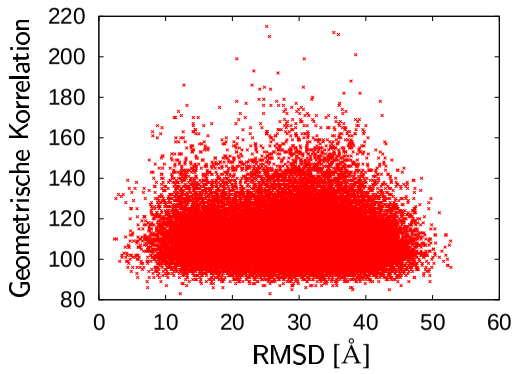


(e) Lückenweite

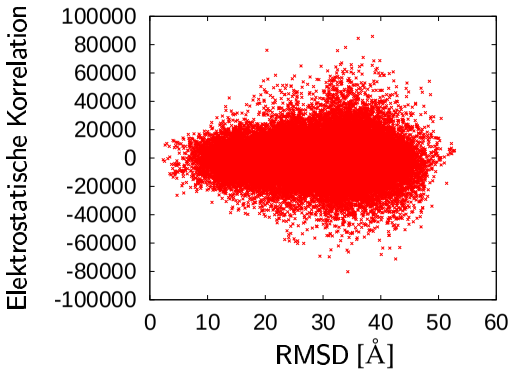


(f) statistisches Potential

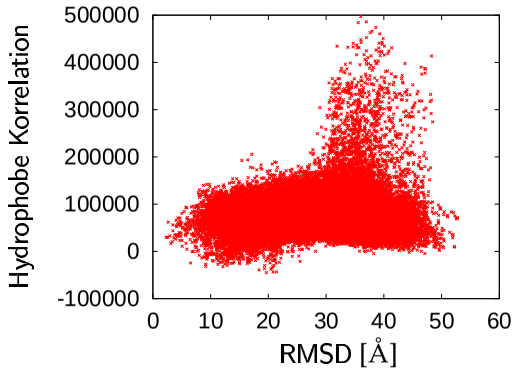
Abbildung A.2: FV Fragment des Antikörpers D1.3 (1vfa) im Komplex mit Hühnerei-Lysozym (1hel), Referenzkomplex 1vfb



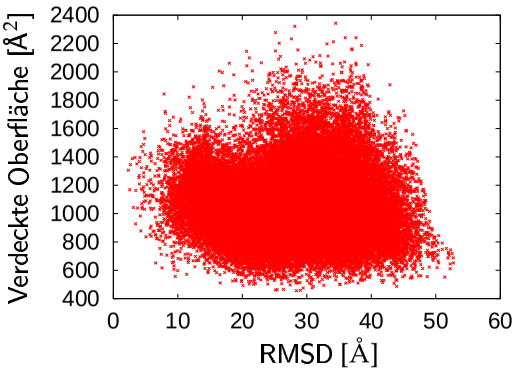
(a) geometrische Korrelation



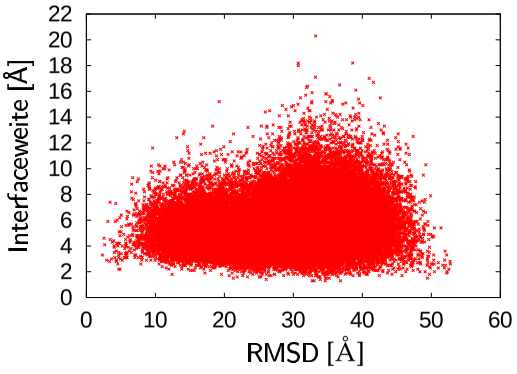
(b) elektrostatische Korrelation



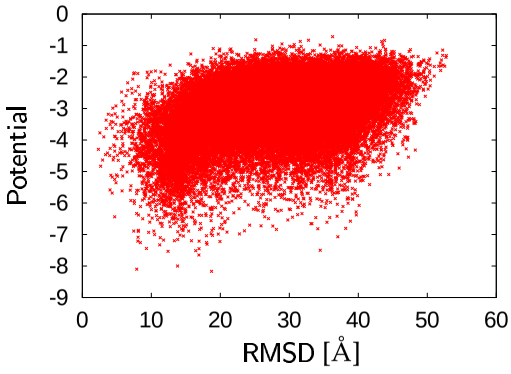
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche

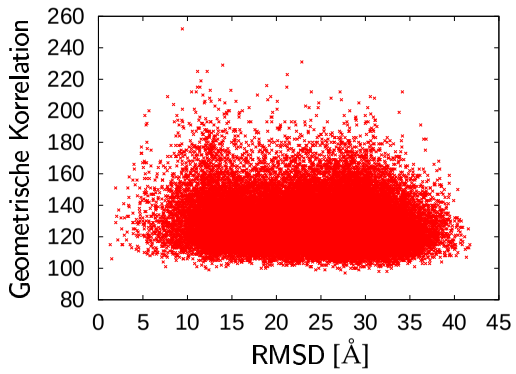


(e) Lückenweite

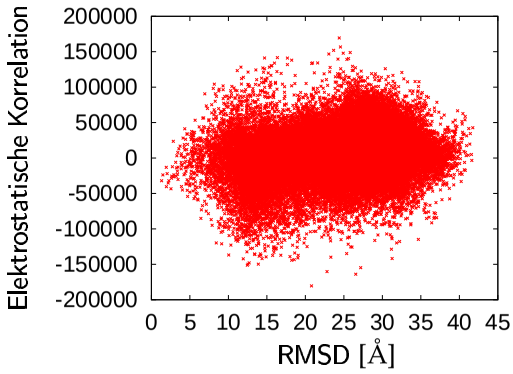


(f) statistisches Potential

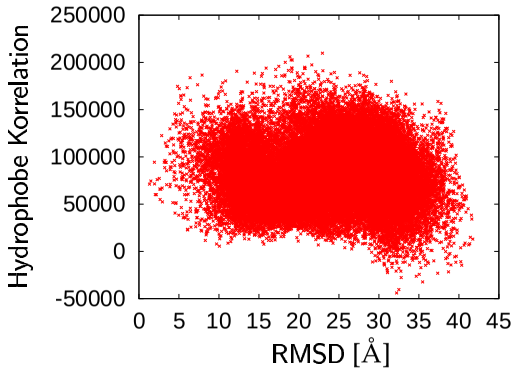
Abbildung A.3: FV Fragment des Antikörpers D1.3 (1vfa) im Komplex mit Hühnerei-Lysozym (1lza), Referenzkomplex 1vfb



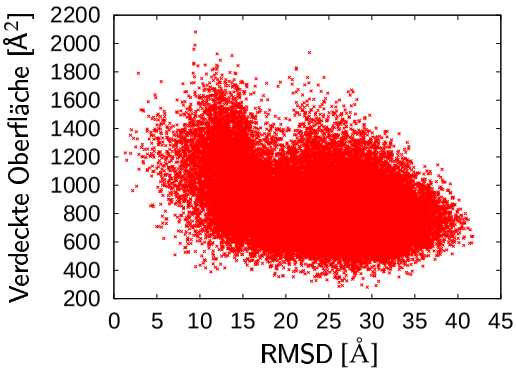
(a) geometrische Korrelation



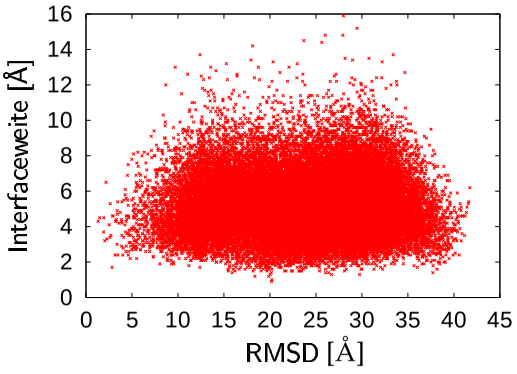
(b) elektrostatische Korrelation



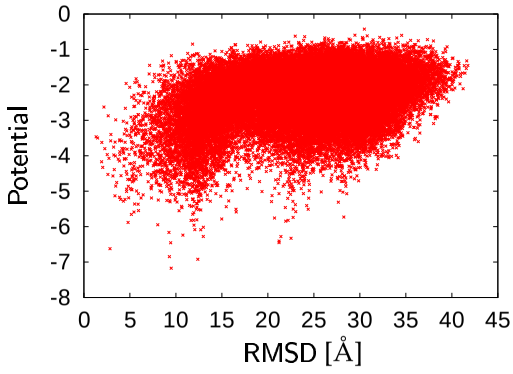
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche

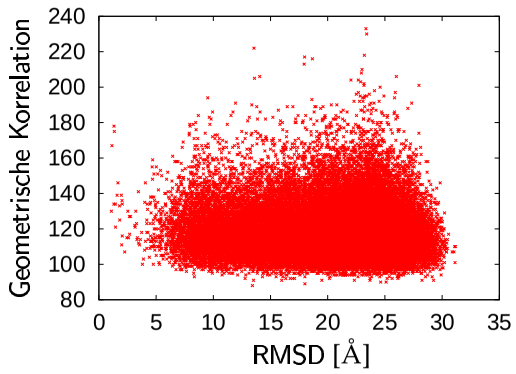


(e) Lückenweite

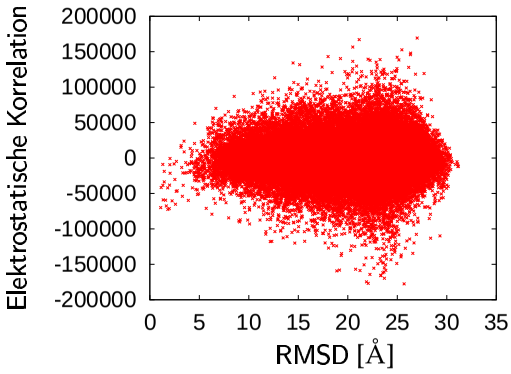


(f) statistisches Potential

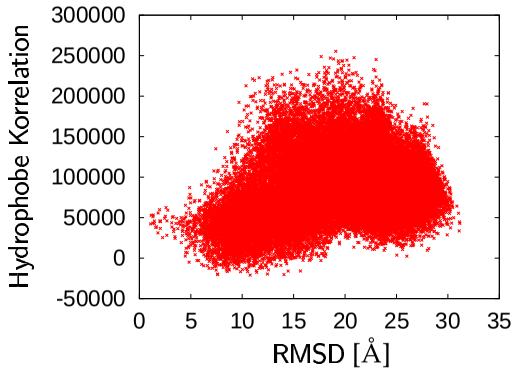
Abbildung A.4: Barnase (1bni) im Komplex mit C40A,C82A Barstar Doppelmutante, Referenzkomplex 1brs



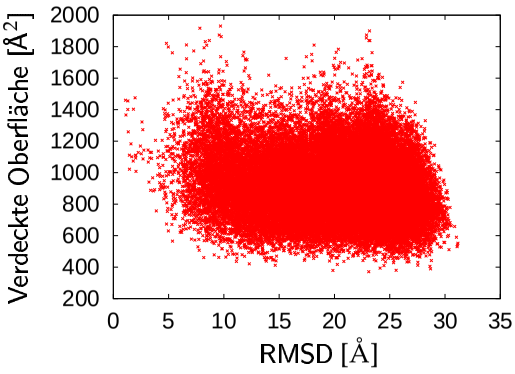
(a) geometrische Korrelation



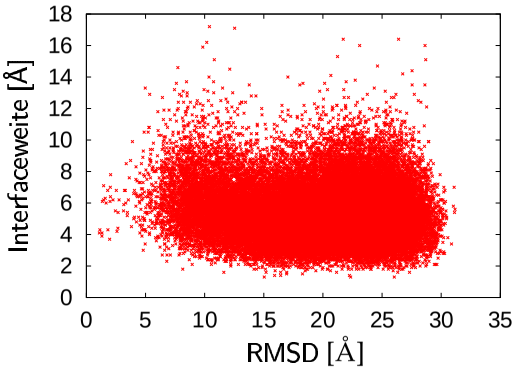
(b) elektrostatische Korrelation



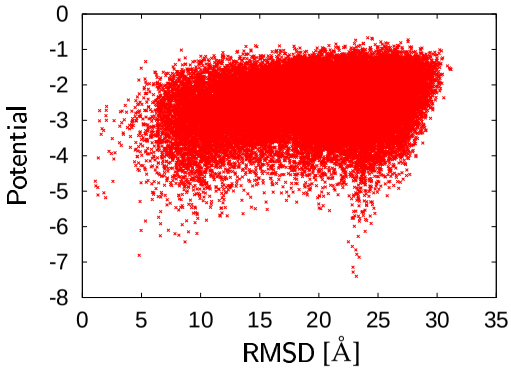
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche

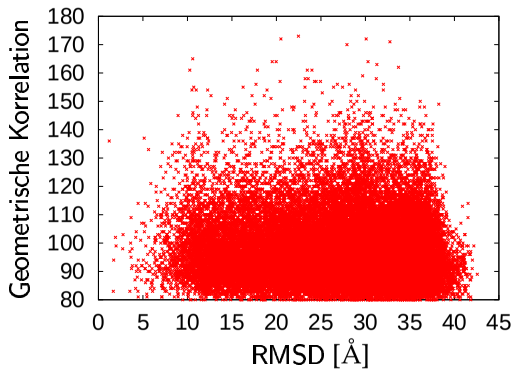


(e) Lückenweite

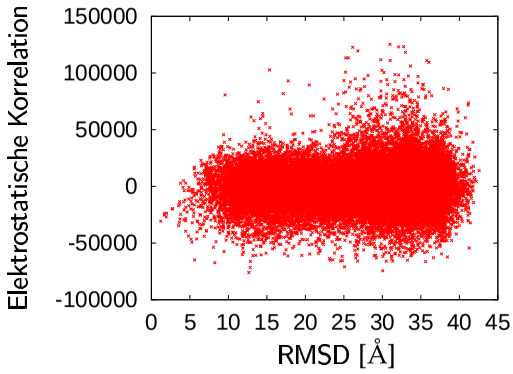


(f) statistisches Potential

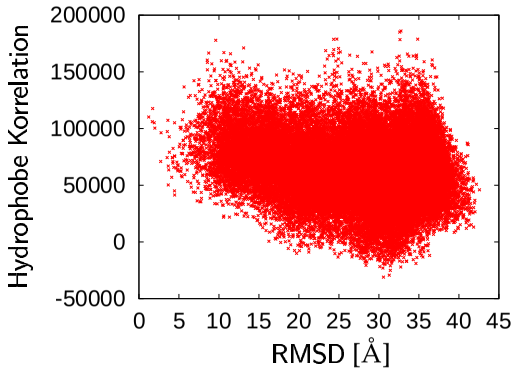
Abbildung A.5: α-Chymotrypsin (5cha) im Komplex mit 3. Domäne des Ovomucoid Proteins aus Truthahn (2ovo), Referenzkomplex 1cho



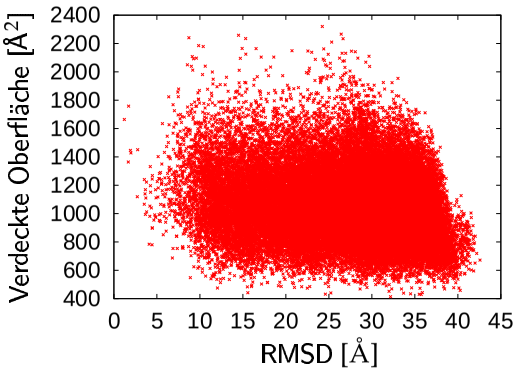
(a) geometrische Korrelation



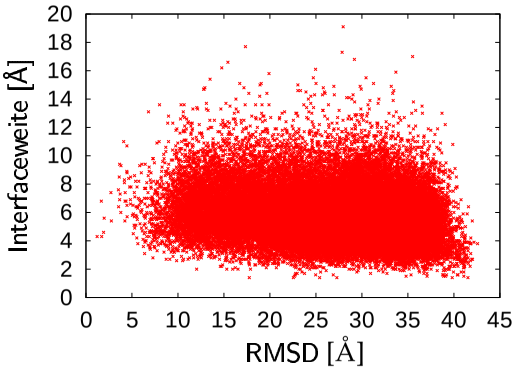
(b) elektrostatische Korrelation



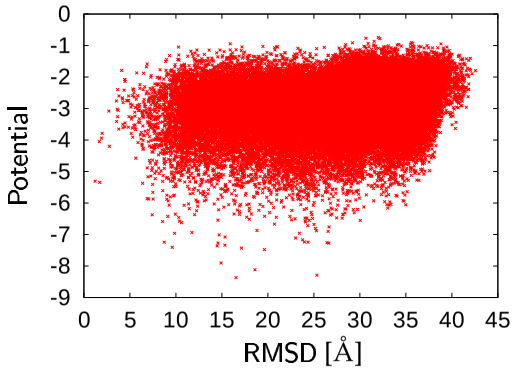
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche

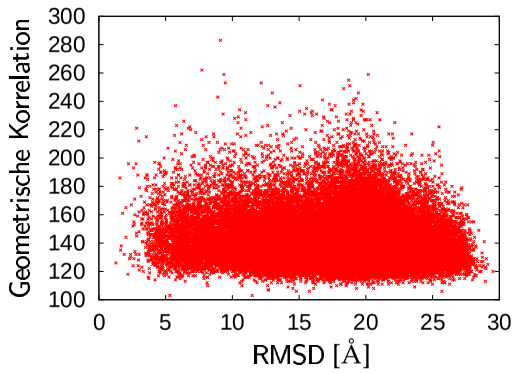


(e) Lückenweite

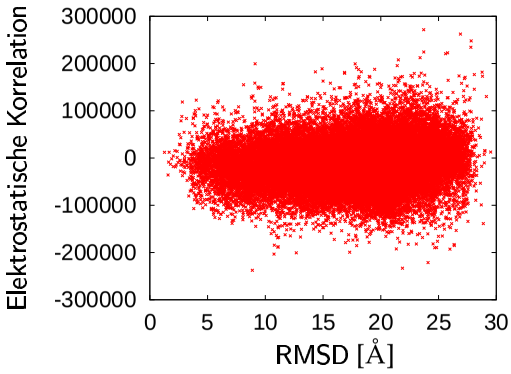


(f) statistisches Potential

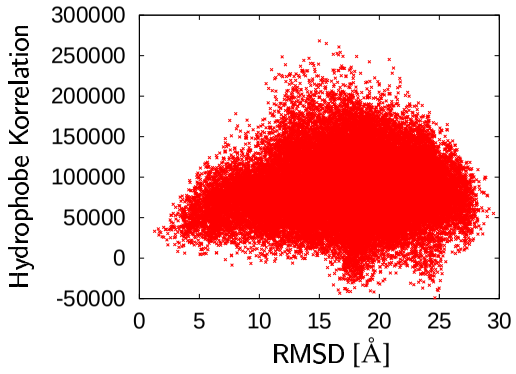
Abbildung A.6: Acetylcholinesterase (2ace) im Komplex mit Fasciculin II (1fsc), Referenzkomplex 1fss



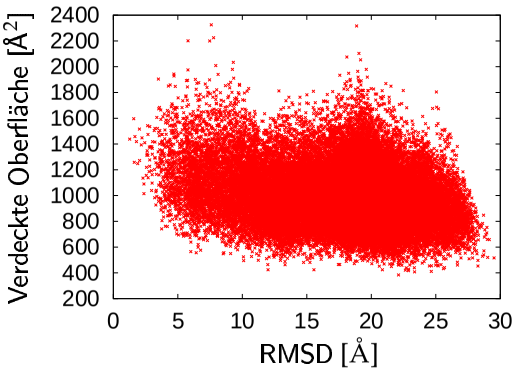
(a) geometrische Korrelation



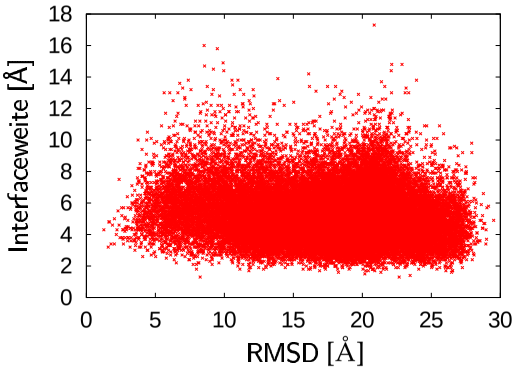
(b) elektrostatische Korrelation



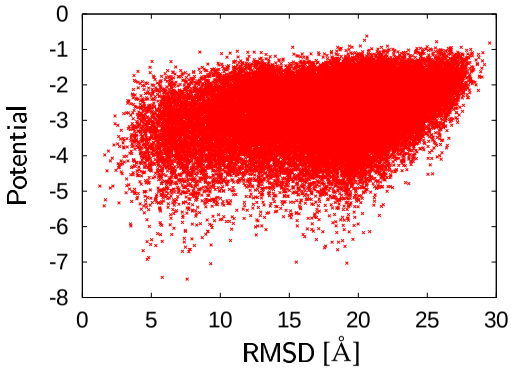
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche

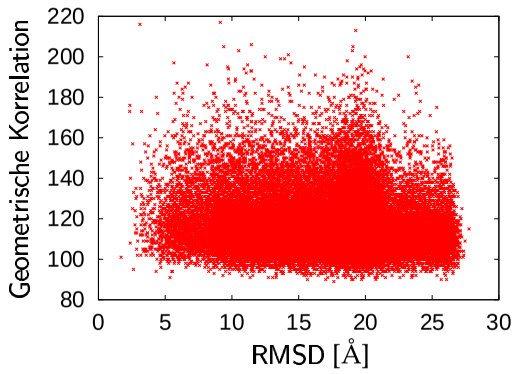


(e) Lückenweite

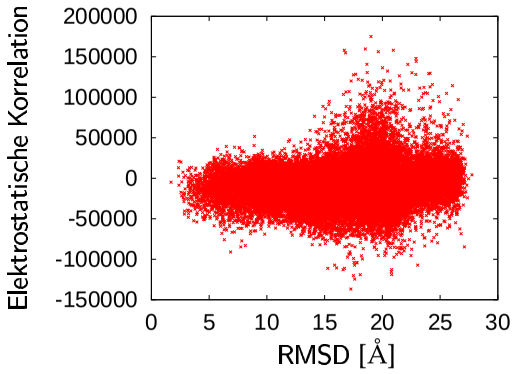


(f) statistisches Potential

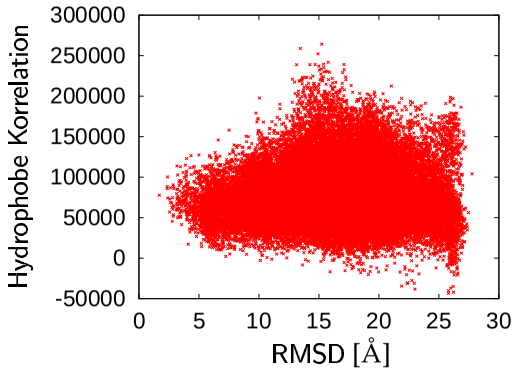
Abbildung A.7: Kallikrein A (2pka) im Komplex mit pankreatischem Trypsin-Inhibitor (1bpi), Referenzkomplex 2kai



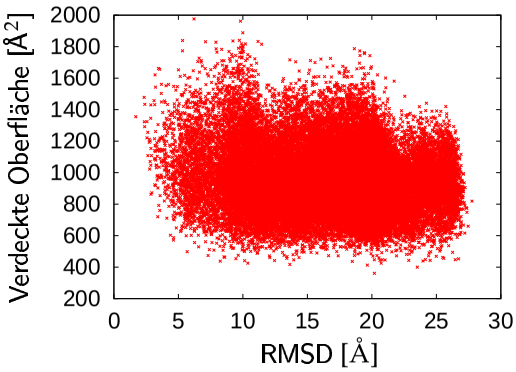
(a) geometrische Korrelation



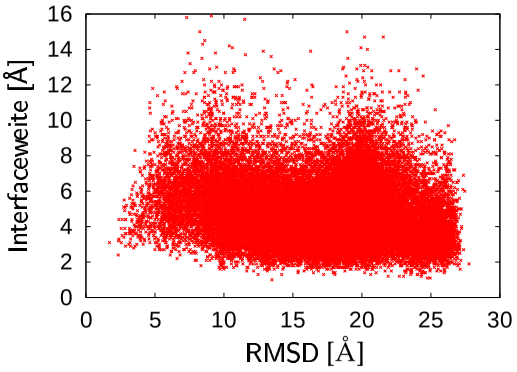
(b) elektrostatische Korrelation



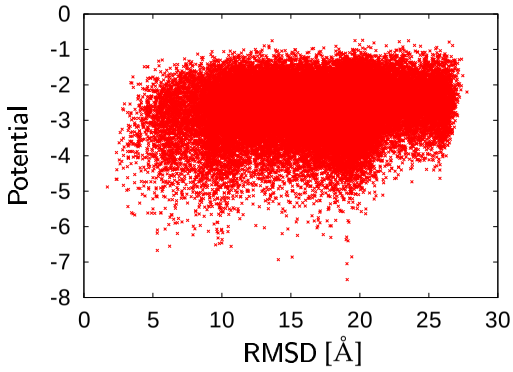
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche

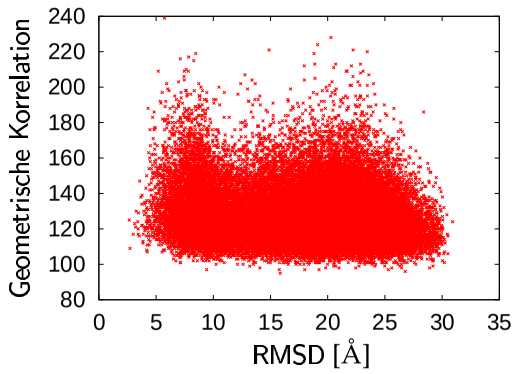


(e) Lückenweite

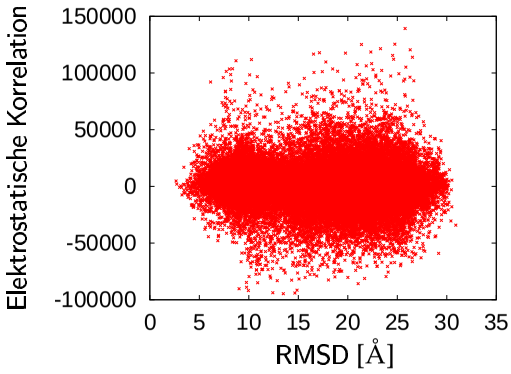


(f) statistisches Potential

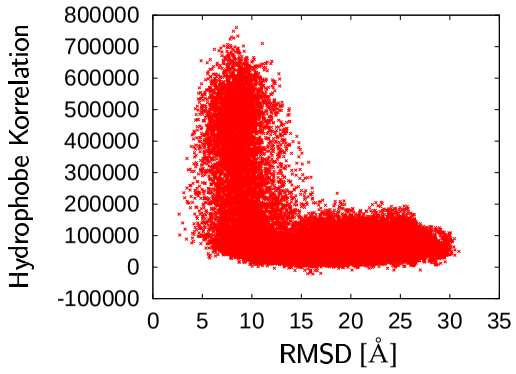
Abbildung A.8: β-Trypsin (2ptn) im Komplex mit pankreatischem Trypsin-Inhibitor (4pti), Referenzkomplex 2ptc



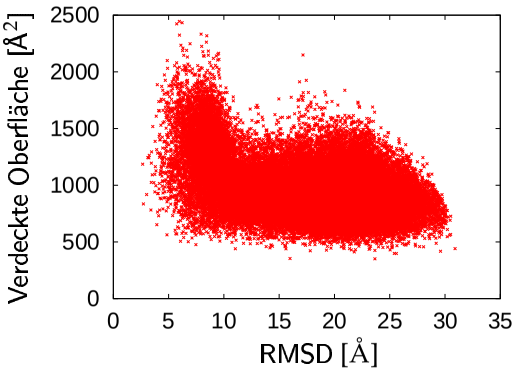
(a) geometrische[Korrelation



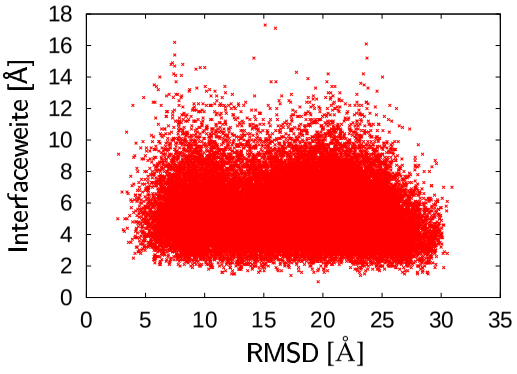
(b) elektrostatische Korrelation



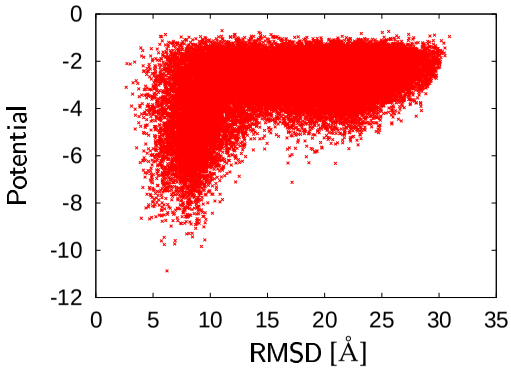
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche

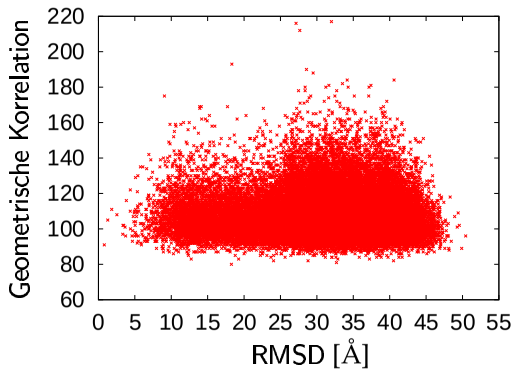


(e) Lückenweite

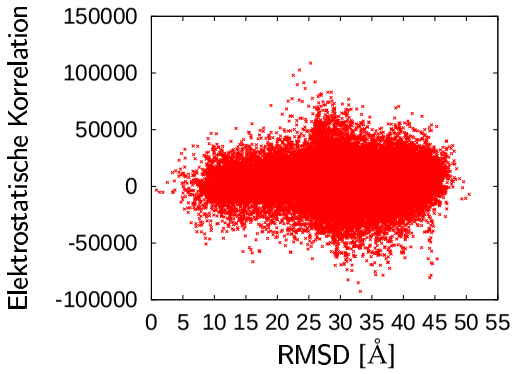


(f) statistisches Potential

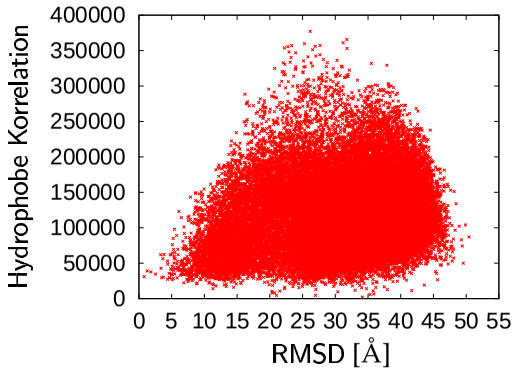
Abbildung A.9: Subtilisin (1sup) im Komplex mit Chymotrypsin-Inhibitor 2 (2ci2), Referenzkomplex 2sni



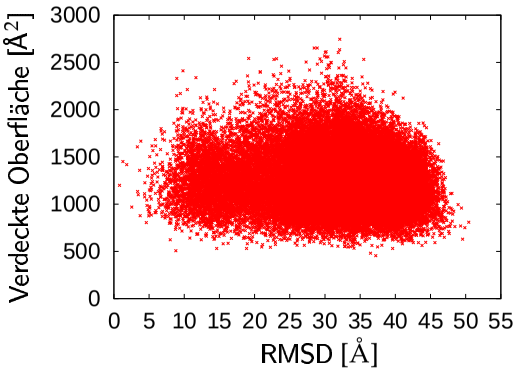
(a) geometrische Korrelation



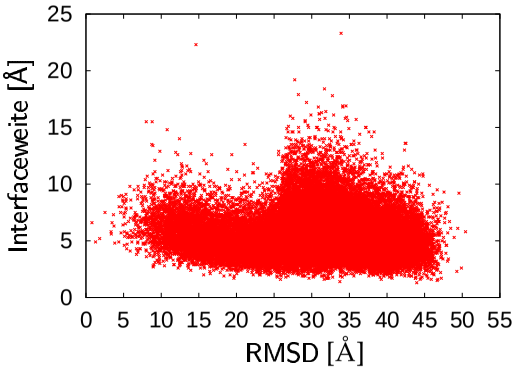
(b) elektrostatische Korrelation



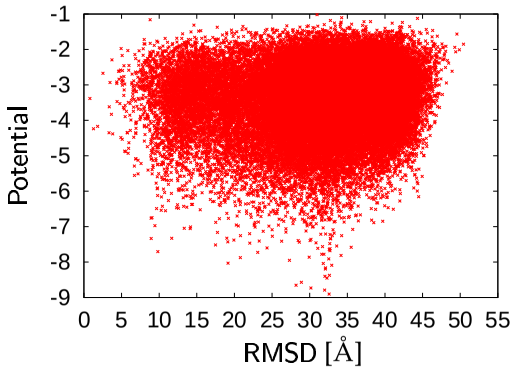
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche

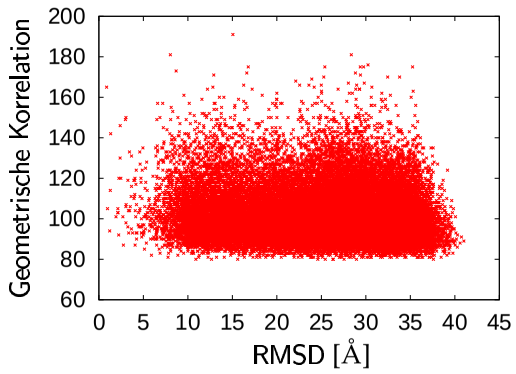


(e) Lückenweite

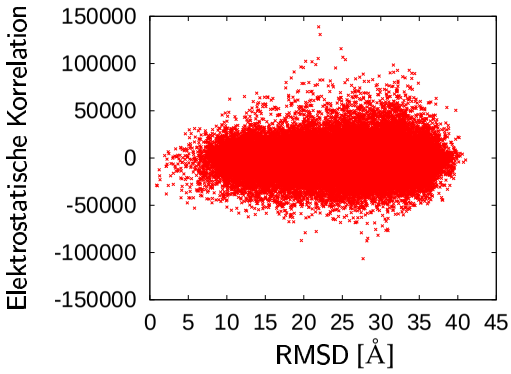


(f) statistisches Potential

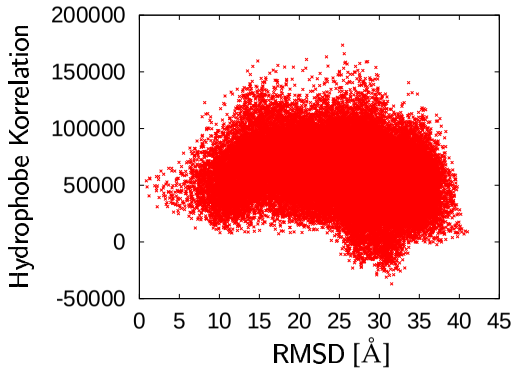
Abbildung A.10: Methylamin-Dehydrogenase (2bbk) im Komplex mit Amicyanin (1aan), Referenzkomplex 1mda



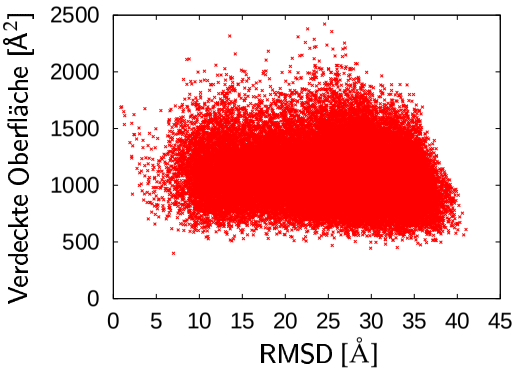
(a) geometrische Korrelation



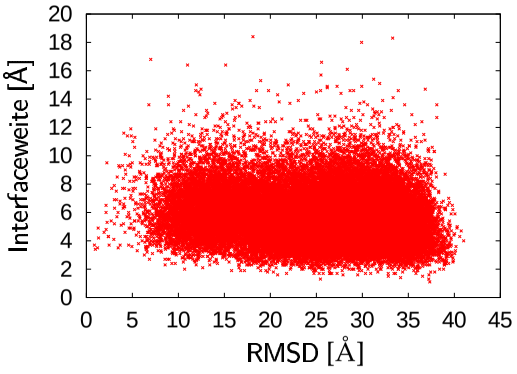
(b) elektrostatische Korrelation



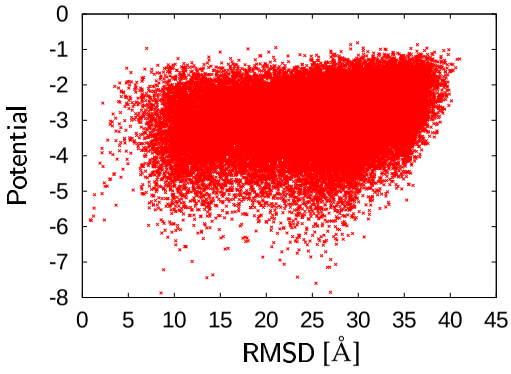
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche

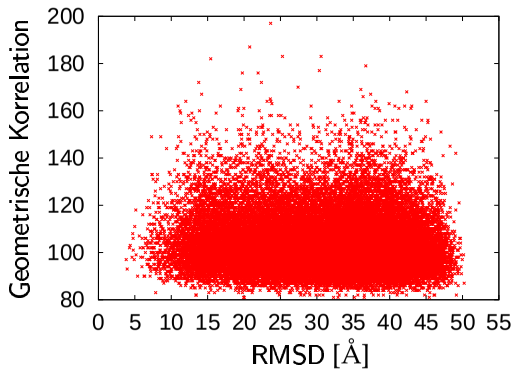


(e) Lückenweite

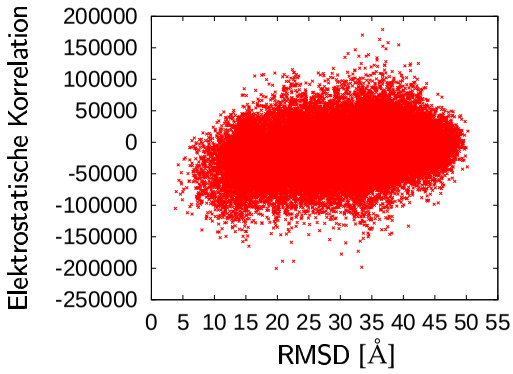


(f) statistisches Potential

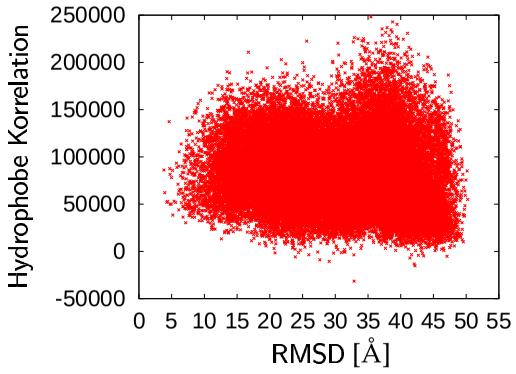
Abbildung A.11: Acetylcholinesterase (1maa) im Komplex mit Fasciculin II (1fsc), Referenzkomplex 1mah



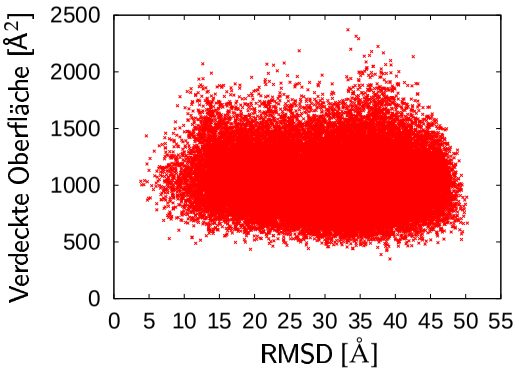
(a) geometrische Korrelation



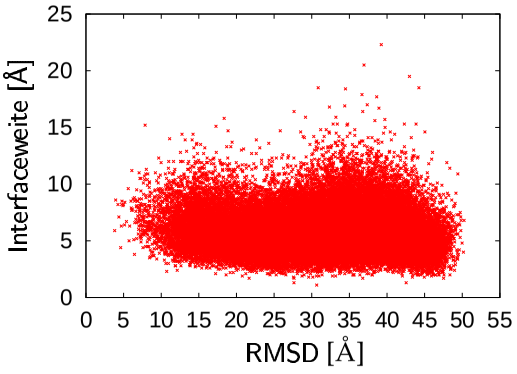
(b) elektrostatische Korrelation



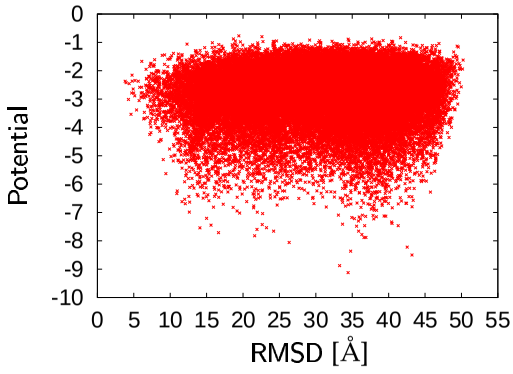
(c) hydrophobe Korrelation



(d) verdeckte Oberfläche



(e) Lückenweite



(f) statistisches Potential

Abbildung A.12: Cytochrom-C Peroxidase (1ccp) im Komplex mit Cytochrom-C (1ycc), Referenzkomplex 2pcc

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Badel-Chagnon, A., Nessi, J., Buffat, L. & Hazout, S. (1994) Iso-depth contour map of a molecular surface. *J Mol Graph* **12**, 162–8, 193.

Bennett, M. J., Choe, S. & Eisenberg, D. (1994) Domain swapping: entangling alliances between proteins. *Proc Natl Acad Sci U S A* **91**, 3127–31.

Bennett, M. J., Schlunegger, M. P. & Eisenberg, D. (1995) 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci* **4**, 2455–68.

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.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002) The Protein Data Bank. *Acta Cryst D Biol Cryst* **58**, 899–907.

Betts, M. J. & Sternberg, M. J. (1999) An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng* **12**, 271–83.

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992) A training algorithm for optimal margin classifiers. In Haussler, D., Hg., *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 144–152. ACM Press.

Bower, M., Cohen, F. E. & Dunbrack, R. L. (1997) Sidechain prediction from a backbone-dependent rotamer library: A new tool for homology modeling. *J Mol Biol* **267**, 1268–1282.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987) Crystallographic R- factor refinement by molecular dynamics. *Science* **235**.

Carugo, O. & Argos, P. (1997) Protein-protein crystal-packing contacts. *Protein Sci* **6**, 2261–3.

Chacon, P. & Wriggers, W. (2002) Multi-resolution contour-based fitting of macromolecular structures. *J Mol Biol* **317**, 375–84.

Chang, C.-C. & Lin, C.-J. (2001) *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

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

Clausen-Schaumann, H., Seitz, M., Krautbauer, R. & Gaub, H. E. (2000) Force spectroscopy with single bio-molecules. *Curr Opin Chem Biol* **4**, 524–30.

Connolly, M. L. (1986) Measurement of protein surface shape by solid angles. *J Mol Graph* **4**, 3–6.

Connolly, M. L. (1992) Shape distributions of protein topography. *Biopolymers* **32**, 1215–36.

Cooley, J. W. & Tukey, J. W. (1965) An Algorithm for the Machine Calculation of Complex Fourier Series. *Math Comp* 297–301.

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.

Delbrück, M. (1963) *Über Vererbungschemie*. Köln : Westdt. Verl.

Dixon, J. S. (1997) Evaluation of the CASP2 docking section. *Proteins* **Suppl 1**, 198–204.

Dong, F., Spott, S., Zimmermann, O., Kisters-Woike, B., Muller-Hill, B. & Barker, A. (1999) Dimerisation mutants of Lac repressor. I. A monomeric mutant, L251A, that binds Lac operator DNA as a dimer. *J Mol Biol* **290**, 653–66.

Edelsbrunner, H. & Mucke, E. (1994) Three dimensional alpha-shapes. *ACM Transactions on Graphics* **13**, 43–72.

Edmonds, D., Rogers, N. K. & Sternberg, M. J. E. (1984) Regular representationof irregular charge distributions. Applications to the electrostatic potential of globular proteins. *Mol Phys* **52**, 1487–1494.

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* **340**, 245–6.

Fischer, E. (1894) Einfluss der Konfiguration auf die Wirkung der Enzyme. *Chem.Ber.* **27**, 2985–2993.

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.

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

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.

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

Green, S. M., Gittis, A. G., Meeker, A. K. & Lattman, E. E. (1995) One-step evolution of a dimer from a monomeric protein. *Nature Structural Biology* **2**, 746 ff.

Grimm, V. (2002) *Untersuchung eines wissensbasierten Potentials zur Bewertung von Protein-Protein-Docking-Studien.* Dissertation, Universität zu Köln.

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.

Heuser, P. (2002) persönliche Mitteilung .

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

Hunter, C. A., Singh, J. & Thornton, J. M. (1991) Pi-pi interactions: the geometry and energetics of phenylalanine-phenylalanine interactions in proteins. *J Mol Biol* **218**, 837–846.

Ingram, V. M. (1956) A specific chemical difference between globins of normal human and sickle cell anaemia hemoglobin. *Nature* **178**, 792.

Jacobson, M. P., Friesner, R. A., Xiang, Z. & Honig, B. (2002) On the role of the crystal environment in determining protein side-chain conformations. *J Mol Biol* **320**, 597–608.

Janin, J. (2001) Structural Principles of Protein-Protein and Protein-DNA Recognition. In *Material zur Winterschule 'In Silico Biomolecular Recognition', Bologna, Italien.*

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

Janin, J. & Rodier, F. (1995) Protein-protein interaction at crystal contacts. *Proteins* **23**, 580–7.

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

Jones, S. & Thornton, J. M. (1997b) Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* **272**, 133–43.

Kabsch, W. & Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–637.

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.

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

Laskowski, R. (1995) SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 323–330.

Lattman, E. E. (1972) *Optimal Sampling of the rotation function.* Appendix. Gordon and Breach, New York.

Leatherbarrow, R. J. & Edwards, P. R. (1999) Analysis of molecular recognition using optical biosensors. *Current Opinion in Cheical Biology* **3**, 544–547.

Lessel, U. & Schomburg, D. (1994) Similarities between protein 3-D structures. *Protein Eng* **7**, 1175–87.

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

Lijnzaad, P. & Argos, P. (1997) Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins* **28**, 333–43.

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

Lorber, D. M., Udo, M. K. & Shoichet, B. K. (2002) Protein-protein docking with multiple residue conformations and residue substitutions. *Protein Sci* **11**, 1393–408.

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.

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.

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.

Müller, H. (1996) *Charakterisierung von Kontaktstellen in Proteinkristallen*. Dissertation, GBF, Braunschweig und Univ.-GH Siegen.

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.

Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997) Correlated mutations contain information about protein-protein interaction. *J Mol Biol* **271**, 511–23.

Peters, K. P., Fauck, J. & Frommel, C. (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J Mol Biol* **256**, 201–13.

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

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. & Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* **17**, 1030–2.

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

Rossmann, M. G. & Blow, D. M. (1962) The Detection of Sub-Units Within Crystallogryphic Asymmetric Unit. *Acta Cryst* **15**, 15–24.

Sandak, B., Wolfson, H. J. & Nussinov, R. (1998) Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. *Proteins* **32**, 159–74.

Schomburg, D. & Reichelt, J. (1988) BRAGI: A Comprehensive Modeling Program System. *J Mol Graph* 161–165.

Schrödinger, E. (1946) *Was ist Leben? Die lebende Zelle mit den Augen des Physikers betrachtet*. Bern: Francke.

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.

Shindyalov, I. N. & Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **11**, 739–47.

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.

Spott, S., Dong, F., Kisters-Woike, B. & Muller-Hill, B. (2000) Dimerisation mutants of Lac repressor. II. A single amino acid substitution, D278L, changes the specificity of dimerisation. *J Mol Biol* **296**, 673–84.

Stryer, L. (1995) *Biochemie*. Spektrum Verlag.

Strynadka, N. C., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B. K., Kuntz, I. D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A.,

Duncan, B., Rao, M., Jackson, R., Sternberg, M. & James, M. N. (1996) Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. *Nat Struct Biol* **3**, 233–9.

Tanford, C. (1970) Protein Denaturation (Part C), Theoretical models for the mechanism of denaturation. *Advan. Protein Chem* **25**, 1–95.

Thornton, J. (1998) www.biochem.ucl.ac.uk/bsm/PP/server .

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. & Yoshikawa, S. (1996) The whole structure of the 13 subunit oxidized cytochrom c oxidase at 2.8 Å. *Science* 272.

Vajda, S., Vakser, I. A., Sternberg, M. J. E. & Janin, J. (2002) Modeling of protein interactions in genomes. *Proteins* **47**, 444–446.

Vakser, I. A. (1995) Protein docking for low-resolution structures. *Prot Eng* **8**, 371–377.

Vakser, I. A. (1996) Low-resolution docking: prediction of complexes for underdetermined 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.

Voet, D. & Voet, J. G. (1992) *Biochemie*. A. Maelicke and W. Müller-Esterl, VCH.

Wüthrich, K. (1989) Determination of three-dimensional protein structures in solution by nuclear magnetic resonance: an overview. *Methods Enzymol* **177**, 125–31.

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.

Xu, D., Tsai, C. J. & Nussinov, R. (1998) Mechanism and evolution of protein dimerization. *Protein Sci* **7**, 533–44.

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

Lebenslauf

25.04.1965	geboren in Remscheid, Staatsangehörigkeit deutsch, ledig
1975-1984	Leibniz Gymnasium, Remscheid
Juni 1984	Abitur
Oktober 1984	Beginn des Biologiestudiums an der Universität zu Köln
1985 - 1986	Zivildienst im Umweltschutz, Wuppertal
Mai 1990	Vordiplomprüfung in Köln
Januar 1996	Diplomprüfung Biologie
1996 - 1997	Diplomarbeit bei Prof. Dr. B. Müller-Hill an der Universität zu Köln Thema: „Untersuchungen zur Spezifität und Stabilität der C-terminalen Dimerisierungsregion des Lac-Repressors aus Escherichia coli“
seit 1997	Doktorarbeit bei Herrn Prof. Dr. D. Schomburg an der Universität zu Köln
Januar 2000	Mitgründung der Bioinformatikfirma SCIENCE FACTORY GmbH