Bioinformatics, 31(16), 2015, 2721–2727
doi: 10.1093/bioinformatics/btv214
Advance Access Publication Date: 21 April 2015
Original Paper

OXFORD

Data and text mining

Protein homology reveals new targets for bioactive small molecules

David Gfeller[1,2,*] and Vincent Zoete[2,*]

[1]Department of Fundamental Oncology, Ludwig Center for Cancer Research, University of Lausanne, 1066 Epalinges, Switzerland and [2]Swiss Institute of Bioinformatics (SIB), Quartier Sorge, Bâtiment Génopode, 1015 Lausanne, Switzerland

*To whom correspondence should be addressed.
Associate Editor: Burkhard Rost

Received on February 12, 2015; revised on March 29, 2015; accepted on April 14, 2015

Abstract

Motivation: The functional impact of small molecules is increasingly being assessed in different eukaryotic species through large-scale phenotypic screening initiatives. Identifying the targets of these molecules is crucial to mechanistically understand their function and uncover new therapeutically relevant modes of action. However, despite extensive work carried out in model organisms and human, it is still unclear to what extent one can use information obtained in one species to make predictions in other species.

Results: Here, for the first time, we explore and validate at a large scale the use of protein homology relationships to predict the targets of small molecules across different species. Our results show that exploiting target homology can significantly improve the predictions, especially for molecules experimentally tested in other species. Interestingly, when considering separately orthology and paralogy relationships, we observe that mapping small molecule interactions among orthologs improves prediction accuracy, while including paralogs does not improve and even sometimes worsens the prediction accuracy. Overall, our results provide a novel approach to integrate chemical screening results across multiple species and highlight the promises and remaining challenges of using protein homology for small molecule target identification.

Availability and implementation: Homology-based predictions can be tested on our website http://www.swisstargetprediction.ch.

Contact: david.gfeller@unil.ch or vincent.zoete@isb-sib.ch.

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

1 Introduction

Small molecules provide remarkable tools to modulate molecular mechanisms in cells, with therapeutic applications in human and veterinary, biotechnological or agricultural applications in many other species from animals, to plants, to microorganisms. Small molecules displaying bioactivity in a given organism typically bind specific proteins and modify their activity. Mapping the actual protein targets of small molecules is therefore a key step toward a better understanding of their mechanisms of action. In addition, several molecules bind to more than one target, which can be unrelated in terms of both sequence and function (Karaman *et al.*, 2008; Mestres *et al.*, 2009). These secondary targets are typically responsible for many favorable or unfavorable side effects of known drugs. Unraveling small molecule secondary targets is helpful to predict and elucidate these side effects (Lounkine *et al.*, 2012; Young *et al.*, 2008). Moreover, it is a promising approach to repurpose existing drugs toward new applications by exploiting their activity on other proteins than those they were initially developed for (Keiser *et al.*, 2009).

Several experimental and computational strategies have been developed to determine the interacting partners and the activity of small molecules (Ziegler *et al.*, 2013). Small molecules can be

screened *in vitro* against large arrays of proteins such as kinases or G protein-coupled receptors (Davis *et al.*, 2011; Karaman *et al.*, 2008). In parallel, *in vivo* screening approaches are increasingly being developed using model organisms such as yeast (Giaever *et al.*, 2004) or zebrafish (Zon and Peterson, 2005). Several of these chemogenomics screens use the power of genetics to provide indirect information about the actual targets of small molecules, for instance by comparing the activity of small molecules across different mutant strains in yeast (Lee *et al.*, 2014). From the computational point of view, most target prediction approaches use similarity relationships between a new molecule and known ligands (Dunkel *et al.*, 2008; Gfeller *et al.*, 2014; Keiser *et al.*, 2007; Liu *et al.*, 2013). There top predicted targets are identified as those with one or more ligands that display high similarity with the query molecule. Different similarity measures between molecules can be used for this purpose (Armstrong *et al.*, 2011; Ballester and Richards, 2007; Gfeller *et al.*, 2013; Rahman *et al.*, 2009; Willett, 2011). Other studies have explored the use of additional information, such as side-effect similarity (Campillos *et al.*, 2008) or gene expression profile similarity (Iorio *et al.*, 2010) to expand the similarity beyond features determined solely by the molecular structure of the compounds. Other groups have also used protein structures to predict small molecule targets and potential binding modes (Gao *et al.*, 2008; Schomburg *et al.*, 2014; Wang *et al.*, 2012a).

Computational target prediction tools are especially useful to analyze the results of high-throughput phenotypic assays that are increasingly being used to identify bioactive compounds in different species but do not provide direct information about their targets (Clemons, 2004; Inglese *et al.*, 2007). Data from thousands of such assays are available in public repositories such as ChemBank (Seiler *et al.*, 2008) or PubChem (Wang *et al.*, 2012b). *In silico* target predictions have been successfully applied to the results of phenotypic screening assays performed in diverse systems ranging from cell lines (Young *et al.*, 2008) to zebrafish (Laggner *et al.*, 2012).

Unfortunately, most current experimental and computational approaches to determine small molecule targets focus on one species such as human, mouse or rat. As such, information about protein orthology relationships to make predictions in less-studied species or improve predictions in model organisms by integrating data from other species has been mostly disregarded. For instance, only a handful of recent studies investigated the properties of ligands binding to related proteins in distinct organisms (Klabunde, 2007; Krüger and Overington, 2012; Paricharak *et al.*, 2013) or within the same organism (Schuffenhauer *et al.*, 2003). This is in stark contrast with many areas of biology and bioinformatics where the ability to transfer results obtained in one organism to others is a central dogma. For instance, mapping protein function based on orthology has proved extremely useful (Loewenstein *et al.*, 2009). Protein structure predictions rely heavily on the existence of homologous proteins with available crystal structures (Kiefer *et al.*, 2009). Protein–protein interaction predictions also strongly benefit from information obtained in orthologous species (Matthews *et al.*, 2001). All these studies have clearly established the use of homology relationships to better predict the properties of proteins.

This lack in conceptual understanding and computational techniques to exploit target homology in small molecule–protein interaction predictions is strongly limiting the scope of chemoinformatics approaches for two main reasons. First, beyond model organisms such as human or rat, very few data are available in public databases for other organisms (e.g. zebrafish and bacteria) and accurate predictions would be very useful in these species. Second, it would be highly desirable to integrate data obtained in close orthologous species to improve existing techniques predicting small molecule–protein interactions, even in well-studied organisms. Understanding how small molecule targets can be mapped across species is also critically important for therapeutic applications since small molecules of therapeutic interest are first tested in model organisms (e.g. mouse or rat) before being considered for clinical trials in human.

Here, we introduce a new strategy to integrate data from different species to improve small molecule-target predictions. Our results reveal that protein orthology leads to improve prediction accuracy and is powerful to uncover new small molecule–protein interactions, especially in species with less experimental data. Interestingly, paralogy relationships do not appear to improve prediction accuracy. Finally, our findings provide a strong basis for expanding small molecule–protein target predictions beyond well-studied organisms.

2 Methods

2.1 Small molecule activity data

Small molecule interactions with protein targets were retrieved from the ChEMBL database (Bento *et al.*, 2014), including only molecules with more than 5 and less than 60 heavy atoms. Interactions are selected as activity relationships between a small molecule and a single protein annotated as binding ('B') in ChEMBL with activity (e.g. K_i, K_d, K_m, IC_{50} or EC_{50}) lower than 10 μM in all assays. In each species, our dataset consists of small molecules with reported activity in ChEMBL18 but absent from ChEMBL16, which is the reference database behind the SwissTargetPrediction method to predict the targets of small molecules. The list of interactions is available at http://www.swisstargetprediction.ch/download.php. Molecular scaffolds were determined using the OPREA definition (Pollock *et al.*, 2008).

2.2 Homology relationships

Four species were considered in this work: human, rat, mouse and cow. The selection was made based on the amount of data present in ChEMBL, requiring especially to have at least 100 different targets with reported ligands to meaningfully evaluate prediction accuracy. Orthology and paralogy relationships were retrieved from Ensembl Compara (Vilella *et al.*, 2009), Treefam (Schreiber *et al.*, 2014) and orthoDB (Waterhouse *et al.*, 2013) databases, considering the union of all three databases and allowing both one-to-one and one-to-many relationships. Including homology relationships results in a much higher number of potential targets (Fig. 2B), especially in rat, mouse and cow. This is because many human targets have orthologs that have no reported ligand in those species. Therefore, considering target homology between these species and human significantly increases the number of potential targets. Homology relationships are available at http://www.swisstargetprediction.ch/download.php.

2.3 Target predictions

Target prediction was carried out by comparing new molecules present only in the recent release of ChEMBL (ChEMBL18) with known ligands in the SwissTargetPrediction database (referred to as 'reference dataset'), which is derived from ChEMBL16 (Gfeller *et al.*, 2014). The method is similar to the one described in our previous work (Gfeller *et al.*, 2013, 2014). In this approach, two kinds of similarity values based on chemical similarity [FP2 Fingerprints (Willett, 2011)] and shape similarity [Electroshape (Armstrong

et al., 2011)] are combined using logistic regression. More precisely, the similarity between each ligand of a target and the query molecule is computed for both similarity measures. The most similar ligand for each kind of similarity measure is used to compute the final score between a ligand and a target. To combine the two similarity values, the logistic regression model that had been trained with the reference dataset derived from ChEMBL16 (Gfeller *et al.*, 2013) has been used here. For this training procedure, the total number of interacting ligand-target pairs was 347'889, and the number of negative data (non-interacting pairs) was five times larger than positive data [see Gfeller *et al.* (2013) for other details]. The final logistic regression scores between a query molecule and their potential targets range between 0 for a mismatch and 1 for a perfect match and can be used to rank the targets and identify the top predicted ones (see below).

Predictions using target homology where implemented as follows (see example in Fig. 1). The query molecule is compared to all ligands in each species. Similarity values with ligands targeting homologous proteins are then mapped by homology to proteins in the species were predictions are made. When testing the influence of orthology relationships, especially interesting molecules are those present in our reference dataset in one of the other organisms considered in this work (see example in Fig. 1C). Such molecules are expected to benefit most from orthology-based predictions, since their targets can be directly mapped across species. Prediction accuracy was therefore computed separately for this subset of molecules (Figs 3B, 4B and 4D).

Including target orthology results in significantly more targets that can be predicted, many of which have not been tested (Table 1). For this reason, we also consider predictions that are restricted to targets that are present in our reference set built from ChEMBL version 16.

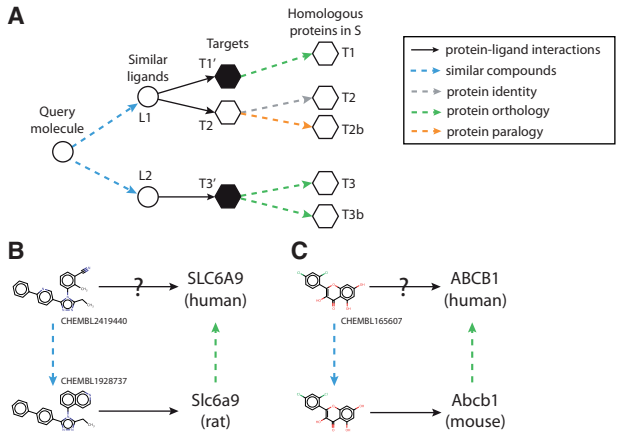

Fig. 1. (**A**) Schematic description of the homology-based prediction method in species S. The most similar ligands (L1 and L2) to a query molecule are first determined (cyan dashed arrows), the targets of these ligands are identified in different species (black arrows) and their homologs in S are determined (green and orange arrows) (see Section 2.3). Here, targets in species S where predictions are carried out are shown in white (T2) and targets in other species are shown in dark (T1′ and T3′). Orthologs (green dashed lines) and paralogs (orange dashed lines) are mapped considering both one-to-one and one-to-many homology relationships. Target prediction without homology results in T2; with orthology in T1, T2, T3, T3b; with paralogy in T2, T2b; with orthology and paralogy in T1, T2, T2b, T3, T3b. (**B**) Example of a successful orthology-based prediction in human using similar ligands of rat proteins (Sugane *et al.*, 2013). (**C**) Example of a successful orthology-based prediction in human when the query molecule itself is a known ligand of a mouse protein in our reference dataset (Boumendjel *et al.*, 2002)

2.4 Performance evaluation

We used two metrics to evaluate the accuracy of the predictions. First, the area under the ROC curve (AUC) was calculated for each molecule in our datasets (i.e. ligands tested in ChEMBL18 but not in ChEMBL16). Negative data were retrieved as interactions with activity higher than 100 μM in all assays. These data were supplemented with randomly selected targets to have five times more negative than positive data for each molecule in our dataset (Gfeller *et al.*, 2013). AUC values were then averaged over the whole dataset in each species. AUC values provide an unbiased global estimate of the prediction accuracy over the entire range of potential targets. However, for practical applications, one is often interested in top-ranking predictions that can typically be tested experimentally. We therefore also used the fraction of ligands in each species with at least one reported target among the top 15 predicted proteins. It is important to note that the number of potential targets is always much larger than 15. Therefore, the probability to obtain a correct target simply by chance is much lower than the numbers reported in Figure 4. However, as noted in the main text, including homology relationships significantly increases the number of targets, which can contribute to the lower values obtained in Figure 4A compared with Figure 4C.

3 Results

The workflow of our method is illustrated on Figure 1A. As in other ligand-based approaches, it relies on the observation that similar bioactive compounds tend to have similar targets. In this framework, the predicted targets for a query molecule are those interacting with ligands displaying the highest similarity with this molecule. To integrate target homology, a small molecule with reported bioactivity in a given species S is compared with all known ligands of proteins from other organisms that have at least one homolog in S. Predicted targets are determined as those with homologous proteins binding to the most similar ligands in other species. In the absence of small molecule–protein interaction data in S, predictions rely only on target orthology. Alternatively, if some ligands are reported in S, ligands of homologous proteins are added to the list of similar molecules. When considering paralogy relationships, potential predicted targets are proteins from S either with known ligands or having a paralog in S with known ligands similar to the query molecule (see example in Fig. 1A). An example of a successful prediction in human using similarity with a ligand binding to a mouse orthologous protein is shown in Figure 1B.

A special situation arises when a query molecule has been already tested in another organism (Fig. 1C). In this case, the most

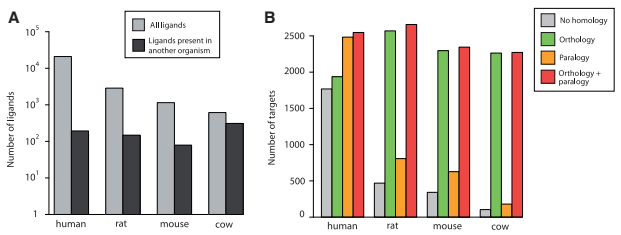

Fig. 2. Distribution of the number of molecules and targets. (**A**) Gray bars show the total number of ligands for each organism in our dataset (i.e. in ChEMBL18 but not in ChEMBL16). Black bars show the number of ligands tested in at least one of the three other organisms in the reference dataset. (**B**) Number of targets available for predictions when considering different target homology relationships. Including target orthology relationships significantly increases the number of potential targets in rat, mouse and cow

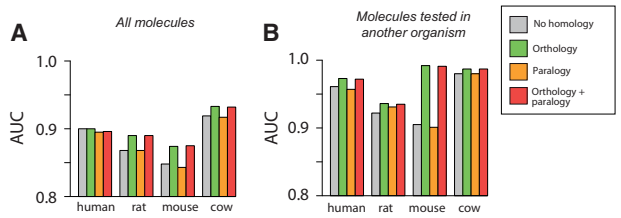

Fig. 3. Average AUC values obtained in our dataset. (**A**) Results with all molecules from our dataset. (**B**) Results obtained with the subset of small molecules in our dataset that have been tested in at least one other organism

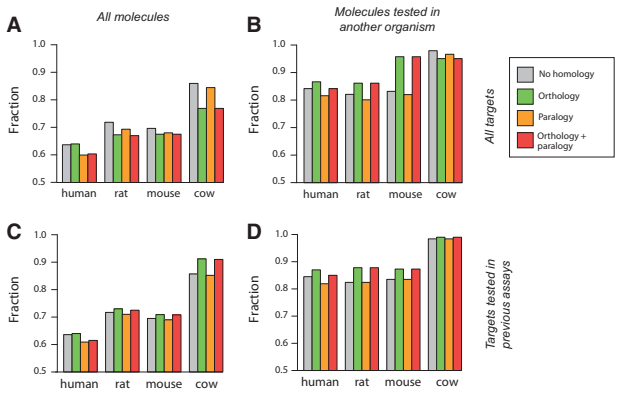

Fig. 4. Fraction of molecules in our dataset for which at least one known target falls in the top 15 predicted ones. (**A**) Results with all molecules from our dataset. (**B**) Results obtained with the subset of small molecules in our dataset that have been tested in at least one other organism. (**C** and **D**) Same as (A) and (B) but considering only targets tested in our reference dataset

similar ligand will be the molecule itself and the top predicted targets will be the orthologs of the targets in the species where the molecule has been tested.

To explore the use of homology-based target predictions, we used the SwissTargetPrediction method to assess similarity between molecules and predict their targets (Gfeller *et al.*, 2013, 2014). This algorithm was built using the ChEMBL database version 16 as a reference dataset of small molecule–protein interactions (Bento *et al.*, 2014) and includes data from human, mouse, rat, cow and horse. Since horse has only three targets with reported ligands, it was not included in this work. Thus, homology-based predictions were tested considering four different organisms: human, rat, mouse and cow. To test the use of homology relationships, we used for each organism separately all compounds in ChEMBL18 that do not appear in our reference dataset derived from ChEMBL16 (see Section 2 for more details). The total number of small molecules in this dataset is displayed in Figure 2A (gray bars). For each species, the number of molecules that were tested in ChEMBL16 in one of the three other organisms is shown in black bars in Figure 2A. Apart from cow, most compounds of our dataset are not found in ChEMBL16 in any of the four organisms.

Predictions were carried out as described in Figure 1 in each organism, considering no homology, orthology, paralogy and both orthology and paralogy relationships. It is important to note that including homology, and especially orthology relationships increases significantly the number of potential targets available for prediction for rat, mouse and cow, as observed in Figure 2B (green bars). This is because the number of targets on which ligands have been tested in these species is much smaller than in human. Therefore, including orthology-based predictions with orthologous human targets expands significantly the number of potential predictable targets. In human, the number of proteins without ligands in our reference dataset but with orthologs in mouse, rat or cow that have known ligands is much smaller. Therefore, including orthology relationships does not increase much the total number of targets available for predictions in human. Paralogy relationships provide on average a 30–40% increase in the number of potential predictable targets (Fig. 2B, orange bars).

For each molecule in our dataset, the performance was assessed using AUC, as well as the fraction of ligands with at least one correct target among the top 15 predictions (see Section 2). In terms of AUC values, Figure 3A indicates that including orthology relationships between targets leads on average to better AUC, especially for rat, mouse and cow. Interestingly, we observe that the use of paralogy relationships in general does not result in better AUC values and most often give rise to lower values. Considering both target orthology and paralogy gives similar performance as when only target orthology is used.

As noted in previous studies, small molecule–protein interaction data display strong biases and redundancies in part because of the systematic exploration of sidechain replacements that preserve the compounds' scaffold in medicinal chemistry experiments. This makes target prediction much easier for molecules belonging to chemical series intensively explored. As a consequence, the AUC value calculated from the entire dataset does not reflect the predictive ability of the approach for molecules belonging to new chemical series (Gfeller *et al.*, 2013; Rohrer and Baumann, 2008). One way of addressing this issue consists in preventing comparisons between ligands with the same scaffold (see Section 2) (Gfeller *et al.*, 2013). As expected, AUC values are lower when preventing comparison between the same scaffolds, but the same trend is visible (Supplementary Fig. S1A).

We further investigate the situation where molecules in our dataset have been previously tested in another species (Fig. 1C). This corresponds to the case where target orthology relationships are expected to be most beneficial. We observe that AUC values are higher and, as before, target orthology leads to improved AUC values, while paralogy relationships result in few changes (Fig. 3B). The same results hold when preventing comparisons between ligands with the same scaffolds (Supplementary Fig. S1B).

AUC values provide an unbiased measure of prediction accuracy. However, for practical purposes, only the top predicted targets are typically tested experimentally. We therefore consider a second measure of performance defined as the fraction of ligands for which at least one of the top 15 predicted targets is a known one (i.e. true positive) in our dataset. This can be seen as the probability of having at least one hit when experimentally testing top predicted targets, which is most useful to guide experimental approaches. In the general case, although orthology relationships gave better AUC values, we observe that the likelihood of obtaining a correct hit among the top predicted targets decreases for rat, mouse and cow and is only very slightly higher in human when considering orthology-based predictions (Fig. 4A). Target paralogy also results in lower performance. However, if we only consider molecules that have been tested in other species (Fig. 4B), the fraction of query molecules with at least one known target in the top 15 is higher for human, rat and mouse when using protein orthology relationships. This result confirms that target orthology relationships are especially useful to map the targets of molecules that have been directly tested in another organisms.

The apparent discrepancy between AUC values and the fraction of compounds with at least one known target among the top

Table 1. Number of targets in different organisms

Organisms	Targets	No homology	Total	Orthology	Total
Human	1132	1028	1768	1032	1937
Rat	179	158	469	168	2569
Mouse	127	105	342	119	2296
Cow	49	46	104	47	2263

Column 2 shows the number of different targets in our dataset. Column 3 shows how many of them are available in our predictions without including homology relationships. Column 4 shows how many in total are available for predictions (Fig. 2B, gray bars). Column 5 shows how many of the targets in our dataset are available in our predictions when including target orthology relationships. Column 6 shows how many in total are available for predictions when considering target orthology relationships (Fig. 2B, green bars)

predicted ones can be understood by observing that the number of potential predictable targets is much larger when orthology-based mapping is allowed (Fig. 2B and Table 1, last column), suggesting that for many new targets that can be predicted by homology, no ligand has simply been tested experimentally. Moreover, most of the targets in our dataset are already present in previous assays even without considering target homology (Table 1, column 3). To explore the effect of this potential bias, we restricted the predictions to targets that are present in our reference dataset. Remarkably, in this case, we see a constant improvement when including orthology relationships for all four organisms (Fig. 4C and D, green versus gray bars). The same results hold when considering the sensitivity (see Supplementary Fig. S2). This suggests that the lower performance of orthology-based predictions observed in Figure 4A also comes from the much larger number of targets that are available for predictions when considering orthology relationships. Because for many of these 'new' targets no ligand has been tested, a larger number of data annotated as negative are present in our dataset, which is known to affect measures considering the top predictions, such as the one used in Figure 4. Overall, these results suggest that orthology-based predictions work especially well for compounds tested in some other organism. In our benchmarks, we further observe that ligands of 'old' targets can be predicted using target orthology. For the other predicted protein targets, it is important to realize that many of them have not been considered in experimental assays. Therefore, some of the orthology-based predictions may be correct but do not appear as true positives in our dataset because of the lack of experimental data for these targets.

To explore the potential of the proposed framework in other species, we display in Figure 5 the number of targets available for predictions in a wide range of organisms. As expected, this number is lower for invertebrates, since there exist less orthologs in these species. Nevertheless, a significant number of proteins can still be accessed by orthology in these distant species. We also note that the higher number of targets in zebrafish compared with other vertebrates likely arises because of the whole genome duplication event in teleosts (Howe *et al.*, 2013; Taylor *et al.*, 2001). Most importantly, Figure 5 indicates that orthology-based predictions have the potential to significantly impact small molecule-target discovery well beyond the four species considered in this work.

4 Discussion

Mapping the properties of proteins across different species using orthology is a well-established way of harnessing the wealth of data available in some organisms to make new and accurate predictions in

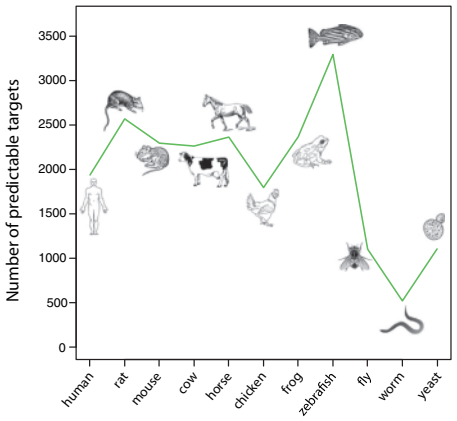

Fig. 5. Number of proteins that can be predicted as small molecule targets in different species using orthology relationships from known human, rat, mouse and cow protein targets

others. However, this property has not been exploited to predict interactions between proteins and small molecule ligands. Here, we observe that target prediction accuracy is improved, both in terms of AUC and of the fraction of molecules with at least one correct target among the top predicted ones, when including data from other organisms in all four species analyzed in this work. This is especially striking when the query molecule has been already tested in some other species, which corresponds to the most straightforward way of mapping small molecule–protein interactions across species (Fig. 1C).

The much larger number of potential targets that can be predicted using orthology (Fig. 2B and Table 1) can sometimes result in lower number of correct predictions among the top 15 predicted targets (Fig. 4A). However, when including only targets on which ligands have been tested in previous assays (here in version 16 of ChEMBL), we observe that target orthology relationships lead to a clear improvement, especially in rat, mouse and cow (Fig. 4C and D). This is likely because targets used in older assays are often reused in new ones, as observed in Table 1. First many of these 'old' targets have been selected based on their therapeutic or biotechnological interest and therefore are more likely to be studied. Second, when planning new experiments, researchers are typically guided by previous successful studies and therefore tend to preferentially study targets with already known ligands. Last, and possibly most importantly, 'old' targets often have well-established experimental assays, possibly commercially available, that are more likely to be reused by experimentalists, rather than setting up a whole new assay for new targets. For these reasons, it is not surprising that orthology-based predictions resulting in many more potential targets give lower performance in terms of the fraction of ligands with validated targets among the top predictions (Fig. 4A and B). However, when correcting for the bias toward reusing the same targets in experimental studies, we see a clear improvement in the predictions (Fig. 4C and D). Based on the improved accuracy when considering only 'old' targets, it is tempting to speculate that many of the orthology-based predictions for new targets, especially in rat, mouse and cow, may actually be correct, but simply have not been tested.

We observed that mapping small molecule targets between paralogs is often detrimental to the predictions. This likely reflects several factors. First, evolutionary studies indicate that paralogs have diverged more than orthologs. As a consequence, protein function is often more conserved among orthologs than paralogs (Altenhoff *et al.*, 2012). In terms of protein–ligand interactions, previous work also suggested that orthologous proteins share more of their ligands

compared with paralogs (Krüger and Overington, 2012). Second, small molecules are often designed to target only some specific members of a protein family (e.g. specific kinase inhibitors). Therefore, mapping their interactions within a family of paralogs may often lead to false positives. Although some of these results might have been guessed, our work demonstrates for the first time that paralogy relationships are in general not appropriate for transferring small molecule–target interactions within an organism.

To validate and compare predictions with and without orthology relationships, we used AUC and the fraction of ligands with at least one correct target among the top 15 predictions. This strategy does not require fixing an arbitrary threshold on the target scores, which may not be the same for all ligands, and takes a more pragmatic approach (correct prediction among the top predicted targets), which corresponds to what can be reasonably tested experimentally. Other studies have used measures such as sensitivity and specificity assuming a fixed and uniform threshold to separate interacting from non-interacting pairs (Liu *et al.*, 2013).

A natural question is whether orthology-based predictions will be as successful in more distant species, such as those displayed in Figure 5. Unfortunately, much less data are available in most species apart from human, rat, mouse and cow. For instance, horse, zebrafish or worm have less than five targets each with reported binding data in ChEMBL. When attempting to check the predictions in these species with little interaction data, we could not find significant validations of the predictions. However, because of the sparsity of data, we expect that this observation may not truly reflect the reality. In particular, orthology-based target predictions are likely to be useful in species like zebrafish, as attested by many biological studies that map small molecule activity across vertebrate species (Gebruers *et al.*, 2013; Ridges *et al.*, 2012).

Our work is the first attempt to establish the use of protein homology for bioactive small molecule target predictions at a large scale. Overall, we observe that clear improvement can be achieved by using target orthology, whereas paralogy relationships do not result in significantly better predictions. In this work, we used our previously published method (Gfeller *et al.*, 2013) to determine ligand similarity and ligand–target interaction scores. However, the idea of integrating orthology relationships into target predictions can be easily generalized to other small molecule similarity metrics. Our analysis also reveals a strong bias in recent small molecule–protein interaction datasets where the same targets are used over and over. This aspect should be considered when benchmarking cross-species target prediction approaches. Although there are many practical reasons for this bias, this work could serve as a guide to develop new assays to test potentially interesting targets in different species. These would increase our sampling of the ligand space for many proteins, which could then further improve the *in silico* predictions from model organisms to humans. Homology-based predictions can be tested at our website: http://www.swisstargetprediction.ch.

Acknowledgements

The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. We thank Olivier Michielin for insightful discussions about the manuscript.

Funding

This work was supported by the Swiss Institute of Bioinformatics and the Solidar-Immun Foundation.

Conflict of Interest: none declared.

References

Altenhoff,A.M. *et al.* (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.*, **8**, e1002514.

Armstrong,M.S. *et al.* (2011) Improving the accuracy of ultrafast ligand-based screening: incorporating lipophilicity into ElectroShape as an extra dimension. *J. Comput. Aided Mol. Des.*, **25**, 785–790.

Ballester,P.J. and Richards,W.G. (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.*, **28**, 1711–1723.

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

Boumendjel,A. *et al.* (2002) Recent advances in the discovery of flavonoids and analogs with high-affinity binding to P-glycoprotein responsible for cancer cell multidrug resistance. *Med. Res. Rev.*, **22**, 512–529.

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

Clemons,P.A. (2004) Complex phenotypic assays in high-throughput screening. *Curr. Opin. Chem. Biol.*, **8**, 334–338.

Davis,M.I. *et al.* (2011) Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.*, **29**, 1046–1051.

Dunkel,M. *et al.* (2008) SuperPred: drug classification and target prediction. *Nucleic Acids Res.*, **36**, W55–W59.

Gao,Z. *et al.* (2008) PDTD: a web-accessible protein database for drug target identification. *BMC Bioinformatics*, **9**, 104.

Gebruers,E. *et al.* (2013) A phenotypic screen in zebrafish identifies a novel small-molecule inducer of ectopic tail formation suggestive of alterations in non-canonical Wnt/PCP signaling. *PLoS One*, **8**, e83293.

Gfeller,D. *et al.* (2013) Shaping the interaction landscape of bioactive molecules. *Bioinformatics*, **29**, 3073–3079.

Gfeller,D. *et al.* (2014) SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res.*, **42**, W32–W38.

Giaever,G. *et al.* (2004) Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proc. Natl. Acad. Sci. USA*, **101**, 793–798.

Howe,K. *et al.* (2013) The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, **496**, 498–503.

Inglese,J. *et al.* (2007) High-throughput screening assays for the identification of chemical probes. *Nat. Chem. Biol.*, **3**, 466–479.

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.

Karaman,M.W. *et al.* (2008) A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.*, **26**, 127–132.

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

Keiser,M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.

Kiefer,F. *et al.* (2009) The SWISS-MODEL repository and associated resources. *Nucleic Acids Res.*, **37**, D387–D392.

Klabunde,T. (2007) Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.*, **152**, 5–7.

Krüger,F.A. and Overington,J.P. (2012) Global analysis of small molecule binding to related protein targets. *PLoS Comput. Biol.*, **8**, e1002333.

Laggner,C. *et al.* (2012) Chemical informatics and target identification in a zebrafish phenotypic screen. *Nat. Chem. Biol.*, **8**, 144–146.

Lee,A.Y. *et al.* (2014) Mapping the cellular response to small molecules using chemogenomic fitness signatures. *Science*, **344**, 208–211.

Liu,X. *et al.* (2013) HitPick: a web server for hit identification and target prediction of chemical screenings. *Bioinformatics*, **29**, 1910–1912.

Loewenstein,Y. *et al.* (2009) Protein function annotation by homology-based inference. *Genome Biol.*, **10**, 207.

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

Matthews,L.R. *et al.* (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res.*, **11**, 2120–2126.

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.

Paricharak,S. *et al.* (2013) Are phylogenetic trees suitable for chemogenomics analyses of bioactivity data sets: the importance of shared active compounds and choosing a suitable data embedding method, as exemplified on kinases. *J. Cheminform.*, **5**, 49.

Pollock,S.N. *et al.* (2008) Scaffold topologies. 1. Exhaustive enumeration up to eight rings. *J. Chem. Inf. Model.*, **48**, 1304–1310.

Rahman,S.A. *et al.* (2009) Small molecule subgraph detector (SMSD) toolkit. *J. Cheminform.*, **1**, 12.

Ridges,S. *et al.* (2012) Zebrafish screen identifies novel compound with selective toxicity against leukemia. *Blood*, **119**, 5621–5631.

Rohrer,S.G. and Baumann,K. (2008) Impact of benchmark data set topology on the validation of virtual screening methods: exploration and quantification by spatial statistics. *J. Chem. Inf. Model.*, **48**, 704–718.

Schomburg,K.T. *et al.* (2014) Facing the challenges of structure-based target prediction by inverse virtual screening. *J. Chem. Inf. Model.*, **54**, 1676–1686.

Schreiber,F. *et al.* (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res.*, **42**, D922–D925.

Schuffenhauer,A. *et al.* (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.*, **43**, 391–405.

Seiler,K.P. *et al.* (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–D359.

Sugane,T. *et al.* (2013) Atropisomeric 4-phenyl-4H-1,2,4-triazoles as selective glycine transporter 1 inhibitors. *J. Med. Chem.*, **56**, 5744–5756.

Taylor,J.S. *et al.* (2001) Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **356**, 1661–1679.

Vilella,A.J. *et al.* (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.

Wang,J.-C. *et al.* (2012a) idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. *Nucleic Acids Res.*, **40**, W393–W399.

Wang,Y. *et al.* (2012b) PubChem's BioAssay Database. *Nucleic Acids Res.*, **40**, D400–D412.

Waterhouse,R.M. *et al.* (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.*, **41**, D358–D365.

Willett,P. (2011) Similarity searching using 2D structural fingerprints. *Methods Mol. Biol.*, **672**, 133–158.

Young,D.W. *et al.* (2008) Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.*, **4**, 59–68.

Ziegler,S. *et al.* (2013) Target identification for small bioactive molecules: finding the needle in the haystack. *Angew. Chem. Int. Ed. Engl.*, **52**, 2744–2792.

Zon,L.I. and Peterson,R.T. (2005) In vivo drug discovery in the zebrafish. *Nat. Rev. Drug Discov.*, **4**, 35–44.