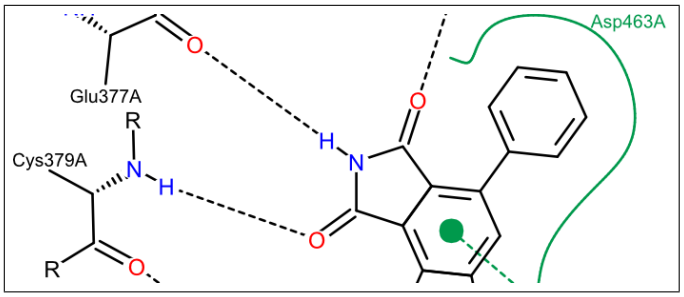
Computerbasierte Methoden zur automatischen Layoutberechnung von zweidimensionalen Darstellungen molekularer Strukturen



Kumulative Dissertation
zur Erlangung des akademischen Grades

Dr. rer. nat.

an der Fakultät
für Mathematik, Informatik und Naturwissenschaften der
Universität Hamburg

eingereicht beim Department Informatik von

Katrin Stierand
aus Aachen

Hamburg, Mai 2011

Erstgutachter: Prof. Dr. Matthias Rarey
Zweitgutachter: Dr. Werner Hansmann
Drittgutachter: Prof. Dr. Oliver Kohlbacher

Vom Department Informatik der Universität Hamburg als Dissertation angenommen
am 14.07.2011

Meinen Söhnen Kilian und Benjamin

Danksagung

Im Verlauf meiner Doktorarbeit habe ich von vielen Menschen Unterstützung erfahren. Ihnen spreche ich an dieser Stelle meinen Dank aus.

Mein besonderer Dank gilt Prof. Dr. Matthias Rarey, der mich als Doktorvater betreut hat. Durch ihn hatte ich die Möglichkeit, ein spannendes Thema zu bearbeiten und es national sowie international auf Konferenzen und in Publikationsn zu präsentieren. Darüberhinaus hat er mich in meinem Bemühen durch fachliche Anleitung und Diskussion immer sehr unterstützt.

Ich bedanke mich auch bei Dr. Werner Hansmann und Prof. Dr. Oliver Kohlbacher für das Begutachten meiner Dissertationsschrift.

Auch meine Kollegen und ehemaligen Kollegen am ZBH danke ich für das gute Arbeitsklima und die vielen anregenden Diskussionen. Dabei gilt mein besonderer Dank Karen Schomburg, die den SMARTSviewer entwickelt und implementiert hat, sowie Patrick Maaß und Matthias Hilbig für die Bereitstellung der Strukturdiagramm-Bibliothek.

Ein wichtiger Beitrag zur internationalen Präsentation meiner Arbeit wurde durch die Integration von PoseView-Bildern in die Website der RCSD PDB geleistet. Dafür bedanke ich mich bei Dr. Peter W. Rose und dem RCSB PDB Staff am San Diego Supercomputer Center.

Der BioSolveIT GmbH danke ich für die Bereitstellung der Flex*-Software-Bibliothek, die Integration von PoseView in LeadIT und die Hilfe beim Anfertigen der Software-Pakete für die PDB.

Das PoseView-Projekt wurde von der Klaus Tschira Stiftung gemeinnützige GmbH finanziert, dafür bedanke ich mich.

Für das Korrekturlesen meiner Dissertationsschrift und wertvolle inhaltliche Vorschläge bedanke ich mich bei meiner Schwester Vera Wimmenauer, bei Marcus Gastreich und bei Holger Schöning.

Mein Dank gilt nicht zuletzt meiner Familie für ihre Unterstützung zu jeder Zeit, allen voran meinen beiden Söhnen Kilian und Benjamin, die mir immer

den nötigen Raum gegeben haben, um diese arbeitsintensive Aufgabe zu erfüllen.

Kurzfassung

Die vorliegende Arbeit ist der allgemeine Teil meiner kumulativen Dissertationsschrift, eingereicht bei der Universität Hamburg im Mai 2011. Sie beschreibt einführend meine wissenschaftliche Tätigkeit von 08/2005 bis 03/2007 und von 05/2008 bis 12/2010 in der Abteilung für Algorithmisches Molekulares Design des Zentrums für Bioinformatik an der Universität Hamburg. Meine Dissertationsschrift besteht darüber hinaus aus fünf wissenschaftlichen Veröffentlichungen, die in einem gesonderten Literaturverzeichnis aufgelistet sind und mit der Bezeichnung D1 - D5 im Text referenziert werden. Aus Gründen des Urheberrechts sind sie jedoch nicht in die Dissertationsschrift eingebunden.

Die zweidimensionale Visualisierung von Protein-Ligand-Komplexen ermöglicht eine schnelle Übersicht über die Beschaffenheit des Wechselwirkungsmusters zwischen den interagierenden Molekülen. Während die computerbasierte Berechnung von Strukturdiagrammen kleiner Moleküle zu einem der ältesten Verfahren in der Chemieinformatik gehört und viele unterschiedliche Ansätze zur Verfügung stehen, gibt es nur wenige Lösungen für die automatische Generierung zweidimensionaler Darstellungen von Protein-Ligand-Komplexen. In der vorliegenden Arbeit wird mit PoseView ein solches Verfahren vorgestellt. Es sind sowohl Algorithmen zur Darstellung einzelner Komplexe als auch zur Darstellung multipler Komplexe, die Serien von unterschiedlichen Liganden gebunden an das gleiche Protein darstellen, entwickelt worden. Die Qualität der resultierenden Diagramme ist vergleichbar mit handgezeichneten Darstellungen in Fachbüchern und wissenschaftlichen Veröffentlichungen; groß angelegte quantitative Studien haben die Robustheit der implementierten Methoden nachgewiesen. PoseView wird in der RCSB Protein Database (PDB) als auch in der Software-Suite LeadIT zur Visualisierung von Komplexen verwendet. Zusätzlich wird es im Rahmen eines Webservices und als alleinstehendes Computerprogramm genutzt.

Abstract

The work in hand is the general part of my cumulative dissertation thesis, submitted to the University of Hamburg in May 2011. In an introductory way it describes my scientific work from 08/2005 to 03/2007 and from 05/2008 to 12/2010 at the Division of Algorithmic Molecular Design of the Center for Bioinformatics, University of Hamburg. The thesis is further composed of five scientific publications which are listed in a separate bibliography and cited in the text with D1 – D5. Due to copyright matters, they are not included in this manuscript.

The two-dimensional visualization of protein-ligand complexes provides a quick insight in the nature of the interaction pattern between the molecules. While the structure diagram computation of small molecules is one of the oldest problems in chemoinformatics and was solved in numerous different software tools, only very few approaches exist for the automatic generation of two-dimensional protein-ligand diagrams. Herein, PoseView, as one of these approaches, is presented. Within the software PoseView, algorithms for the visualization of single complexes as well as algorithms for the visualization of multiple complexes (series of ligands bound to the same protein) were developed. The quality of the resultant diagrams is comparable to hand-drawn examples from textbooks and scientific publications; large-scaled quantitative studies have proven the robustness of the implemented methods. PoseView is used for the two-dimensional visualization of protein-ligand complexes in the RCSB Protein Database (PDB) as well as in the software suite LeadIT. Additionally, it is used as a webservice and as a standalone tool.

Inhaltsverzeichnis

1

Einleitung

Die Suche nach neuen Medikamenten und deren Entwicklung findet heutzutage auf atomarer Ebene und damit im nicht mehr sichtbaren Bereich statt. Das detaillierte Wissen über die molekularen Strukturen und die Wechselwirkungen zwischen unterschiedlichen Molekülen ermöglichen jedoch gezielte Forschungsarbeit sowohl im Labor als auch am Computer [Kle09]. Ein großer Teil solcher Arbeit besteht aus der Suche nach kleinen Molekülen mit ungefähr zehn bis 50 Atomen, die Wechselwirkungen mit einem Protein ausbilden können. Diese kleinen Moleküle werden Wirkstoffe genannt [MGKSK01] und können durch das Ausbilden von (meistens reversiblen) Wechselwirkungen die Funktion der Proteine direkt – kompetitiv – oder indirekt – nicht-kompetitiv – beeinflussen [BTS02] und damit eine biologische Wirkung [MGKSK01] hervorrufen. Die Wirkstoffe können sowohl verstärkend als auch hemmend auf die Funktion des Proteins wirken. Vor allem in der frühen Phase der Medikamentenentwicklung, z.B. auf der Suche nach neuen Leitstrukturen für Wirkstoffe, ist der Computer schon seit einigen Jahrzehnten als Hilfsmittel etabliert [Kle09]. Für den computergestützten Wirkstoffentwurf (Computer Aided Drug Design, CADD) sind im Laufe der Jahre vielfältige Softwareprogramme und Algorithmen entwickelt worden [And03, BMG96, SLT09], die vor allem für die Suche nach hemmenden Wirkstoffen, sogenannten Inhibitoren, eingesetzt werden.

Die Resultate vieler computergestützter Verfahren sind eine Menge von Protein-Ligand-Komplexen – Wirkstoffe, die durch Wechselwirkungen an Proteine gebunden sind – deren Eigenschaften und Anzahl vorher spezifiziert wurden [MG08, Sho04]. Diese Komplexe werden vor ihrer weiteren Verwendung in anderen Computerprogrammen und Laborexperimenten häufig von Wissenschaftlern näher auf das Vorhandensein gewünschter Eigenschaften, z.B. die Ausbildung einer wichtigen Schlüsselwechselwirkung zwischen zwei bestimmten Atomen, und auf das Fehlen unerwünschter Eigenschaften wie Toxizität untersucht. Die Visualisierung der Komplexe stellt bei diesem Arbeitsschritt eine große Erleichterung für den Anwender dar. Einige ausgesuchte Modelle

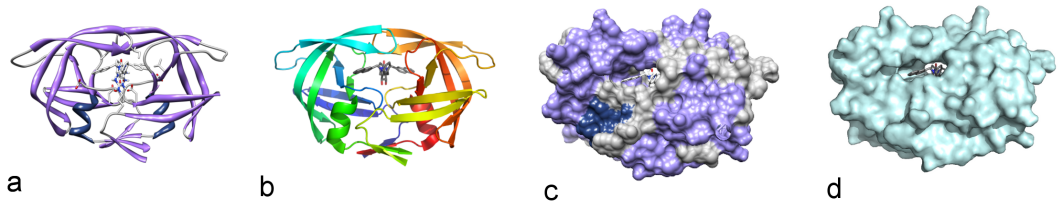


Abbildung 1.1: 3D Visualisierung einer HIV Protease (PDB ID: 1HVR [LJE+94]) in der Ribbon- und der Connolly-Oberflächendarstellung. Die Proteine wurden von zwei unterschiedlichen Programmen gezeichnet: Chimera [PGH+04] (a, c) und PyMol [Sch10] (b, d).

haben sich zur Darstellung etabliert und werden immer wieder verwendet, wie z.B. Visualisierungen der Connolly-Oberfläche [Con83] oder die Ribbon-Darstellung [Ric81], die das Erkennen der Sekundärstruktur eines Proteins erleichtern. Diese etablierten Darstellungen vereinfachen die Kommunikation und Analyse der Resultate stark, denn es besteht eine Ähnlichkeit zwischen den durch unterschiedliche Programme gezeichneten Protein-Ligand-Komplexen, siehe Abbildung 1.1.

Das Spektrum der Visualisierungsmöglichkeiten kann unterteilt werden in dreidimensionale (3D) und zweidimensionale (2D) Darstellungen. Während es für die Generierung von 3D Darstellungen molekularer Komplexe eine große Anzahl an unterschiedlichen Computerprogrammen gibt [OGF+10], ist die Zahl für 2D Darstellungen sehr gering [D5][ZTS09, ZS09]. Eine mögliche Erklärung dafür ist der Grad der Abstraktion: Für eine 3D Visualisierung ist es nicht notwendig, Koordinaten zu berechnen, sie sind durch die Originalstruktur, d.h. die Lage der Atome im Raum, vorgegeben: Hier ist es vielmehr Aufgabe, einen geeigneten Maßstab für die Darstellung und Modelle für die einzelnen Elemente (Atome, Bindungen und Wechselwirkungen) des Komplexes zu finden. Eine 2D Darstellung hingegen erfordert die Neuberechnung aller Koordinaten für ein möglichst kollisionsfreies, planares Layout und somit die Entwicklung geeigneter Algorithmen zur Visualisierung von Graphen unter Beachtung der Vorgaben, die durch den speziellen Anwendungskontext entstehen. Die Bedeutung der zweidimensionalen Visualisierung lässt sich durch die Verwendung von solchen – meist manuell angefertigten – Diagrammen in vielen wissenschaftlichen Veröffentlichungen [BBH+08, DR06, Kub98, LSBL+97] und Lehrbüchern [AS05, Kle09] belegen. Die Anwendung von Hochdurchsatzverfahren in experimentellen sowie in computergestützten Suchen nach Molekülen macht jedoch die Automatisierung des Zeichenprozesses erforderlich.

Über die 2D Darstellung kleiner Moleküle hinaus gibt es auch mehrere Ansätze, um

makromolekulare Strukturen, wie z.B. Proteine und RNA, zu zeichnen. Eine Übersicht über die verfügbaren Ansätze wird in einem Review von Zhou et al. [ZS09] gegeben. Aufgrund der Größe der zugrundeliegenden Strukturen muss der Grad der Abstraktion hier viel höher gewählt werden als bei kleinen Molekülen, um eine übersichtliche Anordnung zu gewährleisten. Der Vollständigkeit halber soll auch noch die planare Darstellung molekularer Netzwerke genannt werden, die aber, genau wie die Darstellung makromolekularer Strukturen, an dieser Stelle nicht weiter vertieft wird. Zu diesen Netzwerken gehören z.B. metabolische Netzwerke, die Stoffwechselwege visualisieren [SDMW09], als auch Wirkstoff-Zielprotein-Netzwerke [YGC+07, VM10], welche Wirkstoffe und die durch sie beeinflussten Strukturen zeigen.

Bei der Visualisierung von Protein-Ligand-Komplexen [WLT95, SMR06, CL07] sind zwar große Moleküle beteiligt, da aber nur die Liganden und die Anteile der Makromoleküle in deren direkter Nachbarschaft abgebildet werden, ist es möglich, dies in atomarer Auflösung zu tun. Bei den Komplexen geht es vor allem um die Darstellung der Beziehung des kleinen Moleküls zum Protein in Form von Wechselwirkungen.

In der vorliegenden Schrift werden Algorithmen zur automatischen Generierung zweidimensionaler Darstellungen von Protein-Ligand-Komplexen beschrieben, im Speziellen PoseView, dessen Algorithmen im Laufe dieser Arbeit entwickelt und implementiert wurden. Zur Zeit existieren drei veröffentlichte Ansätze zur Generierung zweidimensionaler Darstellungen von Protein-Ligand-Komplexen: Als erstes wurde im Jahr 1995 Ligplot publiziert [WLT95], eine Dekade später folgten die Veröffentlichungen von PoseView (2006) [SMR06] und dem 2D-Zeichner für Protein-Ligand-Komplexe eingebettet in die Softwaresuite MOE (2007) [CL07]. Alle drei Programme berechnen automatisch ein 2D Layout für Protein-Ligand-Komplexe für einen gegebenen Satz von 3D Koordinaten der Atome und den Bindungen zwischen ihnen in Form einer Zusammenhangstabelle; sie unterscheiden sich aber stark in der Wahl des zugrundeliegenden Verfahrens und in den resultierenden Layouts.

Im weiteren Verlauf dieses Kapitels werden zunächst Moleküle (Abschnitt 1.1) und dann intermolekulare Wechselwirkungen als wichtiges Konzept für das Zustandekommen von Protein-Ligand-Komplexen (Abschnitt 1.2) einführend erklärt. Anschließend werden noch einige Aspekte zur ästhetischen 2D Darstellung von Graphen erläutert (Abschnitt 1.3).

1.1 Moleküle

Moleküle bestehen aus *Atomen*, die durch *kovalente Bindungen* miteinander verknüpft sind [BTS02]. Aufgrund der kovalenten Bindungen gibt es eine feste topologische An-

ordnung der Atome im Molekül, jedoch mehrere mögliche räumliche Anordnungen der Atome zueinander. Die unterschiedlichen geometrischen Anordnungen werden als *Konformationen* [HK06] bezeichnet und entstehen im Wesentlichen durch Drehung von Einfachbindungen. Durch distanzabhängige Abstoßungs- und Anziehungskräfte zwischen den einzelnen Atomen des Moleküls gibt es energetisch günstige und ungünstige Konformationen; der Wechsel von einer günstigen Konformation in eine andere erfordert häufig die Überwindung einer Energiebarriere [BTS02].

Im computergestützten Wirkstoffentwurf werden hauptsächlich organische Moleküle mit einem Molekulargewicht von ca. 300 - 700 Da betrachtet. Ein weiteres, häufig verwendetes Kriterium für die Auswahl der Moleküle in diesem Forschungsbereich ist die 5er-Regel von Lipinski (engl.: Lipinski's Rule of Five) [LLDF97], mit der die Pharmakokinetik bzw. die ADME-Eigenschaften (Absorption, Distribution, Metabolisierung und Exkretion) des Wirkstoffes eingegrenzt werden sollen, um dessen orale Bioverfügbarkeit zu beschreiben. Dies macht eine Abschätzung möglich, ob eine neu gefundene Verbindung als oraler Arzneistoff dienen kann. Die 5er-Regel wurde von Ghose et al. [GVW99] im Jahr 1999 noch erweitert. Die Erweiterung legt optimale Wertebereiche für die Anzahl der funktionellen Gruppen und der Atome sowie die Molekülmasse und den Oktanol-Wasser-Verteilungskoeffizienten (logP) des Moleküls fest.

Als *Proteine* werden Moleküle bezeichnet, die aus 20 verschiedenen Untereinheiten, den *Aminosäuren*, zusammengesetzt sind [BTS02]. Aminosäuren sind kleine Moleküle, die alle einen gemeinsamen Teil haben, die sogenannten *Rückgratatome*. Diese sind im Protein miteinander verknüpft und bilden so sein *Rückgrat*. Die Reihenfolge der Aminosäuren in einer Kette wird als *Sequenz* bezeichnet. Darüber hinaus hat jede Aminosäure eine für sie charakteristische *Seitenkette*, die ihre chemischen und räumlichen Eigenschaften definiert. Die hier betrachteten Proteine werden aus 20 unterschiedlichen Aminosäuren (L-Aminosäuren) gebildet. Abbildung 1.2 zeigt diese 20 Aminosäuren und Abbildung 1.3 deren Verknüpfung, die Peptidbindung genannt wird. Proteine sind häufig recht große Moleküle, die aus mehreren hundert bis tausend Aminosäuren bestehen können.[1] Sie bestehen oft nicht nur aus einer durchgehenden Aminosäurekette, sondern entweder aus mehreren identischen (Homomultimere) oder unterschiedlichen (Heteromultimere) Ketten. Ihre Form und Funktion werden festgelegt durch die Sequenz der Aminosäuren in den Ketten. Sie sind ein wichtiger Bestandteil des Körpers und nehmen dort unterschiedliche Funktionen ein, wie z.B. Katalyse von Stoffwechselprozessen

[1] Häufig sind Metalle und/oder Kofaktoren an das Protein gebunden, die unerlässlich für seine Funktion sind. Ein Beispiel ist das Protein Hämoglobin, das als Co-Faktor ein Häm gebunden hat, welches wiederum mit einem Eisen-Ion verknüpft ist. Das Eisenion dient zum Sauerstofftransport im Blut. Verallgemeinernd werden diese Moleküle gemeinsam mit den Aminosäuren im Folgenden als *Residuen* des Proteins bezeichnet.

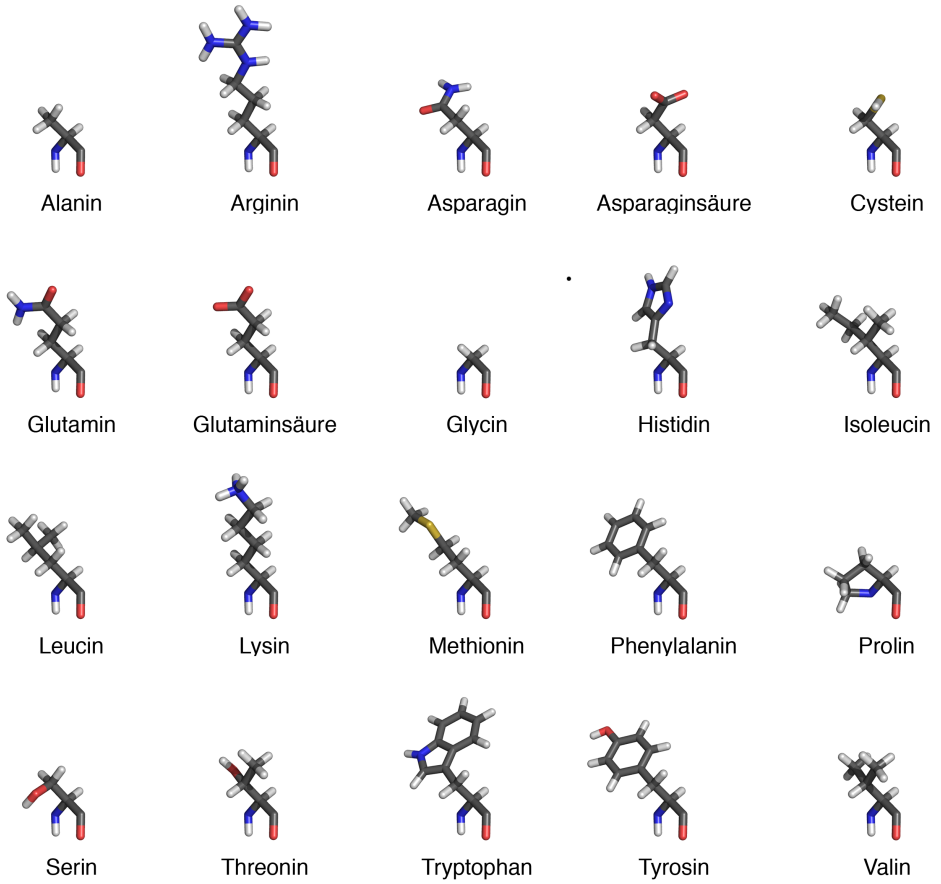


Abbildung 1.2: Darstellung der 20 L-Aminosäuren, gezeichnet mit PyMol [Sch10]. Die Aminosäuren sind alle gleich ausgerichtet, so dass die Seitenkette nach oben zeigt und die Rückgratatome unten zu finden sind.

(Enzyme), Weiterleitung von Informationen (Signalproteine), Bewegung (kontraktile Proteine), Regulation von Stoffwechselprozessen (regulatorische Proteine).

Im weiteren Verlauf wird ein kleines Molekül, das an ein Protein gebunden ist, als dessen *Ligand* bezeichnet. Das *aktive Zentrum* oder die *Bindetasche* des Proteins ist die Stelle, an der der Ligand gebunden ist.

1.2 Intermolekulare Wechselwirkungen

Die Funktion der Proteine ist häufig abhängig von der *Ausbildung von Komplexen* mit anderen Molekülen; es kommen sowohl Komplexbildungen mehrerer Proteine als auch Komplexe aus Proteinen und kleinen Molekülen vor, z.B. von Enzymen mit ihren *Substraten* – ein Substrat ist das Molekül, das im Rahmen der Katalyse umgewandelt wird

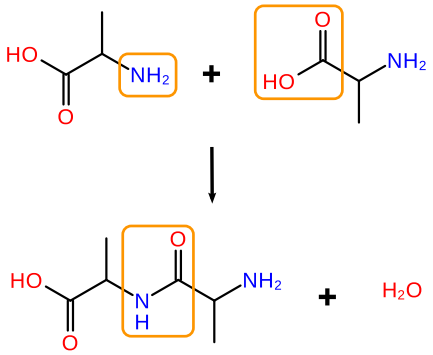


Abbildung 1.3: Bei der Ausbildung einer Peptidbindung wird ein Wasser freigesetzt. Dieses Beispiel zeigt die Verknüpfung von zwei Alaninmolekülen.

– oder mit Wirkstoffen. Die Komplexbildung findet im wässrigen Medium statt, in dem beide Moleküle zunächst solvatisiert vorliegen. Sie ist meist *reversibel* und beruht auf der Ausbildung von *Wechselwirkungen* (auch *Interaktionen* genannt) zwischen den beteiligten Molekülen [BTS02]. Eine wichtige Bedingung für das Zustandekommen eines Komplexes ist die *Komplementarität* der Moleküle[RMF98], sowohl *sterisch* als auch *physiko-chemisch*. Für den Fall, dass eine sterische Komplementarität erst bei der Komplexbildung durch Anpassung der räumlichen Struktur der Moleküle zustande kommt, spricht man von *Induced Fit* [KJ58].

Die treibende Kraft für die Bildung eines Komplexes aus zwei Molekülen ist der *hydrophobe Effekt* [BTS02]. Er beschreibt die Tendenz hydrophober Atome, sich im wässrigen Medium aneinander zu lagern, da der direkte Kontakt mit Wasser energetisch ungünstig ist. Im Gegensatz zu diesen ungerichteten hydrophoben Effekten gibt es noch einige gerichtete Wechselwirkungen, die die Ausrichtung der Moleküle zueinander definieren. Sie sollen im Folgenden erklärt werden. *Wasserstoffbrücken* [BTS02] werden ausgebildet zwischen einem Wasserstoffatom und einem elektronegativen Atom. Sie sind im Gegensatz zu den hydrophoben Kontakten stark gerichtet und bilden damit eine Sonderform der *elektrostatischen Wechselwirkungen*. Diese kommen zustande, wenn sich ein elektrisch geladenes Teilchen im Feld eines anderen befindet. Eine *ionische Wechselwirkung* entsteht auf kurzer Distanz zwischen zwei entgegengesetzt geladenen Atomen. Zusammenfassend werden die Wasserstoffbrücken und die elektrostatischen Wechselwirkungen als *hydrophile Wechselwirkungen* [BTS02] bezeichnet. π*-Interaktionen* spielen wie die ungerichteten hydrophoben Wechselwirkungen eine wichtige Rolle in der Protein-Ligand-Erkennung [MCD03]. Es gibt viele unterschiedliche Formen von π-Interaktionen [AD06], im Zusammenhang mit der vorliegenden Arbeit spielen lediglich Wechselwirkungen zwischen zwei π-Systemen und π-Kation-Wechselwirkungen eine Rolle. Bei Wechsel-

wirkungen zwischen zwei π-Systemen, z.B. Arylgruppen oder Heteroarylgruppen, gibt es unterschiedliche Ausrichtungen der Ringe zueinander [MGR98]: Entweder ein Ring steht senkrecht auf dem anderen oder beide Ringe sind parallel ausgerichtet. Im zweiten Fall sind die Ringzentren meistens parallel verschoben. Alle anderen Winkel der Ringe zueinander können auch auftreten, sind aber weniger häufig. Die π-Kation-Wechselwirkung kommt zustande durch die Anziehung zischen dem elektronenreichen π-System und dem positiv geladenen Ion. Zuletzt seien noch die *kovalenten Bindungen* [BTS02] genannt, die auch schon in Abschnitt 1.1 erwähnt wurden; sie führen zu der Bildung eines irreversiblen Komplexes und sind im Wirkstoffentwurf eher von untergeordnetem Interesse. Ein prominentes Beispiel für solch eine kovalente Wechselwirkung zwischen Arzneistoff und Protein ist die irreversible Acetylierung der Cyclooxygenase-1 (COX-1) durch Acetylsalicylsäure (ASS) [MGKSK01].

1.3 Ästhetik in der Graph-Visualisierung

Die Struktur von Molekülen, bestehend aus Atomen und kovalenten Bindungen, legt die Beschreibung durch Graphen nahe. Die Knoten repräsentieren dabei die Atome, und die Kanten die kovalenten Bindungen, die die Atome miteinander verknüpfen. Dadurch ist es möglich, für Graphen typische Datenstrukturen und Graphen-Algorithmen, wie z.B. die Tiefensuche [Cor01] oder den Dijkstra-Algorithmus zum Finden kürzester Pfade [Dij59], auf diese speziellen Fragestellungen anzuwenden. Die Graphen, die Wirkstoffe repräsentieren, sind vergleichsweise klein und haben einen niedrigen Knotengrad. Dadurch können auch Algorithmen mit hoher Komplexität verwendet werden.

Das Zeichnen von Graphen stellt ein eigenes Forschungsgebiet in der Informatik dar [gra] und hat ein breites Anwendungsfeld: So werden z.B. Netzwerke, Geschäftsprozesse oder Klassendiagramme als Graphen dargestellt. Es existieren verschiedene Zeichenkonventionen [DBETT98], die das Layout – die Anordnung von Knoten und Kanten im Graph – beeinflussen und einzeln oder kombiniert beim Zeichnen beachtet werden müssen. Die Konventionen legen unter anderem die Lage der Knoten (auf einem Gitter oder nicht), Planarität des Graphen und die Art der Kanten (Länge, Vorkommen von Winkeln) fest. Die Wahl eines geeigneten Zeichenalgorithmus [DBETT98] hängt unter anderem von der Kombination der oben genannten Konventionen ab.

Bedingt durch das strenge Regelwerk der IUPAC [Bre08] für die zweidimensionale Darstellung molekularer Strukturen (siehe Abschnitt 2.1) ist es im Fall des Zeichnens von Strukturdiagrammen nicht einfach möglich, die bereits bestehenden Graphzeichen-Algorithmen zu übernehmen. An dieser Stelle sollen jedoch einige Aspekte zur Ästhetik

in der zweidimensionalen Graph-Visualisierung erläutert werden, die auch beim Zeichnen von Strukturdiagrammen einbezogen werden.

Im Bezug auf Graph-Visualisierung wird von Ästhetik gesprochen, wenn eine Darstellung derart ist, dass die enthaltene Information für den Betrachter leicht erkennbar ist [BRSG07]. In diesem Fall spricht man von guter Lesbarkeit eines Graphen, die wiederum die Voraussetzung für das Verständnis dessen ist, was er darstellt. Ästhetik im künstlerischen Sinne ist nicht gemeint. Laut Ware et al. [WPCM02] teilt sich die Generierung eines Graphlayouts in einen syntaktischen (strukturellen) und einen semantischen (domänenspezifischen) Bereich. Der syntaktische Anteil adressiert das Layout des Graphgerüstes, z.B. Kantenlängenunterschiede und Kantenüberschneidungen; für diese Eigenschaften wurden viele empirische Untersuchungen durchgeführt, um Empfehlungen für Voraussetzungen eines guten Layouts geben zu können. Von diesen Empfehlungen werden im Folgenden einige genannt. Der semantische Anteil ist bisher weit weniger beforscht worden und betrifft vor allem die geeignete Darstellung des zugrundeliegenden Datensatzes und die Hervorhebung seiner wichtigen Eigenschaften durch z.B die Färbung und Beschriftung der Knoten oder das zentrale Platzieren wichtiger Knoten.

Zu den empirisch nachgewiesenen syntaktischen Eigenschaften eines Graphen [BRSG07], die eine gute Lesbarkeit erzeugen, gehören das *Gruppieren ähnlicher Knoten*, das *Minimieren von Kantenüberschneidungen*, das *Minimieren von Knicken in Kanten* und im Falle ihres Auftretens die *Einhaltung einer uniformen Platzierung* auf der Kante und eines immer *gleichen Winkels*, die *Orthogonalität der Kanten*, die *Minimierung des Gesamtbereichs*, die der Graph einnimmt (kompakte Darstellung) und die *Maximierung von lokaler Symmetrie*. Eigenschaften, die laut Bennet et al. [BRSG07] nicht empirisch belegt sind, aber in der Darstellung von Strukturdiagrammen zur Anwendung kommen, sind die Einhaltung eines *Minimalabstandes zwischen Knoten und Kanten*, das *Vermeiden von Knotenüberschneidungen* und *Uniformität bezüglich der Kantenlängen*. In Abschnitt 2.1 wird das Layout der Strukturdiagramme unter dem Gesichtspunkt der Graphästhetik diskutiert.

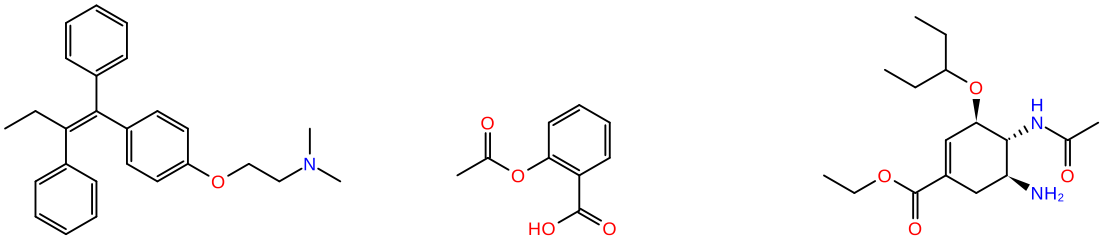
2

Zweidimensionale Visualisierung kleiner Moleküle und molekularer Muster

Die zweidimensionale Visualisierung von Molekülen, molekularen Komplexen und verwandten Strukturen ist ein wichtiges Werkzeug der Chemie. Sie ermöglicht dem Betrachter eine schnelle Übersicht über die dargestellten Eigenschaften der abgebildeten Strukturen.

Neben der Darstellung einzelner Moleküle spielt auch die Abbildung molekularer Fragmente und Muster eine wichtige Rolle, z.B. für die Beschreibung patentierter Molekülgruppen, in denen sich die einzelnen Strukturen häufig nur durch ihre funktionellen Gruppen unterscheiden, aber ein identisches Grundgerüst haben. Hierfür wird in der Regel die Markush-Notation verwendet [LBW81]. Weitaus häufiger werden jedoch Muster im Zusammenhang mit molekularen Substrukturen verwendet. Hier haben sich die SMARTS [Day08] etabliert, die von den SMILES [Wei88], einer eindimensionalen Molekülrepräsentation, abgeleitet sind. Die SMARTS-Sprache ist jedoch vor allem für die computerbasierte Verwendung optimiert und für den Menschen durch die Verschachtelung vieler geklammerter Ausdrücke nur schwer lesbar. Ein aktueller Ansatz zur zweidimensionalen Darstellung von SMARTS ist Teil dieser Dissertationsschrift [D1], siehe Abschnitt 2.2.

In den folgenden Abschnitten dieses Kapitels wird näher auf das Zeichnen von Strukturdiagrammen und die Visualisierung von SMARTS eingegangen, während die Darstellung von Protein-Ligand-Komplexen in späteren Kapiteln als Hauptthema dieser Arbeit eingehend beleuchtet wird.



Tamoxifen **Acetylsalicylsäure** **Oseltamivir**

Abbildung 2.1: Beispiele für Strukturdiagramme von bekannten Wirkstoffen [MGKSK01]: Der Arzneistoff Tamoxifen ist ein Estrogenrezeptomodulator und wird in der Brustkrebsbehandlung eingesetzt. Die Acetylsalicylsäure gehört in die Gruppe der Nichtsteroidalen Antirheumatika und wird unter anderem zur Entzündungshemmung und zur Vorbeugung von Gerinnungsstörungen eingesetzt. Oseltamivir ist ein Neuraminidasehemmer, der zur Behandlung von Grippe verwendet wird.

2.1 Strukturdiagramme

Die moderne 2D Darstellungsform eines Moleküls als Strukturdiagramm hat ihre Wurzeln in der Mitte des 19. Jahrhunderts [Sta58] und wurde durch die International Union of Pure and Applied Chemistry (IUPAC) standardisiert [Bre08]. Schon lange ist die Verwendung von Strukturdiagrammen Standard, sowohl in Lehrbüchern als auch in wissenschaftlichem Kontext, z.B. in Veröffentlichungen und in der täglichen Forschungsarbeit.

Ein Strukturdiagramm ist ein Graph, dessen Kanten den Bindungen und dessen Knoten den Atomen im Molekül entsprechen, für ein Beispiel siehe Abbildung 2.1. Den Konten werden 2D Koordinaten zugewiesen, sie werden nicht explizit gezeichnet, sondern sind implizit durch das Aufeinandertreffen zweier Kanten und dem daraus resultierenden Winkel dargestellt. Knicke innerhalb einer Kante sind nicht erlaubt. Auf den Knoten ist das Element des Atoms notiert, mit Ausnahme von Kohlenstoffen. Da Kohlenstoffe in organischen Molekülen von allen Schweratomen mit Abstand am häufigsten vorkommen, bedeutet eine fehlende Beschriftung, dass sich an dieser Stelle ein Kohlenstoff befindet. Wasserstoffatome, die an Kohlenstoffatome gebunden sind, werden häufig nicht eingezeichnet, da sich ihre Anzahl leicht durch die Anzahl der ausgehenden Bindungen zu Schweratomen an dem entsprechenden Kohlenstoffatom ergibt. Wasserstoffatome, die an ein Heteroatom gebunden sind, werden ohne eigene Bindung mit an das Schweratom geschrieben. Eine Ausnahme bilden die Wasserstoffatome, die an ein chirales Atom gebunden sind und die eingezeichnet werden müssen, um die Chiralität eindeutig darzustellen. Die unterschiedlichen Bindungstypen sind durch die Anzahl von parallelen Linien für eine Kante kodiert: Eine durchgezogene Linie bedeutet eine Ein-

fachbindung, zwei durchgezogene Linien eine Doppelbindung, drei durchgezogene Linien eine Dreifachbindung und eine durchgezogene Linie neben einer gestrichelten Linie das Vorhandensein einer delokalisierten Bindung. Räumliche Anordnungen, die die chemische Bedeutung beeinflussen, werden durch keilförmige Kanten repräsentiert: Gefüllte Keile symbolisieren eine Lage des Atoms vor der Bildeben und gestrichelte die Lage dahinter. Dies gilt immer für das Atom am breiten Ende der keilförmigen Bindung. Alle Kanten sollten nach Möglichkeit die gleiche Länge haben und das Grundmuster der Strukturdiagramme ist wabenförmig, so dass alle Winkel ein Vielfaches von 15° betragen, wobei 120° klar präferiert werden. Unter dem Gesichtspunkt der Graphästhetik (siehe Kapitel 1.3) werden viele Kriterien für ein ästhetisches Layout erfüllt: Es gibt faktisch keine Kantenüberschneidung, viele Kanten sind parallel, die Kantenlängen sind immer gleich, die Winkel zwischen den Kanten, die von einem Knoten ausgehen, sind gleich groß, und die Knoten sind uniform verteilt.[1]

Die computergestützte Berechnung von Strukturdiagrammen wurde bereits in den 1980er Jahren entwickelt und zählt damit zu den ältesten Anwendungen in der Chemieinformatik. Das Spektrum der heute existierenden Programme erstreckt sich von Editoren, mit denen Strukturdiagramme händisch aus gegebenen Bausteinen zusammengesetzt werden können, bis hin zu vollautomatisierten Verfahren [Hel99], die als Eingabe lediglich die Atom- und Bindungsinformation in Form einer Eingabedatei oder einer eindimensionalen Repräsentation verwenden. Letztere sind für das Zeichnen zweidimensionaler Darstellungen von Protein-Ligand-Komplexen mit PoseView von Interesse, und ein exemplarischer Layout-Algorithmus ist als Bestandteil dieser Dissertationsschrift in der Veröffentlichung “Flat and Easy: 2D Depiction of Protein-Ligand Complexes” [D5] beschrieben.

2.2 SMARTS-Visualisierung

Im Vergleich zur graphischen 2D Darstellung kleiner Moleküle sind Visualisierungsansätze für molekulare Muster weit weniger etabliert. Einige Strukturdiagramm-Editoren [IBB09, ST09] bieten eine Darstellung der SMARTS-Zeichenkette [Day08] als Strukturdiagramm an. Die Eigenschaften für die einzelnen Kanten und Knoten werden textuell als Beschriftung hinzugefügt. Dies führt zu Darstellungen, bei denen sehr viele Überschneidungen von Textelementen mit dem Graphen vorliegen.

[1] Bei komplexen Ringsystemen muss manchmal von den Vorgaben abgewichen werden, oft gibt es hier Vorgaben von der IUPAC, ansonsten ist eine Darstellung anzustreben, die möglichst vielen der oben genannten Kriterien genügt.

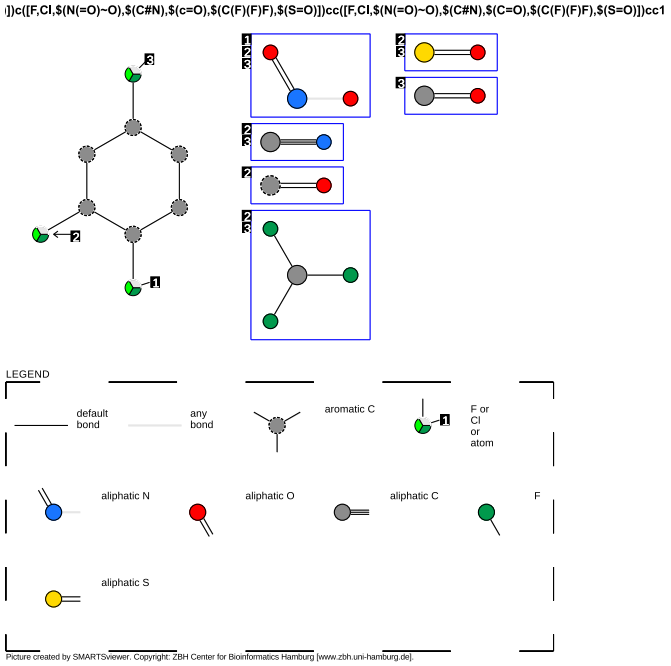


Abbildung 2.2: Visualisierung einer SMARTS-Zeichenkette (oben in der Abbildung) durch den SMARTSviewer

Mit dem SMARTSviewer [D1] wurde ein neues Konzept zur Darstellung erarbeitet und implementiert[1]. Aufgrund der hohen Akzeptanz und Verbreitung von Strukturdiagrammen wurde diese Darstellung als Grundgerüst gewählt. Die Aufgabe lag also weniger im strukturellen Bereich des Graphzeichnens (siehe Abschnitt 1.3) sondern viel mehr im domänenspezifischen Bereich, d.h. im geeigneten Design der einzelnen Knoten und Kanten. Die Zielsetzung war, alle Elemente der SMARTS-Sprache darstellen zu können, die Ästhetik der Strukturdiagramme zu erhalten und gleichzeitig den Lernaufwand für den Betrachter gering zu halten, ohne die Kenntnis der SMARTS-Sprache vorauszusetzen.

Um die SMARTS-Zeichenketten visualisieren zu können, werden diese zunächst mit Hilfe eines SMARTS-Parsers eingelesen und vorverarbeitet. An dieser Stelle besteht auch die Option, die Syntax überprüfen zu lassen und eventuelle Redundanzen zu entfernen (SMARTStrim). Strukturell entsprechen die resultierenden Diagramme den zuvor beschriebenen Strukturdiagrammen, die Differenz liegt in der Darstellung der Knoten

[1]Der SMARTSviewer wurde von Karen Schomburg im Rahmen ihrer Masterarbeit am ZBH entwickelt und von mir mitbetreut. Mein Beitrag zu dieser Arbeit bestand in der Bereitstellung der zugrundeliegenden Zeichenbibliothek. Diese war abgeleitet von der SDG-Bibliothek und wurde von mir für die gestellte Aufgabe angepasst. Desweiteren habe ich an den Diskussionen für das Layoutkonzept aktiv teilgenommen, Vorschläge eingebracht und die anschließende Implementierungsarbeit betreut.

und Kanten. Die zugrundeliegenden Strukturdiagramme sind in ihrer Beschaffenheit in der Regel relativ simpel, die Komplexität der Aufgabe besteht in der geeigneten Anordnung der unterschiedlichen Zeichenelemente für die Eigenschaften der Atome und Bindungen. Die Knoten werden als (meistens farbig gefüllte) Kreise gezeichnet, um den zu notierenden Informationen Platz zu bieten: Gängige Elemente der organischen Verbindungen (siehe Abschnitt 1.1) und Halogene werden durch Farben kodiert, Ladungen durch die Zahl in direkter Nachbarschaft zum Knoten und Valenzen als Punkte innerhalb des Kreises. Eine vollständige Übersicht ist in [D1] zu finden. Der logische Operator NICHT wird durch rote Farbe markiert, während das ODER durch Einzeichnen der Alternativen und die Farbe Blau visualisiert wird. Dieses Konzept ist auch bei den Kanten zu finden. Das Lesen der Diagramme wird vereinfacht durch die automatische Erzeugung einer dynamischen Legende, die die verwendeten Elemente und deren textuelle Beschreibung enthält. Zusätzlich ist eine statische Legende verfügbar.

Auch Rekursionen in den SMARTS-Zeichenketten, die die Umgebung eines Atoms näher definieren, können graphisch dargestellt werden, indem mehrere Diagramme gezeichnet werden. In Abbildung 2.2 ist ein Beispiel für die Visualisierung eines SMARTS gezeigt, die einige der erwähnten Zeichenelemente enthält.

Der SMARTSviewer wurde mit Hilfe eines Testdatensatzes, der aus 762 unterschiedlichen SMARTS-Zeichenketten besteht, validiert. Für jede der Zeichenketten konnte ein Diagramm generiert werden, wobei die einzelnen Zeichenketten stark in der Länge variierten (zwischen zwei und 1008 Zeichen).

3

Zweidimensionale Darstellung einzelner Protein-Ligand-Komplexe

Die zweidimensionale Visualisierung eines Protein-Ligand-Komplexes dient vor allem zur gut erkennbaren Darstellung der Beschaffenheit des Wechselwirkungsmusters zwischen den beteiligten Molekülen eines Komplexes. Deshalb ist es ausreichend, den Ligand und die Residuen in seiner direkten Umgebung abzubilden, das Auswahlkriterium für die zu zeichnenden Residuen kann entweder das Vorhandensein einer Wechselwirkung oder ein gewisser Abstand der Aminosäure zum Ligand sein. Es müssen also vor Beginn des Layoutalgorithmus die Wechselwirkungen in der originalen 3D Struktur berechnet werden. Die Darstellung der Aminosäuren kann im Detailgrad zwischen atomarer und molekularer Auflösung variieren. Wechselwirkungen werden, wenn sie gerichtet sind, als Linien dargestellt, während sie für den Fall, dass sie ungerichtet sind, implizit durch Markierung der beteiligten Ligandatome und Aminosäuren dargestellt werden. Das Ziel bei der Platzierung der einzelnen Elemente des Diagramms[1] ist, sie möglichst überlappungsfrei darzustellen.

Für die automatische Generierung zweidimensionaler Darstellungen von Protein-Ligand-Komplexen gibt es zur Zeit drei publizierte Verfahren: Ligplot [WLT95], PoseView und einen 2D Zeichner eingebettet in die Software-Suite MOE [CL07]. Eine Beschreibung dieser Verfahren wurde im Rahmen eines Reviews angefertigt und ist Teil dieser Dissertationsschrift [D5]. Dort wird sowohl auf die Algorithmen als auch qualitativ vergleichend auf die Ergebnisse der einzelnen Programme eingegangen. Die Heterogenität der unterschiedlichen Ansätze beruht sowohl auf dem gewählten Darstellungsmodus als auch auf den Algorithmen. Hier soll nur kurz der äußerlich sichtbare

[1] Die Elemente eines Diagramms sind die einzelnen Zeicheneinheiten, d.h. Moleküle, die Interaktionslinien, Beschriftungen usw.

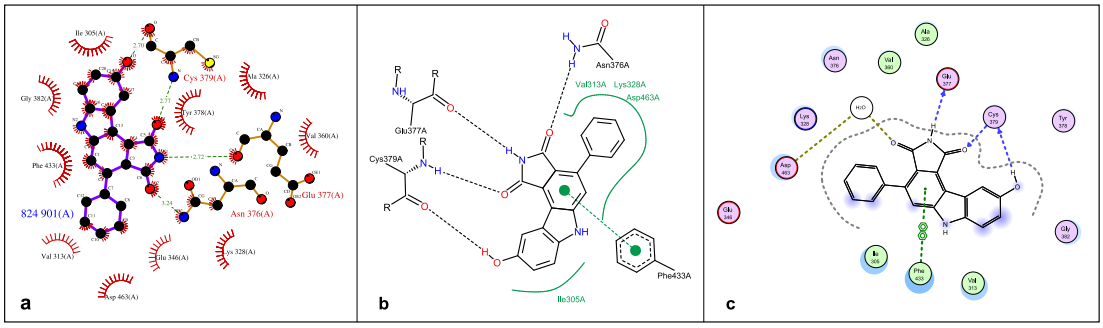


Abbildung 3.1: 2D Visualisierungen der Wee1A Kinase mit dem Inhibitor PD0407824, PDB ID: 1X8B [SDIB05]. Die Zeichnung a wurde mit Ligplot [WLT95] generiert, b mit PoseView [D2] und c mit MOE [CL07].

Unterschied erläutert werden, siehe Abbildung 3.1: Ligplot, als ältester der drei Ansätze, generiert Plots, in denen die Liganden und die durch hydrophile Wechselwirkungen an sie gebundenen Residuen auf atomarer Basis dargestellt werden, jedoch ohne Wasserstoffe und Ladungen. Die Diagramme der einzelnen Moleküle entsprechen nicht den Empfehlungen der IUPAC, da die Koordinaten aus den 3D Koordinaten abgeleitet werden und so ein unregelmäßiges Winkelmuster entsteht. In PoseView werden sowohl die Liganden als auch die Aminosäuren, die gerichtete Wechselwirkungen zum Ligand ausbilden, als Strukturdiagramme entsprechend den Empfehlungen der IUPAC gezeichnet. Der 2D Zeichner in MOE verzichtet auf der Proteinseite auf eine atomare Darstellung, um nicht nur wechselwirkende Residuen darstellen zu können, sondern alle Moleküle innerhalb eines gewissen Abstandes zum Liganden. Diese Moleküle sind in Form von farbig gefüllten Kreisen dargestellt.

Im Folgenden werden die im Rahmen dieser Arbeit entwickelten Algorithmen zur automatischen Berechnung eines 2D Layouts beschrieben. In der vorbereitenden Diplomarbeit wurde ein Prototyp entwickelt, der ein Layout für gegebene hydrophile Interaktionen und die wechselwirkenden Moleküle berechnen konnte. Dieser soll zunächst in Abschnitt 3.1 beschrieben werden. Der Prototyp wurde zunächst um ungerichtete hydrophobe Wechselwirkungen, wie in Abschnitt 3.2 beschrieben, erweitert und reimplementiert. Als nächster Schritt wurde ein eigenes Interaktionsmodell eingeführt (Abschnitt 3.3), das nun auch π-Wechselwirkungen enthält. In Abschnitt 3.4 werden dann die Resultate aus der fertiggestellten Software erläutert.

Der Fokus bei der Bewertung der Ergebnisse ist hauptsächlich auf die Qualität der resultierenden Diagramme gerichtet. Aufgrund eines fehlenden allgemeingültigen Qualitätsmaßes für alle existierenden Ansätze wurde folgendes Bewertungsschema eingeführt: Alle Diagramme, die kollisionsfrei gezeichnet werden können und mindestens eine wech-

selwirkende Aminosäure enthalten, werden als gut bezeichnet. Eine weitere Klasse wird aus Diagrammen gebildet, die verbesserbare Layouts haben, d.h. die prinzipiell kollisionsfrei gezeichnet werden könnten, bei denen jedoch die kollisionsfreie Anordnung der Zeichenelemente durch den Algorithmus nicht gefunden wurde. Desweiteren gibt es Anordnungen von Wechselwirkungen am Liganden, die in der Ebene mit den gegebenen Zeichenregeln nicht planar angeordnet werden können. Diese Kollisionen werden als unlösbar bezeichnet. Zwei weitere Kategorien werden aus Liganden gebildet, für die keine Wechselwirkungen mit dem gegebenen Protein gefunden wurden, und aus Komplexen, bei denen das Programm nicht erfolgreich ausgeführt werden konnte.

3.1 Vorarbeiten in der Diplomarbeit

Die Vorarbeiten in der Diplomarbeit stellen den Ausgangspunkt für die Entwicklung der aktuellen Version von PoseView dar. Der Stand der Arbeit zum Abschlusszeitpunkt der Diplomarbeit wurde 2006 publiziert [SMR06], und im Folgenden wird er zusammenfassend dargestellt, da dies für das Verständnis der weiteren Kapitel notwendig ist.

Die Berechnung des 2D Layouts geschieht unabhängig von den 3D Koordinaten der Eingabe. Als Grundlage wird lediglich die Zusammenhangstabelle der Molekülgraphen und die Wechselwirkungsinformation verwendet. PoseView wurde aufbauend auf die Flex*-Bibliothek [RKLK96] implementiert, so dass das Einlesen und das chemische Modell für die Moleküle übernommen werden konnte. Auch die hydrophilen Wechselwirkungen einer Dockinglösung oder einer Kristallstruktur wurden von FlexX berechnet und als Teil der Eingabe für PoseView verwendet. Die Strukturdiagramme aller Moleküle, die am Komplex beteiligt sind (im folgenden Ensemble genannt), werden unter Zuhilfenahme einer internen SDG-Bibliothek [FGR04] generiert.

Die generierten Komplexdiagramme sollen kollisionsfrei sein, d.h. es dürfen keine Überlappungen von Strukturdiagrammen, Überkreuzungen von Interaktionslinien und Überschneidungen von Kollisionslinien mit Strukturdiagrammen vorliegen. Ein Beispiel für ein Diagramm mit kollisionsfreiem Layout, das jedoch nicht mit dem Prototyp sondern mit der aktuellen PoseView-Version generiert wurde, ist in Abbildung 3.2 zu finden. Der gewählte Ansatz ist ligandzentriert, es wird also zunächst mit der Berechnung eines optimalen Layouts des Ligandstrukturdiagramms begonnen. Es besteht die Möglichkeit, durch Rotationen an Bindungen oder Austauschen von Bindungen, die vom selben Atom ausgehen, das Layout zu verändern. Dabei ist jedoch wichtig zu beachten, dass das resultierende Layout kollisionsfrei und chemisch korrekt bleibt, d.h. es dürfen z.B. nur Bindungen gedreht werden, die in der korrespondierenden Originalstruktur auch rotierbar sind. Um eine überschneidungsfreie Anordnung der Interaktionslinien zu gewährleisten,

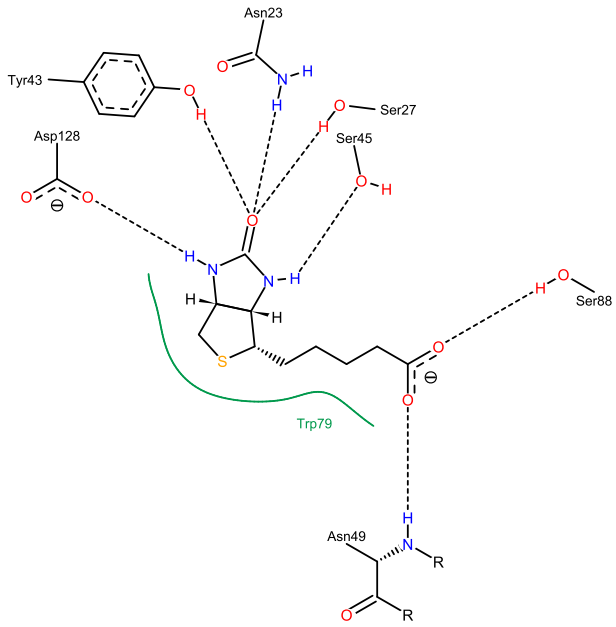


Abbildung 3.2: Das gezeigte von PoseView generierte Layout für den Komplex von Biotin mit Streptavidin, PDB ID: 1STP [WOWS89], ist kollisionsfrei. Die grün gezeichneten hydrophoben Kontakte sind nicht Teil des im Rahmen der Vorarbeiten entwickelten Prototyps.

ist es notwendig, Ligandatome, die Wechselwirkungen mit derselben Aminosäure ausbilden, direkt nebeneinander zu platzieren, ohne dass ein anderes Wechselwirkungsatom dazwischen liegt. Dafür wird zunächst die zirkuläre Anordnung der Wechselwirkungsatome um den Liganden herum berechnet; in Abbildung 3.3 ist die zirkuläre Anordnung durch einen Pfeil graphisch dargestellt. Falls die Anordnung keine überkreuzungsfreie Anordnung der Wechselwirkungslinien zulässt, werden in einem Enumerationsverfahren alle möglichen Kombinationen von Bindungsrotationen und -austauschen aufgezählt. Die resultierenden Ligandlayouts werden in Hinsicht auf die Anzahl der Wechselwirkungsüberkreuzungen bewertet. Die Kombination mit der besten Bewertung wird an die SDG-Bibliothek übergeben. Dort wird das Strukturdiagramm entsprechend modifiziert und, falls es kollisionsfrei gezeichnet werden kann, an PoseView zurückgegeben und als Grundlage für die weiteren Berechnungen verwendet. Die Enumeration kann unterbrochen werden, falls ein zeichenbares Layout gefunden wird, das keine Überkreuzungen der Wechselwirkungslinien verursacht.

Anschließend an die Ligandlayoutberechnung werden die Strukturdiagramme der Residuen initial platziert. Um eine möglichst kollisionsfreie Initialplatzierung zu finden, werden diese radial um den Liganden auf Basis der konvexen Hülle [Jar73] seiner

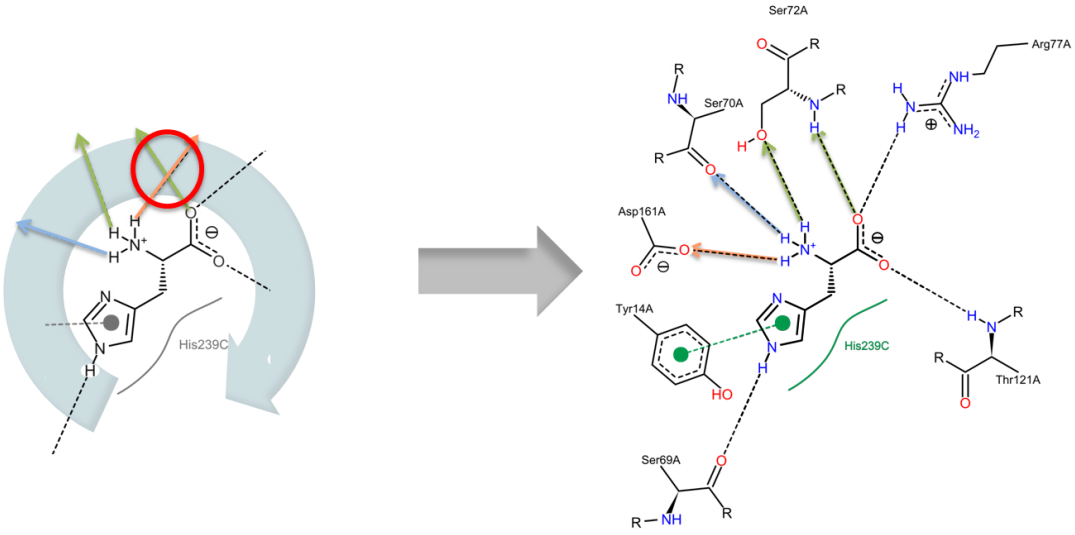


Abbildung 3.3: Ordnung der Wechselwirkungsatome beim Liganden bei initialem Layout und nach erfolgter Modifikation der Bindungen. Durch einen Austausch zweier Bindungen, zwischen den Wasserstoffatomen und dem Stickstoffatom kann ein Ligandlayout generiert werden, das eine überkreuzungsfreie Anordnung der Wechselwirkungslinien ermöglicht.

2D Atomkoordinaten angeordnet. Zunächst wird die Richtung der Wechselwirkungslinien berechnet. Dabei wird zwischen Interaktionslinien, die an einem Atom starten, das ein Eckpunkt der konvexen Hülle bildet, und solchen, die im Inneren der Hülle liegen, unterschieden. Im ersteren Fall wird die Richtung der Interaktion von den Bindungen des Wechselwirkungsatoms abgeleitet: Wenn eine einzelne Bindung zum Atom führt, so wird deren Richtung übernommen, wenn mehrere Bindungen zum Atom führen, wird der Wechselwirkungslinie die resultierende Richtung aus den Einzelrichtungen der Bindungen zugewiesen. Für die Wechselwirkungsatome im Inneren der Hülle wird zunächst auf alle Kanten das Lot gefällt und dann diejenige Kante ausgewählt, zu der der Abstand am kleinsten ist und keine Überschneidungen zwischen Lotstrecke und Ligandstrukturdiagramm entstehen. Die Wechselwirkungsrichtung entspricht dann der Richtung der Lotgeraden. Falls eine Aminosäure mehrere Wechselwirkungen zum Ligand ausbildet, wird für deren Platzierung eine resultierende Hauptrichtung aus den Einzelrichtungen berechnet. Die beschriebene Berechnung wird analog zur Ligandseite für jedes Aminosäurestrukturdiagramm durchgeführt. Anschließend werden die Wechselwirkungsrichtungen von Aminosäure- und Ligandseite entgegengesetzt überlagert. Die Länge der Hauptrichtungen, an deren Endpunkten die Aminosäurestrukturdiagramme platziert werden, entspricht fünf Standardbindungslängen in den Strukturdiagrammen.

Bedingt durch die Tatsache, dass jedes Aminosäurestrukturdiagramm unabhängig

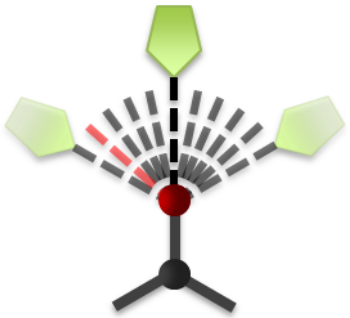


Abbildung 3.4: Mögliche Aminosäurepositionen, die im Rahmen der Kollisionsbehandlung betrachtet werden. Die rote Kugel symbolisiert das Wechselwirkungsatom auf Ligandseite und das grüne Polygon ein Aminosäurestrukturdiagramm.

von den anderen Diagrammen platziert wird, können Kollisionen entstehen. In einem Layoutoptimierungsschritt werden diese Kollisionen detektiert und wenn möglich entfernt. Es soll ein Gesamtlayout gefunden werden, das möglichst wenig von dem initialen Layout abweicht, jedoch kollisionsfrei ist. Dazu werden, ausgehend von der initialen Position, acht weitere alternative Positionen für jedes Aminosäurestrukturdiagramm vor berechnet, indem dieses zusammen mit der Wechselwirkungslinie viermal nach rechts und viermal nach links um jeweils 18° rotiert wird. Als Zentrum der Rotation dient dabei der Startpunkt der Wechselwirkungshauptrichtung auf der Ligandseite, siehe Abbildung 3.4. Das Entfernen der Kollisionen ist realisiert durch ein *Branch-and-Bound-Verfahren*, in dem die möglichen Kombinationen der Aminosäurepositionen aufgezählt und bewertet werden. Die Aufzählung der möglichen Kombinationen wird auf Basis eines k-nären Baumes ausgeführt, der mit Hilfe einer Tiefensuche traversiert wird. k entspricht dabei der Anzahl der unterschiedlichen Positionen, die für eine Aminosäure vorberechnet wurden und die Tiefes des Baumes der Anzahl der Aminosäuren im Diagramm. Im Verlauf der Tiefensuche wird auf jeder Ebene des Baumes eine neue Aminosäure hinzugefügt, deren Platzierung durch die vorher durchlaufene Kante festgelegt ist. Das neu platzierte Strukturdiagramm und die dazugehörenden Wechselwirkungslinien werden auf Kollisionen mit den bereits platzierten Elementen untersucht und das Layout dementsprechend bewertet. Die Bewertung am Blatt eines Baumes entspricht einer Bewertung des Gesamtlayouts. Da das Bewertungsschema additiv gewählt ist, ist es möglich, Subbäume von der Suche auszuschließen, wenn die Bewertung seiner Wurzel die bisher beste Bewertung des Gesamtlayouts überschreitet. Desweiteren kann der Algorithmus abgebrochen werden, wenn an einem Blatt eine kollisionsfreie Lösung gefunden wurde. Die geometrische Abweichung der einzelnen Strukturdiagramme von ihrer Initialplatzierung wird relativ zur Größe des Rotationswinkels bestraft, um eine Bevorzugung der Lösun-

gen mit kleinen Abweichungen sicherzustellen. Dies ist realisiert durch die Reihenfolge, in der die Kanten während der Tiefensuche durchlaufen werden. Diese wird festgelegt durch ein Bewertungsschema, in dem die Summe von zwei kleinen Abweichungen von der initialen Position zweier unterschiedlicher Aminosäurestrukturdiagramme niedriger bestraft wird als eine große Abweichung einer Aminosäure.

3.2 Erweiterung des Layoutalgorithmus

Die Implementation des in Abschnitt 3.1 beschriebenen Prototypen wurde überarbeitet und um neue Graphelemente, die hydrophoben Kontakte, erweitert. Diese Erweiterung wurde in einer wissenschaftlichen Veröffentlichung [D3], die Teil der kumulativen Schrift ist, beschrieben. Die oben genannten Neuerungen führten zu einer Verbesserung der Resultate, was einerseits auf die Einbeziehung von Komplexen, die nur durch hydrophobe Kontakte zwischen Ligand und Protein zustande kommen, und andererseits auf eine robustere Implementierung zurückzuführen ist. Der grundlegende Algorithmus für die Darstellung von hydrophilen Wechselwirkungen wurde nicht verändert. Im Folgenden soll die Berechnung für die Darstellung der hydrophoben Kontakte zusammenfassend beschrieben werden.

Das Konzept für die Darstellung von gerichteten hydrophilen Wechselwirkungen kann auf die hydrophoben Kontakte nicht einfach übertragen werden. Hydrophobe Kontakte sind ungerichtet und entstehen durch die räumliche Nähe von ungeladenen Atomen im wässrigen Medium [BTS02]. Um den Charakter der hydrophoben Kontakte graphisch widerzuspiegeln, wurde eine Darstellung gewählt, die ganze Bereiche des Liganden und der korrespondierenden Aminosäure repräsentiert und nicht einzelne Atome. So werden die hydrophoben Bereiche des Liganden mit Abschnitten eines kubischen Splines hervorgehoben und die Bezeichnungen der Aminosäuren, die hydrophobe Kontakte mit diesem Bereich ausbilden daneben platziert. Ein hydrophober Kontakt wird immer dann angenommen, wenn drei beliebige, als hydrophob festgelegte Ligandatome und drei hydrophobe Atome einer Aminosäure innerhalb einer festgelegten Distanz zueinander liegen. Diese Distanz darf nicht die Summe aus den van-der-Waals-Radien der betrachteten Atome und einem Toleranzwert von 0.8 Å überschreiten. Beispiele für die Darstellung eines hydrophoben Kontaktes sind in Abbildung 3.1b und 3.2 zu finden.

Für die Darstellung der Splineabschnitte wird zunächst ein geschlossener kubischer Spline um den Ligand herum berechnet. Die Stützpunkte für den Spline werden, wie in Abbildung 3.5 gezeigt, von den 2D Koordinaten der Ligandatome im Strukturdiagramm abgeleitet. Für eine detaillierte Beschreibung des Algorithmus und der Kriterien für die

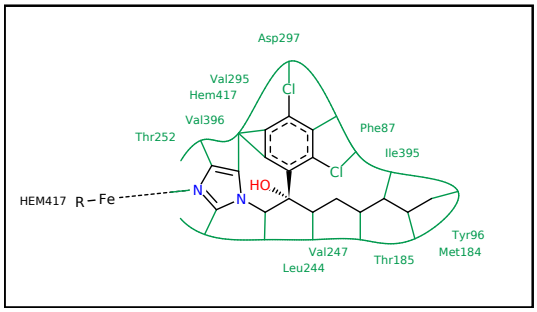


Abbildung 3.5: Die Stützpunkte des Splines, der die hydrophobe Kontaktfläche des Liganden markiert, werden von den 2D Koordinaten der Atome im Strukturdiagramm abgeleitet. Ein Atom wird als Stützpunkt ausgewählt, wenn es entweder endständig ist, wie z.B. die abgebildeten Chloratome oder wenn seine Bindungen einen konvexen Winkel bilden. Tief vergrabene Punkte wie die Hydroxylgruppe werden ausgeschlossen, um Schleifen zu vermeiden. Nach der Splineberechnung werden Abschnitte über hydrophilen Atomen herausgeschnitten. Die Abbildung wurde entnommen aus der Veröffentlichung zur Beschreibung der Layoutberechnung für hydrophobe Kontakte [D3].

Festlegung der sichtbaren Splineabschnitte sei auf die entsprechende Veröffentlichung [D3] verwiesen.

Der Platzbedarf der Elemente, die einen hydrophoben Kontakt visualisieren, ist geringer als für die auf atomarer Ebene dargestellten hydrophilen Wechselwirkungen. Zusätzlich kann die Platzierung der Aminosäurebezeichnungen flexibler gestaltet werden als die der Aminosäurestrukturdiagramme. Deshalb wird der Platzierungsalgorithmus im Anschluss an die Berechnung des Layouts für das Wasserstoffbrückennetzwerk durchgeführt. Die Platzierung geschieht mit Hilfe eines Gitters, an dessen Knoten annotiert ist, ob diese Koordinate bereits von einer Struktur besetzt ist. Durch dieses Verfahren wird sichergestellt, dass das Zeichnen der hydrophoben Kontakte keine neuen Kollisionen im Diagramm erzeugt.

3.3 Interaktionsmodell

Nach Abschluss der Entwicklung des Layoutalgorithmus hat sich die Notwendigkeit eines eigenen Interaktionsmodells herausgestellt. Das Ziel war, ein gut nachvollziehbares Modell zu entwickeln, das im Gegensatz zu dem bisher verwendeten komplexen Bewertungsschema [Böh94] in FlexX [RKLK96] durch wenige geometrische Kriterien das Vorhandensein einer Wechselwirkung abschätzt. Desweiteren sollte ermöglicht werden, neue Wechselwirkungstypen einzuführen, wie π-π-Wechselwirkungen oder π-Kationwechselwirkungen. In der wissenschaftlichen Veröffentlichung [D2], die Teil der

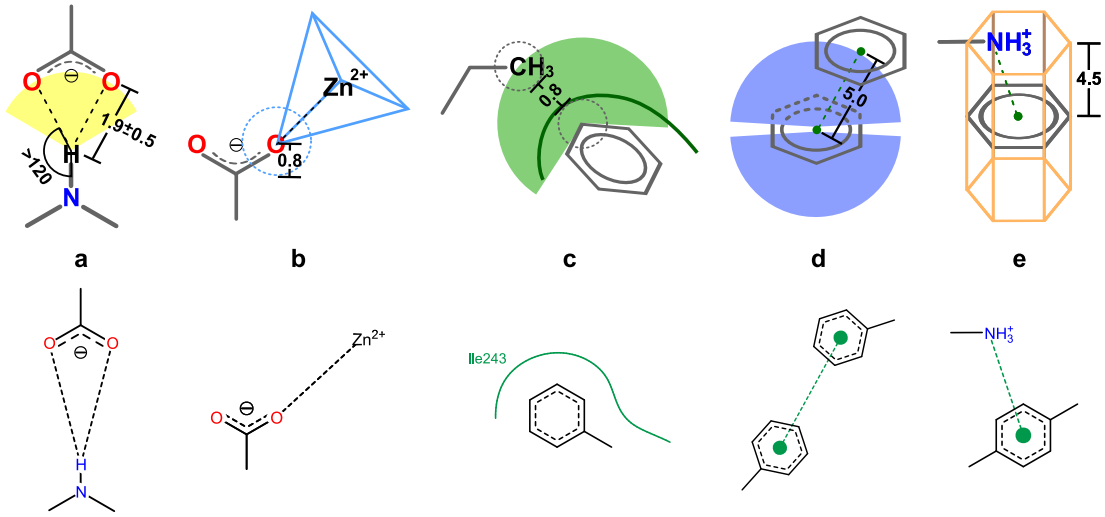


Abbildung 3.6: Die Interaktionstypen des in PoseView enthaltenen Modells: Wasserstoffbrücken (a), Metallwechselwirkungen (b), hydrophobe Kontakte (c), π-π-Wechselwirkungen (d) und π-Kationwechselwirkungen (e). Das Bild wurde entnommen aus Stierand et al. [D2] und anschließend modifiziert.

kumulativen Schrift ist, ist das Modell gemeinsam mit den Ergebnissen einer groß angelegten Anwendungsstudie auf den Daten der PDB [BWF+00] publiziert worden. An dieser Stelle wird das Modell und die Erweiterung des Algorithmus, die sich durch das Hinzufügen von π-Wechselwirkungen ergeben, zusammengefasst. Die Ergebnisse der Studie werden in Abschnitt 3.4 erläutert.

Das Vorhandensein einer Wechselwirkung wird in PoseView auf Basis der dreidimensionalen Eingabekoordinaten und einiger weniger Atomeigenschaften, die festlegen, welche Wechselwirkung ein Atom ausbilden kann, abgeschätzt. Die Atomeigenschaften werden im Rahmen der Initialisierung des Moleküls durch FlexX [RKLK96] auf der Basis des verwendeten Chemiemodells annotiert. Es werden das Element, die Ladung und die von FlexX berechnete Information, ob ein Atom hydrophil oder hydrophob ist, verwendet. Die 3D Koordinaten werden entsprechend der Eingabe verwendet; eine Ausnahme bilden die Wasserstoffatome. Da sie vor allem in Dateien zur Speicherung von Proteinatomkoordinaten häufig nicht definiert sind, werden vor der Berechnung der Wechselwirkungen die Wasserstoffatome von einem externen Programm [LR09] angefügt, bzw. deren Lage optimiert. Die Erfüllung der geometrischen Parameter wird vom PoseView-Algorithmus selbst berechnet. Eine Ausnahme bilden Metallatome: Da diese Atome feste Koordinationsgeometrien haben, die die Lage der potentiellen Wechselwirkungspartner festlegen, werden diese extern mit der von Seebeck et al. [SRKR08] beschriebenen Methode berechnet. Die genaue Beschreibung der Standardeinstellung

für die verschiedenen Parameter ist in der Veröffentlichung zum integrierten Wechselwirkungsmodell in PoseView [D2] zu finden. Eine graphische Darstellung der für die unterschiedlichen Wechselwirkungstypen angewendeten geometrischen Kriterien und jeweils ein Beispiel für deren Darstellung in PoseView wird in Abbildung 3.6 gezeigt.

Die Ausrichtung der Wechselwirkungslinien und die Platzierung der Aminosäurestrukturdiagramme der mit dem neuen Interaktionsmodell eingeführten π-Wechselwirkungen wird vom Algorithmus analog zu der Berechnung bei hydrophilen Wechselwirkungen durchgeführt. Der Unterschied ist in der Platzierung der Start- und Endpunkte der Wechselwirkungslinien zu finden. Da die π-Wechselwirkungen von einem delokalisierten System, das von mehreren Atomen gebildet wird, ausgehen, kann der Start- bzw. Endpunkt der Linie nicht einfach einer Atomkoordinate zugewiesen werden. Da es sich bei den betrachteten π-Systemen immer um aromatische Ringe oder Ringsysteme handelt, wird der Schwerpunkt aller Koordinaten der beteiligten Atome berechnet und als Start-/Endpunkt festgelegt. Um auf die Tatsache hinzuweisen, dass die Wechselwirkung von mehreren Atomen ausgebildet wird, wird als graphisches Element ein gefüllter Kreis auf das Ende der Wechselwirkungslinie, das zu einem π-System gehört, gelegt.

3.4 Ergebnisse

Der entwickelte Algorithmus zur 2D Darstellung von Protein-Ligand-Komplexen wurde auf den Daten, die in der PDB [BWF$^+$00] enthalten sind, getestet. Die Liganden wurden aus der Datenbank LigandExpo [FCM$^+$04] entnommen; sie ist von der PDB abgeleitet und enthält alle dort vorhandenen Liganden. Der Vorteil der Verwendung dieser Daten liegt in deren Aufbereitung: Alle Atom- und Bindungstypen sind definiert und es wurden bereits Protonen hinzugefügt. Das aus diesen Daten abgeleitete Testset bestand aus 201245 Komplexen, die als Eingabe für den PoseView-Algorithmus verwendet wurden. Für 155612 Komplexe konnte ein Diagramm berechnet werden. Sie wurden noch einmal nach den am Anfang des Kapitels beschriebenen Qualitätskriterien in drei Gruppen unterteilt: 80% der Diagramme wiesen ein kollisionsfreies Layout auf, 17% ein verbesserbares Layout, und die restlichen 3% konnten aufgrund der Zeichenkonventionen zweidimensional nicht überkreuzungsfrei gezeichnet werden. Für 32549 Komplexe wurden keine Wechselwirkungen zwischen Ligand und Protein berechnet und damit auch kein Diagramm generiert. 897 Komplexe mit mehr als 18 gerichteten Wechselwirkungen wurden ebenfalls von der Berechnung ausgeschlossen, um zu lange Rechenzeiten zu vermeiden, da in diesem Fall mit vielen Kollisionen und, bedingt durch die hohe Anzahl an Diagrammelementen, eine lange Layoutoptimierungslaufzeit zu erwarten ist. Aufgrund technischer Probleme, z.B. ungültige Dateiformate oder komplexe Ringsysteme, konnte

für weitere 11149 Komplexe kein Diagramm erzeugt werden. In 1038 Fällen wurde die maximale Rechenzeit von 450 Sekunden überschritten. Bei der Auswertung des Tests stellte sich ein starker Zusammenhang zwischen der Anzahl der Interaktionen und der Rechenzeit sowie der Anzahl der Interaktionen und der Layoutqualität heraus. Die Einzelheiten der hier zusammengefassten Anwendungsstudie sind in zugrunde liegenden Veröffentlichung [D2] beschrieben.

4

Zweidimensionale Darstellung multipler Protein-Ligand-Komplexe

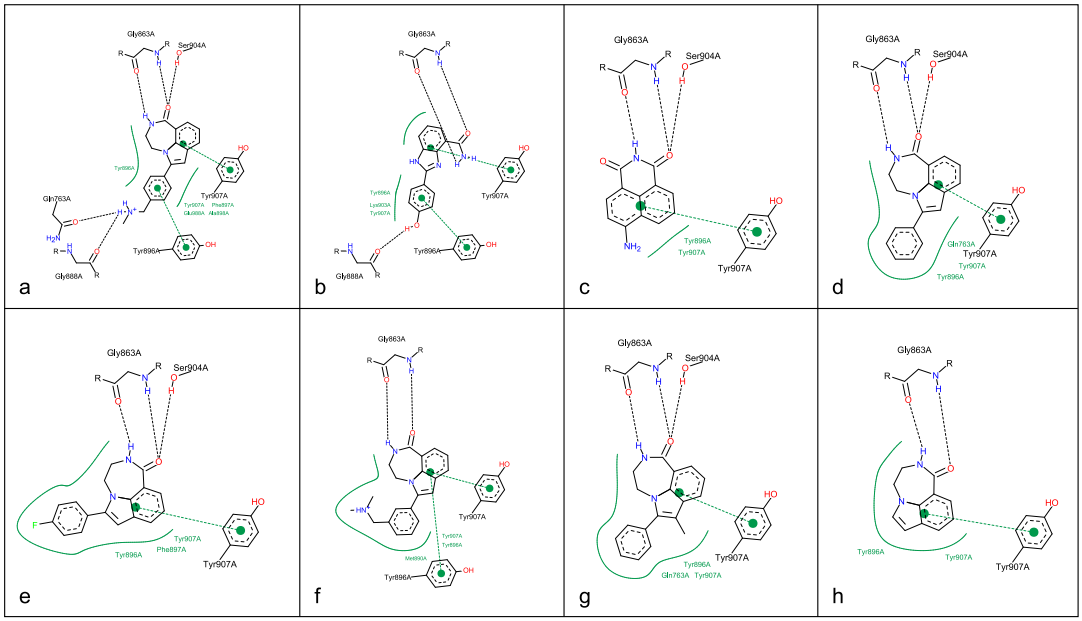
Im Verlauf der Entwicklung des Algorithmus für die 2D Darstellung einzelner Komplexe hat sich der Bedarf an einer Version, die eine konsistente Darstellung von Serien verwandter Komplexe ermöglicht, herausgestellt. CADD-Verfahren, wie Virtuelles Screening, Docking oder Scaffold Hopping, generieren Komplexserien, die dadurch charakterisiert sind, aus unterschiedlichen Liganden gebunden an dasselbe aktive Zentrum eines Proteins zu bestehen. Um die Vergleichbarkeit der Diagramme zu erhöhen, ist es hilfreich für den Betrachter, sich betreffend der Lage der Aminosäurestrukturdiagramme nicht jedes Mal neu orientieren zu müssen, sondern ein konsistentes Layout für die gesamte Serie vorzufinden. Es wurde eine Erweiterung des bestehenden Algorithmus entwickelt, die für eine gegebene Serie von Komplexen mit unterschiedlichen Liganden, die an dasselbe aktive Zentrum eines Proteins gebunden sind, ein konsistentes Layout auf Aminosäureseite berechnet. Dies bedeutete eine Umstellung des Konzeptes von einem ligandzentrierten Ansatz auf einen Ansatz, der das Aminosäurelayout global optimiert, und damit die Einführung zahlreicher neuer Algorithmen. Die wissenschaftliche Veröffentlichung [D4] zu diesem Thema ist Teil der kumulativen Dissertationsschrift.

4.1 Algorithmus

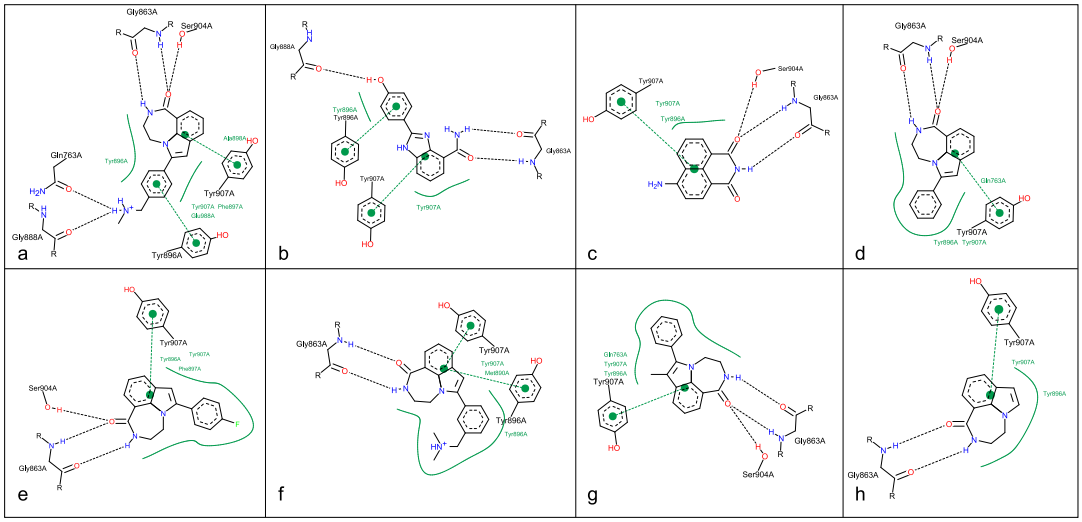
Die Berechnung des Layouts für das Wasserstoffbrückennetzwerk einzelner Komplexe ist ein ligandzentriertes Verfahren, das in dieser Form nicht mehr verwendet werden kann. Für die Darstellung mehrerer Komplexe mit konsistentem Aminosäurelayout wird ein Algorithmus benötigt, der eine Anordnung der Aminosäurestrukturdiagramme berechnet, für die die Anzahl aller Wechselwirkungslinienüberkreuzungen der individuellen Komplexdiagramme minimiert wird. Der Algorithmus berechnet eine globale Position

für jede Aminosäure, die mit mindestens einem Liganden aus der Serie eine gerichtete Wechselwirkung (hydrophile Interaktion oder π-Wechselwirkung) ausbildet. Dabei soll die Summe der Anzahl der überkreuzenden Wechselwirkungslinien für alle Komplexe minimiert werden. Vor Beginn der Layoutberechnung werden zunächst für alle Komplexe die Wechselwirkungen berechnet. Aufgrund der Überlegung, dass eine kollisionsfreie Platzierung aller Diagrammelemente mit steigender Anzahl immer schwieriger wird, werden die Komplexe nach diesem Kriterium sortiert. So wird sichergestellt, dass im Verlauf des Algorithmus immer zuerst die Komplexe mit den meisten Interaktionen betrachtet werden.

Anschließend an die Sortierung der Komplexe beginnt die eigentliche Layoutberechnung, indem eine initiale Reihenfolge für die Aminosäurestrukturdiagramme festgelegt wird. Diese Reihenfolge wird abgeleitet von der 3D Struktur des Proteins, um die Nachbarschaften der Aminosäuren beim Wechsel auf zwei Dimensionen zu erhalten: Entlang der Connolly-Oberfläche des Proteins werden die Distanzen zwischen den Aminosäuren berechnet. Auf Basis des daraus resultierenden vollständigen Graphen, dessen Kanten mit den Distanzen markiert sind, wird der kürzste Hamiltonkreis berechnet. Das Auffinden des küzesten Hamiltonkreises entspricht dem *Traveling Sales Person*-Problem, das in diesem Fall mit der Minimalen Spannbaum-Heuristik (MST-Heuristik) gelöst wird. Die initiale Reihenfolge der Aminosäurestrukturdiagramme entspricht dann dem berechneten Hamiltonkreis. Nun wird für die ersten 20 Liganden der Serie getestet, ob bei der gegebenen Reihenfolge das Ligandlayout so modifiziert werden kann, dass eine überkreuzungsfreie Anordnung der Wechselwirkungslinien möglich ist. Der Test wird nur für 20 Liganden ausgeführt, um die Rechenzeit in diesem Schritt zu beschränken. Kann nicht für alle Liganden eine überkreuzungsfreie Anordnung der Wechselwirkungslinien generiert werden, so wird die Reihenfolge in einem *Simulated Annealing* Verfahren, das die Annahme oder Ablehnung des neuen Zustandes steuert, zufällig modifiziert und der Test erneut durchgeführt. An dieser Stelle ist es von Bedeutung, dass die Komplexe mit den meisten Interaktionen zuerst betrachtet werden, da dieser Test nicht für alle von ihnen durchgeführt wird. Auf die Optimierung der topologischen Ordnung der Aminosäurestrukturdiagramme werden zunächst die Liganddiagramme überlagert. Nachdem alle Liganden entsprechend der vorgegebenen Reihenfolge optimiert wurden, werden sie anhand der Lage ihrer Wechselwirkungsatome überlagert. Für die entstehende Punktewolke aus Interaktionslinienstartpunkten wird die konvexe Hülle berechnet. Basierend auf der konvexen Hülle werden nun die Wechselwirkungslinien analog zum Platzierungsalgorithmus für Einzeldiagramme ausgerichtet. Zuletzt folgt die Kollisionsbehandlung, die ebenfalls der Kollisionsbehandlung für Einzeldiagramme entspricht bis auf die Tatsache, dass die Kollisionen nicht auf atomarer Ebene berechnet werden, sondern auf



(a) Diagramme mit konsistentem Aminosäurelayout



(b) Diagramme derselben Komplexe wie in 4.1(a) mit individuell berechnetem Layout

Abbildung 4.1: Vergleichende Darstellung einer Serie von Diagrammen unterschiedlicher Liganden aus der ZINC Datenbank [HSI06] gebunden an das Protein Poly ADP Ribose Polymerase mit der PDB ID 1EFY [Whi00]. Durch die konstante Platzierung aller Aminosäurestrukturdiagramme in Abbildung 4.1(a) sind die Unterschiede im Wechselwirkungsmuster sofort erkennbar. Im Gegensatz dazu fällt der Vergleich in Abbildung 4.1(b) weniger leicht und ist zeitaufwändiger. Die Abbildungen wurden entnommen aus der Veröffentlichung zu diesem Thema [D4]

.

Basis der bereits erwähnten konvexen Hülle aus den überlagerten Liganden und den konvexen Hüllen für die Aminosäurestrukturdiagramme. Die Platzierung der Aminosäurestrukturdiagramme und die anschließende Kollisionsbehandlung muss nur einmal durchgeführt werden, da sie für die globalen Strukturen berechnet werden.

4.2 Ergebnisse

Die Ergebnisse des Algorithmus für die Berechnung eines konsistenten Layouts wurden aufgrund des protypischen Stadiums ausschließlich qualitativ bewertet. Anhand der in der wissenschaftlichen Veröffentlichung [D4] abgebildeten Beispiele konnte gezeigt werden, dass die Qualität der resultierenden Diagramme für Komplexserien vergleichbar ist mit denen für einzelne Komplexe. Für einige Details muss jedoch noch eine Lösung gefunden werden; dazu gehören die Ausrichtung rotierbarer funktioneller Gruppen auf Aminosäureseite und eine Nachoptimierung des Ligandlayouts, um Überschneidungen von Wechselwirkungslinien mit den Strukturdiagrammen zu vermeiden. Diese Aspekte werden detailliert im Resultate- und Diskussionsteil der zugrunde liegenden Veröffentlichung [D4] besprochen. Eine vergleichende Darstellung einer Komplexserie, einmal mit konsistentem Layout und einmal mit Standardlayout, zeigt den Gewinn, den der beschriebene Algorithmus bringt. Vor allem bei recht unterschiedlichen Liganden fällt es schwer, Gemeinsamkeiten und Unterschiede im Wechselwirkungsmuster schnell zu erkennen, wenn die Positionen der Aminosäurestrukturdiagramme nicht global definiert sind, während diese bei gleich ausgerichteten Diagrammen unmittelbar ins Auge fallen. Abbildung 4.1 zeigt solch eine vergleichende Darstellung.

5

Zusammenfassung und Ausblick

In dieser Arbeit wurden neue Algorithmen für die automatische Generierung zweidimensionaler Darstellungen von Protein-Ligand-Komplexen entwickelt.

Aufbauend auf einen schon bestehenden Prototyp wurde zunächst ein Verfahren für die Berechnung von Diagrammen einzelner Komplexe konzipiert und implementiert. Dieses Verfahren konnte bis zur Anwendung gebracht werden und wird sowohl als alleinstehendes Programm als auch integriert in andere Anwendungen verwendet. Das alleinstehende Programm kann auf dem Rechner lokal installiert oder als Webservice[1] genutzt werden. Die von PoseView generierten Diagramme wurden in der Datenbank PDB[2] eingebunden, um eine graphische Übersicht über die Wechselwirkungen der beinhalteten Liganden mit den Proteinen zu geben. Zusätzlich wurde PoseView in die Software-Suite LeadIT der Firma BiosolveIT[3] eingebunden und dient dort zur ergänzenden Visualisierung neben der 3D Visualisierung von Ergebnissen.

Die Stärke von PoseView im Vergleich zu anderen Verfahren, die ebenfalls zweidimensionale graphische Repräsentationen für Protein-Ligand-Komplexe berechnen, liegt vor allem in der Darstellung der Moleküle auf atomarer Basis als Strukturdiagramme unter Beachtung der für sie geltenden IUPAC-Regeln. Damit sind die Diagramme für Chemiker leicht lesbar, denn diese Darstellungsart von Molekülen ist sehr verbreitet. Auch das in PoseView implementierte Wechselwirkungsmodell ist leicht nachvollziehbar, da es auf einigen wenigen geometrischen Winkel- und Abstandskriterien und chemischen Atomeigenschaften beruht.

Der hier vorgestellte Algorithmus findet immer das Ligandlayout, das eine minimale Anzahl an Überkreuzungen von Wechselwirkungslinien auslöst. Für den anschließenden

[1] http://poseview.zbh.uni-hamburg.de
[2] http://www.rcsb.org/pdb/home/home.do
[3] http://www.biosolveit.de/LeadIT/

Platzierungsalgorithmus der Aminosäurestrukturdiagramme wurde eine Heuristik gewählt, die auf einem diskretisierten Verfahren auf Basis eines Gitters beruht. Dadurch ist es nicht immer möglich eine optimale Anordnung der Diagrammelemente zu generieren. An dieser Stelle könnte das Verfahren noch verbessert werden, z.B. durch die Einführung einer kraftfeldbasierten Methode. Zusätzlich wäre es denkbar, das Gesamtlayout durch kleine Änderungen noch leichter verständlich zu machen. So wäre es unter anderem möglich, das Strukturdiagramm des Liganden durch breitere Linien hervorzuheben oder die Bezeichner der Aminosäuren mit hydrophoben Kontakten zum Ligand an jeden hydrophoben Bereich, mit dem sie interagieren, zu zeichnen, anstatt sie nur an die Stelle mit dem stärksten Kontakt zu platzieren.

Auf der Basis des Algorithmus für die Darstellung einzelner Diagramme wurde eine Methode entwickelt, die für eine gegebene Serie von Komplexen, bei denen das Protein identisch ist, Diagramme mit konsistentem Aminosäurelayout berechnet. Dies erforderte eine Umstellung des ligandzentrierten Ansatzes auf ein Verfahren, das das globale Layout unter Beachtung aller vorkommenden Aminosäuren in der Komplexserie optimiert. Das Konzept wurde in Form eines Software-Prototyps getestet. Es hat sich gezeigt, dass eine feste Position für alle Aminosäurestrukturdiagramme berechnet werden kann und die Qualität der resultierenden Diagramme vergleichbar zu den individuell berechneten Diagrammen ist. Wie jedoch bereits erwähnt, gibt es bei der Layoutgenerierung noch Verbesserungsbedarf. So wäre es hilfreich, nach Abschluss der Platzierung noch eine Optimierung des Ligandlayouts vorzunehmen, um Kollisionen zu vermeiden. Zusätzlich könnte durch eine uniforme Skalierung und das Anzeigen aller Aminosäurediagramme in jedem einzelnen Komplexdiagramm die Darstellung in Hinsicht auf ihre Vergleichbarkeit noch verbessert werden. In diesem Fall könnte man die Diagramme der Aminosäuren, die keine Wechselwirkung zum Ligand ausbilden, in einem hellen Grauton darstellen. Eine konsistente Darstellung der hydrophoben Kontakte ist bis jetzt noch nicht Teil des Konzeptes, das aber noch um diese Diagrammelemente erweitert werden könnte.

Publikationen der kumulativen Dissertationsschrift

[D1] K. Schomburg, H.C. Ehrlich, K. Stierand, and M. Rarey. From Structure Diagrams to Visual Chemical Patterns. *Journal of chemical information and modeling*, 50(9):1529 – 1535, 2010.

[D2] K. Stierand and M. Rarey. Drawing the PDB: Protein-Ligand Complexes in Two Dimensions. *ACS Medicinal Chemistry Letters*, 1(9):540–545, 2010.

[D3] K. Stierand and M. Rarey. From modeling to medicinal chemistry: automatic generation of two-dimensional complex diagrams. *ChemMedChem*, 2(6):853 – 860, 2007.

[D4] K. Stierand and M. Rarey. Consistent Two-Dimensional Visualization of Protein-Ligand Complex Series. *Re-submitted after revision to Journal of Cheminformatics*, 2011.

[D5] K. Stierand and M. Rarey. Flat and Easy: 2D Depiction of Protein-Ligand Complexes. *Molecular Informatics*, 30(1):12 – 19, 2011.

Literaturverzeichnis

[AD06] E.V. Anslyn and D.A. Dougherty. *Modern physical organic chemistry*. University Science Books, 2006. 6

[And03] A.C. Anderson. The process of structure-based drug design. *Chemistry & Biology*, 10(9):787–797, 2003. 1

[AS05] J. Alvarez and B. Shoichet, editors. *Virtual screening in drug discovery*, volume 1. CRC, 2005. 2

[BBH+08] J. Böttcher, A. Blum, A. Heine, W.E. Diederich, and G. Klebe. Structural and Kinetic Analysis of Pyrrolidine-Based Inhibitors of the Drug-Resistant Ile84Val Mutant of HIV-1 Protease. *Journal of Molecular Biology*, 383(2):347–357, 2008. 2

[BMG96] R.S. Bohacek, C. McMartin, and W.C. Guida. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews*, 16(1):3–50, 1996. 1

[Böh94] 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. *Journal of Computer-Aided Molecular Design*, 8(3):243–256, 1994. 22

[Bre08] J. Brecher. Graphical representation standards for chemical structure diagrams (IUPAC Recommendations 2008). *Pure and Applied Chemistry*, 80(2):277–410, 2008. 7, 10

[BRSG07] C. Bennett, J. Ryall, L. Spalteholz, and A. Gooch. The aesthetics of graph visualization. *Computational Aesthetics*, pages 57–64, 2007. 8, 7

[BTS02] J.M. Berg, J.L. Tymoczko, and L. Stryer. *Biochemistry. Fifth Edition*. WH Freeman and Company New York, 2002. 1, 3, 4, 6, 7, 21

[BWF+00] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, TN Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235, 2000. 23, 24, 22

[CL07] A.M. Clark and P. Labute. 2D depiction of protein-ligand complexes. *Journal of Chemical Information and Modeling*, 47(5):1933–1944, 2007. 3, 15, 16, 10

[Con83] M.L. Connolly. Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16(5):548–558, 1983. 2

LITERATURVERZEICHNIS

[Cor01] T.H. Cormen. *Introduction to algorithms*. The MIT press, 2001. 7, 18

[Day08] Daylight Theory Manual, Version 4.9, 2008. 9, 11, 12

[DBETT98] G. Di Battista, P. Eades, R. Tamassia, and I.G. Tollis. *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR Upper Saddle River, NJ, USA, 1998. 7

[Dij59] E.W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959. 7

[DR06] J. Degen and M. Rarey. Flexnovo: Structure-based searching in large fragment spaces. *ChemMedChem*, 1(8):854–868, 2006. 2

[FCM+04] Z. Feng, L. Chen, H. Maddula, O. Akcan, R. Oughtred, H.M. Berman, and J. Westbrook. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, 20(13):2153–2155, 2004. 24

[FGR04] P.C. Fricker, M. Gastreich, and M. Rarey. Automated drawing of structural molecular formulas under constraints. *Journal of Chemical Information and Computer Sciences*, 44(3):1065–1078, 2004. 17

[gra] graphdrawing.org. http://www.graphdrawing.org/. 7

[GVW99] A.K. Ghose, V.N. Viswanadhan, and J.J. Wendoloski. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *Journal of Combinatorial Chemistry*, 1(1):55–68, 1999. 4

[Hel99] H.E. Helson. Structure diagram generation. *Reviews in Computational Chemistry*, 13:313–398, 1999. 11

[HK06] A. Hädener and H. Kaufmann. *Grundlagen der organischen Chemie*. Springer Verlag, 2006. 4, 3

[HSI06] N. Huang, B.K. Shoichet, and J.J. Irwin. Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, 49(23):6789–6801, 2006. 29

[IBB09] W.D. Ihlenfeldt, E.E. Bolton, and S.H. Bryant. The PubChem chemical structure sketcher. *Journal of Cheminformatics*, 1(1):1–9, 2009. 11, 12

[Jar73] RA Jarvis. On the identification of the convex hull of a finite set of points in the plane. *Information Processing Letters*, 2(1):18–21, 1973. 18

[KJ58] D.E. Koshland Jr. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98, 1958. 6

[Kle09] G. Klebe. *Wirkstoffdesign: Entwurf und Wirkung von Arzneistoffen*. Spektrum Akademischer Verlag, 2009. 1, 2

[Kub98] H. Kubinyi. Molekulare Ähnlichkeit. 2. Strukturbasierter Entwurf von Wirkstoffen. *Pharmazie in unserer Zeit*, 27(4):158–172, 1998. 2

[LBW81] M.F. Lynch, J.M. Barnard, and S.M. Welford. Computer storage and retrieval of generic chemical structures in patents. 1. Introduction and general strategy. *Journal of Chemical Information and Computer Sciences*, 21(3):148–150, 1981. 9

[LJE+94] P.Y. Lam, P.K. Jadhav, C.J. Eyermann, C.N. Hodge, Y. Ru, L.T. Bacheler, J.L. Meek, M.J. Otto, M.M. Rayner, Y.N. Wong, et al. Rational design of potent, bioavailable, nonpeptide cyclic ureas as hiv protease inhibitors. *Science*, 263(5145):380, 1994. 2

[LLDF97] C.A. Lipinski, F. Lombardo, B.W. Dominy, and P.J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1-3):3–25, 1997. 4, 3

[LR09] T. Lippert and M. Rarey. Fast automated placement of polar hydrogen atoms in protein-ligand complexes. *Journal of Cheminformatics*, 1(1):13, 2009. 23

[LSBL+97] G. Lange-Savage, H. Berchtold, A. Liesum, K.H. Budt, A. Peyman, J. Knolle, J. Sedlacek, M. Fabry, and R. Hilgenfeld. Structure of HOE/BAY 793 Complexed to Human Immunodeficiency Virus (HIV-1) Protease in Two Different Crystal Forms Structure/Function Relationship and Influence of Crystal Packing. *European Journal of Biochemistry*, 248(2):313–322, 1997. 2

[MCD03] E.A. Meyer, R.K. Castellano, and F. Diederich. Interactions with aromatic rings in chemical and biological recognition. *Angewandte Chemie International Edition*, 42(11):1210–1250, 2003. 6

[MG08] H. Mauser and W. Guba. Recent developments in de novo design and scaffold hopping. *Current Opinion in Drug Discovery & Development*, 11(3):365, 2008. 1

[MGKSK01] E. Mutschler, G. Geisslinger, HK Kroemer, and M. Schäfer-Korting. *Mutschler Arzneimittelwirkungen, Lehrbuch der Pharmakologie und Toxikologie. 8., völlig neu bearbeitete und erweiterte Auflage*. Wissenschaftliche Verlagsgesellschaft mbH Stuttgart, 2001. 1, 7, 10, 6

[MGR98] G.B. McGaughey, M. Gagne, and A.K. Rappe. π-Stacking interactions. *Journal of Biological Chemistry*, 273(25):15458–15463, 1998. 7, 6

[OGF+10] S.I. O'Donoghue, D.S. Goodsell, A.S. Frangakis, F. Jossinet, R.A. Laskowski, M. Nilges, H.R. Saibil, A. Schafferhans, R.C. Wade, E. Westhof, and A.J. Olson. Visualization of macromolecular structures. *Nature Methods*, 7:42–55, 2010. 2

[PGH+04] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. UCSF Chimera – a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004. 2

[Ric81] J.S. Richardson. The anatomy and taxonomy of protein structure. 1981. 2

[RKLK96] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, 261(3):470–489, 1996. 17, 22, 23

LITERATURVERZEICHNIS

[RMF98] H.J. Roth, C.E. Müller, and G. Folkers. *Stereochemie & Arzneistoffe*. Wissenschaftliche Verlagsgesellschaft mbH Stuttgart, 1998. 6

[Sch10] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010. 2, 5, 4

[SDIB05] C.J. Squire, J.M. Dickson, I. Ivanovic, and E.N. Baker. Structure and Inhibition of the Human Cell Cycle Checkpoint Kinase, Wee1A Kinase:: An AtypicalTyrosine Kinase with a Key Role in CDK1 Regulation. *Structure*, 13(4):541–550, 2005. 16

[SDMW09] F. Schreiber, T. Dwyer, K. Marriott, and M. Wybrow. A generic algorithm for layout of biological networks. *BMC Bioinformatics*, 10(1):375, 2009. 3, 10

[Sho04] B.K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004. 1

[SLT09] C. M. Song, S. J. Lim, and J. C. Tong. Recent advances in computer-aided drug design. *Briefings in Bioinformatics*, 10(5):579–591, 2009. 1

[SMR06] K. Stierand, P.C. Maaß, and M. Rarey. Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. *Bioinformatics*, 22(14):1710–1716, 2006. 3, 17, 10

[SRKR08] B. Seebeck, I. Reulecke, A. Kämper, and M. Rarey. Modeling of metal interaction geometries for protein-ligand docking. *Proteins: Structure, Function, and Bioinformatics*, 71(3):1237–1254, 2008. 23

[ST09] CA Symyx Technologies, Inc.: Sunnyvale. Symyx/Draw, Version 3.2, 2009. 11, 12

[Sta58] H.A. Staab. Hundert Jahre organische Strukturchemie. *Angewandte Chemie*, 70(2):37–41, 1958. 10

[VM10] I. Vogt and J. Mestres. Drug-Target Networks. *Molecular Informatics*, 29(1-2):10–14, 2010. 3, 10

[Wei88] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. 9

[Whi00] White, A.W. and Almassy, R. and Calvert, A.H. and Curtin, N. J. and Griffin, R.J. and Hostomsky, Z. and Maegley, K. and Newell, D.R. and Srinivasan, S. and Golding, B.T. Resistance-modifying agents. 9.1 synthesis and biological properties of benzimidazole inhibitors of the dna repair enzyme poly(adp-ribose) polymerase. *Journal of Medicinal Chemistry*, 43(22):4084–4097, 2000. 29

[WLT95] A.C. Wallace, R.A. Laskowski, and J.M. Thornton. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Engineering Design and Selection*, 8(2):127–134, 1995. 3, 15, 16, 10

[WOWS89] P.C. Weber, DH Ohlendorf, JJ Wendoloski, and FR Salemme. Structural origins of high-affinity biotin binding to streptavidin. *Science*, 243(4887):85, 1989. 18

[WPCM02] C. Ware, H. Purchase, L. Colpoys, and M. McGill. Cognitive measurements of graph aesthetics. *Information Visualization*, 1(2):103, 2002. 8, 7

[YGC+07] M.A. Yildirim, K.I. Goh, M.E. Cusick, A.L. Barabasi, and M. Vidal. Drug-target network. *Nature Biotechnology*, 25(10):1119, 2007. 3, 10

[ZS09] P. Zhou and Z. Shang. 2D molecular graphics: a flattened world of chemistry and biology. *Briefings in Bioinformatics*, 10(3):247, 2009. 2, 3, 9

[ZTS09] P. Zhou, F. Tian, and Z. Shang. 2D depiction of nonbonding interactions for protein complexes. *Journal of Computational Chemistry*, 30(6):940–951, 2009. 2

Anhang A

Vorträge und Posterpräsentationen

A.1 Vorträge

1. 25th German Conference on Bioinformatics 2010, Braunschweig: *The Art of Drawing Complexes*

2. 238th ACS Fall National Meeting & Exposition 2009, Washington, DC, USA: *PoseView - 2D visualization of protein-ligand complexes*

3. 1st German Conference on Chemoinformatics 2005, Goslar: *Automated Structure Diagram Generation of Molecular Complexes*

A.2 Posterpräsentationen

1. 5th Joint Sheffield Conference on Chemoinformatics 2010, Sheffield, Uk: *Drawing the PDB: A large-scale application study of the 2D drawing tool PoseView*

2. CHI's Structure-Based Drug Design Conference 2010, Cambridge, MA, USA: *Drawing the PDB: A large-scale application study of the 2D drawing tool PoseView*

3. 5. German Conference on Chemoinformatics - 23. CIC-Workshop 2009: *PoseView - Molecular Interaction Patterns at a Glance*, 2. Posterpreis

4. EuroQSAR 2006 (The 16th European Symposium on Quantitative Structure-Activity Relationships & Molecular Modelling) Civitavecchia, Italy: *PoseView - Protein-Ligand Interactions At First Sight*

5. German Conference on Bioinformatics (GCB 2005), Hamburg: *Poseview - Automated Structure Diagram Generation of Molecular Complexes*

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegeben Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt. Es wurde an keinem anderen Fachbereich ein Antrag auf Eröffnung eines Promotionsverfahrens gestellt.

Hamburg, den 29. Juli 2011

(Katrin Stierand)