Aus dem Institut für Molekularbiologie und Bioinformatik
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

3D Konformationsdatenbanken für das *in silico* Screening

zur Erlangung des akademischen Grades
Doctor rerum medicarum
(Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von
Mathias Dunkel
aus Hoyerswerda

Gutachter: 1. Priv.-Doz. Dr. rer. medic. R. Preißner

2. Prof. Dr. H.-G. Holzhütter

3. Prof. Dr. I. Koch

Datum der Promotion: 17. Oktober 2008

Inhaltsverzeichnis

1. Publikationsübersicht

1.1 Veröffentlichungen als Erstautor

Publikation 1: Bioinformatics. 2005; Vol. 21; 1751-3. IF: 6.701

Titel: SuperDrug: A conformational drug database.

Autoren: Goede A[1], Dunkel M[1], Mester N, Frommel C, Preissner R.

Publikation 2: Nucleic Acids Res. 2006; Vol. 34; D678-83. IF: 7.552

Titel: SuperNatural: A searchable database of available natural compounds.

Autoren: Dunkel M[1], Fullbeck M[1], Neumann S, Preissner R.

1.2 Veröffentlichungen als Co-Autor

Publikation 3: Biosystems. 2005; Vol. 80; 117-22. IF: 1.144

Titel: *In silico* screening of drug databases for TSE inhibitors.

Autoren: Lorenzen S, Dunkel M, Preissner R.

Publikation 4: BMC Bioinformatics. 2005; Vol. 6; 122. IF: 5.423

Titel: SuperLigands - A database of ligand structures derived from the Protein Data Bank.

Autoren: Michalsky E, Dunkel M, Goede A, Preissner R.

Publikation 5: Nat Prod Rep. 2006; Vol. 23; 347-56. IF: 7.89

Titel: Natural products: Sources and databases.

Autoren: Fullbeck M[1], Michalsky E[1], Dunkel M, Preissner R.

Publikation 6: Journal of Integrative Bioinformatics. 2006; Vol. 19; 1-8. [2]

Titel: A structural keystone for drug design.

Autoren: Rother K, Dunkel M, Michalsky E, Trissl S, Goede A, Leser U, Preissner R.

[1] equally contributed
[2] not peer reviewed

Publikation 7:	Nucleic Acids Res. 2007; Vol. 35; D906-10. IF: 7.552
Titel:	SuperHapten: A comprehensive database for small immunogenic compounds.
Autoren:	Günther S, Hempel D, Dunkel M, Rother K, Preissner R.

Publikation 8:	Nucleic Acids Res. 2008; Vol. 36; D919-22. IF: 6.317
Titel:	SuperTarget and Matador: Resources for exploring drug-target relationships
Autoren:	Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales E, Gewiess A, Jensen L, Schneider R, Skoblo R, Russell R, Bourne P, Bork P, Preissner R.

2. Anteilserklärung

Der Promovend hatte folgenden Anteil an den vorgelegten Publikationen:

Publikation 1:	Bioinformatics. 2005; Vol. 21; 1751-3. IF: 6.701
Titel:	SuperDrug: A conformational drug database.
Autoren:	Goede A*, Dunkel M*, Mester N, Frommel C, Preissner R. * equally contributed
Anteil:	50 Prozent
Beitrag im Einzelnen:	Erstellung der Statistik zu Medikamenten der DB, Erstellung der Datenbank und Screening-Anwendungen

Publikation 2:	Nucleic Acids Res. 2006; Vol. 34; D678-83. IF: 7.552
Titel:	SuperNatural: A searchable database of available natural compounds.
Autoren:	Dunkel M*, Fullbeck M*, Neumann S, Preissner R. * equally contributed
Anteil:	60 Prozent
Beitrag im Einzelnen:	Klassifizierung der Naturstoffe, Medikament-Naturstoff-Vergleich, Erstellung der Datenbank, Erstellung von Screening-Anwendungen

Publikation 3:	Biosystems. 2005; Vol. 80; 117-22. IF: 1.144
Titel:	*In silico* screening of drug databases for TSE inhibitors.
Autoren:	Lorenzen S, Dunkel M, Preissner R.
Anteil:	40 Prozent
Beitrag im Einzelnen:	2D und 3D Ähnlichkeitssuchen durchgeführt

Publikation 4:	BMC Bioinformatics. 2005; Vol. 6; 122. IF: 5.423
Titel:	SuperLigands - A database of ligand structures derived from the Protein Data Bank.
Autoren:	Michalsky E, Dunkel M, Goede A, Preissner R.
Anteil:	40 Prozent
Beitrag im Einzelnen:	Liganden-Medikamenten Analyse, Erstellung der Datenbank

Publikation 5:	Nat Prod Rep. 2006; Vol. 23; 347-56. IF: 7.890
Titel:	Natural products: Sources and databases.
Autoren:	Fullbeck M*, Michalsky E*, Dunkel M, Preissner R. * equally contributed
Anteil:	20 Prozent
Beitrag im Einzelnen:	Naturstoffdatenbankschema erstellt, an Zusammenfassung mitgewirkt

Publikation 6:	J. of Integrative Bioinformatics. 2006; Vol. 19; 1-8. (not peer reviewed)
Titel:	A structural keystone for drug design.
Autoren:	Rother K, Dunkel M, Michalsky E, Trissl S, Goede A, Leser U, Preissner R.
Anteil:	30 Prozent
Beitrag im Einzelnen:	Datenanalyse, Liganden-Medikamenten Vergleich, Medikamenten-Datenbank beschrieben

Publikation 7:	Nucleic Acids Res. 2007; Vol. 35; D906-10. IF: 7.552
Titel:	SuperHapten: A comprehensive database for small immunogenic compounds.
Autoren:	Günther S, Hempel D, Dunkel M, Rother K, Preissner R.
Anteil:	30 Prozent
Beitrag im Einzelnen:	Scaffoldanalyse, Klassifizierung der Haptene, Erstellung der Datenbank

Publikation 8:	Nucleic Acids Res. 2008; Vol. 36; D919-22. IF: 6.317
Titel:	SuperTarget and Matador: Resources for exploring drug-target relationships
Autoren:	Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales E, Gewiess A, Jensen L, Schneider R, Skoblo R, Russell R, Bourne P, Bork P, Preissner R.
Anteil:	30 Prozent
Beitrag im Einzelnen:	Erstellung der Datenbank, Literatur-Screening, Textmining, AJAX-Programmierung

3. Zusammenfassung

Schätzungen besagen, dass der virtuelle chemische Raum aus ca. 10^{63} organischen Kleinstrukturen (Moleküle kleiner als 500 Dalton) besteht (1). Davon sind ca. 10 Millionen Strukturen in Datenbanken erfasst, was eine schwer durchsuchbare Menge darstellt. Es existieren verschiedene Substanzbibliotheken mit experimentell ermittelten und virtuell generierten 3D Kleinstrukturen, doch ist die Verfügbarkeit der Substanzen oft unklar und Kataloge verschiedener Hersteller müssen einzeln durchsucht werden. Auch eine chemische und wirkspezifische Klassifikation der Substanzen sowie Referenzierungen zu aktuellen Publikationen sind nur durch hohen manuellen Einsatz möglich. Um diese Probleme zu vermeiden und den Aufwand zu reduzieren, wird in der vorliegenden Arbeit ein Überblick über aktuell verfügbare synthetische Wirkstoffe und Naturstoffe gegeben und biomedizinische Spezialdatenbanken unter Verwendung von cheminformatischen Methoden erstellt. Sowohl chemische Strukturdaten als auch substanzspezifische Informationen wurden dafür zusammengestellt, in einheitliche Formate überführt und in Datenbanken integriert. Statistiken zu physikochemischen Eigenschaften der verschiedenen Datenbanken wurden erstellt und ausgewertet. Wirkstoffe wurden nach strukturellen und chemischen Gesichtspunkten klassifiziert. Neben 2D Ähnlichkeitssuchen wurde auch ein 3D Überlagerungsalgorithmus integriert. Durch Verknüpfungen der Datenbanken untereinander ist in einem integrativen Projekt eine universelle Datenbankanwendung namens SuperTarget entstanden, welche der wissenschaftlichen Gemeinschaft einen umfangreichen Überblick zu Medikament-Zielprotein Interaktionen sowie deren Position im *Pathway* gibt und die Vorhersage neuer Medikament-Protein-*Pathway* Beziehungen ermöglicht. Die Anwendbarkeit der Datenbanken wurde am Beispiel von Ähnlichkeitssuchen mit bekannten Substanzen, die gegen die Ausbreitung von Prionen in Zellassays wirken und des Auffindens von neuen Wirkstoffen gegen die Ausbreitung von Prionen gezeigt. Weitere Inhibitoren zur Hemmung von Lipoxygenase sowie, Apoptose auslösende Stoffe wurden mit Hilfe der Spezialdatenbanken vorgeschlagen und anschließend erfolgreich auf Wirksamkeit getestet.

Die Datenbanken sind online erreichbar:
http://bioinformatics.charite.de/content/databases_and_applications.php

4. Einleitung und Zielstellung

Wirkstoffentwicklung beruht zu großen Teilen auf der Abfrage und Analyse von bereits bestehendem Wissen. Jeden Tag werden tausende neue Moleküle durch parallele bzw. kombinatorische Synthese generiert. Durch die zunehmende Automatisierung werden heute bereits Datenvolumina generiert, die die Zahl früherer Versuchsergebnisse um mehrere Größenordnungen übersteigen. Aufgrund der Größe der anfallenden Datenmengen gestaltet sich das Auffinden relevanter Informationen zunehmend schwieriger.

Eine weitere Herausforderung, auf welche man im *Drug Design* stößt, besteht darin, glaubwürdig vorherzusagen, welche Moleküle aus dem Pool von Millionen Substanzen mit einem Zielprotein von medizinischem oder biologischem Interesse interagiert. Eine große Anzahl an Zielproteinen findet man in Sequenz- und Strukturdatenbanken. Auch für Kleinstrukturen existieren große Datenbanken bestehend aus 3D-Strukturen und chemischen Formeln. Jedoch sind die Daten oft über mehrere Quellen verstreut und in uneinheitlichen Formaten abgelegt. Es existieren kommerziell verfügbare Datenbanken zu bioaktiven Substanzen, experimentellen Medikamenten und patentierten Substanzen, jedoch existiert keine frei verfügbare Quelle, die 3D Konformere beinhaltet.

Um diese Probleme anzugehen, sollen funktionsbasierte Spezialdatenbanken erstellt werden, die dem Thema entsprechend spezifische Eigenschaften der Wirkstoffe berücksichtigen und Konformere beinhalten. Ein weiteres Ziel der Arbeit ist es, spezifische Informationen zu Kleinstrukturen und deren Wirkungen aus wissenschaftlichen Publikationen und Online-Datenbanken so zu verknüpfen, dass neben vorhandenem Wissen auch neue Substanz-Funktions-Beziehungen ableitbar werden. Durch die Integration von verschiedenen 2D und 3D *in silico* Screening-Methoden sollen Wissenschaftlern neue Anwendungen zur Verfügung gestellt werden, die es ihnen erlauben, komplexe biologische Fragestellungen in einfachen Abfragen zu bearbeiten. Die Zusammenführung von niedermolekularen Strukturen, Zielproteinen und Signalkaskaden verschiedener Krankheiten und metabolischer Kreisläufe ist ein wichtiger Meilenstein auf dem Weg von der strukturellen Bioinformatik zur integrativen Systembiologie.

5. Methoden

5.1 Textmining

Mindestens 85% aller erhältlichen Fachinformationen sind im Textformat abgelegt. Ein Großteil der Dokumente, der auch wissenschaftliche Publikationen umfasst, ist relativ

unstrukturiert. Ziel des so genannten Textmining ist es, einzelne strukturierte Informationsbausteine aus der unstrukturierten Textmasse zu extrahieren. Informationen, die aus verschiedenen Dokumenten extrahiert wurden, werden miteinander kombiniert und so neue Beziehungen zwischen Einzelbausteinen identifiziert. Informationen zu chemischen Kleinstrukturen sind sehr komplex und auf verschiedenste Datenquellen verteilt. Mittels Textmining lassen sich Informationen zu Substanzen, wie deren chemischer Aufbau, Einsatzmöglichkeiten, Hersteller oder medizinische Wirkung identifizieren, kombinieren und in neue Spezialdatenbanken überführen. Dieses Verfahren wurde in den letzten Jahren zu einer Textmining-Pipeline weiterentwickelt. Ausgehend von verschiedenen Randbedingungen lassen sich maßgeschneiderte Datensammlungen entwickeln, was im Folgenden schematisch erklärt wird (siehe auch Abbildung 1).

Grundlage des Textminings sind über 16 Millionen medizinisch relevante Artikel, die in Kurzform (Abstracts) in einem Katalog der „National Library of Medicine“ abgelegt sind. Alle Abstracts sind auf einem Server gespiegelt und im Volltext indiziert.

Eine erste einfache Suche nach themenspezifischen Schlagworten, die über Boolesche Operatoren verknüpft werden, reduziert die Anzahl der Artikel deutlich. Die Auswahl der Schlagwörter ist bei diesem Vorgehen von enormer Bedeutung und erfordert meist ein umfangreiches Fachwissen. Ein weiterer wichtiger Filterschritt beinhaltet die Reduktion auf Artikel, in denen Kleinstrukturen erwähnt werden. Dieses erfolgt auf zwei unterschiedlichen Wegen. Zum einen mit Hilfe einer sehr umfangreichen Synonymliste die 4 Millionen Substanzen mit 10 Millionen Synonymen annotiert. Sie wurde im Laufe der letzten Jahre in der AG „Strukturelle Bioinformatik“ zusammengestellt und wird permanent erweitert. Zum anderen werden mit Hilfe von Teil-Strings aus IUPAC-Namen, wie beispielsweise benzyl, nitro, phenoxy usw. chemische Namen identifiziert. Zwar wird so die Menge der relevanten Artikel stark reduziert, allerdings ist die Menge der verbleibenden Abstracts immer noch viel zu groß, um in einer effizienten Zeit manuell bearbeitet werden zu können.

Hier setzen verschiedene Bewertungsverfahren an, die zu einer gefundenen Kleinstruktur einen Relevanzscore ermitteln. Dieser setzt sich aus verschiedenen Einzelkriterien zusammen. Die Gewichtung des Scores aus den Einzelkriterien ist sehr von den verwendeten Schlagworten abhängt. Sie wird in der Regel aus einem Lerndatensatz ermittelt. Dazu wird eine kleine Anzahl aus den extrahierten Abstracts zufällig ausgewählt und manuell in solche unterteilt, die richtige, bzw. keine Treffer enthalten.

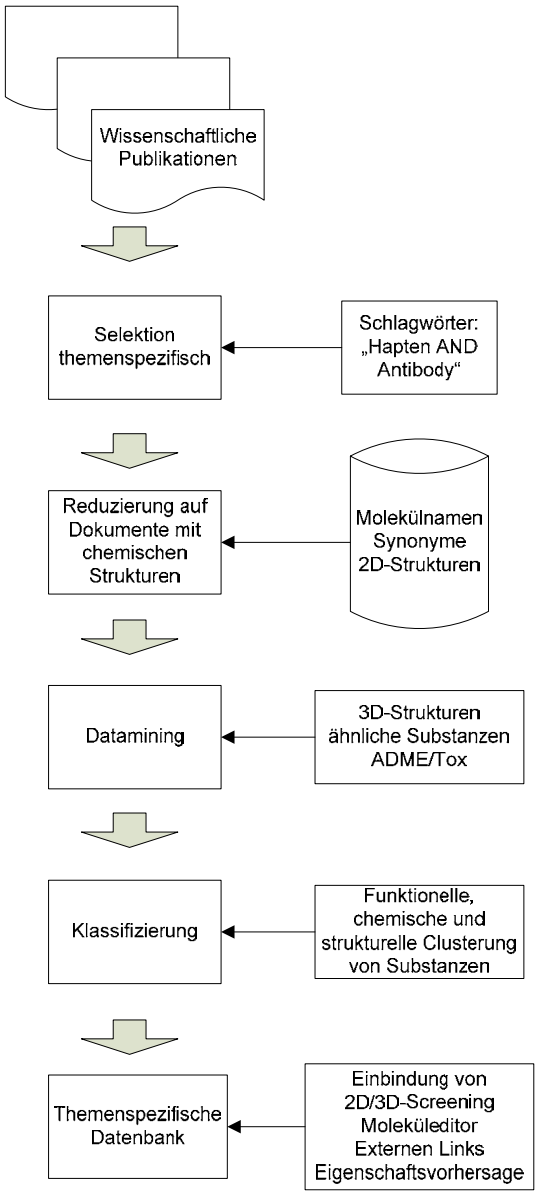

Abbildung 1. Textmining-Pipeline

Bei diesem Arbeitsschritt werden auch neue relevante Schlagworte identifiziert, die eine Funktionsbeziehung beschreiben. Ein automatisiertes Lernverfahren ermittelt dann das optimale Gewicht der oben genannten Einzelkriterien, um die beiden Teilmengen möglichst gut voneinander zu trennen.

5.2 Integration komplexer Daten

Wie in Abbildung 1 dargestellt, lassen sich die aus verschiedenen Datenquellen gewonnen Informationen miteinander kombinieren und so neue Funktionsbeziehungen ermitteln. Die erstellten Datenbanken zeichnen sich dadurch aus, dass sie Informationen enthalten, die aus der Verknüpfung verschiedener Datenquellen generiert wurden und daher nicht auf einfachem

Weg zu ermitteln sind. Welche Informationen das sind, hängt stark davon ab, welche Aufgaben- oder Fragestellung besteht. Aufgrund des Data Warehouse Konzeptes können auf einem effizienten Weg neue Datenquellen (digitalisierte Literatur, Webseiten oder Webdatenbanken) integriert werden. Dabei werden nicht nur weitere Substanzinformationen dem Data Warehouse hinzugefügt, sondern auch bestehenden Einträge aktualisiert und neue Veröffentlichungen eingepflegt. Dadurch ist ein permanent aktueller Stand neuer Erkenntnisse aus verschiedenen wissenschaftlichen Teilgebieten gewährleistet.

Eine weitere Stärke der semi-automatisierten Methodik besteht in der Möglichkeit sehr effizient gewünschte relevante Informationen zusammenzustellen und mit zusätzlichen Datenbank-Applikationen auszustatten. Die Informationen können verschiedenste Bereiche betreffen, wie beispielsweise kommerzielle Verfügbarkeit, Einsatzmöglichkeiten, physikochemische Eigenschaften oder medizinische (Neben-) Wirkungen. Gleiches gilt für die Ausstattung mit verschiedenen Datenbank-Applikationen, die in den letzten Jahren in der AG „Strukturelle Bioinformatik“ entwickelt wurden. So lassen sich beispielsweise 2D und 3D *in silico* screenings mit selbst editierten Substanzen durchführen, die verschiedenen Konformationen einer Substanz kalkulieren oder Bestellinformationen zu Substanzen abrufen.

Die Implementierung ist bewusst so gewählt, dass sie eine hohe Plattformunabhängigkeit gewährleistet, keine Server- oder Softwarelizenzen erfordert und dennoch professionellen Standards entspricht. Die verwendeten Datenbank- und Webserver (MySQL, Apache) sind auf nahezu allen Rechnerarchitekturen lauffähig. Das Interface basiert auf HTML und Javascript und ist damit auf allen Client-Rechnern lauffähig, die über einen Internet-Browser verfügen.

5.3 Berechnen von Konformationen für den 3D Ansatz

Um 3D Überlagerungen sinnvoll durchführen zu können, ist es notwendig, möglichst den gesamten Strukturraum jeder Substanz abzubilden. Dafür wird ein „Poling“ Algorithmus (2) verwendet, um eine hohe Variation der Konformere zu fördern. Dieser Algorithmus erlaubt es ebenfalls, die Redundanz bei der Konformergenerierung zu minimieren und damit den Konformerraum optimal abzudecken. Um sehr ähnliche Konformere zu entfernen, werden alle generierten Strukturen verglichen und Konformere unterhalb einer Ähnlichkeitsgrenze entfernt. Damit wird ein unnötiger Anstieg der Rechenzeit des später einzusetzenden 3D Überlagerungsalgorithmus (3) verhindert.

5.4 2D und 3D Ähnlichkeitssuchen

Die 2D Ähnlichkeitssuche ist ein wichtiges Werkzeug für die Suche nach in Bezug auf die Leitstruktur ähnlichen Substanzen. Mit Hilfe von CDK Fingerprints (4) werden die Moleküle paarweise miteinander verglichen. Als Ähnlichkeitsmaß wurde der Tanimoto-Koeffizient (5) verwendet. Die 2D Treffer bilden anschließend einen Ausgangspunkt für weitere 3D Suchen (3). Damit sollen falschpositive Treffer erkannt und aus der weiteren Analyse ausgeschlossen werden.

5.5 Suche nach Inhibitoren und Validierung des Systems

Die Medikamentendatenbank SuperDrug wurde mit Hilfe von 2D Ähnlichkeitssuchen und 3D Überlagerungen nach Substanzen mit positivem Einfluss auf Transmissible Spongiorme Encephalopathien (TSE) durchsucht. Durch die Kombination beider Methoden konnte eine Liste von 16 Kandidatenmedikamenten erstellt werden. Diese kleine Zahl an möglichen Inhibitoren erlaubt deren zeit- und geldsparende Testung mit dem Vorteil, dass sie bereits bekannt und für die Verwendung im menschlichen Körper geeignet sind.
Weitere Inhibitoren zur Hemmung von Lipoxygenase, sowie Apoptose auslösende Stoffe wurden mit Hilfe der Spezialdatenbanken gefunden und anschließend auf Wirkung getestet.
Die sich ergebenden Hits aus den Suchen werden in Kooperation mit experimentellen Arbeitsgruppen in Labortests überprüft. Für einige Substanzen sind auf Grund ihrer effektiven Wirkung bereits Tierversuche geplant (6).

6. Ergebnisse

6.1 SuperDrug, eine Medikamentendatenbank mit WHO-Klassifikation

Aufbauend auf der **A**natomischen, **T**herapeutischen und **C**hemischen (ATC) Klassifikation (7) ist eine Medikamentendatenbank erstellt worden. Dabei wurden von ca. 2.400 zugelassenen Medikamenten die Strukturen, deren Konformere, physikalische Eigenschaften, wissenschaftliche Referenzen und Anwendungsgebiete aufgenommen. Die entstandene Datenbank ist über das Internet erreichbar und unterstützt 2D und 3D Ähnlichkeitssuchen. Ein Molekül-Editor mit dem der Upload von gezeichneten Strukturen und damit auch *in silico* Screenings möglich sind, wurde integriert. Eine umfangreiche Statistik zur Verteilung der Medikamente innerhalb der Anwendungsgebiete wurde ebenfalls erstellt.
Publikation: „SuperDrug a conformational drug database.“, Bioinformatics. 2005; Vol. 21; 1751-3. IF: 6.701.

6.2 SuperLigands, eine Datenbank zu Ligandenstrukturen aus der Protein Data Bank

Diese Datenbank beinhaltet die Liganden der Protein Data Bank in einem verbreiteten Standardformat für Kleinstrukturen. Im Unterschied zum PDB-Format sind hier auch Informationen über den Bindungstyp abgespeichert. Strukturelle Ähnlichkeit zwischen den Substanzen kann über die Berechnung des Tanimoto-Koeffizientens und eine 3 dimensionale Überlagerung detektiert werden. Auch die Ähnlichkeiten zu Medikamenten der SuperDrug-Datenbank sind abrufbar.

Publikation: „SuperLigands - a database of ligand structures derived from the Protein Data Bank.“, BMC Bioinformatics. 2005; Vol. 6; 122. IF: 5.423.

6.3 SuperNatural, eine Datenbank bestellbarer Naturstoffe

Neben der ausführlichen Charakterisierung bekannter Naturstoffstrukturen wurde eine Sammlung von kommerziell erhältlichen Naturstoffen und Naturstoffanaloga erstellt. Konformere, physikalische Eigenschaften und Herkunft der Substanzen wurden aufgenommen. Schnelle automatisierte Substruktursuchen sowie 2D und 3D Ähnlichkeitssuchen wurden implementiert.

Von ~ 50.000 käuflichen Naturstoffen, Derivaten und Analoga sind die 3D-Strukturen und Konformere sowie deren Hersteller, zusammengefasst. Zu jedem Naturstoff sind ausführliche Informationen bezüglich verschiedener struktureller und chemischer Eigenschaften abgelegt. Die Datenbank enthält ~ 2.500 bekannte Naturstoffe, die durch ihre CAS-Nummer (Chemical Abstracts) charakterisiert sind und als Startpunkt für Ähnlichkeitssuchen dienen können. Durch die Implementierung eines Marvin Applets können eigene Moleküle für Ähnlichkeitssuchen importiert werden

Um mögliche Anwendungsgebiete für die gefundenen Naturstoffe zu bestimmen, wurde eine Suche nach ähnlichen Medikamenten in der SuperDrug-Datenbank ermöglicht.

Publikation: „Supernatural: a searchable database of available natural compounds.“, Nucleic Acids Res. 2006; Vol. 34; D678-83. IF: 7.552.

6.4 SuperHapten, eine Datenbank niedermolekularer immunologischer Substanzen

Immunologisch wirksame Kleinstrukturen wurden in einer Datenbank zusammengefasst. Von den enthaltenen 8.200 Kleinstrukturen sind 80% bei verschiedenen Herstellern bestellbar. Zusatzinformationen wie Carrier-Proteine und bestellbare Antikörper zu den entsprechenden Haptenen steigern den Wert der Datenbank für den Nutzer erheblich.

Publikation: „SuperHapten: A Comprehensive Database for Small Immunogenic Compounds.“, Nucleic Acids Res. 2007; Vol. 35; D906-10. IF: 7.552.

6.5 SuperTarget, eine Datenbank zu Medikament-Zielprotein-Relationen

SuperTarget beinhaltet mehr als 2.500 Target Proteine, die über 7.300 Relationen zu 1.500 Medikamenten zugeordnet sind.

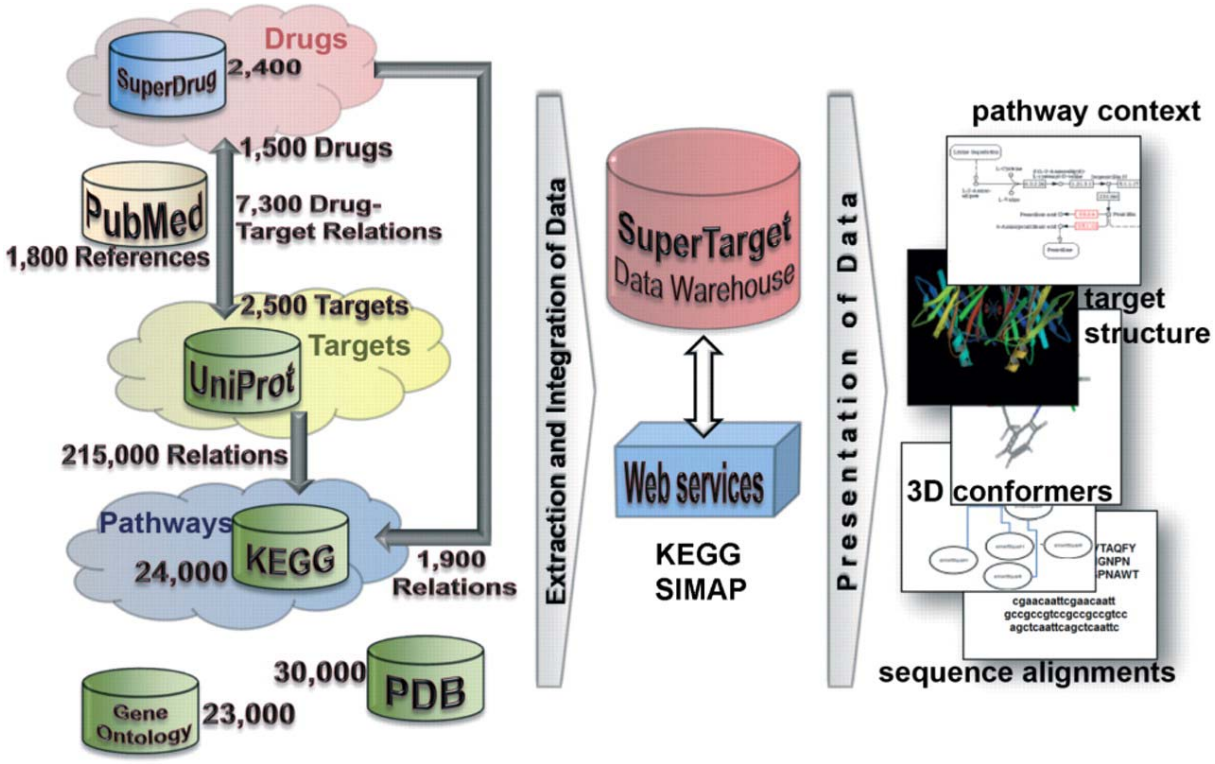

Abbildung 2. Systemarchitektur und Anzahl der Datenbankeinträge der SuperTarget-Datenbank. (entnommen aus SuperTarget, Grafikdatei bereitgestellt von J. Ahmed) Die Datenbank beinhaltet die komplette Uniprot mit mehr als 3 Millionen Einträgen. Neben Zielproteinen, Medikamenten und *Pathways* befinden sich auch 23.000 verschiedene Gen-Ontologie-Terme und Links zu 30.000 Protein-Strukturen in der Datenbank.

Relationen sind mit wissenschaftlichen Referenzen verlinkt. Des Weiteren sind in der Datenbank medikamentenbezogene Informationen über medizinische Indikationsgebiete, Nebenwirkungen, Stoffwechselwege, Signalwege und Gen-Ontologie- (GO) Bezeichnungen der Zielmoleküle integriert. Durch eine einfach zu verwendende Anfrageoberfläche wird dem Nutzer ermöglicht komplexe Anfragen zu erstellen.

Publikation: „SuperTarget and Matador: Resources for exploring drug-target relationships.“, Nucleic Acids Res. 2008; Vol. 36; D919-22. IF: 6.317.

6.6 TSE Inhibitoren, ein in silico screening Ansatz

Mit Hilfe der neuen Konformer-Medikamenten-Datenbank wurden 2D Ähnlichkeitssuchen und 3D Überlagerungssuchen nach Substanzen mit einem positiven Einfluss auf Transmissible Spongiforme Encephalopathien (TSE) durchgeführt. Beide Suchen wurden kombiniert um eine Liste von 16 Medikamenten-Kandidaten abzuleiten. Mit diesem Ansatz zur Neupositionierung von Medikamenten ergeben sich mehrere zeit- und kostensparende Effekte, wie beispielsweise die bereits bestätigte Eignung der Substanzen zur Anwendung im menschlichen Körper, ihr bekanntes Nebenwirkungsprofil und automatisierte Synthesemethoden.

Publikation: „In silico screening of drug databases for TSE inhibitors.", Biosystems. 2005; Vol. 80; 117-22. IF: 1.144.

6.7 Naturstoffe, ein Überblick und deren Verfügbarkeit

Ein genereller Überblick über Naturstoffdatenbanken, Lieferanten und Hersteller wird gegeben. Wissenschaftlern werden die Möglichkeiten aufgezeigt, Informationen über Naturstoffe zu erhalten sowie Anregungen zur Identifizierung von pharmazeutisch relevanten Substanzen gegeben. Darüber hinaus wurden verschiedene Molekülgerüste analysiert und durch Vergleiche mit Medikamenten Rückschlüsse auf deren biologische Aktivität und medizinische Anwendbarkeit geschlossen.

Publikation: „Natural products: Sources and databases.", Nat Prod Rep. 2006; Vol. 23; 347-56. IF: 7.89.

7. Diskussion

Zusammen ergeben die erstellten Datenbanken eine einzigartige Kombination von Hilfsmitteln im Prozess der Medikamentenentwicklung. Die Spezialdatenbanken stellen jeweils in ihren Bereichen neue Screeningwerkzeuge für Wissenschaftler dar. Durch einen integrativen Ansatz bei der Entwicklung der Datenbanken werden systembiologische Zusammenhänge aufgezeigt und verfügbar gemacht.

7.1 Frei verfügbare und kommerzielle Medikamentendatenbanken

SuperDrug stellte als erste Datenbank die WHO klassifizierten Medikamentenstrukturen frei im Internet zur Verfügung und ermöglichte dem Anwender 2D Ähnlichkeitssuchen sowie 3D Überlagerungen. Neben SuperDrug existieren inzwischen verschiedene andere

Medikamentensammlungen, wie Kegg-Drug (8) oder DrugBank (9), jedoch sind dort Konformere nicht enthalten und 3D Überlagerungen nicht möglich.
Als kommerziell verfügbare Ressourcen sind der World-Drug-Index (WDI) und der MDL Drug Data Report (MDDR) zu nennen. Beide stellen umfangreiche Sammlungen von Medikamenten und aktiven Substanzen dar, doch fehlt beiden im Vergleich zur SuperDrug-Datenbank eine Zuordnung der Medikamente zur ATC-Klassifikation, in welcher die Substanzen nach anatomischen, therapeutischen und chemischen Gesichtspunkten unterteilt werden.

7.2 Vergleich zwischen PDB-Liganden und Medikamenten

Die PDB-Liganden-Datenbank SuperLigands beinhaltet alle an Proteine gebundenen Kleinstrukturen aus der Protein Data Bank (PDB). Der Zugriff auf die Funktionalität dieser Datenbank ist direkt über einen Link auf der Webseite der PDB möglich. So können Medikamente der SuperDrug angezeigt werden, die zu den Liganden der PDB ähnlich sind. Es ist allgemein anerkannt, dass ähnliche Substanzen mit einem Tanimoto-Koeffizienten größer als 85% auch eine ähnliche biologische Wirkung aufweisen (10). Der Kreuzvergleich zwischen der SuperLigands-Datenbank und der SuperDrug-Datenbank ergab insgesamt 12 Millionen 2D Scores. Insgesamt konnten 413 der 5.040 PDB-Liganden einem Medikament zugeordnet werden. Eine erweiterte Analyse ergab, dass 1.475 PDB-Liganden ein Pendant auf der Medikamentenseite mit einer Ähnlichkeit von mindestens 90% besitzen. Um die Medikamenten-Ähnlichkeit der PDB-Liganden weiter zu untermauern, wurde die Einhaltung der ‚Rule of five' (11) untersucht. Die Analysen zeigen, dass viele Liganden der PDB entweder Medikamente sind, ihnen strukturell ähnlich sind oder zumindest ähnliche chemische Eigenschaften besitzen. Somit sind sie geeignet mögliche Vorlagen zum *Drug Design* zu bilden.

7.3 Naturstoffe, eine Quelle potentieller Medikamente

Ein Anteil von ca. 50% der durch die „Food and Drug Administration" (FDA) zugelassenenen Medikamente sind Naturstoffe oder Naturstoffanaloga (12). Die chemische Diversität und die einzigartigen Eigenschaften der Naturstoffe ermöglichen einen viel versprechenden Startpunkt zur Entwicklung von Innovationen für wissenschaftliche, medizinische und ökotrophologische Anwendungen. Trotzdem ist der umfassende Zugriff auf Naturstoffe bisher nur durch kommerzielle Lösungen möglich gewesen (13). Ein genereller Überblick zur Verfügbarkeit von Naturstoffen, über Lieferanten und Hersteller wurde in der Publikation "Natural products: sources and databases" gegeben. Die Entwicklung der SuperNatural-

Datenbank ermöglicht einen freien Zugriff auf 3D Strukturen und physikochemische Eigenschaften von ca. 50.000 bestellbaren Naturstoffen, Naturstoff-Analoga und deren Konformeren. In dieser Datenbank ist es möglich, 2D Ähnlichkeitssuchen durchzuführen. Bei einer Analyse wurde festgestellt, dass etwa 300 Naturstoffe der SuperNatural-Datenbank identisch zu bereits zugelassenen Medikamenten sind sowie ca. 3.600 Naturstoffe einen Tanimoto-Wert von mindestens 85% zu Medikamenten aufweisen.

7.4 Haptene, Fluch oder Segen?

Das Immunsystem schützt Organismen vor dem Eindringen von Mikroorganismen, fremden Proteinen, Peptid-Epitopen und einer Vielzahl von chemischen Substanzen. Darunter sind auch kleine Moleküle, sogenannte Haptene, die eine Immunantwort auslösen, wenn sie an sogenannte Carrier-Moleküle gebunden sind. Bekannte Haptene sind Xenobiotika oder Naturstoffe, die Autoimmunkrankheiten wie Kontaktdermatitis oder Asthma auslösen. Haptene werden aber auch für die Entwicklung von Biosensoren, Immunomodulatoren und neuen Impfstoffen verwendet. Obwohl durch Haptene ausgelöste Allergien für 6-10% aller Medikamenten-Nebenwirkungen verantwortlich sind, ist die Korrelation zwischen Struktur und Hapten-ähnlicher Wirkung wenig untersucht. Die Klassifizierung der Haptene nach Grundgerüsten ermöglicht eine tiefergehende Analyse dieser Problematik und verspricht in Kombination mit Medikamentenvergleichen neue Ansatzpunkte im Wirkstoffdesign. Zusammenfassend lässt sich also feststellen, dass Haptene, neben negativen Wirkungen, durch gezielte Anwendungen auch positive Auswirkungen zeigen.

7.5 Integrative Systembiologie mit SuperTarget

Um systembiologische Fragestellungen zu bearbeiten ist die Einbeziehung vieler Ressourcen wie Liganden, Targets, *Pathways*, Gen-Ontologien und wissenschaftlicher Literatur nötig. So müssten beispielsweise für die Beantwortung der in Abbildung 3 gestellten Frage umfangreiche Recherchen angestellt werden.

Durch Verwendung der SuperTarget-Datenbank können derartige interessante wissenschaftliche Probleme auf einer Plattform bearbeitet werden, beispielsweise das Auffinden von Medikamenten, die im gleichen *Pathway* agieren, oder Medikamente, die das gleiche Ziel-Protein adressieren, aber von unterschiedlichen Enzymen abgebaut werden. Alle Ergebnisse einer Datenbankanfrage sind durch wissenschaftliche Referenzen untermauert.

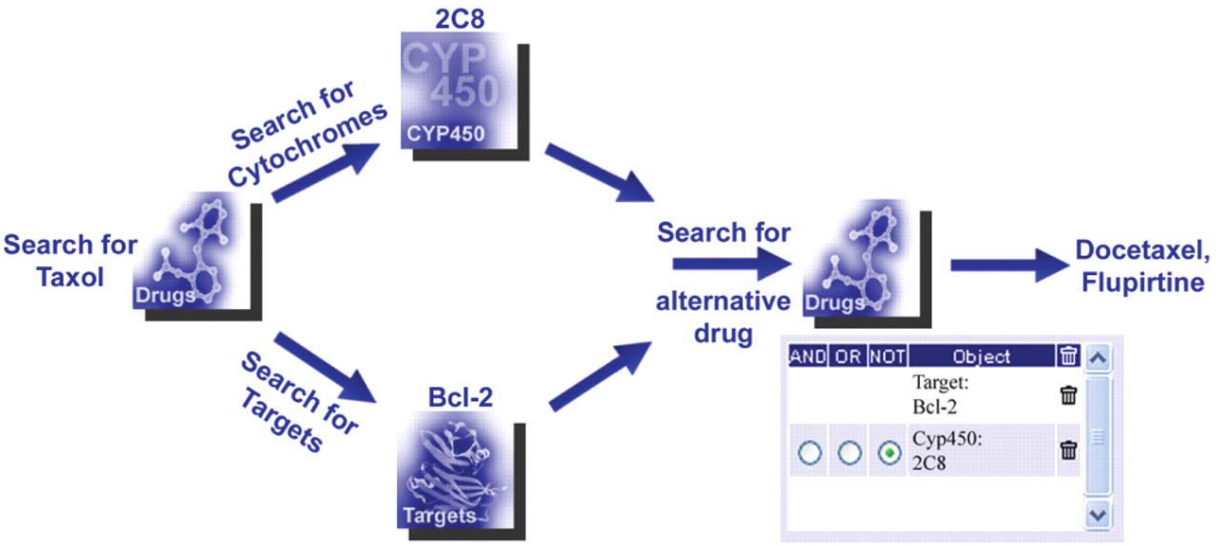

Abbildung 3. Beispiel einer komplexen Anfrage zu Bcl-2 (entnommen aus SuperTarget, Grafikdatei bereitgestellt von J. Ahmed: Suche nach alternativen Medikament zu Taxol, welches das gleiche Zielprotein (Bcl-2) adressiert, aber nicht vom Cytochrom 2C8 metabolisiert wird. Die entstandene Ergebnisliste beinhaltet zwei alternative Wirkstoffe zu Taxol (Docetaxel und Flupirtin).

Neben SuperTarget existieren weitere frei verfügbare Anwendungen zum Thema. Die Therapeutic Target Database (TTD) (14) beinhaltet Targetinformationen zu ca. 1.000 Medikamenten. Eine weitere Anwendung, die DrugBank (15) umfasst 2.600 Medikamenten-Zielprotein-Relationen zu 900 FDA-geprüften Medikamenten. Drugbank stellt nur Referenzen zu den Zielproteinen und nicht zu den Interaktionen bereit, was es erschwert, Informationen über den experimentellen Kontext unter welchem die Interaktionen beobachtet wurden, zu erhalten. Die Medikamente in TTD sind nicht mit PubChem oder CAS-Nummern verlinkt und den Zielproteinen fehlt eine Verknüpfung zu UniProt oder der PDB.

Obwohl die erste Version von SuperTarget mit ihren vielen bioinformatischen Anwendungen und Suchoptionen schon eine umfangreiche Ressource für biologische Suchen und tiefergehende Analysen darstellt, sind die Daten noch nicht vollständig. Durch eine Upload-Funktion wird der wissenschaftlichen Gemeinschaft ermöglicht, neue Medikamenten-Zielprotein-Interaktionen der Datenbank hinzuzufügen.

SuperTarget kann als Wissensquelle, Discovery-Tool oder Training-Set für verschiedene biochemische Anwendungen verwendet werden.

8. Referenzen

1. Bohacek, R.S., McMartin, C. and Guida, W.C. (1996) The art and practice of structure-based drug design: a molecular modeling perspective. *Med Res Rev*, **16,** 3-50.
2. Smellie, A., Teig, S. and Towbin, P. (1995) Poling - Promoting Conformational Variation. *J Comput Chem*, **16,** 171-187.
3. Thimm, M., Goede, A., Hougardy, S. and Preissner, R. (2004) Comparison of 2D similarity and 3D superposition. Application to searching a conformational drug database. *J Chem Inf Comput Sci*, **44,** 1816-1822.
4. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. and Willighagen, E. (2003) The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci*, **43,** 493-500.
5. Delaney, J.S. (1996) Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol Divers*, **1,** 217-222.
6. Füllbeck, M., Huang, X., Dumdey, R., Frommel, C., Dubiel, W. and Preissner, R. (2005) In silico 2D/3D-identification: Novel curcumin- and emodin- related compounds inhibit COP9 signalosome acsociated kinases and include apoptosis in tumor cells.
7. (2006) The selection and use of essential medicines. Report of the WHO expert committee, 2005 (including the 14th model list of essential medicines). *World Health Organ Tech Rep Ser***,** 1-119.
8. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, **36,** D480-484.
9. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. and Woolsey, J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*, **34,** D668-672.
10. Matter, H. (1997) Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J Med Chem*, **40,** 1219-1229.
11. Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*, **46,** 3-26.
12. Newman, D.J. and Cragg, G.M. (2007) Natural products as sources of new drugs over the last 25 years. *J Nat Prod*, **70,** 461-477.
13. Qiao, X., Hou, T., Zhang, W., Guo, S. and Xu, X. (2002) A 3D structure database of components from Chinese traditional medicinal herbs. *J Chem Inf Comput Sci*, **42,** 481-489.
14. Chen, X., Ji, Z.L. and Chen, Y.Z. (2002) TTD: Therapeutic Target Database. *Nucleic Acids Res*, **30,** 412-415.
15. Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B. and Hassanali, M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*, **36,** D901-906.

Lebenslauf

Selbständigkeitserklärung

„Ich, Mathias Dunkel, erkläre, dass ich die vorgelegte Dissertationsschrift mit dem Thema: **3D Konformationsdatenbanken für das *in silico* Screening** selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, ohne die (unzulässige) Hilfe Dritter verfasst und auch in Teilen keine Kopien anderer Arbeiten dargestellt habe."

Berlin, 17.04.08

Unterschrift
Mathias Dunkel