Review

From in silico target prediction to multi-target drug design: Current databases, methods and applications

Alexios Koutsouka, Benjamin Simmsb, Johannes Kirchma, Peter J. Bonda, Alan V. Whitmorec, d, Steven Zimmerc, Malcolm P. Youngc, e, Jeremy L. Jenkinsb, Meir Glickb, Robert C. Glena,⁎, Andreas Bendera,⁎

a Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom
b Lead Discovery Informatics, Center for Proteomic Chemistry, Novartis Institutes for Biomedical Research Inc., 250 Massachusetts Avenue, Cambridge, MA, 02139, USA
c e-Therapeutics plc, Holland Park, Holland Drive, Newcastle upon Tyne, NE2 4LZ, United Kingdom
d The School of Pharmacy, 29-39 Brunswick Square, London WC1N 1AX, United Kingdom
e Institute of Neuroscience, University of Newcastle, Framlington Place, Newcastle upon Tyne, NE2 4HH, United Kingdom

ARTICLE INFO

Article history:
Received 21 January 2011
Accepted 6 May 2011
Available online 18 May 2011

Keywords:
Target prediction
Polypharmacology
Mode of action
Target fishing
In silico
Off targets

ABSTRACT

Given the tremendous growth of bioactivity databases, the use of computational tools to predict protein targets of small molecules has been gaining importance in recent years. Applications span a wide range, from the ‘designed polypharmacology’ of compounds to mode-of-action analysis. In this review, we firstly survey databases that can be used for ligand-based target prediction and which have grown tremendously in size in the past. We furthermore outline methods for target prediction that exist, both based on the knowledge of bioactivities from the ligand side and methods that can be applied in situations when a protein structure is known. Applications of successful in silico target identification attempts are discussed in detail, which were based partly or in whole on computational target predictions in the first instance. This includes the authors’ own experience using target prediction tools, in this case considering phenotypic antibacterial screens and the analysis of high-throughput screening data. Finally, we will conclude with the prospective application of databases to not only predict, retrospectively, the protein targets of a small molecule, but also how to design ligands with desired polypharmacology in a prospective manner.

© 2011 Elsevier B.V. All rights reserved.

Contents

1. Introduction ......................................................... 2555
2. Public and proprietary databases making bioactivity data available. ................................ 2556

⁎ Corresponding authors. Tel.: +44 1223 336 432, +44 1223 762 983.
E-mail addresses: rgc28@cam.ac.uk (R.C. Glen), ab454@cam.ac.uk (A. Bender).

1874-3919/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jprot.2011.05.011
1. Introduction

‘Drug discovery’ has historically been based on phenotypic readouts on the organism level, such as the effect of herbs or other natural remedies on humans. In this case the desired phenotype (usually reversal of the diseased state of the organism back to healthy) was the relevant endpoint being considered, while the mode of action of many compounds (the precise reasons leading to this modulation) remained in the dark. In the more recent past reductionist approaches in the pharmaceutical community have traditionally seen drugs as ‘magic bullets’, a concept first established by Paul Ehrlich [1,2]. This concept states that drugs exert their activities by modulating one target of particular relevance to a disease, the famous idea of one ‘key’ (or ligand) modifying each ‘lock’ (or protein) [3]. This paradigm has guided the pharmaceutical industry throughout approximately the last three decades.

However, in recent years, there is mounting evidence offering a significant challenge to this hypothesis as it has become increasingly obvious that many drugs elicit their therapeutic activities by modulating multiple targets [4–7]. Recently, it has been estimated that each drug on the market possesses bioactivity against, on average, six experimentally confirmed protein targets (based on currently available data) [8]. In fact, in questioning the universal appropriateness of the classical single-target drug-discovery approach one needs to look no further than the large number of drugs that have failed due to insufficient clinical efficacy [9].

While a certain set of targets modulated by a small molecule seems to be in many cases beneficial for efficacy, there are also proteins one would rather avoid modulating (often called ‘antitargets’) [10]. Over the past few decades, a significant number of chemical compounds have failed to get approved and reach the market due to severe clinical side-effects and cross-reactivity that were observed during later-stage clinical trials, increasing attrition rates of new compounds [11,12]. The fact is that the multi-target interactions of drugs are either largely unknown or insufficiently understood in most cases [8], and examples of the truly prospective design of multi-target drugs are still rather scarce [6].

Hence, we need to accept that for a drug to have the desired effect in man we need to modulate a set of targets to achieve efficacy, while avoiding others to reduce the risk of side effects. A promising category of chemical compounds where promiscuity was only relatively recently seen as a virtue is kinase inhibitors [13,14], such as the Novartis compound Gleevec, which was once thought to be highly selective in targeting the Bcr–Abl fusion gene. However, more recently it has turned out that Gleevec (along with many other kinase inhibitors) also inhibits other kinases such as c-kit and PDGFR [15,16], enabling its further use against gastrointestinal stromal tumors [17]. While in this case, different targets are linked to different diseases (but the same compound), the kinase inhibitors Sorafenib and Sunitinib which were recently approved by the FDA, targeting VEGFR, PDGFR, FLT-3 and c-kit, are thought to achieve efficacy in the clinic by modulating different pathways also in the same disease [18]. On the other hand, and as an ‘atypical’ case, the polypharmacology of psychoactive compounds (such as antidepressants or antipsychotics) against a set of mainly dopamine and serotonin receptor subtypes has been known (and accepted as a special case) for some time. Such ‘polypharmacology’ [12] substantially changes the way we perceive the bioactivity of compounds, which, from a medicinal chemist’s point of view, has now evolved from a unidimensional into a multidimensional area to explore and optimize [19,20]. In this context, and this is the subject of this review, the use of target prediction tools can help us to better anticipate the bioactivity spectra of compounds earlier, and in a much more predictive fashion.

Here, computational methods, with the wealth of bioactivity information available today, can help us both to design ligands with a higher likelihood of exhibiting a desired bioactivity profile, as well as to anticipate secondary and alternative bioactivities in compounds at an early stage. Public databases such as ChEMBL [21,22], PubChem [23] and Chem-Bank [24] are continuously increasing in size, with ChEMBL in
its latest release, comprising about 700,000 distinct small molecules and about 2.7 million bioactivity data points. These public initiatives follow the example set by various bioinformatics initiatives globally, of providing open access to essential data; a trend that started in the bioinformatics world several decades ago with sequence databases distributed on magnetic tapes but that has only happened in the cheminformatics world in the last 5 years or so.

Bioactivity data of small molecules can be used in a variety of ways, ‘forwards’ (to predict protein targets of small molecules) as well as ‘backwards’ (to design ligands with desired bioactivity against a set of targets, based on knowledge contained in the databases). The ‘forward’ concept is visualized in Fig. 1, where a user is able to input a molecular structure into a computational bioactivity model, and where – without experiment – based on the database’s ‘educated guesses’, the ‘predicted’ bioactivities of a compound are generated. This review will cover both of the above domains; the rationalization of phenotypic observations related to compound application as well as the prospective design of novel molecules with the desired activities. It is an extension of previous publications in the field [25–27] due to the largely increased availability of data, as well as methods, relating to the area.

In the following, we will firstly (Section 2) summarize which databases are at our disposal at the current stage, an area under tremendous development. We will continue with an overview of ligand-based (Section 3) and structure-based (Section 4) methods which can be used to predict protein targets for small molecule ligands. These sections are followed by published applications of target prediction tools (Section 5), as well as two applications carried out by the authors, namely the prediction of targets for antibacterial compounds in a high-throughput screening settings (Section 6) as well as the application of multi-target affinity fingerprints for scaffold hopping (Section 7). We will conclude with the prospective utilization of bioactivity databases and address the important question how to decide which targets to modulate (Section 8), before concluding this review (Section 9).

2. Public and proprietary databases making bioactivity data available

The field of public, and private, bioactivity databases is large and continuously growing. One cannot claim to present a complete overview of bioactivity data in a review such as this one, however due to the importance of this type of data at least an overview of many of the major public databases shall be listed here, listed in Table 1 (for more information see a recently published review on the field [28]; for a list of commercial databases see a previous review [25]).

A remarkable development in recent years is that databases have not only grown in size, but they have also started to integrate diverse types of data, such as phenotypic data and drug side effects [29], both of which were hardly accessible in electronic form in public previously. Two of the most comprehensive integrated resources available today are probably Chem2Bio2RDF [30] and PROMISCUOUS [31], both linking chemical data to pathway information and phenotypic data as well as merging side effect information within this integrated data resource. Both databases differ in many practical details, such as the data sources used, the database implementation, and the ways in which they can be queried. However, their common aim is the integration of different data types — an area of great interest also to pharmaceutical companies, which is still very much under active investigation, not least due to the heterogeneity of the data at hand. This is closely linked to the development of ontologies such as the Small Molecule Ontology (SMO) [32], which allows for an easier integration of different data sources using relatively recently developed Semantic Web technologies such as the Resource Description Framework (RDF) and Web Ontology Language (OWL). Conversion of formerly relational databases to RDF triples seems to be a general trend in the cheminformatics (and related) fields due to the flexibility of the data that can be represented; however the comparative rigidity of relational databases can also be seen as an advantage in some cases where data needs to adhere to a predefined structure for data mining and similar applications.

![Fig. 1 – An in silico target prediction workflow. In order to enable a computational model to predict protein targets of small molecules, firstly target class models for up to multiple thousand proteins need to be constructed which takes only hours for circular fingerprints in combination with the Naïve Bayes Classifier. Afterwards, a chemical structure input in the model (shown on the left) can be annotated within a fraction of a second with its most likely macromolecular interaction partners (shown on the right).](image)
What is clearly visible from Table 1 is that millions of bioactivity data points are available, which can be used for ligand-based target prediction (Fig. 1). Based on various data sources also internet-based public resources for target prediction are available which are listed in Table 2.

### Table 1 – Overview of some of the major public bioactivity data resources available today. Due to the fast development of the area this list does not claim to be exhaustive in any way.

<table>
<thead>
<tr>
<th>Databases</th>
<th>Data</th>
<th>Web address</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChEMBL [21]</td>
<td>&gt;700k small molecules with &gt;2.7 million bioactivity data points</td>
<td><a href="http://www.ebi.ac.uk/chembl">http://www.ebi.ac.uk/chembl</a></td>
</tr>
<tr>
<td>DrugBank [150]</td>
<td>4800 drug entries including: &gt;1350 FDA-approved drugs and</td>
<td><a href="http://www.drugbank.ca/">http://www.drugbank.ca/</a></td>
</tr>
<tr>
<td>Comparative</td>
<td>6000 compounds, 1.4 million chemical–gene–disease data points</td>
<td><a href="http://ctd.miblb.org/">http://ctd.miblb.org/</a></td>
</tr>
<tr>
<td>Toxicogenomics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Database (CTD) [151]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SuperTarget [152]</td>
<td>1500 drugs, 2500 targets proteins and 7300 drug-target interactions</td>
<td><a href="http://bioinf-tomcat.charite.de">http://bioinf-tomcat.charite.de</a></td>
</tr>
<tr>
<td>MATADOR [152]</td>
<td>Manually annotated compounds from the SuperTarget database</td>
<td><a href="http://matador.embl.de/">http://matador.embl.de/</a></td>
</tr>
<tr>
<td>Therapeutic Target</td>
<td>Contains 1906 targets, including 358 successful, 251 clinical trial,</td>
<td><a href="http://xin.cz3.nus.edu.sg">http://xin.cz3.nus.edu.sg</a></td>
</tr>
<tr>
<td>Database (TTD) [153]</td>
<td>43 discontinued and 1254 research targets, and 5124 drugs,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>including 1511 approved, 1118 clinical trial and 2331 experimental drugs</td>
<td></td>
</tr>
<tr>
<td>BioActivity [23]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BindingDB [39]</td>
<td>&gt;271k compounds, &gt;620k binding affinities against 5526 protein targets</td>
<td><a href="http://www.bindingdb.org">http://www.bindingdb.org</a></td>
</tr>
<tr>
<td>DrugPort (EBI)</td>
<td>Contains 1492 approved drugs and 1664 unique protein targets</td>
<td><a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a></td>
</tr>
<tr>
<td>Potential Drug Target</td>
<td>Contains 1207 entries covering 841 known and</td>
<td><a href="http://www.dddc.ac.cn/pdtd/">http://www.dddc.ac.cn/pdtd/</a></td>
</tr>
<tr>
<td>Database (PDTD) [80]</td>
<td>potential drug targets with structures from the Protein Data Bank (PDB)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;25k compounds, as well as drug side effects and various other data;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>includes protein network visualization</td>
<td></td>
</tr>
<tr>
<td>T3DB [154]</td>
<td>Toxin and Toxin-Target Database; &gt;2900 toxins,</td>
<td><a href="http://www.t3db.org/">http://www.t3db.org/</a></td>
</tr>
<tr>
<td></td>
<td>1300 toxin targets and &gt;33,000 toxin-target associations</td>
<td></td>
</tr>
</tbody>
</table>

What is clearly visible from Table 1 is that millions of bioactivity data points are available, which can be used for ligand-based target prediction (Fig. 1). Based on various data sources also internet-based public resources for target prediction are available which are listed in Table 2.

### 3. Ligand-based methods for protein target prediction

#### 3.1. In silico target prediction based on single structures

The most simple target predictions are based on individual molecules. Ligand-based methods include methods that are based on the principle of chemical similarity, which states that similar chemical structures tend to present similar biological activities more often than not [33,34], although even structurally similar compounds may interact with a protein target in different ways [35]. These methods rely on prior knowledge of bioactive ligands and protein structures and can mine in a short time huge chemical libraries in order to locate the most chemically related structures that share substructural similarities and subsequently interact with similar protein targets. However, they are generally not applicable in cases where no ligands for a given protein exist, since in these situations no training on ligand-based information is possible.

One of the questions that need to be answered is how to describe a molecular structure to a computer — and no universal answer exists here. Recent studies compared [36,37] (and discussed in more detail [38]) currently available molecular descriptors and to which extent they are correlated to bioactivity. Here it was found that circular fingerprints, counts of circular fingerprints, predefined keys, atom pairs and triangles behave significantly distinct — and that in most cases circular
fingerprints in their various definitions seem to have best correlation with bioactivity. Hence, they might also be one of the most appropriate methods to use in target prediction attempts.

Target prediction methods focused on individual molecules are essentially a similarity searching exercise: from a bioactivity database, such as ChEMBL [21] or BindingDB [39], similar structures to the one under consideration are identified. Experimentally measured activities from the database are then used as a guideline for the bioactivity of the molecule under consideration. While this is a rapid method to predict targets, its main shortcoming is that it results in insufficient extrapolation in practice since only molecules as a whole are compared to each other. Let us for this purpose consider the case where a database contains molecules consisting of fragments A–B and fragments C–D as active against a particular protein target. Using similarity searching alone, a new molecule A–C (or A–D, B–C, B–D) would not necessarily be identified as similar to the database-identified molecules. Machine learning models discussed in the following though would likely have higher chances of success to be able to capture this novel combination of features in a previously unseen molecular structure.

3.2. Target prediction based on data mining methods

Another increasingly important field used for bioactivity prediction of chemical compounds is data mining methods. These methods, in contrast to chemical similarity methods which simply identify potential targets based on ligand similarities of molecular substructures, perform pattern recognition in vast chemogenomic space and cluster compounds that present similarities in multidimensional space. These methods steadily gain preference and credibility as they are able to identify and correlate relationships among large numbers of compounds that otherwise would be impossible to connect based substructural similarities alone. Machine learning methods such as Support Vector Machines (SVMs) and Linear Regression Models (LRMs) are the driving force of these approaches. A significant number of methods that employ data mining approaches for mapping the chemogenomic space have been published during the last decade.

Nidhi et al. applied Multiple-Category Bayesian models in order to mine chemogenomics space [40]. The model was trained on extended-connectivity fingerprints of compounds from 964 target classes in the WOMBAT (World Of Molecular BioAcTivity) chemogenomics database and then was applied to predict the 3-top most likely protein targets for all MDDR (MDL Drug Database Report) database compounds. While much of the early work of protein target prediction models was based on the Bayes Classifier [41], more recently, implementations of the Winnow algorithm for target prediction have been reported [42]. While major difference in the overall performance in target prediction were not shown, both algorithms did predict slightly different targets, which demonstrates the importance of having multiple algorithms available. In addition, in situations where new bioactivity data become available, complete retraining of the model is not required in case of the Winnow algorithm, which may well be of relevance where frequent additions of novel data are made to existing models.

A related method – which was one of the first target prediction methods published – is the PASS method [43]. This method is currently able to predict more than 4130 biological activities based on the structural chemical formula of a compound alone. The development of this method required analysis of structure-activity relationships for more than 270,000 biologically active compounds. Currently, this method is reported to predict with an accuracy of about 95%.

A target prediction method that is based on the distribution of features in each bioactivity class of molecules is the Similarity Ensemble Approach (SEA) [5,44], which employs a BLAST-derived algorithm to develop minimal spanning trees considering chemical similarity. SEA estimates target similarity, which is ranked based on their ligands’ chemical similarity. The authors applied the SEA method to large scale study of more than 3000 FDA-approved drugs against hundreds targets and identified 23 new cases of drug-target interactions, five of which were validated in vitro to be potent with affinities less than 100 nM. Moreover, the SEA method was applied to examine for potential off-target inhibition by commercially available drugs against the enzyme farnesyltransferase (PFTase) [45]. Here, loratadine and miconazole were identified and experimentally confirmed as ligands for PFTase.

Other machine learning methods have been applied to fingerprint-based target prediction. Wale [46] used ECFP4 fingerprints on a PubChem-derived dataset comprising 231 targets and 40,170 ligands and compared a Bayes Classifier with binary SVM, a ranking-based SVM, cascaded SVM, Ranking Perceptron and the combination of SVM and a Ranking Perceptron with respect to their ability to predict targets of small molecules. Here it was found that the cascaded SVM methods generally outperformed the Bayes Classifier, while on ligands which belong to multiple categories the combination of SVM and Ranking Perceptron are the best-performing method on this dataset. Given that much target prediction literature was based on the Bayes Classifier alone, this indicates that it would be beneficial to employ – and benchmark – other machine learning methods in the context of the target prediction of small molecules. SVMs were also employed for target prediction in a different recent publication [47], albeit on a smaller dataset comprising only five activity classes. Here, consistently more than 80% sensitivity for the target predictions has been obtained.

3.3. Chemogenomic approaches

Chemogenomic approaches attempt to systematically analyze (‘all’) ligand–target interaction space in toto[48–51]. Frequently, protein targets are, in the spirit of chemogenomics, classified not according to sequence or fold, but according to the similarity of their ligands. Given a ligand-based classification of protein targets, one can analyze which targets are likely to be hit by a ligand, given its structure [52,53].

In this area, recent publications have appeared which have tried to evaluate to what extent novel ligand–protein target pairs could be predicted, based on bioactivity knowledge of a set of ligands and a set of targets on the one hand, or on the relationships of the targets themselves. (As discussed above for the ligand side, also for the protein side no single descriptor exists, so sequences, binding pockets and selected subsets of residues in ligand binding pockets have all been used to
capture the essential properties of a protein for binding ligands.) In this spirit, groups have employed SVMs [54] as well as substructural analysis [53] in order to relate GPCRs from the ligand-based side — and to estimate how well ligand–target pairings could be predicted also in cases where no ligands of a particular target are known. In the latter of the above studies [53] it was found that in those cases, 93% of the targets ligands could be identified with a reliability beyond random even if no ligands were given in the training dataset (this resembles a ‘receptor dephosphonation’ setting in practice), thereby illustrating that not only explicitly known ligand–target pairs can be predicted by computational models.

Linear regression models have also been applied to study drug-target interactions. Zhao et al. [55] developed a computational framework, termed drugCIPHER, which is based on the observed correlation in pharmacological and genomic spaces. They proposed three linear regression models that relate therapeutic similarity, chemical similarity and their combinations to the relevance of the targets based on protein–protein interaction network.

Also in the area of chemogenomics-driven predictive approaches to identifying ligand–target pairs falls a novel low-dimensional protein–ligand fingerprint-based (PLFP) method was recently reported [56] encoding both chemical and biological properties and which is suitable for exploring chemogenomics space. PLFP method captures the pharmacophore properties of ligands and their respective transmembrane binding cavities and was found to show preference for SVM classifiers, where it achieved nearly 90% precision in determining true from false pairs.

Furthermore, topological descriptors based on molecular features have been shown to be especially successful in describing and comparing molecular profiles. Two novel sets of topological molecular descriptors have been recently reported; SHED (SHannon Entropy Descriptors) [57] and RED (Renyi entropy descriptors) [58]. SHED descriptors are calculated from distributions of atom-centered feature pairs extracted from the topology of molecules and their scores reflect the distribution of pharmacophore features in a molecular constitute. Gregori-Puigjané et al. [59] applied SHED descriptors to mine chemogenomics space of more than 700 drugs against 600 targets. Their results conclude that compounds targeting aminergic G protein-coupled receptors (GPCRs) exhibit the most promiscuous pharmacological profiles seen among drugs. On the other hand, RED descriptors are based on generalized Renyi entropy as a variability measure for a feature-pair distribution in contrast to SHED descriptors.

Also scaffold compositions can be used to predict the bioactivity of compounds on protein targets [60]. In this particular work, about 24,000 unique molecular scaffolds were extracted from 458 different bioactivity classes and a hierarchical analysis of scaffolds contained in each bioactivity class was carried out. As a validation of the method, bioactivities for ‘virtual’ scaffolds were performed, and indeed compounds could be found in external tests sets with the predicted bioactivities.

3.4 **Pharmacophore-based methods**

The first software module supporting the profiling of bioactivity spectra using a pharmacophore-based approach was LigandProfiler, distributed with Discovery Studio [61]. Ligand Profiler is based on a Pipeline Pilot [62] workflow for automated parallel screening. The program allows screening of collections of compounds (single- or multi-conformational databases) against a series of pharmacophore models defined by the user (for a recent review see [63]). Technically, the ligand profiler maps each input ligand against all selected pharmacophore models. A fit value is computed that is a measurement of how well the compound maps the chemical function-based features of the pharmacophore and the results can be displayed as a heat map visualizing which compounds are likely to bind to which targets. KNIME is an alternative pipelining software allowing easy integration of such parallel screening workflows [64].

The Inte:Ligand PharmacophoreDB [65] currently includes 2500 annotated structure- and ligand-based pharmacophore models covering 300 unique biological targets. All structure-based models of this database were developed using LigandScout [66] and are provided in Accelrys Catalyst data format.

LigandScout generates pharmacophore models based on a given 3D structure of a ligand–protein complex in a fully automated manner. After clean-up of the binding site, chemical features are identified on the ligand site, while searching for corresponding features in the neighboring protein environment. Pharmacophore features are placed at positions with favorable protein–ligand interactions. Addressing tautomeric effects is a crucial aspect when it comes to the generation of pharmacophore models, in particular with respect to hydrogen bonding patterns [67]. LigandScout examines different tautomeric forms and indicates the most-likely one(s). To account for the shape properties of the binding site, excluded volume spheres can be added on moieties forming the protein surface. This avoids the identification of bulky compounds during virtual screening, which are unlikely to be accommodated in the target interaction site.

Also a free web interface (see Table 2) has been published which employs pharmacophores to predict protein targets of small molecules, entitled PharmMapper [68]. In order to derive a dataset of pharmacophores structures complexed with small molecules were derived from DrugBank, BindingDB, PDBBind and the PDTD database. Subsequently, pharmacophore models were constructed also in LigandScout, resulting in a total of 7302 pharmacophore models, out of which 2241 are covering human proteins. Analyzing the targets for tamoxifen an AUC of the true protein interaction partners of 0.7 could be achieved.

3.5 **How to extend the scope of target prediction models: Extrapolating in ligand space**

When predicting protein targets for a small molecule using ligand-based methods, limitations exist regarding the scope of the training datasets, both in chemical space as well as biological space. This means in practice that not for all chemical classes reliable target predictions can be obtained, and that for proteins, which do not have bioactivity data points available, no target predictions can be made at all. However, both in ligand (chemical) space and protein target (biological) space extrapolations are possible to some extent. Extrapolations in ligand space are largely dependent on the descriptor chosen. In previous work from our groups we explored both two-dimensional descriptors (fingerprints) and
three-dimensional descriptors (FEPOPS) with respect to their ability of predicting protein targets of ligands where no similar structures were contained in the training set. It could be shown that more ‘abstract’ 3D descriptors, such as FEPOPS which just consists of a small number of four pharmacophoric points for each molecule, are able to outperform the more ‘exact’ 2D fingerprints in those cases where no similar ligand-target associations are known [69,70].

Also a hybrid 2D/3D target prediction, ReverseScreen3D, has been published recently [71]. Firstly, 2D fingerprints are used to calculate the similarity of the query molecule to every ligand structure in the ligand–target interaction database, which has in turn been derived from a clustering of the PDB. In an optional second step, the database ligand of each target cluster with the highest similarity to the query molecule is then used as a template for which 3D matching of the query is performed. For 20 compounds a validation procedure was performed and experimental targets could be confirmed in the majority of cases. 2D and 3D predictions often show complementary results, and on a 4-OH tamoxifen case study three other methods (INVDOCK, TarFisDock and PharmMapper) could be outperformed by ReverseScreen3D. The method is also available to the public via a web interface (see Table 2 for details).

3.6. How to extend the scope of target prediction models: Extrapolating in target space

Also on the target side extrapolation abilities are often required — such as in cases where most of the bioactivity data available relate to human proteins, but one is interested in predicting protein targets in a pathogen instead. One way to address this problem is to annotate proteins with protein folds — and map back to pathogen proteins after predicting protein folds targeted by a small molecule. In a recently publication InterPro domains were employed for this purpose [72]. Compounds that bind to a specific protein domain may also bind to other proteins that share the same InterPro domain, even if they share low sequence similarity. This in silico ‘Domain Fishing Model’ (DFM) is capable of extrapolating to targets outside training sets. Furthermore, DFM can be employed to triage the results of affinity chromatography experiments. The DFM was applied by Prathipati et al. [73] to elicit protein targets of 44 antibacterial (antiTB) compounds, where it successfully predicted after validation the targets of two novel antiTB compounds, CBR-2092 and Amiclenomycin.

Another possibility of extrapolating in target space are so-called ‘proteochemometrics’ (PCM) models which include, apart from descriptors of the ligands, descriptors of the target side [74-76]. These methods are related to the chemogenomics methods illustrated above and recently reviewed; [48] however, as opposed to chemogenomics methods ligand and protein, features are explicitly included as input variables in the PCM model, with the bioactivity being the dependent variable. In prospective studies (which are currently being prepared for publication) the authors have also introduced the concept of ‘leave-one-target-out’ cross validation in order to judge the ability of the bioactivity model to predict bioactivities of ligands for related targets — and it was found on the enzyme HIV reverse transcriptase that, indeed, for 12 out of 14 mutants predictivity of the bioactivity model could be obtained that approached the reproducibility of the assay. (In the remaining two cases no related targets were present in the training set; hence, no reliable bioactivity prediction was possible in those cases.)

Also the so-called ‘signature descriptors’ fall into this category of approaches which simultaneously employ ligand as well as target features for enzyme-metabolite interaction modeling [77]. Employing SVMs in combination with Signature Kernels, a five-fold cross validation on 873 drug–target pairs comprising 121 targets as well as 551 drugs yielded interaction accuracies >85%.

As summarized above, a multitude of target prediction methods, based on small-molecule information, exists in the current literature. Given the ever increasing amount of bioactivity data available it can be expected that their performance will continue to increase in the future, as will their applications in predicting targets for compounds with an unknown bioactivity profile.

4. Structure-based methods for protein target prediction

Structure based methods, in the context of drug design, generally describe approaches that exploit protein structural information, combined with scoring functions, in order to predict ligand–target pairs. More conventionally, ligands are identified as interaction partners for a given protein, however more recently this concept has also been reverted to dock one small molecule against a panel of multiple receptors (‘inverse docking’) which is then applied to target prediction. For a more detailed overview of this particular field the reader is also referred to a recent review on the topic [78].

In particular, the docking-based methods INVDOCK [79] and TarFisDock [80] have recently been presented which aim at elucidating the protein target of a small molecule, based on docking against a panel of proteins putatively interacting with the ligand. INVDOCK is able to consider flexibility of the ligand (but not of the protein) and employs a custom-made scoring function, while TarFisDock is based on the software DOCK and is also available publicly in a web interface (see Table 2). Applications of the programs are given in the following section.

More recently, Kellenberger et al. [81] performed a study of currently available methods for in silico reverse screening used for target prediction in order to examine how the algorithms for ranking protein targets perform. As it was found scoring functions still remain the main weakness of virtual screening approaches. Several ranking protocols based on GOLD fitness score and topological molecular interaction fingerprint (IIF) comparison were evaluated in that study, showing that problems associated with accuracy and false positives are present and need to be addressed.

Also for endogenous ligands docking has been used to identify substrate–protein pairs [82]. The conclusion of this work was that proteins (and substrates) are more promiscuous in nature than one might expect. However, this result needs to be considered also in context with the predictivity of current scoring functions, which might not be too reliable an estimate of binding affinities encountered in experiments. A recent review [83] even states that, on the particular dataset chosen and employing a set of 37 scoring functions, “we have demonstrated that for the eight proteins of seven evolutionarily diverse target
types studied in this evaluation, no statistically significant relationship existed between docking scores and ligand affinity performed large scale comparison of available docking programs and scoring functions.” However, it was also found that the docking engines of the programs were in many cases able to generate ligand binding poses, and pick them among the energetically favorable ones, in many cases — hence, while numerically assigning binding affinities through scoring functions seems to be difficult at the current stage, performing classifications into binders and non-binders seems to be more successful in many cases.

While ligand-based methods are fast, docking-based methods take considerably more computational resources for a docking run against hundreds, or even thousands of targets while still not achieving reliable results. Hence, turning to the more computationally intense area of molecular dynamics simulations may in the future, with increasing computer power, become more and more important to predicting ligand–protein binding affinities.

Docking seeks to rank poses ordered by a score, which ideally correlates with free energy of binding. However, only limited flexibility in the protein is considered, and solvent effects are generally excluded. Poisson–Boltzmann based approaches can provide a more rigorous estimate of binding energies, but their accuracy tends to be dependent upon the configuration of the ligand [84]. A more detailed picture of ligand binding may be obtained via simulation-based approaches. All-atom molecular dynamics (MD) simulations provide a means to characterize the motions of biological macromolecular complexes in the context of the ligand-bound and free states, and to explicitly include the effects of water [85]. While simulation approaches tend to be computationally demanding, a number of developments have provided a means to accurately calculate relative and absolute binding free energies and their component contributions in silico, making such methods potentially useful in the latter optimization stages of drug design.

Rigorous simulation-based approaches to the estimation of binding free energies generally rely upon the use of a thermodynamic cycle, in which a change in binding free energy is calculated as the difference between the free energy of converting one ligand to another in the bound state, minus the free energy of the same change in solution [86]. This can also be useful for understanding the ligand binding energetics associated with individual amino acid substitutions, which may be of interest in the context of specificity within a closely related family of proteins (Fig. 2, example taken from [157]). In the free energy perturbation (FEP) approach, the free energy change is measured as one ligand is “alchemically” transformed into another. For a converged estimate, the configurations of the two ligands sampled must be similar (their “phase space” must overlap); as a result, a series of non-physical intermediate “windows” between the two ligands are introduced [87]. Thus, while such methods may be too time consuming for screening large compound databases, they are likely to prove useful toward the end of the drug design process, when attempting to optimize a limited number of ligands rationally [88]. No target prediction methods have been reported yet that make use of large-scale molecular dynamics simulations, however given their accurate results this may change in the foreseeable future.

In conclusion, although structure-based target prediction have made significant advances and improvements over the last decade and can now be applied to real-life problems, there are still significant limitations that prohibit them from attracting a wider acceptance in the pharmaceutical industry. The main current limitations include that expert knowledge is often required to operate the software, they remain computationally very intensive and there are limitations imposed by the availability of experimental data such as X-ray structures for membrane-bound proteins. In addition, scoring functions still need improvement in order to estimate ligand binding affinities reliable. On the other side, structure-based methods are not limited to the chemical ligand classes covered in a training set in the way ligand-based target prediction methods are, so they might very well become more versatile in their applications in the future once the above problems have been alleviated.

5. Applications of target prediction models

A significant number of applications of in silico target prediction models have been published recently, ranging from the mode-of-action elucidation of compounds to the analysis of false positives in reporter gene assays, which we will summarize in the current section.

5.1. Applications of machine learning based target prediction methods

The ‘Similarity Ensemble Approach’ (SEA) described above [5] has recently been applied in a prospective manner [44] to elucidate additional targets for drugs on the market. Calculating the similarity of 3665 FDA-approved and investigational drugs against hundreds of targets a large number of novel association have been uncovered. Out of 30 experimentally tested associations 23 were confirmed, five of which were potent (below 100 nM). This was among others the case for the antagonism of the β1 receptor by fluoxetine (Prozac), the inhibition of the serotonin transporter by ifenprodil (Vadilux) and antagonism of the H4 receptor by delavirdine (Rescriptor). This prospective validation indeed supports hope that, both for predicting off-targets [89,90] as well as novel indications, in silico target prediction tools based on ligand similarity may play a role of growing importance in the future.

Earlier work investigated 130 of the ‘top 200 medicines’ on the market using the PASS method and in 93.2% of the cases known pharmacological effects could be found in the bioactivity spectra [91]. In this case no experimental validation was provided directly in the publication. Shortly afterwards however, PASS was applied to the discovery of new cognition enhancers [92], and in this case even validation using animal models was successfully performed on a novel chemical series. Out of 8 compounds tested in a model of scopolamine-induced passive avoidance reflex (PAR) amnesia all showed cognition enhancement, three at a dose of 1 mg/kg and five at a dose of 10 mg/kg. This is probably one of the few truly prospective validations of novel therapeutic areas of a series of compounds, based on in silico target prediction tools.
Also fingerprint-based target prediction models, based on the WOMBAT database and Naïve Bayes models, have been applied in practice, such as to the rationalization of false positives in reporter gene assays [93]. Here it was found that false positives in reporter gene assays very often are predicted to interact with kinases involved in cell cycle progression — hence, a mechanistic hypothesis for false positives could be generated. This hypothesis was prospectively validated in Fig. 2 – Relative binding affinities from free energy perturbation. A peptide ligand (white cartoon format) is shown in its bound state to an SH2 domain (ice blue cartoon format); note the presence of key water molecules which may affect the binding affinity. In this example, the relative free energy of ligand binding is calculated in the context of an amino acid mutation in the binding pocket, from Tyr (cyan wireframe format) to Leu (green wireframe format). The free energy difference for binding to the protein containing Tyr ($\Delta G_{\text{bind}}^1$) or Leu ($\Delta G_{\text{bind}}^2$) is unknown, but according to the thermodynamic cycle shown, 

$$\Delta \Delta G = \Delta G_{\text{bind}}^2 - \Delta G_{\text{bind}}^1 = \Delta G_{\text{mut}}^2 - \Delta G_{\text{mut}}^1.$$  

To obtain the quantities $\Delta G_{\text{mut}}^1$ and $\Delta G_{\text{mut}}^2$, two alchemical calculations are carried out, during which the Tyr residue is slowly “mutated” to Leu, in the presence and absence of peptide ligand, respectively. In the “dual topology” approach, both amino acids are present at the same time (as shown in the figure), but their interaction with the environment is scaled according to a coupling parameter $\lambda$, which varies between 0 (equivalent to Tyr, the initial state) and 1 (equivalent to Leu, the final state). The change in energy for such a mutation is obtained from a series of $N$ overlapping simulation “windows” as a function of $\lambda$:

$$\Delta G_{\text{Tyr} \rightarrow \text{Leu}} = -k_BT \sum_{i=1}^{N} \ln \left\{ \exp \frac{-E_{\lambda_i} - E_{\lambda_i}}{k_BT} \right\}_{\lambda_i}$$

where (...) implies an ensemble average over configurations of state $i$, $k_B$ is the Boltzmann constant, and $T$ is the temperature.
cytotoxicity assays, where it was indeed shown that >50% of false positives in reporter gene assays possess cytotoxic properties [93].

Another application of this class of models was applied to the analysis of ‘hits’ in High-Content Screening [94]. While providing the screening with a cellular phenotype, high-content screening readouts are very often cryptic with respect to the underlying mode of action which is responsible for the particular phenotype encountered [95,96]. In this recent study [94] high-content screening readouts were combined with in silico target predictions as well as an analysis of the structural similarity of compounds. It was found that all three domains, chemical structure, predicted targets/mode of action, as well as phenotype, are to some degree independent and that no one-to-one mapping between any two of the spaces can be performed, hinting at the importance of using information from all three domains when analyzing the biomodulatory capabilities of a molecular structure.

The recent publication [97] of a set of 13 antipsychotic drugs against 34 protein targets prompted the equivalent analysis using in silico methods. It was found [98] that 65% of the 442 affinities could be predicted within one affinity log unit with the level of precision being above 92%, rendering the computational interaction predictions rather reliable (more than 9 out of 10 predictions would then be confirmed experimentally, with only one false positive prediction present).

For St. John’s Wort (Hypericum perforatum) a recent publication [99] attempted to demonstrate the capabilities of computational methods in the bioactivity spectra prediction of active constituents of plants and, more generally, compounds of natural origin.

One of the earlier applications in the field [100] employed SVMs to generate classification models for a total of 125 molecular functions which were derived from the MDDR database. Models were generated using MDL keys for a total set of 70 molecular actions and 55 therapeutic areas and 871 marketed drugs were analyzed with respect to their in silico bioactivity profile. In addition, adverse effects related to the liver were analyzed with respect to the highest correlated molecular functions, and several of the links were sensible as a first approximation — such as inhibition of HMG-CoA reductase, which is a class of drugs with known association with hepatotoxicity.

Models predicting the interactions of drugs, encoded using presence/absence vectors of functional groups, with proteins, encoded using physicochemical property descriptors have also been published [101]. However, given that only the target family (enzymes, ion channels, G-protein-coupled receptors and nuclear receptors) are predicted as interaction partners for each ligand, the practical applicability of the model will likely be limited. The same four bioactivity classes were used employing bipartite local models, however in this case also individual proteins were predicted as putative targets for a set of the top-scoring ligand–target pairs which is very relevant information to have in practice [102]. In this case, firstly target proteins of a given drugs were predicted, followed by the prediction of drugs of a given target protein. In this way, the performance predicting ligand–target interactions could be improved over previous methods.

Focusing on interactions between metabolites and enzymes, Chen et al. [103] employed a dataset from the KEGG database, where ligands were described by a graph-based similarity measure, and proteins were described using either protein functional domain composition, sequence identity as determined by the BLAST algorithm or a similarity of the GO annotations of each protein. Here it was found that employing multiple nearest-neighbor models using only parts of the negative training data in each case followed by majority voting outperformed each individual target prediction model.

5.2. Applications of pharmacophore-based target prediction methods

Several examples of successful pharmacophore-based bioactivity profiling have been reported (for more details see a recent review [104]). A Pipeline Pilot-based parallel screening system using structure-based pharmacophore models was used to classify 100 antiviral compounds. All of these test compounds are known to exhibit activity on one particular target. 50 pharmacophore models for five different targets (i.e., 10 pharmacophore models per target) were developed using LigandScout. The test compounds were then screened against all pharmacophore models and the prediction of activity of a certain compound on a certain target was based on a majority vote of the respective pharmacophore models. It was found that in 88% of the cases the activity of the test compounds was correctly predicted [105]. In another experiment, the selectivity of pharmacophore models of HIV protease was investigated using a test set of HIV protease inhibitors and inhibitors known to be active on other proteases, as well as inactive compounds. The results indicate that the utilization of ensembles of pharmacophore models and ensemble voting may help to enhance the signal-to-noise ratio. While the retrieval rate of known active compounds may be significantly increased, the selectivity of the pharmacophore-based approach may be preserved [106]. This approach also been employed to screen molecules for interactions with different cytochrome P450 isoforms and other ADME- and anti-targets [107].

Such a parallel screening setup was also applied in a target fishing experiment for 357 known peroxisome proliferator activated receptor (PPAR) agents. While PPAR targets were ranked first among all 181 targets screened, several new potential targets for these PPAR agents were identified. In a related approach, the plant constituents of Prasaplai, a Thai traditional medicine, which consists of 10 plants, camphor and sodium chloride, were analyzed for COX inhibiting compounds using a collection of pharmacophore models. It was shown that compounds known or suspected to be binding to COX could be identified by in silico modeling [108]. A database of Chinese herbal constituents was investigated for activity on four targets involved in inflammation (cyclo-oxygenases 1 and 2, COX; p38 MAP kinase, p38; c-Jun terminal-NH2 kinase, JNK and type 4 cAMP-specific phosphodiesterase, PDE4) using a collection of pharmacophore models. The study revealed several agents that are expected to exhibit biological effects on one or more targets under investigation [109]. 16 main constituents isolated from Ruta graveolens L. were screened against the Inte:Ligand PharmacophoreDB for activity on acetylcholinesterase (AChE), cannabinoid receptor 2 and human rhinovirus (HRV) coat protein. The compounds were also tested experimentally and in the majority of cases the activities were predicted correctly by
5.3. Applications of docking-based target prediction methods

Muller et al. [111] using high-throughput docking screening approach examined more than 2000 druggable active sites from sc-PDB to identify putative biological targets of derivatives of the 1,3,5-triazepan-2,6-dione scaffold. Specifically, they focused on five molecules representatives of this combinatorial library. Among promising biological targets that were identified and verified in vitro was the secreted phospholipase A2 (sPLA2).

Traditional Chinese Medicine (TCM) in recent years has attracted attention of the pharmaceutical community as a potentially alternative option to conventional western-medicine due to its in some cases validated phenotypic efficacy. The main problem associated with chemical compounds found in TCM is that in most cases the biological targets upon which this compounds act are not yet known. Thus, there has been an increased interest exploring these compounds using in silico target prediction strategies. One such large scale study was reported by Chen et al. [112] who mined scientific literature to collect a large amount of available information on TCMs. On the herbal ingredients that have been identified analytically INVDOCK [79,113,114] was applied as an inverse docking algorithm for identifying their potential molecular targets. Their result lead to creation of the Traditional Chinese Medicine Information Database TCM-ID, a web accessible free of charge database that includes information for prescriptions, constituent herbs, herbal ingredients, molecular structure and functional properties of active ingredients, therapeutic and side effects, clinical indication and application and related matters [112]. Another study focused on TCM compounds was conducted by Zahler et al. [115], who applied inverse screening method in order to identify potential kinase targets for three derivatives of indirubin, namely 5-bromo-indirubin-3′-oxime (5BIO), 6-bromo-indirubin-3′-oxime (6BIO), 7-bromo-indirubin-3′-oxime (7BIO). A total number of 84 unique protein kinases were examined in this study, with structural target data being drawn from sc-PDB. Indirubin, an active compound was recently conducted by Zahler et al. [115], who applied inverse screening method in order to identify potential kinase targets for three derivatives of indirubin, namely 5-bromo-indirubin-3′-oxime (5BIO), 6-bromo-indirubin-3′-oxime (6BIO), 7-bromo-indirubin-3′-oxime (7BIO). A total number of 84 unique protein kinases were examined in this study, with structural target data being drawn from sc-PDB. Indirubin, an active compound was recently identified to have potent therapeutic activities against chronic myelogenous leukemia (CML) [116].

INVDOCK was also used in recent work attempting to rationalize adverse drug reactions [117]. In this particular study a total of eleven HIV protease, nucleoside reverse transcriptase and non-nucleoside reverse transcriptase inhibitors were analyzed. Out of the complete set of putative interaction partners several protein targets were consistently predicted to be interaction partners of drugs, such as DNA polymerase beta, DNA topoisomerase 1, FK binding protein 12 and the sterol regulatory element binding protein-1. Overall it was found that more than 86% of the adverse drug reactions predicted by INVDOCK are consistent with the adverse reactions reported in literature. Hence, it could be shown that also docking-based target prediction methods are potentially of use in practice, not only for the anticipation of on-target effects, but also to rationalize (and potentially predict) adverse drug reactions via their association off-targets. TarFisDock on the other hand was applied to predict targets of the toxin 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD), and it was suggested that also aryl hydrocarbon receptor (AhR) independent pathways might play a role in its toxicity [118].

Finally, also integrated approaches for target prediction of small molecules exist, such as those based on gene expression [119] or metabolic profiling data [120]. However, this falls not into the scope of the current review, hence the reader would be referred to the above articles for further information about the field.

6. Novel applications 1: Target identification in phenotypic bacterial growth inhibition screens

We will now present two applications of target prediction tools performed by the authors themselves.

In the last decade many bottlenecks associated with “large scale science” such as HTS have been removed; hence it is no longer an issue to cherry pick (individually select) and assay tens of thousands of compounds in a matter of days. Previously, this was a serious problem so applications such as ‘plate-based diversity selection’ had to be developed to speed up the compound selection process [121]. In tandem, not only the number of compounds as well as the throughput increased, but rather the data associated with each compound increased simultaneously, in proprietary as well as in public databases (see Section 2 for details).

However, the increase of data is not true in all areas, since many pharmaceutical companies abandoned antibacterial research in the past decades, leaving only five players (CSK, Novartis, AZ, Merck and Pfizer) involved in the area [122]. As a result, the knowledge around antibacterial targets is limited. More specifically, the databases are rich in human targets and compounds that modulate these targets, but contain limited information when it comes to bacterial targets. This is unfortunate because the major bottleneck in antibacterial drug discovery is often not discovering hits — in a phenotypic assay, such as a bacterial growth assay, understanding the mode of action of the hits is a challenge. For example, it is unclear if the phenotypic observation is a result of general cytotoxicity or rather target-specific; and if the phenotypic observation is target-specific then which targets are involved. Once the target is elucidated it is much easier to optimize the potency and selectivity against the primary target.

Hence, one way to elucidate the mode of action of compounds tested in antibacterial assays is the application of in silico target prediction models; our workflow to assess hits from phenotypic antibacterial screen is shown in Fig. 3. The figure outlines a computational approach for target elucidation of phenotypic observations taken from the MDDR such as “Antibiotic”, “Antibacterial” or “Antifungal”. After reading all structures they are assessed in silico through a Multiclass Bayes model (implemented in Pipeline Pilot) trained on the GVK bioscience database and employing ECFP circular fingerprints, where the objective is to elucidate the cellular target that the compounds modulate. For each compound all the targets with a Bayes score greater than 30 were considered likely targets and were further analyzed. The most prevalent
targets (those predicted for 100 compounds or more by the model) are visualized in Fig. 4. This knowledge can now be used in real-world situations, such as to annotate the hits and deprioritize the compounds with known mode of action. Specifically, among the most prevalent targets in Fig. 4 are DNA gyrase (GYR), DNA gyrase subunit A (GYRA) and AMPC (beta-lactamase). The compounds predicted to be active against those targets may not be of interest to a team that would like to discover an antibacterial with a novel mode of action. Vice versa, other teams might wish to follow up compounds with known mode of action, but novel chemotypes. All those selections are now possible, based on the annotated putative compound targets.

Interestingly many of the predicted targets are human and not antibacterial. This brings further knowledge to the team both in terms of off target effect and what risks should be mitigated. For example, monoamine oxidase A and B (MAOA and MAOB) are enriched in this analysis (Fig. 4). Another example is cytochrome P450 (CYP3A4); the knowledge that compounds interact (usually as substrates) with this enzyme could be useful to the project team when developing a risk mitigation strategy and interpreting in vivo data such as those referring to a compound half-life. Furthermore, knowledge of the mammalian target could be leveraged to generate a hypothesis with regard for the bacterial target; for example, a target prediction of matrix metalloproteinase 9 (MMP9) may hint that the compounds also may hit a bacterial metalloprotein. Many of the MMP9 inhibitors interact with the zinc atom in the metalloprotein though a hydroxamate moiety that could potentially also bind to other zinc-containing bacterial proteins such as LpxC. As an example (and in line with many other LpxC inhibitors) the antibiotic CHIR-090 contains a hydroxamate moiety coordinating the target interactions.

Fig. 3 – Implementation of target prediction models in PipelinePilot. 7812 structures with annotated antibacterial activity (activity indices 6700 and 68,000) were taken from the MDDR databases and likely targets could be predicted for 6874 of them using an in silico model. Calculations (using fingerprints and a Bayes model, such as here) are computationally efficient, with prediction for all compounds obtained in the order of 10 s.

Fig. 4 – Target predictions for the set of antibacterial compounds analyzed in Fig. 3. It can be seen that gyrase is the most frequently predicted target. Hence, in silico target prediction models can be beneficial when assigning a mode of action to compounds where only a phenotype has been determined experimentally.
catalytic zinc ion [123], and it also exhibits weak activity in our in-house the MMP assay panel. This exemplifies a possible relation, which is not fully understood or obvious, of using chemical probes to associate human and bacterial targets. All this knowledge can now be used to anticipate targets for drugs in development, both in humans and pathogens, based on the initial target prediction obtained in silico.

7. Novel applications 2: HTS fingerprints in target prediction

Traditional ligand-based target prediction methods rely on the commonly accepted principle that small-molecules with similar chemical structures will bind to similar targets [34,124]. Using this principle, target predictions can be made by comparing compounds, using a desired chemical descriptor and similarity metric, to a database of compounds with known activities. This simple descriptor matching can also be further extended using statistical models, such as Bayesian classifiers, trained on large chemogenomic databases [125]. The central caveat in all of these methods is that they rely on chemical structure similarity, rendering it difficult to make predictions for orphan compounds as well as to scaffold-hop to very diverse structures, e.g. from a class I to class II inhibitor of the same kinase. The analysis of “structure–activity cliffs” [126,127] as well as recent ligand-based virtual screening benchmarks against statistically unbiased chemical databases [128], have also revealed that structurally similar compounds may not always be as similar in activity space as once thought. These factors have caused a shift away from traditional structural descriptors toward “biological fingerprints”, which describe small molecules purely by their bioactivity profiles against a panel of targets since the absence of any explicit chemical structure information in these descriptors often allows them to overcome many of the caveats of traditional fingerprints. Overall, this concept has some relation to the ‘affinity fingerprints’ first presented by Kauvar [129], who first mentioned that bioactivity can be described in a ‘universal way’ by using a set of bioactivities against a set of ‘orthogonal’ proteins.

In their work on Biological Spectra Analysis, Fliri et al. showed that by clustering compounds on a “biological spectra” derived from Cerep’s BioPrint database, structure–activity relationships could be found without introducing the bias of any structural descriptor [130]. This work was later extended to analyze and predict adverse drug reactions (ADRs) [131]. Similarly, Plouffe et al. used activity profiles generated from the Genomics Institute of the Novartis Research Foundation’s in-house high-throughput screening data to predict the mechanism of actions of novel anti-malarial compounds [132]. Also an in silico analog of this fingerprint, termed ‘Bayes Affinity Fingerprint’ has recently been presented [129], extending this concept to areas where no experimental bioactivity data of a compound against a particular set of targets is available.

The Novartis Institutes for BioMedical Research has similarly developed a small molecule biological descriptor called High-Throughput Screening (HTS) fingerprints, which are derived from single concentration activity data from 184 high-throughput screens using over 1.6 million compounds (Fig. 5). The current fingerprint is composed of 104 biochemical assays spanning the major drug target classes, as well as 80 cell-based assays or phenotypic screens. The combined dataset creates an information-rich profile that reflects compound pharmacophores and biological effects in cells. Using a systematic network-based clustering approach, we show that compounds with similar HTS fingerprints, while often very structurally diverse, modulate the same broad biological processes and pathways. While in many cases similar profiles will result from compounds having the same target, it is often observed that compounds with similar HTS fingerprints hit different targets in the same pathway. This allows us to not only make single compound-target predictions, but also to group compounds into biologically similar clusters which modulate the same therapeutically relevant processes such as inflammation or autophagy.

In order to compare the HTS fingerprints of two compounds, the similarity metric must take into account the sparseness of the data, i.e., not every compound has been tested in every assay. In order to adjust for this, a modified version of the Pearson correlation that takes into account the number of assays in common between the two compounds was used. Using this metric, an N-by-N HTS fingerprint similarity matrix was computed for a set of 58,056 compounds. The set included all compounds in Novartis’s screening library that could be mapped to human protein targets using public bioactivity data from GVK and ChEMBL, a subset of marketed drugs, and a collection of natural products. This matrix was then used to construct a similarity network by representing each compound as a node and connecting two nodes (compounds) if their similarity score was above a threshold of 0.85. Unconnected nodes were removed, resulting in a network with 8288 nodes and 29,024 edges.

In order to extract meaningful clusters of compounds with similar biological activity from the network, the Molecular Complex Detection (MCODE) clustering algorithm was used. MCODE uses vertex weighting by local neighborhood density to extract highly interconnected clusters of nodes. By examining the protein targets hit by the compounds within each cluster, the statistical enrichment of those targets within certain biological pathways (defined by GeneGo or biological processes (defined by the Gene Ontology (GO) [133]) can be calculated. With alpha equal to 0.05, it was found that 52% of the clusters were enriched for at least one canonical pathway (GeneGo Metacore) and 67% of clusters were enriched for at least one Gene Ontology (GO) biological process. Within these biologically enriched clusters, target predictions can be extended to compounds without known targets using the “guilt-by-association” principle. Using this approach we were able to predict and successfully validate in vitro over 3400 compound activities that were not known in the public domain. The targets included over 100 different proteins, spanning most druggable protein families.

While, as expected, many of the clusters contained analogous compounds which hit the same targets, in many other cases considerable chemical diversity was observed within clusters. Fig. 6 illustrates two well-known, but structurally diverse classes of histone deacetylase (HDAC) inhibitors. This example represents a scaffold hop not only from a synthetic compound to a natural product, but two chemotypes with completely different binding modes. Hydroxamic acids such as
Fig. 5 – Schematic of the HTS fingerprint analysis workflow. Compound HTS fingerprints (a) are compared using a weighted Pearson correlation to create an N-by-N similarity matrix (b). This similarity matrix can then be used to create a network (c) where every node (compound) is connected if their similarity score is above 0.85. The Molecular Complex Detection (MCODE) algorithm can then be used to extract highly interconnected clusters of nodes from the network (d). The targets of the compounds in these clusters are enriched for specific biological pathways and high level cell processes, despite their significant chemical diversity.

Fig. 6 – Two major classes of known HDAC inhibitors. a) Hydroxamic acids. b) Cyclic tetrapeptides. Multiple members of each chemotype were found clustered together in the HTS fingerprint similarity network based on their common biological profiles. It is highly unlikely that a 2D structure-based method would scaffold hop between members of these two chemotypes, based on an average 2D Tanimoto similarity of less that 0.2 (using ECFP4 fingerprints).
trichostatin A reversibly inhibit HDACs by chelating the active-site zinc [134], while cyclic tetrapeptide analogs of trapoxin are covalent irreversible inhibitors [135]. With an average 2D similarity (Tanimoto coefficient of ECFP4) of less than 0.2 between members of the two chemotypes, chemical structure based method would not likely suggest a biological relationship between these compounds. However, due to their shared HDAC activity, the compounds have similarity bioactivity profiles and therefore HTS fingerprint analysis can easily suggest common targets.

8. Prospective utilization of bioactivity data for multi-target drug design: Finding the right target

While predicting individual targets for compounds is already of value, it has recently been discovered that the promiscuity of drugs ‘can also be a virtue’ [136–138]. In this paradigm shift away from single-target drug discovery to multi-target drug discovery [139] we would ideally like to select compounds exhibiting bioactivities against a series of relevant targets, with some targets probably easier to target simultaneously than others [125,140]. While target prediction models – as outline in the preceding sections – can help design suitable chemistry to modulate targets, the question of which targets should be modulated in the first place will be discussed here.

Biological function at a systems level (e.g. switching from one metabolic substrate to another in response to environmental stress; or producing an immune response as a consequence of bacterial invasion) results from the integrative activity of multiple interacting processes that evolution has organized into complex networks. Within such networks the interactors are proteins, lipids, lipoproteins, DNA and RNA, small molecule metabolites, second messengers, antigens and so on. Disease states arise when these interactions are perturbed in specific ways, often by mutations in genes encoding individual proteins. In complex diseases the disturbance is multifactorial [141] and, for the disease phenotype to manifest, multiple simultaneous interacting hits probably need to coincide in space and time, such as genetic background, environmental stress, aging and specific mutation(s). Until recently most effort was focused on trying to understand such interactions and their pathological variants at the pathway level; and most drug discovery efforts have targeted single components within these pathways with drugs intended to exhibit high specificity for those components. This approach has been notable primarily for its failure to generate significant numbers of new drugs to treat complex diseases. In the case of infectious disease, failures frequently arise due to development of resistance; or because of toxicity.

The failure of drugs to yield the desired biological outcome in many cases is the result of two fundamental issues.

1) Biological networks are structurally highly complex and so predicting the functional outcome of interventions or the consequence of mutations that must then be addressed pharmacologically is non-trivial [139]. Biological networks have evolved to be very robust [142]. This is a fundamental organizing principle of life. Biochemical and physiological networks need to be able to continue to function in the face of continual internal and external perturbation. They have the general property of attack tolerance. Random removal of individual components of a biological network has surprisingly little functional consequence for the network [142]. To be effective (or catastrophic depending on one’s perspective) interventions within a complex biological network need to be highly selective.

2) Drugs almost invariably influence more than one target to some extent either as a consequence of structural similarities between the intended target and other protein family members (e.g. β-blockers, SSRIs); or through allosteric effects on other proteins; or through genuine multivalent target binding [7]. In a recent analysis it was found that every drug was, on average, accepted to interact with six different targets [8]. However if this occurs the net result is usually a series of unwanted side effects in addition to the desired effect; or a treatment failure (and often both). Occasionally a drug’s promiscuity and the fortuitous combination of appropriate high value network targets (and minimal unpleasant off-target effects) converge to produce a treatment success. Statins are a good example of this — they appear to work as much for their anti-inflammatory and immunomodulatory properties as for the ability to inhibit HMG-CoA reductase for which they have been deliberately designed [143] (and most people can tolerate the milder side-effects of fatigue, malaise and myalgia, although rhabdomyolysis and hepatotoxicity prevent their use in all cases). The more highly tuned and less promiscuous a drug is for a particular target the more important that target has to be in network terms for it to have a significant effect. However, the fact that only 1 in 10,000 new drug candidates currently makes it to the clinic speaks for itself. So, how can we do better? Is it possible to identify those targets within a complex network whose properties are such that a desired therapeutic consequence can be achieved? And can drugs be designed or repositioned to influence these targets? We believe that the answer to all of these questions is yes.

Networks are amenable to analysis using a branch of combinatorial mathematics termed graph theory whose origins are attributable to Leonhard Euler in the 16th century. Euler devised the first topological tools to address the famous “Seven bridges of Königsberg” problem. Just as for Euler’s bridges and islands along the river, biological networks can be represented as graphs [points [more commonly termed nodes or vertices] connected by lines [called edges]] representing information flow between locations in an interconnected system and the local and global properties of this topological map can be calculated. This information can then be used to identify sets of interventions that are likely to have the desired therapeutic effect and optimal combinations of high value targets computed.

We will use the example of finding an effective antibiotic to treat an imaginary and notoriously drug resistant bacterium, BugX, to illustrate the basics of this approach. Fig. 7a shows a hypothetical (and very much simplified) biological network representing some processes in BugX that we believe to be important for its survival and for which there are no mammalian
homologs (making human toxicity of any drugs designed to perturb this network less likely).

We first identify and represent the components of the network — in this example the key proteins responsible for some critical housekeeping functions: cell wall assembly, intracellular protein transport and post-translational modification of cell wall components that are unlikely to have close mammalian homologs. These proteins form the vertices of our graph. We then connect the vertices with edges representing information flow. For example — Protein A activates Protein B, which in turn modifies Protein C to permit it to be transported by Protein G. Protein G also has a permissive effect on Protein C making it easier for it to bind and be modified by Protein B. Protein D inhibits Protein A. Protein B inhibits Protein D. Properties such as the direction of information flow, its net sign and weight are represented by arrows and by numbers associated with the edges. Finally we compute the importance of the network components in graph theoretical terms based on this connectivity pattern (with appropriate controls).

In this simple example some unique properties of certain proteins can be recognized by eye, although in real networks a mathematical approach is needed. It can be seen that Proteins G and H have high “Betweenness”. In other words they form important conduits between the cluster of proteins producing raw materials (cluster analysis and the identification of functional sub groupings is another important aspect of this approach) and the cluster responsible for assembling the bacterial cell wall. Inactivating these two proteins simultaneously would immediately sever the coupling between the factory and the assembly line as illustrated in Fig. 7b. We can also see that Proteins A, B and J have lots of connections — they have high “Degree”; Proteins F and J have high “Out Degree” — the directions of the arrows is predominantly outward. This type of metric also distinguishes important network nodes — so called “Hubs”.

Having established some high value targets we can then apply various algorithms to calculate the optimal combination of these targets to maximally disrupt the network for minimum cost. “Disruption” can be quantified by many different metrics but connectivity and diameter are commonly used and broadly equate to how many disconnected pieces the network can be fragmented into and the size of the pieces. Obviously this relates directly to functional integrity. Fig. 7c shows the effect of attacking 4 of the 17 possible targets when the targets are all high value (Proteins A, G, H and J) and Fig. 7d when they are a mixture of high, moderate and low value (Proteins B, I, P and Q). A compound, or pair of compounds, that could achieve the first might make an excellent antibiotic with a low probability of subsequent resistance; whereas one that affected the lower value group or individual low value targets (the majority of nodes) would have low efficacy and BugX would likely be able to develop resistance by upregulating alternative redundant routes.

In summary then, the steps required to select protein targets in this example are:

(i) Construct a network model of the processes of interest using current biological and pathophysiological knowledge.
(ii) Subject it to network analysis to identify high value targets (details of the method to be published shortly)
(iii) Compute minimum synergistic combinations of targets (details of the method to be published shortly)
(iv) Find multivalent drugs that bind to these targets to exert the desired effect
   a) by searching chemoproteomic databases for existing compounds with promising binding signatures, or
   b) by searching a “virtual” structure-function space, which also includes the exploration of novel chemical space [19].

Given the growing relevance of multi-target drug discovery, as well as the increasing amount of bioactivity data available both in private companies but also public databases, the authors – along with other researchers in the field [18,139,144–147] – are of the opinion that the prospective design of multi-target drugs will be one of the major developments in the pharmaceutical industry in the near future. Areas such as the consideration of dynamic properties of networks will continue to gain importance in the future [148] and some advances in the field are already being reported [149], however the availability and integration of data remains currently a challenge.

9. Conclusions

The increasing availability of bioactivity data is a unique chance to boost success rates in rational drug development, and to learn from the past in order to make better chemical decisions in the future. This review attempts to summarize available databases, ligand- and structure-based methods to predict the protein targets of small molecules, and to present applications of those tools which have increased in number in recent years. While the volume and quality of information available will continue to grow, we are likely to see more and more successful applications of target prediction algorithms in the future, as well as increased integration of the available data. Given the size of chemical space of about \(10^{42}\) bioactive molecules (estimation of the authors, currently being published elsewhere) it is likely that target prediction methods will be around to supplement experimental methods for the foreseeable future.

Acknowledgments

Unilever is thanked for funding AK, JK, RCG, PJB and AB. MG, JLJ and BS thank Novartis for support.

REFERENCES


[38] Bender A. How similar are those molecules after all? Use two descriptors and you will have three different answers. Exp Op Drug Disc 2010;5:1141–51.


[56] KNIME, Knime.com, Zurich, Switzerland.


