Deciphering the clinical effect of drugs through large-scale data integration

Kjærulff, Sonny Kim

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Kjærulff, S. K. (2013). *Deciphering the clinical effect of drugs through large-scale data integration*. Technical University of Denmark.

Deciphering the clinical effect of drugs through large-scale data integration

Sonny Kim Kjærulff

January 2013

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS **CBS**

Preface

This thesis was prepared at the Center for Biological Sequence Analysis, Department of Systems Biology at the Technical University of Denmark. The work in this thesis was carried out under the supervision of Associate Professor Irene Kouskoumvekaki, Associate Professor Olivier Taboureau and Professor Søren Brunak. This work was funded by Lundbeck foundation and DTU's PhD scholarships.

Kongens Lyngby
January 2013

Sonny Kim Kjærulff

Contents

Abstract

This thesis presents the work carried out at the Center for Biological Sequence Analysis, Technical University of Denmark. The thesis includes four articles describing large-scale data integration and methods for the prediction of drug side-effects.

Chapter 2 presents ChemProt, a novel disease chemical biology database. ChemProt integrates different chemical–protein annotation resources for disease-associated proteins and protein–protein interaction data. ChemProt is developed to assist in silico evaluation of environmental chemicals, natural products and approved drugs, as well as to aid the selection of new compounds based on their activity profiles against biological targets. The latest update of ChemProt database includes a new visual interface, which enables easy navigation through the pharmacological space. Additionally, new search methods for chemical, protein, disease and side-effect data have been implemented.

Chapter 3 presents two articles that showcase the application of systems chemical biology approaches to understand and model drug side-effect data. The first approach applies machine learning methods to cluster side-effects, drugs, proteins and clinical outcomes in networks. This work demonstrates the power of a strategy that uses clinical data mining in association with chemical biology in order to reduce the search space and aid identification of novel drug actions. The second article described in chapter 3 outlines a high confidence side-effect–drug interaction dataset.

We estimated based on the placebo-controlled studies from DailyMed that only approximately 20% of the drug–side-effect associations are significant. With the ChemProt database we linked drugs with their biological targets and applied a scoring function in order to capture frequently encountered side-effect–protein associations. We then built a computational chemical biology model, which revealed side-effect predictive capabilities for 55% of the 133 drugs in the SIDER database. Further validation was performed on withdrawn drugs stored in DrugBank and many side-effects were confirmed through literature search. This work demonstrates the importance of using high-confidence drug–side-effect data in deciphering the effect of small molecules in humans.

In summary, this thesis presents computational systems chemical biology approaches that can help identify clinical effects of small molecules through large-scale data integration. These approaches also serve to pave the way into a variety of chemogenomics, polypharmacology and systems chemical biology studies.

Dansk resumé

Denne afhandling præsenterer arbejde udført på Center for Biologisk Sekvens Analyse, Danmarks Tekniske Universitet. Afhandlingen indeholder fire artikler, der beskriver integration af store data mængder samt metoder til forudsigelse af bivirkninger.

Kapitel 2 præsenterer ChemProt, en ny sygdom-kemikalie-biologi database. ChemProt integrerer forskellige kemikalie-protein annotations resourcer for sygdoms associerede proteiner og protein-protein interaktions data. ChemProt er udviklet til at assistere in silico evaluering af miljø kemikalier, natur produkter og godkendte lægemidler, samt understøtte udvælgelsen af nye kemiske stoffer baseret på deres aktivitets profiler imod biologiske targets. Den seneste opdatering af ChemProt databasen inkluderer en ny visuel interface, som muliggør let navigering i det farmakologiske rum. Ydermere, er der implementeret nye søgemetoder for kemikalie, protein, sygdom og bivirknings data.

Kapitel 3 præsenterer to artikler, der fremviser anvendelsen af system kemisk- biologiske tilgange til at forstå og modelere bivirknings data. Den første tilgang anvender automatiske lærings metoder til at gruppere bivirkninger, kemiske stoffer, proteiner og kliniske resultater i netværk. Dette arbejde demonstrerer styrken i en strategi der anvender klinisk data mining i samarbejde med kemikalie biologi for at reducere søge mulighederne og støtte identificeringen af nye virkemidler. Den anden artikel beskrevet i kapitel 3 skitserer et høj confidens bivirkning-medikament interaktions datasæt.

På baggrund af placebo kontrollerede studier fra DailyMed estimerede vi at kun ca. 20% af lægemiddel bivirkningerne er signifikante. Ved hjælp af ChemProt databasen parrede vi kemiske stoffer med deres biologiske targets og påførte en scoring funktion for at fange de hyppigt forekommende bivirkning-protein associationer. Derefter lavede vi en kemisk biologisk model, som kunne forudsige bivirkninger hos 55% af de 133 lægemidler I SIDER databasen. Yderligere validering blev udført på tilbagetrukne lægemidler fra DrugBank og mange af deres bivirkninger blev bekræftet gennem litteratur søgning. Dette arbejde demonstrerer vigtigheden af at anvende høj confidens lægemiddel- bivirknings data til at decifrere små molekylers effekt på mennesker.

Opsummeret, denne afhandling præsenterer system kemisk biologiske tilgange som kan støtte identificeringen af små molekylers kliniske effekter via integration af store data mængder. Disse tilgange er også med til at bane vejen for en række studier indenfor kemogenomik, polyfarmakologi og system kemisk biologi.

Acknowledgements

I would like to express my gratitude to all who have directly and indirectly contributed to this work. I would especially like to thank:

- My supervisors Irene Kouskoumvekaki, Olivier Taboureau and Søren Brunak for their advices, trust and inspirations. I am very thankful for your supervision.
- Tudor Oprea for his broad knowledge within system chemical biology and for his insightful discussions.
- All members of the Computational Chemical Biology group, especially Louis and Jens for their teamwork during ChemProt 2.
- All members of the Integrative Systems Biology.
- Laura Furlong for lending me a desk at IMIM in Barcelona.
- My colleagues at CBS who I have been blessed to get to know, learn from and discuss with, especially my lunch buddies Marcin, Salvatore, Thomas, Bent, Leon, Edita, Ali, Greg and Kasper.
- The CBS administration for all the help with formalities and the CBS system administration for all their technical support, especially Peter for extending my Perl knowledge.
- My friends, family, my wife Louise and my daughter Nanna. Thank you so much!

Papers included in the thesis

- Olivier Taboureau[†], **Sonny Kim Nielsen**[†], Karine Audouze, Nils Weinhold, Daniel Edsgärd, Francisco S. Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, Søren Brunak and Tudor I. Oprea
  ChemProt: a disease chemical biology database *Nucleic Acids Research, 2011, Vol. 39, Database issue D367–D372 doi:10.1093/nar/gkq906*

- Tudor I. Oprea[†], **Sonny Kim Nielsen**[†], Oleg Ursu[†], Jeremy J. Yang, Olivier Taboureau, Stephen L. Mathias, Irene Kouskoumvekaki, Larry A. Sklar, and Cristian G. Bologa
  Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing. *Molecular Informatics, 30(2-3), 100-111. . doi:10.1002/minf.201100023 2011*

- **Sonny Kim Kjærulff**[†], Louis Wich[†], Jens Kringelum[†], Ulrik P. Jacobsen, Irene Kouskoumvekaki, Karine Audouze, Ole Lund, Søren Brunak, Tudor I. Oprea and Olivier Taboureau.
  ChemProt-2.0: Visual navigation in a disease chemical biology database *Nucleic Acids Research, 2012, 1-6 doi:10.1093/nar/gks1166*

- **Sonny Kim Kjærulff**, Tudor Oprea, Olivier Taboureau and Irene Kouskoumvekaki.
  A system chemical biology approach for the prediction of adverse drug reactions *Manuscript in preperation*

[†] These authors contributed equally.
Changed last name from Nielsen to Kjærulff during this thesis.

Part I

Introduction

Chapter 1

Introduction

1.1 The paradigm in drug discovery

The one-drug-one-target reductionistic approach has been the paradigm in drug discovery for almost a century. However, this approach has reached its limits and is increasingly considered inadequate in the drug industry, especially in cancer drug research [1]. Today, the development of new drugs takes 10–12 years and costs a pharmaceutical company an average of $850 million dollars. The investment in research and development (R&D) has been increasing since the millennium, while the number of new molecular entities (NMEs) approved by the USA Food and Drug Administration (FDA) has remained almost constant at 25 compounds a year (Figure 1.1) [2].

The high cost of drug development is mostly associated with the failure of a large number of promising drug candidates during the final phases of the clinical trails, which is often due to unwanted effects and serious side-effects. A paradigm shift towards a more systems-level approach is therefore warranted in the pharmaceutical industry.

1.2 From systems biology to systems chemical biology

Since the year 2000, the number of articles within systems biology has grown exponentially [3]. Work done within systems biology has led to great advances in our understanding of the biology of complex diseases. It has pushed us away from the 'reductionist approach' where effects of single genes, proteins or phenotypes are studied, and has moved us towards a more

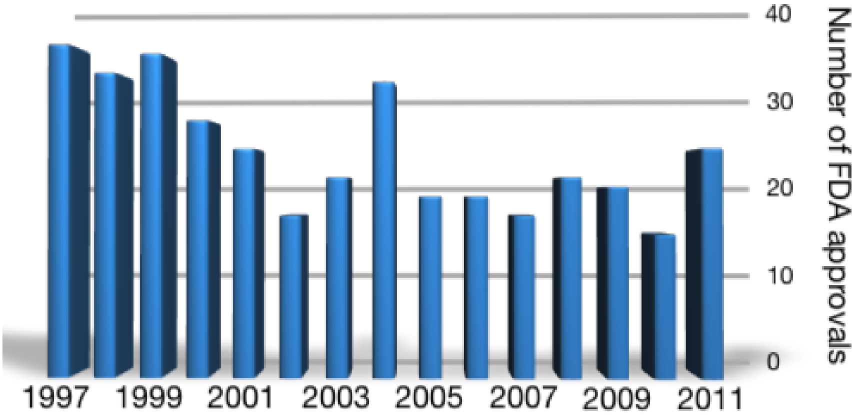


Figure 1.1. Statistic from 1997 to 2011 showing the total number of new molecular entities (NMEs) for each year. From [2]

'holistic approach' where the effects of multiple genes, proteins or phenotypes are studied simultaneously. Even though thousands of genes for human diseases have been discovered using the 'reductionist approach', this approach fails to target complex diseases such as cancer, hypertension, diabetes and mental disorders. Such diseases, due to their inherent multifactorial nature, therefore need to be studied in a systems manner. This involves deciphering and modeling entire networks at, for example, genomic, metabolic or cell signaling networks levels and analyzing them in their totality at different levels of complexities—from molecules to organism [4, 5].

Although the term 'systems biology' has been widely used in life sciences, there is no clear definition of the methods and concepts belonging to it. Instead, it can be regarded as a mixture of interdisciplinary tools/techniques and computational modeling to omics (such as genomics, proteomics, metabolomics) data [3].

Within the last few years, chemical biology has led to the generation of large amounts of data on genes, proteins and their modulation by small molecules. Combining these data types and systems biology approaches allows investigation of the effect of small molecules on the biological system [6]. A new field has thus emerged as a consequence of integrating chemical biology with systems biology, and is described in the literature as systems chemical biology [7], systems pharmacology [8], systems medicine [9], or systems polypharmacology [10]. Although the names are different, they all focus on a systematic understanding of the impact of small molecules on the biological system [2]. Chemoinformatics, a field which initially was introduced in 1995 [11] and now is well integrated in drug discovery [2], is used to retrieve and analyze small molecules from databases.

Combining chemoinformatics tools with biological databases and systems biology methods provides a possibility to understand the complex relationship between chemicals and their effects on living systems.

1.3 Representation and retrieval of chemical compounds

Chemical compounds can be represented in multiple ways, for example, either by trivial names (e.g., Aspirin) or by chemical names (e.g., 2-acetoxybenzoic acid). When using chemical names, chemical nomenclature rules are used to generate systematic names, normally by following the standards defined by the International Union of Pure and Applied Chemistry (IUPAC).

2D graphical representation is the most common mode for chemists to represent a chemical compound. Here, individual atoms are depicted by atomic symbols and the bonding electrons depicted as lines. The 2D graphical representation is a very simplified representation of the molecule; it only explains the topology of the molecule, i.e., which atoms are connected and through which bond types. For more complex representation of the molecules, 3D arrangement (topography) can be used; this requires additional information about the position, angle and distance between each atom. To make it even more complex, electrostatic potential could be added to 3D chemical representations. This type of description is important to depict the region of the chemical significant for its reactivity (Figure 1.2). Such modes of representing chemicals, however, are not computationally suitable for searching through databases and comparing millions of compounds. Instead, standards like SMILES, InChI and Fingerprints, just to mention some of the most common ones, are used to describe the chemical structure information in a highly compressed and simplified notation. SMILES (Simplified Molecular Input Line Entry Specification) was created by David Weininger in 1986 [12] and is now used for comprehensive chemical nomenclature.

The SMILES nomenclature follows 6 basic rules:

1. Atoms are shown by their atomic symbols (C for Carbon, O for Oxygen, etc).
2. Hydrogens are omitted.

3. Neighboring atoms are next to each other.

4. Double and triple bonds are shown by "=" and "#", respectively.

5. Branches are shown by parentheses.

6. Cyclic structures are described by allocating digits to the two "connecting" ring atoms.

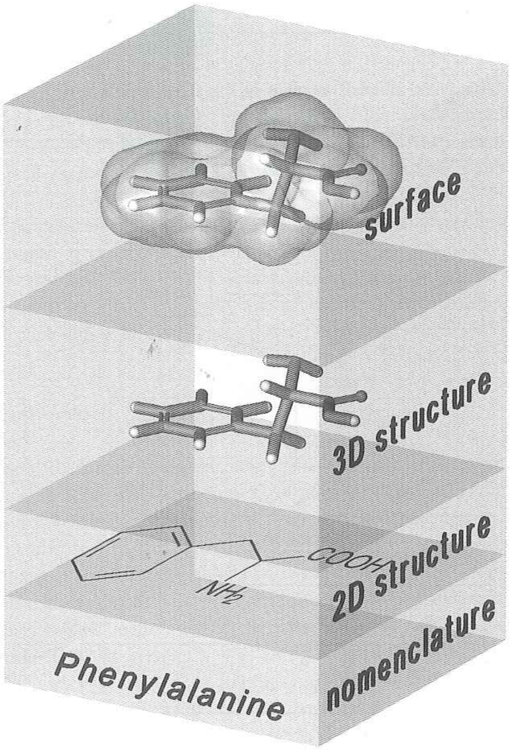


Figure 1.2. Different ways of representing a chemical with different contents of structural information. From Chemoinformatics [13]

Many different SMILES strings can be written for the same structure. This is a major disadvantage when comparing chemicals. For example, Figure 1.3 shows the 2D structure of Aspirin and two SMILES for the same structure. Furthermore, a special extension of SMILES, called USMILES or "Unique SMILES" created by Daylight [14] uses an algorithm independent of the internal atomic numbering and always ensures the same canonical[1]

[1] Canonicalization is to standardize and make rules for numbering of atoms in the chemical to ensure uniqueness. Otherwise, the numbering of the same molecule, in principle, could be in n! different combinations

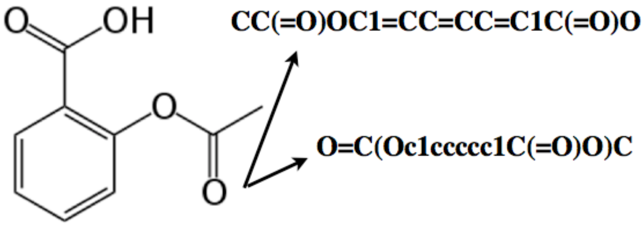


Figure 1.3. Two different ways to write a SMILE string for the 2D chemical structure of Aspirin.

representation of the structure. However, not all chemical databases that include SMILES use the same algorithm, and this can lead to a duplication of the chemical when integrating different databases. The uniqueness of the chemicals can be ensured if their 3D structures are available. 3D coordinates are included in SDF files, but considerable computing power and time would be required to compare thousands of compounds based on their SDF[2] data. Therefore, when dealing with large database integration, another simple representation of the chemicals is needed to make the comparison feasible in a reasonable amount of time.

The IUPAC organization, with an aim to standardize the nomenclature in chemistry, has provided a standard way of encoding molecular information in a textual identifier called InChI (International Chemical Identifier), which is now maintained by the non-profit organization InChI Trust [15]. InChI contains more information about the chemical compared to SMILES and it ensures that every chemical structure is given a unique InChI string via a three-step process: normalization (remove redundant information), canonicalization and serialization (generating the string of characters). The InChI can be compressed down to a fixed length of 25 characters referred to as the InChIKey.

The InChI and InChIKey both contain different blocks of information about the chemical. The first block (14 letters) in InChIKey contains the connectivity information for the chemical. The second block contains the stereochemistry and isotopes information (8 letters), followed by the flag

[2] SDF stands for Structure-Data File and contains each structure in Molfile format with information about the atoms, bonds, connectivity and coordinates.

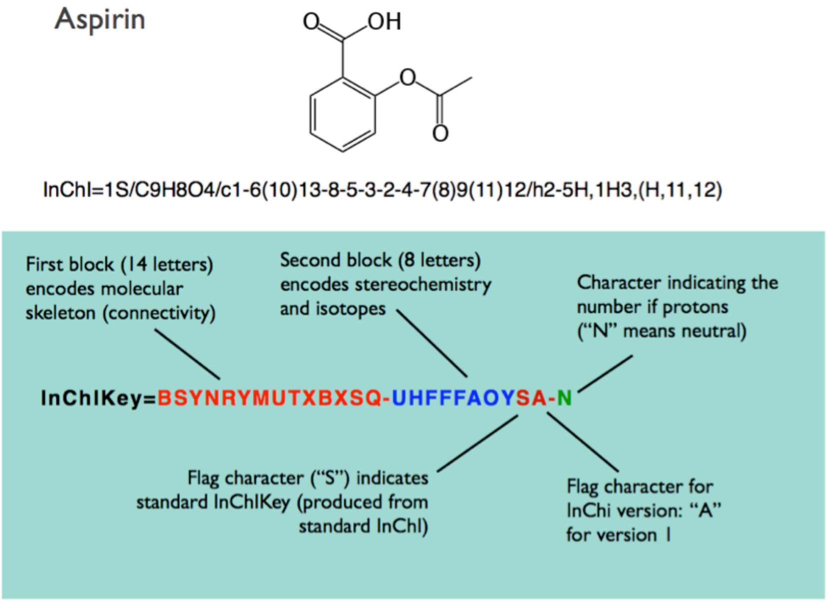

Aspirin
O
OH
O
O
InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)
First block (14 letters) encodes molecular skeleton (connectivity)
Second block (8 letters) encodes stereochemistry and isotopes
Character indicating the number if protons ("N" means neutral)
InChIKey=BSYNRYMUTXBXSQ-UHFFFAOYSA-N
Flag character ("S") indicates standard InChIKey (produced from standard InChI)
Flag character for InChi version: "A" for version 1

Figure 1.4. InChi and InChiKey representation of Aspirin. Figure from IUPAC [16]

character for the InChI type and version number. The last character indicates the number of protons, where "N" stands for neutral (Figure 1.4).

InChIKey is, to date, the most optimal way to search through a large number of chemicals very quickly while ensuring that the compounds are unique. This is a necessary feature when dealing with large chemical database integration.

Evaluation of chemical similarity

Virtual screening of large databases requires a different representation of the chemicals so as to enable fast comparison between chemicals. Fingerprints is an ambiguous way of representing chemical structures and is a widely used method for efficient search of similar compounds. Fingerprints uses a fragment library of chemical structures and features, where each fragment can be presented as a binary string of "1" and ”0", indicating presence or absence of the fragment in the chemical structure, respectively (Figure 1.5).

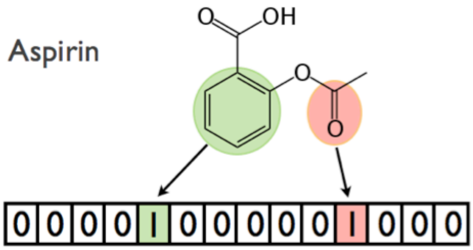


Figure 1.5. Fingerprint representation of Aspirin. Benzene ring and C=O group present in the fragment library indicated with number 1 in the fingerprint string.

The binary string can vary depending on the type of fingerprint and has a typical length between 150–2500 bits. The MACCS fingerprint is a popular 2D structure based fingerprint with a length of 166-bits developed by MDL Information System [17]. As it only has 166 features in the fragment library, it describes the chemicals in a general way. This enables it to capture many similar compounds using a high similarity score. The more features there are in the fingerprint the more specific the search will be, thus depending on the similarity score.

Several similarity measures are available for comparing fingerprints. The Tanimoto coefficient (Tc) is a commonly used measure. The Tc for two binary fingerprints, A and B, is calculated as follows:

$$Tc(A,B) = \frac{N_{AB}}{N_A + N_B - N_{AB}} \tag{1.1}$$

where, N_{AB} is the number of bits that both fingerprints have in common, and N_A and N_B correspond to the number of bits set in A and B, respectively. A Tanimoto coefficient of 1 indicates that the two compounds have identical fingerprints. This suggests that the compounds are identical or very similar. A Tanimoto coefficient of 0 corresponds to non-overlapping fingerprints i.e., very dissimilar compounds.

Based on the principle that similar compounds have similar properties [18] and exhibit similar biological activity, identification of similar compounds is widely used in virtual screening.

The general consensus is that when using a Tanimoto coefficient larger than 0.85 (at least for MACCS fingerprint) as a threshold for considering two compounds as similar, they are very likely to show similar biological activities. This assumption has shown only to be validated in 30% of the cases studied by Martin et al. [19]. However, these results still have great importance for the concept of chemical network biology, where chemicals are linked together based on structure similarity to infer the same biological activity profile.

Similarity ensemble approach (SEA)

SEA is a new approach that relates proteins based on chemical similarity between their ligands. Hundreds of ligand sets can be derived by creating networks of similar compounds linked together based on a tanimoto threshold. The similarities among ligand sets may reveal the pharmacological relationships of the protein targets they modulate. To approximate the similarity between two ligand sets, the similarity scores of the ligand pairs across the sets are summed up. The similarity between two ligand sets returns a raw score which is corrected by the Tanimoto threshold determined by fitting the raw scores to an extreme value distribution. The raw scores are then converted to an E-value that describes the probability of the scores compared to random scores. The smaller the E-values, the more likely the compounds are active against the protein targets [20].

Alternative approaches also exist for comparing chemicals and, in particular, drugs. Drugs are classified according to the Anatomical Therapeutic Chemical (ATC) classification system, which is controlled by WHO Collaborating Centre for Drug Statistics Methodology (WHOCC) [21]. A single drug can have multiple ATC codes assigned to it. The ATC code consists of 5 different levels. The first level indicates the anatomical main group (organ or system); the second, the therapeutic main group; the third, the therapeutic-pharmacological subgroup; the fourth, the chemical/therapeutic/pharmacological subgroup; the final fifth level indicates the chemical substance.

Figure 1.6 shows a disease-disease network, based on drugs grouped together using second level ATC codes, combined with disease information. The disease information was obtained from PharmGKB [22] and the ATC codes from DrugBank [23] and WHOCC [24]. The disease-disease network shows a strong association of schizophrenia with “tobacco use disorder” (see magenta nodes in Figure 1.6). It has also been observed in a number of other studies that people with schizophrenia tend to have a significantly higher consumption of tobacco than the general population [25, 26]. One of these

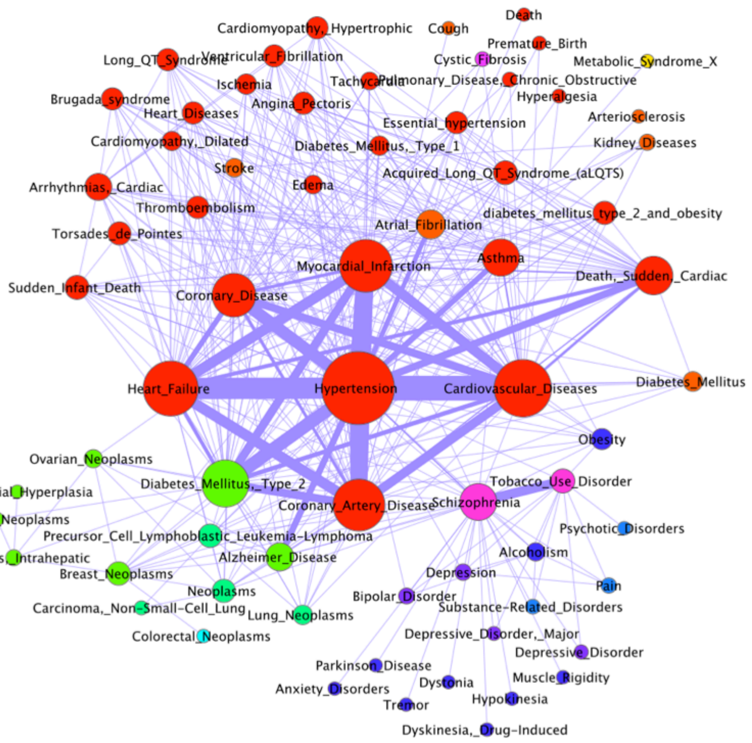


Figure 1.6. Kjaerulff et al. unpublished results presented at QSAR conference, Rhodes 2010. Disease-disease network based on ATC code grouping of drugs. Nodes colored according to ATC classes.

associations is due to nicotine, the addictive substance in tobacco, and has been reported to affect both schizophrenia and antipsychotic medication [27].

Instead of using the therapeutic indication of drugs, like the ATC code, other studies have successfully used the drug side-effects as a similarity method to group drugs [28, 29].

In order to understand the impact that the chemicals have on the biological system, representing and comparing chemicals, as described in this section, is the first step before systems biology methods can be applied.

1.4 Chemical-biological databases

The following section details different chemical-biological databases included in the ChemProt database described in chapter 2.

ChEMBL is a free, public domain chemical database of bioactive molecules from the European Bioinformatics Institute (EBI). It contains 2D structures, calculated properties (logP, molecular weight, Lipinski parameters), and bioactivities (e.g., binding constants, pharmacology, and ADMET data). The bioactivities data are manually curated to ensure normalized uniform sets of end-points and units. ChEMBL version 14 contains 9003 targets, > 1.3 million compounds, and > 10 million activities [30].

DrugBank is a free, publicly available database from University of Alberta. It combines information on drugs (chemical structure, pharmacokinetics data, pharmacological mode of action and pharmaceutical details) with information on drug-targets (target sequence and structure, pathway, splice variants, etc.). The database contains 6711 drug entries including 1447 FDA-approved small molecule drugs along with 4227 non-redundant proteins linked to the drug entries. Each drug entry (DrugCard) contains more than 150 fields of data which cover information on drug/chemical and drug-target/protein interaction [23].

The PDSP Ki database is a free, publicly available resource for psychoactive compounds and their functional activity on cloned CNS receptors, channels and transporters of human or rodent origin. The user interface provides customized data mining tools (Ki graphs, receptor and ligand selectivity mining), and is cross-linked with PubChem and PubMed. Other searchable fields include: receptor name, species name, tissue source, radiolabeled and tested ligands, bibliographic references as well as Ki (inhibition constant) value range [31].

WOMBAT and WOMBAT-PK are two commercial databases from Sunset Molecular Discovery which integrate knowledge from target-driven medicinal chemistry with clinical pharmacokinetics data. WOMBAT version 2012.1 contains > 330000 entries with 1966 unique targets. The information is curated from more than 15000 papers published in medicinal chemistry journals between 1975 and 2009. Additional experimental properties and calculated descriptors are available along with a comprehensive set of keywords related to biology and experimental protocols [32].

BindingDB is a free, public-domain database of measured binding affinities. It focuses chiefly on the interactions of proteins, considered to be drug-targets, with small drug-like molecules. BindingDB contains more than 900000 binding data, for 6263 protein targets and > 370000 small molecules. The data are extracted from scientific literature and focus on drug-target or candidate drug-target proteins that have their 3D structures deposited in the Protein Data Bank [33].

PharmGKB is a free, public-domain pharmacogenomics database which contains clinical information including dosing guidelines and drug labels, potential clinically actionable gene–drug associations and genotype–phenotype relationships. The annotated genetic variants and gene–drug–disease relationships are extracted from a review of scientific literature. PharmGKBs interface tries to integrate clinical interpretations with gene, drug and disease information in a user-friendly layout [22].

PubChem is a free, public-domain database of compounds and their biological activities. The database is maintained by National Center for Biotechnology Information (NCBI). The database includes records for 85 million substances containing 30 million unique chemical structures, and 2.1 million of these substances have bioactivity data in at least one of the 504000 PubChem BioAssays. The database is searchable using text queries as well as structural queries based on chemical SMILES, formulas or chemical structures provided in a variety of formats [34].

The International Union of Basic and Clinical Pharmacology (IUPHAR) database, IUPHAR-DB, is a free, public-domain database providing detailed and expert-driven annotation of human and rodent receptors and other drug targets, along with the substances that act on them. The database includes information on 646 genes from four major protein classes (G protein-coupled receptors, nuclear hormone receptors, voltage- and ligand-gated ion channels) and around 3180 bioactive chemicals that interact with them [35].

The Comparative Toxicogenomics Database (CTD) is a free, public-domain resource that provides information on the interaction of environmental chemicals with gene products, and their effects on human health. The chemical–gene, chemical–disease and gene–disease relationships are manually extracted from scientific literature. The database contains 1.4 million chemical–gene–disease data points. The web interface includes features like GeneComps and ChemComps which find comparable genes and chemicals that share toxicogenomic profiles [36].

STITCH (search tool for interactions of chemicals) is a free, public-domain resource that allows exploration of known and predicted interactions between chemicals and proteins. Chemicals are linked to other chemicals and proteins through evidence derived from pathway and experimental databases and from the literature. The STITCH database contains interactions for more than 300000 chemicals and 2.6 million proteins from 1133 organisms. A confidence score is assigned to the interaction to reflect the level of significance [37].

1.5 Side effects

Side-effects are unintended effects of a drug that are secondary to its therapeutic effect. Normally, side-effects are undesired effects, but they can sometimes also be beneficial. These beneficial effects can often be exploited for drug repurposing. One of the most famous examples of drug repurposing is Pfizer's sildenafil (Viagra) which originally was under development for the treatment of hypertension. A side-effect of the drug was observed in the form of enhancement of penile erections and was found to be caused by the inhibition of cGMP specific phosphodiesterase type 5 (PDE5). This side-effect then became the focus of the study for the treatment of erectile dysfunction [38].

Side-effects may occur either due to the involvement of a single drug target in multiple cellular functions or due to the effect of the drug on multiple drug targets. A study by Brouwers et al. [39] shows that 64% of side-effect similarities were related to overlapping drug targets, while 5.8 % of side-effect similarities were due to protein targets that were closer together in the human protein-protein networks. Protein networks are "small world networks", which means that any two random proteins would be connected via paths of only a few protein interactions. This is important for a fast communication between proteins and also for making the network more resilient (i.e., robust) to changes in the external conditions [40]. With an exception of anti-cancer drugs, it is crucial during drug development to avoid targeting proteins with several interactions (also called protein-hubs) to avoid unwanted side effects [41]. A study by Wang et al. [42] shows that too short a distance between drug-targets and disease-genes in human signaling networks can cause significantly more side-effects and lead to more drugs being withdrawn.

The detection of side-effects by experimental methods requires screening of a large number of potential off-targets, which is both time consuming and costly. However, side-effects can be predicted by network-based methods. A number of side-effect networks and drug-target–-side-effect networks have been constructed in the last few years. Campillos et al. [28] constructed a side-effect similarity network of drugs and used this network to identify novel drug-targets for known drugs. Yang et al. [43] constructed a drug-target–side-effect network using 162 drugs, each causing at least one serious side-effect, and their binding strengths among 845 protein targets. The analysis showed a similar target profile for similar serious side-effects. They confirmed the protein MHC I as the possible target for the sulfonamide-induced side-effect of toxic epidermal necrolysis . Oprea et al. [44] used deep data mining of drug-target interactions combined with text mining of drug package inserts/online repositories to construct side-effect networks with clinical compound profiles. Mizutani et al. [45] used correlation analysis of drug-target binding profiles and side effect profiles to show that the calculated and correlated sets were significantly enriched with proteins that were involved in same biological pathways. In another study, Lounkine et al. [46] constructed a drug-target–side-effect network from chemical similarity based predictions of off-target and side-effects. The network was used to predict the activity of 656 marketed drugs on 73 unintended side-effect targets.

Part II

Publications

Chapter 2

Chemical biology data integration

Large-scale data integration is not a trivial task especially when dealing with different sources that have different standardization. The problem appears when dealing with SMILES, as described in section 1.3, where there are different ways of writing SMILES notation. To ensure uniqueness, we converted all available chemical structures into an InChIKey ID. When the chemical structures were not available, for example as an SDF file, we had to rely on other ways of converting the chemical information to our ID. Here, we relied on mapping the chemical nomenclature, common name or drug name to our ID via other database IDs such as the PubChem ID (CID), DrugBank ID, or PharmGKB ID.

Uniqueness of the chemicals is not the only problem in chemical biology data integration. There are also problems concerning protein and gene IDs. Mapping gene ID to protein ID is not a one-to-one relationship. Genes can encode for multiple proteins by the means of alternative splicing [47]. In cases where we only had gene IDs that were mapped to multiple proteins, we only included the gene ID. However, whenever possible, we mapped both gene and protein IDs.

In the following chapter 2, we describe ChemProt, a database service that integrates large-scale chemical biology data with complex disease networks and is a resource for annotated and predicted chemical-protein interactions. ChemProt not only allows elucidation of the effects of a compound on the disease networks, but also can be useful in identification of genes that may play a role in modulating the chemical response to that compound.

The usefulness of our work can be seen in table 2.2, where 4 articles have cited the use of ChemProt server in their analyses. As of November 2012, ChemProt has also been cited by 14 other articles.

We present an update of the ChemProt database in article 2, where we describe the addition of a new interface that allows visualiztion of the interactions in a heatmap. Other important features added in this update are new search options, interactive complex disease network, SEA implementation and an option to download the extracted data for further analysis. Together, these new features provide a user-friendly interface for data extraction.

Table 2.1 shows the update statistics from ChemProt version 1 to ChemProt version 2. An overview for each database with measurements, interactions, compounds, uniprot and ensembl numbers.

2.1 Protein-Protein Interaction data

Biological research at cell, molecular, structural, biochemistry, and biophysical levels has, for decades, produced indispensable information about functions and properties of individual proteins. These are now stored in extensively curated databases like the Universal Protein Resource (UniProt) [48]. As proteins rarely act alone but rather as complexes of proteins, it is necessary to study these protein-protein interactions (PPIs) in order to understand some of the complex molecular relationships within living organisms. A complete set of PPIs that can occur in a living organism is termed the interactome [49]. An understanding of interactomes would be helpful in revealing protein functions and also in gaining an understanding of phenotypes and complex diseases.

The most commonly used experimental methods to identify PPIs are yeast two-hybrid system (Y2H) [50] and tandem affinity purification (TAB) followed by mass spectrometry (MS) [51].

The human interactome [52], maintained at CBS, combines protein-protein interaction network experiments from both human and model organisms (49), and includes data extracted from databases such as MINT [53], BIND [54], GRID [55], HPRD [56], IntAct [57], DIP [58], PDZbase [59], Reactome [60], and KEGG [61]. False positives in the PPIs are common due to the error-prone nature of the high-throughput experiments used for their detection. Therefore, all data in the CBS interactome has been given scores and validated against a gold standard to remove such false-positive interactions [62].

Human disease complexes

Human disease complexes were created by mining for disease-related proteins from GeneCards, a database of human genes that provides genomic related information on all known and predicted human genes [63] and OMIM, an open source database focusing on the relationship between genotype and phenotype [64]. Subsequently PPIs were applied from the CBS interactome [65]. Lage et al. [52] generated 1524 disease-associated protein complexes and analyzed them through 5 different sources of information. They calculated p-values associated with each disease, which represents the enrichment of proteins from this disease in the particular complex. In BioAlma [66], the relevance scores are based on the co-occurrence (in Medline documents) of complex-related disease-terms with the genes in the complex. The more the gene–disease-term pair co-occurrences observed, the higher the ”weights” value. GO (gene ontology) biological process and GO cellular component sources were used to ensure that the complexes were biologically relevant entities. The enrichment of GO terms (biological process and cellular component) was compared to randomly generated complexes. Human Protein Atlas (HPA) [67] source was used to enrich the proteins co-occurring in the same tissues and was determined using high quality manually curated immunohistochemistry data. The enrichment was again compared to randomly generated complexes. Finally, mRNA expression was used to map complexes to tissues using the expression data from 73 non-diseased tissues from the Novartis research foundation gene expression database [68]. The higher the z-score, the more the tissue affected by the complex of proteins.

	BindingDB	**BindingDB**	**CTD**	**CTD**	**DrugBank**	**DrugBank**	**IUPHARdb**
ChemProt Ver.	**New**	**Old**	**New**	**Old**	**New**	**Old**	**New**
Measurements	651042	62172	30243	33555	13384	10431	3969
Interactions	530981	53966	30243	33555	13384	10431	3037
Compounds	296887	33212	3810	2014	5680	4037	1451
Uniprot	3766	619	2633	3759	3874	4038	201
Ensembl	2314	352	3731	7772	1846	1153	202

	KiDB	**KiDB**	**PharmGKB**	**PharmGKB**	**STITCH**	**STITCH**	**Wombat**
ChemProt Ver.	**New**	**Old**	**New**	**Old**	**New**	**New**	**Old**
Measurements	22626	16204	1608	1615	374661	813100	423261
Interactions	14890	7805	1608	1615	361641	811015	319176
Compounds	3305	1141	383	386	187233	45216	179733
Uniprot	201	295	565	566	4929	12709	2780
Ensembl	229	181	581	582	3576	14575	1760

	Wombat	**ChEMBL**	**ChEMBL**	**PubChem**	**PubChem**	**Total**	**Total**
ChemProt Ver.	**New**	**New**	**Old**	**New**	**Old**		
Measurements	174258	4332258	1389637	2459931	63652	5369210	2563828
Interactions	70444	3433626	940416	2276126	63652	3970866	1932618
Compounds	60281	777776	353249	383974	43700	1157925	729986
Uniprot	755	5656	4444	598	136	15290	18697
Ensembl	415	3621	2088	477	106	8205	15836

Table 2.1. A comparison between version 1 and 2 for the ChemProt database.

Ref.	Acticle title	Comment
[69]	Application of Computational Systems Biology to Explore Environmental Toxicity Hazards	They extracted chemical–protein association networks for each DDT isomer and its metabolites using ChemProt.
[70]	Analysis of Commercial and Public Bioactivity Databases.	Cited
[71]	Assessing Drug Target Association Using Semantic Linked Data	Cited
[44]	Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing	Cited
[72]	Association between chemical pattern in breast milk and congenital cryptorchidism: modelling of complex human exposures.	The top significant chemicals were computationally screened against ChemProt
[73]	Back to the Roots: Prediction of Biologically Active Natural Products from Ayurveda Traditional Medicine	The chemical space of Ayurveda NPs with known biological activity was captured by mapping the compounds against ChemProt
[74]	Chemical structural novelty: on-targets and off-targets.	Cited
[75]	Identifying Novel Drug Indications through Automated Reasoning.	Cited
[76]	DrugLogit: Logistic Discrimination Between Drugs and Non-drugs Including Disease-Specificity by Assigning Probabilities Based on Molecular Properties.	Cited
[77]	Mapping the genome of Plasmodium falciparum on the drug-like chemical space reveals novel anti-malarial targets and potential drug leads	Information on chemical–protein interactions was extracted from ChemProt
[78]	Novel computational approaches to polypharmacology as a means to define responses to individual drugs.	Cited
[79]	Ranking Transitive Chemical-Disease Inferences Using Local Network Topology in the Comparative Toxicogenomics Database.	Cited
[2]	Structure and dynamics of molecular networks: A novel paradigm of drug discovery. A comprehensive review.	Cited
[80]	Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery researc.h	Cited
[81]	The impact of network biology in pharmacology and toxicology.	Cited
[82]	The importance of integrating SNP and cheminformatics resources to pharmacogenomics.	Cited
[83]	The role of computational methods in the identification of bioactive compounds.	Cited
[84]	Virtual Interactomics of Proteins from Biochemical Standpoint.	Cited

Table 2.2. As of November 2012 18 articles have cited the ChemProt database, of which 4 papers have used information in their analysis

Paper I

ChemProt: a disease chemical biology database

Olivier Taboureau[1], Sonny Kim Nielsen[1], Karine Audouze[1], Nils Weinhold[1], Daniel Edsgärd[1], Francisco S. Roque[1], Irene Kouskoumvekaki[1], Alina Bora[2], Ramona Curpan[2], Thomas Skøt Jensen[1], Søren Brunak[1] and Tudor I. Oprea[1,3]

Nucleic Acids Research, (2011) 39 (suppl 1): D367-D372. doi: 10.1093/nar/gkq906

[1]Department of Systems Biology DTU, Building 208, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, 2800, Denmark, [2]Department of Computational Chemistry, Institute of Chemistry, Romanian Academy, Timisoara 300223, Romania and [3]Department of Biochemistry and Molecular Biology, Division of Biocomputing, University of New Mexico School of Medicine, Albuquerque, New Mexico 87131, USA

Abstract

Systems pharmacology is an emergent area that studies drug action across multiple scales of complexity, from molecular and cellular to tissue and organism levels. There is a critical need to develop network-based approaches to integrate the growing body of chemical biology knowledge with network biology. Here, we report ChemProt, a disease chemical biology database, which is based on a compilation of multiple chemical–protein annotation resources, as well as disease-associated protein– protein interactions (PPIs). We assembled more than 700.000 unique chemicals with biological annotation for 30.578 proteins. We gathered over 2 million chemical–protein interactions, which were integrated in a quality scored human PPI network of 428.429 interactions. The PPI network layer allows for studying disease and tissue specificity through each protein complex. ChemProt can assist in the in silico evaluation of environmental chemicals, natural products and approved drugs, as well as the selection of new compounds based on their activity profile against most known biological targets, including those related to adverse drug events. Results from the disease chemical biology database associate citalopram, an antidepressant, with osteogenesis imperfect and leukemia and bisphenol A, an endocrine disruptor, with certain types of cancer, respectively. The server can be accessed at http://www.cbs.dtu.dk/ services/ChemProt/.

Introduction

The old drug design paradigm, i.e. drugs interact selectively with one or two targets (proteins), resulting in treatment and prevention of disease, is now challenged by several studies that show most drugs interacting with multiple targets (‘polypharmacology’) [85, 86]. For example, celecoxib, often considered a selective cyclooxygenase-2 non-steroidal anti-inflammatory drug (NSAID), has been documented to be active on at least two additional targets, namely carbonic anhydrase II and 5-lipoxygenase [87]. Rosiglitazone, which has been used for the treatment of type II diabetes mellitus, not only stimulates the peroxisome proliferator activated receptor g, but also blocks interferon gamma-induced chemokine expression in Graves disease or ophthalmopathy [88]. Polypharmacology is not always beneficial, as it often causes side effects: Cisapride, which acts as a serotonergic 5-HT4 receptor agonist, as well as astemizole, which blocks histamine H1 receptors (H1Rs), have both been withdrawn from all markets due to the risk of fatal cardiac arrhythmia associated with their blockade of the hERG potassium ion channel, an unanticipated and undesirable ‘anti-target’ associated to QT prolongation and ‘torsades de pointes’ [89]. However, ‘target’ and ‘anti-targets’ are dynamic attributes, as exemplified by the case of H1R antagonists and their (in)ability to achieve clinically significant levels in the brain, influenced by the ATP-binding cassette transporter ABCB1 (also known as P-glycoprotein), which effluxes some of these drugs from the brain [90]. Acquiring knowledge of the complete pharmacology profile has inspired new strategies to predict and to characterize drug-target associations in order to improve the success rates of current drug discovery paradigms, i.e. increase the efficacy and reduce toxicity and adverse effects [86].

As large-scale chemical bioactivity databases are being assembled, the polypharmacology (i.e. high affinity bioactivity across related targets) and promiscuity (i.e. low affinity across multiple families) of chemicals are expanding the chemical space for druggable targets [32]. These studies are often focused on specific protein families, such as G-protein coupled receptors [91], nuclear receptors [92] and kinases [93], but global pharmacology profiles of chemicals are considered as well [85, 20]. Recent chemoinformatics advances support the development of polypharmacology data mining, e.g. via iPHACE, an integrative web-based tool that enables pharmacological space navigation for small molecule drugs [94] or based on a Similarity Ensemble Approach (SEA) to relate protein pharmacology by ligand chemistry [20]. Biological information can also be retrieved for a large set of chemical compounds through PubChem [95], CheBI and ChEMBL [96].

Two conceptual developments support polypharmacology: systems pharmacology, aimed at drug actions in the context of regulatory networks [8]; and systems chemical biology [7], which introduces chemical awareness in systems biology. Since proteins rarely operate in isolation inside and outside cells, but rather function in highly interconnected cellular pathways, interactome networks have been developed by data integration. Yildirim et al. [97] combined FDA-approved drugs with a human protein–protein interaction (PPI) network (human interactome) in order to analyze the interrelationships between drug targets and disease–gene products i.e. disease–proteins. Similar work has been based on PubChem bioassays as source of polypharmacology [98]. The use of side-effect similarity has been proposed on the assumption that drugs with similar side-effects are likely to interact with similar target proteins [99]. Recent advances include a protein–protein association network based on the chemical toxicology of environmental chemicals [100] and a human disease network linking disorders and disease genes to various known phenotypes [101].

Our goal in the present work was to develop a disease chemical biology server, called ChemProt, based on the integration of chemical–protein annotation resources that are now accessible from large repositories, and curated disease-linked PPI data [65]. ChemProt is designed to assist the elucidation of drug actions in the context of cellular and disease networks. Further to that, it allows the identification of additional genes that may play major roles in modulating chemical response i.e. to drugs, environmental chemicals and natural products, thus leading to new options in drug discovery and environmental chemical evaluation. Lastly, the ChemProt server could contribute to drug repurposing as well as to the investigation of chemicals related to anti-targets and adverse drug events.

2.2 Implementation

Data sources

We first gathered chemical–protein interaction data from different open source databases i.e. ChEMBL (version chembl_05) [96], BindingDB [33], PDSP Ki Database [31], DrugBank (version 2.5) [102], PharmGKB [103] and two commercial databases, WOMBAT (version 2009) and WOMBAT-PK (version 2008) [32]. Active compounds from the PubChem bioassay (2010) have been collected as well [95]. We considered only active compounds from 'confirmatory' assays in order to capture high-confidence chemical–protein annotations from PubChem. These databases provide experimental evidence of chemical–protein interactions. Drug-target in- formation was collected from DrugBank and PharmGKB. In addition, we integrated chemical–protein associations from CTD (version 2009) [104] and STITCH (version STITCH 2.0) [105]. These last two databases consider the effect or modulation (positive or negative) of a chemical on proteins, other than that defined as binding activity. Examples include gene expression or pathway data, where the deregulation of a gene by a chemical may be not due to a physical interaction between the two entities but a response at a cellular level. Duplicate chemicals from the multiple databases were found by using InChI keys and were merged into a single ChemProt ID. However, the biological information associated to each chemical was conserved for users looking on selective databases. Overall, the final database contains 700000 distinct molecules annotated for 30578 proteins.

Descriptors and similarity measurement

The chemical structure of the molecules was encoded using two rather different types of fingerprints. The 166 MACCS keys, encode the presence or absence of predefined substructural or functional groups [17]. On the other hand, a more complex 3-point pharmacophore fingerprint (GpiDAPH3) is based on an expansion of the PATTY pharmacophore feature recognition scheme of a 2D structure [106]. This scheme assigns one or more pharmacophore feature types to all atoms in a molecule using a predefined list of SMART queries. The list of pharmacophore feature types comprises:

hydrogen-bond donor (D), hydrogen-bond acceptor (A), polar (P) and hydrophobic (H). In addition, an extra label (p or pi) is added to each feature if the originating atom or group is sp2-hybridized or planar for other reasons. The GpiDAPH3 pharmacophore feature scheme is expressed in 2D as triplet feature combinations with a graph based inter-atom distance binning scheme. Both fingerprints are implemented in the Molecular Operating Environment (MOE, version 2008.10) [107]. The similarity between two molecules is measured using the Tanimoto coefficient (Tc), a method of choice for the computation of fingerprint-based similarity [108]. The Tc is defined as the number of bits in common divided by the total number of used bits in both molecules. For any pair of chemicals, Tc assumes values between 0 and 1. A high Tc represents high similarity.

PPI network

The human interactome used is an in-house protein–protein interaction network inferred from experiments in both humans and model organisms [65]. Using an elaborate scoring scheme, all interactions have been validated against a gold standard [62]. The current interactome contains 428.429 unique protein–proteins interactions derived from source databases such as BIND [54], GRID [109], MINT [110], dip_full [58], HPRD [111], intact [112], mppi [113], MPact [114], Reactome [115] and KEGG [116]. Data are transferred between organisms by using the Inparanoid orthology database [117]. In total the human interactome comprises 22.997 genes.

Human disease genes and complexes

Based on a previous study [52], disease-associated protein complexes were associated to the chemical–protein annotation by mining OMIM [118] and GeneCards [63], two data resources for genes association to diseases, we collected a list of 2227 unique disease-related proteins and mapped the complexes of genes to disease. Similarly, complexes of genes were mapped to Gene Ontology (GO) terms [119] and tissues by using the expression data from 73 non-disease tissues from the Novartis Research Foundation Gene Expression Database (GNF) [68] and Human Protein Atlas [120]. Users of ChemProt can thus retrieve gene complexes that are related to a query chemical and visualize the annotations of each complex.

2.3 Applications

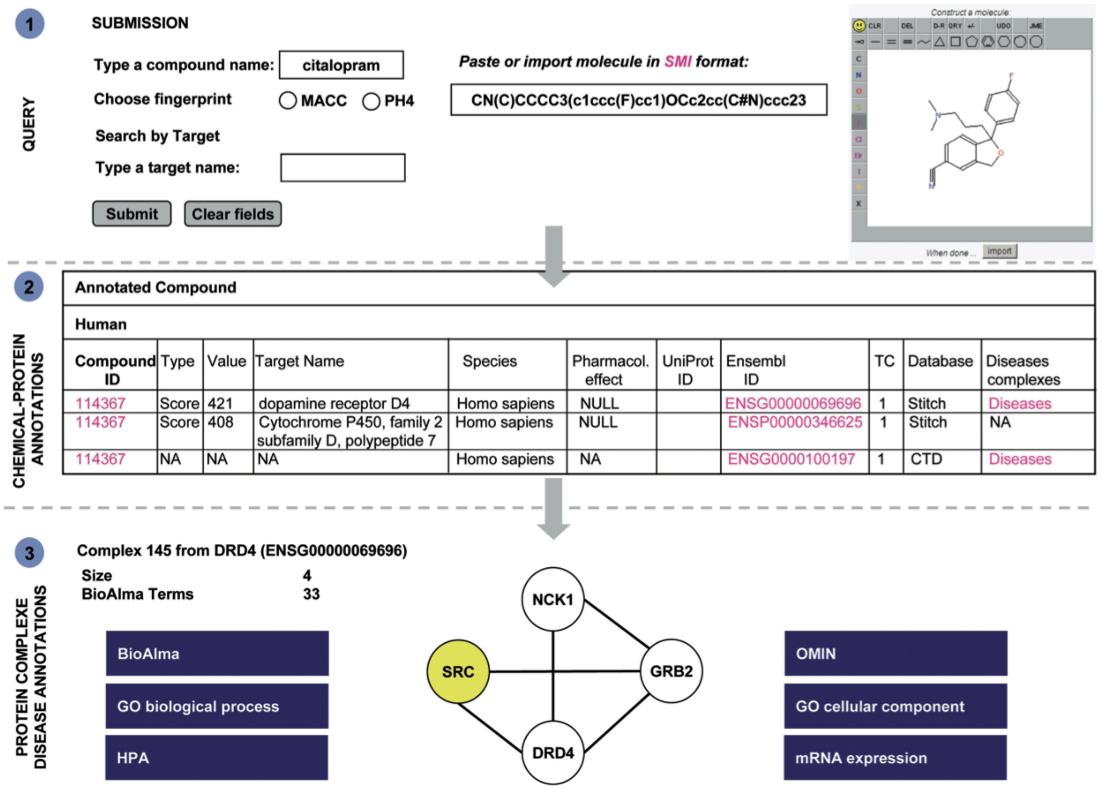
Chemical–protein interactions

Chemicals can be searched using a common name, SMILES and by drawing the 2D structure, or retrieved through their annotation to a protein.

Users can then choose the descriptor space and the Tc threshold to be used for similarity search. Following a successful query, hits grouped by species will be returned, together with computed physico-chemical properties such as Molecular Weight, LogP, the number of hydrogen bond donors and acceptors, the number of rigid bonds and the number of rings, based on the Marvin applet from Chemaxon [121]. Hits are provided separately for known annotations, and for prediction of small molecule bioactivity, respectively. The biochemical and pharmacological effects of a chemical, e.g. substrate, inhibitor, agonist or antagonist, are provided if such information is available, together with hyperlinks to UniProt and Ensembl, which lead to more information on protein sequence and function, respectively.

From chemical–protein interactions to complex protein–disease associations

The unique feature of ChemProt is that it offers the user the possibility to get information at a cellular level, by linking chemically-induced biological perturbations to specific tissues and phenotypes. Proteins that are both affected by a chemical and participate in one or more protein complexes are highlighted in the results table of the ChemProt server. By clicking on the protein, the user is redirected to the 'Disease complexes' server and has to choose which complex to visualize. On the 'Disease complexes' server, size and illustrations of the protein network are provided. Additionally, enrichment analysis results of the proteins in the complex are shown, with respect to disease association (OMIM, BioAlma), GO terms (biological process, cellular component) and tissue specificity (Human Protein Atlas, GNF). To ensure that the complexes were biologically relevant entities, the enrichment of the biological terms (OMIM, GO,..) was compared to randomly generated complexes (1.0e6). The significances were calculated using a hyper-geometric test and the P-value for the most significant enriched term for each of the data types was calculated as previously described [52]. The table presenting the OMIM enrichment results is interactively linked with an illustration of the protein complex where proteins associated with the selected disease are colored yellow. Output of the chemical–proteins interactions and disease complexes can be downloaded from the ChemProt website. In addition, the 'Reflect' service provides further information on chemicals and genes [122]. 'Reflect' tags gene, protein and small molecule names in text and offers the opportunity to quickly view additional information on the ChemProt results, including synonyms, protein sequences, domains, 3D structures and subcellular location.



1
QUERY
2
CHEMICAL-PROTEIN ANNOTATIONS
3
PROTEIN COMPLEXE DISEASE ANNOTATIONS
Construct a molecule:
When done ... import
NCK1
SRC
GRB2
DRD4

SUBMISSION

Type a compound name: citalopram

Choose fingerprint ○ MACC ○ PH4

Search by Target

Type a target name:

Submit | Clear fields

Paste or import molecule in SMI format:

CN(C)CCCC3(c1ccc(F)cc1)OCc2cc(C#N)ccc23

Annotated Compound

Human

Compound ID	Type	Value	Target Name	Species	Pharmacol. effect	UniProt ID	Ensembl ID	TC	Database	Diseases complexes
114367	Score	421	dopamine receptor D4	Homo sapiens	NULL		ENSG00000069696	1	Stitch	Diseases
114367	Score	408	Cytochrome P450, family 2 subfamily D, polypeptide 7	Homo sapiens	NULL		ENSP00000346625	1	Stitch	NA
114367	NA	NA	NA	Homo sapiens	NA		ENSG0000100197	1	CTD	Diseases

Complex 145 from DRD4 (ENSG00000069696)

Size	4
BioAlma Terms	33

BioAlma

GO biological process

HPA

OMIN

GO cellular component

mRNA expression

Figure 2.1. Chemical–protein annotation and disease associations retrieved from ChemProt for the compound citalopram. (1) The compound can be queried using different formats (name, SMILES and structure). (2) A query results in a table showing protein annotations and bioactivity predictions for the compound. (3) Finally, a protein–protein interaction network (protein–complex) for a target protein can be depicted and disease associations (OMIM and BioAlma) and other biological components (GO terms, HPA and mRNA expression) are displayed.

Examples

With the integration of several databases, ChemProt not only provides pharmacological information, but also includes biological data associated to environmental chemicals and natural products. As seen in the examples below, ChemProt can be queried for drugs as well as environmental chemicals. A search for citalopram, an antidepressant, illustrates the complementarity of the integrated databases within ChemProt (Figure 2.1). Marketed as a selective serotonin reuptake inhibitor (SSRI) (DrugBank), this drug displays bioactivity on seven human proteins (ChEMBL). Via ChemProt, four other proteins (DRD3, 5HT1B, 5HT3, ADRA2A) are retrieved from the Ki database. Additional information on drug-target associations is provided by STITCH and CTD. From the first annotation to the D4 dopamine receptor (DRD4), the disease term (under Disease Complexes) is highlighted, indicating that protein–protein interaction information for this protein is available. Using the link to the Disease Complexes server, one finds that DRD4 interacts with three proteins (SRC, GRB2 and NCK1). According to OMIM, this protein network is associated to osteogenesis imperfecta and leukemia and, according to BioAlma, to several psychotic disorders. GO enrichment indicates significant association of the protein complex to signal complex formation and vesicle membrane. Furthermore, tissue annotation suggests that this complex is mainly expressed in follicle and non-follicle cells (HPA) and dentritic cells (GNF). Although it might be surprising to see a connection between antidepressant and leukemia, it has been shown recently that antidepressants such as chlomipramine and fluoxetine reduce the growth of B-cell malignancies in leukemia [123].

The second query, 'bisphenol A' (BPA), is an environmental pollutant used as plasticizer [124]. BPA has biological activity on the estrogen receptor α (ESR1), the androgen receptor (AR) and the estrogen related receptor gamma (ERR3). However, several other proteins are retrieved from CTD and STITCH based on association data with this chemical. Looking at ESR1 in the Disease Complexes server, a complex of 17 proteins is depicted (complex 265) with significant associations to Li-FRAUMENI syndrome, breast cancer and neoplasms. Enrichment analysis indicates that the complex is found in the nucleus (GO cellular component), involved in the regulation of metabolic processes and transcriptionally regulated by the RNA polymerase II promoter (GO biological process). Furthermore, data from immunohistochemistry studies suggest that the complex is mainly located in the endometrium and the cerebral cortex (HPA). The disease chemical biology network for BPA indicates that, under certain conditions, this chemical may be associated with certain types of cancers.

We have illustrated that ChemProt integrates molecular, cellular and phenotypic data associated to small molecules, which can lead to novel links and suggest new avenues for research. We envisage that the ChemProt server will find applications within a variety of chemogenomics, polypharmacology and systems chemical biology studies. ChemProt will be updated once a year with new compounds, new interactions and more sophisticated descriptors.

Acknowledgements

Sunset Molecular Discovery LLC (www.sunsetmolecular .com) contributed with the WOMBAT databases.

Funding

EU (DEER); Innovative Medicines Initiative Joint Undertaking (eTOX); Danish Research Council for Technology and Production Sciences; Lundbeck foundation and the Villum Rasmussen Foundation. Funding for open access charge: DEER.

Conflict of interest statement. None declared.

Paper II

ChemProt 2.0: Visual navigation in a disease chemical biology database

Sonny Kim Kjærulff[1+], Louis Wich[1+], Jens Kringelum[1+], Ulrik P. Jacobsen[1], Karine Audouze[1], Irene Kouskoumvekaki[1], Ole Lund[1], Søren Brunak[1], Tudor I. Oprea[1],[2], Olivier Taboureau[1]

Accepted in *Nucleic Acids Research*, October 2012

[1]Department of Systems Biology DTU, Building 208, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, 2800, Denmark, [2]Department of Computational Chemistry, Institute of Chemistry, Romanian Academy, Timisoara 300223, Romania and [3]Department of Biochemistry and Molecular Biology, Division of Biocomputing, University of New Mexico School of Medicine, Albuquerque, New Mexico 87131, USA

* To whom correspondence should be addressed. Tel: +45 4525 2489; Fax: +45 4593 1585; Email: otab@cbs.dtu.dk (OT)

+ SKK, LW and JK contributed equally to this work.

Abstract

ChemProt-2.0 (http://www.cbs.dtu.dk/services/ChemProt/ChemProt-2.0) is a public available compilation of multiple chemical-protein annotation resources integrated with diseases and clinical outcomes information. The database has been updated to more than 1.15 million compounds with 5.32 millions bioactivity measurements for 15,290 proteins. Each protein is linked to quality-scored human protein-protein interactions (PPI) data based on more than half a million interactions, for studying diseases and biological outcomes (diseases, pathways and GO terms) through protein complexes. In ChemProt-2.0, therapeutic effects as well as adverse drug reactions have been integrated allowing for suggesting proteins associated to clinical outcomes. New chemical structure fingerprints were computed based on the Similarity Ensemble Approach (SEA). Protein sequence similarity search was also integrated to evaluate the promiscuity of proteins, which can help in the prediction of off-target effects. Finally, the database was integrated into a visual interface that enables navigation of the pharmacological space for small molecules. Filtering options were included in order to facilitate and to guide dynamic search of specific queries.

Keywords: systems pharmacology / disease chemical biology / clinical outcomes/ network biology / chemoinformatics /translational informatics

2.4 Introduction

In recent years there has been a shift from the traditionally secret experimental data kept by the pharmaceutical industry to a more open access culture in relation to data sharing [125]. For this reason we have been witnessing a steady increase of public repositories of bioactive small molecules such as ChEMBL [30] and PubChem [34]. However, as public repositories of bioactive small molecules have only just recently been made available, the problem of how to handle chemical entities is still largely unsolved. Pooling data from small molecule databases poses special problems. Even though standards have been widely adopted to describe genes and proteins (eg. Ensembl ID, Entrez ID for genes, UniProt ID for proteins), small molecule identifiers, as well as measures for properties such as biological activities, are not necessarily standardized across different resources [126].

One could claim that the bottleneck in understanding how small molecules perturb biological systems is no longer in the generation, gathering and availability of experimental data but in their organization, presentation and visualization; in other words, in the development of centralized systems that would better enable their exploitation. The problem is not only how to extract data from different (federated) resources, it is also important to provide solutions that facilitate provenance tracking, visualization, uniform and systematic description of data and their integration in ways that can preserve the semantic relationships between the different entities.

Furthermore, the number of failures of drug candidates in advanced stages of clinical trials has increased and the number of submissions for FDA approval has decreased in the last decade. One of the reasons may be our reductionist approach to discovery, whereby a complex system, namely a drug and its metabolites interacting with many proteins across multiple cellular compartments and tissues over time, is reduced to a simplistic ligand-target interaction model. This is probably too crude and emphasizes the need to look at the effects of compounds on global systems aided by the integration of multiple biological and temporal data sources.

With the emerging fields of chemogenomics [127], systems pharmacology [128] and systems chemical biology [7, 80], it becomes feasible to investigate the drug action at different levels from molecular to pathway, cellular, tissues and clinical outcomes [129]. For example, it has become apparent that many common diseases such as cancer, cardiovascular diseases and mental disorders are much more complex than initially anticipated, as they are caused by multiple molecular and cellular dysfunctions rather than being the result of a single defect. Therefore, network-centric therapeutic approaches that consider entire pathways rather than single proteins must be investigated [130].

Among the recent advances in the field of systems chemical biology, servers supporting drug profiling such as STITCH [37], DisGENET [29] or the new database PROMISCUOUS [131] should be mentioned. STITCH3 provides confidence scores that reflect the level of confidence and significance of compound-protein interactions. PROMISCUOUS is a resource focused on drug compounds, including withdrawn and experimental, containing drug-protein interaction and side-effect information. DisGENET is a comprehensive gene-disease association database focused on the current knowledge of human genetic diseases including Mendelian, complex and environmental diseases.

We have previously reported the development of ChemProt, a disease chemical biology database [132]. Compared to other approaches, ChemProt 1.0 offered a high level of integration of chemical and biological data, including internally curated disease-associated PPIs [65]. Here we present the second release of ChemProt, a resource of annotated and predicted disease chemical biology interactions. ChemProt-2.0 can be accessed at http://www.cbs.dtu.dk/services/ChemProt-2.0/. The present release contains a compilation of over 1,100,000 unique chemicals with biological activity for more than 15,000 proteins. We have added a visual interface that supports user-friendly navigation through the data, biological activities and disease associations. ChemProt-2.0 now enables the user to query the database

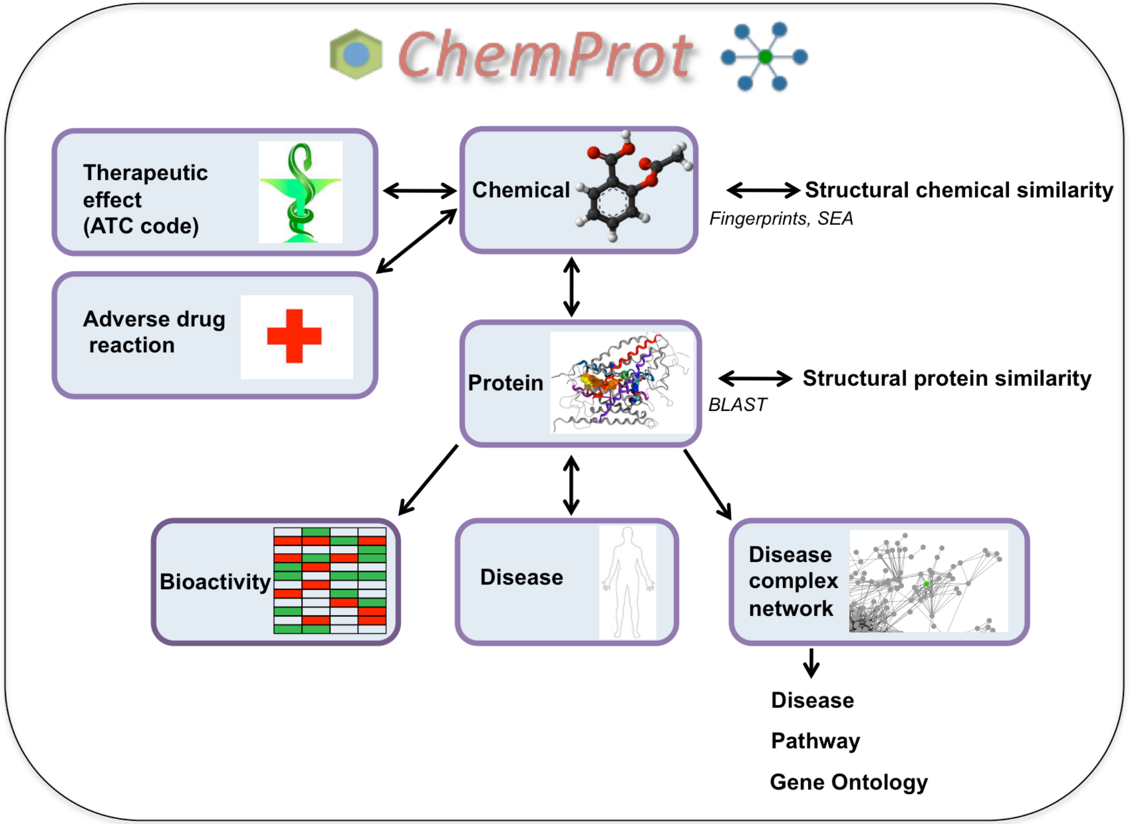


Figure 2.2. A workflow of the functionalities in ChemProt-2.0 is depicted. User can query ChemProt-2.0 using chemical, protein, disease, ATC code and Side Effects. Outcomes from the query are represented with the arrows.

not solely by chemicals or proteins, but also through therapeutic effects, adverse drug reactions and diseases. The Similarity Ensemble Approach (SEA) developed by Keiser et al. [20] has also been implemented, so that protein sequence similarity can be used when examining chemical promiscuity. With these updates, ChemProt-2.0 offers an integrative approach to understand the impact small molecules on biological systems, and contribute to the investigation of molecular mechanisms related to diseases and clinical outcomes. A workflow of the implementation is shown on Figure 2.2.

2.5 Implementation

Data sources

Chemical protein interactions data was gathered in June 2012 from updated open source databases ChEMBL (version 14), BindingDB [33], PDSP Ki database [31], DrugBank (version3.0) [23], PharmGKB [22], active compounds from the PubChem bioassay (2012) targeting human proteins and the two commercial databases, WOMBAT (version 2011) and WOMBAT-PK (version 2011) [32]. The IUPHAR-DB database [35] was also integrated in the new version of ChemProt. Chemical-protein annotations that lack explicit bioactivity data might be of interest in the mining of a large and diverse integrated database. Therefore, we included also data from CTD [36] and STITCH [37]. CTD extract literature data about environmental chemicals and how they modulate gene expression, whereas STITCH provides chemical-protein relationships from text mining the co-occurrence of a chemical term and a protein (gene) term in MEDLINE abstracts. Clinical outcomes were of special interest in this version and we decided to include information from the Anatomical Therapeutic Chemical (ATC) Classification System [133] developed by the World Health Organization (WHO), as well as side effect data from Dailymed (http://dailymed.nlm.nih.gov/dailymed/).

From a biological perspective, we updated our internal human interactome platform to reach 14,421 genes interacting through 507,142 unique PPIs. The updated version of OMIM [64], GeneCards [134], KEGG [61], Reactome [60] and Gene Ontology [48] databases was also downloaded (June 2012), curated and integrated in ChemProt-2.0. Also, the human disease network developed by Goh et al. [101] was integrated, allowing association of proteins to disease categories.

Predictions methods

Based on the assumption that compounds sharing similar structure have potential similar bioactivities, we encoded the chemical structure with two different types of fingerprints, the 166 MACCS key which encode the presence or absence of some predefined substructural or functional groups [17] and the FP2 fingerprints computed with OpenBABEL [135]. Chemical similarity between two compounds is quantitatively assessed using the Tanimoto coefficient. By including the SEA method [20], one can also predict potential new targets for a compound. For the internal development of SEA, compounds with an activity value lower than 100 M were considered (only IC50, EC50, Potency, AC50, Ki values were used). Furthermore, to complete the set of active protein ligands, annotated compound-protein interactions from CTD, DrugBank and PharmGKB were also included, together with annotated protein-compound in the STITCH database. For this dataset, the raw similarity score, i.e. the sum of ligand pair wise Tanimoto coefficients based on the FP2 fingerprint, is 0.44. All proteins with more than five bioactive ligands were considered.

In addition, for all protein targets we operated under the assumption of promiscuity, i.e. proteins with high sequence similarity may share similar functions and may be targeted by the same compound (likely with different bioactivities). Protein sequences were obtained from Uniprot [48] , and sequence comparisons were computed using BLASTP [136]. The similarity of two sequences was assessed using an E-score, an expectation value related to the probability that sequence similarity between two proteins is not achieved by random chance [136]. We filtered the output and proteins with an E-value lower than 10e-10 (as default) are depicted.

With respect to side effects (SE), 988 small molecule drugs were matched against 174 SE as described [44]. Term frequency vectors compiled from Dailymed were integrated in ChemProt-2.0 and proteins associated to each drug are then depicted.

2.6 Visual interface

In ChemProt 2.0, a visual interface was implemented to facilitate the visualization of the results using HTML 5 and JavaScript. The core of the interface has been designed in the form of a heatmap. The chemical-protein associations are depicted in a pie-chart heatmap where each pie corresponds to the database from which we gathered the information. Hovering over the pie charts with the pointer, activity values are then displayed. The user can select different display settings (Circles, Fill and Rectangles). A valuable feature is the handling of multiple activities that have been gathered for a given compound-target pair by selecting “All” values. A color spectrum from blue (low activity) to red (strong activity) is used to indicate the activity (Figure 2.3). It is also possible to select a specific database or/and a specific activity type and define a range of activities (threshold) of interest in order to optimize the query. Results from the SEA approach are also integrated in the “Activity Type”.

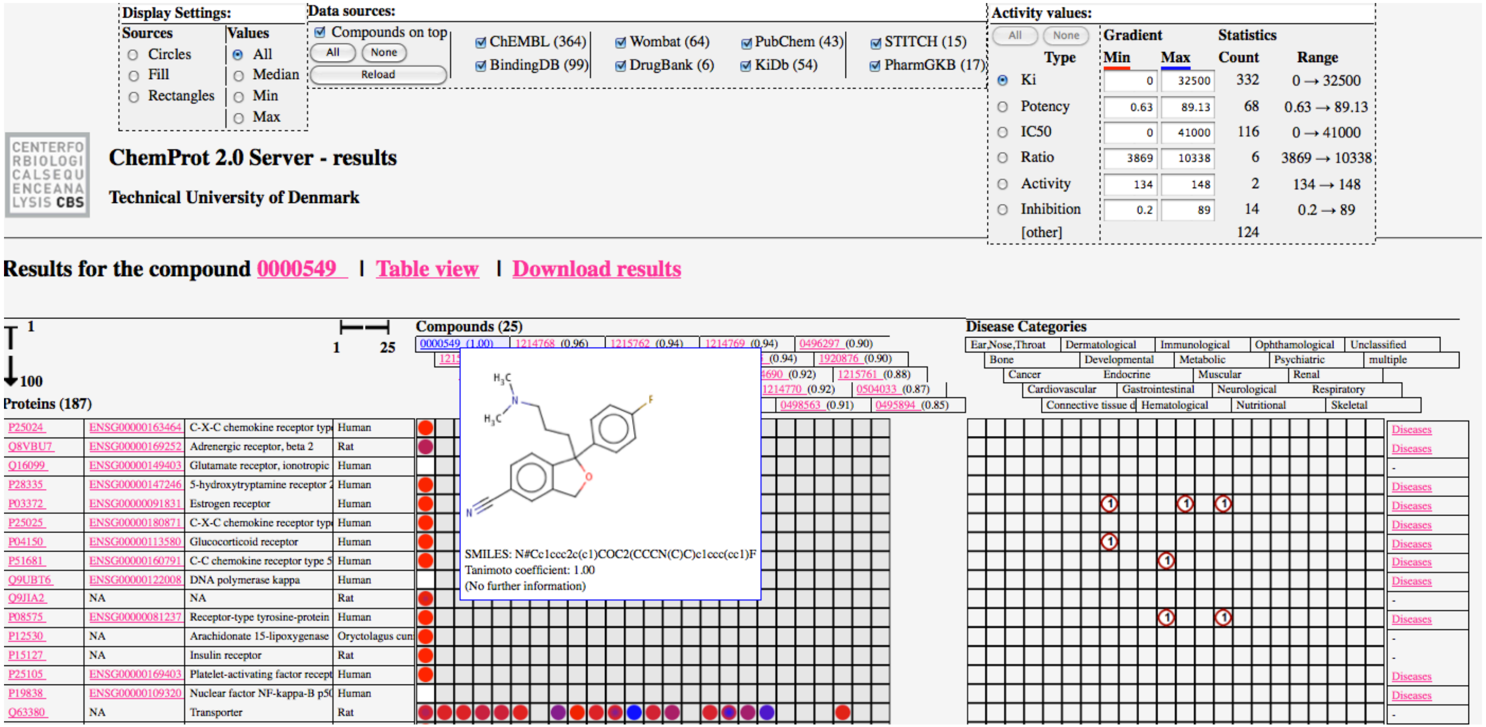
The compound query is always shown in the first column followed by similar compounds (sorted in descending order of similarity) whereas the protein queried is depicted in the first row. To optimize the display, the heatmap is limited to a section of 100 rows x 100 columns. If the chemical-protein matrix is larger, we have included an arrow feature (→) that allows the user to upload the next 100 data items for both axes. The user has still the possibility to view the data in a table format and to download the results in flat-file format. In the table format, display mode the user can dynamically sort and group the activities according to compound, target, species, activity type, etc.

A second heatmap that depicts protein-disease categories is also integrated, which suggests proteins that may be involved in diseases. Next to it, the “Diseases” link redirects the user to the disease-associated proteins complex around the selected protein. A new, dynamic interface has been implemented, where the proteins associated to a biological term are shown when highlighting the term of interest (Figure 2.4).

2.7 Applications

The ChemProt-2.0 database interface is accessible freely online. In addition to the chemical and protein search that was previously implemented, the user can search by diseases, ATC codes and side effects. For example, the query “epilepsy” returns 2,662 compounds active on 13 proteins associated to this disease. Similarly, looking for the side-effect “hallucinations”, 15 drugs (with the term frequency associated to it) active on 470 proteins are displayed. Some of these drugs (ropinirol, pergolide, amantadine, pramipexole) are used for the treatment of Parkinson diseases, by affecting the dopaminergic and serotonergic systems. Interestingly, visual hallucinations are symptoms of the Parkinson’s disease and perturbing the serotonergic system could help to alleviate these symptoms [137]. Another interesting aspect is that these drugs affect several proteins associated to “Bone” and osteoporosis disease. For example, there is a possible association between the polymorphism of the serotonin transporter (HTT) and the development of osteoporosis [138]. Some of these drugs bind to HTT and could thus be potentially investigated for drug repurposing.

Many diseases seem not to be the result of a single defect but are rather caused by multiple molecular and cellular abnormalities. Therefore, observations of a drug not only at the molecular level, but also at cellular and systems levels should guide therapeutic strategies for the development of better



Display Settings:

Sources: ○ Circles ○ Fill ○ Rectangles

Values: ● All ○ Median ○ Min ○ Max

Data sources:

☑ Compounds on top

☑ ChEMBL (364) ☑ BindingDB (99) ☑ Wombat (64) ☑ DrugBank (6) ☑ PubChem (43) ☑ KiDb (54) ☑ STITCH (15) ☑ PharmGKB (17)

Activity values:

	Type	Gradient Min	Max	Statistics Count	Range
●	Ki	0	32500	332	0 → 32500
○	Potency	0.63	89.13	68	0.63 → 89.13
○	IC50	0	41000	116	0 → 41000
○	Ratio	3869	10338	6	3869 → 10338
○	Activity	134	148	2	134 → 148
○	Inhibition	0.2	89	14	0.2 → 89
	[other]			124	

ChemProt 2.0 Server - results

Technical University of Denmark

Results for the compound 0000549 | Table view | Download results

Compounds (25)

0000549 (1.00) | 1214768 (0.96) | 1215762 (0.94) | 1214769 (0.94) | 0496297 (0.90) | 121… | …(0.94) | 1920876 (0.90) | …690 (0.92) | 1215761 (0.88) | 1214770 (0.92) | 0504033 (0.87) | 0498563 (0.91) | 0495894 (0.85)

SMILES: N#Cc1ccc2c(c1)COC2(CCCN(C)C)c1ccc(cc1)F
Tanimoto coefficient: 1.00
(No further information)

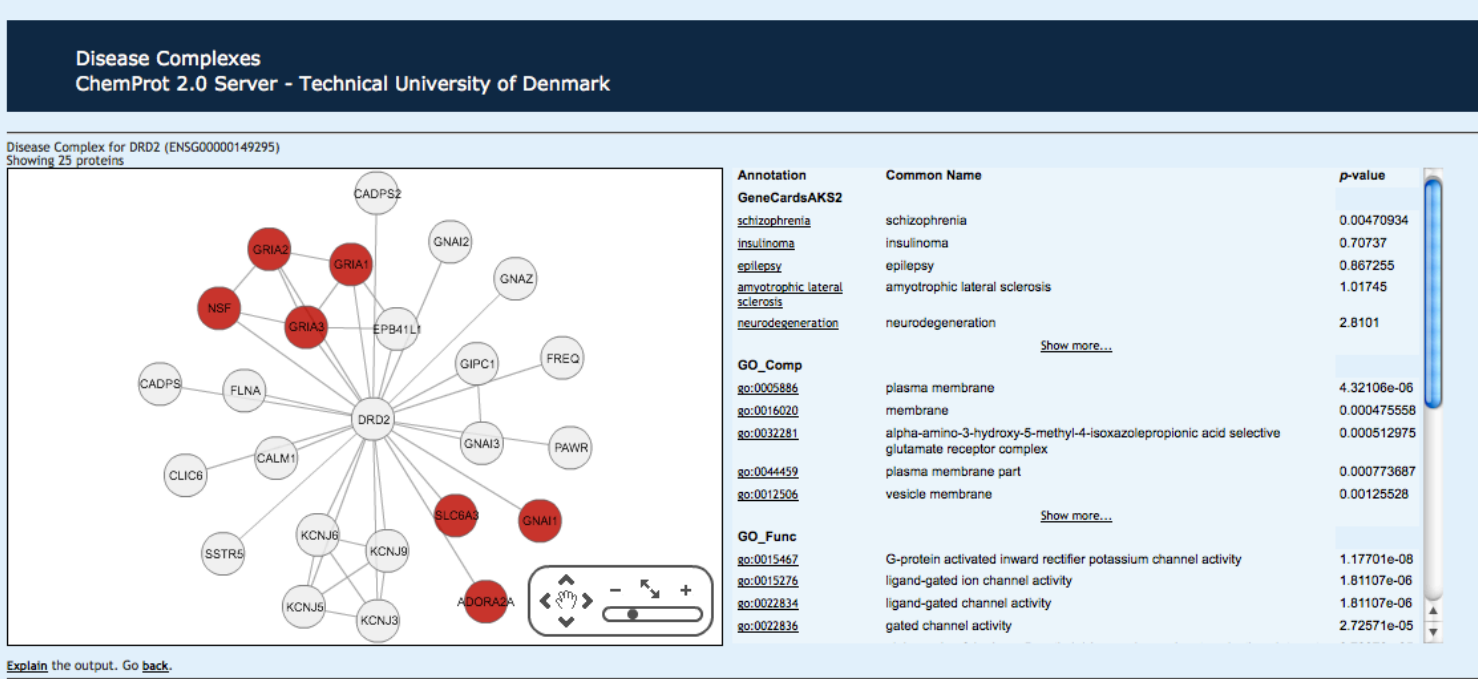
Disease Categories

Ear,Nose,Throat | Dermatological | Immunological | Ophthamological | Unclassified | Bone | Developmental | Metabolic | Psychiatric | multiple | Cancer | Endocrine | Muscular | Renal | Cardiovascular | Gastrointestinal | Neurological | Respiratory | Connective tissue d | Hematological | Nutritional | Skeletal

Proteins (187)

P25024	ENSG00000163464	C-X-C chemokine receptor typ	Human	Diseases
Q8VBU7	ENSG00000169252	Adrenergic receptor, beta 2	Rat	Diseases
Q16099	ENSG00000149403	Glutamate receptor, ionotropic	Human	-
P28335	ENSG00000147246	5-hydroxytryptamine receptor 2	Human	Diseases
P03372	ENSG00000091831	Estrogen receptor	Human	Diseases
P25025	ENSG00000180871	C-X-C chemokine receptor typ	Human	Diseases
P04150	ENSG00000113580	Glucocorticoid receptor	Human	Diseases
P51681	ENSG00000160791	C-C chemokine receptor type 5	Human	Diseases
Q9UBT6	ENSG00000122008	DNA polymerase kappa	Human	Diseases
Q9JIA2	NA	NA	Rat	-
P08575	ENSG00000081237	Receptor-type tyrosine-protein	Human	Diseases
P12530	NA	Arachidonate 15-lipoxygenase	Oryctolagus cun	-
P15127	NA	Insulin receptor	Rat	-
P25105	ENSG00000169403	Platelet-activating factor recept	Human	Diseases
P19838	ENSG00000109320	Nuclear factor NF-kappa-B p50	Human	Diseases
Q63380	NA	Transporter	Rat	-

Figure 2.3. Example of the graphical interface output based on a compound query. On the top, user can specify the query using the Display Settings. The heatmap on the left represent the bioactivities gathered for the input compound (in blue) and structurally similar compounds (in pink) in the X axis and the proteins in the Y axis. A color spectrum from blue (low) to red (high) is used to represent the activity. If several binding data have been measured for the same chemical-protein interaction, intensity of the colors is represented inside the circle. It is shown for example for the dopamine transporter (Q63380). The heatmap on the right describe the disease categories annotated to a protein. The value inside the circle represents the number of diseases associated to a protein.



Disease Complexes
ChemProt 2.0 Server - Technical University of Denmark

Disease Complex for DRD2 (ENSG00000149295)
Showing 25 proteins

Annotation	Common Name	p-value
GeneCardsAKS2		
schizophrenia	schizophrenia	0.00470934
insulinoma	insulinoma	0.70737
epilepsy	epilepsy	0.867255
amyotrophic lateral sclerosis	amyotrophic lateral sclerosis	1.01745
neurodegeneration	neurodegeneration	2.8101
	Show more...	
GO_Comp		
go:0005886	plasma membrane	4.32106e-06
go:0016020	membrane	0.000475558
go:0032281	alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid selective glutamate receptor complex	0.000512975
go:0044459	plasma membrane part	0.000773687
go:0012506	vesicle membrane	0.00125528
	Show more...	
GO_Func		
go:0015467	G-protein activated inward rectifier potassium channel activity	1.17701e-08
go:0015276	ligand-gated ion channel activity	1.81107e-06
go:0022834	ligand-gated channel activity	1.81107e-06
go:0022836	gated channel activity	2.72571e-05

Explain the output. Go back.

Figure 2.4. Example of the disease complexes network representation for the dopamine receptor D2 (DRD2). 25 proteins interact directly to the protein DRD2 and pointing the cursor to "Schizophrenia", 7 genes are associated to this disease.

and safer drugs. ChemProt-2.0 offers the possibility of interrogating multiple layers of information by linking chemically-induced biological perturbations to disease and phenotype. We believe with the advances in proteomics, metabolomics and other –omics sciences, combined with next generation sequencing technologies, we will no longer evaluate the bioactivity profile of a chemical solely at the molecular level, but rather we will investigate biomedical knowledge with the integration of genetic polymorphisms and clinical effects [139].

Chapter 3

Predicting side effects

In the previous chapter we described the importance of large data integration in relation to chemical biology. In this chapter and the two articles included herein, we demonstrate how these data types can be combined with side-effect information to explore and predict the relationship between side-effect, drugs, proteins and clinical outcomes. The following two articles describe two different approaches to explore these relationships.

In the first article we used text-mining techniques to extract as much information as possible from DailyMed records. Thereafter, we applied principal component analysis (PCA) to reduce the dimensionality of the drug–ADR (drug–adverse drug reaction) data followed by application of self-organizing map (SOM) in order to cluster the high-order ADR–drug interactions.

SOM is a type of artificial neural network (ANN), which projects the high dimensional data down to a low-dimensional (e.g. 2D) representation (map) of the input data. The method is an unsupervised learning method that tries to find hidden structures in unlabeled data without any prior knowledge of data grouping. It generates a 2D map of the input data space. The most popular method of displaying SOMs is through a unified distance matrix or U-matrix. It represents the maps as a grid of neurons, where the distance between the adjacent neurons is presented in different colors. Normally black and white colors are used, where a dark color between the neurons corresponds to a large distance and light color indicates close clustering [140].

The second article used the same ADR–drug source, DailyMed. However, only information regarding ADR–drug with placebo-controlled studies were extracted. A high-confidence drug–ADR dataset was created by calculating the significant association between drug and ADR taking into account placebo frequencies and number of cases and controls in groups.

The significance of the drug-ADR associations was calculated based on chi-squared test for a 2x2 contingency table. The chi-squared test gives accurate p-values provided that the number of expected observation is greater than 5. If this is not true, the Fisher's exact test should be used instead [141]. The drug-ADR dataset used in this paper is considered as "large samples" and therefore the chi-squared test can be used. Even though, the p-values calculated are only an approximation of the Fisher's exact test, the approximation provides more than enough confidence for our analysis.

In the first paper we concluded that, at least in part, side effect occurrences can be explained by drug compartmentalization, i.e. the drug is more likely to cause side effects in organ/tissue where it is more likely to accumulate. Based on this observation and from other studies [29, 142], we constructed a ADR-tissue dictionary based on tissue annotations from Human Protein Atlas and side effect terms from MedDRA classification in order to map side effect to tissue.

Paper III

Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing

Tudor I. Oprea[a,b,c], Sonny Kim Nielsen[b], Oleg Ursu[a,c], Jeremy J. Yang[a,c], Olivier Taboureau[b], Stephen L. Mathias[a,c], Irene Kouskoumvekaki[b], Larry A. Sklar[c], and Cristian G. Bologa[a,c],

Molecular Informatics , Volume 30, Issue 2-3, pages 100–111, March 14, 2011, DOI: 10.1002/minf.201100023

[a]Department of Biochemistry and Molecular Biology, Division of Biocomputing, University of New Mexico School of Medicine, Albuquerque, New Mexico 87131, USA [b]Department of Systems Biology DTU, Building 208, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, 2800, Denmark [c]UNM Center for Molecular Discovery, University of New Mexico School of Medicine MSC11 6145, Albuquerque, NM 87131, USA
TIO, SKN and OU contributed equally to this work.

Abstract

Finding new uses for old drugs is a strategy embraced by the pharmaceutical industry, with increasing participation from the academic sector. Drug repurposing efforts focus on identifying novel modes of action, but not in a systematic manner. With intensive data mining and curation, we aim to apply bioand cheminformatics tools using the DRUGS database, containing 3837 unique small molecules annotated on 1750 proteins. These are likely to serve as drug targets and antitargets (i.e., associated with side effects, SE). The academic community, the pharmaceutical sector and clinicians alike could benefit from an integrated, semantic-web compliant computer-aided drug repurposing (CADR) effort, one that would enable deep data mining of associations between approved drugs (D), targets (T), clinical outcomes (CO) and SE. We report preliminary results from text mining and multivariate statistics, based on 7684 approved drug labels, ADL (Dailymed) via text mining. From the ADL corresponding to 988 unique drugs, the "adverse reactions" section was mapped onto 174 SE, then clustered via principal component analysis into a 5x5 selforganizing map that was integrated into a Cytoscape network of SE-D-T-CO. This type of data can be used to streamline drug repurposing and may result in novel insights that can lead to the identification of novel drug actions.

3.1 Computer-Aided Drug Repurposing

The pharmaceutical industry is subject to an "innovation deficit" [143], which is often expressed as the widening gap between productivity (new molecular entities, NMEs, approved each year) and the annual R&D budget. The number of NMEs approved has declined from mid-40s in the early nineties [144], to under 15 in recent years. The price of drug innovation estimates [145] place the cost of a new drug anywhere between $500 million to over $2 billion, depending on the therapy area and the developing firm [146]. These trends indicate a sharp decline in research productivity across the entire pharmaceutical sector, with the exception of biologics. Currently, major pharmaceutical houses seek to increase short-term profitability via mergers and acquisitions, drastic reductions in research personnel and an increased outsourcing effort. It is therefore not surprising that the National Institutes of Health (NIH) is emerging as a leader not only in the arena of early drug discovery via MLI, the Molecular Libraries Initiative, [147] but also in the area of translational medicine via the Clinical and Translational Science Awards (CTSA) initiative [148]. Academic investigators are now more effective in "de-risking" compounds of industrial interest [149].

Examples

Genomics, pharmacovigilance and side-effects evaluation [150, 151], screening drug libraries against neglected diseases [152], data mining for drug side-effects [28] and finding novel targets using in silico tools [86] are equally valid strategies to identify novel uses for old drugs. The concept of drug repurposing [153] is not novel to the pharmaceutical industry: One of the oldest semi-synthetic drugs, acetylsalicylic acid, an anti-inflammatory drug formulated as 500-mg tablets launched in 1896 as Aspirin, was recently repositioned as daily-dose “baby Aspirin” (75-mg tablets in Europe, 81-mg tablets in the US), for cardiovascular disease prevention [154]. Pfizer combines cetirizine, a histamine H1 receptor antago-nist (approved in 1987 as Zyrtec) with pseudoephedrine (a sympathomimetic approved in 1975 as Novafed, now discontinued) as a new drug, “Zyrtec-D 12 hour” (launched in 2001), which contains cetirizine 5 mg and pseudoephedrine 120 mg per tablet, for the symptomatic relief of seasonal allergies [155]. Caffeine, a naturally-occurring CNS stimulant [156] used in combination with multiple API to increase alertness and diuresis, [157] was approved as “Cafcit ” (caffeine citrate, injection for intravenous administration) in 2000 by the U.S. Food and Drug Administration (FDA), and in 2007 by the European Medicines Agency (EMA) for the short-term treatment of apnea of prematurity in newborn infants between 28 and 33 weeks gestational age [158]. Other examples, including duloxetine and thalidomide, are reviewed elsewhere [153].

Critical Barriers

escribed in section 505(b)(2)[159] of the Federal Food, Drug, and Cosmetic Act, the process of drug repurposing is made possible by the Drug Price Competition and Patent Term Restoration Act of 1984 (also known as the Hatch-Waxman Amendments[160]), which enables the applicant for a new drug application (NDA) to reference investigations of safety and effectiveness where at least some of the information required for approval comes from studies not conducted by or for the applicant and for which the applicant has not obtained a right of reference. Section 505(b)(2) offers patent protection (hence market monopoly) for NMEs, new dosage forms (e.g., “baby Aspirin”), new administration routes (e.g., oral vs. intra-venous caffeine citrate), new indications, and for new NME combinations (e.g., Zyrtec-D). While the expectation is that fewer clinical studies are required for repositioning a drug, this has no impact when the drug is repurposed for a medical condition that previously lacked drug therapy. The burden is even higher when therapeutic agents already exist, i.e., the petitioner needs to prove the therapeutic advantage offered by repurposed drugs. Although the process is expected to last considerably less compared to an all-new NME effort [153], the applicant must nevertheless conduct clinical trials with respect to efficacy (e.g., for new indications), as well as safety (e.g., for higher doses). This financial burden blocks drug repurposing efforts because clinical research can quickly reach the multi-year, multi-million dollar range.

CTSA

Having recognized the gap between basic and clinical science, the NIH launched efforts to bridge the repurposing “valley of death” by fostering translational research via the CTSA (Clinical and Translational Science Awards) initiative [148]. CTSA has an online collection of research volunteers[161], an index of CTSA technologies and intellectual properties [162], as well as a portal partnering academics and pharmaceutical companies for no-longer developed molecules [163].

Viability

Against the backdrop of increased difficulties in taking NMEs into the clinic, one ought to consider the merits of “drug repurposing” (also termed drug repositioning, or drug re-profiling) as a viable option. First, our level of knowledge in the polypharmacology [85, 97] of drugs has reached a good degree of maturity [164], because of an increased in-depth profiling effort, in particular for novel drugs. More importantly, our level of knowledge, addressing data completeness gaps [87], is increasing for the older drugs as well: This is, to a large extent, due to the availability of screening data in public sources such as PubChem [165] for out-of-patent drugs, as incorporated for example in the Prestwick Chemical Library [166].

Approach

We envision a computer-aided drug repurposing (CADR) platform as being a semantic-web service that would rely on factual associations between drugs, targets and clinical outcomes. The CADR platform would provide in-depth integration for these four categories :

- drugs (D), i.e., the active pharmaceutical ingredients (API) and their active metabolites, with initial focus on small molecule APIs;
- targets (T), macromolecules perturbed by API that lead to a clinical outcome;
- positive clinical outcomes (CO), i.e., the intended therapeutic effects of drugs, as specified on the approved drug labels (ADL) under “Indications”

- negative clinical outcomes, often referred to as "adverse events" or drug side effects, SE.

To establish such factual associations, a two-pronged approach is needed: (i) deep data mining of D-T interactions, including indexing, cross-referencing, processing and curation of the molecular, pharmacological and biochemical aspects of drug-target interactions; and (ii) text mining of ADL and clinical research documents using controlled vocabularies, which would be used to extensively process the "adverse events" and "indications" sections of medical package inserts or on-line repositories such as DailyMed[167]. This approach is conceptually built on prior work, which inferred novel drug targets starting from a combination of chemical and phenotypic side-effect similarities[28].

Potential

The CADR platform, while built upon open-access resources such as DrugBank[168] and DailyMed, can provide improvements in two directions: (i) Semantic web[169] compliance, which aims to provide structured drug-related information (D-T-CO and D-T-SE relationships), with associated sets of inference rules in the form of RDF (Resource Description Framework) triples that computers will use to conduct automated reasoning; and (ii) systematic mapping of SE (at the symptom level wherever possible) with targets and antitargets [170], that would overlap symptoms related to unmet clinical needs (e.g., for rare and neglected diseases[149]) with SE and CO relationships. As it increases its coverage, the CADR platform may lead to systemic analyses of both clinical and basic science data, and may reduce the impact of the accidental discovery (i.e., serendipity). This prospective review describes preliminary steps taken towards assembling the requisite elements for a viable CADR platform: First, we address efforts in developing an exhaustive knowledge base for D-T interactions. Then we discuss preliminary results based from SE data modeling, as extracted from ADL. Finally, network-based associations between D-T pairs and clinical outcomes are evaluated from the perspective of side-effect inter-relationships.

3.2 Data Collection and Analysis

Small Molecule API Interaction Annotations (D-T)

With the final goal being data completeness [87], we identified and curated information from multiple databases referring to API in order to create a comprehensive repository of drugs, beginning with small molecules. The focus of the DRUGS database is to capture and integrate target bioactivity information for all small molecule drugs, i.e., unique API that have obtained marketed drug status for human use, regardless of country of approval. In its current form, DRUGS has 3837 unique chemicals (December 2010). DRUGS was primarily built using data from WOMBAT-PK,[32] PDSP[171] and DrugBank, with additional information collected from publications.[172, 173] DRUGS stores accurate chemical structures which were independently verified across several sources including ADL and SciFinder [174], generic names and common synonyms. Chemical structures were subject to standardization (salt removal, charge neutralization, and aromatization) prior to identity searching. For unique targets, we used UniProt [48] identifiers and ontologies to uniquely map the proteins in DRUGS, observing compatibility with the disease chemical biology approach, ChemProt.[175] As of December 2010, DRUGS had 3837 unique API, 19 593 D-T interactions, and 1750 unique targets.

Numerical Values

However, not all of the D-T are ascertained to be clinically relevant, nor have the D-T values been validated for having an affinity that is linked to a clinical outcome. Because massive amounts of D-T interaction data are sometimes available from on-line resources such as IUPHAR-DB[176, 177], ChEMBL[178], PDSP and PubChem, in particular for older drugs, this may result in a plurality of information that is sometimes contradictory : e.g. , the same API is indexed on the same target from the same species, but numerical values differ by 2 orders of magnitude or more. Furthermore, numerical values attributed to biological activities are also subject to “temporal drift”: For example, in the 1960s propranolol had an affinity of 31 nM for the betaadrenergic receptor[179] (only one was known), but is now annotated with affinities of 2 nM, 5 nM and 600 nM for the b1, b2 and b3 adrenergic receptors, respectively [176], in addition to the serotonergic 5-HT1A receptor (30 nM) [32]. Both accuracy of detection and our understanding of targets improve with time.

Therefore, we have implemented a process to eliminate duplicated D-T-bioactivity pairs, giving higher priority to expert-curated (e.g., IUPHAR-DB or “PDSP certified”) data wherever possible. Median values for bioactivity data (excluding highest and lowest value) were used wherever 5 or more values for the same biological end-point were available. A confidence score (e.g., 1.0 for highly trusted sources, and 0.5 for lack of numerical data) was implemented. For the purpose of this report, the DRUGS database relied on data from IUPHAR-DB, PDSP and WOMBAT-PK, which were converted into a format amenable for further processing via in-house data conversion and cheminformatics tools (i.e. , JChem[180] and OpenEye[181] software).

Visual Mapping

Earlier efforts resulted in the integration of 739 molecules (out of 2300) from IUPHAR-DB into iPHACE [94, 182], a webbased tool built around in extenso pharmacological annotations from IUPHAR-DB and PDSP that has the capability to visualize the interaction on many drugs on many targets (Figure 3.1). The IUPHAR website has linked individual records back into iPHACE [183]. This technology, currently extended to DRUGS, will help us evaluate the complexity of data parsing and extraction as well as the degree of automation achievable for future updates.

Approved Drug Labels Availability

There are currently a number of on-line resources that contain relevant information related to package inserts and ADL: DailyMed and the (related) U.S. FDA [184], the EMA [185], the World Health Organization, WHO [186], the Australian Therapeutic Goods Authority, TGA[187] – which are open access, as well as the for-fee resources Physician Desk Reference, PDR [188], Martindale,[189] and the American Hospital Formulary Service (AHFS) [190], among others. The process of deep data mining for ADL starts with data capture and mapping for API indexed in DRUGS. Priority is given to "clinical pharmacology", "indications and usage", "contraindications", "adverse reactions" and "description". However, as other sections may contain pertinent information, they are typically stored (unprocessed) for later use. Particular care needs to be given to standardize, catalog and process clinical outcomes and drug side effects via extensively annotated vocabularies. Compatibility with the side-effect resource (SIDER)[99] is likely to build on SE frequency for the available D-SE pairs. The CADR platform is likely to take advantage of this D-SE mapping to enable inferences combining drug and target information over an enormous, presently sparsely mapped space of drug-target-clinical outcome assertions that would not otherwise be possible.

ADL Text Mining

We processed all Dailymed (XML format) records (May 2010 version) in order to evaluate (i) the number of unique drugs present and (ii) the relationship between these drugs and SE. Dailymed entities contain a surprising number of API duplicates, e.g., over 90 drug entries contain "estradiol". We performed de-duplication in order to simplify, structure and streamline this dataset. We flagged duplicates both at the API level, e.g., where API names are identical, and at the generic drug name level. De-duplication by active moiety can be illustrated for gentamicin: while its correct chemical name is "gentamicin C1 sulfate", the following synonyms were identified in DailyMed: "Gentamicin

Sulfate in Sodium Chloride", "Gentamicin Sulfate in Sodium Chloride Injection", "GENTAK", "Isotonic Gentamicin Sulfate", "Gentamicin Sulfate" and "GENTAMICIN SULFATE". These trade names are listed in DailyMed as separate drugs (which they are), but cannot be easily mapped onto a unique API.

Our two-step procedure reduced the dataset from 7684 to 1768 unique entities, or 77% reduction. We further removed 261 animal drug products as well as allergenic therapeutics lacking specific chemical API information (e.g. "Cat pelt" and other animal extracts). This process yielded 1329 entries. After manual curation, the dataset was found to include 1021 small molecules, of which 20 were duplicates not detected via automation (e.g., "acetate hydrocortisone" and "hydrocortisone acetate"); 243 small molecule mixtures (e.g., simvastatin and niacin), 3 undefined mixtures (omega-3-acid ethyl esters, perflutren and sinecatechins), 28 proteins, 26 monoclonal antibodies, three non-drugs with therapeutic use (sodium acetate, sodium bicarbonate and tromethamine), one parasite extract (Trichophyton) and one insect extract (Sitotroga), respectively.

The final XML files for 988 small molecule drugs (which is what DailyMed contains) were processed with the Python xml.dom.minidom package for SE word frequency and association using the text mining TM package from the R statistical software[191]. Term-frequency vectors, term-document matrices, and distance matrices were generated and used to analyze SE similarity and groupings. In particular, we subjected the frequency matrix containing 174 SE columns for 988 rows (drugs) to PCA, principal component analysis PCA[192] using the Simca package.[193] Data were then visualized using a self-organizing map[140] via Spotfire [194]. Each SE was manually associated with a specific tissue or organ, wherever possible (see also Figure 3.2).

Associating Drugs, Targets and Clinical Outcomes

The D-SE sparse matrix (29263 SE occurrences, or 16.81 % occupancy) yielded a 10-dimensional model (missing data were attributed a 0 value). The cumulative fraction of the variation of the X variables explained by the 10-PC model, $R^2VX(cum) = 0.365$, with a cross-validated cumulative predicted fraction of the variation of the X variables, $Q^2VX(cum) = 0.171$. Additional principal components produced eigen values under the 5% tolerance limit and were therefore not considered. Although clearly incomplete in terms of SE coverage, we wanted to examine the biomedi-cal relevance of the potential D-SE associations uncovered by this model. The PCA model is graphically summarized in Figure 3.2, color-coded by tissue or organ: 14 such categories, plus "systemic" were added to the set, but not used in the PCA model.

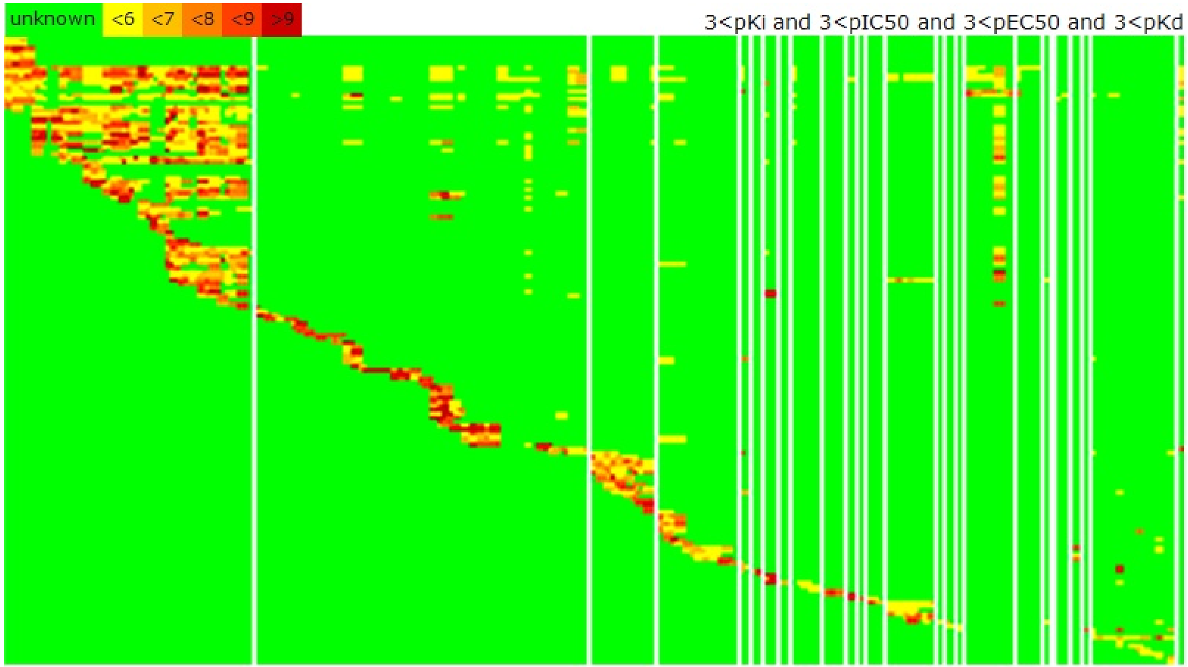


Figure 3.1. Visual summary of the IUPHAR/PDSP databases, showing 4033 interactions for 736 drugs (rows) and 178 targets (columns). The heatmap is color coded: bioactivity is higher from yellow to dark red; green indicates absent data. Targets are clustered by family (e.g., Gprotein coupled Class A amine receptors are in the far left column). The density of bioactivity in the top left corner reflects the promiscuity of this class of receptors, which includes dopaminergic, muscarinic and serotonergic receptors.

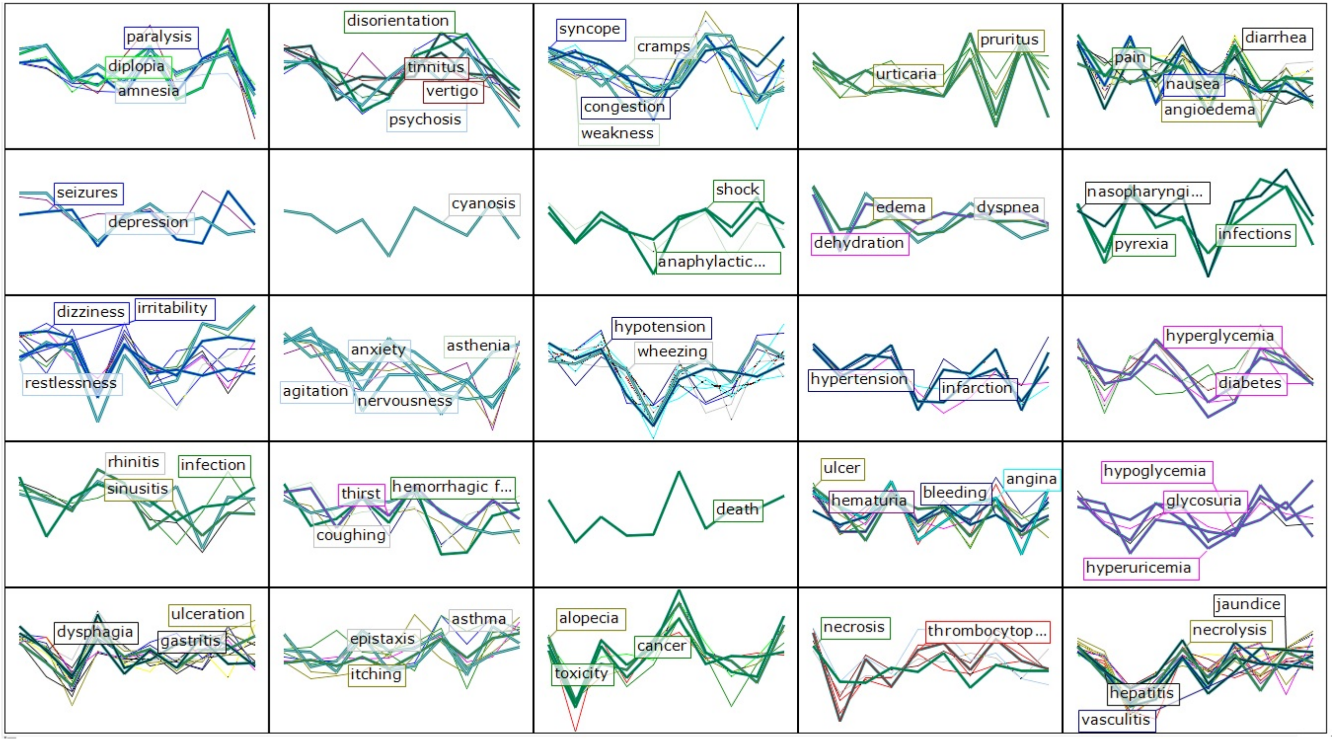


Figure 3.2. Self-organizing map (SOM) for 174 SE, based on the 10-PCA model derived from frequency of occurrence in DailyMed ADL from 988 small molecule drugs; the SOM clustering was mapped onto 25 cells. Colors encode tissue information; most frequent are skin-mucosa (brown; 27), CNS (azure blue; 17), digestive tract (black; 15), “metabolic imbalances” (magenta; 15), respiratory (pink, 13) and vascular (dark purple; 11). SE labels are representative of each cluster.

The emerging clusters indicate that ADL co-occurrence is far from random. For example, rhinitis, sinusitis and infection are related (column 1, row 4, or “cluster_1_4”); ulcer, hematuria, angina and bleeding (cluster 4_4) are close to “death” (the singleton cluster_3_4), which in turn neighbors alopecia, cancer and toxicity (cluster_3_5). Yet other members of cluster_3_5, e.g., stomatitis, fever and lacrimation, relate to some members of cluster_2_5, such as itching, connectivitis, eruptions and itching, or to the more severe asthma and allergic. Toxic and immunotoxic reactions that manifest on the dermal and mucosal layers or in the digestive tract include nausea, diarrhea, vomiting, angioedema, rashes, dyspepsia and flatulence, and are co-clustered with joint-and-muscular pain symptoms such as arthralgia, myal-gia, headache, pain, as well as cough (cluster_5_1). Cardiovascular and respiratory SE associations include arrhythmia, bradycardia, tachycardia, fibrillation, hypotension and phlebitis, and bronchospasm, wheezing and apnea, respectively, as well as sedation (cluster_3_3). All eight blood-located SE are in the bottom row (except thrombosis, row 4), whereas most of the CNS and mental SE are on the top left part of this SOM.

Two of the three ophthalmic SE, diplopia (double vision) and photophobia respectively, but not lacrimation, co-occur with CNS and mental SE, namely coma, paralysis, convulsions, amnesia, confusion and ataxia (cluster_1_1). While clinically associated with the eye organ, diplopia and photophobia are not really ophthalmic dysfunctions; rather, it is our perception that is altered; therefore, their clustering association to the CNS/mental neighborhood is quite appropriate. Though based on a limited data set, we conclude that, at least in part, SE occurrences can be explained by drug compartmentalization, i.e., the drug is more likely to cause side effects in the organ/tissue where it is more likely to accumulate. This result, albeit intuitive, is quite surprising, since it stems from a generic text analytics tool that lacks medical context. It matches observations from co-occurrence pharmacovigilance processing of electronic health records for seven drugs [195].

Limitations of the PCA Model

For all its potential merit, this SE-based PCA model is by no means directly usable within the CADR platform: First, automated text mining yielded a rather limited (174) set of adverse reactions, which is significantly smaller than the side effects from SIDER [99]. Second, the PCA model covers only 36.5 % of the relationship between these adverse events and the drugs included in this model, based on an already sparse matrix. Finally, the relationship between drugs and side effects depends on dosage, which requires more in-depth analysis of ADL data. Some of these aspects (SE incidence, dosage and relative risk) were detailed elsewhere [76]. Despite its limitations,

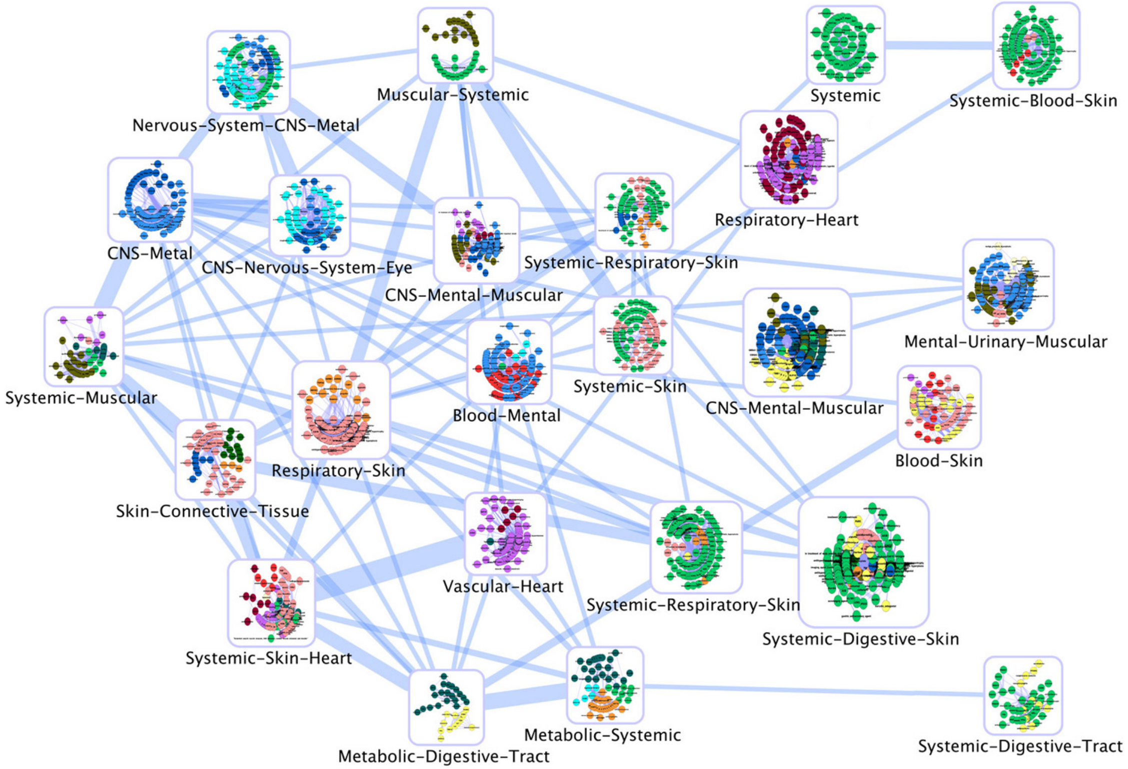


Figure 3.3. The SE-based D-T-CO network, showing the inter-dependence between drugs, targets and clinical outcomes. Color codes are as follows: Blood, red; CNS, blue; connective tissue, green; digestive tract, yellow; eye, medium purple; heart, dark red; mental, light blue; metabolic, dark cyan; muscular, olive; nervous system, cyan; respiratory, orange; skin and mucosa, pink; systemic, lime green; urinary, light yellow; vascular, magenta. Edge thickness in this network is based on the 10 dimensional PC model, where the centroids of the 25 SOM clusters were used to calculate the Euclidean distances between clusters. This was further projected on two dimensions to map the relative position among clusters.

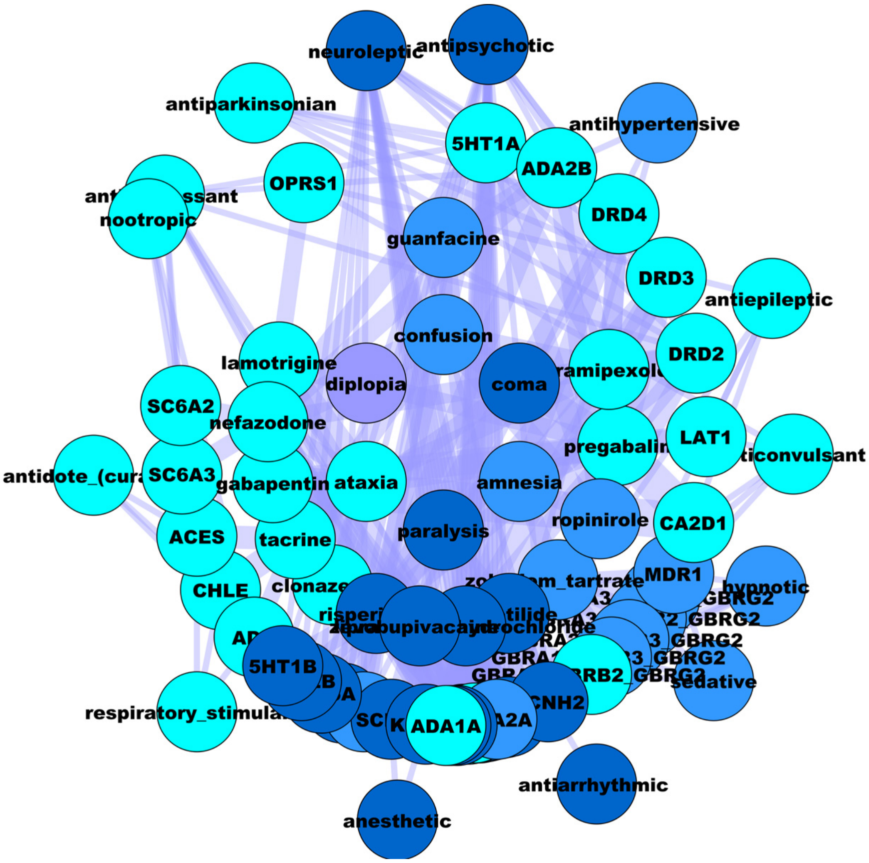


Figure 3.4. The mental-CNS-deficiency network based on SE (inner layer), the associated drugs (inner middle layer), their confirmed targets (outer middle layer), and intended clinical outcomes (outer layer). Edge thickness in this network is proportional with the strength of the DT interaction. Color codes are discussed in Figure 3.3.

the SE drug matrix allows us to conclude that text occurrences from the ADL “adverse reactions” section co-emerge in clusters by (possible) mechanism of action and topicality (i.e., organ or tissue where the effect occurs).

Exploring the SE-D-T-CO Relationship

The availability of tools that enable biomedical data visualization such as Cytoscape,[196, 197] can be used to associate D-SE data with target information via biological network analyses. Through its plug-in architecture, we extended Cytoscape to display and analyze SE-D-T-CO networks for small

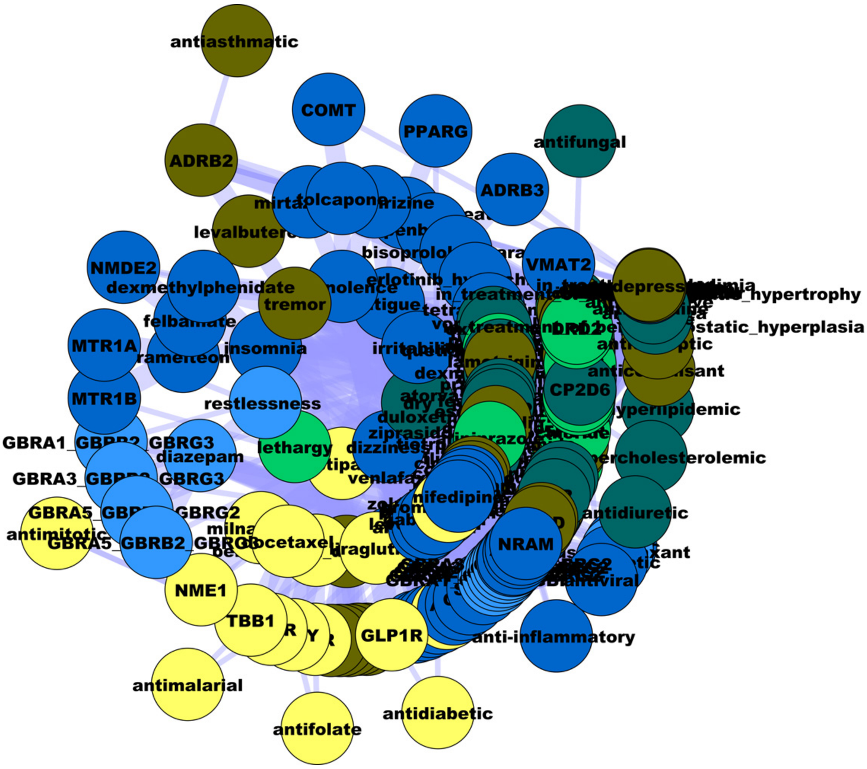


Figure 3.5. The CNS-mental-disorder network, which includes non-CNS acting drugs in addition to CNS drugs. Layer position and edges are similar to Figure 3.4. Color codes are discussed in Figure 3.3.

molecule API. Integrating the results from the ADL-based D-SE PCA, we generated a comprehensive network, where each D-SE cluster (shown in Figure 3.2) was related to the other clusters based on the inter-cluster relationship given by p-scores (from PCA), and by using available D-T-CO information from WOMBAT-PK. This Cytoscape visualization was intentionally designed to maintain the organ-based topicality observed earlier, while at the same time highlighting the possible associations between often-unrelated (at least from a clinical standpoint) drugs and their modes of action.

Knowledge mining of D-T-CO data requires controlled and structured information, where drugs and targets can be nouns, their relationship can be described as immediate interactions, and pharmacokinetics and pharmacodynamics can be described via more comprehensive associations (i.e., clinical outcomes). The complexity of this interdependence is conceptualized in Figure 3.3 Each node of this Cytoscape plot is a network in itself, and nodes are maintained separated for visual clarity. The network visualized here is based on 307 drugs, which were selected based on their high affinity (better than 1 mM) for each of the targets within the network node. Two of these nodes are shown in Figures 3.4 and 3.5, and discussed further. Cytoscape session files, complete with network images for each node given in Figure 3.5, as well as instructions on how to upload them in Cytoscape, are available as Supporting Information. The complexity of these relationships is further illustrated in Figure 3.4, which shows the SE-D-T-CO network based on Cluster_1_1, focused on drugs associated with diplopia (e.g. , lamotrigine, zonisamide, phenythoin, pregabalin and topiramate). This network primarily includes drugs and targets associated with anticonvulsant, antiparkinsonian and nootropic activities, further supporting the earlier observation that drug-induced diplopia is not an eye-related dysfunction.

Another CNS-related association, given in Figure 3.5, shows the SE-D-T-CO network based on Cluster_3_1, which includes CNS-related SE for non-CNS drugs (anti-inflammatory, antidiuretic, immunosuppressant, anti-allergic, etc.). This network associates insomnia, dizziness and somnolence with tremor, restlessness, agitation and nervousness, respectively. Although having opposite clinical meaning, it is likely that these SE derive from interactions with the same targets, and that non-CNS drugs penetrate (at least to some extent) the blood-brain barrier.

3.3 Conclusions

In this report we illustrated some of the inherent difficulties in developing the required elements for a viable CADR platform. These steps are necessary, but not sufficient: First, we discussed our efforts in developing a comprehensive evidence-based system for D-T interactions; the DRUGS database attempts to collect public knowledge detailing biochemical and pharmacological interactions between drugs and (potential) targets. Then we discussed our first foray into automated text mining for side effects, one that strictly looks at word associations; these results offer some insight, supported by our 10-PC model and the SE associations.

Although in its early stages, the CADR platform illustrates how knowledge can evolve given deep data mining and tight integration. We showed that content related to clinical (e.g. , DailyMed) and chemogenomic data (e.g. , DRUGS) can be seamlessly processed and evaluated. Our preliminary PCA model mapped 988 unique drug ADL from DailyMed onto 174 SE. We concluded that adverse reactions can be explained by compartmentalization, i.e., the drug is more likely to cause side effects in the organ/tissue where it accumulates. Carefully associated DT-CO and DT-SE networks are likely to morph into RDF-based knowledge mining, perhaps via Cytoscape. These

RDF triples can further lead to computer assertions, i.e., computer-aided drug repurposing. This may be accomplished via Chem2Bio2RDF, a semantic framework developed for linking drug-target information,[198] or perhaps via deep knowledge mining processing systems [199]. Even at this preliminary stage, we have showed how D-T pairs and clinical outcomes can be associated within a recursive network-of-networks system. Such recursive system flexibility is likely to be required within the RDF framework itself: the "DT" triple ("drug A inhibits target X") is in itself the subject of another triple, "A-inhibiting-X" causes "CO/SE", which is probably the RDF equivalent of a phrase.

When developing DT-CO associations, evidence-based examples, where drug "A" binds to targets $X_1...X_n$ resulting in clinical outcomes $CO_1...CO_m$ (this is rarely a $1 : 1$ relationship) will need to be given priority, since this allows computer assertions with relative ease. STITCH,[200, 201] an online tool for the exploration of biological networks, does exactly that, i.e., it ranks D-T interactions based on scoring literature co-occurrence data starting with chemical-protein interactions. The ChemProt server[132] can also serve as D-T validation tool. However, many of these relationships are likely to require a certain degree of manual intervention. Tissue location, the presence of active metabolites and additional information related to CO and SE needs to be used for complex cases.

These are likely to result in novel insights that may lead to the identification and assertion of novel "off-target" or "off-label" drug actions. As knowledge bases asymptotically approach completeness, the CADR platform will become more amenable to deep knowledge mining and systemic analyses, integrating basic and translational science with clinical data, which may reduce the impact of the accidental discovery. It will provide to the scientific community, basic scientists and clinicians alike, a new tool to map the clinical, biological and medicinal chemistry space for small molecule drugs, effectively bridging often separate knowledge domains in a multi-disciplinary manner.

Are the factual associations assembled via the CADR platform enough to build a strong case for drug repurposing? With the expectation that it could automatically lead to an NDA, the answer is most likely negative. Such a system could rank with higher priority those cases that are more likely to result in clinically beneficial applications. However, the CADR platform is unlikely to serve as an automated drug repurposing tool in the immediate future. The plethora of DT-CO and DT-SE associations can be mined via automated reasoning, which will narrow down the search space. Yet, humans will remain center stage: Toxicity, efficacy and dosage, as well as alternate therapies (e.g., surgery) are likely to require individual decisions.

Acknowledgements

This work was supported, in part, by NIH Grants 1R21GM095952-01 and 5U54MH084690-03 (TIO, OU, JJY, SLM, LAS, CGB), by the Lundbeck Foundation Grant R32A3932 (SKN) and by the Villum Foundation CDSB (TIO). TIO, SKN and OU contributed equally to this work. Supporting Information, including the Cytoscape session used to create Figure 3.5, is available.

Paper IV

A systems chemical biology approach for the prediction of drug side effects

Sonny Kim Kjærulff[a], Tudor I. Oprea[a,b], Olivier Taboureau[a*], and Irene Kouskoumvekaki[a*]

[a]Department of Systems Biology DTU, Building 208, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, 2800, Denmark [b]Department of Biochemistry and Molecular Biology, Division of Biocomputing, University of New Mexico School of Medicine, Albuquerque, New Mexico 87131, USA

Abstract

Adverse drug reactions (ADRs) are a major problem both for patients and the pharmaceutical industry. An understanding of the underlying mechanisms behind these will, therefore, benefit both. Here, we present a novel method for associating ADRs to drugs and their protein targets. First, we created a high-confidence drug–ADR dataset by calculating significance of the associations between drugs and ADRs based on placebo-controlled studies stored in DailyMed. From this, we determined that only approximately 20% of the drug–ADR associations in DailyMed are significant. Subsequently, we linked drugs with their biological targets through ChemProt-2.0 database and applied a scoring function to capture frequently encountered ADR–target associations. Based on these associations, we developed a model for the prediction of ADRs for drugs and drug candidates. We validated the model on a set of 133 drugs from the SIDER 2 database and successfully predicted the highest ADR frequencies for 55% of the dataset. Further validation was performed on withdrawn drugs stored in DrugBank and we were able to confirm many predictions through the published literature. Our work demonstrates the importance of using high-confidence drug–ADR data in the development of methods that aim to elucidate, at the molecular level, the emergence of ADRs and in prediction of ADRs for existing and future drugs.

3.4 Introduction

The occurrence of adverse drug reactions (ADRs) is a huge problem for both pharmaceutical companies and patients, with an estimated cost of several billion dollars every year [202, 203]. As high as 30% of all drug candidates fail during clinical trails due to toxicity and other unwanted drug reactions [204]. Understanding the underlying mechanisms behind these unwanted drug reactions will help reduce the costs for both pharmaceutical companies and patients alike. As many drugs interact with multiple proteins and thereby perturbe the protein networks [128], there is a need for system-wide approaches to capture the effects that the drugs have on the system.

In recent years, researchers have proposed a variety of methods for linking side effects to drug actions. The commonly used approach is based on correlating chemical structure information with side effects using different frameworks [205, 206, 207, 208]. These methods have proven successful to some extent; these, however, lack the biological interpretation of side effect emergence, due to their sole reliance on chemical similarity. Unrelated chemical structures have also been shown to share similar side effects by sharing the same off-targets [209]. Campillos et al., [28] proposed a method to use side effect similarity to predict drug pairs with common protein targets. By adopting a systems biology approach, Fliri et al., [210] showed that drugs with similar bioactivity profiles tend to cause similar side effects. In a more recent study, Bauer–Mehren et al., [29] integrated data from multiple sources, however the causality relationship between the drug and the ADR can only be judged by experts after reading the full text article in order to determine the correctness and significance of the associations.

The most common method to represent the ADR profile of a drug is to use a binary association matrix, where "1" denotes presence and "0" denotes absence of an ADR term. Representing ADR–drug interactions as a binary association matrix together with machine learning techniques has proven successful in many studies [208, 211, 45, 44]. Campillos et al., [28] took into account that side effects vary greatly in their abundance and are not always independent of each other, and based on this, they weighted the side effects to identify drug-targets based on side effect similarity. Lounkine et al., [46] developed an enrichment score that associated targets with ADRs based on the likelihood of the target–ADR pairs co-occuring as compared to random and then applied the SEA method [20] to predict new ADR–targets. Garcia-Serna and Mestres [76] assigned a strength score between drug and ADR depending on the reporting frequency among the five ADR sources used in the study, where "1" denoted presence of ADR in all sources and "0.2" denoted presence in only one source.

Detection of significant drug–ADR associations in large datasets that include adverse event reports (e.g., from postmarketing surveillance) is prone to factors such as the reporting rate of ADRs or subject selection bias [212]. Tatonetti et al., [212] therefore, proposed an adaptive data-driven approach to correct the different factors in cases where the covariates are either not measured or are unknown.

In this article, we present a novel method for associating ADRs to proteins and for the prediction of potential ADRs based on their targets. This method uses statistically significant drug–ADR associations derived from placebo-controlled drug trials [167] and drug targets from ChemProt 2.0 database [213]. We developed a scoring scheme to weigh the interactions between ADR–drugs and drug–protein targets in order to associate ADRs to proteins. We validated the method using drug–ADR pairs from SIDER 2 [99]. A set of withdrawn drugs from DrugBank [23] were used to further validate the method.

3.5 Results

We created a high-confidence drug–ADR dataset using a novel methodology. The workflow of our strategy is presented in Figure 3.6.

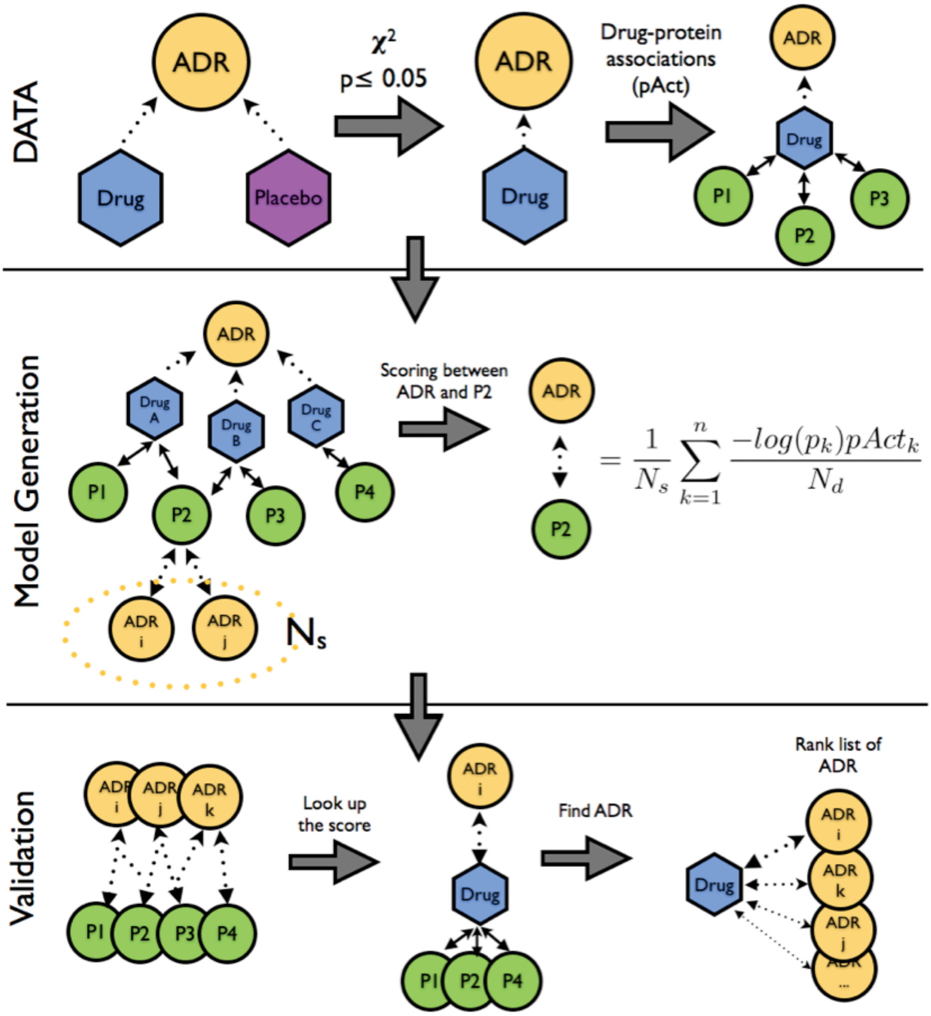


Figure 3.6. Workflow for dataset development, model generation and validation. DATA: Drug – side-effects data were extracted from DailyMed®. Only χ^2 significant (p-value < 0.05) drug–ADR associations were used further. Drug–target association data were integrated from ChemProt-2.0 database with an activity filter of pAct > 4 (where pAct is the drug-target activity value). The target proteins are labeled P1– P3. MODEL GENERATION: An ADR is associated with Drugs A, B and C which bind proteins P1–P4. Protein P2 is indirectly associated to ADR_i and ADR_j via interactions with other drugs (not shown in the figure). The score between this ADR and P2 is calculated as shown in the formula, where P_k is the p-value of the ADR–Drug A association, pAct is the drug–target activity value, N_d is the number of targets Drug A and Drug B share, and Ns is the total number of ADRs connected to P2. VALIDATION: For a given protein a list of ADRs and an association score are obtained. For a given drug, its targets are looked up in the list, the ADR score is multiplied with the bioactivity and the product is summed for each ADR. Lastly, the ADRs are ranked according to the final score.

Threshold p-value	**Number of drugs**	**Number of ADRs**	**Number of Drug-ADRs pairs**	**Percent %**
1	183	615	4515	100 %
<0.05	142	303	908	20 %
placebo (<0.05)	34	53	78	2 %

Table 3.1. The number of drugs, ADRs, and drug-ADR pairs with different levels of significance. The first row summarizes all the interactions. The second row shows the significant drug-ADR association with a p-value $<$ 0.05. The last row shows special cases where the placebo frequencies are greater than the drug frequencies and thus significant with a p-value < 0.05.

3.6 Dataset

Drug-ADR associations

183 drugs with 615 ADRs were extracted from DailyMed which contains information on placebo-controlled trials. Applying χ^2 statistics on the 2x2 contingency table for drug and placebo allowed us to calculate the statistical significance of the drug–ADR associations (see Materials and Methods section for details). Applying the significance level of < 0.05 to the dataset as a filter, we were able to reject 80% of the total 4515 drug-ADR pairs, which reduced the number of ADRs from 615 to 303 and the number of drugs from 183 to 142 (Table 3.1)

Quite surprisingly we also identified that 2% of all associations were significant ADR-placebo associations (Table 3.1), which means that in these cases the placebo caused more ADRs compared to the drug. We determined 78 significant ADR–placebo associations from 34 drugs and 53 ADRs. The most frequent side effects from these interactions were headache, somnolence, nausea and fatigue. Table 3.2 shows some examples of filtering associations by p-value where the placebo resulted in higher incidence of ADRs than the drug itself. For example, use of placebo showed a significantly higher incidence of cardiac chest pain as an ADR as compared to the drug Pravastatin, a cholesterol-lowering agent used for hypercholesterolemia to reduce the risk of cardiovascular disease. Cardiac chest pain is a pharmacologically indicated ADR for Pravastatin. Similar was the case with Terazosin, a drug used for treatment of benign prostatic hyperplasia (BPH), an ADR of which can increase the risk of urinary tract infections. We noted again that there was a significantly higher incidence of urinary tract infections in people taking placebo as compared to those taking Terazosin. Such significant ADR–placebo associations indicated that filtering by p-value using drug–placebo information

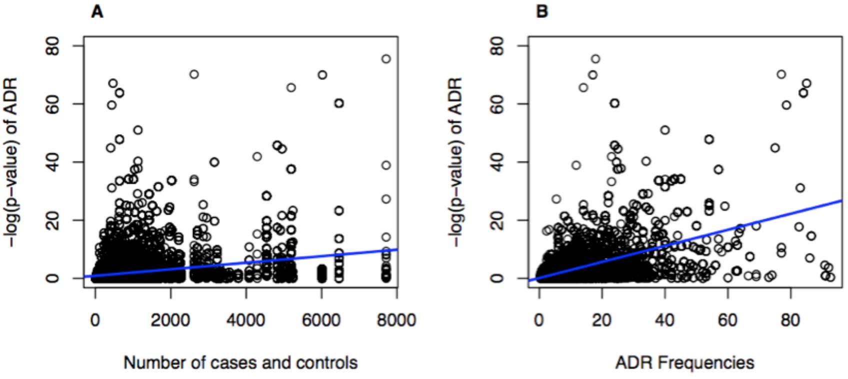


Figure 3.7. A. Negative logarithm of p-value vs. the total number of cases and controls. The linear correlation is 0.09. B. Negative logarithm of p-value vs. the frequencies of the ADRs. The graph shows a correlation of 0.32.

resulted in meaningful associations. This highlights the importance of taking into account placebo effects.

It was important to rule out any bias due to an unequal number of participants in each drug-placebo trial. We therefore plotted the p-value of ADR against the total number of patients in the trials (Figure 3.7A). The plot shows the linear correlation coefficient to be only 0.09, which indicated that the size of the trials had not biased our model. A positive correlation coefficient of 0.32 is obtained between the ADR frequencies and the p-values (Figure 3.7B), which indicated that the higher frequencies, to some extent, lead to more significant p-values.

Table 3.3 shows the correlation between the top 15 ADR frequencies for a drug and the significance of the ADRs with a threshold of p-value < 0.05. For example, the highest ADR frequency was found to be significant with a p-value < 0.05 in 88 of 183 (48%) drugs. The second highest ADR frequency was found to be significant in 40% of all drugs. There was a clear indication that the highest frequencies were also significant for each drug. The significance dropped below 20% at the 8th highest frequency, which is the same significance as the overall interactions (with p-value < 0.05) calculated in Table 3.1.

DrugBank ID	**Drug name**	**ADR**	**Indication**	**p-value**
DB00175	Pravastatin	Cardiac chest pain	Hypercholesterolemia	1.6E-07
DB01162	Terazosin	Urinary tract infection	Benign Prostatic Hyperplasia (BPH)	7.8E-03
DB00177	Valsartan	Dizziness	Hypertension	3.2E-34
DB00332	Ipratropium bromide	Upper respiratory tract infection	Bronchospasm	4.0E-02
DB00472	Fluoxetine	Anxiety nervousness	An antidepressant agent	6.3E-04 3.2E-02
DB00482	Celecoxib	Headache	Pain	2.5E-03
DB00590	Doxazosin	Hypotension micturition frequency	Mild to moderate hypertension and urinary obstruction symptoms caused by BPH.	2.5E-02

Table 3.2. Examples of ADRs significantly associated with the placebo treatment

Top ranking ADR frequencies	**Number of significant drugs (p-value < 0.05)**	**Percent %**
1	88	48
2	74	40
3	60	33
4	54	30
5	49	27
6	46	25
7	40	22
8	35	19
9	28	15
10	21	11
11	21	11
12	27	15
13	23	13
14	25	14
15	21	11

Table 3.3. Rank list of ADR frequencies ordered by highest for each drug. A total of 183 drugs. The ranking ADR frequency within each drug is significant with a p-value < 0.05 in 88 of 183 drug (48%). The second highest ADR frequency within each drug is significant in 40% of all drugs.

Drug-target associations

We integrated data from 8 different databases in ChemProt and enriched the number of drug-targets based on drug–target annotations to a total of 6494 associations with an activity $< 100\ \mu M$. Figure 3.8 shows the enrichment from each database.

Model generation and comparison

First, we filtered ADR–drug associations by calculating the p-value between drug–placebo, and then we combined the drug targets obtained with drug–ADR pairs and developed a quantitative score that associated proteins with ADRs (Figure 3.6; for details see Materials and Methods section). We generated different models by changing the ADR–drug p-value threshold and by categorizing the ADR into organ/tissue compartment and thereby only allowing proteins expressed in the same organ/tissue as the ADR. Table 3.4 shows the comparison between 14 different models thus generated. As we were interested in finding the models with high performances within the highest ADR frequencies, Using comparison of the highest predicted ADR

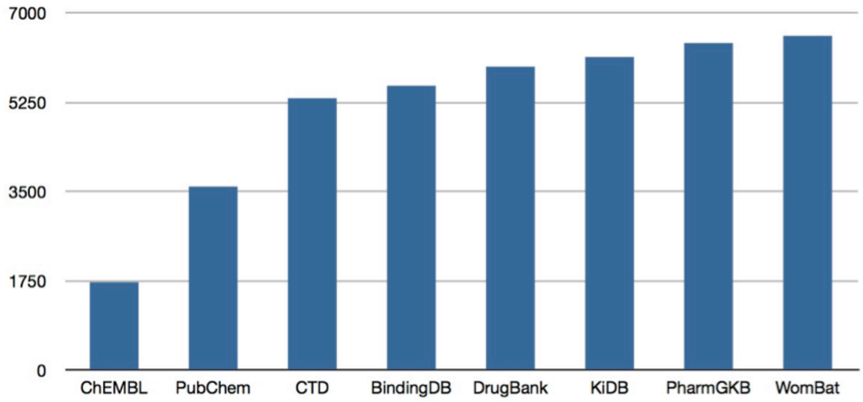


Figure 3.8. Drug-target cumulative enrichment from the 8 databases. A total of 6494 drug-target interactions from 142 drugs. We removed PubChem database from ChEMBL to clearly see the contribution from each database.

score to the top 10 ADR frequencies iteratively for each drug, we evaluated the performance of the models.

Effect of p-value

Model_0.5 gave the best performance with 76% correctly classified ADRs (Table 3.4). This model used a threshold of 0.5 for ADR–drug interaction and included all the top 10 frequencies. From the top 10th frequency in Table 3.3, we inferred that as many as 89% of ADR–drug pairs could be non-significant associations. Therefore, this model was not deemed suitable for validation against ADR frequencies and only Model_0.001 seemed to be the best candidate, since it had the best performance over the top 6 frequencies at a threshold of 0.001. We chose the model that validated against the frequencies within the top 4 because at this level at least 30% of the ADR–drug pairs would be significant as seen from the significance ranks in Table 3.3 (the model is marked with blue color in Table 3.4). We found that this model could correctly classify 55% of the drugs using the top predicted ADR compared to the 4th highest ADR frequencies within each drug.

Model evaluation

To evaluate the significance of the model performance we randomized the drug-targets and ran the model 10000 times to estimate the p-value of the performance. An asterisk in Table 3.4 indicates significant performances.

Effect of tissue specificity

Applying tissue filter to the data, only 5 models out of 70 were found to be significant, whereas 52 models were found to be significant without the use of tissue filter. Applying the tissue filter, thus, reduced the number of proteins in the model by a factor of 3.z

3.7 Validation

We validated the models using information on drug–ADR associations extracted from package inserts from the SIDER 2 database, [28]. We extracted 316 drugs where we could get ADR–drug frequencies and removed the 183 drugs that were in our training set, thus ending up with a set of 133 drugs. Furthermore, we removed 2 drugs which we could not associate any proteins to via ChemProt database and 7 other drugs as they did not have a single ADR within our training set. A total of 124 drugs were used for the validation of the models. Table 3.5 shows the evalidation results for these 124 drugs using ADR–drug frequencies from SIDER. Overall, we saw the same trend that using tissue filter gives lower numbers of correctly classified associations compared to not using it. The selected model Model_0.001 (blue) from the comparison in Table 3.4 was able to classify 54.9% of the interactions correctly. This model was, however, the second best when compared to 4th highest ADR frequencies. The best model was Model_5e-05 with 55.7% correctly classified. It is worth noting that the best performer from the model comparison, Model_0.5, which used all 10 highest frequencies, could only classify 57.4% which is less than the Model_0.001 that only included the 5th highest frequencies (Table 3.5).

Examples of withdrawn drugs from DrugBank

We further validated our model on the withdrawn drugs deposited in DrugBank and searched for their protein targets in the ChemProt 2 server using $< 100\ \mu M$ as an activity filter. We obtained 68 withdrawn drugs from DrugBank of which 51 drugs had protein targets and were not used to build our model. None of these drugs were mentioned in DailyMed or the SIDER database and therefore all the predicted ADRs had to be validated from literature instead. Using the best model from our last validation we found several ADRs associated with Cisapride (Table 3.6) that was withdrawn due to long QTs and arrhythmias in patients [214]. Somnolence, dry mouth and nausea,

	Top frequencies	Significant p-values				Include non-significant p-values		
	of ADRs	**Model_5e-05**	**Model_1e-04**	**Model_0.001**	**Model_0.05**	**Model_0.1**	**Model_0.5**	**Model_1.0**
Without Tissue filter	1	0.225*	0.218*	0.239*	0.218 *	0.211*	0.183*	0.204*
	2	0.444*	0.430*	0.437*	0.380 *	0.345*	0.296*	0.310*
	3	0.514*	0.514*	0.437*	0.486 *	0.451 *	0.444*	0.423*
	4	**0.592***	**0.577***	**0.592***	0.556 *	0.521*	0.535*	0.514*
	5	0.627	0.620	0.620*	0.606 *	0.542*	0.577*	0.599*
	6	0.655	0.662	**0.641***	0.669 *	0.641 *	0.662*	0.655*
	7	0.704	0.718	0.683	0.697 *	0.704*	0.690*	0.690*
	8	0.711	0.732	0.718	0.725*	0.711 *	0.732*	0.711*
	9	0.718	0.732	0.732	0.739 *	0.725*	0.761*	0.718*
	10	0.725	0.739	0.739	**0.746***	**0.732***	**0.761***	**0.732***
With Tissue filter	1	**0.246***	**0.225***	0.218*	0.134	0.120	0.120	0.127
	2	0.387	0.345	0.394	0.268	0.246	0.239	0.254
	3	0.465	0.451	0.493*	0.338	0.303	0.275	0.296
	4	0.542	0.528	**0.556***	0.423	0.380	0.338	0.338
	5	0.570	0.563	0.606	0.465	0.437	0.408	0.387
	6	0.585	0.585	0.627	0.521	0.486	0.444	0.430
	7	0.613	0.620	0.662	0.549	0.514	0.486	0.458
	8	0.641	0.634	0.669	0.563	0.528	0.507	0.486
	9	0.648	0.641	0.683	0.599	0.577	0.563	0.521
	10	0.648	0.641	0.683	0.627	0.613	0.592	0.549

Table 3.4. Comparing models. Models were split depending on whether tissue filter was used or not. These included 4 significant p-values (5e-05,1e-04,1e-03 and 0.05) and 3 non-significant p-values (0.1,0.5 and 1.0). Each of these models was tested against 1 to 10 highest frequencies. An asterisk (*) indicates significant model based on permutation test. Best significant models are indicated with number in bold font and number in bold blue font indicates the optimal model

	Top frequencies	Significant p-values			Include non-significant p-values			
	of ADRs	**Model_5e-05**	**Model_1e-04**	**Model_0.001**	**Model_0.05**	**Model_0.1**	**Model_0.5**	**Model_1.0**
Without Tissue filter	1	0.205*	0.230*	0.180*	0.164 *	0.098*	0.115*	0.115*
	2	0.426*	0.418*	0.434*	0.369 *	0.270*	0.254*	0.238*
	3	0.475*	0.475*	0.475*	0.426 *	0.352 *	0.361	0.328*
	4	**0.557***	**0.549***	**0.549***	0.484 *	0.385*	0.385*	0.344*
	5	0.598	0.607	0.590*	0.500 *	0.475*	0.459*	0.418*
	6	0.598	0.615	**0.615***	0.557 *	0.492 *	0.475*	0.443*
	7	0.639	0.639	0.631	0.582 *	0.541*	0.500*	0.451*
	8	0.656	0.656	0.631	0.607*	0.557 *	0.549*	0.508*
	9	0.656	0.656	0.648	0.623 *	0.574*	0.574*	0.549*
	10	0.664	0.664	0.648	**0.623***	**0.582***	**0.574***	**0.549***
With Tissue filter	1	0.205	0.197	0.189	0.164	0.107	0.107	0.082
	2	0.393	0.385	0.361	0.262	0.189	0.197	0.164
	3	0.451	0.459	0.467	0.352	0.254	0.254	0.221
	4	0.484	0.484	0.484	0.385	0.270	0.311	0.254
	5	0.516	0.516	0.508	0.410	0.320	0.344	0.270
	6	0.541	0.541	0.516	0.434	0.328	0.377	0.328
	7	0.549	0.541	0.533	0.467	0.369	0.410	0.344
	8	0.557	0.549	0.557	0.475	0.377	0.426	0.369
	9	0.566	0.566	0.566	0.492	0.402	0.443	0.393
	10	0.566	0.566	0.590	0.500	0.410	0.459	0.393

Table 3.5. Model validation. Each model for the training set was validated against 124 drugs from the SIDER database. The optimal model from our training could predict 54.9 % (blue bold font). The highest best performance model was Model_0.05 with 62.3% correctly predicted ADRs, which however used all 10 highest frequencies

DrugBank ID	Drug name	Score	ADR	Literature
DB00604	Cisapride	21.85	somnolence	yes
		21.54	dry mouth	yes
		16.66	nausea	yes
		14.30	Dyskinesias	no
			(20 ADRs)	
DB01107	Methyprylon	1.01	somnolence	yes
		0.95	headache	yes
			(2 ADRs)	
DB00901	Bitolterol	1.43	bradycardia	yes
		1.32	fatigue	yes
		1.14	Constipation	no
		0.97	nausea	yes
			(4 ADRs)	
DB04815	Clioquinol	3.36	nausea	yes
		2.65	exfoliation	yes*
		2.34	diarrhea	yes
		1.86	Tremors	no
		1.77	Skin warm	yes
		1.76	Burning sensation	yes
		1.55	dry mouth	yes
			(15 ADRs)	
DB04817	Dipyrone	1.29	abdominal pain	yes
		1.21	Flatulence	yes
		0.98	dyspepsia	yes
			(3 ADRs)	

Table 3.6. Predicting ADRs from withdrawn drugs from DrugBank. Among the predicted ADRs some are supported by literature, which highlight the usefullness of the model

the three top scoring ADRs that are associated with Cisapride are described in the literature [215, 216]. We were not able to predict these ADRs, because they were not part of the high confidence drug–ADR dataset.

We looked at another withdrawn drug, Methyprylon, a sedative used to treat insomnia which now has been replaced by newer drugs like benzodiazepines that have fewer side effects. We could predict ADRs such as somnolence and headache, which were also supported by literature [217]. Somnolence (drowsiness) is a common side effect for sedative medicines and to some extent it is also the therapeutic effect of the drug, which indicates that we captured meaningful interactions. Another interesting finding is the association between Bitolterol (a drug for treatment of asthma) and bradycardia.

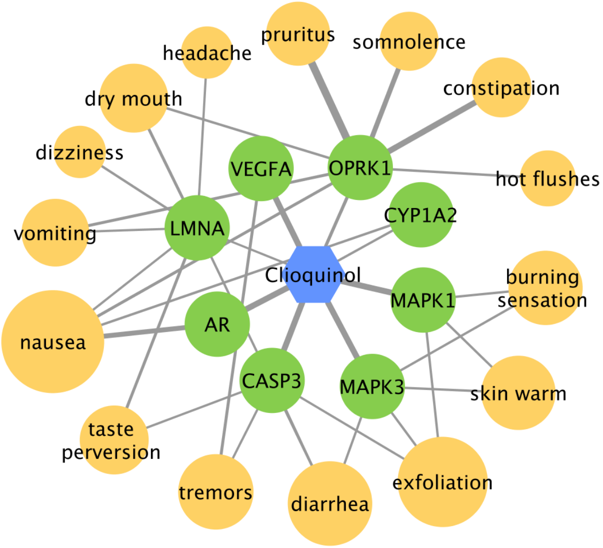


Figure 3.9. Drug-target-ADR network. Prediction of ADR based on the drug targets.

Bradycardia is an ADR connected to a very slow heart rhythm which may cause cardiac arrest. This finding is supported by a study [218] which has shown that Bitolterol could cause cardiac ADRs like rapid heart rate, heart palpitations and irregular heartbeats, which may also be related to bradycardia symptoms.

Case study on Clioquinol

One interesting feature of our models is that it suggests which combination of proteins would be susceptible to generate an ADR. To illustrate how our model works, we have created a network for the antifungal and antiprotozoal drug Clioquinol, its targets and the predicted ADRs associated with it (Figure 3.9). The thickness of the edges between Clioquinol and the targets indicate the strength of bioactivity while the thickness of the edges between targets and ADRs indicate strength of their association based on prediction from the model. The size of the ADR nodes relates to the predicted score and is the summation of the combined interactions from the contributing targets. Nausea is the highest predicted ADR in this exampleas illustrated by the largest node in Figure 3.9, and is caused due to the additive effect from four targets, namely the androgen receptor (AR), Lamin A/C (LMNA), opioid

	Model 5e-05	**Model 1e-04**	**Model 0.001**	**Model 0.05**	**Model 0.1**	**Model 0.5**	**Model 1.0**
Drugs	44	46	64	125	137	167	171
Protein	126	128	168	280	303	447	485
ADR	24	26	34	101	132	232	258
Prediction	0.549	0.549	0.549	0.484	0.385	0.385	0.349

Table 3.7. An overview of the different models. Comparing the number of drug, protein, ADR and the prediction for each models.

receptor (OPRK1), and cytochrome P450 1A2 (CYP1A2). Constipation also has a relatively high ADR score due to OPRK1 target, but due to the missing additive effect from other targets and low bioactivity value of Clioquinol on OPRK1, the total score for the association of this drug to constipation is lower. Clioquinol is also/now marketed as a lotion that is applied to the skin to treat infections and could cause burning sensation on the skin, itching, redness, and swelling [219], which are in our top predicted ADRs (Table 3.6) with an exception of tremors, diarrhea and nausea. However, tremors could be related its neurotoxic side effects, due to which Clioquinol was withdrawn.

3.8 Discussion

We propose a novel approach that takes into account that not all drug–ADR associations are "true" associations. We calculated significant drug–ADR associations, based on the placebo-controlled trials data from the DailyMed database, to create a high confidence drug–ADR dataset. We combined drug–target associations from ChemProt 2 database with the drug–ADR dataset to build a model that could predict ADRs based on drug targets. The results from the analysis of withdrawn drugs show the usefulness of our model in predicting side effects for drugs that are not included in databases such as DailyMed and SIDER 2. This method could be used to give indications about possible side effects for a given compound as long as its bioactivity profile is known.

Our method is, for the moment, limited to the 169 proteins and 35 ADRs collected from DailyMed (Table 3.7). However based on our dataset and the validation method we used, we can conclude that including more proteins with weak bioactivity and non-significant ADRs would only bring more noise to the model. For example, the model (Model_1.0) with 486 proteins and 259 ADRs only predicted 34.9% associations compared to our optimal model (Model_0.001) could predict 54.9% associations. To enrich the protein target space we decided to include cross-species drug-targets from animal studies.

Using the same model parameters from our previous best model and integrating them with the cross-species drug-targets, we were able to improve the performance of the model on the validation set by 2.5% (from 0.549 to 0.574). The full comparison is provided in the supplementary material (Table 3.9 and 3.10). However, although integrating extra information does improve the performance, it also increases the noise in the model, as cross-species drug-targets do not always reflect the same response in human targets. It is yet unclear how much would we gain or lose in the overall performance by including cross-species data. Data completeness is still a problem in drug target networks [87]. Our study focused on clinical ADR reports and some ADRs can be reported after the drug has entered the market. As a perspective, such postmarket ADRs could be integrated for a more accurate prediction and better drug safety assessment.

3.9 Materials and Methods

Data set sources

We downloaded all adverse event reports, as of June 4, 2012, in XML format from DailyMed. XML files were processed using Perl XML and HTML modules and tables summarizing the frequency of adverse events reported in the trials were extracted. We only considered tables that contained ADRs along with placebo frequencies and number of patients used in the trials. We were able to extract 183 drugs with placebo frequency information. The drug names were annotated to a DrugBank ID [23]. The ADRs terms were mapped to MedDRA preferred terms (PT) [220] and to Unified Medical Language System (UMLS) [221]. From version 2 of the ChemProt database, which is a collection of multiple open source and commercial chemical-protein database [213], we retrieved 6494 drug targets, using a relatively low binding affinity threshold (100 μM). The Human Protein Atlas (HPA) version 9.0 [67] was used to annotate drug targets to human tissues. To validate our method we used version 2 of SIDER database released on March 16, 2012 [99].

Estimation of statistical significance

We made a 2x2 contingency table for each ADR / placebo pair by converting the frequencies into a case and a control. We then applied the χ^2 statistic to the 2x2 contingency tables and calculated the p-values. An example is the drug Carduran, which had a 0.099 (9.9%) frequency of causing headache (ADR) whereas the control (placebo) showed a frequency of 0.09 (9%). In order to estimate whether Carduran's association with headache was a chance event, we also needed to take into account the number of patients that underwent the trial. In this example there were 665 patients in the case group and

	Headache	non-Headache
Carduran	65	600
Placebo	81	819

Table 3.8. A 2x2 contingency table showing Carduran vs. placebo.

900 patients in the control group (Table 3.8). With the calculated p-value of 0.66 which was greater than the highest acceptable significance level of 0.05, we rejected Carduran's association with headache .

Estimation of scoring function

We combined the data into a ADR–drug–protein network and applied a weight scheme for the ADR–protein associations which we defined as:

$$= \frac{1}{N_s} \sum_{k=1}^{n} \frac{-log(P_k)pAct_k}{N_d} \tag{3.1}$$

where P_k is the p-value for the ADR, pAct is the negative logarithm of the target activity measured in nM, Nd is the number of proteins associated with the ADR–drug pair and Ns is the total number of ADRs associated with a protein. The term Nd was introduced to avoid scoring function bias towards ADR–protein pairs, where multiple drugs have shown activities for the same target (promisquous targets). The term Ns weighs down the score when the protein is associated with multiple ADRs.

This scoring function enabled us to score each protein–ADR pair and to predict proteins related to ADRs. This in turn enabled the creation of a model that predicted and ranked ADR for a given drug by its bioactivity profile.

Tissue filter

The tissue filter was constructed using protein–tissue annotation with the highest confidence score from HPA. We grouped the tissue names from HPA into 12 main tissue categories using the System Organ Class (SOC) terms from MedDRA to create an ADR–tissue dictionary (Supplementary Table 3.11). Using this dictionary we filtered the proteins from ADR–drug–protein associations to sort them into different tissue categories before calculating the score using Eq. 3.1.

Optimization of significance and score thresholds

We generated a list of ADR–drug interactions by varying the significance level of the p-value in the range of 5e-05 to 0.05. For each generated list we ran the model iteratively over possible scores with a 0.1 increment and saved the best score that could correctly assign the ADRs to the 142 drugs in the training set. We created 14 different models: 7 using the tissue filter and 7 without . Each of the models has one of the 4 significant p-values: (5e-5, 1e-4, 1e-3 and 0.05) or 3 non-significant p-values: (0.1, 0.5 and 1.0). During p-value optimization we ran each model 10 times: First, the highest ADR score was compared to the highest ADR frequency; in the second comparison, the highest ADR score was compared to either of the two highest ADR frequencies; the process was continued until the 10th highest frequency. We evaluated the performance of the models by comparing the highest predicted ADR score to highest ADR frequencies within each 142 drugs in the training set. The model performance was given by the correctly predicted ADRs out of the total number of drugs. The score threshold was subsequently optimized by iterating through the ADR–protein scores to find the one that gave the best performance. Permutation test was used to evaluate the significance of the model. The drug-targets were randomized and each model was run 10000 times to estimate the p-value. An asterisk in Tables 3.4 and 3.5 indicate significant models with a p-value less than 0.05.

Validation

We validated the method using 133 drugs, not included in our dataset, along with information on ADR frequencies from SIDER 2 database. For these 133 drugs we extracted the drug-targets from ChemProt 2 database and retrieved 2115 drug–protein pairs with binding affinity $< 100\ \mu M$. We applied the optimal score calculated from the training set to each validation model.

Supplementary

Model comparison

	Top highest	Significant p-values				Include non-significant p-values		
	Frequencies	Model_5e-05	Model_1e-04	Model_0.001	Model_0.05	Model_0.1	Model_0.5	Model_1.0
Without Tissue filter	1	0.225*	0.225*	0.225*	0.268 *	0.211*	0.204	0.225
	2	0.444*	0.345*	0.451*	0.430 *	0.359	0.324*	0.352
	3	**0.507***	0.451*	0.549*	0.521	0.472 *	0.472*	0.458*
	4	0.599	0.528*	**0.613***	0.585	0.535*	0.563*	0.549
	5	0.634	**0.563***	**0.641***	0.655 *	0.592*	0.606*	0.620
	6	0.669	0.585	0.655	0.676 *	0.655 *	0.676*	0.655
	7	0.704	0.62	0.697	0.690	0.690*	0.704*	0.697*
	8	0.711	0.634	0.732	**0.704***	0.718 *	0.739*	0.718*
	9	0.718	0.641	0.746	0.732	**0.725***	0.761*	0.732*
	10	0.725	0.641	0.754	0.746	0.732	**0.761***	**0.739***
With Tissue filter	1	**0.246***	**0.225***	0.218*	0.134	0.120	0.120	0.127
	2	0.387	0.345	0.394	0.268	0.246	0.239	0.254
	3	0.465	0.451	0.493*	0.338	0.303	0.275	0.296
	4	0.542	0.528	**0.556***	0.423	0.380	0.338	0.338
	5	0.570	0.563	0.606	0.465	0.437	0.408	0.387
	6	0.585	0.585	0.627	0.521	0.486	0.444	0.430
	7	0.613	0.620	0.662	0.549	0.514	0.486	0.458
	8	0.641	0.634	0.669	0.563	0.528	0.507	0.486
	9	0.648	0.641	0.683	0.599	0.577	0.563	0.521
	1	0.648	0.641	0.683	0.627	0.613	0.592	0.549

Table 3.9. Comparing models including cross-species protein targets. Models were split between tissue filter and non-tissue filter. They included 4 significant p-values (5e-05,1e-04,1e-03 and 0.05) and 3 non-significant p-values (0.1,0.5 and 1.0). Each model were tested against 1 to 10 highest frequencies. An asterisk (*) indicate significant model based on permutation. Best significant models are indicated with bold number and bold blue number indicate our optimal model.

	Top highest	**Significant p-values**				**Include non-significant p-values**		
	Frequencies	**Model_5e-05**	**Model_1e-04**	**Model_0.001**	**Model_0.05**	**Model_0.1**	**Model_0.5**	**Model_1.0**
Without Tissue filter	1	0.221	0.238	0.213	0.18	0.172	0.074	0.115
	2	0.481	0.451	0.426	0.41	0.377	0.213	0.295
	3	0.467	0.525	0.484	0.467	0.443	0.393	0.402
	4	0.557	**0.623**	**0.574**	0.533	0.484	0.434	0.426
	5	0.598	0.615	0.607	0.582	0.508	0.516	0.467
	6	0.607	0.623	0.648	0.623	0.557	0.508	0.492
	7	0.615	0.639	0.615	0.648	0.574	0.549	0.508
	8	0.631	0.648	0.615	0.672	0.59	0.615	0.566
	9	0.631	0.648	0.68	0.68	0.598	0.639	0.574
	10	0.639	0.656	0.68	0.697	0.598	0.615	0.582
With Tissue filter	1	0.205	0.197	0.189	0.164	0.107	0.107	0.082
	2	0.393	0.385	0.361	0.262	0.189	0.197	0.164
	3	0.451	0.459	0.467	0.352	0.254	0.254	0.221
	4	0.484	0.484	0.484	0.385	0.270	0.311	0.254
	5	0.516	0.516	0.508	0.410	0.320	0.344	0.270
	6	0.541	0.541	0.516	0.434	0.328	0.377	0.328
	7	0.549	0.541	0.533	0.467	0.369	0.410	0.344
	8	0.557	0.549	0.557	0.475	0.377	0.426	0.369
	9	0.566	0.566	0.566	0.492	0.402	0.443	0.393
	10	0.566	0.566	0.590	0.500	0.410	0.459	0.393

Table 3.10. Model validation including cross-species protein targets. Each model for the training set were validated against 124 drugs from the SIDER database. The optimal model from our training could predict 54.9 % (blue bold number). The highest best performance model were model_0.05 with 62.3% correctly predicted ADRs.

SOC	**HPA**
Blood and lymphatic system disorders	Blood and immune system (Hematopoietic)
Cardiac disorders	Cardiovascular system (Heart and blood vessels)
Congenital, familial and genetic disorders	
Ear and labyrinth disorders	
Endocrine disorders	Endocrine glands
Eye disorders	
Gastrointestinal disorders	Digestive tract (GI-tract)
General disorders and administration site conditions	-
Immune system disorders	Blood and immune system (Hematopoietic) Liver and pancreas Placenta
Infections and infestations	Blood and immune system (Hematopoietic)
Injury, poisoning and procedural complications	Blood and immune system (Hematopoietic)
Investigations	
Metabolism and nutrition disorders	Blood and immune system (Hematopoietic)
Musculoskeletal and connective tissue disorders	Skin and soft tissues
Neoplasms benign, malignant and unspecified (incl cysts and polyps)	
Nervous system disorders	Central nervous system (Brain)
Psychiatric disorders	Central nervous system (Brain)
Renal and urinary disorders	Urinary tract (Kidney and bladder)
Reproductive system and breast disorders	Breast and female reproductive system (Female tissues) Male reproductive system (Male tissues)
Respiratory, thoracic and mediastinal disorders	Respiratory system (Lung)
Skin and subcutaneous tissue disorders	Skin and soft tissues
Social circumstances	Central nervous system (Brain)
Surgical and medical procedures	
Vascular disorders	Cardiovascular system (Heart and blood vessels)

Table 3.11. The System Organ Class (SOC) terms from MedDRA grouped with the 12 main tissue categories from Human Protein Atlas (HPA)

Part III

Epiloque

Chapter 4

Epilogue

In this thesis, I have presented and discussed the integration of chemical–protein annotation resources with complex disease-linked PPI data and applications of side-effect prediction based on system chemical biology approaches. In chapter 2, I presented ChemProt, a disease chemical biology database that integrates 9 different chemical–protein databases. The unique feature of ChemProt is an ability to retrieve information at a cellular level by associating the proteins affected by a chemical to specific tissues and phenotypes. We improved and updated ChemProt (ChemProt 2.0) to include helpful visualization tools like heatmap and interactive complex disease network, which ease the navigation of the pharmacological space for small molecules. In addition to the chemical search, the users of ChemProt 2.0 can now search by diseases, ATC codes and side-effects and thereby retrieve annotation for chemicals and proteins associated with these search queries. All together we hope that this web application framework will assist researchers' in-silico evaluation of the effect of small molecules on proteins, diseases, tissues and adverse drug events.

In chapter 3, I included two articles on side-effect prediction. In the first article, we presented a computer-aided drug repurposing (CADR) method, which was centered on side-effects. The CADR method illustrates how knowledge can evolve from data mining and data integration.

In the second article, we presented a systems chemical biology approach to predict side-effects, which was based on a high confidence drug–ADR dataset. We linked drug–ADR with chemical–protein annotation from

ChemProt and built a model to capture frequently encountered ADR–target associations. We used this model to predict side-effects of withdrawn drugs stored in DrugBank and confirmed the results from the literature.

Perspectives

New joint project collaboration initiatives between the the pharmaceutical industry, academia and other small businesses, for example, OpenPHACTS [222], eTOX [223] and INBIOMED [224], have now started to emerge. These initiatives are the first steps in the right direction for sharing data and information from an otherwise very protected and closed pharmaceutical industry, which will help researchers decipher the clinical effects of drugs.

There are numerous ways in which the work in this thesis could continue as it seems to only scratch the surface of possibilities that could be integrated in systems chemical biology approaches. Pharmacogenomics and personalized medicine could be the next challenges in large-scale data integration, genetic variation from single nucleotide polymorphisms (SNPs) and text mining patient records could reveal useful information.

Data incompleteness of drug-targets is another limiting factor for prediction of novel drug-targets and side-effects [87]. There is a need for full human protein screening programs where compounds can be tested on the whole human proteome.

In order to change the drug discovery paradigm where the disease "one-effect/one-cause/one-target" can be cured by a magic bullet "the drug" we need to develop and optimize network-aided drug development together with human creativity and background knowledge [2].

Bibliography

[1] R L Ho and C A Lieu. Systems biology: an evolving approach in drug discovery and development. *Drugs in R&D*, 9(4):203–216, 2008. PMID: 18588352. 3

[2] P. Csermely, T. Korcsmáros, H. J. M. Kiss, G. London, and R. Nussinov. Structure and dynamics of molecular networks: A novel paradigm of drug discovery. a comprehensive review. *arXiv preprint arXiv:1210.0330*, 2012. 3, 4, 5, 23, 88

[3] H. Y. Chuang, M. Hofree, and T. Ideker. A decade of systems biology. *Annual review of cell and developmental biology*, 26:721, 2010. 3, 4

[4] Hiroaki Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, March 2002. 4

[5] Eberhard Voit, Ana Rute Neves, and Helena Santos. The intricate side of systems biology. *Proceedings of the National Academy of Sciences of the United States of America*, 103(25):9452–9457, June 2006. PMID: 16766654. 4

[6] Leroy Hood and Roger M. Perlmutter. The impact of systems approaches on biological problems in drug discovery. *Nature Biotechnology*, 22(10):1215–1217, 2004. 4

[7] Tudor I. Oprea, Alexander Tropsha, Jean-Loup Faulon, and Mark D. Rintoul. Systems chemical biology. *Nature chemical biology*, 3(8):447–450, August 2007. PMID: 17637771 PMCID: PMC2734506. 4, 27, 36

[8] Seth I Berger and Ravi Iyengar. Network analyses in systems pharmacology. *Bioinformatics (Oxford, England)*, 25(19):2466–2472, October 2009. PMID: 19648136. 4, 27

[9] Charles Auffray, Zhu Chen, and Leroy Hood. Systems medicine: the future of medical genomics and healthcare. *Genome Medicine*, 1(1):2, January 2009. 4

[10] Aislyn D W Boran and Ravi Iyengar. Systems approaches to polypharmacology and drug discovery. *Current opinion in drug discovery & development*, 13(3):297–309, May 2010. PMID: 20443163. 5

[11] F. Brown. Editorial opinion: Chemoinformatics–a ten year update. *Curr Opin Drug Discov Devel*, 8:298–302, 2005. 5

[12] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, February 1988. 5

[13] J. Gasteiger and T. Engel. *Chemoinformatics*. Wiley-VCH, 2006. 6

[14] Daylight. Daylight - http://www.daylight.com/smiles/. 6

[15] www.inchi trust.org. http://www.inchi-trust.org/. 7

[16] The IUPAC Homepage. is http://www.iupac.org/. 8

[17] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002. 9, 28, 39

[18] A.H.-L. and Y.G.S. Concepts and applications of molecular similarity : edited by M.A. johnson and G.M. maggiora; wiley interscience publication, john wiley & sons, new york, 1990, pp. xix + 393, price £51.35. *Journal of Molecular Structure*, 269(3–4):376–377, June 1992. 9

[19] Yvonne C. Martin, James L. Kofron, and Linda M. Traphagen. Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, 45(19):4350–4358, September 2002. 10

[20] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet. Relating protein pharmacology by ligand chemistry. *Nature biotechnology*, 25(2):197–206, 2007. 10, 27, 38, 39, 66

[21] WHO. WHOCC - http://www.whocc.no/. 10

[22] Ellen M McDonagh, Michelle Whirl-Carrillo, Yael Garten, Russ B Altman, and Teri E Klein. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomarkers in medicine*, 5(6):795–806, December 2011. PMID: 22103613. 10, 13, 38

[23] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, An Chi Guo, and David S Wishart. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(Database issue):D1035–1041, January 2011. PMID: 21059682. 10, 12, 38, 66, 79

[24] WHO. WHO | essential medicines. http://www.who.int/topics/essential_medicines/en/. 10

[25] JR Hughes, DK Hatsukami, JE Mitchell, and LA Dahlgren. Prevalence of smoking among psychiatric outpatients. *Am J Psychiatry*, 143(8):993–997, August 1986. 10

[26] Ciara Kelly and Robin G. McCreadie. Smoking habits, current symptoms, and premorbid characteristics of schizophrenic patients in nithsdale, scotland. *Am J Psychiatry*, 156(11):1751–1757, November 1999. 10

[27] Edward R. Lyon. A review of the effects of nicotine on schizophrenia and antipsychotic medications. *Psychiatr Serv*, 50(10):1346–1350, October 1999. 11

[28] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, July 2008. 11, 14, 49, 51, 65, 66, 73

[29] Anna Bauer-Mehren, Erik M. van Mullingen, Paul Avillach, María del Carmen Carrascosa, Ricard Garcia-Serna, Janet Piñero, Bharat Singh, Pedro Lopes, José L. Oliveira, Gayo Diallo, Ernst Ahlberg Helgee, Scott Boyer, Jordi Mestres, Ferran Sanz, Jan A. Kors, and Laura I. Furlong. Automatic filtering and substantiation of drug safety signals. *PLoS Computational Biology*, 8(4):e1002457, April 2012. 11, 36, 46, 66

[30] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012. 12, 35

[31] B. L. Roth, E. Lopez, S. Beischel, R. B. Westkaemper, J. M. Evans, et al. Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacology & therapeutics*, 102(2):99, 2004. 12, 28, 38

[32] M. Olah, R. Rad, L. Ostopovici, A. Bora, N. Hadaruga, D. Hadaruga, R. Moldovan, A. Fulias, M. Mractc, and T. I. Oprea. WOMBAT and WOMBAT-PK: bioactivity databases for lead and drug discovery. *Chemical Biology: From Small Molecules to Systems Biology and Drug Design, Volume 1-3*, page 760–786, 2008. 12, 27, 28, 38, 52

[33] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl 1):D198–D201, 2007. 13, 28, 38

[34] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, and S. Federhen. Database resources of the national center for biotechnology information. *Nucleic acids research*, 39(suppl 1):D38–D51, 2011. 13, 35

[35] J. L. Sharman, C. P. Mpamhanga, M. Spedding, P. Germain, B. Staels, C. Dacquet, V. Laudet, A. J. Harmar, et al. IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. *Nucleic acids research*, 39(suppl 1):D534–D538, 2011. 13, 38

[36] Allan Peter Davis, Cynthia Grondin Murphy, Robin Johnson, Jean M Lay, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L King, Michael C Rosenstein, Thomas C Wiegers, and Carolyn J Mattingly. The comparative toxicogenomics database: update 2013. *Nucleic acids research*, October 2012. PMID: 23093600. 13, 38

[37] M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, L. J. Jensen, and P. Bork. STITCH 3: zooming in on protein–chemical interactions. *Nucleic Acids Research*, 40(D1):D876–D880, 2012. 14, 36, 38

[38] A. Morales, C. Gingell, M. Collins, P. A. Wicker, and I. H. Osterloh. Clinical safety of oral sildenafil citrate (VIAGRA) in the treatment of erectile dysfunction. *International Journal of Impotence Research*, 10(2):69, 1998. 14

[39] Lucas Brouwers, Murat Iskar, Georg Zeller, Vera van Noort, and Peer Bork. Network neighbors of drug targets contribute to drug side-effect similarity. *PLoS ONE*, 6(7):e22187, July 2011. 14

[40] A. L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. 14

[41] Takeshi Hase, Hiroshi Tanaka, Yasuhiro Suzuki, So Nakagawa, and Hiroaki Kitano. Structure of protein interaction networks and their implications on drug design. *PLoS Comput Biol*, 5(10):e1000550, October 2009. 14

[42] Juan Wang, Zhi-xin Li, Cheng-xiang Qiu, Dong Wang, and Qing-hua Cui. The relationship between rational drug design and drug side effects. *Briefings in Bioinformatics*, 13(3):377–382, May 2012. 14

[43] L. Yang, J. Chen, and L. He. Harvesting candidate genes responsible for serious adverse drug reactions from a chemical-protein interactome. *PLoS computational biology*, 5(7):e1000441, 2009. 15

[44] T. I. Oprea, S. K. Nielsen, O. Ursu, J. J. Yang, O. Taboureau, S. L. Mathias, I. Kouskoumvekaki, L. A. Sklar, and C. G. Bologa. Associating drugs, targets and clinical outcomes into an integrated network affords a new platform for computer-aided drug repurposing. *Molecular informatics*, 30(2-3):100–111, 2011. 15, 23, 39, 66

[45] S. Mizutani, E. Pauwels, V. Stoven, S. Goto, and Y. Yamanishi. Relating drug–protein interaction network with drug side effects. *Bioinformatics*, 28(18):i522–i528, 2012. 15, 66

[46] E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, and S. Côté. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, 486(7403):361–367, 2012. 15, 66

[47] Douglas L. Black. Mechanisms of alternative pre-messenger rna splicing. *Annual Review of Biochemistry*, 72(1):291–336, 2003. PMID: 12626338. 19

[48] E. C. Dimmer, R. P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M. J. Martin, B. Bely, P. Browne, W. M. Chan, R. Eberhardt, et al. The UniProt-GO annotation database in 2011. *Nucleic acids research*, 40(D1):D565–D570, 2012. 20, 38, 39, 52

[49] M. E. Cusick. Interactome: gateway into systems biology. *Human Molecular Genetics*, 14(suppl_2):R171–R181, October 2005. 20

[50] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein protein interactions. *, Published online: 20 July 1989; | doi:10.1038/340245a0*, 340(6230):245–246, July 1989. 20

[51] R. Aebersold, M. Mann, et al. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003. 20

[52] Kasper Lage, Niclas Tue Hansen, E Olof Karlberg, Aron C Eklund, Francisco S Roque, Patricia K Donahoe, Zoltan Szallasi, Thomas Skøt Jensen, and Søren Brunak. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 105(52):20870–20875, December 2008. PMID: 19104045. 20, 21, 29, 30

[53] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(D1):D857–D861, November 2011. 20

[54] Gary D Bader, Doron Betel, and Christopher W V Hogue. BIND: the biomolecular interaction network database. *Nucleic acids research*, 31(1):248–250, January 2003. PMID: 12519993. 20, 29

[55] B. J. Breitkreutz, C. Stark, M. Tyers, et al. The GRID: the general repository for interaction datasets. *Genome Biol*, 4(3):R23, 2003. 20

[56] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein reference database–2009 update. *Nucleic Acids Research*, 37(Database):D767–D772, January 2009. 20

[57] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. IntAct–open source resource for molecular interaction data. *Nucleic Acids Research*, 35(Database):D561–D565, January 2007. 20

[58] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic acids research*, 32(Database issue):D449–451, January 2004. PMID: 14681454. 20, 29

[59] Thijs Beuming, Lucy Skrabanek, Masha Y. Niv, Piali Mukherjee, and Harel Weinstein. PDZBase: a protein–protein interaction database for PDZ-domains. *Bioinformatics*, 21(6):827–828, March 2005. 20

[60] Lisa Matthews, Gopal Gopinath, Marc Gillespie, Michael Caudy, David Croft, Bernard de Bono, Phani Garapati, Jill Hemish, Henning Hermjakob, Bijay Jassal, Alex Kanapin, Suzanna Lewis, Shahana Mahajan, Bruce May, Esther Schmidt, Imre Vastrik, Guanming Wu, Ewan Birney, Lincoln Stein, and Peter D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic acids research*, 37(Database issue):D619–622, January 2009. PMID: 18981052. 20, 38

[61] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, January 2012. PMID: 22080510 PMCID: PMC3245020. 20, 38

[62] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S Goldberg, Lan V Zhang, Sharyl L Wong, Giovanni Franklin, Siming Li, Joanna S Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S Sikorski,

Jean Vandenhaute, Huda Y Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E Cusick, David E Hill, Frederick P Roth, and Marc Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, October 2005. PMID: 16189514. 21, 29

[63] Marilyn Safran, Irina Solomon, Orit Shmueli, Michal Lapidot, Shai Shen-Orr, Avital Adato, Uri Ben-Dor, Nir Esterman, Naomi Rosen, Inga Peter, Tsviya Olender, Vered Chalifa-Caspi, and Doron Lancet. GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics (Oxford, England)*, 18(11):1542–1543, November 2002. PMID: 12424129. 21, 29

[64] Joanna Amberger, Carol Bocchini, and Ada Hamosh. A new face and new challenges for online mendelian inheritance in man (OMIM®). *Human mutation*, 32(5):564–567, May 2011. PMID: 21472891. 21, 38

[65] Kasper Lage, E. Olof Karlberg, Zenia M. Størling, Páll Í Ólason, Anders G. Pedersen, Olga Rigina, Anders M. Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, Yves Moreau, and Søren Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309–316, 2007. 21, 27, 29, 36

[66] BioAlma. GeneCards - www.genecards.org/. 21

[67] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, and S. Hober. Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248–1250, 2010. 21, 79

[68] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P Cooke, John R Walker, and John B Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, April 2004. PMID: 15075390. 21, 29

[69] K. Audouze and P. Grandjean. Application of computational systems biology to explore environmental toxicity hazards. *Environmental health perspectives*, 119(12):1754, 2011. 23

[70] P. Tiikkainen and L. Franke. Analysis of commercial and public bioactivity databases. *Journal of chemical information and modeling*, 52(2):319–326, 2011. 23

[71] B. Chen, Y. Ding, and D. J. Wild. Assessing drug target association using semantic linked data. *PLoS Computational Biology*, 8(7):e1002574, 2012. 23

[72] K. Krysiak-Baltyn, J. Toppari, N. E. Skakkebaek, T. S. Jensen, H. E. Virtanen, K. W. Schramm, H. Shen, T. Vartiainen, H. Kiviranta, and O. Taboureau. Association between chemical pattern in breast milk and congenital cryptorchidism: modelling of complex human exposures. *International journal of andrology*, 2012. 23

[73] H. Polur, T. Joshi, C. T. Workman, G. Lavekar, and I. Kouskoumvekaki. Back to the roots: Prediction of biologically active natural products from ayurveda traditional medicine. *Molecular Informatics*, 30(2-3):181–187, 2011. 23

[74] E. R. Yera, A. E. Cleves, and A. N. Jain. Chemical structural novelty: on-targets and off-targets. *Journal of medicinal chemistry*, 54(19):6771–6785, 2011. 23

[75] L. Tari, N. Vo, S. Liang, J. Patel, C. Baral, and J. Cai. Identifying novel drug indications through automated reasoning. *PloS one*, 7(7):e40946, 2012. 23

[76] Ricard Garcia-Serna and Jordi Mestres. Anticipating drug side effects by comparative pharmacology. *Expert Opinion on Drug Metabolism & Toxicology*, 6(10):1253–1263, October 2010. 23, 57, 66

[77] K. Jensen, D. Plichta, G. Panagiotou, and I. Kouskoumvekaki. Mapping the genome of plasmodium falciparum on the drug-like chemical space reveals novel anti-malarial targets and potential drug leads. *Molecular BioSystems*, 2012. 23

[78] L. Xie, L. Xie, S. L. Kinnings, and P. E. Bourne. Novel computational approaches to polypharmacology as a means to define responses to individual drugs. *Annual review of pharmacology and toxicology*, 52:361–379, 2012. 23

[79] B. L. King, A. P. Davis, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly. Ranking transitive chemical-disease inferences using local network topology in the comparative toxicogenomics database. *PloS one*, 7(11):e46524, 2012. 23

[80] D. J. Wild, Y. Ding, A. P. Sheth, L. Harland, E. M. Gifford, and M. S. Lajiness. Systems chemical biology and the semantic web: what they mean for the future of drug discovery research. *Drug discovery today*, 2011. 23, 36

[81] G. Panagiotou and O. Taboureau. The impact of network biology in pharmacology and toxicology. *SAR and QSAR in Environmental Research*, 23(3-4):221–235, 2012. 23

[82] H. W. Chang, L. Y. Chuang, M. T. Tsai, and C. H. Yang. The importance of integrating SNP and cheminformatics resources to pharmacogenomics. *Current Drug Metabolism*, 13(7):991–999, 2012. 23

[83] M. Glick and E. Jacoby. The role of computational methods in the identification of bioactive compounds. *Current Opinion in Chemical Biology*, 15(4):540–546, 2011. 23

[84] J. Kubrycht, K. Sigler, and P. Sou\vcek. Virtual interactomics of proteins from biochemical standpoint. *Molecular Biology International*, 2012, 2012. 23

[85] Gaia V. Paolini, Richard H. B. Shapland, Willem P. van Hoorn, Jonathan S. Mason, and Andrew L. Hopkins. Global mapping of pharmacological space. *Nature Biotechnology*, 24(7):805–815, 2006. 26, 27, 50

[86] Michael J. Keiser, Vincent Setola, John J. Irwin, Christian Laggner, Atheir I. Abbas, Sandra J. Hufeisen, Niels H. Jensen, Michael B. Kuijer, Roberto C. Matos, Thuy B. Tran, Ryan Whaley, Richard A. Glennon, Jérôme Hert, Kelan L. H. Thomas, Douglas D. Edwards, Brian K. Shoichet, and Bryan L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, November 2009. 26, 27, 49

[87] Jordi Mestres, Elisabet Gregori-Puigjané, Sergi Valverde, and Ricard V. Solé. Data completeness—the achilles heel of drug-target networks. *Nature Biotechnology*, 26(9):983–984, 2008. 26, 50, 51, 79, 88

[88] A. Antonelli, S. M. Ferrari, P. Fallahi, S. Piaggi, A. Paolicchi, S. S. Franceschini, M. Salvi, and E. Ferrannini. Cytokines (interferon-< i> γ</i> and tumor necrosis factor–< i> α</i>)-induced nuclear factor–< i> κ</i> b activation and chemokine (CXC motif) ligand 10 release in graves disease and ophthalmopathy are modulated by pioglitazone. *Metabolism*, 60(2):277–283, 2011. 26

[89] R. J. Vaz, T. Klabunde, and R. Mannhold. *Antitargets: Prediction and prevention of drug side effects*. Wiley-VCH Weinheim, Germany, 2008. 26

[90] F. Broccatelli, E. Carosati, G. Cruciani, and T. I. Oprea. Transporter-mediated efflux influences CNS side effects: ABCB1, from antitarget to target. *Molecular informatics*, 29(1-2):16–26, 2010. 27

[91] N. Weill and D. Rognan. Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to g protein-coupled receptors and their ligands. *Journal of chemical information and modeling*, 49(4):1049, 2009. 27

[92] J. Mestres, L. Martín-Couce, E. Gregori-Puigjané, M. Cases, and S. Boyer. Ligand-based approach to in silico pharmacology: nuclear receptor profiling. *Journal of chemical information and modeling*, 46(6):2725–2736, 2006. 27

[93] Zachary A Knight, Henry Lin, and Kevan M Shokat. Targeting the cancer kinome through polypharmacology. *Nature reviews. Cancer*, 10(2):130–137, February 2010. PMID: 20094047. 27

[94] Ricard Garcia-Serna, Oleg Ursu, Tudor I. Oprea, and Jordi Mestres. iPHACE: integrative navigation in pharmacological space. *Bioinformatics*, 26(7):985–986, April 2010. 27, 53

[95] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Lewis Y Geer, Yuri Kapustin, Oleg Khovayko, David Landsman, David J Lipman, Thomas L Madden, Donna R Maglott, James Ostell, Vadim Miller, Kim D Pruitt, Gregory D Schuler, Edwin Sequeira, Steven T Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Roman L Tatusov, Tatiana A Tatusova, Lukas Wagner, and Eugene Yaschenko. Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(Database issue):D5–12, January 2007. PMID: 17170002. 27, 28

[96] Paula de Matos, Rafael Alcántara, Adriano Dekker, Marcus Ennis, Janna Hastings, Kenneth Haug, Inmaculada Spiteri, Steve Turner, and Christoph Steinbeck. Chemical entities of biological interest: an update. *Nucleic acids research*, 38(Database issue):D249–254, January 2010. PMID: 19854951. 27, 28

[97] Muhammed A Yıldırım, Kwang-Il Goh, Michael E Cusick, Albert-László Barabási, and Marc Vidal. Drug—target network. *Nature Biotechnology*, 25(10):1119–1126, October 2007. 27, 50

[98] Bin Chen, David Wild, and Rajarshi Guha. PubChem as a source of polypharmacology. *Journal of chemical information and modeling*, 49(9):2044–2055, September 2009. PMID: 19708682. 27

[99] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6:343, 2010. PMID: 20087340. 27, 53, 57, 66, 79

[100] K. Audouze, A. S. Juncker, F. J. Roque, K. Krysiak-Baltyn, N. Weinhold, O. Taboureau, T. S. Jensen, and S. Brunak. Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks. *PLoS computational biology*, 6(5):e1000788, 2010. 27

[101] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007. 27, 38

[102] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(Database issue):D668–672, January 2006. PMID: 16381955. 28

[103] Micheal Hewett, Diane E Oliver, Daniel L Rubin, Katrina L Easton, Joshua M Stuart, Russ B Altman, and Teri E Klein. PharmGKB: the pharmacogenetics knowledge base. *Nucleic acids research*, 30(1):163–165, January 2002. PMID: 11752281. 28

[104] Allan Peter Davis, Cynthia G Murphy, Cynthia A Saraceni-Richards, Michael C Rosenstein, Thomas C Wiegers, and Carolyn J Mattingly. Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic acids research*, 37(Database issue):D786–792, January 2009. PMID: 18782832. 28

[105] Michael Kuhn, Damian Szklarczyk, Andrea Franceschini, Monica Campillos, Christian von Mering, Lars Juhl Jensen, Andreas Beyer, and Peer Bork. STITCH 2: an interaction network database for small molecules and proteins. *Nucleic acids research*, 38(Database issue):D552–556, January 2010. PMID: 19897548. 28

[106] B. L. Bush and R. P. Sheridan. PATTY: a programmable atom type and language for automatic classification of atoms in molecular databases. *Journal of chemical information and computer sciences*, 33(5):756–762, 1993. 28

[107] MOE. (version 2007.09), chemical computing group, montreal, canada. [(29 september 2010, date last accessed)]. www.chemcomp.com. 29

[108] Peter Willett. Similarity-based virtual screening using 2D fingerprints. *Drug discovery today*, 11(23-24):1046–1053, December 2006. PMID: 17129822. 29

[109] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database issue):D535–539, January 2006. PMID: 16381927. 29

[110] Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, and Gianni Cesareni. MINT: a molecular INTeraction database. *FEBS letters*, 513(1):135–140, February 2002. PMID: 11911893. 29

[111] Gopa R Mishra, M Suresh, K Kumaran, N Kannabiran, Shubha Suresh, P Bala, K Shivakumar, N Anuradha, Raghunath Reddy, T Madhan Raghavan, Shalini Menon, G Hanumanthu, Malvika Gupta, Sapna Upendran, Shweta Gupta, M Mahesh, Bincy Jacob, Pinky Mathew, Pritam Chatterjee, K S Arun, Salil Sharma,

K N Chandrika, Nandan Deshpande, Kshitish Palvankar, R Raghavnath, R Krishnakanth, Hiren Karathia, B Rekha, Rashmi Nayak, G Vishnupriya, H G Mohan Kumar, M Nagini, G S Sameer Kumar, Rojan Jose, P Deepthi, S Sujatha Mohan, T K B Gandhi, H C Harsha, Krishna S Deshpande, Malabika Sarker, T S Keshava Prasad, and Akhilesh Pandey. Human protein reference database–2006 update. *Nucleic acids research*, 34(Database issue):D411–414, January 2006. PMID: 16381900. 29

[112] Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, Hanah Margalit, John Armstrong, Amos Bairoch, Gianni Cesareni, David Sherman, and Rolf Apweiler. IntAct: an open source molecular interaction database. *Nucleic acids research*, 32(Database issue):D452–455, January 2004. PMID: 14681455. 29

[113] Philipp Pagel, Stefan Kovac, Matthias Oesterheld, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Pekka Mark, Volker Stümpflen, Hans-Werner Mewes, Andreas Ruepp, and Dmitrij Frishman. The MIPS mammalian protein-protein interaction database. *Bioinformatics (Oxford, England)*, 21(6):832–834, March 2005. PMID: 15531608. 29

[114] Ulrich Güldener, Martin Münsterkötter, Matthias Oesterheld, Philipp Pagel, Andreas Ruepp, Hans-Werner Mewes, and Volker Stümpflen. MPact: the MIPS protein interaction resource on yeast. *Nucleic acids research*, 34(Database issue):D436–441, January 2006. PMID: 16381906. 29

[115] G Joshi-Tope, M Gillespie, I Vastrik, P D'Eustachio, E Schmidt, B de Bono, B Jassal, G R Gopinath, G R Wu, L Matthews, S Lewis, E Birney, and L Stein. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(Database issue):D428–432, January 2005. PMID: 15608231. 29

[116] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research*, 34(Database issue):D354–357, January 2006. PMID: 16381885. 29

[117] Kevin P O'Brien, Maido Remm, and Erik L L Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic acids research*, 33(Database issue):D476–480, January 2005. PMID: 15608241. 29

[118] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(Database issue):D514–517, January 2005. PMID: 15608251. 29

[119] Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. *Nucleic acids research*, 32(Database issue):D262–266, January 2004. PMID: 14681408. 29

[120] F Pontén, K Jirström, and M Uhlen. The human protein atlas–a tool for pathology. *The Journal of pathology*, 216(4):387–393, December 2008. PMID: 18853439. 29

[121] Marvin. , version5.3. [(29 september 2010, date last accessed)]. http://www.chemaxon.com/. 30

[122] Evangelos Pafilis, Seán I O'Donoghue, Lars J Jensen, Heiko Horn, Michael Kuhn, Nigel P Brown, and Reinhard Schneider. Reflect: augmented browsing for the life scientist. *Nature biotechnology*, 27(6):508–510, June 2009. PMID: 19513049. 30

[123] Anita Chamba, Michelle J Holder, Ruth F Jarrett, Lesley Shield, Kai M Toellner, Mark T Drayson, Nicholas M Barnes, and John Gordon. SLC6A4 expression and anti-proliferative responses to serotonin transporter ligands chlomipramine and fluoxetine in primary b-cell malignancies. *Leukemia research*, 34(8):1103–1106, August 2010. PMID: 20363025. 32

[124] Rolf U Halden. Plastics and health risks. *Annual review of public health*, 31:179–194, 2010. PMID: 20070188. 32

[125] Asher Mullard. Accelerated approval dust begins to settle. *Nature Reviews Drug Discovery*, 10(11):797–798, November 2011. 35

[126] A. J. Williams, S. Ekins, and V. Tkachenko. Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug discovery today*, 2012. 35

[127] D. Rognan. Chemogenomic approaches to rational drug design. *British journal of pharmacology*, 152(1):38–52, 2009. 36

[128] A. L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690, 2008. 36, 65

[129] G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman. Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27(13):1741–1748, 2011. 36

[130] N. T. Hansen, S. Brunak, and R. B. Altman. Generating genome-scale candidate gene lists for pharmacogenomics. *Clinical Pharmacology & Therapeutics*, 86(2):183–189, 2009. 36

[131] J. von Eichborn, M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne, and R. Preissner. PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic acids research*, 39(suppl 1):D1060–D1066, 2011. 36

[132] O. Taboureau, S. K. Nielsen, K. Audouze, N. Weinhold, D. Edsgärd, F. S. Roque, I. Kouskoumvekaki, A. Bora, R. Curpan, T. S. Jensen, et al. ChemProt: a disease chemical biology database. *Nucleic acids research*, 39(suppl 1):D367–D372, 2011. 36, 62

[133] P. De Smet. New applications of the ATC/DDD methodology in the netherlands. part 1.(ATC/DDD principles and computerized medication surveillance). *INTERNATIONAL PHARMACY JOURNAL*, 7:196–196, 1993. 38

[134] G. Stelzer, I. Dalah, T. I. Stein, Y. Satanower, N. Rosen, N. Nativ, D. Oz-Levi, T. Olender, F. Belinky, I. Bahir, et al. In-silico human genomics with GeneCards. *Human Genomics*, 5(6):709–717, 2011. 38

[135] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):1–14, 2011. 39

[136] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997. 39

[137] M. Politis and C. Loane. Serotonergic dysfunction in parkinson's disease and its relevance to disability. *TheScientificWorldJOURNAL*, 11:1726, 2011. 40

[138] J. T. Ferreira, P. Q. Levy, C. R. Marinho, M. P. Bicho, and M. R. Mascarenhas. Association of serotonin transporter gene polymorphism 5HTTVNTR with osteoporosis. *Acta reumatológica portuguesa*, 36(1):14, 2011. 40

[139] T. I. Oprea, O. Taboureau, and C. G. Bologa. Of possible cheminformatics futures. *Journal of computer-aided molecular design*, page 1–6, 2012. 43

[140] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464 –1480, September 1990. 45, 54

[141] Kinley Larntz. Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association*, 73(362):253–263, 1978. 46

[142] R. L. Chang, L. Xie, L. Xie, P. E. Bourne, and B. Ø Palsson. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Computational Biology*, 6(9):e1000938, 2010. 46

[143] J. Drews. Innovation deficit revisited: reflections on the productivity of pharmaceutical R&D. *Drug Discovery Today*, 3(11):491–494, 1998. 48

[144] T Olsson and T I Oprea. Cheminformatics: a tool for decision-makers in drug discovery. *Current opinion in drug discovery & development*, 4(3):308–313, May 2001. PMID: 11560063. 48

[145] Joseph A DiMasi, Ronald W Hansen, and Henry G Grabowski. The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 22(2):151–185, March 2003. 48

[146] Christopher P. Adams and Van V. Brantner. Estimating the cost of new drug development: Is it really $802 million? *Health Affairs*, 25(2):420–428, March 2006. 48

[147] Christopher P. Austin, Linda S. Brady, Thomas R. Insel, and Francis S. Collins. NIH molecular libraries initiative. *Science*, 306(5699):1138–1139, November 2004. 48

[148] CTSA. : http ://www.ncrr.nih.gov/clinical_research_resources/ clinical_and_translational_science_awards/. 48, 50

[149] Francis S. Collins. Opportunities for research and NIH. *Science*, 327(5961):36–37, January 2010. 48, 51

[150] Mark S. Boguski, Kenneth D. Mandl, and Vikas P. Sukhatme. Repurposing with a difference. *Science*, 324(5933):1394–1395, June 2009. 49

[151] Jeffrey H. Toney, Jeffry I. Fasick, Sonal Singh, Chris Beyrer, and David J. Sullivan. Purposeful learning with drug repurposing. *Science*, 325(5946):1339–1340, September 2009. 49

[152] Curtis R. Chong and David J. Sullivan. New uses for old drugs. *Nature*, 448(7154):645–646, August 2007. 49

[153] Ted T. Ashburn and Karl B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, August 2004. 49

[154] Smyth S Campbell CL. Aspirin dose for the prevention of cardiovascular disease: A systematic review. *JAMA: The Journal of the American Medical Association*, 297(18):2018–2024, May 2007. 49

[155] The FDA approved label for Zyrtec D 12 hour. can be accessed at http ://www.accessdata.fda.gov/drugsatfda_docs/label/ 2004/19835slr016,21150slr005,30346slr011_zyrtec_lbl.pdf. 49

[156] Diogo R. Lara. Caffeine, mental health, and psychiatric disorders. *Journal of Alzheimer's Disease*, 20(0):239–248, January 2010. 49

[157] Michael J. Glade. Caffeine—Not just a stimulant. *Nutrition*, 26(10):932–938, October 2010. 49

[158] David J Henderson-Smart and Antonio G De Paoli. Methylxanthine treatment for apnoea in preterm infants. *Cochrane database of systematic reviews (Online)*, (12):CD000140, 2010. PMID: 21154343. 49

[159] The FDA Guidance. for the industry with respect to 505(b)(2) is available at http ://www.fda.gov/downloads/Drugs/Guidance ComplianceRegulatoryInformation/Guidances/ucm079345.pdf. 49

[160] The Hatch-Waxman Amendments. are at http ://www.fdli.org/ pubs/Journal online/54_2/art2.pdf. 49

[161] CTSA research volunteers. : https ://www.researchmatch.org/. 50

[162] CTSA IP. : http ://www.ctsaip.org/. 50

[163] CTSA Portal. : http ://www.ctsapharmaportal.org/. 50

[164] Ingo Vogt and Jordi Mestres. Drug-target networks. *Molecular Informatics*, 29(1-2):10–14, 2010. 50

[165] The PubChem Portal. is http ://pubchem.ncbi.nlm.nih.gov/. 50

[166] Prestwick Chemical Library information. at http ://www. prestwickchemical.fr/index.php ?pa = 26. 50

[167] DailyMed. can be accessed at http ://dailymed.nlm.nih.gov/ dailymed/. 51, 66

[168] DrugBank. is available at http ://www.drugbank.ca/. 51

[169] T. Berners-Lee, J. Hendler, O. Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001. 51

[170] VAZ. Front matter. In Roy J. Vaz and Thomas Klabunde, editors, *Antitargets*, page I–XXIV. Wiley-VCH Verlag GmbH & Co. KGaA, 2008. 51

[171] PDSP. is available at http://pdsp.med.unc.edu/. 52

[172] John R. Proudfoot. The evolution of synthetic oral drug properties. *Bioorganic & Medicinal Chemistry Letters*, 15(4):1087–1090, February 2005. 52

[173] Michal Vieth and Jeffrey J. Sutherland. Dependence of molecular properties on proteomic family for marketed oral drugs. *Journal of Medicinal Chemistry*, 49(12):3451–3453, June 2006. 52

[174] SciFinder. is available at https ://scifinder.cas.org/scifinder/. 52

[175] O. Taboureau, S. K. Nielsen, K. Audouze, N. Weinhold, D. Edsgard, F. S. Roque, I. Kouskoumvekaki, A. Bora, R. Curpan, T. S. Jensen, S. Brunak, and T. I. Oprea. ChemProt: a disease chemical biology database. *Nucleic Acids Research*, 39(Database):D367–D372, October 2010. 52

[176] A. J. Harmar, R. A. Hills, E. M. Rosser, M. Jones, O. P. Buneman, D. R. Dunbar, S. D. Greenhill, V. A. Hale, J. L. Sharman, T. I. Bonner, W. A. Catterall, A. P. Davenport, P. Delagrange, C. T. Dollery, S. M. Foord, G. A. Gutman, V. Laudet, R. R. Neubig, E. H. Ohlstein, R. W. Olsen, J. Peters, J.-P. Pin, R. R. Ruffolo, D. B. Searls, M. W. Wright, and M. Spedding. IUPHAR-DB: the IUPHAR database of g protein-coupled receptors and ion channels. *Nucleic Acids Research*, 37(Database):D680–D685, January 2009. 52

[177] The IUPHAR Database. is http ://www.iuphar-db.org/index.jsp. 52

[178] The ChemblDB Databases. are at http ://www.ebi.ac.uk/ chembldb/index.php. 52

[179] J W Black, W A Duncan, and R G Shanks. Comparison of some properties of pronethalol and propranolol. *British journal of pharmacology and chemotherapy*, 25(3):577–591, December 1965. PMID: 4380598. 52

[180] JChem. and other ChemAxon software are available from ChemAxon at http ://www.chemaxon.com/product/jc_base. html. 52

[181] OEChem. and other OpenEye software are available from OpenEye scientific software at http ://www.eyesopen.com/. 52

[182] The iPHACE interface from IMIM. , http ://cgl.imim.es/iphace/ main.php, is mirrored at UNM, http ://agave.health.unm.edu/ iphace/main.php. 53

[183] The IUPHAR Database linked to Wikipedia. , iPHACE and PharmKGB in may 2010. 53

[184] The U.S. Food and drug administration on-line search inter- face, “Drugs@FDA” service, can be accessed at http://www. accessdata.fda.gov/scripts/cder/drugsatfda/. 53

[185] EMA. products authorized for human use can be accessed at http ://www.ema.europa.eu/htms/human/epar/a.htm. 53

[186] WHO. Essential medicines are listed at http://www.who.int/topics/essential_medicines/en/. 53

[187] TGA Medicines. are found at http ://www.tga.gov.au/. 53

[188] PDR. is available at http ://www.pdr.net/. 53

[189] Martindale. is available at http ://medicinescomplete.com/mc/ martindale/current/. 53

[190] AHFS. is available at http ://medicinescomplete.com/mc/ahfs/ current/. 53

[191] I. Feinerer, K. Hornik, and D. Meyer. Text mining infrastructure in r. *Journal of Statistical Software*, 25(5):1–54, 2008. 54

[192] J. Edward Jackson. *A User's Guide to Principal Components.* John Wiley & Sons, March 1991. 54

[193] Simca software. is available from umetrics at http://www. umetrics.com/. 54

[194] The Spotfire Data Analysis Module. is available from TIBCO at http ://spotfire.tibco.com/. 54

[195] Xiaoyan Wang, George Hripcsak, Marianthi Markatou, and Carol Friedman. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. *Journal of the American Medical Informatics Association*, 16(3):328–337, May 2009. 57

[196] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, November 2003. 59

[197] Cytoscape. can be downloaded from http ://www.cytoscape. org. 59

[198] Bin Chen, Xiao Dong, Dazhi Jiao, Huijun Wang, Qian Zhu, Ying Ding, and David J Wild. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, 11(1):255, 2010. 62

[199] H.W. Mewes, B. Wachinger, and V. Stümpflen. Perspectives of a systems biology of the synapse: How to transform an indefinite data space into a model? *Pharmacopsychiatry*, 43(S 01):S2–S8, May 2010. 62

[200] M. Kuhn, D. Szklarczyk, A. Franceschini, M. Campillos, C. von Mering, L. J. Jensen, A. Beyer, and P. Bork. STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Research*, 38(Database):D552–D556, November 2009. 62

[201] STITCH. , http://stitch.embl.de/. 62

[202] D. C. Classen, S. L. Pestonik, R. Scott Evans, J. F. Lloyd, and J. P. Burke. Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality. *Obstetrical & gynecological survey*, 52(5):291, 1997. 65

[203] D. W. Bates, D. J. Cullen, N. Laird, L. A. Petersen, S. D. Small, D. Servi, G. Laffel, B. J. Sweitzer, B. F. Shea, and R. Hallisey. Incidence of adverse drug events and potential adverse drug events. *JAMA: the journal of the American Medical Association*, 274(1):29–34, 1995. 65

[204] I. Kola and J. Landis. Can the pharmaceutical industry reduce attrition rates? *Nature reviews Drug discovery*, 3(8):711–716, 2004. 65

[205] Josef Scheiber, Jeremy L. Jenkins, Sai Chetan K. Sukuru, Andreas Bender, Dmitri Mikhailov, Mariusz Milik, Kamal Azzaoui, Steven Whitebread, Jacques Hamon, Laszlo Urban, Meir Glick, and John W. Davies. Mapping adverse drug reactions in chemical space. *Journal of Medicinal Chemistry*, 52(9):3103–3107, May 2009. 65

[206] Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. 26(12):i246–i254, June 2010. PMID: 20529913 PMCID: 2881361. 65

[207] A. Bender, J. Scheiber, M. Glick, J. W. Davies, K. Azzaoui, J. Hamon, L. Urban, S. Whitebread, and J. L. Jenkins. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem*, 2(6):861–873, 2007. 65

[208] N. Atias and R. Sharan. An algorithmic framework for predicting side effects of drugs. *Journal of Computational Biology*, 18(3):207–218, 2011. 65, 66

[209] Keith Finlayson, Harry J. Witchel, James McCulloch, and John Sharkey. Acquired QT interval prolongation and HERG: implications for drug discovery and development. *European Journal of Pharmacology*, 500(1–3):129–142, October 2004. 65

[210] Anton F Fliri, William T Loging, and Robert A Volkmann. Analysis of system structure-function relationships. *ChemMedChem*, 2(12):1774–1782, December 2007. PMID: 17952882. 66

[211] M. Takarabe, M. Kotera, Y. Nishimura, S. Goto, and Y. Yamanishi. Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*, 28(18):i611–i618, 2012. 66

[212] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman. Data-driven prediction of drug effects and interactions. *Sci Transl Med*, 4:125ra31, 2012. 66

[213] Sonny Kim Kjærulff, Louis Wich, Jens Kringelum, Ulrik P. Jacobsen, Irene Kouskoumvekaki, Karine Audouze, Ole Lund, Søren Brunak, Tudor I. Oprea, and Olivier Taboureau. ChemProt-2.0: visual navigation in a disease chemical biology database. *Nucleic Acids Research*, November 2012. 66, 79

[214] Ilari Paakkari. Cardiotoxicity of new antihistamines and cisapride. *Toxicology letters*, 127(1-3):279–284, February 2002. PMID: 12052668. 73

[215] Dr Enrico Corazziari Md, Immacolata Bontempo Md, and Fiorella Anzini. Effects of cisapride on distal esophageal motility in humans. *Digestive Diseases and Sciences*, 34(10):1600–1605, October 1989. 76

[216] Lucena R, Monteiro L, and Melo A. [cisapride related movement disorders]. *Jornal de pediatria*, 74(5):416, October 1998. 76

[217] Louis Lasagna. A study of hypnotic drugs in patients with chronic diseases: Comparative efficacy of placebo; methyprylon (noludar); meprobamate (miltown, equanil) pentobarbital; phenobarbital; secobarbital. *Journal of Chronic Diseases*, 3(2):122–133, February 1956. 76

[218] J L Pinnas, B D Bhatt, S C Campbell, J P Kemp, and D G Tinkelman. Dose-response study of nebulized bitolterol mesylate solution in asthmatic patients. *CHEST Journal*, 91(4):533–539, April 1987. 77

[219] Xinliang Mao and Aaron D. Schimmer. The toxicology of clioquinol. *Toxicology Letters*, 182(1–3):1–6, November 2008. 78

[220] E. G. Brown, L. Wood, and S. Wood. The medical dictionary for regulatory activities (MedDRA). *Drug Safety*, 20(2):109–117, 1999. 79

[221] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(Database issue):D267–270, January 2004. PMID: 14681409. 79

[222] OpenPHACTS. OpenPHACTS - http://www.openphacts.org/. 88

[223] www.etoxproject.eu. http://www.etoxproject.eu/. 88

[224] INBIOMEDvision Promoting and Monitoring Biomedical Informatics in Europe > HOMEPAGE. http://www.inbiomedvision.eu/index.html. 88