Bioinformatics, 31(24), 2015, 3922–3929
doi: 10.1093/bioinformatics/btv477
Advance Access Publication Date: 18 August 2015
Original Paper

OXFORD

Structural bioinformatics

An integrated structure- and system-based framework to identify new targets of metabolites and known drugs

Hammad Naveed[1,2], Umar S. Hameed[3], Deborah Harrus[4,5], William Bourguet[4,5], Stefan T. Arold[2,3,*] and Xin Gao[1,2,*]

[1]Computer, Electrical and Mathematical Sciences and Engineering Division, [2]Computational Bioscience Research Center, [3]Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, [4]Inserm U1054, Centre de Biochimie Structurale and [5]CNRS UMR5048, Universités Montpellier 1 & 2, Montpellier, France

*To whom correspondence should be addressed.
Associate Editor: Anna Tramontano

Received on April 18, 2015; revised on July 16, 2015; accepted on August 8, 2015

Abstract

Motivation: The inherent promiscuity of small molecules towards protein targets impedes our understanding of healthy versus diseased metabolism. This promiscuity also poses a challenge for the pharmaceutical industry as identifying all protein targets is important to assess (side) effects and repositioning opportunities for a drug.

Results: Here, we present a novel integrated structure- and system-based approach of drug-target prediction (iDTP) to enable the large-scale discovery of new targets for small molecules, such as pharmaceutical drugs, co-factors and metabolites (collectively called 'drugs'). For a given drug, our method uses sequence order–independent structure alignment, hierarchical clustering and probabilistic sequence similarity to construct a probabilistic pocket ensemble (PPE) that captures promiscuous structural features of different binding sites on known targets. A drug's PPE is combined with an approximation of its delivery profile to reduce false positives. In our cross-validation study, we use iDTP to predict the known targets of 11 drugs, with 63% sensitivity and 81% specificity. We then predicted novel targets for these drugs—two that are of high pharmacological interest, the peroxisome proliferator-activated receptor gamma and the oncogene B-cell lymphoma 2, were successfully validated through *in vitro* binding experiments. Our method is broadly applicable for the prediction of protein-small molecule interactions with several novel applications to biological research and drug development.

Availability and implementation: The program, datasets and results are freely available to academic users at http://sfb.kaust.edu.sa/Pages/Software.aspx.

Contact: xin.gao@kaust.edu.sa and stefan.arold@kaust.edu.sa

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Most metabolites and pharmaceutical drugs bind to more than one protein (Reddy and Zhang, 2013), resulting in a phenotype composed of many molecular (side) effects. For the pharmaceutical industry, predicting and minimizing off-target effects is important because they are the source of low efficacy and high toxicity that result in a high failure rate of new drugs in clinical trials (Arrowsmith, 2011a,b; Liebler and Guengerich, 2005). Recent studies estimate

that each drug on average binds to at least six known and several unknown targets (Lounkine *et al.*, 2012; Mestres *et al.*, 2009). Thus, knowledge of off-target effects can help reduce drug resistance and provide opportunities for multi-target drug development (Peters, 2013). Moreover, off-target ligands for a given drug may inspire 'drug repositioning', where a drug already approved for one condition is redirected to treat another condition, thereby overcoming delays and costs associated with clinical trials and drug approval (Ashburn and Thor, 2004). Therefore, predicting off-target binding sites to comprehensively understand the side effects of drugs and exploit drug repositioning opportunities is central for rapid, cost-efficient drug development.

In addition to drug development, identifying all cellular targets of a given biological cofactor, metabolite or other small molecules is of great importance for understanding cellular function and dysfunction in general (e.g. metabolome-target interactions and associated diseases) (Alam *et al.*, 2014). Finally, identifying possible targets of environmental pollutants may help us to understand and avoid health hazards from released chemicals.

Computational methods to predict new targets for existing endogenous or administered small-molecule compounds are, therefore, of high biological and pharmacological value. (For simplicity, we herein refer to all these compounds collectively as 'drugs', meaning 'a small-molecule chemical substance with effects on a biological system'.) These methods can be divided into three broad categories: structure based, expression based and ligand based. Structure-based methods utilize information from drug targets by employing binding site similarity or molecular docking (Chang *et al.*, 2010; Engin *et al.*, 2012; Kinnings *et al.*, 2009; Li *et al.*, 2011); expression-based methods exploit molecular activity perturbation signatures that result from drug activity (Chen *et al.*, 2009; Emig *et al.*, 2013; Hu and Agarwal, 2009; Iorio *et al.*, 2010; Lamb *et al.*, 2006; Suthram *et al.*, 2010; Wei *et al.*, 2006); and ligand-based methods utilize the chemical and structural properties of a drug to discover new targets (Keiser *et al.*, 2009; Noeske *et al.*, 2006; Qu *et al.*, 2009). In addition to these methods, previously unknown targets for existing drugs have also been predicted using side effect similarity (Campillos *et al.*, 2008), genome-wide association studies (Sanseau *et al.*, 2012) and medical genetics (Wang and Zhang, 2013). Recently, methods that combine information from multiple sources have been introduced and will likely become the preferred approach (Napolitano *et al.*, 2013); however, because most of these methods were not benchmarked on drugs with known targets (sensitivity analysis), it is difficult to evaluate their success rate. To date, the only studies to report a true positive prediction rate have done so at relatively low rates (29% and 49%, respectively) (Chang *et al.*, 2010; Li *et al.*, 2011). Moreover, a high-throughput framework based on structural information remains unavailable, and current methods do not satisfactorily capture the structural flexibility of drugs that can adopt several conformations, allowing them to interact differently with different targets.

In this study, we propose a novel computational drug target prediction method that integrates structural signatures of small-molecule compounds with their tissue delivery profiles. iDTP incorporates four major developments: (i) Unlike previous methods, this framework is generic and does not target a specific drug. (ii) iDTP uses the probabilistic pocket ensemble (PPE) to capture the promiscuous nature of different binding pockets for the same drug. (iii) iDTP uses the approximated drug delivery profile (aDDP) of the respective drug to predict biologically relevant targets. The drug delivery profile (DDP) is defined as the distribution of drug concentrations in different tissues after circulation. Since such information is not directly available, we approximated the DDP, denoted as aDDP, as the average of the mRNA expression of the known drug targets in 79 human tissues. Thus, an aDDP is a vector of length 79. (iv) iDTP has a performance guarantee supported by (i) cross validation on a benchmark dataset; (ii) *in vitro* binding experiments; and (iii) large-scale text mining. Application of iDTP allowed us to propose a novel cellular target for coenzyme A (CoA), a novel druggable pocket and lead compound for Bcl-2, and plausible mechanistic information for the inhibition of CYP2E1 by Trolox.

2 Methods

2.1 Dataset

We extracted the approved/experimental drugs from the DrugBank database (version 3) (Knox *et al.*, 2011). Eleven drugs selected for use in this study had at least one 3D structure of a drug–protein complex and more than 40 known drug targets with solved apo structures (Supplementary Table S1). During the *in silico* validation of our method, a 5-fold cross-validation is done on the known targets of each drug to evaluate how well our method can recover the known targets. For example, for a drug with 40 known targets, only 32 structures are used to construct the PPE for each fold. When our method is used to predict new targets, all the known targets are used to construct the PPE for a drug. Therefore, in this study, our dataset contains drugs with 40 or more known targets because our experiments involve 5-fold cross-validation. In practical use of our method, 30 known targets are sufficient. We expect this number to be further reduced in the future. We expect that our method is also suitable for much larger sets of drugs established for proprietary research that are not detailed in public databases.

Protein structures have more than 30 pockets on average (some structures have >100 pockets), and a majority of the small-molecule protein interactions occur in the three largest pockets (Huang and Schroeder, 2006). A typical pocket involved in small-molecule protein interactions (also known as a druggable pocket) has characteristic values for pocket solvent accessible surface area (300–600 $Å^2$) and pocket volume (400–600 $Å^3$) (Gao and Skolnick, 2013; Pérot *et al.*, 2010). We hypothesize that a protein structure that has a minimal (i.e. less than three) number of pockets and no druggable pocket is unlikely to interact with a drug. To form the negative dataset, we extracted the protein structures with fewer than three pockets from the CASTp database (Dundas *et al.*, 2006). We extracted the surface residues (>70% of the solvent-accessible surface area) of these protein structures using POPS after redundancy reduction using the PISCES webserver (<60% pairwise sequence identity) (Cavallo *et al.*, 2003; Wang and Dunbrack, 2003). We further removed the NMR structures, the structures with cocrystallized DNA, RNA or ligands (excluding ions like Zn^{2+}, Cl^-), and the structures with druggable pockets. Our negative dataset contained a total of 63 protein structures (a detailed list of PDB IDs can be found in the Supplementary file NEG dataset). Note that our negative dataset thus can still contain protein structures with less than three pockets. However, if they do so, none of the pockets can be druggable. The surface residues of the structures in the negative dataset were aligned with the PPEs of all drugs in the dataset by CPalign (Dundas *et al.*, 2007). The alignments were then scored by the scoring function defined in Section 2.3. Ideally, each of these alignments should have a bad (i.e. high) score, so it will not be predicted to be a drug target. Otherwise, it is counted as a false-positive prediction when we evaluate the specificity of our method (see Section 3.2).

2.2 Constructing a PPE

The overview flowchart of our method is shown in Supplementary Figure S1. After identifying the drug–target complex for each drug in the dataset, we extracted the pocket that the drug binds to in the protein structure using the CASTp webserver (Dundas *et al.*, 2006), which we refer to as the 'bound pocket'. To identify the drug-binding site in apo structures of known drug targets, we extracted the three largest pockets from the first chain of their respective 3D structures. We used sequence order-independent alignment to choose the pocket most similar to the bound pocket (Cui *et al.*, 2015; Dundas *et al.*, 2011). Dundas *et al.* (2011) established a method to construct the structural signatures for enzyme binding pockets that require high-quality, manually curated enzyme binding sites and is, therefore, not suitable for high-throughput studies. However, we reduced this requirement by using just one manually curated pocket (except for nicotinamide-adenine-dinucleotide where we used both bound structures that are available) and predicting the binding pockets on the rest of the targets from their apo (unbound) structures to construct the PPE for each drug. Conversely, Dundas *et al.* manually searched the literature to find residues that are important for the interaction and mapped them back onto the apo structures. The PPE represents a unified set of individual pockets that potentially bind to several conformations of the drug.

Extraction of the common structural features from the set of binding pockets is ideally performed using a multiple structure alignment method. However, because no such method currently exists that can handle our dataset, we followed Dundas *et al.* (2011) by first using pairwise sequence order-independent structure alignment of surface pockets and then using hierarchical clustering based on the pairwise similarities. Dundas *et al.* constructed several structural signatures corresponding to the different ligand/ligand binding site conformations at a predefined specific level of the hierarchical tree. In most cases, identifying this cutoff is nontrivial and requires in-depth knowledge about the different conformations of the ligand/ligand binding site. In contrast, we constructed the structural signature at the root of the tree. The hierarchical tree is used as a guide to recursively combine sibling pockets along the paths from leaf nodes to the root. A signature pocket is computed as the average of two child (signature) pockets, and the two original child nodes are replaced with a new single leaf node on the hierarchical tree.

As a result, the structural signature is an ensemble of more than one unique pocket (corresponding to distinct branches in the hierarchical tree). Each position in the PPE has a preservation ratio (how often that particular atom was present in the underlying set of pockets) (Dundas *et al.*, 2011) of at least 0.5 (each atom was present in at least half of the structures that have an atom present after alignment at this position). To achieve a minimalistic ensemble and reduce the computational time, we increased the preservation ratio cutoff to 0.6 if the number of atoms in the PPE was greater than 110. For specific drugs, a stricter conservation ratio of atoms essential for binding action of the respective drugs can be readily incorporated in our method.

2.3 Calculating distance to the PPE

Each position in the structural signature may be occupied by more than one type of atom (which can be from different residues). Therefore, we formulated a probabilistic distance function to accommodate this property. The distance function of a query protein to the already constructed PPE has both structural and sequence components. The structural component follows Dundas *et al.*'s approach, while the sequence component is based on maximum likelihood.

$$Score = Structural\ score + \alpha\ *\ Sequence\ score$$
$$Structural\ score = RMSD * N^{(-1/3)}$$
$$Sequence\ score = 1 - (Sequence\ similarity/Best\ sequence\ similarity)$$
$$Sequence\ similarity = \Sigma_i\ (AtomFreq_i + ResFreq_i)$$
$$Best\ sequence\ similarity = \Sigma_i\ (MaxAtomFreq_i + MaxResFreq_i),$$

where the value for α is set to 1.2 following Dundas *et al.* (2011), RMSD is the root mean square distance after the alignment, N is the number of positions aligned, $AtomFreq_i/ResFreq_i$ is the frequency of aligned atom/residue at position i, $MaxAtomFreq_i/MaxResFreq$ is the highest frequency of any atom/residue at position i, and their summation is over all the aligned positions. An empirical distance cutoff of 0.85 that maps to an RMSD of 0.7 Å and pocket sequence similarity of 60% for a sequence order-independent alignment of 12–15 atoms is used in this study. An alignment should also contain at least five atoms.

2.4 Integrating aDDP

We included an approximation for the DDP as an orthogonal source of structure-independent information. The aDDP for each drug is calculated by averaging the mRNA expression of known drug targets over 79 human tissues from Su *et al.* (2004). We mapped each drug–target structure to a gene using Uniprot ID mapping service (Wu *et al.*, 2006); the gene was then searched in the tissue expression dataset compiled by Su *et al.* (2004). Because a protein structure could be mapped to more than one gene, the average expression of all the mapped genes was used. We classified expression in 79 human tissues into three classes (low, medium and high) based on empirical cutoffs for mRNA expression (<300, <1000 and ≥ 1000, respectively). The new drug target list is reordered by including the drug–target tissue expression term in the distance function as follows:

$$Score = Structural\ similarity + \alpha\ *\ (1 - Sequence\ similarity) + \beta\ *\ Tissue\ expression$$
$$Tissue\ expression = 1 - (Number\ of\ tissue\ with\ matching\ expression/Total\ number\ of\ tissues),$$

where β is empirically set to 0.4. If a gene is not present in the dataset compiled by Su *et al.*, we set (β*Tissue expression) to 0.2. Our method is not sensitive to the specific value of β in the range of 0.3–0.5. β plays an important role in differentiating true targets from the false targets and the most promising targets from less promising ones. However, it usually does not play an important role in ranking the top 10 targets as the top 10 predicted targets have an almost perfect match between their mRNA expression profile and that of the estimated DDP.

2.5 *In vitro* experimental setup

Detailed sections on protein expression and purification; ligands and peptides; differential scanning fluorimetry; differential static light scattering; fluorescence anisotropy (FA) measurements; and intrinsic tryptophan fluorescence quenching binding assay can be found in the Supplementary information.

3 Results

3.1 Probabilistic pocket ensemble

An inherently promiscuous drug can bind to different protein pockets that have a range of features, making it difficult to establish a general description of a drug's possible binding sites. To capture the essential binding site features of a promiscuous drug, we developed

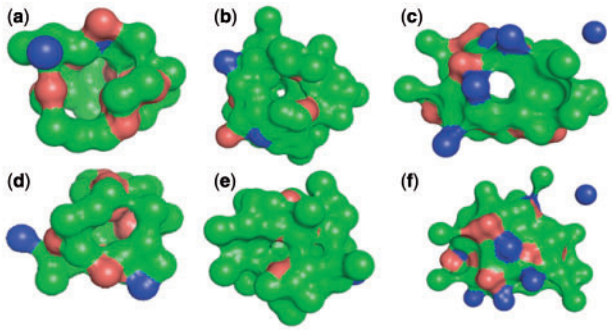

Fig. 1. The PPE of formic acid (**a, d**), β-D-glucose (**b, e**) and phosphoaminophosphonic acid-adenylate ester (**c, f**) (Top view: a–c, Side view: d–f). Each position is labeled with the atom of highest frequency. The PPE represents a unified set of individual pockets that potentially bind to several conformations of the drug. The atoms are color coded as C: green, O: red and N: blue

a method to construct its PPE (see Section 2 for details). The PPE represents a unified set of individual pockets that potentially bind to several conformations of the drug. Each position in the PPE can consist of a number of atoms from different residues. The frequency of the atoms and residues at each position is recorded and used to construct a maximum likelihood sequence similarity scoring function. This probabilistic scoring method adequately accounts for the fact that a drug can bind several pockets and a pocket can bind several drugs (Gao and Skolnick, 2013). The PPEs of formic acid, β-D-glucose and phosphoaminophosphonic acid-adenylate ester are shown in Figure 1; each position in the PPE is labeled with the atom of highest frequency.

3.2 Evaluation of the PPE and the probabilistic scoring function

To investigate the capacity of PPE to retrieve structurally similar drug-binding pockets, we compared the protein pocket bound by 2′-monophosphadenosine 5′-diphosphoribose with the predicted binding pockets of the top 10 predicted targets of this compound using sequence order-independent and sequence order-dependent alignments (see Supplementary Section S1). Our results suggest that a combination of the minimalistic PPE and sequence order-independent alignment is more powerful in identifying new drug targets than the combination of the complete binding pocket with sequence order-dependent structure alignment. Moreover, we showed that the probabilistic similarity function performs better than the deterministic similarity function used by Dundas *et al.* (2011) (see Supplementary Section S1). The PPE is able to extract non-trivial sequence and structure signatures that are necessary for capturing the promiscuous process of a drug binding to multiple sites and the sites binding to multiple drugs. Most of the predicted targets spatially align with distinct parts of a respective drug's PPE (Supplementary Figure S2), suggesting that the PPE is indeed an ensemble of several pockets and, therefore, can accommodate different conformations of each drug. In contrast to previous studies (Dundas *et al.*, 2011; Tseng and Liang, 2006), these results suggest that multiple structural signatures may not be optimal for capturing different drug conformations, but instead, this can be achieved by the incorporation of a probabilistic scoring function in structural signatures. Moreover, our methodology does not require in-depth details about the number of binding conformations or the number of structural signatures.

In the next step, we used cross-validation (see Supplementary section S3 for details) to assess whether the PPE for each drug captures the essential features for the drug–protein interaction. We were able to predict the interaction of drugs with known targets to an average sensitivity of 63% (Supplementary Table S1). We also constructed a negative dataset (a benchmark dataset is not available for such studies) to assess the specificity of the methodology. Construction of such a negative dataset is non-trivial because of the inherent promiscuity of drug binding sites and incomplete knowledge of drug targets (see Section 2 for details). We found the average specificity of this method to be 81% (Supplementary Table S1).

3.3 Integrating the DDP to reduce false positives

To identify new drug targets, we downloaded the CASTp database (Dundas *et al.*, 2006), which consists of 75 000 protein structures and their pockets. We extracted the three largest pockets, reported to account for more than 80% of a protein's small-molecular binding sites (Huang and Schroeder, 2006; Laskowski, 1995; Liang *et al.*, 1998; Peters *et al.*, 1996), from each of these protein structures. We aligned the pockets of each protein structure with the PPE (constructed using all the known drug targets) of each drug in our dataset using sequence order-independent structure alignment. This resulted in several thousand hits, a number similar to those of other drug repositioning studies (Keiser *et al.*, 2009). Although our method has high specificity for the curated dataset, the false-positive rate is expected to be higher in a general database search because our construction of the PPE is minimalistic, and therefore it can align with several unrelated protein surfaces randomly (Dundas *et al.*, 2011; Watson *et al.*, 2005).

To reduce the false positive rate of our method, we included an approximation for the aDDP as an orthogonal source of structure-independent information. Given that the actual tissue delivery profile for a specific drug is generally not available, we reasoned that the intracellular delivery profile of this drug has to be compatible with the mRNA expression profile of its established targets. In other words a protein can only be a target of a given drug if the drug is delivered into (or produced in) the tissues in which the protein is expressed at significant levels. For each candidate drug, we therefore approximate its delivery profile by averaging the mRNA expression profiles of all its known targets in a set of 79 human tissues (Su *et al.*, 2004). For the drugs tested, the mRNA expression profiles of the known target proteins are similar for the same drug (e.g. the Pearson correlation coefficient of aDDP of CoA with its known targets is 0.56), but are different between different drugs (e.g. the average Pearson correlation coefficient of aDDP of CoA with the aDDP's of the rest of the drugs is 0.44) as shown in Figure 2. The mRNA expression profiles not only provide information about protein localization but also provide information about protein–protein interactions and pathways (Jansen *et al.*, 2002). Thus, the comparison of the average tissue expression profile of the established drug targets with the expression profiles of the predicted target is expected to reflect the likelihood for drug–target interactions in a particular set of tissues and hence can be used as a proxy for the drug delivery.

3.4 Validation using *in silico* experiments

We used text mining to investigate the capacity of an aDDP to predict drug–target interactions. A cocitation index finds the association between two terms (in this case the name of a drug and a gene) by comparing the number of times the two terms appear in the abstract of studies in the PubMed library when compared with two random terms (Qiao *et al.*, 2013). We found that, when the aDDP was combined with the PPE, the number of predictions with a

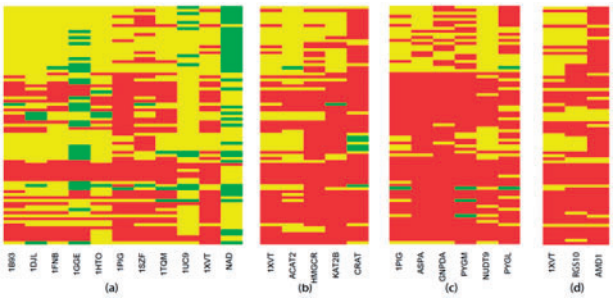

Fig. 2. (**a**) aDDPs of 11 drugs investigated in this study. (**b**) aDDP of CoA (1XVT) and the mRNA expression profile of four known CoA targets (ACAT2, HMGCR, KAT2B, CRAT), the Pearson correlation coefficient of aDDP of CoA with its known targets is 0.56. (**c**) aDDP of β-D-glucose (1PIG) and the mRNA expression profile of five known β-D-glucose targets (ASPA, GNDPA, PYGM, NUDT9, PYGL). (**d**) The mRNA expression profile of RGS10 matches the aDDP of CoA in 65/79 tissues, while the mRNA expression profile of AMD1 matches the aDDP of CoA in 46/79 tissues. In this case, RGS10 will be preferred over AMD1 as the predicted target of CoA. Color code: Red (low expression), Yellow (medium expression) and Green (high expression). *Y*-axis has the 79 human tissues

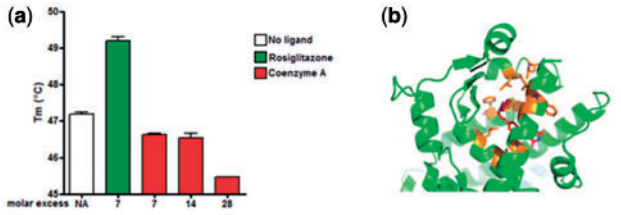

Fig. 3. (**a**) Thermal shift assays on hPPARγ-LBD. Melting temperatures (T_m) calculated from thermal denaturation curves of hPPARγ-LBD in the presence of varying molar excess of Rosiglitazone or CoA. Rosiglitazone displays a protective effect (increased T_m) against thermal denaturation, while CoA displays a destabilizing effect (decreased T_m). (**b**) The predicted CoA binding site overlaps with the ligand-binding site on hPPARγ-LBD. The figure is based on the crystal structure of hPPARγ-LBD (green) bound to rosiglitazone (red; PDB ID 4EMA). The predicted CoA binding pocket is shown in orange

statistically significant cocitation index was 2- to 4-fold higher than using aDDP alone (see Supplementary Section S2 and Table S2). The top 10 predicted targets (<60% sequence similarity with any of the known drug targets) of β-D-glucose using the combined approach aligned extremely well with its PPE (Supplementary Table S3). Similar results were observed for all the drugs in the dataset (see Supplementary file PredTargets). Moreover, six of the top ten predictions for β-D-glucose and four predictions for CoA have a statistically significant cocitation index (P-value < 0.05). We found a total of 34 predicted targets with cocitation index values of statistical significance (P-value < 0.05) for all the drugs in our dataset. For the top 10 predicted targets of all the drugs in the dataset, the range of the average sequence-similarity score between the PPEs and the predicted pocket on the targets was between 80 and 87%, the range of the average RMSD between the PPEs and the predicted pocket on the targets was between 0.62 and 0.66 Å, the range of the average match of mRNA expression profile between 72 and 76 out of the 79 tissues and the range of the average final score was between 0.45 and 0.56 (compared to our cutoff value of 0.85 in our cross-validation study). These results support that the combination of PPE and aDDP into iDTP allows identifying novel proteins that have high structural (binding site) and system-level similarity with known drug targets.

3.5 Validation using *in vitro* binding experiments

To provide an experimental assessment of the performance of iDTP, we chose to test the predicted targets for CoA because it has the least number of known binding proteins in our dataset and hence has the least well-defined PPE. We used multiple *in vitro* binding experiments to test binding site and affinity for two top hits from iDTP, namely the peroxisome proliferator-activated receptor gamma (PPARγ) and B-cell lymphoma 2 (Bcl-2).

PPARγ is a nuclear hormone receptor that regulates numerous biological functions including adipogenesis and cell differentiation. Its dysregulation is involved in the onset of diabetes and obesity (Swedenborg *et al.*, 2009). The interaction of the ligand binding domain (LBD) of human PPARγ (hPPARγ-LBD) with CoA is one of our most promising predictions because this interaction achieved a high iDTP score and a statistically significant cocitation index (Supplementary Table S4). The pocket predicted by iDTP to bind CoA overlaps with the known ligand binding site of PPARγ. To test this prediction *in vitro*, we used differential scanning fluorimetry to measure the melting temperature (T_m) of hPPARγ-LBD in the presence or absence of CoA or rosiglitazone, an antidiabetic drug known to act as a ligand of hPPARγ-LBD (Fig. 3a and Supplementary Table S5). At a molar excess of seven, rosiglitazone had a protective effect by raising the T_m of hPPARγ-LBD's by 2°C from that of the apo protein, while CoA displayed a destabilizing effect, lowering the T_m by 0.8°C from that of the apo protein, suggesting a direct interaction with hPPARγ-LBD. Next, we used fluorescence anisotropy (FA) to characterize the interaction between hPPARγ-LBD and its natural partner proteins upon binding with CoA. We measured K_d between hPPARγ-LBD and fluorescein-labeled peptides derived from a coactivator protein (PGC1) and two corepressor proteins (NCoR and SMRT). These experiments were performed in the presence or absence of increasing molar excess of CoA or the reference hPPARγ-LBD agonist rosiglitazone or antagonist CD5477 (LeMaire *et al.*, 2009). If an increasing molar excess of the ligand causes the fluorescently labeled coactivator/corepressor K_d to increase, we can infer that ligand binding is taking place because the ligand disturbs binding to coactivators or corepressors. The nature of the ligand-hPPARγ-LBD interaction can also be inferred: an agonist ligand enhances binding to a coactivator and decreases binding to a corepressor; an inverse agonist causes the opposite effect; and a neutral antagonist decreases binding for both coactivators and corepressors. Accordingly, adding a 2–10 M excess of the agonist rosiglitazone hPPARγ-LBD raised the affinity of hPPARγ-LBD for the coactivator PGC1 (Figs 4a and b, Supplementary Table S6). Conversely, a 2 M excess of the antagonist CD5577 lowered the affinity of hPPARγ-LBD for the coactivator PGC1 (Figs 4a and b, Supplementary Table S6), whereas the addition of 2–10 M excess of CoA lowered the affinity of hPPARγ-LBD for both coactivator PGC1 (Fig. 4b) and corepressors NCoR and SMRT (Figs 4c and d), Supplementary Table S6). Collectively, our experiments confirm a direct interaction between CoA and hPPARγ-LBD in which CoA behaves as a neutral antagonist. From its potency in competing with known ligands and its dose-dependent stabilization of hPPARγ-LBD, we estimate an apparent K_d of $<500\,\mu$M.

We also tested direct binding of CoA to recombinant Bcl-2 (see Supplementary file PredTargets). Bcl-2, the founding member of the Bcl-2 family of proteins that control cell death, is an important anti-apoptotic protein and is classified as an oncogene. Using differential static light scattering, we observed that the aggregation temperature T_{agg} for 0.5 mg/ml apo Bcl-2 was ~57°C. Four hundred nanomoles of the known ligand, Bax-BH3, significantly increased the T_{agg} to 67°C, whereas the presence of 1 μM of the scrambled LD4 peptide

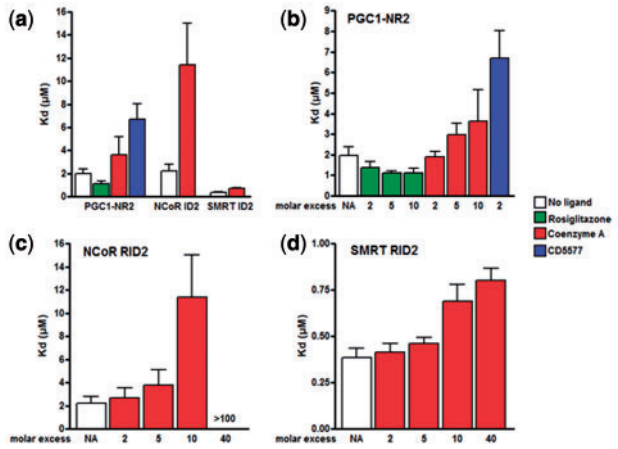

Fig. 4. FA on hPPARγ-LBD. Dissociation constants (K_d) measured from FA titrations between fluorescein-labeled PGC1-NR2, N-CORNR2 (NCoR RID2) or S-CORNR2 (SMRT RID2) peptides and hPPARγ-LBD in the absence of a ligand or in the presence of (**a**) a 10 M excess and (**b–d**) increasing molar excess of Rosiglitazone, CoA or CD5477, respectively.

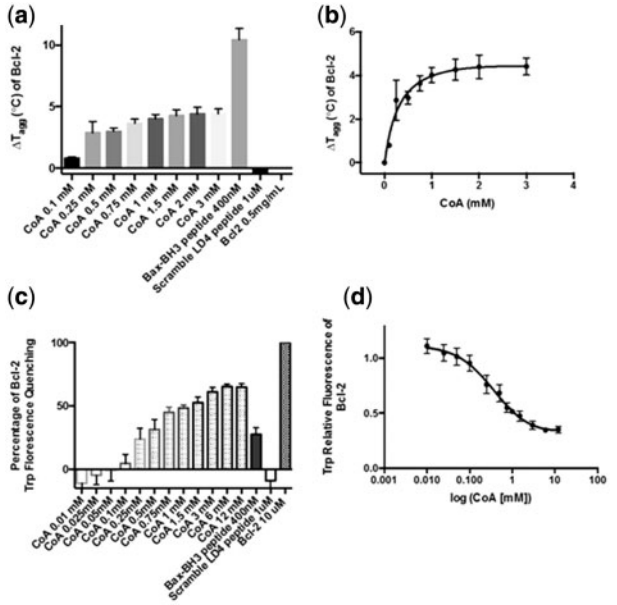

Fig. 5. (**a**) Change in the aggregation temperature ΔT_{agg} of Bcl-2 in the presence of the Bax-BH3 peptide (as a positive control), the scrambled LD4 peptide (as negative control) and CoA at various concentrations. (**b**) The change in ΔT_{agg} plotted against the concentration of CoA was used to determine an apparent K_d of 0.32 ± 0.13 mM using the single-binding-site model. (**c**) Comparison of tryptophan fluorescence quenching by the Bax-BH3 peptide, scrambled LD4 and various concentrations of CoA. CoA (0.25 mM) was as effective in quenching tryptophan fluorescence as 400 nM Bax-BH3 peptide. (**d**) Tryptophan relative fluorescence of Bcl-2 in the presence of increasing concentrations of CoA. Using a single-binding-site model the K_d was 0.38 ± 0.08 mM

(a negative control) did not alter T_{agg}. Increasing concentrations of CoA increased T_{agg} for Bcl-2 up to 62°C, suggesting a direct interaction (Figs 5a and b, Supplementary Table S7). By measuring the quenching of intrinsic tryptophan fluorescence of Bcl-2 in the presence of increasing CoA concentrations, we established the K_d of the CoA:Bcl-2 interaction to be 0.38 mM (Figs 5c and d), whereas the K_d of a the fluorescent-labeled Bax-BH3 peptide was 128 ± 21 nM (Supplementary Fig. 3a). According to iDTP, the CoA binding pocket is adjacent to the Bax-BH binding site with no notable overlap (Supplementary Fig. 3b). In agreement even 4.7 mM CoA did not reduce the FA of Bax-BH3 (at a concentration of 20 nM, 235 000 times less than CoA), supporting the prediction that the binding sites of CoA and Bax-BH do not overlap (Supplementary Fig. 4). Thus, our *in vitro* binding experiments strongly support our computational predictions.

4 Discussion

We have developed a computational method to extract implicit structural signatures of a drug binding site from an ensemble of structures of proteins to which this drug binds. We showed that such a PPE, can be built using as few as one structure of a drug–protein complex and a set of apo structures of other known drug-binding proteins. The PPE of a given drug is constructed using sequence order–independent alignments and a probabilistic scoring function, which allows weakly conserved but significant structural patterns of the interactions between the drug and its several target proteins to emerge and be quantified. Thus, our PPE is able to encode features related to promiscuous target interactions and structural flexibility of a drug. The validity of 11 PPEs was confirmed by illustrating that they reliably identify known targets of the respective drugs. We found that by combining a PPE with an aDDP as an orthogonal source of structure-independent information, the resulting method, iDTP, enables large-scale prediction of novel drug targets.

The challenge of identifying new drug–target pairs *in silico* has attracted significant interest from the computational community. However, compared with other algorithms, iDTP includes unprecedented features, because no previous studies have combined sequence order-independent alignment and probabilistic scoring function to model the drug–protein interaction, nor have they employed the aDDP to filter out false positive predictions. Most previous studies have not assessed the performance of their methodologies by exploiting known drug targets, as we did here to validate the success rate of PPE. Because other studies used considerably different datasets and their programs are not publicly available, a direct comparison among methodologies is unfortunately impossible. However, compared with iDTP, other existing methods including conventional docking or structure-based virtual screening, share one or more of the following limitations: (i) they are known to scale poorly with the size and complexity of drugs and drug binding sites (Diller and Li, 2003) and (ii) their algorithms do not appropriately account for the different conformations of both drug and binding site residues. Our method addresses these concerns by constructing a structural signature from a set of binding sites, instead of a single binding site, and by using a probabilistic sequence similarity function that allows accounting for the different conformations of drugs and binding site residues. (The improvement expected from this methodology is analogous to the improvement from a multiple sequence alignment compared to a pairwise alignment.) We also incorporated the aDDP to identify relevant new targets.

The predictive power of iDTP was supported by both computational cross-validation and text mining. Additionally, we validated two of our predicted interactions by *in vitro* experiments. First, we showed that CoA bound to hPPARγ-LBD with an apparent K_d of less than 500 μM, displaying characteristics of a neutral antagonist. CoA is a ubiquitous cofactor that can reach high concentrations in eukaryotes depending on cell type and subcellular localization (~0.14, 0.7 and 5 mM in animal cytosol, peroxysomes and mitochondria, respectively) (Leonardi *et al.*, 2005). It is therefore possible that this predicted interaction plays a currently unrecognized

biological role in fatty acid signaling and metabolism. iDTP predicted that the CoA binding site on hPPARγ-LBD is the receptor's ligand-binding pocket, which also binds rosiglitazone and CD5477. Indeed, the ligand-binding pocket of hPPARγ is one of the largest among the nuclear receptor protein family (Li *et al.*, 2003), allowing hPPARγ to bind a variety of ligands. Thus, CoA may trigger a conformational change that disrupts or unsettles the binding surface of both coactivators and corepressors, producing the characteristics of a neutral antagonist. However, we cannot strictly rule out that CoA binds to the surface where coactivators and corepressors would normally bind, creating competition for the binding site.

Second, we verified another CoA interaction predicted by iDTP by showing that CoA binds *in vitro* to recombinant Bcl-2 with a K_d of $\sim$350 μM. The predicted CoA binding pocket on Bcl-2 is adjacent to the known binding site for Bax-BH3. Because we showed that CoA binds Bcl-2 without displacing Bax-BH3, we can indeed infer non-competitive binding. The predicted binding pocket of CoA is interesting for drug design purposes, because it is located adjacent to the well-explored Bax-BH3 binding pocket (Ku *et al.*, 2011); therefore, it may provide an alternative target site with possible synergetic effects.

Beyond validating our computational predictions, our *in vitro* experiments also suggest the usefulness of iDTP for various applications: The case of the CoA-hPPARγ illustrates how iDTP might be used to reveal biologically relevant interactions between small molecules (ligands, cofactors or metabolites) and cellular proteins. Thus, our method could help establish metabolite-protein pairs for large-scale metabolic analyses or for predicting possible targets for chemical small-molecule pollutants such as bisphenols. The interaction between CoA and Bcl-2 illustrates how iDTP could be used for drug discovery by suggesting possible lead compounds and novel druggable protein binding pockets. In addition, iDTP could provide insight into the binding mechanisms of known drugs for which the drug–target complex has not yet been determined. For example, our results suggest that formic acid binds to CYP2E1 (Supplementary file PredTargets). CYP2E1 is an enzyme known to interact with more than 70 small drugs and xenobiotic compounds (Ogu and Maxa, 2000). Induction of CYP2E1 has been shown to cause oxidative stress and alcohol-induced liver injury in mouse models (McGehee *et al.*, 1994; Nanji *et al.*, 1994); however, Trolox[6-hydroxy,2,5,7,8-tetramethylchroman-2-carboxylic acid], a drug that contains the formic acid structure, has been shown to reduce the aforementioned toxicity (Wu and Cederbaum, 2000, 2002). Hence, our results suggest a direct interaction between the Trolox formic acid moiety and CYP2E1 that results in reduced toxicity.

To further evaluate the usefulness of iDTP for pharmaceutical purposes, we identified the genetic diseases associated with the predicted target proteins for each drug using the databases—Online Mendelian Inheritance in Man (Hamosh *et al.*, 2000) and Human Gene Mutation Database (Stenson *et al.*, 2014). A cocitation index with high statistical significance ($P < 0.005$) was found for 16 predicted drug–target pairs (including CoA-hPPARγ). The predicted drug targets were associated with major human diseases, such as cancer, heart problems and metabolic dysfunctions (Supplementary Table S4), making these results a potentially valuable basis for drug discovery and repositioning. However, the use of iDTP for drug repositioning, in the strict sense of re-using a FDA-approved chemical compound, remains currently limited, as a relatively large set of 3D structures of known targets is required to construct a high-confidence PPE. The rapid pace of experimental determination of protein structures will reduce this limitation in the future.

Acknowledgements

The authors acknowledge the technical support from the Molecular and Systems Computational Bioengineering Lab at UIC.

Funding

This research was supported by competitive research funding from King Abdullah University of Science and Technology, and Fondation ARC pour la recherche sur le cancer.

Conflict of Interest: none declared.

References

Alam,T. *et al.* (2014) How to find a leucine in a haystack? Structure, ligand recognition and regulation of Leucine-Aspartic acid (LD) motifs. *Biochem. J.*, **460**, 317–329.

Arrowsmith,J. (2011a) Trial watch: phase III and submission failures: 2007–2010. *Nat. Rev. Drug Discov.*, **10**, 87.

Arrowsmith,J. (2011b) Trial watch: Phase II failures: 2008–2010. *Nat. Rev. Drug Discov.*, **10**, 328–329.

Ashburn,T. and Thor,K. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.*, **3**, 673–683.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Cavallo,L. *et al.* (2003) POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res.*, **31**, 3364–3366.

Chang,R. *et al.* (2010) Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput. Biol.*, **6**, e1000938.

Chen,B. *et al.* (2009) Pubchem as a source of polypharmacology. *J. Chem. Inf. Model.*, **49**, 2044–2055.

Cui, X. *et al.* (2015) Finding optimal interaction interface alignments between biological complexes. *Bioinformatics*, **31**(12): i133–i141.

Diller,D. and Li,R. (2003) Kinases, homology models, and high throughput docking. *J. Med. Chem.*, **46**, 4638–4647.

Dundas,J. *et al.* (2006) CASTp: computed atlas of surface topography of proteins with structural and topographicalmapping of functionally annotated residues. *Nucleic Acids Res.*, **34**, W116–W118.

Dundas,J. *et al.* (2007) Topology independent protein structural alignment *BMC Bioinformatics*, **8**, 388.

Dundas,J. *et al.* (2011) Structural signatures of enzyme binding pockets from order-independent surface alignment: a study of metalloendopeptidase and NAD binding proteins. *J. Mol. Biol.*, **406**, 713–729.

Emig,D. *et al.* (2013) Drug target prediction and repositioning using an integrated network-based approach. *PLoS One*, **8**, e60618.

Engin,H. *et al.* (2012) A strategy based on protein-protein interface motifs may help in identifying drug off-targets. *J. Chem. Inf. Model.*, **52**, 2273–2286.

Gao,M. and Skolnick,J. (2013) A comprehensive survey of small-molecule binding pockets in proteins. *PLoS Comput. Biol.*, **9**, e1003302.

Hamosh,A. *et al.* (2000) Online mendelian inheritance in man (OMIM). *Hum. Mutat.*, **15**, 57–61.

Hu,G. and Agarwal,P. (2009) Human disease-drug network based on genomic expression profiles. *PLoS One*, **4**, e6536.

Huang,B.D. and Schroeder,M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.*, **6**, 19.

Iorio,F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl Acad. Sci. USA*, **107**, 14621–14626.

Jansen,R. *et al.* (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, **12**, 37–46.

Keiser,M. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature*, **462**, 175–181.

Kinnings,S. *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.

Knox,C. *et al*. (2011) Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.

Ku,B. *et al*. (2011) Evidence that inhibition of BAX activation by BCL-2 involves its tight and preferential interaction with the BH3 domain of BAX. *Cell Res.*, **21**, 627–641.

Lamb,J. *et al*. (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Laskowski,R.A. (1995) SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330.

LeMaire,A. *et al*. (2009) Activation of RXR-PPAR heterodimers by organotin environmental endocrine disruptors. *EMBO Rep.*, **10**, 367–373.

Leonardi,R. *et al*. (2005) Coenzyme A: back in action. *Progr. Lipid Res.*, **44**, 125–153.

Li,Y. *et al*. (2003) Activation of nuclear receptors: a perspective from structural genomics. *Structure*, **11**, 741–746.

Li,Y.Y. *et al*. (2011) A computational approach to finding novel targets for existing drugs. *PLoS Comput. Biol.*, **7**, e1002139.

Liang, J. *et al*. (1998) Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Prot. Sci.*, **7**, 1884–1897.

Liebler,D. and Guengerich,F. (2005) Elucidating mechanisms of drug-induced toxicity. *Nat. Rev. Drug Discov.*, **4**, 410–420.

Lounkine,E. *et al*. (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature*, **486**, 361–367.

McGehee,R. Jr., *et al*. (1994) Characterization of cytochrome p450 2e1 induction in a rat hepatoma FGC-4 cell model by ethanol. *Biochem. Pharmacol.*, **48**, 1823–1833.

Mestres,J. *et al*. (2009) The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst.*, **5**, 1051–1057.

Nanji,A. *et al*. (1994) Markedly enhanced cytochrome p450 2e1 induction and lipid peroxidation is associated with severe liver injury in fish oil-ethanol-fed rats. *Alcohol Clin. Exp. Res.*, **18**, 1280–1285.

Napolitano,F. *et al*. (2013) Drug repositioning: a machine-learning approach through data integration. *J. Cheminform.*, **5**, 30.

Noeske,T. *et al*. (2006) Predicting compound selectivity by self-organizing maps: cross-activities of metabotropic glutamate receptor antagonists. *ChemMedChem*, **1**, 1066–1068.

Ogu,C. and Maxa,J. (2000) Drug interactions due to cytochrome p450. *Proc. Bayl. Univ. Med. Cent.*, **13**, 421–423.

Pérot,S. *et al*. (2010) Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Disc. Today*, **15**, 656–667.

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

Peters,K. *et al*. (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.*, **256**, 201–213.

Qiao,N. *et al*. (2013) Cociter: an efficient tool to infer gene function by assessing the significance of literature co-citation. *PLoS One*, **8**: e74074.

Qu,X. *et al*. (2009) Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. *BMC Bioinformatics*, **10** (Suppl. 5): S4.

Reddy,H. and Zhang,S. (2013) Polypharmacology: drug discovery for the future. *Expert Rev. Clin. Pharmacol.*, **6**, 41–47.

Sanseau,P. *et al*. (2012) Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.*, **30**, 317–320.

Stenson,P. *et al*. (2014) The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9.

Swedenborg,E. *et al*. (2009) Endocrine disruptive chemicals: mechanisms of action and involvement in metabolic disorders. *J. Mol. Endocrinol.*, **43**, 1–10.

Su,A. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl. Acad. Sci. USA*, **101**, 6062–6067.

Suthram,S. *et al*. (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput. Biol.*, **6**, e1000662.

Tseng,Y. and Liang,J. (2006) Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach. *Mol. Biol. Evol.*, **23**, 421–436.

Wang,G. and Dunbrack,R.Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wang,Z. and Zhang,H. (2013) Rational drug repositioning by medical genetics. *Nat. Biotechnol.*, **31**, 1080–1082.

Watson,J. *et al*. (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.*, **15**, 275–284.

Wei,G. *et al*. (2006) Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell*, **10**, 331–342.

Wu,C. *et al*. (2006) The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–91.

Wu,D. and Cederbaum,A. (2000) Ethanol and arachidonic acid produce toxicity in hepatocytes from pyrazole-treated rats with high levels of CYP2E1. *Mol. Cell Biochem.*, **204**, 157–167.

Wu,D. and Cederbaum,A. (2002) Cyclosporine a protects against arachidonic acid toxicity in rat hepatocytes: role of CYP2E1 and mitochondria. *Hepatology*, **35**, 1420–1430.