GMD Research Series

GMD –
Forschungszentrum
Informationstechnik
GmbH

Holger Claußen

Effizientes Protein-Ligand-Docking mit flexiblen Proteinstrukturen

N° 11/2001

GMD – Forschungszentrum Informationstechnik GmbH
Schloß Birlinghoven
D-53754 Sankt Augustin
Germany
Telefon +49 -2241 -14 -0
Telefax +49 -2241 -14 -2618
http://www.gmd.de

Anschrift des Verfassers/Address of the author:
Holger Claußen
Institut für Algorithmen und Wissenschaftliches Rechnen
GMD – Forschungszentrum Informationstechnik GmbH
Schloß Birlinghoven
D-53754 Sankt Augustin
E-mail: Holger.Claussen@gmd.de

D5

Die Deutsche Bibliothek - CIP-Einheitsaufnahme:

Claußen, Holger:
Effizientes Protein-Ligand-Docking mit flexiblen Proteinstrukturen /
Holger Claußen. GMD — Forschungszentrum Informationstechnik GmbH. -
Sankt Augustin : GMD — Forschungszentrum Informationstechnik, 2001
(GMD Research Series ; 2001, No. 11)
Zugl.: Bonn, Univ., Diss., 2001
ISBN 3-88457-394-2

ISSN 1435-2699
ISBN 3-88457-394-2

Kurzfassung

Medikamente greifen in das komplexe Netzwerk von Regel- und Stoffwechselprozessen eines Organismus ein, meist indem sie an bestimmte Zielmoleküle binden und ihre Funktion beeinflussen. Die Suche nach geeigneten Wirkstoffen bedient sich heute ausgefeilter experimenteller und computergestützter Verfahren, um neue Verbindungen von therapeutischer Relevanz zu finden. Eine solche computerbasierte Methode ist das *molekulare Docking*. Diese Methode versucht auf Grundlage der dreidimensionalen Struktur eines Zielproteins zu entscheiden, ob gegebene Verbindungen (Liganden) eine hohe Bindungsaffinität zum Zielprotein besitzen, und zwar bevor die Substanzen synthetisiert werden.

Die meisten Ansätze für das Docking-Problem berücksichtigen die Flexibilität der Liganden. Dagegen wird das Protein vor allem aus Effizienzgründen in der Regel als starres Objekt behandelt. Aus experimentellen Beobachtungen weiß man jedoch, daß es auch bei den Proteinen Konformationsänderungen aufgrund der Bindung verschiedener Liganden gibt (*induced fit*), die kritisch für das Docking sein können.

Die hier vorgestellte Arbeit präsentiert einen neuen, effizienten Ansatz für das Docking-Problem, der in der Lage ist, Variationen der Proteinkonformation beim Docking zu berücksichtigen. Die Proteinflexibilität wird dabei durch eine Menge (Ensemble) möglicher Proteinstrukturen repräsentiert, die während des Dockingprozesses zu neuen gültigen Proteinstrukturen rekombiniert werden können. Die Abhängigkeiten zwischen den Konformationen der einzelnen Strukturen verwaltet ein sog. Kompatibilitätsgraph, auf den Graphsuchalgorithmen angewendet werden.

Die Methode wurde anhand von zehn Ensembles experimentell bestimmter Protein-Ligand-Komplexe evaluiert, bei denen die Lage der Liganden im Protein, sowie Konformationen der Bindungspartner bekannt sind.

Schlagworte: strukturbasierter Wirkstoffentwurf, computerbasierter Wirkstoffentwurf, Protein-Ligand-Wechselwirkungen, flexibles molekulares Docking, Proteinflexibilität

Abstract (English)

Drugs interfere with the complex network of regulatoric and metabolic processes of an organism, mostly by binding to certain target molecules and affecting their function. Today, the search for appropriate agents uses sophisticated experimental and computer-aided procedures in order to find novel compounds of therapeutical relevance. One such computer based method is called *molecular docking*. This method tries to decide whether given compounds possess a high binding affinity to the given target protein using its known three-dimensional structure, before the substance is synthezied actually.

Most approaches for the docking problem take into account the flexibility of ligands. In contrast, the protein is usually treated as a rigid object largely for efficiency reasons. However, it is known from experimental data that conformational changes due to the binding to different ligands (induced fit) do also appear in proteins, which may be critical for docking.

The work presented here, describes a novel and efficient approach for the docking problem, which is able to take into account the protein flexibility. The flexibility of a protein is represented by a set (ensemble) of protein structures, which may be recombined to new valid protein structures during the docking process. The dependencies between the conformations of the particular structures are described by a so-called compatibility graph, to which graph search algorithms are applied.

The method has been evaluated on the basis of ten ensembles of experimentally determined protein-ligand complexes, for which the position of the ligands within the protein and the conformations of the binding partners are known.

Keywords: structure-based drug design, computer-based drug design, protein-ligand interactions, flexible molecular docking, protein flexibility

Vorwort

Diese Arbeit faßt die Ergebnisse meiner Tätigkeit an der GMD – Forschungszentrum Informationstechnik GmbH im Institut für Algorithmen und wissenschaftliches Rechnen (SCAI) zusammen. Das vorgestellte Docking-Programm FlexE entstand im Rahmen des Projektes *Rezeptormodellierung und De-Novo-Design kombinatorischer Bibliotheken* (RELIMO). Die Partner dieses vom Bundesministerium für Bildung und Forschung (bmb+f) geförderten Verbundprojektes[1] sind das Institut für Pharmazeutische Chemie der Philipps-Universität Marburg, die Firmen Merck KGaA, Darmstadt, und Boehringer Ingelheim Pharma KG, Ingelheim, sowie die GMD Institute IPSI[2] und SCAI[3]. Bei der Entwicklung von FlexE konnte ich außerdem auf viele Resultate des Vorgängerprojektes *Berechnung und Vorhersage von Rezeptor-Ligand-Wechselwirkungen* (RELIWE)[4], insbesondere auf das Docking-Programm FlexX zurückgreifen.

Mein besonderer Dank gilt meinem Doktorvater Thomas Lengauer, der mir nach meinem Studium in Bielefeld in seiner Arbeitsgruppe die Möglichkeit gab, meine Kenntnisse der Molekularen Bioinformatik weiter zu vertiefen. Er hat meine Arbeit stets mit großem Interesse und zahlreichen konstruktiven Kommentaren begleitet und damit im entscheidenden Maße zum Gelingen dieser Arbeit beigetragen. Ferner möchte ich mich bei Matthias Rarey bedanken, der mir den Code des Docking-Programms FlexX zur Verfügung gestellt hat. Er hatte ebenso wie Christian Lemmen immer ein offenes Ohr für meine Fragen. Beide haben mir in zahllosen Diskussionen eine Menge hilfreicher Tips gegeben.

Diese Arbeit ist im interdisziplinären Bereich zwischen Chemie und Informatik angesiedelt. Deshalb war es eine große Hilfe für mich, immer wieder auf die Erfahrung der Chemiker Bernd Kramer und Christian Buning zurückgreifen zu können, die an der GMD im RELIMO-Projekt mitgearbeitet haben. Auch für die Validierung von FlexE waren sie eine große Unterstützung ebenso wie die pharmazeutischen Projektpartner.

Die Arbeitsgruppe an der GMD und das RELIMO-Projekt insgesamt haben sich als ein gutes Forschungsumfeld erwiesen. Dafür möchte ich mich neben den bereits erwähnten Personen bei allen (ehemaligen) Kollegen und Projektmitgliedern und herzlich bedanken: Karl Aberer, Soheila Anzali, Joannis Apostolakis, Gehard Barnickel, Hans Briem, Holger Gohlke, Helmuth Griebler, Sven Grüneberg, Judith Günther, Sandra Handschuh, Klemens Hemm, Manfred Hendlich, Sally Hindle, Daniel Hoffmann, Gerhard Klebe, Oliver Krämer, Uta Lessel, Astrid Maaß, Ingo Macherius, Günther Metz, Theo Mevissen, Willem Nissink, Friedrich Rippmann, Joachim Selbig, Andrea Schafferhans, Stefan Schmitt, Eberhard Schrüfer, Christoph Sotriffer, Ingolf Sommer, Ralf Zimmer, Marc Zimmermann.

Schließlich möchte ich noch Alberto D. Podjarny (UPR de Biologie Structurale IGBMC, Illkirch, Frankreich) und Michael van Zandt (Inst. for Diabetes Discovery, Branford, CT, USA) erwähnen, die mir nach Vermittlung von Gerhard Klebe und Oliver Krämer freundlicherweise das Homologiemodell der Aldose Reduktase zur Verfügung gestellt haben.

Schloß Birlinghoven, den 06. Mai 2001 Holger Claußen

[1] Förderkennzeichen: 0311620, Laufzeit: 01.04.1998 – 31.03.2001
[2] Institut für Integrierte Publikations- und Informationssysteme, Darmstadt
[3] Institut für Algorithmen und wissenschaftliches Rechnen, St. Augustin
[4] Förderkennzeichen: 01 IB 302 A, Laufzeit: 01.04.93 – 31.03.97

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

Die Wirkung von Medikamenten beruht im allgemeinen auf einem Eingriff in die komplexen Regelungsmechanismen und Stoffwechselprozesse des Körpers, die durch eine Vielzahl unterschiedlicher Moleküle, oft Proteine, gesteuert und katalysiert werden. Wirkstoffmoleküle beeinflussen diese *regulatorischen* und *metabolischen Netzwerke*, indem sie mit bestimmten *Zielmolekülen* interagieren. Wesentliche Voraussetzung für eine solche Interaktion ist eine hohe *Bindungsaffinität* zwischen Wirkstoff- und Zielmolekül. Über die Höhe dieser Affinität entscheiden dabei sowohl die geometrische als auch die chemische Komplementarität der beiden Moleküle.

Die Suche nach neuen Medikamenten gegen eine bestimmte Krankheit beginnt heute in der Regel mit der Identifikation eines geeigneten Zielmoleküls (meist ein Makromolekül) mit regulatorischen bzw. metabolischen Aufgaben, das durch den neuen Wirkstoff beeinflußt werden soll. Dies setzt ein gutes Verständnis des Krankheitsbildes auf biomolekularer Ebene voraus. Der Vergleich der exprimierten Gene in gesunden und kranken Zellen mit Hilfe von sog. DNA-Chips [1, 2] sowie die vollständige Aufklärung des menschlichen Genoms [3, 4] sind dabei wichtige neue Informationsquellen.

Wenn ein geeignetes Zielmolekül identifiziert ist, beginnt die Suche nach sog. *Liganden*, in der Regel kleine Moleküle, die an das Zielmolekül binden. Findet sich ein Ligand mit hoher Affinität zum Zielmolekül, so dient er als *Leitstruktur* für die Entwicklung des neuen Medikaments [5]. Die Leitstruktursuche kann experimentell durch Messen von Bindungsaffinitäten oder virtuell mit Computermethoden erfolgen. Oft werden Kombinationen beider Verfahren eingesetzt.

Vollautomatisierte Robotersysteme sind in der Lage, mehr als zehntausend Verbindungen pro Tag experimentell daraufhin zu überprüfen, ob sie an ein Zielmolekül binden [6]. Die Substanzen werden dabei auf etwa 10 cm x 20 cm großen Platten mit bis zu 1536 Vertiefungen zusammengebracht, wobei jede Vertiefung eine andere Substanz enthält. Solche *High-Throughput-Screenings (HTS)* setzt die pharmazeutische Industrie heutzutage routinemäßig ein, um bei der Suche nach Leitstrukturen große Bibliotheken von $10^5 - 10^6$ Verbindungen zu durchmustern.

Bei der virtuellen Leitstruktursuche gibt es zwei Situationen: Ist die dreidimensionale Struktur des Zielmoleküls unbekannt, so lassen sich durch die Analyse von bekannten Liganden des Zielmoleküls Rückschlüsse auf den Bindungsmechanismus ziehen. Man kann nach ähnlichen Verbindungen suchen und mit QSAR-Modellen (quantitative structure-activity relationships, [7, 8]) versuchen, deren Affinität vorherzusagen. Man spricht deshalb

vom *ligandbasierten Wirkstoffdesign*. Beim *strukturbasierten Wirkstoffdesign* ist die dreidimensionale Struktur des Zielmoleküls dagegen verfügbar. Hier besteht die Möglichkeit, entweder neue Liganden ganz gezielt zu entwerfen (*De-Novo-Design*) oder für gegebene Verbindungen zu entscheiden, ob sie eine gute sterische und chemische Komplementarität zur gegebenen Struktur des Zielmoleküls besitzen (*Docking*). Durchmustert man mit diesen Verfahren große Mengen von Liganden, so spricht man von *virtuellem Screening*. Ziel ist es dabei, die Bindungsaffinität von potentiellen Liganden abzuschätzen, noch bevor sie synthetisiert werden.

Damit aus einer Leitstruktur ein Wirkstoff wird, müssen weitere physiko-chemische Eigenschaften optimiert werden. Denn neben einer hohen Affinität zum Zielmolekül muß ein Medikament weitere Anforderungen erfüllen: Es sollte gut synthetisierbar und stabil sein, die Verbindung darf weder selbst toxisch sein, noch im Körper in toxische Substanzen umgewandelt werden. Das Wirkstoffmolekül sollte sehr spezifisch an das Zielmolekül binden, um möglichst wenige Nebenwirkungen zu haben. Außerdem muß der Wirkstoff *bioverfügbar* sein, das heißt, er muß nach der Aufnahme unverändert und in möglichst hoher Konzentration zu seinem Wirkort gelangen und dort eine gewisse Zeit zur Verfügung stehen, bevor er abgebaut und ausgeschieden wird (*Pharmakokinetik*). Die Modifikationen, die aus diesen Gründen an der Leitstruktur vorgenommen werden, führen aber in der Regel zu einer Verbindung mit geringerer Affinität zum Zielmolekül. Deshalb versucht man neuerdings, die physiko-chemischen Eigenschaften bereits bei der Leitstruktursuche zu berücksichtigen.

In der *klinischen Prüfung* wird der neue Wirkstoff schließlich in mehreren Phasen zunächst an gesunden und später an erkrankten Patienten erprobt, bevor er als neues Medikament zugelassen werden kann. Diese klinische Prüfung ist der zeit- und kostenintensivste Schritt bei der Entwicklung neuer Medikamente.

Das experimentelle Durchmustern großer Substanzbibliotheken *High-Troughput-Screening* bei der Suche nach einer geeigneten Leitstruktur ist technisch und logistisch aufwendig. Denn alle Verbindungen, die getestet werden sollen, müssen verfügbar sein, das heißt, man muß sie entweder synthetisieren oder kaufen und für spätere Tests lagern. Dabei dürfen sich die Moleküle während der Lagerung nicht verändern. Die Messungen müssen mit kleinsten Mengen von Substanzen erfolgen. Dies erfordert eine hohe Präzision, ist aber auch Ursache von Meßfehlern. Darüber hinaus können Verunreinigungen, Nebenreaktionen und gealterte Verbindungen zu falschen Ergebnissen führen.

Virtuelle Verfahren haben gegenüber experimentellen den Vorteil, daß die Moleküle nicht real verfügbar sein müssen. Die Informationen über die Verbindungen liegen in elektronischen Datenbanken vor, so daß man keine Substanzen zu kaufen, zu synthetisieren, zu lagern und während der Messungen zu handhaben braucht. Deshalb sind auch die Kosten dieser computerbasierten Methoden erheblich geringer. Außerdem ist bei In-silico-Tests immer klar, welche Verbindung gerade bearbeitet wird, denn es gibt keine Seiteneffekte durch Verunreinigungen, Nebenreaktionen oder Alterung bei der Lagerung.

Da die zugrunde liegenden Modelle oft Vereinfachungen erforderlich machen, sind die virtuellen Vorhersagen allerdings nicht so exakt, daß sie Messungen ersetzen könnten. Aus diesem Grund verwendet man das virtuelle Screening vor allem, um die Zahl der Experimente zu reduzieren, indem man z.B. möglichst diverse Substanzkollektionen zusammenstellt, im voraus mit hoher Wahrscheinlichkeit nicht bindende Substanzen verwirft oder für ein gegebenes Zielmolekül maßgeschneiderte Bibliotheken entwirft (*focused libraries*).

Diese Arbeit beschäftigt sich mit einem speziellen Verfahren aus dem Bereich des virtuellen strukturbasierten Wirkstoffdesigns, und zwar dem *molekularen Protein-Ligand-Docking*. Dabei geht es um die Vorhersage eines Komplexes zwischen Proteinen und Liganden.

Die Liganden können sehr unterschiedliche Konformationen annehmen und nur in ganz bestimmten Konformationen in die Bindetasche des Proteins eingepaßt werden. Deshalb wird diese Flexibilität der Liganden heute von den meisten Docking-Programmen berücksichtigt. Proteine werden dagegen vor allem von den Programmen, die auf das Screening größerer Datenmengen ausgelegt sind, vor allem aus Effizienzgründen als starr betrachtet, obwohl es durchaus auch Konformationsänderungen auf der Proteinseite gibt. Deshalb können Programme, die starre Proteinstrukturen benutzen, nicht in jedem Falle den korrekten Bindungsmodus von Liganden vorhersagen, wenn die Proteinkonformation von der vorgegebenen Struktur abweicht.

Das Problem der Proteinflexibilität ist in erster Linie kombinatorischer Natur. Die Zahl der möglichen Proteinkonformationen wächst exponentiell mit der Zahl der drehbaren Bindungen im Protein. Allerdings sind viele dieser Konformationen aufgrund der dichten Packung der Atome im Protein gar nicht realisierbar und man hat beobachtet, daß die einzelnen Aminosäuren, aus denen ein Protein aufgebaut ist, in der Regel nur eine kleine Zahl unterschiedlicher Konformationen einnehmen (*Rotamere*). Diese Einschränkungen sowie die Annahme, daß die Tertiärstruktur und insbesondere die Form des aktiven Zentrums des Proteins im großen und ganzen erhalten bleiben muß, um die Funktionalität des Proteins zu erhalten, bieten die Möglichkeit, die Proteinflexibilität als ein diskretes kombinatorisches Optimierungsproblem zu formulieren.

Das Ziel dieser Arbeit ist die Entwicklung neuer Datenstrukturen und Algorithmen für das effiziente molekulare Protein-Ligand-Docking mit flexiblen Proteinstrukturen, die zu anwendbaren Programmen führen. Deshalb stehen Robustheit und praktische Laufzeit- und Speichereffizienz im Vordergrund. Alle hier vorgestellten Ansätze sind in dem neuen Programm FlexE implementiert, das zur Zeit von den Kooperationspartnern in der Praxis evaluiert wird.

Das folgende Kapitel gibt zunächst einen Überblick über das molekulare Protein-Ligand-Docking und stellt kurz das Docking-Programm FlexX vor, das starre Proteinstrukturen verwendet und auf dem diese Arbeit aufbaut. Danach wird am Beispiel der Aldose-Reduktase die Problematik flexibler Proteine beim Docking erläutert, ein Überblick über die Behandlung der Proteinflexibilität beim Docking gegeben sowie verschiedene Ansätze aus der Literatur zur Behandlung dieses Problems vorgestellt und verglichen. Kapitel 4 beschreibt das hier verwendete Modell der Proteinflexibilität und die Modifikationen, die sich daraus für die Modellierung der Wechselwirkungen, der Oberfläche und für die Bewertungsfunktion ergeben. Eine Diskussion der algorithmischen Aspekte des Verfahrens findet man im darauffolgenden Kapitel. Das sechste Kapitel geht auf verschiedene Methoden zur Evaluation eines Docking-Programms ein und erklärt, welche Verfahren hier angewendet werden. Die Ergebnisse der Evaluierung von FlexE an realen Daten und einen Vergleich zwischen FlexE und FlexX befinden sich in Kapitel 7. Anschließend werden einige Limitierungen der entwickelten Methode aufgezeigt und Lösungsansätze für deren Überwindung skizziert. Eine Zusammenfassung und ein Ausblick auf denkbare Erweiterungsmöglichkeiten schließen die Arbeit ab.

Kapitel 2

Protein-Ligand-Docking

Beim molekularen Protein-Ligand-Docking geht es darum, die Geometrie eines Komplexes und die Bindungsaffinität zwischen einem Makromolekül (Protein) und einem kleinen Molekül vorherzusagen, das in der Regel als *Ligand* bezeichnet wird. Zur Lösung dieses Problems werden also eine geeignete Modellierung von Protein und Ligand, ein Algorithmus, um den Liganden in der Bindetasche des Proteins zu plazieren, und eine Bewertungsfunktion für die vorhergesagten Plazierungen benötigt.

Proteine und Liganden haben unterschiedliche Eigenschaften. Sie werden hier kurz charakterisiert. Insbesondere unterscheiden sich diese Moleküle in ihrer Flexibilität. Die weniger flexiblen Proteine werden von den aktuellen Docking-Programmen in der Regel als starre Strukturen modelliert, während diese Programme vielfach die größere Flexibilität der Liganden berücksichtigen. Die verschiedenen Ansätze zur Modellierung der Ligandflexibilität werden knapp zusammengefaßt.

Zentrales Element aller Plazierungsalgorithmen ist die Überlagerung zweier Mengen korrespondierender Punkte. Sie sollen so aufeinander gelegt werden, daß ihre mittlere Abweichung minimal ist. Dieses Teilproblem kann man z.B. mit dem Kabsch-Algorithmus [9], den viele Docking-Programme verwenden [10], optimal lösen. Deshalb wird hier nicht weiter auf die Plazierungsalgorithmen eingegangen. Bei den Bewertungsfunktionen gibt es dagegen verschiedene Ansätze, deren wesentlichen Merkmale erläutert und diskutiert werden.

Ein aktuelles Docking-Programm ist FlexX [11]. Es kann flexible Liganden in starre Proteinstrukturen plazieren. Das Programm wird hier vorgestellt, weil diese Arbeit darauf aufbaut. Abgeschlossen wird dieses Kapitel durch eine Definition des Docking-Problems unter besonderer Berücksichtigung der Proteinflexibilität. Aktuelle Überblicksartikel zum Protein-Ligand-Docking im allgemeinen findet man in [10, 12].

2.1 Proteine

Proteine sind große Polymere aus mehreren tausend Atomen. Sie sind aus einem Satz von 20 Aminosäuren aufgebaut. Alle in Proteine eingebauten Aminosäuren haben dieselbe Grundstruktur: Sie bestehen aus einem zentralen C_α-Atom einer Amino- und einer Carboxylatgruppe sowie einem variablen Rest R, der auch als *Seitenkette* bezeichnet wird (s.a. Abb. 5.1). Amino- und Carboxylatgruppen verschiedener Aminosäuren können unter Wasserabspaltung verbunden werden (*Peptidbindungen*, vgl. Abb. 2.1). Eine lange Kette

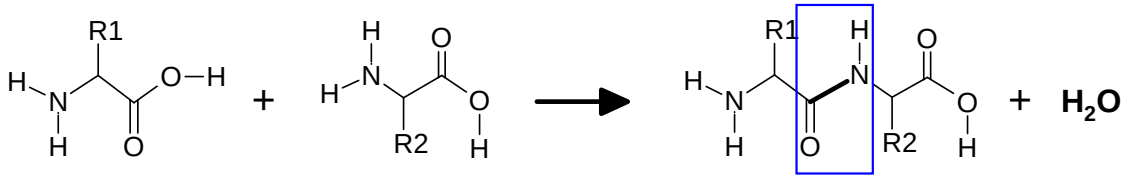


Abbildung 2.1: Peptidbindung. Aminosäuren können unter Wasserabspaltung zu Peptiden verknüpft werden. Die gebildete Peptidbindung (blauer Kasten) ist planar, das heißt, die vier Atome der Peptidbindung liegen in einer Ebene.

so verknüpfter Aminosäuren bildet das Rückgrat (*Backbone*) eines Proteins.

Die Peptidbindung ist infolge einer Elektronendelokalisation eine Einfachbindung mit partiellem Doppelbindungscharakter. Sie ist deshalb planar. Die Bindungen zum zentralen C_α-Atom und zu den Seitenketten sind dagegen prinzipiell frei drehbar, aber aus sterischen Gründen treten auch hier nur ganz bestimmte Torsionswinkel auf. Typische Seitenkettenkonformationen werden als *Rotamere* bezeichnet [13].

Eine spezielle Aminosäuresequenz (*Primärstruktur*) ist charakteristisch für ein Protein und bestimmt unter gegebenen äußeren Bedingungen seine räumliche Gestalt. Man unterscheidet dabei mehrere Ebenen: Die *Sekundärstruktur* wird vor allem durch Wasserstoffbrücken stabilisiert. Es bilden sich α-Helizes und β-Faltblätter, die durch Schleifen (*Loops*) miteinander verbunden sind. Die Abfolge mehrerer Sekundärstrukturelemente wird als *Motiv* bezeichnet. Mehrere Motive falten zu *Domänen* und ergeben die *Tertiärstruktur*, die vor allem von hydrophoben Wechselwirkungen und Disulfidbrücken zusammengehalten wird. Häufig sind die Domänen Träger einer bestimmten Funktion. Viele Proteine enthalten nur eine Domäne, komplexe Proteine können aber auch aus mehreren Domänen aufgebaut sein. Wenn ein Protein aus mehreren Aminosäureketten besteht, so wird deren Formation als *Quartärstruktur* bezeichnet [5, 14].

Proteine sind in der Lage, andere Moleküle sehr spezifisch zu erkennen und mit ihnen in Wechselwirkung zu treten. Dabei bindet der Ligand in der sog. *Bindetasche* des Proteins, die auch als *aktives Zentrum* bezeichnet wird. Dies ist in aller Regel eine deutliche Vertiefung in der sonst mehr oder minder konvexen Oberfläche des Proteins. Bindet ein Ligand, so bildet sich ein Komplex, der im Gegensatz zu einer kovalenten Bindung in der Regel reversibel ist. Wesentliche Voraussetzungen für die Komplexbildung sind eine sowohl chemische als auch geometrische Komplementarität der beiden Moleküle. Deshalb sprach man lange Zeit vom *Schlüssel-Schloß-Prinzip* [15, 16]. Dieses Paradigma wurde jedoch vom Prinzip des *induced fit* [5, 17] abgelöst, weil sich herausstellte, daß Proteine ihre Struktur bei der Bindung eines Liganden verändern können. Diese Proteinflexibilität wird jedoch bis jetzt von den meisten Docking-Programmen vernachlässigt.

2.2 Liganden

Im Gegensatz zu den Proteinen können Liganden sehr flexibel sein. Sie lassen sich aus einer Vielzahl unterschiedlicher Gruppen zusammensetzen und haben deshalb keine typische Struktur. Die Moleküle können flexible Ringsysteme enthalten, sehr große Makrozyklen sind aber nur in Ausnahmefällen für das Wirkstoff-Design von Interesse. Es ist davon

auszugehen, daß Liganden in einer Konformation an das Protein binden, die, ähnlich wie die Konformation in Lösung, energetisch günstig ist. Denn eine hohe Spannungsenergie würde einen großen Teil der Bindungsaffinität kompensieren.

Liganden können im Prinzip beliebige Moleküle sein, von einzelnen Metallionen bis hin zu ganzen Proteinen. Im allgemeinen sind aber nur nicht-peptidische organische Moleküle, die aus höchstens ein paar hundert Atomen bestehen, pharmazeutisch interessant [5, 18]. Für die Eignung als Medikament müssen die Verbindungen *bioverfügbar* sein, das heißt, sie müssen das Protein, an das sie binden sollen, unverändert erreichen. Peptide sind deshalb zumindest für die orale Verabreichung ungeeignet, weil sie im Verdauungstrakt abgebaut werden. Ebenso sind Substanzen, die mehr als fünf Wasserstoffbrücken-Donatoren, mehr als zehn Wasserstoffbrücken-Akzeptoren, ein Molekulargewicht von mehr als 500 Dalton haben oder zu lipophil sind ($logP > 5$), in der Regel als Wirkstoffe nicht geeignet („rule of five“ [19]).

2.2.1 Modellierung der Ligandflexibilität

Die Ligandflexibilität wird von den heutigen Docking-Programmen in der Regel berücksichtigt, weil die Liganden aufgrund der Vielzahl von Freiheitsgraden die unterschiedlichsten Konformationen bei der Bindung an ein Protein annehmen können. Einen ausführlicheren Überblick über die Modellierung der Ligandflexibilität beim Docking findet man bei Rarey [10].

Die einfachste Möglichkeit die Ligandflexibilität zu modellieren besteht darin, ein Ensemble von verschiedenen Ligandkonformationen zu verwenden, die separat in die Proteinstruktur gedockt werden [20, 21, 22]. Andere Ansätze zerlegen den Liganden in mehr oder minder große Fragmente, die man als starr betrachten kann. Diese starren Bausteine lassen sich entweder einzeln im aktiven Zentrum des Proteins plazieren und anschließend zu kompletten Molekülen zusammensetzen [23, 24, 25] oder aber der Ligand wird – ausgehend von einem oder mehreren Basisfragmenten – inkrementell in der Bindetasche aufgebaut [11, 26, 27, 28, 29, 30].

Genetische Algorithmen [31, 32, 33, 34, 35, 36, 37] orientieren sich an der Strategie der Evolution und versuchen, eine Menge von initial plazierten Ligandkonformationen durch zufällige Veränderungen und iteratives Rekombinieren von Teilkonformationen bezüglich einer sog. Fitness-Funktion zu optimieren. Abschnitt 3.3.3 geht etwas ausführlicher auf diese Art der Algorithmen ein.

Andere diskrete Modellierungen versuchen die Plazierungen zu optimieren, indem sie iterativ die aktuell beste Lösung auf andere Weise zufällig variieren [38, 39]. Schließlich gibt es die Distanzgeometriemethoden [40, 41, 42], die darauf basieren, alle potentiellen Konformationen eines Liganden durch die jeweils minimal und maximal möglichen Abstände der Ligandatome zu beschreiben. Innerhalb dieser Distanzintervalle wird eine Anordnung der Atome gesucht, die den durch das Protein bestimmten Randbedingungen entspricht. Auf diese Distanzgeometriemethoden wird hier nicht näher eingegangen.

Einen ganz anderen Ansatz verfolgen die Verfahren, die mit Moleküldynamik-Rechnungen [43, 44, 45, 46, 47], Monte-Carlo-Methoden [48, 49, 50, 51, 52, 53] oder Simulated-Annealing-Protokollen [54] versuchen, die Komplexbildung zu simulieren, indem sie die Energie einer oder mehrerer orientierter Startkonformationen des Liganden durch gezielte bzw. zufällige Konfigurationsänderungen minimieren (vgl. Abs. 3.3.4). Diese Verfahren sind allerdings in der Regel wesentlich zeitaufwendiger als die zuvor beschriebenen.

2.3 Bewertungsfunktionen

Bei der Vorhersage von Protein-Ligand-Komplexen ist die Bewertungsfunktion die zu optimierende Größe. Sie stellt eine heuristische Approximation der freien Bindungsenergie ΔG dar, weil deren exakte Berechnung heute auch mit aufwendigeren Methoden nicht möglich ist. Da die Bindungsenergie bei der Komplexbildung frei wird, ist sie per Definition mit negativem Vorzeichen versehen. Die Optimierung entspricht also einer Minimierung der geschätzten freien Bindungsenergie. Die Aufgabe der Bewertungsfunktion besteht darin, energetisch günstige, das heißt in der Realität mögliche Komplexe von energetisch ungünstigen Vorhersagen zu diskriminieren. Dazu sind die verschiedenen Lösungen bezüglich ihrer Energie richtig zu ordnen. Ausführliche Überblicksartikel über Bewertungsfunktionen für das Docking-Problem sind [55, 56, 57, 58, 59, 60].

Für das Protein-Ligand-Docking haben sich zwei Typen von Funktionen für die Bewertung herauskristallisiert: die sog. *empirischen* und die *wissensbasierten* Bewertungsfunktionen.

Empirische Bewertungsfunktionen [11, 35, 61, 62, 63, 64, 65, 66, 67, 68] versuchen, alle wesentlichen Beiträge zur freien Bindungsenergie durch gewichtete Terme zu erfassen. Hauptsächlich werden dabei Wasserstoffbrückenbindungen, Kontaktflächen zwischen Protein und Ligand sowie entropische Effekte berücksichtigt. Die Gewichte der Terme beziehen sich auf den Typ und die Geometrie der Wechselwirkungen und werden anhand experimentell bestimmter Protein-Ligand-Komplexe geeicht.

Die wissensbasierten Bewertungsfunktionen [69, 70, 71, 72, 73, 74] beruhen auf Statistiken über typische Abstände von Protein- und Ligandatomen in Komplexen. Sie gehen davon aus, daß häufig beobachtete Abstände energetisch günstig sind, und berechnen deshalb für jedes Paar von Atomen aus dem Histogramm ihrer Abstände ein Pseudopotential. Die Energie eines Protein-Ligand-Komplexes läßt sich dann durch Addition der Potentiale aller Atompaare berechnen. Diese Methode stammt ursprünglich aus der Proteinstrukturvorhersage [75, 76].

Keine dieser Bewertungsfunktionen kann jedoch die freie Bindungsenergie exakt vorhersagen und die Vorhersagequaltität der einzelnen Funktionen variiert für verschiedene Proteine, bei denen unterschiedliche Wechselwirkungen bei der Ligandbindung dominieren. Deshalb wurde vorgeschlagen, die Bewertung mehrerer Funktionen in einem sog. *consensus scoring* [77, 78] zu vereinigen.

2.4 Das Docking-Programm FlexX

FlexX von Rarey et al. [11] ist ein Computerprogramm zur Vorhersage von Protein-Ligand-Wechselwirkungen, das dem aktuellen Stand der Entwicklung entspricht. Ausgehend von einer dreidimensionalen Struktur eines Proteins und einem kleinen Ligandmolekül sagt FlexX die Geometrie des Protein-Ligand-Komplexes vorher und schätzt die Bindungsenergie ab. Die Ligandflexibilität wird dabei durch einen inkrementellen Aufbau des Liganden in der Bindetasche modelliert, während das Programm die Proteinstruktur als starr betrachtet und nur eine Proteinkonformation berücksichtigt. Zur Bewertung der Lösungen wird eine modifizierte empirische Bewertungsfunktion von Böhm [61] verwendet.

Evaluiert wurde FlexX anhand eines Testdatensatz mit 200 Komplexen [79] aus der PDB [80] und durch Blindvorhersagen bei CASP II (Critical Assesment of Structure Pre-

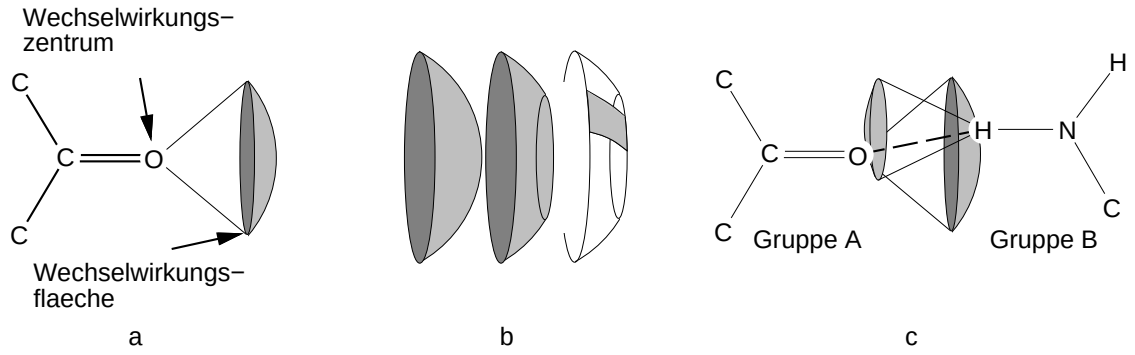


Abbildung 2.2: Modell der Wechselwirkungen. (a) Eine Wechselwirkungsgeometrie besteht aus einem Zentrum und einer Wechselwirkungsfläche. (b) Verschiedene Wechselwirkungsflächen: Kugelkappe, Kugelkappenstumpf und sphärisches Rechteck. (c) Eine Wechselwirkung kommt zustande, wenn die Wechselwirkungszentren jeweils näherungsweise auf den Wechselwirkungsflächen der Gegengruppe plaziert sind.

diction, zweite Runde) [30, 81]. Für den Testdatensatz findet FlexX in rund 70% der Fälle den richtigen Bindungsmodus, der bei gut 46% aller Liganden auch als beste Lösung bewertet wird. Die Vorhersage eines Protein-Ligand-Komplexes dauert dabei im Durchschnitt rund 90 Sekunden auf einer Sun Ultra 30.

Dieser Abschnitt faßt nur kurz die Modelle und wesentlichen algorithmischen Konzepte zusammen, auf denen FlexX beruht. Für eine detailierte Beschreibung der Methoden [11, 30, 82, 83, 84, 85], der Evaluierungen [79, 86, 87] und der aktuellen Weiterentwicklungen [78, 88, 89], wird auf die Originalarbeiten verwiesen.

2.4.1 Ligandflexibilität

Die konformative Flexibilität des Liganden wird durch eine diskrete Menge von bevorzugten Torsionswinkeln an den azyklischen Einzelbindungen und durch mehrere Konformationen für Ringsysteme modelliert. Die Torsionswinkel an Mehrfachbindungen, die Bindungslängen und die Bindungswinkel werden so verwendet, wie sie in der Eingabestruktur vorgegeben sind. Deshalb sollten energieminimierte Geometrien verwendet werden. Die Torsionswinkel stammen aus der MIMUMBA Datenbank, die ca. 900 Molekülfragmente mit einer zentralen Einzelbindung enthält und die von Klebe und Mietzner [90] aus der Cambridge-Structure-Database (CSD) [91] abgeleitet wurde. Durch diese Methode können jeder Einzelbindung bis zu 12 Torsionswinkel mit niedrigerer Energie zugewiesen werden.

Für Ringsysteme werden mehrere Konformationen mit dem Programm CORINA [92, 93, 94] berechnet. Die Zahl der Ringatome eines Elementarrings ist auf acht beschränkt. Größere Ringe werden als starr behandelt, und es wird die Eingabestruktur benutzt.

2.4.2 Wechselwirkungen

Das Modell der molekularen Wechselwirkungen, das FlexX verwendet, wurde von Böhm [95, 96] und Klebe [97] übernommen. Jeder Gruppe, die Wechselwirkungen ausbilden kann, weist man eine Wechselwirkungsgeometrie zu, die aus der Position des Zentrums und der Gestalt der Wechselwirkungsfläche besteht. Sie sind jeweils Teile einer Kugeloberfläche. Zwei Gruppen interagieren miteinander, wenn die Wechselwirkungszentren beider Grup-

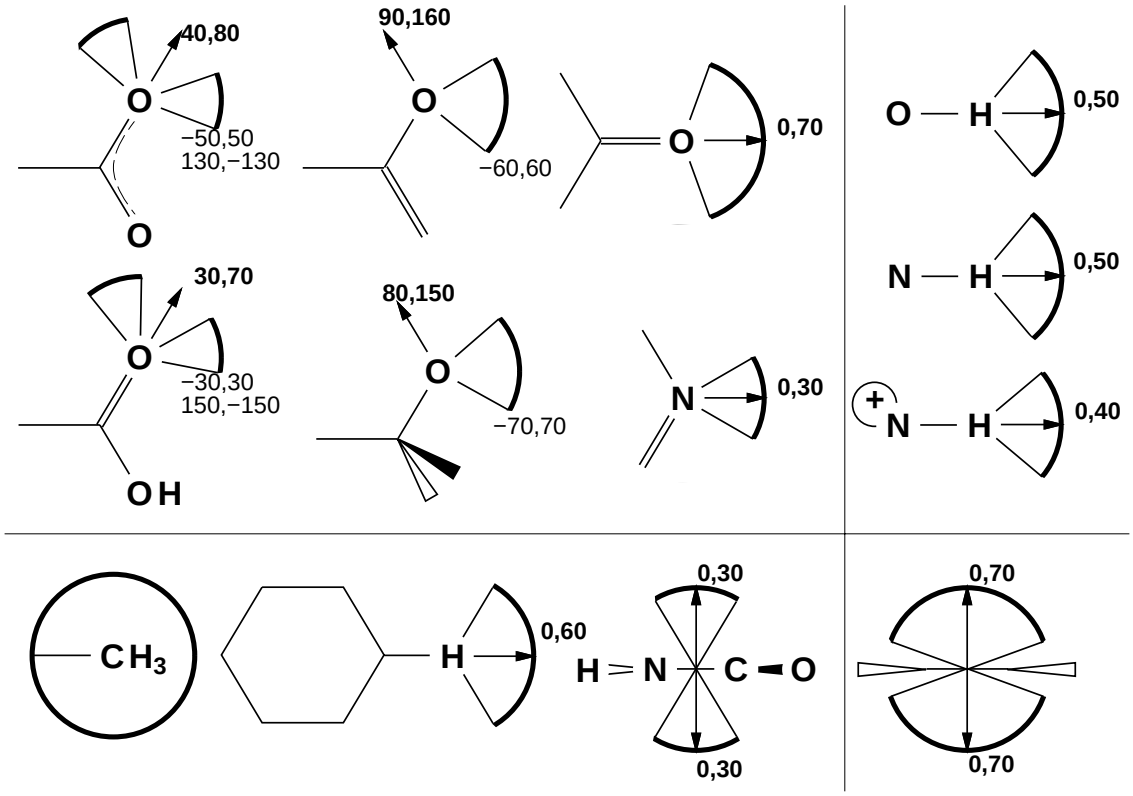


Abbildung 2.3: Wechselwirkungsgeometrien. Die Abbildung zeigt die in FlexX verwendeten Wechselwirkungsgeometrien: H-Akzeptoren (Stufe 3, links oben), H-Donatoren (Stufe 3, rechts oben), H-Atome in aromatischen Ringen, Methyl- und Amidgruppen (Stufe 2, links unten) sowie Zentren aromatischer Ringe (Stufe 2, rechts unten). Wechselwirkungsgeometrien von Metallen (Stufe 3) sind immer vollständige Kugeln und hier nicht dargestellt. Wie in Tabelle 2.1, in der die verwendeten Abstände und Energieparamenter der Geometrien angegeben sind, sind die kompatiblen Wechselwirkungstypen nebeneinander und die der verschiedenen Stufen übereinander angeordnet. Die fett gedruckten Winkelintervalle sind relativ zu den angegebenen Vektoren in der Zeichenebene, die übrigen senkrecht zu dieser Ebene gemessen. Der aromatische Ring (rechts unten) steht aus der Zeichenebene heraus. Die skizzierten Geometrien werden für verschiedene funktionelle Gruppen verwendet, so wird z.B. die Wechselwirkungsgeometrie des Carbonyl-Sauerstoffs auch für sp^2 Sauerstoffe an Schwefel und Phosphor benutzt.

pen (in etwa) auf der Wechselwirkungsfläche der Gegengruppe liegen (s. Abb. 2.2). Für jede Wechselwirkungsgeometrie ist eine Menge von Referenzatomen definiert, die jeweils ein (orthogonales) Bezugssystem festlegen, in dem die Lage der Geometrie eindeutig beschrieben werden kann. Abbildung 2.3 zeigt alle derzeit in FlexX verwendeten Wechselwirkungsgeometrien.

Die Wechselwirkungen lassen sich in drei Typen unterteilen: von Stufe 3 stark gerichtete Wechselwirkungen, wie z.B. Wasserstoffbrücken, bis zu Stufe 1 für ungerichtete Wechselwirkungen, wie z.B. hydrophobe Wechselwirkungen (vgl. Tab. 2.1). Die Wechselwirkungen höherer Stufen werden bevorzugt bei der Auswahl der Basisfragmente (s. Abs. 2.4.4) und bei der Plazierung der Liganden verwendet. Nur wenn es nicht ausreichend gerichtete Wechselwirkungen gibt, greift der Algorithmus auf solche niedrigerer Stufen zurück [85].

Kompatible Wechselwirkungstypen		Abstand	$\Delta G_{neutral}$	ΔG_{ionic}	Stufe
H-Akzeptor	H-Donor	1.9 Å	−4.7 kJ/mol	−8.3 kJ/mol	3
Metal Akzeptor	Metal	2.0 Å	−4.7 kJ/mol	−8.3 kJ/mol	3
Atome ar. Ring	Zentrum				
Methyl, Amide	ar. Ring	4.5 Å	−0.7 kJ/mol	–	2
aliphatische und aromatische					
Kohlenstoff- u. Schwefelatome		4.5 Å	–	–	1

Tabelle 2.1: Wechselwirkungstypen und Energieparameter von FLEXX. Für alle Paare von Wechselwirkungstypen sind der Abstand zwischen den interagierenden Atomen, die Energiebeiträge für neutrale und geladene Wechselwirkungen sowie die Stufe angegeben, zu der die Wechselwirkung gehört (s. Text).

FLEXX basiert auf der Annahme, daß lediglich Atome an der Oberfläche des Proteins mit einem Liganden in Wechselwirkung treten können. Deshalb werden nur Atomen an der Oberfläche im aktiven Zentrum des Proteins Wechselwirkungsgeometrien zugeordnet. Die Connolly Oberfläche [98] wird berechnet, um zu entscheiden, ob ein Atom des Proteins für Wasser zugänglich ist. Diese Oberfläche entspricht der Fläche, auf der sich der Mittelpunkt einer über das Protein gerollten Probekugel bewegt. Diese Kugel hat mit 1.4 Å in etwa den Radius eines Wassermoleküls.

Die Wechselwirkungsflächen lassen sich auf der Proteinseite durch eine endliche Menge sog. *Wechselwirkungspunkte* annähern. Da die Wechselwirkungsflächen bzw. die Wechselwirkungspunkte die Region beschreiben, in die Ligandatome plaziert werden sollen, können alle Punkte verworfen werden, auf die man kein Ligandatom plazieren kann, weil es dort mit dem Protein überlappen würde. Aus diesem Grund wird auf alle Wechselwirkungspunkte eine Testkugel mit dem van-der-Waals-Radius der potentiellen Gegengruppe gelegt und das Überschneidungsvolumen zum Protein bestimmt. Ist diese Überlappung zu groß, kann der betroffene Punkt ausgeschlossen werden.

2.4.3 Bewertungsfunktion

Die FLEXX-Bewertungsfunktion basiert auf einer Modifikation der empirischen Bewertungsfunktion von Böhm [61]:

$$\begin{aligned}
\Delta G &= \Delta G_0 + \Delta G_{rot} \times N_{rot} && (2.1)\\
&+ \Delta G_{hb} \sum_{neutral\ H-bonds} f(\Delta R, \Delta\alpha) && (2.2)\\
&+ \Delta G_{io} \sum_{ionic\ int.} f(\Delta R, \Delta\alpha) && (2.3)\\
&+ \Delta G_{aro} \sum_{aro\ int.} f(\Delta R, \Delta\alpha) && (2.4)\\
&+ \Delta G_{lipo} \sum_{lipo.\ cont.} f^*(\Delta R) && (2.5)
\end{aligned}$$

Der globale Grundterm $\Delta G_0(P, L)$ hängt nur vom Liganden ab und besteht aus einer Konstanten ΔG_0 und einem Faktor ΔG_{rot}, der proportional zur Zahl N_{rot} der drehbaren Bindungen im Liganden ist. Er berücksichtigt den Entropieverlust durch die Bindung des Liganden aufgrund der Einschränkung von drehbaren Bindungen.

Die weiteren Terme (2.2-2.5) summieren über alle paarweisen Wechselwirkungen. Der letzte Term der Funktion bewertet die Atom-Atom-Kontakte zwischen Protein und Ligand, die hydrophob oder zu dicht sind (Überlappung). Die Funktionen $f, f*$ sind heuristische distanz- und winkelabhängige Straffunktionen, die die Abweichung von der Idealgeometrie berücksichtigen. Die Parameter sind in Tabelle 2.1 angegeben.

2.4.4 Plazierungs-Algorithmus

Der Algorithmus zur Plazierung eines Liganden besteht aus drei Phasen (vgl. Abb. 2.4):

1. Auswahl eines bzw. mehrerer geeigneter Basisfragmente
2. Plazierung der Basisfragment in der Bindetasche
3. inkrementeller Aufbau des gesamten Liganden

Auswahl der Basisfragmente
Der zu plazierende Ligand wird durch Schneiden an allen azyklischen Einfachbindungen in Fragmente zerlegt, die als starr betrachtet werden. Unter Berücksichtigung ihrer möglichen Wechselwirkungen lassen sich aus dieser Menge von Fragmenten automatisch sog. *Basisfragmente* bestimmen, die als erste in die Bindetasche plaziert werden. Die Basisfragmente können aus einzelnen Fragmenten oder Kombinationen bestehen. Sie werden so ausgewählt, daß sie möglichst gut über das gesamte Molekül verteilt sind [30].

Basisplazierung
Die Basisfragmente werden mit Hilfe von zwei verschiedenen Algorithmen, die auf dem Kabsch-Algorithmus [9] basieren, im aktiven Zentrum plaziert. Der erste Algorithmus (triangle matching) überlagert Tripel von Wechselwirkungszentren des Basisfragments mit Tripeln von geeigneten Wechselwirkungspunkten in der Bindetasche, auf die mit Hilfe einer Hashtabelle (s. nächste Seite) effizient zugegriffen werden kann. Wenn ein Basisfragment weniger als drei Wechselwirkungszentren besitzt oder wenn die Zahl der Plazierungen kleiner als 100 ist, dann wird der zweite Algorithmus (line matching) gestartet. Er überlagert Paare von Wechselwirkungszentren mit Paaren von Wechselwirkungspunkten. Wegen der geometrischen Mehrdeutigkeit werden durch Rotation um die Achse, die durch die Wechselwirkungspunkte bzw. -zentren definiert ist, mehrere Plazierungen erzeugt. Beide Basisplazierungs-Algorithmen generieren typischerweise eine Vielzahl von Lösungen. Deshalb erfolgt eine Reduktion der Lösungsmenge durch Clashtests und Clustern der Plazierungen.

Inkrementeller Aufbau
Ausgehend von den verschiedenen Basisplazierungen wird der gesamte Ligand durch Anhängen der verbleibenden Fragmente Stück für Stück entsprechend der Torsionsdatenbank aufgebaut (s. Abs. 2.4.1). Nachdem ein neues Fragment hinzugefügt worden ist, wird nach neuen Wechselwirkungen gesucht und die Bewertungsfunktion verwendet, um die besten

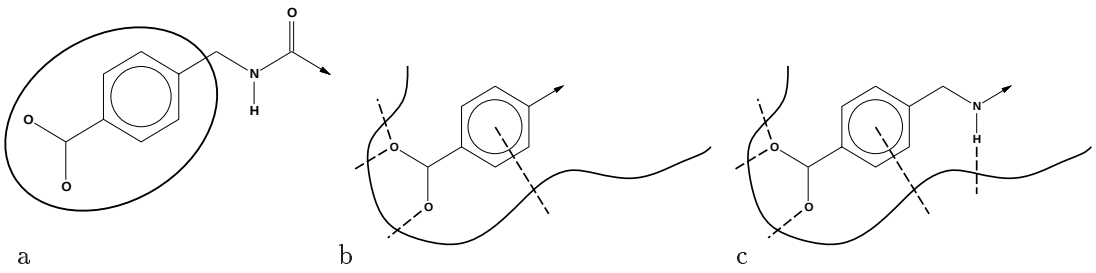


Abbildung 2.4: Inkrementeller Aufbau des Liganden. Die Plazierung des Liganden besteht aus drei Schritten: (a) Auswahl der Basisfragmente, (b) Basisplazierung und (c) inkrementeller Aufbau des Liganden in der Bindetasche.

(Teil-) Lösungen für den nächsten Aufbauschritt auszuwählen bzw. die vollständigen Vorhersagen zu sortieren. Plazierungen, bei denen die durchschnittliche Überlappung oder das Schnittvolumen einzelner Atome von Protein und Ligand zu groß ist, werden verworfen. Die maximale Anzahl von Lösungen, die in der nächsten Iteration berücksichtigt werden, ist $400 + 100n_f$, wobei n_f die Zahl der verschiedenen Basisfragmente angibt.

Hashtabelle
Um bei der Basisplazierung effizient auf die Wechselwirkungspunkte zugreifen zu können, werden in einer Vorverarbeitungsphase Paare von Wechselwirkungspunkten, deren Abstand in vorgegebenen Grenzen liegt, in einer Hashtabelle abgelegt. Adressiert werden die Paare über die Typen der Wechselwirkungen und den Abstand der Punkte, wobei man äquidistante Abstands-Intervalle, sog. *Buckets*, verwendet. In den Buckets werden die Punktpaare in sortierten, doppelt verzeigerten Listen abgelegt (hashing with chaining). Aus den Punktpaaren werden online die Punkttripel generiert, die zu den Anfragedreiecken des Basisfragments passen, dabei werden auch benachbarte Buckets berücksichtigt.

2.5 Definition des Docking-Problems mit Proteinflexibilität

Bisher wurde in diesem Kapitel der aktuelle Stand bei der Modellierung des molekularen Protein-Ligand-Docking-Problems zusammengefaßt und mit FlexX ein Docking-Programm vorgestellt, das diesem neusten Entwicklungsstand entspricht. Während man das Plazierungsproblem lösen kann und verschiedene, geeignete Modelle für flexible Liganden zur Verfügung stehen, bleibt die Proteinflexibilität bis jetzt bei den meisten Docking-Programmen unberücksichtigt.

Ziel dieser Arbeit ist es deshalb, ein Docking-Verfahren zu entwickeln, das alternative Proteinkonformationen schon bei der Plazierung verschiedener flexibler Liganden berücksichtigt. Dieses Problem wird hier als *Protein-Ligand-Docking-Problem mit Proteinflexibilität* bezeichnet und im folgenden genauer definiert. Dabei handelt es sich nicht um eine formale, in sich abgeschlossene Definition, sondern vielmehr um eine knappe übersichtliche Beschreibung des Problems, das hier gelöst werden soll.

Definition 1 (Konformation) *Unter einer Konformation versteht man die genaue räumliche Anordnung von Atomen oder Atomgruppen eines Moleküls. Verschiedene Konforma-*

tionen werden durch Rotation um Einfachbindungen erzeugt und lassen sich nicht zur Deckung bringen [13].

Definition 2 (Konformere) *Konformere oder Konformationsisomere sind verschiedene Konformationen eines Moleküls, die approximativ lokalen Energieminima entsprechen. Bei offenkettigen Verbindungen spricht man auch von Rotameren [13].*

Definition 3 (Bindungsmodus) *Ein Bindungsmodus beschreibt die Lage eines Liganden im aktiven Zentrum eines Proteins. Charakteristisch für den Bindungsmodus eines Liganden sind das besetzte Volumen in der Bindetasche, die abgedeckte Proteinoberfläche und die Aminosäurereste, mit denen der Ligand interagiert [5].*

Seien $\mathcal{L}(L)$ alle Konformationen eines Liganden L und $\mathcal{P}(P)$ alle Konformationen eines Proteins P sowie $l \in \mathcal{L}(L)$ und $p \in \mathcal{P}(P)$ einzelne Konformationen, dann läßt sich der Bindungsmodus in Form von Plazierung folgendermaßen modellieren:

Definition 4 (Plazierung) *Eine Plazierung $B(l,p)$ beschreibt die Anordnung eines Liganden in der Konformation $l \in \mathcal{L}(L)$ relativ zur Proteinkonformation $p \in \mathcal{P}(P)$.*

Diese Definition ist strikter als die des Bindungsmodus, denn leicht unterschiedliche Protein- und Ligandkonformationen mit geringfügig verschiedenen Plazierungen können denselben Bindungsmodus repräsentieren. Da ein optimaler Bindungsmodus aber minimale Energie haben sollte, kann man das allgemeine Docking-Problem wie folgt formulieren:

Definition 5 (Allgemeines Docking-Problem)

Gegeben: *ein Protein P und ein Ligand L*

Gesucht: *die Konformationen $l \in \mathcal{L}(L)$, $p \in \mathcal{P}(P)$ und ihre Plazierung $B(l,p)$ mit minimaler Energie*

Werden anstelle der kontinuierlichen Konformationsräume $\mathcal{L}(L)$ für den Liganden L und $\mathcal{P}(P)$ für das Protein P jeweils diskrete Teilmengen $\mathbf{L}(L) \subset \mathcal{L}(L)$ bzw. $\mathbf{P}(P) \subset \mathcal{P}(P)$ von Konformeren (Def. 2) verwendet, so ergibt sich das diskrete Docking-Problem:

Definition 6 (Diskretes Docking-Problem)

Gegeben: *eine Menge von Proteinkonformeren $\mathbf{P}(P)$ und eine Menge von Ligandkonformeren $\mathbf{L}(L)$*

Gesucht: *die Konformationen $l \in \mathbf{L}(L)$, $p \in \mathbf{P}(P)$ und ihre Plazierung $B(l,p)$ mit minimaler Energie*

Bei der Evaluation von Docking-Methoden geht man davon aus, daß der Bindungsmodus eines gemessenen Komplexes eine minimale Energie hat, das heißt, daß die Ligand- und Proteinkonformation sowie ihre Plazierung optimal sind. Die vorhergesagten Komplexe können deshalb über die RMS-Abweichung zur experimentellen Struktur bewertet werden. Das ist vor allem deshalb notwendig, weil die verwendeten Bewertungsfunktionen den experimentell bestimmten Komplex nicht unbedingt am besten bewerten.

Definition 7 (RMS-Abweichung) *Gegeben seien zwei im Raum orientierte Konformationen $m_1, m_2 \in \mathcal{M}(M)$ eines Moleküls M. Seien $c_a^{m_x}$ die Koordinaten des Atoms a der orientierten Konformation m_x, dann ist die RMS-Abweichung definiert als:*

$$RMSD(m_1, m_2) \quad = \quad \sqrt{\frac{1}{\mid \{a\} \mid} \sum_a (c_a^{m_1} - c_a^{m_2})^2}$$

In dieser Arbeit wird die Abweichung einer Plazierung über die RMS-Abweichung zwischen den Orientierungen des Liganden im vorhergesagten Komplex und in der Kristallstruktur (*Referenzstruktur*) bestimmt. Um Plazierungen eines Liganden in einer Proteinstruktur, aus der er nicht extrahiert wurde, bestimmen zu können, werden die Backbone-Atome dieser Struktur und der Originalstruktur mit minimaler RMS-Abweichung überlagert und die Referenzstruktur entsprechend transformiert.

Kapitel 3

Proteinflexibilität beim Docking

Bevor in diesem Kapitel die Behandlung der Proteinflexibilität beim Docking diskutiert wird und einige Ansätze dazu aus der Literatur vorgestellt werden, soll die Problematik am Beispiel des Proteins Aldose-Reduktase erläutert werden. Dieses Beispiel soll belegen, daß die Proteinflexibilität beim Docking kein rein akademisches Problem ist, sondern auch in der Praxis der Wirkstofforschung bei der Suche nach neuen Leitstrukturen von Bedeutung ist. Denn das Enzym Aldose-Reduktase ist auf der einen Seite sehr flexibel und auf der anderen Seite pharmazeutisch interessant, weil es einen Therapieansatz des Diabetes mellitus eröffnet, vor allem im Hinblick auf eine Verhinderung von Spätschäden.

3.1 Beispiel: Aldose-Reduktase

Der Diabetes mellitus ist die häufigste und zugleich bedeutendste Stoffwechselstörung [102]. Dabei kann es trotz Insulintherapie immer wieder zu unphysiologisch hohen Glukosekonzentrationen in Zellen kommen, deren Glukoseaufnahme insulinunabhängig erfolgt. Die Aldose-Reduktase katalysiert die Reduktion der überschüssigen Glukose zu Sorbitol. Die Anreicherung des polaren Sorbitols führt zu Zellschädigungen, die sich klinisch in den sogenannten diabetischen Spätschäden manifestieren. Inhibitoren der Aldose-Reduktase stellen somit einen Therapieansatz zur Vermeidung dieser gravierenden Komplikationen dar [100, 103]. Bisher bekannte Inhibitoren besitzen nicht die gewünschte Wirkstärke und verursachen Nebenwirkungen. Ausgehend von der dreidimensionalen Struktur der Aldose-Reduktase wird daher nach neuen Leitstrukturen gesucht.

Abbildung 3.1 zeigt die überlagerten aktiven Zentren von vier Aldose-Reduktase-Strukturen, und zwar den drei Kristallstrukturen 1ah0, 1ah3 und 1ah4 aus der PDB [80] sowie einem Homologiemodell [99, 100, 101]. Die Strukturen 1ah0, 1ah3 und das Modell enthalten die Inhibitoren Sorbinil, Tolrestat bzw. Zopolrestat, wohingegen das native Holoenzym der Struktur 1ah4 nur den natürlichen Cofaktor NADPH gebunden hat, der an einer anderen Stelle im Protein bindet und nicht dargestellt ist.

Zwei Bereiche des aktiven Zentrums der Aldose-Reduktase sind an der Bindung der Liganden beteiligt: eine hydrophobe Kontaktzone zwischen den Tryptophanen TRP 20, TRP 79, TRP 111 und TRP 219 sowie ein hydrophiler Bereich, in dem Wasserstoffbrückenbindungen zu den Aminosäuren THR 48, HIS 110 und TRP 111 ausgebildet werden können. Während der letztgenannte Bereich über die vier Strukturen geometrisch konserviert ist, treten im hydrophoben Teil unterschiedliche Seitenkettenkonformationen und sogar Ab-

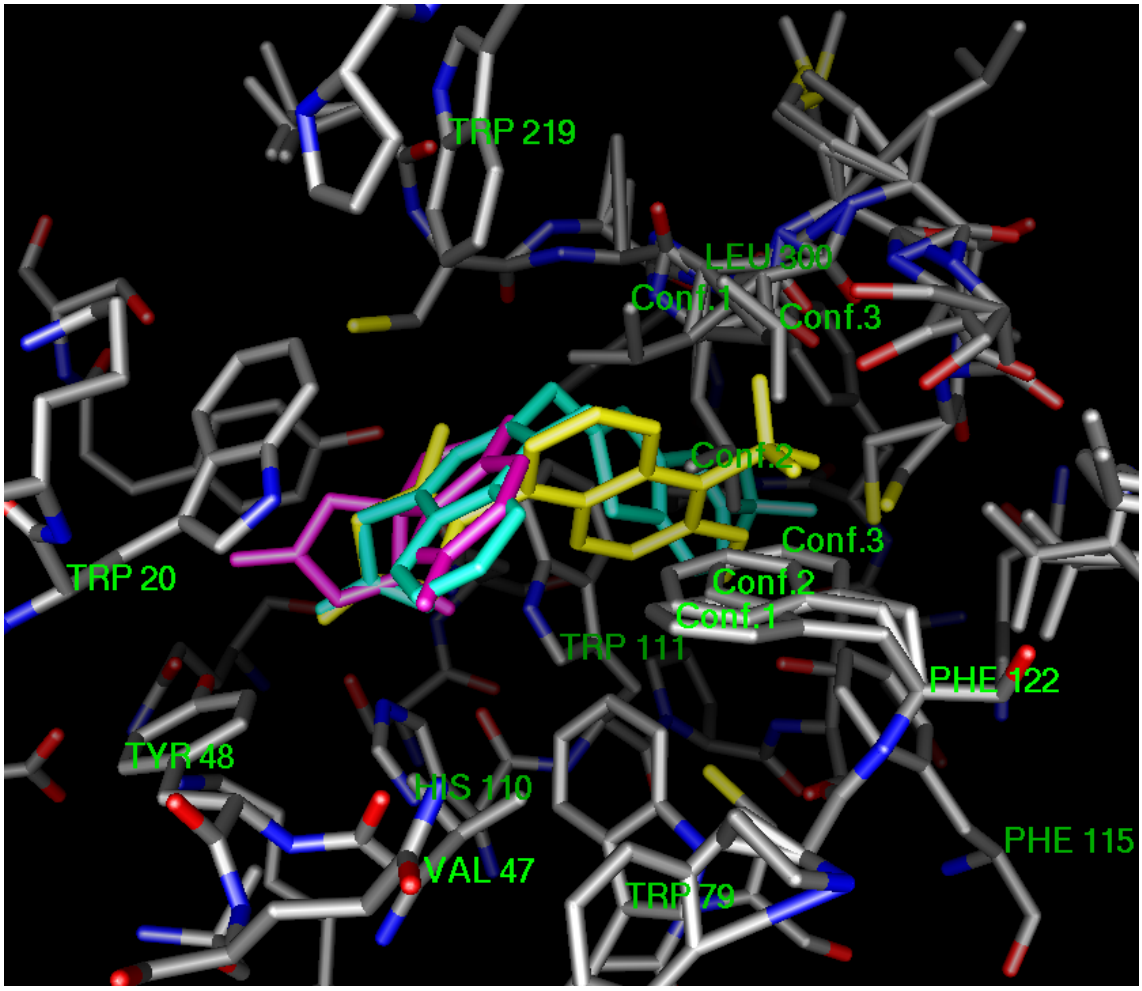


Abbildung 3.1: Beispiel: Aldose-Reduktase. Gezeigt sind die überlagerten aktiven Zentren der vier Kristrallstrukturen 1ah0, 1ah3 und 1ah4 aus der PDB [80, 99] sowie ein Homologiemodell [100, 101], die die Inhibitoren Sorbinil (1ah0, violett), Tolrestat (1ah3, gelb) bzw. Zopolrestat (Modell, türkis) gebunden haben. Die verschiedenen Inhibitoren induzieren sehr verschiedene Konformationen bei der Aldose-Reduktase.

weichungen im Backbone auf (Loop: ALA 299 – CYS 303). Hier befindet sich die Spezifitätstasche, der flexible Teil des aktiven Zentrums der Aldose-Reduktase. Diese Tasche wird von PHE 122 und LEU 300 verschlossen, wenn Sorbinil gebunden wird, und öffnet sich auf zwei verschiedene Weisen, um sich den Liganden Tolrestat bzw. Zopolrestat anzupassen. Die verschiedenen Proteinkonformationen variieren dabei so stark, daß die beiden größeren Liganden Tolrestat und Zopolrestat aus sterischen Gründen jeweils nicht in die anderen Strukturen passen. Diese Flexibilität könnte die große Zahl von sehr unterschiedlichen natürlichen Substraten der Aldose-Reduktase erklären [99].

Um Leitstrukturen für die Entwicklung neuer Inhibitoren der Aldose-Reduktase zu suchen, reicht also ein virtuelles Screening mit nur einer starren Proteinstruktur nicht aus. Zu leicht könnten dabei potentielle Liganden übersehen werden, weil sie nicht in die gegebene Proteinstruktur passen, sondern nur in eine andere Konformation des Enzyms.

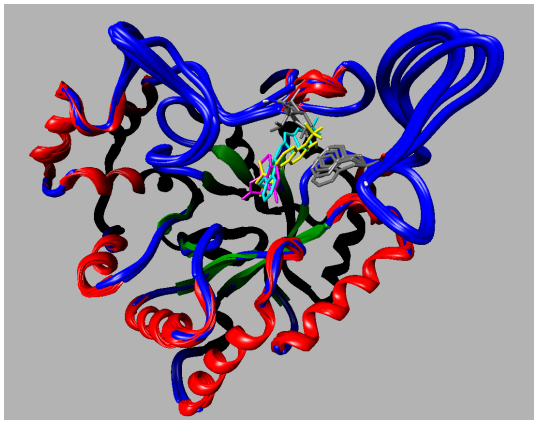


Abbildung 3.2: Konformationsänderungen. Die Abbildung zeigt den Backbone mehrerer experimentell bestimmter Aldose-Reduktase-Komplexe aus der PDB [80] sowie die beiden Seitenketten PHE 122 und LEU 300 und die Liganden Sorbinil (violett), Tolrestat (gelb) bzw. Zopolrestat (türkis) aus dem vorherigen Beispiel. Die dargestellten Seitenketten richten sich bei der Bindung der Liganden unterscheidlich aus. Oben rechts in der Abbildung sind außerdem unterschiedliche Loopkonformationen zu erkennen, die durch verschiedene Liganden induziert werden.

3.2 Behandlung der Proteinflexibilität beim Docking

Proteine sind nicht so flexibel wie Liganden, weil die charakteristische Faltung ihrer Aminosäureketten weitgehend die *Tertiärstruktur* dieser Makromoleküle bestimmt [14] (vgl. Abs. 2.1). Das gilt zwar streng genommen nur im Inneren konkaver Bindetaschen, aber Proteine werden deshalb von den diskreten Docking-Algorithmen im wesentlichen als starre Objekte modelliert [11, 20, 22, 25, 31, 33, 28]. Mehrere Argumente lassen diesen Ansatz, der in der Praxis oft zu guten Ergebnissen führt, sinnvoll erscheinen, obwohl die Funktion von Proteinen oft gerade auf Konformationsänderungen beruhen:

1. Die dichte Packung der Atome im Protein erschwert Konformationsänderungen aus sterischen und physiko-chemischen Gründen.

2. Die Spezifität des aktiven Zentrums wird vor allem durch seine Form bestimmt. Konformationsänderungen würden die Funktionalität des Proteins verändern und sind deshalb unwahrscheinlich.

3. Die Proteinstrukturen, die dem Docking zugrunde liegen, stammen meist aus bekannten Protein-Ligand-Komplexen, z.B. mit dem natürlichen Substrat, und liegen deshalb in einer Konformation vor, die für die Komplexbildung besonders geeignet ist.

4. Starre Proteinstrukturen lassen sich leichter modellieren und effizienter handhaben.

5. Im Zusammenhang mit einer ungenauen Bewertungsfunktion führt ein flexibles Protein zu vielen falsch-positiven Plazierungen.

Allerdings ist schon seit den fünfziger Jahren das Prinzip des *induced fit* [5, 17, 104, 105, 106, 107], also die Adaption der Proteinstruktur an die Bindung eines Liganden, bekannt (vgl. Abb. 3.2). Auch beim Vergleich von Komplexen eines Proteins mit verschiedenen Liganden wurde beobachtet, daß das Protein unterschiedliche Konformationen annimmt [108, 109, 110, 111], die, wie das Beispiel im nächsten Abschnitt zeigt (s.a. Abb. 3.1), so unterschiedlich sein können, daß ein Ligand, der an die eine Proteinkonformation bindet, aus sterischen Gründen in die andere Proteinkonformation nicht eingepaßt werden kann [99, 100]. Die Flexibilität der Proteine reicht dabei von der Rotation endständiger Gruppen

über die Konformationsänderung von Seitenketten und die Anpassungen einzelner Loops bis hin zu großen Bewegungen ganzer Domänen [112].

Ein weiterer Typ der Proteinflexibilität ergibt sich für modellierte Proteinstrukturen. Es gibt zahlreiche Sequenzen, für die bis jetzt keine dreidimensionale Struktur bekannt ist. Zwar können einige dieser Proteinstrukturen mit Hilfe homologer Strukturen modelliert werden [113, 114], aber dabei gibt es oft mehrere Möglichkeiten für die Plazierung von Seitenketten und Unsicherheiten bei der Modellierung von Loops [115, 116, 117, 118, 119]. Wenn z.B. eine drehbare Seitenkette eine Subtasche des aktiven Zentrums ausfüllt, in die ein Ligand gedockt werden müßte, so kann der richtige Bindungsmodus nicht vorhergesagt werden, ohne die Seitenkette wegzudrehen. Dieser Effekt kann unter Umständen sogar dazu führen, daß der Ligand überhaupt nicht in die modellierte Proteintasche eingepaßt werden kann.

Kleine Variationen der Proteinkonformation verursachen in der Realität nur geringe Unterschiede in der Bindungsaffinität zwischen Protein und Ligand, denn ein realer Komplex kann gegebenenfalls relaxieren. Beim virtuellen Docking können sich dagegen schon aufgrund leichter Variationen der Konformation große Unterschiede bei der vorhergesagten Bindungsenergie ergeben [109], z.B. wenn die Rotation einer endständigen Gruppe die Bildung einer Wasserstoffbrücke zwischen Protein und Ligand verhindert oder eine Überlappung der Bindungspartner verursacht [84]. Der Grund dafür ist unter anderem bei der Bewertungsfunktion zu suchen, die die Realität nur unzulänglich erfaßt. FlexX [11] erlaubt zwar, geeignete Torsionswinkel für solche Gruppen manuell vorzugegeben, eine automatische Einstellung für verschiedene Liganden ist aber nicht möglich. Der genetische Docking-Algorithmus, auf dem das Programm GOLD [31, 32] basiert, läßt die Einzelbindungen zu endständigen Wasserstoffakzeptoren bzw. -donatoren frei drehbar, so daß geeignete Konformationen eingestellt werden können. Der Rest des Proteins ist aber auch hier starr. Weitere Ansätze, die Proteinflexibiliät bei den diskreten Docking-Werkzeugen zu berücksichtigen, lassen einfach flexible Seitenketten unberücksichtigt [120], lösen Kollisionen zwischen Protein und Ligand flexibel auf [121], mitteln über mehrere bekannte Proteinstrukturen [122, 123] oder durchsuchen den Konformationsraum der Seitenketten [124, 125]. Auf diese Verfahren wird in den Abschnitten 3.3.2 u. 3.3.3 näher eingegangen.

Die volle Proteinflexibilität kann zur Zeit nur von solchen Docking-Methoden behandelt werden, die die Komplexbildung simulieren [46, 52, 126, 127, 128] (s. Abs. 3.3.4). Die Plazierungen und Bindungsenergien, die diese Methoden vorhersagen, sind oft genauer als die der diskreten Verfahren. Allerdings haben sie eine hohe Laufzeit im Stundenbereich und eignen sich daher vor allem für das genauere Studium weniger Protein-Ligand-Komplexe, nicht aber für das Durchmustern großer Datenmengen (Screening). Oft wird daher auch hier versucht, die Simulation dadurch zu beschleunigen, daß man Teile des Proteins fixiert [46, 126] oder sogar das ganze Protein starr hält [53, 129].

Hybridansätze [130, 131] kombinieren die Vorteile der diskreten Methoden und der Simulationen, indem sie zunächst mit einem der schnelleren Verfahren eine Menge von Grobplazierungen generieren, die anschließend energieminimiert werden. Solche Verfahren bleiben jedoch zeitaufwendig und sehr stark abhängig von der Vorhersagekraft der diskreten Plazierungsmethode. Denn die Energiemininierung versagt, wenn aufgrund der starren Proteinkonformation für einen Liganden keine geeignete Startposition gefunden wird.

3.3 Ansätze aus der Literatur

Es gibt eine Vielzahl von Software-Werkzeugen für das Docking-Problem, die hier nicht alle vorgestellt werden sollen, weil bereits eine Menge Übersichtsartikel zum strukturbasierten Wirkstoffdesign [132, 133, 134, 135, 136, 137, 138, 139, 140, 141] sowie zum Docking im speziellen [10, 12, 142, 143, 144, 145, 146] existieren und auch zahlreiche Vergleiche zwischen verschiedenen Docking-Programmen [77, 81, 147, 148, 149, 150] veröffentlicht worden sind. Stattdessen konzentriert sich dieser Abschnitt nur auf solche Ansätze aus der Literatur, die Proteinflexibilität beim Protein-Ligand bzw. Protein-Protein-Docking berücksichtigen. Einen kurzen Überblick über den aktuellen Stand der Behandlung von Proteinflexibilität im Wirkstoffdesign geben Carlson und McCammon [151].

Prinzipiell kann jedes Docking-Verfahren zur Berücksichtigung der Proteinflexibilität erweitert werden, indem es zwar starre Proteinstrukturen verwendet, aber die Liganden in mehrere verschiedene Proteinkonformationen plaziert. Daher werden zunächst zwei solche Docking-Studien vorgestellt und anschließend Docking-Verfahren aufgeführt, die die Flexibilität der Proteine per Design berücksichtigen. Die Docking-Verfahren lassen sich dabei in drei methodische Klassen unterteilen, die jedoch nicht völlig orthogonal zueinander sind:

1. diskrete Docking-Methoden, die den Konformationsraum diskretisieren, um darin gezielt nach Konformationen mit niedriger Energie zu suchen,

2. genetische Algorithmen, die die Evolution nachbilden, indem sie eine initiale Menge von Plazierungen durch Rekombination und zufällige Veränderungen bezüglich einer Fitness-Funktion optimieren,

3. Simulationsverfahren, die versuchen, die Energie einer oder mehrerer Startkonfigurationen durch zufällige oder gezielte lokale Veränderung der Konfiguration zu verringern.

Ein qualitativer Vergleich der vorgestellten Verfahren bezüglich Laufzeit und Vorhersagequalität schließt diese Übersicht ab.

3.3.1 Kreuz-Docking mit starren Proteinstrukturen

Es gibt eine Vielzahl von Proteinen in der PDB [80], die im Komplex mit verschiedenen Liganden vermessen worden sind. Bei diesen Proteinen kann man versuchen, die Liganden in die Proteinstrukturen zu docken, die im Komplex mit anderen Liganden bestimmt worden sind. Diese Proteinstrukturen werden im folgenden auch *fremde* Strukturen genannt. Wenn alle möglichen Kombinationen probiert werden, wird das Experiment als *Kreuz-Docking* bezeichnet. Kreuz-Docking-Experimente mit Docking-Programmen, die das Protein starr halten, weisen auf die Notwendigkeit hin, die Proteinflexibilität beim Docking zu berücksichtigen, weil bereits kleine Variationen der Proteinstrukturen zu großen Abweichungen bei der Plazierung der Liganden führen können.

- Kramer et. al [79] haben für sieben Proteine jeweils 3 – 9 Komplexe mit verschiedenen Liganden mit FlexX [11] untersucht. Sie kommen zu dem Ergebnis, daß die Liganden in den meisten Fällen auch in die fremden Strukturen plaziert werden können. In einigen Fällen haben diese Plazierungen sogar eine geringere RMS-Abweichung als die, die mit der Originalstruktur vorhergesagt werden. Es gibt aber auch Fälle, in

denen der Ligand in einen Teil der fremden Proteinstrukturen nicht gedockt werden kann. Die Autoren raten deshalb, mehrere Proteinstrukturen für ein Screening nach neuen Liganden zu verwenden.

- Murray et al. [109] führen mit ihrem Programm PRO_LEADS [38] für drei Proteine ein Kreuz-Docking mit jeweils 6 – 9 Komplexen durch. Dabei kann PRO_LEADS in 76% der Fälle die korrekte Plazierung auf Rang 1 vorhersagen, wenn die originale Proteinstruktur zugrunde gelegt wird. Die korrekte Lösung findet sich aber nur in 49% der Fälle auf Rang 1, wenn das Docking auf einer fremden Proteinstruktur beruht. Sie machen Seitenkettenbewegungen zum Teil in Verbindung mit C_α-Verschiebungen, Backbone-Bewegungen sowie die Positionsänderung von Metallatomen für diesen Unterschied verantwortlich. Diese Veränderungen können, selbst wenn sie teilweise nur sehr gering sind, die Bildung bzw. Bewertung von Wasserstoffbrückenbindungen erheblich beeinflussen.

3.3.2 Diskrete Algorithmen für das Protein-Ligand-Docking

Es gibt nur relativ wenige diskrete Docking-Algorithmen, die die Flexibilität des Proteins wenigstens zum Teil berücksichtigen. Die Ansätze reichen von der Vernachlässigung einzelner Seitenketten [120] bis hin zur vollständigen Suche im Konformationsraum der Seitenketten [124, 125]. Andere übertragen das Konzept der Ligandflexibilität auf das Protein, müssen dafür aber den Liganden starr halten [152], verwenden Mittelwerte über Ensembles von Proteinstrukturen [122, 123] oder verlassen sich auf eine Nachoptimierung des Komplexes [121]. Die verschiedenen Methoden werden im folgenden kurz beschrieben. Angaben über die Vorhersagequalität und die Laufzeit der Verfahren sind in Tabelle 3.1 gegenübergestellt.

- Sobolev et al. [120, 153, 154] plazieren starre Liganden mit ihrem Docking-Programm LIGIN, indem sie die Oberflächenkomplementarität zwischen Protein und Ligand optimieren. Dazu werden die Atome in acht Klassen eingeteilt und Paarkontakte zwischen den Oberflächen der Atome entsprechend der Klassen als legal oder illegal klassifiziert. Die Summe der legalen Kontakte abzüglich der Summe der illegalen Kontakte und eines repulsiven Terms für die Überlappung zwischen Protein und Ligand bestimmt die Komplementarität. Atome, die nicht klassifiziert werden können, werden bei der Berechnung der Komplementarität nicht berücksichtigt. Eine, wenn auch eingeschränkte, Proteinflexibilität wird dadurch modelliert, daß der Algorithmus automatisch bis zu zwei Seitenketten des Proteins, mit denen der Ligand kollidiert, von der Berechnung der Komplementarität ausschließen kann. Für diese Aminosäuren werden nur die Backbone-Atome und das C_β-Atom berücksichtigt. Daher läßt sich dieser Ansatz nicht auf alle Seitenketten verallgemeinern.

- Desmet et al. [125] docken Peptide, die aus bis zu 18 Aminosäuren bestehen, indem sie die Peptide, ausgehend von einer manuell ausgewählten Aminosäure, inkrementell in der Bindetasche des Proteins aufbauen. Dabei modellieren sie die Flexibilität des Backbones und der Seitenketten des Peptids sowie die der Seitenketten des Proteins in der Bindetasche mit Rotamerbibliotheken, die sie kombinatorisch nach der optimalen Konformation durchsuchen. Der Backbone des Protein bleibt unverändert. Die mögliche Translation und Rotation der Liganden werden durch ein sphärisches

Gitter mit 4.0 Å Radius und diskrete Rotationen um die drei Raumachsen erfaßt, die insgesamt zu 6636 relativen Orientierungen führen. Um den Suchraum einzuschränken, verwenden sie die Dead-End-Elimination [155, 156, 157, 158], die energetisch ungünstige Rotamere frühzeitig von der weiteren Suche ausschließt, und eine Greedy-Strategie, bei der jeweils im nächsten Aufbauschritt nur die Plazierungen berücksichtigt werden, deren Bindungsenergie nicht mehr als 5–7 kcal/mol geringer ist als die der bis dahin besten Lösung.

- Leach [124, 159] untersucht den Konformationsraum der Proteinseitenketten innerhalb eines bestimmten Energie-Cutoffs, um die Kombination von Aminosäureseitenketten und Ligandkonformation zu finden, die zu einem globalen Energieminimum führt. Er geht dabei von einer vorgegebenen Ligandposition aus und hält den Backbone starr. Die Suche im Konformationsraum basiert auf dem A*-Algorithmus [160, 161] mit Dead-End-Elimination [155, 156, 157, 158]. Die Untersuchung der Konformationen nahe des globalen Energieminimums zeigt, daß die meisten Unterschiede auf isolierten Konformationsänderungen einzelner Aminosäuren beruhen und nur wenige aus der gekoppelten Bewegung mehrerer Residuen resultieren.

- Sandak et al. [24, 25, 152, 162] modellieren Flexibilität mit dem Konzept der Gelenkbewegung (hinge-bending) – einer Methode, die zur Objekterkennung in der Bildverarbeitung für flexible Objekte verwendet wird, die aus mehreren starren Teilen zusammengesetzt sind. Dieser Ansatz wurde ursprünglich benutzt, um die Ligandflexibilität zu modellieren. Die Rollen von Ligand und Protein können aber vertauscht werden, weil das zugrundeliegende mathematische Problem symmetrisch ist. Die Flexibilität beider Bindungspartner kann jedoch nicht gleichzeitig modelliert werden. Die Gelenke müssen manuell definiert werden. Da ihre Anzahl auf eine geringe Zahl beschränkt ist, eignet sich dieser Ansatz nicht zur Modellierung von Seitenkettenflexibilität.

- Knegtel et al. [122] benutzen energie- und geometriegewichtete Mittel über Ensembles von Proteinstrukturen, um die Flexibilität von Proteinen zu beschreiben. Die Ensembles sind Mengen von Kristall- bzw. NMR-Strukturen aus der PDB [80]. Die Flexibilität des Liganden wird in diesem Ansatz nicht berücksichtigt. Die Kraftfeldterme der einzelnen Strukturen werden zu einem einzigen Gitter kombiniert, in welches die Liganden mit dem Programm DOCK 3.5 unter Verwendung von SPHGEN [163, 164, 165, 166] plaziert werden. Die repulsiven Potentiale werden nur berücksichtigt, wenn an einem Gitterpunkt die Potentiale aller Strukturen repulsiv sind. Obwohl mehrere Proteinstrukturen bei diesen Berechnungen verwendet werden, wird in diesem Ansatz die kombinatorische Natur des Problems vernachlässigt. Vielmehr werden die Potentiale der einzelnen Konformationen vermischt, was zu unrealistischen Maxima führen kann, und das aktive Zentrum wird durch die Vernachlässigung der repulsiven Terme vergrößert. Dennoch kann diese Methode für alle 15 Testliganden Plazierungen mit weniger als 2.0 Å RMS-Abweichung vorhersagen, was zum einen daran liegt, daß die Liganden starr gehalten werden, und zum anderen daran, daß Liganden mit einer hohen Bindungsaffinität gerade in solchen Bindungsmodi vorliegen, die mit Potentialmaxima verschiedener Proteinkonformationen korrespondieren. Das Verfahren ist aufgrund der Proteinbeschreibung, die während des Docking-Prozesses nicht mehr angepaßt werden muß, relativ schnell.

- Broughton [123] verwendet eine sehr ähnliche Methode wie Knegtel [122], um den FLOG [20] Ansatz zu erweitern. Er erzeugt mit Hilfe von Moleküldynamik-Simulationen eine größere Menge von Proteinkonformationen und faßt die Potentialgitter der einzelnen Strukturen zu einem neuen gewichteten Gitter zusammen, wobei die neuen Gewichte vom Mittelwert und der Standardabweichung der Gitterwerte abhängen. Dadurch wird der Einfluß der repulsiven Terme verringert, die nur in einigen Konformationen auftreten, ohne daß sie vollständig vernachlässigt werden, wie das bei Knegtel [122] der Fall ist. Die Maxima des neugewichteten Potentialgitters bilden die Referenzpunkte, für die analog zum FLOG Ansatz mit einem Cliquesuchalgorithmus passende Liganden aus einer Datenbank gesucht werden, die pro Ligand 250 verschiedene Konformationen enthält (Flexibases [21]).

- Schnecke et al. [121, 167, 168, 169] verwenden das Docking als schnelle Screening-Methode. Ziel ihrer Screening-Tools SPECTITOPE und SLIDE ist ein sehr schnelles und grobes Docking vieler Liganden aus einer großen Datenbank. Sie benutzen Multi-Level-Hashing, um das sog. Ankerfragment des Liganden mit den sog. Schablonenpunkten zur Deckung zu bringen, die das aktive Zentrum des Proteins beschreiben. Das Ankerfragment ist der Teil des Liganden, der, ausgehend von der in der Datenbank gegebenen Konformation des Liganden, auf die Schablonenpunkte paßt. Alle drehbaren Bindungen innerhalb des Ankerfragments werden starr gehalten. Daher hängt dieser Ansatz sehr stark von den Ausgangskonformationen der Liganden in der Datenbank und des Proteins ab, von dem die Schablonenpunkte abgeleitet werden. Die Flexibilität von Ligand und Protein wird als Nachoptimierung dadurch modelliert, daß Kollisionen zwischen dem plazierten Ligand und dem Protein durch disrekte Rotationen von Einzelbindungen im flexiblen Teil des Liganden bzw. der Seitenketten aufgelöst werden. Welche Bindungen dazu gedreht werden, wird mit Hilfe der Mean-field-Theorie [170, 171, 172] entschieden.

3.3.3 Genetische Docking-Algorithmen

Genetische Docking-Algorithmen basieren auf einer Menge (Population) von Plazierungen (Individuen), deren Eigenschaften in Form von Bit-Strings (Chromosomen) repräsentiert werden. Durch Rekombination (Cross Over) und zufällige Veränderungen (Mutation) der Chromosomen erzeugt man iterativ neue Populationen (Generationen). Dabei entscheidet eine Fitness-Funktion, welche Individuen ihre Eigenschaften in die nächste Generation vererben. Wieviele Iterationen des Verfahrens notwendig sind, hängt unter anderem von der Zahl der Freiheitsgrade ab, die optimiert werden. Als Faustregel gilt, je mehr Möglichkeiten es gibt, desto mehr evolutionären Zyklen sind erforderlich.

Zur Behandlung der Proteinflexibilität werden Freiheitsgrade für endständige Gruppen [31] oder wenige Seitenkettenkonformationen [173] eingeführt. Dieser Ansatz könnte zwar theoretisch zur vollen Seitenkettenflexibilität erweitert werden, in der Praxis würde das aber zu unakzeptabel hohen Laufzeiten führen, weil die Zahl der evolutionären Zyklen entsprechend der größeren Anzahl von Freiheitsgraden erhöht werden müßte.

Im folgenden werden zwei Docking-Verfahren vorgestellt, die auf genetischen Algorithmen beruhen. Da bis jetzt nur von Jones et al. [32] Angaben über die Laufzeit bei Testsystemen mit Proteinflexibilität veröffentlicht worden sind, ist nur diese Methode in der Übersicht (Tab. 3.1) enthalten.

- Jones et al. [31, 32] verwenden für ihr Docking-Programm GOLD einen genetischen Algorithmus, der neben allen drehbaren Bindungen innerhalb des Liganden auch die endständigen Wasserstoffakzeptoren und -donatoren des Proteins drehen kann.

- Olson et al. [37, 173] haben für das Programm AutoDock einen sog. Lamarck'schen genetischen Algorithmus entwickelt. Bei dieser Form des genetischen Algorithmus wird in jeder Generation die Energie der einzelnen Individuen minimiert und ihre neue Konformation in die genetische Beschreibung zurückübersetzt, bevor aus ihnen eine neue Generation erzeugt wird (Vererbung von gelernten Eigenschaften). Neben der Ligandkonformation können in der neusten, Version von AutoDock (4.0.i), die bisher nur den Programmentwicklern zugänglich ist, auch die Torsionswinkel einiger Seitenketten variiert werden. Ihre Zahl ist jedoch wegen der wachsenden Komplexität des Problems begrenzt [173].

3.3.4 Simulation des Protein-Ligand-Dockings

Simulationsverfahren gehen von einer oder mehreren Startkonfigurationen aus und versuchen, durch lokale Veränderung der Konfiguration die Energie des Systems zu verringern. Bei *Monte-Carlo*-Verfahren (MC) erfolgen diese Konfigurationsänderungen zufällig. Bei Metropolis-MC werden neue Konfigurationen mit niedrigerer Energie immer, solche mit höherer nur mit einer Wahrscheinlichkeit akzeptiert, die aus der temperaturabhängigen Boltzmann Statistik abgeleitet wird (*Metropolis-Kriterium* [174]). Wird dabei die Temperatur über die Zeit abgesenkt, so spricht man von *Simulated-Annealing*. Im Gegensatz dazu ergeben sich bei *Molekül̈dynamik-Simulationen* (MD) die lokalen Änderungen aus der Integration der Newtonschen Bewegungsgleichung. Oft werden diese beiden Simulationsverfahren auch in der Art kombiniert, daß eine zufällig generierte Konfigurationsänderung mit einem kurzen MD-Lauf minimiert wird, bevor man die neue Konformation energetisch bewertet.

Im Prinzip können die Simulationsverfahren die volle Proteinflexibilität behandeln. Oft wird die Flexibilität des Proteins aber nur für die Seitenketten und/oder das aktive Zentrum berücksichtigt. Dafür gibt es zwei Gründe. Zum einen reduziert diese Einschränkung die Laufzeit und zum anderen besteht so nicht die Gefahr, daß sich das Protein entfaltet. Da die Energieberechnungen dieser Verfahren im allgemeinen sehr viel aufweniger sind und die Konfigurationsänderungen in Abhängigkeit von der Zahl der Freiheitsgrade mehrfach iteriert werden müssen, haben die Simulationverfahren eine längere Laufzeit als die diskreten Docking-Algorithmen.

Hier wird nur eine Auswahl von Veröffentlichungen über Docking-Simulationen vorgestellt, die sich explizit mit der Modellierung von Proteinflexibilität beim Docking beschäftigen, um einen Eindruck von der Qualität und Komplexität dieser Verfahren zu geben. Auch für diese Methoden enthält Tabelle 3.1 Angaben über die Vorhersagequalität und die Laufzeit.

Monte-Carlo- / Simulated-Annealing-Verfahren mit lokaler Minimierung

- Abagyan's und Totrov's [49, 126] ICM-DOCK beruht auf einem Monte-Carlo-Verfahren und lokaler Energieminimierung mit dem ECEPP3 Kraftfeld nach jeder Konformationsänderung. Sie verwenden interne Koordinaten zur Beschreibung der Komplexe, die optimiert werden. Neben den sechs Freiheitsgraden für die Ligandposition,

können die Torsionswinkel im Liganden, die der Seitenketten sowie die des Backbones in Loopbereichen verändert werden. Dabei verwenden sie für die Seitenketten des Proteins Rotamere, die mit hoher Wahrscheinlichkeit auftreten [175].

- Apostolakis et al. [52] generieren zunächst 1000 zufällige Ligandplazierungen im aktiven Zentrum des Proteins. Diese Startkonformationen werden mit einem modifizierten CHARMM Kraftfeld minimiert, bei dem zunächst weichere Potentiale verwendet werden, die schrittweise zu ihrer normalen Form zurückkehren. Die so minimierten Strukturen werden unter Berücksichtigung von Solvatationseffekten bewertet und die besten 20 Lösungen werden noch einmal mit einem Monte-Carlo-Verfahren mit lokaler Energieminimierung relaxiert. Bei allen Minimierungen können alle Torsionswinkel im Protein sowie in Liganden verändert werden.

Moleküldynamik-Verfahren

- Luty et al. [46] führen ausgehend von 100 verschiedenen Startplazierungen Moleküldynamik-Simulationen von 20 ps Länge aus, um den Liganden zu plazieren. Dabei verwenden sie ein implizites Modell für die Solvatationsenergie und berechnen die Energiebeiträge des Proteinkerns auf einem Gitter vor. Die Aminosäuren im aktiven Zentrum sind vollständig flexibel und durch eine weniger flexible Pufferzone vom starren Proteinkern getrennt.

- Mangoni et al. [45, 128] bestimmen die Lage eines Liganden in der Bindetasche eines Proteins mit Hilfe einer 100 ps Moleküldynamik-Simulation mit dem Simulationspaket GROMOS87. Um die Simulation effizienter zu machen, verwenden sie dabei unterschiedliche Temperaturen für die Bewegung des Massenschwerpunktes des Liganden und der Wassermoleküle sowie für die internen Freiheitsgrade des Liganden und des Proteins.

- Zacharias et al. [176] schlagen vor, die Rezeptorflexibilität mit Hilfe der Normalmoden zu beschreiben, die kleine Eigenwerte in der Hesseschen Matrix haben, weil sie Bewegungen mit geringen Energiebarrieren entsprechen. Sie studieren diesen Ansatz an einem kleinen DNA-Ligand-Komplex, bei dem sie den Liganden starr halten. Der Ansatz eignet sich prinzipiell auch für das Protein-Ligand-Docking, jedoch ist die Berechnung und Diagonalisierung der Hesseschen Matrix für große Systeme mit vielen Atomen sehr aufwendig.

3.3.5 Proteinflexibilität beim Protein-Protein-Docking

Neben dem Protein-Ligand-Docking, bei dem ein kleines Molekül in der Bindetasche eines Proteins plaziert wird, gibt es ein verwandtes Docking-Problem, nämlich die Vorhersage von Protein-Komplexen, bei denen zwei Proteine miteinander einen Komplex bilden. Die computerbasierte Vorhersage solcher Komplexe wird im allgemeinen als Protein-Protein-Docking oder kurz als Protein-Docking bezeichnet. Einen Überblick über dieses Gebiet geben die Artikel [177, 178, 179, 180].

Natürlich treten auch bei der Bildung von Protein-Komplexen Konformationsänderungen auf [181]. Allerdings werden diese Konformationsänderungen abgesehen von einigen Monte-Carlo- [48, 182] und Simulated-Annealing-Ansätzen [183], die zum Teil Kenntnisse über die Lage der Bindestelle verwenden [48, 183], in der Regel erst bei der Optimierung

von Rigid-Body-Plazierungen berücksichtigt, weil das kombinatorische Problem ungleich komplexer als beim Protein-Ligand-Docking ist [180, 184].

Für die Rigid-Body-Plazierungen werden verschiedene Algorithmen verwendet. Einer sucht nach geometrischer und chemischer Komplementarität [185, 186, 187]. Die meisten approximieren aber die Proteine mit diskreten Punkten auf kubischen Gitter, die sie entweder direkt [184] oder fourier-transformiert korrelieren [188, 189, 190, 191, 192]. Dabei verwenden sie weiche Potentiale [172, 191, 192], sehr grobe Gitter [189, 190] oder lassen für besonders flexible Seitenketten eine größere Überlappung zu [184], um der Proteinflexibilität Rechnung zu tragen.

Die Optimierung der Seitenkettenkonformationen erfolgt anschließend mit ähnlichen Methoden, wie sie bereits beim Protein-Ligand-Docking vorgestellt wurden: Mean-Field-Methode [170, 171, 172, 192] und systematische Suche [193, 187] mit Dead-End-Elimination [155, 156, 157, 158, 187]. Althaus et al. [187] vergleichen dabei einen heuristischen Greedy-Ansatz mit einer vollständigen Branch-&-Cut-Suche basierend auf linearer Integer-Programmierung.

3.3.6 Vergleich der Protein-Ligand-Docking-Verfahren

Ein direkter Vergleich der vorgestellten Protein-Ligand-Docking-Methoden bezüglich der Laufzeit und der Vorhersagequalität ist schwierig, weil die Verfahren, die Testsysteme, die einfließende Vorabinformation und die verwendete Hardware zu unterschiedlich sind. So sind die Liganden bei einigen Verfahren flexibel, bei anderen werden dagegen zufällige, starre Konformationen benutzt oder es wird sogar die Konformation verwendet, die im Komplex vorliegt. Außerdem werden die Liganden teilweise in der Bindetasche vororientiert. Dies hat nicht nur erhebliche Auswirkungen auf die Komplexität des Problems und damit auf die Laufzeit des Algorithmus, sondern auch auf die zu erwartende RMS-Abweichung. Hier spielt auch die Definition des Referenzsystems und die Modellierung der Proteinflexibilität eine Rolle. Einige Methoden basieren z.B. auf Rotamerbibliotheken, aus denen nicht unbedingt die exakte Proteinkonformation reproduziert werden kann. Andere beruhen dagegen auf Ensembles von Proteinstrukturen, die die korrekte Konformation enthalten. Darüber hinaus werden die Strukturen manchmal energieminimiert, was ebenfalls zu leicht veränderten RMS-Abweichungen führt.

Da aber Aussagen über die Vorhersagequalität und die Laufzeit zum Vergleich der Verfahren und zur Einordnung einer neuen Methode unerläßlich sind, stellt Tabelle 3.1 die wichtigsten Kenndaten der vorgestellten Verfahren zum Protein-Ligand-Docking gegenüber. Dabei werden keine Einzelergebnisse verglichen, sondern es wird versucht, qualitative Aussagen über die Größenordnung der RMS-Abweichung (RMSD) und der Laufzeit zu machen, wobei man bei der letzteren natürlich auch die verwendete Hardware berücksichtigen muß. Die Tabelle enthält für jedes Verfahren die Zahl der Testsysteme, eine Charakterisierung der berücksichtigten Flexibilität, die Größenordnung der RMS-Abweichung (RMSD) und der Laufzeit sowie die verwendete Hardware.

Erstautor (für Vergleich)	Ref.	# Testsys. Lig.	Pro.	Flexibilität Lig.	Protein	RMSD Lig. [Å]	Laufzeit pro Ligand	Hardware (Prozessor)
Sobolev	[120]	4	4	starr	Sk. ignor.	0.4–1.4	3 min	DEC Turbo Laser
Desmet	[125]	4	4	flex.[1]	Rotamere	0.6–1.9	k.A.	SGI Indy (R4600PC)
Leach	[124]	2	2	flex.[2]	Rotamere	0.7–2.5	1 u. 9 Tage	SGI Indigo (R3000)
Sandak	[152]	3	3	starr	2 Gelenke	3.5–7.2[3]	2.5–39 min	SGI (R10000)
Knegtel	[122]	15	4 E	starr	gew. Mittel	0.4–1.6	1.5 min	k.A.
Broughton	[123]	∼7000	2 E	Ens.	gew. Mittel	k.A.[4]	2.5-17 s[6]	Cray J-90 (16 CPU's)
Schnecke	[121]	175000	3	Aufl. v. Kollisionen		k.A.[5]	3–350 ms[7]	Pentium II (450MHz)
Jones	[32]	100	100	flex.	endst. Grp.	66%<2.0	20x 12 min	SGI IndigoII(R4400)
Totrov	[126]	5(8)[8]	5(8)[8]	flex.	flex. Sk.[9]	1.8–7.8[8]	5–15 h	k.A.
Apostolakis	[52]	3	3	flex.	flex.	0.9–1.4	5 Tage	SGI Chall. (R4400)
Luty	[46]	1	1	flex.	flex.	∼1.0	∼1 Tag	SGI Challenge
Mangoni	[128]	1	1	flex.	flex.	k.A.[10]	6.5 h	SGI P.Chall.(R10000)

Tabelle 3.1: Vergleich der Protein-Ligand-Docking-Verfahren. Für jedes Verfahren sind der Erstautor und die Referenz, die Zahl der Testsysteme (Liganden u. Proteine bzw. Ensemble (E)), der Grad der berücksichtigten Flexibilität bei Liganden u. Proteinen , die Größenordnung der RMS-Abweichung der plazierten Liganden (RMSD), die Laufzeit, die für das Docking eines Liganden im Durchschnitt benötigt wird, sowie die verwendete Hardware angegeben. Eine genaue Beschreibung der Methoden findet man in den Abschnitten 3.3.2 – 3.3.4.
[1]) z.T vorplaziert; [2]) vorplaziert; [3]) des Proteins; [4]) Evaluation über Anreicherungsfaktoren; [5]) Evaluation über Rang bekannter Liganden; [6]) 5 – 32 CPU Stunden für die gesamte Datenbank; [7]) 9 min – 17 h für gesamte Datenbank, dabei wird ein Großteil der Liganden gar nicht erst gedockt; [8]) Ergebnisse von CASP2, bei der RMS-Abweichung wurden drei Komplexe nicht berücksichtigt, weil sie kovalent gebunden (2) bzw. auf einer falschen Ligandstruktur (1) beruhen; [9]) Verwendung von Rotameren; [10]) Evaluation über Wechselwirkungslängen.
Das Programm AutoDock [37, 173] fehlt in dieser Übersicht, weil bis jetzt noch keine Daten über die Lauftzeit bei Testsytemen mit Proteinflexibilität veröffentlicht worden sind. Der Ansatz von Zacharias et al. [176] fehlt ebenfalls, weil Angaben über die Gesamtlaufzeit des Verfahrens und die Qualität der Lösungen fehlen.

Kapitel 4

Modellierung der Proteinflexibilität beim Docking

Die charakteristischen Eigenschaften von Proteinen wurden bereits in Abschnitt 2.1 zusammengefaßt und die mögliche Proteinflexibilität sowie ihre Auswirkung auf das Docking-Problem im vorherigen Kapitel ausführlich diskutiert. Hier soll nun eine Modellierung der Proteinflexibilität vorgestellt werden, die auf der einen Seite die physiko-chemischen Eigenschaften, die beim Docking eine Rolle spielen, möglichst genau repräsentiert und auf der anderen Seite algorithmisch effizient zu handhaben ist. Dabei wird von einer gegebenen Proteinstruktur ausgegangen, deren dreidimensionale Struktur im großen und ganzen erhalten bleibt. Die Modellierung der Proteinflexibilität beschränkt sich somit auf die Seitenkettenflexibilität und die Anpassung von Loops an den Liganden.

Die Modelle der Ligandflexibilität und der Wechselwirkungen sowie die Bewertungsfunktion kann man von FlexX übernehmen. Sie wurden in Abschnitt 2.4 kurz skizziert und werden deshalb hier nicht noch einmal erläutert. Allerdings macht die Repräsentation der Proteinflexibilität Modifikationen an diesen Modellen notwendig, die in den nachfolgenden Abschnitten dargestellt werden.

4.1 Proteinflexibilität

Proteine haben eine definierte dreidimensionale Struktur. Sie können andere Moleküle aufgrund chemischer und geometrischer Komplementarität sehr spezifisch erkennen. Zwar können sich die Proteine unterschiedlichen Liganden anpassen, aber dabei muß ein charakteristisches Wechselwirkungsmuster erhalten bleiben, um die Spezifität zu gewährleisten. Deshalb basiert das hier vorgestellte Modell der Proteinflexibilität auf der Annahme, daß Konformationsunterschiede bei der Bindung verschiedener Liganden vor allem für einzelne Seitenketten oder in den Loop-Bereichen auftreten, während dagegen der Backbone-Verlauf im großen und ganzen erhalten bleibt. Diese Annahme wird einerseits durch experimentelle Beobachtungen gestützt [108] andererseits ist sie notwendig, um die Komplexität des Problems einzuschränken und die Anzahl der falsch-positiven Lösungen zu begrenzen.

Diese Arbeit geht davon aus, daß es eine Menge möglicher Proteinkonformationen gibt, die z.B. bei der Bindung verschiedener Liganden beobachtet wurden, die aufgrund von Moleküldynamik-Simulationen [123] sinnvoll erscheinen oder die alternative Homologiemodelle eines Proteins sind. Ein weitgehend identischer Backbone-Verlauf ist charakteristisch für

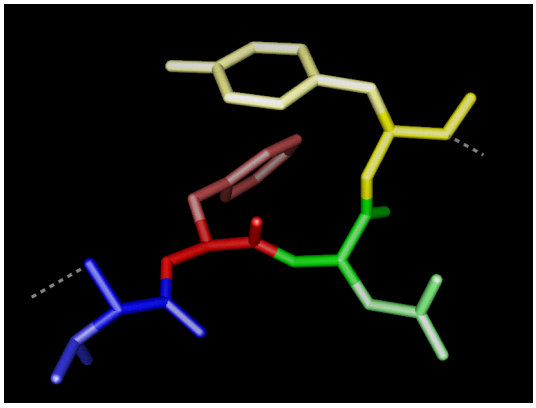


Abbildung 4.1: Segmentierung der Aminosäurekette. Das Bild zeigt die Segmentierung einer Aminosäurekette an einem Ausschnitt, der aus vier Residuen besteht. Die Backbone- und Seitenketten- Abschnitte sind jeweils ein einzelnes Segment und farblich anders eingefärbt.

eine solche Menge von Strukturen, die hier als *Ensemble* bezeichnet wird. Zwar handelt es sich dabei nicht um ein Ensemble im strengen thermodynamischen Sinn, aber eine solche Menge von Proteinstrukturen enthält vor allem Proteinkonformationen mit niedriger Energie, die auch ein thermodynamisches Ensemble im Gleichgewicht dominieren. In diesem Sinn sind diese beiden Begriffe verwandt.

Definition 8 (Ensemble) *Unter einem Ensemble wird im Rahmen dieser Arbeit eine Menge von Proteinstrukturen verstanden, die die Variabilität eines Proteins in Form alternativer Proteinkonformationen und Punktmutationen repräsentiert. Charakteristisch für die Strukturen eines solchen Ensembles sind verschiedene Seitenkettenrotamere bzw. Mutationen und unterschiedliche Loop-Konformationen, während der übrige Verlauf des Backbones weitgehend konserviert ist.*

Die globale Struktur des Proteins ist durch ein Ensemble von Proteinstrukturen festgelegt. Es kommt bei der Modellierung der Proteinflexibilität also nicht darauf an, völlig neue Proteinkonformationen zu generieren, sondern eine Menge gegebener Konformationen effizient zu verwalten, wobei neue Kombinationen von Teilkonformationen möglich sein sollten.

Analog zum Modell der Ligandflexibilität basiert die hier vorgestellte Modellierung der Proteinflexibilität auf einer geeigneten Zerlegung des Proteins in Teilstrukturen, die als starre Bausteine betrachtet werden, aus denen man das Protein wieder zusammensetzen kann. Eine Proteinstruktur wird jedoch anders als die Liganden nicht an allen azyklischen Einfachbindungen zerschnitten, sondern entsprechend der natürlichen Unterteilung der Proteine an den Peptid-Bindungen zwischen den Aminosäuren. Zusätzlich werden die einzelnen Aminosäuren in einen Backbone-Abschnitt und eine Seitenkette zerlegt. Abbildung 4.1 illustriert diese Segmentierung einer Aminosäurekette.

Definition 9 (Segment) *Ein Segment s umfaßt alle Atome, die zu einem Backbone-Abschnitt oder einer Seitenkette einer Aminosäure eines Proteins gehören.*

Da die Peptidbindungen, die die Backbone-Abschnitte miteinander verbinden, planar und starr sind (vgl. Abs. 2.1 u. Abb 2.1), gibt es nur einen möglichen Torsionswinkel für die Verknüpfung. Unterschiedliche Backbone-Verläufe ergeben sich durch verschiedene Konformationen der Backbone-Abschnitte.

Für die Seitenketten der verschiedenen Aminosäuren gibt es jeweils eine kleine Menge typischer Konformationen (Rotamere). Aus der Proteindatenbank kann man sog. *Rotamer-Bibliotheken* ableiten, die jedem Rotamer eine Wahrscheinlichkeit für sein Auftreten zuordnen [194, 195]. Diese Rotamer-Bibliotheken werden bei der Modellierung von Proteinstrukturen eingesetzt [114, 196, 197]. Sie können aber auch zur Generierung alternativer Proteinkonformationen verwendet werden.

Anstelle verschiedener Konformationen einer Seitenkette kann manchmal auch der Austausch einer Seitenkette durch eine andere Seitenkette beim Docking-Problem von Interesse sein. Eine solche Punktmutation kann durch alternative Seitenketten ebenfalls leicht modelliert werden.

Eine spezielle Konformation oder Mutation eines Segments wird hier *Instanz* genannt:

Definition 10 (Instanz) *Eine Instanz i repräsentiert eine spezielle Konformation oder Mutation eines Segments des Proteins.*

Ein Protein besteht in dem beschriebenen Modell aus einer Folge von Backbone-Segmenten, die jeweils mit einem Seitenketten-Segment verknüpft sind. Bei einer einzelnen Proteinstruktur liegt jedes dieser Segmente in einer speziellen Konformation vor, das heißt, jedes Segment hat genau eine Instanz.

Für das Docking von Liganden kann ein flexibles Protein in diesem Modell durch eine Menge disjunkter, alternativer Instanzen für jedes Segment beschrieben werden, die bei der Plazierung eines Liganden zu einer geeigneten Proteinkonformation kombiniert werden können. Die alternativen Instanzen werden von der sog. *vereinigten Proteinbeschreibung* verwaltet.

4.1.1 Vereinigte Proteinbeschreibung

In einem Ensemble von Proteinstrukturen haben viele korrespondierende Segmente, insbesondere die Backbone-Abschnitte, sehr ähnliche Konformationen. Um redundante Konformere zu entfernen, wird eine *vereinigte Proteinbeschreibung* erzeugt, die die Backbone-Abschnitte bzw. die Seitenketten mit ähnlichen Konformationen zusammenfaßt und nur unterschiedliche Instanzen als getrennte Alternativen behandelt (vgl. Abb. 4.2).

Da die Strukturen des Ensembles nach Voraussetzung (vgl. Def. 8) einen weitgehend identischen Backbone-Verlauf haben, können die Backbone-Atome der Ensemblemitglieder überlagert und die Instanzen der sich entsprechenden Segmente verglichen werden. Als Maß für die Ähnlichkeit zweier Instanzen dienen ihre RMS-Distanz sowie der maximale Abstand der korrespondierenden Atome. Sind sie beide kleiner als ein gegebener Schwellwert, werden die Instanzen zu einer neuen Instanz verschmolzen, wobei die Positionen der Atome aus dem Mittelwert der alten Atomkoordinaten entstehen. Beschreiben zwei Instanzen zwei Seitenketten verschiedener Aminosäurereste (Mutation), werden sie grundsätzlich getrennt behandelt. Die RMS-Distanz zwischen zwei unterschiedlichen Aminosäureresten kann deshalb als unendlich betrachtet werden.

Definition 11 (Verschmelzen von Instanzen) *Sei RMSD(i,j) die RMS-Distanz zwischen den Instanzen i, j (vgl. Def. 7), c_a^i die Koordinaten des Atoms a der Instanz i und $t \in \{1, 2, .., 20\}$ die verschiedenen Aminosäuretypen sowie $T(i)$ der Typ der Instanz i mit*

$$T(i) \quad = \quad \begin{cases} 0 & : \quad \textit{wenn i ein Backbone-Abschnitt ist} \\ t & : \quad \textit{wenn i eine Seitenkette einer Aminosäure vom Typ t ist} \end{cases}$$

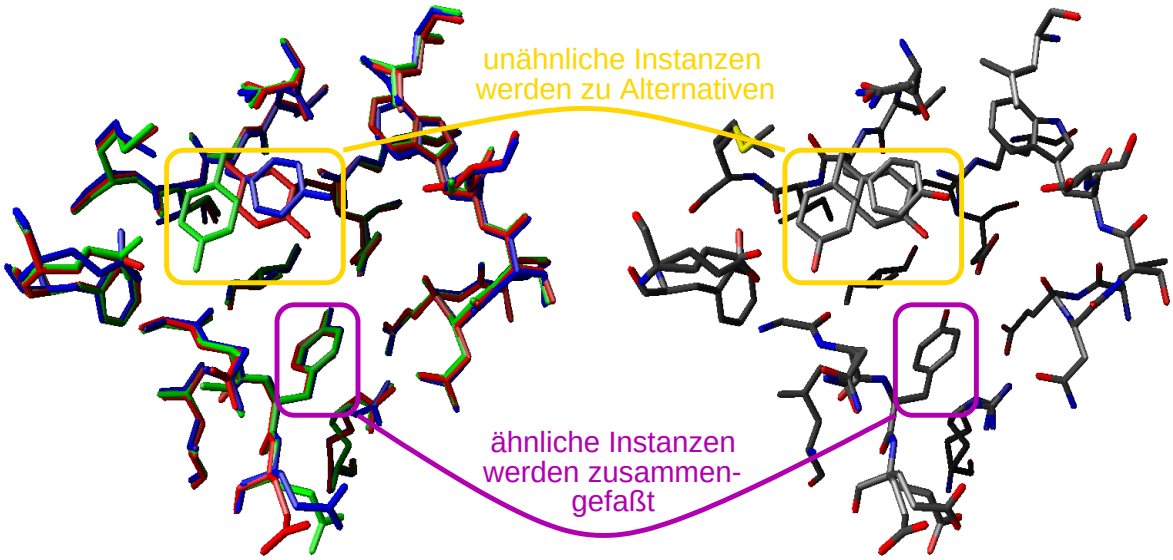


Abbildung 4.2: Vereinigte Proteinbeschreibung. Die vereinigte Proteinbeschreibung (rechts) wird aus den überlagerten Ensemblestrukturen (links) erzeugt, indem ähnliche Instanzen zusammengefaßt werden und deutlich verschiedene Konformationen als separate Alternativen behandelt werden.

Eine Menge Instanzen $I \subseteq I_s$ eines Segments s wird genau dann verschmolzen, wenn

$$\forall i,j \in I \qquad T(i) = T(j) \quad \wedge \quad RMSD(i,j) < \delta_{max} \quad \wedge \quad \max_a(\|c_a^i - c_a^j\|) < \delta_{max}$$

Die neuen Koordinaten ergeben sich aus dem Mittelwert der alten Koordinaten:

$$a_n^{neu} = \frac{1}{|I|} \sum_{i \in I} c_a^i \quad \forall a$$

Der Schwellwert δ_{max} darf nicht zu groß gewählt sein, weil nur solche Instanzen zusammengefaßt werden sollen, die so ähnlich sind, daß man sie durch eine einzige Instanz repräsentieren kann. In dieser Arbeit wird ein Schwellwert von $\delta_{max} = 1.0\,\text{Å}$ verwendet (vgl. Abs. 7.1). Die Verzerrung der verschmolzenen Instanzen und der Bindungen zwischen benachbarten Segmenten sollte deshalb so gering ausfallen, daß man sie beim Docking durch geringe Toleranzen bei der Suche nach Wechselwirkungen und ihrer Bewertung sowie bei der Überlappung zwischen Protein und Ligand ausgleichen kann.

Jedes Segment $s \in S$ der vereinigten Proteinbeschreibung entspricht einer Menge alternativer Instanzen I_s für die jeweiligen Backbone-Abschnitte bzw. Seitenketten. Die Instanzen $i_s \in I_s$ können entweder direkt, das heißt unverändert aus einer der zugrundeliegenden Ensemblestrukturen stammen, oder durch das Verschmelzen sehr ähnlicher Instanzen verschiedener Ensemblemitglieder entstanden sein. Deshalb enthält jedes Segment mindestens eine Instanz, wenn alle Strukturen des Ensembles für ein Segment sehr ähnliche Konformationen haben, und maximal N_{Ens} Instanzen, wobei N_{Ens} die Anzahl

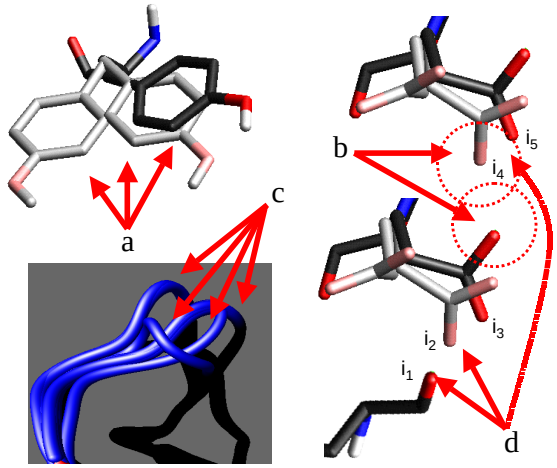


Abbildung 4.3: Inkompatibilität von Instanzen. Die Darstellumg illustriert die verschiedenen Typen von Inkompatibilität: (a) logisch: die Instanzen sind Alternativen voneinander; (b) geometrisch: die Instanzen überlappen; (c) strukturell: Instanzen, die zu verschiedenen Loop-Konformationen gehören, dürfen nicht kombiniert werden. (d) indirekt: $i_1 \not\sim i_2 \wedge i_2 \not\sim i_3 \wedge i_3 \not\sim i_4 \Rightarrow i_1 \not\sim i_4$.

der Ensemblestrukturen ist, falls alle korrespondierenden Segmente verschiedene Konformationen oder Mutationen aufweisen.

An dieser Stelle muß auf einen Sonderfall hingewiesen werden, der durch die Mutation eines Glycins entsteht. In diesem Fall muß man als Alternative für die Seitenketteninstanz(en) der Mutation eine leere Instanz erzeugen, die die nicht vorhandene Seitenkette des Glycins repräsentiert. Solche Instanzen werden in dieser Arbeit als *Pseudoinstanzen* bezeichnet. Pseudoinstanzen sind volumenlos und haben keine Koordinaten.

Für das aktive Zentrum der vereinigten Proteinbeschreibung müssen bei dieser Modellierung der Proteinflexibilität immer alle Instanzen eines Segments berücksichtigt werden, da man bei der Auswahl von Instanzen immer auch alle Alternativen betrachten muß. Darum gehören alle Segmente, in denen es mindestens eine Instanz gibt, die zu einem aktiven Zentrum einer Ensemblestruktur gehört, komplett zum aktiven Zentrum der vereinigten Proteinbeschreibung.

Die Zahl der Instanzen eines Segments der vereinigten Proteinbeschreibung kann zusätzlich durch Rotamere aus einer Bibliothek erweitert werden, wenn es nicht genügend alternative Konformationen im Ensemble von Proteinstrukturen gibt, oder nur eine einzige Proteinkonformation zu Verfügung steht.

4.1.2 Inkompatibilität von Instanzen

Nicht alle Instanzen der vereinigten Proteinstruktur können beliebig miteinander zu Proteinstrukturen kombiniert werden. Einige Instanzen schließen sich gegenseitig aus. Sie werden als *inkompatibel* bezeichnet. Instanzen, die gleichzeitig in einer Proteinkonformation auftreten können, sind dagegen *kompatibel* miteinander. Es lassen sich vier Typen von Inkompatibilität unterscheiden (vgl. Abb. 4.3):

- logische Inkompatibilität: zwei Instanzen sind Alternativen voneinander
- geometrische Inkompatibilität: zwei Instanzen überlappen sich
- strukturelle Inkompatibilität: zwei Instanzen einer Kette sind nicht verbunden
- indirekte Inkompatibilität: zwei Instanzen schließen sich durch Abhängigkeiten aus

Die *logische Inkompatibilität* ergibt sich aus der Definition der vereinigten Proteinbeschreibung: Zwei Instanzen des gleichen Segments sind inkompatibel, weil sie Alternativen voneinander sind, von denen jeweils nur eine verwendet werden darf.

Zwei Instanzen können auch nicht gleichzeitig in einer Proteinkonformation enthalten sein, wenn sie sich räumlich überlappen. Diese Instanzen sind *geometrisch inkompatibel*. Da die Atome einerseits als starre Kugeln modelliert werden und andererseits die Konformationen durch das Verschmelzen von Instanzen leicht verändert sein können, muß man jedoch ein geringes Überlappungsvolumen von 5.5 $Å^3$ tolerieren (vgl. Abs. 7.1). Instanzen von kovalent gebundenen Segmenten dürfen sich uneingeschränkt durchdringen. Diese Bindungen treten zwischen benachbarten Backbone-Segmenten, dem Backbone- und dem Seitenketten-Segment einer Aminosäure, beim Ringschluß der Proline und zwischen zwei Cysteinen (Disulfid-Bindung) auf. Pseudoinstanzen haben kein Volumen und überlappen deshalb mit keiner anderen Instanz.

Die einzelnen Aminosäureketten eines Proteins dürfen in einer möglichen Proteinkonformation nicht unterbrochen sein. Deshalb müssen zwei Instanzen einer solchen Kette jeweils direkt oder über eine Folge kompatibler Instanzen miteinander verbunden sein. Ist das nicht der Fall, sind diese beiden Instanzen *strukturell inkompatibel*. Die Verknüpfungen zwischen den Instanzen entsprechen entweder den Peptidbindungen im Backbone oder der $C_\alpha - C_\beta$-Bindung zwischen dem Backbone und einer Seitenkette. Da diese Bindungen ebenfalls durch das Verschmelzen leicht verzerrt sein können, muß auch hier eine Toleranz zur theoretischen Bindungslänge [198] erlaubt werden. Somit gelten zwei Instanzen i_s, j_t als direkt verbunden ($i_s \bowtie j_t$), wenn der Abstand der zu verknüpfenden Atome in etwa der theoretischen Bindungslänge entspricht (± 1.0 Å, vgl. Abs. 7.1). Die Instanzen i_s, j_t sind indirekt verbunden, wenn es eine Folge von direkt verbundenen Instanzen $k_m, .., k_n$ gibt, so daß $i_s \bowtie k_m \bowtie k_{m+1} ... k_n \bowtie j_t$. Durch diese Modellierung wird insbesondere vermieden, daß eine rekombinierte Proteinstruktur zwischen zwei alternativen Loop-Konformationen hin und her wechselt.

Zwei Segmente s, t sind verknüpft, wenn es mindestens ein Paar von Instanzen (i_s, j_t) gibt, die miteinander verbunden sind $i_s \bowtie j_t$. Für alle Seitenketteninstanzen sollte es mindestens eine Verbindung zu einer Backbone-Instanz geben, das heißt, alle Seitenkettensegmente sollten mit einem Backbone-Segment verknüpft sein. Dagegen können jedoch zwischen zwei Backbone-Segmenten Unterbrechungen auftreten, z.B. wenn ein Protein aus mehreren Aminosäureketten besteht oder wenn einzelne Atome oder Aminosäuren in der PDB-Datei fehlen. Eine Folge von verknüpften Backbone-Segmenten wird als *Segmentkette* bezeichnet.

Gibt es mehrere Segmentketten, läßt sich nicht ohne weiteres entscheiden, welche Instanzen der beiden Segmente der Kettenenden man verbinden darf. Da das Modell der Proteinflexibilität aber von einem im großen und ganzen ähnlichen Backbone-Verlauf ausgeht, sollten in der Regel alle Verknüpfungen zwischen den Instanzen dieser beiden Backbone-Segmente möglich sein. Deshalb werden die Instanzen zweier Backbone-Segmente, die nicht zu derselben Segmentkette gehören, immer als strukturell kompatibel betrachtet.

Diese technisch motivierte Definition hat vor allem die Aufgabe, Backbone-Unterbrechungen außerhalb des erweiterten aktiven Zentrums behandeln zu können. Insbesondere bietet sie die Möglichkeit, Teile des Proteinkerns wegzulassen, die nicht an der Ligandbindung beteiligt sind. Wenn eine Unterbrechung der Segmentkette jedoch in einem Loop innerhalb des aktiven Zentrums auftritt, führt diese Definition unter Umständen zu falschen

Ergebnissen, weil sie einen Wechsel zwischen verschiedenen Loop-Konformationen zuläßt. Backbone-Unterbrechungen innerhalb eines Loops sind jedoch in der Regel auf fehlende Atome oder Aminosäuren zurückzuführen und Strukturen, bei denen das aktive Zentrum unvollständig ist, sind ohnehin als Basis für Docking-Anwendungen ungeeignet.

Schließlich gibt es die *indirekte Inkompatibilität*, die aufgrund fortgesetzter Abhängigkeiten zwischen Instanzen auftreten kann, die ansonsten logisch, geometrisch und strukturell kompatibel sind. Diese Art der Inkompatibilität kann sehr weit reichen und ist für den allgemeinen Fall sehr aufwendig zu berechnen, weil es sich dabei nicht einfach um einen transitiven Abschluß der Inkompatibilität für eine Menge von Instanzen handelt. Um das darstellen zu können, müssen jedoch zunächst die Begriffe *gültige Proteinkonformation* und *Inkompatibilitätsgraph* eingeführt werden. Deshalb wird dieser Effekt erst in Abschnitt 4.2.2 näher erklärt.

Definition 12 (Inkompatibilität von Instanzen) *Zwei Instanzen i, j sind genau dann inkompatibel $i \not\sim j$, wenn sie aus logischen, geometrischen oder strukturellen Gründen oder aufgrund indirekter Abhängigkeiten nicht gleichzeitig in einer Proteinkonformation enthalten sein können.*

Definition 13 (Kompatibilität von Instanzen) *Zwei Instanzen i, j sind genau dann kompatibel $i \sim j$, wenn sie nicht inkompatibel sind:*

$$i \sim j \quad \Leftrightarrow \quad \neg(i \not\sim j)$$

4.1.3 Gültige Proteinkonformationen

Eine in der Realität mögliche Proteinkonformation besteht in dem hier beschriebenen Modell aus einer Menge von kompatiblen Instanzen. Diese Menge muß für jedes Segment genau eine Instanz enthalten, denn jedes Segment repräsentiert eine Auswahl von alternativen Konformationen, von denen genau eine in einer *gültigen Proteinkonformation* enthalten sein muß.

Definition 14 (Gültige Proteinkonformation) *Eine gültige Proteinkonformation I^* ist eine Teilmenge von kompatiblen Instanzen, die genau eine Instanz aus jedem Segment enthält:*

$$I^* \subseteq \bigcup_s I_s \quad mit \quad i \sim j \quad \forall i, j \in I^* \quad \wedge \quad |I^* \cap I_s| = 1 \quad \forall s \in S$$

Aus der vereinigten Proteinbeschreibung lassen sich gegebenenfalls eine Vielzahl unterschiedlicher gültiger Proteinkonformationen konstruieren. Die Menge aller dieser Konformationen ist die diskrete Teilmenge des gesamten Konformationsraums des Proteins, die von dem hier beschriebenen Modell der Proteinflexibilität bei der Plazierung eines Liganden erfaßt wird. Somit ist die Anzahl der berücksichtigten Proteinkonformationen im allgemeinen erheblich größer als die Zahl der Ensemblestrukturen.

4.2 Kompatibilitätsgraph

Die Abhängigkeiten zwischen den einzelnen Instanzen repräsentiert der sog. *Kompatibilitätsgraph*, in ihm sind je zwei paarweise kompatible Instanzen mit einer Kante verbunden:

Definition 15 (Kompatibilitätsgraph) *Die Instanzen bilden die Knoten des Kompatibilitätsgraphen. Die Kanten verbinden die Instanzen, die miteinander kompatibel sind:*

$$\begin{aligned} \mathcal{K} &= (I, E) \\ E &= \{\{i, j\} \subseteq I \mid i \sim j\} \end{aligned}$$

Da alle Instanzen einer gültigen Proteinkonformation paarweise kompatibel sein müssen, entspricht eine gültige Proteinkonformation im Kompatibilitätsgraphen einer Menge von vollständig miteinander verbundenen Knoten, einer sog. *Clique* [199], die genau eine Instanz aus jedem Segment enthalten muß:

Lemma 1 (Gültige Proteinkonformation im Kompatibilitätsgraphen) *Eine gültige Proteinkonformation I^* entspricht einer Clique im Kompatibilitätsgraphen mit genau einem Knoten aus jedem Segment s:*

$$\{i, j\} \in E \quad \forall i, j \subseteq I^* \quad \wedge \quad |I^* \cap I_s| = 1 \quad \forall s \in S$$

Für einige Betrachtungen ist es einfacher, den Graphen zugrunde zu legen, der aus der komplementären Kantenmenge $\bar{E} = \{\{i, j\} \subseteq I\}/E$ entsteht. Er wird in in dieser Arbeit als *Inkompatibilitätsgraph* bezeichnet. Da die Zahl der Kanten $|\bar{E}|$ im Inkompatibilitätsgraphen wesentlich kleiner ist als im Kompatibilitätsgraphen, eignet sich dieser Graph besser zur Visualisierung. Segmente werden in den Abbildungen durch Kreise um die Instanzen/Knoten dargestellt (s. Abb. 4.4).

Definition 16 (Inkompatibilitätsgraph) *Die Instanzen bilden die Knoten des Inkompatibilitätsgraphen. Kanten verbinden die Instanzen, die miteinander inkompatibel sind:*

$$\begin{aligned} \bar{\mathcal{K}} &= (I, \bar{E}) \\ \bar{E} &= \{\{i, j\} \subseteq I \mid i \not\sim j\} \end{aligned}$$

Aus einer Clique im Kompatibilitätsgraphen wird im Inkompatibilitätsgraphen eine *unabhängige Menge* von Knoten [199], das heißt, eine Menge von Knoten, zwischen denen keine Kanten existieren. Entsprechend wird eine gültige Proteinkonformation durch eine unabhängige Menge repräsentiert, die genau eine Instanz aus jedem Segment enthält.

Lemma 2 (Gültige Proteinkonformation im Inkompatibilitätsgraphen) *Eine gültige Proteinkonformation I^* entspricht einer unabhängigen Menge von Knoten im Inkompatibilitätsgraphen mit genau einem Knoten aus jedem Segment s:*

$$\{i, j\} \notin \bar{E} \quad \forall \{i, j\} \subseteq I^* \quad \wedge \quad |I^* \cap I_s| = 1 \quad \forall s \in S$$

Die Segmente selbst bilden Cliquen im Inkompatibilitätsgraphen bzw. unabhängige Mengen im Kompatibilitätsgraphen, weil alle Instanzen eines Segments paarweise inkompatibel sind. Deshalb enthält eine gültige Proteinkonformation bereits per Definition maximal einen Knoten aus jedem Segment. Die zusätzliche Bedingung besteht also nur darin, daß alle Segmente überdeckt werden müssen.

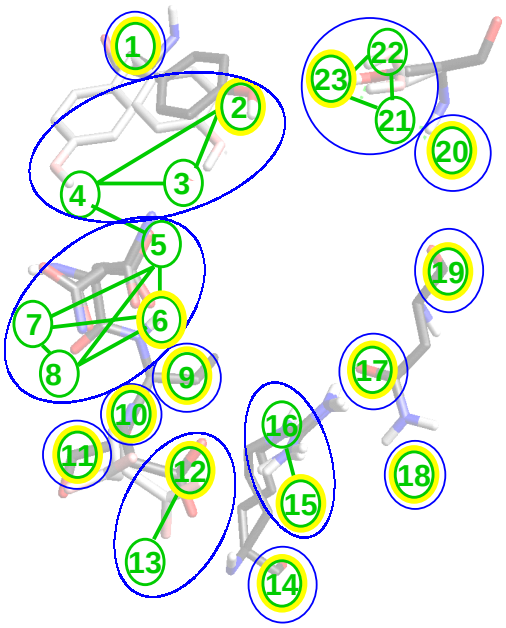


Abbildung 4.4: Inkompatibilitätsgraph. Die Abbildung zeigt den Inkompatibilitätsgraphen einer Proteinstruktur, die im Hintergrund zu sehen ist. Die Knoten und Kanten sind grün und die Segmente blau dargestellt. Eine gültige Proteinkonformation entspricht einer unabhängigen Menge (gelb).

4.2.1 Zusammenhangskomponenten des Inkompatibilitätsgraphen

Der Inkompatibilitätsgraph zerfällt wie jeder Graph in *Zusammenhangskomponenten* [199], das sind maximale zusammenhängende Teilgraphen (vgl. Abb. 4.5). Da es zwischen Knoten verschiedener Zusammenhangskomponenten nach Definition keine Kanten gibt, die Segmente aber Cliquen bilden, besteht jede Zusammenhangskomponente im Inkompatibilitätsgraphen aus vollständigen Segmenten, das heißt, die Instanzen eines Segments sind in keinem Fall auf zwei Zusammenhangskomponenten verteilt.

Lemma 3 (Vereinigung unabhängiger Mengen) *Es seien I alle Knoten und S alle Segmente im Inkompatibilitätsgraphen, der in eine Menge von Zusammenhangskomponenten $Z = \{z_1, z_2, .., z_n\}$ zerfällt. Seien I_s alle Instanzen des Segments s sowie S_z die Menge der Segmente, I_z alle Knoten und I_z^* eine unabhängige Menge der Zusammenhangskomponente z, die S_z abdeckt, dann bildet die Vereinigung aller dieser unabhängigen Mengen I_z^* wiederum eine unabhängige Menge I^* auf dem gesamten Graphen, die genau einen Knoten aus jedem Segment s enthält:*

$$\begin{aligned} &\forall z \;\; \forall \{i,j\} \subseteq I_z^* \quad \{i,j\} \notin \bar{E} \quad \wedge \quad |I_z^* \cap I_s| = 1 \quad \forall s \in S_z \\ &\Rightarrow \quad \bigcup_z I_z^* = I^* \quad \wedge \quad \forall \{i,j\} \subseteq I^* \quad \{i,j\} \notin \bar{E} \quad \wedge \quad |I^* \cap I_s| = 1 \quad \forall s \in S \end{aligned}$$

Beweis: Die Vereinigung der unabhängigen Mengen aller Zusammenhangskomponenten des Inkompatibilitätsgraphen ist selbst eine unabhängige Menge im gesamten Graphen, denn nach den Definitionen der unabhängigen Mengen und der Zusammenhangskomponenten gibt es keine Kanten zwischen den Knoten derselben unabhängigen Menge oder zwischen den Knoten verschiedener Zusammenhangskomponenten.

Wenn alle unabhängigen Mengen der Zusammenhangskomponenten genau einen Knoten aus jedem Segment der jeweiligen Zusammenhangskomponente enthalten, umfaßt die

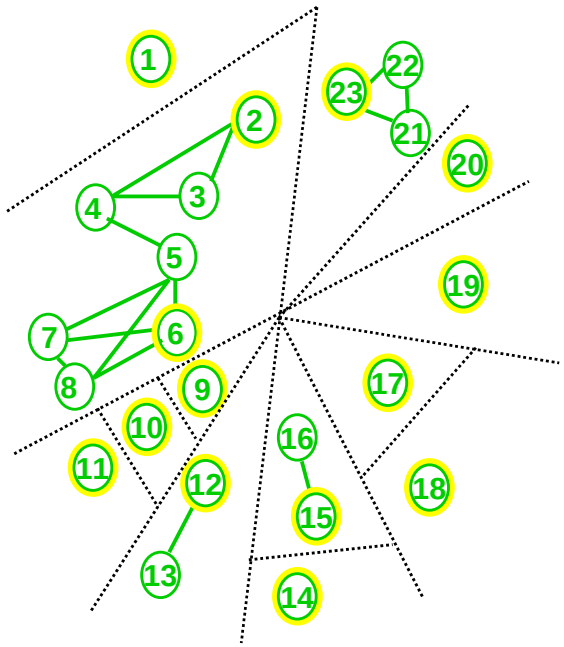


Abbildung 4.5: Zusammenhangskomponenten. Das Bild zeigt die Zerlegung des Inkompatibilitätsgraphen in Zusammenhangskomponenten. Die unabhängigen Mengen (gelb) der einzelnen Zusammenhangskomponenten bilden auch auf dem gesamten Graphen eine unabhängige Menge.

Vereinigung der unabhängigen Mengen aller Zusammenhangskomponenten auch für den gesamten Graphen genau einen Knoten aus jedem Segment. Denn gäbe es ein Segment, aus dem kein Knoten der Vereinigungsmenge stammt, dann würde entweder eine Zusammenhangskomponente in der Vereinigung fehlen oder es gäbe eine Zusammenhangskomponente, in der ein Segment nicht abgedeckt wäre. Beides ist nach Voraussetzung nicht erlaubt.

Es kann auch kein Segment geben, in das zwei Knoten gehören, denn dann hätte entweder bereits eine unabhängige Menge einer einzelnen Zusammenhangskomponente ein solches Segment besessen, was nach Voraussetzung nicht möglich ist, oder die beiden Knoten müßten aus verschiedenen Zusammenhangskomponenten stammen. Aber auch das ist unmöglich, weil alle Knoten eines Segments paarweise verbunden sind. Die Segmente fallen deshalb vollständig in eine Zusammenhangskomponente. □

Somit können die unabhängigen Mengen der einzelnen Zusammenhangskomponenten unabhängig voneinander bestimmt werden. Die Suche nach einer geeigneten gültigen Proteinkonformation bei der Plazierung eines Liganden in die vereinigte Proteinstruktur, die in dieser Modellierung der Suche nach einer geeigneten unabhängigen Menge im Inkompatibilitätsgraph entspricht, zerfällt durch die Zerlegung in Zusammenhangskomponenten in unabhängige Teilprobleme.

4.2.2 Indirekte Inkompatibilität in Zusammenhangskomponenten

Die indirekte Inkompatibilität (vgl. Abs. 4.1.2) entsteht durch die Abhängigkeit zwischen der Inkompatibilität und der Bedingung, daß eine gültige Proteinkonformation genau eine Instanz aus jedem Segment enthalten muß. Sie kann deshalb nur zwischen Instanzen innerhalb einer Zusammenhangskomponente auftreten. Abbildung 4.6 veranschaulicht das an einem sehr einfachen Beispiel.

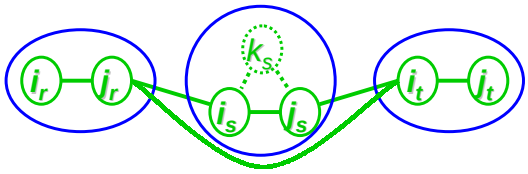


Abbildung 4.6: Indirekte Inkompatibilität. Die Darstellung zeigt drei Segmente $r, s, t \in S$ (blau) mit je zwei Instanzen i, j und ihre Inkompatibilitäten (grün) zuächst ohne die Instanz k_s. Wird k_s berücksichtigt, so entfällt die Kante $j_r \not\sim i_t$.

Gegeben sind drei Segmente $r, s, t \in S$ mit je zwei Instanzen i, j. Die Instanzen der Segmente sind jeweils logisch inkompatibel $i_x \not\sim j_x \ \forall x \in \{r, s, t\}$. Außerdem sind die Instanzen $j_r \not\sim i_s$ und $j_s \not\sim i_t$ jeweils paarweise inkompatibel (z.B. geometrisch). Daraus folgt, daß auch $j_r \not\sim i_t$. inkompatibel sind. Denn wenn j_r für eine gültige Proteinkonformation ausgewählt wird, kann i_s nicht benutzt werden. Da aber eine Instanz aus s verwendet werden muß, muß man j_s in die Proteinkonformation aufnehmen. Damit ist jedoch i_t ausgeschlossen. Hätte das Segment s noch eine weitere Instanz k_s, die mit allen Instanzen aus r und t kompatibel wäre, träte dieser Effekt allerdings nicht auf. Ferner sind i_r und j_t in jedem Fall kompatibel. Also entspricht die indirekte Inkompatibilität nicht einfach einem transitiven Abschluß der Inkompatibilität für eine Menge von Instanzen. Ein solcher transitiver Abschluß würde in einer Zusammenhangskomponente zu einer Clique führen.

Die Inkompatibilität bezieht sich jeweils nur auf Paare von Instanzen, sie ist abgesehen von der strukturellen Inkompatibilität nicht transitiv. Die indirekte Wirkung entsteht allein durch die zusätzliche Bedingung, daß genau eine Instanz aus jedem Segment ausgewählt werden muß. Aus diesem Grund ist die indirekte Inkompatibilität nur sehr aufwendig zu bestimmen, indem man den Inkompatibilitätsgraphen für jeden Knoten erneut analysiert. Dabei muß dieses Verfahren jeweils für alle Knoten wiederholt werden, wenn man eine neue Inkompatibilitätskante findet. In dieser Arbeit wird deshalb unter Ausnutzung der Transitivität nur die indirekte strukturelle Inkompatibilität bestimmt (vgl. Abs. 5.2).

4.3 Wechselwirkungen

Die Wechselwirkungen, die die Instanzen im aktiven Zentrum der vereinigten Proteinbeschreibung ausbilden können, werden mit den Wechselwirkungsgeometrien modelliert, die auch FlexX verwendet. Ebenso kann die Unterteilung in unterschiedlich stark gerichtete Wechselwirkungen und die Diskretisierung der Wechselwirkungsflächen in eine Menge von Punkten unverändert übernommen werden (vgl. Abs. 2.4.2).

Die Wechselwirkungspunkte repräsentieren mögliche Positionen von Ligandatomen. Deshalb werden Testkugeln mit den Van-der-Waals-Radien der zu plazierenden Atome an die Position der Punkte gelegt und das Überlappungsvolumen mit dem Protein bestimmt. Übersteigt dieses Volumen einen Schwellwert $V_{max} = 5.5\,\text{Å}^3$ (vgl. Abs. 7.1), so kann man an dieser Stelle kein Ligandatom plazieren. Der Punkt ist unzugänglich und kann verworfen werden.

In der vereinigten Proteinbeschreibung kann ein Wechselwirkungspunkt jedoch nur dann sicher verworfen werden, wenn es keine einzige gültige Proteinkonformation gibt, in der der Punkt zugänglich ist. Dies festzustellen ist aber sehr aufwendig, weil dazu alle Proteinkonformationen erzeugt, das heißt alle Cliquen des Kompatibilitätsgraphen aufgezählt werden müßten. Deshalb geht diese Arbeit von der folgenden vereinfachten

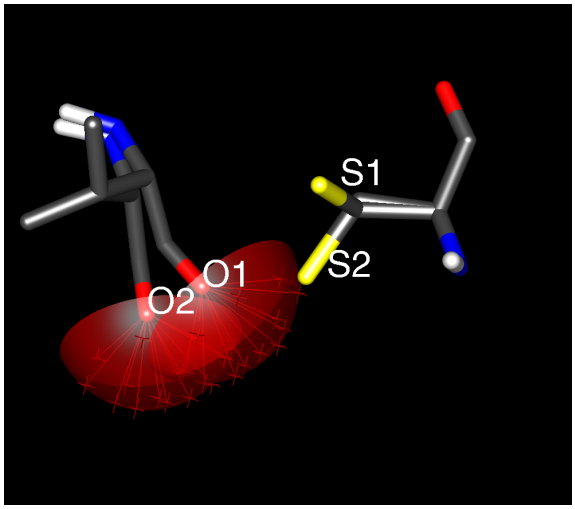


Abbildung 4.7: Unzugängliche Wechselwirkungspunkte. Ein Wechselwirkungspunkt gilt nur dann als unzugänglich, wenn es in einem Segment keine kompatible Instanz gibt, mit der er nicht überlappt. In der Abbildung sind die Wechselwirkungsgeometrien zweier Instanzen $O1, O2$ dargestellt. Ihre Wechselwirkungspunkte werden durch die jeweils alternative Instanz nicht beeinflußt, weil $O1$ und $O2$ inkompatibel sind. Außerdem sind die zwei Instanzen $S1$ und $S2$ eines anderen Segments dargestellt, die jeweils kompatibel zu $O1$ und $O2$ sind. Obwohl einige Wechselwirkungspunkte von $O1$ mit $S2$ überlappen, wird keiner dieser Punkte verworfen, weil es in diesem Segment auch die Instanz $S1$ gibt, die keinen der Punkte schneidet.

Definition aus, um zu entscheiden, ob ein Wechselwirkungspunkt verworfen werden kann (vgl. Abb. 4.7).

Definition 17 (Unzugänglicher Wechselwirkungspunkt) *Seien V_{j_t} das Volumen einer Instanz $j \in I_t$ des Segments $t \in S$ und $V_{p_{i_s}}$ das Volumen einer Testkugel, die an der Position des Wechselwirkungspunkts p_{i_s} einer Wechselwirkungsgeometrie, die zu der Instanz $i \in I_s$ des Segments $s \in S$ gehört. p_{i_s} ist genau dann unzugänglich, wenn*

$$\exists t \in S \quad \forall j' \in \{j \in I_t \mid j_t \sim i_s\} \quad V_{j'_t} \cap V_{p_{i_s}} > V_{max}$$

Die zu i inkompatiblen Instanzen dürfen nicht berücksichtigt werden, weil sie nach Definition der Kompatibilität nicht gleichzeitig mit i verwendet werden können. Falls p_i innerhalb der Instanz i selbst liegt, ist er unzugänglich, weil i die einzige Instanz des Segments s_x ist, mit der i kompatibel ist.

Für die Auswahl der Wechselwirkungspunkte nach dieser Definition ist es nicht notwendig, alle Proteinkonformationen zu erzeugen. Allerdings kann es theoretisch durch indirekte Inkompatibilitäten zwischen den Instanzen Wechselwirkungspunkte geben, die fälschlich für zugänglich gehalten werden (falsch-positiv). Diese Fälle sind in der Praxis aber eher selten und führen zu Plazierungen, bei denen der Ligand mit dem Protein kollidiert, so daß dann die Vorhersage verworfen wird. Deshalb führt auch dieser Fall nicht zu unzulässigen Lösungen.

4.4 Oberfläche

Nur die Atome, die an der Oberfläche eines Proteins liegen, können mit dem Liganden in Wechselwirkung treten. Nur diesen Atomen brauchen deshalb Wechselwirkungsgeometrien zugeordnet werden.

Der Anteil einer Instanz i an der Gesamtoberfläche eines Proteins hängt nicht allein von der Konformation von i ab, sondern auch von der Lage der räumlich benachbarten Instanzen $i_m, .., i_n \in I^*$ in einer gültigen Proteinkonformation. Deshalb kann dieser Oberflächenanteil von i bei verschiedenen Proteinkonformationen variieren und die Proteinober-

fläche kann nicht einfach additiv aus konstanten Beiträgen der Instanzen zusammengesetzt werden.

Aus diesem Grund läßt sich nicht ohne weiteres entscheiden, ob sich ein Atom einer Instanz in irgendeiner gültigen Proteinkonformation an der Oberfläche befindet. Um alle potentiellen Oberflächenatome zu finden, müßten alle zulässigen Cliquen des Kompatibilitätsgraphen aufgezählt und die Oberflächen der entsprechenden Proteinkonformationen berechnet werden. Auch wenn jeweils nur die Änderungen der Oberfläche neu berechnet würden, wäre dies ein sehr aufwendiges Verfahren.

Eine naheliegende Vereinfachung besteht darin, statt der Oberflächen aller möglichen Proteinkonformationen, nur die Oberflächen der zugrunde liegenden Ensemblestrukturen zu verwenden, die eine repräsentative Auswahl von Konformationen darstellen sollten. Allerdings können dabei mögliche Wechselwirkungen übersehen werden, und zwar solche an Atomen, die erst durch eine Rekombination der Ensemblestrukturen an die Oberfläche gelangen.

Eine andere Möglichkeit besteht darin, überhaupt keine Oberfläche zu berechnen, sondern stattdessen zunächst allen Atomen, die zum aktiven Zentrum gehören, geeignete Wechselwirkungsgeometrien zuzuordnen und die Auswahl von Wechselwirkungspunkten dem im vorherigen Abschnitt beschriebenen Überlappungstest zu überlassen.

Da sich bei praktischen Tests zwischen den beiden Ansätzen kaum Unterschiede zeigten, wird in dieser Arbeit aus folgenden Gründen der zweite Ansatz verwendet:

1. Er ist schneller. Denn es werden überhaupt keine Oberflächen berechnet und der Überlappungstest muß für den größten Teil der Wechselwirkungspunkte in jedem Fall durchgeführt werden.

2. Er ist unabhängig von den Ensemblestrukturen. Das heißt, es macht keinen Unterschied, wie die Konformationen der Instanzen auf die Ensemblestrukturen verteilt sind.

3. Er erzeugt unter Umständen einige wenige falsch-positive, aber keine falsch-negativen Wechselwirkungspunkte. Dadurch werden keine möglichen Wechselwirkungen übersehen. Wechselwirkungspunkte, die eigentlich unmöglich wären, führen dagegen zu Plazierungen mit Überlappungen zwischen Ligand und Protein, die dann nicht weiter berücksichtigt werden.

4. Er eignet sich auch für den Fall, daß die Instanzen der Segmente mit Hilfe von Rotamerbibliotheken angereichert werden. Da es für die Rotamere keine vollständigen Ensemblestrukturen gibt, wäre hier eine Oberflächenberechnung nicht ohne weiteres möglich.

4.5 Bewertungsfunktion

Bewertungsfunktionen, die für das Docking-Problem eingesetzt werden, bestehen im allgemeinen aus einem Grundterm $\Delta G_0(P, L)$, der global vom Protein P und vom Liganden L abhängen kann, sowie mehreren unabhängigen additiv verknüpften Termen für die verschiedenen Wechselwirkungstypen W. Die Beiträge der einzelnen Wechselwirkungen $w \in W$ zur Gesamtenergie werden dabei mit den Funktionen $f_W(w)$ gewichtet (vgl. Abs. 2.3).

$$\Delta G = \Delta G_0(P,L) + \sum_W \Delta G_W \sum_{w \in W} f_W(w) \tag{4.1}$$

Aufgrund der Additivität der Bewertungsfunktion können die Wechselwirkungen zwischen einem Liganden und den einzelnen Instanzen i einer gültigen Konformation I^* separat bestimmt und anschließend zur Gesamtbewertung addiert werden:

$$\Delta G_{ges} = \Delta G_0(P,L) + \sum_{i \in I^*} \Delta G_i \tag{4.2}$$

$$= \Delta G_0(P,L) + \sum_{i \in I^*} \left(\sum_W \Delta G_W \sum_{w_i \in W} f_W(w_i) \right) \tag{4.3}$$

Für eine gegebene Ligandkonformation ist die Proteinkonformation I^*_{opt} optimal, die diesen Ausdruck minimiert. Es kann auch mehrere solcher Konformationen I^*_{opt} geben, die jeweils einer unabhängigen Menge von Instanzen im Inkompatibilitätsgraphen entsprechen.

$$\Delta G_{min} = \Delta G_0(P,L) + \sum_{i \in I^*_{opt}} \left(\sum_W \Delta G_W \sum_{w_i \in W} f_W(w_i) \right) \leq \Delta G_{ges}(I^*) \quad \forall I^* \tag{4.4}$$

Die Zusammenhangskomponenten $z \in Z$ des Inkompatibilitätsgraphen können unabhängig voneinander nach optimalen Teilkonformationen, das heißt nach unabhängigen Mengen von Instanzen I^*_z mit minimaler Energie durchsucht werden (Lemma 3, Abs. 4.2.1). Die Gesamtenergie läßt sich als Summe der minimalen Bewertungen der einzelnen Zusammenhangskomponenten $z \in Z$ berechnen:

$$\Delta G_{min} = \Delta G_0(P,L) + \sum_{z \in Z} \min_{I^*_z} \left\{ \sum_{i \in I^*_z} \left(\sum_W \Delta G_W \sum_{w_i \in W} f_W(w_i) \right) \right\} \tag{4.5}$$

Da die Instanzen der verschiedenen Segmente nicht unabhängig voneinander sind, kann man jedoch nicht lokal für jedes Segment s die Instanz i mit minimaler Energie auswählen, um das globale Minimum zu finden. Die Summe dieser Energien ist aber eine untere Schranke für die minimale Bindungsenergie:

$$\Delta G_{us} = \Delta G_0(P,L) + \sum_{s \in S} \min_{i \in I_s} \left(\sum_W \Delta G_W \sum_{w_i \in W} f_W(w_i) \right) \leq G_{min} \tag{4.6}$$

Im Rahmen dieses allgemeinen additiven Bewertungschemas kann man verschiedene Bewertungsfunktionen realisieren (vgl. Abs. 2.3). In dieser Arbeit wird die empirische Funktion von Böhm [61] benutzt, die für FlexX modifizierte wurde (vgl. Abs. 2.4.3). Somit ergeben sich für Gleichung 4.2 die folgenden Terme:

$$\Delta G_0(P,L) = \Delta G_0 + \Delta G_{rot} \times N_{rot} \tag{4.7}$$

$$\Delta G_i = \Delta G_{hb} \sum_{neutral\ H-bonds} f(\Delta R, \Delta\alpha) \tag{4.8}$$

$$+ \Delta G_{io} \sum_{ionic\ int.} f(\Delta R, \Delta\alpha) \tag{4.9}$$

$$+ \Delta G_{aro} \sum_{aro\ int.} f(\Delta R, \Delta\alpha) \tag{4.10}$$

$$+ \Delta G_{lipo} \sum_{lipo.\ cont.} f^*(\Delta R) \tag{4.11}$$

Der globale Grundterm $\Delta G_0(P,L)$ hängt nur vom Liganden ab und besteht aus der Konstanten ΔG_0 und dem Faktor ΔG_{rot}, der proportional zur Zahl N_{rot} der drehbaren Bindungen im Liganden ist. Er berücksichtigt den Entropieverlust durch die Bindung des Liganden aufgrund der Einschränkung von drehbaren Bindungen.

Die folgenden Terme (4.8-4.11) summieren für alle Instanzen über alle paarweisen Wechselwirkungen zwischen der Instanz und dem Liganden. Dabei werden die gleichen Parameter verwendet wie für FlexX (s. Tab. 2.1, Abs. 2.4.3).

Kapitel 5

Algorithmische Konzepte zur Behandlung der Proteinflexibilität

Das Modell der Proteinflexibilität basiert zum einen auf der vereinigten Proteinbeschreibung und zum anderen auf dem Konzept der Inkompatibilität. Diese Datenstrukturen müssen erzeugt werden. Außerdem muß man die Algorithmen zur Plazierung der Liganden, die FlexX verwendet (s. Abs. 2.4.4), für die Behandlung alternativer Proteinkonformationen anpassen.

Dieses Kapitel beschreibt zunächst, wie die vereinigte Proteinbeschreibung und der Inkompatibilitätsgraph erzeugt werden. Danach stellt es ein Verfahren zur Reduktion der Hashtabelle vor, weil die vereinigte Proteinbeschreibung aufgrund der alternativen Instanzen sehr viele Wechselwirkungspunkte enthalten kann. Anschließend wird erläutert, wie die Instanzen für eine gültige Proteinkonformation während der Plazierung eines Liganden ausgewählt werden.

5.1 Aufbau der vereinigten Proteinbeschreibung

Die vereinigte Proteinbeschreibung ist die zentrale Datenstruktur zur Modellierung der Proteinflexibilität. Sie verwaltet die alternativen Instanzen der verschiedenen Ensemblestrukturen. Dabei werden sehr ähnliche Instanzen der einzelnen Ensemblestrukturen zusammengefaßt und nur deutlich verschiedene Instanzen getrennt behandelt. Um die einzelnen Ensemblestrukturen vergleichen zu können, muß man sie zunächst überlagern und, falls nötig, eine Symmetriekorrektur für einzelne Seitenketten durchführen. Für jedes Segment werden die entsprechenden Instanzen der Ensemblestrukturen separat geclustert. Die so erzeugten Cluster bilden die Instanzen der vereinigten Proteinbeschreibung.

5.1.1 Überlagerung der Proteinstrukturen

Das Modell der Proteinflexibilität basiert auf einem im großen und ganzen sehr ähnlichen Backbone-Verlauf der einzelnen Ensemblemitglieder. Deshalb kann man die Strukturen relativ leicht anhand der Backbone-Atome überlagern. Die erste Ensemblestruktur dient als Referenz, auf die man alle anderen Strukturen unabhängig voneinander mit minimaler RMS-Abweichung plaziert. Dabei werden die Ensemblestrukturen starr gehalten, um die Proteinkonformationen nicht zu verändern.

Die Backbone-Atome, die überlagert werden sollen, müssen manuell definiert werden. Im allgemeinen benutzt man den gesamten Backbone bzw. die Backbone-Atome im aktiven Zentrum. Wenn jedoch Atome in einzelnen Strukturen fehlen, muß man die entsprechenden Atome auch in den anderen Strukturen wegen der fehlenden Referenz auslassen. Da die Strukturen immer als Ganzes überlagert werden, können die fehlenden Atome in der vereinigten Proteinbeschreibung aber später durch die entsprechenden Atome aus anderen Strukturen ersetzt werden.

Mathematisch besteht die Aufgabe darin, zwei Mengen von Punkten so zu überlagern, daß die RMS-Abweichung minimal ist. Kabsch [9] konnte dieses Problem analytisch lösen. Man kann diese analytische Lösung entweder einmal berechnen, um die Überlagerungen zu erhalten, oder man iteriert diese Berechnung, wobei alle Paare von Atomen, deren Abstand größer als ein benutzerspezifizierter Schwellwert ist, im nächsten Schritt ignoriert werden. Wiederholt wird dieses Verfahren, solange sich die Menge der ausgewählten Atompaare von einem Schritt zum nächsten ändert und eine maximale Anzahl von Schritten noch nicht erreicht ist. In der Regel konvergiert die Berechnung aber bereits nach wenigen Schritten. Diese iterative Prozedur betont die Unterschiede und verbessert die Paßgenauigkeit der konservierten Regionen der Strukturen [200, 201].

Statt der unabhängigen Überlagerungen der Ensemblestrukturen auf die Referenzstruktur könnte man auch versuchen, eine global optimale Überlagerung zu finden, bei der die paarweise RMS-Abweichung zwischen allen Paaren von Ensemblestrukturen minimal ist. Für die Berechnung einer solchen Abbildung stand jedoch keine geschlossene Lösung zur Verfügung, sie hätte deshalb nur iterativ approximiert werden können. Aufgrund der Annahme, daß die Backbone-Verläufe der Ensemblestrukturen insgesamt recht ähnlich sind, sollte sich eine solche global optimierte Überlagerung aber nicht wesentlich von der Superposition auf eine Referenzstruktur unterscheiden. Wegen der beim Clustern der Instanzen und beim Docking der Liganden verwendeten Toleranzen ist deshalb eine global optimale Überlagerung der Ensemblestrukturen als Basis für die vereinigte Proteinbeschreibung nicht erforderlich.

5.1.2 Symmetriekorrektur

Die Schweratome der Aminosäuren des Proteins sind in der PDB-Datei mit eindeutigen Namen versehen. Die Backbone-Atome heißen C, N, O und CA für das C_α-Atom. Die Namen der Seitenkettenatome bestehen aus dem Elementnamen (C,N,O,S) gefolgt von der Transliteration des griechischen Buchstabens, der die Entfernung vom C-Atom der Carbonyl-Gruppe angibt ($\beta \rightarrow$B, $\gamma \rightarrow$G, $\delta \rightarrow$D, $\epsilon \rightarrow$E, $\zeta \rightarrow$Z). Bei Verzweigungen oder in Ringen in den Seitenketten gibt es mehrere Atome, die gleichweit vom C_α-Atom entfernt sind. In diesen Fällen werden die gleichweit entfernten Atome mit derselben Nummer versehen. Abbildung 5.1 zeigt diese Benennung der Seitenkettenatome für alle Aminosäuren.

Die Aminosäuren ARG, ASP, GLU, LEU, PHE, TYR und VAL sind symmetrisch, das heißt, man erhält chemisch äquivalente Strukturen, wenn man konsistent alle Zahlen der Zweige bzw. Ringhälften vertauscht. Da die Zuordnung der Atome aus den verschiedenen Ensemblestrukturen anhand dieser Namen erfolgt, können beim Clustern symmetrische Instanzen entstehen, die sich nur durch die Numerierung der Zweige, aber nicht in ihren chemischen Eigenschaften unterscheiden. Darum werden die Zweige der symmetrischen Seitenketten der überlagerten Ensemblestrukturen intern entsprechend der folgenden Regel eindeutig numeriert, um solche Artefakte zu vermeiden:

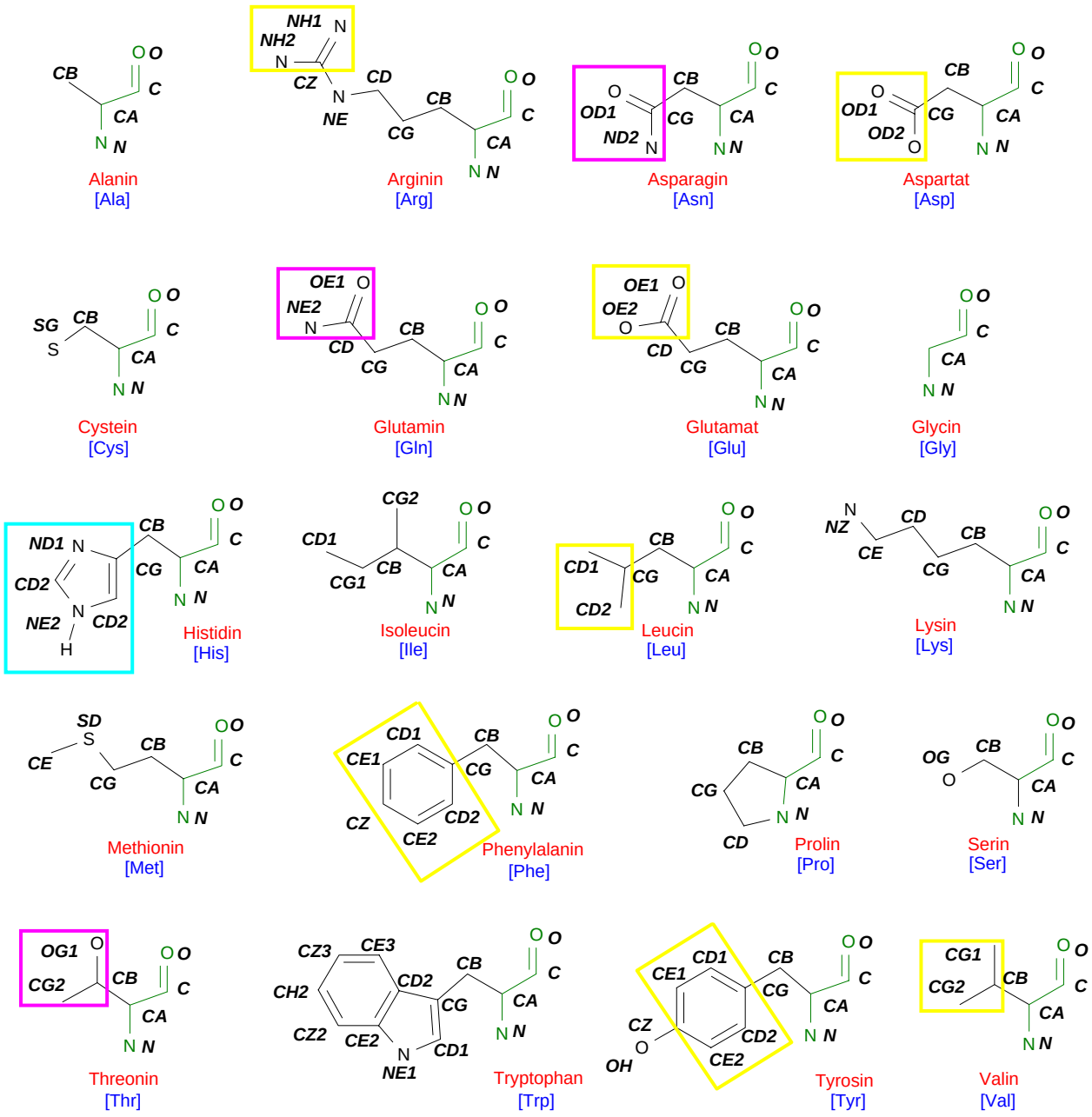


Abbildung 5.1: PDB-Nomenklatur der Aminosäuren. Die Abbildung zeigt für die 20 Aminosäuren, aus denen sich natürliche Proteine zusammensetzen, die eindeutigen Atombezeichnungen (fett schwarz), die in der PDB [80] verwendet werden. Unter den Strukturen stehen in rot die Namen der Aminosäuren und in blau die üblicherweise verwendeten Drei-Buchstaben-Abkürzungen. Die Backboneatome sind in grün dargestellt. Gelb sind die Seitenketten markiert, bei denen eine Vertauschung der Numerierung zu symmetrischen Strukturen mit identischen chemischen Eigenschaften führt. Bei den violett gekennzeichneten Seitenketten ergeben sich durch diese Vertauschung unterschiedliche Rotamere und beim Histidin (hellblauer Kasten) entsteht dadurch ein anderes Tautomer.

Definition 18 (Numerierung symmetrieäquivalenter Seitenkettenzweige) *Seien die Atomkoordinaten der jeweils ersten Atome nach der Verzweigung für den ersten Zweig* (x_1, y_1, z_1) *und für den zweiten* (x_2, y_2, z_2)*, dann gilt für die Tripel die lexikographische Ordnung:*

$$(x_1 < x_2) \quad \vee \quad ((x_1 = x_2) \wedge (y_1 < y_2)) \quad \vee \quad ((x_1 = x_2) \wedge (y_1 = y_2) \wedge (z_1 < z_2))$$

Eine relative große Toleranz für die Identität der Werte bei diesem Vergleich sorgt dafür, daß das Verfahren nur die Numerierung von tatsächlich gedrehten Zweigen vertauscht.

Neben den symmetrischen Aminosäuren gibt es auch solche, bei denen eine Vertauschung der Numerierung zu Rotameren führt, die um 180° gedreht sind (ASN, GLN, HIS, THR). Diese Aminosäuren bleiben in der aktuellen Implementation unverändert. Auf Basis der eindeutigen Numerierung könnte man jedoch auch systematisch alle alternativen Instanzen einschließlich der Tautomere des Histidins generieren.

5.1.3 Clustern von Instanzen

In der vereinigten Proteinbeschreibung sollen sehr ähnliche Instanzen zusammengefaßt werden. Diese Aufgabe entspricht einem Clustern der Instanzen bezüglich ihrer paarweisen RMS-Abweichung.

Mit einem Graphen läßt sich das Cluster-Problem relativ allgemein beschreiben: Die zu clusternden Objekte werden als Knoten und ihre Abstände als gewichtete Kanten dargestellt. Die Zielfunktionen beim Clustern können sehr unterschiedlich sein. So kann man beispielsweise die Anzahl, die Größe oder den Abstand der Cluster beschränken. Außerdem gibt es verschiedene Definitionsmöglichkeiten für die Distanz zwischen den Clustern bzw. ihrer Elemente, die vom zugrundeliegenden Problem abhängt. Die meisten Clusterprobleme führen in diesem Graph-Modell zu einem $\mathcal{NP}$-schweren Graph-Partitionierungsproblem, so daß diese Probleme oftmals nur approximativ oder heuristisch gelöst werden können. Einen Überblick über Cluster-Algorithmen findet man z.B. in [202, 203, 204, 205].

Hierarchische Cluster-Algorithmen sind eine häufig eingesetzte Heuristik. Ihre Hauptstrategie besteht darin, zwei Cluster mit minimalem Abstand iterativ zu verschmelzen, solange dieser Abstand kleiner als ein vorgegebener Schwellwert ist. In der Regel wird die Distanz zwischen zwei Clustern über ihre Elemente definiert. Hierarchische Cluster-Algorithmen lassen sich entsprechend der verwendeten Distanzmaße folgendermaßen klassifizieren:

- $d(C_1, C_2) = min_{c_1 \in C_1, c_2 \in C_2} d(c_1, c_2)$ — *nearest neighbor* oder *single linkage*
- $d(C_1, C_2) = max_{c_1 \in C_1, c_2 \in C_2} d(c_1, c_2)$ — *furthest neighbor* oder *complete linkage*
- $d(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{c_1 \in C_1, c_2 \in C_2} d(c_1, c_2)$ — *average linkage*

Die Cluster von Instanzen beschreiben Mengen von Proteinkonformationen, die so ähnlich sind, daß sie durch eine einzige gemittelte Konformation repräsentiert werden können. Somit sollte bei diesen Clustern der maximale Abstand der Elemente beschränkt sein. Zum Clustern der Instanzen wird deshalb ein hierarchischer Complete-Linkage-Cluster-Algorithmus verwendet, der von Rarey [84] zum Clustern der Ligandplazierungen in FlexX entwickelt wurde. Unter der Voraussetzung, daß die Abstände zwischen den zu clusternden

Elementen sortiert vorliegen, hat dieser Algorithmus eine Laufzeit von $O(m)$, wobei m die Anzahl von zu verschmelzenden Elementpaaren ist.

Die symmetriekorrigierten Instanzen aus den verschiedenen Ensemblestrukturen werden getrennt nach korrespondierenden Segmenten gemäß Definition 11 zusammengefaßt, das heißt, es werden nur Instanzen vom gleichen Typ verschmolzen, wenn sowohl die RMS-Abweichung als auch der maximale Abstand zwischen zwei einzelnen Atomen unterhalb des vorgegebenen Schwellwertes von 1.0 Å liegen (vgl. Abs. 7.1).

Die Instanzen des jeweiligen Segments sind die Elemente der Cluster, und die Distanz zwischen zwei Elementen ist der RMS-Abstand zwischen den Atomen der Instanzen. Falls nur der Abstand eines einzelnen Atompaares über dem Schwellwert liegt, die mittlere Abweichung aber nicht, wird die Distanz künstlich auf einen Wert über dem Schwellwert angehoben. Wenn die Instanzen nicht vom gleichen Typ sind, das heißt, wenn sie Seitenketteninstanzen von verschiedenen Aminosäuren sind, wird anstelle des RMS-Abstands eine große Konstante verwendet, die verhindert, daß die Instanzen zusammengefaßt werden. Mit der Benennung aus Definition 11 ist die Distanz zwischen zwei Elementen bzw. Instanzen i, j also folgendermaßen definiert:

$$d(i,j) \quad = \quad \begin{cases} \max(RMSD(i,j), \max_a(\|c_a^i - c_a^j\|)) & : \quad \text{wenn } T(i) = T(j) \\ Konst. > V_{max} & : \quad \text{wenn } T(i) \neq T(j) \end{cases}$$

5.2 Berechnung der Kompatibilität

Die Kompatibilität zwischen den Instanzen ist ein zentrales Konzept im Modell der Proteinflexibilität. Deshalb treten im Verlauf des Docking-Algorithmus häufig zwei verschiedene Fragen auf:

1. Sind die Instanzen i und j kompatibel?

2. Welche Instanzen sind zur Instanz i (in)kompatibel?

Um diese Fragen möglichst effizient beantworten zu können, werden parallel zwei verschiedene Repräsentationen des (In)kompatibilitätsgraphen verwendet:

Zum einen erlaubt die Adjazenzmatrix des Kompatibilitätsgraphen, die erste Frage in $O(1)$ zu beantworten. Diese Matrix wird auch als *Kompatibilitätsmatrix* bezeichnet. Zum anderen besitzt jede Instanz eine Liste mit den zu ihr inkompatiblen Instanzen. Dies entspricht einer Adjazenzlisten-Darstellung des Inkompatibilitätsgraphen. Die inkompatiblen Instanzen werden verwendet, weil die Zahl der zu einer Instanz inkompatiblen Instanzen erheblich kleiner als die der kompatiblen ist. Implizit sind damit auch die kompatiblen Instanzen bekannt.

Um die Kompatibilitätsmatrix und die Listen inkompatibler Instanzen zu erzeugen, werden alle geordneten Instanzpaare (i, j) mit $i < j$ aufgezählt und ihre Kompatibilität in den im folgenden dargestellten Stufen entsprechend Definition 12 aus Abschnitt 4.1.2 überprüft:

Instanzen, die zu demselben Segment gehören, sind inkompatibel und brauchen nicht weiter untersucht zu werden (logische Inkompatibilität).

Für alle Instanzen sind die kleinsten umschließenden Kugeln vorberechnet, das sind minimale Kugeln, die alle Atome der Instanz umschließen, die in Form von Kugeln mit

Van-der-Waals-Radius repräsentiert sind. Nur wenn diese Kugeln überlappen, ist ein detaillierter Überlappungstest erforderlich, bei dem das Überlappungsvolumen paarweise für alle Atome der beiden Instanzen aufsummiert wird, bis es gegebenenfalls den vorgegebenen Schwellwert übersteigt. Instanzen von Segmenten, zwischen denen kovalente Bindungen auftreten, sind von diesem Test der geometrischen Inkompatibilität ausgenommen.

Strukturell kompatible Instanzen müssen direkt oder über eine Folge kompatibler Instanzen verbunden sein. Für zwei Instanzen i_s, j_t benachbarter Segmente s, t muß man deshalb prüfen, ob der Abstand der zu verknüpfenden Atome in etwa der theoretischen Bindungslänge [198] entspricht. Um festzustellen, ob zwei Instanzen einer Kette indirekt über weitere Instanzen verbunden sind, kann mit dynamischer Programmierung auf die bereits berechnete Kompatibilitätsmatrix zurückgegriffen werden. Jede Instanz besitzt eine Liste ihrer Vorgänger. Das sind bei Backbone-Instanzen die Backbone-Instanzen des Vorgängersegments und bei Seitenketten-Instanzen die Backbone-Instanzen der entsprechenden Aminosäure. Da für diese Vorgängerinstanzen die Kompatibilität bereits berechnet und in der Kompatibilitätsmatrix eingetragen ist, braucht für ein Instanzenpaar (i_s, j_t) nur geprüft zu werden, ob es eine Instanz k_{t-1} im Vorgängersegment $t-1$ von t gibt, die mit j_t verbunden und mit i_s kompatibel ist. Denn es gilt die Transitivität der strukturellen Inkompatibilität:

Lemma 4 (Transitivität der strukturellen Inkompatibilität) *Für zwei logisch und geometrisch kompatible Instanzen i_s, j_t der Segmente s und t mit $s < t$ gilt:*

$$j_t \bowtie k_{t-1} \quad \wedge \quad k_{t-1} \sim i_s \quad \Rightarrow \quad j_t \sim i_s$$

Beweis: Aus $k_{t-1} \sim i_s$ folgt insbesondere $k_{t-1} \bowtie i_s$. Da $j_t \bowtie k_{t-1}$ folgt $j_t \bowtie k_{t-1} \bowtie i_s$, das heißt $j_t \bowtie i_s$ und da die Instanzen i_s und j_t nach Voraussetzung auch logisch und geometrisch kompatibel sind, sind sie insgesamt kompatibel: $j_t \sim i_s$ □

Die indirekte Inkompatibilität wird nur zum Teil vorberechnet. Es wird festgestellt, ob es ein Segment s mit nur einer Instanz i_s gibt, die inkompatibel mit einer anderen Instanz j_t eines anderen Segments t ist. Tritt ein sog. *Konflikt* auf, wird die Instanz j_t völlig von den Berechnungen ausgeschlossen, weil sie ohnehin nie verwendet werden darf. Die Suche wird wiederholt, falls das Segment dann nur noch eine einzige Instanz enthält. Solche Konflikte sollten nach der Definition des Ensembles eigentlich nicht auftreten, weil die Instanzen mindestens mit allen Instanzen ihrer eigenen Ensemblestruktur kompatibel sein sollten. Tatsächlich treten Konflikte aufgrund des Clusterns der Instanzen gelegentlich auf. Sind es zu viele, ist dies ein Hinweis darauf, daß entweder der Schwellwert für das Clustern zu groß, oder die Toleranz für das Überlappungsvolumen zu klein ist.

Darüber hinaus werden keine indirekten Inkompatibilitäten zwischen Instanzen bestimmt, weil dazu jeder einzelne Knoten des Inkompatibilitätsgraphen unter Umständen mehrfach analysiert werden muß (vgl. Abs. 4.2.2). Da die indirekte Inkompatibilität zwischen zwei Instanzen auch ohne das Vorhandensein einer expliziten Kante im Inkompatibilitätsgraphen verhindert, daß diese beiden Instanzen gemeinsam in einer unabhängigen Menge enthalten sein können, kann man auf diese aufwendige Suche verzichten und die Bestimmung der indirekten Inkompatibilität dem Algorithmus zum Auffinden der optimalen unabhängigen Menge überlassen.

5.2.1 Kompatibilität der Wechselwirkungspunkte

Den Instanzen im aktiven Zentrum der vereinigten Proteinstruktur werden Wechselwirkungsgeometrien zugeordnet, die jeweils in einem Bezugssystem aus Referenzatomen eindeutig definiert sind. Die Wechselwirkungsgeometrien lassen sich durch eine Menge diskreter Wechselwirkungspunkte approximieren. Über eine vorberechnete Hashtabelle kann dann der Basisplazierungsalgorithmus effizient auf Paare solcher Punkte zugreifen. Diese Methode stammt aus FlexX (vgl. Abs. 2.4.4).

Wechselwirkungspunkte, auf die kein Ligandatom plaziert werden kann, weil es dann im Inneren des Proteins läge, dürfen nicht verwendet werden. Deshalb werden solche Punkte entfernt, die nach Definition 17 unzugänglich sind. Dazu muß man für jeden Wechselwirkungspunkt p_{i_s} jeder Instanz i_s feststellen, ob es ein Segment t gibt, bei dem eine an die Position von p_{i_s} plazierte Testkugel mit dem Van-der-Waals-Radius des entsprechenden Ligandatoms innerhalb aller Instanzen j_t liegt, die mit i_s kompatibel sind.

Dieser Überlappungstest läßt sich effizient mit den dreidimensionalen Raumabfragen realisieren, die für die Bestimmung des Überlappungsvolumens zwischen Protein und Ligand für FlexX entwickelt worden sind [84]. Bei dieser Methode wird der Raum in achsenparallele, gleich große Kuben zerlegt. Alle Kuben, die z.B. die Mittelpunkte von Proteinatomen enthalten, werden in einer Hashtabelle abgelegt, die man über die Indizes der Kuben adressiert. Für eine Anfragekugel lassen sich dann alle Kuben bestimmen, die diese Kugel schneidet. So kann man effizient alle Punkte finden, die innerhalb der Anfragekugel liegen.

Auf diese Weise lassen sich für einen Wechselwirkungspunkt p_{i_s} sehr schnell alle Proteinatome a_{j_t} bestimmen, die in einem gewissen Radius um p_{i_s} zu finden sind. Für jedes dieser Atome a_{j_t} muß man zunächst überprüfen, ob die Instanzen s_i und j_t kompatibel sind. Ist das der Fall, wird das Überlappungsvolumen $V(p_{i_s}, a_{j_t})$ zwischen dem Atom a_{j_t} und der Testkugel, die p_{i_s} repräsentiert, bestimmt und zu dem Gesamtvolumen V_{j_t} der Instanz j_t addiert. Alle Wechselwirkungspunkte p_{i_s}, für die folgende Bedingung gilt, werden entfernt:

$$\exists\, t \in S \quad \forall j' \in \{j \in I_t \mid j_t \sim i_s\} \quad V_{j'_t} > V_{max}$$

Man muß dabei nur die Segmente j überprüfen, die mindestens eine Instanz enthalten, die ein Überlappungsvolumen V_{j_t} mit p_{i_s} besitzen. Deshalb sind bereits bei der Berechnung der Überlappungsvolumen Zeiger auf diese Segmente festzuhalten.

Für die vereinigte Proteinbeschreibung brauchen keine Paare von Wechselwirkungspunkten in die Hashtabelle aufgenommen zu werden, die zu inkompatiblen Instanzen gehören, weil diese Punktpaare ohnehin nicht für die Plazierung verwendet werden dürfen. Generell macht es deshalb keinen Sinn, Wechselwirkungspunkte von alternativen Instanzen, die dicht beieinander liegen zusammenzufassen. Denn sie sind nicht unbedingt mit denselben Punkten kompatibel, so daß sich unterschiedliche Punktpaare für die Hashtabelle ergeben. Die Kompatibilität zweier Wechselwirkungspunkte bzw. der zugehörigen Instanzen läßt sich durch einfache Abfrage der zuvor berechneten Kompatibilitätsmatrix bestimmen.

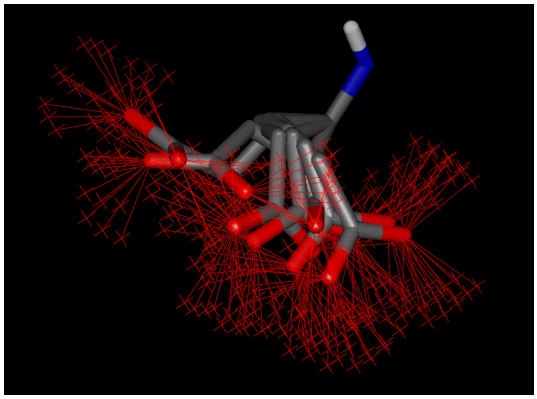


Abbildung 5.2: Überlappende Wechselwirkungsgeometrien alternativer Instanzen. In der Abbildung sind sieben Instanzen eines Segments mit ihren diskretisierten Wechselwirkungsgeometrien dargestellt, die alle denselben Typ haben, weil sie Alternativen voneinander sind. Wegen der räumlichen Nähe der Instanzen überlappen die verschiedenen Geometrien und ihre Wechselwirkungspunkte liegen gehäuft in einem Gebiet.

5.3 Reduktion der Hashtabelle durch Clustern

Die Vorverarbeitung legt Paare von Wechselwirkungspunkten in einer Hashtabelle ab, damit die Basisplazierung effizient darauf zugreifen kann. In dieser Hashtabelle werden die Wechselwirkungspunktpaare mit derselben Adresse in einer sortierten Liste gehalten. Dieses Verfahren wurde von FlexX übernommen (vgl. Abs. 2.4.4).

Wenn die Distanzen der einzutragenden Punktpaare gleichmäßig verteilt sind und die Anzahl der Paare nicht wesentlich größer als die Zahl der Buckets der Hashtabelle ist, sind die zu sortierenden Listen kurz und der zusätzliche Aufwand für das Sortieren kann praktisch vernachlässigt werden. Die Zeit für den Aufbau der Hashtabelle wächst dann mit $O(N)$, wobei N der Anzahl der einzutragenden Wechselwirkungspunktpaare entspricht. Dies ist bei FlexX der Fall.

Berücksichtigen muß man den Sortieraufwand dagegen, wenn es viele Punktpaare mit ähnlichem Abstand gibt, weil dann lange Listen an den Buckets entstehen, die sortiert werden müssen. Viele Paare mit ähnlicher Distanzen ergeben sich automatisch, wenn es viele Punkte im Verhältnis zur Zahl der Buckets gibt, oder wenn Gebiete im Raum mit einer großen Zahl von Wechselwirkungspunkten vom gleichen Typ existieren, denn alle Kombinationen von Punkten zweier solcher Gebiete haben in etwa den gleichen Abstand. Deshalb wächst die Zeit für den Aufbau der Hashtabelle mit in diesem Fall mit $O(N\,logN)$. Außerdem steigt natürlich auch der benötigte Speicherplatz.

In der vereinigten Proteinbeschreibung gibt es im Vergleich zu einer normalen Proteinstruktur insgesamt mehr Wechselwirkungspunkte und es treten auch oft Häufungen von Punkten des gleichen Typs auf, weil alternative Instanzen Wechselwirkungen vom gleichen Typ besitzen und deren Geometrien in der Regel nicht weit voneinander entfernt sind und oft auch überlappen (vgl. Abb. 5.2). Darum dauert der Aufbau der Hashtabelle für eine vereinigte Proteinbeschreibung ungleich länger als für eine einfache Proteinstruktur. In einigen Fällen wird die Hashtabelle dabei so groß, daß man sie nicht mehr effizient erzeugen kann. Insbesondere müssen Maschinen, die nicht über einen ausreichend großen Hauptspeicher verfügen, ein Teil des Hauptspeichers und damit der Hashtabelle auf die Festplatte auslagern (swapping). Diese Plattenzugriffe führen nicht nur beim Aufbau der Hashtabelle, sondern auch bei späteren Anfragen zu erheblichen Laufzeitverzögerungen, die unter ungünstigen Umständen das gesamte Verfahren unpraktikabel werden lassen.

Aus diesem Grund wurde folgendes Verfahren entwickelt, um die Zahl der Punkte, die effektiv in die Hashtabelle eintragen werden, ohne Informationsverlust zu reduzieren.

1. Die Wechselwirkungspunkte werden unter Berücksichtigung der Wechselwirkungstypen geclustert. Als Repräsentanten dieser Cluster dienen ihre geometrischen Mittelpunkte.

2. Anstatt für die Wechselwirkungspunkte selbst erzeugt man dann die Hashtabelle für die Cluster-Repräsentanten.

3. Bei der Abfrage der Hashtabelle werden anstatt eines Punktpaares alle Paare von ursprünglichen Wechselwirkungspunkten zurückgegeben, die die beiden Cluster repräsentieren.

Zum Clustern wird wiederum der hierarchische Complete-Linkage-Cluster-Algorithmus von Rarey [84] (s. Abs. 5.1.3) verwendet, weil es auch hier vor allem darauf ankommt, die maximale Distanz zwischen den Wechselwirkungspunkten eines Clusters zu begrenzen. Die Kompatibilität der Wechselwirkungspunkte wird beim Clustern nicht berücksichtigt. Denn ein Hauptgrund für die Vielzahl von dicht beieinander liegenden Punkten sind alternative Instanzen. Ihre Wechselwirkungspunkte sind aber miteinander inkompatibel und würden dann nicht zusammengefaßt. Die Kompatibilität der Punkte wird deshalb erst bei der Rückgabe der Punktpaare im dritten Schritt überprüft.

Dieses Verfahren benötigt zwar zusätzlich Laufzeit für das Clustern der Wechselwirkungspunkte und das Erzeugen aller Punktpaare bei einer Anfrage an die Hashtabelle sowie zusätzlichen Speicher für die Cluster und ihre Repräsentanten, dafür spart es aber Laufzeit und Speicherplatz beim Aufbau der Hashtabelle. Dabei ist zu beachten, daß alle Punkte in die Hashtabelle eingetragen werden müssen, also die Laufzeit für das Sortieren der Listen in jedem Fall anfällt, während im Falle der Komprimierung die zusätzliche Laufzeit für das Aufzählen aller Punktpaare bei einer Anfrage nur für diejenigen Cluster erforderlich ist, auf die man tatsächlich zugreift. Diese Laufzeit ist jedoch nicht viel größer als die, die für die Rückgabe der gespeicherten Liste ohne Clustern benötigt wird.

Welche Faktoren überwiegen, hängt im Einzelfall von der Dichte der Wechselwirkungspunkte mit gleichem Typ ab. Enthält im schlechtesten Fall jeder Cluster nur einen Wechselwirkungspunkt, so sind die Gesamtkosten größer als ohne das Reduktionsverfahren. Umfassen die Cluster aber mehrere Punkte, so ist die Laufzeit für das Clustern geringer als die Zeit, die beim Aufbau der Hashtabelle eingespart wird. Auch Speicherplatz kann eingespart werden, denn der zusätzliche Speicherbedarf des Clusterns wächst linear mit der Zahl der Wechselwirkungspunkte, während der Platzverbrauch der Hashtabelle mit dem Quadrat der eingetragenen Punkte steigt. Aus diesem Grund sollte man den Schwellwert für das Clustern der Wechselwirkungspunkte nicht zu gering wählen, damit die Cluster nicht zu klein ausfallen. Zu große Cluster führen aber auf der anderen Seite dazu, daß Punkte falsch in die Hashtabelle eingeordnet werden. Als geeigneter Schwellwert wurde empirisch 0.6 Å ermittelt (s. Abs. 7.4).

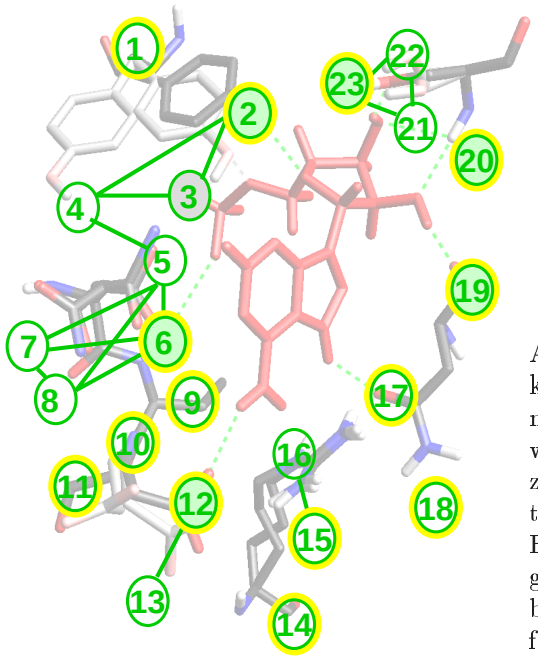


Abbildung 5.3: Auswahl von Instanzen beim inkrementellen Aufbau. Das Bild gibt die Situation nach der Plazierung eines (Teil-)Liganden wieder: Es wird für alle Instanzen die Wechselwirkungsenergie zum Liganden bestimmt, die als Gewicht für die Knoten (grüne, Kreise) dient (mit Farbe hinterlegt, falls Energie ungleich Null). Dann wird die unabhängige Menge von Knoten (gelb) mit minimaler Energie bestimmt. Sie entspricht einer gültigen Proteinkonformation, die im Hintergrund dargestellt ist.

5.4 Plazierung der Liganden

Der Plazierungs-Algorithmus von FlexX besteht aus den drei Phasen, die in Abschnitt 2.4.4 erläutert wurden: der Auswahl der Basisfragmente, der Basisplazierung und dem inkrementellem Aufbau der Liganden.

Da die Auswahl der Basisfragmente unabhängig vom Protein erfolgt, wird sie durch die Modellierung der Proteinflexibilität nicht beeinflußt und kann unverändert übernommen werden.

Die Basisplazierung ordnet mit Hilfe der Hashtabelle Tripel bzw. Paare von Wechselwirkungspunkten des Proteins Wechselwirkungzentren der Basisfragmente des Liganden zu. Dabei darf man nur kompatible Wechselwirkungspunkte gleichzeitig verwenden (s. Abs. 5.2.1). Dies kann dadurch erreicht werden, daß nur Paare von kompatiblen Wechselwirkungspunkten in die Hashtabelle aufgenommen werden und die Kompatibilität der Punkte beim Erzeugen der Rückgabedreiecke berücksichtigt wird. Die Überlagerung der korrespondierenden Punkte erfolgt dann analog zu FlexX.

Das Zusammensetzen der Ligandfragmente entsprechend der Torsionsdatenbank im Zuge des inkrementellen Aufbaus kann wiederum ohne Änderungen von FlexX übertragen werden.

Größere Modifikationen sind dagegen für die Bewertung der Plazierungen erforderlich. Denn um die (Teil-)Lösungen bewerten zu können, müssen nach jedem Konstruktionsschritt alle möglichen Wechselwirkungen und die Überlappungen zwischen dem (zum Teil) plazierten Liganden und allen Instanzen der vereinigten Proteinbeschreibung einzeln bestimmt werden. Die Bewertungsfunktion schätzt separat für jede einzelne Instanz die Wechselwirkungsenergie zum Liganden (vgl. Abs. 4.5). Diese Energie dient als Gewicht der entsprechenden Knoten des Inkompatibilitätsgraphen (s. Abb. 5.3).

Aufgrund der Abhängigkeiten zwischen den verschiedenen Instanzen kann man jedoch nicht einfach die Knoten mit den niedrigsten Gewichten aus jedem Segment kombinieren,

sondern muß eine unabhängige Menge von Knoten im Inkompatibilitätsgraphen bestimmen, die zur minimalen Bindungsenergie führt. Sie repräsentiert eine gültige Proteinkonformation, die im Hinblick auf die Bewertungsfunktion am besten zu dem (zum Teil) plazierten Liganden paßt. Der Score dieser optimalen unabhängigen Menge, der der Summe der Einzelenergien der Knoten entspricht, wird als Gesamtenergie der (Teil-)Lösung verwendet. Die Bewertung einer Plazierung wird damit zur Suche nach einer optimalen gewichteten unabhängigen Menge im Inkompatibilitätsgraphen. Diese Suche wird nach jedem Anbauschritt für die gesamte (Teil-)Plazierung wiederholt, weil sich die Auswahl der Instanzen nach dem Hinzufügen eines Fragmentes auch für die zuvor plazierten Fragmente verändern kann.

Die Suche nach optimalen gewichteten unabhängigen Mengen ist, wie die Bestimmung von unabhängigen Mengen selbst, $\mathcal{NP}$-vollständig [206, 207, 208] und läßt sich auch nur schwer approximieren [209]. Allerdings kann man das Problem durch die Zerlegung des Inkompatibilitätsgraphen in Zusammenhangskomponenten in kleinere unabhängige Teilprobleme aufteilen (vgl. Lemma 3, Abs. 4.2). Dabei müssen nur die Zusammenhangskomponenten berücksichtigt werden, die wenigstens zum Teil im aktiven Zentrum liegen und mindestens einen Knoten enthalten, der einen Score ungleich Null hat, denn die anderen Zusammenhangskomponenten tragen nicht zum Gesamtscore bei. Dennoch bleibt die zeitintensive Suche nach einer optimalen unabhängigen Menge der zeitkritische Schritt des Docking-Algorithmus, weil diese Suche nach jedem Aufbauschritt für jede Plazierung erneut durchgeführt werden muß.

5.4.1 Zerlegung des Kompatibilitätsgraphen

Da die Liganden in das aktive Zentrum der vereinigten Proteinstruktur plaziert und nur den Instanzen innerhalb der Bindetasche Wechselwirkungsgeometrien zugeordnet werden, können nur Instanzen, die sich in der Bindetasche befinden, zur Gesamtbindungsenergie beitragen. Darum brauchen bei der Suche nach der optimalen unabhängigen Menge nur die Zusammenhangskomponenten berücksichtigt werden, bei denen zumindest ein Teil der Knoten im aktiven Zentrum liegt. Alle Knoten dieser Zusammenhangskomponenten bilden zusammen das erweiterte aktive Zentrum, auf das die Suche bei der Plazierung eines Liganden beschränkt werden kann.

Da sich die Konnektivität des Inkompatibilitätsgraphen während der Ligandplazierung nicht ändert, kann man dieses erweiterte aktive Zentrum bereits in der Vorverarbeitungsphase bestimmen. Dabei läßt sich zwischen Segmenten, die nur eine Instanz enthalten, und solchen, die aus mehreren Instanzen bestehen, unterscheiden.

Die Einzelinstanzen bilden jeweils einelementige Zusammenhangskomponenten, da andere Instanzen, die mit ihnen in Konflikt stehen, von der weiteren Berechnung ausgeschlossen wurden (Abs. 5.2). Diese Einzelinstanzen tragen in jedem Fall zur Gesamtenergie bei, weil sie nach Voraussetzung Bestandteil einer gültigen Proteinstruktur sein müssen. Eine Auswahl ist deshalb bei diesen Instanzen nicht erforderlich.

Für die übrigen Instanzen lassen sich die Zusammenhangskomponenten mit folgendem Verfahren bestimmen: Ausgehend von einer Instanz des aktiven Zentrums werden im Inkompatibilitätsgraphen alle mit ihr verbundenen Instanzen gesucht und mit der Nummer der Zusammenhangskomponente markiert. Weitere Zusammenhangskomponenten entstehen, indem man die Suche mit einer noch nicht markierten Instanz des aktiven Zentrums wiederholt, bis es keine solche Instanz mehr gibt. Diese mehrelementigen Zusam-

menhangskomponenten können unabhängig voneinander nach optimalen unabhängigen Mengen durchsucht werden.

Durch diese Zerlegung wird die Suche nach optimalen unabhängigen Mengen aus zwei Gründen erheblich beschleunigt: Erstens braucht eine Vielzahl von Zusammenhangskomponenten überhaupt nicht durchsucht zu werden, und zwar die einelementigen und die, die keine Instanz mit einem Score ungleich Null enthalten. Vor allem aber bleiben die Zusammenhangskomponenten unberücksichtigt, die nicht zum erweiterten aktiven Zentrum gehören. Zweitens wird die Größe des Suchraums verringert, denn alle Kombinationen von Instanzen aus zwei verschiedenen Zusammenhangskomponenten müssen nicht bewertet werden.

5.4.2 Suche nach der optimalen gewichteten unabhängigen Menge

Die Suche nach einer optimalen gewichteten unabhängigen Menge im Inkompatibilitätsgraphen verfolgt im wesentlichen zwei Ziele: Erstens soll sie feststellen, ob die gefundene Ligandplazierung überhaupt möglich ist, das heißt, ob es eine gültige Proteinkonformation gibt, die möglichst viele der Knoten/Instanzen enthält, die Wechselwirkungen ausbilden. Zweitens soll eine unabhängige Menge mit minimaler Energie bestimmt werden, weil ihre Energie als Bewertung der Plazierung dient.

Die erste Frage: „Gibt es eine unabhängige Menge?“, ist die wichtigere, denn es dürfen nur Plazierungen verwendet werden, die auf gültigen Proteinstrukturen basieren. Dabei wird zusätzlich gefordert, daß die unabhängige Menge einen Knoten aus jedem Segment enthält. Hier dürfen durch eine mögliche Approximation keine Inkompatibilitätskanten übersehen werden, weil damit die Lösung ungültig würde.

Bei der zweiten Frage: „Wie groß ist die Energie einer optimalen unabhängigen Menge?“, können dagegen kleinere Fehler toleriert werden. Sie stören zwar die Ordnung der Lösungsmenge, führen aber nicht direkt zum Ausschluß einer Plazierung. Insbesondere spielt es keine Rolle, ob es unter Umständen weitere unabhängige Mengen mit derselben Energie gibt.

Verschiedene unabhängige Mengen stellen unterschiedliche Auswahlen von Instanzen dar, die gültige Proteinkonformationen für dieselbe Ligandplazierung repräsentieren. Haben sie die gleiche Gesamtenergie, führen sie zu derselben Bewertung der Ligandplazierung. Dagegen bedeuten zwei unabhängige Mengen mit unterschiedlichem Score, daß dieselbe Plazierung eines Liganden unterschiedlich bewertet wird. In diesem Fall reicht es aus, die bessere Bewertung zu berücksichtigen. Denn nach jedem Anbau eines weiteren Fragments werden die unabhängigen Mengen sowieso für jede (Teil-)Plazierung insgesamt neu bestimmt, weil sich aufgrund der Abhängigkeiten zwischen den Instanzen durch das Anfügen eines zusätzlichen Fragments auch die Auswahl der Instanzen in dem Bereich der Bindetasche ändern kann, in dem der erste Teil des Liganden bereits plaziert worden war. Aus diesem Grunde braucht man für jede Plazierung nur eine optimale unabhängige Menge zu finden.

Das Problem, die größte unabhängige Menge zu finden, bzw. das entsprechende Cliquenproblem, tritt nicht nur beim Docking [20, 95, 163, 210, 211, 212], sondern auch bei anderen Fragestellungen aus dem Bereich der Biochemie und Bioinformatik immer wieder auf [213, 214, 215, 216, 217]. Es gibt verschiedene Algorithmen, die das Problem generell lösen [207, 208, 218, 219, 220], die Eigenschaften bestimmter Spezialfälle ausnutzen [221, 222, 223, 224] oder versuchen, es heuristisch zu approximieren [206, 209, 225, 226, 227]. In

allen Fällen bleiben die Verfahren recht aufwendig und dauern für die wiederholte Suche für jede Plazierung nach jedem Konstruktionsschritt zu lange. Abschnitt 7.3.4 enthält eine Statistik über die Anzahl und Größe der Zusammenhangskomponenten, die bei den realen Testsystemen auftreten. Außerdem ist es für die heuristischen Verfahren zum Teil schwierig, die zusätzliche Bedingung, daß ein Knoten aus jedem Segment verwendet werden muß, effizient zu integrieren. Deshalb wird in dieser Arbeit das folgende Verfahren verwendet, um eine optimale unabhängige Menge in den mehrelementigen Zusammenhangskomponenten zu bestimmen.

Für Zusammenhangskomponenten, die nur zwei oder drei Instanzen enthalten, läßt sich eine optimale unabhängige Menge durch direkten Vergleich ihrer Energien und der Kompatibilität sehr schnell bestimmen, zumal es sich hierbei meist um die Instanzen aus ein und demselben Segment handelt, die Cliquen im Inkompatibilitätsgraphen bilden. Die optimale unabhängige Menge ist in diesem Fall die Instanz mit der niedrigsten Energie.

Zusammenhangskomponenten, die mehr als drei Knoten enthalten, lassen sich mit einer Tiefensuche auf Basis des rekursiven Aufzählungsschemas des Bron-Kerbosch-Algorithmus [218] durchsuchen. Aus Zeitgründen wird die erste unabhängige Menge, die dieses Verfahren findet, als Lösung verwendet. Diese Methode ist extrem schnell und findet in der Regel gute Lösungen, obwohl das natürlich mit dieser stark heuristischen Greedy-Strategie in keiner Weise garantiert werden kann.

Der originale Bron-Kerbosch-Algorithmus [218] zählt rekursiv alle Cliquen in einem ungerichteten Graphen auf. Dieses Problem ist komplementär zu der Suche nach unabhängigen Mengen. In jedem Schritt wird die aktuelle partielle Clique P, die initial leer ist, durch einen Knoten k erweitert, der aus einer Kandidatenmenge C ausgewählt wird. Diese Auswahl ist kritisch für die Leistung des Algorithmus, weil sie bestimmt, wieviele Knoten vor der nächsten Rekursion ausgeschlossen werden können. Es ist daher sinnvoll, die Kandidatenmenge nach dem Grad der Knoten zu sortieren [218].

Bron-Kerbosch-Algorithmus [218] zum Aufzählen unabhängiger Mengen

1. Wähle einen Kandidatenknoten k aus C.

2. Erweitere die aktuelle unabhängige Menge P durch k.

3. Erzeuge N' und C' aus N bzw. C durch Entfernen
   aller Knoten k', die mit k verbunden sind.

4. Wenn $N' = C' = \emptyset$ schreibe P (eine unabhängige Menge),
   andernfalls starte die Rekursion erneut mit N' und C' in Schritt 1

5. Entferne k aus der aktuellen unabhängigen Menge P und füge ihn in N ein.

Die Mengen N und C sind dabei als ein fortlaufendes Feld, die aktuelle unabhängige Menge P als Stack implementiert. Auf diese Weise bedeutet die Verschiebung von k aus C nach N, daß nur ein Feldindex um 1 erhöht werden muß.

Die Tiefensuche beruht auf folgenden Modifikationen dieses Basisalgorithmus: Erstens werden die Knoten in der Kandidatenliste C aufsteigend nach der geschätzten Bindungsenergie der Instanzen sortiert. Auf diese Weise lassen sich unabhängige Mengen mit niedrigerer Energie früher erzeugen.

Zweitens wird nach jeder Auswahl eines Knotens k überprüft, ob die Vereinigung $P \cup C$ der aktuellen unabhängigen Menge P und der Kandidatenmenge C immer noch mindestens eine Instanz für jedes Segment enthält. Falls das nicht der Fall ist, kann der Algorithmus die Rekursion frühzeitig beenden, weil die unabhängige Menge nach Voraussetzung genau eine Instanz pro Segment enthalten muß.

Kapitel 6

Methoden der Evaluierung

Theoretische Aussagen über die Leistungsfähigkeit eines neuen Docking-Programms sind aufgrund der vielen verwendeten Heuristiken und Approximationen nicht möglich. Daher bestimmt man die Qualität der Lösungen und den Laufzeitbedarf im allgemeinen experimentell auf der Basis realer Testdaten, bei denen die Lösungen bereits bekannt sind. Für das Docking-Problem haben sich dabei verschiedene Experimente etabliert [10], auf die im folgenden eingegangen wird.

6.1 Redocking

Der erste Test für ein neues Docking-Programm ist in der Regel ein sog. *Redocking*-Experiment. Dazu werden experimentell bestimmte Protein-Ligand-Komplexe zerlegt und durch Docking wieder reproduziert [10]. Da bei diesem Experiment die korrekte Position des Liganden in der Bindetasche bekannt ist, können die Lösungen über die RMS-Abweichung der Liganden zu ihrer Referenzstruktur bewertet werden. In der Literatur werden dabei üblicherweise Plazierungen mit einer RMS-Abweichung von weniger als 2.0 Å als akzeptable Lösungen betrachtet [32, 79].

6.1.1 Abhängigkeiten vom Testdatensatz

Beim Redocking besteht die Gefahr, daß Informationen über den Komplex in die Aufbereitung der Testdaten einfließen, z.B. durch die Definition des aktiven Zentrums, die Protonierung oder die Wahl der Ligandkonformation. Evaluationen mit großen Testdatensätzen, wie sie beispielweise für die Programme GOLD (100 Komplexe) [32] und FlexX (200 Komplexe) [79] durchgeführt worden sind, können dieses Problem zwar mindern, aber nicht völlig ausschließen. Dies ist nur mit sog. *Blindvorhersagen* möglich, bei denen der Komplex zum Zeitpunkt der Docking-Rechnung noch nicht bekannt ist. Deshalb kann man sicher sein, daß keine speziellen Informationen über den Komplex in die Vorhersage eingehen. Solche Blindvorhersageexperimente sind aber organisatorisch sehr aufwendig, weil sie von unabhängigen Dritten kontrolliert werden müssen. Sie wurden für das Protein-Ligand-Docking bisher nur auf einzelnen Beispielen im Rahmen des sog. CASP-II-Experiments [81] durchgeführt.

Um eine Abhängigkeit von FlexE von speziellen Testdaten soweit wie möglich zu vermeiden, wurde deshalb ein Testdatensatz zusammengestellt, der zumindest mehrere verschiedene Ensembles und Liganden umfaßt.

6.1.2 Vorgabe der Proteinkonformation

Die Redocking-Experimente übernehmen im allgemeinen die unveränderte Proteinkonformation eines Komplexes. Das bedeutet zwar unter Umständen eine Vereinfachung der Aufgabe, ist aber nur für Docking-Programme legitim, die das Protein als starr behandeln, weil hier die Proteinkonformation ohnehin nicht vorhergesagt wird. *Kreuzdocking-Experimente* zeigen darüber hinaus, daß sich die so ermittelte Vorhersagequalität oft auch auf leicht veränderte Proteinstrukturen übertragen läßt. In Ausnahmefällen findet man dabei sogar bessere Plazierungen in anderen als den originalen Proteinstrukturen [79].

Für Docking-Programme, die die Proteinflexibilität berücksichtigen, ist das Verwenden der originalen Proteinkonformation dagegen problematisch, weil man damit Komplexinformationen vorgibt, die vorhergesagt werden sollen. Auf der anderen Seite können diskrete Docking-Programme nur Konformationen reproduzieren, die z.B. in Form von Rotamerbibliotheken oder alternativen Proteinstrukturen vorgegeben worden sind. Strukturen, die über die Rekombination der gegebenen Konformere hinausgehen, kann man nur mit Simulations-Verfahren finden.

Ziel des in dieser Arbeit beschriebenen Ensembleansatzes ist es, aus einer Menge von Instanzen die gültige Proteinkonformation effizient zusammenzusetzen, die zur besten Plazierung eines Liganden führt. Das kann jedoch nur gelingen, wenn die richtigen Instanzen oder zumindest ähnliche Instanzen in der Menge aller Instanzen enthalten sind. Fehlen die korrekten Instanzen, kann man von diesem Ansatz keine gute Plazierung erwarten.

Um sicher zu sein, daß die korrekte Proteinkonformation überhaupt in der Lösungsmenge enthalten ist, wird deshalb die Originalproteinstruktur als eine Ensemblestruktur verwendet. Sie ist aber in keiner Weise gekennzeichnet. Damit bleibt immer noch das Problem, diese richtige Konformation als solche zu identifizieren.

Bei größeren Ensembles kann man davon ausgehen, daß es unter den Ensemblestrukturen mehrere ähnliche Instanzen gibt, die in geeigneter Weise kombiniert werden können, um einen Liganden korrekt zu plazieren. In diesem Fall ist die Originalstruktur nicht erforderlich. Diese Situation entspricht einer realen Anwendung, bei der man zwar die konkrete Proteinkonformation nicht kennt, wohl aber ähnliche Strukturen in Komplexen mit anderen Liganden beobachtet hat.

6.1.3 Bewertung der Vorhersage

Die meisten Docking-Programme generieren eine Menge von Lösungen, die entsprechend der Bewertung sortiert ist. Im Idealfall sollte die Plazierung auf dem ersten Rang auch die niedrigste RMS-Abweichung zur Referenzstruktur haben. Zumindest sollte die erste Vorhersage aber eine niedrige Abweichung haben. Aufgrund der Heuristiken, auf denen die Bewertungsfunktionen beruhen (vgl. Abs. 4.5), ist dies aber oft nicht der Fall [59, 60].

Es gibt deshalb verschiedene Kriterien, anhand derer man unterschiedliche Docking-Programme bewerten und vergleichen kann. Sie werden im folgenden diskutiert. Dabei gilt im Rahmen dieser Arbeit die Plazierung mit der geringsten Abweichung zur Referenzstruktur als die *beste* Lösung, weil sie die korrekte Lösung am besten reproduziert. Qualitative Bewertungen beziehen sich immer auf diese korrekte Plazierung. Die Lösung, die sich auf dem ersten Rang befindet, also vom Programm als optimal bewertet wird, wird dagegen als die *erste* Lösung bezeichnet. Die Ordnung in der Liste wird mit Zahlwörtern beschrieben (erste, zweite, ..., letzte Plazierung).

Lösung auf dem ersten Rang
Die RMS-Abweichung der ersten Lösung wird oft verwendet, um die Leistungsfähigkeit eines Docking-Programms zu beurteilen, weil sie zum einen im Idealfall auch die beste oder zumindest eine gute Lösung sein sollte und weil man zum anderen bei Screening-Anwendungen nur eine geringe Zahl von Lösungen haben möchte.

Diese erste Lösung hängt neben dem Plazierungsalgorithmus auch sehr stark von der Bewertungsfunktion ab. Denn wenn der Algorithmus die richtige Plazierung findet und sie in der Lösungsmenge enthalten ist, entscheidet die Bewertung, ob die Lösung auch auf den ersten Rang fällt, wobei unter Umständen minimale Unterschiede in der geschätzten Bindungsenergie den Ausschlag geben können.

Beste Lösung auf den ersten 10 Rängen
Berücksichtigt man statt der ersten Lösung die beste Lösung auf den ersten zehn Rängen, so wird die gerade beschriebene Abhängigkeit der Vorhersage von der Bewertungsfunktion etwas gemindert, ohne daß dabei die Zahl der Lösungen zu stark anwächst.

In dieser Arbeit geht es in erster Linie um einen neues Verfahren, das bei der Plazierung eines Liganden Variationen der Proteinstruktur berücksichtigen kann. Dabei wird eine bekannte Bewertungsfunktion verwendet, die im Rahmen dieser Arbeit jedoch nicht verbessert werden sollte. Sie kann gegebenenfalls auch gegen eine andere ausgetauscht oder durch ein Consensus-Scoring ersetzt werden (vgl. Abs. 2.3). Deswegen ist es für die Evaluation des Verfahrens wichtiger, daß eine gute Plazierung in der Lösungsmenge enthalten ist, als daß sich diese Lösung auf Rang Eins befindet. Aus diesem Grund wird hier vor allem die beste Lösung auf den ersten 10 Rängen untersucht.

Beste Lösung unabhängig vom Rang
Über die RMS-Abweichung der besten Lösung läßt sich abschätzen, welche Vorhersagequalität mit einer idealen Bewertungsfunktion zu erreichen wäre. Allerdings ist auch diese Vorhersage bei vielen Verfahren nicht völlig unabhängig von der Bewertungsfunktion, weil oft bereits Zwischenergebnisse bewertet werden müssen. FlexE und FlexX verwenden diese Funktion beispielsweise während des inkrementellen Aufbaus der Liganden, um die Teillösungen zu sortieren und diejenigen auszuwählen, an die die nächsten Fragmente anhängt werden.

Rang der besten Lösung
Der Rang der besten Lösung ist ein Indikator für die Qualität der Bewertungsfunktion. Im Prinzip ist er umso niedriger, je besser die Funktion zwischen guten und schlechten Plazierungen diskriminieren kann. Aber auch die Gesamtzahl der Lösungen beeinflußt diesen Rang. Denn wenn sich eine gute Plazierung weit vorn in der Liste der Vorhersagen befindet, ist die Wahrscheinlichkeit, eine unter Umständen nur minimal bessere Lösung weiter hinten zu finden, bei längeren Listen größer.

Rang der ersten Lösung mit RMS-Abweichung unter 2.0 Å
Der Rang der ersten Lösung mit einer RMS-Abweichung von unter 2.0 Å besagt, wieviele Lösungen man hätte durchgehen müssen, um eine gute Plazierung zu finden. Für praktische Anwendungen ist dieser Rang deshalb wesentlich wichtiger, als der Rang der besten Lösung, die unter Umständen wesentlich weiter hinten in der Liste liegen kann.

Da FlexE die Instanzen der Ensemblestrukturen clustert, was die Proteinstruktur unter Umständen leicht verzerrt, werden in dieser Arbeit zusätzlich die Ränge für einen etwas höherer Schwellwert von 2.5 Å angegeben.

Laufzeit
Die Laufzeit eines Docking-Programms entscheidet vor allem über die Größe der Datensätze, die noch in akzeptabler Zeit bewältigt werden können, und damit über das Anwendungsgebiet für das Programm. Schnelle Verfahren kann man für das Screening von Datenbanken einsetzen, während sich langsamere Methoden, die unter Umständen bessere Vorhersagen machen, eher für das Studium einzelner Komplexe eignen, z.B. im Rahmen einer Leitstrukturoptimierung.

6.2 Anreicherungsfaktoren beim Screening

Ein typisches Anwendungsgebiet für schnelle Docking-Programme ist das Durchsuchen von Datenbanken nach möglichen Bindungspartnern für ein bestimmtes Protein (Screening) [228]. Hier kommt es vor allem darauf an, möglicherweise aktive Verbindungen von definitiv inaktiven zu trennen. Nicht zuletzt diese Anwendung war der Anlaß, die Proteinflexibilität beim Docking zu berücksichtigen, weil man nicht unbedingt davon ausgehen kann, daß die Konformation eines Proteins bei der Bindung verschiedener Liganden unverändert bleibt. Es können vielmehr Konformationsänderungen auftreten, die unter Umständen die Bindungseigenschaften des Proteins verändern.

Das Maß für die Güte eines Verfahrens ist in diesem Fall der *Anreicherungsfaktor*, der angibt, wievielmal mehr aktive Liganden sich tatsächlich unter den von der Docking-Methode als aktiv vorhergesagten Liganden befinden, verglichen mit einer gleich großen zufälligen Auswahl von Liganden aus der Datenbank. Die Bestimmung dieses Faktors ist nicht ganz einfach, weil man dazu wissen muß, wieviele aktive und inaktive Verbindungen die Datenbank insgesamt enthält. Der Einfachheit halber geht man deshalb in der Regel davon aus, daß nur die bekannten Liganden eines Proteins auch wirklich aktiv sind. Alle übrigen Liganden werden als inaktiv betrachtet [10]. In aller Regel muß man dabei nur mit wenigen falsch-negativen rechnen.

Ein solches Screening-Experiment ist aus zwei Gründen mit FlexE bisher nicht durchgeführt worden: Erstens macht diese Experiment erst Sinn, wenn das Verfahren hinreichend robust geworden ist und ausreichend Erfahrungen mit dem Redocking gesammelt worden sind. Zweites erfordert ein solches Screening sehr spezielle Testdaten, nämlich ein Ensemble von Proteinstrukturen, bei dem tatsächlich Konformationsänderungen aufgrund der Bindung verschiedener Liganden erwartet werden und dazu eine größere Menge von bekannten Liganden vorhanden ist. Diese Daten standen jedoch bisher nicht zur Verfügung.

6.3 Vergleich mit anderen Docking-Programmen

Da in den Publikationen über verschiedene Docking-Verfahren ganz unterschiedliche Testdatensätze und Methoden zur Evaluation herangezogen werden, ist ein detailierter Vergleich der Verfahren nicht praktikabel (vgl. auch Abs. 3.3.6). Ein solcher Vergleich kann deshalb nur die Methoden und Modelle gegenüberstellen, sowie globale Aussagen über die Vorhersagequalität und Laufzeit machen.

Vergleichbare Resultate bekommt man nur, wenn die verschiedenen Algorithmen auf denselben Datensatz angewendet werden, der nicht zu klein sein darf, um die Signifikanz des Vergleichs zu gewährleisten (s. Abs. 6.1.1). Ein solcher Vergleich sollte aber nicht von den Entwicklern der Programme selbst, sondern von unabhängigen Anwendern durchgeführt werden, die für alle Verfahren dasselbe Vorwissen mitbringen.

6.3.1 Vergleich von FlexE und FlexX

Verschiedene Proteinkonformationen können mit einem Docking-Programm, das die Proteinstrukturen starr hält, nur durch sequentielles Docking in alle Einzelstrukturen berücksichtigt werden (Kreuz-Docking [79, 109]). Die Lösungen der einzelnen Docking-Läufe muß man in einer automatisch reproduzierbaren Weise, die keinen Gebrauch von der korrekten Plazierung macht, zu einer Ergebnisliste kombinieren. Denn es ist unzulässig, jeweils die Proteinstruktur, die zur besten Vorhersage führt herauszugreifen, und zur Grundlage für den Vergleich zu machen, weil dies in der realen Anwendung auch nicht möglich ist.

Darum werden für den Vergleich von FlexE und FlexX alle Plazierungen, die FlexX für einen Liganden in den einzelnen Strukturen vorhersagt, zusammengefaßt und entsprechend der geschätzten Bindungsenergie neu geordnet. Damit liegen der FlexX-Vorhersage zwar insgesamt mehr Lösungen zugrunde, was aber vertretbar ist, weil insbesondere die ersten Lösungen verglichen werden sollen.

Da beide Programme die Liganden absolut gleich behandeln und auch dieselbe Bewertungsfunktion benutzen, konzentriert sich dieser Vergleich insbesondere auf die Unterschiede bei der Behandlung der Proteinflexibilität.

Kapitel 7

Ergebnisse

Die in den vorherigen Kapiteln dargestellten Modelle und Methoden sind in dem Docking-Programm FlexE realisiert, das auf FlexX basiert. So sind die grundlegenden algorithmischen Konzepte und die Behandlung der Liganden sowie die Benutzungsschnittstelle und die Visualisierung übernommen worden. Aktuelle Weiterentwicklungen von FlexX wie das Docking in kombinatorische Bibliotheken [89], neue Bewertungsfunktionen oder die Verwendung von Pharmakophor-Randbedingungen können daher wegen der engen Verwandtschaft der Programme in Zukunft leicht adaptiert werden.

Dieses Kapitel erläutert zunächst die Parametrisierung der benutzten Docking-Programme FlexE und FlexX und stellt dann den Testdatensatz vor, auf dem das Redocking-Experiment mit FlexE und der Vergleich zwischen FlexE und FlexX basieren. Nachdem die Ergebnisse dieser beiden Tests zusammengefaßt worden sind, wird auf zwei Ensembles etwas mehr im Detail eingegangen: Das eine ist das Beispiel der Aldose-Reduktase vom Anfang (Abs. 3.1), das zweite ein verkleinertes Ensemble für die Dihydrofolat-Reduktase, das nicht für alle gedockten Liganden die Originalproteinstruktur enthält. Ein Vergleich von FlexE mit anderen Ansätzen aus der Literatur, der an Abschnitt 3.3 anknüpft, schließt dieses Kapitel ab.

7.1 Parametrisierung

Zur Modellierung der Proteinflexibilität wurden in FlexE einige Parameter neu eingeführt bzw. in ihrer Bedeutung modifiziert. Die nachfolgende Übersicht gibt die Belegung dieser neuen Parameter an und erläutert kurz, wie die gewählten Werte zustande kommen. Die Belegungen aller weiteren Parameter sind für beide Programme identisch und entsprechen den Defaulteinstellungen von FlexE/X, die sich bereits für FlexX bewährt haben [79]. Die wichtigsten Parameter sind in Tabelle 7.1 zusammengefaßt.

- **Schwellwert für Cluster von Wechselwirkungspunkten (Abs. 5.3): 0.6 Å**
  Je größer dieser Wert ist, desto mehr Punkte werden zusammengefaßt. Ist der Wert aber größer als die Hälfte der Bucketgröße der Hashtabelle, werden Punktpaare falsch einsortiert. Man kann jedoch einen Schwellwert benutzen, der etwas größer als die Hälfte der aktuellen Bucketgröße von 0.9 Å ist, weil die Abfrage benachbarte Buckets berücksichtigt. Dieser Wert wurde empirisch ermittelt (vgl. Abs. 7.4). Von diesem Versuch abgesehen, wurden **bei den folgenden Tests keine Wechselwirkungspunkte geclustert!**

- **Schwellwert für Cluster von Instanzen (Abs. 4.1.1 u. 5.1.3):** **1.0 Å**
  Dieser Schwellwert stellt eine Abwägung zwischen den konkurrierenden Zielen dar, möglichst viele ähnliche Instanzen zusammenzufassen und die Verzerrung der Proteinstruktur gering zu halten. Da eine gewisse Verzerrung durch die Toleranzen bei der Plazierung und der Bewertung ausgeglichen wird, kann man diesen relativ großen Abstand verwenden.

- **Toleranz für Bindungslänge (strukt. Inkompatibilität, Abs. 4.1.2):** **1.0 Å**
  Da davon auszugehen ist, daß die Bindungslängen in den einzelnen Ensemblestrukturen korrekt sind, hat dieser Parameter vor allem die Aufgabe, Fehler auszugleichen, die durch das Clustern der Instanzen entstehen. Deshalb entspricht die Bindungslängentoleranz dem Schwellwert für das Clustern von Instanzen.

- **Toleranz für Überlappung (geom. Inkompatibilität, Abs. 4.1.2):** **5.5 Å^3**
  Dieser Parameter wurde so eingestellt, daß es zu keiner geometrischen Inkompatibilität innerhalb einer einzelnen experimentell bestimmten Proteinstruktur kommt.

- **Toleranz für Überlappung zwischen Ligand u. Protein (Abs. 5.4):** **5.5 Å^3**
  FlexE bestimmt die Überlappung zwischen Ligandatomen und Instanzen. Darum wurde das maximal zulässige Schnittvolumen der Toleranz bei der geometrischen Inkompatibilität angepaßt. FlexX berechnet dagegen die Überschneidung zwischen einem Ligandatom und einzelnen Proteinatomen und erlaubt dafür ein geringeres Volumen von 2.5 Å^3. Beide Programme verwenden als maximale durchschnittliche Überlappung 40% ihres jeweiligen Maximums (s. Abs. 2.4.4).

  FlexX 2.5 Å^3

- **Toleranz für Überlappung d. Wechselwirkungspunkte (Abs. 4.3):** **5.5 Å^3**
  Dieser Parameter hängt eng mit dem vorherigen zusammen und sollte deshalb denselben Wert haben. FlexX berechnet die Überlappung der Punkte wiederum mit einzelnen Proteinatomen und toleriert dabei, wie auch bei den Ligandatomen, maximal 2.5 Å^3 Überschneidung.

  FlexX 2.5 Å^3

Die im folgenden vorgestellten Resultate beruhen auf den Programmversionen FlexE 0.8.0, FlexX 1.8.1 [11, 84] und CORINA 2.4 [92, 93, 94], das zur Berechnung von Ringkonformationen eingesetzt wird.

7.2 Testdatensatz

Der Testdatensatz für diese Arbeit umfaßt 10 Proteinstrukturensembles mit insgesamt 105 experimentell bestimmten Kristallstrukturen aus der PDB [80] und ein Homologiemodell sowie 60 Liganden, die in den gemessenen Komplexen dieser Proteine vorliegen, so daß ihre Bindungsmodi bekannt sind (vgl. Tab. 7.2 u. 7.3).

Da weder das vorgestellte Modell der Proteinflexibilität noch die verwendeten Algorithmen auf experimentell bestimmte Strukturen beschränkt sind, liegt der Grund für die starke Ausrichtung des Testdatensatzes auf Kristallstrukturen allein in der Methodik der Evaluation, die das Wissen über den korrekten Bindungsmodus benötigt, um die vorhergesagten Plazierungen beurteilen zu können.

Parameter	Wert
Faktor für die durchschnittliche Überlappung (vom Maximum s. Text)	0.4
Maximale Anzahl automatisch generierter Basisfragmente	4
Maximale Anzahl von Konformationen pro Basisfragment	30
Maximale Anzahl von Lösungen für den nächsten Anbauschritt	400
Zusätzliche Anzahl von Lösungen je Basisfragment	100
Bucketgröße in der Hashtabelle	0.9 Å
Minimaler Abstand der Wechselwirkungspunkte in der Hashtabelle	0.5 Å
Maximaler Abstand der Wechselwirkungspunkte in der Hashtabelle	10.0 Å
Schwellwert für das Clustern von Dreiecken	1.1 Å
Schwellwert für das Clustern von Wechselwirkungspaaren	0.4 Å
Schwellwert für das Clustern von Plazierungen	0.7 Å

Tabelle 7.1: Parameter für FlexE und FlexX. Die Tabelle enthält die verwendeten Werte der wichtigsten Parameter. Sie entsprechen den Defaultwerten der beiden Programme.

7.2.1 Auswahl der Testdaten

Die Auswahl der Ensembles ist in keiner Weise vollständig oder gar repräsentativ für die PDB. Sie bildet lediglich einen ersten, zufälligen Testdatensatz, der als Basis zur Entwicklung der Methode dient. Um eine zu starke Fokussierung des Algorithmus auf ein spezielles Ensemble zu vermeiden, wurden von Beginn an mehrere Testensembles verwendet.

Es ist schwierig, geeignete Ensembles in der PDB zu finden, denn oft enthält sie nur eine Struktur für ein Protein, und wenn es verschiedene Strukturen für dasselbe Protein gibt, sind diese meist sehr ähnlich, das heißt, ihre Seitenketten haben eine RMS-Abweichungen von weniger als 0.5 Å und der Backbone ist noch stärker konserviert. Eine automatisierte vollständige Suche nach Strukturen mit Konformationsänderungen ist sehr aufwendig, weil man dazu alle Strukturen für ein Protein überlagern und analysieren muß.

Deshalb wurde von Hand in der Datenbank ReliBase [229] nach Proteinstrukturen mit einer Sequenzidentität von mehr als 90% gesucht und manuell geprüft, ob die Strukturen die folgenden Anforderungen erfüllen. Die Suche wurde nicht auf hundertprozentig identische Sequenzen beschränkt. Denn erstens sind auch Mutationen von Interesse und zweitens fehlen in einzelnen Strukturen Aminosäuren, die z.B. nicht aufgelöst werden konnten.

- Die Strukturen sollten einen weitgehend gleichen Backbone-Verlauf haben (mittlere RMS-Abweichung $\leq$ 1.0 Å).
- Es sollten mindestens drei verschiedene Liganden vorhanden sein.
- Es sollten verschiedene Seitenkettenkonformationen vertreten sein, die auf die Bindung unterschiedlicher Liganden zurückzuführen sind. Die Rotamere sollten dabei im Mittel um mindestens 1.0–2.0 Å variieren.
- Es sollte mindestens eine Struktur geben, die ohne Ligand kristallisiert und vermessen wurde.
- Punktmutationen sind erlaubt. Sie sind insbesondere dann interessant, wenn sie die Bindung verschiedener Liganden beeinflussen.
- Flexible Loops sollten maximal zehn Aminosäuren lang sein.

Nicht alle Ensembles des Testdatensatzes erfüllen alle Kriterien. Zwar haben alle Mitglieder eines Ensembles einen sehr ähnlichen Backbone-Verlauf, verschiedene Seitenkettenkonformationen oder Punktmutationen und leicht unterschiedliche Loop-Verläufe, aber aufgrund der beschränkten Datenlage in der PDB sind die Liganden oft sehr ähnlich und manchmal auch identisch.

Für die Aldose-Reduktase wurden drei PDB-Strukturen des Enzyms aus den Linsen von Schweineaugen mit einem Homologiemodell von menschlicher Aldose-Reduktase kombiniert [100], weil es für dieses Protein nur zwei verschiedene Liganden in der PDB gab, als der Datensatz zusammengestellt wurde. In diesem Fall beträgt die Sequenzidentität zwar nur etwa 86%, aber das aktive Zentrum ist sehr stark konserviert.

Tabelle 7.2 gibt einen Überblick über die einzelnen Ensembles. Dort sind für jedes Ensemble die PDB-Kürzel der Strukturen aufgelistet, aus denen es sich zusammensetzt, wobei die PDB-Strukturen markiert sind, deren Liganden als Referenz verwendet werden. Das Intervall der Backbone-Abweichung und die Angabe der Aminosäuren, bei denen es alternative Instanzen bzw. Mutationen im aktiven Zentrum gibt, vermitteln einen Eindruck von der strukturellen Variabilität der Ensembles. Sie bestimmen im wesentlichen die Größe des Suchraums, die hier durch die Zahl der potentiellen Konformationen nach oben abgeschätzt wird. Diese Zahl stellt eine obere Schranke dar, weil ihre Berechnung die Abhängigkeiten zwischen den Instanzen vernachlässigt. Im Anhang findet man darüber hinaus für alle Ensembles Abbildungen des aktiven Zentrums der vereinigten Proteinbeschreibung.

Es wurden nicht alle Liganden verwendet, die in den PDB-Einträgen enthalten sind. Zu kleine (weniger als vier Atome) oder kovalent gebundene Moleküle blieben unberücksichtigt. Die wichtigsten Eigenschaften der übrigen 60 Liganden sind in Tabelle 7.3 zusammengefaßt, Abbildungen ihrer Strukturformeln findet man ebenfalls im Anhang.

7.2.2 Aufbereitung der Daten

Bevor die Ensembles von Proteinstrukturen in FlexE bzw. FlexX verwendet werden können, müssen die Liganden aus den Komplexen extrahiert, minimiert und im mol2-Format [230] gespeichert werden. Die Ensembles werden in sog. EDF-Dateien (ensemble description files) definiert, die die Bindetasche bestimmen, Mehrdeutigkeiten der PDB-Dateien auflösen und festlegen, wie FlexE die einzelnen Proteinstrukturen kombinieren soll. Dieses Format ist an das Format der RDF-Dateien (receptor description file) von FlexX angelehnt. Daher können für FlexX aus einer EDF-Datei automatisch eine Menge von RDF-Dateien erzeugt werden, die die entsprechenden Information für die einzelnen Strukturen eines Ensembles enthalten.

Tabelle 7.2: Übersicht über die Ensembles. Die Tabelle gibt für alle Ensembles die PDB-Codes [80] der Ensemblemitglieder an. Die Liganden der markierten Strukturen (*) dienen als Referenz. Außerdem enthält sie das Intervall der mittleren Backbone Abweichung im aktiven Zentrum und die Anzahl der potentiellen Konformationen im aktiven Zentrum (RMSD) sowie insgesamt. Die Zahl ergibt sich aus dem Produkt der Anzahl von Instanzen pro Segment. Ferner sind die Segmente aufgeführt, die im aktiven Zentrum mehr als eine Instanz umfassen. Bei Mutationen sind beide Kürzel angegeben.

Ensemble	PDB Codes*	RMSD [Å]	# pot. Konf.	Flexible Seg. / Mutationen
Aldose-Reduktase (4 Struk., 3 Lig.)	1ah0* 1ah3* 1ah4 homology model*	0.2 – 0.6	$2.0\ 10^{6}$ ($2.5\ 10^{57}$)	PHE 122, LEU 124, VAL 130, VAL 297 – ALA/THR 304
Alpha-Momorcharin (7 Struk., 4 Lig.)	1mri 1ahc 1mrh* 1mrg* 1aha* 1ahb* 1mom	0.2 – 0.2	$9.6\ 10^{4}$ ($2.4\ 10^{47}$)	GLU 85, GLU 112, ARG 122, LEU/VAL 64, ASN/ASP 110, ASN 68 – MET 72
Carboanhydrase II (16 Struk., 8 Lig.)	2cbb 1h9n 1hec 1mua 1uga 1ugd 1bcd* 1cil* 1cnw* 1cnx* 1cra* 1ray 1cam* 1caz* 1ugb 1zsb*	0.1 – 0.3	$1.5\ 10^{5}$ ($8.0\ 10^{36}$)	HIS 64, ASN 67, GLN 92, LYS 133, VAL 135, GLN 136, CYS 206, ALA/PHE/SER 65, GLN/GLU 106, ASP/HIS 119, HIS/LEU 198, ALA/THR 199, ALA/PRO 202
Carboxypeptidase (16 Struk., 7 Lig.)	5cpa 1arl 1yme 3cpa* 4cpa* 6cpa* 7cpa* 8cpa 1bavA 1bavB 1bavC 1bavD 1cbx* 1cps* 2ctb 2ctc*	0.1 – 0.5	$4.2\ 10^{16}$ ($4.8\ 10^{112}$)	HIS 69, ARG 71, ARG 127, ARG 145, LEU 203, LEU 206, TYR 208, ILE 243, THR 268, GLU 270, THR 274, SER 162 – THR 164, SER 194 – SER 199, ILE 247 – ILE 255,
Dihydrofolat-Reduktase (12 Struk., 12 Lig.)	1dyh* 1dyi* 1dyj* 1jol* 1ra2* 1ra3* 3drc* 1dhj* 1dra* 1drb* 2drc* 4dfr*	0.1 – 0.7	$8.3\ 10^{10}$ ($1.1\ 10^{43}$)	LEU 4, LEU 8, ILE 14, GLY 15 ALA 19, MET 20, TRP 30, LYS 32, ARG33, PHE/TRP 22 ASP/CYS/SER 27, HIS45–LEU54
Isocitrat-Dehyrogenase (14 Struk., 6 Lig.)	1idf 1idd 6icd 7icd 1ika 4icd 5icd* 9icd 1ide* 1idc* 1gro* 1grp* 8icd* 1iso	0.3 – 0.6	$3.6\ 10^{12}$ ($4.5\ 10^{270}$)	ARG 119, ARG 129, ASN 155, ASN 303, LEU 304, ASP 307, SER 310, ASP 311, GLU 336, PHE/TYR 160, ASP/GLU/SER 113 – ALA 117 LEU 103 – THR 105 THR 338 – GLY 340
Mandelat-Racemase (6 Struk., 4 Lig.)	2mnr 1mdr* 1mns 1mdl* 1mra* 1dtn*	0.1 – 1.1	$9.3\ 10^{10}$ ($1.0\ 10^{96}$)	VAL 29, ASN 197, ASN 248, LEU 321, ARG/LYS 166, ASN/ASP 270, GLN/GLU 317, VAL 22 – ALA 25
Ricin (9 Struk., 3 Lig.)	1rtc 1obs 2aai 1fmp* 1apg* 1ifu 1ifs 1ift 1obt*	0.3 – 0.4	$3.2\ 10^{8}$ ($3.8\ 10^{88}$)	ASN 78, TYR 80, VAL 82, ASP 96, ASN 122, ASP 124, ARG 125, GLN 173, ILE 205, THR 206, ASN 209, SER 210, ARG 213, ARG 258 ARG/GLY/HIS 180
Seryl-T-RNA-Synthetase (6 Struk., 4 Lig.)	1sryA 1sryB 1sesA* 1sesB* 1setA* 1setB*	0.3 – 0.5	$4.4\ 10^{20}$ ($6.2\ 10^{162}$)	ARG 157, GLN 211, GLU 227, ARG 256, GLU 258, GLU 345, SER 348, ASN 378, ASN 379, LEU 382, ARG 386, ILE 387, MET 270 – GLU 279
Trypsin (16 Struk., 9 Lig.)	3ptn 2ptc 1tld 1tpo 1taw 1max 1ppc* 1pph* 1tng* 1tnh* 1tni* 1tnj* 1tnk* 1tnl* 1tpp 3ptb*	0.2 – 0.3	$2.2\ 10^{6}$ ($1.2\ 10^{56}$)	ASN 97, THR 98, LEU 99 GLN 175, LYS 188, GLN 192 SER 217, GLN 221, LYS 224 TYR 228

Ensemble	Ligand	Name	Anzahl A	Anzahl B	Anzahl R	Wechselwirk. S3	Wechselwirk. S2	Wechselwirk. S1	Anzahl Frag.	Anzahl Konf.	min. rmsd
Aldose-Reduktase	SBI.1ah0	Sorbinil	17	1	1	6	4	5	2	8	0.18
	TOL.1ah3	Tolrestat	24	6	1	3	9	4	7	23000	0.32
	ZST.model	Zopolrestat	29	5	2	5	9	8	6	29000	0.28
Alpha-Momorcharin	FMC.1mrh	Formycin	19	5	2	13	2	8	6	7400	0.46
	ADN.1mrg	Adenin	10	0	1	6	4	3	1	1	0.14
	ADE.1aha	Adenin	10	0	1	6	4	3	1	1	0.07
	FMP.1ahb	Formycin-5-Monophosphat	23	5	2	15	2	8	6	7400	0.23
Carbo-anhydrase II	FMS.1bcd	Trifuormethansulfonamid	8	2	0	5	0	0	3	3	0.12
	ETS.1cil	Inhibitor ETS	19	4	1	9	2	5	5	1900	0.33
	EG1.1cnw	Benzolsulfonamid EG1	26	15	1	13	7	9	16	$1.1\ 10^{12}$	0.43
	EG2.1cnx	Benzolsulfonamid EG2	22	12	1	11	6	8	13	$3.9\ 10^{09}$	0.23
	TRI.1cra	1,2,4-Triazol	5	0	1	3	2	0	1	1	0.02
	BCT.1cam	Bicarbonat	4	1	0	4	0	0	2	4	0.02
	ACY.1caz	Essigsäure	4	0	0	2	1	0	1	1	0.17
	AZM.1zsb	Acetazolamid	13	3	1	8	2	1	4	290	0.35
Carboxy-peptidase	G-Y.3cpa	Glycyl-Tyrosin	17	8	1	9	6	5	9	130000	0.23
	GLY.4cpa	Glycin	5	2	0	5	0	1	3	9	0.29
	ZAF.6cpa	Phosphonat ZAA=P=(O)F	33	14	2	10	16	7	15	$1.3\ 10^{10}$	0.46
	FVF.7cpa	(BZ-Phe-Val=P=(O)-Phe)	41	17	3	10	22	10	18	$3.5\ 10^{12}$	0.47
	BZS.1cbx	Benzyl-Bersteinsäure	15	5	1	4	6	4	6	86000	0.21
	CPM.1cps	Sulfonamid Inhibitor CMP	16	5	1	7	7	4	6	8600	0.28
	LOF.2ctc	3-Phenyl-Milchsäure	12	4	1	4	6	3	5	450	0.24
Dihydrofolat-Reduktase	DZF.1dyh	5-Deaza-Folsäure	32	11	2	14	10	10	12	$2.2\ 10^{8}$	0.40
	FOL.1dyi	Folsäure	32	11	2	15	9	9	12	$2.4\ 10^{8}$	0.39
	DDF.1dyj	5,10-Dideaza-4H-Folsäure	32	10	2	11	7	10	11	$7.6\ 10^{7}$	0.36
	FFO.1jol	5-Formyl-6-Hydro-Folsäure	34	12	2	15	7	11	13	$2.2\ 10^{9}$	0.49
	FOL.1ra2	Folsäure	32	11	2	15	9	9	12	$9.4\ 10^{8}$	0.35
	MTX.1ra3	Methotrexat	33	10	2	14	10	9	11	$3.9\ 10^{7}$	0.27
	MTX.3drc	Methotrexat	33	10	2	14	10	9	11	$3.9\ 10^{7}$	0.32
	MTX.1dhj	Methotrexat	33	10	2	14	10	9	11	$3.9\ 10^{7}$	0.30
	MTX.1dra	Methotrexat	33	10	2	14	10	9	11	$3.9\ 10^{7}$	0.30
	MTX.1drb	Methotrexat	33	10	2	14	10	9	11	$3.9\ 10^{7}$	0.43
	MTX.2drc	Methotrexat	33	10	2	14	10	9	11	$3.9\ 10^{7}$	0.40
	MTX.4dfr	Methotrexat	33	10	2	14	10	9	11	$3.9\ 10^{7}$	0.42
Isocitrat-Dehyrogenase	ICT.5icd	Isocitronensäure	13	6	0	8	0	3	7	49000	0.25
	ICT.1ide	Isocitronensäure	13	6	0	8	0	3	7	49000	0.22
	OXS.1idc	Oxalbernsteinsäure	13	5	0	7	0	2	6	22000	0.26
	ICT.1gro	Isocitronensäure	13	6	0	8	0	3	7	49000	0.19
	ICT.1grp	Isocitronensäure	13	6	0	8	0	3	7	49000	0.19
	ICT.8icd	Isocitronensäure	13	6	0	8	0	3	7	49000	0.20
Mandelat-Racemase	SAA.1mdr	2-Phenyl-Milchsäure	12	3	1	4	7	1	4	90	0.21
	SMN.1mdl	2-Phenyl-Glycolsäure	11	3	1	4	6	2	4	180	0.14
	SAA.1mra	2-Phenyl-Milchsäure	12	3	1	4	7	1	4	90	0.21
	SAA.1dtn	2-Phenyl-Milchsäure	12	3	1	4	7	1	4	90	0.20
Ricin	FMP.1fmp	Formycin-5'-Monophosphat	23	5	2	15	2	8	6	7400	0.57
	A-G.1apg	Adenyl(3'->5')Guanosin	42	12	4	25	7	13	13	$4.1\ 10^{8}$	0.79
	AMP.1obt	Adenosin-Monophosphat	23	6	2	14	4	8	7	6800	0.52
Seryl-T-RNA-Synthetase	AHX.1ses	Seryl-Hydroxamat-AMP	30	13	2	21	5	10	14	$4.4\ 10^{9}$	0.61
	AMP.1ses	Adenosin-Monophosphat	23	6	2	14	4	8	7	6800	0.62
	SSA.1set	} 5'-O-(N-(L-Seryl)-	29	12	2	21	4	10	13	$9.6\ 10^{8}$	0.61
	SSA.1set	} Sulfamoyl)-Adenosin	29	12	2	20	5	10	13	$3.4\ 10^{9}$	0.54
Trypsin	NAS.1ppc	NAPAP	37	11	3	11	15	11	12	$3.3\ 10^{8}$	0.39
	TOS.1pph	3-TAPAP	30	8	3	9	11	11	9	$2.1\ 10^{6}$	0.51
	AMC.1tng	Aminomethylcyclohexan	8	1	1	3	0	7	2	18	0.12
	FBA.1tnh	4-Fluorbenzylamin	9	1	1	3	5	3	2	6	0.09
	PBN.1tni	4-Phenylbutylamin	11	4	1	3	6	5	5	1500	0.32
	PEA.1tnj	2-Phenylethylamin	9	2	1	3	6	3	3	30	0.46
	PRA.1tnk	3-Phenylpropylamin	10	3	1	3	6	4	4	210	0.49
	TPA.1tnl	Phenylcyclopropylamin	10	1	2	3	6	4	2	6	0.12
	BEN.3ptb	Benzamidin	9	1	1	4	6	1	2	4	0.06

Aufbereitung der Liganden

Die Liganden werden für FlexE und FlexX unter Verwendung des Programmpakets SYBYL [230] nach exakt dem gleichem Verfahren wie folgt aufgearbeitet: Zunächst werden die Ligandkoordinaten der Nicht-Wasserstoffatome aus dem Original-PDB-Eintrag extrahiert. Sie dienen später als Referenz für die Berechnung der RMS-Abweichung. In den Fällen, in denen es gleiche Liganden in verschiedenen PDB-Einträgen gibt, werden alle Liganden als separate Referenzen verwendet, weil sie sich zum Teil in ihrer Position leicht unterscheiden. Eine Eingabedatei für die Liganden erhält man, indem man die korrekten Atomtypen (einschließlich der Hybridisierung) und die korrekten Bindungstypen definiert, Wasserstoffatome hinzufügt, jedem Atom Formalladungen zuweist und die Struktur schließlich energieminimiert. Im allgemeinen werden dabei alle Carboxylat- und alle Phosphorsäuregruppen mit einer negativen Ladung versehen, wohingegen alle Amino-, Amidino- und Guanidinogruppen, aber keine Amidgruppen protoniert werden.

Die Energieminimierung garantiert eine Konformation mit niedriger Energie, deren Bindungslängen und -winkel nahe bei den theoretischen Werten liegen [198]. Diese neue Geometrie und die Tatsache, daß die minimierte Struktur nicht entsprechend der Überlagerung der Ensemblestrukturen verschoben wird, garantieren, daß die Eingabedatei für den Liganden keine impliziten Docking-Informationen über den Komplex enthält.

Aufbereitung der Ensemblestrukturen

Die EDF-Dateien enthalten für jede Struktur eines Ensembles die Definition der Proteinatome (über Kettenbezeichner und Heterogruppen), die beim Docking berücksichtigt werden sollen. Das sind in der Regel alle Atome. Von Multimeren, bei denen das aktive Zentrum vollständig in einem Monomer liegt, wird allerdings nur das entsprechende Monomer verwendet. Bei zwei Ensembles (Carboxypeptidase, Seryl-T-RNA-Synthetase) werden die verschiedenen Monomere eines PDB-Eintrags überlagert und als alternative Konformationen benutzt.

In einigen PDB-Einträgen sind alternativer Koordinaten für bestimmte Atome angegeben. Hier muß man festgelegen, welche Position für eine bestimmte Ensemblestruktur verwendet wird. Im Zweifelsfall wurden beide Konformationen alternativ benutzt.

Die Zuordnung der Wasserstoffpositionen erfolgt im allgemeinen auf der Basis von Default-Regeln. Ausgenommen sind hier nur die Torsionswinkel der Hydroxylgruppen der Aminosäuren Serin, Threonin und die Wasserstoffposition der Histidinseitenkette. In diesen Fällen werden die Torsionswinkel (0°, 180° für Tyrosin; 60°, 180°, 300° für Serin u. Threonin) und die Tautomerie manuell so festgelegt, wie für das Ausbilden von Wasserstoffbrücken am günstigsten ist. Für FlexE werden gegebenenfalls mehrere Alternativen verwendet. Die Seitenketten von Lysin und Arginin werden protoniert und die Carboxylatgruppen von Asparagin und Glutamin werden ionisiert. Wassermoleküle, die in der PDB-Datei enthalten sind, werden entfernt.

Tabelle 7.3: Übersicht über die Liganden. Die Tabelle führt für alle Ensembles die in dieser Arbeit verwendeten Kürzel für die Liganden auf. Für die Liganden sind die (Trivial-) Namen, die Anzahl der Schweratome (A), der Bindungen (B) und die Anzahl der Ringe (R) sowie die Zahl der Wechselwirkungen getrennt nach den Stufen (S1, S2, S3) angegeben. Außerdem findet man die Zahl der Fragmente und die Zahl potentieller Konformationen. Die letzte Spalte gibt die RMS-Abweichung zwischen der Konformation des Liganden in der Kristallstruktur und der Konformation wieder, die der gemessenen Struktur unter Verwendung der Torsionsdatenbank am nächsten kommt.

Um das aktive Zentrum der Proteine zu bestimmen, werden alle Mitglieder eines Ensembles zusammen mit den entsprechenden Referenzliganden überlagert. Alle Atome, die weniger als 6.5 Å von einem der Referenzligandatome entfernt sind, gehören zum aktiven Zentrum. Auf diese Weise wird die Bindetasche durch die Vereinigung aller Referenzliganden definiert. Zusätzlich wird die komplette Aminosäure ausgewählt, wenn wenigstens ein Atom im aktiven Zentrum liegt.

Die überlagerten Proteinstrukturen und Referenzliganden dienen als Input für das Kreuz-Docking-Experiment mit FlexX, für das die entsprechenden RDF-Dateien aus den EDF-Dateien erzeugt werden. Somit verwenden beide Programme für den Vergleich identisch aufgearbeitete Proteinstrukturen mit derselben Definition des aktiven Zentrums.

7.3 Redocking mit FlexE

Die Liganden aller Ensembles des Testdatensatzes werden mit FlexE jeweils in das aktive Zentrum der vereinigten Proteinbeschreibung plaziert, die aus allen Proteinstrukturen des entsprechenden Ensembles entsteht. Dann werden die symmetriekorrigierten RMS-Abweichung aller Vorhersagen und die Laufzeiten der verschiedenen Phasen des Verfahrens bestimmt. Als Referenz für die RMS-Abweichung dient jeweils die Position des Liganden in der überlagerten, experimentell bestimmten Proteinstruktur.

7.3.1 Qualität der Plazierungen

Wie in Abschnitt 6.1.3 bereits ausgeführt wurde, gilt im Rahmen dieser Auswertung die Plazierung mit der geringsten Abweichung zur Referenzstruktur als die *beste* Lösung, weil sie die korrekte Lösung am besten reproduziert. Dagegen wird die Lösung, die sich auf dem ersten Rang befindet, als *erste* Lösung bezeichnet.

Die einzelnen Ergebnisse des Redockings mit FlexE stellt Tabelle 7.4 dar. Sie zeigt für jeden Liganden die Anzahl der vorhergesagten Plazierungen, die RMS-Abweichung der ersten Lösung, der besten Lösung unter den ersten zehn Vorhersagen und der besten Lösung von allen Plazierungen unabhängig vom Rang. Zusätzlich sind der Rang der besten Lösung und der Rang der ersten Lösung mit einer RMS-Abweichung von weniger als 2.0 Å bzw. 2.5 Å aufgelistet. Eine Statistik über die jeweilige Zahl der Lösungen mit einer RMS-Abweichung bis zu 1.0 Å, 1.5 Å, 2.0 Å bzw. 2.5 Å findet man in Tabelle 7.5.

FlexE findet für etwa ein Drittel aller Liganden auf dem ersten Rang eine akzeptable Plazierung mit einer RMS-Abweichung von bis zu 2.0 Å. Die Anzahl erhöht sich auf zwei Drittel, wenn man die ersten zehn Lösungen betrachtet. Unter den Plazierungen mit der geringsten Abweichung zur Referenzstruktur befinden sich rund 80% Lösungen mit einer RMS-Abweichung von weniger als 2.0 Å. Im Durchschnitt befinden sich diese Vorhersagen allerdings auf Rang 25. Für einen Liganden (AMC.1tng) belegt die beste Plazierungen auch den ersten Rang und weicht dabei lediglich um 0.7 Å von der Kristallposition ab.

Tabelle 7.4: FlexE-Vorhersagen. Für jeden Liganden sind die Anzahl der vorhergesagten Plazierungen, die RMS-Abweichung der ersten Lösung, der besten Lösung unter den ersten zehn Vorhersagen und der besten Lösung von allen Plazierungen unabhängig vom Rang sowie der Rang der ersten Lösung mit einer RMS-Abweichung von weniger als 2.0 Å bzw. 2.5 Å angegeben.

Ensemble	Ligand	Anz. Lös.	(Min.) RMSD [Å] Rang 1	Rang 10	Beste Lösung RMSD [Å]	Rang	< 2.0 Å Rang	< 2.5 Å Rang
Aldose-	SBI.1ah0	153	0.58	0.54	0.54	2	1	1
Reduktase	TOL.1ah3	116	1.09	1.05	1.05	5	1	1
	ZST.model	195	6.74	6.72	0.64	133	133	133
Alpha-	FMC.1mrh	270	2.13	1.75	0.80	119	7	1
Momorcharin	ADN.1mrg	161	4.30	1.11	0.97	13	10	10
	ADE.1aha	163	3.27	0.85	0.76	13	10	7
	FMP.1ahb	445	1.60	1.42	0.85	116	1	1
Carbo-	FMS.1bcd	175	3.33	1.47	1.47	2	2	2
anhydrase II	ETS.1cil	348	2.76	2.69	2.35	95	-	95
	EG1.1cnw	178	7.86	7.86	6.83	111	-	-
	EG2.1cnx	284	4.70	4.70	4.10	65	-	-
	TRI.1cra	502	5.84	0.82	0.79	82	5	5
	BCT.1cam	163	6.82	2.51	1.96	12	12	11
	ACY.1caz	285	2.18	1.18	0.96	41	2	1
	AZM.1zsb	368	6.48	3.16	1.85	302	240	82
Carboxy-	G-Y.3cpa	233	1.81	1.69	1.05	210	1	1
peptidase	GLY.4cpa	500	3.20	3.17	1.63	402	237	42
	ZAF.6cpa	176	7.35	7.33	7.31	53	-	-
	FVF.7cpa	122	6.51	5.37	5.08	24	-	-
	BZS.1cbx	247	6.40	6.03	1.53	67	16	16
	CPM.1cps	492	4.97	1.02	1.00	21	2	2
	LOF.2ctc	402	2.44	2.32	1.72	140	140	1
Dihydrofolat-	DZF.1dyh	111	2.21	2.00	1.86	60	4	1
Reduktase	FOL.1dyi	117	2.17	1.84	1.81	57	4	1
	DDF.1dyj	468	5.43	1.63	1.58	123	8	8
	FFO.1jol	67	8.49	5.37	5.21	43	-	-
	FOL.1ra2	111	2.30	1.91	1.91	4	4	1
	MTX.1ra3	284	1.50	0.81	0.53	59	1	1
	MTX.3drc	231	1.21	0.50	0.50	7	1	1
	MTX.1dhj	301	1.12	0.67	0.46	11	1	1
	MTX.1dra	290	1.10	0.70	0.59	51	1	1
	MTX.1drb	284	1.23	0.98	0.64	65	1	1
	MTX.2drc	289	1.05	0.97	0.61	36	1	1
	MTX.4dfr	284	1.32	0.66	0.65	60	1	1
Isocitrat-	ICT.5icd	258	4.35	1.47	1.12	21	8	8
Dehyrogenase	ICT.1ide	257	4.29	1.39	1.01	77	7	7
	OXS.1idc	277	3.77	1.91	1.68	16	9	9
	ICT.1gro	255	4.29	1.53	1.05	99	9	9
	ICT.1grp	252	4.38	2.91	0.89	90	11	11
	ICT.8icd	255	4.25	1.67	1.00	109	7	7
Mandelat-	SAA.1mdr	307	1.85	0.99	0.56	99	1	1
Racemase	SMN.1mdl	286	2.54	1.45	0.83	175	8	8
	SAA.1mra	307	1.08	0.82	0.54	27	1	1
	SAA.1dtn	303	1.95	0.86	0.30	168	1	1
Ricin	FMP.1fmp	469	5.42	2.71	1.29	200	40	11
	A-G.1apg	230	8.22	4.44	3.26	91	-	-
	AMP.1obt	541	3.11	3.11	1.50	298	238	38
Seryl-T-RNA-	AHX.1ses	186	3.28	2.62	2.46	149	-	47
Synthetase	AMP.1ses	466	5.87	2.48	2.45	58	-	8
	SSA.1set	185	2.35	1.95	1.92	20	5	1
	SSA.1set	184	2.54	2.28	2.12	14	-	2
Trypsin	NAS.1ppc	405	3.04	1.98	1.10	16	4	2
	TOS.1pph	173	3.75	3.75	1.03	33	33	33
	AMC.1tng	464	0.70	0.70	0.70	1	1	1
	FBA.1tnh	235	0.55	0.48	0.48	3	1	1
	PBN.1tni	375	0.96	0.96	0.84	15	1	1
	PEA.1tnj	500	1.86	0.61	0.61	2	1	1
	PRA.1tnk	350	1.97	0.57	0.57	8	1	1
	TPA.1tnl	200	0.93	0.89	0.89	2	1	1
	BEN.3ptb	500	0.63	0.27	0.27	5	1	1

A kompletter Datensatz

	erster Rang				
rms≤	1.0	1.5	2.0	2.5	min
#	6	15	21	28	60
%	10.0	25.0	35.0	46.7	100.0

	ersten 10 Ränge				
rms≤	1.0	1.5	2.0	2.5	min
#	20	29	40	43	60
%	33.3	48.3	66.7	71.7	100.0

	unabhängig vom Rang				
rms≤	1.0	1.5	2.0	2.5	min
#	29	39	50	54	60
%	48.3	65.0	83.3	90.0	100.0
∅	24.0	25.2	24.7	11.9	73.3
σ	37.8	60.6	60.2	25.0	80.8

B bereinigter Datensatz

	erster Rang				
rms≤	1.0	1.5	2.0	2.5	min
#	6	9	13	19	45
%	13.3	20.0	28.9	42.2	100.0

	ersten 10 Ränge				
rms≤	1.0	1.5	2.0	2.5	min
#	12	19	27	29	45
%	26.7	42.2	60.0	64.4	100.0

	unabhängig vom Rang				
rms≤	1.0	1.5	2.0	2.5	min
#	19	26	36	39	45
%	42.2	57.8	80.0	86.7	100.0

Die Statistik zu den Rängen wurde nicht berechnet, weil dazu Lösungen hätten zusammengefaßt werden müssen.

Tabelle 7.5: FlexE-Ergebnisse. Die Tabellen fassen die Anzahl der Lösungen mit einer RMS-Abweichung unter 1.0, 1.5, 2.0, 2.5 Å bzw. mit minimaler Abweichung in absoluten Zahlen (#) und in Prozent (%) jeweils für den ersten Rang, die ersten zehn Ränge und für alle Ränge zusammen. Für den letzten Fall sind außerdem der durchschnittliche (∅) Rang der ersten Lösung unterhalb des Schwellwertes und die dazugehörige Standardabweichung (σ) angegeben. Die Tabellen unter A beziehen sich auf den gesamten Testdatensatz, während die Tabellen unter B nur den um identische Liganden bereinigten Datensatz berücksichtigen.

Um die einzelnen Docking-Ergebnisse besser nachvollziehen zu können, werden die Qualität der Vorhersagen von FlexE im folgenden in knapper Form mit den Eigenschaften der einzelnen Ensembles (vgl. Tab. 7.2) und denen ihrer Liganden (vgl. Tab. 7.3) in Beziehung gesetzt und Besonderheiten erläutert. Abbildungen der aktiven Zentren der vereinigten Proteinbeschreibungen und der Ligandstrukturen befinden sich im Anhang.

Aldose-Reduktase
Das Ensemble der Aldose-Reduktase besteht aus drei Kristallstrukturen und einem Homologiemodell, deren Konformationen sich im hydrophoben Teil der Bindetasche stark unterscheiden. Während das Sorbinil (SBI.1ah0) und das Tolrestat (TOL.1ah3) von FlexE gut gedockt werden können, findet man für das Zopolrestat (ZST.model) erst auf Rang 133 eine akzeptable Lösung, weil die relativ enge Tasche in diesem Fall zu einem großen Strafterm für die Überlappung zwischen Protein und Ligand führt.

Alpha-Momorcharin
Das Alpha-Momorcharin hat ein relativ kleines aktives Zentrum mit einigen alternativen Seitenkettenkonformationen und Mutationen im Ensemble. Für alle Liganden gibt es unter den ersten zehn Plazierungen Lösungen mit RMS-Abweichungen von unter 1.5 Å zur Lage des Moleküls im experimentell bestimmten Komplex. Die besten Vorhersagen liegen bei 0.54 Å, 1.05 Å und 0.64 Å.

Carboanhydrase II
Das aktive Zentrum der Carboanhydrase II ist groß und strukturell recht stark konserviert.

Die Alternativen bestehen in diesem Fall vor allem in Mutationen. An diesem Beispiel zeigt sich, daß sowohl extrem kleine Liganden (BCT.1cam) als auch extrem große Moleküle (EG1.1cnw, EG2.1cnx) für FLEXE problematisch sind. Denn für die kleinen Liganden gibt es in einer großen Tasche viele verschiedene Plazierungen, die energetisch günstig sind, während bei den großen Molekülen die Greedy-Aufbaustrategie den Suchraum offenbar zu stark einschränkt. Dementsprechend findet sich für EG1.1cnw und EG2.1cnx keine Lösung unter 2.5 Å und für den kleinen Liganden BCT.1cam sind 1.96 Å für die beste Plazierung (Rang 12) viel relativ zu seiner Größe.

Carboxypeptidase
Die Carboxypeptidase hat das größte aktive Zentrum aller Testensembles mit einer großen strukturellen Variabilität. Dadurch ergeben sich eine Vielzahl von Kombinationsmöglichkeiten und zahlreiche unterschiedliche Plazierungen mit ähnlicher Energie. Aus diesem Grunde findet man die Vorhersagen, die der Referenzstruktur nahe kommen, in einigen Fällen erst sehr weit hinten in der Liste der Lösungen. Daß für die Liganden ZAF.6cpa und FVF.7cpa überhaupt keine akzeptable Lösung gefunden wird, liegt an ihrer Größe, da sie aus 15 bzw. 18 Fragmenten bestehen.

Dihydrofolat-Reduktase
Die Strukturunterschiede bei der Dihydrofolat-Reduktase sind nicht sehr groß, allerdings gibt es einen beweglichen Loop (HIS45–LEU54) und einige Mutationen. Die Liganden dieses Ensembles bilden zwei Gruppen: Die erste umfaßt die Folsäure und ihre Derivate. Die zweite besteht aus den Methotrexat-Molekülen (MTX) der verschiedenen PDB-Einträge, die sich, wenn auch nur geringfügig, in ihrer Position in der Bindetasche unterscheiden. Außerdem weisen die Proteine in den dazu gehörigen Komplexe unterschiedliche Konformationen auf.

FLEXE sagt für die Heterozyklen beider Gruppen ähnliche Plazierungen vorher, obwohl die Pteridinringe der Methotrexat-Moleküle in den Kristallstrukturen relativ zu denen der Folsäure(derivate) um 180° gedreht sind (s. Abb. 7.1). Die Ursache für diese falsche Vorhersage liegt darin, daß die Pteridinringe der Folsäure(derivate) in der protonierten Enolform gedockt wurden. Diese Form wurde verwendet, weil das Wassermolekül, das im Falle der Folsäure(derivate) eine Wasserstoffbrücke vermittelt, bei der Plazierung des Methotrexates aber nicht gebraucht wird, nicht berücksichtigt werden konnte, da FLEXE in der getesteten Version noch keine optionalen Wassermoleküle behandeln kann.

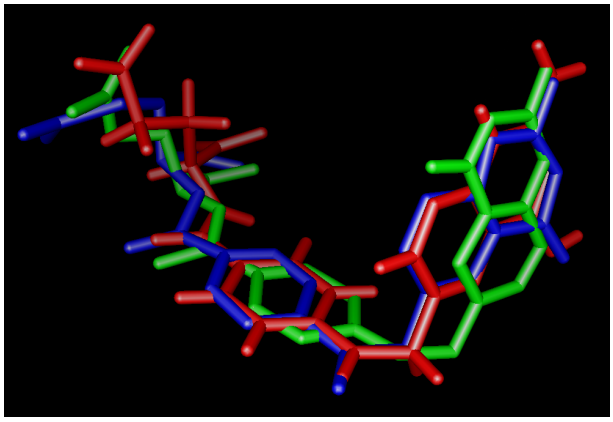


Abbildung 7.1: Bindungsmodus von Folsäure und Methotrexat in Dihydrofolat-Reduktase. Der Heterozyklus der Folsäure (rot) wird in einer ähnlichen Orientierung plaziert, wie der Heterozyklus des Methotrexats (blau), der jedoch im Vergleich zur Kristallstruktur der Folsäure (grün) um 180° gedreht ist.

Isocitrat-Dehydrogenase
Die Isocitrat-Dehydrogenase-Strukturen variieren vor allem in den Loop-Konformationen. Dadurch ergibt sich eine relativ große Zahl potentieller Konformationen (10^{12}). Zwar findet man auf dem ersten Rang in keinem Fall eine annehmbare Lösung, aber unter den ersten 11 Vorhersagen ist in allen Fällen eine Plazierung, die um weniger als 2.0 Å abweicht, und die besten Lösungen liegen fast alle im Bereich von etwa 1.0 Å.

Mandelat-Racemase
Unter den Mandelat-Racemase-Strukturen treten ebenfalls große Backbone-Abweichungen auf. Deshalb gibt es trotz der kleinen Bindetasche ziemlich viele alternative Konformationen. Dennoch können alle Liganden gut plaziert werden und unter den ersten zehn Vorhersagen befindet sich jeweils eine Lösung mit einer Abweichung von weniger als 1.5 Å zur Referenzstruktur.

Ricin
Dieses Ensemble enthält mit 6 Mitgliedern relativ viele Proteinstrukturen, die an keinen Liganden gebunden sind. FlexE findet zwar für die beiden kleineren Liganden FMP.1fmp und AMP.1obt gute Lösungen mit 1.29 Å bzw. 1.50 Å RMS-Abweichung, aber sie befinden sich sehr weit hinten in der Liste. Das Dinukleotid A-G.1apg kann dagegen gar nicht plaziert werden. Der Grund dafür liegt vor allem darin, daß sich die eine Hälfte des Liganden in der Kristallstruktur gar nicht in der Bindetasche, sondern im umgebenden Lösungsmittel befindet. Dies könnte zwar auch ein Artefakt sein, da bei allen verwendeten PDB-Strukturen die B-Kette fehlt, die unter Umständen die Bindetasche abdeckt. In jedem Fall kann diese Plazierung aber so nicht reproduziert werden.

Seryl-T-RNA-Synthetase
Auch für die Liganden der Seryl-T-RNA-Synthetase kann FlexE die Kristallpositionen nur schlecht vorhersagen. Das liegt einerseits an der Konformation der Liganden, die im Rahmen des Modells nur mit rund 0.6 Å Abweichung reproduziert werden kann, andererseits an der Variabilität in der Bindetasche, die zu extrem vielen (10^{20}) potentiellen Kombinationsmöglichkeiten führt.

Trypsin
Die Ensemblemitglieder sind beim Trypsin strukturell ziemlich ähnlich, so daß sich trotz der 16 Strukturen nur 10^6 potentielle Konformationen für das aktive Zentrum ergeben. Die kleinen Liganden des Ensembles lassen sich alle aufgrund ihrer charakteristischen Amidingruppe gut plazieren. Deshalb findet FlexE für sie bereits auf dem ersten Rang Lösungen mit niedrigen RMS-Abweichungen. Für die beiden größeren Liganden NAPAP (NAS.1ppc) und 3-TAPAP (TOS.1pph) befinden sich dagegen die jeweils ersten Lösungen mit einer RMS-Abweichung unter 2.0 Å erst auf dem vierten bzw. 33. Rang.

7.3.2 Gleiche Liganden für ein Ensemble

Einige Ensembles des Testdatensatzes enthalten mehrfach den gleichen Liganden. Das Ensemble der Dihydrofolat-Reduktase umfaßt beispielsweise sieben Proteinstrukturen, die alle im Komplex mit Methotrexat bestimmt worden sind. Nach der Überlagerung der Strukturen sind die Positionen dieser Liganden zwar sehr ähnlich, aber nicht identisch. Man kann dies in Tabelle 7.3 daran erkennen, daß die RMS-Abweichungen zwischen der

Referenzstruktur und der Ligandkonformation, die dieser im Rahmen des Konformationsmodells am nächsten kommt, unterschiedlich ausfallen.

Die RMS-Abweichungen der vorhergesagten Plazierungen variieren deshalb für diese Liganden ebenfalls, weil für die Bestimmung der Abweichung jeweils die Position des Liganden als Referenz dient, die er in der überlagerten Proteinstruktur einnimmt, aus der er extrahiert wurde. Im allgemeinen beträgt diese Streuung der RMS-Abweichung etwa 0.5 Å für einen Liganden. So haben alle Methotrexatplazierungen auf Rang 1 eine RMS-Abweichung im Bereich von 1.0 Å bis 1.5 Å; die der jeweils besten Lösung der ersten zehn Ränge liegt zwischen 0.5 Å und 1.0 Å (vgl. Tab. 7.4). Sie fallen damit für die Statistik über die jeweilige Zahl der Lösungen mit einer RMS-Abweichung bis 1.0 Å, 1.5 Å, 2.0 Å bzw. 2.5 Å in dieselben Intervalle. Da die Zahl der Lösungen nicht in allen Intervallen gleichmäßig anwächst, beeinflussen die mehrfach vorhandenen Liganden diese Statistik.

Die Tabelle 7.5B zeigt daher noch einmal die RMS-Statistik, wobei gleiche Liganden nur einmal berücksichtigt wurden. Insgesamt verringert sich durch diese veränderte Zählweise der Anteil der Lösungen für fast alle Schwellwerte, wobei der Effekt stärker ausfällt, wenn man nur die ersten zehn Ränge oder nur den ersten Rang betrachtet. Allein die Anzahl der Plazierungen auf Rang Eins mit einer RMS-Abweichung bis 1.0 Å bleibt konstant, wodurch sich der prozentuale Anteil dieser Lösungen leicht erhöht.

7.3.3 Laufzeitverhalten von FLEXE

Alle in dieser Arbeit vorgestellten Berechnungen wurden auf einer Sun Ultra 10 durchgeführt, die über einen Ultra SPARC 2e Prozessor mit einer Taktfrequenz von 440 MHz und 512 MB Hauptspeicher verfügt. Tabelle 7.6 enthält für alle Ensembles die Zahl der Wechselwirkungspunkte und deren Paare als Maß für die Größe der Hashtabelle sowie, aufgeschlüsselt nach den verschiedenen Phasen des Algorithmus, die Laufzeiten, die FLEXE im Durchschnitt für die Plazierung eines Liganden in die vereinigte Proteinbeschreibung benötigt.

Die hier aufgeführte Zeit für die Vorverarbeitung bezieht sich nur auf die Proteinseite. Das Einlesen der Liganden und die externe Berechnung ihrer Ringkonformationen, die im allgemeinen nur wenige Sekunden dauert, ist nicht berücksichtigt. Die Preprozessingphase kann man noch einmal in zwei Schritte unterteilen: erstens die Vorbereitung der vereinigten Proteinbeschreibung, die das Einlesen und Überlagern der Ensemblestrukturen, das Clustern der Instanzen und das Erzeugen und Zerlegen des Inkompatibilitätsgraphen umfaßt; zweitens den Aufbau der Hashtabelle aus Paaren von Wechselwirkungspunkten. Beide Schritte braucht man nur einmal durchführen, wenn mehrere Liganden gedockt werden. Somit erhöht sich die Laufzeit für die Vorverarbeitung im Falle mehrerer Liganden nicht.

Die Vorbereitung der vereinigten Proteinbeschreibung benötigt in der Regel ebenfalls nur wenige Sekunden. Die Laufzeit hängt im wesentlichen von der Größe und den Verknüpfungen des Inkompatibilitätsgraphen ab. Diese Parameter können jedoch a priori nur schlecht abgeschätzt werden, weil sie in komplexer Weise von verschiedenen Faktoren, wie beispielsweise der Zahl von Strukturen, dem Konservierungsgrad, der Verteilung von alternativen Instanzen und der Größe des aktiven Zentrums beeinflußt werden. Im allgemeinen kann man aber davon ausgehen, daß die Vorbereitung großer Bindetaschen wie z.B. bei der Carboxypeptidase oder bei Strukturen mit mehreren Loop-Konformationen (Isocitrat-Dehydrogenase) sowie in Fällen mit vielen alternativen Instanzen (Seryl-T-RNA-Synthetase) mehr Zeit braucht.

Ensemble	#	Hashtabelle Punkte	Paare	Vorverarb. [h:m:s] Prep.	Hashtab.	Docking [m:s] Basisp.	Aufbau
Aldose-Reduktase	4	7099	7 600 514	6.22	22:07.57	4:50.41	16.64
Alpha-Momorcharin	7	4278	3 430 526	6.32	7:33.73	2:07.55	15.35
Carboanhydrase II	16	3712	1 810 278	13.23	3:37.67	2:20.58	19.12
Carboxypeptidase	16	13338	30 665 256	22.42	2:59:42.58	12:52.40	1:12.84
Dihydrofolat-Reduktase	12	5531	3 478 479	7.43	11:55.56	1:28.47	52.76
Isocitrat-Dehydrogenase	14	6419	7 430 572	38.74	25:58.31	5:56.50	42.59
Mandelat-Racemase	6	4701	5 496 722	9.75	21:18.54	2:28.57	38.81
Ricin	9	5817	5 009 379	9.80	22:49.63	1:28.61	51.26
Seryl-T-RNA-Synthetase	6	5377	5 124 088	23.69	24:06.88	8:06.86	8:58.61
Trypsin	16	6546	3 198 583	11.00	10:36.76	1:09.72	10.36
Durchschnitt	10.6	6282	7 324 440	14.86	32:58.72	4:16.97	1:25.83

Tabelle 7.6: FLEXE-Laufzeit. Die Tabelle zeigt für jedes Ensemble die Anzahl der Strukturen (#) und der Wechselwirkungspunkte sowie die daraus resultierende Zahl von Punktpaaren, die in der Hashtabelle abgelegt werden. Dem gegenüber gestellt sind die durchschnittlichen Laufzeiten für einen Liganden für die verschiedenen Phasen des Algorithmus: die Vorverarbeitung bestehend aus Preprozessing (Einlesen u. Überlagern der Strukturen etc.) und dem Aufbau der Hashtabelle sowie dem eigentlichen Docking, das die Basisplazierung und den inkrementellen Aufbau des Liganden umfaßt. Die Zeiten wurden auf einer Sun Ultra 10 (Ultra SPARC 2e, 440 MHz, 512 MB) gemessen.

Der Zeitbedarf für den Aufbau der Hashtabelle wächst stärker als linear mit der Anzahl der Wechselwirkungspunktpaare im aktiven Zentrum der vereinigten Proteinstruktur, die in die Hashtabelle aufgenommen werden, weil die Paare mit derselben Adresse in sortierten Listen abgelegt werden (vgl. Abs. 5.3). Darum braucht FLEXE fast drei Stunden, um die Hashtabelle für die Carboxypeptidase zu erstellen, deren aktives Zentrum mehr als 13 000 Wechselwirkungspunkte und 30 Mio Punktpaare enthält. Erschwerend kommt hinzu, daß dafür rund 1 GB Speicherplatz benötigt wird, so daß der Rechner Daten auf die Festplatte auslagern muß.

Für die Plazierung der Basisfragmente sind im Durchschnitt gut vier Minuten erforderlich. Bei der Laufzeit der Basisplazierungen spielen viele Parameter eine Rolle. Auf der Ligandseite sind dies vor allem die Zahl der Basisfragmente, die Anzahl und Verteilung von Wechselwirkungszentren sowie die Zahl von Konformationen der Basisfragmente. Auf der Proteinseite kommen die Anzahl der gefundenen Dreiecke sowie die Größe und Zahl von Zusammenhangskomponenten des Inkompatibilitätsgraphen hinzu, die zur Bewertung der Plazierung nach optimalen unabhängigen Mengen durchsucht werden müssen. Die Vielzahl von Wechselwirkungspunktpaaren bei der Carboxypeptidase ist deshalb der Grund für die extrem lange Laufzeit, während bei der Seryl-T-RNA-Synthetase eine sehr große Zusammenhangskomponente dafür verantwortlich ist (s. Abs. 7.3.4).

Der Aufbau der Liganden in der Bindetasche erfolgt mit zwei Ausnahmen innerhalb einer Minute und ist damit wesentlich schneller als die Basisplazierung. Die Zeit, die diese Phase in Anspruch nimmt, wird im wesentlichen von zwei Faktoren bestimmt: von der Anzahl der anzubauenden Fragmente und von der Größe und Zahl von Zusammenhangskomponenten des Inkompatibilitätsgraphen, die nach jedem Anbauschritt durchsucht werden müssen. Die sehr große Zusammenhangskomponente bei der Seryl-T-RNA-Synthetase führt deshalb auch beim Ligandaufbau zu einer Laufzeit von fast neun Minuten.

Ensemble	einel. Zhgk.	mehrelem. Zhgk. Z	S	I	I/S	Anzahl Zhg.-Komponenten mit ... Knoten 2	3	4	5	6	7	8	9	10	<10
Aldose-Reduktase	63	10	17	41	2.4	7	2	-	-	-	-	-	-	-	1 (21)
Alpha-Momorcharin	50	11	11	34	3.1	6	1	1	3	-	-	-	-	-	-
Carboanhydrase II	50	13	13	34	2.6	8	2	3	-	-	-	-	-	-	-
Carboxypeptidase	46	28	33	122	3.7	15	4	3	1	3	1	-	1	-	1 (29)
Dihydrofolat-Reduktase	54	19	30	71	2.4	14	2	1		1	-	-	-	-	1 (27)
Isocitrat-Dehydrogenase	24	24	27	86	3.6	10	7	1	2	1	1	1	-	1	-
Mandelat-Racemase	34	8	25	72	2.9	4	1	2	-	-	-	-	-	-	1 (53)
Ricin	48	16	18	59	3.3	9	1	1	-	3	-	2	-	-	-
Seryl-T-RNA-Synthetase	61	17	37	147	4.0	10	6	-	-	-	-	-	-	-	1 (109)
Trypsin	58	10	11	50	4.6	4	-	3	-	-	-	1	-	1	1 (12)

Tabelle 7.7: FLEXE: Zusammenhangskomponenten. Die Tabelle zeigt für jedes Ensemble die Anzahl der ein- und mehrelementigen Zusammenhangskomponenten (Zhgk.) des erweiterten aktiven Zentrums. Die einelementigen Zusammenhangskomponenten enthalten jeweils genau eine Instanz aus einem Segment, das nur eine Instanz umfaßt. Für die mehrelementigen Zusammenhangskomponenten sind neben ihrer Anzahl (Z) auch die Zahl der Segmente (S) und Instanzen (I) sowie deren Quotient (I/S) angegeben, die diese Zusammenhangskomponenten insgesamt umfassen. Ferner ist eine Statistik über die Anzahl der Knoten der mehrelementigen Zusammenhangskomponenten dargestellt. Für die Zusammenhangskomponenten mit mehr als zehn Instanzen ist deren Anzahl in Klammern angegeben.

7.3.4 Anzahl und Größe der Zusammenhangskomponenten

Die Anzahl und vor allem die Größe der Zusammenhangskomponenten des Inkompatibilitätsgraphen haben einen entscheidenden Einfluß auf die Laufzeit des Docking-Algorithmus. Tabelle 7.7 zeigt für alle Ensembles des Testdatensatzes eine Statistik über die Anzahl der Zusammenhangskomponenten und die Zahl der Knoten, aus denen die einzelnen Komponenten bestehen. Außerdem gibt die Tabelle an, wieviele Segmente jeweils von allen Zusammenhangskomponenten des erweiterten aktiven Zentrums erfaßt werden.

Die einelementigen Zusammenhangskomponenten, deren Anzahl in der zweiten Spalte angegeben ist, enthalten jeweils genau eine Instanz eines Segments, das ebenfalls jeweils genau eine Instanz umfaßt. Da für eine gültige Proteinkonformation aus jedem Segment genau eine Instanz ausgewählt werden muß, sind diese Instanzen Bestandteil aller gültigen Proteinkonformationen und ihre Energiebeiträge fließen in jedem Fall in die Gesamtbewertung ein. Eine Auswahl von Instanzen ist deshalb bei den einelementigen Zusammenhangskomponenten nicht erforderlich, so daß ihre Anzahl die Laufzeit der Suche nach unabhängigen Mengen nicht beeinflußt.

Die folgenden Spalten beziehen sich alle auf mehrelementige Zusammenhangskomponenten. Angegeben sind die Anzahl der Komponenten (Z), die Zahl der Segmente (S), die sie überdecken, und die Gesamtzahl ihrer Knoten bzw. Instanzen (I) sowie der Quotient I/S. Danach folgt ein Histogramm über die Zahl der Knoten der einzelnen Zusammenhangskomponenten.

Das Verhältnis I/S von Instanzen (I) und überdeckten Segmenten (S) gibt an, wieviele alternative Instanzen es im Durchschnitt pro Segment gibt, wobei nur die Segmente berücksichtigt werden, bei denen es überhaupt Alternativen gibt. Dieser Quotient ist unabhängig von der Größe der Ensembles und liegt in allen Fällen zwischen 2.4 und 4.6.

Wenn die Anzahl der Zusammenhangskomponenten gleich der Zahl der überdeckten Segmente ist, bedeutet das, daß die einzelnen Komponenten jeweils nur aus den Cliquen im Inkompatibilitätsgraphen bestehen, die die Instanzen der Segmente bilden. Dementsprechend gibt es in diesen Fällen keine großen Zusammenhangskomponenten (Alpha-Momorcharin, Carboanhydrase II). Übersteigt die Zahl der überdeckten Segmente dagegen die Anzahl der Komponenten deutlich, so existieren große Zusammenhangskomponenten (Aldose-Reduktase, Carboxypeptidase, Dihydrofolat-Reduktase, Mandelat-Racemase, Seryl-T-RNA-Synthetase). Dabei fällt auf, daß es in den vereinigten Proteinbeschreibungen dieser Ensembles neben mehreren mittelgroßen immer eine besonders große Komponente mit mehr als zwanzig Knoten gibt. Diese größte Zusammenhangskomponente bestimmt wesentlich die Gesamtlaufzeit des Docking-Algorithmus. Denn für den Aufwand bei der Suche nach unabhängigen Mengen ist nicht die Gesamtzahl der Instanzen im erweiterten aktiven Zentrum entscheidend, sondern die Zahl der Knoten in den einzelnen Zusammenhangskomponenten. Während sich in den zwei- und dreielementigen Komponenten eine optimale unabhängige Menge sehr schnell durch direkten Vergleich der Instanzenergien bestimmen läßt, werden die Zusammenhangskomponenten, die mehr als drei Knoten enthalten, mit der in Abschnitt 5.4.2 beschriebenen Tiefensuche auf Basis des Bron-Kerbosch-Algorithmus [218] durchsucht. Die lange Laufzeit beim Docking in die Seryl-T-RNA-Synthetase, dessen Ensemble nur aus sechs Strukturen besteht, läßt sich daher vor allem mit der extrem großen Zusammenhangskomponente aus 109 Instanzen erklären.

7.4 Clustern von Wechselwirkungspunkten

In Abschnitt 5.3 wurde ein Verfahren zur Reduktion der Hashtabelle vorgestellt, das die Wechselwirkungspunkte clustert. Um den optimalen Schwellwert für das Clustern experimentell zu ermitteln, wurde dieser Parameter in 0.2 Å Schritten bis zum eineinhalbfachen der Bucketgröße der Hashtabelle erhöht. Tabelle 7.8 gibt für die verschiedenen Werte die, über alle zehn Testensembles gemittelte, Anzahl von Wechselwirkungspunkten und Punktpaaren, die durchschnittlichen Laufzeiten sowie die Zahl der Lösungen mit RMS-Abweichungen unter 2.0 Å für die verschiedenen Ränge wieder.

Zunächst fällt auf, das die Zahl der in die Hashtabelle eingetragenen Punktpaare beim Cluster-Schwellwert 0.2 Å steigt, obwohl die Anzahl der Wechselwirkungspunkte durch das Clustern abnimmt. Das liegt daran, daß Paare von Cluster-Repräsentanten unabhängig von ihrer Kompatibilität in die Hashtabelle aufgenommen werden. Denn die Kompatibilität der Punkte wird erst bei der Rückgabe der Punktpaare überprüft (vgl. Abs. 5.3).

Generell sinkt die Zahl der Wechselwirkungspunkte und Punktpaare mit steigendem Schwellwert. Dem entsprechend wird der Zeitbedarf für den Aufbau der Hashtabelle insgesamt etwa um den Faktor 30 geringer, und die Basisplazierung wird circa doppelt so schnell, weil sich die Anfragezeit an die Hashtabelle verringert. Die Laufzeit für das Preprozessing und den inkrementellen Aufbau der Liganden, ändern sich dagegen nicht, weil diese beiden Schritte unabhängig vom Clustern der Wechselwirkungspunkte sind.

Auffällig ist, daß die Zahl der Plazierungen mit einer RMS-Abweichung unter 2.0 Å, auf dem ersten wie auf den ersten zehn Rängen bei den kleinen Schwellwerten 0.2 Å und 0.4 Å, die kleiner als die halbe Bucketgröße (0.45 Å) sind, größer ist als ohne Clustern.

Die Zahl der Lösungen mit einer RMS-Abweichung von unter 2.0 Å variiert zwar etwas für die verschiedenen Schwellwerte, in der Gesamttendenz nimmt ihre Anzahl aber

δ	Hashtabelle		Vorver. [h:m:s]		Docking [m:s]		RMS-Abweichung ≤ 2.0 Å [#] [%]					
[Å]	Pkt.	Paare	Prep.	Hasht.	Basisp.	Aufbau	Rang 1		Rang 10		bst. Lsg.	
—	6282	7 324 440	14.86	32:58.72	4:16.97	1:25.83	21	35.0	40	66.7	50	83.3
0.2	5966	8 634 541	14.80	36:04.77	5:28.19	1:25.64	23	38.3	43	71.7	50	83.3
0.4	5162	5 777 156	14.89	17:41.42	4:27.53	1:25.49	26	43.3	41	68.3	50	83.3
0.6	4217	3 654 880	14.85	8:32.63	3:02.72	1:22.30	20	33.3	35	58.3	50	83.3
0.8	3437	2 352 598	14.72	4:11.87	2:30.33	1:20.90	23	38.3	36	60.0	48	80.0
1.0	2983	1 675 424	14.68	2:29.13	2:15.90	1:22.69	21	35.0	35	58.3	48	80.0
1.2	2383	1 097 026	14.73	1:31.36	2:17.27	1:23.02	17	28.3	33	55.0	46	76.7
1.4	2079	816 535	15.06	1:00.70	2:09.10	1:23.40	17	28.3	31	51.7	46	76.7

Tabelle 7.8: FLEXE: Clustern von Wechselwirkungspunkten. Die Tabelle zeigt für verschiedene Schwellwerte (δ) die durchschnittliche Anzahl der Wechselwirkungspunkte und die daraus resultierende Zahl von Punktpaaren, die im Mittel in der Hashtatabelle abgelegt werden. Dem gegenüber gestellt sind die durchschnittlichen Laufzeiten für einen Liganden für die verschiedenen Phasen des Algorithmus: die Vorverarbeitung bestehend aus Preprozessing (Einlesen u. Überlagern der Strukturen etc.) und dem Aufbau der Hashtabelle sowie dem eigentlichen Docking, das die Basisplazierung und den inkrementellen Aufbau des Liganden umfaßt. Außerdem sind die Anzahl der Lösungen mit einer RMS-Abweichung unter 2.0 Å in absoluten Zahlen (#) und in Prozent (%) jeweils für den ersten Rang, die ersten zehn Ränge und der besten Lösung von allen Ränge angegeben. Die Zeiten wurden auf einer Sun Ultra 10 (Ultra SPARC 2e, 440 MHz, 512 MB) gemessen.

mit größeren Schwellwerten bei allen Rang-Intervallen ab. Denn bei zu großen Cluster-Schwellwerten geben die Abstände der Cluster-Repräsentanten die Distanzen zwischen den Wechselwirkungspunkten falsch wieder, so daß die Paare nicht richtig in der Hashtabelle abgelegt werden. Ist der Schwellwert größer als 0.6 Å werden deshalb auch insgesamt auf allen Rängen weniger Lösungen gefunden.

Ein Cluster-Schwellwert von 0.6 Å wird als optimal betrachtet, weil er die Laufzeit der Vorverarbeitung etwa um den Faktor vier und die Basisplazierung immerhin um rund 25% beschleunigt, ohne dabei Lösungen zu verlieren, wenn man alle Ränge zugrunde legt. Erlaubt man den Verlust einiger Vorhersagen, so kann der Schwellwert bis etwa auf die Bucketgröße (0.9 Å) erhöht werden, den bei Cluster-Schranken von 0.8 Å bzw. 1.0 Å werden insgesamt nur zwei Lösungen nicht gefunden und auf dem ersten Rang finden sich mit 23 bzw. 21 Lösungen vergleichbar viele Plazierungen mit einer RMS-Abweichung unter 2.0 Å, wie im Fall ohne Clustern. Wesentlich größer als die Bucketgröße sollte der Schwellwert aber nicht gewählt werden, da ab 1.2 Å in allen betrachteten Rang-Intervallen mehr als vier Lösungen (6.6%) verloren gehen.

7.5 Vergleich von FLEXE und FLEXX

Um die Resultate von FLEXX mit den dargestellten Ergebnissen, die FLEXE erzielt, vergleichen zu können, wird folgendes Docking-Protokoll benutzt: Zunächst werden alle Liganden aller Ensembles mit FLEXX sequentiell in alle Proteinstrukturen der entsprechenden Ensembles gedockt (Kreuz-Docking) und die RMS-Abweichungen sowie die Laufzeiten bestimmt. Dabei verwendet man für beide Programme dieselbe Größe des aktiven Zentrums und die gleiche Definition der Referenzstrukturen. Anschließend werden für alle Liganden

die Lösungen, die FlexX für die einzelnen Ensemblestrukturen generiert hat, zu einer Lösungsmenge zusammengefaßt und anhand ihres Scores neu sortiert.

Da die Vorverarbeitungszeit für einen Liganden erheblich kürzer als für eine Proteinstruktur ist, werden in der praktischen Durchführung dieses Protokolls alle Proteinstrukturen nur einmal vorbereitet und die Liganden mehrfach eingelesen. Die Proteinoberflächen werden für FlexX vorberechnet, weil FlexE völlig ohne Oberflächen arbeitet.

Aus den 106 Proteinstrukturen und den 60 Liganden des Testdatensatzes ergeben sich auf diese Weise insgesamt 727 einzelne Komplexe mit separaten Plazierungen, die man dann zu 60 Lösungsmengen für die einzelnen Liganden zusammenfaßt.

7.5.1 Plazierungen

Tabelle 7.9 stellt die einzelnen Ergebnisse der jeweils zusammengefaßten Lösungsmengen des Kreuz-Dockings-Experiments mit FlexX dar. Sie zeigt wiederum für jeden Liganden die Anzahl der vorhergesagten Plazierungen, die RMS-Abweichung der ersten Lösung, der besten Lösung unter den ersten zehn Vorhersagen und der besten Lösung von allen Plazierungen unabhängig vom Rang. Außerdem sind der Rang der besten Lösung und der Rang der ersten Lösung mit einer RMS-Abweichung von weniger als 2.0 Å bzw. 2.5 Å angegeben. Auch hier gilt wieder die Plazierung mit der geringsten Abweichung zur Referenzstruktur als die *beste* Lösung, während die *erste* Lösung diejenige ist, die sich auf dem ersten Rang befindet.

Die Statistiken über die jeweilige Zahl der Lösungen mit einer RMS-Abweichung bis zu 1.0 Å, 1.5 Å, 2.0 Å bzw. 2.5 Å zeigt die Tabelle 7.10A. Der Vollständigkeit halber findet man in Tabelle 7.10B auch die entsprechenden Statistiken für den Datensatz, der durch Entfernen der identischen Liganden bereinigt wurde. Auf diese zweite Statistik, die die gleichen Effekte wie bei FlexE zeigt, wird hier aber nicht weiter eingegangen, weil für einen Vergleich zwischen zwei Docking-Programmen nur entscheidend ist, daß er auf derselben Testmenge beruht.

FlexX findet für die Hälfte aller Liganden auf dem ersten Rang eine akzeptable Plazierung mit einer RMS-Abweichung von bis zu 2.0 Å. Diese Zahl erhöht sich, wenn man die ersten zehn Lösungen betrachtet, auf knapp zwei Drittel, und unter den Plazierungen mit geringster Abweichung zur Referenzstruktur befinden sich zu 90% Lösungen mit einer RMS-Abweichung von weniger als 2.0 Å. Mit einem durchschnittlichen Rang von 104 sind sie allerdings wesentlich weiter hinten in der Plazierungsliste zu finden als bei FlexE, wo sie im Mittel auf Rang 25 liegen. Der Grund für den im allgemeinen schlechteren Rang der besten Lösung von FlexX besteht darin, daß man durch das Zusammenfassen der Einzelplazierungen wesentlich mehr Vorhersagen pro Ligand erhält als bei FlexE. Darum sind die besten Lösungen von FlexX oft auf Rängen zu finden, die die Zahl der von FlexE vorhergesagten Plazierungen übersteigen.

Tabelle 7.9: FlexX-Vorhersagen, zusammengefaßte Lösungen. Für jeden Liganden sind die Anzahl der vorhergesagten Plazierungen, die RMS-Abweichung der ersten Lösung, der besten Lösung unter den ersten zehn Vorhersagen und der besten Lösung von allen Plazierungen unabhängig vom Rang sowie der Rang der ersten Lösung mit einer RMS-Abweichung von weniger als 2.0 Å bzw. 2.5 Å angegeben. Die Rangangabe ist eingeklammert, wenn der Rang größer ist als die Gesamtzahl der Lösungen, die von FlexE vorhergesagt werden.

Ensemble	Ligand	Anz. Lös.	(Min.) RMSD [Å] Rang 1	Rang 10	Beste Lösung RMSD [Å]	Rang	< 2.0 Å Rang	< 2.5 Å Rang
Aldose-	SBI.1ah0	248	7.67	0.56	0.41	16	9	9
Reduktase	TOL.1ah3	619	3.25	0.72	0.72	2	2	2
	ZST. model	849	0.75	0.75	0.75	1	1	1
Alpha-	FMC.1mrh	1818	1.47	1.47	0.75	265	1	1
Momorcharin	ADN.1mrg	334	0.72	0.72	0.72	1	1	1
	ADE.1aha	340	3.24	0.67	0.49	24	3	3
	FMP.1ahb	3028	1.71	1.26	0.37	403	1	1
Carbo-	FMS.1bcd	790	1.75	0.87	0.52	(589)	1	1
anhydrase II	ETS.1cil	5628	2.65	2.45	0.94	(2032)	18	9
	EG1.1cnw	2953	10.66	10.64	2.92	(2026)	-	-
	EG2.1cnx	3865	6.47	5.85	1.29	(743)	145	145
	TRI.1cra	4734	1.82	1.77	0.66	(1868)	1	1
	BCT.1cam	1669	2.15	1.82	1.30	(243)	7	1
	ACY.1caz	2215	2.06	1.88	0.42	232	5	1
	AZM.1zsb	5501	6.00	2.95	1.15	(1413)	345	230
Carboxy-	G-Y.3cpa	3520	7.48	2.03	1.17	(280)	19	8
peptidase	GLY.4cpa	5371	2.73	2.70	1.54	(5132)	(2540)	86
	ZAF.6cpa	5799	4.97	4.74	2.48	(307)	-	(307)
	FVF.7cpa	2454	4.31	4.15	2.74	(1689)	-	-
	BZS.1cbx	4386	1.35	1.00	0.76	76	1	1
	CPM.1cps	8411	0.78	0.78	0.78	1	1	1
	LOF.2ctc	4638	2.04	0.51	0.51	313	2	1
Dihydrofolat-	DZF.1dyh	1397	2.58	2.16	1.77	(721)	57	2
Reduktase	FOL.1dyi	1361	2.22	2.12	1.79	(303)	31	1
	DDF.1dyj	4482	3.86	2.58	1.78	(614)	(614)	78
	FFO.1jol	1325	3.27	2.94	2.40	(299)	-	(219)
	FOL.1ra2	1440	2.80	2.42	1.79	(365)	94	7
	MTX.1ra3	3305	1.58	1.14	0.53	156	1	1
	MTX.3drc	2931	1.31	0.70	0.53	38	1	1
	MTX.1dhj	3310	1.12	0.68	0.57	184	1	1
	MTX.1dra	3415	1.23	0.72	0.53	25	1	1
	MTX.1drb	3307	1.29	0.88	0.64	113	1	1
	MTX.2drc	3546	1.42	0.74	0.60	83	1	1
	MTX.4dfr	3452	1.48	0.78	0.62	78	1	1
Isocitrat-	ICT.5icd	3318	1.50	0.72	0.71	26	1	1
Dehyrogenase	ICT.1ide	3350	1.46	1.07	0.83	(1400)	1	1
	OXS.1idc	3352	3.37	3.30	1.50	(895)	234	106
	ICT.1gro	3340	1.15	0.33	0.33	5	1	1
	ICT.1grp	3344	1.42	0.69	0.59	27	1	1
	ICT.8icd	3313	1.24	1.24	0.39	29	1	1
Mandelat	SAA.1mdr	1416	1.16	0.68	0.48	91	1	1
Racemase	SMN.1mdl	1765	2.31	1.08	0.50	14	5	1
	SAA.1mra	1417	1.28	0.73	0.47	104	1	1
	SAA.1dtn	1417	1.06	0.70	0.61	(794)	1	1
Ricin	FMP.1fmp	3961	3.48	2.19	1.30	(1261)	14	10
	A-G.1apg	1792	4.72	4.57	2.91	137	-	-
	AMP.1obt	4227	3.62	3.01	1.66	(1158)	(1158)	(1134)
Seryl-T-RNA-	AHX.1ses	1420	7.93	4.05	2.48	(930)	-	(930)
Synthetase	AMP.1ses	3205	5.67	5.56	1.12	(3192)	24	24
	SSA.1set	1199	2.62	2.48	1.72	(282)	118	2
	SSA.1set	1346	1.41	1.41	1.39	40	1	1
Trypsin	NAS.1ppc	6591	2.72	1.42	0.73	(635)	2	2
	TOS.1pph	3293	5.48	5.46	1.05	(389)	91	91
	AMC.1tng	6756	0.78	0.37	0.37	9	1	1
	FBA.1tnh	1437	0.45	0.43	0.20	14	1	1
	PBN.1tni	5749	2.37	2.34	0.56	308	37	1
	PEA.1tnj	5891	1.84	0.73	0.55	19	1	1
	PRA.1tnk	5930	1.92	1.31	0.63	223	1	1
	TPA.1tnl	895	1.46	0.55	0.55	10	1	1
	BEN.3ptb	3214	0.52	0.32	0.20	19	1	1

A kompletter Datensatz

	erster Rang				
rms≤	1.0	1.5	2.0	2.5	min
#	6	24	30	36	60
%	10.0	40.0	50.0	60.0	100.0

	ersten 10 Ränge				
rms≤	1.0	1.5	2.0	2.5	min
#	26	35	38	46	60
%	43.3	58.3	63.3	76.7	100.0

	unabhängig vom Rang				
rms≤	1.0	1.5	2.0	2.5	min
#	38	47	54	57	60
%	63.3	78.3	90.0	95.0	100.0
∅	110.5	50.5	103.8	60.4	544.1
σ	343.7	173.3	380.6	196.2	877.2

B bereinigter Datensatz

	erster Rang				
rms≤	1.0	1.5	2.0	2.5	min
#	6	13	18	24	45
%	13.3	28.9	40.0	53.3	100.0

	ersten 10 Ränge				
rms≤	1.0	1.5	2.0	2.5	min
#	16	22	25	31	45
%	35.6	48.9	55.6	68.9	100.0

	unabhängig vom Rang				
rms≤	1.0	1.5	2.0	2.5	min
#	25	34	39	42	45
%	55.6	75.6	86.7	93.3	100.0

Die Statistik zu den Rängen wurde nicht berechnet, weil dazu Lösungen hätten zusammengefaßt werden müssen.

Tabelle 7.10: FlexX-Ergebnisse, zusammengefaßte Lösungen. Die Tabellen fassen die Anzahl der Lösungen mit einer RMS-Abweichung unter 1.0, 1.5, 2.0, 2.5 Å bzw. mit minimaler Abweichung in absoluten Zahlen (#) und in Prozent (%) jeweils für den ersten Rang, die ersten zehn Ränge und für alle Ränge zusammen. Für den letzten Fall sind außerdem der durchschnittliche (∅) Rang der ersten Lösung unterhalb des Schwellwertes und die dazugehörige Standardabweichung (σ) angegeben. Die Tabellen unter A beziehen sich auf den gesamten Testdatensatz, während die Tabellen unter B nur den um identische Liganden bereinigten Datensatz berücksichtigen.

In drei Fällen (GLY.4cpa, DDF.1dyi, AMP.1obt) trifft dies auch für die erste Lösung mit einer RMS-Abweichung unter 2.0 Å zu. Läßt man deshalb diese drei Lösungen beim Vergleich mit FlexE unberücksichtigt, so gibt es bei FlexX mit 51 Lösungen in etwa genauso viele Vorhersagen mit RMS-Abweichung bis 2.0 Å wie bei FlexE mit 50 solcher Plazierungen. Auch auf den ersten zehn Rängen wird von beiden Programmen mit rund einem Drittel die gleiche Zahl von akzeptablen Plazierungen gefunden. Nur auf dem ersten Rang unterscheidet sich die Anzahl mit 30 (50%, FlexX) bzw. 21 (35%, FlexE) Lösungen deutlich. Ebenso findet FlexX für die kleineren Schwellwerte von 1.0 Å und 1.5 Å in allen drei Statistiken über die verschiedenen Ränge wesentlich mehr Plazierungen als FlexE. Verantwortlich dafür sind zwei Effekte: Erstens werden ähnliche Instanzen bei der Erzeugung der vereinigten Proteinbeschreibung von FlexE geclustert, was zu einer geringfügigen Verzerrung der Proteinstruktur führt. Zweitens dockt FlexX einen Liganden in eine genau definierte Struktur, während FlexE alle Alternativen gleichzeitig verwendet. Da in verschiedenen Bereichen der Bindetasche unter Umständen Instanzen aus verschiedenen Ensemblemitgliedern am besten bewertet werden, kann auch dies die Plazierung leicht verschieben. Beide Effekte können, wie in Abschnitt 7.3.2 diskutiert, zu einer Störung der Plazierung von ca. 0.5 Å führen und erklären die beobachtete Verlagerung der Lösungen innerhalb der Statistik.

Die Zahl der Liganden, bei denen das Docking-Protokoll mit FlexX die beste Lösung auch tatsächlich auf Rang Eins findet, ist wie auch bei FlexE gering. Nur in drei Fällen (ZST.model, ADN.1mrg, CPM.1cps) belegen die besten Plazierungen mit Abweichungen von 0.72–0.78 Å den ersten Rang.

Ensemble	#	Wechselwirkungspunkte FlexX	FlexE	F	$F/\#$	Wechselwirkungspunktpaare FlexX	FlexE	F	$\sqrt{F}$
Aldose-Reduktase	4	2228	7099	3.2	0.80	681 758	7 600 514	11.1	3.3
Alpha-Momorcharin	7	1132	4278	3.8	0.54	270 633	3 430 526	12.7	3.6
Carboanhydrase II	16	1297	3712	2.9	0.18	263 588	1 810 278	6.9	2.6
Carboxypeptidase	16	1954	13338	6.8	0.43	513 536	30 665 256	59.7	7.7
Dihydrofolat-Reduktase	12	1974	5531	2.8	0.23	532 223	3 478 479	6.5	2.5
Isocitrat-Dehydrogenase	14	1077	6419	6.0	0.43	236 937	7 430 572	31.4	5.6
Mandelat-Racemase	6	933	4701	5.0	0.84	199 151	5 496 722	27.6	5.2
Ricin	9	1273	5817	4.6	0.51	290 601	5 009 379	17.2	4.1
Seryl-T-RNA-Synthetase	6	1286	5377	4.2	0.70	274 816	5 124 088	18.6	4.3
Trypsin	16	1700	6546	3.9	0.24	391 106	3 198 583	8.2	2.7
Durchschnitt	10.6	1506	6282	4.2	0.39	362 982	7 324 440	20.2	4.5

Tabelle 7.11: Vergleich der Anzahl von Wechselwirkungspunkten. Die Tabelle stellt für alle Ensembles die Anzahl der Wechselwirkungspunkte, die von FlexE und FlexX zugewiesen werden, gegenüber und gibt den Zuwachsfaktor $F = E/X$ sowie den Quotienten $F/\#$ von diesem Faktor (F) und der Zahl der Ensemblestrukturen (#) an. Außerdem sind die Anzahl der Wechselwirkungspunktpaare in der Hashtabelle, der entsprechende Faktor $F = E/X$ sowie die Wurzel $\sqrt{F}$ angegeben.

Die Vorhersagequalität unterscheidet sich für verschiedene Ensembles sowohl bei FlexE als auch bei FlexX. Gründe dafür wurden für FlexE in Abschnitt 7.3.1 diskutiert. Sie gelten in vielen Fällen auch für FlexX, denn extrem große oder kleine Liganden (Carboanhydrase, Carboxypeptidase), eine Vielzahl von ähnlich bewerteten Lösungen (Carboxypeptidase) oder Moleküle, die nicht vollständig in der Bindetasche liegen (Ricin) sind auch für FlexX problematisch. Auch die falsche Plazierung der Heterozyklen der Folsäure(derivate) in der Dihydrofolat-Reduktase (vgl. Abs. 7.3.1) tritt bei FlexX genauso wie bei FlexE auf, weil auch bei FlexX aus Gründen der Vergleichbarkeit keine Wassermoleküle beim Docking verwendet wurden. Auf zwei der Ensembles gehen die Abschnitte 7.6 und 7.7 detailierter ein.

Die Ergebnisse aller separaten Kreuz-Docking-Experimente sind im Anhang in farbcodierten Matrizen dargestellt. Darin entspricht jede Zeile einem Liganden und jede Spalte einer Proteinstruktur. Die erste Spalte enthält die Ergebnisse von FlexE, die zweite die Resultate der zusammengefaßten Lösungsmengen und die übrigen Spalten die Vorhersagen von FlexX für die einzelnen Proteinstrukturen.

7.5.2 Wechselwirkungspunkte und Hashtabellen

Tabelle 7.11 zeigt für alle Ensembles die durchschnittliche Zahl von Wechselwirkungspunkten im aktiven Zentrum, die von FlexX je Ensemblestruktur generiert werden, und vergleicht sie mit der Gesamtzahl der entsprechenden Punkte in der vereinigten Proteinbeschreibung von FlexE. Ferner sind jeweils die Anzahl von Wechselwirkungspunktpaaren aufgeführt, die die beiden Programme in den Hashtabellen ablegen.

Die Zunahme der Wechselwirkungspunkte von FlexX zu FlexE fällt sehr unterschiedlich aus. Das Verhältnis des Zuwachses zur Zahl der Proteinstrukturen im Ensemble spiegelt dabei im wesentlichen wieder, wie stark sich die Instanzen im aktiven Zentrum

Ensemble	#	Vorverarbeitungszeit [h:m:s] FlexX einzel	FlexX akkum.	FlexE Ens.	E/X Fak.
Aldose-Reduktase	4	23.15	1:32.60	22:13.79	14.4
Alpha-Momorcharin	7	8.42	58.93	7:40.05	7.8
Carboanhydrase II	16	8.42	2:14.79	3:50.90	1.7
Carboxypeptidase	16	26.01	6:56.23	3:00:05.00	26.0
Dihydrofolat-Reduktase	12	26.72	5:20.64	12:02.99	2.3
Isocitrat-Dehydrogenase	14	9.70	2:15.74	26:37.05	11.8
Mandelat-Racemase	6	7.24	43.43	21:28.29	29.7
Ricin	9	9.86	1:28.71	22:59.43	15.5
Seryl-T-RNA-Synthetase	6	14.93	1:29.59	24:30.47	16.4
Trypsin	16	16.54	4:24.67	10:47.76	2.4
Durchschnitt	10.6	15.10	2:44.53	33:13.58	12.1

Tabelle 7.12: Vergleich: Laufzeit Vorverarbeitung. Die Tabelle zeigt für jedes Ensemble die Anzahl der Strukturen (#) sowie die Laufzeiten von FlexE und FlexX, die während der Vorverarbeitung für das Preprozessing (Einlesen u. Überlagern der Strukturen etc.) und den Aufbau der Hashtabelle benötigt werden. Dabei sind für FlexX sowohl die durchschnittlichen Zeiten für eine einzelne Proteinstruktur als auch die akkumulierte Zeit für alle Strukturen eines Ensembles angegeben. Auf die akkumulierte Zeit bezieht sich der Quotient der Laufzeiten von FlexE und FlexX (E/X). Die Zeiten wurden auf einer Sun Ultra 10 (Ultra SPARC 2e, 440 MHz, 512 MB) gemessen.

der Ensemblestrukturen unterscheiden. Denn ähnliche Instanzen werden in der vereinigten Proteinbeschreibung zusammengefaßt. Dadurch fallen ihre Wechselwirkungsgeometrien zusammen und die redundanten Wechselwirkungspunkte weg. Deshalb kann man aus dieser Gegenüberstellung ablesen, daß die Ensembles der Aldose-Reduktase, der Mandelat-Racemase und der Seryl-T-RNA-Synthetase relativ stark variieren, während die Konformationen bei der Carboanhydrase und dem Trypsin recht gut konserviert sind.

Die Hashtabellen wachsen in etwa mit dem Quadrat der Wechselwirkungspunkte. Dieser Zusammenhang ist nicht exakt, weil nur die Paare von Punkten in die Hashtabellen aufgenommen werden, deren Distanz innerhalb des Intervalls von 0.5 Å und 10.0 Å liegen, außerdem bei FlexE zusätzlich auch nur diejenigen Punktpaare, die miteinander kompatibel sind.

7.5.3 Laufzeiten

Um die Proteinflexibilität mit FlexX zu berücksichtigen, werden die Liganden sequentiell in alle Ensemblestrukturen plaziert. Darum muß man die Zeit, die FlexE für das Docken eines Liganden benötigt, mit der akkumulierten Laufzeit von FlexX vergleichen, das heißt, man muß die Zeit für FlexX über alle Einzelläufe aufsummieren. Die Tabellen 7.12 und 7.13 stellen die Laufzeiten von FlexE und FlexX für die Vorverarbeitung bzw. das eigentliche Docking gegenüber. Sie enthalten für jedes Ensemble die Zeiten, die bei FlexE und FlexX im Durchschnitt für die Plazierung eines Liganden erforderlich sind. Dabei sind für FlexX neben den akkumulierten Zeiten zum Vergleich auch die Mittel der Einzellaufzeiten angegeben.

		Basisplazierung [m:s]				Komplexaufbau [m:s]			
Ensemble	#	FlexX einzel	FlexX akkum.	FlexE Ens.	E/X Fak.	FlexX einzel	FlexX akkum.	FlexE Ens.	E/X Fak.
Aldose-Reduktase	4	3:12.91	12:51.64	4:50.41	0.4	8.57	34.27	16.64	0.5
Alpha-Momorcharin	7	17.57	2:02.97	2:07.55	1.0	6.26	43.84	15.35	0.4
Carboanhydrase II	16	38.49	10:15.84	2:20.58	0.2	10.19	2:43.02	19.12	0.1
Carboxypeptidase	16	48.02	12:48.27	12:52.40	1.0	15.64	4:10.23	1:12.84	0.3
Dihydrofolat-Reduktase	12	51.85	10:22.18	1:28.47	0.1	30.54	6:06.51	52.76	0.1
Isocitrat Dehydrogenase	14	19.28	4:29.94	5:56.50	1.3	11.19	2:36.70	42.59	0.3
Mandelat-Racemase	6	53.34	5:20.03	2:28.57	0.5	3.70	22.18	38.81	1.7
Ricin	9	39.48	5:55.31	1:28.61	0.2	25.94	3:53.43	51.26	0.2
Seryl-T-RNA-Synthetase	6	25.18	2:31.06	8:06.86	3.2	33.81	3:22.89	8:58.61	2.7
Trypsin	16	27.72	7:23.50	1:09.72	0.2	5.80	1:32.86	10.36	0.1
Durchschnitt	10.6	51.38	7:24.07	4:16.97	0.6	15.16	2:36.59	1:25.83	0.5

Tabelle 7.13: Vergleich: Laufzeit Docking. Die Tabelle zeigt für jedes Ensemble die Anzahl der Strukturen (#) sowie die durchschnittlichen Laufzeiten von FlexE und FlexX für das Docking eines Liganden, das aus der Basisplazierung und dem inkrementellen Komplexaufbau besteht. Dabei sind für FlexX sowohl die durchschnittlichen Zeiten für eine einzelne Proteinstruktur als auch die akkumulierte Zeit für alle Strukturen eines Ensembles angegeben. Auf die akkumulierte Zeit bezieht sich der Quotient der Laufzeiten von FlexE und FlexX (E/X). Die Zeiten wurden auf einer Sun Ultra 10 (Ultra SPARC 2e, 440 MHz, 512 MB) gemessen.

Für die Vorverarbeitung der Liganden benutzen FlexE und FlexX dieselben Routinen. Deshalb ist die Zeit, die im Mittel für das Einlesen und Zerlegen eines Liganden benötigt wird, bei beiden Programmen identisch. Allerdings muß FlexX in dem verwendeten Docking-Protokoll die Liganden für jede Ensemblestruktur erneut einlesen. Deswegen multipliziert sich der Zeitbedarf für die Vorbereitung der Liganden bei FlexX mit der Anzahl der Liganden eines Ensembles.

Die in Tabelle 7.12 angegebenen Vorverarbeitungzeiten beziehen sich wie auch schon in Abschnitt 7.3.3 nur auf das Protein. Das Preprozessing besteht bei FlexE, wie bereits beschrieben, aus dem Einlesen der Ensemblestrukturen, dem Aufbau der vereinigten Proteinbeschreibung einschließlich des Inkompatibilitätsgraphen sowie aus dem Aufbau der Hashtabelle, der den größten Teil der Zeit beansprucht. Bei FlexX werden in dieser Phase die Proteinstrukturen und die vorberechneten Oberflächen eingelesen sowie die Hashtabelle aufgebaut, die im Vergleich zu FlexE wesentlich kleiner ist. Darum benötigt FlexX mit ein paar Minuten deutlich weniger Zeit für die Vorverarbeitung als FlexE. Daß die Zunahme der Vorbereitungszeiten dabei von Ensemble zu Ensemble differiert, liegt vor allem daran, daß die Zahl der Wechselwirkungspunkte und damit die Größe der Hashtabelle, wie im vorherigen Abschnitt gezeigt, beim Übergang von FlexX zu FlexE unterschiedlich stark wächst.

Für die Basisplazierung benötigt FlexE in zwei Fällen eine längere Laufzeit und ist in sechs Fällen wesentlich schneller als FlexX, obwohl die Hashtabellen von FlexE größer sind und außerdem für jede Plazierung die Zusammenhangskomponenten nach unabhängigen Mengen durchsucht werden müssen. Das liegt jedoch daran, daß FlexE insgesamt weniger Basisplazierungen erzeugt als FlexX.

Beim Komplexaufbau ist FlexE in 8 von 10 Fällen deutlich schneller als FlexX und

benötigt in vielen Fällen nur etwa die doppelte Laufzeit eines Einzeldockings mit FlexX. Grund dafür ist vor allem die parallele Verarbeitung der verschiedenen Alternativen. Daß FlexE in zwei Fällen langsamer ist als FlexX, kann insbesondere bei der Seryl-T-RNA-Synthetase auf das Durchsuchen der Zusammenhangskomponenten nach jedem Anbauschritt zurückgeführt werden.

7.6 Beispiel Aldose-Reduktase

Zu Beginn dieser Arbeit (Abs. 3.1) wurde das Enzym Aldose-Reduktase als ein Beispiel genannt, bei dem die Proteinflexibilität dazu führen kann, daß mögliche Liganden bei einer Screening-Anwendung übersehen werden, wenn nur eine einzige starre Proteinstruktur zugrunde gelegt wird. Diese Aussage belegt das Kreuz-Docking-Experiment mit FlexX, dessen Lösungen zu einer Ergebnisliste zusammengefaßt werden. Deshalb geht dieser Abschnitt noch einmal ausführlich auf das Beispiel der Aldose-Reduktase ein.

Das Ensemble der Aldose-Reduktase besteht aus vier Proteinstrukturen. Drei dieser Strukturen (1ah0,1ah3, 1ah4) sind Kristallstrukturen des Enzyms der Schweinelinsen, die Urzhumtsev et al. im Komplex mit den potenten Inhibitoren Sorbinil (1ah0) und Tolrestat (1ah3) sowie allein mit dem Co-Faktor (1ah4) experimentell bestimmt haben [99]. Diese Strukturen wurden mit einem Homologiemodell der menschlichen Aldose-Reduktase kombiniert, das den Liganden Zopolrestat enthält. [100]. Für diesen Komplex sind bisher nur die Koordinaten der C_α-Atome veröffentlicht (1mar, [101]). Die menschliche Aldose-Reduktase hat eine Sequenzidentität von etwa 86% zu der vom Schwein, aber das aktive Zentrum ist sehr stark konserviert.

Abbildung 7.2a zeigt das aktive Zentrum der vereinigten Proteinbeschreibung mit den Referenzstrukturen der drei Inhibitoren, aber in diesem Fall sind nur die Aminosäuren dargestellt, die direkte Wechselwirkungen zum Liganden ausbilden (vgl. Abb. 3.1, Abs. 3.1). Die stark konservierte, hydrophile Bindungsregion befindet sich zwischen den Aminosäuren TRP 20, THR 48, HIS 110 und TRP 111, die Wasserstoffbrücken zum Liganden ausbilden können, während die hydrophobe Kontaktzone im wesentlichen aus den Tryptophanen TRP 20, TRP 79, TRP 111, TRP 219 besteht. Die flexible Spezifitätstasche liegt zwischen den Residuen PHE 122 und LEU 300. Diese Tasche ist verschlossen (Konformation 2), wenn Sorbinil gebunden wird und öffnet sich auf zwei verschiedene Weisen, um sich Tolrestat (Konformation 1) bzw. Zopolrestat (Konformation 3) anzupassen.

Abbildung 7.3 stellt die RMS-Abweichungen der jeweils besten Lösung dar, die von FlexE bzw. FlexX auf einem beliebigen Rang vorhergesagt wurden. Daher ist ausgeschlossen, daß es eine Lösung mit einer geringeren Abweichung auf einem niedrigeren Rang gibt. Da die Spezifitätstasche nicht an der Bindung von Sorbinil beteiligt ist, kann FlexX diesen Liganden in alle Ensemblestrukturen mit einer RMS-Abweichung von deutlich unter 1.0 Å plazieren. Der Originalkomplex kann sogar mit einer Abweichung von nur 0.43 Å reproduziert werden. Aufgrund der unterschiedlichen Konformationen der Spezifitätstasche bei der Bindung von Tolrestat und Zopolrestat kann FlexX diese beiden Liganden nur in die Strukturen korrekt docken, aus denen sie extrahiert wurden. Die besten Lösungen in fremden Strukturen haben Abweichungen von mehr als 3.5 Å beim Tolrestat und 6.0 Å beim Zopolrestat. Das macht deutlich, daß man möglicherweise geeignete Inhibitoren nicht gefunden hätte, wenn man nur eine Ensemblestruktur für ein Screening verwendet hätte, weil einige Liganden nur in eine andere Konformation der Aldose-Reduktase passen. Dieses

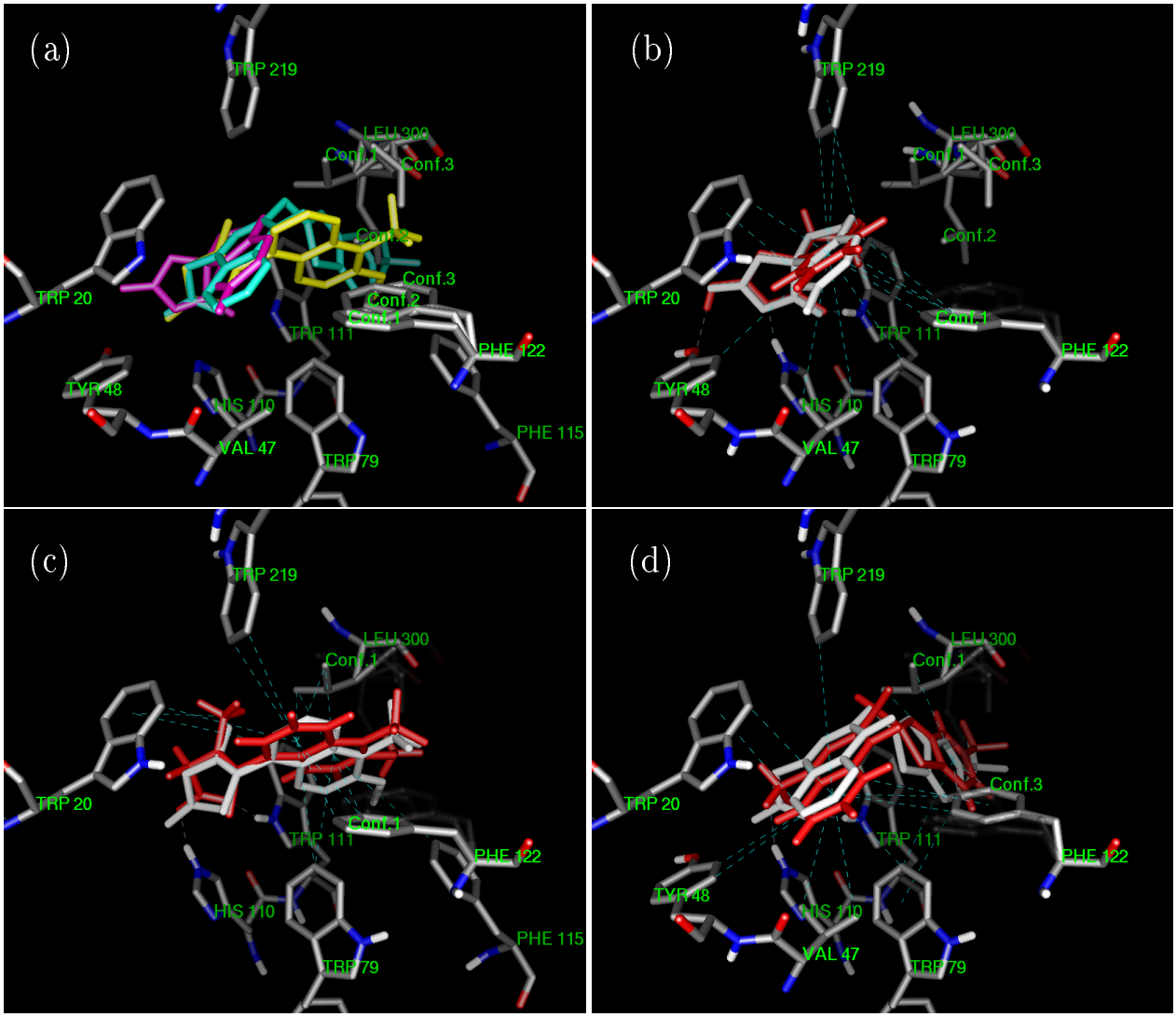


Abbildung 7.2: Beste Vorhersagen Aldose-Reduktase. Die Abbildungen zeigen die vereinigte Proteinbeschreibung der Aldose-Reduktase, die (a) die Referenzstrukturen der drei Inhibitoren Sorbinil (violett), Tolrestat (gelb) und Zopolrestat (türkis) sowie (b-c) jeweils die besten Vorhersagen (rot) zusammen mit diesen Referenzstrukturen (grau) für Sorbinil (b), Tolrestat (c) und Zopolrestat (d) enthält. Es sind nur die Aminosäuren dargestellt, die Wechselwirkungen mit den Liganden ausbilden, nicht verwendete alternative Instanzen sind ausgeblendet.

Beispiel zeigt die Notwendigkeit, mehrere verschiedene Proteinstrukturen beim Docking zu berücksichtigen.

Die vereinigte Proteinbeschreibung von FlexE wie auch das Zusammenfassen der Lösungen beim Kreuz-Docking mit FlexX sind verschiedene Arten, solche Variationen der Proteinstruktur zu berücksichtigen. Beide Ansätze wählen geeignete Konformationen aus und finden für alle drei Inhibitoren gute Plazierungen, die nur um rund 1.0 Å oder weniger von der Lage im Kristall abweichen.

Die besten Lösungen, die von FlexE vorhergesagt wurden, sind in den Abbildungen 7.2b–c zusammen mit dem jeweiligen Referenzliganden dargestellt. Auf diesen Abbildun-

	ENS.	MRGD	FlexX 1ah0	FlexX 1ah3	FlexX 1ah4	FlexX model
SBI.1ah0	0.54	0.41	0.43	0.92	0.41	0.56
TOL.1ah3	1.05	0.72	3.71	0.72	3.82	4.16
ZST.model	0.64	0.75	6.14	6.41	6.84	0.75

Abbildung 7.3: Kreuz-Docking der Aldose-Reduktase. Die farbcodierte Tabelle zeigt die RMS-Abweichung in Å der besten Vorhersage unabhängig vom Rang der Lösung. Jede Zeile enthält die Ergebnisse für einen Liganden, die von FlexE (FlexE, 1. Spalte), den zusammengefaßten Lösungsmengen (MRGD, 2. Spalte) und von FlexX (übrige Spalten) für die einzelnen Proteinstrukturen vorhergesagt wurden.

gen sind jeweils nur die Instanzen gezeigt, die Wechselwirkungen mit dem Liganden ausbilden. Dazu alternative Konformationen, die für diese speziellen Lösungen nicht verwendet werden, sind ausgeblendet. Wenn sich FlexE bei einem Segment nicht auf eine Instanz festlegen mußte, sind alle Konformationen dargestellt (z.B. LEU 300 in Abb. 7.2b).

Sorbinil (Abb.7.2b), das aus drei verbundenen Ringen besteht, ist weitgehend starr. Die Wasserstoffbrücken zu THR 48, HIS 110 und TRP 111 wurden gefunden und der aromatische Ring liegt in der hydrophoben Region. Eine gute Lösung mit 0.58 Å findet sich auf Rang 1 und die beste Lösung mit 0.54 Å auf Rang 2. Bei FlexX findet man die beste Lösung auf Rang 16 der vereinigten Vorhersageliste.

Im Fall des Tolrestat (Abb.7.2c) werden zwei Wasserstoffbrücken zwischen seinen Carboxylatgruppen und den Aminosäuren HIS 110 und TRP 111 gefunden, während es zwischen dem Naphtylring und TRP 20, TRP 79 TRP 111 TRP 219, PHE 115 PHE 122 und LEU 300 hydrophobe Kontakte gibt. FlexE benutzt sowohl für PHE 122 als auch für LEU 300 jeweils die Konformation 1, die im Originalkomplex vorliegen, so daß die Trifluormethylgruppe und die Methoxygruppe des Inhibitors in Übereinstimmung mit der experimentell bestimmten Struktur in den oberen Teil der Spezifitätstasche plaziert werden. Wiederum findet sich eine gute Lösung (1.09 Å) auf dem ersten Rang und eine etwas bessere Lösung (1.05 Å) auf Rang fünf. Die beste Lösung der zusammengefaßten Ergebnisliste von FlexX befindet sich auf dem zweiten Rang.

Wie auch das Tolrestat, bildet das Zopolrestat (Abb.7.2d) zwei Wasserstoffbrücken zwischen seinen Carboxylatgruppen und HIS 110 bzw. TRP 111 aus. Der Phthalazionring liegt in der konservierten hydrophoben Zone und der Benzothiazolring füllt den tieferen Teil der Spezifitätstasche, die von LEU 300 geöffnet wird. Für diese Seitenkette verwendet FlexE die Konformation, die aus dem Komplex mit Tolrestat stammt (Konf. 1) anstelle der Konformation drei. Deshalb gibt es eine größere Überlappung zwischen dem Stickstoff des Benzothiazolrings und LEU 300, die jedoch von FlexE toleriert wird. Beim PHE 122 verwendet FlexE die Konformation drei, deren Orientierung aus dem modellierten Komplex zwischen Zopolrestat und Aldose-Reduktase stammt. Somit kann PHE 122 hydrophobe Wechselwirkungen sowohl zum Benzothiazolring als auch zum Phthalazionring des Zopolrestats ausbilden. Die besten Lösungen befinden sich auf Rang 133 bei FlexE und auf dem ersten Rang bei FlexX. Der Grund dafür liegt vermutlich bei der leicht unterschiedlichen Bewertung des Protein-Ligand-Überlappungsvolumens durch FlexE bzw. FlexX, die sich bei dieser relativ engen Tasche deutlich bemerkbar macht.

		FlexX												
	ENS.	MRGD	1dyh	1dyi	1dyj	1jol	1ra2	1ra3	3drc	1dhj	1dra	1drb	2drc	4dfr
DZF.1dyh	2.00	2.16	1.97	1.94	2.00	1.91	1.93	2.07	2.07	3.11	2.03	3.72	2.15	2.16
FOL.1dyi	1.84	2.12	1.94	2.10	2.10	1.94	1.87	2.04	2.08	3.93	1.95	3.47	1.94	2.12
DDF.1dyj	1.63	2.59	3.54	2.78	2.59	4.56	2.81	3.23	2.89	2.83	3.01	4.13	3.06	2.80
FFO.1jol	5.37	2.94	3.25	2.90	2.67	5.36	3.06	4.41	4.37	4.16	2.77	4.97	4.54	3.68
FOL.1ra2	1.91	2.41	2.01	2.09	2.22	2.02	1.95	2.06	2.04	4.26	1.92	7.55	2.12	2.63
MTX.1ra3	0.81	1.19	1.38	1.37	1.39	3.85	1.19	0.86	1.11	2.70	1.14	2.73	1.11	1.14
MTX.3drc	0.50	0.70	0.89	0.87	1.32	1.60	0.93	0.72	0.63	2.60	0.70	2.60	0.70	0.77
MTX.1dhj	0.67	0.68	0.91	0.88	1.20	1.24	0.96	0.72	0.66	2.59	0.59	2.58	0.68	0.82
MTX.1dra	0.70	0.72	0.97	0.86	1.17	3.06	1.05	0.81	0.66	2.60	0.57	2.60	0.71	0.53
MTX.1drb	0.98	0.88	1.06	0.99	1.01	1.45	0.94	0.75	0.87	2.55	0.69	2.59	0.68	0.87
MTX.2drc	0.97	0.74	1.13	1.16	0.95	3.50	0.97	0.73	1.09	2.64	0.92	3.42	0.74	0.81
MTX.4dfr	0.66	0.78	0.98	1.19	1.20	3.67	1.03	0.76	0.65	2.56	0.78	4.03	0.78	0.84

Abbildung 7.4: Kreuz-Docking der Dihydrofolat-Reduktase. Die farbcodierte Tabelle zeigt die RMS-Abweichung in Å der besten Vorhersage unter den ersten zehn Lösungen. Jede Zeile enthält die Ergebnisse für einen Liganden, die von FlexE (ENS., 1. Spalte), den zusammengefaßten Lösungsmengen (MRGD, 2. Spalte) und von FlexX (übrige Spalten) für die einzelnen Proteinstrukturen vorhergesagt wurden.

7.7 Verkleinertes Ensemble für Dihydrofolat-Reduktase

Das Ensemble für die Dihydrofolat-Reduktase besteht aus zwölf Kristallstrukturen, die alle im Komplex mit kleineren Molekülen experimentell bestimmt worden sind. Alle diese Liganden sind im Testdatensatz enthalten. Ihre Bindungsmodi lassen sich, abgesehen von dem bereits beschriebenen Problem bei der Plazierung der Heteroringe (vgl. Abs. 7.3.1 u. 7.5.1), sowohl mit FlexE in der vereinigten Proteinbeschreibung als auch mit dem Kreuz-Docking-Protokoll für FlexX vorhersagen. Allerdings enthält das Ensemble für alle Liganden die Originalstruktur.

Abbildung 7.4 zeigt die farbcodierte Matrix, die für die ersten zehn Ränge die besten Ergebnisse für FlexE und für die zusammengefaßten Lösungen vom FlexX sowie aller einzelnen Kreuz-Docking-Experimente wiedergibt. Die entsprechenden Matrizen für den

	complete		reduced		ensemble FlexX		
	ENS.	MRGD	ENS	MRGD	1dhj	1drb	1jol
DZF.1dyh	2.00	2.16	2.01	2.09	3.11	3.72	1.91
FOL.1dyi	1.84	2.12	2.07	1.94	3.93	3.47	1.94
DDF.1dyj	1.63	2.59	3.14	2.83	2.83	4.13	4.56
FFO.1jol	5.37	2.94	3.95	5.36	4.16	4.97	5.36
FOL.1ra2	1.91	2.41	2.10	2.02	4.26	7.55	2.02
MTX.1ra3	0.81	1.19	1.06	2.70	2.70	2.73	3.85
MTX.3drc	0.50	0.70	0.80	2.60	2.60	2.60	1.60
MTX.1dhj	0.67	0.68	0.92	1.24	2.59	2.58	1.24
MTX.1dra	0.70	0.72	0.74	2.60	2.60	2.60	3.06
MTX.1drb	0.98	0.88	0.96	2.56	2.55	2.59	1.45
MTX.2drc	0.97	0.74	0.80	2.64	2.64	3.42	3.50
MTX.4dfr	0.66	0.78	0.90	2.56	2.56	4.03	3.67

Abbildung 7.5: Kreuz-Docking der Dihydrofolat-Reduktase, verkleinertes Ensemble. Die farbcodierte Tabelle stellt die Docking-Ergebnisse für das vollständige und das verkleinerte Ensemble aus drei Proteinstrukturen der Dihydrofolat-Reduktase gegenüber. Angegeben sind die RMS-Abweichung in Å der besten Vorhersage unter den ersten zehn Lösungen. Jede Zeile enthält die Resultate für einen Liganden. Die ersten beiden Spalten geben die Vorhersagen von FlexE (ENS.) und dem Merging-Ansatz (MRGD) mit FlexX für das vollständige Ensemble wieder. Die folgenden beiden Spalten zeigen die entsprechenden Ergebnisse für das verkleinerte Ensemble und die übrigen drei Spalten stellen die Plazierungen für die einzelnen Proteinstrukturen dar.

ersten Rang und die jeweils besten Lösungen unabhängig vom Rang sehen ähnlich aus (s. Anhang). Insbesondere für die Methotrexat-Moleküle aus den verschiedenen PDB-Einträgen werden von beiden Programmen Plazierungen mit geringer Abweichung zur Referenzstruktur vorhergesagt.

Beim separaten Docking in die einzelnen Ensemblemitglieder mit FlexX zeigt sich, daß drei Proteinstrukturen (1dhj, 1drb, 1jol) zu signifikant schlechteren Resultaten führen als alle anderen Strukturen. Während bei 1jol offenbar nur eine ungünstige Proteinkonformation vorliegt, gibt es bei den beiden anderen Strukturen an der Position 27 jeweils eine Mutation von Aspartat nach Serin bzw. Cystein, die sich negativ auf die Vorhersage auswirken.

Um zu analysieren, ob FlexE immer noch gute Vorhersagen macht, wenn nicht für alle Liganden die Originalstruktur im Ensemble enthalten ist, wurde das Ensemble für Dihydrofolat-Reduktase verkleinert, nur die drei PDB-Strukturen 1dhj, 1dbr und 1jol ausgewählt und der Vergleich zwischen FlexE und FlexX wiederholt. Die Ergebnisse sind in Abbildung 7.5 den ursprünglichen Ergebnissen gegenübergestellt.

Die Vorhersagequalität von FlexE mit dem verkleinerten Ensemble ist etwas schlechter als mit dem vollständigen Ensemble aus Dihydrofolat-Reduktase-Strukturen, wohingegen die Vorhersage von FlexX insbesondere für die Methotrexat-Moleküle deutlich schlechter wird. In vier Fällen (MTX.1ra3, MTX.1dra, MTX.2drc, MTX.4dfr) findet FlexE

auf den ersten zehn Rängen Plazierungen mit einer RMS-Abweichung von rund 1.0 Å, während alle Vorhersagen von FlexX um mehr als 2.5 Å von der Referenz abweichen.

Von FlexX werden dagegen bei zwei Liganden (MTX.3drc, MTX.1drb) die guten Plazierungen in einzelnen Strukturen so schlecht bewertet, daß sie sich nach dem Zusammenfassen der Lösungen nicht mehr unter den ersten zehn Vorhersagen befinden.

Dieses Experiment zeigt, daß FlexE die verschiedenen Ensemblestrukturen, die alle für sich problematisch sind, zu neuen Proteinstrukturen kombinieren kann, die besser für das Docking eines speziellen Liganden geeignet sind, wohingegen das nachträgliche Zusammenfassen der Lösungen bei FlexX dazu nicht in der Lage ist. FlexE kann bereits bei der Plazierung des Liganden durch Rekombination lokale Probleme ausgleichen, während dies bei einzelnen starren Strukturen nicht möglich ist. Damit geht der Konformationsraum, der von FlexE berücksichtigt wird, wesentlich über die Proteinkonformationen hinaus, die von den einzelnen Ensemblestrukturen repräsentiert werden.

7.8 Vergleich von FlexE mit anderen Ansätzen

Ein Vergleich verschiedener Docking-Programme anhand der Literatur ist schon deshalb schwierig, weil die Veröffentlichungen auf unterschiedlichen Testdatensätzen basieren. Dennoch lassen sich aufgrund eines solchen Vergleichs einige grundsätzliche Aussagen über die Leistungsfähigkeit von FlexE machen. Anknüpfend an Abschnitt 3.3 nimmt deshalb Tabelle 7.14 die Tabelle 3.1 (Abs. 3.3.6) wieder auf und ordnet FlexE darin ein.

FlexE ist ein diskreter Ansatz für das Docking-Problem, der im Rahmen dieser Arbeit anhand von zehn Ensembles und 60 Liganden getestet wurde. Dieser Datensatz ist im Vergleich zu den Testfällen, an denen die anderen Methoden validiert wurden, relativ groß und vermeidet dadurch eine zu starke Fokussierung auf einzelne Spezialfälle.

Für 80% der Liganden des Testdatensatzes findet FlexE eine Lösung mit einer RMS-Abweichung von weniger als 2.0 Å. Damit liefert FlexE ebenso gute Resultate wie das Zusammenfassen und Neu-Sortieren der Lösungen von FlexX, wenn die Original-Proteinstruktur Bestandteil des Ensembles ist. Diese Vorhersagequalität ist auch mit der Plazierungsgenauigkeit anderer Methoden vergleichbar, die in Tabelle 7.14 angegeben sind. Allerdings sind viele dieser Verfahren nur mit einer geringen Zahl von Liganden getestet worden.

Neu an FlexE ist, daß es, anders als beim Zusammenfassen der Lösungen von FlexX möglich, die Rekombination der gegebenen Ensemblestrukturen erlaubt. Deshalb kann FlexE, wie das Beispiel des verkleinerten Ensembles für die Dihydrofolat-Reduktase belegt, auch noch gute Vorhersagen machen, wenn ein Ligand in keine der vorgegebenen Proteinstrukturen selbst paßt. FlexE ist also in der Lage, für die Plazierung eines Liganden geeignete Konformere des Proteins beim Docking zu einer neuen Proteinkonformation zu kombinieren und dadurch die Flexibilität des Proteins effektiv zu modellieren.

Einige Ansätze aus der Literatur berücksichtigen nur eine sehr eingeschränkte Proteinflexibilität (Jones [32], Sobolev [120], Sandak [152]) oder benutzen starre Liganden (Sobolev [120], Sandak [152], Knegtel [122]) und zum Teil wird die Komplexität des Problems durch eine Vorplazierung der Liganden eingeschränkt (Desmet [125], Leach [124]). Bei FlexE sind diese Einschränkungen nicht erforderlich. Die vorgestellte Modellierung berücksichtigt sowohl die Flexibilität der Liganden (inkrementeller Aufbau) als auch die der Proteine (Rekombination der Strukturen), ohne daß eine Vorplazierung nötig ist.

Das Modell der Proteinflexibilität besteht darin, aus einer Menge gegebener diskreter Konformationen für die einzelnen Aminosäuren diejenige auszuwählen, die am besten zu dem speziellen Liganden paßt. Dabei ist FlexE nicht auf experimentelle Daten beschränkt. Insofern ist dieser Ansatz mit Methoden, die Rotamer-Bibliotheken benutzen (Desmet [125], Leach [124], Schnecke [121]) vergleichbar. Bei diesen Verfahren werden Seitenkettenkonformationen aus der Menge von Rotameren ausgewählt, die man typischerweise in Proteinstrukturen findet. Nur die Simulations-Verfahren (Totrov [126], Apostolakis [52], Luty [46], Mangoni [128]), die die Konformationen von Ligand und Protein dynamisch anpassen, sind in der Lage völlig neue Konformationen vorherzusagen, wobei diese Simulationen allerdings für die allermeisten Docking-Anwendungen aus Gründen der Praktikabilität wegen langer Rechenzeiten ausscheiden.

Die Modellierung der Proteinflexibilität bei FlexE stellt also prinzipiell keine stärkere Einschränkung dar als die Rotamer-Ansätze, die zudem in der Regel auf einem starrem Backbone basieren (Desmet [125], Leach [124]). Im Gegenteil: FlexE bietet die Möglichkeit, parallel zu unterschiedlichen Seitenketten-Konformationen auch verschiedene Loop-Konformationen bei der Plazierung von flexiblen Liganden zu handhaben.

Die Zahl der Kombinationsmöglichkeiten bei der Rekombination der Ensemblestrukturen in FlexE ist zwar bis jetzt nicht so groß wie bei Rotamer-Bibliotheken, dafür sind die vorgegebenen Konformationen aber unter Umständen besser an das Docking-Problem angepaßt, weil nur Alternativen betrachtet werden, die man bei dem speziellen Protein beobachtet hat, die sich in einer Moleküldynamik-Simulation verändern oder deren Position in einem Homologiemodell unklar ist. Eine Erweiterung des Ensemble-Ansatzes auf die Verwendung von Rotamer-Bibliotheken ist vom Konzept her leicht möglich, so daß man die Vorteile der großen Auswahl von Rotamer-Ansätzen und die Möglichkeit des Ensemble-Ansatzes, Eigenschaften des speziellen Docking-Problems einfließen lassen zu können sowie Loop-Flexibilität zu modellieren, in FlexE kombinieren kann.

Eine ganz andere Art die Proteinflexibilität zu modellieren, stellen die Verfahren dar, die über eine Menge von Proteinstrukturen mitteln und daraus eine konstante Proteinbeschreibung generieren, die während des Dockings nicht mehr an verschiedene Liganden angepaßt wird (Knegtel [122], Broughten [123]). Auch diese Modelle basieren auf einer Menge bekannter Konformationen und können deshalb nicht extrapolieren. Diese Verfahren sind mit einer Laufzeit von 1–2 Minuten relativ schnell, weil sie die kombinatorische Natur des Problems umgehen. Denn es werden immer alle Alternativen gleichzeitig betrachtet, weil die Potentiale der einzelnen Strukturen durch Mittelung zusammengefaßt werden. Dadurch entsteht jedoch ein völlig unrealistisches Bild der Bindetasche, denn entweder ist die Tasche zu eng, wenn alle repulsiven Terme berücksichtigt werden, oder aber zu groß, wenn man diese Terme unberücksichtigt läßt. Auch in den Potentialen der übrigen Wechselwirkungen können Maxima an Positionen auftreten, die eigentlich nicht gleichzeitig vorhanden sein oder von keiner real möglichen Proteinkonformation erreicht werden können. Allerdings werden Liganden mit hoher Bindungsaffinität gerade solche Bindungsmodi bevorzugen, die mit Potentialmaxima verschiedener Proteinkonformationen korrespondieren.

Die von FlexE vorhergesagten Plazierungen beruhen dagegen immer auf gültigen Proteinstrukturen, weil durch das Konzept der Kompatibilität sichergestellt ist, daß keine Instanzen kombiniert werden, die nicht gleichzeitig in einer Proteinkonformation auftreten können. Dabei wird die kombinatorische Natur des Problems berücksichtigt. Trotzdem

benötigt FlexE im Durchschnitt nur eine Rechenzeit von 5.5 Minuten zum Docken eines Liganden.

Die Geschwindigkeit einer Methode bestimmt seine Einsatzmöglichkeit beim Screening von Datenbanken. Aufgrund ihrer langen Rechenzeiten von mehreren Stunden bis Tagen für das Plazieren eines Liganden, eignen sich die Simulationsverfahren für diese Anwendung nicht. Ebenso kommen die diskreten Docking-Algorithmen, die den Konformationsraum vollständig durchsuchen (Desmet [125], Leach [124]) mit Laufzeiten in derselben Größenordnung für das Screening nicht in Frage.

Die schnellsten diskreten Docking-Methoden benötigen dagegen nur wenige Sekunden bis Minuten, um ein Molekül zu docken. Sie arbeiten mit starren Liganden (Sobolev [120], Knegtel [122]) und/oder mit konstanten Proteinmodellen (Knegtel [122], Broughten [123]). Das Verfahren von Schnecke [121] ist im Mittel vor allem deswegen so schnell, weil es mit effizienten Vorfiltern eine Vielzahl von Molekülen aussortieren kann, so daß nur ein Teil der Liganden gedockt werden muß. Das Docking basiert dabei im wesentlichen auf einem Matching zwischen einer starren Proteinbeschreibung und der gegebenen Ligandkonformation. Nur wenn diese wenigstens zum Teil zusammenpassen, wird in einer Nachoptimierung versucht, Kollisionen zwischen Protein und Ligand flexibel aufzulösen.

FlexE berücksichtigt im Gegensatz dazu die Flexibilität von Ligand und Protein bereits bei der Plazierung und eignet sich mit einer Laufzeit von durchschnittlich 5.5 Minuten zum Durchmustern mittelgroßer Datenbanken von bis zu 50 000 Molekülen. Durch effektive Vorfilter ließe sich auch bei FlexE der mögliche Durchsatz von Liganden beim Screening weiter erhöhen.

Erstautor (für Vergleich)	Ref.	# Testsys. Lig.	Pro.	Flexibilität Lig.	Protein	RMSD Lig. [Å]	Laufzeit pro Ligand	Hardware (Prozessor)
Sobolev	[120]	4	4	starr	Sk. ignor.	0.4–1.4	3 min	DEC Turbo Laser
Desmet	[125]	4	4	flex.[1]	Rotamere	0.6–1.9	k.A.	SGI Indy (R4600PC)
Leach	[124]	2	2	flex.[2]	Rotamere	0.7–2.5	1 u. 9 Tage	SGI Indigo (R3000)
Sandak	[152]	3	3	starr	2 Gelenke	3.5–7.2[3]	2.5–39 min	SGI (R10000)
Knegtel	[122]	15	4 E	starr	gew. Mittel	0.4–1.6	1.5 min	k.A.
Broughton	[123]	∼7000	2 E	Ens.	gew. Mittel	k.A.[4]	2.5-17 s[6]	Cray J-90 (16 CPU's)
Schnecke	[121]	175000	3	Aufl. v.	Kollisionen	k.A.[5]	3–350 ms[7]	Pentium II (450MHz)
Claußen	**[231]**	**60**	**10 E**	**flex.**	**rkb. Ens.**	**83%<2.0**	**5.5 min**	**Sun Ultra 10**
Jones	[32]	100	100	flex.	endst. Grp.	66%<2.0	20x 12 min	SGI Indigo II (R4400)
Totrov	[126]	5(8)[8]	5(8)[8]	flex.	flex. Sk.[9]	1.8–7.8[8]	5–15 h	k.A.
Apostolakis	[52]	3	3	flex.	flex.	0.9–1.4	5 Tage	SGI Chall. (R4400)
Luty	[46]	1	1	flex.	flex.	∼1.0	∼1 Tag	SGI Challenge
Mangoni	[128]	1	1	flex.	flex.	k.A.[10]	6.5 h	SGI P.Chall.(R10000)

Tabelle 7.14: Vergleich der Protein-Ligand-Docking-Verfahren II. Für jedes Verfahren sind der Erstautor und die Referenz, die Zahl der Testsysteme (Liganden u. Proteine bzw. Ensemble (E)), der Grad der berücksichtigten Flexibilität bei Liganden u. Proteinen , die Größenordnung der RMS-Abweichung der plazierten Liganden (RMSD), die Laufzeit, die für das Docking eines Liganden im Durchschnitt benötigt wird, sowie die verwendete Hardware angegeben. Eine genaue Beschreibung der Methoden findet man in den Abschnitten 3.3.2 – 3.3.4. Diese Tabelle erweitertTabelle 3.1 (Abs. 3.3.6) um den hier vorgestellen Ansatz (fett hervorgehoben).
[1]) z.T vorplaziert; [2]) vorplaziert; [3]) des Proteins; [4]) Evaluation über Anreicherungsfaktoren; [5]) Evaluation über Rang bekannter Liganden; [6]) 5 – 32 CPU Stunden für die gesamte Datenbank; [7]) 9 min – 17 h für gesamte Datenbank, dabei wird ein Großteil der Liganden gar nicht erst gedockt; [8]) Ergebnisse von CASP2, bei der RMS-Abweichung wurden drei Komplexe nicht berücksichtigt, weil sie kovalent gebunden (2) bzw. auf einer falschen Ligandstruktur (1) beruhen; [9]) Verwendung von Rotameren; [10]) Evaluation über Wechselwirkungslängen.
Das Programm AutoDock [37, 173] fehlt in dieser Übersicht, weil bis jetzt noch keine Daten über die Lauftzeit bei Testsytemen mit Proteinflexibilität veröffentlicht worden sind. Der Ansatz von Zacharias et al. [176] fehlt ebenfalls, weil Angaben über die Gesamtlaufzeit des Verfahrens und die Qualität der Lösungen fehlen.

Kapitel 8

Limitierungen und Lösungsansätze

Die computergerechte Modellierung eines komplexen Problems wie die des Protein-Ligand-Dockings erfordert eine Vielzahl von Annahmen und Vereinfachungen. Diese Heuristiken limitieren die Vorhersagequalität des vorgestellten Docking-Verfahrens. In diesem Kapitel werden einige dieser Einschränkungen aufgezeigt und Ansätze diskutiert, sie zu umgehen.

8.1 Modellierung der Proteinflexibilität

8.1.1 Drehbare endständige Gruppen

Endständige Gruppen der Aminosäuren im Protein sind meist frei drehbar, das heißt, sie sind in der Regel nicht sterisch behindert und können sich zur Bildung einer Wechselwirkung optimal zum Liganden ausrichten.

Solche Gruppen werden im Moment nicht gesondert behandelt. Alternative Konformationen dieser Gruppen kann man in Form verschiedener Instanzen für die betreffende Seitenkette modellieren. Dabei sind mehrere Instanzen notwendig, um die volle Rotation abzudecken. Dieses Vorgehen erzeugt jedoch aus zwei Gründen eine Vielzahl redundanter Wechselwirkungspunkte, die die Hashtabelle unnötig belasten: Erstens wird auch der Teil der Seitenkette, der unverändert bleibt, und damit gegebenenfalls auch seine Wechselwirkungsgeometrien dupliziert. Zweitens überlappen die Wechselwirkungsgeometrien der alternativen Konformationen der endständigen Gruppe.

Das erste Problem könnte relativ leicht gelöst werden, indem man die Seitenkette noch einmal in zwei Segmente unterteilt, die jeweils alle Instanzen des Seitenkettenstamms bzw. der endständigen Gruppe umfassen. Eine geeignete Erweiterung des Konzepts der strukturellen Kompatibilität würde dabei die Verknüpfung der Instanzen aus diesen beiden Segmenten kontrollieren, wenn es mehrere Konformationen oder Mutationen für die gesamte Seitenkette gibt. Bei diesem Ansatz würde es allerdings ebenfalls zu überlappenden Wechselwirkungsgeometrien der endständigen Gruppe kommen.

Ein Modell, das redundante Wechselwirkungspunkte aufgrund dieser Überlappung vermeidet, könnte etwa folgendermaßen aussehen: Anstelle der bisherigen Wechselwirkungsgeometrie ordnet man einer frei drehbaren endständigen Gruppe eine Art Torus zu, der durch die Rotation der Geometrie um die Drehachse der endständigen Gruppe entsteht. Dieser Torus wird genauso wie die übrigen Wechselwirkungsgeometrien mit Punkten diskretisiert (vgl. Abb 8.1). Bei der Verwendung dieser Wechselwirkungspunkte müßten dann

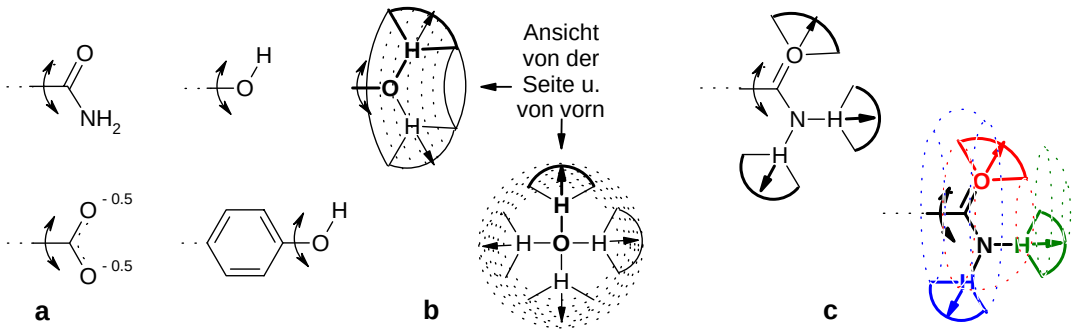


Abbildung 8.1: Drehbare endständige Gruppen. (a) Beispiele für drehbare endständige Gruppen von Aminosäuren. (b) Die Rotation von Wechselwirkungsgeometrien kann man durch eine Art Torus beschreiben, der sich durch diskrete Punkte approximieren läßt. (c) Wenn mehrere Wechselwirkungsgeometrien von der Rotation betroffen sind, muß man auch räumliche Abhängigkeiten zwischen den verschiedenen Wechselwirkungen beachten: Hier müssen z.B. die ausgewählten Wechselwirkungspunkte der beiden Wasserstoffe in etwa den gleichen Winkel in ihrem jeweiligen Torus haben, während die Punkte des Sauerstoffs in dem um 180° gedrehten Winkelbereich liegen müssen.

Abhängigkeiten etwa in der Form beachtet werden, daß nur Punkte, die in einem bestimmten Winkelbereich des Torus liegen, gleichzeitig verwendet werden dürfen. Bei endständigen Gruppen, die mehrere Wechselwirkungen ausbilden können, müßte außerdem sicher gestellt werden, daß die jeweils ausgewählten Punkte der verschiedenen Geometrien kompatibel sind. Dafür ist bei dieser Modellierung eine zusätzliche Segmentierung der Seitenkette nicht erforderlich.

8.1.2 Untypische Konformationen und dynamische Bewegungen

Die hier verwendete Modellierung der Proteinflexibilität basiert im wesentlichen auf der Auswahl einer Proteinkonformation aus der Menge diskreter Alternativen, die man durch die Rekombination der Ensemblestrukturen erzeugen kann. Die ausgewählte Proteinkonformation sollte dabei für die Bindung des speziellen Liganden am besten geeignet sein. Völlig neue Proteinkonformationen kann dieser Ansatz deshalb nicht vorhersagen.

Die Zahl der alternativen Konformationen für einzelne Seitenketten könnte man zwar durch die Verwendung von Rotamerbibliotheken erhöhen, aber diese Bibliotheken umfassen nur Seitenkettenkonformationen, die man typischerweise in Proteinstrukturen findet. Sehr selten auftretende Konformationen sind auch in Rotamerbibliotheken nicht enthalten. Deshalb kann man auch mit Hilfe dieser Bibliotheken keine ungewöhnlichen oder unerwarteten Seitenkettenpositionen vorhersagen.

Die Menge der diskreten alternativen Konformationen ist ein statisches Abbild der Proteinflexibilität. Die Dynamik der Flexibilität wird von diesem Modell nicht erfaßt. Der in dieser Arbeit vorgestellte Ansatz konzentriert sich somit, wie alle diskreten Docking-Programme, vor allem auf die Frage, *wo* ein Ligand in der Bindetasche plaziert ist.

8.1.3 Bewegungen von Domänen

FlexE kann einen *Induced-Fit*, bei dem sich ganze Domänen des Proteins bewegen, nicht handhaben. Prinzipiell könnten Domänenbewegungen zwar in dem vorgestellten Proteinmodell durch ein Ensemble von Strukturen beschrieben werden, die Zwischenstufen zwischen der geöffneten und dem geschlossen Form des Protein darstellen, aber die daraus resultierende vereinigte Proteinbeschreibung wäre aus folgenden Gründen zu komplex, um noch effizient behandelt werden zu können:

1. Es wäre eine sehr große Zahl von Ensemblestrukturen erforderlich, weil es auch für jede Zwischenstufe alternative Seitenkettenkonformationen gibt. Daraus ergeben sich sehr viele Instanzen je Segment.
2. Alle Instanzen einer Zwischenstufe der beweglichen Domäne sind mit allen Instanzen aller anderen Zwischenstufen strukturell inkompatibel. Dadurch entsteht ein Inkompatibilitätsgraph mit sehr großen Zusammenhangskomponenten und das erweiterte aktive Zentrum dehnt sich auf die gesamte bewegliche Domäne aus.

8.1.4 Technische Einschränkungen

Aus technischer Sicht gibt es vor allem zwei kritische Größen für das Verfahren: die Zahl der Wechselwirkungspunkte, die die Größe der Hashtabelle bestimmt, sowie die Größe, die Anzahl und die Gestalt der Zusammenhangskomponenten im Inkompatibilitätsgraphen. Der Aufbau großer Hashtabellen dauert sehr lange und erfordert viel Speicherplatz, während bei komplexen Zusammenhangskomponenten mehr Zeit für die Suche nach geeigneten unabhängigen Mengen nach jedem Aufbauschritt benötigt wird. Vor allem der zweite Parameter kann a priori nur sehr schlecht abgeschätzt werden.

Zur Reduktion der Hashtabelle wurde bereits in Abschnitt 5.3 ein Verfahren zum Clustern der Wechselwirkungspunkte vorgestellt.

Um den Einfluß der Zusammenhangskomponenten auf die Laufzeit der Plazierung zu reduzieren, gibt es im wesentlichen zwei Möglichkeiten: Auf der einen Seite könnte man versuchen, die Zahl der Suchen in den Zusammenhangskomponenten durch Vorberechnung von unabhängigen (Teil-)Mengen oder durch das Wiederverwenden von Lösungen zu verringern. Auf der anderen Seite ließe sich unter Umständen die Komplexität der Zusammenhangskomponenten durch eine geeignete Zerlegung der Komponenten vermindern.

8.2 Bewertungsfunktion

8.2.1 Qualität der Bewertungsfunktion

Die vorhergesagten Plazierungen des Docking-Algorithmus sind bezüglich der implementierten Bewertungsfunktion optimiert, die eine heuristische Approximation der freien Bindungsenergie darstellt, denn eine exakte Berechnung der freien Bindungsenergie ist praktisch nicht möglich (vgl. Abs. 4.5). Damit hat die Güte der Bewertungsfunktion großen Einfluß auf die Vorhersagequalität des Docking-Verfahrens.

Aufgrund der Vereinfachungen, auf denen die Bewertungsfunktionen basieren, bewerten sie nicht immer die Plazierungen mit minimaler RMS-Abweichung am besten. Dies ist eine prinzipielle Einschränkung der Leistungsfähigkeit, der jeder Docking-Algorithmus unterliegt. Eine Verbesserung von Bewertungsfunktionen war nicht das Ziel dieser Arbeit.

Das hier vorgestellte Verfahren zum Docken in flexible Proteinstrukturen führt einen zusätzlichen Freiheitsgrad ein, und zwar den der verschiedenen Proteinkonformationen. Die Bewertungsfunktion wird deshalb auch zur Auswahl der Instanzen benutzt, die energetisch am besten zum plazierten Liganden passen. Damit hängt von der Qualität der Bewertungsfunktion ebenfalls ab, ob die Abschätzung der Energie überhaupt auf der richtigen Proteinkonformation beruht. Da die Unterschiede zwischen den alternativen Konformationen aber unter Umständen nicht sehr groß sind, muß man davon ausgehen, daß die heuristische Funktion auch hier Fehler machen kann, die die Vorhersagequalität des Docking-Verfahrens limitieren.

8.2.2 Intramolekulare Wechselwirkungen

Auswahlkriterium für die Instanzen ist im Moment nur die Wechselwirkungsenergie der jeweiligen Instanz zum Liganden. Die Energie der Instanzen selbst wird nicht berücksichtigt. Die Wechselwirkungen zwischen den verschiedenen ausgewählten Instanzen fließen nur in Form der strukturellen und geometrischen Kompatibilität ein. Das heißt, man betrachtet alle gültigen Proteinkonformationen als energetisch gleichwertig. Diese sehr starke Vereinfachung kann jedoch nicht mehr verwendet werden, wenn man die Anzahl der alternativen Konformationen mit Hilfe von Rotamerbibliotheken anreichert. In diesem Fall müssen intramolekulare Wechselwirkungen im Protein bei der Bewertung der Komplexe berücksichtigt werden, um zwischen besser und schlechter geeigneten Rotameren unterscheiden zu können.

8.3 Sonstige Limitierungen

8.3.1 Wasser im aktiven Zentrum

Bei der Bildung von Protein-Ligand-Komplexen spielt Wasser eine wichtige Rolle. Im ungebundenen Zustand füllen Wassermoleküle das aktive Zentrum des Proteins aus. Sie werden bei der Bindung eines Liganden nicht immer vollständig verdrängt. Einzelne Moleküle können in der Bindetasche verbleiben und Wasserstoffbrückenbindungen zwischen dem Protein und dem Liganden vermitteln. Ob und wo solche Wassermoleküle in einem Komplex auftreten, hängt dabei im Einzelfall von den Liganden ab und kann bei verschiedenen Liganden variieren.

Für FlexX wurde zur Vorhersage dieser optionalen Wassermoleküle das sog. *Partikelkonzept* entwickelt [88]. Es berechnet nach jedem Anbauschritt eine Menge von günstigen Positionen für Wassermoleküle und bezieht diese in die Bewertung der Teillösungen mit ein. Im weiteren Verlauf des inkrementellen Aufbaus des Liganden, können diese Wassermoleküle gegebenenfalls auch wieder entfernt werden, wenn an derselben Stelle ein Ligandfragment plaziert werden kann.

FlexE kann optionale Wassermoleküle zur Zeit noch nicht behandeln. Aber aufgrund der engen Verwandtschaft von FlexE und FlexX könnte das Partikelkonzept in FlexE leicht übernommen werden. Eine andere Möglichkeit besteht darin, günstige Wasserpositionen als Instanzen in den Inkompatibilitätsgraphen aufzunehmen. Die Optionalität kann dabei wie bei der Mutation eines Glycins durch Pseudoinstanzen realisiert werden (vgl. Abs. 4.1.1).

8.3.2 Große Liganden

Die vorgestellten Ergebnisse zeigen, daß die Vorhersage des korrekten Bindungsmodus bei großen Liganden sowohl für FlexE als auch für FlexX problematisch ist. Man kann davon ausgehen, daß FlexX Liganden nur bis zu einer Größe von etwa 15 Fragmenten plazieren kann [79]. Verantwortlich dafür ist die Greedy-Strategie des inkrementellen Aufbaus der Liganden. Sie beruht auf dem Prinzip der lokalen Optimalität, das heißt, die Verknüpfung lokal optimal plazierter Fragmente sollte auch eine global optimale Lösung ergeben.

Bei großen Liganden ist der Bindungsmodus aber ein energetischer Kompromiß, der es ermöglicht, alle Fragmente gleichzeitig zu plazieren. Darum kann es für die Einzelfragmente jeweils energetisch günstigere Plazierungen geben, die jedoch zusammen mit den anderen Fragmenten nicht realisiert werden können. Aus diesem Grund kann man lokal optimal plazierte Basisfragmente nicht unbedingt zur richtigen Lösung expandieren. Dieser Effekt verstärkt sich unter Umständen, wenn es alternative Proteinkonformationen gibt. Denn hier werden möglicherweise Alternativen für die Plazierung des einzelnen Fragments benutzt, die später mit anderen Teilen des Liganden inkompatibel sind.

Mit einer vorausschauenden Bewertung, die abschätzt, ob die gegebene Teillösung auch noch die Plazierung des restlichen Liganden erlaubt, ließe sich hier Abhilfe schaffen, weil so eine Funktion Teillösungen verwerfen könnte, die zwar für sich optimal sind, aber nicht expandiert werden können.

Kapitel 9

Zusammenfassung und Ausblick

Diese Arbeit befaßt sich mit einem speziellen Aspekt des molekularen Docking-Problems, und zwar der Modellierung von Proteinflexibilität bei der computerbasierten Vorhersage von Protein-Ligand-Komplexen, die auch die Flexibilität der Liganden berücksichtigt.

Den Ausgangspunkt bildet das Docking-Programm FlexX, von dem das hier beschriebene Verfahren neben der Repräsentation von Wechselwirkungen und der Bewertungsfunktion vor allem die Behandlung der Liganden übernimmt. Das Modell der Ligandflexibilität beruht dabei auf einem inkrementellen Aufbau dieser Moleküle im aktiven Zentrum des Proteins. Die Proteinstrukturen werden dagegen von FlexX als starr betrachtet. Diese Annahme kann aber dazu führen, daß sich Liganden, die Konformationsänderungen im Protein induzieren, nicht in der Bindetasche plazieren lassen. Dies wurde anhand des Beispiels der Aldose-Reduktase belegt.

Es gibt bisher nur wenige Verfahren, die Proteinflexibilität beim Protein-Ligand-Docking berücksichtigen. Allerdings ist die modellierte Flexibilität bei diesen Verfahren zum Teil sehr eingeschränkt, ersetzt die Ligandflexibilität oder ist als Nachoptimierung realisiert.

Der Ansatz dieser Arbeit basiert darauf, die Proteinflexibilität, Punktmutationen oder alternative Modellierungen eines Proteins durch ein Ensemble von möglichen Proteinstrukturen zu repräsentieren, das nicht auf experimentell bestimmte Konformere beschränkt ist. Diese Strukturen lassen sich während des Docking-Prozesses zu neuen gültigen Proteinkonformationen kombinieren, die in bezug auf die Bewertungsfunktion optimal zu der jeweiligen Ligandplazierung passen.

Zur Verwaltung der alternativen Proteinkonformationen (Instanzen) und ihrer Abhängigkeiten, wurden das Konzept der Kompatibilität eingeführt und geeignete Datenstrukturen entworfen, um dieses in Form eines Graphen darzustellen. Die Möglichkeit, alternative Instanzen rekombinieren zu können, machte außerdem Modifikationen bei der Zuordnung von Wechselwirkungen, der Definition der Oberfläche sowie bei der Anwendung der Bewertungsfunktion erforderlich.

Es gelang, die Auswahl der optimalen gültigen Proteinkonformationen bei der Plazierung eines Liganden auf die Suche nach unabhängigen Mengen im Inkompatibilitätsgraphen zu übertragen. Durch die Zerlegung dieses Graphen in Zusammenhangskomponenten und die Beschränkung der Suche auf das erweiterte aktive Zentrum ließ sich darüber hinaus die Bestimmung der optimalen Proteinkonformationen erheblich beschleunigen.

Die beschriebenen Modelle und Algorithmen wurden in dem Docking-Programm FlexE realisiert und empirisch getestet.

Basis für die Evaluierung war ein Testdatensatz bekannter Protein-Ligand-Komplexe, der aus zehn Ensembles mit insgesamt 106 Proteinstrukturen und 60 Liganden besteht.

Alle Komplexe wurden mit FlexE reproduziert (Redocking) und die RMS-Abweichungen zu den Referenzstrukturen bestimmt. Außerdem wurde ein Vergleich mit einem Kreuz-Docking-Experiment mit FlexX durchgeführt, das die Liganden sequentiell in die einzelnen Ensemblestrukturen plaziert und die Lösungen zu einer Liste zusammenfaßt.

Die Ergebnisse des Redockings mit FlexE belegen, daß der Ensemble-Ansatz in der Lage ist, mehrere Seitenkettenkonformationen und sogar Loop-Bewegungen effektiv zu behandeln. Denn FlexE findet in 83% aller Testfälle akzeptable Ligandplazierungen mit RMS-Abweichungen von weniger als 2.0 Å.

Wenn sich unter den Ensemblestrukturen eine Proteinkonformation befindet, in die der Ligand plaziert werden kann, treten im Vergleich von FlexE zu FlexX keine signifikanten Unterschiede der Vorhersagequalität auf. Beide Programme finden auf den ersten zehn Rängen für rund zwei Drittel der Testliganden eine akzeptable Lösung.

Der Vorteil von FlexE wird erst deutlich, wenn es keine Ensemblestruktur gibt, in die der Ligand paßt: Während man durch das nachträgliche Zusammenfassen von FlexX-Lösungen keine richtigen Plazierungen findet, ist FlexE noch immer in der Lage, gute Vorhersagen zu machen, weil es die gegebenen Proteinkonformationen in geeigneter Weise beim Docking rekombinieren kann. Dies konnte am Beispiel des verkleinerten Ensembles für die Dihydrofolat-Reduktase nachgewiesen werden.

Die Rechenzeit von FlexE beträgt im Durchschnitt 5.5 Minuten auf einer gängigen Workstation, das bedeutet, FlexE ist um den Faktor zwei schneller als FlexX (akkumulierte Zeit) und gehört zu den schnelleren der aktuellen Docking-Programme. Mit dieser Laufzeit eignet sich FlexE auch für ein Screening mittelgroßer Datenbanken von bis zu etwa 50 000 Verbindungen.

Ein solches Screening mit FlexE konnte bis jetzt wegen fehlender Daten noch nicht durchgeführt werden. Es sollte deshalb einer der nächsten Tests für das neue Verfahren sein. Ferner sollte der Testdatensatz für das Redocking durch eine systematische Suche in der PDB erweitert werden.

Es wurden einige Limitierungen des Verfahrenes, sowie Lösungsansätze zu ihrer Umgehung aufgezeigt. Diese Ansätze können die Grundlage für eine Weiterentwicklung des vorgestellten Docking-Algorithmus bilden. Insbesondere die geschlossene Modellierung von drehbaren endständigen Gruppen, die Integration von Rotamer-Bibliotheken und die Berücksichtigung intramolekularer Wechselwirkungsenergien im Protein sollten das Anwendungsspektrum von FlexE wesentlich erweitern.

Weitere offene Punkte sind eine Verbesserung der Bewertungsfunktion und die Entwicklung effizienter Vorfilter, die bei Screening-Anwendungen die Zahl der Moleküle, die gedockt werden müssen, effektiv reduzieren.

Insgesamt hat sich FlexE aber schon jetzt als ein effizientes Verfahren mit guter Vorhersagequalität für das Protein-Ligand-Docking mit flexiblen Proteinstrukturen erwiesen.

Anhang A

Ligandstrukturen und Kreuz-Docking-Matrizen

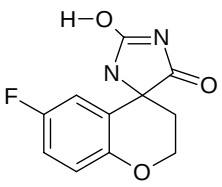
Dieser Anhang enthält für alle Ensembles eine Abbildung des aktiven Zentrums der vereinigten Proteinbeschreibung und die Strukturen der Liganden aus dem Testdatensatz, die als Referenzen für den korrekten Bindungsmodus dienen. Außerdem sind für alle Liganden die RMS-Abweichungen der jeweils ersten Vorhersage, der besten unter den ersten zehn Lösungen und die beste Plazierung unabhängig vom Rang angegeben, die von FlexE, durch das Zusammenfassen der FlexX-Lösungen bzw. von FlexX für die einzelnen Ensemblestrukturen vorhergesagt wurden.

Es wurden nicht alle Liganden verwendet, die in den PDB-Einträgen des Testdatensatzes enthalten sind. Zu kleine oder kovalent gebundene Moleküle blieben unberücksichtigt. Die wichtigsten Eigenschaften der Ensembles und der 60 ausgewählten Liganden fassen die Tabellen 7.2 und 7.3 in Abschnitt 7.2.1 zusammen. Die Abbildungen der vereinigten Proteinbeschreibungen wurden mit FlexV [232] produziert. Die Ligandstrukturen wurden mit dem Programm ISIS_Draw [233] gezeichnet.

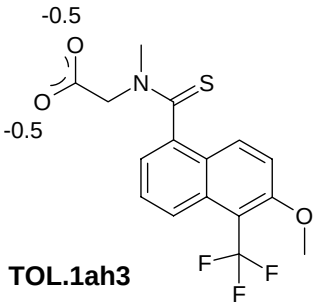
Die Liganden aller Ensembles des Testdatensatzes werden mit FlexE jeweils in das aktive Zentrum der vereinigten Proteinbeschreibung plaziert, die aus allen Proteinstrukturen des entsprechenden Ensembles entsteht. Dann werden die symmetriekorrigierten RMS-Abweichungen aller Vorhersagen bestimmt. Als Referenz für die RMS-Abweichung diente jeweils die Position des Liganden in der überlagerten, experimentell bestimmten Proteinstruktur. Zum Vergleich werden alle Liganden aller Ensembles mit FlexX sequentiell in alle Proteinstrukturen der entsprechenden Ensembles gedockt (Kreuz-Docking) und die RMS-Abweichungen bestimmt. Dabei verwendet man dieselbe Größe des aktiven Zentrums und die gleiche Definition der Referenzstrukturen wie bei FlexE. Anschließend werden für alle Liganden die Lösungen, die FlexX für die einzelnen Ensemblestrukturen generiert hat, zu einer Lösungsmenge zusammengefaßt und anhand ihres Scores neu sortiert.

Die RMS-Abweichungen in Å aller separaten Kreuz-Docking-Experimente sind in farbcodierten Matrizen dargestellt. Darin entspricht jede Zeile einem Liganden und jede Spalte einer Proteinstruktur. Die erste Spalte enthält die Ergebnisse von FlexE, die zweite die Resultate der zusammengefaßten Lösungsmengen (MRGD) und die übrigen Spalten die Vorhersagen von FlexX für die einzelnen Proteinstrukturen. Die Matrizen wurden mit dem Programm MATLAB [234] erzeugt.

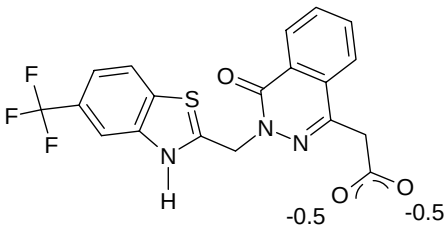
A.1 Aldose-Reduktase



SBI.1ah0



TOL.1ah3



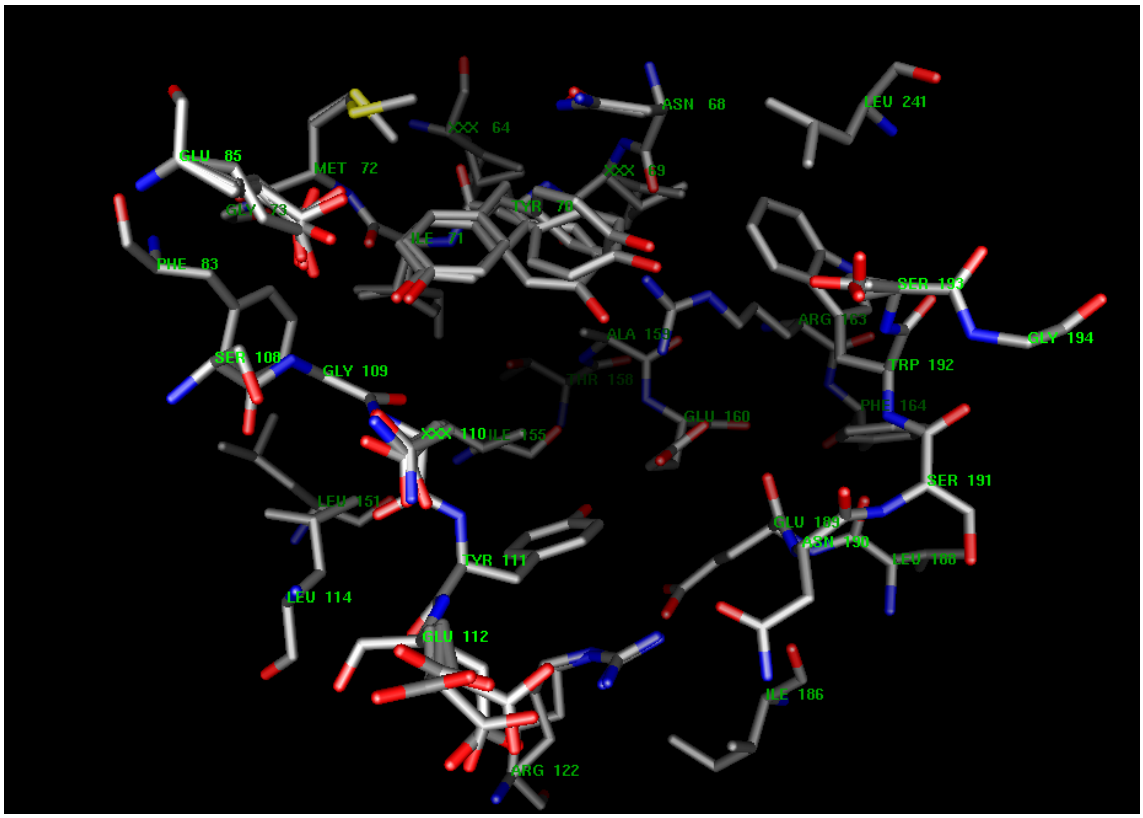
ZST.model

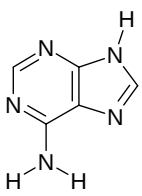
1. Rang	FlexE	MRGD	FlexX 1ah0	FlexX 1ah3	FlexX 1ah4	FlexX model
SBI.1ah0	0.58	7.66	7.66	0.92	7.58	0.56
TOL.1ah3	1.09	3.25	6.23	3.25	4.12	6.29
ZST.model	6.74	0.75	9.85	6.71	9.72	0.75

10 Ränge	FlexE	MRGD	FlexX 1ah0	FlexX 1ah3	FlexX 1ah4	FlexX model
SBI.1ah0	0.54	0.56	0.76	0.92	0.41	0.56
TOL.1ah3	1.05	0.72	3.83	0.72	4.12	5.38
ZST.model	6.72	0.75	7.95	6.56	7.56	0.75

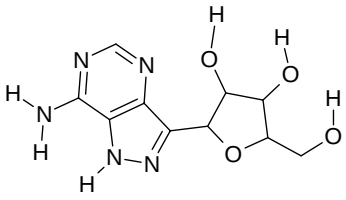
alle Ränge	FlexE	MRGD	FlexX 1ah0	FlexX 1ah3	FlexX 1ah4	FlexX model
SBI.1ah0	0.54	0.41	0.43	0.92	0.41	0.56
TOL.1ah3	1.05	0.72	3.71	0.72	3.82	4.16
ZST.model	0.64	0.75	6.14	6.41	6.84	0.75

A.2 Alpha-Momorcharin

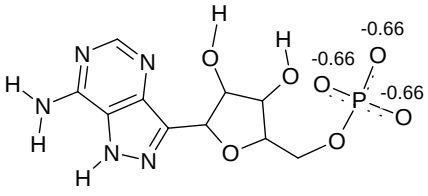




ADE.1aha /ADE.1mrg



FMC.1mrh



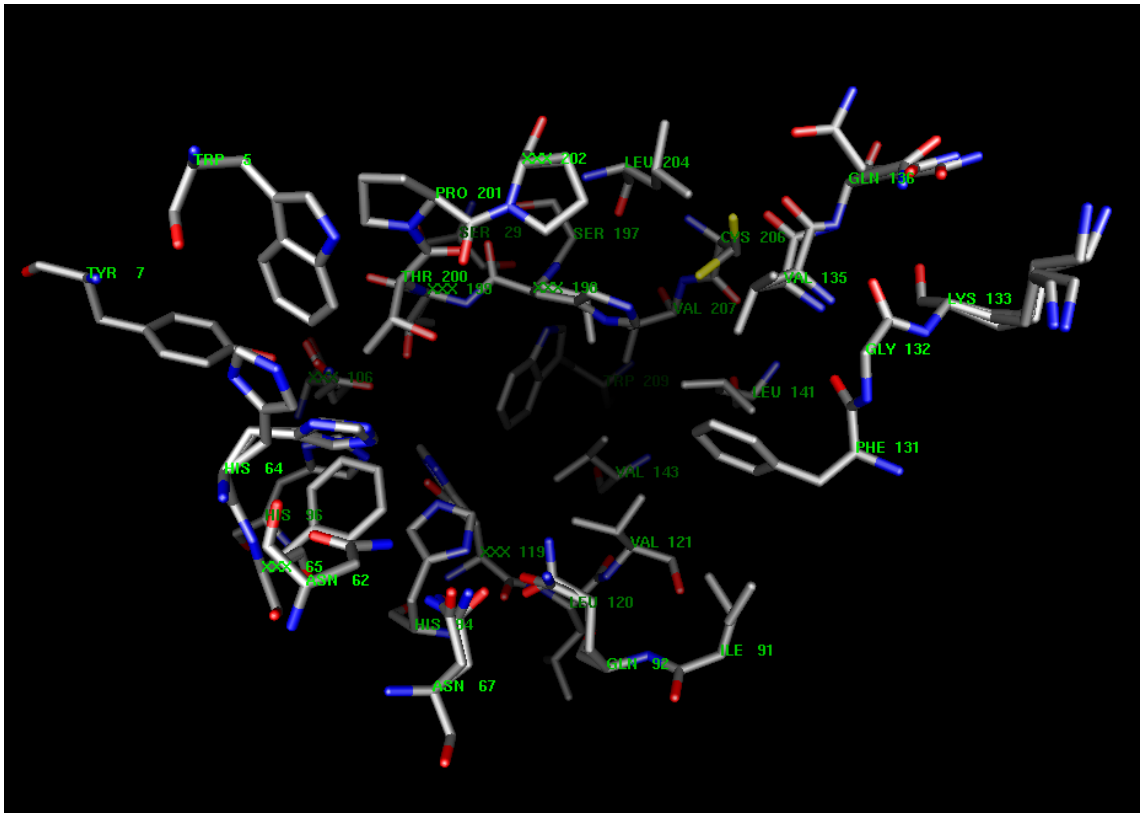
FMP.1ahb

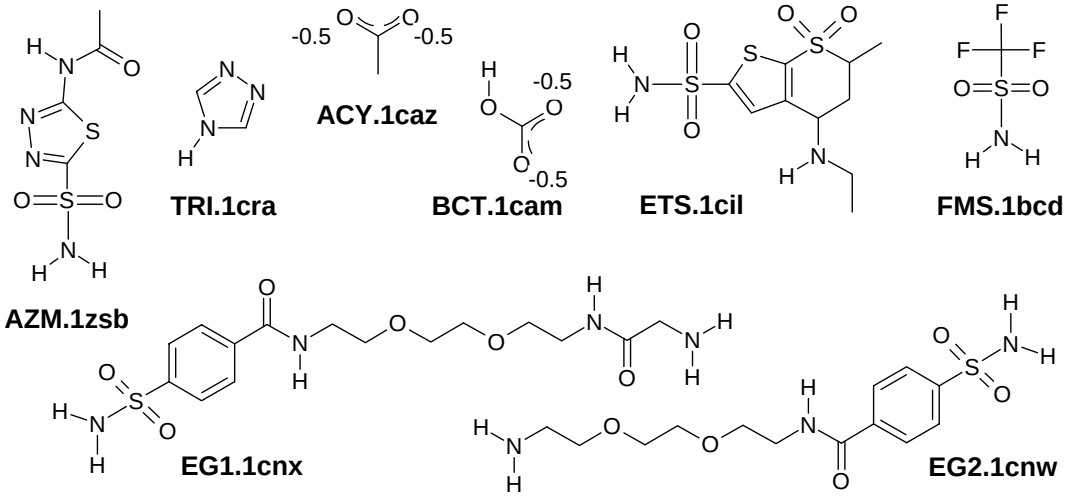
1. Rang	FlexE	FlexX MRGD	FlexX 1mri	FlexX 1ahc	FlexX 1mrh	FlexX 1mrg	FlexX 1aha	FlexX 1ahb	FlexX 1mom
FMC.1mrh	2.13	1.47	6.87	2.04	2.22	2.01	1.95	1.47	2.30
ADN.1mrg	4.30	0.72	4.57	4.55	0.75	4.33	4.28	0.72	1.01
ADE.1aha	3.27	0.67	3.45	3.34	0.70	3.29	3.24	0.82	0.67
FMP.1ahb	1.60	1.71	3.67	3.62	1.71	1.76	1.35	1.71	1.75

10 Ränge	FlexE	FlexX MRGD	FlexX 1mri	FlexX 1ahc	FlexX 1mrh	FlexX 1mrg	FlexX 1aha	FlexX 1ahb	FlexX 1mom
FMC.1mrh	1.75	1.47	1.97	1.35	1.27	1.61	1.43	1.32	1.31
ADN.1mrg	1.11	0.72	1.06	0.90	0.75	0.89	0.81	0.72	1.01
ADE.1aha	0.85	0.67	0.82	0.64	0.70	0.66	0.49	0.51	0.67
FMP.1ahb	1.42	1.23	2.00	1.54	1.46	1.46	0.97	1.26	1.48

alle Ränge	FlexE	FlexX MRGD	FlexX 1mri	FlexX 1ahc	FlexX 1mrh	FlexX 1mrg	FlexX 1aha	FlexX 1ahb	FlexX 1mom
FMC.1mrh	0.80	0.75	1.66	0.88	0.97	0.86	1.10	0.91	0.75
ADN.1mrg	0.97	0.72	1.06	0.90	0.75	0.89	0.81	0.72	1.01
ADE.1aha	0.76	0.49	0.82	0.64	0.70	0.66	0.49	0.51	0.67
FMP.1ahb	0.85	0.37	1.44	0.78	0.71	0.59	0.77	0.53	0.37

A.3 Carboanhydrase II





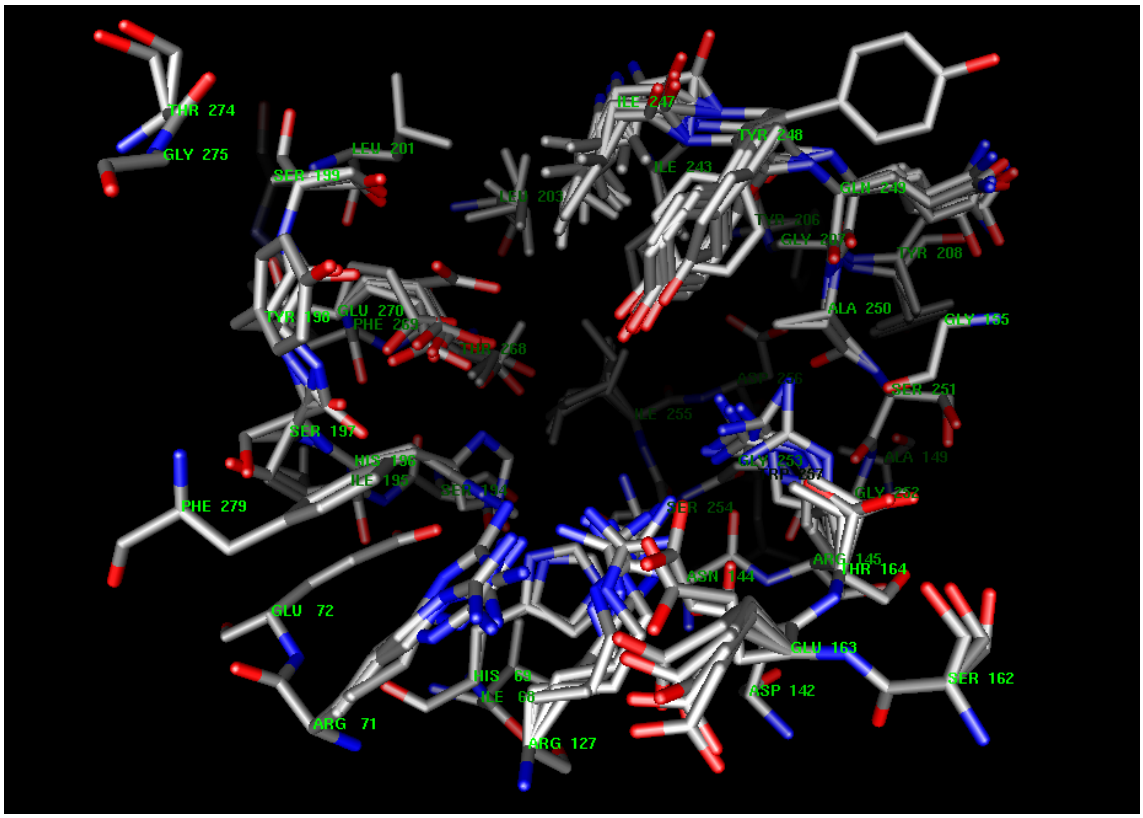
[0]Die Sulfonamide binden in einfach deprotonierter Form an die Carboanhydrase. Die Liganden wurden jedoch in neutraler Form gedockt, um eine spezielle Aufarbeitung für das Zielprotein zu vermeiden, da dies bei Datenbanken ebenfalls nicht möglich ist.

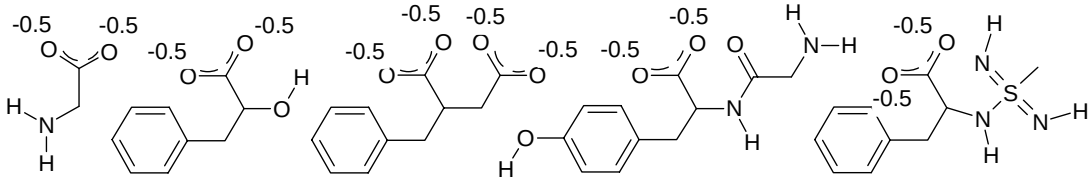
1. Rang	FlexE	FlexX																
		MRGD	2cbb	1h9n	1hec	1mua	1uga	1ugd	1bcd	1cil	1cnw	1cnx	1cra	1ray	1cam	1caz	1ugb	1zsb
FMS.1bcd	3.33	1.75	1.64	5.85	2.90	1.75	2.68	0.87	2.13	2.78	3.28	3.01	1.75	1.64	1.51	1.67	2.87	3.17
ETS.1cil	2.76	2.65	2.75	3.53	2.05	1.89	2.34	2.98	2.85	2.80	3.21	2.45	2.65	2.55	5.71	1.89	2.96	3.41
EG1.1cnw	7.86	10.7	11.4	11.4	11.2	10.9	7.52	11.5	10.7	7.72	11.4	11.3	11.8	7.78	6.02	10.6	11.0	11.8
EG2.1cnx	4.70	6.47	6.58	8.12	5.62	4.81	6.18	6.52	1.29	4.75	6.64	6.52	6.70	5.14	5.45	6.26	6.47	7.41
TRI.1cra	5.84	1.82	4.92	2.20	2.24	1.82	1.82	1.79	1.32	5.60	2.10	2.25	1.26	2.32	1.82	5.36	1.80	1.84
BCT.1cam	6.82	2.15	2.03	2.28	2.15	2.14	2.05	2.11	2.09	2.00	2.23	2.23	2.10	2.03	2.26	2.06	2.00	2.09
ACY.1caz	2.18	2.06	1.21	2.03	1.21	1.00	1.16	1.16	1.11	2.02	2.06	1.24	1.17	1.16	1.06	1.24	1.24	1.20
AZM.1zsb	6.48	6.00	6.43	6.00	2.95	3.21	3.05	4.65	5.96	3.27	3.29	6.33	6.03	2.97	6.33	3.74	3.21	3.12

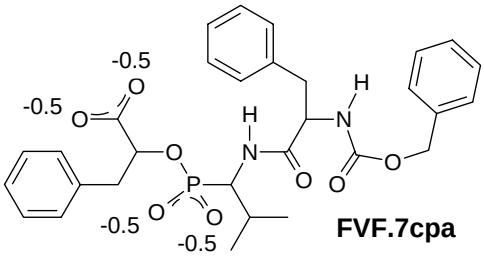
10 Ränge	FlexE	FlexX																
		MRGD	2cbb	1h9n	1hec	1mua	1uga	1ugd	1bcd	1cil	1cnw	1cnx	1cra	1ray	1cam	1caz	1ugb	1zsb
FMS.1bcd	1.47	0.87	1.25	2.90	0.57	1.75	1.48	0.87	0.94	1.31	2.45	2.66	1.39	1.64	1.51	1.31	2.05	1.97
ETS.1cil	2.69	2.45	1.68	1.99	2.05	1.89	1.74	1.89	2.21	2.42	2.08	2.16	1.93	1.89	2.14	1.78	2.95	3.41
EG1.1cnw	7.86	10.6	7.98	11.4	7.21	10.9	5.95	11.4	10.6	7.60	7.17	7.50	11.3	6.52	5.68	10.6	10.9	11.8
EG2.1cnx	4.70	5.85	5.84	4.74	5.58	4.53	6.01	5.69	1.29	4.50	5.90	5.52	6.43	5.14	5.29	6.26	5.85	6.16
TRI.1cra	0.82	1.77	1.48	2.20	1.97	1.82	1.68	1.77	0.67	5.08	1.89	1.73	1.20	1.69	1.26	1.71	1.65	1.84
BCT.1cam	2.51	1.82	1.68	1.85	1.82	1.80	1.31	1.87	1.99	1.89	1.89	1.91	1.86	1.69	1.60	1.89	1.89	1.82
ACY.1caz	1.18	1.88	1.21	1.57	1.04	0.91	1.01	1.03	1.09	1.86	1.06	0.85	1.02	1.07	0.90	1.15	0.80	1.04
AZM.1zsb	3.16	2.85	5.45	3.60	2.95	2.92	1.87	3.03	3.16	3.03	3.15	3.45	3.11	2.97	2.57	3.11	2.85	2.85

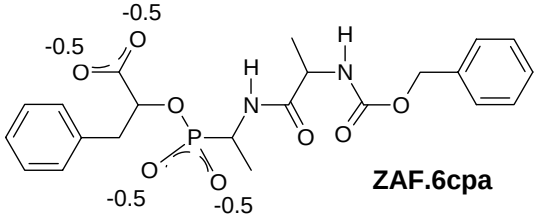
alle Ränge	FlexE	FlexX																
		MRGD	2cbb	1h9n	1hec	1mua	1uga	1ugd	1bcd	1cil	1cnw	1cnx	1cra	1ray	1cam	1caz	1ugb	1zsb
FMS.1bcd	1.47	0.52	0.60	0.64	0.57	1.42	1.48	0.52	0.81	1.19	1.31	0.82	1.36	0.86	1.51	1.12	1.32	1.16
ETS.1cil	2.35	0.94	1.62	1.79	1.72	1.89	1.74	1.72	1.78	0.94	1.91	1.64	1.60	1.75	2.05	1.53	1.61	1.55
EG1.1cnw	6.83	2.92	6.43	6.75	4.30	6.87	4.85	7.12	5.84	7.30	5.18	2.92	6.57	6.06	5.09	5.65	7.08	6.86
EG2.1cnx	4.10	1.29	5.68	4.55	4.24	4.33	5.80	5.56	1.29	3.33	5.51	5.47	5.90	5.05	5.16	5.09	4.65	5.32
TRI.1cra	0.79	0.66	0.87	1.67	0.88	1.04	0.84	0.70	0.67	0.77	0.86	0.90	0.78	0.81	0.73	0.66	0.82	0.75
BCT.1cam	1.96	1.31	1.38	1.65	1.52	1.65	1.31	1.65	1.62	1.70	1.52	1.74	1.69	1.62	1.39	1.34	1.36	1.62
ACY.1caz	0.96	0.42	1.05	0.46	1.04	0.91	1.01	1.03	1.09	0.42	0.54	0.85	1.02	1.07	0.90	1.15	0.80	1.04
AZM.1zsb	1.85	1.15	1.80	1.74	1.53	1.79	1.44	2.02	2.18	1.32	1.75	1.15	1.81	1.85	1.98	1.80	1.69	2.85

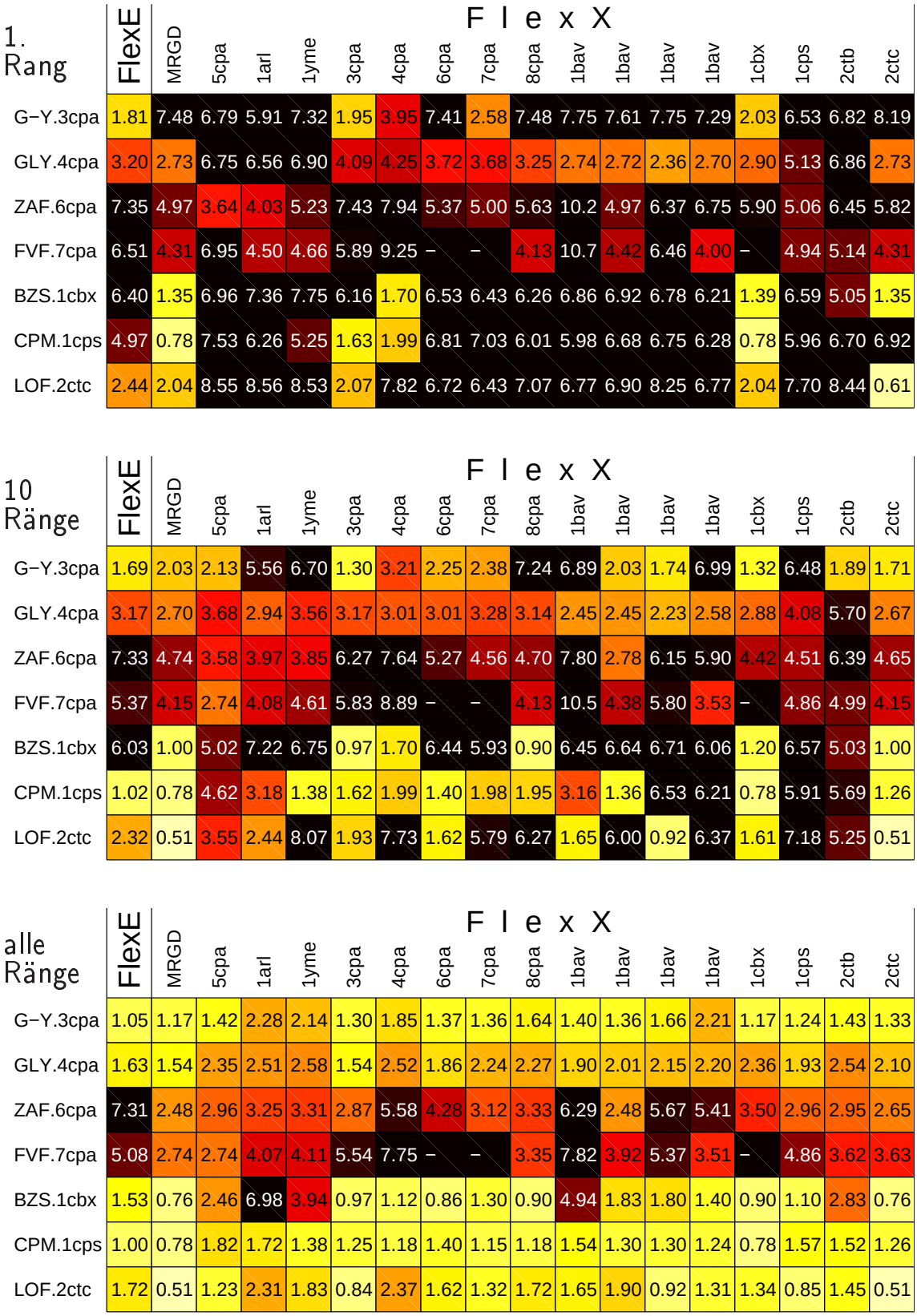
A.4 Carboxypeptidase



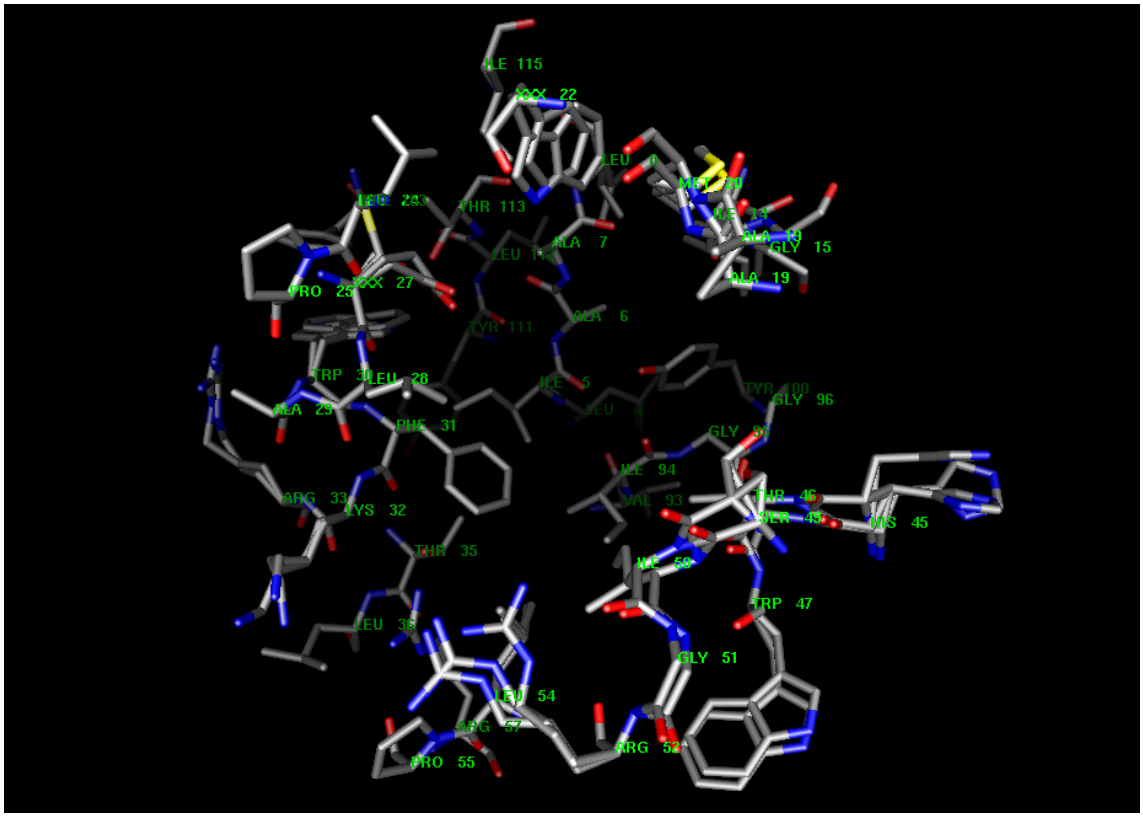


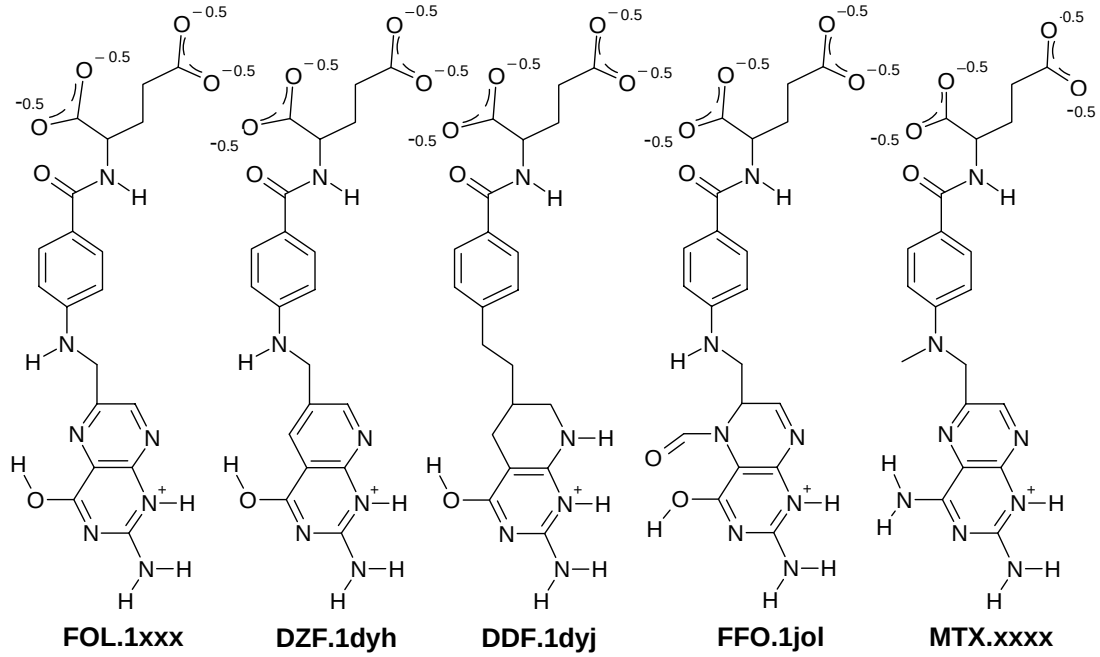





1. Rang
FlexE
FlexX
MRGD 5cpa 1arl 1yme 3cpa 4cpa 6cpa 7cpa 8cpa 1bav 1bav 1bav 1bav 1cbx 1cps 2ctb 2ctc
G−Y.3cpa 1.81 7.48 6.79 5.91 7.32 1.95 3.95 7.41 2.58 7.48 7.75 7.61 7.75 7.29 2.03 6.53 6.82 8.19
GLY.4cpa 3.20 2.73 6.75 6.56 6.90 4.09 4.25 3.72 3.68 3.25 2.74 2.72 2.36 2.70 2.90 5.13 6.86 2.73
ZAF.6cpa 7.35 4.97 3.64 4.03 5.23 7.43 7.94 5.37 5.00 5.63 10.2 4.97 6.37 6.75 5.90 5.06 6.45 5.82
FVF.7cpa 6.51 4.31 6.95 4.50 4.66 5.89 9.25 – – 4.13 10.7 4.42 6.46 4.00 – 4.94 5.14 4.31
BZS.1cbx 6.40 1.35 6.96 7.36 7.75 6.16 1.70 6.53 6.43 6.26 6.86 6.92 6.78 6.21 1.39 6.59 5.05 1.35
CPM.1cps 4.97 0.78 7.53 6.26 5.25 1.63 1.99 6.81 7.03 6.01 5.98 6.68 6.75 6.28 0.78 5.96 6.70 6.92
LOF.2ctc 2.44 2.04 8.55 8.56 8.53 2.07 7.82 6.72 6.43 7.07 6.77 6.90 8.25 6.77 2.04 7.70 8.44 0.61
10 Ränge
FlexE
FlexX
MRGD 5cpa 1arl 1yme 3cpa 4cpa 6cpa 7cpa 8cpa 1bav 1bav 1bav 1bav 1cbx 1cps 2ctb 2ctc
G−Y.3cpa 1.69 2.03 2.13 5.56 6.70 1.30 3.21 2.25 2.38 7.24 6.89 2.03 1.74 6.99 1.32 6.48 1.89 1.71
GLY.4cpa 3.17 2.70 3.68 2.94 3.56 3.17 3.01 3.01 3.28 3.14 2.45 2.45 2.23 2.58 2.88 4.08 5.70 2.67
ZAF.6cpa 7.33 4.74 3.58 3.97 3.85 6.27 7.64 5.27 4.56 4.70 7.80 2.78 6.15 5.90 4.42 4.51 6.39 4.65
FVF.7cpa 5.37 4.15 2.74 4.08 4.61 5.83 8.89 – – 4.13 10.5 4.38 5.80 3.53 – 4.86 4.99 4.15
BZS.1cbx 6.03 1.00 5.02 7.22 6.75 0.97 1.70 6.44 5.93 0.90 6.45 6.64 6.71 6.06 1.20 6.57 5.03 1.00
CPM.1cps 1.02 0.78 4.62 3.18 1.38 1.62 1.99 1.40 1.98 1.95 3.16 1.36 6.53 6.21 0.78 5.91 5.69 1.26
LOF.2ctc 2.32 0.51 3.55 2.44 8.07 1.93 7.73 1.62 5.79 6.27 1.65 6.00 0.92 6.37 1.61 7.18 5.25 0.51
alle Ränge
FlexE
FlexX
MRGD 5cpa 1arl 1yme 3cpa 4cpa 6cpa 7cpa 8cpa 1bav 1bav 1bav 1bav 1cbx 1cps 2ctb 2ctc
G−Y.3cpa 1.05 1.17 1.42 2.28 2.14 1.30 1.85 1.37 1.36 1.64 1.40 1.36 1.66 2.21 1.17 1.24 1.43 1.33
GLY.4cpa 1.63 1.54 2.35 2.51 2.58 1.54 2.52 1.86 2.24 2.27 1.90 2.01 2.15 2.20 2.36 1.93 2.54 2.10
ZAF.6cpa 7.31 2.48 2.96 3.25 3.31 2.87 5.58 4.28 3.12 3.33 6.29 2.48 5.67 5.41 3.50 2.96 2.95 2.65
FVF.7cpa 5.08 2.74 2.74 4.07 4.11 5.54 7.75 – – 3.35 7.82 3.92 5.37 3.51 – 4.86 3.62 3.63
BZS.1cbx 1.53 0.76 2.46 6.98 3.94 0.97 1.12 0.86 1.30 0.90 4.94 1.83 1.80 1.40 0.90 1.10 2.83 0.76
CPM.1cps 1.00 0.78 1.82 1.72 1.38 1.25 1.18 1.40 1.15 1.18 1.54 1.30 1.30 1.24 0.78 1.57 1.52 1.26
LOF.2ctc 1.72 0.51 1.23 2.31 1.83 0.84 2.37 1.62 1.32 1.72 1.65 1.90 0.92 1.31 1.34 0.85 1.45 0.51

A.5 Dihydrofolat-Reduktase





[0]Das Methotrexat wurde protoniert, weil daduch eine Wasserstoffbrücke zum ASP 27 augebildet werden kann. Die Folsäure(derivate) wurden in der Enolform gedockt, obwohl sie eigentlich in der Ketoform binden (vgl. Abs. 7.3.1).

1. Rang

	FlexE	MRGD	1dyh	1dyi	1dyj	1jol	1ra2	1ra3	3drc	1dhj	1dra	1drb	2drc	4dfr
DZF.1dyh	2.21	2.58	3.50	2.34	2.10	2.24	2.41	2.52	2.58	3.56	2.47	3.73	2.46	2.16
FOL.1dyi	2.17	2.22	2.10	3.50	3.55	2.23	2.37	2.08	2.35	7.95	2.23	5.45	2.22	2.12
DDF.1dyj	5.43	3.86	3.54	5.07	2.59	4.56	2.81	4.00	2.89	3.86	4.70	5.50	5.77	4.19
FFO.1jol	8.49	3.27	3.33	4.98	3.11	5.39	3.27	4.45	5.21	8.28	2.94	6.05	5.18	3.71
FOL.1ra2	2.30	2.80	2.41	2.28	2.27	2.32	2.23	2.06	2.36	4.64	2.02	8.30	2.42	2.80
MTX.1ra3	1.50	1.58	1.42	1.74	1.79	3.85	1.19	2.40	1.58	2.99	1.38	3.03	1.54	1.37
MTX.3drc	1.21	1.31	0.92	1.41	1.65	1.73	1.03	2.47	1.13	2.68	1.19	2.85	1.31	0.93
MTX.1dhj	1.12	1.12	1.04	1.43	1.62	1.26	1.02	1.02	1.12	2.94	0.84	2.72	1.27	1.14
MTX.1dra	1.10	1.23	1.04	1.47	1.59	3.06	1.46	2.51	1.17	2.87	0.82	2.82	1.23	0.90
MTX.1drb	1.23	1.23	1.10	1.48	1.49	1.45	1.30	0.87	1.27	2.73	1.02	2.79	1.16	1.29
MTX.2drc	1.05	1.42	1.26	1.51	0.95	3.50	1.05	0.98	1.14	2.74	1.09	3.83	0.74	1.42
MTX.4dfr	1.32	1.48	1.19	1.55	1.39	3.67	1.06	1.54	1.54	2.65	1.11	6.14	0.78	1.48

	FlexE	MRGD	1dhj	1drb	1jol
DZF.1dyh	2.16	3.56	3.56	3.73	2.24
FOL.1dyi	2.60	2.23	7.95	5.45	2.23
DDF.1dyj	3.63	3.86	3.86	5.50	4.56
FFO.1jol	4.78	5.39	8.28	6.05	5.39
FOL.1ra2	2.56	2.32	4.64	8.30	2.32
MTX.1ra3	1.76	2.99	2.99	3.03	3.85
MTX.3drc	1.46	2.68	2.68	2.85	1.73
MTX.1dhj	1.48	1.26	2.94	2.72	1.26
MTX.1dra	1.41	2.87	2.87	2.82	3.06
MTX.1drb	1.30	2.79	2.73	2.79	1.45
MTX.2drc	1.41	2.74	2.74	3.83	3.50
MTX.4dfr	1.46	2.65	2.65	6.14	3.67

10 Ränge

	FlexE	MRGD	1dyh	1dyi	1dyj	1jol	1ra2	1ra3	3drc	1dhj	1dra	1drb	2drc	4dfr
DZF.1dyh	2.00	2.16	1.97	1.94	2.00	1.91	1.93	2.07	2.07	3.11	2.03	3.72	2.15	2.16
FOL.1dyi	1.84	2.12	1.94	2.10	2.10	1.94	1.87	2.04	2.08	3.93	1.95	3.47	1.94	2.12
DDF.1dyj	1.63	2.59	3.54	2.78	2.59	4.56	2.81	3.23	2.89	2.83	3.01	4.13	3.06	2.80
FFO.1jol	5.37	2.94	3.25	2.90	2.67	5.36	3.06	4.41	4.37	4.16	2.77	4.97	4.54	3.68
FOL.1ra2	1.91	2.41	2.01	2.09	2.22	2.02	1.95	2.06	2.04	4.26	1.92	7.55	2.12	2.63
MTX.1ra3	0.81	1.19	1.38	1.37	1.39	3.85	1.19	0.86	1.11	2.70	1.14	2.73	1.11	1.14
MTX.3drc	0.50	0.70	0.89	0.87	1.32	1.60	0.93	0.72	0.63	2.60	0.70	2.60	0.70	0.77
MTX.1dhj	0.67	0.68	0.91	0.88	1.20	1.24	0.96	0.72	0.66	2.59	0.59	2.58	0.68	0.82
MTX.1dra	0.70	0.72	0.97	0.86	1.17	3.06	1.05	0.81	0.66	2.60	0.57	2.60	0.71	0.53
MTX.1drb	0.98	0.88	1.06	0.99	1.01	1.45	0.94	0.75	0.87	2.55	0.69	2.59	0.68	0.87
MTX.2drc	0.97	0.74	1.13	1.16	0.95	3.50	0.97	0.73	1.09	2.64	0.92	3.42	0.74	0.81
MTX.4dfr	0.66	0.78	0.98	1.19	1.20	3.67	1.03	0.76	0.65	2.56	0.78	4.03	0.78	0.84

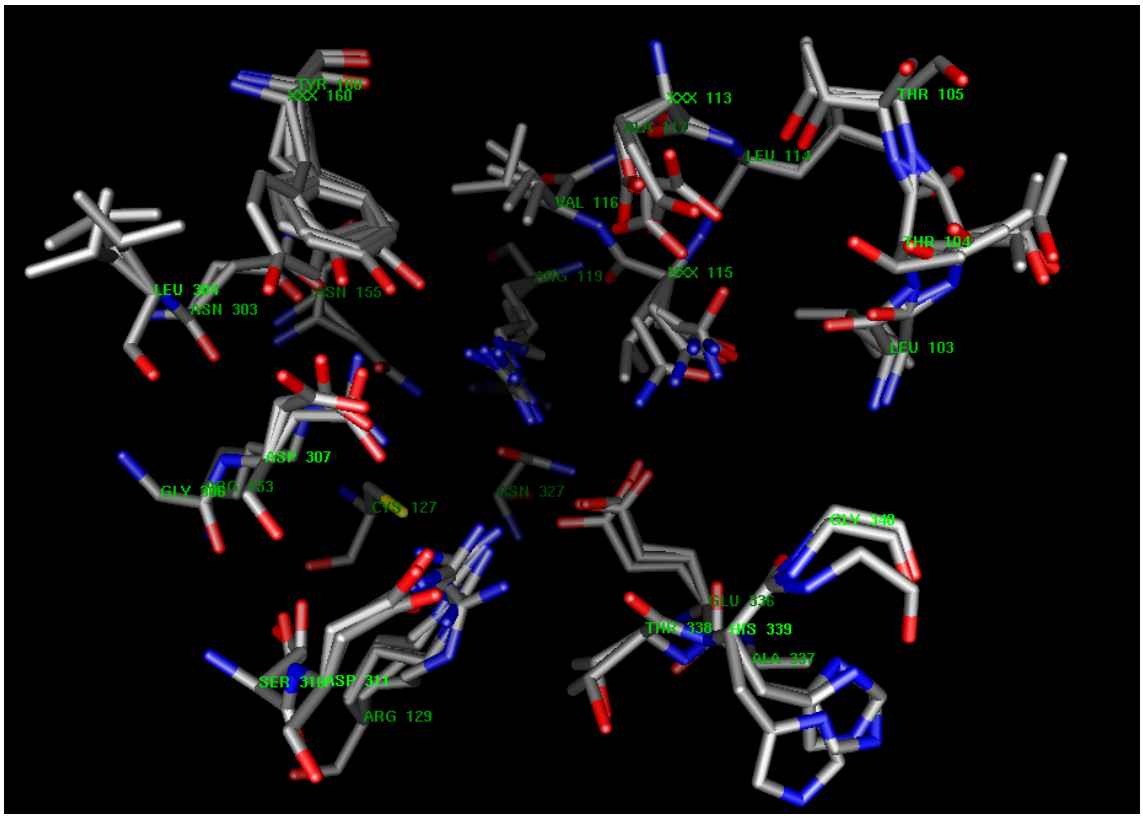
	FlexE	MRGD	1dhj	1drb	1jol
DZF.1dyh	2.01	2.09	3.11	3.72	1.91
FOL.1dyi	2.07	1.94	3.93	3.47	1.94
DDF.1dyj	3.14	2.83	2.83	4.13	4.56
FFO.1jol	3.95	5.36	4.16	4.97	5.36
FOL.1ra2	2.10	2.02	4.26	7.55	2.02
MTX.1ra3	1.06	2.70	2.70	2.73	3.85
MTX.3drc	0.80	2.60	2.60	2.60	1.60
MTX.1dhj	0.92	1.24	2.59	2.58	1.24
MTX.1dra	0.74	2.60	2.60	2.60	3.06
MTX.1drb	0.96	2.56	2.55	2.59	1.45
MTX.2drc	0.80	2.64	2.64	3.42	3.50
MTX.4dfr	0.90	2.56	2.56	4.03	3.67

alle Ränge

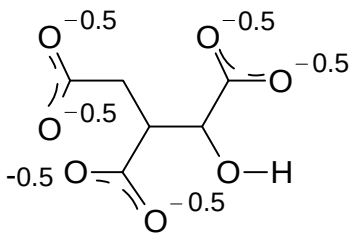
	FlexE	MRGD	1dyh	1dyi	1dyj	1jol	1ra2	1ra3	3drc	1dhj	1dra	1drb	2drc	4dfr
DZF.1dyh	1.86	1.77	1.86	1.84	1.80	1.91	1.77	1.85	1.81	3.08	1.88	3.57	1.86	1.88
FOL.1dyi	1.81	1.79	1.82	1.85	1.88	1.90	1.80	1.92	1.79	3.84	1.81	3.47	1.84	1.83
DDF.1dyj	1.58	1.96	3.16	2.77	2.56	4.13	2.81	3.23	1.96	2.22	3.01	3.15	2.64	2.78
FFO.1jol	5.21	2.40	2.98	2.88	2.40	4.21	3.06	4.41	4.37	3.83	2.64	4.97	3.89	3.53
FOL.1ra2	1.91	1.79	1.89	1.90	2.00	1.95	1.84	1.94	1.79	4.26	1.88	7.52	1.85	1.91
MTX.1ra3	0.53	0.53	0.76	0.79	1.06	3.85	0.99	0.55	0.57	2.59	0.58	2.51	0.73	0.53
MTX.3drc	0.50	0.53	0.74	0.83	0.86	1.41	0.92	0.72	0.57	1.56	0.59	2.55	0.58	0.53
MTX.1dhj	0.46	0.57	0.82	0.85	0.81	1.24	0.96	0.72	0.66	1.53	0.59	2.42	0.57	0.60
MTX.1dra	0.59	0.53	0.82	0.84	0.87	3.06	0.89	0.67	0.66	1.53	0.57	2.60	0.70	0.53
MTX.1drb	0.64	0.64	0.89	0.75	1.01	1.45	0.94	0.73	0.76	2.46	0.64	2.51	0.68	0.68
MTX.2drc	0.61	0.60	0.85	0.93	0.95	3.12	0.97	0.64	0.60	1.59	0.61	2.50	0.60	0.60
MTX.4dfr	0.65	0.62	0.93	0.78	1.08	3.67	0.98	0.66	0.62	1.66	0.67	2.51	0.63	0.80

	FlexE	MRGD	1dhj	1drb	1jol
DZF.1dyh	1.96	1.91	3.08	3.57	1.91
FOL.1dyi	1.84	1.90	3.84	3.47	1.90
DDF.1dyj	2.74	2.22	2.22	3.15	4.13
FFO.1jol	3.22	3.83	3.83	4.97	4.21
FOL.1ra2	1.97	1.95	4.26	7.52	1.95
MTX.1ra3	0.82	2.51	2.59	2.51	3.85
MTX.3drc	0.70	1.41	1.56	2.55	1.41
MTX.1dhj	0.73	1.24	1.53	2.42	1.24
MTX.1dra	0.70	1.53	1.53	2.60	3.06
MTX.1drb	0.79	1.45	2.46	2.51	1.45
MTX.2drc	0.76	1.59	1.59	2.50	3.12
MTX.4dfr	0.80	1.66	1.66	2.51	3.67

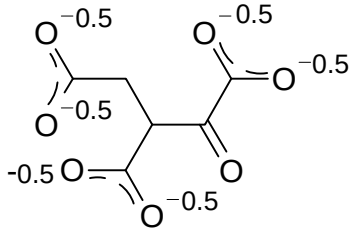
A.6 Isocitrat-Dehydrogenase



OXS.1icd



ICT.xxxx

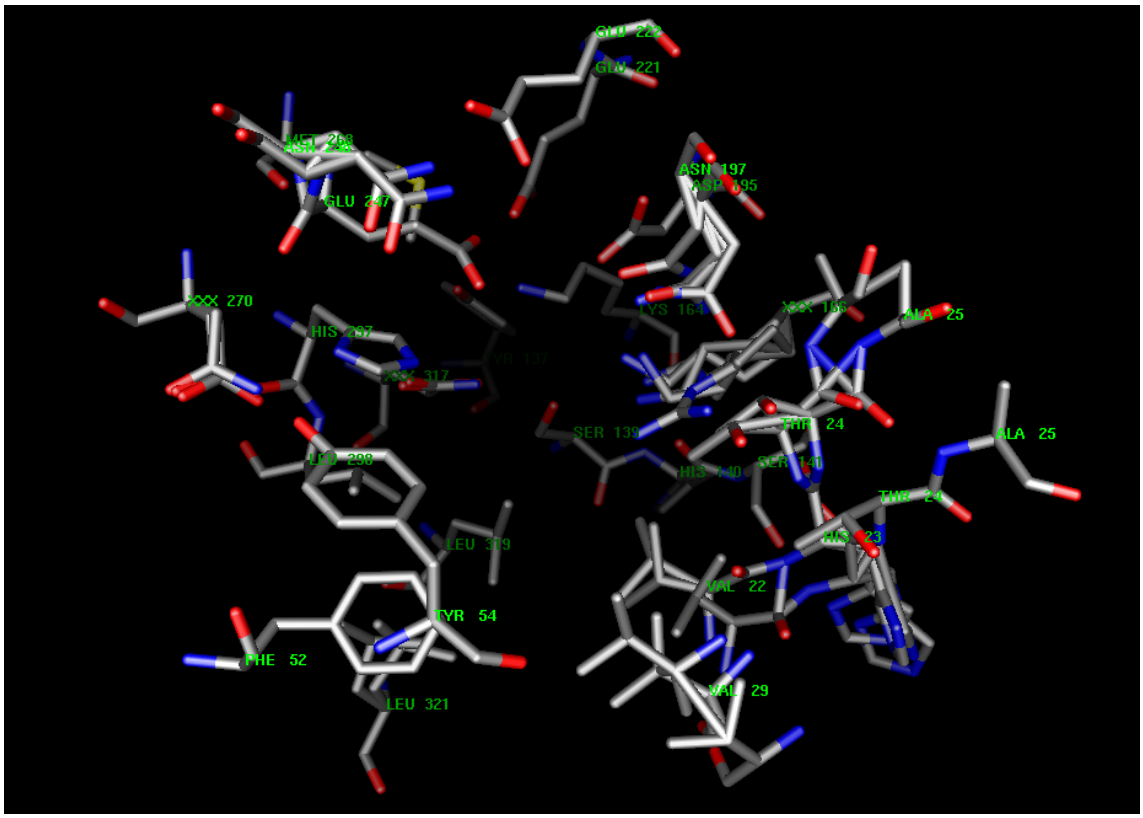


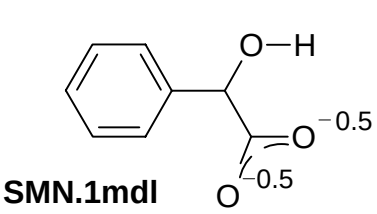
1. Rang	FlexE	FlexX														
		MRGD	1idf	1idd	6icd	7icd	1ika	4icd	5icd	9icd	1ide	1idc	1gro	1grp	8icd	1iso
ICT.5icd	4.35	1.50	3.34	4.22	4.22	3.78	3.15	3.81	3.24	3.76	3.12	2.98	2.49	1.50	0.74	2.98
ICT.1ide	4.29	1.46	3.51	3.94	4.07	3.30	2.64	4.05	2.86	4.05	3.00	3.53	2.27	1.46	1.10	2.96
OXS.1idc	3.77	3.37	3.82	4.95	4.19	4.11	3.54	3.94	3.37	3.89	3.96	3.87	4.08	4.06	3.78	3.57
ICT.1gro	4.29	1.15	3.28	4.13	4.14	3.74	3.21	3.82	2.94	3.76	2.94	3.82	3.43	1.15	0.41	2.96
ICT.1grp	4.38	1.42	3.38	4.27	4.27	3.83	3.13	3.83	2.96	3.84	3.12	3.91	3.54	1.42	0.72	2.98
ICT.8icd	4.25	1.24	3.26	4.08	4.14	3.68	3.27	3.73	2.99	3.66	2.95	4.01	3.35	1.24	0.39	2.97

10 Ränge	FlexE	FlexX														
		MRGD	1idf	1idd	6icd	7icd	1ika	4icd	5icd	9icd	1ide	1idc	1gro	1grp	8icd	1iso
ICT.5icd	1.47	0.72	3.24	3.87	1.32	3.20	2.18	3.61	3.01	3.28	1.90	1.98	2.48	1.08	0.72	2.70
ICT.1ide	1.39	1.07	3.21	3.63	1.62	2.81	2.09	3.74	2.86	3.21	1.79	2.34	2.20	1.24	1.07	2.54
OXS.1idc	1.91	3.30	3.43	3.59	3.45	3.41	3.50	3.68	3.16	3.40	2.03	3.81	3.92	3.46	3.19	3.00
ICT.1gro	1.53	0.33	3.28	3.84	3.31	3.04	2.14	3.64	2.92	3.24	1.89	2.78	3.00	0.74	0.33	2.67
ICT.1grp	2.91	0.69	3.31	3.91	2.95	3.19	2.22	2.39	2.96	3.30	1.89	2.88	3.10	0.97	0.69	2.70
ICT.8icd	1.67	1.24	3.20	3.78	3.02	3.11	2.72	3.49	2.91	2.05	1.91	2.13	3.06	1.24	0.39	2.70

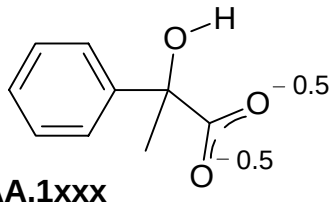
alle Ränge	FlexE	FlexX														
		MRGD	1idf	1idd	6icd	7icd	1ika	4icd	5icd	9icd	1ide	1idc	1gro	1grp	8icd	1iso
ICT.5icd	1.12	0.71	1.70	3.44	1.07	1.20	1.45	2.29	0.81	1.67	1.49	1.75	1.37	0.71	0.72	1.80
ICT.1ide	1.01	0.83	1.72	3.41	1.32	0.83	1.35	1.94	1.33	1.53	1.30	2.15	1.49	1.17	0.90	1.23
OXS.1idc	1.68	1.50	2.47	3.36	2.25	2.89	2.53	1.62	1.55	2.26	1.76	1.90	2.03	2.93	1.50	2.88
ICT.1gro	1.05	0.33	1.81	3.34	0.89	1.21	1.32	2.05	0.80	1.36	1.44	1.70	1.32	0.38	0.33	1.81
ICT.1grp	0.89	0.59	1.62	3.58	1.04	1.23	1.42	2.11	0.82	1.52	1.44	1.72	1.37	0.59	0.69	1.78
ICT.8icd	1.00	0.39	1.89	3.43	1.10	1.25	1.38	2.10	0.81	1.42	1.49	1.71	1.29	0.48	0.39	1.90

A.7 Mandelat-Racemase





SMN.1mdl



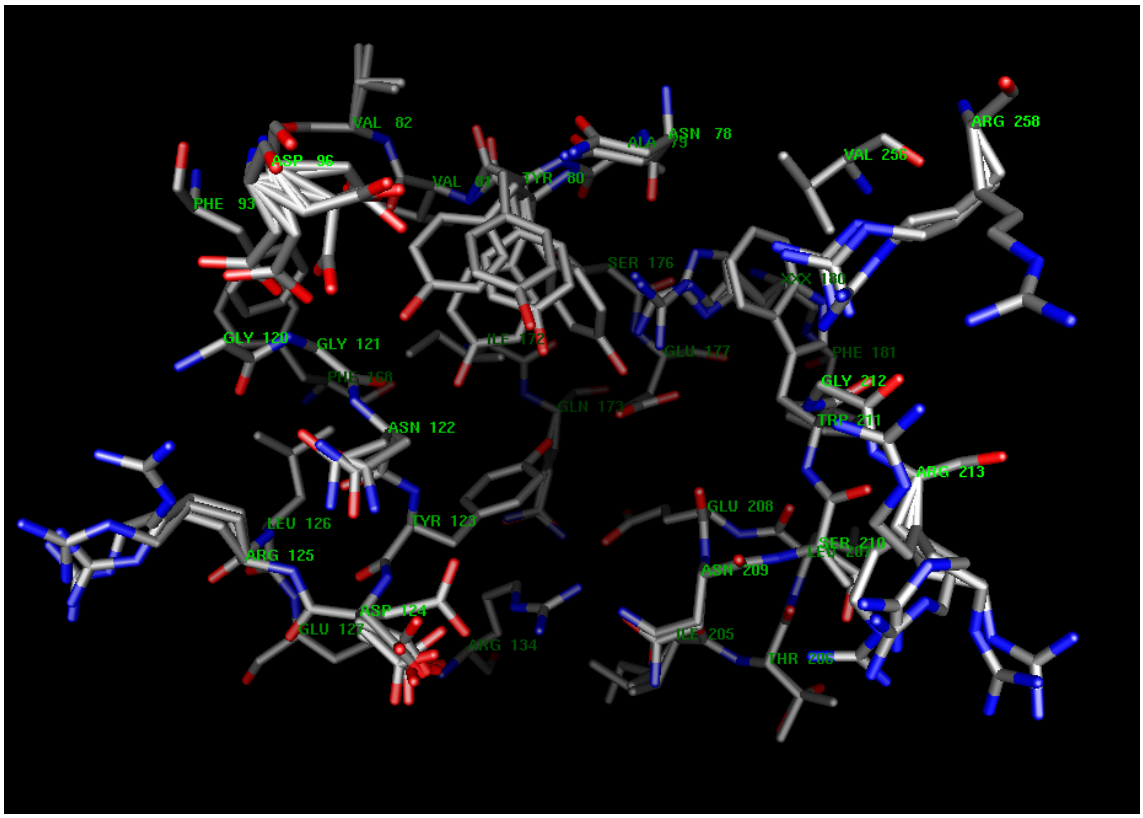
SAA.1xxx

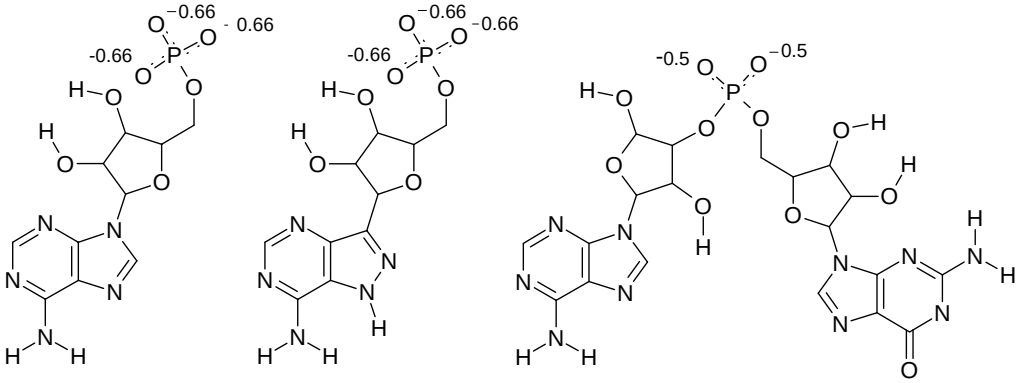
1. Rang	FlexE	MRGD	2mnr	FlexX 1mdr	1mns	1mdl	1mra	1dtn
SAA.1mdr	1.85	1.16	2.48	1.43	1.93	1.16	0.81	1.25
SMN.1mdl	2.54	2.31	2.02	1.31	1.30	1.02	1.23	2.31
SAA.1mra	1.08	1.28	2.45	1.56	1.93	1.28	0.80	1.77
SAA.1dtn	1.95	1.06	2.51	1.67	1.92	1.06	0.70	1.03

10 Ränge	FlexE	MRGD	2mnr	FlexX 1mdr	1mns	1mdl	1mra	1dtn
SAA.1mdr	0.99	0.68	1.53	0.64	1.89	0.95	0.48	1.25
SMN.1mdl	1.45	1.08	1.43	1.00	0.92	1.02	0.85	1.08
SAA.1mra	0.82	0.73	1.54	0.60	1.86	0.99	0.47	1.11
SAA.1dtn	0.86	0.70	1.62	0.76	1.88	0.92	0.64	1.03

alle Ränge	FlexE	MRGD	2mnr	FlexX 1mdr	1mns	1mdl	1mra	1dtn
SAA.1mdr	0.56	0.48	1.47	0.64	1.66	0.81	0.48	0.96
SMN.1mdl	0.83	0.50	1.04	0.88	0.92	0.62	0.61	0.50
SAA.1mra	0.54	0.47	1.49	0.60	1.73	0.85	0.47	0.93
SAA.1dtn	0.30	0.61	1.52	0.76	1.65	0.87	0.61	0.80

A.8 Ricin





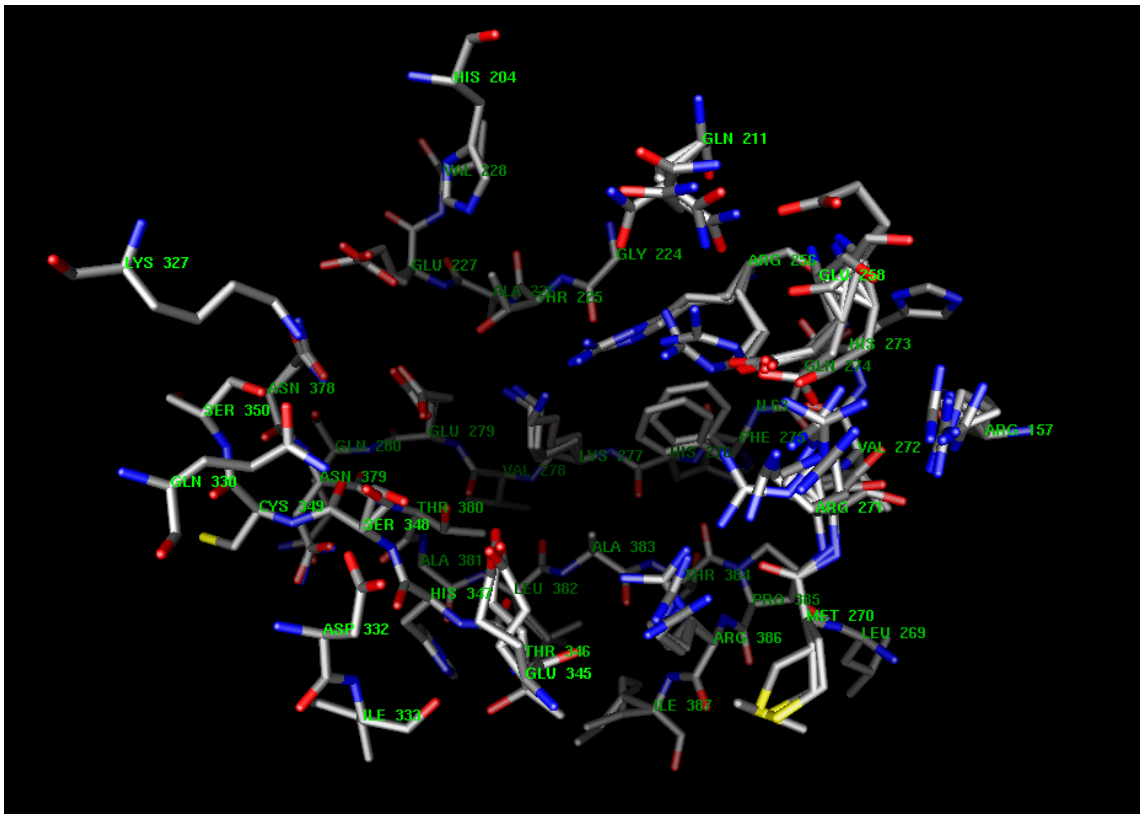
AMP.1obt **FMP.1fmp** **A-G.1apg**

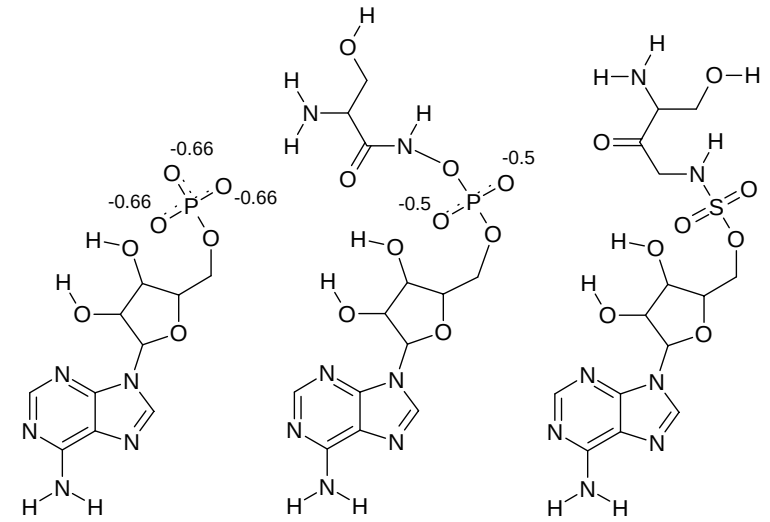
1. Rang	FlexE	FlexX									
		MRGD	1rtc	1obs	2aai	1fmp	1apg	1ifu	1ifs	1ift	1obt
FMP.1fmp	5.42	3.48	6.08	6.15	6.14	3.48	2.62	6.09	3.27	5.90	6.49
A−G.1apg	8.22	4.72	9.39	7.54	6.84	4.72	6.51	6.79	6.86	7.98	11.45
AMP.1obt	3.11	3.62	5.97	6.04	4.27	3.01	3.62	5.61	5.78	5.85	6.07

10 Ränge	FlexE	FlexX									
		MRGD	1rtc	1obs	2aai	1fmp	1apg	1ifu	1ifs	1ift	1obt
FMP.1fmp	2.71	2.19	5.36	3.30	5.48	2.19	1.31	3.30	1.78	5.62	6.00
A−G.1apg	4.44	4.57	8.35	6.12	4.05	4.57	3.66	5.64	6.72	6.72	4.84
AMP.1obt	3.11	3.01	5.40	4.37	4.08	2.93	2.90	4.47	5.08	4.53	5.54

alle Ränge	FlexE	FlexX									
		MRGD	1rtc	1obs	2aai	1fmp	1apg	1ifu	1ifs	1ift	1obt
FMP.1fmp	1.29	1.30	4.57	3.06	4.52	1.30	1.31	3.30	1.78	4.71	4.14
A−G.1apg	3.26	2.91	5.52	5.59	4.05	2.91	3.17	5.63	4.88	4.19	4.54
AMP.1obt	1.50	1.66	3.98	3.54	3.30	2.08	1.66	3.73	4.37	3.79	2.44

A.9 Seryl-T-RNA-Synthetase





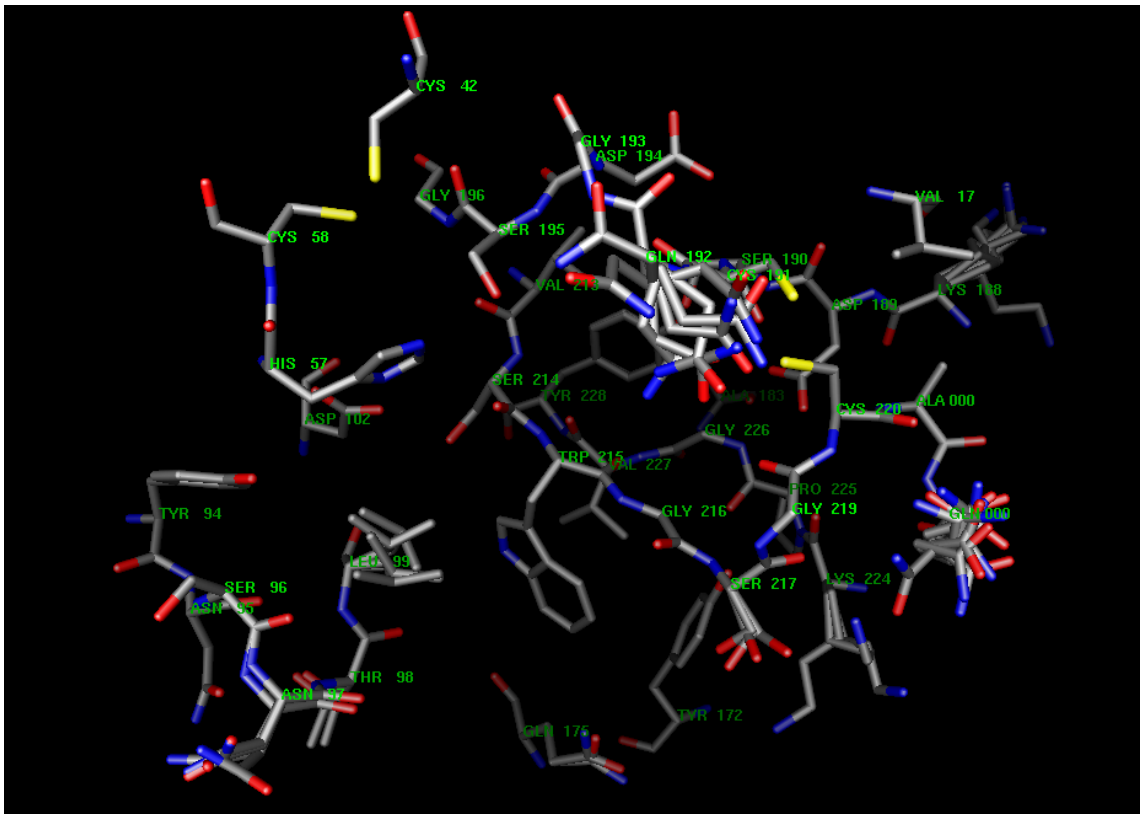
AMP.1ses **AHX.1ses** **SSA.1set**

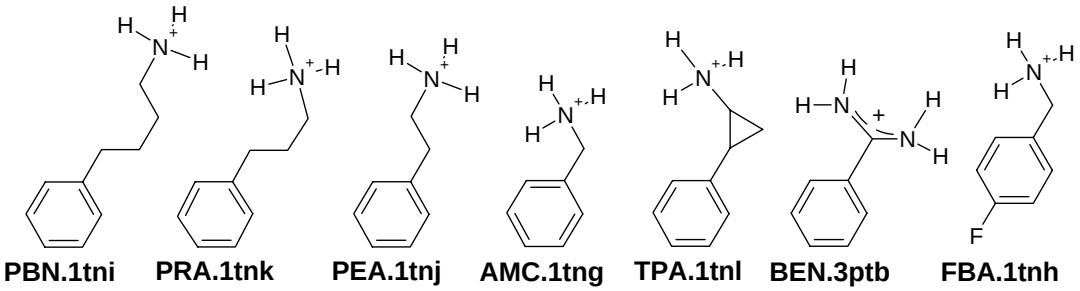
1. Rang	FlexE	MRGD	FlexX 1sry	FlexX 1sry	FlexX 1ses	FlexX 1ses	FlexX 1set	FlexX 1set
AHX.1sesA	3.28	7.93	8.05	7.48	5.78	7.93	4.05	8.49
AMP.1sesB	5.87	5.67	10.99	10.81	5.67	5.90	8.05	5.70
SSA.1setA	2.35	2.62	5.63	2.72	4.17	8.14	2.62	3.14
SSA.1setB	2.54	1.41	10.12	7.12	4.67	8.26	5.05	1.41

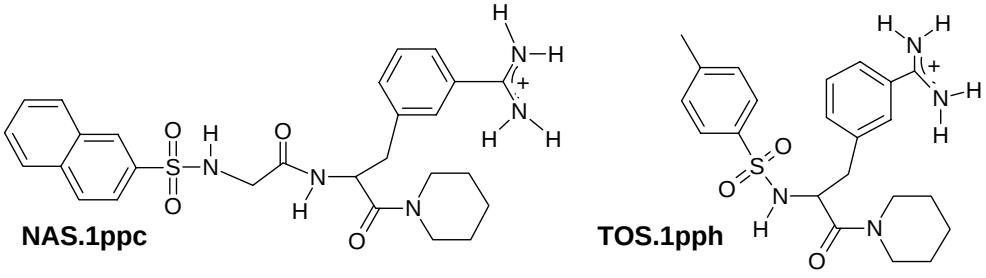
10 Ränge	FlexE	MRGD	FlexX 1sry	FlexX 1sry	FlexX 1ses	FlexX 1ses	FlexX 1set	FlexX 1set
AHX.1sesA	2.62	4.05	6.24	3.03	4.93	3.10	2.59	2.80
AMP.1sesB	2.48	5.56	5.58	5.88	5.65	1.24	2.89	5.70
SSA.1setA	1.95	2.47	5.37	2.72	4.17	7.73	2.47	3.14
SSA.1setB	2.28	1.41	7.09	6.44	3.68	6.58	4.94	1.39

alle Ränge	FlexE	MRGD	FlexX 1sry	FlexX 1sry	FlexX 1ses	FlexX 1ses	FlexX 1set	FlexX 1set
AHX.1sesA	2.46	2.48	4.02	2.74	3.10	3.10	2.48	2.80
AMP.1sesB	2.45	1.12	3.31	2.85	2.50	1.12	2.68	3.12
SSA.1setA	1.92	1.72	3.77	2.72	4.17	1.72	2.44	3.14
SSA.1setB	2.12	1.39	3.92	3.40	3.49	3.06	3.67	1.39

A.10 Trypsin


CYS 42
GLY 193
ASP 194
GLY 196
SER 195
CYS 58
GLN 192
SER 190
VAL 213
VAL 17
LYS 188
ASP 189
HIS 57
SER 214
TYR 228
CYS 220
ALA 000
ASP 102
GLY 226
TRP 215
VAL 227
PRO 225
GLY 219
GLY 216
TYR 94
SER 96
SER 217
LYS 224
THR 98
TYR 172


PBN.1tni
PRA.1tnk
PEA.1tnj
AMC.1tng
TPA.1tnl
BEN.3ptb
FBA.1tnh


NAS.1ppc
TOS.1pph

1. Rang

	FlexE	FlexX																
		MRGD	3ptn	2ptc	1tld	1tpo	1taw	1max	1ppc	1pph	1tng	1tnh	1tni	1tnj	1tnk	1tnl	1tpp	3ptb
NAS.1ppc	3.04	2.72	6.04	3.38	2.87	2.81	2.88	1.23	1.42	2.72	2.50	2.26	1.88	1.87	7.86	1.70	1.91	3.38
TOS.1pph	3.75	5.48	4.23	4.37	2.96	3.72	5.48	4.14	2.88	5.12	5.91	4.58	5.98	5.22	4.56	5.09	3.92	5.15
AMC.1tng	0.70	0.78	0.97	0.37	1.68	1.65	1.26	1.85	1.57	0.88	1.94	0.72	0.61	0.99	1.65	1.13	0.69	0.78
FBA.1tnh	0.55	0.45	0.70	0.37	0.45	0.80	0.70	0.70	0.45	0.41	0.90	0.54	0.54	0.59	0.43	0.53	1.10	0.63
PBN.1tni	0.96	2.37	2.79	2.79	3.18	3.09	3.17	2.85	2.91	3.49	2.50	2.65	2.34	2.37	2.24	2.26	3.27	2.45
PEA.1tnj	1.86	1.84	0.73	1.03	0.86	1.77	1.03	0.97	2.46	2.57	1.24	1.34	0.99	1.84	1.78	1.35	2.75	2.00
PRA.1tnk	1.97	1.92	1.74	1.46	1.55	1.64	1.92	1.49	1.49	1.98	1.65	1.57	1.91	1.31	1.77	1.41	2.05	2.05
TPA.1tnl	0.93	1.46	1.07	0.69	1.02	0.95	0.80	1.41	0.75	1.29	0.81	0.54	0.80	1.46	1.48	0.84	1.12	4.22
BEN.3ptb	0.63	0.52	3.78	0.77	0.42	0.64	0.52	0.67	0.33	0.31	0.54	0.87	0.61	0.76	0.78	0.76	0.64	0.63

10 Ränge

	FlexE	FlexX																
		MRGD	3ptn	2ptc	1tld	1tpo	1taw	1max	1ppc	1pph	1tng	1tnh	1tni	1tnj	1tnk	1tnl	1tpp	3ptb
NAS.1ppc	1.98	1.42	4.50	1.61	1.09	1.38	1.85	1.23	1.08	1.36	1.53	1.47	1.18	1.13	7.84	1.11	1.01	1.67
TOS.1pph	3.75	5.46	4.23	4.09	1.26	1.21	5.46	3.53	2.66	2.88	1.25	4.03	1.08	4.38	1.36	4.03	3.56	3.55
AMC.1tng	0.70	0.37	0.51	0.37	0.49	0.56	1.26	0.86	0.95	0.88	0.53	0.72	0.61	0.87	0.54	0.60	0.69	0.78
FBA.1tnh	0.48	0.43	0.64	0.37	0.45	0.39	0.65	0.69	0.45	0.41	0.52	0.40	0.54	0.47	0.20	0.53	0.72	0.63
PBN.1tni	0.96	2.34	2.08	2.53	1.28	2.40	2.80	2.21	1.30	2.52	0.80	0.81	2.09	2.15	0.56	2.23	2.16	2.30
PEA.1tnj	0.61	0.73	0.73	0.58	0.86	0.55	0.98	0.90	1.61	1.73	1.18	0.79	0.96	0.58	1.21	0.73	1.27	1.54
PRA.1tnk	0.57	1.31	1.57	1.26	1.06	1.49	1.45	1.49	1.24	1.71	0.63	0.78	0.83	1.05	0.83	1.41	1.06	1.55
TPA.1tnl	0.89	0.54	1.07	0.64	0.84	0.75	0.80	1.33	0.66	1.03	0.77	0.54	0.73	0.60	0.77	0.84	1.12	2.17
BEN.3ptb	0.27	0.32	1.26	0.41	0.32	0.37	0.52	0.67	0.33	0.21	0.53	0.56	0.58	0.70	0.60	0.64	0.42	0.42

alle Ränge

	FlexE	FlexX																
		MRGD	3ptn	2ptc	1tld	1tpo	1taw	1max	1ppc	1pph	1tng	1tnh	1tni	1tnj	1tnk	1tnl	1tpp	3ptb
NAS.1ppc	1.10	0.73	4.11	1.07	1.05	0.76	1.41	0.73	1.08	1.36	0.94	0.89	1.11	1.13	7.84	1.11	0.99	1.09
TOS.1pph	1.03	1.05	1.22	1.21	1.26	1.12	2.34	3.02	2.06	1.92	1.24	1.26	1.06	1.07	1.17	1.05	2.25	1.05
AMC.1tng	0.70	0.37	0.51	0.37	0.49	0.56	1.26	0.67	0.57	0.62	0.53	0.57	0.54	0.52	0.54	0.60	0.69	0.78
FBA.1tnh	0.48	0.20	0.64	0.37	0.45	0.39	0.65	0.69	0.45	0.41	0.52	0.40	0.54	0.47	0.20	0.53	0.60	0.63
PBN.1tni	0.84	0.56	1.32	1.94	1.23	0.90	1.56	1.34	1.30	1.62	0.80	0.81	1.00	0.74	0.56	1.00	1.23	1.89
PEA.1tnj	0.61	0.55	0.73	0.58	0.86	0.55	0.98	0.90	1.19	1.41	0.62	0.77	0.69	0.58	0.79	0.63	1.22	1.49
PRA.1tnk	0.57	0.63	1.55	0.84	0.66	0.68	0.93	1.17	0.92	1.33	0.63	0.71	0.79	0.76	0.69	0.71	1.06	0.98
TPA.1tnl	0.89	0.54	1.07	0.64	0.84	0.71	0.80	1.33	0.66	1.03	0.77	0.54	0.73	0.60	0.77	0.84	1.12	2.17
BEN.3ptb	0.27	0.21	1.26	0.41	0.32	0.37	0.52	0.67	0.33	0.21	0.53	0.56	0.58	0.70	0.60	0.64	0.42	0.42

Literaturverzeichnis

[1] C. Debouck and P.N. Goodfellow. Dna microarrays in drug discovery. *Nature Genetics*, 21:48–50, 1999.

[2] P.O. Brown and D. Botstein. Exploring the new wordl of the genome with dna microarrays. *Nature Genetics*, 21:33–37, 1999.

[3] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[4] J.C. Venter et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[5] H.-J. Böhm, Klebe G., and Kubinyi H. *Wirkstoffdesign*. Spektrum Akademischer Verlag GmbH, Heidelberg, 1996.

[6] G.S. Sittampalam, S.D. Kahl, and W.P. Janzen. High-troughput screening: advances in assy technologies. *Curr. Opin. Chem. Biol.*, 1:384–391, 1997.

[7] Hugo Kubinyi, editor. *3D QSAR in Drug Design. Theory, Methods and Applications*. ESCOM Science Publishers, Leiden, 1993.

[8] H.-D. Höltje and W. Sippl, editors. *Rational Approaches to Drug Design*. Prous Science, Barcelona, 2001.

[9] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, A32:922–923, 1976.

[10] M. Rarey. Protein-ligand docking in drug design. In T. Lengauer, editor, *Bioinformatics - From Genomes to Drugs*. Wiley-VCH, Heidelberg, 2001.

[11] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261(3):470–489, 1996.

[12] M. Vieth, J.D. Hirst, B.N. Dominy, H. Daigler, and C.L. Brooks III. Assessing search strategies for flexible docking. *J. Comput. Chem.*, 19:1623–1631, 1998.

[13] J. Falbe and M. Regitz, editors. *Römpp Chemie Lexikon*. Thieme, 1995.

[14] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, Inc., New York, 1991.

[15] P. Ehrlich. Address in pathology on chemotherapeutics: Scientific principles, methods, and results. *Lancet II*, pages 445–451, 1913.

[16] Hugo Kubinyi. Der Schlüssel zum Schloß: I. Grundlagen der Arzneimittelwirkung. *Pharmazie in unserer Zeit*, 23(3):158–168, 1994.

[17] D.E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.*, 44:98–104, 1958.

[18] R.B. Silverman. *Medizinische Chemie für Organiker, Biochemiker und pharmazeutische Chemiker*. VCH, Weinheim, 1994.

[19] C.A. Lipinski, F. Lombardo, B.W. Dominy, and P.J. Feeney. Experimental and computational approaches to estimate solubility and permeability. *Adv. Drug Deliv. Rev.*, 23:3–25, 1997.

[20] M.D. Miller, Kearsley S.K., D.J. Underwood, and R.P. Sheridan. FLOG: A system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Des.*, 8:153–174, 1994.

[21] S.K. Kearsley, D.J. Underwood, R.P. Sheridan, and M.D. Miller. Flexibases: A way to enhance the use of molecular docking methods. *J. Comput.-Aided Mol. Des.*, 8:565–582, 1994.

[22] D.M. Lorber and B.K. Shoichet. Flexible ligand docking using conformational ensembles. *Protein Sci.*, 7:938–950, 1998.

[23] R.L. DesJarlais, R.P. Sheridan, J.S. Dixon, I.D. Kuntz, and R. Venkataraghavan. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.*, 29:2149–2153, 1986.

[24] B. Sandak, R. Nussinov, and H.J. Wolfson. 3-D flexible docking of molecules. In A. Califano, editor, *Shape and Pattern Matching in Computational Biology: Procedings of IEEE Workshop 1994, Seattle, W*, pages 41–54. Plenum Press, 1996.

[25] B. Sandak, R. Nussinov, and H.J. Wolfson. An automated computer vision and robotics-based technique for 3-D flexible biomolecular docking and matching. *Comput. Appl. Biosci.*, 11(1):87–99, 1995.

[26] J.B. Moon and W.J. Howe. Computer design of bioactive molecules: A method for receptor-based de novo ligand design. *Proteins*, 11:314–328, 1991.

[27] A.R. Leach and I.D. Kuntz. Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comput. Chem.*, 13:730–748, 1992.

[28] W. Welch, J. Ruppert, and A.N. Jain. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. & Biol.*, 3:449–462, 1996.

[29] S. Makino and I.D. Kuntz. Automated Flexible Ligand Docking Method and Its Application for Database Search. *J. Comput. Chem.*, 18, 1997.

[30] M. Rarey, B. Kramer, and T. Lengauer. Multiple automatic base selection: Protein-ligand docking based on incremental construction without manual intervention. *J. Comput.-Aided Mol. Des.*, 11:369–384, 1997.

[31] G. Jones, P. Willett, and R.C. Glen. Molecular Recognition of Receptor Sites using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.*, 245:43–53, 1995.

[32] G. Jones, P. Willett, R.C. Glen, A.R. Leach, and R.Taylor. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.*, 267:727–748, 1997.

[33] C.M. Oshiro, I.D. Kuntz, and J.S. Dixon. Flexible ligand docking using a genetic algorithm. *J. Comput.-Aided Mol. Des.*, 9:113–130, 1995.

[34] K.P. Clark and Ajay. Flexible ligand docking without parameter adjustment across four ligand-receptor complexes. *J. Comput. Chem.*, 16(10):1210–1226, 1995.

[35] D.K. Gelhaar, G. Verkhivker, P.A. Rejto, D.B. Fogel, L.J. Fogel, and S.T. Freer. Docking conformationally flexible small molecules into a protein binding site through evolutionary programming. In J.R. McDonnell, R.G. Reynolds, and D.B. Fogel, editors, *Proceedings of the Fourth Annual Conference on Evolutionary Programming*, pages 615–627, 1995.

[36] D.K. Gehlhaar, G.M. Verkhivker, P.A. Rejto, C.J. Sherman, D.B. Fogel, L.J. Fogel, and S.T. Freer. Molecular recognition of the inhibitor ag-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. & Biol.*, 2:317–324, 1995.

[37] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19:1639–1662, 1998.

[38] C.A. Baxter, C.W. Murray, D.E. Clark, D.R. Westhead, and M.D. Eldridge. Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins*, 33:367–382, 1998.

[39] L. David, R. Luo, and M.K. Gilson. Ligand-receptor docking with the mining minima optimizer. *J. Comput.-Aided Mol. Des.*, 15:157–171, 2001.

[40] A.K. Ghose and G.M. Crippen. Geometrically feasible binding modes of a flexible ligand molecule at the receptor site. *J. Comput. Chem.*, 6(5):350–359, 1985.

[41] M. Billeter, T.F. Havel, and I.D. Kuntz. A new approach to the problem of docking two molecules: The ellipsoid algorithm. *Biopolymers*, 26:777–793, 1987.

[42] A.S. Smellie, G.M. Crippen, and W.G. Richards. Fast drug-receptor mapping by site-directed distances: A novel method of predicting new pharmacological leads. *J. Chem. Inf. Comput. Sci.*, 31:386–392, 1991.

[43] J.E.B. Platzer, F.A. Momany, and H.A. Scheraga. Conformational energy calculations of enzyme-substrate interactions. *Int. J. Pept. Protein Res.*, 4:187–200, 1972.

[44] J.E.B. Platzer, F.A. Momany, and H.A. Scheraga. Conformational energy calculations of enzyme-substrate interactions. *Int. J. Pept. Protein Res.*, 4:201–219, 1972.

[45] A. Di Nola, D. Roccatano, and H.J.C. Berendsen. Molecular dynamics simulation of the docking of substrates to proteins. *Proteins*, 19:174–182, 1994.

[46] B.A. Luty, Z.R. Wasserman, P.F.W. Stouten, C.N. Hodge, M. Zacharias, and J.A. McCammon. A molecular mechanics/grid method for evaluation of ligand-receptor interactions. *J. Comput. Chem.*, 16(4):454–464, 1995.

[47] J.A. Given and M.K. Gilson. A hierarchical method for generating low-energy conformers of a protein-ligand complex. *Proteins*, 33:475–495, 1998.

[48] T.N. Hart and R.J. Read. A multiple-start monte carlo docking method. *Proteins*, 13:206–222, 1992.

[49] R. Abagyan, M. Totrov, and D. Kuznetsov. ICM – a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.*, 15(5):488–506, 1994.

[50] A. Wallqvist and D.G. Covell. Docking enzyme-inhibitor complexes using a preference-based free-energy surface. *Proteins*, 25:403–419, 1996.

[51] C. McMartin and R.S. Bohacek. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.*, 11:333–344, 1997.

[52] J. Apostolakis, A. Plückthun, and A. Caflisch. Docking Small Ligands in Flexible Binding Sites. *J. Comput. Chem.*, 19:21–37, 1998.

[53] J.Y. Trosset and H.A. Scheraga. Reaching the global minimum in docking simulations: A Monte Carlo energy minimization approach using Bezier splines. *Proc. Natl. Acad. Sci. U.S.A.*, 95:8011–8015, 1998.

[54] D.S. Goodsell and A.J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins*, 8:195–202, 1990.

[55] Ajay, M.A. Murcko, and P.F.W. Stouten. Recent advances in the prediction of binding free energy. In P.S. Charifson, editor, *Practical Application of Computer-Aided Drug Design*, pages 355–410. Marcel Dekker, 1997.

[56] J.D. Hirst. Predicting ligand binding energies. *Curr. Opin. Drug Discovery and Development*, 1:28–33, 1998.

[57] M. Vieth, J.D. Hirst, A. Kolinsky, and C.L. Brooks III. Assessing energy functions for flexible docking. *J. Comput. Chem.*, 19:1612–1622, 1998.

[58] R.M.A. Knegtel and P.D.J. Grootenhuis. Binding affinities and non-bonded interaction energies. *Perspectives in Drug Discovery and Design*, 9/10/11:99–114, 1998.

[59] J.R.H. Tame. Scoring functions: A view from the bench. *J. Comput.-Aided Mol. Des.*, 13:99–108, 1999.

[60] H.-J. Böhm and M. Stahl. Rapid empirical scoring functions in virtual screening applications. *Medicinal Chemistry Research*, 9:445–462, 1999.

[61] H.-J. Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.*, 8:243–256, 1994.

[62] R.D. Head, M.L. Smythe, T.I. Oprea, C.L. Waller, S.M. Green, and G.R. Marshall. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.*, 118:3959–3969, 1996.

[63] M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini, and R.P. Mee. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.*, 11:425–445, 1997.

[64] C.W. Murray, T.R. Auton, and M.D. Eldridge. Empirical scoring functions. ii. the testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of bayesian regression to improve the quality of the model. *J. Comput.-Aided Mol. Des.*, 12:503–519, 1998.

[65] Y. Takamutsu and A. Itai. A new method for predicting binding free energy between receptor and ligand. *Proteins*, 33:62–73, 1998.

[66] R. Wang, L. Liu, L. Lai, and Y. Tang. SCORE: A new empirical method for estimating the binding affinity of a protein-ligand complex. *J. Mol. Model.*, 4:379–394, 1998.

[67] M. Stahl and H.J. Böhm. Development of filter functions for protein-ligand docking. *J. Mol. Graph. and Model.*, 16:121–132, 1998.

[68] M. Stahl. Modifications of the scoring function in flexx for virtual screening applications. *Perspectives in Drug Discovery and Design*, 20:83–98, 2000.

[69] J.B.O. Mitchell, R.A. Laskowski, A. Alex, and J.M. Thornton. BLEEP – a potential of mean force describing protein-ligand interactions: I. generating the potential. *J. Comput. Chem.*, 20:1165–1177, 1999.

[70] J.B.O. Mitchell, R.A. Laskowski, A. Alex, M.J. Forster, and J.M. Thornton. BLEEP – a potential of mean force describing protein-ligand interactions: Ii. calculation of binding energies and comparison with experimental data. *J. Comput. Chem.*, 20:1177–1185, 1999.

[71] I.Muegge and Y.C.Martin. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *Journal of Medicinal Chemistry*, 42(5):791–804, 1999.

[72] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. Poster presented at the 12th. European QSAR Symposium in Kopenhagen, Denmark, September 22-27, 1998.

[73] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 295:337–356, 2000.

[74] G. Gohlke, M. Hendlich, and G. Klebe. Predicting binding modes and binding affinities and hot spots for protein-ligand complexes using a knowledge-based scoring function. *Perspectives in Drug Discovery and Design*, 20:115–144, 2000.

[75] M.J. Sippl. Calculation of conformational ensembles from potentials of mean force - An approach to knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883, 1990.

[76] M.J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures. *J. Comput.-Aided Mol. Des.*, 7:474–501, 1993.

[77] P.S. Charifson, J.J. Corkery, M.A. Murcko, and Walters W.P. Consensus Scoring: A Method for obtaining Improved Hit Rates from Docking Databases of Thress-Dimensional Structures into Proteins. *J. Med. Chem.*, 42:5100–5109, 1999.

[78] M. Stahl, M.and Rarey. Detaild analysis of scoring functions for virtual screening. *J. Med. Chem.*, ?:in press, 2001.

[79] B. Kramer, M. Rarey, and T. Lengauer. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins*, 37:228–241, 1999.

[80] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Jr. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.

[81] J.S. Dixon. Evaluation of the casp2 docking section. *Proteins*, Suppl 1:1(1):198–204, 1997.

[82] M. Rarey, S. Wefing, and T. Lengauer. Placement of medium-sized molecular fragments into active sites of proteins. *J. Comput.-Aided Mol. Des.*, 10:41–54, 1996.

[83] M. Rarey, B. Kramer, and T. Lengauer. Time-efficient docking of flexible ligands into active sites of proteins. In C. Rawlings et al., editor, *Proceedings of the Third International Conference on Intelligent Systems in Molecular Biology*, pages 300–308. AAAI Press, Menlo Park, California, 1995.

[84] M. Rarey. *Rechnergestützte Vorhersage von Rezeptor-Ligand-Wechselwirkungen*, volume 268 of *GMD-Bericht*. R.Oldenbourg Verlag, 1996.

[85] M. Rarey, B. Kramer, and T. Lengauer. Docking of hydrophobic ligands with interaction-based matching algorithms. *Bioinformatics*, 15:243–250, 1999.

[86] B. Kramer, M. Rarey, and T. Lengauer. Casp-2 experiences with docking flexible ligands using flexx. *Proteins*, Suppl 1:1(1):221–225, 1997.

[87] B. Kramer, G. Metz, M. Rarey, and T. Lengauer. Ligand docking and screening with FlexX. *Medicinal Chemistry Research*, 7/8:463–478, 1999.

[88] M. Rarey, B. Kramer, and T. Lengauer. The particle concept: Placing discrete water molecules during protein-ligand docking predictions. *Proteins*, 34(1):17–28, 1999.

[89] M. Rarey and T. Lengauer. A recursive algorithm for efficient combinatorial library docking. *Perspectives in Drug Discovery and Design*, 20:63–81, 2000.

[90] G. Klebe and T. Mietzner. A fast and efficient method to generate biologically relevant conformations. *J. Comput.-Aided Mol. Des.*, 8:583–606, 1994.

[91] F.H. Allen, S. Bellard, M.D. Brice, B.A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink-Peters, O. Kennard, W.D.S. Motherwell, J.R. Rodgers, and D.G. Watson. The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Cryst.*, B35:2331–2339, 1979.

[92] J. Gasteiger, C. Rudolph, and J. Sadowski. Automatic generation of 3d-atomic coordinates for organic molecules. *Tetrahedron*, 3:537–547, 1990.

[93] J. Sadowski and J. Gasteiger. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.*, 93:2567–2581, 1993.

[94] J. Sadowski, J. Gasteiger, and G. Klebe. Comparison of automatic three-dimensional model builders using 639 x-ray structures. *J. Chem. Inf. Comput. Sci.*, 34:1000–1008, 1994.

[95] H.-J. Böhm. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.*, 6:61–78, 1992.

[96] H.-J. Böhm. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput.-Aided Mol. Des.*, 6:593–606, 1992.

[97] G. Klebe. The use of composite crystal-field enviroments in molecular recognition and the de-novo design of protein ligands. *J. Mol. Biol.*, 237:221–235, 1994.

[98] M.L. Connolly. Analytical molecular surface calculation. *J. Appl. Crystallogr.*, 16:548–558, 1983.

[99] A. Urzhumtsev, F. Tête-Favier, A. Mitschler, J. Barbanton, L. Barth, P. Urzhumtseva, J.-F. Biellmann, Podjarny A.D., and Moras D. A 'specificity' pocket inferred from the crystal structure of the complexes of aldose reductase with the pharmaceutically important inhibitors tolrestat and sorbinil. *Structure*, 5:601–612, 1997.

[100] A.D. Podjarny, M. van Zandt, O. Krämer, and G. Klebe. Model of human aldose reductase in complex with zopolrestat. personal communication.

[101] D. K. Wilson, I. Tarle, J. M. Petrash, and F. A. Quicho. Refined 1.8 angstroms structure of human aldose reductase complexed with the potent inhibitor zopolrestat. *Proc. Natl. Acad. Sci. U.S.A.*, 90:9847–9851, 1993.

[102] E. Mutschler. *Arzneimittelwirkungen*. Wissenschaftliche Verlagsgesellschaft mbh, Stuttgart, 1987.

[103] O. Krämer, M. Böhm, M. Schlitzer, and G. Klebe. 3D OSAR Analysis in Case of a Flexible Protein: COSIMA Model for a Series of Aldose Reductase Inhibitors with Various Binding Modes, Poster. Poster presented at the 13th European Symposium on Quantitative Structure-Activity Relationships, Düsseldorf, Germany, 2000.

[104] D.E. Koshland and K.E. Neet. The catalytic and regulatory properties of enzymes. *Annu. Rev. Biochem.*, 37:359–410, 1968.

[105] S.H. Done, J.A. Brannigan, P.C.E. Moody, and R.E. Hubbard. Ligand-induced conformational changes in penicillin acylase. *J. Mol. Biol.*, 284:463–475, 1998.

[106] J.A. McCammon. Theory of biomolecular recognition. *Curr. Opin. Struct. Biol.*, 8:245–249, 1998.

[107] A. Varrot, M. Schülein, and G.J. Davies. Insights into ligand-induced conformational change in cel5a from *bacillus agaradhaerens* revealed by a catlytically actiove crystal form. *J. Mol. Biol.*, 297:819–828, 2000.

[108] R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman. Side-Chain Flexibility in Proteins Upon Ligand Binding. *Proteins*, 39:261–268, 2000.

[109] C.W. Murray, C.A. Baxter, and A.D. Frenkel. the sensitivity of the results of molecular docking to induced fit effects: Application to thrombin, thermolysin and neuraminidase. *J. Comput.-Aided Mol. Des.*, 13:547–562, 1999.

[110] T. Klabunde, H.M. Tetrassi, V.B. Oza, P. Raman, JU.W. Kelly, and J.C. Sacchettini. Rational design of potent human transtyretin amyloid desease inhibitors. *Nat. Struct. Biol.*, 7:312–321, 2000.

[111] A.M. Davis and S.J. Teague. Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. *Angew. Chem. Int. Ed.*, 38:736–749, 1999.

[112] M.Gerstein and W.G. Krebs. A database of macromolecular motions. *Nucleic Acids Res*, 26:4280–90, 1998.

[113] A. Šali and T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234:779–815, 1993.

[114] M.J. Bower, F.E. Cohen, and R.L. Dunbrack. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J. Mol. Biol.*, 267:1268–1282, 1997.

[115] M.J.E. Sternberg, P.A. Bates, L.A. Kelley, and R.M. MacCallum. Progress in protein structure prediction: assessment of casp3. *Curr. Opin. Struct. Biol.*, 9:368–373, 1999.

[116] A. Šali and J. Kuriyan. Challenges at the frontiers of structural biology. *Trends Biochem. Sci.*, 24:M20–M24, 1999.

[117] M. Vásquez. Modeling side–chain conformation. *Curr. Opin. Struct. Biol.*, 6:217–221, 1996.

[118] D.T. Jones. Protein structure prediction in the postgenomic era. *Curr. Opin. Struct. Biol.*, 10:371–379, 2000.

[119] A. Fiser, R. Kinh Gian Do, and A. Šali. Modeling of loops in protin structures. *Protein Sci.*, 9:1753–1773, 2001.

[120] V. Sobolev, R.C. Wade, G. Vriend, and M. Edelman. Molecular docking using surface complementarity. *Proteins*, 25:120–129, 1996.

[121] V. Schnecke and L.A. Kuhn. Virtual screening with solvation and ligand-induced complementarity. *Perspectives in Drug Discovery and Design*, 20:171–190, 2000.

[122] R.M.A. Knegtel, I.D. Kuntz, and C.M. Oshiro. Molecular Docking to Ensembles of Protein Structures. *J. Mol. Biol.*, 266:424–440, 1997.

[123] H.B. Broughton. A method for including protein flexibility in protein-ligand docking. *J. Mol. Graph. and Model.*, 18:247–257, 2000.

[124] A.R. Leach. Ligand Docking to Proteins with Discrete Side–chain Flexibility. *J. Mol. Biol.*, 235:345–356, 1994.

[125] J. Desment, I.A. Wilson, M. Joniau, M. de Maeyer, and I. Lasters. Computations of the binding of fully flexible peptides to proteins with flexible side chains. *FASEB*, 11:164–172, 1997.

[126] M. Totrov and R. Abagyan. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins*, Suppl 1:1(1):210–214, 1997.

[127] J.-Y. Trosset and H.A. Scheraga. PRODOCK: Software package for protein modeling and docking. *J. Comput. Chem.*, 20:412–427, 1999.

[128] M. Mangoni, D. Roccatano, and A. Di Nola. Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins*, 35:153–162, 1999.

[129] J.-Y. Trosset and H.A. Scheraga. Flexible docking simulations: Scaled collective variable monte carlo minimization approach using bezier splines, and comparison with a standard monte carlo algorithm. *J. Comput. Chem.*, 20:244–252, 1999.

[130] J. Wang, P.A. Kollman, and I.D. Kuntz. Flexible ligand docking: A multistep approach. *Proteins*, 36:1–19, 1999.

[131] Daniel Hoffmann, Bernd Kramer, Takumi Washio, Torsten Steinmetzer, Matthias Rarey, and Thomas Lengauer. Two-stage method for protein-ligand docking. *J. Med. Chem.*, 42:4422–4433, 1999.

[132] I.D. Kuntz. Structure-based strategies for drug design and discovery. *Science*, 257:1078–1082, 1992.

[133] I.D. Kuntz, E.C. Meng, and B.K. Shoichet. Structure-based molecular design. *Acc. Chem. Res.*, 27(5):117–123, 1994.

[134] W.C. Guida. Software for structure-based drug design. *Curr. Opin. Struct. Biol.*, 4:777–781, 1994.

[135] P.M. Colman. Structure-based drug design. *Curr. Opin. Struct. Biol.*, 4:868–874, 1994.

[136] H.-J. Böhm. Current computational tools for de novo ligand design. *Curr. Opin. Biotechnol.*, 7:433–436, 1996.

[137] P.S. Charifson and I.D. Kuntz. Recent Successes and Continuing Limitations in Computer Aided Drug Design. In P.S. Charifson, editor, *Practical Applications of Computer–Aided Drug Design*, pages 1–37. Marcel Dekker Inc., New York, 1997.

[138] H. Kubinyi. Structure-based design of enzyme inhibitors and receptor ligands. *Curr. Opin. Drug Discovery and Development*, 1:4–15, 1998.

[139] P.W. Finn and L.E. Kavraki. Computational approaches to drug design. *Algorithmica*, 25:347–371, 1999.

[140] P.J. Gane and P.M. Dean. Recent advances in structure-based rational design. *Curr. Opin. Struct. Biol.*, 10:401–404, 2000.

[141] A. Tropsha. Recent trends in computer-aided drug discovery. *Curr. Opin. Drug Discovery and Development*, 3:310–313, 2000.

[142] J.M. Blaney and J.S. Dixon. A good ligand is hard to find: Automated docking methods. *Perspectives in Drug Discovery and Design*, 1:301–319, 1993.

[143] T.P. Lybrand. Ligand-protein docking and rational drug design. *Curr. Opin. Struct. Biol.*, 5:224–228, 1995.

[144] R. Rosenfeld, S. Vajda, and C. DeLisi. Flexible docking and design. *Annu. Rev. Biophys. Biomol. Struct.*, 24:677–700, 1995.

[145] P. Bamborough and F.E. Cohen. Modeling protein-ligand complexes. *Curr. Opin. Struct. Biol.*, 6:236–241, 1996.

[146] T. Lengauer and M. Rarey. Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.*, 6:402–406, 1996.

[147] T.J.A. Ewing and I.D. Kuntz. Critical Evaluation of Search Algorithms for Automated Molecular Docking and Database Screening. *J. Comput. Chem.*, 18(9):1175–1189, 1997.

[148] D.J. Diller and C.L.M. Verlinde. A critical evaluation of several global optimization algorithms for the purpose of molecular docking. *J. Comput. Chem.*, 20:1740–1751, 1999.

[149] R.M.A. Knegtel, D.M. Bayada, R.A. Engh, W. von der Saal, V.J van Geerestein, and P.D.J. Grootenhuis. Comparison of two simplementations of the incremental construction algorithm in flexible docking of thrombin inhibitors. *J. Comput.-Aided Mol. Des.*, 13:167–183, 1999.

[150] S. Ha, R. Andreani, A. Robbins, and I. Muegge. Evaluation of docking/scoring approaches: A comparative study based on mmp3 inhibitors. *J. Comput.-Aided Mol. Des.*, 14:435–448, 2000.

[151] H.A. Carlson and J.A. McCammon. Accommodating Protein Flexibility in Computational Drug Design. *Mol. Pharmacol.*, 57:213–218, 1999.

[152] B. Sandak, R. Nussinov, and H.J. Wolfson. A method for biomolecular structural recognition and docking allowing conformational flexibility. *J. Comput. Biol.*, 5:631–654, 1998.

[153] V. Sobolev, T.M. Moallem, R.C. Wade, G. Vriend, and M. Edelman. Casp2 molecular docking predictions with the ligin software. *Proteins*, Suppl 1:1(1):210–214, 1997.

[154] R.C. Wade, V. Sobolev, A.R. Ortiz, and G. Peters. Computational approaches to modeling receptor flexibility upon binding: Application to interfacially activated enzymes. *NATO ASI Ser., Ser. E (Structure Based Desgin)*, 352:223–232, 1998.

[155] J. Desmet, M. DeMaeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.

[156] I. Lasters and J. Desmet. The fuzzy-end elimination theorem: Correctly implementing the side chsin placement algorithm based on the dead-end elimination theorem. *Protein Eng.*, 6:717–722, 1993.

[157] J. Desmet, M. De Maeyer, and I. Lasters. The "dead-end elimination" theorem: A new approach to the side chain packing problem. In K. Merz, Jr., and S. LeGrand, editors, *The Protein Folding Problem and Tertiary Structure Prediction*, pages 307–337, Bosten, 1994. Birkhauser.

[158] I. Lasters, M. De Maeyer, and J. Desmet. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Eng.*, 8(8):815–822, 1995.

[159] A.R. Leach and A.P. Lemon. Exploring the Conformational Space of Protein Side Chains Using Dead-End Elimination and the A* Algorithm. *Proteins*, 33:227–239, 1998.

[160] P.E. Hart, N.J. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. on SSC*, 4:100–114, 1968.

[161] N.J. Nilsson. *Principles of Aritifical Intelligence*. Springer-Verlag, Berlin, 1982.

[162] B. Sandak, H.J. Wolfson, and R. Nussinov. Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. *Proteins*, 32:159–174, 1998.

[163] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R.L. Langridge, and T.E. Ferrin. A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.*, 161:269–288, 1982.

[164] E.C. Meng, Gschwend D.A., J.M. Blaney, and I.D. Kuntz. Orientational sampling and rigid-body minimization in molecular docking. *Proteins*, 17:266–278, 1993.

[165] E.C. Meng, B.K. Shoichet, and I.D. Kuntz. Automated docking with grid-based energy evaluation. *J. Comput. Chem.*, 13(4):505–524, 1992.

[166] B.K. Shoichet, D.L. Bodian, and I.D. Kuntz. Molecular docking using shape descriptors. *J. Comput. Chem.*, 13(3):380–397, 1992.

[167] V. Schnecke, C.A. Swanson, E.D. Getzoff, J.A. Tainer, and L.A. Kuhn. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins*, 33:74–87, 1998.

[168] V. Schnecke and L.A. Kuhn. Flexible screening for molecules interacting with proteins. In M.F. Thorpe and P.M. Duxbury, editors, *Rigidity in Theory and Applications*, pages 385–400. Plenum Publishing, New York, 1999.

[169] V. Schnecke and L.A. Kuhn. Database screening for hiv protease ligand: The influence of binding-site conformations and representation on ligand selectivity. In T. Lengauer, R. Schneider, P. Bork, D. Brutlag, W.-W. Mewes, and R. Zimmer, editors, *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 242–251. AAAI Press, Menlo Park, CA, 1999.

[170] P. Koehl and M. Delarue. Application of a Self-consisten Mean Field Theory to Predict Protein Side-chains Conformations and Estimate Their Conformational Entropy. *J. Mol. Biol.*, 239:249–275, 1994.

[171] P. Koehl and M. Delarue. Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.*, 6:222–226, 1996.

[172] R.M. Jackson, H.A. Gabb, and M.J.E. Sternberg. Rapid refinement of protein interfaces incorporating solvation: Application to the docking problem. *J. Mol. Biol.*, 276:265–285, 1998.

[173] A.J. Olson. The autodock suite of programms. Lecture at the Bologna winter school: IN SILICO BIOMOLECULAR RECOGNITION, Bologna, 2001.

[174] N.A. Metropolis, A.W. Rosenbluth, N.M.. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *J. Chem. Phys.*, 21:1087–1097, 1953.

[175] R. Abagyan and M. Totrov. Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, 235:983–1002, 1994.

[176] M. Zacharias and H. Sklenar. Harmonic modes as variables to approximately account for receptor flexibility in ligand-receptor docking simulations: Application to dna minor groove ligand complex. *J. Comput. Chem.*, 20:287–300, 1999.

[177] J. Cherfils and J. Janin. Protein docking algorithms: simulating molecular recognition. *Curr. Opin. Struct. Biol.*, 3:265–269, 1993.

[178] J. Janin. Protein-protein recognition. *Prog. Biophys. Mol. Biol.*, 64:145–166, 1995.

[179] B.K. Shoichet and I.D. Kuntz. Predicting the structure of protein complexes: a step in the right direction. *Chem. & Biol.*, 3:151–156, 1996.

[180] M.J.E. Sternberg, H.A. Gabb, and R.M. Jackson. Predictive docking of protein-protein and protein-dna complexes. *Curr. Opin. Struct. Biol.*, 8:250–256, 1999.

[181] M.J. Betts and M.J.E. Sternberg. An analysis of conformational changes on protein-protein association implications: for predictive docking. *Protein Eng.*, 12:271–283, 1999.

[182] M. Totrov and R. Abagyan. Detailed ab initio prediction of lysozym-antibody complex with 1.6 *Å* accuracy. *Struct. Biol.*, 1(4):259–263, 1994.

[183] B.L. Stoddard and D.E. Koshland Jr. Molecular recognition analyzed by docking simulations: The aspartate receptor and isocitrate dehydrogenase from escheria coli. *Proc. Natl. Acad. Sci. U.S.A.*, 90:1146–1153, 1993.

[184] P.N. Palama, L. Krippakl, J.E. Wampler, and J.J.G. Moura. Bigger: A new (soft) docking algorithm for predicting protein interactions. *Proteins*, 39:372–384, 2000.

[185] H.-P. Lenhof. An algorithm for the protein docking problem. In D. Schomburg and U. Lessel, editors, *Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism*, pages 125–139. VCH Weinheim, 1995.

[186] H.-P. Lenhof. New contact measures for the protein docking problem. In C. Waterman et al., editor, *Proceedings of the First Annual International Conference on Computational Molecular Biology*, pages 182–191. ACM Press, 1997.

[187] E. Althaus, O. Kohlbacher, H.-P. Lenhof, and P. Müller. A combinatorial approach to protein docking with flexible side-chains. In R. Shamier, S. Miyano, S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceedings of the Fourth Annual International Conference on Molecular Biology*, pages 15–24. RECOMB 2000, ACM Press, New York, 2000.

[188] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, and I.A. Vakser. Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U.S.A.*, 89:2195–2199, 1992.

[189] I.A. Vakser. Protein docking for low–resolution structures. *Protein Eng.*, 8:371–377, 1995.

[190] I.A. Vakser. Evaluation of GRAMM Low-Resolution Docking Methodology on the Hemagglutinin-Antibody Complex. *Proteins*, Suppl. 1:226–230, 1997.

[191] H.A. Gabb, R.M. Jackson, and M.J.E. Sternberg. Modelling protein docking using shape complementarity, electrostatics, and biochemical information. *J. Mol. Biol.*, 272:106–120, 1997.

[192] M.J.E. Sternberg, H.A. Gabb, R.M. Jackson, G. Moont, and E. Querol. A computational system for modelling flexible protein-protein and protein-DNA docking. In *Proceedings of the International Conference on Inteligent Systems in Molecular Bilogy 98 (ISMB 98)*, pages 183–192, Cambridge, UK, 1998.

[193] Z. Weng, S. Vajda, and C. Delisi. Prediction of protein complexes using empirical free energy functions. *Protein Sci.*, 5:614–626, 1996.

[194] J.W. Ponder and F.M. Richards. Tertiary templates for proteins. *J. Mol. Biol.*, 193:775–791, 1987.

[195] Jr. R. L. Dunbrack and M. Karplus. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.*, 230:534–574, 1993.

[196] G. Chinea, G. Padron, R.W.W.Hooft, C.Sander, and G.Vriend. The use of position specific rotamers in model building by homology. *Proteins*, 23:415–421, 1995.

[197] J. Mendes, A.M. Baptista, M.A. Carrondo, and C.M. Soares. Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins*, 37:503–543, 1999.

[198] R.A. Engh and R. Huber. Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement. *Acta Cryst.*, A47:392–400, 1991.

[199] R. Diestel. *Graphentheorie*. Springer-Verlag, Heidelberg, 2000.

[200] R.B. Altman and M.Gerstein. Finding an average core structure: Application to the globins. In *Proc. Second Int. Conf. Intell. Sys. Mol. Biol.*, pages 19–27, Menlo Park, CA, 1994. AAAI Press.

[201] M.Gerstein and R.B. Altman. Average Core Structure and Variability Measures for Protein Families: Application to the Immunoglobulins. *J. Mol. Biol.*, 251:161–175, 1995.

[202] U. Dorndorf and E. Pesch. Fast clustering algorithms. *ORSA Journal on Computing*, 6(2):141–153, 1994.

[203] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., New York, 1973.

[204] B. Everitt. *Cluster Analysis*. Halsted Press, Devision of John Wiley & Sons, Inc., New York, 1980.

[205] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983.

[206] R. Boppana and M.M. Halldórsson. Approximating maximum independent set by excluding subgraphs. *Bit*, 32:180–196, 1992.

[207] A. Babel. A fast algorithm for the maximum weight clique problem. *Computing*, 52:31–38, 1994.

[208] D.R. Wood. An algorithm for finding a maximum clique in a graph. *Operat. Res. Lett.*, 21:211–217, 1997.

[209] H.L. Bodlaender, D.M. Thilikos, and K. Yamazaki. It is hard to know when greedy is good for finding independent sets. *Inf. Process. Lett.*, 61:101–106, 1997.

[210] F.S. Kuhl, G.M. Crippen, and D.K. Friesen. A combinatorial algorithm for calculating ligand binding. *J. Comput. Chem.*, 5(1):24–34, 1984.

[211] M.C. Lawrence and P.C. Davis. CLIX: A search algorithm for finding novel ligands capable of binding proteins of known three-dimensional structure. *Proteins*, 12:31–41, 1992.

[212] M.Y. Mizutani, N. Tomioka, and A. Itai. Rational automatic search method for stable docking models of protein and ligand. *J. Mol. Biol.*, 243:310–326, 1994.

[213] A. Brint and P. Willett. Algorithms for the identification of three-dimensional maximal common substructures. *J. Chem. Inf. Comput. Sci.*, 27:152–158, 1987.

[214] I. Koch, T. Lengauer, and E. Wanke. An algorithm for finding maximal common subtopologies in an set of protein sturctures. *J. Comput. Biol.*, 3:289–306, 1996.

[215] Gardiner E.J., P.J. Artymiuk, and P. Willett. Clique-detection algorithms for matching three-dimensional molecular structures. *J. Mol. Graph. and Model.*, 15:245–253, 1997.

[216] R. Samudrala and J. Moult. A Graph-theoretic Algorithm for Comparative Modeling of Protein Structure. *J. Mol. Biol.*, 279:287–302, 1998.

[217] G. Caporossi, D. Cvetkovi—'c, I. Gutman, and P. Hansen. Variable neighborhood search for extremal graphs. 2. finding graphs with extremal energy. *J. Chem. Inf. Comput. Sci.*, 39:984–996, 1999.

[218] C. Bron and J. Kerbosch. Finding all cliques of an undirected graph [H]. *Commun. ACM*, 16(9):575–577, 1973.

[219] E. Tomita, K. Imamatsu, Y. Kohata, and M. Wakatsuki. A simple and efficient branch and bound algorithm for finding a maximum clique with experimental evaluation. *Systems and Computers in Japan*, 28:60–67, 1997.

[220] P.M. Pardalos, J. Rappe, and M.G.C. Resende. An exact parallel algorithm for the maximum clique problem. In R. De Leone et al., editor, *High performance algorithms and software in nonlinear optimisation*, pages 279–300. Kluwer, 1999.

[221] H. Galeana-Sánchez and H.A. Rincón-Mejia. Independent sets which meet all longest paths. *Discrete Math.*, 152:141–145, 1996.

[222] B.K. Bhattacharya and D. Kaller. An O(m + n log n) algorithm for the maximum-clique problem in circular-arc graphs. *J. Algorithms*, 25:336–358, 1997.

[223] D. Kagaris and S. Tragoudas. Maximum weighted independent sets on transitive graphs and applications. *INTEGRATION*, 27:77–86, 1999.

[224] F. Gavril. Maximum wight independent sets and cliques in intersection graphs of filaments. *Inf. Process. Lett.*, 73:181–188, 2000.

[225] R. Battiti and M. Protasi. Reactive local search for the maximum clique problem. *Algorithmica*, 29:610–637, 2001.

[226] D.S. Hochbaum. Approximating clique and biclique problems. *J. Algorithms*, 29:174–2000, 1998.

[227] N. Alon, Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. In *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 594–598, 1999.

[228] M. Stahl, M. Rarey, and G. Klebe. Screeing of drug databases. In T. Lengauer, editor, *Bioinformatics - From Genomes to Drugs*. Wiley-VCH, Heidelberg, 2001.

[229] M. Hendlich. Databases for protein-ligand complexes. *Acta Cryst.*, D54:1178–1182, 1998.

[230] Tripos Associates, Inc., St. Louis, Missouri, USA. *SYBYL Molecular Modeling Software Version 6.x*, 1994.

[231] H. Claußen, C. Buning, M. Rarey, and T. Lengauer. FlexE: Efficient Molecular Docking Considering Protein Structure Variations. *J. Mol. Biol.*, in press.

[232] M. Rarey. *FlexV 1.5.0 User Guide*. GMD – German National Research Center for Information Technology, 1997.

[233] MDL Informations Systems, Inc., San Leandro, CA, USA. *ISIS_Draw Version 2.3*, 1990-2000.

[234] The MathWorks, Inc, Natick, MA, USA. *MATLAB Version 6.0.0.88 Release 12*, 1984-2000.