ELSEVIER

DRUG DISCOVERY TODAY TECHNOLOGIES

Drug Discovery Today: Technologies

Vol. 3, No. 4 2006

In silico techniques

In silico target fishing: Predicting biological targets from chemical structure

Jeremy L. Jenkins*, Andreas Bender, John W. Davies

Novartis Institutes for BioMedical Research, Discovery Technologies, Lead Discovery Center, 250 Massachusetts Avenue, Cambridge, MA 02139, USA

***In silico* target fishing is an emerging technology that enables the prediction of biological targets of compounds on the basis of chemical structure by using information from increasingly available biologically annotated chemical databases. We provide a comparative review of recent studies in which data mining, similarity, or docking of chemical structures is used to elucidate target class or mechanism of action. These approaches reverse the paradigm of finding compounds for targets to finding targets for compounds.**

Introduction

Orphan compounds may be defined as compounds that modulate cellular phenotypes but have unknown macromolecular targets. For example, many natural products have a clinically proven therapeutic use in humans yet their mechanism-of-action is undetermined. Orphan compounds can also be the endpoint of phenotypic screens, necessitating subsequent target identification experiments [1], such as affinity chromatography of proteins (Fig. 1). By contrast, an emergent computational approach analogous to experimental target finding is to map the compound structure into known structure–activity relationship (SAR) space by mining large biologically annotated chemical databases (a.k.a chemogenomics databases). In this vein, there are an increasing number of examples where chemical structures were linked to targets or bioactivities using cheminformatics methods. Such approaches are relevant for rapid prediction of primary targets, potential off-target effects, and even selectivity among target families.

*Corresponding author: J.L. Jenkins (jeremy.jenkins@novartis.com)

Key technologies for *in silico* target fishing

Table 1 summarizes four related computational approaches that make possible the prediction of target or mechanism-of-action (MOA) from chemical structure: chemical similarity searching, data mining/machine learning, bioactivity spectra, and panel docking. Although each of these technologies were traditionally developed to find new compounds for known targets, this review focuses on the reverse use of these technologies to link targets to compounds (Fig. 1).

Similarity searching in databases

Chemical similarity searching for target prediction simply compares an orphan compound structure to a database of compounds with known targets. Any chemical descriptor or similarity metric can be used and more importantly similarity searching does not require a well-curated database of normalized target names. In this sense, any database can be queried: the orphan compound is input and the similarity

 DOI: 10.1016/j.ddtec.2006.12.008

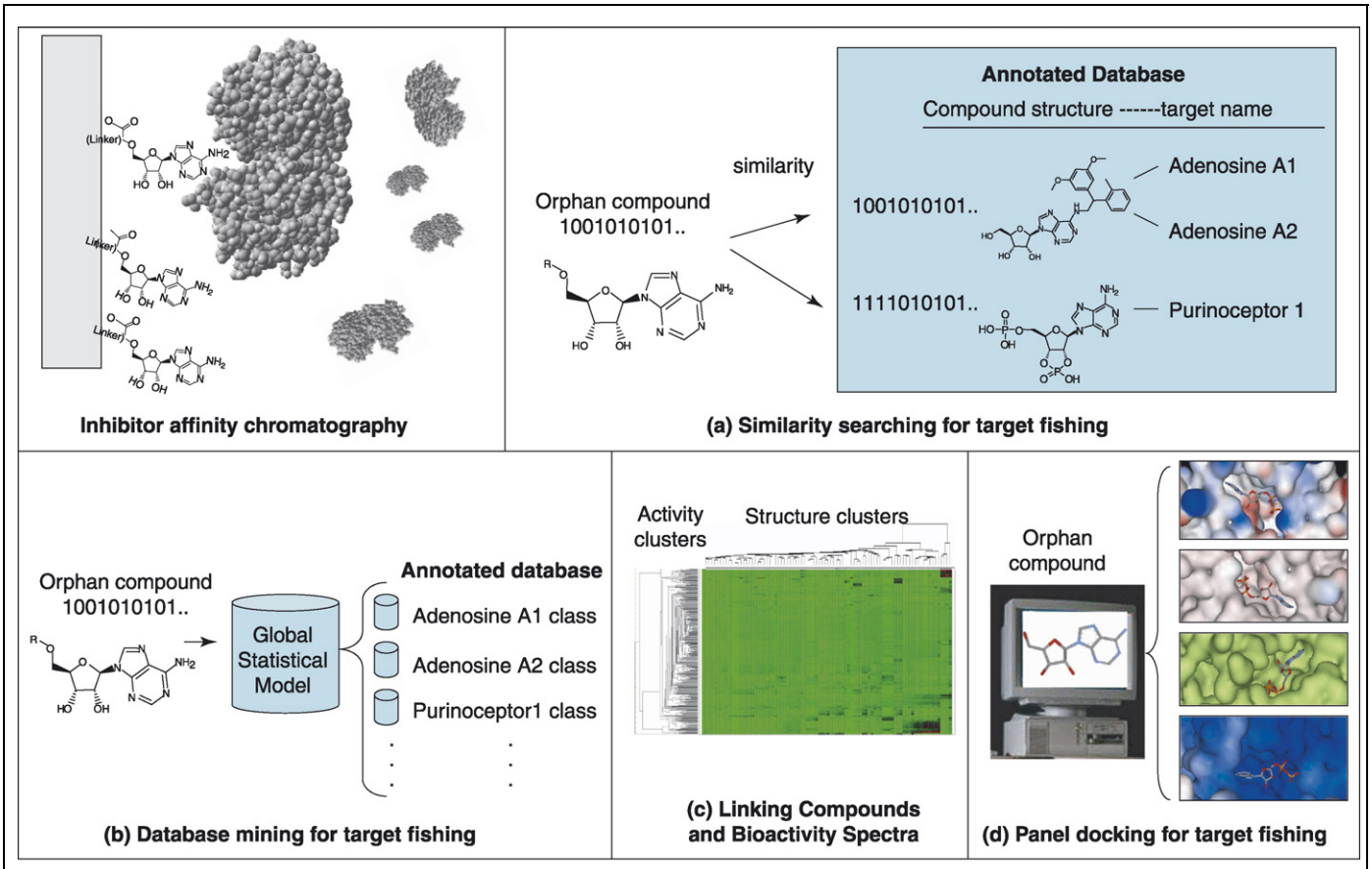

Figure 1. *In silico* analogies to the experimental target fishing approach of affinity chromatography ('pulling-down' proteins from cell extract with compounds displayed on resin). (a) Similarity searching pulls protein names from databases by matching orphan chemical fingerprints to biologically annotated compound fingerprints. (b) Data mining approaches use target class models trained by machine learning on target class fingerprints. (c) The behavior of chemical structures across multiple targets creates a spectra that can be used for *a priori* activity predictions. (d) Compounds with unknown targets can be virtually docked to active sites of databases of 3D protein structures for target prediction or to compute interaction fingerprints for target selectivity studies.

matches point to potential target classes. (The reader is referred to [2] for a general review of chemical similarity searching.) Similarity and substructure searching have been used for many years for target prediction – albeit unsystematically – to assess patent coverage around chemotypes with tools such as SciFinder (http://www.cas.org/SCIFINDER/). In more recent years, web-based search engines have become available for finding chemically similar bioactive structures. Table 2 lists websites where the basic task of web-based similarity searching in a database of bioactives is possible. In addition, a small number of commercial companies are taking on *in silico* target prediction, either through their software or services [e.g. Inpharmatica (http://www.inpharmatica.co.uk/) and Eidogen-Sertanty (http://www.eidogen-sertanty.com/)]. Although the searches themselves may take seconds in practice, the amount of follow-up reading needed to align potential targets with compound phenotype can be time-intensive (imagine using SciFinder to predict targets for 1000 hits from a cell-based screen). Further, how does one go about ranking the targets? Should only the target associated with the most chemically similar compound be considered or also less similar compounds? Is target enrichment among similar compounds most important given that some target classes are more represented than others? All of these questions need to be addressed to establish a reliable target identification metric. Further, the technical disadvantage of similarity searching is that prior knowledge of target class information is typically not incorporated in a way that focuses or improves the search. However, there are recent notable exceptions where search performance is improved by weighting molecular fingerprints with target class knowledge [3–6].

Similarity searching for target fishing can also be performed with 3D chemical descriptors. For example, Cleves and Jain have demonstrated predictive ability of 3D morphological descriptors [7]. The authors of this review have also found that while 2D descriptors are powerful for similarity searching in annotated databases, 3D descriptors may be more appropriate when the orphan compound has low 2D similarity to all database molecules [8].

Table 1. Comparison summary of technologies for *in silico* target fishing[a]

Method	Chemical similarity searching in annotated databases	Data mining in annotated chemical databases	Linking chemical structures to biological activity spectra	Protein panel docking
Specific technologies	Substructure searches, chemical fingerprints (e.g. Daylight, UNITY, MDL Keys, E-states, topological descriptors, Similogs), ReliBase, 3D morphological and surface similarity	Regression, probabilistic neural networks, trend vector analysis, text influenced molecular indexing, multiple category Bayesian models, self-organizing maps	Affinity fingerprints, 3D pharmacophore fingerprints and neighborhood behavior models, various clustering methods, 'SAR similarity', structure activity target (SAT), coincidence scores	SIFts, INVDOCK, AutoDock and other docking methods
Pros	Does not require a training set or a well-annotated database	Fast, robust, accurate, provides a ranked list of targets	Links structure to biological response across multiple targets	Does not depend on ligand data
Cons	May ignore pre-existing target class knowledge, problem with ranking targets, may ignore prior target frequency in database	Cannot predict targets outside of training set, requires accurate annotation in training set (normalized target names)	Complete activity matrices are optimal for usage, proven mainly for focused areas (e.g. safety pharmacology, kinases)	Slow, works best for kinase panels, limited to targets with resolved structures, intensive data preparation, tricky to normalize docking scores for target ranking
References	[3–9,36,37]	[11–17,38]	[19-25,39,40]	[30–33]

[a] Daylight: http://www.daylight.com/; UNITY: http://www.tripos.com/index.php?family=modules,SimplePage,sybyl_unity; MDL Keys: http://www.mdl.com/solutions/white_papers/JCICS-2002-42-keys.jsp; E-states: electrotopological state indices; Similogs [9]; ReliBase: relibase.ebi.ac.uk/; SIFTs: structural–interaction fingerprints [33]; INVDOCK: inverse docking [30]; AutoDock: http://autodock.scripps.edu/.

In addition to chemical similarity searching for target fishing, target ontologies may be exploited to find new targets for compounds. In this case, compounds are not orphans, but rather a target is known; the goal then is to link new related targets to a compound on the basis of sequence similarity to the known target. Schuffenhauer *et al.* demonstrated the relationship between similar compounds and similar targets in similarity searching [9] and Sheinerman *et al.* explored the relationship between kinase sequences and kinase inhibitor selectivity [10].

Data mining in annotated chemical databases

Generally speaking, data mining is the automated extraction of patterns and associations from large databases. Given a large, diverse database of compounds with annotated target names, data mining is an ideal approach for target prediction. First, associations between target names and chemical substructures can be extracted automatically across target class sets with inductive machine learning. Chemical features correlated with specific target binding are then stored in the form of multiple-target models. The target fishing problem is thus one of compound classification on a grand scale involving thousands of individual target class models. By comparing orphan compound features with correlated features in each target class, target prediction can be achieved with scalability and speed far beyond that of, say, docking compounds across multiple targets. By contrast to similarity searching in databases, which uses all bits in chemical fingerprint to find similar ligands, models built from machine learning retain only the relevant bits and ignore bits common to both actives and inactives. Table 3 provides a comparison of large, diverse databases on which target class model building is possible. Other databases exist that document current drugs, although they are less amenable to biological target fishing due to their small size; however, drug databases may be useful for target fishing if combined with larger databases. Some examples are Comprehensive Medicinal Chemistry (http://www.mdl.com/products/knowledge/medicinal_chem/index.jsp), Ashgate Drugs (http://www.cambridgesoft.com/databases/details/?db=2), and World Drug Index (http://scientific.thomson.com/products/wdi/).

In the mid 1990s, V. Poroikov, D. Filimonov, and others pioneered *in silico* prediction of activity spectra for substances (PASS) by training models on the chemical features of activity classes. The PASS application as well as the multilevel neighborhood of atoms descriptor they developed have been described [11,12]. PASS predictions are incorporated into the NCI database browser and recent successes were reported using the PASS technology to guide medicinal chemistry, including in the design of novel cognition enhancers [13]. Niwa [14] later explored the use of probabilistic neural networks in combination with atom-type descriptors to predict targets for compounds. Bayesian modeling on chemoge-

Table 2. Large web databases for *in silico* target fishing via similarity searching

Site	Description	URL
PubChem	Part of the National Institutes of Health Molecular Libraries Roadmap Initiative. Provides smiles, similarity, sub- and super-structure searches and bioactivity information of hits. Combines a large number of bioactivity and screening databases	http://pubchem.ncbi.nlm.nih.gov/
ChemBank	A public, web-based service from the Broad Institute's Chemical Biology Program. Consists of cell-based and other screening data	http://chembank.broad.harvard.edu/
Query Chem	Searches Chembank, PubChem, eMolecules with Google's Web Application Program Interface	http://llama.med.harvard.edu/~jklekota/QueryChem.html
Enhanced NCI Database Browser 2	SMILES, similarity and substructure searching in 250K public NCI structures. Provides cancer, yeast, and HIV activity data, PASS predictions	http://cactus.nci.nih.gov/
ChemIDPlus	National Library of Medicine hosted. Substructure, similarity searching in 108K compounds. 'Classification codes' link provides MOA	http://chem.sis.nlm.nih.gov/chemidplus/
LIGAND	Compilation of public databases [ChemBank, ChemPDB, Kyoto Encylopedia of genes and genomes (KEGG), NCI, and others] with ~750K records. Some information on biological activity, FDA drug status, or anti-HIV activity, among others	http://ligand.info/
Relibase	Free online version searches the ligands in the Protein Data Bank by chemical similarity and substructure	http://relibase.rutgers.edu http://relibase.ccdc.cam.ac.uk
Miscellaneous Smaller databases	Sites where URL links provided for smaller databases	http://cactus.nci.nih.gov/ncidb2/chem_www.html http://www.cheminformatics.org/
PASS	Prediction of activity spectra for substances. Commercial website that predicts ~1000 bioactivities, targets, MOAs, toxicities	http://www.akosgmbh.de/pass/index.htm

nomics database targets [15] has been carried out by Nidhi *et al*. [16] using extended connectivity fingerprints (http://www.scitegic.com). The authors demonstrated that compounds with known therapeutic activities could be systematically deconvoluted to specific protein targets. On a grander scale, several commercial and in-house databases containing 4.8 million compounds and 2876 targets were combined to create a global pharmacology map that was also explored by multiple-category Bayesian modeling [17]. One advantage of creating models on chemical fingerprints is the interpretability: substructures correlated with target binding can be back-projected onto orphan compound structure. For example, Fig. 2 shows the COX (GenBank accession nos. AF440204, U04636) inhibitor Pamicogrel (Kanebo, EP-00159677, http://www.kanebo.co.jp/english/company/pharm_about.html) where features correlated with target binding are mapped onto it (red bonds). While COX is the highest scoring target in the Bayesian model as would be expected, the off-target prediction of peripheral benzodiazepine receptor (PBR, GenBank accession no. AY998017) is interesting due to a previously suggested link between PBR ligands and COX inhibition [18]. The off-target prediction of leukotriene B4 (LTB4, GenBank accession no. BC004545) receptor is fitting with the role of both COX and LTB4 in eicosanoid-related pathways and inflammation. These examples demonstrate the power of mining chemogenomics databases to link chemistry and biology.

A word on chemogenomics database format

To be useful for data mining and model building, the user must have access to the entire database (it is either purchasable or downloadable). The minimum requirements for a chemogenomics database amenable to target fishing include chemical structures (atomic coordinates or SMILES-type representations) and their associated targets or bioactivities, where the target nomenclature is standardized and consistent and free of spelling or typing errors. Data is clearly more reliable for compound-target pairings derived from secondary assays where activity values (IC_{50}, EC_{50}, or K_d for example) are calculated from multiple concentration points, as opposed to single data points from high-throughput screening. Ideally, the database would have additional layers of annotation describing target family ontology (to truly be a chemogenomics database), normalized activity values, inactive as well as active compounds, and data source (journal or patent citation). Each record of a compound-target pairing in fact describes a molecular event; thus the assay conditions and the type of cell, tissue, or organism used are all influential on the activity measurement and should be included with the record. Also of value are unique identifiers that enable linkage

Table 3. Chemogenomics databases suitable for data mining and *in silico* target fishing[a]

Database	URL	Description	Size	Data source coverage	Breadth of target coverage	Target names standardized for data mining?	Target ontology	Gene ID	Amino acid seq.?	GUI
Target inhibitor database (GVK Bio[b]**)**	http://gvkbio.com/informatics/dbprod.htm	Large, diverse collection of chemical series from medchem literature and patents	1.8 Mio entries, 500K cmpd records, 1.5K targets	Journals and patents	Wide (enzymes, ion channels, GPCRs, nuclear hormone receptors)	Yes, official gene name of protein	Yes	Yes	No	ISIS base, Oracle, MS-Access
MedChem (GVK Bio[b]**)**	http://gvkbio.com/informatics/dbprod.htm	Large, diverse collection of chemical series from medchem literature	750K cmpd records, 607K unique cmpds, 4.9K targets	64 medicinal chemistry journals	Wide (enzymes, ion channels, GPCRs, nuclear hormone receptors)	Yes, official gene name of protein	Yes	Yes	No	ISIS base, Oracle, MS-Access
AurSCOPE (Aureus)	http://www.aureus-pharma.com/Pages/Products/Aurscope.php	Collection of 'knowledgebases': GPCR, Kinase, ion channel, hERG, ADMET	GPCR, 152Kcmpds, 635K activities; kinases, 51.8Kcmpds, 163.7K activities; ion channel, 58.4Kcmpds, 217.6K activities	~370 journals and thousands of patents	Focused on large target classes (general 'enzymes' or 'nuclear receptors not included yet)	Yes	Yes	Yes	No	AurQuest or in-house systems
stARLITe (Inpharmatica)	http://www.inpharmatica.co.uk/StARLITe/Index.htm	Large, diverse collection of chemical series from medchem literature	300K cmpds, ~5K targets, 1.3 million datapoints	Medchem journals (1980–present)	Wide (enzymes, ion channels, GPCRs, nuclear receptors)	Yes, RefSeq or ACCESSION and some standardized gene names	Yes	No	Yes	Chematica or Oracle 9i
ChemBioBase Suite (Jubilant BioSys)	http://www.jubilantbiosys.com/chemo.htm	Combined target-centric ligand databases: Kinase, GPCR, nuclear receptor (NR), ion channel, protease, phosphodiesterase	~1020 total targets. Kinase, 319K cmpds; GPCR 400K cmpds; NR 150K cmpds; Ion Channel, 100K cmpds, Protease, 400K cmpds	Standard medicinal chemistry journals and patents	Focused on large target classes (general 'enzymes' not included)	Yes	Yes	No	No	Isis base, Oracle, among others
BioPrint (Cerep)	http://www.cerep.fr/cerep/users/pages/collaborations/bioprint.asp	Focused pharmacology/ ADME profiling database. A full data matrix.	180 diverse targets and 2.5K drugs, >1Mio records. Also an in vivo dataset. Adverse drug reactions	Experimentally determined *in vitro* data	Wide, but small number	Yes	?	No	No	BioPrint GUI
WOMBAT 2006.1 (Sunset Molecular)	http://sunsetmolecular.com/products/?id=4	Chemical series published in medchem literature	154K entries, 136K unique compounds, 308K total activities, 1320 protein targets	Four journals (6791 papers)	Wide (enzymes, ion channels, GPCRs, nuclear receptors)	Yes, normalized names and ACCESSION	Yes	No	No	ISIS base
MDL drug database report	http://www.mdl.com/products/knowledge/drug_data_report/index.jsp	Drugs launched or under development	~160K entries, 123.7K unique compounds, ~700 targets, bioactivities, chemical classes	Patents, conf. proceedings, and other sources. 1998-present	Wide (enzymes, ion channels, GPCRs, nuclear receptors)	Somewhat, but with exceptions and overlaps	No	No	No	ISIS base

[a] Table information is extracted from the URL websites in August 2006 or product literature.
[b] Other specialized databases are offered.

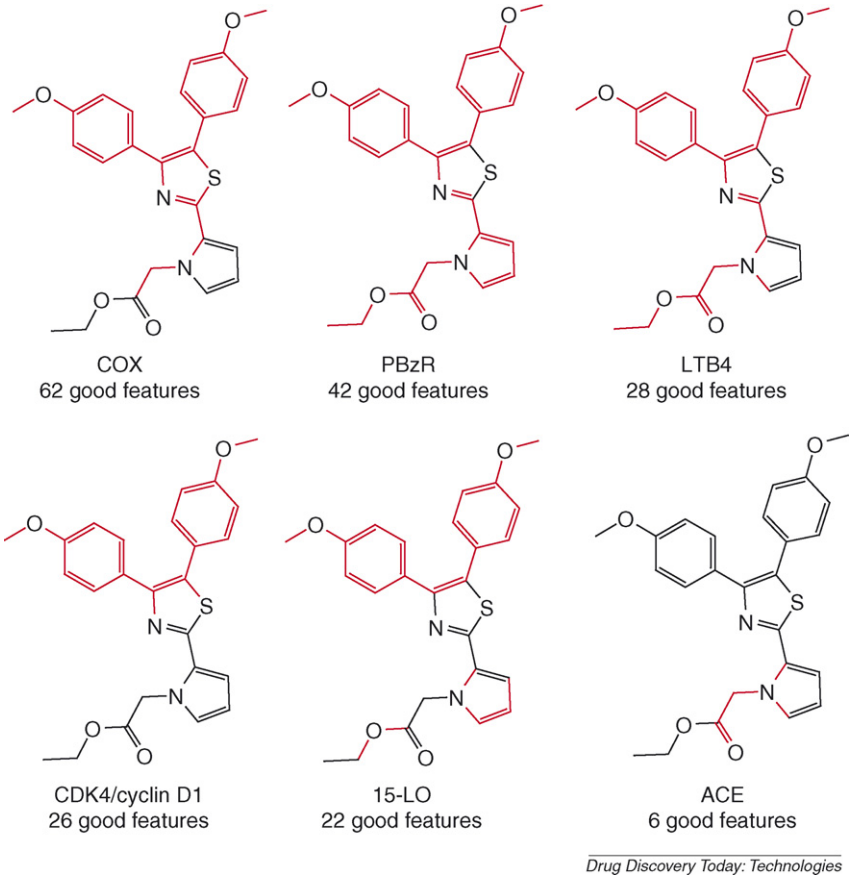

Figure 2. Back-projection of features correlated with target class inhibition for selected targets in a multiple category Bayesian model trained on the WOMBAT database. The COX inhibitor Pamicogrel is predicted to inhibit COX, and potential off-target binding is predicted for peripheral benzodiazepine receptor (due to similarity to some benzothiazepine inhibitors and the leukotriene B4 receptor (also in the arachidonic acid cascade)).

of the target to external databases (e.g. bioinformatics-oriented databases), such as Entrez GeneID, Uniprot-Swissprot Accession Number, or NCBI RefSeq number.

Data mining on bioactivity spectra

The activities of a compound across a protein panel, cell line panel, HTS screening panel, or DNA microarray can also be a type of signature, which has been termed the 'biological activity spectra', 'bioactivity spectra', or just 'biospectra'. The biospectra of a compound is related to chemical structure, and therefore can be used predictively in either direction – predicting activities for compounds or predicting compounds for activities. The behavior of compounds across targets enables prediction of structure–property associations and provides probabilistic SAR.

Early work in this area by Kauvar *et al.* [19] showed that one could predict binding of compounds to a new target if they are first screened against a reference panel of proteins, and then a small, diverse subset of those compounds are screened against the new target. The binding signature of the diverse subset is an affinity fingerprint that can be compared to the panel binding of the whole compound set with stepwise linear regression to predict binding of the whole set to the new targets [19]. Similar compounds will have similar affinity fingerprints, although interestingly, there are cases where structurally dissimilar compounds are not distant in affinity fingerprint space and vice versa [20]. Therefore, bioactivity spectra provide foresight that does not entirely overlap with structural predictions.

Bioactivity spectra have also shown great promise with respect to mining pharmacology data and predicting adverse drug reactions (ADRs) rather than primary targets, especially in the case of the BioPrint database (Cerep). For example, ADRs can be predicted for a compound on the basis of its 'profile similarity' to other compounds with known ADRs, where the profile is determined in ligand-binding assays against a panel of targets [21,22].

Cytotoxicity data across multiple cell lines is another type of biospectra; extensive research has been carried out at the National Cancer Institute to deconvolute cytotoxicity and gene expression data to specific chemotypes, targets and modes-of-action using self-organizing maps [23,24]. The question of whether or not compound activities across multiple cell-based screens are the result of primary target binding

or off-target binding was addressed by Klekota *et al.* by applying an entropy-based score to structurally clustered compounds to see whether their biospectra statistically reflected a single-target mechanism [25].

In the simplest scenario compound activity refers to inhibition measurements or protein binding; however, the notion of compound activity could also be expanded to include its effect on gene expression patterns. The trend in linking chemical structure to mRNA profiles from microarray gene expression data has recently been emerged as a tool to drive post-genomics drug development [26]. For example, compound selection or design on the basis of similarity to other compounds with a desired global effect on cellular gene expression is now possible, a true clinical application of systems biology. In another intriguing study, Rosania showed that chemical substructures can be predictive of subcellular distribution, which could be highly relevant to the current topic of target prediction [27].

The main strength of the biospectra approach is also its main disadvantage: the *in silico* predictions initially require experimental data collected across a matrix of targets or assays, which can be difficult to obtain, and is typically specialized in nature (e.g. kinase or pharmacology targets, cytotoxicity assays).

Docking to protein databases

The target prediction methods described above are based on small molecule information alone to infer targets of new compounds. An alternative approach is to include 3D information about the target protein and to perform ligand–target docking [28], to a wide panel of proteins to determine which one is – according to the scoring function – the most likely interaction partner *in vivo*. Compared to ligand-based approaches, this puts tremendous strain on computational resources since docking needs to be performed on hundreds of proteins for a given small molecule, including the necessary (often manual) preparation steps. Also, since binding affinities are compared across different proteins (and classes of proteins) normalization of those affinities, to establish the 'absolute' best binding partner, is a non-trivial exercise. The scoring function, which assigns a numerical affinity value to a ligand–target interaction, lies at the heart of docking-based target identification approaches and its values are crucial to the predictions. Thus, it needs to be kept in mind that current scoring functions might be a bit crude for this task, since a recent comparative review even states 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' [29]. Nonetheless, some successful applications of docking in this area exist.

The first application of docking for target identification, which is known to the authors involves 4H-tamoxifen and vitamin E [30]. An inverse docking procedure (one ligand, multiple targets instead of one target, multiple ligands) termed INVDOCK was used to predict protein targets. For 4H-tamoxifen several known targets such as the estrogen receptor, protein kinase C, collagenase, 17β-hydroxysteroid dehydrogenase, alcohol dehydrogenase and prostaglandin synthase were predicted. Experiments also confirmed the predicted involvement of 4H-tamoxifen on DHFR and immunoglobulin levels. More flexible known targets such as calmodulin were not predicted. Overall, for 4H-tamoxifen and vitamin E, about 50% of the predicted targets are either known targets or they are conformed by experiments. While this is a starting point, further validation of the method is needed.

A large amount of docking-based target prediction has been performed on the kinase family of proteins. Since not always experimental crystal structures are available, in a recent study [31] homology models were generated across the family and under the constraints that both backbone conformation of the kinase and binding mode of the inhibitors are fixed across the kinase panel. Using these simplifying assumptions, AutoDock was able to reproduce experimental affinity profiles to a surprising extent. One of the important findings was that to reproduce selectivities the inhibitor needed to remain fixed in a given binding orientation.

Using the same basic approach recently the question was investigated as to what the 'true' targets of ligands of the Protein Data Bank (PDB) actually are, which is whether they would be able to bind to other proteins with similar or even higher affinity as well [32]. For four unrelated ligands, docking was performed to a subset of the PDB containing 2148 protein structures, which fit certain quality criteria. In all cases, the true target was recovered in the top 1% of the ranked targets which was not true for a generic ligand, leading to the conclusion that inverse docking is rather more suited to selective compounds.

Overall, while target determination via docking shows some successes, there are many issues that need to be addressed. Today, establishing kinase selectivity by interaction fingerprints [33] for example seems to be a more efficient approach. Other areas that need to be improved upon are the low-throughput scoring function accuracy and normalization, and the large amount of manual data preparation required.

Conclusions

Chemical similarity searching for *in silico* target fishing is a good first pass, but it is currently not systematic in how targets are ranked and it does not generally incorporate target class knowledge. Models built with machine learning methods have the capacity to rapidly and automatically predict targets. Target prediction by linking structures to biological activity spectra is a proven approach given one has in hand the experimental data to carry out this analysis. (While much

of the kinome data is public, the great majority of pharmacology data is proprietary and expensive.) Target panel docking is a computationally expensive approach and perhaps not yet a mature enough technology to widely support target fishing.

In silico target fishing is an emerging field that exploits the growing volume of available SAR data; however, outstanding issues remain (Outstanding issues). Although novel compound-target pairings may be predicted, *in silico* target fishing in databases is still derivative in the sense that new targets outside the scope of the training set are not proposed; however, it is possible to point to novel targets if the target record from the database contains additional layers of annotation. For example, prediction of the correct target family is easier to achieve than fishing the true target [16].

There are several practical applications for *in silico* target prediction. The predictions may reveal intellectual property coverage around a chemotype or point to established synthetic chemistry routes. The predictions may be used to annotate hits from a cell-based screen [34] or to map compounds into protein–protein interaction pathways [35]. In the context of a modeling application with chemical intelligence and multi-optimization capabilities, the target predictions could be used for rational design. Finally, *in silico* predictions are fully complementary to experimental target fishing methodologies; in cases where chemical proteomics experiments point to multiple protein targets, consensus *in silico* predictions will add greater confidence in the targets or target families fished out.

Outstanding issues

- Databases need both cheminformatics and bioinformatics annotations.
- No unified target ontology exists that is used universally, making the merging of databases a laborious task.
- Currently, the best IC50 data is sold commercially (Table 3) and is often expensive, although public resources such as PubChem may fulfill this need in the future.
- *In silico* target fishing is not well-integrated with experimental target fishing approaches.

Acknowledgements

We thank Meir Glick, Jim Nettles, Zhan Deng, Nidhi, Thomas Crisman, Ansgar Schuffenhauer, John Peter Priestle, Kamal Azzaoui, Edgar Jacoby, John Tallarico, Bernd Rohde, Dmitri Mikhailov of NIBR and Paul Clemons (the Broad Institute) and Brian Schoichet (UCSF) for helpful discussions on *in silico* target fishing.

References

1 Hart, C.P. (2005) Finding the target after screening the phenotype. *Drug Discov. Today* 10, 513–519
2 Bender, A. *et al.* (2006) Molecular similarity: advances in methods, applications, and validations in virtual screening and QSAR. In *Annual Reports In Computational Chemistry*, (Vol. 2) (Spellmeyer, D.C., ed.), pp. 141–168, Elsevier
3 Birchall, K. *et al.* (2006) Training similarity measures for specific activities: application to reduced graphs. *J. Chem. Inf. Model.* 46, 577–586
4 Eckert, H. *et al.* (2006) Mapping algorithms for molecular similarity analysis and ligand-based virtual screening: design of DynaMAD and comparison with MAD and DMC. *J. Chem. Inf. Model.* 46, 1623–1634
5 Bender, A. *et al.* (2006) 'Bayes affinity fingerprints' improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multi-target drugs a feasible concept? *J. Chem. Inf. Model* 46, 2445–2456
6 Stiefl, N. and Zaliani, A. (2006) A knowledge-based weighting approach to ligand-based virtual screening. *J. Chem. Inf. Model* 46, 587–596
7 Cleves, A.E. and Jain, A.N. (2006) Robust ligand-based modeling of the biological targets of known drugs. *J. Med. Chem.* 49, 2921–2938
8 Nettles, J.N. *et al.* (2006) Bridging chemical and biological space: 'target fishing' using 2D and 3D molecular descriptors. *J. Med. Chem.* 49, 6802–6810
9 Schuffenhauer, A. *et al.* (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* 43, 391–405
10 Sheinerman, F.B. *et al.* (2005) High affinity targets of protein kinase inhibitors have similar residues at the positions energetically important for binding. *J. Mol. Biol.* 352, 1134–1156
11 Poroikov, V.V. and Filimonov, D.A. (2002) How to acquire new biological activities in old compounds by computer prediction. *J. Comput. Aided Mol. Des.* 16, 819–824
12 Lagunin, A. *et al.* (2000) PASS: prediction of activity spectra for biologically active substances. *Bioinformatics* 16, 747–748
13 Geronikaki, A.A. *et al.* (2004) Design of new cognition enhancers: from computer prediction to synthesis and biological evaluation. *J. Med. Chem.* 47, 2870–2876
14 Niwa, T. (2004) Prediction of biological targets using probabilistic neural networks and atom-type descriptors. *J. Med. Chem.* 47, 2645–2650
15 Olah, M. *et al.* (2004) WOMBAT: World of Molecular Bioactivity. In *Cheminformatics in Drug Discovery* (Oprea, T.I., ed.), Wiley-VCH, New York
16 Nidhi, *et al.* (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* 46, 1124–1133
17 Paolini, G.V. *et al.* (2006) Global mapping of pharmacological space. *Nat. Biotechnol.* 24, 805–815
18 Choppin, A. and Berry, C.N. (1995) Peripheral benzodiazepine ligands inhibit aggregation and thromboxane synthesis induced by arachidonic acid in rabbit platelets *in vitro*. *Thromb. Res.* 78, 293–302
19 Kauvar, L.M. *et al.* (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* 2, 107–118
20 Beroza, P. *et al.* (2002) Chemoproteomics as a basis for post-genomic drug discovery. *Drug Discov. Today* 7, 807–814
21 Krejsa, C.M. *et al.* (2003) Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discov. Dev.* 6, 470–480
22 Fliri, A.F. *et al.* (2005) Biological spectra analysis: linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U S A* 102, 261–266
23 Covell, D.G. *et al.* (2005) Linking tumor cell cytotoxicity to mechanism of drug action: an integrated analysis of gene expression, small-molecule screening and structural databases. *Proteins* 59, 403–433
24 Rabow, A.A. *et al.* (2002) Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. *J. Med. Chem.* 45, 818–840
25 Klekota, J. *et al.* (2006) Using high-throughput screening data to discriminate compounds with single-target effects from those with side effects. *J. Chem. Inf. Model.* 46, 1549–1562
26 Fischer, H.P. and Heyse, S. (2005) From targets to leads: the importance of advanced data analysis for decision support in drug discovery. *Curr. Opin. Drug Discov. Dev.* 8, 334–346
27 Rosania, G.R. (2003) Supertargeted chemistry: identifying relationships between molecular structures and their sub-cellular distribution. *Curr. Top. Med. Chem.* 3, 659–685

28 Glen, R.C. and Allen, S.C. (2003) Ligand–protein docking: cancer research at the interface between biology and chemistry. *Curr. Med. Chem.* 10, 763–777
29 Warren, G.L. *et al.* (2006) A critical assessment of docking programs and scoring functions. *J. Med. Chem.* 49, 5912–5931
30 Chen, Y.Z. and Zhi, D.G. (2001) Ligand–protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* 43, 217–226
31 Rockey, W.M. and Elcock, A.H. (2005) Rapid computational identification of the targets of protein kinase inhibitors. *J. Med. Chem.* 48, 4138–4152
32 Paul, N. *et al.* (2004) Recovering the true targets of specific ligands by virtual screening of the protein data bank. *Proteins* 54, 671–680
33 Deng, Z. *et al.* (2004) Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein–ligand binding interactions. *J. Med. Chem.* 47, 337–344
34 Oprea, T.I. *et al.* (2005) Post-high-throughput screening analysis: an empirical compound prioritization scheme. *J. Biomol. Screen.* 10, 419–426
35 Bredel, M. and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* 5, 262–275
36 Jain, A.N. (2000) Morphological similarity: A 3D molecular similarity method correlated with protein–ligand recognition. *J. Comput. Aid. Mol. Des* 14, 199–213
37 Sheridan, R.P. and Shpungin, J. (2004) Calculating similarities between biological activities in the MDL drug data report database. *J. Chem. Inf. Comput. Sci.* 44, 727–740
38 Huang, R. *et al.* (2006) Assessment of *in vitro* and *in vivo* activities in the National Cancer Institute's anticancer screen with respect to chemical structure, target specificity, and mechanism of action. *J. Med. Chem.* 49, 1964–1979
39 Fliri, A.F. *et al.* (2005) Biospectra analysis: model proteome characterizations for linking molecular structure and biological response. *J. Med. Chem.* 48, 6918–6925
40 Wallqvist, A. *et al.* (2006) Evaluating chemical structure similarity as an indicator of cellular growth inhibition. *J. Chem. Inf. Model.* 46, 430–437