Aus dem Institut für Physiologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Analyse und Vorhersage von Interaktionen von kleinen organischen Molekülen und ihren medizinischen Zielmolekülen

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

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Janette Nickel-Seeber

aus Langen (Hessen)

Datum der Promotion: 25.06.2017

Inhaltsverzeichnis

1 Abstrakt

Ein wichtiger Aspekt während der präklinischen Entwicklung potenzieller neuer Wirkstoffe ist die Vorhersage unerwünschter Nebenwirkungen der zu untersuchenden kleinen organischen Moleküle. Oftmals interagieren kleine organische Moleküle nicht nur mit einem Zielprotein, sondern mit multiplen Proteinen. Die Modulation multipler Proteine kann jedoch zu unerwünschten Nebenwirkungen führen. Die vorliegende Arbeit untersucht Ligand-Protein-Interaktionen mit der Zielsetzung der Entwicklung eines *in silico* Target-Vorhersagealgorithmus. Dieser soll die Vorhersage von Interkationen zwischen kleinen organischen Molekülen und ihren Zielmolekülen, auch Targets genannt, erlauben. Die Target-Vorhersage dient allerdings nicht nur der Nebenwirkungsanalyse, sondern kann zusätzlich Verwendung im Repositioning von Wirkstoffen finden. Darüber hinaus werden Ligand-Protein-Interaktionen der vom Markt genommenen Wirkstoffe analysiert, um die zugrunde liegenden Toxizitäten zu untersuchen, die zur Arzneimittelrücknahme führten. Die gleichzeitige Einnahme verschiedener zugelassener Wirkstoffe kann ebenfalls zu nicht erwünschten Nebenwirkungen führen. Die simultane Metabolisierung der Wirkstoffe durch die gleichen Enzyme während der Biotransformation kann in einer Verstärkung oder Verringerung der Wirkstoffstärke resultieren und ungewollte Nebenwirkungen hervorrufen. Derselbe Effekt kann durch die Inhibition oder Induktion der Enzyme durch Wirkstoffe auftreten, welche die Metabolisierung weiterer Wirkstoffe beeinflussen. Die Entwicklung eines Algorithmus zur Optimierung von Kombinationstherapien hilft unerwünschte Nebenwirkungen zu vermeiden, indem alternative Wirkstoffe angeboten werden. Die angebotenen alternativen Wirkstoffe umgehen dabei die oben dargestellte Problematik. Grundlage zur Erstellung der Algorithmen ist die Generierung integrativer Datenbestände anhand von verschiedenen veröffentlichten Datenbanken sowie mithilfe von Text Mining. Die Integration verschiedener Datenquellen stellt in diesem Zusammenhang eine Herausforderung dar. Die Aufarbeitung der Daten sollte auf eine Weise erfolgen, die eine hohe Zuverlässigkeit gewährleistet und Redundanz innerhalb des integrativen, standardisierten Datenbestandes ausschließt. Auf Basis dieser Datenbestände können Algorithmen zur Identifizierung potenzieller medizinischer Zielmoleküle etabliert werden. Die *in silico* Target-Vorhersagemethode wird durch die Ähnlichkeitsverteilung der bekannten Liganden für ein biologisches Target realisiert. Um die extrahierten sowie vorhergesagten Ligand-Zielmolekül-Interaktionen in einem systembiologischen Ansatz zu verstehen, werden diese auf Stoff- und Signalwege projiziert. Eine Kombination der projizierten Relationen mit Mutations- und Expressionsdaten verschiedener Krebsgewebe erlaubt den Schritt zur personalisierten Medizin.

2 Abstract

Prediction of unwanted adverse drug reactions (ADR) induced by small molecules is an important aspect during preclinical development of new drugs. Usually small molecules interact not only with their main protein target but with multiple protein targets. Modulation of multiple protein targets can lead to unwanted ADR. The present thesis investigates ligand-protein interactions with the aim of developing an *in silico* target prediction algorithm. The algorithm is to predict interactions between small molecules and their target proteins. Target predictions serves not only for analyzing ADR but can also be used for repositioning of drugs. Furthermore, ligand-protein interactions of withdrawn drugs were analyzed to investigate underlying toxicities which led to the drugs' withdrawal. Simultaneous drug ingestion of various approved drugs can induce unwanted ADR likewise. A simultaneous metabolism of drugs by the same enzyme during biotransformation can result in an increase or decrease of a drug's efficacy and thereby induce unwanted ADR. The same effect can emerge through an inhibition or induction of enzymes by drugs. This can influence the metabolism of further drugs. Development of an algorithm to optimize combination therapies aids to avoid unwanted ADR by proposing alternative drugs. The proposed alternative drugs are able to bypass the beforehand described problems. The basis for generating such algorithms is the generation of an integrative dataset established on different published databases and the results of a text mining approach. Integration of various data sources is challenging in this context. Processing of the data should be done in a way that results in a high confidence and excludes redundancy within the integrative, standardized dataset. Algorithms for identifying potential medical targets are established on the basis of the underlying integrative and standardized datasets. The *in silico* target prediction method is realized by the similarity distribution of the known ligands of a biological target. Compound-target-interactions are projected to signaling pathways to understand the extracted and predicted compound-target-interactions in a systems biological approach. A combination of the projected relations with mutation- and expression data of various cancer tissues allows the step to personalized medicine.

3 Einleitung

3.1 Polypharmakologie und Repositioning von Wirkstoffen

Das Paradigma „Ein Wirkstoff - eine Zielstruktur - eine Krankheit“ war lange das gängige Konzept der Wirkstoffentwicklung [1]. Dieses Paradigma besagt, dass ein selektiver Wirkstoff spezifisch für eine individuelle molekulare Zielstruktur, auch „Target“ genannt, entwickelt wird. Dieses Target, das normalerweise ein Protein darstellt, ist wiederum spezifisch mit einem bestimmten Krankheitsbild assoziiert. Komplexe Krankheitsbilder können wie im Fall von Krebserkrankungen multiple molekulare Abnormalitäten aufweisen. Aufgrund dessen tendiert das primäre Paradigma dazu, in Richtung der Paradigmen „Ein Wirkstoff - eine Krankheit - multiple Zielstrukturen“ oder „Multiple Wirkstoffe - eine Krankheit - multiple Zielstrukturen“ verschoben zu werden und dadurch mehr einem systembiologischen Ansatz zu folgen [2].

Polypharmakologie bezeichnet die Fähigkeit eines kleinen organischen Moleküls, auch „Compound“ genannt, mit multiplen Targets zu interagieren. Dies kann sowohl mit therapeutischen Vorteilen als auch Nachteilen einhergehen [3]. Bei einem komplexen Krankheitsbild wie Krebs sind Wirkstoffe, die multiple Targets adressieren, oftmals von Vorteil. Tumore, bei denen nur eine einzelne Kinase inhibiert wird, entwickeln häufig Mechanismen, um der Inhibition entgegenzuwirken. Unter anderem werden Resistenzmutationen ausgebildet [4] oder Ersatzkinasen anstelle der inhibierten Kinase aktiviert [5]. Dies führt dazu, dass die Modulation eines einzelnen Targets keine Wirkung mehr erzielt, falls der „Ein-Target“-Therapie entgegenwirkende Mechanismen auftreten. Wirkstoffe, die simultan unterschiedliche Targets modulieren, werden verwendet, um diese Limitationen zu umgehen. Als Beispiel soll hierbei der Wirkstoff Sunitinib genannt werden. Er interagiert mit multiplen Targets und greift dadurch in diverse Signalwege des Tumors ein [6]. Polypharmakologie kann jedoch auch zu unerwünschten Arzneimittelwirkungen führen. Ungewollte Interaktionen eines Wirkstoffes mit Neben-Zielstrukturen, den sogenannten „Off-Targets“, können häufig zu Nebenwirkungen führen. Der 5-HT_{2B}-Serotoninrezeptor [7] und der Kaliumkanal hERG [8] repräsentieren solche bekannten Off-Targets. Der Appetithemmer Fenfluramin-Phentermin wurde aufgrund der Aktivierung des 5-HT_{2B}-Serotoninrezeptors durch einen seiner Metabolite, Norfenfluramin, im Jahr 1997 vom Markt genommen, da er Herzklappenschäden induziert [7]. Inhibition des Off-Target-Kaliumkanals hERG durch das Antihistaminikum Terfenadin verursacht Herzrhythmusstörungen [8].

Repositioning, auch Repurposing genannt, von Wirkstoffen bezeichnet die Identifizierung wie auch Umpositionierung bekannter Wirkstoffe für neue therapeutische Indikationen. Wirkstoffe,

die für das Repositioning herangezogen werden können, lassen sich in drei Gruppen einteilen: Erstens, Wirkstoffe, die für den Markt zugelassen sind. Zweitens, Wirkstoffe, die bereits für den Markt zugelassen waren, aber aufgrund unterschiedlicher Gründe vom Markt genommen wurden. Drittens, Wirkstoffe, die nach Evaluation innerhalb der klinischen Phasen nicht zugelassen werden konnten [9]. Vorteile dieses Verfahrens ergeben sich durch reduzierte Risiken, da die Repositioning-Kandidaten meistens schon in mehreren klinischen Phasen geprüft wurden [10]. Hierdurch stellt das Repositioning von Wirkstoffen zur gängigen Wirkstoffentwicklung eine kosten- und zeitsparende Alternative dar. Diese Alternative wird durch bereits bekannte Sicherheits-, therapeutische Toleranz- und Wirksamkeitsprofile der Wirkstoffe ermöglicht [11]. Das Konzept der Polypharmakologie findet Verwendung im Repositioning von Wirkstoffen. Hierdurch eröffnen sich, neben der eigentlichen therapeutischen Indikation, Möglichkeiten der Anwendung für weitere Indikationen. Ein bekanntes Beispiel für das Repositioning von Wirkstoffen ist das Schlafmittel Thalidomid. Ursprünglich wurde es u. a. gegen Morgenübelkeit während der Schwangerschaft eingesetzt. Jedoch führte es zu dramatischen Fehlbildungen der Kinder, so dass es vom Markt genommen wurde. Einige Jahre später wurde seine Wirkung zufällig bei Leprakomplikationen nachgewiesen, für die Thalidomid nun wieder vermarktet wird [9]. Neben polypharmakologischen Ansätzen kann auch das Anatomisch-Therapeutisch-Chemische Klassifikationssystem (ATC) der Weltgesundheits-organisation (WHO) zum Repositioning von Wirkstoffen herangezogen werden. Dieses System basiert sowohl auf den therapeutischen, pharmakologischen und chemischen Eigenschaften eines Wirkstoffes als auch auf seinem Applikationsgebiet [12].

3.2 Polypharmazie und ihre klinische Relevanz

Der Begriff der Polypharmazie wird in der Fachliteratur unterschiedlich definiert. Polypharmazie kann als die Applikation multipler Arzneimittel oder die Administration von mehr Arzneimitteln als klinisch indiziert verstanden werden [13]. Die Weltgesundheitsorganisation (WHO) definiert Polypharmazie als gleichzeitige Applikation von vielen Arzneimitteln oder Applikation einer exzessiven Anzahl an Arzneimitteln [14]. Eine weitere Definition beschreibt die regelmäßige und gleichzeitige Einnahme von fünf oder mehr Arzneimitteln [15]. Insbesondere in der älteren Bevölkerungsschicht tritt Polypharmazie aufgrund multipler vorliegender Erkrankungen der älteren Patienten häufig auf. Perspektivisch führt der demographische Trend, der sich momentan in der Gesamtbevölkerung andeutet, zu einem Anstieg der Lebenserwartung und folglich auch zu einer Zunahme der Multimorbidität sowie der Polypharmazie [16].

Eine Vielzahl von Arzneistoffen durchläuft den Vorgang der Biotransformation. Während der Biotransformation von mehreren gleichzeitig applizierten Arzneistoffen können Interaktionen auftreten, welche die Wirkstoffstärke der Arzneistoffe beeinflussen wie auch unerwünschte Nebenwirkungen verursachen. Die Biotransformation wird in unterschiedliche Phasen eingeteilt und dient der Umwandlung von nicht ausscheidbare in ausscheidbare Stoffe. In Phase-I-Reaktionen werden funktionelle Gruppen in die Substanzen eingeführt oder freigesetzt [17]. Diese Reaktion wird hauptsächlich durch Enzyme der Cytochrom P450-Familie katalysiert. In Phase-II-Reaktionen hängen Transferasen polare endogene Moleküle an die Xenobiotika an [18]. Anhand beider Reaktionen wird eine erhöhte Hydrophilie der Substanzen erzielt, aufgrund derer sie aus der Zelle transportiert werden müssen, um schließlich final aus dem Körper ausgeschieden zu werden. Wenn Arzneistoffe als Prodrugs vorliegen, werden diese erst durch Reaktionen der Biotransformation in ihre reaktive Metaboliten überführt [19]. Verschiedene Interaktionen können einen Einfluss auf die Metabolisierung der Substanzen haben. Zum einen können exogene Substanzen durch Inhibition oder Induktion von Enzymen der Biotransformation auf die Metabolisierungsrate weiterer Substanzen einwirken. Zum anderen kann die Konkurrenz von Substanzen um das gleiche Enzym die Metabolisierung beeinflussen. Das Prokinetikum Cisaprid kann Arrhythmie auslösen, wenn es gleichzeitig mit dem Antibiotikum Erythromycin appliziert wird [20]. Beide Xenobiotika werden über das Cytochrom P450 3A4 metabolisiert, wobei der Erythromycin-Metabolit einen Komplex mit dem Cytochrom P450 3A4-Enzym bildet und es auf diese Weise inhibiert [21]. Aufgrund dessen steigt die Serumkonzentration von Cisaprid, die eine Inhibierung des Kaliumkanals hERG nach sich zieht und somit eine schwerwiegende Nebenwirkung, eine Arrhythmie, hervorrufen kann [22].

3.3 Personalisierte Krebstherapie

Ursprünglich sind Krebstherapien nur auf Grundlage des Ursprungsgewebes des Tumors, der Größe und des Gewichts eines Patienten an diesen angepasst. Im Gegensatz dazu, bieten personalisierte Therapien die Möglichkeit einer individuell an Patientengruppen angepassten Therapie [23]. Sie basieren mitunter auf der genetischen Zusammenstellung des Tumors, wie auch auf vererbten genetischen Varianten der Personen, welche die Wirksamkeit von Wirkstoffen und folglich die klinische Reaktion des Tumors beeinflussen können. Hierzu werden experimentelle Genomdaten herangezogen sowie analysiert, um Auswirkungen der identifizierten Alterationen auf die Krebsbehandlung vorherzusagen. Des Weiteren können vergleichende Analysen zwischen Tumor und normalem gesunden Gewebe durchgeführt werden. Die Erkenntnisse, die aus diesen Analysen gewonnen werden, können für eine

effektivere, gezieltere und nebenwirkungsärmere Behandlung der jeweiligen Patienten verwendet werden [24][25]. So wurde für Mutationen im Protoonkogen KRAS nachgewiesen, dass diese ein Marker für die Resistenz von EGFR-spezifischen Antikörpern Cetuximab und Panitumumab sind [26]. Andererseits erhöht Vemurafenib, ein Inhibitor der Proteinkinase BRAF, die Überlebensrate von Personen, die an Melanom erkrankt sind und die BRAF-Mutation V600E besitzen [27].

3.4 Arzneimittelrücknahmen

Bevor ein experimenteller Wirkstoff für den Markt zugelassen werden kann, durchläuft er mehrere klinische Studien zur Überprüfung seiner Wirksamkeit und Unbedenklichkeit. Nach kritischer Beurteilung von Nutzen und Schaden des Wirkstoffes wird eine Zulassung zur Vermarktung des Wirkstoffes ausgesprochen oder verwehrt. Nach Vermarktungsfreigabe treten häufig unvorhergesehene Nebenwirkung auf, die nicht während der klinischen Studien zu beobachten waren [28]. Die meisten Wirkstoffe werden in klinischen Studien an etwa 1.000 bis 5.000 ausgewählten Patienten untersucht. Die Chance seltene, aber wichtige unerwünschte Arzneimittelwirkungen im Vergleich zu Kontrollgruppen zuverlässig zu beobachten, ist anhand der begrenzten Patientenkohorten gering [29]. Allerdings können verschiedene Maßnahmen ergriffen werden, wenn Nebenwirkungen des bereits zugelassenen Arzneistoffes auftreten: die neue Kennzeichnung der Arzneistoffe mit spezifischen Warnhinweisen, der Vermerk neuer Kontraindikationen, die Verbreitung internationaler Mitteilungen über Arzneimittelrisikos durch „Direct Healthcare Professional Communication (DHCP)“, die individuelle Entscheidung des Patienten, ob der Arzneistoff genommen wird und die Revision der Vermarktung des Arzneistoffes bei schwerwiegenden Nebenwirkungen [30]. Analysen der vom Markt zurückgezogenen Wirkstoffe und der zugrunde liegenden Mechanismen, die Nebenwirkungen auslösen können, bieten Möglichkeiten zur frühzeitigen Optimierung von experimentellen Wirkstoffen innerhalb der präklinischen Phase.

4 Zielstellung

Ziel dieser Dissertation ist die Entwicklung von Methoden zur Vorhersage von Interaktionen von kleinen organischen Molekülen mit ihren medizinischen Zielmolekülen. Grundlage für die Entwicklung der Algorithmen ist die Analyse der Substanz-Protein-Interaktionen wie auch die Erzeugung eines umfangreichen, integrativen Datenbestandes, der über aufbereitete, standardisierte und gefilterte Daten verfügt. *In silico* Methoden, die Target-Vorhersagen für

kleine organische Moleküle ermöglichen, bieten eine Unterstützung während der präklinischen Entwicklung möglicher Wirkstoffe. Auf diese Weise sollen etwaige unerwünschte Nebenwirkungen durch unerwartete Wechselwirkungen der experimentellen Wirkstoffe mit Off-Targets frühzeitig vermieden werden, da in deren Folge etwa 20 Prozent der Wirkstoffe nicht zugelassen werden können [31]. Die Verwendung computergestützter Methoden während der präklinischen Phase lässt eine zeit-, arbeits- und finanziell effizientere Entwicklung von Arzneimitteln zu. Dies geschieht indem Compounds in einer frühen Entwicklungsphase ausgeschlossen oder optimiert werden können, um dementsprechend Wechselwirkungen mit unerwünschten Off-Targets zu verhindern. Der Einsatz von *in silico* Methoden zur Target-Vorhersage kann allerdings auch bei der Identifizierung gewünschter Off-Targets eines potenziellen Wirkstoffes mit dem Ziel einer verbesserten therapeutischen Wirksamkeit und Sicherheit Verwendung finden [32]. Des Weiteren soll die entwickelte Methode im Repositioning von Wirkstoffen von Nutzen sein. Die Entwicklung eines Algorithmus zur Optimierung von Kombinationstherapien soll sich unterstützend auf die Pharmakotherapie auswirken, mit dem Ziel, unerwünschte Nebenwirkungen zu reduzieren. Interaktionen, die Auswirkungen auf die Metabolisierung von Wirkstoffen haben, werden erkannt und alternative Wirkstoffe können für die Arzneimittel der bestehenden Medikamentenkombination angeboten werden. Integrative Datenbestände sollen für Methoden zur personalisierten Krebstherapie erstellt werden. Mit ihnen soll ein Abgleich von Expressions- und Mutationsprofilen z. B. von Krebszelllinien mit denen in der Datenbank erhältlichen Daten ermöglicht werden. Dies erlaubt die Identifizierung sowohl ähnlicher Krebszelllinien als auch effektiver Wirkstoffe. Zusätzlich soll eine Datengrundlage für weiterführende Analysen der vom Markt zurückgezogenen Wirkstoffe, assoziierten Toxizitäten sowie Interaktionen generiert werden. Diese Informationen können in den Wirkstoffentwicklungsprozess eingebaut werden und frühzeitig die Identifizierung von Toxizitäten von experimentellen Wirkstoffen ermöglichen. Die integrativen Datenbestände und entwickelten *in silico* Methoden sollen der wissenschaftlichen Gemeinschaft online über eigens realisierte Internetseiten zur Verfügung gestellt werden.

5 Methodik

5.1 Textmining – Wissensgewinnung durch Extraktion von Informationen

Textmining ist ein Verfahren, um Informationen aus unstrukturierten Texten zu extrahieren. Das Textmining-Verfahren wurde für die Erstellung des Primärdatensatzes für die Originalarbeit der Transformer Datenbank [33] verwendet. Hierfür wurde die Literaturdatenbank PubMed/Medline

als Datenquelle verwendet, um alle Abstracts (ca. 21 Millionen; Stand Sept. 2013) nach Interaktionen zwischen Wirkstoffen und Enzymen der Biotransformation zu durchsuchen. Dies geschah anhand von vordefinierter Suchbegriffe wie Wirkstoff- und Enzymnamen, deren Synonymen wie auch anderweitigen Suchbegriffen, die Relationen zwischen den Wirkstoffen und Enzymen darstellen. Die im XML-Format vorliegenden Abstracts wurden anhand der Apache Lucene (http://lucene.apache.org) Programmbibliothek prozessiert. Weiterführend konnten die Daten mit LingPipe (http://alias-i.com/lingpipe) indiziert werden, so dass Relationen innerhalb des Datensatzes identifiziert werden konnten. Anschließend wurden die Ergebnisse mithilfe der Abstracts oder der Volltexte der Veröffentlichungen manuell validiert.

5.2 Datenintegration zur Erstellung anwendungsorientierter Wissensrepräsentationen

Aufgrund hochmoderner Hochdurchsatz-Screening-Verfahren steigt die Anzahl der Informationen über Compound-Target-Interaktionen. Diese Informationen sind dezentral in verschiedenen öffentlich zugänglichen Datenbanken vorzufinden. Sowohl die Erstellung anwendungsorientierter und umfangreicher Wissensrepräsentationen für spezifische Fragestellungen als auch die Anwendung durch Integration dieser Informationen ermöglicht die Bereitstellung eines flächendeckenden Wissens der öffentlich existierenden Informationen. Die Schwierigkeit bei der Integration aus verschiedenen Datenquellen ergibt sich aus der Heterogenität und Überlappung der Daten. Um Redundanz in dem integrativen Datenbestand zu vermeiden, ist eine Vorverarbeitung der Daten notwendig. Die zugrunde liegenden Daten aus Datenbanken wie ChEMBL [34], SuperTarget [35], BindingDB [36] oder DrugBank [37] wurden zunächst standardisiert. Hierzu wurden die Compounds normalisiert und anhand des International Chemical Identifiers (InChI) identifiziert, um sie datenbankübergreifend miteinander zu vergleichen sowie identische Strukturen zu bestimmen. Targets wurden anhand des Entrez Gene Index des National Center for Biotechnology Information (NCBI) vereinheitlicht. Eine Zusammenführung der Daten unter Berücksichtigung von identischen Interaktionen vermeidet das Auftreten von Redundanz innerhalb des Datensatzes. Weitere Schritte beinhalten die Filterung des Datensatzes, die je nach Fragestellung unterschiedlich durchgeführt wurde. Mitunter wurde nach Bioaktivitätstypen, Organismen und molekularen Target-Typen gefiltert. Genexpressionsdaten der Konsortien COSMIC (Catalogue Of Somatic Mutations In Cancer) [38], CCLE (Cancer Cell Line Encyclopedia) [39] und der CellMiner Datenbank [40] wurden skaliert und zentriert zusammengeführt. Weiterführende Details sind in den Originalarbeiten [41] - [43] nachzulesen.

5.3 Target-Vorhersage anhand eines Ensemble-Ansatzes

Strukturell ähnliche Compounds haben laut des „Similar Property Principle“ [44] häufig ähnliche Eigenschaften, die zu einer ähnlichen biologischen Aktivität führen. Dieses Prinzip hat man sich für die Target-Vorhersage zu eigen gemacht. Ein Compound wird bei der Target-Vorhersage anhand eines Ensemble-Ansatzes mit Ligandengruppen von Targets verglichen und unter Anwendung statistischer Verfahren diesen Targets als potenzieller Ligand zugewiesen. Für jedes Target wurde hierfür ein sogenanntes Target-Set an Liganden erstellt. Dazu wurden Compound-Target-Interaktionen aus verschiedenen Datenbanken integriert wie auch gefiltert (siehe Originalarbeit [41]). Compounds, die mit einem Target interagieren, bilden das Target-Set für dieses Target. Zur Bestimmung der zwei-dimensionalen strukturellen Ähnlichkeit des Ensembles wurde für jedes Compound ein Extended-Connectivity Fingerprint mit einem Radius von vier (ECFP4) [45] berechnet, um den Tanimoto Koeffizienten [46] der paarweisen Vergleiche zu bestimmen. Für jedes Target-Set wurden schließlich die Tanimoto Koeffizienten summiert, die oberhalb eines Grenzwertes von 0,45 liegen (Raw Score). Anschließend wurde diese Summe anhand der Anzahl der Liganden normalisiert, um Vergleiche zwischen unterschiedlich großen Target-Sets zu ermöglichen. Um die Signifikanz dieser Werte zu prüfen, wurden Z-Scores und Erwartungswerte (E-Werte) berechnet. Da der Z-Score sich bei diversen Target-Sets wie bei zufälligen Vorhersagen verhält, wurde ein Gewichtungsfaktor (λ) eingeführt.

$$Z_A = \frac{\left(\frac{Raw\ Score_A}{N_A} - \mu\right) \exp(0{,}335 \ln(N_A)}{\sigma} * \lambda_A$$

Hier repräsentiert A das Target, N_A die Ligandenanzahl des Target-Sets, μ den Mittelwert der normalisierten Raw Scores und σ die Standardabweichung des Z-Scores.

$$\lambda_A = \exp(0{,}335\ ln\left(\frac{Raw\ Score_{AA}}{N_{AA}}\right))$$

Hierbei stellt $Raw\ Score_{AA}$ die Summe der Tanimoto-Koeffizienten oberhalb des Grenzwertes von 0,45 innerhalb eines Target-Sets und N_{AA} die Anzahl der Liganden des Target-Sets dar.
Der E-Wert beschreibt die Anzahl der vorhergesagten Targets, die zufällig vorhergesagt werden. E-Wert und Z-Score reagieren gegensätzlich. Der E-Wert fällt, wenn der Z-Score ansteigt. Die Vorhersage ist signifikanter, je niedriger der E-Wert ist, wobei ein E-Wert von größer als eins auf eine zufällige Vorhersage hindeutet.

5.4 Wirkstoffklassifizierung durch Ähnlichkeitsuntersuchungen

Um ATC-Codes für (potenzielle) Wirkstoffe vorherzusagen, wurden Methoden unter Berücksichtigung von zweidimensionaler, Fragment- und dreidimensionaler Ähnlichkeit zur Bestimmung von strukturellen Ähnlichkeiten angewandt und miteinander kombiniert. Ein Datensatz von 2.650 Wirkstoffen mit zugewiesenen ATC-Codes der Transformer Datenbank [33] wurde hierfür verwendet. Zum Vergleich der zweidimensionalen strukturellen Ähnlichkeit wurden drei molekulare Fingerprints für die Compound-Strukturen berechnet sowie verglichen. Die zweidimensionale strukturelle Ähnlichkeit wurde anhand des Tanimoto-Koeffizienten [46] festgestellt.

$$Tanimoto - Koeffizient_{AB} = \frac{AB}{A + B - AB}$$

Der Tanimoto-Koeffizient verwendet Bits, die in den Fingerprints auf 1 gesetzt wurden. *AB* stellt die Anzahl der Bits dar, die in den Fingerprints von Molekül *A* sowie *B* auf 1 gesetzt wurden. *A* stellt die Anzahl der Bits in dem Fingerprint von Molekül *A* und *B* stellt die Anzahl der Bits in dem Fingerprint von Molekül *B* dar, die auf 1 gesetzt wurden.

Darüber hinaus wurden alle 2.650 Strukturen entsprechend der Linker Regel [47] fragmentiert. Zur Bestimmung des Tanimoto-Koeffizienten der Fragmente zweier Moleküle *A* und *B*, wobei *n* und *m* nichtredundanten Fragmente sind, wurde eine Ähnlichkeitsmatrix erstellt, die $\binom{n}{m}$ mögliche Kombinationen enthält. Die Summe der Tanimoto-Koeffizienten der Ähnlichkeitsmatrix aus dem Vergleich der nichtredundanten Fragmente zweier Moleküle wurde anschließend durch die kleinere Anzahl der Fragmente eines der beiden Moleküle normalisiert.

Die dreidimensionale Ähnlichkeit von Molekülen wurde anhand eines Superimpositionsalgorithmus bestimmt [48]. Die Qualität der Überlagerung wird durch den „root-mean-square-deviation“ (RMSD) beschrieben. Folgende Formel definiert die beste Überlagerung/Superimposition:

$$3D - Score = \frac{N_S}{\max(N_A \times N_B)} \exp(-RMSD)$$

N_S beschreibt die Anzahl der überlagerten Atome, N_A die Anzahl der Atome von Molekül *A* und N_B die Anzahl der Atome von Molekül *B*.

Der Konsensus der drei Methoden wird zur ATC-Vorhersage verwendet. Wenn zwei der drei Methoden dieselbe ATC-Klasse bestimmen, wird diese für das Molekül vorhergesagt. Wenn alle drei Methoden unterschiedliche ATC-Klassen berechnen, so wird ein methodenabhängiger Grenzwert zur Bestimmung der ATC-Klasse verwendet.

6 Ergebnisse

6.1 Target-Vorhersage und Repositioning von Wirkstoffen

Zur Vorhersage von Off-Targets wird eine breite, qualitativ hochwertige Datengrundlage benötigt. Dies macht eine Datenintegration von Interaktionsdaten unterschiedlicher Quellen unabdingbar. Basierend auf den etwa 12,5 Millionen Interaktionen, die aus den Datenbanken ChEMBL [34], SuperTarget [35] und BindingDB [36] zusammengetragen wurden, konnte ein integrativer, standardisierter und nichtredundanter Datensatz erstellt werden. Dieser Datensatz enthält nach Filterung der Daten etwa 665.000 Interaktionen, 341.000 Compounds und 1.800 Targets. Auf Grundlage dieses Datensatzes wurde die Target-Vorhersage implementiert, die eine Vorhersagerate von 92,8 Prozent erzielt. Diese kann sogar auf 94,1 Prozent gesteigert werden, wenn ein Grenzwert von eins für den E-Wert eingeführt wird, da ein E-Wert von über eins eine zufällige Vorhersage kennzeichnet. Die entwickelte *in silico* Target-Vorhersage wurde über ein eigens entwickeltes Web-Interface der wissenschaftlichen Gemeinschaft zur Verfügung gestellt. Der skizzierte Ablauf der Target-Vorhersage wird in Abbildung 1 dargestellt.

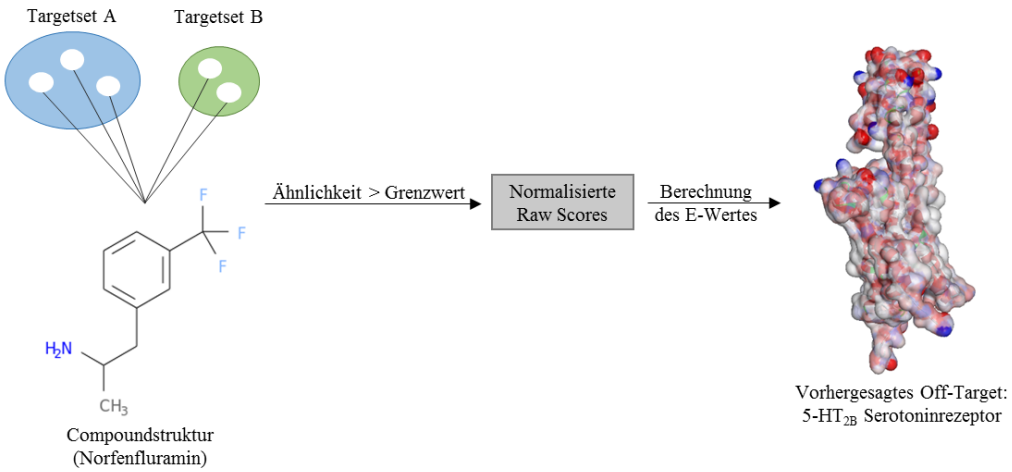

Abbildung 1: Skizzierter Ablauf der Target-Vorhersage am Beispiel des Fenfluramin-Phentermin Metaboliten Norfenfluramin. Nach Berechnung des ECFP4 Fingerprints für die Compoundstruktur Norfenfluramin wird die paarweise zweidimensionale Ähnlichkeit zwischen der Compoundstruktur und den Liganden der Target-Sets der in der Datenbank vorhandenen Proteine durchgeführt. Die Ähnlichkeiten oberhalb eines definierten Grenzwertes werden zum Raw Score summiert sowie normalisiert. Anhand des Raw Scores werden Erwartungswerte für die jeweiligen Proteine berechnet. Als Off-Target wird der 5-HT_{2B}-Serotoninrezeptor vorhergesagt. Die Aktivierung des Off-Targets 5-HT_{2B}-Serotoninrezeptor durch den Fenfluramin-Phentermin Metabolit Norfenfluramin induziert Herzklappenschäden und führte zur Arzneimittelrücknahme des Appetithemmers Fenfluramin-Phentermin [7].

Für ein Compound werden Informationen zu physikalischen und chemischen Eigenschaften berechnet sowie schon bekannte Compound-Target-Interaktionen und vorhergesagte Compound-Target-Interaktionen samt E-Wert aufgelistet. Zusätzlich werden detaillierte Informationen zu den Targets angegeben. Darüber hinaus bietet das Web-Interface die Möglichkeit Vorhersagen zu der ATC-Klassifizierung bis zum vierten Level und somit auch zu der Indikation des Compounds zu tätigen. Hierdurch wird das Repositioning anhand von Neuklassifizierungen von (experimentellen) Wirkstoffen unterstützt. Die ATC-Klassifizierungsvorhersage beruht auf einem Datensatz, der 2.650 Wirkstoffe inklusive zugewiesener ATC-Klassifizierung enthält. Die Vorhersagerate der Wirkstoffklassifizierung beträgt 75,1 Prozent für die kombinierte Methode, die zweidimensionale, Fragment- und dreidimensionale Ähnlichkeitsberechnungen vereint. Sie erreicht somit um eine bis sieben Prozent exaktere Vorhersage als die jeweiligen Ähnlichkeitsmethoden separat. Der schematische Ablauf der Vorhersage ist in Abbildung 2 abgebildet. Auf dem Web-Interface werden neben der Vorhersagegenauigkeit, der vorhergesagten ATC-Klassifizierung für das Compound, Informationen zu physikalischen und chemischen Eigenschaften und die Erfüllung der Kriterien der „Lipinski Rule of Five“ [49] angezeigt.

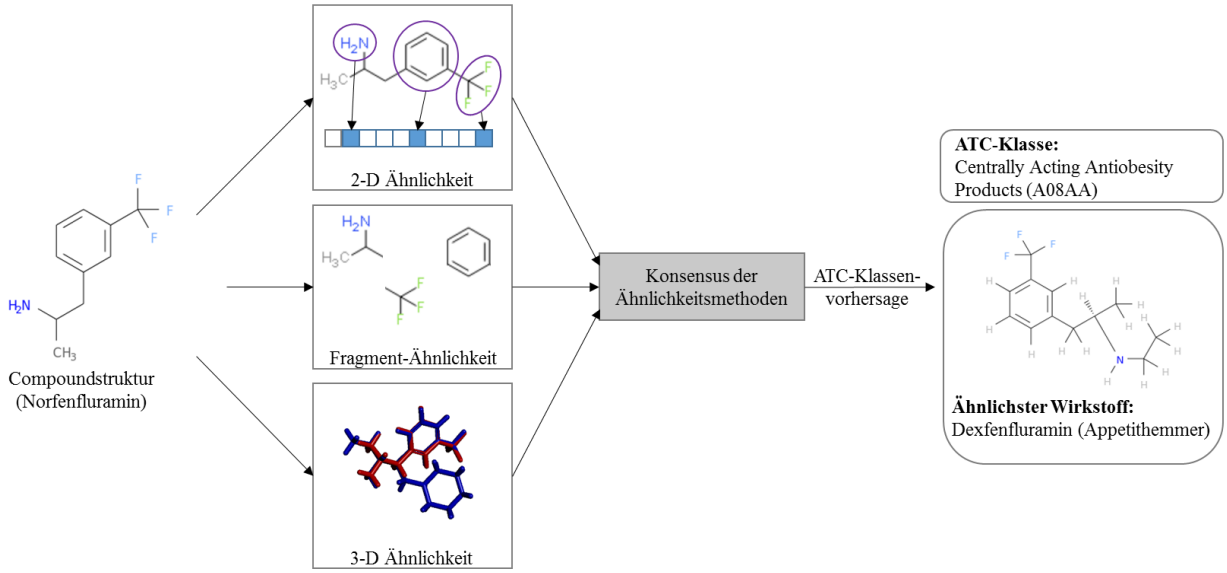

Abbildung 2: Skizzierter Ablauf der ATC-Vorhersage am Beispiel des Fenfluramin-Phentermin Metaboliten Norfenfluramin. Drei Methoden kommen zur Bestimmung der ATC-Klasse zum Einsatz: 1. Berechnung der paarweisen zweidimensionalen Ähnlichkeit des Compounds zu 2.650 Wirkstoffen. 2. Berechnung der Ähnlichkeit der Fragmente zwischen dem Compound und den in der Datenbank enthaltenen Wirkstoffen. 3. Berechnung der paarweisen dreidimensionalen Ähnlichkeit zwischen dem Compound und den 2.650 Wirkstoffen. Der gebildete Konsensus dieser drei Methoden wird zur Vorhersage der ATC-Klasse verwendet. Am Beispiel von Norfenfluramin wird der Wirkstoff Dexfenfluramin als ähnlichster Wirkstoff und die ATC-Klasse „A08AA“ identifiziert.

Die Lipinski Rule of Five bestimmt, ob ein Compound als oral verabreichtes Arzneimittel Verwendung finden kann. Hierzu wird das Molekulargewicht (< 500 g/mol), der logarithmierte Oktanol-Wasser-Verteilungskoeffizient (≤ 5), die Anzahl der Wasserstoffbrückenakzeptoren (≤ 10) und die Anzahl der Wasserstoffbrückendonoren (≤ 5) berücksichtigt. Der Benutzer hat die Möglichkeit, die fünf ähnlichsten Wirkstoffe und zugewiesenen ATC-Klassifizierungen zu betrachten, die anhand der kombinierten Ähnlichkeitsmethode eruiert wurden. Diese können Erkenntnisse über neue Indikationen der Compoundstruktur aufzeigen. Des Weiteren bietet das Web-Interface die Möglichkeit, alle ATC-Klassifizierungen der, in der Datenbank enthaltenen, Wirkstoffe mithilfe eines ATC-Baumes zu betrachten.

6.2 Optimierung von Pharmakotherapien

Nach manueller Validierung der Textmining-Ergebnisse konnten 5.595 Relationen zwischen 2.801 Wirkstoffen, und 155 Proteinen der Biotransformation identifiziert werden. Zusätzlich wurden 350 Nahrungsmittel identifiziert, die einen Einfluss auf die Biotransformation haben. Alle Daten wurden in die Datenbank Transformer aufgenommen und der wissenschaftlichen Gemeinschaft über ein Web-Interface zur Verfügung gestellt. Zu jedem Wirkstoff werden physikalische wie auch chemische Eigenschaften und ATC-Klassifizierungen der Wirkstoffe angegeben. Halbwertszeiten und Q_0-Werte wurden, wenn vorhanden, ebenfalls in die Datenbank aufgenommen. Neben Wirkstoffen sind auch 125 Prodrugs in der Datenbank enthalten, die erst nach Metabolisierung in den aktiven Metaboliten umgewandelt werden und Informationen zur ihrer Aktivierung aufführen. Ebenso werden Informationen zu den Enzymen der Biotransformation zur Verfügung gestellt. Darüber hinaus werden Relationen zwischen den Enzymen und den Wirkstoffen inklusive der jeweiligen Referenzen gelistet, aus denen die Interaktionen extrahiert wurden. Anhand der extrahierten Wirkstoff-Protein-Interaktionsinformationen konnte eine Methode zur Optimierung von Kombinationstherapien entwickelt werden, die über das dynamisch agierende „Cocktail-Tool“ des Web-Interfaces verfügbar ist (Abbildung 3).

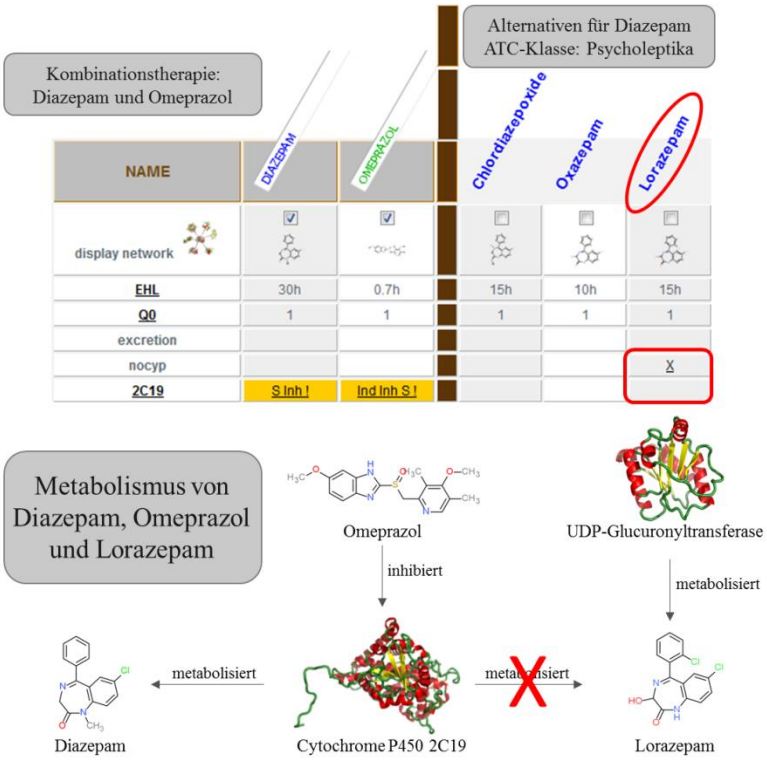

Abbildung 3: Vereinfachte Darstellung der Funktionalität des „Cocktail-Tools“ anhand eines aus zwei Medikamenten bestehenden Medikamentencocktails. Diazepam wird hauptsächlich durch das Cytochrom P450 2C19-Enzym (CYP2C19) metabolisiert. Bei gleichzeitiger Applikation mit Omeprazol sinkt die Plasmakonzentration des Diazepam-Metaboliten Desmethyldiazepam. Dies ist auf die Inhibition des CYP2C19 durch Omeprazol zurückzuführen. Diazepam- und Omeprazol-Interaktionen mit CYP2C19 werden farblich in Gelb hervorgehoben. Alternativen können anhand von Indikationen durch ATC-Klassen gewählt werden. Lorazepam wird nicht durch Cytochrom P450-Enzyme metabolisiert (siehe roter Kasten), sondern durch UDP-Glucuronyltransferasen. Aufgrund dessen ist es nicht durch die CYP2C19-Inhibition betroffen. Metabolismusbeispiel herausgenommen aus [50].

Bei Konflikten besteht die Möglichkeit, für jede Indikation interaktiv alternative Wirkstoffe aus denselben ATC-Klassifizierungsgruppen zu wählen, die über andere Enzyme der Biotransformation metabolisiert werden. Ein Konflikt besteht einerseits wenn Wirkstoffe Biotransformationsenzyme inhibieren oder induzieren und einer dadurch ausgelösten veränderten Metabolisierungsrate weiterer Wirkstoffe. Andererseits besteht ein Konflikt, wenn Wirkstoffe um das gleiche Enzym konkurrieren, wodurch Metabolisierungsraten der Wirkstoffe beeinträchtigt werden können.

6.3 Ein Schritt zur personalisierten Krebstherapie

Da Krebs eine genetische Erkrankung ist, können sich die Charakteristika der Mutationsprofile von Tumoren selbst in ein und derselben Krebsart von Patient zu Patient unterscheiden. Aufgrund dieser genomischen Alterationen kann die Wirkungskraft eines Wirkstoffes beeinträchtigt sein oder sogar zu Resistenzen führen und somit das klinische Ansprechen

beeinflussen. Die Datenbank CancerResource enthält Informationen zu Compound-Protein-Interaktionen, Genexpressions- und Mutationsdaten von Krebszelllinien. Sensitivitätsprofile von Compounds auf Krebszelllinien wurden zur Erstellung eines zellulären Fingerprints verwendet [51]. Der zelluläre Fingerprint dient der Ermittlung von alternativen und effizienten Substanzen. Die Targets und Compounds wurden auf krebsrelevante Stoff- und Signalwege der KEGG PATHWAY Datenbank [52] projiziert, um sie in einem kontextabhängigen, systembiologischen Ansatz interpretieren zu können. Zusätzlich werden Genexpressions- und Mutationsprofile der Targets in Krebszelllinien dargestellt. Eine Besonderheit stellt die Möglichkeit dar, Samples von Genexpressions- oder Mutationsprofilen hochzuladen, um ähnliche Krebszelllinien durch einen Datenbankabgleich zu bestimmen. Informationen zu effektiven Compounds und Wirkstoffen werden für die ähnlichen Krebszelllinien des hochgeladenen Samples zur Verfügung gestellt (Abbildung 4).

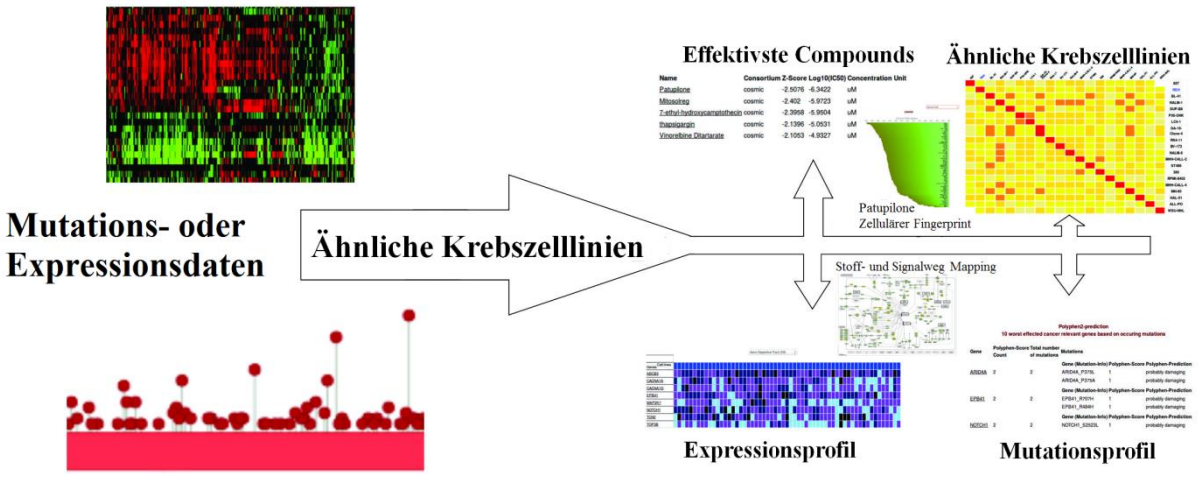

Abbildung 4: CancerResource-Uploadfunktion von Mutations- oder Expressionsprofilen zur Ermittlung ähnlicher Krebszelllinien. Basierend auf den ähnlichen Krebszelllinien werden die effektivsten Compounds, ähnliche Krebszelllinien, Expressions- und Mutationsprofile zu anderen Krebszelllinien, Stoff- und Signalwege dargestellt. Grafik angepasst aus [42].

6.4 Arzneimittelrücknahmen und unerwünschte Nebenwirkungen

Arzneimittelrücknahmen treten bei etwa einem Prozent der bereits zugelassenen Arzneimittel aufgrund unerwünschter Arzneimittelwirkungen auf [31]. Nach einer umfassenden Literatur- und Datenbankrecherche konnten 578 vom Markt zurückgenommene Wirkstoffe identifiziert und in die Datenbank WITHDRAWN integriert werden. Der Fokus der Datenbank liegt auf Wirkstoffen, die aufgrund von Nebenwirkungen sowie toxischen Effekten in mindestens einem Land zurückgezogen wurden. Die Datenbank soll zum Verständnis der Toxizitätsmechanismen beitragen, um diese schon während der Entwicklung von potenziellen Wirkstoffen vermeiden zu

können. Diesbezüglich wurden 946 humane therapeutische Targets und Off-Targets, genetische Variationen der Targets, 14 unterschiedliche Toxizitätstypen und 149 biologische Signalwege aufgenommen, die mit den zurückgezogenen Wirkstoffen assoziiert sind. Genetische Variationen, die eine Rolle bei der Ausbildung von Toxizitäten spielen können, wurden in Form von Einzelnukleotid-Polymorphismen (SNP) von der dbSNP Datenbank [53] extrahiert. Es konnten 27.790 Einzelnukleotid-Polymorphismen für 889 humane Targets identifiziert werden. Metadaten wie ATC-Klassifizierungen der zurückgezogenen Wirkstoffe oder Identifizierer für die Targets wurden, wenn vorhanden, den Daten hinzugefügt.

7 Diskussion

Hochmoderne Hochdursatz-Screening-Verfahren liefern eine stetig steigende Anzahl an Daten über Compound-Target-Interaktionen. Zusätzlich lassen sich Compound-Target-Interaktionen im Überfluss in veröffentlichten Publikationen finden. Diese Datenbestände wurden in unterschiedliche Datenbanken aufgenommen, um eine bessere Auffindbarkeit und Nutzbarkeit dieser Daten zu ermöglichen. Beispielsweise verfügt ChEMBL [34] über fast 14 Millionen Bioaktivitätsdaten, 1,6 Millionen Compounds sowie etwa 11.000 Targets. Die erste öffentlich zugängliche Ligand-Protein-Datenbank BindingDB [36] enthält etwa 1,2 Millionen Bioaktivitätsdaten für 500.000 Compounds und 6.000 Targets (Stand: Juni 2016). Da auch immer mehr Daten publiziert werden, bemühen sich ChEMBL und BindingDB darum, sich bezüglich ihrer Textmining-Unternehmungen besser zu koordinieren und sich folglich auf unterschiedliche wissenschaftliche Zeitschriften zu konzentrieren [54]. Die Fülle an Informationen, die diese und noch weitere öffentliche Datenbanken bieten, kann für die präklinische Entwicklung von Wirkstoffen von Nutzen sein. Mithilfe dieser Informationen können integrative und umfangreiche Datengrundlagen geschaffen werden, die ihren Schwerpunkt auf spezifische Fragestellungen legen. Allerdings ist man bei der Integration von Interaktionsdaten aus unterschiedlichen Datenbeständen vor einige Schwierigkeiten gestellt. Die zu integrierenden Datenmengen können über doppelte Einträge verfügen, die zunächst identifiziert und zusammengeführt werden müssen. Zusätzlich können multiple Messungen zu derselben Compound-Target-Interaktion ambivalente Ergebnisse liefern. Beispielsweise ist die Bestimmung von IC_{50} Werten (mittlere inhibitorische Konzentration) Assay-spezifisch, so dass unter bestimmten, voneinander unabhängigen Bedingungen Messwerte für dieselbe untersuchte Compound-Protein-Interaktion voneinander abweichen können [55]. Häufig ist eine Reproduzierbarkeit von Resultaten zwischen Laboren nicht gegeben. Das TDR (Special Program

for Research and Training in Tropical Diseases) der WHO hat aufgrund dessen Leitlinien zur sogenannten „Guten Laborpraxis“ (GLP, engl.: Good Laboratory Practice) festgelegt. Eine umfassende, zuverlässige Integration und Filterung der Daten ist demnach essentiell, um spezialisierte Wissensrepräsentationen und darauf fundierte Methoden zu entwickeln. Mithilfe spezialisierter Wissensrepräsentationen konnten im Rahmen dieser Dissertation Methoden zur Vorhersage von (Off-)Targets implementiert und evaluiert werden. Die Entwicklung eines Wirkstoffes bis zu seiner Zulassung kann zwischen zehn und 17 Jahren dauern. Eine frühzeitige Identifizierung unerwünschter Nebenwirkungen, die durch den potenziellen Wirkstoff ausgelöst werden können, liegt demnach im ökonomischen Interesse der Pharmaindustrie. Etwa 20 Prozent der experimentellen Wirkstoffe können infolge von in klinischen Prüfungen festgestellten Nebenwirkungen nicht zugelassen werden [31]. Vorhersagemöglichkeiten von Off-Targets könnten Nebenwirkungen vermeiden, indem sie die Möglichkeit bieten, sicherere Compounds in der präklinischen Forschung zu wählen. Die in dieser Arbeit vorgestellte Methode zur Target-Vorhersage beruht auf der chemischen Ähnlichkeit der Liganden, die mit einem Target interagieren. Die Validierung der Methode mit internen und externen Datensätzen demonstriert ihre Stärke zur Vorhersage von Compound-Target-Interaktionen im Vergleich zu anderen bekannten Methoden. Die Methode kann sowohl zur Off-Target-Vorhersage als auch zur Bestimmung neuer therapeutischer Targets für einen (experimentellen) Wirkstoff verwendet werden. Dadurch ist eine Grundlage zum Repositioning von bereits zugelassenen Arzneimitteln ermöglicht worden. Repositioning von bereits zugelassenen Wirkstoffen stellt eine alternative und effiziente Anwendung in der Wirkstofffindung dar, da hier bereits klinische Prüfungen des Wirkstoffes erfolgt sind. Der Ansatz der Target-Vorhersage auf der Grundlage der chemischen Ähnlichkeit von Ligandenensembles verfügt allerdings über Restriktionen. So kann es trotz hoher struktureller Ähnlichkeit von Compounds in einigen Fällen unerwartet zu großen Aktivitätsunterschieden kommen, den sogenannten „activity cliffs“. Eine Unterscheidung der Liganden in Agonisten und Antagonisten spielt eine wichtige Rolle im Prozess der Wirkstoffentwicklung und sollte daher berücksichtigt werden. Des Weiteren verzichtet eine einseitige ligandbasierte Betrachtung auf die zur Verfügung stehenden Target-Informationen. Neben ligandbasierten Methoden besteht die Möglichkeit, Vorhersagen auch auf Grundlage von targetbasierten Methoden zu etablieren. So können Ähnlichkeiten zwischen Targets, deren Sequenzen, dreidimensionale Strukturen, Domänen oder Bindestellen der Liganden ebenfalls für Vorhersagen verwendet werden [56]. Eine Kombination aller zur Verfügung stehenden Informationen würde eine Target-Vorhersage für Compounds komplettieren und womöglich präzisere Vorhersagen ermöglichen. Eine Verbesserung von Vorhersageraten durch die

Kombination unterschiedlicher Methoden konnte für die Vorhersage der ATC-Klassen gezeigt werden.

Textmining von Publikationstexten wurde zur Generierung eines Datensatzes bezüglich Interaktionen von Proteinen der Biotransformation und Xenobiotika durchgeführt. Auf Basis dieser Wissensrepräsentationen konnte ein Algorithmus entwickelt werden, der die Möglichkeit bietet, Pharmakotherapien zu optimieren. Angezeigte problematische Medikamenten-Interaktionen können durch die Empfehlung alternativer Medikamente vermieden werden. Diese Empfehlung sollte allerdings durch den Kliniker unter Berücksichtigung weiterer klinisch relevanter Faktoren abgewogen werden. Darüber hinaus könnten Medikamenten-Interaktionen durch Polymorphismen der Proteine der Biotransformation beeinflusst werden. Dies wird allerdings nicht in der vorliegenden Arbeit berücksichtigt, sollte aber in zukünftigen Untersuchungen einbezogen werden. Die Miteinbeziehung von Polymorphismen würde einen Schritt in Richtung der personalisierten Medizin bedeuten. Polymorphismen können zu einer verlangsamten oder beschleunigten Metabolisierung von Arzneistoffen führen, oder sogar dazu, dass diese gänzlich nicht metabolisiert werden. Dies macht Dosisanpassungen oder die Auswahl alternativer Wirkstoffe nötig. Ein Schritt in Richtung der personalisierten Krebstherapie wurde durch die Einbeziehung genomischer Informationen im Rahmen dieser Dissertation aufgezeigt. Krebserkrankungen desselben Gewebes können sich durch genetische Alterationen unterscheiden. Dies beeinflusst die Reaktion des Tumors auf Therapieansätze. Die Option, eigene Mutations- oder Expressionsdaten hochzuladen, erlaubt die Identifikation ähnlicher Krebszelllinien samt der für sie effizienten Substanzen. Neben somatischen Mutationen sollen in Zukunft ebenfalls epigenetische Veränderungen wie Methylierungen aufgenommen werden, um die Etablierung neuer Ansätze der personalisierten Medizin zu ermöglichen und voranzutreiben.

Zugelassene Arzneistoffe werden häufig aufgrund unerwünschter Nebenwirkung vom Markt genommen [31]. Oftmals werden Nebenwirkungen durch Interaktionen zwischen dem Wirkstoff und Off-Targets ausgelöst. Analysen der zurückgezogenen Medikamente, ihrer (Off-)Targets, Polymorphismen und zugrunde liegenden Stoff- und Signalwege soll zu einem vertieften Verständnis der durch zurückgezogene Medikamente ausgelösten Toxizitäten beitragen. Das daraus gewonnene Wissen kann frühzeitig im Wirkstoffentwicklungsprozess angewendet werden, um für den Menschen sicherere Wirkstoffe zu entwickeln.

8 Literaturverzeichnis

1. Hopkins,A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, **4**, 682–90.
2. Sakharkar,K.R., Sakharkar,M.K. and Chandra,R. (2015) Post-genomic Approaches in Drug and Vaccine Development 1st ed. River Publisher, Aalborg, Denmark.
3. Boran,A.D. and Iyengar,R. (2010) Systems approaches to polypharmacology and drug discovery. *Curr. Opin. Drug Discov. Devel.*, **13**, 297–309.
4. Gorre,M.E., Mohammed,M., Ellwood,K., Hsu,N., Paquette,R., Rao,P.N. and Sawyers,C.L. (2001) Clinical Resistance to STI-571 Cancer Therapy Caused by BCR-ABL Gene Mutation or Amplification. *Science*, **293**, 876–80.
5. Engelman,J.A., Zejnullahu,K., Mitsudomi,T., Song,Y., Hyland,C., Park,J.O., Lindeman,N., Gale,C.-M., Zhao,X., Christensen,J., *et al.* (2007) MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 Signaling. *Science*, **316**, 1039–43.
6. Faivre,S., Demetri,G., Sargent,W. and Raymond,E. (2007) Molecular basis for sunitinib efficacy and future clinical development. *Nat. Rev. Drug Discov.*, **6**, 734–45.
7. Rothman,R.B., Baumann,M.H., Savage,J.E., Rauser,L., McBride,A., Hufeisen,S.J. and Roth,B.L. (2000) Evidence for Possible Involvement of 5-HT2B Receptors in the Cardiac Valvulopathy Associated With Fenfluramine and Other Serotonergic Medications. *Circulation*, **102**, 2836–41.
8. Roy,M., Dumaine,R. and Brown,A.M. (1996) HERG , a primary human ventricular target of the nonsedating antihistamine terfenadine. *Circulation*, **94**, 817–23.
9. Boguski,M.S., Mandl,K.D. and Sukhatme,V.P. (2009) Drug discovery. Repurposing with a difference. *Science*, **324**, 1394–95.
10. Ashbum,T.T. and Thor,K.B. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.*, **3**, 673–83.
11. Wang,W., Yang,S., Zhang,X. and Li,J. (2014) Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, **30**, 2923–30.
12. Wang,Y.C., Chen,S.L., Deng,N.Y. and Wang,Y. (2013) Network predicting drug's anatomical therapeutic chemical code. *Bioinformatics*, **29**, 1317–24.
13. Hajjar,E.R., Cafiero,A.C. and Hanlon,J.T. (2007) Polypharmacy in elderly patients. *Am. J. Geriatr. Pharmacother.*, **5**, 345–51.
14. Walckiers,D., Van Der Heyden,J. and Tafforeau,J. (2015) Factors associated with excessive polypharmacy in older people. *Arch. Public Heal.*, **73**, 50.
15. Hovstadius,B. and Petersson,G. (2012) Factors leading to excessive polypharmacy. *Clin. Geriatr. Med.*, **28**, 159–72.
16. Birriel,T.J., Uchino,R., Barry,N., Butryn,T., Sabol,D.M., Valenza,P.L. and Stawicki,S.P. (2015) Adverse drug reactions in the era of multi-morbidity and polypharmacy. *J. Basic Clin. Pharm.*, **6**, 122–23.
17. Guengerich,F.P. (2001) Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem. Res. Toxicol.*, **14**, 611–50.
18. Jancova,P., Anzenbacher,P. and Anzenbacherova,E. (2010) Phase II drug metabolizing enzymes. *Biomed. Pap. Med. Fac. Univ. Palacky. Olomouc. Czech. Repub.*, **154**, 103–16.
19. Ortiz de Montellano,P.R. (2013) Cytochrome P450-activated prodrugs. *Future Med. Chem.*, **5**, 213–28.
20. Westphal,J.F. (2000) Macrolide - induced clinically relevant drug interactions with cytochrome P-450 (CYP) 3A4: an update focused on clarithromycin, azithromycin and dirithromycin. *Br. J. Clin. Pharmacol.*, **50**, 285–95.
21. Thummel,K.E. and Wilkinson,G.R. (1998) In vitro and in vivo drug interactions involving human CYP3A. *Annu. Rev. Pharmacol. Toxicol*, **38**, 389–430.

22. Rampe,D., Roy,M.-L., Dennis,A. and Brown,A.M. (1997) A mechanism for the proarrhythmic effects of cisapride (Propulsid): high affinity blockade of the human cardiac potassium channel HERG. *FEBS Lett.*, **417**, 28–32.
23. Hayes,D.F., Markus,H.S., Leslie,R.D. and Topol,E.J. (2014) Personalized medicine: risk prediction, targeted therapies and mobile health technology. *BMC Med.*, **12**, 37.
24. Verma,M. (2012) Personalized medicine and cancer. *J. Pers. Med.*, **2**, 1–14.
25. Watkins,J., Marsh,A., Taylor,P.C. and Singer,D.R.J. (2010) Personalized medicine: the impact on chemistry. *Ther. Deliv.*, **1**, 651–65.
26. Salari,K., Watkins,H. and Ashley,E.A. (2012) Personalized medicine: hope or hype? *Eur. Heart J.*, **33**, 1564–70.
27. Chapman,P.B., Hauschild,A., Robert,C., Haanen,J.B., Ascierto,P., Larkin,J., Dummer,R., Garbe,C., Testori,A., Maio,M., *et al.* (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.*, **364**, 2507–16.
28. Furberg,C.D. (2011) Understanding drug safety and how to maximize it for patients. *Off. J. Am. Acad. Physician Assist.*, **24**, 16.
29. Berlin,J.A., Glasser,S.C. and Ellenberg,S.S. (2008) Adverse event detection in drug development: recommendations and obligations beyond phase 3. *Am. J. Public Health*, **98**, 1366–71.
30. Onakpoya,I.J., Heneghan,C.J. and Aronson,J.K. (2016) Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Med.*, **14**, 10.
31. Altman,R.B., Flockhart,D. and Goldstein,D.B. (2012) Principles of Pharmacogenetics and Pharmacogenomics 1st ed. Cambridge University Press, Cambridge, England.
32. Lavecchia,A. and Cerchia,C. (2016) In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov. Today*, **21**, 288–98.
33. Hoffmann,M.F., Preissner,S.C., Nickel,J., Dunkel,M., Preissner,R. and Preissner,S. (2014) Transformer database: biotransformation of xenobiotics. *Nucleic Acids Res.*, **42**, D1113–7.
34. Gaulton,A., Bellis,L.J., Bento,A.P., Chambers,J., Davies,M., Hersey,A., Light,Y., McGlinchey,S., Michalovich,D., Al-Lazikani,B., *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–07.
35. Hecker,N., Ahmed,J., Von Eichborn,J., Dunkel,M., Macha,K., Eckert,A., Gilson,M.K., Bourne,P.E. and Preissner,R. (2012) SuperTarget goes quantitative: Update on drug-target interactions. *Nucleic Acids Res.*, **40**, D1113–7.
36. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–201.
37. 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**, D668–72.
38. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S., *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–11.
39. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehár,J., Kryukov,G. V, Sonkin,D., *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature*, **483**, 603–7.
40. Reinhold,W.C., Sunshine,M., Liu,H., Varma,S., Kohn,K.W., Morris,J., Doroshow,J. and Pommier,Y. (2012) CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res.*, **72**, 3499–511.
41. Nickel,J., Gohlke,B.-O., Erehman,J., Banerjee,P., Rong,W.W., Goede,A., Dunkel,M. and Preissner,R. (2014) SuperPred: update on drug classification and target prediction. *Nucleic*

Acids Res., **42**, W26–31.
42. Gohlke,B.-O., Nickel,J., Otto,R., Dunkel,M. and Preissner,R. (2016) CancerResource--updated database of cancer-relevant proteins, mutations and interacting drugs. *Nucleic Acids Res.*, **44**, D932–7.
43. Siramshetty,V.B., Nickel,J., Omieczynski,C., Gohlke,B.-O., Drwal,M.N. and Preissner,R. (2016) WITHDRAWN—a resource for withdrawn and discontinued drugs. *Nucleic Acids Res.*, **44**, D1080–6.
44. Johnson,M.A. and Maggiora,G.M. (1990) Concepts and Applications of Molecular Similarity 1st ed. John Wiley & Sons, New York, United States of America.
45. Rogers,D. and Hahn,M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **50**, 742–54.
46. Willett,P., Barnard,J.M. and Downs,G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Model.*, **38**, 983–96.
47. Ahmed,J., Worth,C.L., Thaben,P., Matzig,C., Blasse,C., Dunkel,M. and Preissner,R. (2011) FragmentStore-A comprehensive database of fragments linking metabolites, toxic molecules and drugs. *Nucleic Acids Res.*, **39**, D1049–54.
48. Bauer,R.A., Bourne,P.E., Formella,A., Frömmel,C., Gille,C., Goede,A., Guerler,A., Hoppe,A., Knapp,E.-W., Pöschel,T., *et al.* (2008) Superimposé: a 3D structural superposition server. *Nucleic Acids Res.*, **36**, W47–54.
49. 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.
50. Zhou,Q., Li,W., Zeng,S. and Yu,L.S. (2013) Pharmacokinetic drug interaction profile of omeprazole with adverse consequences and clinical risk management. *Ther. Clin. Risk Manag.*, **9**, 259–71.
51. Füllbeck,M., Dunkel,M., Hossbach,J., Daniel,P.T. and Preissner,R. (2009) Cellular fingerprints: a novel approach using large-scale cancer cell line data for the identification of potential anticancer agents. *Chem. Biol. drug desing*, **74**, 439–48.
52. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
53. Sherry,S.T., Ward,M.-H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–11.
54. Nicola,G., Liu,T. and Gilson,M.K. (2012) Public domain databases for medicinal chemistry. *J. Med. Chem.*, **55**, 6987–7002.
55. Kalliokoski,T., Kramer,C., Vulpetti,A. and Gedeck,P. (2013) Comparability of mixed IC50 data - a statistical analysis. *PLoS One*, **8**, e61007.
56. Wang,C., Liu,J., Luo,F., Deng,Z. and Hu,Q.-N. (2015) Predicting target-ligand interactions using protein ligand-binding site and ligand substructures. *BMC Syst. Biol.*, **9**, S2.

9 Eidesstaatliche Versicherung

„Ich, Janette Nickel-Seeber, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: „Analyse und Vorhersage von Interaktionen von kleinen organischen Molekülen und ihren medizinischen Zielmolekülen“ selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche in korrekter Zitierung (siehe „Uniform Requirements for Manuscripts (URM)“ des ICMJE -*www.**icmje**.org*) kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) entsprechen den URM (s.o) und werden von mir verantwortet.

Meine Anteile an den ausgewählten Publikationen entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Betreuer/in, angegeben sind. Sämtliche Publikationen, die aus dieser Dissertation hervorgegangen sind und bei denen ich Autor bin, entsprechen den URM (s.o) und werden von mir verantwortet.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156,161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

_____

Datum Unterschrift

10 Anteilserklärung

Janette Nickel-Seeber hatte folgenden Anteil an den folgenden Publikationen:

Publikation 1:

Janette Nickel, Björn-Oliver Gohlke, Jevgeni Erehman, Priyanka Banerjee, Wen Wei Rong, Andrean Goede, Mathias Dunkel, Robert Preissner

SuperPred: update on drug classification and target prediction

Nucleic Acids Research, 2014

Anteil: 35 %

Beitrag im Einzelnen:

Entwicklung des Konzeptes; Literaturrecherche, Erstellung des Datensatzes (Datenintegration und Filterung des Protein-Compound-Interaktionsdatensatzes, Mapping der Compounddaten); Entwicklung, Ausarbeitung und Auswertung der Algorithmen, der zugrunde liegenden Daten und der Datenbankstruktur; Datenbankaufbau; Entwicklung und Erstellung des Web-Interfaces; Co-Betreuung der Masterarbeit von Frau Wen Wei Rong; Verfassen des Manuskripts; Reviewprozess.

Publikation 2:

Michael F. Hoffmann, Sarah C. Preissner, **Janette Nickel**, Mathias Dunkel, Robert Preissner, Saskia Preissner

The Transformer database: biotransformation of xenobiotics

Nucleic Acids Research, 2014

Anteil: 20 %

Beitrag im Einzelnen:

Mitarbeit an der Entwicklung des Konzeptes; Datenintegration und Filterung des Datensatzes; Datenbankaufbau; Erstellung des Web-Interfaces.

Publikation 3:

Björn-Oliver Gohlke, **Janette Nickel**, Raik Otto, Mathias Dunkel, Robert Preissner

CancerResource – updated database of cancer-relevant proteins, mutations and interacting drugs

Nucleic Acids Research, 2016

Anteil: 25 %

Beitrag im Einzelnen:

Mitarbeit an der Entwicklung des Konzeptes; Literaturrecherche, Ausarbeitung und Auswertung der zugrunde liegenden Expressionsdaten; Datenintegration der Expressionsdaten; Datenintegration und Filterung des Protein-Compound-Interaktionsdatensatzes; Normalisierung und Mapping der Compounddaten; Mitarbeit bei der Ausarbeitung der Datenbankstruktur; Mitarbeit an der Erstellung des Web-Interfaces; Mitarbeit beim Verfassen des Manuskriptes; Reviewprozess.

Publikation 4:

Vishal B. Siramshetty, **Janette Nickel**, Christian Omieczynski, Björn-Oliver Gohlke, Malgorzata N. Drwal, Robert Preissner

WITHDRAWN – a resource for withdrawn and discontinued drugs

Nucleic Acids Research, 2016

Anteil: 20 %

Beitrag im Einzelnen:

Mitarbeit an der Entwicklung des Konzeptes; Literaturrecherche; Datenintegration und Filterung des Protein-Compound-Interaktionsdatensatzes; Normalisierung und Mapping der Compounddaten; Mitarbeit beim Verfassen des Manuskriptes; Reviewprozess.

Unterschrift der Doktorandin

_____

11 Druckexemplare der ausgewählten Publikationen

W26–W31 Nucleic Acids Research, 2014, Vol. 42, Web Server issue *Published online 30 May 2014*
doi: 10.1093/nar/gku477

SuperPred: update on drug classification and target prediction

Janette Nickel[1,2,†], Bjoern-Oliver Gohlke[1,3,†], Jevgeni Erehman[1,3], Priyanka Banerjee[1,4], Wen Wei Rong[1], Andrean Goede[1], Mathias Dunkel[1] and Robert Preissner[1,*]

[1]Charité—University Medicine Berlin, Structural Bioinformatics Group, Institute of Physiology & Experimental Clinical Research Center, Berlin 13125, Germany, [2]Charité—University Medicine Berlin, Division of General Pediatrics, Department of Pediatric Oncology and Hematology, Berlin 13353, Germany, [3]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany and [4]Graduate School of Computational System Biology, Berlin 10115, Germany

Received February 28, 2014; Revised April 23, 2014; Accepted May 13, 2014

ABSTRACT

The SuperPred web server connects chemical similarity of drug-like compounds with molecular targets and the therapeutic approach based on the similar property principle. Since the first release of this server, the number of known compound–target interactions has increased from 7000 to 665 000, which allows not only a better prediction quality but also the estimation of a confidence. Apart from the addition of quantitative binding data and the statistical consideration of the similarity distribution in all drug classes, new approaches were implemented to improve the target prediction. The 3D similarity as well as the occurrence of fragments and the concordance of physico-chemical properties is also taken into account. In addition, the effect of different fingerprints on the prediction was examined. The retrospective prediction of a drug class (ATC code of the WHO) allows the evaluation of methods and descriptors for a well-characterized set of approved drugs. The prediction is improved by 7.5% to a total accuracy of 75.1%. For query compounds with sufficient structural similarity, the web server allows prognoses about the medical indication area of novel compounds and to find new leads for known targets. SuperPred is publicly available without registration at: http://prediction.charite.de.

INTRODUCTION

The Anatomical Therapeutic Chemical (ATC) classification system of the World Health Organization (WHO) is currently the most prevalent system to characterize drugs. This system is divided into several hierarchical categories differentiating between anatomical, therapeutic, pharmacological and chemical properties (1). Drug utilization can be investigated using the ATC classification system. Therefore, comparing the drugs' structural and physico-chemical features by means of ATC codes offers a possibility to gain knowledge for drug repositioning and predicting new medical indications as well as classifying yet unclassified compounds. The established 'similarity property principle' (2) is based on the assumption that structurally similar molecules exhibit similar biological activity (3). Various 2D methods have been developed to search for similarity between compounds (4). Among others, topological descriptors like 2D fingerprints (5) or BCUT descriptors (6) are often applied in similarity searching. Although 2D fingerprints are widely used for various applications like virtual screening, similarity searching and clustering, several problems can occur. For instance, the molecular size of a compound can affect the similarity calculations as well as a folding of fixed-length bit strings which can result in the negligence of functional and structural features. To overcome these interferences, the SuperPred update (SuperPred II) does not only consider 2D similarity methods but also fragment and 3D similarity searching. Recently, some attempts have been undertaken to address the ATC prediction problem. Gurulingappa *et al.* used a combination of information extraction and machine learning techniques for classifying yet unclassified drugs into ATC classes (7). To verify their method, they used classified drugs with an indication on the cardiovascular system (ATC class 'C'). Another approach by Chen *et al.* joins chemical–chemical interaction with chemical–chemical similarity information (8) to classify drugs. Using this approach, the authors analyzed the identification of drugs among the 14 main ATC classes. Furthermore, Wang *et al.* presented NetPredATC, a drug–target network based on support vector machines for predicting the ATC class of a compound (9). They assume that drugs with similar chem-

*To whom correspondence should be addressed. Tel: +49 30 450 540 755; Fax: +49 30 450 540 955; Email: robert.preissner@charite.de
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

ical structures or target proteins share common ATC codes. Based on their assumption, they integrated the compounds chemical similarity with target information and used a support vector machine approach for the ATC code prediction. The method validation was carried out using four different drug datasets which include enzymes, ion channels (IC), G-protein coupled receptors (GPCR) and nuclear receptors (NR) as target proteins.

Recently, drug promiscuity has become an important issue in drug discovery. It was observed that drugs show a more promiscuous way of binding than it was assumed in the past (10). Due to the more complex nature of drug binding, the view of drugs as specific ligands to targets had to be reconsidered. Drug promiscuity, which entails unwanted side effects due to binding to off-targets (11), is considered as one of the main reasons for failure and withdrawal of marketed drugs. A case example represents the withdrawal of the drug combination fenfluramine/phentermine (fen–phen) because of inducing valvular heart diseases (12). Predicting targets as well as off-targets for drugs or drug candidates might help avoiding unwanted side effects as well as facilitating drug-repositioning. Several approaches have been introduced for predicting drug–target interactions. Network-based methods have been proposed to identify protein targets for drugs (13–15). Moreover, the similarity ensemble approach (16) has been proposed. The method is based on the stochastic analysis of the 2D similarity between ligands that bind to the same target and predicts ligand–target interactions adapting concepts of the basic local alignment search tool (BLAST) algorithm (17). Another method for predicting compound–target interactions is SPiDER (18). It addresses the issue of predicting targets for *de novo* designed molecules and drugs using two self-organizing maps (SOM) differing in the molecular representations for the SOM projections. The resulting two confidence scores are converted into a consensus score and contemplated in a statistical analysis to indicate the significance of the prediction.

The SuperPred web server comprises two methods, one for drug classification based on approved drugs classified by WHO (1) and one for target prediction based on compound–target interaction data. The drug classification method takes into account 2D- and fragment-similarity, and a method for 3D superposition of small molecules. The method for target prediction uses the similarity distribution among ligands for estimating the targets' individual thresholds and probabilities to avoid false positive predictions.

MATERIALS AND METHODS

Data set for drug classification

For drug classification, a dataset containing 2650 drugs is taken from Transformer (19). To ensure comparability between SuperPred I (20) and SuperPred II, the dataset (1035 drugs) described in SuperPred I was used for evaluation of the drug classification method.

Based on the actual drugs classified by WHO, an external dataset containing 190 novel drugs was created for validation of the drug classification method.

Data set for the target prediction

The dataset for target prediction was created by extracting compound–target interaction data from SuperTarget, ChEMBL and BindingDB (21–23). Those databases offer a huge amount on publicly available ligand–target interaction data. To integrate the extracted data into one consistent set, several normalization steps were accomplished concerning compound and target entities and interaction data.

First, compound structures were normalized using JChem (Instant JChem 6.2.0 (January 2014), ChemAxon (http://www.chemaxon.com)). Normalization steps involved isolation of the largest fragment in the structure, removal of salts and explicit hydrogens and the standardization of stereochemical and charge information using the JChem standardization protocol. Furthermore, the structures were aromatized and formal charges were removed. For compound unification, International Chemical Identifiers (InChI) were calculated using Open Babel (http://openbabel.org/) and compounds having identical InChI were merged. Second, non-molecular target types (ChEMBL) like cell-lines, tissues and organisms and molecular target types like deoxyribonucleic acid as well as non-mammal enzymes and proteins were removed, yielding a target dataset of mammal proteins only. The remaining targets were unified using the Entrez Gene Index from NCBI (National Center for Biotechnology Information) (24) and those mapping to the same gene were merged. Third, the interaction data was filtered for certain binding types (e.g. IC50, Ki and KD), resulting in 1 900 000 interactions. Additionally, interactions described by binding affinities weaker than 10 000 nm were removed. Finally, targets having interaction data for less than five compounds were removed, resulting in a dataset consisting of ~341 000 compounds, ~1800 targets and ~665 000 compound–target interactions.

For the evaluation of the target prediction, the dataset was restricted to 'successful targets' from the Therapeutic Target Database (TTD) (25) narrowing the set to 221 targets, 95 000 compounds and 174 000 compound–target interactions.

Drug classification pipeline

The drug classification pipeline is a combination of three different structure based methods, considering 2D, fragment and 3D similarity, described in detail below. This combination ensures an optimal coverage of the structural features represented by a final score. The consensus of these methods is taken into account. If at least two methods predict the same ATC class, that class is considered as final prediction. If three different ATC classes are predicted, a threshold for every method is used to decide for the most probable ATC class (Figure 1: left).

2D similarity searching

In order to select the optimal fingerprint for the 2D similarity comparison, several fingerprints have been compared (Table 1). The extended-connectivity fingerprints (ECFP) (26) exhibit the best performance for our dataset and hence, have been used in the prediction pipeline. The fingerprints

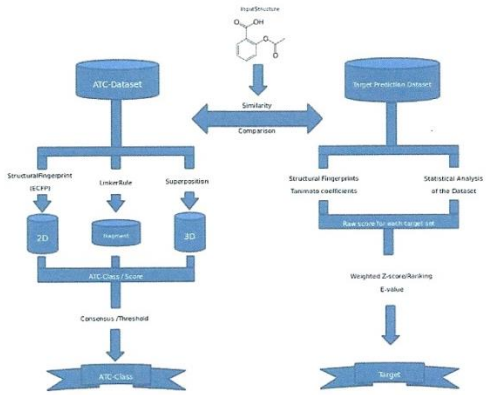

Figure 1. This diagram illustrates the drug classification pipeline (left) and target prediction pipeline (right). The drug classification is carried out in three steps. In the first step the input compound is compared with the ATC dataset by the following methods namely 2D, fragment and 3D similarity searching. In the second step the ATC-class and the corresponding score is calculated for each method. The last step ensembles the predicted ATC-classes according to the score and predict the final ATC-class. Similarly, the target prediction is also carried out in three main steps. In the first step, the input compound is compared based on structural similarity (2D). The second step analyzed the statistical significance of the similarity score in comparison with precalculated statistics of the dataset. The last step computes the raw score for each target and finally the target is predicted with consideration of the weighted Z-score and E-value threshold.

Table 1. Comparison of fingerprints and their attained prediction rate for the evaluation dataset of 1035 compounds

Fingerprint	2D prediction rate
FP24	62.6
MDL(166)	72.3
ECFP4	74.1

belong to the class of radial fingerprints and are generated by a modified version of the Morgan Algorithm (27). The calculated fingerprints were subsequently compared by the Tanimoto similarity measure for bit strings (28).

Fragment similarity searching

All 2650 drugs from the prediction dataset have been fragmented according to the linker rule (29). This method preferentially generates cyclic fragments by removing the linker atoms between ring structures. All non-redundant fragments which were produced by the fragmentation method are considered for comparison. While comparing the fragments of two small molecules (A and B) having n and m fragments, a similarity matrix with $n \times m$ fields is constructed. Each field contains the Tanimoto coefficient of the particular fragment comparison. The matrix is used to calculate $\binom{n}{m}$ possible fragment combinations. For each combination, a final Tanimoto score is calculated by summing up its Tanimoto coefficients from the matrix. The final similarity score is further divided by the smaller number of fragments belonging to one of the molecules.

3D similarity searching

The superimposition of one molecule to a reference molecule structure is done by mapping atoms with optimal distances. In order to reduce time complexity, only 100 low-energy conformations are generated for the two molecules to be compared and pairwise comparisons of all possible conformations are performed. Hence, given a molecule pair, a maximum of 10 000 comparisons take place (30). The first step of the algorithm normalizes the set of atoms into a new coordinate system. Based on these coordinates, the centers of mass for both conformers are calculated and superimposed. Then, the principal axes of inertia are estimated and aligned. Thereby, the possible rotations are strongly reduced and only four orientations have to be considered. For every orientation, a mapping of atom pairs is performed whereupon atoms are fitted to each other with the smallest possible distance. A maximal distance threshold is applied for atom pair assignment, therefore not every atom is assigned. The rotation matrix with the highest amount of mapped pairs was used for further calculations. The normalized variant with the minimal distance is chosen if more than one rotation with the same amount of mapped atom pairs exists. For this mapping a root-mean-square-deviation (rmsd) was calculated. To find the best superposition of two molecules, the number of superposed atoms and the corresponding rmsd value are taken into account by the following formula:

$$\text{3D-score} = \frac{N_S}{\max(N_A \times N_B)} \exp(-\text{rmsd})$$

where N_S is the number of superposed atoms, N_A the number of atoms of molecule A and N_B the number of atoms of molecule B.

Target prediction method

The method for the drug–target prediction takes into consideration the 2D similarity between the query compound and the ligands associated to their respective targets (target sets). For each target set, the summation of all Tanimoto coefficients above a threshold of 0.45 is considered as raw score. To achieve comparability between raw scores of small and large target sets, the raw scores are normalized by dividing them by the number of ligands of the corresponding target. To further evaluate the specificity of a prediction, Z-scores and E-values are computed. The Z-score is calculated by the formula:

$$Z_A = \frac{\left(\frac{(\text{raw score}_A)}{N_A} - \mu\right) \exp(0.335 \ln(N_A))}{\sigma}$$

where A is a target set and N_A represents the number of ligands of target set A. Similar to BLAST (17) μ and σ describe the random background noise of the database.

The E-value describes the number of predicted targets one can expect to see by chance, thereby it depends on the size of the dataset. The E-value decreases exponentially as the Z-score of the prediction increases. The lower the E-value, the more significant is the prediction (17). (For further details and formulas please see the FAQ section on our SuperPred website).

For diverse target sets, Z-scores tend to behave like high random scores. Therefore, a weighting factor λ_A is introduced which indicates the average similarity between the ligands within each target set:

$$\lambda_A = \exp\left(0.335 \ln\left(\frac{\text{raw score}_{AA}}{N_{AA}}\right)\right)$$

The weighting factor ranges between almost one for very uniform target sets to more than ten for very diverse target sets. The target prediction results are ranked according to the weighted Z-scores (Figure 1: right).

Input and output options

There are four input options available for drug classification and target prediction. First, via the ChemDoodle tool (http://www.chemdoodle.com/), an upload function for MOL files is provided. Second, it is possible to draw a structure using the ChemDoodle editor. Third, a PubChem (31) name search option is provided and fourth, a molecule can be searched by its Simplified Molecular Input Line Entry Specification (SMILES) (http://daylight.com/smiles/).

The output for the 'Drug Classification' and the 'Target-Prediction' displays the input compound's properties and its molecular structure. In case of the 'Drug Classification' result site the prediction accuracy, the ATC-class and information about similar drugs, that have ATC-codes assigned, is given. Furthermore, Lipinski-rule of five properties (32) for the uploaded compound are also shown. In addition, the statistics for physico-chemical properties for the predicted ATC class are presented. Moreover, a button is provided to start the target prediction for the input compound likewise the 'Target-Prediction' result site offers a button for starting the drug classification. Furthermore, it displays known and predicted targets for the input compound and provides detailed information about the targets. Links to other databases as well as available PDB structures are given.

RESULTS AND DISCUSSION

Drug classification

The implementation of the new prediction pipeline (Figure 1: left) consisting of 2D, fragment and 3D similarity searching methods results in a higher prediction rate for ATC classes compared to SuperPred I based on the evaluation dataset from SuperPred I. In SuperPred II, the prediction accuracy has increased to 75.1% for the validation set (Table 2). SuperPred I, taking only into account the 2D similarity of compounds, showed a prediction accuracy of 67.6%. The prediction rates' distribution values of the correctly predicted ATC codes are shown in Table 3. For a consensus score range between 0.8 and 0.9, a prediction rate of 88.9% is achieved. Furthermore, a cumulative recall graph is shown in Figure 2 representing the fraction of correct ATC class predictions in dependency of the quantity of retrieved molecules. By taking into account the three most similar structures, a prediction rate of 80.3% is reached whereas a recall of maximal 88% is reached by taking at least 16 similar compounds into consideration. For further validation of the prediction pipeline, we utilized the 190 new drugs contained in the evaluation dataset. The prediction accuracy for this dataset is 72.1% .

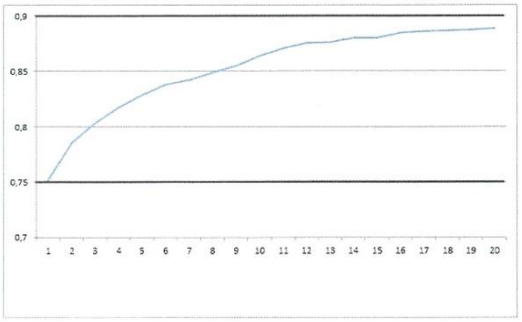

Figure 2. Cumulative recall graph for ATC-prediction relative to the rank of retrieved compounds.

Table 2. Prediction accuracy overview of the individual drug classification methods (2D, fragment, 3D similarity methods) as well as for the combined pipeline (consensus) for the evaluation dataset

Method	Prediction accuracy (%)
2D	74.1
Fragment	69.4
3D	67.7
Consensus	75.1

Table 3. Prediction rate distribution for correctly predicted ATC codes

Range of consensus score	Number of hits/misses	Prediction rate (%)
0.0–0.1	0/0	0
0.1–0.2	2/4	33.3
0.2–0.3	6/40	13.0
0.3–0.4	14/22	38.9
0.4–0.5	20/43	31.8
0.5–0.6	50/32	61.0
0.6–0.7	88/24	78.6
0.7–0.8	164/36	82.0
0.8–0.9	233/29	88.9
0.9–1.0	199/27	88.1

The distribution is based on the evaluation dataset, which contains 1035 drugs. This table shows the number of right (hits) and wrong (misses) predictions for a specific consensus score range.

Comparison to other drug classification methods

In comparison to other drug classification methods, SuperPred II yields the best prediction rate (Table 4). Chen *et al.* have analyzed their prediction performance of the first level of ATC codes (13). They have combined chemical–chemical interaction with chemical–chemical similarity information. Based on this, their method reaches a prediction rate of 73.25% for prediction of the 14 main ATC classes. To make our method comparable, we modified our prediction to the first level ATC class. This resulted in a prediction rate of 80.9% for identifying the right ATC classes among the 14 main classes.

Table 4. Comparison of the SuperPred update with other ATC prediction methods

	Total accuracy [%]	Comment
SuperPred (2008)	67.6	Overall prediction
NetPredATC (2013)	74.0	GPCR
Chen et. Al (2012)	73.25	Main ATC classes
SuperPred (2014)	75.1	Overall prediction

The comment column indicates how the prediction accuracy was achieved.

Furthermore, we compared our method with NetPredATC from Wang *et al.* The accuracy of NetPredATC lies between 74 and 76.5% according to the previously mentioned subsets belonging to single target classes like GPCR, NR, IC and enzymes.

Target prediction

The target prediction method, results in a prediction accuracy of 91.2% without the use of the weight function (λ). Considering the weight function, the prediction rate increases to 92.8%. Additionally, the *E*-value is used as a threshold: an *E*-value above 1 is an indication of random prediction. Considering this threshold, it was observed that the prediction rate further increases to 94.1%. However, about 9400 compounds were not considered for prediction because they were above the *E*-value threshold. It was also observed that the target groups which have more diverse compounds show lower prediction rates. This could be caused by multiple binding sites on the target or by targets with different domains that have different properties and bind different types of ligands, resulting in subsets of related compounds inside its target set.

CONCLUSION

In comparison to the ATC prediction method described by Chen *et al.*, the SuperPred II method is able to produce a higher prediction accuracy of ~8%. Wang *et al.* perform their prediction on four relatively small benchmark datasets whereas SuperPred II considers a wide range of target classes and produces a comparable prediction accuracy. For further improvement of the drug classification method an integration of drug-protein networks could increase the prediction accuracy as drug pairs having the same ATC code may bind to the same targets (33).

To improve the target prediction method, target groups with diverse compounds due to multiple binding sites or different domains will be considered as independent target sets as agonists and antagonists bind to different binding sites or domains and cause different pharmacological effects.

ACKNOWLEDGMENT

The authors thank Malgorzata Drwal for many helpful discussions and her support.

FUNDING

German José Carreras Leukaemia Foundation [DJCLS R 12/05]; German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; German Federal Ministry of Education and Research (e:Top); European Union Seventh Framework Programme SYNSYS (Synaptic Systems: dissecting brain function in health and disease), HEALTH-2009-2.1.2-1, under grant agreement no [242167]; Immunotox project (BMBF) [031A268B]; Berlin-Brandenburg research platform BB3R (BMBF) [031A262C]; German Research Foundation (DFG) [GRK1772, GRK1360]. Funding for open access charge: German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany.

Conflict of interest statement. None declared.

REFERENCES

1. The selection and use of essential medicines. Report of the WHO expert committee, 2005 (including the 14th model list of essential medicines). (2003) *World Health Organ. Tech. Rep. Ser.*, 1–119.
2. Ugi,I., Wochner,M., Fontain,E., Bauer,J., Gruber,B., Karl,R.E., Johnson,A. and Maggiora,G. (1990) *Concepts and applications of molecular similarity*.Wiley Nova Iorque , New York.
3. Martin,Y.C., Kofron,J.L. and Traphagen,L.M. (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**, 4350–4358.
4. Sheridan,R.P. and Kearsley,S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discov. Today*, **7**, 903–911.
5. Duan,J., Dixon,S.L., Lowrie,J.F. and Sherman,W. (2010) Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.*, **29**, 157–170.
6. Pirard,B. and Pickett,S.D. (2000) Classification of kinase inhibitors using BCUT descriptors. *J. Chem. Inf. Comput. Sci.*, **40**, 1431–1440.
7. Gurulingappa,H., Kolarik,C., Hofmann-Apitius,M. and Fluck,J. (2009) Concept-based semi-automatic classification of drugs. *J. Chem. Inf. Model.*, **49**, 1986–1992.
8. Chen,L., Zeng,W.M., Cai,Y.D., Feng,K.Y. and Chou,K.C. (2012) Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLoS One*, **7**, e35254.
9. Wang,Y.C., Chen,S.L., Deng,N.Y. and Wang,Y. (2013) Network predicting drug's anatomical therapeutic chemical code. *Bioinformatics*, **29**, 1317–1324.
10. Hopkins,A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, **4**, 682–690.
11. Yamanishi,Y., Pauwels,E. and Kotera,M. (2012) Drug side-effect prediction based on the integration of chemical and biological spaces. *J. Chem. Inf. Model.*, **52**, 3284–3292.
12. Rothman,R.B., Baumann,M.H., Savage,J.E., Rauser,L., McBride,A., Hufeisen,S.J. and Roth,B.L. (2000) Evidence for possible involvement of 5-HT(2B) receptors in the cardiac valvulopathy associated with fenfluramine and other serotonergic medications. *Circulation*, **102**, 2836–2841.
13. van Laarhoven,T., Nabuurs,S.B. and Marchiori,E. (2011) Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics*, **27**, 3036–3043.
14. Chen,X., Liu,M.X. and Yan,G.Y. (2012) Drug-target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.*, **8**, 1970–1978.
15. Emig,D., Ivliev,A., Pustovalova,O., Lancashire,L., Bureeva,S., Nikolsky,Y. and Bessarabova,M. (2013) Drug target prediction and repositioning using an integrated network-based approach. *PLoS One*, **8**, e60618.
16. Keiser,M.J., Roth,B.L., Armbruster,B.N., Ernsberger,P., Irwin,J.J. and Shoichet,B.K. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.
17. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
18. Reker,D., Rodrigues,T., Schneider,P. and Schneider,G. (2014) Identifying the macromolecular targets of de novo-designed chemical

entities through self-organizing map consensus. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 4067–4072.
19. Hoffmann,M.F., Preissner,S.C., Nickel,J., Dunkel,M., Preissner,R. and Preissner,S. (2014) The Transformer database: biotransformation of xenobiotics. *Nucleic Acids Res.*, **42**, D1113–D1117.
20. Dunkel,M., Gunther,S., Ahmed,J., Wittig,B. and Preissner,R. (2008) SuperPred: drug classification and target prediction. *Nucleic Acids Res.*, **36**, W55–W59.
21. Hecker,N., Ahmed,J., von Eichborn,J., Dunkel,M., Macha,K., Eckert,A., Gilson,M.K., Bourne,P.E. and Preissner,R. (2012) SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res.*, **40**, D1113–D1117.
22. Gaulton,A., Bellis,L.J., Bento,A.P., Chambers,J., Davies,M., Hersey,A., Light,Y., McGlinchey,S., Michalovich,D., Al-Lazikani,B. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.
23. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
24. Coordinators,N.R. (2013) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**, D8–D20.
25. Zhu,F., Shi,Z., Qin,C., Tao,L., Liu,X., Xu,F., Zhang,L., Song,Y., Liu,X., Zhang,J. *et al.* (2012) Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.*, **40**, D1128–D1136.
26. Rogers,D. and Hahn,M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **50**, 742–754.
27. Morgan,H.L. (1965) The generation of a unique machine description for chemical structures–a technique developed at chemical abstracts service. *J. Chem. Doc.*, **5**, 107–113.
28. Willett,P., Barnard,J.M. and Downs,G.M.D. (1998) Chemical similarity searching. *J. Chem. Inf. Model.*, **38**, 983–996.
29. Ahmed,J., Worth,C.L., Thaben,P., Matzig,C., Blasse,C., Dunkel,M. and Preissner,R. (2011) FragmentStore–a comprehensive database of fragments linking metabolites, toxic molecules and drugs. *Nucleic Acids Res.*, **39**, D1049–D1054.
30. Bauer,R.A., Bourne,P.E., Formella,A., Frommel,C., Gille,C., Goede,A., Guerler,A., Hoppe,A., Knapp,E.W., Poschel,T. *et al.* (2008) Superimpose: a 3D structural superposition server. *Nucleic Acids Res.*, **36**, W47–W54.
31. Bolton,E.E., Wang,Y., Thiessen,P.A. and Bryant,S.H. (2008) Chapter 12 PubChem: integrated platform of small molecules and biological activities. *Ann. Rep. Comput. Chem.*, **4**, 217–241.
32. 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.
33. Brouwers,L., Iskar,M., Zeller,G., van Noort,V. and Bork,P. (2011) Network neighbors of drug targets contribute to drug side-effect similarity. *PLoS One*, **6**, e22187.

Published online 10 December 2013 *Nucleic Acids Research, 2014, Vol. 42, Database issue* ***D1113–D1117***
doi:10.1093/nar/gkt1246

The Transformer database: biotransformation of xenobiotics

Michael F. Hoffmann, Sarah C. Preissner, Janette Nickel, Mathias Dunkel, Robert Preissner and Saskia Preissner*

Structural Bioinformatics Group, Institute for Physiology & ECRC, Charité – Universitätsmedizin Berlin, Lindenberger Weg 80, 13125 Berlin, Germany and Department of Operative and Preventive Dentistry, Charité - Universitätsmedizin Berlin, Assmannshauser Str. 4-6, 14197 Berlin, Germany

Received August 16, 2013; Revised November 7, 2013; Accepted November 9, 2013

ABSTRACT

As the number of prescribed drugs is constantly rising, drug–drug interactions are an important issue. The simultaneous administration of several drugs can cause severe adverse effects based on interactions with the same metabolizing enzyme(s). The Transformer database (http://bioinformatics.charite.de/transformer) contains integrated information on the three phases of biotransformation (modification, conjugation and excretion) of 3000 drugs and >350 relevant food ingredients (e.g. grapefruit juice) and herbs, which are catalyzed by 400 proteins. A total of 100 000 interactions were found through text mining and manual validation. The 3D structures of 200 relevant proteins are included. The database enables users to search for drugs with a visual display of known interactions with phase I (Cytochrome P450) and phase II enzymes, transporters, food and herbs. For each interaction, PubMed references are given. To detect mutual impairments of drugs, the drug-cocktail tool displays interactions between selected drugs. By choosing the indication for a drug, the tool offers suggestions for alternative medications to avoid metabolic conflicts. Drug interactions can also be visualized in an interactive network view. Additionally, prodrugs, including their mechanisms of activation, and further information on enzymes of biotransformation, including 3D models, can be viewed.

INTRODUCTION

The number of prescribed drugs is rising (1). A study revealed that 87.1% of people >50 years of age take at least one drug per day, and 43.3% take >5 (2). Polypharmacy, which is defined as the regular use of five or more drugs, leads to an increased risk of adverse drug reactions (ADRs). The frequency of ADR is associated with the number of drugs prescribed (3). Among hospitalized patients, ADRs have an incidence of 6.7% and are the fifth commonest cause of death (4). One possible cause for ADR might be the individual variance of drug metabolism (5), and age-related changes make elderly patients more sensitive to ADRs (6). The information is widely scattered over the scientific literature. A knowledge base of xenobiotic metabolism and the effect of polymorphisms could prevent ADR and cases of death.

Xenobiotic metabolism and detoxification (especially for drugs) are separated into three different phases of reaction. Only a few xenobiotics are excreted unchanged in urine or feces without any metabolic degradation.

Phase I and phase II reactions convert compounds to more water-soluble and often less active derivatives to increase excretion. Thereby, phase I reflects the production of reactive groups through oxidation and is primarily managed by the Cytochrome P450 family (CYP) of enzymes (7). Subsequently, the reactive groups are used to conjugate small polar molecules (phase II) to increase the polarity. Six enzyme families that provide the detoxification and excretion of xenobiotics mainly realize the conjugation (8).

Transporters (phase III) play a crucial role in pharmacokinetics by enabling the migration of hydrophilic molecules, which cannot penetrate cellular membranes. Kell *et al.* showed that the majority of drugs enter cells through at least one transporter (9). Those proteins form a transmembrane channel lined with hydrophilic amino acid side chains spanning the lipid bilayer (10). Two major protein superfamilies are known: 49 ATP-binding cassette transporters (ABC) (11) and 362 solute carriers (SLC) (12). These are important for absorption, distribution and

*To whom correspondence should be addressed. Tel: +49 30 450 562 675; Fax: +49 30 450 562 932; Email: saskia.preissner@charite.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

excretion of drugs (13) and are involved in a broad range of physiological processes (10).

Eukaryotic ABC transporters are predominantly exporters, which require energy released by ATP hydrolysis. One problem is multidrug resistance, which is caused by active transporters. Unfortunately, 40% of human tumors develop resistance to chemotherapeutics by overexpressing ABC proteins (14). The SLC transporters facilitate passive diffusion along the concentration gradient or use concentration gradients from other substrates as a symporter or antiporter (12).

Another issue related to drug metabolism and ADR is prodrugs. Prodrugs have to be converted to active drugs by metabolic conversion (15). In general, prodrugs are non-toxic and need to have their chemical structure changed to enable their inherent medical capability. However, problems in conversion can also lead to undesired side effects. For example, the antihistamine terfenadine is a potent hERG blocker as a prodrug and a slow conversion can cause cardiac toxicity (16).

Prodrugs can be activated by photo irradiation (17), a change in pH (18) and enzymes, such as esterases or CYPs (19,20). Many prodrugs are activated by hydrolysis with the aid of esterases or phosphatases. Thereby, gastric intestinal tolerance and pharmacokinetics can be improved, but the targeting of drugs to specific cells or tissues cannot. The activation of prodrugs by CYPs might be a better approach (21).

Not only can drugs participate in the alteration of drug metabolism but food and herbs also have a proven influence; e.g. furanocoumarins in grapefruit inhibit intestinal CYP3A4 and organic anion-transporting polypeptides 1A2 (22,23).

More than 350 ingredients in food and drink, such as broccoli, alcohol and char-grilled meat, as well as herbal medicine, such as St John's wort, are known to alter drug responses.

A comprehensive resource that combines scientific information on phase I and phase II enzymes, transporter enzymes, prodrugs, food and herbs could help to improve research in this field and prevent ADR.

MATERIALS AND METHODS

Text mining

We created a text mining approach using semantic web standards. To develop a specialized text mining pipeline, we first downloaded Medline/PubMed data from the NCBI FTP site in xml-format. Using the search engine library Apache Lucene (http://lucene.apache.org) and a tool kit for processing text with computational linguistics (http://alias-i.com/lingpipe), the data was indexed. The search engine comprises comprehensive lists of chemical compounds and drug names (24), metabolic enzymes (25) and transporters (26), including their various synonyms. Additionally, we added a list of common interaction terms, such as 'activate', 'inhibit', and 'metabolize'. The search engine, written in Java, dynamically queries the indexed data and produces a structured query language (SQL) file containing the text mining hits. A query example is:

```
(DrugSynonym [TI] AND TransformerSynonym
  [TI]) OR
(DrugSynonym [abstract] AND TransformerSynonym
  [abstract]) OR
(DrugSynonym [abstract] AND InteractionTerm AND
  TransformerSynonym [abstract])
```

The positional distance between the different terms had to be restricted to reduce false-positive hits, when terms occurred far from each other in the abstract. The 22 500 records found were scored as rule-based. Duplicates were removed and a team of scientists manually processed 12 427 articles found in PubMed. Further details about the text mining approach can be found on the Web site in the frequently asked questions (FAQs) section.

Database

The database was designed as a relational database on a MySQL server. To allow chemical functionality, such as handling chemical data within MySQL, the MyChem package was included.

Information about ~3000 CYP drug interactions and 2000 polymorphisms were extracted from the SuperCYP database (27). SuperCYP is a database with a focus on human CYPs. However, there are many other important enzymes in the metabolism of xenobiotics, such as transporters or phase II enzymes.

DATABASE FEATURES

Over 100 000 interactions were revealed. In the 12 427 articles found in PubMed, 769 drugs were attributed to those phase II enzymes that are involved in drug metabolism. Text mining was also performed for prodrugs, transporters and food. We found 125 prodrugs described in 890 PubMed articles together with their mechanism of activation, accompanying enzymes, chemical structure and identification numbers. Furthermore, ~500 drug-transporters and 150 food interactions were identified.

Additionally, ~200 3D structures were collected for transporters, CYPs and phase II enzymes.

The database includes four main functionalities. To provide an overview of these, a comprehensive FAQs section was created, which is helpful for first-time users. Depending on the user's interest or needs (e.g. clinicians, researchers), different ways to browse the data were enabled.

Prodrugs

Prodrugs can be identified directly by entering the name, PubChemID, CAS number or ATC code, as well as by choosing a mechanism of activation, such as ring opening or carboxylation.

Drugs

To view the metabolism of particular drugs, users can search directly by entering the name, PubChemID or

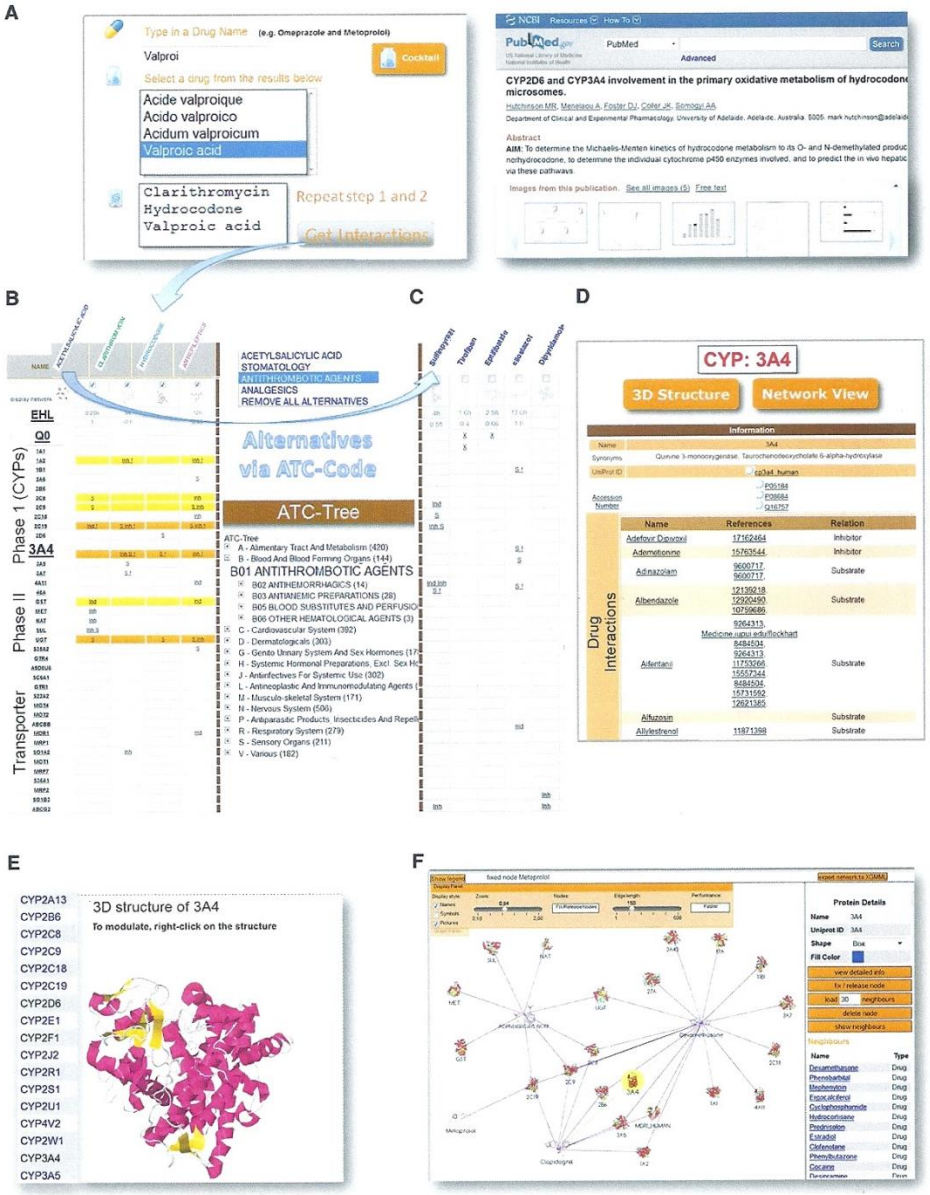

Figure 1. Functionalities of the 'Transformer database'. (**A**) Composition of a drug-cocktail. An example of a PubMed reference is shown. PubMed references can be viewed by clicking on 'S', 'Inh' or 'Ind' in the result table. (**B**) Clicking on 'Get Interactions' leads to a result table, which shows the interactions between the drugs. (**C**) By choosing the indications of the drugs via ATC-code the user receives specific alternatives. (**D**) By clicking on an enzyme (e.g. CYP3A4), detailed information on the enzyme, including drug interactions, are shown. (**E**) The 3D structures of all enzymes can be viewed (e.g. CYP3A4). (**F**) Network views are provided for each enzyme and compound.

CAS number. Based on the WHO classification system, which classifies drugs into different groups according to Anatomical site of action, Therapeutical effect and Chemical structure (ATC), a tree with all of the drugs contained in the database can be viewed in their ATC group.

Cocktail

This tool enables users to see drug interactions of an individually composed drug cocktail (Figure 1A). If >1 drug interacts with the same enzyme, lines of the interaction table are shown in yellow, orange, red and dark red (Figure 1B). In the header of each column, the indication for the drugs can be chosen and the database will provide alternative drugs that are metabolized by different enzymes (Figure 1C). PubMed references are available by clicking on the interaction. Additionally, food interactions, as well as elimination half-life (EHL) times and Q_0 values are displayed. A Q_0 value (extrarenal excretion) of <0.3 is shown in green because those drugs are, to a large extent, excreted in the unchanged form. Clicking 'Display network' presents these interactions in a network view based on Cobweb (28).

Biotransformation

To find drugs that are metabolized by specific phase I, phase II or transporter enzymes, users can perform a search by clicking on 'Biotransformation'. This page provides (homology modeled) 3D structures of all enzymes (Figure 1E). Furthermore, a list of interacting drugs can be viewed in a table (Figure 1D) or in a network view (Figure 1F).

DATABASE USAGE

The following case illustrates the need to detect interactions with the help of the Transformer database. A five-year-old child died from a fatal opioid toxicity. She was inadvertently administered a high dose of hydrocodone (an antitussive drug) while suffering from a cold. Additionally, she was administered clarithromycin for an ear infection and valproic acid for seizures. The postmortem blood screen revealed an excessively high-hydrocodone level and, in contrast, barely measurable hydromorphone (biotransformation metabolite of hydrocodone) concentration (29). Hydrocodone is metabolized by CYP2D6, CYP3A4 and afterwards by UGT. There were three reasons for the low metabolism rate:

(1) CYP-polymorphism: the child was found to be a CYP2D6 poor metabolizer,
(2) inhibition of CYP3A4 by clarithromycin and
(3) inhibition of UGT by valproic acid.

Figure 1B shows the Transformer database results for this drug combination. All interactions described in this case are displayed and colored because of enzyme overload. Nevertheless, parts of the available information of drug–enzyme interactions are experimental data and offer no evidence for drug interactions in humans and clinical work, although Q_0 and elimination half-life times could be relevant. The Transformer database, however, provides a platform for detecting mutual drug impairments and could help to appraise the drug response. The database is a comprehensive resource on drug enzyme/transporter interactions and could be a sound starting-point for further research.

The database will be updated yearly to add new drugs/compounds and interactions.

AVAILABILITY

The Transformer database is publicly available via http://bioinformatics.charite.de/transformer and should be used under the terms of the Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.

FUNDING

Funding for open access charge: SynSys (EU Framework 7) and DKTK (BMBF), BMBF Immunotox.

Conflict of interest statement. None declared.

REFERENCES

1. Banerjee,A., Mbamalu,D., Ebrahimi,S., Khan,A.A. and Chan,T.F. (2011) The prevalence of polypharmacy in elderly attenders to an emergency department—a problem with a need for an effective solution. *Int. J. Emerg. Med.*, **4**, 22.
2. Morgan,T.K., Williamson,M., Pirotta,M., Stewart,K., Myers,S.P. and Barnes,J. (2012) A national census of medicines use: a 24-hour snapshot of Australians aged 50 years and older. *Med. J. Aust.*, **196**, 50–53.
3. Sato,I. and Akazawa,M. (2013) Polypharmacy and adverse drug reactions in Japanese elderly taking antihypertensives: a retrospective database study. *Drug Healthc. Patient Saf.*, **5**, 143–150.
4. Lazarou,J., Pomeranz,B.H. and Corey,P.N. (1998) Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA*, **279**, 1200–1205.
5. Phillips,K.A., Veenstra,D.L., Oren,E., Lee,J.K. and Sadee,W. (2001) Potential role of pharmacogenomics in reducing adverse drug reactions: a systematic review. *JAMA*, **286**, 2270–2279.
6. Marusic,S., Bacic-Vrca,V., Obreli Neto,P.R., Franic,M., Erdeljic,V. and Gojo-Tomic,N. (2013) Actual drug-drug interactions in elderly patients discharged from internal medicine clinic: a prospective observational study. *Eur. J. Clin. Pharmacol.*, **69**, 1717–1724.
7. Guengerich,F.P. (2001) Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem. Res. Toxicol.*, **14**, 611–650.
8. Jakoby,W.B. and Ziegler,D.M. (1990) The enzymes of detoxication. *J. Biol. Chem.*, **265**, 20715–20718.
9. Kell,D.B., Dobson,P.D., Bilsland,E. and Oliver,S.G. (2013) The promiscuous binding of pharmaceutical drugs and their transporter-mediated uptake into cells: what we (need to) know and how we can do so. *Drug Discov. Today*, **18**, 218–239.
10. Yan,Q. (2003) Pharmacogenomics of membrane transporters an overview. *Methods Mol. Biol.*, **227**, 1–20.
11. Velamakanni,S., Wei,S.L., Janvilisri,T. and van Veen,H.W. (2007) ABCG transporters: structure, substrate specificities and physiological roles:a brief overview. *J. Bioenerg. Biomembr.*, **39**, 465–471.
12. He,L., Vasiliou,K. and Nebert,D.W. (2009) Analysis and update of the human solute carrier (SLC) gene superfamily. *Hum. Genomics*, **3**, 195–206.
13. Ishikawa,T., Tamura,A., Saito,H., Wakabayashi,K. and Nakagawa,H. (2005) Pharmacogenomics of the human ABC

transporter ABCG2: from functional evaluation to drug molecular design. *Naturwissenschaften*, **92**, 451–463.
14. Landry,Y. and Gies,J.P. (2008) Drugs and their molecular targets: an updated overview. *Fundam. Clin. Pharmacol.*, **22**, 1–18.
15. Albert,A. (1958) Chemical aspects of selective toxicity. *Nature*, **182**, 421–422.
16. Yap,Y.G. and Camm,A.J. (2002) Potential cardiac toxicity of H1-antihistamines. *Clin. Allergy Immunol.*, **17**, 389–419.
17. Dorman,G. and Prestwich,G.D. (2000) Using photolabile ligands in drug discovery and development. *Trends Biotechnol.*, **18**, 64–77.
18. Simplicio,A.L., Clancy,J.M. and Gilmer,J.F. (2007) Beta-aminoketones as prodrugs with pH-controlled activation. *Int. J. Pharm.*, **336**, 208–214.
19. Huttunen,K.M., Mahonen,N., Raunio,H. and Rautio,J. (2008) Cytochrome P450-activated prodrugs: targeted drug delivery. *Curr. Med. Chem.*, **15**, 2346–2365.
20. Topf,N., Worgall,S., Hackett,N.R. and Crystal,R.G. (1998) Regional 'pro-drug' gene therapy: intravenous administration of an adenoviral vector expressing the *E. coli* cytosine deaminase gene and systemic administration of 5-fluorocytosine suppresses growth of hepatic metastasis of colon carcinoma. *Gene Ther.*, **5**, 507–513.
21. Ortiz de Montellano,P.R. (2013) Cytochrome P450-activated prodrugs. *Future Med. Chem.*, **5**, 213–228.
22. Bressler,R. (2006) Grapefruit juice and drug interactions. Exploring mechanisms of this interaction and potential toxicity for certain drugs. *Geriatrics*, **61**, 12–18.
23. Uno,T. and Yasui-Furukori,N. (2006) Effect of grapefruit juice in relation to human pharmacokinetic study. *Curr. Clin. Pharmacol.*, **1**, 157–161.
24. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
25. Sim,S.C. and Ingelman-Sundberg,M. (2010) The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. *Hum. Genomics*, **4**, 278–281.
26. Saier,M.H. Jr, Yen,M.R., Noto,K., Tamang,D.G. and Elkan,C. (2009) The Transporter Classification Database: recent advances. *Nucleic Acids Res.*, **37**, D274–D278.
27. Preissner,S., Kroll,K., Dunkel,M., Senger,C., Goldsobel,G., Kuzman,D., Guenther,S., Winnenburg,R., Schroeder,M. and Preissner,R. (2010) SuperCYP: a comprehensive database on Cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions. *Nucleic Acids Res.*, **38**, D237–D243.
28. von Eichborn,J., Bourne,P.E. and Preissner,R. (2011) Cobweb: a Java applet for network exploration and visualisation. *Bioinformatics*, **27**, 1725–1726.
29. Madadi,P., Hildebrandt,D., Gong,I.Y., Schwarz,U.I., Ciszkowski,C., Ross,C.J., Sistonen,J., Carleton,B.C., Hayden,M.R., Lauwers,A.E. *et al.* (2010) Fatal hydrocodone overdose in a child: pharmacogenetics and drug interactions. *Pediatrics*, **126**, e986–e989.

Nucleic Acids Research Advance Access published November 20, 2015

Nucleic Acids Research, 2015 **1**
doi: 10.1093/nar/gkv1283

CancerResource—updated database of cancer-relevant proteins, mutations and interacting drugs

Bjoern-Oliver Gohlke[1,2], Janette Nickel[1,2], Raik Otto[1], Mathias Dunkel[1] and Robert Preissner[1,2,*]

[1]Charité – University Medicine Berlin, Structural Bioinformatics Group, Institute of Physiology & Experimental Clinical Research Center, Berlin 13125, Germany and [2]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany

Received September 18, 2015; Revised October 23, 2015; Accepted November 04, 2015

ABSTRACT

Here, we present an updated version of CancerResource, freely available without registration at http://bioinformatics.charite.de/care. With upcoming information on target expression and mutations in patients' tumors, the need for systems supporting decisions on individual therapy is growing. This knowledge is based on numerous, experimentally validated drug-target interactions and supporting analyses such as measuring changes in gene expression using microarrays and HTS-efforts on cell lines. To enable a better overview about similar drug-target data and supporting information, a series of novel information connections are established and made available as described in the following. CancerResource contains about 91 000 drug-target relations, more than 2000 cancer cell lines and drug sensitivity data for about 50 000 drugs. CancerResource enables the capability of uploading external expression and mutation data and comparing them to the database's cell lines. Target genes and compounds are projected onto cancer-related pathways to get a better overview about how drug-target interactions benefit the treatment of cancer. Features like cellular fingerprints comprising of mutations, expression values and drug-sensitivity data can promote the understanding of genotype to drug sensitivity associations. Ultimately, these profiles can also be used to determine the most effective drug treatment for a cancer cell line most similar to a patient's tumor cells.

INTRODUCTION

According to the World Health Organisation cancer is one of the most common causes for human death and has been responsible for about 8.2 million cases of death worldwide in the year 2012 (http://www.who.int/mediacentre/factsheets/fs310/en/index312.html). To overcome difficulties in cancer therapy and to develop new methods for cancer diagnosis and treatment a huge amount of information is generated in cancer research experiments like in drug-target assays, high-throughput screenings on cancer cell lines or large-scale cancer genomics projects including next-generation sequencing studies (1–3).

In 2002 after the sequencing of the human genome, Hopkins and Groom established the term 'druggable genome' which comprises proteins that are known (or predicted) to interact with drugs. In their study they reveal an amount of 3051 druggable targets (4). Since then, novel drug targets have been identified that are relevant for cancer and which could be bound by compounds to provoke an activating or inhibiting molecular reaction, e.g. Superoxide dismutase 1 (SOD1). The overexpression of SOD1 results in lung cancer cells' growth and reduces apoptosis (5,6). It could be shown that its enzymatic activity was inhibited by compounds in lung cancer cells leading to growth inhibition of the cancer cells suggesting it as a promising target for cancer therapy (5).

Application of microarray-based gene expression data for cancer research is a broadly used method for identifying significant differentially expressed genes, compared to normal tissue or other cancer tissues, or for profiling cancer signatures, which can be associated with clinical outcome (7–9). Microarray-based gene expression data can even be considered for identifying new therapeutic targets (10) or biomarkers for specific cancer types (11).

Nowadays, as a result of the establishment of next-generation sequencing technologies and improved bioinformatical evaluation a better understanding of the genomic

*To whom correspondence should be addressed. Tel: +49 30 450 540 755; Fax: +49 30 450 540 955; Email: robert.preissner@charite.de

foundation of cancer was achieved (12). Given that cancer is a genetic disease, mutational characteristics of a cancer type can vary from patient to patient even though if the patients are affected by the apparently same cancer type. These genomic alterations might affect an anti-cancer drug's efficacy on the tumor and influence the clinical response. For instance, the anti-cancer drug vemurafenib improves the overall survival rate of patients having the BRAF V600E mutation (13). Consideration of genomic alterations in patients is part of personalized medicine (14) and enables the opportunity of an improved cancer diagnosis and anti-cancer therapy. Nevertheless, the analysis of these data and the understanding of the genotype-phenotype relationship between genomic alterations and anti-cancer drug response remains a major challenge in cancer research (15). The Cancer Genome Atlas (TCGA) project focuses on generating large-scale cancer genomics data sets which are stored by the cBio Cancer Genomics Portal (cBioPortal) which also provides further analysis tools (16).

To support, promote and gain a better insight into these data the updated CancerResource database links gene expression values, gene mutations as well as drug-sensitivity data to cell lines related to cancer. From the inclusion of the data from the consortia 'catalogue of somatic mutations in cancer' (CoSMIC) (17), 'Cancer Cell Line Encyclopedia' (CCLE) (18) and the CellMiner database (17) an explorative data analysis is enabled helping to achieve a better understanding of specific drug response in cancer.

MATERIALS AND METHODS

Gene expression, mutation and drug sensitivity data

The cancer cell line expression, mutation and drug sensitivity data are provided by the CCLE (18), CoSMIC (19) and CellMiner (17) websites, respectively.

All expression data are based on the Affymetrix HG-U133 Plus 2.0 technology. In order to increase the comparability of the expression data from three different sources, all gene expression values were scaled and centered. To determine the similarity between cancer cell lines two similarity measurements are used: first, the Pearson correlation distance; second, the percentage of a categorical classification of genes based on the fold change between the genes of two cancer cell lines.

Only somatic mutations were included in the similarity analyses and germline variants such as single nucleotide polymorphisms (SNPs) were excluded. The mutational status of a gene was modeled boolean: gene is mutated, a one was set; in case a wild-type was present, a zero was set. Similarity of two cancer cell lines based on their mutations is the amount of shared mutations, divided by all observed mutations of both cancer cell lines. The similarity calculations for mutations are accordingly based on the Tanimoto coefficient calculation (20).

Drug sensitivity data provided by CellMiner were measured as growth inhibition (GI50). Data provided by CCLE and CosMIC are given by IC50 and EC50 values. In case the primary target of a drug is known, the consortia provide information about them. The activity data of the consortia were scaled and centered to obtain a uniform format of the data. The normalized data were used to create the cellular fingerprint for every compound considering each consortium separately when a compound occurred in more than one consortium (21).

All cancer cell lines from one consortium have been compared to each other. The *P*-value calculations are based on a normal distribution to the observed similarity for all cancer cell lines of the same consortium. Cancer cell lines, whose observed similarity differs from the expected percentage of similarity in either direction with a *P*-value of less or equal to 0.05 were counted as significantly similar or dissimilar.

Compound mapping and target identification

Protein targets for the CancerResource update have been obtained from the ChEMBL database v. 19 (22). Those targets were filtered using the following criteria: First, all interactions with an activity comment 'inactive', 'inconclusive' or 'not active' were removed. Second, only 'homo sapiens' was considered as organism and third, only protein target types were extracted. The structures of the CancerResource compounds and the ChEMBL compounds have been standardized using JChem (Instant Jchem version 14 10.27.0, ChemAxon (http://www.chemaxon.com/)) for identifying CancerResource compounds in ChEMBL. The standardization steps included aromatization of the structures and addition of explicit hydrogens. Furthermore, solvents and salts were removed. Additionally, 3D structures were generated. For comparing the standardized CancerResource compounds with the equally standardized ChEMBL compounds InChiKeys were calculated and used for compound identification. Furthermore, drug-target information from CTD (23), TTD (24), PharmGKB (25) and DrugBank (26) has been added to the database. The final data set includes 91 000 interactions whereat 11 000 compounds and 3400 targets are involved. Where available, 3D structures are linked via Cancer3D (27).

Compound similarity

The structural similarity search for uploaded structures is based on extended-connectivity fingerprints (ECFP). For the computation of the circular topological fingerprints the diameter, which defines the circular neighborhood considered for each atom, was set to 4 (ECFP4). The calculation of these fingerprints was performed by the cheminformatics toolkit of ChemAxon (JChem compr (14.10.20.0), 201n (2014), ChemAxon (http://www.chemaxon.com/)). All other parameters provided by ChemAxon were used in default configuration. To determine the similarity between the compounds stored in the database, the Tanimoto coefficient is calculated. The Tanimoto calculation on the website is performed by MyChem (http://mychem.sourceforge.net/).

Pathways

To achieve a better understanding of drug-target interactions at molecular level, KEGG (signaling) pathways (28) were analyzed according to their relevance in cancer emergence and cancer development. This set comprises cancer-specific pathways, pathways related to cell-cycle regulation, replication, immune response and drug metabolism. Pathway maps are dynamically retrieved via Web service from

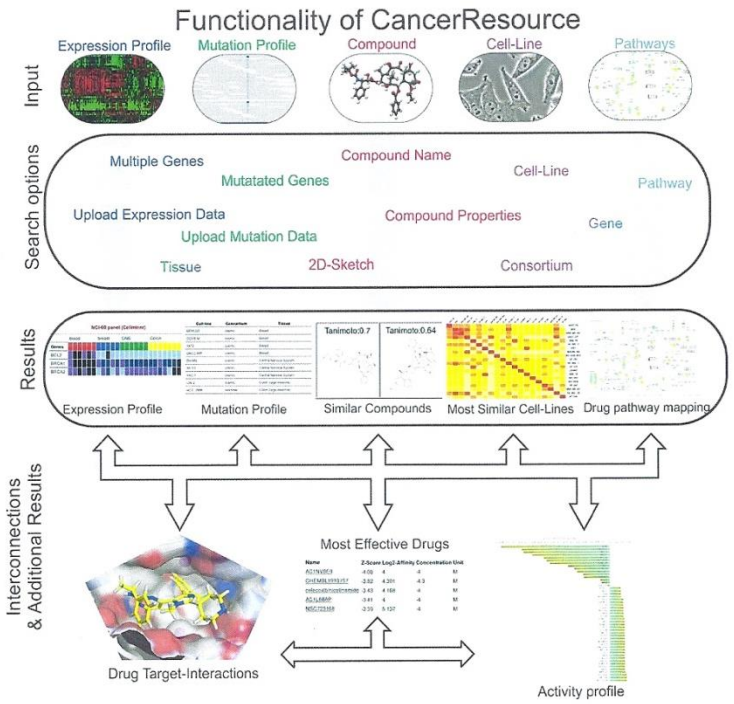

Figure 1. Overview of the functionality of CancerResource illustrating input variants, search options and result variants of the database.

KEGG facultative with highlighted expression data if gene expression is computed online before.

Server, database and system requirements

CancerResource is based on a relational MySQL database (http://www.mysql.com/). The database is normalized to the third normal form, for which large tables were split into smaller ones to minimize redundancy and dependency. The website of CancerResource is build using PHP (http://www.php.net/), JavaScript (http://www.java.com/), Ajax and web access is enabled via an Apache HTTP Server (http://apache.org/). For optimal usage we strongly recommend the latest version of Mozilla Firefox, Google Chrome or Safari browser and Internet Explorer, in descending order, with JavaScript option enabled.

RESULTS

CancerResource is comprised of about 50 000 compounds with detailed information like synonyms, structure identifier (SMILES, InChiKeys) as well as hydrogen bonds, molecular weight and log*P* values. Additionally, the database provides links to PubChem. Furthermore, about 3400 protein targets could be identified for the compounds stored in the database. This results in about 91 000 compound-target interactions. By integrating CCLE and CoSMIC into CancerResource the total number of cancer cell lines now exceeds 2000. Mutation information for 19 834 genes, expression values for 23 016 genes and about 872 658 mutations from the consortia that were collected and included and are now available in CancerResource.

The CancerResource website provides different possibilities for the user to start using the database. Regardless of how the user begins a search, all results are interconnected via different joining's. An overview of the multiple search forms is displayed in Figure 1. These search options are described in more detail in the following sections.

Compound search for alternative, most effective drugs

CancerResource can easily be searched for compounds. For this purpose, two search categories are available. On the one hand, the database can be browsed by the compound's properties like name, molecular weight, number of atoms or log*P* value. On the other hand, a connection to PubChem is established by which the user can search by compound name or SMILES. If no structures could be found via PubChem, a structure can be drawn by the molecule sketching tool. If the structure is available in the database detailed information about physicochemical properties, drug-target interactions, pathway information and an activity profile to all available

cancer cell lines are presented. Furthermore, the ten most similar compounds are listed for which detailed information can be displayed interactively.

Gene/target search

Targets can be found via different gene identifiers. Extended information about the target is displayed including expression profiles and drugs interacting with the target. Information about cancer cell lines, where the gene is mutated is also provided (mutation profile). If target-pathway mapping is possible, cancer relevant pathways are displayed.

Cell line/expression profile search

To search CancerResource for mRNA expression profiles of genes of interest different variants are prepared for the user. Besides searching for mRNA expression profiles of several selected genes between different cancer cell lines, the user can also upload an external expression profile to compare it to the database. As a result, the most similar cancer cell lines are presented to the user for which again detailed information can be displayed interactively. The results are accessible for at least one month by accessing the data via a bookmark of the web page.

A direct search for a specific cancer cell line or tissue type is also provided on the website. The results also embrace the most effective drugs against the cancer cell line as well as a tabular presentation of the similarity to other cancer cell lines based on a compound's activity-, mutation- and expression-fingerprint, respectively.

Mutation profile search

Searching for mutation profiles is enabled on the website. On the one hand, the user can search for gene mutations that occur in cancer-relevant genes. On the other hand, the user can search for gene mutations that occur in a cancer cell line of interest. Additionally, a tissue specific search is included. Furthermore, the opportunity to upload a mutation profile of user-provided tumor cells was implemented to compare them to well-characterized cancer cell lines and to identify the most similar cancer cell line based on mutations.

Pathway search

To provide a detailed insight into cancer relevant pathways a search by pathway names is provided. For this purpose pathway maps were extracted from the KEGG database. All targets of those maps for which compound-target interactions are available in CancerResource are highlighted. A mouse-over for those targets is made available displaying the binding compounds in a pop-up. Based on the chosen option, all cancer cell lines with certain mutated gene(s) or all mutations occurring in one specific cancer cell line are displayed.

From this site interactive browsing to detailed information, to cancer cell lines or genes is available to collect further information like most effective drugs or to compare the cancer cell line to others.

Similarity comparison

The user interface provides four ways to measure similarity of cancer cell lines. Similarity is calculated by the activity profile of compounds for two cancer cell lines, a Pearson distance correlation of expression values of genes, percentage similarity of known mutations of genes and a similarity measurement of a categorical classification of genes based on the fold changes between the genes of two cancer cell lines. Additionally, two options are provided to the user to calculate the similarity of own data to cancer cell line data from the three consortia. The results are presented in heat maps and differentially expressed genes are displayed.

Comparison to other databases

A detailed comparison to the original CancerResource (29), canSAR (30) and CancerDR (31) is represented in Table 1. The upload of external mutation data and expression values for cell lines or patient data to find the most similar cancer cell lines in the database is considered as a unique feature of the new version of CancerResource. Another feature of CancerResource is the mapping and annotation of cancer relevant protein targets to KEGG pathways. An additional extension of the updated CancerResource compared to other databases is the integration of mutation, expression and drug sensitivity data from the CCLE, CoSMIC and CellMiner consortia and to provide additionally a dynamic drug sensitivity comparison for external mutation or expression data. By this tool the user is enabled to create own hypotheses, which might possibly not have been developed by taking only one of the consortia into account.

USE CASES

Upload of own gene expression or mutation data to find an alternative, most effective drug for a tumor similar to a cancer cell line

CancerResource can be used to identify the most similar cancer cell line of an external tissue sample by using either gene expression or mutation data. Therefore, normalized data from Human Genome U133A or Human Genome U133 Plus 2.0 microarray chips from Affymetrix can be queried either by Affymetrix probe set IDs or HGNC gene symbols. Based on the most similar cancer cell line the most effective drug for the external tissue is determined. To identify the most similar cancer cell line by applying gene expression data a Pearson correlation distance and additionally the percentage of the categorical classification of genes based on the fold change between the uploaded data and all cancer cell lines stored in CancerResource are calculated. Interconnections to most effected genes in the resulting cell line and additional similar cell lines are displayed as heatmaps. Alternatively, gene identifiers of mutated genes of the external tissue can be used to query the database for the most similar cancer cell line of the respective consortia. The next step gives an overview of the most effective drugs for the determined similar cancer cell line of the input data. A visualization of this use case is shown in Figure 2.

Table 1. Comparison of the updated CancerResource database with the original CancerResource, canSAR and CancerDR databases

	Update Cancer Resource	Cancer Resource (29)	canSAR (30)	CancerDR (31)
Expression, Mutation and Drug Sensitivity	All	Expression and drug sensitivity	All	Only mutation and drug sensitivity
Cellminer, CCLE and CoSMIC	All	Only CellMiner	All	Only CCLE and CoSMIC
Dynamic Drug Sensitivity Comparison	Yes	Yes	No	No
Pathways	Yes	Yes	Yes	Overview
Cell lines	2037	60	11 000	952
Drugs with Drug Sensitivity Data	48 404	≈40 000	16 000	148
Mutated Genes	19 799	No	Yes	Yes
No. of Genes	23 016	11 964	3466 studies	
Protein Targets	3387	2392	All	116
Integrated Similarity Measurements	Yes	Expression	No	No
Upload of external mutation data or expression values	Yes	Only expression	No	No

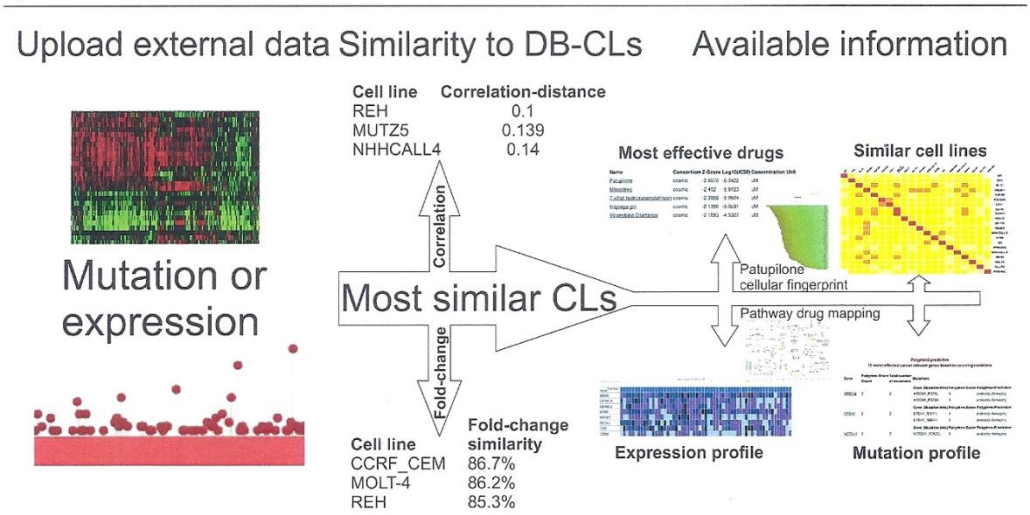

Figure 2. Use case—upload of external mutation or mRNA expression data to find similar cancer cell lines and alternative/most effective drugs for the external sample. Expression and mutation profiles for selected genes are available and in addition a mapping of genes to cancer relevant pathways is enabled.

Cancer cell line compound response

Almost 50 000 compounds that are stored in CancerResource have been screened against about 2000 cancer cell lines to determine their drug sensitivity. These data have been made available in the database. To identify which cancer cell line has a high or low sensitivity towards a compound, a similarity search is implemented within the database comparing a query compound to all compounds of CancerResource. If the query compound is found to be identical to a database compound, the user will directly be passed to the compound's details page. Otherwise, the ten most similar compounds found in the database are displayed. The 'similar property principle' (32) forms the bases for the similarity search and states that similar compounds might have similar properties in this case similar cancer cell line compound responses. By choosing one of the similar compounds the user will be forwarded to the compound's details page. On the details page of the identical or similar compound(s) the cellular fingerprint of the compound is displayed showing high and low responding cancer cell lines.

DISCUSSION AND OUTLOOK

Cancers, even from the same tissue, are extremely divergent in terms of gene alterations and therapy resistance. Therefore, individual therapy is required and will be made possible by understanding the entirety of single nucleotide polymorphisms (SNPs), complete or partial gene deletions, copy number variations, gene aberrations, gene fusions etc. All those issues may cause substantial dysfunctions or defected genes that have influence on gene regulation. The heterogeneity of tumors (33) and their reaction on chemotherapy (or other treatments) causes a further challenge, which will be addressed in a new release of CancerResource. Recently, tumor stratification not only based on somatic mutations but also on epigenetic changes like methylation has been

proven successful (34) and should become part of a future update to further support the development of personalized therapies. The data content of CancerResource is going to be updated in a yearly pattern based on the regular updates of the source data, which occur to be in a time period between three months and two years.

FUNDING

Funding for open access charge: Berlin-Brandenburg research platform BB3R, Federal Ministry of Education and Research (BMBF), Germany [031A262C]; Immunotox project, Federal Ministry of Education and Research (BMBF), Germany [031A268B]; German Cancer Consortium (DKTK).
Conflict of interest statement. None declared.

REFERENCES

1. Li,T., Ziniel,P.D., He,P.Q., Kommer,V.P., Crowther,G.J., He,M., Liu,Q., Van Voorhis,W.C., Williams,D.L. and Wang,M.W. (2015) High-throughput screening against thioredoxin glutathione reductase identifies novel inhibitors with potential therapeutic value for schistosomiasis. *Infect. Dis. Poverty*, **4**, 40.
2. Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–823.
3. Tomczak,K., Czerwinska,P. and Wiznerowicz,M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contempor. Oncol. (Poznan, Poland)*, **19**, A68–A77.
4. Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.
5. Somwar,R., Erdjument-Bromage,H., Larsson,E., Shum,D., Lockwood,W.W., Yang,G., Sander,C., Ouerfelli,O., Tempst,P.J., Djaballah,H. *et al.* (2011) Superoxide dismutase 1 (SOD1) is a target for a small molecule identified in a screen for inhibitors of the growth of lung adenocarcinoma cell lines. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 16375–16380.
6. Papa,L., Manfredi,G. and Germain,D. (2014) SOD1, an unexpected novel target for cancer therapy. *Genes Cancer*, **5**, 15–21.
7. Axelsen,J.B., Lotem,J., Sachs,L. and Domany,E. (2007) Genes overexpressed in different human solid cancers exhibit different tissue-specific expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 13122–13127.
8. Chibon,F. (2013) Cancer gene expression signatures—the rise and fall? *Eur. J. Cancer*, **49**, 2000–2009.
9. Sotiriou,C. and Pusztai,L. (2009) Gene-expression signatures in breast cancer. *N. Engl. J. Med.*, **360**, 790–800.
10. Sandhu,R., Parker,J.S., Jones,W.D., Livasy,C.A. and Coleman,W.B. (2010) Microarray-Based Gene Expression Profiling for Molecular Classification of Breast Cancer and Identification of New Targets for Therapy. *LabMedicine*, **41**, 364–372.
11. Kafetzopoulou,L.E., Boocock,D.J., Dhondalay,G.K., Powe,D.G. and Ball,G.R. (2013) Biomarker identification in breast cancer: Beta-adrenergic receptor signaling and pathways to therapeutic response. *Comput. Struct. Biotechnol. J.*, **6**, e201303003.
12. Meyerson,M., Gabriel,S. and Getz,G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.*, **11**, 685–696.
13. Chapman,P.B., Hauschild,A., Robert,C., Haanen,J.B., Ascierto,P., Larkin,J., Dummer,R., Garbe,C., Testori,A., Maio,M. *et al.* (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.*, **364**, 2507–2516.
14. Roychowdhury,S., Iyer,M.K., Robinson,D.R., Lonigro,R.J., Wu,Y.M., Cao,X., Kalyana-Sundaram,S., Sam,L., Balbin,O.A., Quist,M.J. *et al.* (2011) Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci. Transl. Med.*, **3**, 111–121.
15. Vuong,H., Cheng,F., Lin,C.C. and Zhao,Z. (2014) Functional consequences of somatic mutations in cancer using protein pocket-based prioritization approach. *Genome Med.*, **6**, 81.
16. Cerami,E., Gao,J., Dogrusoz,U., Gross,B.E., Sumer,S.O., Aksoy,B.A., Jacobsen,A., Byrne,C.J., Heuer,M.L., Larsson,E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.
17. Reinhold,W.C., Sunshine,M., Liu,H., Varma,S., Kohn,K.W., Morris,J., Doroshow,J. and Pommier,Y. (2012) CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res.*, **72**, 3499–3511.
18. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehar,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
19. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
20. Willett,P., Barnard,J.M. and Downs,G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, **38**, 983–996.
21. Fullbeck,M., Dunkel,M., Hossbach,J., Daniel,P.T. and Preissner,R. (2009) Cellular fingerprints: a novel approach using large-scale cancer cell line data for the identification of potential anticancer agents. *Chem. Biol. Drug Design*, **74**, 439–448.
22. Bento,A.P., Gaulton,A., Hersey,A., Bellis,L.J., Chambers,J., Davies,M., Kruger,F.A., Light,Y., Mak,L., McGlinchey,S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.
23. Davis,A.P., Grondin,C.J., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., King,B.L., Wiegers,T.C. and Mattingly,C.J. (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, **43**, D914–D920.
24. Zhu,F., Shi,Z., Qin,C., Tao,L., Liu,X., Xu,F., Zhang,L., Song,Y., Liu,X., Zhang,J. *et al.* (2012) Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.*, **40**, D1128–D1136.
25. Whirl-Carrillo,M., McDonagh,E.M., Hebert,J.M., Gong,L., Sangkuhl,K., Thorn,C.F., Altman,R.B. and Klein,T.E. (2012) Pharmacogenomics knowledge for personalized medicine. *Clin. Pharm. Ther.*, **92**, 414–417.
26. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y., Maciejewski,A., Arndt,D., Wilson,M., Neveu,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.
27. Porta-Pardo,E., Hrabe,T. and Godzik,A. (2015) Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res.*, **43**, D968–D973.
28. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
29. Ahmed,J., Meinel,T., Dunkel,M., Murgueitio,M.S., Adams,R., Blasse,C., Eckert,A., Preissner,S. and Preissner,R. (2011) CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic Acids Res.*, **39**, D960–D967.
30. Bulusu,K.C., Tym,J.E., Coker,E.A., Schierz,A.C. and Al-Lazikani,B. (2014) canSAR: updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res.*, **42**, D1040–D1047.
31. Kumar,R., Chaudhary,K., Gupta,S., Singh,H., Kumar,S., Gautam,A., Kapoor,P. and Raghava,G.P. (2013) CancerDR: cancer drug resistance database. *Sci. Rep.*, **3**, 1445.
32. Johnson,M.A. and Maggiora,G.M. (1990) *Concepts and applications of molecular similarity*. John Willey & Sons, NY.
33. Marusyk,A. and Polyak,K. (2010) Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta*, **1805**, 105–117.
34. Cebrian,V., Fierro,M., Orenes-Pinero,E., Grau,L., Moya,P., Ecke,T., Alvarez,M., Gil,M., Algaba,F., Bellmunt,J. *et al.* (2011) KISS1 methylation and expression as tumor stratification biomarkers and clinical outcome prognosticators for bladder cancer patients. *Am. J. Pathol.*, **179**, 540–546.

Nucleic Acids Research Advance Access published November 8, 2015

Nucleic Acids Research, 2015 1
doi: 10.1093/nar/gkv1192

WITHDRAWN—a resource for withdrawn and discontinued drugs

Vishal B. Siramshetty[1], Janette Nickel[2,3], Christian Omieczynski[2], Bjoern-Oliver Gohlke[2,3], Malgorzata N. Drwal[2,*] and Robert Preissner[2,3,4,*]

[1]Structural Bioinformatics Group, ECRC Experimental and Clinical Research Center, Charité – University Medicine Berlin, 13125 Berlin, Germany, [2]Structural Bioinformatics Group, Institute of Physiology, Charité – University Medicine Berlin, 13125 Berlin, Germany, [3]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany and [4]BB3R – Berlin Brandenburg 3R Graduate School, Freie Universität Berlin, 14195 Berlin, Germany

Received August 14, 2015; Accepted October 25, 2015

ABSTRACT

Post-marketing drug withdrawals can be associated with various events, ranging from safety issues such as reported deaths or severe side-effects, to a multitude of non-safety problems including lack of efficacy, manufacturing, regulatory or business issues. During the last century, the majority of drugs voluntarily withdrawn from the market or prohibited by regulatory agencies was reported to be related to adverse drug reactions. Understanding the underlying mechanisms of toxicity is of utmost importance for current and future drug discovery. Here, we present WITHDRAWN, a resource for withdrawn and discontinued drugs publicly accessible at http://cheminfo.charite.de/withdrawn. Today, the database comprises 578 withdrawn or discontinued drugs, their structures, important physico-chemical properties, protein targets and relevant signaling pathways. A special focus of the database lies on the drugs withdrawn due to adverse reactions and toxic effects. For approximately one half of the drugs in the database, safety issues were identified as the main reason for withdrawal. Withdrawal reasons were extracted from the literature and manually classified into toxicity types representing adverse effects on different organs. A special feature of the database is the presence of multiple search options which will allow systematic analyses of withdrawn drugs and their mechanisms of toxicity.

INTRODUCTION

Efficacy and safety are two decisive factors that affect the viability of a chemical entity while furthering in the drug discovery pipeline. Consequently, the financial burden on pharmaceutical companies grows higher when the chemical entities tend to fail in late stages of clinical trials (1). However, a significant number of new chemical entities (NCEs) were recalled from the market post to their regulatory approval due to various reasons ranging from inefficiency to severe side-effects to financial and regulatory concerns. Adverse drug reactions (ADRs) not only account for market withdrawals but also for changes in labels or introduction of new black-box warnings for prescription drugs (2). ADRs can be interpreted either as primary effects elicited after modulation of the therapeutic (or primary) target or unintended effects due to interactions with off-targets. In few instances, the primary target is expressed in multiple organs and simultaneously targeted, leading to the therapeutic effect in the target tissue and unwanted effects in other tissues.

A well-known class of drugs that cause adverse reactions due to their activity at primary target are antiarrhythmic drugs, the benefits of which are, in few cases, hindered due to aggravation of arrhythmia which is the indication being treated (3). This effect is due to modulation of the alpha subunit of a potassium ion channel (human Ether-à-go-go-related gene, hERG), which is primarily associated with regulation of cardiac action potentials (4). The hERG channel is also a prominent off-target example whose unintended modulation can cause severe side-effects. This has ultimately lead to market withdrawal of drugs inhibiting the hERG channel, a classical example being the withdrawal of the antihistaminic drug terfenadine due to severe arrhythmias and death (5).

Although there is much progress in elucidation and understanding of the mechanisms leading to drug related toxic effects, gaining clearer insights about these effects at cellular and biochemical level is much needed to appropriately adjust or reinvent the development strategies so as to overcome the attrition during clinical trial phases of drug

*To whom correspondence should be addressed. Tel: +4930450540755; Fax: +4930450540955; Email: malgorzata.drwal@charite.de or robert.preissner@charite.de

discovery and withdrawal after drug approval (6–8). This toxicological knowledge could be used to develop a panel of relevant *in vitro* assays that could mechanistically examine the effects and profile the propensity of drugs to cause ADRs (9). In contrast to the majority of ADR cases which are relatively frequent and mostly dose-dependent, few side-effects are idiosyncratic drug reactions (IADRs), i.e. the extremely rare drug reactions which occur unpredictably in a population. The target organs that are most commonly associated with idiosyncratic events include liver, cardiovascular and central nervous systems (10–12). Hepatocellular and cholestatic drug-induced liver injury (DILI), liver failure and hepatic necrosis are the common patterns of IADRs associated with the liver. Limited knowledge exists to understand the underlying mechanisms of such IADRs. However, it is apparent that IADRs develop via complex mechanisms which are subjective to both differential patient responses and drug combination effects that result from simultaneous triggering of multiple off-targets (13). Factors associated with differential patient responses include genetic attributes like single nucleotide polymorphisms (SNPs) and mutations, and non-genetic attributes such as gender, age and co-treatments (14). Drug-induced events are a result of various effects ranging from direct activity on organs (e.g. on cardiovascular systems) to reactivity of active metabolites of drugs to interactions with biological transporters (15).

Over the decades, drug regulatory agencies, pharmaceutical companies and various clinical studies have reported the events of drug withdrawals due to side-effects (16–18). About 2.3 million adverse event reports were collected against ~6000 marketed drugs between 1969 and 2002 (19). Yet, only a small proportion (75 drugs; ~1%) of these marketed drugs were withdrawn during this period. Another study reported that ~95 drugs were documented to be withdrawn due to death as the primary reason between 1950 and 2013 (17). However, not all of these drugs were withdrawn world-wide. Most drugs were reported to be withdrawn in the Unites States and European countries.

Several public resources contain information relevant to drug withdrawals (e.g. websites from regulatory agencies, World Health Organization's consolidated list for withdrawn drugs and scientific literature). However, in many cases, the information is hidden in regulatory documents and not easily accessible, impeding comprehensive analyses. Furthermore, there exists no single resource reporting a complete list of drugs withdrawn due to safety concerns. In order to allow access to a variety of information related to drug withdrawals as well as shed light on the mechanisms of ADRs, we here present WITHDRAWN—a resource for withdrawn and discontinued drugs. We collected a list of more than 500 drugs/drug products, which were withdrawn or discontinued in at least one country, and assembled information regarding their molecular targets, pathways and toxicities. For approximately half of the drugs, extensive literature search revealed that toxic events are associated with the withdrawal. Thus, WITHDRAWN can be seen as a platform to understand the mechanisms for severe ADRs due to primary and off-target interactions of drugs, simultaneous perturbation of complex biological pathways and genetic polymorphisms (SNPs). Furthermore, it provides multiple search options to systematically analyse molecules of interest by performing different types of molecular similarity search across the database's drugs and can be a valuable resource for scientists in the drug development and toxicity prediction field.

MATERIALS AND METHODS

Withdrawn and discontinued drugs

A number of resources including the drug collections from the U.S. Food and Drug Administration (FDA; http://www.fda.gov/), the European Medicines Agency (EMA; http://www.ema.europa.eu/ema/), peer-reviewed literature (17), public databases such as DrugBank (20), e-Drug3D (21) and text-books (16) were searched in order to extract information on drug withdrawals. Monoclonal antibodies and substance combinations were removed from the dataset. Currently, the database comprises two sets of drugs: withdrawn and discontinued. A total of 270 drugs, that were identified to be withdrawn or recalled in at least one country/market due to safety issues are included in the former set while the latter consists of 308 drugs that were suspended or discontinued in at least one market due to unclear reasons. The chemical structures of the withdrawn/discontinued drugs were standardized using the JChem Suite (Instant JChem version 14.10.27.0, ChemAxon (http://www.chemaxon.com)). The standardization steps included aromatization of the structures, addition of explicit hydrogens, removal of salts, and generation of 3D structures. InChIKeys were calculated for the standardized structures and used to join structures from different datasets and to remove duplicates. In addition to InChIKeys, the set was scanned for duplicates using chemical names, canonical smiles and external identifiers.

In many cases, the reason(s) for withdrawal and associated toxicity was directly provided by the source. The reasons were manually extracted for the remaining drugs by performing literature search. Furthermore, the years of first approval, first and last withdrawal, and the year of first reported death for all the withdrawn drugs and most of the discontinued drugs were extracted from the literature. Additionally, the Anatomical Therapeutic Chemical (ATC) codes and external chemical identifiers were collected to link the drugs to the public databases WHO ATC index (http://www.whocc.no/atc_ddd_index/), ChEMBL (22) and PubChem (23), respectively. External identifiers were extracted using the PubChem Identifier Exchange Service (https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi) whereas the ATC codes were collected by looking for drug names in the WHO ATC index. For those drugs without an ATC code assigned by the WHO, pseudo-ATC class names were assigned based on their primary indication areas. The acute oral toxicity class was calculated for each drug using the ProTox webserver (24). The toxicity classes (ranging from 1 to 6) are based on the Globally Harmonized System of Classification and Labelling of Chemicals (GHS; https://www.osha.gov/dsg/hazcom/ghs.html) which classifies compounds using their median lethal doses (LD_{50}). Drugs that demonstrated very low structural similarity to the ProTox dataset were assigned to the class 0.

Protein targets

Human protein targets for withdrawn and discontinued drugs were obtained from the Comparative Toxicogenomics Database (CTD) (25) and the ChEMBL database v. 19 (22). The targets from CTD were filtered to obtain only interactions with the interaction types involving activity, binding, transport or metabolic processing. The ChEMBL targets were filtered using the following criteria, adapted from the recommendations on search criteria by Bajorath *et al.* (26). First, all interactions with an activity comment 'inactive', 'inconclusive' or 'not active' were removed. Second, only interactions with nanomolar (nM) standard units were kept. Third, all interactions with a confidence score below 4 were deleted to remove all non-protein targets. Fourth, only interactions with standard activity relations '=', '<', '<<', '<=', '==' and those without a standard activity relation were kept. In the last step, all interactions marked with target types as cell-line and ADMET were omitted to retain only interactions those with protein targets measured in functional or binding assays. As a result, we retained a total of 1.4 million compound-target interactions. Target interactions were assigned to the withdrawn/discontinued drugs by mapping the ChEMBL/CTD compound identifiers which resulted in a total of 20,558 drug-target interactions. These involved 327 drugs and 946 distinct human protein targets. To provide additional information concerning adverse effects, drug-target interactions were classified into therapeutic and potential off-targets. Therapeutic or primary drug targets were identified using mechanism of action information from ChEMBL (22), primary target information from PDB (27), pharmacological action from Drugbank (20) as well as the Therapeutic Target Database TTD (28). Information regarding targets considered as off-targets was gathered from the Novartis Safety Panel list published by Lounkine *et al.* (29).

Enriched pathways

In order to emphasise the interpretation of drug-target interactions at molecular level, we enriched the biological pathways from ConsensusPathDB (30) using the human protein targets from our database. A total of 149 KEGG pathways were enriched with an enrichment P-value > 0.01 while ensuring that at least two protein targets are involved in each pathway. The 149 enriched pathways comprise different signaling, metabolic and biochemical pathways in addition to the drug-target interaction pathways. Altogether, 703 human protein targets were found to be involved in the enriched pathways.

Genetic variations

Information on genetic variations, or widely known as single nucleotide polymorphisms (SNPs), were extracted from the dbSNP database (31). To extract the SNP information from dbSNP for the human protein targets within our database, the BioMart R package (32) was used. The human genome assembly GRCh38.p3, provided by the Ensembl database (33), was used as a reference genome. SNP information extraction started with a collection of gene symbols or names as defined by the HUGO Gene Nomenclature Committee (HGNC) database (34). The EnsemblMart was queried for HGNC symbols and the corresponding Ensembl transcript identifiers were extracted for each gene. The chromosomal position was identified for each transcript and SNP identifiers were used to get additional information including minor allele frequency (MAF) and function predictions from SNP-Mart. This information was mapped to the genes queried for on Ensemble-Mart using the SNP identifiers and transcript identifiers. In order to identify the most important variations, only those SNPs located within the coding region of a protein and marked as missense variants with an MAF value were retained. A total of 889 human protein targets were identified to be associated with 27 790 unique SNP identifiers. In total, 1731 SNPs have a MAF >1%.

Toxicity types

A total of 14 categories of toxicity types were defined based on the adverse effects associated with drug withdrawal. These include the following toxicity types: hepatic, cardiovascular, haematological, dermatological, carcinogenic, neurological, renal, gastrointestinal, ophthalmic, muscular, reproductive and respiratory toxicity as well as the type 'multiple toxicities' comprising compounds with observed multiple organ failure as well as 'unknown toxicity' where no specific toxic effect could be identified, although a safety issue was associated with the withdrawal. The toxicity types were manually assigned based on the reasons available and also the reasons extracted from the literature. The number of withdrawn/discontinued drugs associated with each toxicity type is summarized in Figure 1 and Supplementary Table S1.

Server, database and system requirements

WITHDRAWN is based on a relational MySQL database (http://www.mysql.com/). All data is stored on the MySQL database and WITHDRAWN is hosted as a Java web application on a Linux virtual server, accessible at http://cheminfo.charite.de/withdrawn. We strongly recommend using a latest Mozilla Firefox, Google Chrome or Safari browser, with JavaScript options enabled, to access the website.

DATABASE SEARCH OPTIONS

The data presented by WITHDRAWN can be queried via multiple search forms, as summarized in Figure 2. A quick and simple way is to browse through the lists of withdrawn and discontinued drugs. Different search options available on the database include.

Drug search

Drugs can be searched using multiple options. In case a direct match by name or synonym is not possible, the structure of the queried name is obtained from PubChem and five most similar withdrawn/discontinued drugs will be identified and displayed to the users. When providing a structure

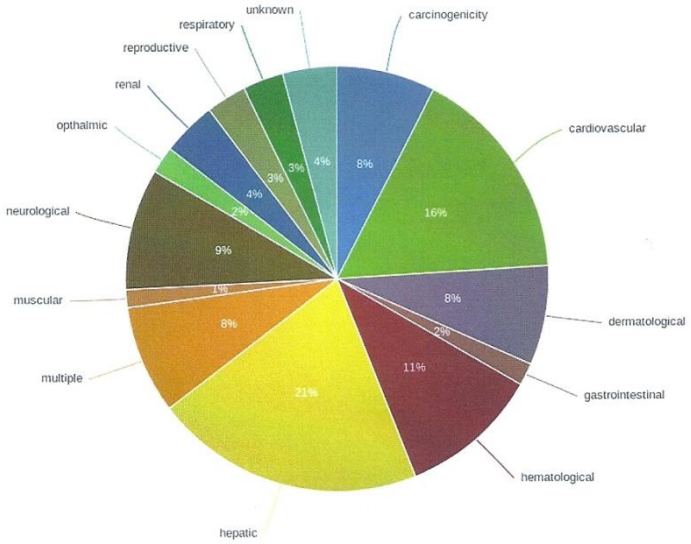

Figure 1. Overview of toxicity types associated with drug withdrawals.

input via the molecule sketching tool, the user has the flexibility to search for database compounds at different levels of Tanimoto similarity (fingerprint similarity using MACCS keys) and also to adjust the number of results to be displayed. In addition, a sub-structure search, using Ullmann's algorithm for subgraph isomerism (35), was implemented to provide an option to lookup for withdrawn/discontinued drugs that contain the query structure. Additionally, drugs can be searched using ATC codes. A detailed drug record displays information about drug withdrawal, physicochemical properties and links to external databases. The users can also view the target interactions of the selected drug. Two separate tables for ChEMBL and CTD interactions are displayed. ChEMBL interactions can additionally be filtered using different activity value cutoffs.

Target search

The users can search for protein targets by providing a gene name, UniProt entry number or UniProt entry name (36) as query in the target search form. In addition, it is possible to browse protein targets using their ChEMBL classification. The resulting target record displays various protein identifiers, PDB (http://www.rcsb.org) structures, and links to external target databases. In addition, the interactions of the target with withdrawn/discontinues drugs can be viewed in the same page. The information includes activity types, units and values as well as the organism and information source. Furthermore, the information on biological pathways and SNPs, including amino acid changes, peptide positions, MAFs, PolyPhen scores (37) and links to dbSNP, were added in the detailed record of a target.

Pathway search

To provide clear insights on withdrawn drug-target interaction effects, the pathway maps were extracted from the KEGG database (38,39) for all the enriched biological pathways. In every pathway map, the targets that have an interaction with withdrawn drugs are highlighted. Pathways can be accessed via a selection list. Additionally, the targets highlighted within the map are listed below to provide a link to interacting drugs.

Toxicity type search

Alternatively, the drugs can be browsed by toxicity type. An interactive wheel was designed to visualize different toxicity types using the open source D3 visualization libraries (http://d3js.org/). The users can see number of drugs in each toxicity type as well as the distribution of the drugs into different ATC classes within each toxicity type. Furthermore, the list of drugs classified in each toxicity type can be exclusively viewed by clicking on the toxicity type. Major withdrawal reasons under each toxicity type are summarized in Figure 1 and Supplementary Table S1.

USE CASE

The following use case, represented in Figure 3, illustrates the utility of WITHDRAWN as a knowledge-base to understand the mechanism of adverse drug reactions associated with drug withdrawals:

A search for the drug sibutramine, originally developed by Knoll Pharmaceuticals, as an appetite suppressant for treatment of exogenous obesity reveals that it was recalled in the USA in 2010 due to adverse cardiovascular events

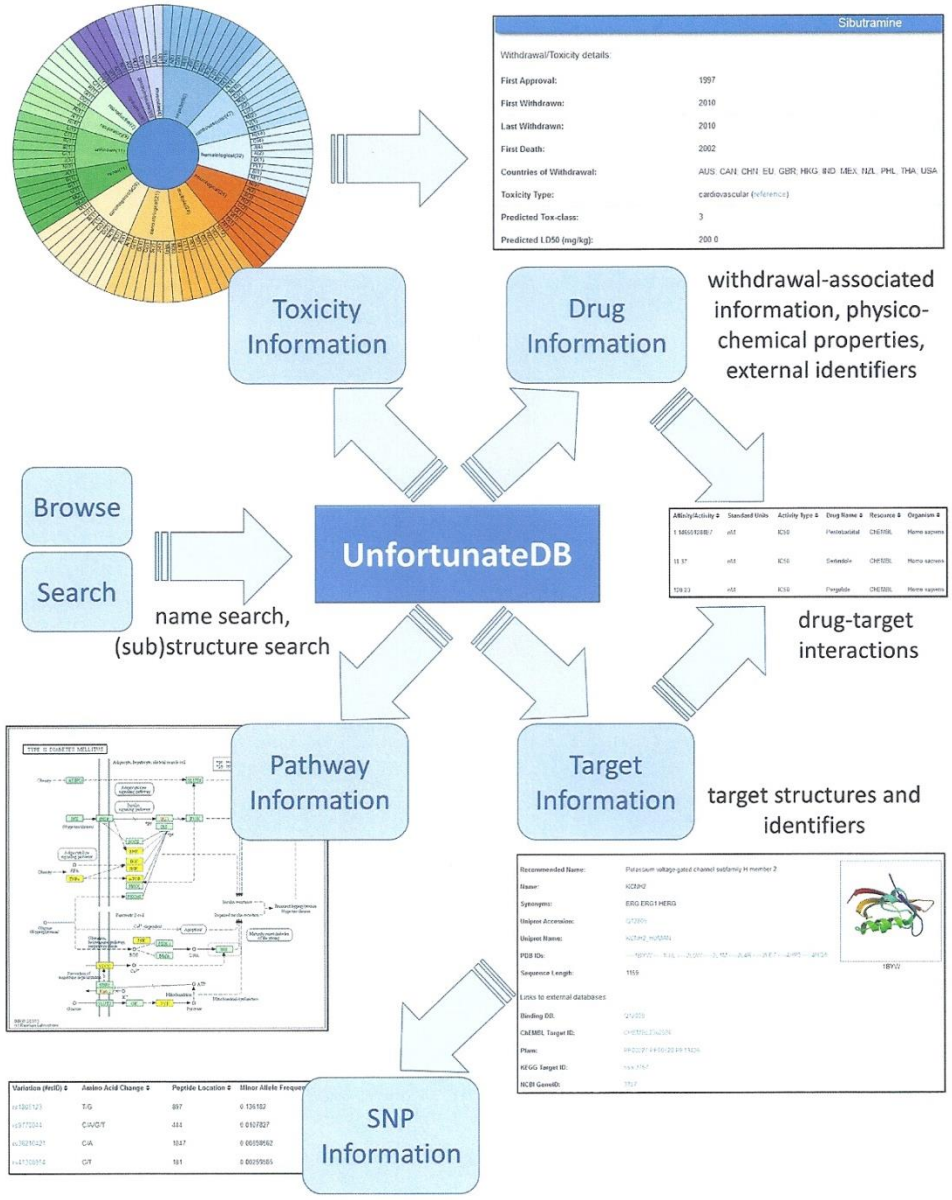

Figure 2. Schematic representation of WITHDRAWN: various search options and different entity types: drugs, targets, pathways, toxicity types and SNPs.

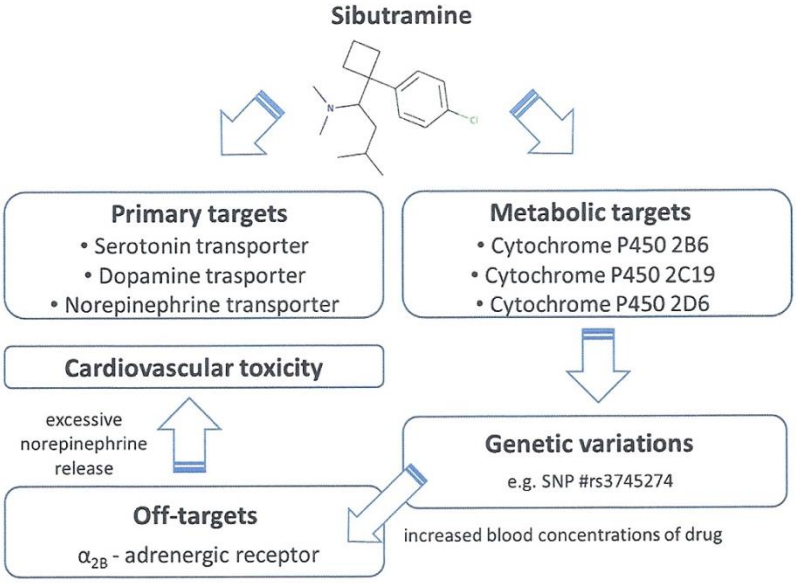

Figure 3. Case study—use of WITHDRAWN in connecting links between drugs, targets and SNPs in toxicological context.

including myocardial infarctions and stroke (40). Sibutramine is a non-selective inhibitor that acts by inhibiting the reuptake of the three monoamine neurotransmitters: serotonin, dopamine and norepinephrine. By searching for sibutramine targets in WITHDRAWN, the drug record shows additional drug-target interactions including the cytochromes CYP2B6, CYP2C19 and CYP2D6 as well as the α_{2B}-adrenergic receptor (ADRA2B) where sibutramine exhibits similar activity as at the primary targets. WITHDRAWN shows four genetic variants for CYP2B6 with a MAF above 1% (rs3745274, rs3211371, rs8192709 and rs28399499). Indeed, it has been shown that CYP2B6 variations, particularly rs3745274, may lead to a significant increase in the blood concentration of sibutramine and its active metabolites (41,42). As summarized by Zhang *et al.* (43), the increased drug concentration could result in an increased off-target activity at ADRA2B which, through an increased norepinephrine release, can lead to increased blood pressure and adverse cardiovascular events. The example emphasizes the importance of considering extensive drug-target and pharmacogenetics studies during drug development.

CONCLUSIONS

WITHDRAWN is a rich resource of withdrawn or discontinued drugs. Due to a relatively small number of drugs withdrawn per year (~10), we will update the database annually to ensure good coverage and high standard. The database not only contains information related to drug withdrawals and associated adverse drug reactions but also drug-target interactions and genetic variations of the protein targets. The drug-target interaction information is mapped to biological context by enriching the relevant pathways. The illustrated case study proves that, connecting links between drugs, targets and SNPs may explain the underlying mechanisms of toxicity. The knowledge presented in the database can improve the insights of drug-target interactions in toxicological context and provide the rationale for further off-target profiling and enhanced pharmacogenetics studies in different populations.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors kindly acknowledge Jevgeni Erehman, Mathias Dunkel and Herbert Schulz for their input and support. *Author Contributions*: Implementation of website: V.B.S.; curation of drug information: V.B.S., C.O., M.N.D., J.N.; curation of target information: J.N., V.B.S., M.N.D., C.O.; curation of pathway information: V.B.S.; curation of SNP information: B.O.G.; writing of manuscript: mainly V.B.S., input from M.N.D., R.P., J.N., B.O.G.; project coordination: R.P., M.N.D.

FUNDING

Berlin-Brandenburg research platform BB3R, Federal Ministry of Education and Research (BMBF), Germany

[031A262C] and the Immunotox project, Federal Ministry of Education and Research (BMBF), Germany [031A268B]. Funding for open access charge: Berlin-Brandenburg research platform BB3R, Federal Ministry of Education and Research (BMBF), Germany [031A262C].
Conflict of interest statement. None declared.

REFERENCES

1. DiMasi,J.A., Hansen,R.W. and Grabowski,H.G. (2003) The price of innovation: new estimates of drug development costs. *J. Health Econ.*, **22**, 151–185.
2. Lasser,K.E., Allen,P.D., Woolhandler,S.J., Himmelstein,D.U., Wolfe,S.M. and Bor,D.H. (2002) Timing of new black box warnings and withdrawals for prescription medications. *JAMA*, **287**, 2215–2220.
3. Podrid,P.J. (1984) Can antiarrhythmic drugs cause arrhythmia? *J. Clin. Pharmacol.*, **24**, 313–319.
4. Heijman,J., Voigt,N., Carlsson,L.G. and Dobrev,D. (2014) Cardiac safety assays. *Curr. Opin. Pharmacol.*, **15**, 16–21.
5. Roy,M., Dumaine,R. and Brown,A.M. (1996) HERG, a primary human ventricular target of the nonsedating antihistamine terfenadine. *Circulation*, **94**, 817–823.
6. Giacomini,K.M., Krauss,R.M., Roden,D.M., Eichelbaum,M., Hayden,M.R. and Nakamura,Y. (2007) When good drugs go bad. *Nature*, **446**, 975–977.
7. Arrowsmith,J. (2011) Trial watch: phase III and submission failures: 2007–2010. *Nat. Rev. Drug Discov.*, **10**, 87.
8. Arrowsmith,J. and Miller,P. (2013) Trial watch: phase II and phase III attrition rates 2011–2012. *Nat. Rev. Drug Discov.*, **12**, 569.
9. Thompson,R.A., Isin,E.M., Li,Y., Weaver,R., Weidolf,L., Wilson,I., Claesson,A., Page,K., Dolgos,H. and Kenna,J.G. (2011) Risk assessment and mitigation strategies for reactive metabolites in drug discovery and development. *Chem. Biol. Interact.*, **192**, 65–71.
10. Patel,H., Bell,D., Molokhia,M., Srishanmuganathan,J., Patel,M., Car,J. and Majeed,A. (2007) Trends in hospital admissions for adverse drug reactions in England: analysis of national hospital episode statistics 1998–2005. *BMC Clin. Pharmacol.*, **7**, 9.
11. Edwards,I.R. and Aronson,J.K. (2000) Adverse drug reactions: definitions, diagnosis, and management. *Lancet*, **356**, 1255–1259.
12. Hussaini,S.H. and Farrington,E.A. (2007) Idiosyncratic drug-induced liver injury: an overview. *Expert Opin. Drug Saf.*, **6**, 673–684.
13. Ulrich,R.G. (2007) Idiosyncratic toxicity: a convergence of risk factors. *Annu. Rev. Med.*, **58**, 17–34.
14. Lucena,M.I., Andrade,R.J., Kaplowitz,N., Garcia-Cortes,M., Fernandez,M.C., Romero-Gomez,M., Bruguera,M., Hallal,H., Robles-Diaz,M., Rodriguez-Gonzalez,J.F. *et al.* (2009) Phenotypic characterization of idiosyncratic drug-induced liver injury: the influence of age and sex. *Hepatology*, **49**, 2001–2009.
15. Greer,M.L., Barber,J., Eakins,J. and Kenna,J.G. (2010) Cell based approaches for evaluation of drug-induced liver injury. *Toxicology*, **268**, 125–131.
16. Waller,P. (2004) *Stephens' Detection of New Adverse Drug Reactions*. 5th Edition ed. John Wiley & Sons, Ltd, Chichester, West Sussex, England.
17. Onakpoya,I.J., Heneghan,C.J. and Aronson,J.K. (2015) Delays in the post-marketing withdrawal of drugs to which deaths have been attributed: a systematic investigation and analysis. *BMC Med.*, **13**, 26.
18. Jefferys,D.B., Leakey,D., Lewis,J.A., Payne,S. and Rawlins,M.D. (1998) New active substances authorized in the United Kingdom between 1972 and 1994. *Br. J. Clin. Pharmacol.*, **45**, 151–156.
19. Wysowski,D.K. and Swartz,L. (2005) Adverse drug event surveillance and drug withdrawals in the United States, 1969–2002: the importance of reporting suspected reactions. *Arch. Intern. Med.*, **165**, 1363–1369.
20. 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.
21. Pihan,E., Colliandre,L., Guichou,J.F. and Douguet,D. (2012) e-Drug3D: 3D structure collections dedicated to drug repurposing and fragment-based drug design. *Bioinformatics*, **28**, 1540–1541.
22. Bento,A.P., Gaulton,A., Hersey,A., Bellis,L.J., Chambers,J., Davies,M., Kruger,F.A., Light,Y., Mak,L., McGlinchey,S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–1090.
23. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J., Zhou,Z., Han,L., Karapetyan,K., Dracheva,S., Shoemaker,B.A. *et al.* (2012) PubChem's BioAssay Database. *Nucleic Acids Res.*, **40**, D400–412.
24. Drwal,M.N., Banerjee,P., Dunkel,M., Wettig,M.R. and Preissner,R. (2014) ProTox: a web server for the in silico prediction of rodent oral toxicity. *Nucleic Acids Res.*, **42**, W53–58.
25. Davis,A.P., Grondin,C.J., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., King,B.L., Wiegers,T.C. and Mattingly,C.J. (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, **43**, D914–920.
26. Hu,Y. and Bajorath,J. (2014) Influence of search parameters and criteria on compound selection, promiscuity, and pan assay interference characteristics. *J. Chem. Inf. Model.*, **54**, 3056–3066.
27. Rose,P.W., Prlic,A., Bi,C., Bluhm,W.F., Christie,C.H., Dutta,S., Green,R.K., Goodsell,D.S., Westbrook,J.D., Woo,J. *et al.* (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, **43**, D345–D356.
28. Qin,C., Zhang,C., Zhu,F., Xu,F., Chen,S.Y., Zhang,P., Li,Y.H., Yang,S.Y., Wei,Y.Q., Tao,L. *et al.* (2014) Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res.*, **42**, D1118–D1123.
29. Lounkine,E., Keiser,M.J., Whitebread,S., Mikhailov,D., Hamon,J., Jenkins,J.L., Lavan,P., Weber,E., Doak,A.K., Cote,S. *et al.* (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, **486**, 361–367.
30. Kamburov,A., Stelzl,U., Lehrach,H. and Herwig,R. (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D793–D800.
31. Smigielski,E.M., Sirotkin,K., Ward,M. and Sherry,S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.*, **28**, 352–355.
32. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–598.
33. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–669.
34. Gray,K.A., Yates,B., Seal,R.L., Wright,M.W. and Bruford,E.A. (2015) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.*, **43**, D1079–D1085.
35. Ullmann,J.R. (1976) Algorithm for Subgraph Isomorphism. *J. ACM*, **23**, 31–42.
36. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*, bar009.
37. Adzhubei,I., Jordan,D.M. and Sunyaev,S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.*, Chapter 7, Unit7 20.
38. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
39. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
40. James,W.P., Caterson,I.D., Coutinho,W., Finer,N., Van Gaal,L.F., Maggioni,A.P., Torp-Pedersen,C., Sharma,A.M., Shepherd,G.M., Rode,R.A. *et al.* (2010) Effect of sibutramine on cardiovascular outcomes in overweight and obese subjects. *N. Engl. J. Med.*, **363**, 905–917.
41. Bae,S.K., Cao,S., Seo,K.A., Kim,H., Kim,M.J., Shon,J.H., Liu,K.H., Zhou,H.H. and Shin,J.G. (2008) Cytochrome P450 2B6 catalyzes the formation of pharmacologically active sibutramine (N-{1-[1-(4-chlorophenyl)cyclobutyl]-3-methylbutyl}-N,N-dimethylamine) metabolites in human liver microsomes. *Drug Metab. Dispos.*, **36**, 1679–1688.
42. Chung,J.Y., Jang,S.B., Lee,Y.J., Park,M.S. and Park,K. (2011) Effect of CYP2B6 genotype on the pharmacokinetics of sibutramine and active metabolites in healthy subjects. *J. Clin. Pharmacol.*, **51**, 53–59.
43. Zhang,W., Roederer,M.W., Chen,W.Q., Fan,L. and Zhou,H.H. (2012) Pharmacogenetics of drugs withdrawn from the market. *Pharmacogenomics*, **13**, 223–231.

12 Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

13 Komplette Publikationsliste

Originalarbeiten (Peer-reviewed);
Kumulativer Impact Factor (2016): 45,56

Marie Chantal Lemfack*, **Janette Nickel***, Mathias Dunkel, Robert Preissner, Birgit Piechulla (* geteilte Erstautorenschaft)
mVOC: a database of microbial volatiles
Nucleic Acids Research, 2014; 42: D744 – D748
Impact Factor: 9,112 (5-jähriger Impact Factor: 8,867)

Michael F. Hoffmann, Sarah C. Preissner, **Janette Nickel**, Mathias Dunkel, Robert Preissner, Saskia Preissner
The Transformer database: biotransformation of xenobiotics
Nucleic Acids Research, 2014; 42: D1113 – D1117
Impact Factor: 9,112 (5-jähriger Impact Factor: 8,867)

Janette Nickel*, Björn-Oliver Gohlke*, Jevgeni Erehman, Priyanka Banerjee, Wen Wei Rong, Andrean Goede, Mathias Dunkel, Robert Preissner (* geteilte Erstautorenschaft)
SuperPred: update on drug classification and target prediction
Nucleic Acids Research, 2014; 42: W26 – W31
Impact Factor: 9,112 (5-jähriger Impact Factor: 8,867)

Vishal B. Siramshetty, **Janette Nickel**, Christian Omieczynski, Björn-Oliver Gohlke, Malgorzata N. Drwal, Robert Preissner
WITHDRAWN – a resource for withdrawn and discontinued drugs
Nucleic Acids Research, 2016; 44: D1080 – D1086
Impact Factor: 9,112 (5-jähriger Impact Factor: 8,867)

Björn-Oliver Gohlke, **Janette Nickel**, Raik Otto, Mathias Dunkel, Robert Preissner
CancerResource – updated database of cancer-relevant proteins, mutations and interacting drugs
Nucleic Acids Research, 2016; 44: D932 – D937
Impact Factor: 9,112 (5-jähriger Impact Factor: 8,867)

14 Danksagung

An dieser Stelle möchte ich die Gelegenheit nutzen meinen Dank an besondere Menschen zu richten, die mich während der letzten Jahre unterstützt haben.

Ich danke meinem Supervisor PD Dr. Robert Preißner für die Möglichkeit diese Dissertation in seiner Arbeitsgruppe, der Strukturellen Bioinformatik, anzufertigen und für die vielen Erfahrungen, die ich aus dieser Zeit mitnehmen kann. Ich bedanke mich auch bei PD Dr. Saskia Preißner für die Zweitbetreuung meiner Arbeit.

Einen lieben Dank an alle meine Kollegen und Kolleginnen, die mit mir an interessanten Projekten gearbeitet haben und für die vielen netten Gespräche und Diskussionen: Björn-Oliver Gohlke, Dr. Mathias Dunkel, Dr. Andrean Goede, Sophie Dabbert, Jevgeni Erehman, Priyanka Banerjee, Vishal Siramshetty, Dr. Malgorzata Drwal, Wen Wei Rong, Dr. Franziska Niehr, Christoph Hapke, Dr. Thomas Heiden und Prof. Dr. Dr. Karlheinz Seeger.

Ich danke außerdem der Berlin-Brandenburger Forschungsplattform BB3R und dem Deutschen Krebsforschungszentrum, die es mir ermöglicht haben an Konferenzen und unterschiedlichen Lehrveranstaltungen teilzunehmen.

Ein großer Dank gilt meinen Freunden und meiner Familie, die mich in jeglicher Weise unterstützt, ermutigt und immer an mich geglaubt haben. Vielen lieben Dank an meine Eltern, meine Schwester und meinen Schwager, meinen Schwiegereltern und ganz besonders an meinen Ehemann Sven.