research articles Journal of proteome research

Use of Ligand Based Models for Protein Domains To Predict Novel Molecular Targets and Applications To Triage Affinity Chromatography Data

Andreas Bender,[*,†,#,+] Dmitri Mikhailov,[†] Meir Glick,[†] Josef Scheiber,[†] John W. Davies,[†]
Stephen Cleaver,[‡] Stephen Marshall,[‡] John A. Tallarico,[§] Edmund Harrington,[§]
Ivan Cornella-Taracido,[§] and Jeremy L. Jenkins[*,†,#]

Center for Proteomic Chemistry, Lead Discovery Informatics, Developmental and Molecular Pathways, and Global Discovery Chemistry, Chemogenetics and Proteomics, Novartis Institutes for BioMedical Research, Inc., 250 Massachusetts Avenue, Cambridge, Massachusetts 02139

Received February 9, 2009

The elucidation of drug targets is important both to optimize desired compound action and to understand drug side-effects. In this study, we created statistical models which link chemical substructures of ligands to protein domains in a probabilistic manner and employ the model to triage the results of affinity chromatography experiments. By annotating targets with their InterPro domains, general rules of ligand−protein domain associations were derived and successfully employed to predict protein targets *outside* the scope of the training set. This methodology was then tested on a proteomics affinity chromatography data set containing 699 compounds. The domain prediction model correctly detected 31.6% of the experimental targets at a specificity of 46.8%. This is striking since 86% of the predicted targets are not part of them (but share InterPro domains with them), and thus could not have been predicted by conventional target prediction approaches. Target predictions improve drastically when significance (FDR) scores for target pulldowns are employed, emphasizing their importance for eliminating artifacts. Filament proteins (such as actin and tubulin) are detected to be 'frequent hitters' in proteomics experiments and their presence in pulldowns is not supported by the target predictions. On the other hand, membrane-bound receptors such as serotonin and dopamine receptors are noticeably absent in the affinity chromatography sets, although their presence would be expected from the predicted targets of compounds. While this can partly be explained by the experimental setup, we suggest the computational methods employed here as a complementary step of identifying protein targets of small molecules. Affinity chromatography results for gefitinib are discussed in detail and while two out of the three kinases with the highest affinity to gefitinib in biochemical assays are detected by affinity chromatography, also the possible involvement of NSF as a target for modulating cancer progressions *via* beta-arrestin can be proposed by this method.

Keywords: chemogenomics • gefitinib • Iressa • resistance • pulldown • selectivity • kinases • chemoproteomics • in silico • cheminformatics • protein domains • target prediction • extrapolation

Introduction

Orphan compounds are compounds that still modulate biological processes *via* an unknown mechanism and for which no macromolecular targets are known. This is the case for molecules in all phases of drug development, even for compounds already on the market. Early medicines, such as Aspirin and its precursors, have been used for thousands of years. While their mode of action has often only been established recently,[1] their relatively safe use has been established empirically over long periods of time. The situation where targets for compounds are unknown (i.e., 'off-targets') also applies to some current drugs on the market such as the Bcr-Abl/c-KIT inhibitor Gleevec.[2,3] Interestingly, promiscuity has in some cases been increasingly viewed in a positive light since modulation of multiple points in signaling pathways may be more efficient than modulating single targets. This multipronged approach may also limit the extend of resistance development.[4] On the other hand, undesired promiscuity always bears the disad-

[*] To whom correspondence should be addressed. E-mail: (A.B.) bendera@lacdr.leidenuniv.nl; (J.L.J.) Jeremy.Jenkins@novartis.com.

[†] Center for Proteomic Chemistry, Lead Discovery Informatics, Novartis Institutes for BioMedical Research, Inc.

[#] These authors contributed equally to the work.

[+] Current address: Leiden/Amsterdam Center for Drug Research, Division of Medicinal Chemistry, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands.

[‡] Developmental and Molecular Pathways, Novartis Institutes for BioMedical Research, Inc.

[§] Global Discovery Chemistry, Chemogenetics and Proteomics, Novartis Institutes for BioMedical Research, Inc.

10.1021/pr900107z CCC: $40.75

Published on Web 03/09/2009

vantage of increased likelihood of interfering with undesired pathways.[5,6]

Computational approaches for target identification have been used increasingly in parallel with experimental approaches.[7] While docking-based approaches to the prediction of targets for small molecules have been published,[8−10] in practice, this approach faces some hurdles. Often the target structure is not known, such as in case of G-protein coupled receptors. Also, the computational complexity of the problem is considerable and scoring functions are still not optimal.[11] On the other hand, ligand-based approaches have been published more recently which do not require the target structure to be known since they are purely based on small-molecule information.[12−14] Ligand-based models are computationally fast, rendering feasible the *in silico* profiling of hundreds of thousands of compounds against thousands of targets. In this case, circular fingerprints were employed which are known to be rich in information content relevant to bioactivity.[15,16] Mathematical modeling was performed employing a Naïve Bayes Classifier, a simple yet surprisingly well-performing model generation method. Applications of ligand-based target prediction models range from the prediction of false positives in reporter gene assays[17] to the prediction of promiscuity profiles employed in preclinical profiling.[5,6]

While on the ligand side, descriptor definitions can be varied to achieve improved prediction of novel scaffolds,[18] the question appears as to how to generalize to unknown *targets*, that is, those targets for which no ligands whatsoever are given in the training set. A step toward answering this question is attempted in the current work (of which a short description has been published previously in the context of a review article[19]) by annotating protein targets with their respective InterPro (IPR) domains, which represent a 'generalized', more abstract representation of an individual protein target. Our hypothesis is that if we can properly associate ligands with protein domains, we can extrapolate to potential targets outside of our training set which also share those domains. In a related manner, Strombergsson et al. used probabilistic linkages of (ligand) chemical features to protein folds to create association rules in order to predict receptor−ligand binding affinities and also to derive the responsible receptor−ligand interacting motifs.[20,21] This work is intriguing in that it uses machine learning to find *a priori* protein features involved in ligand binding. Small-molecule binding site annotation for protein sequences is still not well-documented, but an effort has been made to compile known small-molecule binding domains from X-ray crystal structures into a Small Molecule Interaction Domain (SMID) database.[22] In this previous work, however, information from a small set of 124 crystal structures was employed which covered only 104 different enzymes; these numbers are greatly extended in the current work by covering about 1300 targets with 150 000 related ligands.

We added InterPro (IPR) domain annotations on the targets in a database of ligand−target pairs (WOMBAT[23]), followed by the generation of multicategory Bayesian models on chemical substructures (circular fingerprints) of ligands associated with each IPR domain. The InterPro database integrates multiple databases with various protein signature detecting techniques. The source databases are compilations of sequence motifs and clusters signifying protein domains, folds, functional sites, and families, many of which are nested within a parent−child hierarchy.[24−26] In the present work, we have used all hierarchical levels of protein domain annotations for model building. Once ligand chemical structures for targets are linked to the target InterPro domains, machine learning approaches can be used to discover tens of thousands of chemical substructures statistically associated (or not associated) with thousands of protein domains, which can, as the target prediction models themselves, be used to predict *protein folds* that may be targeted by a test compound. Notably, associations learned by our Bayesian classifier may not reflect an actual small-molecule binding interaction; however, as observed by Strombergsson et al.,[20] even local protein substructures *not* in direct contact with the ligand often encode information crucial for the feasibility of a certain ligand−target interaction which might also include novel allosteric modulation modes. The relationship between InterPro domains within a given protein is not random, but rather it is a pre-established functional linkage that is likely to co-occur in other proteins. Thus, the association of a given target's ligands to all of its domains, while not physically accurate, still translates (as shown in this work) to the correct probabilistic target predictions for test molecules.

One of the applications of such an *in silico* "domain fishing" (domain target prediction) approach will be presented after validation of the model, namely, its application to *in silico* guided chemical proteomics. Chemical proteomics experiments are generally used to evaluate the interaction partners of immobilized small molecules *via* affinity chromatography.[27−31] After attaching a small molecule ligand to a resin and running a cell lysate over the ligand, analysis of the proteins bound to the ligand is in our case performed by combination of liquid chromatography followed by mass spectroscopy (LC/MS). However, oftentimes small-molecule proteomics experiments pull down a very large number of proteins (such as 285 in a recent typical study[30]) which represent both direct interaction partners as well as indirect (protein−protein) interaction partners of the proteins sticking to the ligand. Both the large number of proteins pulled down and the large number of false positive interaction partners often make analysis of the resulting proteomics data difficult. One important source of false positive and false negative readouts is also the expression levels of the respective proteins in the cell; as we will see in the following, highly expressed proteins are ubiquitous in small molecule affinity chromatography experiments, and proteins expressed at a low level can often be missed. Cases such as 2,6,9-trisubstituted purines which achieve relatively clear pull-downs due to scaffold selectivity (for which CDK2 was identified as a target[32]) are the exception rather than the norm. Frequent false-positive readouts consist for example of filament proteins such as actin and tubulin due to their abundance in the cell and simple physicochemical reasons such as their hydrophobicity which causes them to stick to any kind of hydrophobic region.[33,34] Those findings led to the development of tools to estimate which interactions are true positive observations, such as the false discovery rate (FDR) used later, or tools such as the Bayesian Estimator of Protein-Protein Association Probabilities (BEPro^3).[35] Most recently, also the introduction of quantitative chemical proteomics techniques will alleviate many of the false positive proteins pulled down from cell lysate.[3]

In practice, the question remains in proteomics studies as to which targets are the 'true' targets of the ligands and which ones are false-positive artifacts. In the current work, we create and validate an *in silico* domain prediction model, followed by a demonstration of its applicability to predict targets for 699 unique compounds profiled in proteomics experiments, The

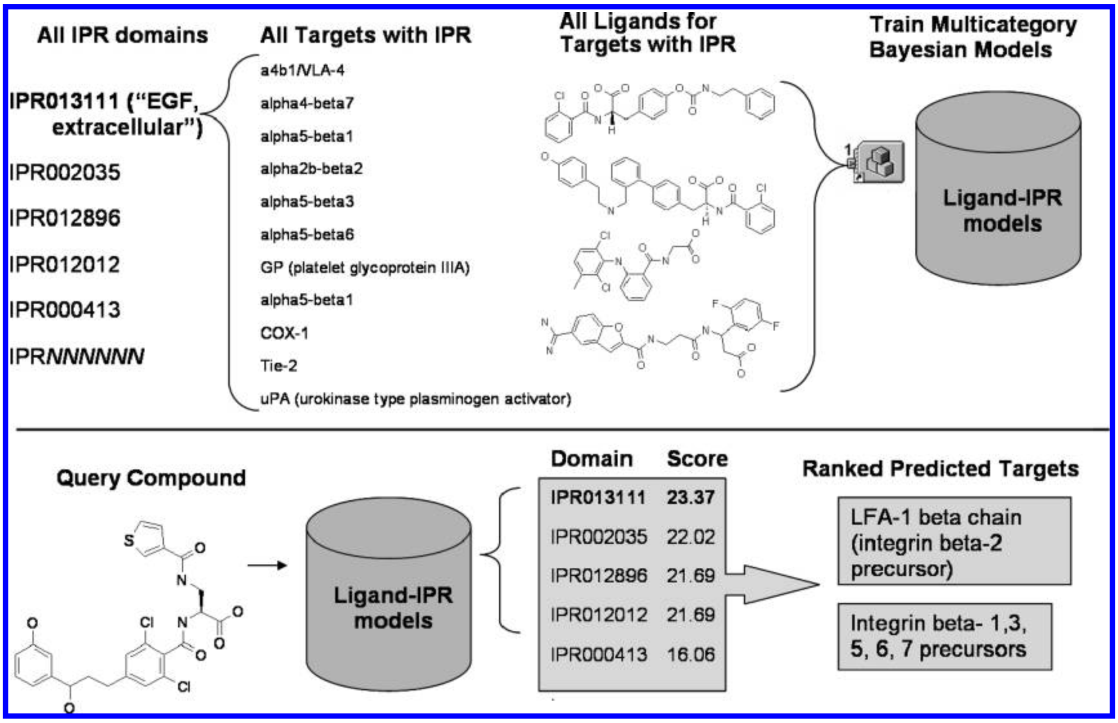


Figure 1. Schematic of the domain prediction workflow. In the training stage (top panel), multicategory Bayesian models are created for each Interpro domain associated with ligand sets for the targets that contain the IPRs. In the application stage (lower panel), the model predicts InterPro domains for each compound—and thus targets by association—the ligands are likely to bind. By exploiting chemogenomics principles, InterPro domains are used as an "aggregate target abstraction", enabling models to extrapolate to targets with no known ligands (but with InterPro domains shared with proteins having known ligands).

results show how computational target predictions can be used as an orthogonal (as well as a complementary) approach to experimental affinity chromatography studies.

Materials and Methods

1. Data Set for Domain Prediction Model. For training the domain prediction model, all compounds with activity values better than 10 μM were extracted from the WOMBAT 2006.01 database,[23] comprising about 1300 targets and 150 000 distinct structures. Compounds were pretreated using StandardizeStereo and StandardizeCharges before calculation of ECFP_4 fingerprints in PipelinePilot 6.0.[36] The Sequence Retrieval System[37] (SRS) was employed to annotate InterPro[24,26] domains onto WOMBAT targets. Merging both identifiers was performed using the SWISSPROT IDs present in WOMBAT or, in cases where SWISPROT IDs were not available, by mapping InterPro domains associated with targets in the same Enzyme Classification (E.C.) code. KEGG and PPI annotations were mapped onto InterPro domains using the BioMining and Text Analytics modules implemented in PipelinePilot.

Multicategory Bayes Models were generated on all InterPro domains associate with compounds of the WOMBAT data set. This model generation step (Figure 1) was analogous to this step in a previous study, and for detailed information, the reader is referred to this publication.[13] A 'Cumulative Domain Score' for all possible proteins is used to rank them as likely targets. The Cumulative Domain Score is the sum of the Bayes scores for every occurrence of the predicted domains found in a given protein. Sorting proteins according to their Cumulative Domain Scores ranks targets according to their likelihood of being interaction partners of a given ligand.

The domain model was generated on ~1300 targets in WOMBAT, covered by ~150 000 distinct molecular entities. Those ~1300 targets corresponded to 2443 unique InterPro Bayesian Models. A total of 1403 of the InterPro domains were identified using the Swiss-Prot accession numbers provided in the WOMBAT database directly. The remaining 1040 InterPro domains were obtained by indirect mapping of target Enzyme Classification (E.C.) numbers to InterPro (IPR) domains. The median number of structures associated with an InterPro domain in the training set is 83 with an average number of 808 (meaning that most of the domains have few associated compounds, while a small number of domains with a large number of domains raises the average number considerably). The minimum number of associated compounds with a given InterPro domain was one and the maximum was 77 981 (for the rhodopsin-like GPCR superfamily). The distribution of domains is shown in Supplementary Figure 1 in Supporting Information.

Given that we are trying to investigate whether extrapolation of target predictions to novel targets (with shared InterPro domains) is possible, we further investigated the overlap between targets given in WOMBAT and those pulled down by proteomics experiments. For compound-target pairings with associated FDR scores of >2 (see below for a description of FDR scores), 564 InterPro domains are found in common between WOMBAT and proteomics data sets based on WOMBAT Swiss-Prot annotations. When we extend WOMBAT target annotations to include EC mappings, the overlap is 800 InterPro domains. This represents 42% of all possible InterPro domains from the affinity chromatography data set with FDR scores >2. Stated another way, based on WOMBAT and InterPro domain associations, we were able to generate models for 42% of the protein domains found in proteins pulled down by proteomics experiments. Note that this number can actually be interpreted as quite a large number: WOMBAT uses only data abstracted from literature, and this literature data set covers nearly half of all protein domains which can be pulled down *via* a large set of affinity chromatography experiments.

2. Benchmark Data Set for Domain Target Prediction. Ten sets of activity classes from the MDDR Database that were previously employed in target prediction studies[13] were used here to compare performance to established tools. The activity classes (which cover several target families) and their respective sizes, as well as their InterPro identifiers, are given in Table 1.

3. Database with Experimental Affinity Chromatography Binding Data. The affinity chromatography data set contained 699 unique compounds with SMILES strings (valid entries) and with 692 compounds with MW > 150 (with a mean molecular weight 385). A total of 207 607 entries were present in this data set which constituted putative ligand−target interactions detected by small molecule affinity chromatography experiments usually performed in HeLa cells. Out of this number of entries, in 200 047 cases, both the International Protein Index (IPI) provided by the proteomics experiments as well as the SMILES representation of the probe were defined. A total of 8417 unique protein targets with valid associated chemical structures for their bait molecules were contained in the data set.

The raw proteomics data needed to be subjected to statistical analysis since some targets are easier to identify in the LC/MS detection than others. For this purpose, the FDR (False Discovery Rate) score was calculated for each data point. The FDR score[38] is the $-\log_{10}$ of the False Discovery Rate (FDR) *q*-value based on the Fisher exact *p*-value. In effect, this means that an FDR score of 2 equals a false discovery rate of 1%. The FDR score was calculated using a custom R script which employed both human and mouse proteins for its determination. In our data set, a total of 28 660 ligand−protein pairings yielded an FDR score ≥2 (which was the reliability cutoff in the current study), covering 3127 unique targets. Out of those targets, 2291 proteins were identified only once. For the interested reader, the idea of promiscuity[6,39] and the discovery of "true targets" of endogenous ligands[10] has been the subject of some discussion in recent literature.

On average, 273 targets were associated with the given compounds with a median of 106 targets and a minimum of 4 targets. The total number of InterPro domains in all proteins was 518 492, of which 2967 were unique. There were 468 unique compounds with at least 1 FDR score larger than 2. For those compounds, on average 367 targets were pulled down per compounds with a median of 122 targets and a minimum of 3 targets, and a total of 74 221 InterPro domains, of which 1899 were unique.

4. Computational Details and Performance Evaluation. The proteomics data set described previously was used to investigate the agreement between experimental and predicted interaction data. The fraction of true targets (relative to the number for which domain models could be constructed based on WOMBAT) was calculated, as well as the number of overpredicted domains.

Table 1. Details of the Benchmark Data Set Used for Domain Prediction[a]

activity class name	MDDR activity index	data set size	InterPro ID
ACE Inhibitors	31410	574	IPR001548 IPR006025
Acetylcholinesterase Inhibitors	09221	830	IPR000997 IPR002018 IPR014788
Angiotensin II AT1 Antagonists	31432	2192	IPR000190 IPR000248 IPR001186 IPR001277 IPR000276
Angiotensin II AT2 Antagonists	31433	59	IPR000147 IPR000248 IPR001186 IPR000276
Cyclooxygenase 1 Inhibitors	78453	101	IPR006210 IPR000742 IPR006209 IPR013032 IPR002007
Cyclooxygenase 2 Inhibitors	78454	1115	IPR006210 IPR000742 IPR006209 IPR013032 IPR002007
HIV 1 Protease Inhibitors	71523	1086	IPR000477 IPR000721 IPR001037 IPR001584 IPR003308 IPR000071 IPR001969 IPR009007 IPR001995 IPR008916 IPR008919 IPR002156 IPR010659 IPR010661 IPR001878
H^+/K^+ ATPase Inhibitors	54112	769	IPR006069 IPR005775 IPR001757 IPR006068 IPR004014 IPR005834 IPR008250 IPR015127
Phosphodiesterase IV Inhibitors	78418	2231	IPR003607 IPR002073
HIV Reverse Transcriptase Inhibitors	71522	945	IPR000477 IPR000721 IPR001037 IPR001584 IPR003308 IPR000071 IPR001969 IPR009007 IPR001995 IPR008916 IPR008919 IPR002156 IPR010659 IPR010661 IPR001878

[a] It is based on a previously visited data set derived from the MDL Drug Data Repository. Shown are activity classes with their respective number of data points as well as the InterPro IDs associated with each target.

Results and Discussion

1. Validation of the Domain Prediction Model on a Benchmark Data Set. The performance of the domain prediction model to infer the correct protein folds of experimentally established targets for a previously visited benchmark data set[13] described in Table 1 is shown in Figure 2. Model performance varies between 21.38% and 97.76% and correctly identified targeted InterPro domains in the first three predictions of the model with an average of 60.68% of the targets captured. This compared to an average of 77% of the targets of the same compound set captured in target prediction models trained on individual targets[13] as in the current study (instead on protein folds, i.e., groups of targets). Overall, while the domain prediction model presented here shows good results, it is generally

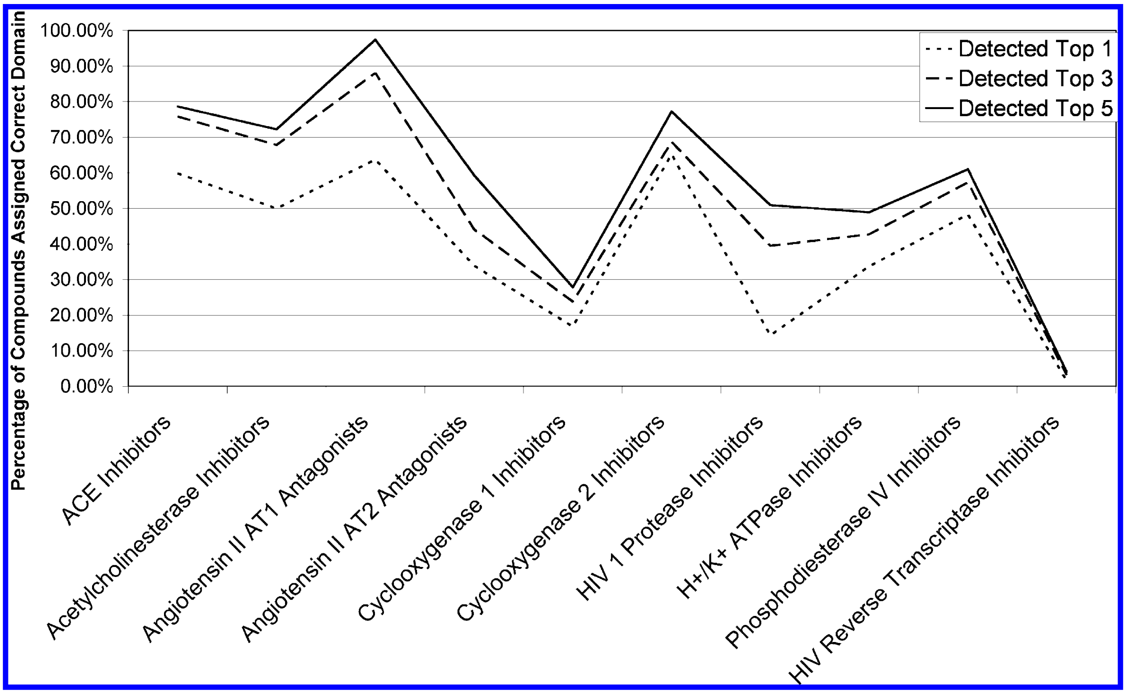


Figure 2. Performance of the domain target prediction tools on previously visited activity classes. While domain prediction shows good results, they are generally inferior to single-target based models.[13] This is understandable due to a variety of reasons, such as increasing fuzziness (but also generalizability!) of the models if multiple activity classes targeting one domain are involved in the model generation step.

Table 2. Prediction Performance of the Domain Prediction Model on the Full Affinity Chromatography Data Set[a]

Bayes cutoff	FDR range	sensitivity	selectivity	% correctly predicted targets not in training set
1	FDR > 2	42.3% (2138/5056)	17.6% (2138/12143)	86% (1836/2138)
	FDR < 2	14% (6728/48581)	6.5% (6728/102814)	88% (5948/6728)
5	FDR > 2	38.3% (1937/5056)	29.0% (1937/6686)	85% (1652/1937)
	FDR < 2	11.0% (5325/48581)	13.7% (5325/38909)	87% (4641/5327)
10	FDR > 2	31.8% (1607/5056)	47.0% (1607/3417)	86% (1379/1607)
	FDR < 2	8.1% (3959/48581)	25.9% (3959/15306)	87% (3461/3971)

[a] It can be seen that selectivity increases while sensitivity decreases with higher Bayes scores. The FDR cutoff has a considerable influence on prediction performance and seems to be crucial to judge results from pulldown experiments.

inferior to target-based models. This is expected because the protein fold-based models are ‘fuzzier’ than their single-target counterparts: ligands binding to multiple targets are used in the training set for each fold and thus a larger area of chemical space will be statistically associated with binding to a certain InterPro domain. Also it should be noted that the comparison is only approximate since the number of target proteins in the single-target model was 964 targets,[13] while the number of InterPro domains predicted by the current model is 2443 domains, translating to several thousand possible protein targets. With this perspective, we feel that identifying 61% of the known targets of a benchmark ligand data set is a satisfying value.

2. Predicting Target Domains Pulled Down by Proteomics Experiments. We next used our domain prediction model to predict protein targets for the affinity chromatography pull down experiments. The purpose of this part of the study was to investigate whether our domain prediction model would indeed be able to predict protein targets that are not present in the training set, and to get statistically significant results on a large data set. This analysis was performed separately for the protein targets which are contained in the training (WOMBAT) data set, as well as those targets which only share InterPro domains with the training data sets. The first scenario confirms whether our model is able to predict the correct protein domains for targets which are in the training data set (using different compounds as probes, with the particular compound-fold pairing excluded in the training stage). The second scenario investigates the ability of the model to extrapolate to novel targets outside the training data set, which share InterPro domains with targets contained in this set. Results of this benchmark are given in Table 2.

For compound−target pairings with FDR scores smaller than 2 (indicating nonspecific interactions between compound and target) only 14% sensitivity and 6.5% specificity were achieved at a Bayes cutoff for positive predictions of 1. This means that only about 1 out of 7 targets detected by affinity chromatography was predicted, and conversely, only about 1 out of 15 predicted targets was experimentally found. Although at first pass these results appear to indicate a bad performing model, this is due to the use of the raw experimental data without employing confidence cutoffs, as we will see in the following. If only ligand−target parings with FDR scores of larger than 2 are considered, as shown in Table 2, both performance measures improved considerably, namely, to a sensitivity of 42.3% (from before 14%) and a specificity of 17.6% (from before 6.5%). Thus, nearly one out of two protein targets detected by affinity chromatography is predicted by our model, with about

4 out of 5 predictions still being 'false positive' predictions. The improvement in agreement between experimentally detected proteins and predicted targets certainly underlines the importance of using significance scores in experimental proteomics settings. Because we are mainly interested in prioritizing proteins pulled down by affinity chromatography, the specificity of our method is less important than the sensitivity in the current application; only proteins in common between the pulldown experiments and the *in silico* predictions are further considered, making the 'false positive' *in silico* predictions less of a consideration in practice.

Next, we investigated the influence of model training parameters to improve the sensitivity and specificity of our model. As seen above, a Bayes cutoff score for positive predictions of 1.0 yields a very high false positive rate, indicated by a specificity of 6.5% and 17.6%, for FDR < 2 and FDR > 2, respectively. Thus, we varied the Bayes score needed for making 'positive' predictions to 5 and 10, respectively. The results for this threshold variation are given also in Table 2 and visualized in Supplementary Figure 2 in Supporting Information. As expected, when considering confident ligand−target pairings with FDR > 2, sensitivity slightly deteriorates with a higher Bayes cutoff, from 42.3% (threshold of 1) over 38.3% (threshold of 5) to 31.6% (threshold of 10). This means that instead of detecting nearly 1 out of 2 proteins by our predictions, we are only able to predict a little less than every third target with a more stringent cutoff. On the other hand, as desired, specificity improves greatly in this case. Starting from a specificity of 17.6% at a threshold of 1, we achieve over 29.0% specificity at a threshold of 5 and 46.8% specificity at a threshold of 10 for positive model predictions. These numbers effectively mean that at the highest cutoff (of a Bayes score of 10) for making positive predictions, nearly every second predicted target is indeed a true positive target, while at the same time, we capture nearly one-third of the experimental targets in the predictions.

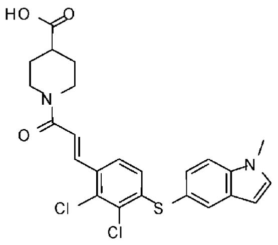
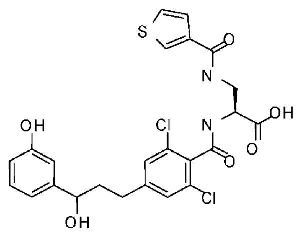
Supplementary Figure 2 in Supporting Information emphasizes the development of sensitivity and specificity upon variation of the Bayes score threshold used for positive predictions. Also shown is the approximate average random expectation value of sensitivity and specificity for a Bayes cutoff of 1. The actual recall rate achieved was 42%, showing about 6-fold enrichment in sensitivity. For specificity, we show approximately 3-fold better performance than random. Overall, the predictions of this model are thus about 18-fold better than random ('enrichment of 18-fold'). Note that better predictions than random are only achieved for the ligand−target pairings with FDR scores above 2, while for those pairings below the threshold, predictions are of the quality of random predictions (Supplementary Figure 2), again confirming the importance of employing false discovery rates as confidence scores in proteomics experiments.

To what extent now does our target InterPro annotations allow extrapolation of targets, that is, the prediction of targets for which no ligands whatsoever were provided in the training set? Strikingly, nearly 87% of the proteins pulled down in the entire data set are not actually present in the WOMBAT database, meaning if we were to perform a straightforward target prediction approach, we could at best be successful for only 13% of the proteomics data set. Clearly, however, because our models were trained on protein domains rather than proteins themselves, we are able to predict much more than 13% of the targets pulled down. Table 2 shows that of the 2138 proteins we correctly predicted at a Bayes score cutoff of 1, 86% of these are not in the source database for model training. Thus, using the chemogenomics principle that similar compounds bind to similar proteins, in this case proteins with the same domains, we are able to predict nascent protein targets for compounds. Furthermore, in many of the cases, the proteins predicted to be targeted by the ligands used have no known small-molecule inhibitors in published or patent literature.

3. Discussion of Individual Compound Results. To better understand how domain prediction works at an individual compound level, we first present a case study previously demonstrated for LFA-1 inhibitors,[19] followed by an example from the present proteomics data set. In Table 3, compounds **19** and **20** (to be consistent with previous work[40]) are known inhibitors of the LFA-1/ICAM protein−protein interaction (PPI). Both chemicals bind to leukocyte function associated molecule 1 (LFA-1) at allosteric sites that do not directly interfere with the LFA-1/ICAM interface. Compound **19** binds to the I-domain in the α-chain of LFA-1, while compound **20** binds at the interface of the β and α chains ('allosteric site'). Extrapolation is required to predict the target of this compound correctly since the training set for the models (WOMBAT[23]) contains inhibitors of the target LFA-1, but this comprises only compounds that bind to the I-domain such as compound **19**, and not allosteric compounds such as compound **20**. Applying the domain prediction model on compound **19**, six IPR domains are predicted as likely targets, where the protein with the highest Cumulative Domain Score is the α chain of LFA-1, as would be expected. For compound **20** on the other hand, for which the seven most likely InterPro domains are also listed in Table 3, the Cumulative Domain Score correctly assigns the most probable target as the Integrin beta-2 precursor. Algorithmically, this extension in predictive ability occurs because other targets in WOMBAT contain some of the domains found in the LFA-1 β chain, and their ligands resemble compound **20**. Both of those examples illustrate that the domain fishing method presented in this work is able to correctly pull out a protein target that is not present in the training set.

An example of analyzing individual compounds from our experimental affinity chromatography dataset is shown in Table 4 for the compound gefitinib. Listed are the proteins binding to a linkered version of gefitinib probe detected with FDR scores >2 that were also correctly predicted by our *in silico* model. The known primary target of gefitinib, EGFR (epidermal growth factor receptor), was detected correctly, along with nine other proteins that overlapped with our *in silico* predictions. Five of the targets are kinases (EGFR, EPHB4, GAK, MAPK9 and RIPK2) while the remaining ones are an ABC efflux transporter (ABCE1), an ATPase (NSF), a plastin (actin-binding protein, PLS3), parts of the proteasome (PSMD3) and a hydrolase (USP24). Out of those proteins which are detected experimentally as well as predicted *in silico*, the family of kinases would be expected to contribute to a large fraction of the detected proteins; however, the variety of kinase 'classes' covered is of interest: tyrosine kinases (EGFR, EPHB4), a CMGC kinase (MAPK9), a tyrosine-kinase like kinase (RIPK2) and a kinase belonging to the 'other' group (GAK), according to the classification employed by Manning et al.[41] Compared to recent work on analyzing kinase inhibitors on a larger scale,[2,42] our affinity experiments agree with the detection of the kinases exhibiting high affinity to gefitinib, although both false-negatives and false-positives occur. Namely, EGFR, the primary target on gefitinib, with a reported K_d value between 1.8 nM[2] and 1 nM[42] was correctly detected *via* affinity chromatography.

Table 3. Target Prediction via Probabilistic Association of Chemical Features with InterPro Domains in Proteins[a]

Compound	Target in Training Set?	Structure	Known Target Binding Site	Domains Predicted and Associated Bayes Score	Predicted Domain Descriptions	Highest Scoring Target Extrapolated	Number Predicted Domains Found in Target	Cumulative Domain Score
19	YES		Alpha subunit of LFA-1 in I-domain	1) IPR000413 (43.99) 2) IPR013513 (43.99) 3) IPR013517 (43.99) 4) IPR013519 (43.99) 5) IPR013649 (43.99) 6) IPR002035 (41.11)	1) Integrins alpha chain 2) Integrin alpha chain, C-terminal cytoplasmic region 3) FG-GAP 4) Integrin alpha beta-propellor 5) Integrin alpha-2 6) von Willebrand factor, type A	Integrin alpha-L precursor (Leukocyte adhesion glycoprotein LFA-1 alpha chain), Swiss-Prot P20701	6	261.06
20	NO		Between beta and alpha subunits of LFA-1 in I-like domain	1) IPR013111 (23.37) 2) IPR002035 (22.02) 3) IPR012896 (21.69) 4) IPR012012 (21.69) 5) IPR003659 (21.69) 6)IPR002369 (21.69) 7) IPR001169 (21.69)	1) EGF, extracellular 2) von Willebrand factor, type A 3) Integrin beta tail 4) Integrin, beta subunit 5) Plexin / semaphorin / integrin 6) Integrin, beta chain N-terminal 7) Integrin beta, C-terminal	Integrin beta-2 precursor (Cell surface adhesion glycoproteins LFA-1/CR3/p150,95 subunit beta) (Complement receptor C3 subunit beta) (CD18 antigen), Swiss-Prot P05107	7	153.84

[a] Compound IDs are in line with the original publication. For compound **19**, the correct target is predicted *via* the prediction of its interaction domains. Ligands of this target have been known before, so in principle also a model based on ligand−target pairings could have been used. For compound **20**, the correct target is predicted *without* having the target in the training set. This extrapolation is achieved by generating models on *protein folds* instead of individual targets.

Table 4. Proteins Detected *via* Affinity Chromatography with Gefitinib Which Were Also Predicted Correctly by Our *in Silico* Domain Fishing tool[a]

interactor protein accession number	name	description
IPI00303207.3	ABCE1	ATP-binding cassette subfamily e member 1
IPI00018274.1	EGFR	splice isoform 1 of epidermal growth factor receptor precursor
IPI00289342.1	EPHB4	ephrin type-B receptor 4 precursor
IPI00298949.1	GAK	cyclin G-associated kinase
IPI00303550.2	MAPK9	splice isoform beta-2 of mitogen-activated protein kinase 9
IPI00006451.5	NSF	vesicle-fusing ATPase
IPI00216694.2	PLS3	plastin 3 variant (fragment)
IPI00011603.2	PSMD3	26S proteasome non-ATPase regulatory subunit 3
IPI00021917.1	RIPK2	splice isoform 1 of receptor-interacting serine/threonine-protein kinase 2
IPI00398505.5	USP24	ubiquitin carboxyl-terminal hydrolase 24

[a] The known primary target of gefitinib, EGFR, was detected correctly, along with other proteins such as the efflux transporter ABCE1. There is evidence that the related transporter ABCG2 is likely responsible for tumor resistance against gefitinib in patient subgroups.

Among the other four kinases we detected, EPHB4, GAK, RIPK2 and MAPK9, two show submicromolar K_d values in biochemical assays, namely, GAK with a $K_d = 13$ nM and RIPK2 with a $K_d = 530$ nM.[42] On the other hand, while measured in biochemical assays, the two additional kinases detected by our experiments were found to be inactive *in vitro*, which are EPHB4 and MAPK9.[42] The affinity constant of EPHA6 was determined to be 590 nM, so detecting EPHB4 might or might not be an artifact. MAPK9 was in the first part of the study found to be a 'frequent hitter' in proteomics experiments so it might also here be a 'false positive', in agreement with the biochemical results as shown above. The other kinases with submicromolar affinity as measured in biochemical assays[42] but not detected *via* affinity chromatography are ERBB4 ($K_d = 410$ nM), LCK (630 nM), MKNK1 (290 nM), SLK (920 nM) and CSNK1E (430 nM). It is worth noting that in subsequent more robust, quantitative affinity chromatography experiments using free gefitinib to compete with linkered gefitinib, only kinase targets consistently competed off with high strength and specificity scores (data not shown). One explanation for this finding may be that the proteins in Table 4 are only weak interactors with gefitinib. Nevertheless, we look more closely at two proteins pulled down which may warrant further experimental validation.

With respect to the nonkinases detected, the ATPase NSF (*N*-ethylmaleimide-sensitive factor) is not an unreasonable potential target due to the ATP binding site and the fact that gefitinib is targeted to kinase ATP binding sites. In addition to giving a reasonable target prediction, a hypothesis for the mode of action of gefitinib could be posed. It is known that NSF binds beta-arrestin,[43] and, in turn, beta-arrestin is known to be involved in colorectal cancer.[44−46] The hypothesis that this might be an unknown mode of action of gefitinib would require further investigation and validation.

Also of interest is the detection of ABCE1 with the gefitinib pulldown. ABCE1 also contains ATPase domains, and is an evolutionarily highly conserved member of the ABC cassette transporter family. Its critical functions include translation initiation, ribosome biogenesis, inhibition of RNase L, and human immunodeficiency virus capsid assembly. It has been shown previously that a related transporter protein, ABCG2, might be involved in resistance development against gefitinib due to efflux transport,[47,48] a finding that may have been predicted by extrapolation from the affinity chromatography experiment. Erlotinib on the other hand was found to inhibit the function of ABCB1 as well as ABCG2, in turn leading to higher intracellular levels of the drug and likely increasing compound efficacy.[48] However, while computational methods are able to generate hypotheses such as the ones above, it should be kept in mind that experimental validation would be necessary for their verification or falsification.

4. Comparison of Experimental Proteomics Data with the Predicted Compound Targets. To investigate the performance and a possible method-inherent target bias of the affinity chromatography experiments, we analyzed the distribution of targets pulled down by our compounds and compared them to the predicted targets. It must be emphasized that the types of proteins pulled down are, of course, dependent on the linkered chemical structures; thus, any global patterns could be influenced by the nature of the drug discovery projects that submitted compounds for affinity chromatography experiments. Also important to keep in mind is that the database we used for target predictions, WOMBAT, contains roughly 50% GPCR ligands; thus, this target class as a chemically well-annotated target class has a certain advantage over other protein families. Nevertheless, some important trends can be discerned.

First, we investigated whether targets from certain cellular compartments are favored by affinity chromatography experiments. For this purpose, we used all experimental targets with FDR scores larger than 2 which were also contained in the WOMBAT database. Those *experimental* targets were annotated with the GO Cellular Component annotation as implemented in the BioMining Module of PipelinePilot 6.0. The same annotation was performed for the top 10 targets as *predicted* by our domain prediction tool. The top 10 predicted target domains had usually Bayes scores greater (in most cases much greater) than 10, meaning that in our experience they should possess a higher likelihood of indeed interacting with the ligand used. Figure 3 and Supplementary Table 1 in Supporting Information show the distribution of the GO "Cellular Component" annotations for the targets pulled down experimentally (inner circle) and those predicted for the compounds (outer circle). The most significant positive deviations between predicted and expected cellular components are observed for the intracellular components. It was found that intracellular proteins are detected in affinity chromatography experiments nearly twice as often as theoretically predicted (16.9% of the targets belonged to this class, instead of a predicted 9.4%). On the other hand, membrane parts are present in affinity chromatography data sets less than half the expected time - only 4.1% of the targets belong to this class, instead of 9.6% of the predicted targets (47.5%).

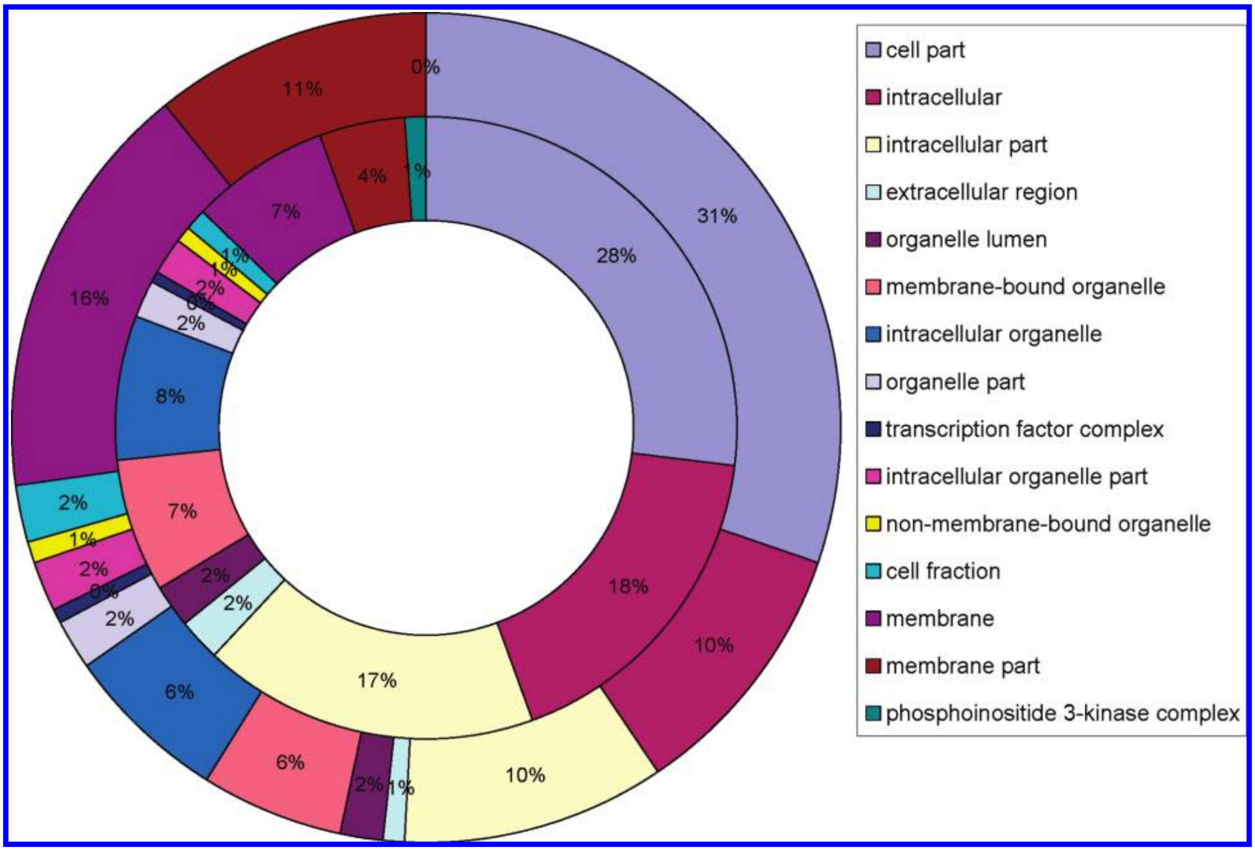


Figure 3. Distribution of the GO "Cellular Component" annotations for the targets pulled down experimentally (inner circle) and those predicted for the compounds (outer circle; using only FDR scores larger 2 and targets present in WOMBAT). Intracellular parts and organelles are enriched in the experimental pulldown while membrane parts are unenriched in the experimental results.

We then extended this type of analysis to the individual protein targets encountered in small molecule affinity chromatography experiments. Supplementary Table 2 in Supporting Information lists the individual targets most frequently detected in affinity chromatography experiments, compared to their predicted frequencies. Three different criteria were employed to generate the list of protein targets:

(a) The raw affinity chromatography data set was used (containing all targets, without employing any FDR cutoff to assign a confidence threshold to the predicted proteins),

(b) The data set was limited to targets with FDR > 2 (i.e., those we deemed significant) and

(c) Only those targets with FDR > 2 which were also present in WOMBAT were analyzed.

Without filtering the list of targets (item (a) in the above list) we found that the major histocompatibility complex, class I, A (HLA-A), tublin alpha 4 (TUBA4) and tropomyosin 4 (TPM4) are detected most frequently in affinity chromatography experiments at frequencies of 0.64%, 0.44% and 0.40% of all proteins detected, respectively. Only the histocompatibility complex was predicted as a target in 0.07% of all predictions, while all other targets in this part of the table are never predicted to the targeted by our *in silico* analysis. The frequent presence of tubulin in this kind of experiments is consistent with previous observations and it can be explained by the large absolute amount of tubulin present in the cell as well to its largely lipophilic character which leads to unspecific binding.

If filtering for FDR scores is employed (item (b) in the above list), all three targets are removed from the list, indicating their nature as "frequent hitters" with no statistical significance. In this case, immunoglobulin heavy constant alpha 1 (IGHA1), major histocompatibility complex, class I, B (HLA-B) and ferrochelatase (FECH) are the most frequently detected proteins with frequencies of 0.43%, 0.40% and 0.40% of all detections, respectively, while never being predicted targets of the ligands employed in this study. The presence of a relatively large number of proteins involved in immune response (IGHA1, HLA-B) is striking in this part of the analysis. In a similar vein, ALA-B is on the exterior part of the membrane, which is not the most prominent cellular location retrieved in our affinity chromatography experiments.

If additional filtering to those targets present in WOMBAT is employed (item (C) in the above list), the bias in proteins detected *via* affinity chromatography versus the predicted targets is shifted largely toward kinases. For our data set of 699 compounds, the proteins most often detected in this case are mitogen-activated protein kinase 9 (MAPK9; also known as JNK2, 4.13% of all detected proteins), cholesterol acyltransferase (SOAT1, 3.64%) and mitogen-activated protein kinase 8 (MAPK8; also known as JNK1, 3.40%) which are only predicted at frequencies of 0.16%, 0.00% and 0.55%, respectively.

In all three different data sets considered (a–c), it can be seen that proteomics experiments possess a considerable bias toward both intracellular components of the cell, and also toward particular target types. This bias needs to be kept in mind in order to eliminate as many false-positive targets from affinity chromatography results as possible.

Conversely, we investigated which targets were less frequently detected in affinity chromatography experiments than they were predicted. Results of this analysis are given in Supplementary Table 3 in Supporting Information for a data set with FDR > 2 where targets are present in WOMBAT. Most of the targets less frequently detected than expected from target predictions are membrane bound proteins such as the families of 5HT receptors (here HTR1B, HTR2A, HTR3A, HTR2C, HTR1A), dopamine receptors (DRD5, DRD4, DRD2) and adrenoreceptors (ADRB1, ADRA2A, ADRA1A). A large fraction of the other targets less easy to detect in this type of experiment are kinases, such as conserved helix−loop−helix ubiquitous kinase (CHUK; also IKK1 or IKKA), inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta (IKBKB; also IKK2 or IKKB), cyclin-dependent kinase 2 (CDK2), mitogen activated protein kinase kinase kinase 1 (MAP3K1), glycogen-synthase kinase 3D (GSK3B), mitogen activated protein kinase 1 (MAPK1) and 3-phosphoinositide dependent protein kinase-1 (PDPK1; also PDK1). Multiple reasons can contribute here, and among the most important are the cell type used for generating cell lysate and the variable endogenous activation states of kinases. The question of which proteins are expressed while performing proteomics experiments would need to be addressed separately; however, it is a major determinant of the experimental outcome. Also, in practice, subcellular fractionation would be employed in order to enrich membrane proteins in protein pulldown experiments; however, as shown above, if no fractionation is employed, experimentally detected ligand−protein interaction pairs show a severely distorted picture of the true binding capabilities of a compound, an effect which is quantified in this work.

In general, the presence of a large number of membrane proteins in this list of targets which are much less frequently detected than expected, many of them G-protein coupled receptors, confirms the 'pro-intracellular, antimembrane bias' of chemical proteomics experiments under the conditions used in the present study. However, it should be kept in mind that the results of a particular SMAC experiment are hugely dependent on chromatographic conditions (such as the columns used, temperature, buffers, etc.), steric hindrance due to ligand immobilization; in case of membrane proteins also nonspecific interactions with membrane components, the presence of detergents, the necessity of cofactors such as GDP and GTP for G-protein coupled receptors and the presence of multiple protein complexes as well as multiple states of the protein itself (just to name a few of the variables). On the other hand, the bias in favor of ubiquitous, hydrophobic proteins such as tubulin and actin and against membrane-bound receptors is profound as we have seen in the first part of this analysis. Individual proteins of the expressed proteome are present in various concentrations in cells and their concentrations are clearly dependent on the cell line and the preparation of the cell lysate. It is clear that proteins expressed at high concentrations will be easier to capture by SMAC than lower expressed proteins. Similarly, highly abundant cytosolic proteins are easier to capture than less abundant, highly localized proteins. Detecting such proteins may require, for example, special centrifugation techniques to isolate enough material to capture by SMAC, as may be the case for proteins specific to organelles, nucleus, or membranes.

Conclusions

In the present study, statistical models were generated which link small molecule substructures to protein domains in a probabilistic manner. By annotating targets with their InterPro domains, we can derive general rules of ligand−protein domain associations that were successfully employed to predict protein targets both inside and outside the model training set. For a proteomics affinity chromatography data set, 31.6% sensitivity was achieved at a specificity of 46.8%, which is a considerable achievement for two reasons. First, 86% of the predicted targets lie outside the scope of the training set and would never be accessible for conventional target prediction models. Second, the domain prediction model employed contains information about a total of 2443 InterPro domains, making the number of targets correctly predicted by chance very small, and rendering the number above an 18-fold enrichment over random selection of targets. The fact that the domain prediction models were not as accurate on individual targets in our benchmark data set as our regular target prediction models suggests the optimal way to use them is when we are dealing with data sets that may lie outside known ligand−target space, such as in triaging affinity chromatography experiments or de-orphanizing novel ligands as well as receptors.

Target predictions improved significantly when significance (FDR) scores for target pulldowns are used, outlining their importance. In fact, when omitting FDR scores, no correlation between predicted targets and experimentally detected proteins was observed. We have shown that filament proteins (such as actin and tubulin) were 'frequent hitters' in proteomics experiments and their presence in pulldowns is not expected at all from the target predictions performed. On the other hand, membrane-bound receptors are mostly absent in the affinity chromatography sets, although their presence would be expected from the predicted targets of compounds. While partly also due to the particular experimental conditions used, we conclude that affinity chromatography data sets are by nature biased. In addition to improvements in experimental protocols (e.g., quantitative formats, cell fractionations), *in silico* domain prediction is a valuable orthogonal method for prioritizing proteins identified in affinity chromatography experiments.

Acknowledgment. A.B. and J.S. thank the Education Office of NIBR for postdoctoral fellowships, and J. Klekota for supporting data. The anonymous referees are thanked for constructive comments on the original version of the manuscript.

Supporting Information Available: Supplementary Figure 1, distribution of InterPro domains in the data set; Supplementary Figure 2, dependence of the domain prediction performance on the Bayes score threshold used to indicate "positive" predictions; Supplementary Table 1, distribution of the GO "Cellular Component" annotations for the targets pulled down experimentally and those predicted for the compounds, using only FDR scores larger 2 and targets present in WOMBAT; Supplementary Table 2, analysis of affinity chromatography data set; Supplementary Table 3, further analysis of the affinity chromatography data set. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Simmons, D. L.; Botting, R. M.; Hla, T. Cyclooxygenase isozymes: The biology of prostaglandin synthesis and inhibition. *Pharmacol. Rev.* **2004**, *56*, 387–437.

(2) Fabian, M. A.; Biggs, W. H.; Treiber, D. K.; Atteridge, C. E. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329–336.

(3) Bantscheff, M.; Eberhard, D.; Abraham, Y.; Bastuck, S. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat. Biotechnol.* **2007**, *25*, 1035–1044.

(4) Hampton, T. "Promiscuous" anticancer drugs that hit multiple targets may thwart resistance. *JAMA* **2004**, *292*, 419–422.

(5) Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* **2007**, *2*, 861–873.

(6) Azzaoui, K.; Hamon, J.; Faller, B.; Whitebread, S. Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem* **2007**, *2*, 874–880.

(7) Jenkins, J. L.; Bender, A.; Davies, J. W. In silico target fishing: Predicting biological targets from chemical structure. *Drug Discovery Today: Technol.* **2007**, *3*, 413–421.

(8) Chen, Y. Z.; Zhi, D. G. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* **2001**, *43*, 217–226.

(9) Rockey, W. M.; Elcock, A. H. Rapid computational identification of the targets of protein kinase inhibitors. *J. Med. Chem.* **2005**, *48*, 4138–4152.

(10) Paul, N.; Kellenberger, E.; Bret, G.; Muller, P.; Rognan, D. Recovering the true targets of specific ligands by virtual screening of the Protein Data Bank. *Proteins* **2004**, *54*, 671–680.

(11) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(12) Cleves, A. E.; Jain, A. N. Robust ligand-based modeling of the biological targets of known drugs. *J. Med. Chem.* **2006**, *49*, 2921–2938.

(13) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124–1133.

(14) Faulon, J. L.; Misra, M.; Martin, S.; Sale, K.; Sapra, R. Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics* **2008**, *24*, 225–233.

(15) Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L. Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9*, 199–204.

(16) Faulon, J. L.; Visco, D. P., Jr.; Pophale, R. S. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707–720.

(17) Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J. Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data. *J. Chem Inf. Model.* **2007**, *47*, 1319–1327.

(18) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49*, 6802–6810.

(19) Bender, A.; Young, D. W.; Jenkins, J. L.; Serrano, M. Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. *Comb. Chem. High Throughput Screening* **2007**, *10*, 719–731.

(20) Strombergsson, H.; Kryshtafovych, A.; Prusis, P.; Fidelis, K. Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures. *Proteins* **2006**, *65*, 568–579.

(21) Strombergsson, H.; Prusis, P.; Midelfart, H.; Lapinsh, M. Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions. *Proteins* **2006**, *63*, 24–34.

(22) Snyder, K. A.; Feldman, H. J.; Dumontier, M.; Salama, J. J.; Hogue, C. W. V. Domain-based small molecule binding site annotation. *BMC Bioinf.* **2006**, *7*, 152.

(23) WOrld of Molecular BioAcTivity (WOMBAT), available from Sunset Molecular Discovery LLC, http://www.sunsetmolecular.com/.

(24) Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **2001**, *29*, 37–40.

(25) Mulder, N. J.; Apweiler, R.; Attwood, T. K.; Bairoch, A. InterPro, progress and status in 2005. *Nucleic Acids Res.* **2005**, *33*, D201–205.

(26) Mulder, N. J.; Apweiler, R.; Attwood, T. K.; Bairoch, A. New developments in the InterPro database. *Nucleic Acids Res.* **2007**, *35*, D224–228.

(27) Mano, N.; Sato, K.; Goto, J. Specific affinity extraction method for small molecule-binding proteins. *Anal. Chem.* **2006**, *78*, 4668–4675.

(28) Szardenings, K.; Li, B.; Ma, L.; Wu, M. Fishing for targets: novel approaches using small molecule baits. *Drug Discovery Today: Technol.* **2004**, *1*, 9–15.

(29) Brown, D.; Superti-Furga, G. Rediscovering the sweet spot in drug discovery. *Drug Discovery Today* **2003**, *8*, 1067–1077.

(30) Oda, Y.; Owa, T.; Sato, T.; Boucher, B. Quantitative chemical proteomics for identifying candidate drug targets. *Anal. Chem.* **2003**, *75*, 2159–2165.

(31) Kuster, B.; Schirle, M.; Mallick, P.; Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 577–583.

(32) Knockaert, M.; Gray, N.; Damiens, E.; Chang, Y. T. Intracellular targets of cyclin-dependent kinase inhibitors: identification by affinity chromatography using immobilised inhibitors. *Chem. Biol.* **2000**, *7*, 411–422.

(33) Katayama, H.; Oda, Y. Chemical proteomics for drug discovery based on compound-immobilized affinity chromatography. *J. Chromatogr., B* **2007**, *855*, 21–27.

(34) Shiyama, T.; Furuya, M.; Yamazaki, A.; Terada, T.; Tanaka, A. Design and synthesis of novel hydrophilic spacers for the reduction of nonspecific binding proteins on affinity resins. *Bioorg. Med. Chem.* **2004**, *12*, 2831–2841.

(35) Gilmore, J. M.; Auberry, D. L.; Sharp, J. L.; White, A. M. A Bayesian estimator of protein-protein association probabilities. *Bioinformatics* **2008**, *24*, 1554–1555.

(36) PipelinePilot 5.1, available from Scitegic. http://www.scitegic.com/.

(37) SRS (Sequence Retrieval System), http://srs6.ebi.ac.uk.

(38) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc., Ser. B* **1995**, *57*, 289–300.

(39) Macchiarulo, A.; Nobeli, I.; Thornton, J. M. Ligand selectivity and competition between enzymes in silico. *Nat. Biotechnol.* **2004**, *22*, 1039–1045.

(40) Arkin, M. R.; Wells, J. A. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat. Rev. Drug Discovery* **2004**, *3*, 301–317.

(41) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912–1934.

(42) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P. A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2008**, *26*, 127–132.

(43) McDonald, P. H.; Cote, N. L.; Lin, F. T.; Premont, R. T. Identification of NSF as a beta-arrestin1-binding protein. Implications for beta2-adrenergic receptor regulation. *J. Biol. Chem.* **1999**, *274*, 10677–10680.

(44) Buchanan, F. G.; DuBois, R. N. Emerging roles of beta-arrestins. *Cell Cycle* **2006**, *5*, 2060–2063.

(45) Sun, Y.; Cheng, Z.; Ma, L.; Pei, G. Beta-arrestin2 is critically involved in CXCR4-mediated chemotaxis, and this is mediated by its enhancement of p38 MAPK activation. *J. Biol. Chem.* **2002**, *277*, 49212–49219.

(46) Buchanan, F. G.; Gorden, D. L.; Matta, P.; Shi, Q. Role of beta-arrestin 1 in the metastatic progression of colorectal cancer. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 1492–1497.

(47) Elkind, N. B.; Szentpetery, Z.; Apati, A.; Ozvegy-Laczka, C. Multidrug transporter ABCG2 prevents tumor cell death induced by the epidermal growth factor receptor inhibitor Iressa (ZD1839, Gefitinib). *Cancer Res.* **2005**, *65*, 1770–1777.

(48) Li, J.; Cusatis, G.; Brahmer, J.; Sparreboom, A. Association of variant ABCG2 and the pharmacokinetics of epidermal growth factor receptor tyrosine kinase inhibitors in cancer patients. *Cancer Biol. Ther.* **2007**, *6*, 432–438.

PR900107Z