Bioinformatics Advance Access published November 4, 2015

Bioinformatics, 2015, 1–8
doi: 10.1093/bioinformatics/btv597
Advance Access Publication Date: 26 October 2015
Original Paper

Databases and ontologies

PDID: database of molecular-level putative protein–drug interactions in the structural human proteome

Chen Wang[1], Gang Hu[2], Kui Wang[2], Michal Brylinski[3], Lei Xie[4] and Lukasz Kurgan[1,5,*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada T6G 2V4, [2]School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, People's Republic of China, [3]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA, [4]Department of Computer Science, Hunter College, City University of New York (CUNY), New York, NY 10065, USA and [5]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on June 25, 2015; revised on September 24, 2015; accepted on October 12, 2015

Abstract

Motivation: Many drugs interact with numerous proteins besides their intended therapeutic targets and a substantial portion of these interactions is yet to be elucidated. Protein–Drug Interaction Database (PDID) addresses incompleteness of these data by providing access to putative protein–drug interactions that cover the entire structural human proteome.
Results: PDID covers 9652 structures from 3746 proteins and houses 16 800 putative interactions generated from close to 1.1 million accurate, all-atom structure-based predictions for several dozens of popular drugs. The predictions were generated with three modern methods: ILbind, SMAP and eFindSite. They are accompanied by propensity scores that quantify likelihood of interactions and coordinates of the putative location of the binding drugs in the corresponding protein structures. PDID complements the current databases that focus on the curated interactions and the BioDrugScreen database that relies on docking to find putative interactions. Moreover, we also include experimentally curated interactions which are linked to their sources: DrugBank, BindingDB and Protein Data Bank. Our database can be used to facilitate studies related to polypharmacology of drugs including repurposing and explaining side effects of drugs.
Availability and implementation: PDID database is freely available at http://biomine.ece.ualberta.ca/PDID/.
Contact: lkurgan@vcu.edu

1 Introduction

Majority of the molecular targets of drugs are proteins (Overington *et al.*, 2006; Rask-Andersen *et al.*, 2014) and there are several databases of the already characterized protein–drug interactions. DrugBank (Law *et al.*, 2014; Wishart *et al.*, 2006) provides access to biochemical and pharmacological information about a large set of 7759 drugs, including 1600 Food and Drug Administration (FDA)-approved compounds, and their known 4104 protein targets. Therapeutic Target Database (Zhu *et al.*, 2010, 2012) offers a comprehensive coverage of over 20 000 drugs, including close to 15 000 experimental drugs, and their interactions with 2360 protein targets. This database also links targets and drugs to about 900 diseases. Other databases expand beyond the drug molecules to cover small drug-like ligands. BindingDB (Liu *et al.*, 2007) gives experimentally measured binding affinities between about 7000 known protein targets and a large set of almost half a million of small ligands. ChEMBL (Bento *et al.*, 2014; Gaulton *et al.*, 2012) contains structures, physicochemical properties and bioactivity (e.g. binding constants, pharmacology data) of drug-like small molecules. The current release of ChEMBL incorporates 1.7 million distinct compounds and 13.5 million bioactivity data points which are mapped to over 10 thousand protein targets, where the corresponding binding sites are defined at varying levels of granularity (protein, protein domain or residue level). SuperTarget (Hecker *et al.*, 2012) includes about 6200 protein targets from several dozens of species and close to 200 000 drug-like compounds. It integrates drug-related information from BindingDB, DrugBank and the SuperCyp database of cytochrome–drug interactions (Preissner *et al.*, 2010), adverse drug effects from SIDER (Kuhn *et al.*, 2010a), drug metabolism and pathways and Gene Ontology (GO) terms for the target proteins. The PROMISCUOUS database (von Eichborn *et al.*, 2011) integrates data from DrugBank, SuperTarget and SuperCyp and covers about 6500 protein targets and over 25 thousands drug-like compounds that are annotated with side effects. This database also provides facilities that can be used to predict novel targets based on structural similarity between drugs and between side effect profiles of drugs. STITCH (Kuhn *et al.*, 2010b, 2014) combines information from many sources of experimentally and manually curated interactions between small ligands and proteins including ChEMBL, Protein Data Bank (PDB), DrugBank, Therapeutic Target Database, text mining of articles from MEDLINE and PubMed and several other resources. It currently houses data on 390 000 chemicals and 3.6 million proteins. The recently released IntSide database (Juan-Blanco *et al.*, 2015) links about 1000 drugs with their human protein targets collected from DrugBank and STITCH, and with close to 1200 side effects and other annotations of associated diseases, pathways and cellular functions. Although most of these resources summarize the interactions at the protein or residue level, scPDB (Desaphy *et al.*, 2015; Meslamani *et al.*, 2011) includes molecular-level (all-atom) information for native binding sites in proteins structures collected from PDB (Berman *et al.*, 2000) that are suitable for docking of drug-like ligands. It includes molecular-level details of about 9200 binding sites (all-atom annotation of binding sites and list of ligand-binding residues grouped by various types of bonds) and binding modes (all-atom position of ligand inside the site) in 3600 proteins, and summary of physicochemical properties of approximately 5600 drug-like ligands.

However, many of the established drugs interact not only with the intended therapeutic target protein(s) but also with other protein targets (off-targets). Individual compounds were shown to on average target 6.3 proteins (Hu and Bajorath, 2013; Mestres *et al.*, 2008). Given a high degree of incompleteness of this information (Mestres *et al.*, 2008; Peters, 2013), the number of off-targets is likely substantially higher. To compare, DrugBank includes 15 199 protein–drug interactions for 7759 drugs with the average number of targets per drug at 1.96, which further substantiates incompleteness of the currently available data. Moreover, this polypharmacology can be both beneficial if a given drug can be repurposed for a different disease and harmful, leading to side effects (Peters, 2013). A couple of high-profile examples include imatinib that was repurposed for treatment of gastrointestinal stromal tumors (Hirota *et al.*, 1998) and sorafenib for the kidney and liver cancers (Wilhelm *et al.*, 2006). The incompleteness of the data combined with the importance of polypharmacology motivates research toward elucidation of novel protein–drug interactions. Conventional (non-computational) methods for the identification of novel off-targets rely on an in vitro counter-screen of a given drug against a 'large' set of enzymes and receptors (Bass *et al.*, 2004). Recognizing corresponding implications related to side effects, pharmaceutical companies have implemented screening protocols for the drugs that they currently develop. For instance, Novartis screens against interactions with a panel of 24 targets associated with serious side effects and high hit rates (Urban, 2012), Pfizer screens against between 15 and 30 targets (Wang and Greene, 2012), and Roche uses a panel of 48 targets (Bendels, 2013).

Compared with the experimental screens, computational methods that find novel drug targets are more cost- and time-effective, allow screening of a larger number of targets and provide insights into the molecular-level mechanisms of protein–drug interactions (MacDonald *et al.*, 2006). These *in silico* methods are successful in the context of drug repositioning and identification of off-targets (Liu *et al.*, 2013). A couple of databases that focus on the putative protein–drug and druggable protein–protein interactions (PPIs) were recently released. BioDrugScreen (Li *et al.*, 2010) stores results of docking of about 1600 small drug-like molecules against 1589 known proteins targets in human, which were annotated based on DrugBank and HCPIN (Huang *et al.*, 2008) databases. Docking was ran for close to 2000 cavities on the surfaces of these proteins, for the total of about 3 million receptor–ligand complexes. Druggable Protein–Protein Interaction Assessment System (Dr. PIAS) (Sugaya and Furuya, 2011; Sugaya *et al.*, 2012) is a database of druggable PPIs predicted by a machine learning method. This database lists druggable interactions predicted from over 83 thousand PPIs in human, mouse and rat but they are not associated with specific compounds.

We developed Protein–Drug Interaction Database (PDID) that complements existing repositories and addresses the lack of access to a comprehensive set of putative protein–drug interactions. Based on close to 1.1 million of all-atom predictions over the entire structural human proteome (10 thousand structures for over 3700 proteins), PDID provides access to all putative targets (between 4444 and 7184, depending on the prediction method used) of several dozens of popular drugs. Unique features of our database are:

- It incorporates accurate predictions generated by three methods, ILbind (Hu *et al.*, 2012), SMAP (Xie and Bourne, 2008) and eFindSite (Brylinski and Feinstein, 2013; Feinstein and Brylinski, 2014), which are complementary and independent of docking that was used in the BioDrugScreen database
- It uniformly covers the entire structural human proteome
- It includes molecular-level information on localization of the putative binding sites in the structures of the corresponding protein targets
- It includes comprehensive annotations of known drug targets that are linked to their sources: DrugBank, BindingDB and PDB

The methods that we use were shown empirically to provide high-quality predictions of drug targets (Hu *et al.*, 2012) and their results were already successfully used to predict novel off-targets. Examples include applications to find new off-targets of estrogen receptor modulators (Xie *et al.*, 2007), cholesteryl ester transfer protein inhibitors (Xie *et al.*, 2009b), comtan (Kinnings *et al.*, 2009), inhibitors of Trypanosoma brucei RNA editing ligase 1 (Durrant *et al.*, 2010), nelfinavir (Xie *et al.*, 2011), raloxifene (Sui *et al.*, 2012) and cyclosporine A (Hu *et al.*, 2014b).

2 Methods

2.1 Datasets

We collected the structural human proteome from PDB by removing low resolution structures (>3 Å) and following Hu *et al.* (2014) and

Table 1. Compounds included in the current release 1.1 of PDID

Drug name	Formula	Drugbank ID	PDB ID	# complexes in PDB	Primary use
acetazolamide	C4 H6 N4 O3 S2	DB00819	AZM	22	Treatment of glaucoma, edema and epilepsy
acyclovir	C8 H11 N5 O3	DB00787	AC2	5	Antiviral for herpes, chickenpox, and shingles
adenosine	C10 H13 N5 O4	DB00640	ADN	107	Treatment of cardiac arrhythmia
alendronate	C4 H9 N O7 P2 -4	DB00630	AHD	3	Treatment of osteoporosis
ampicillin	C16 H19 N3 O4 S	DB00415	AIC	8	Antibiotic
bepridil	C24 H34 N2 O	DB01244	BEP	2	Treatment of angina
caffeine	C8 H10 N4 O2	DB00201	CFF	10	Stimulant
captopril	C9 H15 N O3 S	DB01197	MCO	5	Treatment of hypertension
cerulenin	C12 H19 N O3	DB01034	CER	8	Antibiotic
chloramphenicol	C11 H12 CL2 N2 O5	DB00446	CLM	16	Antibiotic
chloroquine	C18 H26 CL N3	DB00608	0TX	1	Treatment of malaria
clavulanate	C8 H9 N O5	DB00766	J01	4	Antibiotic
cyanocobalamin	C63 H88 CO N14 O14 P1	DB00115	CNC	10	Vitamin B12 activity
cyclosporin A	C62 H111 N11 O12	DB00091	CSA	30	Immunosuppressant
didanosine	C10 H12 N4 O3	DB00900	2DI	1	Antiviral for HIV
dopamine	C8 H11 N O2	DB00988	LDP	9	Treatment of hypotension and cardiac arrest
efavirenz	C14 H9 CL F3 N O2	DB00625	EFZ	6	Antiviral for HIV
erlotinib	C22 H23 N3 O4	DB00530	AQ4	3	Anticancer
ertapenem	C22 H27 N3 O7 S	DB00303	1RG	3	Antibiotic
erythromycin	C37 H67 N O13	DB00199	ERY	9	Antibiotic
estradiol	C18 H24 O2	DB00783	EST	28	Hormonal contraception
exemestane	C20 H24 O2	DB00990	EXM	1	Anticancer
furosemide	C12 H11 CL N2 O5 S	DB00695	FUN	3	Treatment of hypertension and edema
gemcitabine	C9 H11 F2 N3 O4	DB00441	GEO	3	Anticancer
ibuprofen	C13 H18 O2	DB01050	IBP	9	Anti-inflammatory
imipenem	C12 H19 N3 O4 S	Db01598	IM2	12	Antibiotic
indomethacin	C19 H16 CL N O4	DB00328	IMN	24	Anti-inflammatory
isoflurane	C3 H2 CL F5 O	DB00753	ICF	2	Anesthetic
kanamycin	C18 H36 N4 O11	DB01172	KAN	21	Antibiotic
l-carnitine	C7 H16 N O3 1	DB00583	152	8	Treatment of heart attack and heart failure
mercaptopurine	C5 H4 N4 S	DB01033	PM6	2	Immunosuppressant
naproxen	C14 H14 O3	DB00788	NPS	4	Anti-inflammatory
niflumic acid	C13 H9 F3 N2 O2	DB04552	NFL	2	Anti-inflammatory
nitroxoline	C9 H6 N2 O3	DB01422	HNQ	1	Antibiotic
pentamidine	C19 H24 N4 O2	DB00738	PNT	7	Antimicrobial
pioglitazone	C19 H20 N2 O3 S	DB01132	P1B	2	Treatment of diabetes
ponatinib	C29 H27 F3 N6 O	DB08901	0LI	3	Anticancer
prednisone	C21 H26 O5	DB00635	PDN	8	Immunosuppressant
progesterone	C21 H30 O2	DB00396	STR	15	Hormone replacement therapy
rifampin	C43 H58 N4 O12	DB01045	RFP	7	Antibiotic
ritonavir	C37 H48 N6 O5 S2	DB00503	RIT	12	Antiviral for HIV
salicyclic acid	C7 H6 O3	DB00936	SAL	36	Treatment of acne
saxagliptin	C18 H25 N3 O2	DB06335	BJM	1	Treatment of diabetes
streptomycin	C21 H39 N7 O12	DB01082	SRY	14	Antibiotic
sulindac	C20 H17 F O3 S	DB00605	SUZ	7	Anti-inflammatory
suramin	C51 H40 N6 O23 S6	DB04786	SVR	12	Antimicrobial
tobramycin	C18 H37 N5 O9	DB00684	TOY	6	Antibiotic
tretinoin	C20 H28 O2	DB00755	REA	30	Treatment of acne
vidarabine	C10 H13 N5 O4	DB00194	RAB	2	Antibiotic
zidovudine	C10 H13 N5 O4	DB00495	AZZ	4	Antiviral for HIV
zoledronate	C5 H10 N2 O7 P2	DB00399	ZOL	12	Treatment of osteoporosis

Xie *et al*. (2007) we kept proteins for which sequences were mapped to human proteins in Ensembl (Hubbard *et al*., 2002). More specifically, structures of chains with at least 90% sequence identity quantified using BLAST (Altschul *et al*., 1990) with default parameters to any human protein from 68th release of Ensembl were selected. As a result, we include total of 9652 human and human-like high resolution structures that correspond to 3746 unique human proteins; the structures are listed at http://biomine-ws.ece.ualberta.ca/PDID/files/list_proteome.txt. Protein chains that correspond to PDB structures were mapped to UniProt (Consortium, 2012) to facilitate mapping of proteins between PDID, PDB, DrugBank and BindingDB.

The database includes drugs which were solved structurally in complex with at least one protein; this is necessary to predict targets. There are 355 such drugs in PDB which we extracted with PDBsum (de Beer *et al*., 2014). The current release 1.1 includes 51 drugs, compared with the release 1.0 that had 26 drugs. These compounds are listed in Table 1 and include popular antibiotics, anti-inflammatory, antiviral and anticancers agents, immunosuppressants and drugs for the treatment of osteoporosis, diabetes, heart attack, hypertension, edema, angina, glaucoma and other diseases. The currently included compounds comprehensively sample the structural drug space; we clustered structures of the 355 drugs using their structural fingerprint expressed with Tanimoto coefficient and sampled at least one drug from each of the resulting 25 clusters to select the 51 compounds.

2.2 Putative protein–drug interactions

Prediction of binding sites from protein structures for a given ligand (drug) are done by searching for sites that are similar to the known sites of this ligand, which are extracted from the structure(s) of the protein–drug complex(es) or by docking the ligand to all binding sites. There are three classes of prediction methods that implement different trade-offs between accuracy and computational cost. These methods are based on searching for the similar sites using a reduced representation of protein structure or complete all-atom structure of protein, and by docking the all-atom structure of ligand into the all-atom structure of the target proteins.

The fastest class of methods utilizes the reduced representation, usually in a form of a numeric vector that summarizes geometry and physicochemical properties of binding sites. Representative examples of such methods that find similar binding sites are PatchSurfer (Hu *et al*., 2014a; Zhu *et al*., 2015) and method by Tomii's group (Ito *et al*., 2012a). The latter algorithm was recently used to create the PoSSuM database (Ito *et al*., 2012b, 2015) that includes 49 million pairs of similar binding sites computed from the known binding sites of 194 drug-like molecules over all protein structures from PDB. Given the large number of these putative sites it is likely that many of them are false positives and would have to be further screened via a more advanced method.

The second class of methods that is characterized by a lower throughput performs docking of a given compounds into protein structures to find which proteins harbor binding sites that are complementary to the given ligand. An example platform that utilizes such type of docking to find targets of a given ligand is INVDOCK (Ji *et al*., 2006). Given the relatively high computational cost of docking, we highlight the availability of the BioDrugScreen database (Li *et al*., 2010). This database stores results of docking with AutoDock and scores these putative interactions based on several scoring functions, such as AutoDock, GoldScore, X-Score, ChemScore, PMF and DFIRE. This docking-based database covers about 1600 drug-like molecules and 2000 cavities on the surfaces of close to 1600 human proteins. However, these results are limited to interactions that are localized in pockets/cavities on the protein surface rather than exploring the whole surface. This is motivated by prohibitively high computational costs of searching the entire surface. BioDrugScreen uses Relibase+ algorithm (Hendlich *et al*., 2003) to identify pockets of interest, while INVDOCK uses an older algorithm by Kuntz *et al*. (1982).

Our database takes advantage of the third class of methods that are complementary to docking. These methods are not constrained to surface pockets and produce accurate predictions of the protein–drug binding at the molecular level. They implement inverse ligand binding where structure(s) of known protein–drug complex(es), called template(s), is used to predict other protein targets together with the corresponding binding sites for the same drug. There are two ways to find novel binding sites based on similarity to known binding sites, one based on the similarity of the corresponding protein fold and another based on similarity of binding pockets. The first approach is implemented by the eFindSite method (Brylinski and Feinstein, 2013; Feinstein and Brylinski, 2014) and the other approach by the SMAP algorithm (Xie and Bourne, 2008). The eFindSite predictor is an improved version of FINDSITE method (Brylinski and Skolnick, 2008; Skolnick and Brylinski, 2009) that uses meta-threading with eThread (Brylinski and Lingam, 2012) and the Affinity Propagation clustering algorithm (Frey and Dueck, 2007) to optimize selection of the ligand-bound templates for a given query structure. It was empirically shown to outperform FINDSITE and several geometrical methods for detection of pockets (Brylinski and Feinstein, 2013). SMAP is based on a sequence order independent profile–profile alignment (SOIPPA) which finds evolutionary and functional relationships across the space of protein structures (Xie and Bourne, 2007, 2008; Xie *et al*., 2009a). SMAP utilizes a shape descriptor to characterize the structure of the protein template and the SOIPPA algorithm to detect and align similar pockets between the query and template proteins. We also include results from a novel meta-method ILbind (Hu *et al*., 2012), which is a machine learning-based consensus of 15 support vector machines that combines prediction scores generated by SMAP and FINDSITE. Details concerning how predictions are performed with SMAP, FINDSITE and ILbind are given in Hu *et al*. (2012). Our recent article shows that ILbind, SMAP and FINDSITE accurately predict targets even when the corresponding structure of the query protein and the template(s) are substantially different, i.e. they are from different Structural Classification of Proteins (SCOP) folds. The corresponding average (over three tested ligands) areas under the receiver operating characteristic (ROC) curve (AUCs) equal 0.727, 0.693 and 0.687 for ILbind, SMAP and FINDSITE, respectively (Hu *et al*., 2012). These results justify our use of the three predictors on the proteome scale.

The PDID database provides access to precomputed results of computationally expensive all-atom predictions by eFindSite and SMAP. Their average runtime for a single protein structure and a given drug is about 30 min on a single processor; the runtime of ILbind is negligible since it is based a consensus of results generated by the two predictors. This high computational cost makes *ad hoc* predictions for a given user query (a given drug or a given protein) computationally impractical.

3 Results

3.1 Assessment of predictive quality

We assessed predictive performance of ILbind, SMAP and eFindSite on a set of 25 representative drugs that are included in PDID. These compounds were selected from 25 clusters of chemically similar drug structures (one compound from each cluster) that were generated from the 355 drugs that can be found in complex with proteins in PDB. The evaluation follows the protocol from (Hu *et al*., 2014b). Briefly, native targets of the 25 drugs were collected from PDB, BindingDB and DrugBank, and we compare predictions from the three methods on the structural human proteome against these native targets. We clustered proteins in the structural human proteome at 90% identity using BLASTCLUST and evaluated the results on the corresponding clusters, i.e. a given cluster is considered to be a native target of given drug (predicted to bind the drug) if at least one protein in this cluster shares at least 90% identity with a native target of that drug (at least one protein in this cluster is predicted to bind that drug). The clustering assures that the evaluation is not biased toward targets that are overrepresented with many structures of similar folds.

Empirical results demonstrate that the three methods are characterized by high predictive quality. The average AUCs over the 25 drugs of eFindSite, SMAP and ILbind equal 0.630, 0.740 and 0.761, respectively (Fig. 1A). Although ILbind outperforms the other two methods, which is expected from this meta-method and consistent with results in Hu *et al*. (2012), different methods perform better for different ligands. More specifically, eFindSite provides the highest

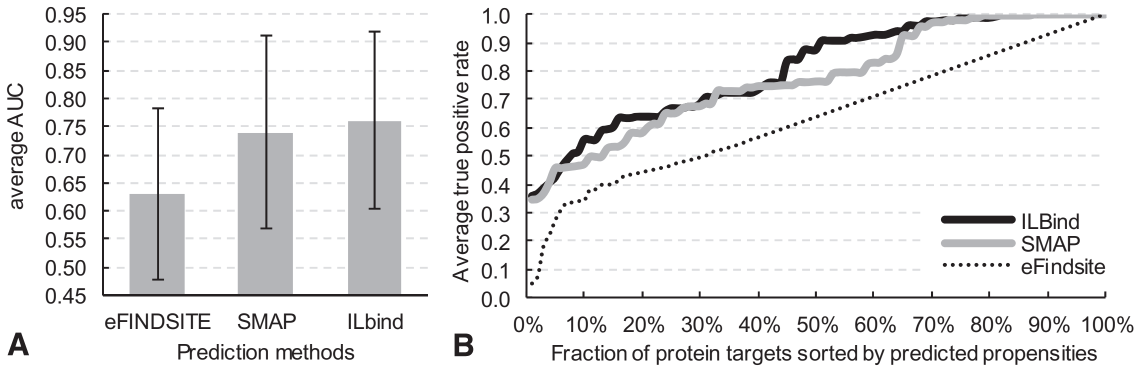

Fig. 1. Predictive quality of eFindSite, SMAP and ILbind for the 25 representative drugs. Panel **A** shows the average AUC computed over the 25 drugs; error bars give the corresponding standard deviations. Panel **B** shows average true positive rate (fraction of correctly predicted native targets) computed over the 25 drugs in the function of the ranking of predictions; the *x*-axis shows fraction of predicted protein targets sorted in the descending order by the predicted propensities for the interaction

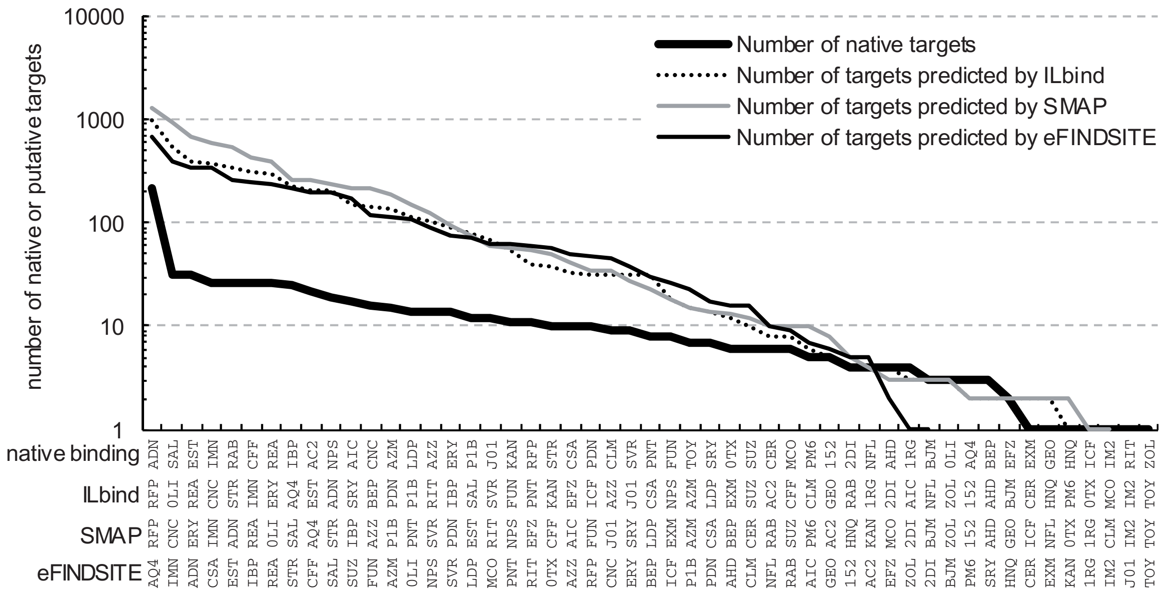

Fig. 2. Number of native and putative targets for the considered 51 drugs. The native targets are based on annotations from PDB, DrugBank and BindingDB. The predictions were generated by ILbind, SMAP and eFindSite. The drugs, which are shown on the *x*-axis, are sorted by their corresponding number of targets in the descending order and separately for each of the four annotations

AUC for 5 drugs, SMAP for 6 drugs and ILbind for the remaining 14 drugs. Figure 1B gives average true positive rates (fractions of correctly predicted native targets) in the function of the fraction of predicted protein targets sorted in the descending order by the propensities for the interaction generated by each of the three predictors. It shows that 40% of the native targets (true positive rate = 0.4) are found in the top 4% of predictions from ILbind and SMAP and in top 14% of predictions from eFindSite.

We note that predictive performance varies between compounds and primarily depends on their size. Higher AUCs are characteristic for medium sized drugs (with molecular weight between 200 and 400 g/mol) and lower AUCs for either small (below 200 g/mol) or large (over 400 g/mol) drugs. To compare, the average AUCs for the small/medium/large drugs for eFindSite, SMAP and ILbind are 0.56/0.68/0.58, 0.7/0.83/0.58 and 0.7/0.86/0.59, respectively. Example small and large compounds for which predictive quality is relatively low are salicyclic acid (138.1 g/mol; average AUC over the three methods of 0.50), isoflurane (184.5 g/mol; 0.60), suramin (1297.3 g/mol; 0.55) and cyanocobalamin (1355.4 g/mol; 0.57). Example drugs for which prediction are more accurate are naproxen (230.3 g/mol; 0.88), furosemide (330.7 g/mol; 0.94) and prednisone (358.4 g/mol; 0.87).

3.2 Database contents and availability

PDID is freely available at http://biomine.ece.ualberta.ca/PDID/. The backend is implemented with the relational MS MySQL database and webpages use PHP script. Protein targets are linked to PDB, UniProt, BindingDB and DrugBank. Drugs are linked to the corresponding records in PDB, BindingDB and DrugBank. Protein and drugs are linked with each other through their known and putative interactions. The interactions are defined at molecular level, i.e. coordinates of the location of the drug in the protein structure file are included. Besides displaying this information in the browser window, PDID allows to download the source files with the sequence and structure of the target proteins. We also offer download of the parsable raw source datasets in text format under the Section 2.1 on the main page. They include the current version of the structural human proteome (IDs of all considered protein structures), list of drugs and predicted targets for each drug together with scores from each of the three prediction methods and the corresponding coordinates of the putative binding sites.

The current version of PDID includes results of about 1.1 million predictions of targets over the 10 thousand structures and 51 drugs with the corresponding 5172, 7184 and 4444 putative targets generated by ILbind, SMAP and eFindSite. It also includes 730 known targets of the 51 drugs mapped from and linked to the corresponding records in DrugBank, BindingDB and PDB. Figure 2 shows the number of native and putative targets for each drug. The median number of putative protein–drug interactions equals 23, 30 and 31 for SMAP, eFindSite and ILbind, respectively, compared with the median of eight based on the known interactions collected from DrugBank, BindingDB and PDB.

The database will be updated semiannually by adding additional drugs and proteins. The initial version 1.0 that included 26 drugs was released in October 2014 and the current version 1.1 in April 2015. This schedule is consistent with other related resources, e.g. scPDB is updated annually, ChEMBL is updated twice a year and DrugBank was recently updated in April 2015 (version 4.2), May 2014 (version 4.1) and December 2013 (version 4.0).

3.3 User interface

The main page includes overview of the contents of the database, access to three available search types (by drug name, by ID of the protein target and by sequence of the protein target), links to the source datasets and related resources and date of the last update. It also includes link to the 'About' page that explains contents of the database and introduces related methods and the 'Help & Tutorial' page that explains the interface of the main page and the three types of output pages that correspond to the three search types.

The search by drug name returns a table with details of known and putative targets including links to the corresponding records in PDB, DrugBank and BindingDB, links to files with structure and sequence of each target and propensities for binding outputted by ILbind, SMAP and eFindSite (Fig. 3A). Targets are sorted by the number of methods that predict them as binding (propensities shown in green font indicate prediction of binding) and by the scores generated by the most accurate ILbind when the number is the same. Detailed description of the formatting and contents of this output page can be found at http://biomine-ws.ece.ualberta.ca/PDID/help.html#drug_page. Each target protein is available as a link that leads to a webpage with the summary of results for this target.

The search by protein ID returns a webpage that maps this ID into corresponding UniProt protein (quality of mapping is annotated using sequence similarity), gives links to the sequence and structure files, provides customizable visualization of the structure together with the localization of the putative (red dots) and known (blue sticks) ligands, and a table that summarizes information about drugs that are known and predicted to bind this protein (Fig. 3B). This information includes color-coded scores generated by each methods that generated prediction and the corresponding predicted location of the drug in the protein structure. We use JSmol (Hanson *et al.*, 2013) to visualize structures and BLAST to compute sequence similarity. Detailed description of this webpage is available at http://biomine-ws.ece.ualberta.ca/PDID/help.html# prot_page.

The search based on protein sequence invokes BLAST that compares the input chain with the target sequences included in the databases. The most similar target is selected given that its similarity

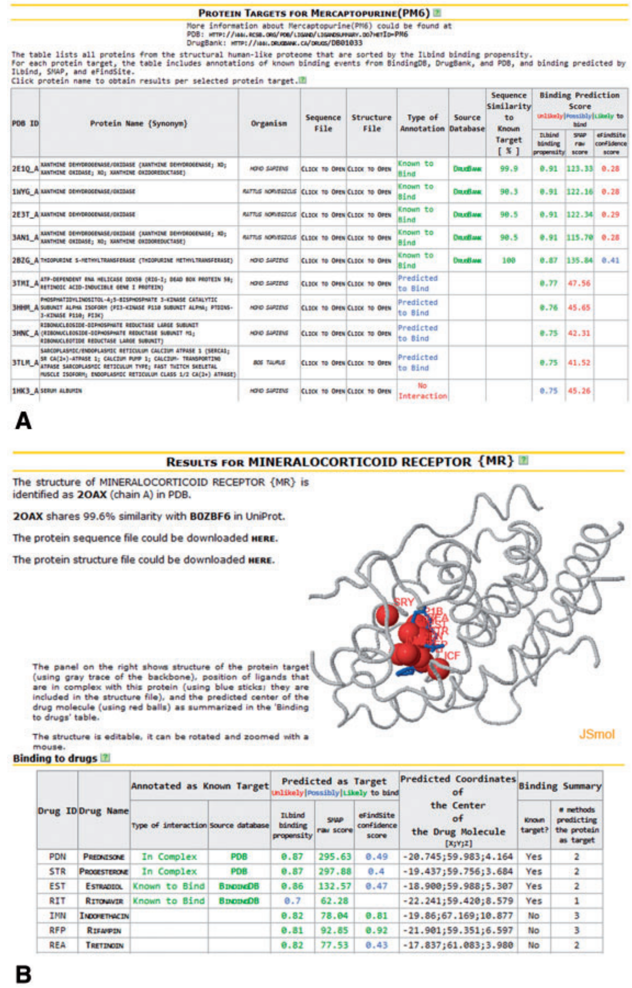

Fig. 3. Results of queries against the PDID database. Panel **A** shows results for a query for mercaptopurine. Detailed description of this webpage is given at http://biomine-ws.ece.ualberta.ca/PDID/help.html#drug_page. Panel **B** gives results form a query for mineralocorticoid receptor protein. Detailed explanations of contents of this page are available at http://biomine-ws.ece.ualberta.ca/PDID/help.html#prot_page.'?' symbol opens the corresponding help page

quantified with the *e*-value is better than a user-defined cutoff; default *e*-value cutoff equals 0.001. The resulting webpage displays the alignment of the query and target proteins and the summary of results for the aligned target protein; the format of the summary is the same as for the query based on the protein ID.

4 Discussion

Numerous drugs are highly promiscuous and we do not know many of their targets. PDID database addresses this issue by providing access to a complete set of putative protein–drug interactions and a set of known protein–drug interactions in the structural human proteome. Our database includes data that otherwise would be accessible only to individuals and research groups with significant computational expertise and resources. The putative interactions were generated by three accurate predictors, ILbind, SMAP and eFindSite, that were shown to produce results that led to finding new drug targets (Durrant *et al.*, 2010; Hu *et al.*, 2014b; Kinnings *et al.*, 2009; Sui *et al.*, 2012; Xie *et al.*, 2007, 2009, 2011) and which complement the existing BioDrugScreen database that relies on docking. The database also integrates annotations of known protein targets collected across DrugBank, BindingDB and PDB, links proteins to the corresponding records in UniProt and provides coordinates of the location of binding sites in the structures of the putative drug targets.

PDID can be used to systematically catalog protein–drug interactions and to facilitate various studies related to polypharmacology of drugs (Xie, 2012), such as explaining side effects caused by interactions with off-targets and for the drug repurposing. Relevant recent examples include use of predictions with ILbind to find three novel off-targets of cyclosporine A that explain nephrotoxicity associated with use of this immunosuppressant (Hu *et al.*, 2014b). Another example involves repurposing of raloxifene, which is used for prevention and treatment of osteoporosis, as a potential compound to treat *Pseudomonas aeruginosa* infections based on predictions with the SMAP method (Sui *et al.*, 2012).

Funding

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) Canada (Discovery grant 298328 to L.K.); the National Natural Science Foundation of China (31050110432 to L.K., 31150110577 to L.K., 11101226 to G.H. and 11301286 to K.W.) and the National Institute of Health (LM011986 to L.X.).

Conflict of Interest: none declared.

References

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bass,A. *et al*. (2004) Origins, practices and future of safety pharmacology. *J. Pharmacol. Toxicol. Methods*, **49**, 145–151.

Bendels, S., *et al*. (2013) Safety screening in early drug discovery: An improved assay profile. *Gordon Research Conference on Computer Aided Drug Design*. Mount Snow, VT. July 21-26, 2013.

Bento,A.P. *et al*. (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Brylinski,M. and Feinstein,W.P. (2013) eFindSite: improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *J. Comput. Aided Mol. Des.*, **27**, 551–567.

Brylinski,M. and Lingam,D. (2012) eThread: a highly optimized machine learning-based approach to meta-threading and the modeling of protein tertiary structures. *PloS One*, **7**, e50200.

Brylinski,M. and Skolnick,J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. USA*, **105**, 129–134.

Consortium,U. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.

de Beer,T.A. *et al*. (2014) PDBsum additions. *Nucleic Acids Res.*, **42**, D292–D296.

Desaphy,J. *et al*. (2015) sc-PDB: a 3D-database of ligandable binding sites–10 years on. *Nucleic Acids Res.*, **43**, D399–D404.

Durrant,J.D. *et al*. (2010) A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. *PLoS Comput. Biol.*, **6**, e1000648.

Feinstein,W.P. and Brylinski,M. (2014) eFindSite: enhanced fingerprint-based virtual screening against predicted ligand binding sites in protein models. *Mol. Inform.*, **33**, 135–150.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Gaulton,A. *et al*. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.

Hanson,R.M. *et al*. (2013) JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. *Isr. J. Chem.*, **53**, 207–216.

Hecker,N. *et al*. (2012) SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res.*, **40**, D1113–D1117.

Hendlich,M. *et al*. (2003) Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.*, **326**, 607–620.

Hirota,S. *et al*. (1998) Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. *Science*, **279**, 577–580.

Hu,B. *et al*. (2014a) PL-PatchSurfer: a novel molecular local surface-based method for exploring protein-ligand interactions. *Int. J. Mol. Sci.*, **15**, 15122–15145.

Hu,G. *et al*. (2012) Finding protein targets for small biologically relevant ligands across fold space using inverse ligand binding predictions. *Structure*, **20**, 1815–1822.

Hu,G. *et al*. (2014b) Human structural proteome-wide characterization of Cyclosporine A targets. *Bioinformatics*, **30**, 3561–3566.

Hu,Y. and Bajorath,J. (2013) Compound promiscuity: what can we learn from current data? *Drug Discov. Today*, **18**, 644–650.

Huang,Y.J. *et al*. (2008) Targeting the human cancer pathway protein interaction network by structural genomics. *Mol. Cell. Proteomics*, **7**, 2048–2060.

Hubbard,T. *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Ito,J. *et al*. (2012a) PDB-scale analysis of known and putative ligand-binding sites with structural sketches. *Proteins*, **80**, 747–763.

Ito,J. *et al*. (2012b) PoSSuM: a database of similar protein-ligand binding and putative pockets. *Nucleic Acids Res.*, **40**, D541–D548.

Ito,J. *et al*. (2015) PoSSuM v.2.0: data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs. *Nucleic Acids Res.*, **43**, D392–D398.

Ji,Z.L. *et al*. (2006) In silico search of putative adverse drug reaction related proteins as a potential tool for facilitating drug adverse effect prediction. *Toxicol. Lett.*, **164**, 104–112.

Juan-Blanco,T. *et al*. (2015) IntSide: a web server for the chemical and biological examination of drug side effects. *Bioinformatics*, **31**, 612–613.

Kinnings,S.L. *et al*. (2009) Drug discovery using chemical systems biology: repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.*, **5**, e1000423.

Kuhn,M. *et al*. (2010a) A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, **6**, 343.

Kuhn,M. *et al*. (2010b) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*, **38**, D552–D556.

Kuhn,M. *et al*. (2014) STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res.*, **42**, D401–D407.

Kuntz,I.D. *et al*. (1982) A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, **161**, 269–288.

Law,V. *et al*. (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.

Li,L. *et al*. (2010) BioDrugScreen: a computational drug design resource for ranking molecules docked to the human proteome. *Nucleic Acids Res.*, **38**, D765–D773.

Liu,T. *et al*. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.

Liu,X. *et al*. (2013) Predicting targeted polypharmacology for drug repositioning and multi- target drug discovery. *Curr. Med. Chem.*, **20**, 1646–1661.

MacDonald,M.L. *et al*. (2006) Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nat. Chem. Biol.*, **2**, 329–337.

Meslamani,J. *et al*. (2011) sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics*, **27**, 1324–1326.

Mestres,J. *et al*. (2008) Data completeness - the Achilles heel of drug-target networks. *Nat. Biotechnol.*, **26**, 983–984.

Overington,J.P. *et al*. (2006) Opinion - how many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.

Peters,J.U. (2013) Polypharmacology - foe or friend? *J. Med. Chem.*, **56**, 8955–8971.

Preissner,S. *et al*. (2010) SuperCYP: a comprehensive database on Cytochrome P450 enzymes including a tool for analysis of CYP-drug interactions. *Nucleic Acids Res.*, **38**, D237–D243.

Rask-Andersen,M. *et al.* (2014) The druggable genome: evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annu. Rev. Pharmacol.*, **54**, 9–26.

Skolnick,J. and Brylinski,M. (2009) FINDSITE: a combined evolution/structure-based approach to protein function prediction. *Brief. Bioinform.*, **10**, 378–391.

Sugaya,N. and Furuya,T. (2011) Dr. PIAS: an integrative system for assessing the druggability of protein-protein interactions. *BMC Bioinformatics*, **12**, 50.

Sugaya,N. *et al.* (2012) Dr. PIAS 2.0: an update of a database of predicted druggable protein-protein interactions. *Database*, **2012**, bas034.

Sui,S.J.H. *et al.* (2012) Raloxifene attenuates *Pseudomonas aeruginosa* pyocyanin production and virulence. *Int. J. Antimicrob. Agents*, **40**, 246–251.

Urban,L. (2012) Translational value of early target-based safety assessment and associated risk mitigation. *4th Annual Predictive Toxicology Summit*. London, UK. February 15-16, 2012.

von Eichborn,J. *et al.* (2011) PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Res.*, **39**, D1060–D1066.

Wang,X.Y. and Greene,N. (2012) Comparing measures of promiscuity and exploring their relationship to toxicity. *Mol. Inform.*, **31**, 145–159.

Wilhelm,S. *et al.* (2006) Discovery and development of sorafenib: a multikinase inhibitor for treating cancer. *Nat. Rev. Drug Discov.*, **5**, 835–844.

Wishart,D.S. *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.

Xie,L. and Bourne,P.E. (2007) A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics*, **8 (Suppl. 4)**, S9.

Xie,L. and Bourne,P.E. (2008) Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc. Natl. Acad. Sci. USA*, **105**, 5441–5446.

Xie,L. *et al.* (2007) In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comput. Biol.*, **3**, e217.

Xie,L. *et al.* (2009a) A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*, **25**, i305–i312.

Xie,L. *et al.* (2009b) Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput. Biol.*, **5**, e1000387.

Xie,L. *et al.* (2011) Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. *PLoS Comput. Biol.*, **7**, e100203.

Xie,L. *et al.* (2012) Predicting the polypharmacology of drugs: identifying new uses through chemoinformatics, structural informatics, and molecular modeling-based approaches. In: Barratt,M.J. and Faril,D.E. (eds.) *Drug Repositioning: Bringing New Life to Shelved Assets and Existing Drugs*. Wiley, Hoboken, NJ, USA, 163-205.

Zhu,F. *et al.* (2010) Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.*, 38, D787–D791.

Zhu,F. *et al.* (2012) Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.*, **40**, D1128–D1136.

Zhu,X. *et al.* (2015) Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2.0. *Bioinformatics*, **31**, 707–713.